Matrix-Calibration-Based Cascaded Channel Estimation for Reconfigurable Intelligent Surface Assisted Multiuser MIMO

Hang Liu[ID], *Graduate Student Member, IEEE*, Xiaojun Yuan[ID], *Senior Member, IEEE*, and Ying-Jun Angela Zhang, *Fellow, IEEE*

Abstract—Reconfigurable intelligent surface (RIS) is envisioned to be an essential component of the paradigm for beyond 5G networks as it can potentially provide similar or higher array gains with much lower hardware cost and energy consumption compared with the massive multiple-input multiple-output (MIMO) technology. In this paper, we focus on one of the fundamental challenges, namely the channel acquisition, in a RIS-assisted multiuser MIMO system. The state-of-the-art channel acquisition approach in such a system with fully passive RIS elements estimates the cascaded transmitter-to-RIS and RIS-to-receiver channels by adopting excessively long training sequences. To estimate the cascaded channels with an affordable training overhead, we formulate the channel estimation problem in the RIS-assisted multiuser MIMO system as a matrix-calibration based matrix factorization task. By exploiting the information on the slow-varying channel components and the hidden channel sparsity, we propose a novel message-passing based algorithm to factorize the cascaded channels. Furthermore, we present an analytical framework to characterize the theoretical performance bound of the proposed estimator in the large-system limit. Finally, we conduct simulations to verify the high accuracy and efficiency of the proposed algorithm.

Index Terms—Channel estimation, reconfigurable intelligent surface, multiuser MIMO, matrix factorization, matrix calibration, message passing, replica method.

I. INTRODUCTION

AS ONE of the key techniques of the fifth-generation (5G) wireless networks, massive multiple-input multiple-output (MIMO) can greatly enhance system throughput and enlarge cell coverage. However, expensive hardware cost and high power consumption are two unresolved challenges in current massive MIMO systems [2]. Substitute technologies have been explored to achieve more sustainable and more reliable communications for the next generation mobile networks. Among these new technologies, reconfigurable intelligent surface (RIS, a.k.a. passive holographic MIMO surface [3]) assisted MIMO [4] is deemed very promising to realize similar or even higher array gains with significant cost reduction compared with massive MIMO. Comprising a large number of reconfigurable reflecting elements that can induce adjustable and independent phase shifts on the incident signals, a RIS is able to constructively combine the reflected signals to achieve a high level of energy focusing at the receiver side. Due to the passive and low-cost nature of the reflecting elements, the RIS requires very low energy consumption and is easy to be integrated into the existing wireless systems [5].

In recent years, the design of RISs in assisting wireless communications has attracted extensive attention. To name a few, the authors in [5]–[7] employ RISs to control the propagation environment and improve the coverage for indoor communications. In [8]–[12], the authors propose various methods to configure the RIS phase shifts in outdoor communications, in order to optimize different communication utilities. It is worth noting that accurate channel state information (CSI) is critical in optimizing the RIS parameters. Unfortunately, all the aforementioned work assumes perfect CSI without considering its acquisition difficulty. In fact, channel estimation in a RIS-assisted wireless system is much more challenging than in a conventional system. This is because the passive RIS elements are incapable of sensing and estimating channel information. Therefore, we shall rely on the receiver to estimate both the transmitter-to-RIS and RIS-to-receiver channels by observing only a noisy cascade of the two channels.

To address the channel estimation challenge in RIS-assisted communication systems, some pioneering work has recently emerged. For example, Ref. [13] assumes that the RIS elements are fully active and are connected to a signal processing unit to perform channel estimation. Similarly, Ref. [14] requires a portion of the RIS elements to be active so that the channels of the passive elements can be inferred via a compressed-sensing based approach. Compared with active RIS elements, purely passive RIS elements are undoubtedly

Hang Liu and Ying-Jun Angela Zhang are with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: lh117@ie.cuhk.edu.hk; yjzhang@ie.cuhk.edu.hk).

Xiaojun Yuan is with the Center for Intelligent Networking and Communications, University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: xjyuan@uestc.edu.cn).

more appealing due to their extremely low hardware and deployment costs. Ref. [15], [16] show that channel estimation in passive-RIS-assisted systems can be converted into a sequence of conventional MIMO channel estimation problems by turning on one RIS element at a time. However, the training overhead of this method is proportional to the size of the RIS and may be prohibitively large as the RIS typically comprises a large number of elements. As such, the following question arises naturally: *How to estimate the two cascaded channels with purely passive reflecting elements and an affordable training overhead?* Preliminary answers to this question appear in the recent papers [17]–[20]. In [17], the authors develop a cascaded channel estimation algorithm for a RIS-assisted *single-user* MIMO system. Specifically, [17] formulates the cascaded channel estimation problem as a combination of sparse matrix factorization and low-rank matrix completion by leveraging the programmable property of the RIS and the low-rankness of the propagation channel. In [18], the authors sequentially estimate the cascaded channels for users. Since users share the same RIS-to-receiver channel, the required training overhead is largely reduced by exploiting such channel correlations among users. In [19], the authors exploit the sparsity of the transmitter-to-RIS-to-receiver channels and estimate the cascaded channels based on compressed sensing. In [20], the authors employ parallel factor decomposition to alternatively estimate the cascaded channels.

In this paper, we consider the cascaded channel estimation problem for an uplink RIS-assisted *multiuser* MIMO system, where a fully passive RIS is used to assist the communication. In practice, the RIS can be coated onto a wall, a ceiling, or a furniture in an indoor environment; or mounted on a building facade, an advertising panel, a traffic sign, or a highway fence in an outdoor environment. In a typical application scenario, both the base station (BS) and the RIS rarely move after deployment. As a result, the channel between the BS and the RIS can be modelled as a quasi-static end-to-end MIMO channel, in which most of the channel components evolve much more slowly compared with conventional mobile communication channels. Meanwhile, a small portion of the channel components of the BS-to-RIS channel may experience sudden changes. For example, an opening/closing of a door in an indoor scenario or a moving car in an outdoor scenario may change the scattering geometry. By modeling both the fast-varying and slow-varying channel components, we formulate the CSI acquisition problem as a *matrix-calibration-based matrix factorization* task. Then, we propose a novel message-passing based algorithm to effectively estimate the two cascaded channels. Besides, we present an analytical framework to analyze the performance of the considered system. The main contributions of this paper are summarized as follows.

- We characterize the slow-varying and fast-varying channel components of the RIS-to-BS channel by the Rician fading model. Unlike [17] that assumes the low-rank channel matrix, we exploit the information on the slow-varying channel components for channel estimation. Specifically, we assume that the slow-varying channel components can be estimated by long-term channel

averaging prior to the RIS channel estimation procedure. Based on this assumption, we formulate the cascaded channel estimation problem as a joint task of the RIS-to-BS channel matrix calibration and the RIS-to-user channel matrix estimation.

- We develop the posterior mean estimators under the Bayesian inference framework to infer the two cascaded channels and adopt the sum-product message passing algorithm to approximately compute the estimators. We further show that direct computation of the messages leads to prohibitively high complexity. To tackle this challenge, we introduce additional approximations to the messages based on the approximate message passing (AMP) framework [21]–[23]. The proposed algorithm only needs to update the means and variances of the messages and hence avoids high-dimensional integrations in the canonical message passing algorithm. Furthermore, different from [19] that exploits the sparsity of the joint transmitter-to-RIS-to-receiver channels, our work exploits the sparsity inherent in the separated transmitter-to-RIS and the RIS-to-receiver channels to enhance the estimation accuracy.

- Based on the replica method from statistical physics [24], we analyze the asymptotic performance of the posterior mean estimators in the large-system limit. We show that, with perfect knowledge of the channel prior distributions and under some appropriately chosen angle sampling bases, the mean square errors (MSEs) of the posterior mean estimators can be determined by the fixed point of a set of scalar equations. We further show by numerical simulations that the proposed approximate channel estimators can closely approach the replica method bound.

As a final remark, we note that the approximations in our proposed algorithm share the same spirit with AMP. However, in this work, the cascaded channel coefficient matrices cannot be directly modelled with independent and identically distributed (i.i.d.) elements, and shall be described by sparse representations under the domain transformed by angular sampling bases. In contrast, AMP [21], [22] requires only one matrix to be unknown, and its extension to bilinear matrix factorization in [23] requires the two unknown matrices to be multiplied without domain transformations. Therefore, the existing work on AMP [21]–[23] cannot be directly applied here.

The remainder of this paper is organized as follows. In Section II, we describe the RIS-assisted multiuser MIMO system model and the RIS channel estimation protocol. In Section III, we discuss the channel model and formulate the cascaded channel estimation problem. In Section IV, we develop the channel estimation algorithm and introduce additional approximations to reduce the computational complexity. In Section V, we perform asymptotic analysis based on the replica method. Section VI presents extensive numerical results of the proposed method, and the paper concludes in Section VII.

Notation: Throughout, we use $\mathbb{R}$ and $\mathbb{C}$ to denote the real and complex number sets, respectively. Regular letters, bold small letters, and bold capital letters are used to denote scalars, vectors, and matrices, respectively. We use $j \triangleq \sqrt{-1}$
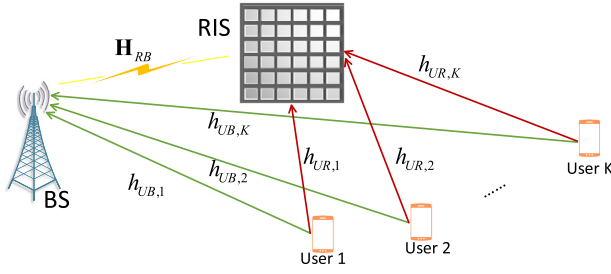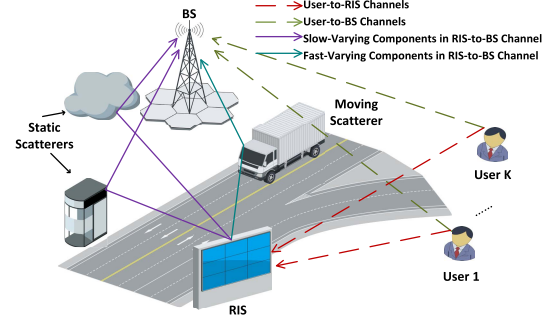
Fig. 1.   A RIS-assisted multiuser system.

Fig. 2.   An example of the outdoor propagation geometry. The propagation path through a moving truck belongs to the fast-varying channel components between the BS and the RIS, while the paths from static scatterers belong to the slow-varying channel components.

to denote the imaginary unit. We use $(\cdot)^\star$, $(\cdot)^T$, and $(\cdot)^H$ to denote the conjugate, the transpose, and the conjugate transpose, respectively. We use x_{ij} to denote the (i, j)-th entry of $\mathbf{X}$. We use $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{CN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote that $\mathbf{x}$ follows the real normal and the circularly-symmetric normal distributions with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, respectively. We use $\mathrm{tr}(\mathbf{X})$ to denote the trace of $\mathbf{X}$, $\mathbf{I}$ to denote the identity matrix with an appropriate size, $\mathrm{diag}(\mathbf{x})$ to denote a diagonal matrix with the diagonal entries specified by $\mathbf{x}$, and $\mathbf{1}$ to denote the all-one vector with an appropriate size. We use $\|\cdot\|_p$ to denote the ℓ_p norm, $\|\cdot\|_F$ to denote the Frobenius norm, $\delta(\cdot)$ to denote the Dirac delta function, $\propto$ to denote equality up to a constant multiplicative factor, and $\mathbb{E}[\cdot]$ to denote the expectation operator.

II. SYSTEM MODEL

A. RIS-Assisted Multiuser MIMO

Consider a single-cell RIS-assisted multiuser MIMO system depicted in Fig. 1. Assume that K single-antenna users simultaneously communicate with an M-antenna BS. A RIS comprising L phase-shift elements is deployed to assist the communication between the users and the BS. A uniform linear array (ULA) is adopted at the BS, and the passive reflecting elements in the RIS are arranged in the form of an $L_1 \times L_2$ uniform rectangular array (URA) with $L_1 L_2 = L$.

We assume a quasi-static flat fading channel model, where the channel coefficients remain invariant within the coherence time. The coefficient vectors/matrix of the k-th-user-to-BS channel, the k-th-user-to-RIS channel, and the RIS-to-BS channel are denoted by $\mathbf{h}_{UB,k} \in \mathbb{C}^{M \times 1}$, $\mathbf{h}_{UR,k} \in \mathbb{C}^{L \times 1}$, and $\mathbf{H}_{RB} \in \mathbb{C}^{M \times L}$, respectively. Moreover, we assume that the RIS elements induce independent phase shifts on the incident signals. We denote the RIS phase-shift vector at time t by $\boldsymbol{\psi}(t) \triangleq [\varpi_1(t)e^{j\psi_1(t)}, \varpi_2(t)e^{j\psi_2(t)}, \cdots, \varpi_L(t)e^{j\psi_L(t)}]^T$, where $\varpi_l(t) \in \{0, 1\}$ represents the on/off state of the l-th RIS element, and $\psi_l(t) \in [0, 2\pi)$ represents the phase shift of the l-th RIS element.

B. RIS Channel Estimation Protocol

For the studied system, we aim to estimate the CSI of the channels $\mathbf{H}_{RB}$, $\{\mathbf{h}_{UR,k}\}$, and $\{\mathbf{h}_{UB,k}\}$. We assume that all the direct channels $\{\mathbf{h}_{UB,k}\}$ are accurately estimated at first and focus our discussions on estimating the RIS channels

$\mathbf{H}_{RB}$ and $\{\mathbf{h}_{UR,k}\}$.[1] To facilitate RIS channel estimation, the users simultaneously transmit training sequences with length T to the BS. Denote by $\mathbf{x}_k = [x_{k1}, \cdots, x_{kT}]^T$ the training sequence of user k, where x_{kt} is the training symbol of user k in time slot t. We assume that the users transmit at constant power τ_X, i.e., $\mathbb{E}[|x_{kt}|^2] = \tau_X, \forall k, t$.

Over the time duration T, all the RIS elements are turned on and are set to have the same phase shift. Without loss of generality, we assume that $\boldsymbol{\psi}(t) = \mathbf{1}, 1 \leq t \leq T$. Then, the received signal at the BS in time slot t is given by

$$\mathbf{y}_0(t) = \sum_{k=1}^K (\mathbf{h}_{UB,k} + \mathbf{H}_{RB}\mathbf{h}_{UR,k})x_{kt} + \mathbf{n}(t), \quad 1 \leq t \leq T,$$

$$(1)$$

where $\mathbf{n}(t)$ is an additive white Gaussian noise (AWGN) vector following the distribution of $\mathcal{CN}(\mathbf{n}(t); \mathbf{0}, \tau_N \mathbf{I})$.

III. RIS CHANNEL MODEL AND PROBLEM FORMULATION

In this section, we discuss the models of the RIS-to-BS channel matrix $\mathbf{H}_{RB}$ and the user-to-RIS channel vectors $\{\mathbf{h}_{UR,k}\}$. Then, we show that the RIS channel estimation problem can be formulated as a sparse matrix factorization problem by employing virtual channel representations.

A. Channel Model

Consider the RIS-to-BS channel $\mathbf{H}_{RB}$. On one hand, since the BS and the RIS rarely move after deployment, a majority of the channel paths evolve very slowly, as compared with the channel coherence time (determined by the movement of mobile ends). We refer to these paths as the slow-varying channel components of $\mathbf{H}_{RB}$, denoted by $\bar{\mathbf{H}}_{RB}$. On the other hand, a small portion of the propagation paths may experience fast changes due to the change of the propagation geometry. We refer to them as the fast-varying channel components of $\mathbf{H}_{RB}$, denoted by $\widetilde{\mathbf{H}}_{RB}$. To illustrate this, consider an outdoor scenario depicted in Fig. 2, where a RIS is mounted on an advertising panel by the road. The propagation paths

---

[1]By turning off the RIS reflecting elements, the estimation of $\{\mathbf{h}_{UB,k}\}$ can be done by using conventional channel estimation methods for multiuser MIMO systems.
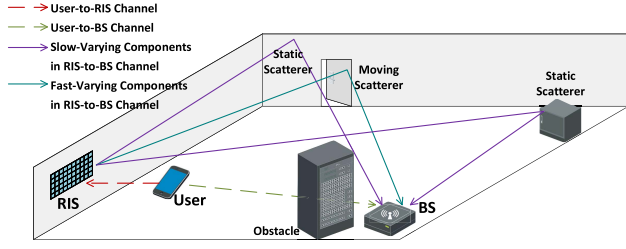
Fig. 3. An example of the indoor propagation geometry. The propagation path through a closing door belongs to the fast-varying channel components between the BS and the RIS, while the paths from static scatterers belong to the slow-varying channel components.

from static scattering clusters between the BS and the RIS changes very slowly and $\bar{\mathbf{H}}_{RB}$ sums up the paths from all the static scattering clusters. Meanwhile, suppose, for example, that a truck is moving past the RIS. The path through the truck evolves fast, as the corresponding propagation geometry is quickly changing. This fast-varying path is modelled in $\widetilde{\mathbf{H}}_{RB}$. As another example, consider an indoor environment depicted in Fig. 3, where a RIS is coated on a wall to assist the communication between the user and the BS. Most of the propagation paths are from static scattering clusters and are modelled in the slow-varying component matrix $\bar{\mathbf{H}}_{RB}$. Meanwhile, the paths through non-static scatterers (e.g., the closing door in Fig. 3) are modelled in the fast-varying component matrix $\widetilde{\mathbf{H}}_{RB}$.

To capture $\bar{\mathbf{H}}_{RB}$ and $\widetilde{\mathbf{H}}_{RB}$, we model $\mathbf{H}_{RB}$ by the MIMO Rician fading model as [25]

$$\mathbf{H}_{RB} = \sqrt{\frac{\kappa}{\kappa+1}}\bar{\mathbf{H}}_{RB} + \sqrt{\frac{1}{\kappa+1}}\widetilde{\mathbf{H}}_{RB}, \tag{2}$$

where κ is the Rician factor denoting the power ratio between $\bar{\mathbf{H}}_{RB}$ and $\widetilde{\mathbf{H}}_{RB}$.

By specifying the paths in the prorogation environment, $\bar{\mathbf{H}}_{RB}$ and $\widetilde{\mathbf{H}}_{RB}$ are given by

$$\bar{\mathbf{H}}_{RB} = \sqrt{\beta_0}\sum_{p=1}^{\bar{P}_{RB}}\alpha_p\mathbf{a}_B(\theta_p)\mathbf{a}_R^H(\phi_p,\sigma_p), \tag{3}$$

$$\widetilde{\mathbf{H}}_{RB} = \sqrt{\beta_0}\sum_{p=1}^{\widetilde{P}_{RB}}\alpha_p\mathbf{a}_B(\theta_p)\mathbf{a}_R^H(\phi_p,\sigma_p), \tag{4}$$

where β_0 is the large-scale path gain encompassing distance-dependent path loss and shadowing; $\bar{P}_{RB}$ (or $\widetilde{P}_{RB}$) is the number of the slow-varying (or fast-varying) paths; α_p is the corresponding complex-valued channel coefficient of the p-th path; θ_p is the corresponding azimuth angle-of-arrival (AoA) to the BS; ϕ_p (or σ_p) is the corresponding azimuth (or elevation) angle-of-departure (AoD) from the RIS; and $\mathbf{a}_B$ (or $\mathbf{a}_R$) is the steering vector associated with the BS (or RIS) antenna geometry. Specifically, they are given by

$$\mathbf{a}_B(\theta) = \mathbf{f}_M(\sin(\theta)), \tag{5a}$$

$$\mathbf{a}_R(\phi,\sigma) = \mathbf{f}_{L_2}(-\cos(\sigma)\cos(\phi)) \otimes \mathbf{f}_{L_1}(\cos(\sigma)\sin(\phi)), \tag{5b}$$

where $\otimes$ denotes the Kronecker product; and

$$\mathbf{f}_N(x) \triangleq \frac{1}{\sqrt{N}}\left[1, e^{-j\frac{2\pi}{\varrho}dx}, \cdots, e^{-j\frac{2\pi}{\varrho}d(N-1)x}\right]^T. \tag{6}$$

In (6), ϱ denotes the carrier wavelength; and d denotes the distance between any two adjacent antennas. Here, we set $d/\varrho = 1/2$ for simplicity.

Similarly to (3)–(4), we represent $\mathbf{h}_{UR,k}$ as

$$\mathbf{h}_{UR,k} = \sqrt{\beta_k}\sum_{p=1}^{P_k}\alpha_p\mathbf{a}_R(\phi_p,\sigma_p), \tag{7}$$

where β_k is the large-scale path gain for the channel between the k-th user and the RIS; and P_k is the number of paths between the k-th user and the RIS.

Finally, we assume without loss of generality that $\mathbb{E}[\|\bar{\mathbf{H}}_{RB}\|_F^2] = \mathbb{E}[\|\widetilde{\mathbf{H}}_{RB}\|_F^2] = \beta_0\,ML$ and $\mathbb{E}[\|\mathbf{h}_{UR,k}\|_2^2] = \beta_k L, \forall k$. In other words, β_0 (or β_k) can be regarded as the average power attenuation for each RIS-BS (or user-RIS) antenna pair.

B. Virtual Channel Representation

We assume that the slow-varying component matrix $\bar{\mathbf{H}}_{RB}$ in (3) keeps static over a time interval much larger than the coherence block length. As a consequence, $\sqrt{\kappa/(\kappa+1)}\bar{\mathbf{H}}_{RB}$ can be accurately estimated by long-term channel averaging prior to the RIS channel estimation procedure.[2]

Furthermore, as illustrated above, the fast-varying component matrix $\widetilde{\mathbf{H}}_{RB}$ contains a limited number of paths, i.e., $\widetilde{P}_{RB}$ in (4) is small. Following [26], we employ a pre-discretized sampling grid ϑ with length $M' (\geq M)$ to discretize $\{\sin(\theta_p)\}_{1\leq p\leq \widetilde{P}_{RB}}$ over $[0,1]$. Similarly, we employ two sampling grids φ with length $L_1' (\geq L_1)$ and ς with length $L_2' (\geq L_2)$ to discretize $\{\cos(\sigma_p)\sin(\phi_p)\}_{1\leq p\leq \widetilde{P}_{RB}}$ and $\{-\cos(\sigma_p)\cos(\phi_p)\}_{1\leq p\leq \widetilde{P}_{RB}}$, respectively. Then, we represent the fast-varying component matrix $\widetilde{\mathbf{H}}_{RB}$ under the angular bases as [26]

$$\sqrt{\frac{1}{\kappa+1}}\widetilde{\mathbf{H}}_{RB} = \mathbf{A}_B\mathbf{S}\left(\underbrace{\mathbf{A}_{R,v}\otimes\mathbf{A}_{R,h}}_{=\mathbf{A}_R}\right)^H, \tag{8}$$

where $\mathbf{A}_B \triangleq [\mathbf{f}_M(\vartheta_1), \cdots, \mathbf{f}_M(\vartheta_{M'})]$ is an over-complete array response with $\mathbf{f}_M$ defined in (6); $\mathbf{A}_{R,h} \triangleq [\mathbf{f}_{L_1}(\varphi_1), \cdots, \mathbf{f}_{L_1}(\varphi_{L_1'})]$ (or $\mathbf{A}_{R,v} \triangleq [\mathbf{f}_{L_2}(\varsigma_1), \cdots, \mathbf{f}_{L_2}(\varsigma_{L_2'})]$) is an over-complete horizontal (or vertical) array response; and $\mathbf{S} \in \mathbb{C}^{M'\times L'}$ is the corresponding channel coefficient matrix in the angular domain with $L' = L_1'L_2'$. Note that the (i,j)-th entry of $\mathbf{S}$ corresponds to the channel coefficient of $\widetilde{\mathbf{H}}_{RB}$ along the path specified by the i-th AoA steering vector in $\mathbf{A}_B$ and the j-th AoD steering vector in $\mathbf{A}_R$. Since the total number of paths $\widetilde{P}_{RB}$ is small, only a few entries of $\mathbf{S}$ are nonzero with each corresponding to a channel path. That is, $\mathbf{S}$ is a sparse matrix.

---

[2]The assumption of the perfect knowledge of $\sqrt{\kappa/(\kappa+1)}\bar{\mathbf{H}}_{RB}$ does not lose any generality since any possible error in the channel averaging process can be absorbed into $\sqrt{1/(\kappa+1)}\widetilde{\mathbf{H}}_{RB}$.

Similarly to (8), $\mathbf{h}_{UR,k}$ can be represented as

$$\mathbf{h}_{UR,k} = \mathbf{A}_R \mathbf{g}_k, \tag{9}$$

where $\mathbf{g}_k \in \mathbb{C}^{L' \times 1}$ represents the channel coefficients of $\mathbf{h}_{UR,k}$ in the angular domain. Experimental studies have shown that the propagation channel often exhibits limited scattering geometry [27]. As a consequence, $\mathbf{g}_k$ is also sparse, where each nonzero value corresponds to a channel path. We note that the sparsity of $\mathbf{S}$ and $\{\mathbf{g}_k\}$ plays an important role in our channel estimation design.

C. Cascaded Channel Estimation

With the above channel representations, we can rewrite (1) in a matrix form, as shown in the following proposition.

Proposition 1: With (2), (8), and (9), the model in (1) is equivalent to

$$\mathbf{Y} = \left(\mathbf{H}_0 + \mathbf{A}_B \mathbf{S} \mathbf{R}\right) \mathbf{G} \mathbf{X} + \mathbf{N}, \tag{10}$$

where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_K]^T$; $\mathbf{N} = [\mathbf{n}(1), \cdots, \mathbf{n}(T)]$, $\mathbf{Y} = [\mathbf{y}_0(1), \cdots, \mathbf{y}_0(T)] - [\mathbf{h}_{UB,1}, \cdots, \mathbf{h}_{UB,K}]\mathbf{X}$; $\mathbf{H}_0 \triangleq \sqrt{\kappa/(\kappa+1)}\bar{\mathbf{H}}_{RB}\mathbf{A}_R \in \mathbb{C}^{M \times L'}$; $\mathbf{R} \triangleq \mathbf{A}_R^H \mathbf{A}_R \in \mathbb{C}^{L' \times L'}$; and $\mathbf{G} \triangleq [\mathbf{g}_1, \cdots, \mathbf{g}_K] \in \mathbb{C}^{L' \times K}$.

Proof: See Appendix A.

Upon the reception of $\mathbf{Y}$ in (10), the BS aims to factorize the channel matrices $\mathbf{S}$ and $\mathbf{G}$ with the knowledge of the training signal matrix $\mathbf{X}$. Once the angular bases $\mathbf{A}_B$ and $\mathbf{A}_R$ are predetermined and the slow-varying component matrix $\sqrt{\kappa/(\kappa+1)}\bar{\mathbf{H}}_{RB}$ is given, the sensing matrices $\mathbf{A}_B$, $\mathbf{H}_0$, and $\mathbf{R}$ are also known to the BS. We refer to the above problem as *matrix-calibration* based cascaded channel estimation, since this problem bears some similarity to the blind matrix calibration (BMC) problem [28].

Remark 1: The BMC problem in [28] is defined as to calibrate $\mathbf{A}$ and estimate $\mathbf{B}$ from a noisy observation of the product $\mathbf{AB}$, where the dictionary $\mathbf{A}$ is partially known. That is, $\mathbf{A} = \mathbf{A}_0 + \mathbf{A}_1$ with the known part $\mathbf{A}_0$ and the unknown perturbation $\mathbf{A}_1$. Rigorously speaking, the problem in (10) is not exactly a BMC problem. The difference is twofold: First, in (10), the observation of the matrix product $\mathbf{Z} = (\mathbf{H}_0 + \mathbf{A}_B \mathbf{S} \mathbf{R})\mathbf{G}$ is through a linear system $\mathbf{Y} = \mathbf{ZX} + \mathbf{N}$. In BMC, $\mathbf{X}$ is limited to $\mathbf{X} = \mathbf{I}$. Second, BMC often assumes an i.i.d. Gaussian distribution for perturbation $\mathbf{A}_1$ [28], [29], whereas we exploit the hidden sparsity of the perturbation matrix under the domain transformed by $\mathbf{A}_B$ and $\mathbf{R}$ in (10). As aforementioned, such a sparse representation of the perturbation models the fast channel variation in the RIS-to-BS channel.

IV. MATRIX-CALIBRATION BASED CASCADED CHANNEL ESTIMATION ALGORITHM

In this section, we first derive the posterior mean estimators of $\mathbf{S}$ and $\mathbf{G}$ given $\mathbf{Y}$ in (10) under the Bayesian inference framework. We then resort to sum-product message passing to compute the estimators. To reduce the computational complexity, we impose additional approximations to simplify the message updates in the large-system limit. Finally, we analyze the computational complexity of the proposed algorithm.

TABLE I
NOTATION OF FACTOR NODES

Factor	Distribution	Exact Form
$p(s_{m'l'})$	$p(s_{m'l'})$	$(1-\lambda_S)\delta(s_{m'l'})+$ $\lambda_S \mathcal{CN}\left(s_{m'l'};0,\tau_S\right)$
$p(g_{lk})$	$p(g_{lk})$	$(1-\lambda_G)\delta(g_{lk})+$ $\lambda_G \mathcal{CN}\left(g_{lk};0,\tau_G\right)$
ws_{ml}	$p(w_{ml}\mid s_{m'l'},\forall m',l')$	$\delta(w_{ml}-h_{0,ml}-$ $\sum_{m',l'}a_{B,mm'}s_{m'l'}r_{l'l})$
zwg_{mk}	$p(z_{mk}\mid w_{ml},g_{lk},\forall l)$	$\delta(z_{mk}-\sum_{l=1}^{L'}w_{ml}g_{lk})$
qz_{mt}	$p(q_{mt}\mid z_{mk},\forall k)$	$\delta(q_{mt}-\sum_{k=1}^{K}z_{mk}x_{kt})$
$p(y_{mt}\mid q_{mt})$	$p(y_{mt}\mid q_{mt})$	$\mathcal{CN}(y_{mt};q_{mt},\tau_N)$

A. Bayesian Inference

Define $\mathbf{W} \triangleq \mathbf{H}_0 + \mathbf{A}_B \mathbf{S} \mathbf{R}$, $\mathbf{Z} \triangleq \mathbf{WG}$, and $\mathbf{Q} \triangleq \mathbf{ZX}$. Under the assumption of AWGN, we have

$$p(\mathbf{Y}|\mathbf{Q}) = \prod_{m=1}^{M}\prod_{t=1}^{T}\mathcal{CN}\left(y_{mt};q_{mt},\tau_N\right). \tag{11}$$

Motivated by the sparsity of $\mathbf{S}$ and $\mathbf{G}$, we employ Bernoulli-Gaussian distributions to model their prior distributions as

$$p(\mathbf{S}) = \prod_{m'=1}^{M'}\prod_{l'=1}^{L'}(1-\lambda_S)\delta(s_{m'l'}) + \lambda_S \mathcal{CN}\left(s_{m'l'};0,\tau_S\right), \tag{12}$$

$$p(\mathbf{G}) = \prod_{l=1}^{L'}\prod_{k=1}^{K}(1-\lambda_G)\delta(g_{lk}) + \lambda_G \mathcal{CN}\left(g_{lk};0,\tau_G\right), \tag{13}$$

where λ_S (or λ_G) is the corresponding Bernoulli parameter of $\mathbf{S}$ (or $\mathbf{G}$); and τ_S (or τ_G) is the variance of the nonzero entries of $\mathbf{S}$ (or $\mathbf{G}$).

With the prior distributions (11)–(13), the following proposition states the minimum MSEs (MMSEs) of $\mathbf{S}$ and $\mathbf{G}$.

Proposition 2: The posterior distribution is given by

$$p(\mathbf{S},\mathbf{G}|\mathbf{Y}) = \frac{1}{p(\mathbf{Y})}p(\mathbf{Y}|\mathbf{S},\mathbf{G})p(\mathbf{S})p(\mathbf{G}), \tag{14}$$

where $p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{S},\mathbf{G})p(\mathbf{S})p(\mathbf{G})\mathrm{d}\mathbf{S}\mathrm{d}\mathbf{G}$. Furthermore, the MMSEs of $\mathbf{S}$ and $\mathbf{G}$ are given by

$$\text{MMSE}_{\mathbf{S}} = \frac{1}{M'L'}\mathbb{E}\left[\|\mathbf{S}-\hat{\mathbf{S}}\|_F^2\right], \tag{15a}$$

$$\text{MMSE}_{\mathbf{G}} = \frac{1}{L'K}\mathbb{E}\left[\|\mathbf{G}-\hat{\mathbf{G}}\|_F^2\right], \tag{15b}$$

where the expectations are taken over the joint distribution of $\mathbf{S}$, $\mathbf{G}$, and $\mathbf{Y}$; and $\hat{\mathbf{S}} = [\hat{s}_{m'l'}]$ (or $\hat{\mathbf{G}} = [\hat{g}_{lk}]$) is the posterior mean estimator of $\mathbf{S}$ (or $\mathbf{G}$), given by

$$\hat{s}_{m'l'} = \int s_{m'l'}p(s_{m'l'}|\mathbf{Y})\mathrm{d}s_{m'l'}, \hat{g}_{lk} = \int g_{lk}p(g_{lk}|\mathbf{Y})\mathrm{d}g_{lk}. \tag{16}$$

In the above, $p(s_{m'l'}|\mathbf{Y}) = \int\int p(\mathbf{S},\mathbf{G}|\mathbf{Y})\mathrm{d}\mathbf{G}\mathrm{d}(\mathbf{S}\setminus s_{m'l'})$ and $p(g_{lk}|\mathbf{Y}) = \int\int p(\mathbf{S},\mathbf{G}|\mathbf{Y})\mathrm{d}\mathbf{S}\mathrm{d}(\mathbf{G}\setminus g_{lk})$ are the marginal distributions with respect to $s_{m'l'}$ and g_{lk}, respectively, where $\mathbf{X}\setminus x_{ij}$ means the collection of the elements of matrix $\mathbf{X}$ except for the (i,j)-th one.

Fig. 4. An illustration of the factor graph representation for $M = M' = K = 3$ and $T = L' = 2$, where blank circles and black squares represent variable nodes and factor nodes, respectively.

Proof: The result in (14) follows from Bayes' theorem, and (16) follows from [30, Page 143].

Exact evaluation of $\hat{\mathbf{S}}$ and $\hat{\mathbf{G}}$ are generally intractable due to the high-dimensional integrations involved in the marginalization. In the following, we provide an approximate solution by following the message passing principle.

B. Message Passing for Marginal Posterior Computation

Plugging (11)–(13) into (14), we obtain

$$
p(\mathbf{S}, \mathbf{G}|\mathbf{Y}) = \frac{1}{p(\mathbf{Y})} \left(\prod_{m=1}^{M} \prod_{t=1}^{T} p(y_{mt}|q_{mt}) p(q_{mt}|z_{mk}, \forall k) \right)
$$
$$
\times \left(\prod_{m=1}^{M} \prod_{k=1}^{K} p(z_{mk}|w_{ml}, g_{lk}, 1 \leq l \leq L') \right)
$$
$$
\times \left(\prod_{l=1}^{L'} \prod_{k=1}^{K} p(g_{lk}) \right)
$$
$$
\times \left(\prod_{m=1}^{M} \prod_{l=1}^{L'} p(w_{ml}|s_{m'l'}, \forall m', l') \right)
$$
$$
\times \left(\prod_{m'=1}^{M'} \prod_{l'=1}^{L'} p(s_{m'l'}) \right), \tag{17}
$$

where the factorizable distributions are defined in Table I.

We construct a factor graph to represent (17) and apply the canonical message passing algorithm to approximately compute the estimators in (16). The factor graph is depicted in Fig. 4. The variables $\mathbf{S}$, $\mathbf{G}$, $\mathbf{W}$, $\mathbf{Z}$, and $\mathbf{Q}$ are represented by the variable nodes $\{s_{m'l'}\}_{1 \leq m' \leq M', 1 \leq l' \leq L'}$, $\{g_{lk}\}_{1 \leq l \leq L', 1 \leq k \leq K}$, $\{w_{ml}\}_{1 \leq m \leq M, 1 \leq l \leq L'}$, $\{z_{mk}\}_{1 \leq m \leq M, 1 \leq k \leq K}$, and $\{q_{mt}\}_{1 \leq m \leq M, 1 \leq t \leq T}$, respectively. The factorizable pdfs in (17), represented by factor nodes $\{p(s_{m'l'})\}$, $\{p(g_{lk})\}$, $\{ws_{ml}\}$, $\{zwg_{mk}\}$, $\{qz_{mt}\}$, and $\{p(y_{mt}|q_{mt})\}$, are connected to their associated arguments. We summarize the notation of the factor nodes in Table I. Denote by $\Delta_{a \to b}^{i}(\cdot)$ the message from node a to b in iteration i, and by $\Delta_{c}^{i}(\cdot)$ the marginal message computed at variable node c in iteration i. Applying the sum-product rule, we obtain the following messages:

1) Messages Between $\{qz_{mt}\}$ and $\{z_{mk}\}$: For $1 \leq m \leq M, 1 \leq t \leq T, 1 \leq k \leq K$,

$$
\Delta_{qz_{mt} \to z_{mk}}^{i}(z_{mk}) \propto \int \mathcal{CN}\left(y_{mt}; \sum_{k=1}^{K} z_{mk} x_{kt}, \tau_N \right)
$$
$$
\times \prod_{j \neq k} \left(\Delta_{z_{mj} \to qz_{mt}}^{i}(z_{mj}) \mathrm{d}z_{mj} \right), \tag{18}
$$

$$
\Delta_{z_{mk} \to qz_{mt}}^{i+1}(z_{mk}) \propto \mathcal{P}_{z_{mk}}^{i}(z_{mk}) \prod_{j \neq t} \Delta_{qz_{mj} \to z_{mk}}^{i}(z_{mk}), \tag{19}
$$

where the auxiliary distribution $\mathcal{P}_{z_{mk}}^{i}(z_{mk})$ is defined as

$$
\mathcal{P}_{z_{mk}}^{i}(z_{mk})
$$
$$
\propto \int p(z_{mk}|w_{ml}, g_{lk}, 1 \leq l \leq L')
$$
$$
\times \prod_{l=1}^{L'} \left(\Delta_{w_{ml} \to zwg_{mk}}^{i}(w_{ml}) \Delta_{g_{lk} \to zwg_{mk}}^{i}(g_{lk}) \mathrm{d}w_{ml} \mathrm{d}g_{lk} \right). \tag{20}
$$

2) Messages Between $\{g_{lk}\}$ and $\{zwg_{mk}\}$: For $1 \leq l \leq L', 1 \leq m \leq M, 1 \leq k \leq K$,

$$
\Delta_{zwg_{mk} \to g_{lk}}^{i}(g_{lk})
$$
$$
\propto \int \mathrm{d}z_{mk} \, p(z_{mk}|w_{ml}, g_{lk}, 1 \leq l \leq L')
$$
$$
\times \prod_{t=1}^{T} \Delta_{qz_{mt} \to z_{mk}}^{i}(z_{mk}) \prod_{j \neq l} \left(\Delta_{g_{jk} \to zwg_{mk}}^{i}(g_{jk}) \mathrm{d}g_{jk} \right)
$$
$$
\times \prod_{l=1}^{L'} \left(\Delta_{w_{ml} \to zwg_{mk}}^{i}(w_{ml}) \mathrm{d}w_{ml} \right), \tag{21}
$$

$$
\Delta_{g_{lk} \to zwg_{mk}}^{i+1}(g_{lk}) \propto p(g_{lk}) \prod_{j \neq m} \Delta_{zwg_{jk} \to g_{lk}}^{i}(g_{lk}). \tag{22}
$$

3) Messages Between $\{w_{ml}\}$ and $\{zwg_{mk}\}$: For $1 \leq l \leq L', 1 \leq m \leq M, 1 \leq k \leq K$,

$$
\Delta_{zwg_{mk} \to w_{ml}}^{i}(w_{ml})
$$
$$
\propto \int \mathrm{d}z_{mk} \, p(z_{mk}|w_{ml}, g_{lk}, 1 \leq l \leq L')
$$
$$
\times \prod_{t=1}^{T} \Delta_{qz_{mt} \to z_{mk}}^{i}(z_{mk}) \prod_{l=1}^{L'} \left(\Delta_{g_{lk} \to zwg_{mk}}^{i}(g_{lk}) \mathrm{d}g_{lk} \right)
$$
$$
\times \prod_{j \neq l} \left(\Delta_{w_{mj} \to zwg_{mk}}^{i}(w_{mj}) \mathrm{d}w_{mj} \right), \tag{23}
$$

$$
\Delta_{w_{ml} \to zwg_{mk}}^{i+1}(w_{ml}) \propto \mathcal{P}_{w_{ml}}^{i}(w_{ml}) \prod_{j \neq k} \Delta_{zwg_{mj} \to w_{ml}}^{i}(w_{ml}), \tag{24}
$$

where

$$
\mathcal{P}_{w_{ml}}^{i}(w_{ml}) \propto \int \prod_{m'=1}^{M'} \prod_{l'=1}^{L'} \left(\Delta_{s_{m'l'} \to ws_{ml}}^{i}(s_{m'l'}) \mathrm{d}s_{m'l'} \right)
$$
$$
\times p(w_{ml}|s_{m'l'}, \forall m', l'). \tag{25}
$$

TABLE II
NOTATIONS OF MEANS AND VARIANCES FOR MESSAGES

Message	Mean	Variance
$\Delta^i_{g_{lk} \to zwg_{mk}}(g_{lk})$	$\hat{g}_{lk,m}(i)$	$v^g_{lk,m}(i)$
$\Delta^i_{s_{m'l'} \to ws_{ml}}(s_{m'l'})$	$\hat{s}_{m'l',ml}(i)$	$v^s_{m'l',ml}(i)$
$\Delta^i_{w_{ml} \to zwg_{mk}}(w_{ml})$	$\hat{w}_{ml,k}(i)$	$v^w_{ml,k}(i)$
$\Delta^i_{z_{mk} \to qz_{mt}}(z_{mk})$	$\hat{z}_{mk,t}(i)$	$v^z_{mk,t}(i)$
$\Delta^i_{g_{lk}}(g_{lk})$	$\hat{g}_{lk}(i)$	$v^g_{lk}(i)$
$\Delta^i_{s_{m'l'}}(s_{m'l'})$	$\hat{s}_{m'l'}(i)$	$v^s_{m'l'}(i)$
$\Delta^i_{w_{ml}}(w_{ml})$	$\hat{w}_{ml}(i)$	$v^w_{ml}(i)$
$\Delta^i_{z_{mk}}(z_{mk})$	$\hat{z}_{mk}(i)$	$v^z_{mk}(i)$

4) Messages Between $\{s_{m'l'}\}$ and $\{ws_{ml}\}$: For $1 \le l, l' \le L', 1 \le m' \le M', 1 \le m \le M$,

$$\Delta^i_{ws_{ml} \to s_{m'l'}}(s_{m'l'})$$
$$\propto \int \mathrm{d}w_{ml}\, p(w_{ml}|s_{m'l'}, \forall m', l')$$
$$\times \prod_{k=1}^K \Delta^i_{zwg_{mk} \to w_{ml}}(w_{ml})$$
$$\times \prod_{(j,n) \ne (m',l')} \left(\Delta^i_{s_{jn} \to ws_{ml}}(s_{jn}) \mathrm{d}s_{jn} \right), \qquad (26)$$

$$\Delta^{i+1}_{s_{m'l'} \to ws_{ml}}(s_{m'l'})$$
$$\propto p(s_{m'l'}) \prod_{(j,n) \ne (m,l)} \Delta^i_{ws_{jn} \to s_{m'l'}}(s_{m'l'}). \qquad (27)$$

5) Marginal Messages at Variable Nodes: For $1 \le l, l' \le L', 1 \le m' \le M', 1 \le m \le M, 1 \le k \le K$,

$$\Delta^{i+1}_{z_{mk}}(z_{mk}) \propto \mathcal{P}^i_{z_{mk}}(z_{mk}) \prod_{t=1}^T \Delta^i_{qz_{mt} \to z_{mk}}(z_{mk}), \qquad (28)$$

$$\Delta^{i+1}_{w_{ml}}(w_{ml}) \propto \mathcal{P}^i_{w_{ml}}(w_{ml}) \prod_{k=1}^K \Delta^i_{zwg_{mk} \to w_{ml}}(w_{ml}), \qquad (29)$$

$$\Delta^{i+1}_{g_{lk}}(g_{lk}) \propto p(g_{lk}) \prod_{m=1}^M \Delta^i_{zwg_{mk} \to g_{lk}}(g_{lk}), \qquad (30)$$

$$\Delta^{i+1}_{s_{m'l'}}(s_{m'l'}) \propto p(s_{m'l'}) \prod_{m=1}^M \prod_{l=1}^{L'} \Delta^i_{ws_{ml} \to s_{m'l'}}(s_{m'l'}). \qquad (31)$$

C. Approximations for Message Passing

The messages in (18)–(31) are computationally intractable in general due to the high-dimensional integrations and normalizations therein. To tackle this, we simplify the calculation of (18)–(31) by following the idea of AMP [21], [22] in the large-system limit, i.e., $M, M', K, L, L', T, \tau_N \to \infty$ with the ratios M/K, M'/K, L/K, L'/K, T/K, and τ_N/K^2 fixed. For ease of notation, we define the means and variances of the messages as in Table II. In the following, we sketch the main approximations involved in the algorithm design, while the detailed derivations can be found in Appendix B.

- A second-order Taylor expansion is adopted to approximate $\mathcal{P}^i_{z_{mk}}$ as a Gaussian distribution. Moreover, the CLT argument [21] approximates $\prod_{j \ne k} \Delta^i_{z_{mj} \to qz_{mt}}$ as a

---

Algorithm 1: The Proposed Algorithm

Input: $\mathbf{Y}; \mathbf{A}_B; \mathbf{H}_0; \mathbf{R}; \mathbf{X}; \tau_N; \lambda_S; \lambda_G; \tau_S; \tau_G$.
Initialization: $\hat{\gamma}_{mt}(0) = \hat{\xi}_{mk}(0) = \hat{\alpha}_{ml}(0) = 0$;
$\hat{s}_{m'l'}(1)$ randomly drawn from $p(s_{m'l'})$;
$\hat{w}_{ml}(1) = h_{0,ml} + \sum_{m',l'} a_{B,mm'} \hat{s}_{m'l'}(1) r_{l'l}$;
$\hat{g}_{lk}(1) = \hat{z}_{mk}(1) = 0$;
$v^s_{m'l'}(1) = v^g_{lk}(1) = v^w_{ml}(1) = v^z_{mk}(1) = 1$.
  for $i = 1, 2, \cdots, I_{\max}$ **do**
    For $\forall m, k$, update $\hat{z}_{mk}$ and v^z_{mk} by (50), (49), (56), and (48c)–(48d).
    For $\forall l, k$, update $\hat{g}_{lk}$ and v^g_{lk} by (54), (53a)–(53b), and (55).
    For $\forall m, l$, update $\hat{w}_{ml}$ and v^w_{ml} by (53c)–(53d), (61a)–(61b), and (63).
    For $\forall m', l'$, update $\hat{s}_{m'l'}$ and $v^s_{m'l'}$ by (62), (61c)–(61d), and (64).
    if $\sqrt{\frac{\sum_{m'} \sum_{l'} |\hat{s}_{m'l'}(i+1) - \hat{s}_{m'l'}(i)|^2}{\sum_{m'} \sum_{l'} |\hat{s}_{m'l'}(i)|^2}} \le \epsilon$ **and**
    $\sqrt{\frac{\sum_l \sum_k |\hat{g}_{lk}(i+1) - \hat{g}_{lk}(i)|^2}{\sum_l \sum_k |\hat{g}_{lk}(i)|^2}} \le \epsilon$ **then**
      Stop;
    end if
  end for
Output: $\hat{s}_{m'l'}$ and $\hat{g}_{lk}$.

Gaussian distribution. Based on these two approximations, we characterize both $\Delta^{i+1}_{z_{mt} \to qz_{mk}}$ and $\Delta^{i+1}_{z_{mk}}$ as Gaussian distributions with tractable means and variances.

- To further reduce the computational complexity, we show that $\Delta^{i+1}_{z_{mk} \to qz_{mt}}$ differs from $\Delta^{i+1}_{z_{mk}}$ in only one term, and this term vanishes in the large-system limit. As a consequence, we use the mean and variance of $\Delta^{i+1}_{z_{mk}}$ (i.e., $\hat{z}_{mk}(i+1)$ and $v^z_{mk}(i+1)$) to approximate those of $\Delta^{i+1}_{z_{mk} \to qz_{mt}}$, and obtain closed-loop updating formulas for $\hat{z}_{mk}$ and v^z_{mk}.

- With the tractable form of $\Delta^{i+1}_{z_{mk}}$, we show in a similar way that $\Delta^{i+1}_{w_{ml}}$, $\prod_m \Delta^i_{zwg_{mk} \to g_{lk}}$, and $\prod_{m,l} \Delta^i_{ws_{ml} \to s_{m'l'}}$ can be approximated as Gaussian distributions. Plugging these expressions into (30)–(31) and taking the prior information (12)–(13) into account, we obtain the closed-loop updating formulas for $\hat{g}_{lk}$, v^g_{lk}, $\hat{s}_{m'l'}$ and $v^s_{m'l'}$.

We summarize the resultant algorithm in Algorithm 1. In Algorithm 1, we employ the adaptive damping technique [23] to improve the convergence. The details are omitted due to space limitation. Besides, we terminate Algorithm 1 either when the maximum allowable number of iterations $I_{\max}$ is reached or when the change between any two consecutive iterations is smaller than ϵ for ϵ being a small positive value. The MATLAB code is available at https://github.com/liuhang1994/Matrix-Calibration-Based-Cascaded-Channel-Estimation.

D. Computational Complexity

In this subsection, we analyze the computational complexity of the proposed algorithm. Recall that each iteration of Algorithm 1 consists of four steps. The computational complexity of each step is detailed as follows. The complexity

of computing $\hat{z}_{mk}$ and v_{mk}^z is $\mathcal{O}(MKT)$; the complexity of computing $\hat{g}_{lk}$ and v_{lk}^g is $\mathcal{O}(MKL')$; the complexity of computing $\hat{w}_{ml}$ and v_{ml}^w is $\mathcal{O}(MKL')$; and the complexity of computing $\hat{s}_{m'l'}$ and $v_{m'l'}^s$ is $\mathcal{O}\left(MM'(L')^2\right)$. As a consequence, the overall complexity of the proposed algorithm is $\mathcal{O}\left(I(MKT + 2MKL' + MM'(L')^2)\right)$, where $I \leq I_{\max}$ is the number of iterations in Algorithm 1.

When the ratios M/K, M'/K, L/K, L'/K, and T/K are fixed, the overall complexity of the proposed algorithm can be simplified as $\mathcal{O}\left(IK^4\right)$ with a large K. In contrast, both the MMSE estimators (16) and the canonical message passing algorithm (18)–(31) involve integrations with respect to $\mathbf{S}$ or $\mathbf{G}$. Therefore, the corresponding computational complexity grows exponentially with K^2. We conclude that our approximations in Algorithm 1 significantly reduce the computational complexity compared with the conventional estimators in (16) and (18)–(31).

V. MSE ANALYSIS BY REPLICA METHOD

As discussed in Section IV-A, the MMSEs in (15) are difficult to evaluate in general. In this section, we derive an asymptotic performance bound of the MSEs under some mild assumptions by employing the replica method from statistical physics [24]. Although some basic assumptions used in the replica method cannot be rigorously justified, this method has proved successful in analyzing bilinear matrix factorization problems [28], [29]. Moreover, the result in [29] is adopted in analyzing the symbol error rate of the massive MIMO systems with low-precision analog-to-digital convertors in [31]. The analysis in this section can be regarded as an extension of the replica framework in [29] to the considered matrix-calibration-based cascaded channel estimation problem in (10).

A. Asymptotic MSE Analysis

We assume that the prior distributions (11)–(13) are perfectly known. Furthermore, throughout this subsection, we restrict our analysis to the scenario where both the sampling grids $\boldsymbol{\varphi}$ and $\boldsymbol{\varsigma}$ uniformly cover $[-1, 1]$ with $L_1' = L_1$ and $L_2' = L_2$. As a consequence, the array response matrices $\mathbf{A}_{B,v}$ and $\mathbf{A}_{B,h}$ in (8) become two normalized discrete Fourier transform (DFT) matrices and we have $\mathbf{R} = \mathbf{I}$. Discussions on the case of over-complete sampling bases with $L_1' > L_1$ and/or $L_2' > L_2$ can be found in the next subsection. Under these assumptions, the analysis is conducted by evaluating the average free entropy in the large-system limit, i.e., $M, M', K, L, L', T, \tau_N \rightarrow \infty$ with the ratios M/K, M'/K, L/K, L'/K, T/K, and τ_N/K^2 fixed. In the sequel, we use $K \rightarrow \infty$ to denote this limit for convenience. We define the average free entropy as

$$\mathcal{F} \triangleq \lim_{K \to \infty} \frac{1}{K^2} \mathbb{E}_{\mathbf{Y}}\left[\ln p(\mathbf{Y})\right]. \tag{32}$$

As shown in [28], the MMSEs in (15) correspond to a stationary point of $\mathcal{F}$. Following the argument in [31, eq. (35)], we first transform (32) as

$$\mathcal{F} = \lim_{n \to 0} \frac{\partial}{\partial n} \lim_{K \to \infty} \frac{1}{K^2} \ln \mathbb{E}_{\mathbf{Y}}\left[p^n(\mathbf{Y})\right]. \tag{33}$$

With (33), we compute $\ln \mathbb{E}_{\mathbf{Y}}\left[p^n(\mathbf{Y})\right]/K^2$ as $K \rightarrow \infty$ and then compute the derivative with respect to n with $n \rightarrow 0$. Finally, we compute the stationary point of $\mathcal{F}$ to obtain the values of the MMSEs.

To facilitate the computation, we define

$$Q_S = \lambda_S \tau_S, Q_G = \lambda_G \tau_G, \tag{34a}$$

$$Q_W = \frac{M'}{M} Q_S + \tau_{H_0}, Q_Z = L' Q_W Q_G, \tag{34b}$$

where $\tau_{H_0} \triangleq \mathbb{E}[|h_{0,ml}|^2]$.

Besides, we define two *scalar* AWGN channels as

$$Y_S = \sqrt{\widetilde{m}_S} S + N_S, Y_G = \sqrt{\widetilde{m}_G} G + N_G, \tag{35}$$

where $N_S, N_G \sim \mathcal{CN}(\cdot; 0, 1)$; $S \sim p(S) \triangleq (1 - \lambda_S)\delta(S) + \lambda_S \mathcal{CN}(S; 0, \tau_S)$; $G \sim p(G) \triangleq (1 - \lambda_G)\delta(G) + \lambda_G \mathcal{CN}(G; 0, \tau_G)$. The parameters $\widetilde{m}_S, \widetilde{m}_G$ will be specified later in (37).

From Bayes' theorem, we obtain the posterior distributions of S and G as $p(S|Y_S) \propto p(S)p(Y_S|S)$ and $p(G|Y_G) \propto p(G)p(Y_G|G)$. Similarly to (16), the corresponding posterior mean estimators are given by $\hat{S} = \int S p(S|Y_S)\mathrm{d}S$ and $\hat{G} = \int G p(G|Y_G)\mathrm{d}G$. Moreover, these estimators achieve the following MSEs:

$$\mathrm{MSE}_S = \mathbb{E}_{S,Y_S}\left[|S - \hat{S}|^2\right], \tag{36a}$$

$$\mathrm{MSE}_G = \mathbb{E}_{G,Y_G}\left[|G - \hat{G}|^2\right]. \tag{36b}$$

With these definitions, we show in Appendix C that the MMSEs of $\mathbf{S}$ and $\mathbf{G}$ in (15) converge to MSE_S and MSE_G as $K \rightarrow \infty$, where $(\mathrm{MSE}_S, \mathrm{MSE}_G)$ is the non-trivial solution to the following fixed-point equations:

$$\widetilde{m}_Z = \frac{T \tau_X}{\tau_N + K \tau_X (Q_Z - m_Z)}, \tag{37a}$$

$$\widetilde{m}_W = \frac{K m_G}{1/\widetilde{m}_Z + Q_Z - L' m_W m_G}, \tag{37b}$$

$$\widetilde{m}_G = \frac{M m_W}{1/\widetilde{m}_Z + Q_Z - L' m_W m_G}, \tag{37c}$$

$$\widetilde{m}_S = \frac{1}{1/\widetilde{m}_W + Q_W - \tau_{H_0} - M' m_S/M}, \tag{37d}$$

$$m_Z = Q_Z - \frac{Q_Z - L' m_W m_G}{1 + \widetilde{m}_Z (Q_Z - L' m_W m_G)} \tag{37e}$$

$$m_W = Q_W - \frac{Q_W - \tau_{H_0} - M' m_S/M}{1 + \widetilde{m}_W (Q_W - \tau_{H_0} - M' m_S/M)}, \tag{37f}$$

$$m_G = (1 - \lambda_G)\mathbb{E}_{N_G}\left[\left|f_G\left(\frac{N_G}{\sqrt{\widetilde{m}_G}}, \frac{1}{\widetilde{m}_G}\right)\right|^2\right] + \lambda_G \mathbb{E}_{N_G}\left[\left|f_G\left(\frac{N_G \sqrt{\widetilde{m}_G + 1}}{\sqrt{\widetilde{m}_G}}, \frac{1}{\widetilde{m}_G}\right)\right|^2\right], \tag{37g}$$

$$m_S = (1 - \lambda_S)\mathbb{E}_{N_S}\left[\left|f_S\left(\frac{N_S}{\sqrt{\widetilde{m}_S}}, \frac{1}{\widetilde{m}_S}\right)\right|^2\right] + \lambda_S \mathbb{E}_{N_S}\left[\left|f_S\left(\frac{N_S \sqrt{\widetilde{m}_S + 1}}{\sqrt{\widetilde{m}_S}}, \frac{1}{\widetilde{m}_S}\right)\right|^2\right], \tag{37h}$$

$$\mathrm{MSE}_G = Q_G - m_G, \tag{37i}$$

$$\mathrm{MSE}_S = Q_S - m_S, \tag{37j}$$

In (37g)–(37h), we define

$$f_G(x, y) = \frac{\int G p(G) \mathcal{CN}(G; x, y)\mathrm{d}G}{\int p(G) \mathcal{CN}(G; x, y)\mathrm{d}G}, \tag{38a}$$

$$f_S(x,y) = \frac{\int S p(S) \mathcal{CN}(S;x,y)\mathrm{d}S}{\int p(S)\mathcal{CN}(S;x,y)\mathrm{d}S}. \qquad (38\mathrm{b})$$

As a result, we can obtain MSEs in (36) by computing the fixed-point solution of (37), which asymptotically describes the MMSEs in (15). This can be efficiently implemented by iteratively updating $\{\mathrm{MSE}_S, \mathrm{MSE}_G, m_o, \widetilde{m}_o \; : \; o \in \{Z, W, S, G\}\}$ following (37) until convergence.

B. Further Discussions

It is worth noting that the replica method in Appendix C adopts the CLT to approximate $w_{ml}^{(a)}$, $z_{mk}^{(a)}$, and $q_{mt}^{(a)}$ as Gaussian distributions in (73), where $x^{(a)}, 0 \leq a \leq n$ is the a-th replica of random variable x and follows the same distribution as x. To apply the CLT, we restrict the sampling bases $\mathbf{A}_{B,v}$ and $\mathbf{A}_{B,h}$ to be two normalized DFT matrices in Section V-A. In a more general case where $\mathbf{A}_{B,v}$ and $\mathbf{A}_{B,h}$ are two over-complete bases, the Gaussian approximations in (73) may be inaccurate and consequently affect the accuracy of the replica method. In this case, MSE_S and MSE_G in (36) may not exactly correspond to the stationary point of the average free entropy in (32). As a consequence, the performance bound derived in Section V-A may become loose, as verified by the numerical results presented in the next section.

VI. NUMERICAL RESULTS

A. Simulation Results Under Channel Generation Model (11)–(13)

We conduct Monte Carlo simulations to verify the analysis in Section V. In this subsection, we assume that the channel is generated according to the prior distributions in (11)–(13). This allows us to calculate the MMSEs of the cascaded channel estimation problem defined in (15) by using the replica method as in Section V. Note that MSE_S and MSE_G in (36) derived by the replica method will be used as a benchmark to evaluate the performance of our proposed algorithm (i.e., Algorithm 1).[3]

We first verify the efficiency of Algorithm 1 over the conventional MMSE estimators. Table III lists the running time for Algorithm 1 and the MMSE estimators (16) with $M = M' = L = L' = K$, where $I_{\max}$ is set to 2000 for Algorithm 1 and the integrations in computing the MMSE estimators are implemented by numerical quadratures with 2 integration points per dimension.[4] The simulation platform is a Dell Optiplex 9010 Desktop with an i7–3770/3.40GHz Quad-Core CPU and 16GB RAM. All results are obtained by averaging over 100 Monte Carlo trials. We see that the proposed algorithm runs much faster than the MMSE estimators. This is because the proposed approximations avoid the high-dimensional integrations in directly computing (16). In addition, the running time of the MMSE estimators sharply increases as K becomes large. This is consistent with the analysis in Section IV-D.

---

[3]Although the replica analysis in Section V is valid only in the large-system limit, we still use the derived performance bound as a benchmark even when we conduct simulations with finitely large systems.

[4]Note that a larger number of integration points will improve the integration accuracy but increase the running time.

TABLE III
RUNNING TIME IN SECONDS

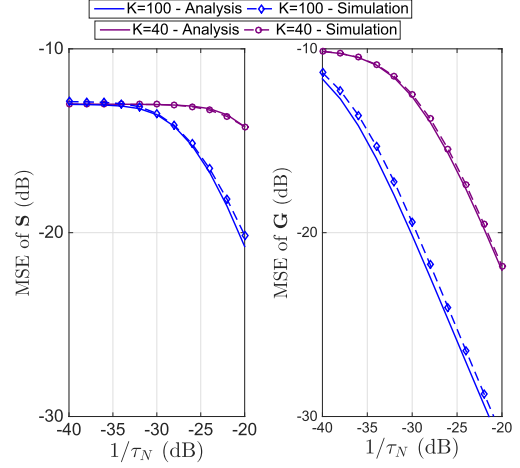| System Size	Algorithm 1	MMSE Estimators (16)
$K = 1$	$1.81 * 10^{-3}$	$6.75 * 10^{-1}$
$K = 2$	$2.33 * 10^{-3}$	$1.02 * 10^{4}$
$K = 3$	$2.82 * 10^{-3}$	$\approx 4.43 * 10^{10}$

Fig. 5. MSEs versus τ_N under normalized DFT bases $\mathbf{A}_{B,v}$ and $\mathbf{A}_{B,h}$. For $K = 40$, we set $L_1 = L_1' = 4$ and $L_2 = L_2' = 5$; For $K = 100$, we set $L_1 = L_1' = 10$ and $L_2 = L_2' = 5$.
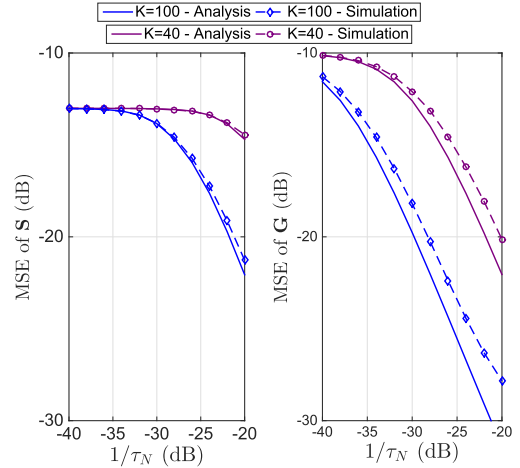

Fig. 6. MSEs versus τ_N under over-complete bases $\mathbf{A}_{B,v}$ and $\mathbf{A}_{B,h}$. For $K = 40$, we set $L_1 = L_1' = 4$, $L_2 = 5$, and $L_2' = 7$; For $K = 100$, we set $L_1 = L_1' = 10$, $L_2 = 5$, and $L_2' = 7$.

Then, we investigate the performance of the proposed algorithm with normalized DFT bases $\mathbf{A}_{B,v}$ and $\mathbf{A}_{B,h}$. We set $M = 1.28K$, $M' = 1.6K$, $T = 1.5K$, $L = L' = 0.5K$, $\lambda_G = 0.1$, $\lambda_S = 0.05$, and $\tau_S = \tau_G = \tau_{H_0} = \tau_X = 1$. Fig. 5 plots the MSE performance versus noise power for $K = 40$ and $K = 100$. The simulation results are obtained by averaging over 5000 Monte Carlo trials. We see that the performance of the proposed message passing algorithm closely matches the analytical bound derived by the replica method.

Next, we simulate on the scenario with an over-complete $\mathbf{A}_{B,v}$. Specifically, we set $L' = 0.7K$ and keep the other parameters unchanged. From Fig. 6, we find that the analytical
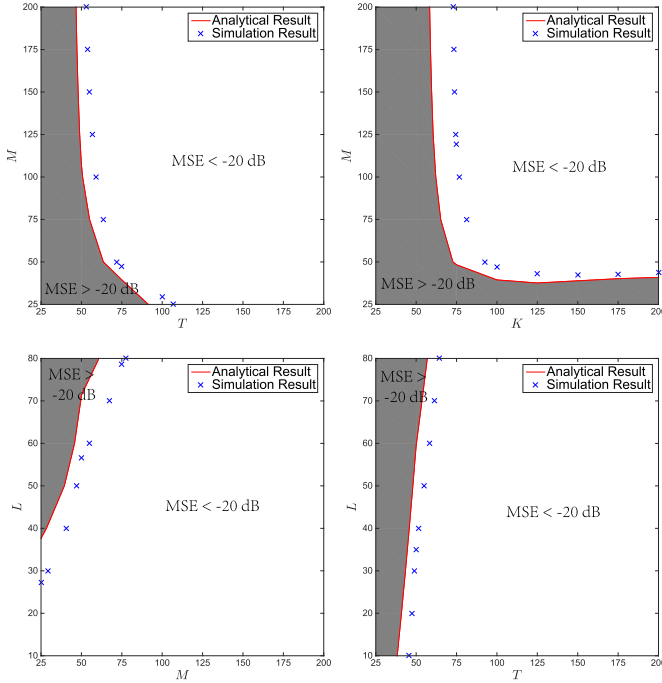
Fig. 7.   Phase diagrams with various M, K, T, and L. Unless otherwise specified, we set $\lambda_S = 0.05$, $\lambda_G = 0.1$, $\tau_N = 15$dB, $K = 100$, $M = 150$, $M' = 1.25M$, $T = 75$, $L = 50$ and $L' = L$. The red line and the blue cross markers represent the settings that achieve -20dB MSEs for the analytical result and Algorithm 1, respectively.

result has a small gap from the simulation result. As discussed in Section V-B, this is because the Gaussian approximations in (73) are less accurate, and hence the performance bound derived by the replica method is not as tight as in Fig. 5.

Next, we study the phase transitions of the cascaded channel estimation problem. Here, we say that the estimation of $\mathbf{S}$ and $\mathbf{G}$ is successful if $\text{MSE}_{\mathbf{S}} < -20$ dB and $\text{MSE}_{\mathbf{G}} < -20$ dB. In Fig. 7, we plot the phase transition diagrams for the analytical MSEs (36) and the empirical MSEs of Algorithm 1. We find sharp phase transitions for the analytical MSEs, and Algorithm 1 requires slightly larger parameters to achieve successful estimation. We conclude that the performance of the proposed algorithm is close to the theoretical performance bound in all the simulated settings.

Finally, we investigate the required training length T for RIS channel estimation in the noiseless case (i.e., $\tau_N = 0$). In Table IV, we list the minimum training length T of the proposed algorithm to achieve near-zero estimation errors ($\text{MSE}_{\mathbf{S}} < -50$ dB and $\text{MSE}_{\mathbf{G}} < -50$ dB), where we set $\lambda_S = 0.05$, $\lambda_G = 0.1$, $M' = 1.25M$, and $L' = L$. The performance bound derived from the replica analysis is included as well. Besides, we compare the proposed algorithm with the methods in [15], [18]. Specifically, when $\tau_N = 0$, the training length T in [18] is $L + \max\{K - 1, \lceil (K-1)L/M \rceil\}$, where $\lceil \cdot \rceil$ is the ceiling function. And the method in [15] requires $T = LK$. It can be seen that our proposed algorithm requires a smaller T compared with the two baselines. In particular, for the considered settings, the proposed algorithm only needs about 30% of the training overhead of [18] and about 1% of the overhead of [15]. This is because our algorithm exploits

the information on the slowing-varying channel components and the channel sparsity in the angular domain, which significantly reduces the number of variables to be estimated.

B. Simulation Results Under a More Realistic Channel Generation Model

We now consider a more realistic channel generation model as follows. The large-scale fading component is given by $\beta_i = \beta_{\text{ref}} \cdot d_i^{-\alpha_i}$, $0 \le i \le K$, where β_{ref} is the reference path loss at the distance 1 meter (m); d_i is the corresponding link distance; and α_i is the corresponding pass loss exponent. We set $\beta_{\text{ref}} = -20$ dB for all the channel links; $\alpha_0 = 2$; $\alpha_k = 2.6, 1 \le k \le K$; $d_0 = 50$ m; and d_k uniformly drawn from $[10 \text{ m}, 12 \text{ m}]$.

Moreover, we generate $\bar{\mathbf{H}}_{RB}$ by (3) with 20 clusters of paths and 10 subpaths per cluster. We draw the central azimuth AoA at the BS of each cluster uniformly over $[-90°, 90°]$; draw the central azimuth (or elevation) AoD at the RIS of each cluster uniformly over $[-180°, 180°]$ (or $[-90°, 90°]$); and draw each subpath with a $10°$ angular spread. The channels $\widetilde{\mathbf{H}}_{RB}$ and $\mathbf{h}_{UR,k}$ are generated by (4) and (7) in a similar way both with a cluster of 10 subpaths. Moreover, every α_p is drawn from $\mathcal{CN}(\alpha_p; 0, 1)$ and is normalized to satisfy $\|\mathbf{H}_{RB}\|_F^2 = \beta_0 ML$ and $\|\mathbf{h}_{UR,k}\|_2^2 = \beta_k L$. We set $K = 20$, $M = 60$, $T = 35$, $L_1 = L_2 = 4$ (i.e., $L = 16$), $\tau_X = 1$, and $\kappa = 9$. For the proposed algorithm, we set $I_{\max} = 2000$, $\epsilon = 10^{-4}$, and the over-complete bases $\boldsymbol{\vartheta}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\varsigma}$ in (8)–(9) to be uniform sampling grids covering $[-1, 1]$. The lengths of the sampling grids are set to have a fixed ratio to the antenna dimensions, i.e., $M'/M = L_1'/L_1 = L_2'/L_2 = 2$, unless otherwise specified. All the results in the sequel are conducted by averaging over 1500 Monte Carlo trials.

Apart from the proposed algorithm, the following baselines are involved for comparisons:

- Concatenate linear regression (LR): By setting aside $\widetilde{\mathbf{H}}_{RB}$, we first set the estimate of $\mathbf{H}_{RB}$ as $\hat{\mathbf{H}}_{RB} = \sqrt{\kappa/(\kappa + 1)} \bar{\mathbf{H}}_{RB}$, which is assumed to be deterministic. From (10), we obtain the following linear regression problem:

$$\text{vec}(\mathbf{Y}) = (\mathbf{X}^T \otimes \hat{\mathbf{H}}_{RB} \mathbf{A}_R) \text{vec}(\mathbf{G}) + \mathbf{n}', \quad (39)$$

where $\text{vec}(\cdot)$ represents the vectorization operator, and $\mathbf{n}'$ represents the effective AWGN. We then infer $\hat{\mathbf{G}}$ from (39) by employing generalized AMP (GAMP) [22]. Finally, we employ GAMP to estimate $\widetilde{\mathbf{H}}_{RB}$ with the estimated $\hat{\mathbf{G}}$.

- Oracle bound with $\mathbf{H}_{RB}$ known: Assume that an oracle gives the accurate value of $\mathbf{H}_{RB}$. Similarly to (39), we employ GAMP to obtain $\hat{\mathbf{G}}$ with $\hat{\mathbf{H}}_{RB} = \mathbf{H}_{RB}$.

Besides, the following channel estimation methods are included as baselines in the sequel.

- Method in [15]: RIS channels are estimated sequentially. From time slots $(l - 1)K + 1$ to lK, $1 \le l \le L$, we turn off all the RIS elements but the l-th one. With orthogonal training symbols from the users, the BS computes the linear MMSE estimators of the channel coefficients associated with the l-th RIS element.

TABLE IV
MINIMUM TRAINING LENGTH T REQUIRED IN DIFFERENT METHODS

System Size	Algorithm 1	Replica Analysis	Method in [18]	Method in [15]
$K = 100,\ L = 50,\ M = 128$	**44**	11	149	5000
$K = 100,\ L = 100,\ M = 100$	**60**	26	199	10000
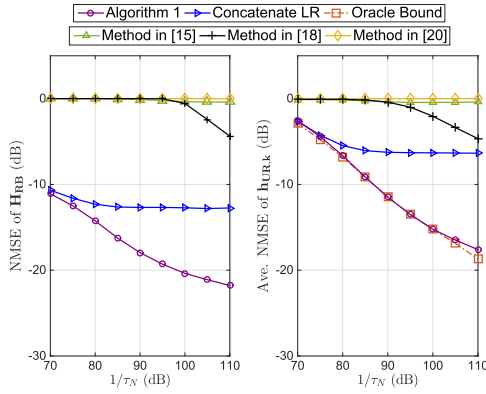
Fig. 8. The NMSE performance versus noise power with $T = 35$.

Fig. 9. The NMSE performance versus sampling resolution η, where we set $\tau_N = -95$ dB.

- Method in [18]: First, the first user sends an all-one training sequence with length no less than L to the BS, and the BS estimates the channel coefficients of this user (i.e., $\mathbf{H}_{RB} \operatorname{diag}(\mathbf{h}_{UR,1})$). Then, the other users sequentially send a training symbol to the BS, and the BS estimates $\{\mathbf{H}_{RB} \operatorname{diag}(\mathbf{h}_{UR,k})\}_{2 \leq k \leq K}$ by exploiting the correlations with the channel of the first user.
- Method in [20]: RIS Channel estimation is divided into P phases with $P = \lfloor T/K \rfloor \leq L$, where $\lfloor \cdot \rfloor$ is the floor function. In the p-th phase, the RIS phase-shift vector $\boldsymbol{\psi}_p$ is set as the p-th column of an $L \times L$ DFT matrix. The users send orthogonal training sequences with length no less than K to the BS. The BS collects the received signals in P phases and estimates the cascaded channels $\mathbf{H}_{RB}$ and $\{\mathbf{h}_{UR,k}\}$ alternatively by using parallel factor decomposition.

The estimates of all the three baselines suffer from scaling ambiguities. *For the comparison purpose*, we remove the scaling ambiguities by assuming the perfect knowledge of $\mathbf{h}_{UR,1}$. Besides, we set the total training length T to be the same for all the algorithms.

We use the normalized MSEs (NMSEs) of $\mathbf{H}_{RB}$ and $\{\mathbf{h}_{UR,k}\}$ to evaluate the performance of the cascaded channel estimation algorithms. Specifically, they are given by

$$\text{NMSE of } \mathbf{H}_{RB} = \frac{\|\hat{\mathbf{H}}_{RB} - \mathbf{H}_{RB}\|_F^2}{\|\mathbf{H}_{RB}\|_F^2}, \quad (40a)$$

$$\text{Average NMSE of } \mathbf{h}_{UR,k} = \frac{1}{K} \sum_{k=1}^{K} \frac{\|\hat{\mathbf{h}}_{UR,k} - \mathbf{h}_{UR,k}\|_2^2}{\|\mathbf{h}_{UR,k}\|_2^2}. \quad (40b)$$

Fig. 8 investigates the RIS channel estimation performance as τ_N varies. It can be seen that 1) the proposed algorithm achieves an NMSE of $\mathbf{h}_{UR,k}$ that is very close to the oracle
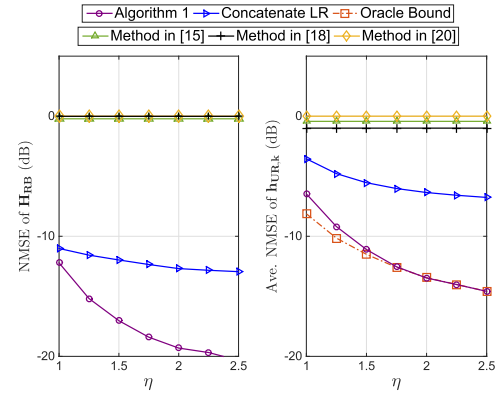
bound, which assumes NMSE of $\mathbf{H}_{RB}$ is zero; 2) our algorithm outperforms the baselines, especially when the noise power is large. This is because our proposed algorithm exploits the information on the slowing-varying channel components and the hidden channel sparsity; 3) the NMSEs of concatenate LR does not decrease when $\tau_N \leq -90$ dB. The reason is that the effective noise $\mathbf{n}'$ in (39) is a combination of the original AWGN noise $\operatorname{vec}(\mathbf{N})$ and the error resulted from the model mismatch, i.e., the ignored term $\left(\mathbf{X}^T \otimes \sqrt{1/(\kappa+1)} \widetilde{\mathbf{H}}_{RB} \mathbf{A}_R\right) \operatorname{vec}(\mathbf{G})$. Essentially, the latter term strongly correlates with the variable to be estimated in (39) (i.e., $\mathbf{G}$). When the model mismatch error dominates the effective noise in the low noise power regime, the correlation issue compromises the convergence of GAMP. On the contrary, the proposed method avoids this problem since the estimate of $\widetilde{\mathbf{H}}_{RB}$, or equivalently $\mathbf{S}$, are updated iteratively during the message passing iteration.

Next, we study the effect of the grid lengths M', L_1', and L_2'. We use η to represent the ratio between the grid length and the number of antennas, i.e., $M'/M = L_1'/L_1 = L_2'/L_2 = \eta$. Fig. 9 plots the NMSEs of the channel estimation algorithms under various η, where τ_N is fixed to -95 dB. We have the following observations: 1) the NMSEs of the proposed algorithm, concatenate LR and the oracle estimator decrease as η increases, since increasing the sampling grid length leads to a higher angle resolution and hence sparser $\mathbf{S}$ and $\mathbf{G}$; 2) the baselines [15], [18], [20] do not exploit the channel sparsity in the angular domain and its performance is invariant to η; 3) the proposed method substantially improves the estimation performance compared to the other algorithms and closely approaches the oracle bound.

In Fig. 10, we study the impact of the Rician factor κ on the channel estimation performance. As κ increases,
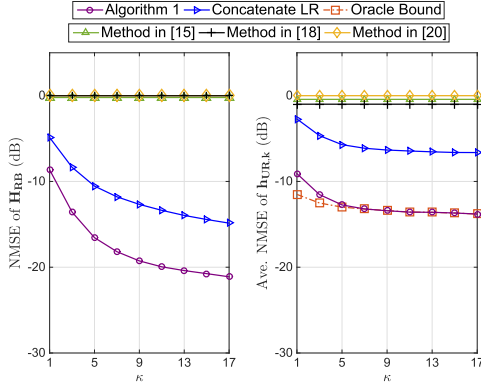
Fig. 10.   The NMSE performance versus Rician factor κ with $\tau_N = -95$ dB and $\eta = 2$.
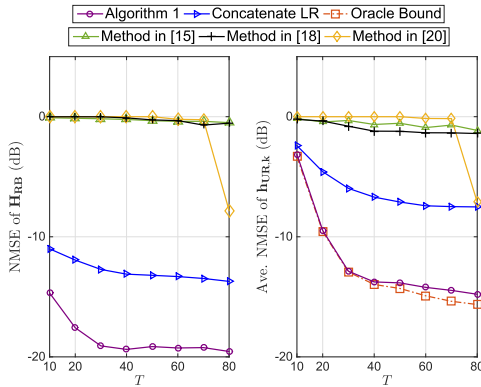

Fig. 11.   The NMSE performance versus the training length T with $\tau_N = -95$ dB.

the portion of the known information in $\mathbf{H}_{RB}$ increases, and hence the NMSEs of the proposed algorithm decrease. Moreover, similarly to Fig. 9, we find the proposed method generally achieves better performance compared to the other channel estimation algorithms.

Finally, Fig. 11 plots the NMSEs versus the training length T. We conclude that 1) increasing T from a small value leads to more measurements in $\mathbf{Y}$, and hence provides sharp improvements to the proposed algorithm; 2) when T is very large, increasing T only provides more redundant measurements and hence slightly improves the estimation performance; 3) the NMSEs for the baselines [15], [18] are only slightly improved with a larger T, as they suffer from large estimation errors; 4) the baseline [20] works with reasonably good performance when $T > 70$, which is much larger than the required T of the proposed algorithm ($T \approx 35$).

VII. CONCLUSION

In this paper, we studied the channel estimation problem in the RIS-assisted multiuser MIMO system. We formulated the cascaded channel estimation task as a matrix-calibration based sparse matrix factorization problem by exploiting the knowledge of the slow-varying channel components and the hidden channel sparsity in the angular domain. Then, we proposed a novel message-passing based algorithm to infer the

cascaded BS-to-RIS and RIS-to-user channels. Furthermore, we developed a framework to analyze the performance of the considered system. Finally, we used numerical results to confirm the performance improvement of the proposed algorithm compared to the state-of-the-art approach.

APPENDIX A
PROOF OF PROPOSITION 1

Subtracting the term $\sum_{k=1}^{K} \mathbf{h}_{UB,k} x_{kt}$ in both sides of (1) and letting $\mathbf{y}(t) \triangleq \mathbf{y}_0(t) - \sum_k \mathbf{h}_{UB,k} x_{kt}$, we obtain

$$\mathbf{y}(t) = \sum_{k=1}^{K} \mathbf{H}_{RB} \mathbf{h}_{UR,k} x_{kt} + \mathbf{n}(t), \quad 1 \le t \le T. \quad (41)$$

With the definitions of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{N}$, we rewrite (41) as

$$\mathbf{Y} = \mathbf{H}_{RB} \mathbf{H}_{UR} \mathbf{X} + \mathbf{N}, \quad (42)$$

where $\mathbf{H}_{UR} = [\mathbf{h}_{UR,1}, \cdots, \mathbf{h}_{UR,K}]$. Plugging (2), (8), and (9) into (42), we obtain the system model as

$$\mathbf{Y} = \left(\sqrt{\frac{\kappa}{\kappa+1}} \bar{\mathbf{H}}_{RB} + \mathbf{A}_B \mathbf{S} \mathbf{A}_R^H \right) \mathbf{A}_R \mathbf{G} \mathbf{X} + \mathbf{N}$$
$$= (\mathbf{H}_0 + \mathbf{A}_B \mathbf{S} \mathbf{R}) \mathbf{G} \mathbf{X} + \mathbf{N}. \quad (43)$$

APPENDIX B

1) Approximation for (18)–(20)*:* We apply the Fourier inversion theorem to rewrite (20) as

$$\mathcal{P}_{z_{mk}}^i(z_{mk}) \propto \int dt e^{jt z_{mk}} \prod_{l=1}^{L'} (e^{-jt w_{ml} g_{lk}} \Delta_{w_{ml} \to z w g_{mk}}^i(w_{ml})$$
$$\times \Delta_{g_{lk} \to z w g_{mk}}^i(g_{lk}) dw_{ml} dg_{lk}). \quad (44)$$

Following the steps in [29, eqs. (51)–(53)], we expand the exponential in the bracket in (44) to the second order and simplify (44) as $\mathcal{P}_{z_{mk}}^i(z_{mk}) \approx \mathcal{CN}(z_{mk}; \hat{p}_{mk}(i), v_{mk}^p(i))$, where

$$v_{mk}^p(i) = \sum_{l=1}^{L'} \left(|\hat{w}_{ml,k}(i)|^2 v_{lk,m}^g(i) \right.$$
$$\left. + v_{ml,k}^w(i) |\hat{g}_{lk,m}(i)|^2 + v_{ml,k}^w(i) v_{lk,m}^g(i) \right), \quad (45a)$$

$$\hat{p}_{mk}(i) = \sum_{l=1}^{L'} \hat{w}_{ml,k}(i) \hat{g}_{lk,m}(i). \quad (45b)$$

Then, we approximate $q_{mt,k} \triangleq \sum_{j \ne k} z_{mj} x_{jt} \sim \prod_{j \ne k} \Delta_{z_{mj} \to q z_{mt}}^i$ by the CLT as a Gaussian random variable with mean $\hat{q}_{mt}(i) - \hat{z}_{mk,t}(i) x_{kt}$ and variance $v_{mt}^q(i) - v_{mk,t}^z(i)|x_{kt}|^2$, where $\hat{q}_{mt}(i) \triangleq \sum_{k=1}^{K} \hat{z}_{mk,t}(i) x_{kt}$ and $v_{mt}^q(i) \triangleq \sum_{k=1}^{K} v_{mk,t}^z(i)|x_{kt}|^2$. Therefore,

$$\Delta_{q z_{mt} \to z_{mk}}^i(z_{mk})$$
$$\propto \int dq_{mt,k} \mathcal{CN}(y_{mt}; q_{mt,k} + z_{mk} x_{kt}, \tau_N)$$
$$\times \mathcal{CN}\left(q_{mt,k}; \hat{q}_{mt}(i) - \hat{z}_{mk,t}(i) x_{kt}, v_{mt}^q(i) - v_{mk,t}^z(i)|x_{kt}|^2\right)$$
$$= \mathcal{CN}(y_{mt}; \hat{q}_{mt}(i) + (z_{mk} - \hat{z}_{mk,t}(i)) x_{kt},$$
$$\tau_N + v_{mt}^q(i) + (|z_{mk}|^2 - v_{mk,t}^z(i))|x_{kt}|^2). \quad (46)$$

Following the steps in [32, eqs. (A.6)–(A.16)], we have

$$\Delta_{qz_{mt} \to z_{mk}}^{i+1}(z_{mk}) = \mathcal{CN}\left(z_{mk}; \hat{z}_{mk,t}(i+1), v_{mk,t}^z(i+1)\right), \tag{47a}$$

$$\Delta_{z_{mk}}^{i+1}(z_{mk}) = \mathcal{CN}\left(z_{mk}; \hat{z}_{mk}(i+1), v_{mk}^z(i+1)\right), \tag{47b}$$

$$\prod_{t=1}^{T} \Delta_{qz_{mt} \to z_{mk}}^{i}(z_{mk}) = \mathcal{CN}(z_{mk}; \hat{e}_{mk}(i), v_{mk}^e(i)), \tag{47c}$$

where

$$\hat{z}_{mk,t}(i+1) \approx \hat{z}_{mk}(i+1) - v_{mk}^z(i+1)x_{kt}\hat{\gamma}_{mt}(i), \tag{48a}$$

$$v_{mk,t}^z(i+1) \approx v_{mk}^z(i+1), \tag{48b}$$

$$v_{mk}^z(i+1) = \frac{v_{mk}^p(i)v_{mk}^e(i)}{v_{mk}^p(i) + v_{mk}^e(i)}, \tag{48c}$$

$$\hat{z}_{mk}(i+1) = \frac{v_{mk}^p(i)\hat{e}_{mk}(i) + \hat{p}_{mk}(i)v_{mk}^e(i)}{v_{mk}^p(i) + v_{mk}^e(i)}. \tag{48d}$$

In the above, we introduce the auxiliary variables as

$$v_{mt}^\gamma(i) = \left(v_{mt}^\beta(i) + \tau_N\right)^{-1}, \tag{49a}$$

$$\hat{\gamma}_{mt}(i) = v_{mt}^\gamma(i)\left(y_{mt} - \hat{\beta}_{mt}(i)\right), \tag{49b}$$

$$v_{mk}^e(i) = \left(\sum_{t=1}^{T} v_{mt}^\gamma(i)|x_{kt}|^2\right)^{-1}, \tag{49c}$$

$$\hat{e}_{mk}(i) = \hat{z}_{mk}(i) + v_{mk}^e(i)\sum_{t=1}^{T} x_{kt}^\star \hat{\gamma}_{mt}(i), \tag{49d}$$

where

$$v_{mt}^\beta(i) = \sum_{k=1}^{K} v_{mk}^z(i)|x_{kt}|^2, \tag{50a}$$

$$\hat{\beta}_{mt}(i) = \sum_{k=1}^{K} \hat{z}_{mk}(i)x_{kt} - v_{mt}^\beta(i)\hat{\gamma}_{mt}(i-1). \tag{50b}$$

2) Approximation for (21)–(24) *and* (30)*:* Plugging (47c) into (21) and (23), we have

$$\int \mathrm{d}z_{mk} \prod_{t=1}^{T} \Delta_{qz_{mt} \to z_{mk}}^{i}(z_{mk})p(z_{mk}|w_{ml}, g_{lk}, \forall l)$$

$$= \mathcal{CN}\left(\sum_{l=1}^{L'} w_{ml}g_{lk}; \hat{e}_{mk}(i), v_{mk}^e(i)\right). \tag{51}$$

With (51), the forms of (21) and (23) match eq. (13) of [23]. Following the steps in [23, Sec. II-D–Sec. II-E], we obtain

$$\Delta_{g_{lk}}^{i+1}(g_{lk}) \approx p(g_{lk})\mathcal{CN}(g_{lk}, \hat{b}_{lk}(i), v_{lk}^b(i)), \tag{52a}$$

$$\prod_{k=1}^{K} \Delta_{zwg_{mk} \to w_{ml}}^{i}(w_{ml}) = \mathcal{CN}(w_{ml}, \hat{c}_{ml}(i), v_{ml}^c(i)). \tag{52b}$$

In (52), we introduce

$$v_{lk}^b(i) = \left(\sum_{m=1}^{M} |\hat{w}_{ml}(i)|^2 v_{mk}^\xi(i)\right)^{-1}, \tag{53a}$$

$$\hat{b}_{lk}(i) = \left(1 - v_{lk}^b(i)\sum_{m=1}^{M} v_{ml}^w(i)v_{mk}^\xi(i)\right)\hat{g}_{lk}(i)$$

$$+ v_{lk}^b(i)\sum_{m=1}^{M} \hat{w}_{ml}^\star(i)\hat{\xi}_{mk}(i), \tag{53b}$$

$$v_{ml}^c(i) = \left(\sum_{k=1}^{K} |\hat{g}_{lk}(i)|^2 v_{mk}^\xi(i)\right)^{-1}, \tag{53c}$$

$$\hat{c}_{ml}(i) = \left(1 - v_{ml}^c(i)\sum_{k=1}^{K} v_{lk}^g(i)v_{mk}^\xi(i)\right)\hat{w}_{ml}(i)$$

$$+ v_{ml}^c(i)\sum_{k=1}^{K} \hat{g}_{lk}^\star(i)\hat{\xi}_{mk}(i), \tag{53d}$$

where

$$v_{mk}^\xi(i) = \frac{v_{mk}^p(i) - v_{mk}^z(i)}{(v_{mk}^p(i))^2}, \tag{54a}$$

$$\hat{\xi}_{mk}(i) = \frac{\hat{z}_{mt}(i) - \hat{p}_{mt}(i)}{v_{mk}^p(i)}. \tag{54b}$$

Plugging (13) into (52a), we compute the mean and variance of $\Delta_{g_{lk}}^{i+1}$ as

$$\hat{g}_{lk}(i+1) = \int g_{lk}\Delta_{g_{lk}}^{i+1}(g_{lk})\mathrm{d}g_{lk}, \tag{55a}$$

$$v_{lk}^g(i+1) = \int g_{lk}^2 \Delta_{g_{lk}}^{i+1}(g_{lk})\mathrm{d}g_{lk} - |\hat{g}_{lk}(i+1)|^2. \tag{55b}$$

Furthermore, following the steps in [23, Sec. II-F], we compute (45) as

$$v_{mk}^p(i) = \sum_{l=1}^{L'} \left(|\hat{w}_{ml}(i)|^2 v_{lk}^g(i) + v_{ml}^w(i)|\hat{g}_{lk}(i)|^2\right.$$

$$\left. + v_{ml}^w(i)v_{lk}^g(i)\right), \tag{56a}$$

$$\hat{p}_{mk}(i) = \sum_{l=1}^{L'} \hat{w}_{ml}(i)\hat{g}_{lk}(i) - \hat{\xi}_{mk}(i-1)$$

$$\times \sum_{l=1}^{L'} \left(|\hat{w}_{ml}(i)|^2 v_{lk}^g(i) + v_{ml}^w(i)|\hat{g}_{lk}(i)|^2\right). \tag{56b}$$

3) Approximation for (25)–(29) *and* (31)*:* From (52b), we obtain

$$\int \mathrm{d}w_{ml} \prod_{k=1}^{K} \Delta_{zwg_{mk} \to w_{ml}}^{i}(w_{ml})p(w_{ml}|s_{m'l'}, \forall m', l')$$

$$= \mathcal{CN}\left(\hat{c}_{ml}(i); h_{0,ml} + \sum_{m'=1}^{M'}\sum_{l'=1}^{L'} a_{B,mm'}s_{m'l'}r_{l'l}, v_{ml}^c(i)\right). \tag{57}$$

Plugging (57) into (26), we obtain

$$\Delta_{ws_{ml} \to s_{m'l'}}^{i}(s_{m'l'})$$

$$\propto \int \prod_{(j,n)\neq(m',l')} \left(\Delta_{s_{jn} \to ws_{ml}}^{i}(s_{jn})\mathrm{d}s_{jn}\right)$$

$$\times \mathcal{CN}\left(\hat{c}_{ml}(i); h_{0,ml} + \sum_{m',l'} a_{B,mm'}s_{m'l'}r_{l'l}, v_{ml}^c(i)\right). \tag{58}$$

Define $\mu_{ml,m'l'} \triangleq \sum_{(j,n)\neq(m',l')} a_{B,mj} s_{jn} r_{nl}$. By the CLT, we obtain that $\mu_{ml,m'l'}$ is Gaussian distributed with mean $\hat{\mu}_{ml}(i) - a_{B,mm'}\hat{s}_{m'l',ml}(i)r_{l'l}$ and variance $v_{ml}^{\mu}(i) - a_{B,mm'}v_{m'l',ml}^{s}(i)r_{l'l}$, where

$$\hat{\mu}_{ml}(i) = \sum_{m'=1}^{M'}\sum_{l'=1}^{L'} a_{B,mm'}\hat{s}_{m'l',ml}(i)r_{l'l}, \tag{59a}$$

$$v_{ml}^{\mu}(i) = \sum_{m'=1}^{M'}\sum_{l'=1}^{L'} |a_{B,mm'}|^2 v_{m'l',ml}^{s}(i)|r_{l'l}|^2. \tag{59b}$$

From (59) and following the steps in [32, eqs. (A.6)–(A.16)], we obtain

$$\prod_{m',l'} \Delta_{s_{m'l'}\to ws_{ml}}^{i}(s_{m'l'}) = \mathcal{CN}\left(\mu_{ml}; \hat{\mu}_{ml}(i), v_{ml}^{\mu}(i)\right), \tag{60a}$$

$$\prod_{m',l'} \Delta_{ws_{ml}\to s_{m'l'}}^{i}(s_{m'l'}) = \mathcal{CN}\left(s_{m'l'}; \hat{d}_{m'l'}(i), v_{m'l'}^{d}(i)\right), \tag{60b}$$

where

$$v_{ml}^{\mu}(i) = \sum_{m'=1}^{M'}\sum_{l'=1}^{L'} |a_{B,mm'}|^2 v_{m'l'}^{s}(i)|r_{l'l}|^2, \tag{61a}$$

$$\hat{\mu}_{ml}(i) = \sum_{m'=1}^{M'}\sum_{l'=1}^{L'} a_{B,mm'}\hat{s}_{m'l'}(i)r_{l'l} - v_{ml}^{\mu}(i)\hat{\alpha}_{ml}(i-1), \tag{61b}$$

$$v_{m'l'}^{d}(i) = \left(\sum_{m=1}^{M}\sum_{l=1}^{L'} |a_{B,mm'}|^2 v_{ml}^{\alpha}(i)|r_{l'l}|^2\right)^{-1}, \tag{61c}$$

$$\hat{d}_{m'l'}(i) = \hat{s}_{m'l'} + v_{m'l'}^{d}(i)\sum_{m=1}^{M}\sum_{l=1}^{L'} a_{B,mm'}^{\star}\hat{\alpha}_{ml}(i)r_{l'l}^{\star}. \tag{61d}$$

In (61), we define

$$v_{ml}^{\alpha}(i) = \left(v_{ml}^{\mu}(i) + v_{ml}^{c}(i)\right)^{-1}, \tag{62a}$$

$$\hat{\alpha}_{ml}(i) = v_{ml}^{\alpha}(i)\left(\hat{c}_{ml}(i) - h_{0,ml} - \hat{\mu}_{ml}(i)\right). \tag{62b}$$

Plugging (60a) into (25) and combining (52b), we obtain

$$v_{ml}^{w}(i+1) = \frac{v_{ml}^{\mu}(i)v_{ml}^{c}(i)}{v_{ml}^{\mu}(i) + v_{ml}^{c}(i)}, \tag{63a}$$

$$\hat{w}_{ml}(i+1) = \frac{v_{ml}^{\mu}(i)\hat{c}_{ml}(i) + v_{ml}^{c}(i)\hat{\mu}_{ml}(i) + v_{ml}^{c}(i)h_{0,ml}}{v_{ml}^{\mu}(i) + v_{ml}^{c}(i)}. \tag{63b}$$

Similarly to (55), by plugging (60b) into (31), we obtain $\Delta_{s_{m'l'}}^{i+1}(s_{m'l'}) \propto p(s_{m'l'})\mathcal{CN}\left(s_{m'l'}; \hat{d}_{m'l'}, v_{m'l'}^{d}\right)$ and

$$\hat{s}_{m'l'}(i+1) = \int s_{m'l'}\Delta_{s_{m'l'}}^{i+1}(s_{m'l'})\mathrm{d}s_{m'l'}, \tag{64a}$$

$$v_{m'l'}^{s}(i+1) = \int s_{m'l'}^{2}\Delta_{s_{m'l'}}^{i+1}(s_{m'l'})\mathrm{d}s_{m'l'} - |\hat{s}_{m'l'}(i+1)|^2. \tag{64b}$$

APPENDIX C

First, note that the replicate partition function $\mathbb{E}_{\mathbf{Y}}[p^n(\mathbf{Y})]$ is given by

$$\mathbb{E}_{\mathbf{Y}}[p^n(\mathbf{Y})] = \mathbb{E}_{\boldsymbol{\mathcal{S}},\boldsymbol{\mathcal{G}},\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{Q}}}\left[\int \mathrm{d}\mathbf{Y}\prod_{a=0}^{n} p\left(\mathbf{Y}|\mathbf{Q}^{(a)}\right)\right], \tag{65}$$

where $x^{(a)}, 0 \le a \le n$ is the a-th replica of x and follows the same distribution as x; $\boldsymbol{\mathcal{S}} \triangleq \{\mathbf{S}^{(a)} : \forall a\}$; $\boldsymbol{\mathcal{G}} \triangleq \{\mathbf{G}^{(a)} : \forall a\}$; $\boldsymbol{\mathcal{W}} \triangleq \{\mathbf{W}^{(a)} : \forall a\}$; $\boldsymbol{\mathcal{Z}} \triangleq \{\mathbf{Z}^{(a)} : \forall a\}$; and $\boldsymbol{\mathcal{Q}} \triangleq \{\mathbf{Q}^{(a)} : \forall a\}$.

Following [29], we define the $(n+1) \times (n+1)$ covariance matrix of $\boldsymbol{\mathcal{S}}$ as $\mathbf{C}_S \triangleq [C_S^{ab}]_{0\le a,b,\le n}$, where $C_S^{ab} = \sum_{m',l'} (s_{m'l'}^{(a)})^{\star} s_{m'l'}^{(b)}/(M'L')$. Similarly, we define $\mathbf{C}_G = [C_G^{ab}]$, $\mathbf{C}_W = [C_W^{ab}] = [M'C_S^{ab}/M + \tau_{H_0}]$, and $\mathbf{C}_Z = [C_Z^{ab}] = [L'C_W^{ab}C_G^{ab}]$ as the covariance matrices of $\boldsymbol{\mathcal{G}}$, $\boldsymbol{\mathcal{W}}$, and $\boldsymbol{\mathcal{Z}}$, respectively.

Following [29, eqs. (219)–(221)], we obtain

$$\mathbb{E}_{\mathbf{Y}}[p^n(\mathbf{Y})] = \int \mathbb{E}_{\boldsymbol{\mathcal{Q}}|\mathbf{C}_Z}\left[\int \mathrm{d}\mathbf{Y}\prod_{a=0}^{n} p(\mathbf{Y}|\mathbf{Q}^{(a)})\right]\mathrm{d}\mu_S(\mathbf{C}_S) \times \mathrm{d}\mu_G(\mathbf{C}_G)\mathrm{d}\mu_W(\mathbf{C}_W)\mathrm{d}\mu_Z(\mathbf{C}_Z), \tag{66}$$

where

$$\mu_S(\mathbf{C}_S) = \mathbb{E}_{\boldsymbol{\mathcal{S}}}\left[\prod_{0\le a\le b}^{n} \delta\left(\sum_{m',l'} \left(s_{m'l'}^{(a)}\right)^{\star} s_{m'l'}^{(b)} - M'L'C_S^{ab}\right)\right], \tag{67a}$$

$$\mu_G(\mathbf{C}_G) = \mathbb{E}_{\boldsymbol{\mathcal{G}}}\left[\prod_{0\le a\le b}^{n} \delta\left(\sum_{l,k} \left(g_{lk}^{(a)}\right)^{\star} g_{lk}^{(b)} - L'KC_G^{ab}\right)\right], \tag{67b}$$

$$\mu_W(\mathbf{C}_W) = \mathbb{E}_{\boldsymbol{\mathcal{W}}}\left[\prod_{0\le a\le b}^{n} \delta\left(\sum_{m,l} \left(w_{ml}^{(a)}\right)^{\star} w_{ml}^{(b)} - ML'C_W^{ab}\right)\right], \tag{67c}$$

$$\mu_Z(\mathbf{C}_Z) = \mathbb{E}_{\boldsymbol{\mathcal{Z}}}\left[\prod_{0\le a\le b}^{n} \delta\left(\sum_{m,k} \left(z_{mk}^{(a)}\right)^{\star} z_{mk}^{(b)} - MKC_Z^{ab}\right)\right]. \tag{67d}$$

By introducing auxiliary matrices $\widetilde{\mathbf{C}}_o = [\widetilde{C}_o^{ab}], o \in \{S, G, W, Z\}$ and applying the saddle point method [29, eq. (223)], we obtain

$$\frac{1}{K^2}\ln \mathbb{E}_{\mathbf{Y}}[p^n(\mathbf{Y})] = \mathrm{extr}_{\mathbf{C}_o,\widetilde{\mathbf{C}}_o, o\in\{S,G,W,Z\}} \mathcal{F}_n, \tag{68}$$

with

$$\mathcal{F}_n \triangleq \frac{1}{K^2}(\mathcal{I}_Q + \mathcal{I}_Z + \mathcal{I}_W + \mathcal{I}_G + \mathcal{I}_S), \tag{69}$$

where extr is the operation of extremization,

$$\mathcal{I}_Q = MT \cdot \ln \mathbb{E}_{\{q^{(a)}\}|\mathbf{C}_Z}\left[\int \mathrm{d}y \prod_{a=0}^{n} \mathcal{CN}\left(y; q^{(a)}, \tau_N\right)\right], \tag{70a}$$

$$\mathcal{I}_Z = \ln \mathbb{E}_{\boldsymbol{\mathcal{Z}}}\left[e^{\mathrm{tr}(\widetilde{\mathbf{C}}_Z(\mathbf{Z}')^H\mathbf{Z}')}\right] - MK\,\mathrm{tr}(\widetilde{\mathbf{C}}_Z\mathbf{C}_Z), \tag{70b}$$

$$\mathcal{I}_W = \ln \mathbb{E}_{\boldsymbol{\mathcal{W}}} \left[e^{\mathrm{tr}(\widetilde{\mathbf{C}}_W (\mathbf{W}')^H \mathbf{W}')} \right] - M L' \, \mathrm{tr}(\widetilde{\mathbf{C}}_W \mathbf{C}_W), \tag{70c}$$

$$\mathcal{I}_G = \ln \mathbb{E}_{\boldsymbol{\mathcal{G}}} \left[e^{\mathrm{tr}(\widetilde{\mathbf{C}}_G (\mathbf{G}')^H \mathbf{G}')} \right] - L' K \, \mathrm{tr}(\widetilde{\mathbf{C}}_G \mathbf{C}_G), \tag{70d}$$

$$\mathcal{I}_S = \ln \mathbb{E}_{\boldsymbol{\mathcal{S}}} \left[e^{\mathrm{tr}(\widetilde{\mathbf{C}}_S (\mathbf{S}')^H \mathbf{S}')} \right] - M' L' \, \mathrm{tr}(\widetilde{\mathbf{C}}_S \mathbf{C}_S). \tag{70e}$$

In (70), we define $\mathbf{Z}' = [z_{mk}^{(a)}] \in \mathbb{C}^{MK \times (n+1)}$, $\mathbf{W}' = [w_{ml}^{(a)}] \in \mathbb{C}^{ML' \times (n+1)}$, $\mathbf{G}' = [g_{lk}^{(a)}] \in \mathbb{C}^{L'K \times (n+1)}$, and $\mathbf{S}' = [s_{m'l'}^{(a)}] \in \mathbb{C}^{M'L' \times (n+1)}$ as the matrix representations of $\mathcal{Z}, \mathcal{W}, \mathcal{G}$, and $\mathcal{S}$, respectively.

Plugging (68) into (33), we have

$$\mathcal{F} = \lim_{n \to 0} \frac{\partial}{\partial n} \, \mathrm{extr}_{\mathbf{C}_o, \widetilde{\mathbf{C}}_o} \, \mathcal{F}_n. \tag{71}$$

To obtain an explicit expression of (71), we further impose the restrictions on $\mathbf{C}_o, \widetilde{\mathbf{C}}_o, o \in \{S, G, W, Z\}$ following the replica symmetry ansatz [29], as $\mathbf{C}_o = (Q_o + m_o)\mathbf{I} - m_o \mathbf{1}\mathbf{1}^T$ and $\widetilde{\mathbf{C}}_o = (\widetilde{Q}_o + \widetilde{m}_o)\mathbf{I} - \widetilde{m}_o \mathbf{1}\mathbf{1}^T$. Plugging $\mathbf{C}_o$ and $\widetilde{\mathbf{C}}_o$ into (71), we obtain

$$\mathcal{F} = \lim_{n \to 0} \frac{\partial}{\partial n} \, \mathrm{extr}_{\{Q_o, m_o, \widetilde{Q}_o, \widetilde{m}_o\}} \, \mathcal{F}_n. \tag{72}$$

Furthermore, we approximate $w_{ml}^{(a)} = h_{0,ml} + \sum_{m',l'} a_{B,mm'} s_{m'l'}^{(a)} r_{l'l}$, $z_{mk}^{(a)} = \sum_l w_{ml}^{(a)} g_{lk}^{(a)}$, and $q_{mt}^{(a)} = \sum_k z_{mk}^{(a)} x_{kt}$ as Gaussian random variables by applying the CLT under the large-system limit. Specifically, we have

$$w_{ml}^{(a)} = \sqrt{Q_W - \tau_{H_0} - \frac{M' m_S}{M}} u_W^{(a)} + \sqrt{\tau_{H_0} + \frac{M' m_S}{M}} v_W, \tag{73a}$$

$$z_{mk}^{(a)} = \sqrt{Q_Z - L' m_W m_G} u_Z^{(a)} + \sqrt{L' m_W m_G} v_Z, \tag{73b}$$

$$q_{mt}^{(a)} = \sqrt{K \tau_X (Q_Z - m_Z)} u_Q^{(a)} + \sqrt{K \tau_X m_Z} v_Q, \tag{73c}$$

where $u_W^{(a)}, u_Z^{(a)}, u_Q^{(a)}, v_W, v_Z, v_Q \sim \mathcal{CN}(\cdot; 0, 1)$.

Substituting (73) into (72) and following [31, eqs. (75)–(80)], we compute the explicit expression of $\mathcal{F}_n$. By noting $\lim_{n \to 0} \mathbb{E}_{\mathbf{Y}}[p^n(\mathbf{Y})] = 1$, we obtain (34) and $\widetilde{Q}_o = 0$. Moreover, setting the derivative of $\mathcal{F}_n$ with respect to m_o and $\widetilde{m}_o$, we obtain (37a)–(37f), $m_G = \mathbb{E}_{G,Y_G}[|\hat{G}|^2]$ and $m_S = \mathbb{E}_{S,Y_S}[|\hat{S}|^2]$, where $\hat{S}$ and $\hat{G}$ are defined in (36). Substituting $p(G)$, $p(S)$, $p(Y_G|G)$, and $p(S|Y_S)$ from (35), we obtain (37g) and (37h). Finally, aligned with the argument in [29, eq. (241)], we have (37i) and (37j).

As a result, taking the partial derivative of $\mathcal{F}_n$ with respect to n and letting $n \to 0$, we obtain

$$\mathcal{F} = \frac{2MT}{K^2} \mathcal{J}_Q + \frac{M}{K}(\mathcal{J}_Z - \widetilde{m}_Z m_Z) + \frac{ML'}{K^2}(\mathcal{J}_W - \widetilde{m}_W m_W)$$
$$+ \frac{L'}{K}(\mathcal{J}_G - \widetilde{m}_G m_G) + \frac{M'L'}{K^2}(\mathcal{J}_S - \widetilde{m}_S m_S), \tag{74}$$

where

$$\mathcal{J}_Q = \ln \mathbb{E}_{v,u',y} \left[\ln \mathbb{E}_u \left[\mathcal{N} \left(y; \sqrt{c_1} u + \sqrt{c_2} v, \tau_N \right) \right] \right], \tag{75a}$$

with $v, u', u \sim \mathcal{N}(\cdot; 0, 1)$; $c_1 \triangleq K \tau_X (Q_Z - m_Z)$; $c_2 \triangleq K \tau_X m_Z$; and $y \sim \mathcal{N}(y; \sqrt{c_1} u' + \sqrt{c_2} v, \tau_N)$;

$$\mathcal{J}_Z = \mathbb{E}_{Y_Z} \left[\ln \mathbb{E}_{Z^{(1)}} \left[e^{-\widetilde{m}_Z |Z^{(1)}|^2 + 2\sqrt{\widetilde{m}_Z} \, \mathrm{Re}(Y_Z^\star Z^{(1)})} \right] \right], \tag{75b}$$

with $Y_Z \sim \mathcal{CN}(Y_Z; \sqrt{\widetilde{m}_Z} Z^{(0)}, 1)$ and $Z^{(0)}, Z^{(1)}$ defined as in (73b);

$$\mathcal{J}_W = \mathbb{E}_{Y_W} \left[\ln \mathbb{E}_{W^{(1)}} \left[e^{-\widetilde{m}_W |W^{(1)}|^2 + 2\sqrt{\widetilde{m}_W} \, \mathrm{Re}(Y_W^\star W^{(1)})} \right] \right], \tag{75c}$$

with $Y_W \sim \mathcal{CN}(Y_W; \sqrt{\widetilde{m}_W} W^{(0)}, 1)$ and $W^{(0)}, W^{(1)}$ defined as in (73a);

$$\mathcal{J}_G = \mathbb{E}_{Y_G} \left[\ln \mathbb{E}_{G^{(1)}} \left[e^{-\widetilde{m}_G |G^{(1)}|^2 + 2\sqrt{\widetilde{m}_G} \, \mathrm{Re}(Y_G^\star G^{(1)})} \right] \right], \tag{75d}$$

$$\mathcal{J}_S = \mathbb{E}_{Y_S} \left[\ln \mathbb{E}_{S^{(1)}} \left[e^{-\widetilde{m}_S |S^{(1)}|^2 + 2\sqrt{\widetilde{m}_S} \, \mathrm{Re}(Y_S^\star S^{(1)})} \right] \right], \tag{75e}$$

with Y_G, Y_S defined in (35); $G^{(1)} \sim p(G)$; and $S^{(1)} \sim p(S)$.

Finally, by evaluating the stationary points of (74), we obtain the MSEs defined in (36) as the asymptotic MMSEs.

REFERENCES

[1] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Message-passing based channel estimation for reconfigurable intelligent surface assisted MIMO," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, to be published.

[2] 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.

[3] C. Huang *et al.*, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," 2019, *arXiv:1911.12296*. [Online]. Available: http://arxiv.org/abs/1911.12296

[4] C. Huang, A. Zappone, M. Debbah, and C. Yuen, "Achievable rate maximization by passive intelligent mirrors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 3714–3718.

[5] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 162–169, Sep. 2018.

[6] X. Tan, Z. Sun, D. Koutsonikolas, and J. M. Jornet, "Enabling indoor mobile millimeter-wave networks based on smart reflect-arrays," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun.*, Apr. 2018, pp. 270–278.

[7] S. Nie, J. M. Jornet, and I. F. Akyildiz, "Intelligent environments based on ultra-massive MIMO platforms for wireless communication in millimeter wave and terahertz bands," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7849–7853.

[8] 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.

[9] 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.

[10] Q. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M. Alouini, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Trans. Wireless Commun.*, early access, Apr. 14, 2020, doi: 10.1109/TWC.2020.2986438.

[11] W. Yan, X. Yuan, and X. Kuai, "Passive beamforming and information transfer via large intelligent surface," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 533–537, Apr. 2020.

[12] 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.

[13] M. Jung, W. Saad, and G. Kong, "Performance analysis of large intelligent surfaces (LISs): Uplink spectral efficiency and pilot training," 2019, *arXiv:1904.00453*. [Online]. Available: http://arxiv.org/abs/1904.00453

[14] A. Taha, M. Alrabeiah, and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," 2019, *arXiv:1904.10136*. [Online]. Available: http://arxiv.org/abs/1904.10136

[15] Q.-U.-U. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Intelligent reflecting surface assisted wireless communication: Modeling and channel estimation," 2019, *arXiv:1906.02360*. [Online]. Available: http://arxiv.org/abs/1906.02360

[16] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 4659–4663.

[17] 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.

[18] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications," 2019, *arXiv:1911.03084*. [Online]. Available: http://arxiv.org/abs/1911.03084

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

[20] L. Wei, C. Huang, G. C. Alexandropoulos, and C. Yuen, "Parallel factor decomposition channel estimation in RIS-assisted multi-user MISO communication," 2020, *arXiv:2001.09413*. [Online]. Available: http://arxiv.org/abs/2001.09413

[21] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.

[22] S. Rangan, P. Schniter, E. Riegler, A. K. Fletcher, and V. Cevher, "Fixed points of generalized approximate message passing with arbitrary matrices," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 7464–7474, Dec. 2016.

[23] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part I: Derivation," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014.

[24] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (International Series of Monographs on Physics (111)). Oxford, U.K.: Oxford Univ. Press, 2001.

[25] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge Univ. Press, 2005.

[26] X. Li, J. Fang, H. Li, and P. Wang, "Millimeter wave channel estimation via exploiting joint sparse and low-rank structures," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1123–1133, Feb. 2018.

[27] A. F. Molisch, A. Kuchar, J. Laurila, K. Hugl, and R. Schmalenberger, "Geometry-based directional model for mobile radio channels? Principles and implementation," *Eur. Trans. Telecommun.*, vol. 14, no. 4, pp. 351–359, 2003.

[28] F. Krzakala, M. Mezard, and L. Zdeborova, "Phase diagram and approximate message passing for blind calibration and dictionary learning," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 659–663.

[29] Y. Kabashima, F. Krzakala, M. Mezard, A. Sakata, and L. Zdeborova, "Phase transitions and sample complexity in Bayes-optimal matrix factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4228–4265, Jul. 2016.

[30] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York, NY, USA: Springer-Verlag, 1994.

[31] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-optimal joint channel- and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, May 2016.

[32] J. Ma, J. Xu, and A. Maleki, "Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2-regularization," 2018, *arXiv:1801.01170*. [Online]. Available: http://arxiv.org/abs/1801.01170

Hang Liu (Graduate Student Member, IEEE) received the B.Sc. degree in mathematics and information engineering from The Chinese University of Hong Kong, Hong Kong, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include statistical signal processing and machine learning in massive MIMO and intelligent communications.
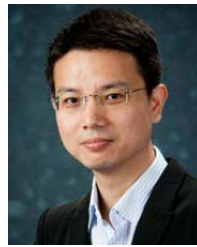
Xiaojun Yuan (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from The City University of Hong Kong in 2008.

From 2009 to 2011, he was a Research Fellow with the Department of Electronic Engineering, City University of Hong Kong. He was also a Visiting Scholar with the Department of Electrical Engineering, University of Hawaii at Manoa, from Spring 2009 to Summer 2010. From 2011 to 2014, he was a Research Assistant Professor with the Institute of Network Coding, The Chinese University of Hong Kong. From 2014 to 2017, he was an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University. He is currently a Professor with the Center for Intelligent Networking and Communications, University of Electronic Science and Technology of China. His research interests cover a broad range of signal processing, machine learning, and wireless communications, including but not limited to multi-antenna and cooperative communications, sparse and structured signal recovery, Bayesian approximate inference, network coding, and so on. He has published over 160 peer-reviewed research articles in the leading international journals and conferences in the related areas. He has served on a number of technical programs for international conferences. He has been an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS since 2017 and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2018. He was a co-recipient of the Best Paper Award of IEEE International Conference on Communications (ICC) 2014 and the Best Journal Paper Award of IEEE Technical Committee on Green Communications and Computing (TCGCC) 2017.

Ying-Jun Angela Zhang (Fellow, IEEE) is currently with the Department of Information Engineering, The Chinese University of Hong Kong. Her research interests include mainly wireless communications systems and smart power systems, in particular optimization techniques for such systems. She is a fellow of IET and a Distinguished Lecturer of the IEEE ComSoc. She was a co-recipient of the 2011 IEEE Marconi Prize Paper Award on Wireless Communications, the 2013 IEEE SmartgridComm Best Paper Award, the 2014 IEEE ComSoc APB Outstanding Paper Award, and the 2011 Young Researcher Award from The Chinese University of Hong Kong. As the only winner from engineering science, she has won the Hong Kong Young Scientist Award 2006, conferred by the Hong Kong Institution of Science. She served as the Chair for the Executive Editor Committee of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2018 to 2019. She has also served many years on the Editorial Boards for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON COMMUNICATIONS, and *Security and Communications Networks* (Wiley). She has been on the organizing committees of major IEEE conferences, including ICC, GLOBECOM, SmartgridComm, VTC, CCNC, ICCC, and MASS. She was the Founding Chair of the IEEE ComSoc Technical Committee on Smart Grid Communications.