Passive Beamforming and Information Transfer Design for Reconfigurable Intelligent Surfaces Aided Multiuser MIMO Systems

Wenjing Yan, *Graduate Student Member, IEEE*, Xiaojun Yuan, *Senior Member, IEEE*, Zhen-Qing He, *Member, IEEE*, and Xiaoyan Kuai, *Member, IEEE*

***Abstract*— This paper investigates the passive beamforming and information transfer (PBIT) technique for multiuser multiple-input multiple-output (Mu-MIMO) systems with the aid of reconfigurable intelligent surfaces (RISs), where the RISs enhance the primary communication via passive beamforming (P-BF) and at the same time deliver additional information by the on-off reflecting modulation (in which the RIS information is carried by the on/off state of each reflecting element). For the P-BF design, we propose to maximize the achievable user sum rate of the RIS-aided Mu-MIMO channel and formulate the problem as a two-step stochastic program. A sample average approximation (SAA) based iterative algorithm is developed for the efficient P-BF design of the considered scheme. To strike a balance between complexity and performance, we further propose a simplified P-BF algorithm by approximating the stochastic program as a deterministic alternating optimization problem. For the receiver design, the signal detection at the receiver is a bilinear estimation problem since the RIS information is multiplicatively modulated onto the reflected signals of the reflecting elements. To solve this bilinear estimation problem, we develop a turbo message passing (TMP) algorithm in which the factor graph associated with the problem is divided into two modules: one for the estimation of the user signals and the other for the estimation of the on-off state of each RIS element. The two modules are executed iteratively to yield a near-optimal low-complexity solution. Furthermore, we extend the design of the Mu-MIMO PBIT scheme from single-RIS to multi-RIS, by leveraging the similarity between the single-RIS and multi-RIS system models. Extensive simulation results are provided to demonstrate the advantages of our P-BF and receiver designs.**

***Index Terms*— Reconfigurable intelligent surface (RIS), intelligent reflecting surface (IRS), large intelligent surface (LIS), passive beamforming and information transfer (PBIT), two-stage stochastic programming, bilinear signal estimation, turbo message passing (TMP).**

Manuscript received October 1, 2019; revised January 15, 2020; accepted February 17, 2020. Date of publication June 15, 2020; date of current version August 20, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1801105, in part by the 111 Project under Grant B20064, and in part by the National Natural Science Foundation of China under Grant 61801084 and Grant 61801083. This article was presented in part at the IEEE International Conference on Communications (ICC), Dublin, Ireland, June 2020. *(Corresponding author: Xiaojun Yuan.)*

The authors are with the Center for Intelligent Networking and Communications, National Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: wjyan@std.uestc.edu.cn; xjyuan@uestc.edu.cn; zhenqinghe@uestc.edu.cn; xy_kuai@uestc.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2020.3000811

I. Introduction

With the rapid development of communication and computer technologies in recent years, many new applications (such as virtual reality, artificial intelligence, 3D media, ultra-high definition video, etc.) have emerged, resulting in an exploding demand of wireless services for the fifth-generation (5G) and beyond wireless networks. Although the contemporary advanced technologies, such as millimetre wave (mmWave), massive multiple-input multiple-output (MIMO), ultra-dense deployments, etc., have greatly improved the spectral efficiency of wireless networks, their extensive implementations are still constrained by the bottlenecks such as hardware cost and energy consumption [2]. Reconfigurable intelligent surface (RIS), as a very recent new hardware technology to improve the spectrum efficiency of wireless networks with minimal energy consumption, has been regarded as a promising solution to enhance the 5G and beyond wireless networks [3]–[6].

A RIS is an electromagnetic two-dimensional surface, composed of a large number of low-cost nearly-passive reconfigurable reflecting elements [7], [8]. As a prominent feature, the RIS can be flexibly implemented in practical communication scenarios (no matter the outdoor by installing it onto the facades of buildings, or the indoor by installing it onto the ceilings and walls of rooms) [9]. Equipped with a smart controller, the RIS is able to intelligently adjust the phases of incident electromagnetic waves to increase the received signal energy, expand the coverage region, and alleviate interference, so as to enhance the communication quality of the wireless networks [5], [10]. Besides, the RIS can also deliver additional information by adopting the on-off modulation on the index of the reflecting elements [4], [11].

The RIS is advantageous over the existing related technologies in many aspects, such as multi-antenna relay [12], active intelligent surface [13], and backscattering [14]. For example, compared to multi-antenna relays and active intelligent surfaces, passive RISs do not consume any energy in processing or retransmitting radio frequency signals; compared to backscatters, RISs are able to conduct passive beamforming (P-BF), i.e., to judiciously adjust the phases of the reflected electromagnetic waves to focus energy in the desired spatial directions, so as to significantly improve the energy-efficiency of wireless networks.

However, the incorporation of RISs into wireless networks poses a number of unprecedented challenges to the transceiver and RIS design [5], [6], [10]. For example, the authors in [7] studied the joint active and P-BF design to minimize the total transmit power, where both the active beamforming matrix at the transmitter and the P-BF matrix at the RIS are taken into account in the optimization. In [8], transmit power allocation and P-BF were jointly designed to maximize energy/spectral efficiency. The optimization of P-BF requires the knowledge of the channel state information (CSI) of both the transmitter-RIS link and the RIS-receiver link. The CSI acquisition in a RIS-aided wireless communication network is a particularly challenging task due to the limited processing capability of the RIS. In this regard, the authors in [15] assumed that a small portion of the RIS elements are "active" and are able to conduct baseband signal processing. Then, a channel estimation approach based on compressive sensing and deep learning was proposed for a RIS-aided multiple-input multiple-output (MIMO) channel. The channel estimation problem for a fully passive RIS was first tackled in [16], where a two-stage algorithm was developed to estimate the RIS-aided MIMO channel by utilizing sparse matrix factorization and matrix completion techniques.

Recently, a new RIS-aided wireless communication scheme, termed P-BF and information transfer (PBIT), was proposed to require that the RIS simultaneously enhances the primary communication (via P-BF) and delivers additional information to the receiver in a passive manner (by adopting the on-off reflecting modulation on the RIS elements) [11]. There are many potential sources of RIS information. For example, the wireless control link requires the RIS to acknowledge its status for synchronisation and packed delivery; the RIS installed on a smart building is required to upload environmental data to the wireless network; or the CSI measured at the RIS (e.g., following the approach in [15]) is required to be uploaded to a control center for assisting global resource allocation [17]. In the PBIT scheme, the reflecting coefficients of the RIS contain randomness since they carry the additional information delivered by the RIS, and so the P-BF design for a PBIT scheme generally involves difficult stochastic optimization. To simplify the problem, the authors in [11] proposed to maximize a heuristic performance metric, i.e., the average receive signal-to-noise ratio (SNR), and a semidefinite relaxation method was developed for the P-BF design of the single-input multiple-output (SIMO) PBIT system, where a single-antenna user communicates with a multi-antenna base station (BS) with the help of a RIS. Furthermore, the receiver of the SIMO PBIT scheme is required to reliably recover the information from both the user and the RIS, which gives rise to a bilinear detection problem. Efficient detection algorithms were developed in [11] by exploiting the rank-1 property of the received signal matrix.

In this paper, we study the design of the PBIT scheme for a multiuser MIMO (Mu-MIMO) system, where a number of single-antenna users communicate with a multi-antenna base station (BS) via the help of a RIS. Due to the multiplexing effect of the Mu-MIMO system, it is not appropriate to characterize the system performance by using a single SNR, and so the P-BF design in [11] is no longer applicable to this new scenario. A natural P-BF design criterion is to maximize the sum capacity of the channel. However, the sum capacity of a Mu-MIMO PBIT system is difficult to characterize since it is a multiplicative multiple access channel. Considering that the information rate of the RIS is typically much lower than those of the users, we ignore the RIS information rate and propose to maximize the achievable sum rate of the users. The user sum rate maximization problem, however, is still difficult to solve since it involves taking expectation over the RIS information. That is, this problem is generally a stochastic program. The state-of-the-art studies on the design of the RIS-aided wireless system mostly involve deterministic optimization problems; see, for example, the weighted sum-rate optimization for RIS-aided multiple-input single-output (MISO) system [18], the channel capacity maximization for RIS-aided MIMO system [19], and the weighted sum-rate optimization for RIS-aided multicell MIMO systems [20]. Unlike these approaches, we formulate the user sum rate maximization problem as a two-step stochastic program. A sample average approximation (SAA) based iterative algorithm is developed for the efficient P-BF design of the Mu-MIMO PBIT system. However, as a common issue of stochastic programming, the convergence speed of the SAA-based P-BF algorithm is relatively slow. To strike a balance between complexity and performance, we further propose a simplified P-BF algorithm by approximating the stochastic program as a deterministic alternating optimization problem. We prove that both the proposed P-BF algorithms are guaranteed to converge to a local minimum.

Also due to the multiplexing effect, the received signal matrix at the BS of the Mu-MIMO system is not rank-one, and therefore the rank-1 matrix factorization techniques developed in [11] are no longer applicable. The receiver in the Mu-MIMO PBIT system aims to retrieve the information from both the users and the RIS, which is a bilinear detection problem. Message passing is a powerful technique to yield near-optimal low complexity solutions to sophisticated inference problems [21]. Particularly, the parametric bilinear generalized approximate message passing (PBiGAMP) algorithm [22] is designed to handle bilinear inference problems such as the one considered in this work. However, the complexity of the PBiGAMP algorithm is quadratic to the transmission block length, which is unaffordable when the block length is large. More importantly, in simulation, we observe that the PBiGAMP algorithm performs very poor in our scheme. The reason is that the PBiGAMP algorithm is specifically designed for measurement matrices composed of independent and identically distributed (i.i.d.) Gaussian entries, whereas the measurement matrixes for our problem are low-rank matrices far from i.i.d. Gaussian. To address these two issues, we develop a turbo message passing (TMP) algorithm to retrieve the information from both the users and the RIS in Mu-MIMO PBIT system. Specifically, we represent the inference problem by a factor graph and divide the whole factor graph into two modules, namely, one for the estimation of the user signals and the other for the estimation of the RIS's on-off reflecting states (that carries the RIS information). The performance

of the proposed TMP algorithm is evaluated by numerical results.

Furthermore, we extend the design of the Mu-MIMO PBIT scheme from single-RIS to multi-RIS, where a number of RISs are installed in different locations to cooperatively enhance the user-BS communications as well as the RIS-BS information transfer. We observe the similarity between the multi-RIS system model and the single-RIS system model. Based on the model similarity, we show that the P-BF and receiver designs for the single-RIS case straightforwardly carry over to the multi-RIS case, except for some minor modifications to the prior distribution of the on-off states of the RIS elements.

The main contributions of this paper are summarized as follows.

- This is the first paper to study the design of the PBIT scheme for a RIS-aided Mu-MIMO system, which mainly involves the P-BF design and the receiver design. The P-BF design for the Mu-MIMO PBIT system is required to solve a challenging stochastic optimization problem. The receiver design in Mu-MIMO PBIT system is required to solve a bilinear detection problem.
- An SAA-based iterative algorithm is developed for the efficient P-BF of the Mu-MIMO PBIT system. A simplified P-BF algorithm is further developed to strike a balance between complexity and performance for the P-BF. The convergence of both the proposed P-BF algorithms is proved.
- A TMP algorithm is developed to retrieve the information from both the users and the RIS in the Mu-MIMO PBIT system. We show that the proposed TMP algorithm outperforms the PBiGAMP algorithm not only in the computational complexity but also in the signal detection performance. We also show that the TMP algorithm is able to closely approach the performance lower bounds (obtained by assuming either known user data or known RIS data).

Notation: For any matrix $\boldsymbol{A}$, $\boldsymbol{a}_i$ refers to the ith column of $\boldsymbol{A}$, and a_{ij} refers to the (i,j)th entry of $\boldsymbol{A}$. $\mathbb{C}$ denotes the complex field; $\mathbb{R}$ denotes the real field; $\mathcal{S}$ denotes a set, and $|\mathcal{S}|$ represents the the cardinality of $\mathcal{S}$. $|x|$ represents the absolute value of x; $\|\cdot\|_2$ represents the ℓ_2-norm; $\|\cdot\|_{\mathrm{F}}$ represents the Frobenius norm. The superscripts $(\cdot)^{\mathrm{T}}$, $(\cdot)^*$, $(\cdot)^{\mathrm{H}}$, $(\cdot)^{-1}$ represent the transpose, the conjugate, the conjugate transpose, and the inverse of a matrix, respectively. $\odot$ represents the Hadamard product; $\otimes$ represents the Kronecker product. $\mathbb{E}(\cdot)$ and $\mathbb{V}\mathrm{ar}(\cdot)$ represent the expectation and the variance, respectively. $\delta(\cdot)$ represents the Dirac delta function. $\mathrm{diag}\{\boldsymbol{A}\}$ represents the vector composed of the diagonal elements of those of $\boldsymbol{A}$. $\mathrm{diag}\{\boldsymbol{a}\}$ and $[\boldsymbol{A}]_{\mathrm{diag}}$ represent the diagonal matrixes with the diagonal specified by $\boldsymbol{a}$ and the diagonal elements of $\boldsymbol{A}$, respectively. For any integer N, $\mathcal{I}_N$ denotes the set of integers from 1 to N; $\boldsymbol{I}$ is the identity matrix with an appropriate size; $\boldsymbol{1}_N$ represents the N-dimensional all-one vector. $\mathcal{CN}(\cdot;\mu,\nu)$ represents a complex Gaussian distribution with mean μ, covariance ν, and relation zero. $\mathrm{vec}(\boldsymbol{A})$ represents the vector obtained by stacking the columns of the matrix $\boldsymbol{A}$ sequentially. $\boldsymbol{A}\succeq\boldsymbol{0}$ means that $\boldsymbol{A}$ is a semidefinite matrix.

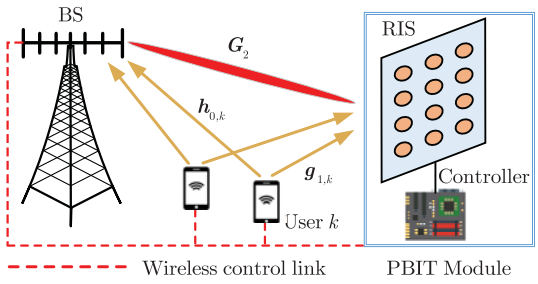


Fig. 1. A PBIT-enhanced uplink MIMO system with a RIS.

II. System Model and Problem Description

A. System Model

As illustrated in Fig. 1, we consider a PBIT-enhanced uplink Mu-MIMO system, where K single-antenna users communicate with a BS equipped with M antennas via the help of a RIS. Beyond traditional end-to-end wireless communications, the system adds an additional PBIT module to simultaneously enhance the active user-BS communication and achieve the passive RIS-BS communication. The PBIT module is an intelligent reflecting device that consists of a RIS equipped with N passive reflecting elements, and a controller to adaptively adjust the on-off states and the phase shifts of the reflecting elements. To enhance the active user-BS communication, the RIS reflects the incident signals transmitted from the users by the turn-on reflecting elements. Further, the phases of the reflected signals can be adjusted by the controller to optimize the user-BS communication performance. To achieve the passive RIS-BS communication, the controller adjusts the on-off reflecting state of each RIS element according to the RIS data. (i.e., the on-off reflecting states of the RIS elements carry the RIS information).

Denote by $\boldsymbol{H}_0=[\boldsymbol{h}_{0,1},\ldots,\boldsymbol{h}_{0,K}]\in\mathbb{C}^{M\times K}$ the baseband equivalent channel of the user-BS link, where $\boldsymbol{h}_{0,k}\in\mathbb{C}^{M\times 1}$ is the channel coefficient vector of user k. Denote by $\boldsymbol{G}_1=[\boldsymbol{g}_{1,1},\ldots,\boldsymbol{g}_{1,N}]^{\mathrm{T}}\in\mathbb{C}^{N\times K}$ the baseband equivalent channel of the user-RIS link, where $\boldsymbol{g}_{1,n}\in\mathbb{C}^{K\times 1}$ is the channel coefficient vector between all users and the nth reflecting element of the RIS. Let s_n be the state of the nth passive reflecting element, where s_n takes the value 0 or 1 to represent the "on" or "off" reflecting state of the ith element. $\boldsymbol{S}=\mathrm{diag}\{\boldsymbol{s}\}$ is the diagonal state matrix of the RIS, where $\boldsymbol{s}=[s_1,s_2,\ldots,s_N]^{\mathrm{T}}$ carries the information from the RIS. Since the RIS information is modulated by turning on/off the reflecting elements, we refer to the modulation of RIS as on-off reflecting modulation. Let φ_n be the phase shift of the nth passive reflecting element, where $\varphi_n\in[0,2\pi]$, and $\gamma\in[0,1]$ be the amplitude reflection coefficient.[1] Then $\boldsymbol{\Theta}=\mathrm{diag}\{\boldsymbol{\theta}\}$ is the diagonal phase-shift matrix of the RIS, where $\boldsymbol{\theta}=[\theta_1,\theta_2,\ldots,\theta_N]^{\mathrm{T}}$ and $\theta_n=\gamma e^{j\varphi_n}$.[2] Denote by $\boldsymbol{G}_2=[\boldsymbol{g}_{2,1},\ldots,\boldsymbol{g}_{2,N}]\in\mathbb{C}^{M\times N}$ the baseband equivalent

[1]We assume that all RIS elements share a common amplitude reflection coefficient since they have the same physical structure.

[2]In the sequel of this paper, we follow [7] to only consider the setting of $\gamma=1$. Note that the results in this paper can be readily extended to any value of γ.

channel of the RIS-BS link, where $\boldsymbol{g}_{2,n} \in \mathbb{C}^{M\times 1}$ is the channel coefficient vector between the nth reflecting element of the RIS and the BS. We neglect the signal power reflected by the RIS for two or more times due to severe path loss.

Assume that the channel is block-fading with each transmission block consisting of T time slots. Denote by x_{kt} the transmit symbol of user k in time slot t. Then, the observed signal at the BS in time slot t is

$$\boldsymbol{y}_t = (\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}\boldsymbol{G}_1 + \boldsymbol{H}_0)\boldsymbol{x}_t + \boldsymbol{w}_t, \tag{1}$$

where $\boldsymbol{x}_t = [x_{1t}, \ldots, x_{Kt}]^{\mathrm{T}} \in \mathbb{C}^{K\times 1}$ is the transmit signal vector and $\boldsymbol{w}_t \in \mathbb{C}^{M\times 1}$ is an additive white Gaussian noise (AWGN) with the elements independently drawn from $\mathcal{CN}(0, \sigma_w^2)$. Assume that the diagonal state matrix $\boldsymbol{S}$ and the diagonal phase-shift matrix $\boldsymbol{\Theta}$ of the RIS remain fixed over each transmission block of length T. Then, the observed signal matrix in a transmission block, denoted by $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$, can be expressed as

$$\boldsymbol{Y} = (\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}\boldsymbol{G}_1 + \boldsymbol{H}_0)\boldsymbol{X} + \boldsymbol{W}, \tag{2}$$

where $\boldsymbol{X} \triangleq [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$ and $\boldsymbol{W} \triangleq [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T]$.

Each entry of $\boldsymbol{X}$ is independently drawn from a distribution $p_x(x)$. Then, the probability density function (PDF) of $\boldsymbol{X}$ is given by

$$p_{\boldsymbol{X}}(\boldsymbol{X}) = \prod_{k=1}^{K}\prod_{t=1}^{T} p_x(x_{kt}). \tag{3}$$

Note that the information rate of each user is $-\mathbb{E}_{p_x}[\log p_x(x)]$. The transmitted signals of each user k in a transmission block are power-constrained as

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|x_{kt}|^2 \leqslant P_k, \quad \forall k \in \mathcal{I}_K, \tag{4}$$

where P_k is the power budget of user k.

On-off reflecting modulation is applied by manipulating the on-off state of each reflecting element. Specifically, we assume that $s_n, \forall n \in \mathcal{I}_N$, independently takes the value of 1 (meaning that the state of the nth passive element is "on") with probability ρ and the value of 0 (meaning that the state of the nth passive element is "off") with probability $1-\rho$, yielding

$$p(\boldsymbol{s}) = \prod_{n=1}^{N} p(s_n) = \prod_{n=1}^{N}\left((1-\rho)\delta(s_n) + \rho\delta(s_n - 1)\right), \tag{5}$$

where ρ is the sparsity of the RIS. Note that each $s_n, \forall n \in \mathcal{I}_N$ remains unchanged over each transmission block of length T. The total amount of information carried by the on-off states of the RIS elements in each transmission block is $H(\rho) = -N(\rho\log_2\rho + (1-\rho)\log_2(1-\rho))$ bits, and hence the RIS information rate is $H(\rho)/T$.

B. Problem Description

In the PBIT-enhanced uplink MIMO system as described above, the receiver is required to retrieve the information from both the users and the RIS (i.e., $\boldsymbol{X}$ and $\boldsymbol{s}$) under the assumption of perfect channel state information (CSI).[3] Furthermore, the phase shifts of the passive reflecting elements in the RIS need to be carefully adjusted to enhance the recovery performance at the receiver. A natural design criterion is to maximize the sum channel capacity. From information theory, the sum channel capacity of the PBIT system is given by the mutual information $I(\boldsymbol{x}_t, \boldsymbol{s}; \boldsymbol{y}_t)$[4] [23]. Then, our design problem can be divided into two subproblems: One is the P-BF design, i.e., to maximize $I(\boldsymbol{x}, \boldsymbol{s}; \boldsymbol{y})$ over the phase shift matrix $\boldsymbol{\Theta}$; and the other is the transceiver design, i.e., to design the signaling of $\boldsymbol{x}$ and $\boldsymbol{s}$ at the users and the RIS as well as the receiver at the BS to achieve the maximized $I(\boldsymbol{x}, \boldsymbol{s}; \boldsymbol{y})$.

We first consider the P-BF design to maximize $I(\boldsymbol{x}, \boldsymbol{s}; \boldsymbol{y})$ over $\boldsymbol{\Theta}$. However, $I(\boldsymbol{x}, \boldsymbol{s}; \boldsymbol{y})$ is difficult to evaluate since (1) is a complicated model involving the multiplication of $\boldsymbol{s}$ and $\boldsymbol{x}$. Note that the RIS data are from environmental monitoring or the control signaling, the required information rate of the RIS is typically much lower than that of a cellular user. As such, we propose the following design metric:

$$I(\boldsymbol{x}, \boldsymbol{s}; \boldsymbol{y}) = I(\boldsymbol{x}; \boldsymbol{y}|\boldsymbol{s}) + I(\boldsymbol{s}; \boldsymbol{y}) \approx I(\boldsymbol{x}; \boldsymbol{y}|\boldsymbol{s}), \tag{6}$$

where the first step follows from the chain rule of mutual information, and the second step is from the assumption of $I(\boldsymbol{s}; \boldsymbol{y}) \ll I(\boldsymbol{x}; \boldsymbol{y}|\boldsymbol{s})$. Then, the P-BF design problem is converted to maximizing $I(\boldsymbol{x}; \boldsymbol{y}|\boldsymbol{s})$ over $\boldsymbol{\Theta}$. Note that $I(\boldsymbol{x}; \boldsymbol{y}|\boldsymbol{s})$ is exactly the maximum achievable rate of $\boldsymbol{x}$ [23].

For the transceiver design, we will focus on the design of the receiver at the BS to reliably recover both $\boldsymbol{X}$ and $\boldsymbol{s}$ from the received signal $\boldsymbol{Y}$ (for a given $\boldsymbol{\Theta}$). From information theory, besides the receiver design, we also need to design signal shaping and channel coding at the transmitter, so as to approach the channel capacity. The signal shaping and channel coding design is, however, out of the scope of this paper.

III. Beamforming Design

We now consider the P-BF design of $\boldsymbol{\Theta}$ to maximize the conditional mutual information as

$$\min_{\boldsymbol{\Theta}} \; I(\boldsymbol{x}; \boldsymbol{y}|\boldsymbol{s}) = \mathbb{E}_{\boldsymbol{s}}\left(\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}} \log \frac{p_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{s})}{p_{\boldsymbol{x}}(\boldsymbol{x})}\right) \tag{7a}$$

$$\text{s.t. } |\theta_n| = 1, \quad \forall n \in \mathcal{I}_N \tag{7b}$$

where $p_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{s})$ is the conditional distribution of $\boldsymbol{x}$ conditioned on $\boldsymbol{y}$ and $\boldsymbol{s}$. This problem can be recast as a two-stage stochastic program:

$$\min_{\boldsymbol{\Theta}} \; \mathbb{E}_{\boldsymbol{s}}\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}) \tag{8a}$$

$$\text{s.t. } |\theta_n| = 1, \forall n \in \mathcal{I}_N, \tag{8b}$$

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}) = \min_{\boldsymbol{\Phi}, \boldsymbol{\Sigma}\succeq\boldsymbol{0}} \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}\Big(\log\det(\boldsymbol{\Sigma}) + \left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\Phi}\boldsymbol{y})\right\|_2^2\Big) \tag{8c}$$

where $\boldsymbol{\Phi} \in \mathbb{C}^{K\times M}$ and $\boldsymbol{\Sigma} \in \mathbb{C}^{K\times K}$ are auxiliary parameter matrices. Note that $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s})$ is a variational representation of

[3]The CSI acquisition techniques of the RIS-aided MIMO channel can be found, e.g., in [15] and [16].

[4]Since $I(\boldsymbol{x}_t, \boldsymbol{s}; \boldsymbol{y}_t)$ is invariant to index t, we henceforth omit the index t by simply writing $I(\boldsymbol{x}, \boldsymbol{s}; \boldsymbol{y})$.

$\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}} \log \frac{p_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})}{p_{\boldsymbol{x}}(\boldsymbol{x})}$. The equivalence between (7) and (8) can be found in Appendix A. To calculate the expectation in (8a), we need to enumerate all the 2^N scenarios of $\boldsymbol{s}$, which is difficult especially when N is large. Thus, we follow the SAA method [24] to approximate (8) by independently generating $\ell_{\boldsymbol{s}}$ replications $\{\boldsymbol{s}_{[1]}, \ldots, \boldsymbol{s}_{[\ell_{\boldsymbol{s}}]}\}$ of $\boldsymbol{s}$ based on the probability model in (5). It has been shown that the solution of the SAA approximation approaches the optimal solution of the original stochastic program with probability one for a sufficiently large sample size [25], [26]. Then, the stochastic program in (8) is approximated by

$$\min_{\boldsymbol{\Theta}} \ \frac{1}{\ell_{\boldsymbol{s}}} \sum_{i=1}^{\ell_{\boldsymbol{s}}} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}_{[i]}) \tag{9a}$$

$$\text{s.t. } |\theta_n| = 1, \forall n \in \mathcal{I}_N \tag{9b}$$

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}_{[i]}) = \min_{\boldsymbol{\Phi}, \boldsymbol{\Sigma} \succeq \boldsymbol{0}} \mathcal{J}_{[i]}(\boldsymbol{\Phi}, \boldsymbol{\Sigma}), \quad \forall i \in \mathcal{I}_{\ell_{\boldsymbol{s}}} \tag{9c}$$

where

$$\mathcal{J}_{[i]}(\boldsymbol{\Phi}, \boldsymbol{\Sigma}) = \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}_{[i]}} \Bigg(\log\det(\boldsymbol{\Sigma}) + \left\| \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{x} - \boldsymbol{\Phi} \boldsymbol{y}_{[i]} \right) \right\|_2^2 \Bigg), \quad \forall i \in \mathcal{I}_{\ell_{\boldsymbol{s}}}. \tag{10}$$

In the above, $\boldsymbol{y}_{[i]} = (\boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{S}_{[i]} \boldsymbol{G}_1 + \boldsymbol{H}_0)\boldsymbol{x} + \boldsymbol{w}$ is the replication of $\boldsymbol{y}$ corresponding to $\boldsymbol{s}_{[i]}$. From [25] and [26], the solution of (9) converges to the solution of (8) exponentially fast with the increase of the sample size $\ell_{\boldsymbol{s}}$. Thus, a moderate $\ell_{\boldsymbol{s}}$ is sufficient to find a relatively accurate solution of (8). The optimization problem in (9) is non-convex due to the non-convex unit modulus constraint on $\theta_n, \forall n \in \mathcal{I}_N$. To solve this problem, we propose a two-step alternating optimization method: First optimize $\boldsymbol{\Theta}$ for given $\{\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}_{[i]})\}$, and then solve $\{\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}_{[i]})\}$ for given $\boldsymbol{\Theta}$.

A. Optimization of $\boldsymbol{\Theta}$ for Given $\{\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}_{[i]})\}$

For given $\{\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}_{[i]})\}$, the problem in (9) reduces to

$$\min_{\boldsymbol{\Theta}} \ \frac{1}{\ell_{\boldsymbol{s}}} \sum_{i=1}^{\ell_{\boldsymbol{s}}} \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}_{[i]}} \left\| \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{x} - \boldsymbol{\Phi} \boldsymbol{y}_{[i]} \right) \right\|_2^2 \tag{11a}$$

$$\text{s.t. } |\theta_n| = 1, \quad \forall n \in \mathcal{I}_N. \tag{11b}$$

From the model in (1), we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}_{[i]}} \left\| \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{x} - \boldsymbol{\Phi} \boldsymbol{y}_{[i]} \right) \right\|_2^2 \\
&= \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}_{[i]}} \left\| \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{x} - \boldsymbol{\Phi}((\boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{S}_{[i]} \boldsymbol{G}_1 + \boldsymbol{H}_0)\boldsymbol{x} + \boldsymbol{w}) \right) \right\|_2^2 \\
&= \mathrm{tr} \Big\{ \boldsymbol{\Sigma}^{-1} \Big(\boldsymbol{Q} - \boldsymbol{\Phi} \boldsymbol{H}_0 \boldsymbol{Q} - (\boldsymbol{\Phi} \boldsymbol{H}_0 \boldsymbol{Q})^{\mathrm{H}} + \boldsymbol{\Phi} \boldsymbol{H}_0 \boldsymbol{Q} \boldsymbol{H}_0^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} \\
&\quad + \sigma_w^2 \boldsymbol{\Phi} \boldsymbol{\Phi}^{\mathrm{H}} + \boldsymbol{\Phi} \boldsymbol{G}_2 \boldsymbol{S}_{[i]} \boldsymbol{\Theta} \boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{G}_1^{\mathrm{H}} \boldsymbol{\Theta}^{\mathrm{H}} \boldsymbol{S}_{[i]}^{\mathrm{H}} \boldsymbol{G}_2^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} \\
&\quad + \boldsymbol{\Phi} \boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{S}_{[i]} \boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{H}_0^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} + (\boldsymbol{\Phi} \boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{S}_{[i]} \boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{H}_0^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}})^{\mathrm{H}} \\
&\quad - \boldsymbol{\Phi} \boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{S}_{[i]} \boldsymbol{G}_1 \boldsymbol{Q} - (\boldsymbol{\Phi} \boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{S}_{[i]} \boldsymbol{G}_1 \boldsymbol{Q})^{\mathrm{H}} \Big) \Big\}.
\end{aligned} \tag{12}$$

To simplify the expression in (12), we decompose $\boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{G}_1^{\mathrm{H}}$ as

$$\boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{G}_1^{\mathrm{H}} = \sum_{k=1}^{K} p_k \boldsymbol{g}_{1,k} \boldsymbol{g}_{1,k}^{\mathrm{H}}. \tag{13}$$

Algorithm 1 MM Algorithm

1: Initialize $\boldsymbol{\theta}$ randomly
2: **repeat**
3: Compute $\boldsymbol{q} = (\lambda_{\max} \boldsymbol{I}_N - \boldsymbol{\Lambda})\boldsymbol{\theta} - \boldsymbol{\alpha}$
4: Update $\boldsymbol{\theta} = e^{j \arg(\boldsymbol{q})}$
5: **until** the objective function in (18a) is reduced by less than ϵ_{MM}, or the iteration index reaches $iter_{MM}$

Plugging (13) into (12), we obtain

$$\begin{aligned}
&\min_{\boldsymbol{\Theta}} \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}_{[i]}} \left\| \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{x} - \boldsymbol{\Phi} \boldsymbol{y}_{[i]} \right) \right\|_2^2 \\
&= \min_{\boldsymbol{\Theta}} \sum_{k=1}^{K} p_k \boldsymbol{\theta}^{\mathrm{H}} \mathrm{diag}\{\boldsymbol{g}_{1,k}\}^{\mathrm{H}} \boldsymbol{S}_{[i]}^{\mathrm{H}} \boldsymbol{G}_2^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \boldsymbol{S}_{[i]} \\
&\quad \times \mathrm{diag}\{\boldsymbol{g}_{1,k}\} \boldsymbol{\theta} \\
&\quad + \boldsymbol{\theta}^{\mathrm{T}} \mathrm{diag} \left\{ \boldsymbol{S}_{[i]} \boldsymbol{G}_1 \boldsymbol{Q} (\boldsymbol{H}_0^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} - \boldsymbol{I}) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \right\} \\
&\quad + \left(\boldsymbol{\theta}^{\mathrm{T}} \mathrm{diag} \left\{ \boldsymbol{S}_{[i]} \boldsymbol{G}_1 \boldsymbol{Q} (\boldsymbol{H}_0^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} - \boldsymbol{I}) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \right\} \right)^{\mathrm{H}},
\end{aligned} \tag{14}$$

where the terms irrelevant to $\boldsymbol{\Theta}$ are omitted, and in the derivation we use the facts that $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$, $\mathrm{tr}\left(\mathrm{diag}\{\boldsymbol{\alpha}\}^{\mathrm{H}} \boldsymbol{A} \mathrm{diag}\{\boldsymbol{\alpha}\}\right) = \boldsymbol{\alpha}^{\mathrm{H}} [\boldsymbol{A}]_{\mathrm{diag}} \boldsymbol{\alpha}$, and $\mathrm{tr}(\mathrm{diag}\{\boldsymbol{\alpha}\} \boldsymbol{A}) = \boldsymbol{\alpha}^{\mathrm{T}} \mathrm{diag}\{\boldsymbol{A}\}$. Denote

$$\boldsymbol{\Lambda}_{[i]} = \sum_{k=1}^{K} p_k \mathrm{diag}\{\boldsymbol{g}_{1,k}\}^{\mathrm{H}} \boldsymbol{S}_{[i]}^{\mathrm{H}} \boldsymbol{G}_2^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \boldsymbol{S}_{[i]} \mathrm{diag}\{\boldsymbol{g}_{1,k}\}, \tag{15}$$

$$\boldsymbol{\alpha}_{[i]} = \left(\mathrm{diag} \left\{ \boldsymbol{S}_{[i]} \boldsymbol{G}_1 \boldsymbol{Q} (\boldsymbol{H}_0^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} - \boldsymbol{I}) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \right\} \right)^*. \tag{16}$$

We further define

$$\boldsymbol{\Lambda} = \frac{1}{\ell_{\boldsymbol{s}}} \sum_{i=1}^{\ell_{\boldsymbol{s}}} \boldsymbol{\Lambda}_{[i]} \quad \text{and } \boldsymbol{\alpha} = \frac{1}{\ell_{\boldsymbol{s}}} \sum_{i=1}^{\ell_{\boldsymbol{s}}} \boldsymbol{\alpha}_{[i]}. \tag{17}$$

Then, the optimization problem in (11) is converted to

$$\min_{\boldsymbol{\Theta}} \ \left(\boldsymbol{\theta}^{\mathrm{H}} \boldsymbol{\Lambda} \boldsymbol{\theta} + \boldsymbol{\alpha}^{\mathrm{H}} \boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{H}} \boldsymbol{\alpha} \right) \tag{18a}$$

$$\text{s.t. } |\theta_n| = 1, \quad \forall n \in \mathcal{I}_N \tag{18b}$$

where (18a) utilizes the fact that $\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{\theta}^* + \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\alpha}^* = \boldsymbol{\alpha}^{\mathrm{H}} \boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{H}} \boldsymbol{\alpha}$. The optimization problem in (18) is a quadratically constrained quadratic program (QCQP) and can be solved by the majorization-minimization (MM) algorithm proposed in [20]. The MM algorithm is given in Algorithm 1, where $\lambda_{\max}$ in Line 3 is the maximum eigenvalue of $\boldsymbol{\Lambda}$, and $\arg(\cdot)$ in Line 4 returns the argument(s) of the input in an elementwise manner. As shown in [20], the converged solution of the MM algorithm satisfies the Karush-Kuhn-Tucker optimality conditions of Problem (18). We omit the details here due to space limitation.

B. Solve $\{\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}_{[i]})\}$ for Given $\boldsymbol{\Theta}$

We now solve $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{s}_{[i]}) = \min_{\boldsymbol{\Phi}, \boldsymbol{\Sigma} \succeq \boldsymbol{0}} \mathcal{J}_{[i]}(\boldsymbol{\Phi}, \boldsymbol{\Sigma}), \forall i \in \mathcal{I}_{\ell_{\boldsymbol{s}}}$ for given $\boldsymbol{\Theta}$.

Algorithm 2 SAA-Based P-BF Algorithm

1: Initialize $\boldsymbol{\Theta}$ randomly
2: **repeat**
3: Generate $\{\boldsymbol{s}_{[1]},\ldots,\boldsymbol{s}_{[\ell_s]}\}$ independently based on (5)
4: Compute $\{(\boldsymbol{\Phi}_{[i]},\boldsymbol{\Sigma}_{[i]})\}$ based on (19) and (20)
5: Compute $\boldsymbol{\Lambda}_{[i]}$ and $\boldsymbol{\alpha}_{[i]}$ based on (15) and (16)
6: Compute $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}$ based on (17)
7: Update $\boldsymbol{\theta}$ by the MM algorithm given in Algorithm 1
8: **until** the objective function in (9a) is reduced by less than ϵ_{bf1}, or the iteration index reaches $iter_{bf1}$

Proposition 1: The close-form solution of $\min_{\boldsymbol{\Phi},\boldsymbol{\Sigma}\succeq\mathbf{0}} \mathcal{J}_{[i]}(\boldsymbol{\Phi},\boldsymbol{\Sigma}),\forall i\in\mathcal{I}_{\ell_s}$ *is given by*

$$\boldsymbol{\Phi}_{[i]} = \boldsymbol{\mathcal{C}}_{\boldsymbol{xy}|\boldsymbol{s}_{[i]}}\boldsymbol{\mathcal{C}}^{-1}_{\boldsymbol{yy}|\boldsymbol{s}_{[i]}},\forall i\in\mathcal{I}_{\ell_s}, \tag{19}$$

$$\boldsymbol{\Sigma}_{[i]} = \boldsymbol{\mathcal{C}}_{\boldsymbol{xx}} - \boldsymbol{\mathcal{C}}_{\boldsymbol{xy}|\boldsymbol{s}_{[i]}}\boldsymbol{\mathcal{C}}^{-1}_{\boldsymbol{yy}|\boldsymbol{s}_{[i]}}\boldsymbol{\mathcal{C}}_{\boldsymbol{yx}|\boldsymbol{s}_{[i]}},\forall i\in\mathcal{I}_{\ell_s}, \tag{20}$$

where $\boldsymbol{\mathcal{C}}_{\boldsymbol{xy}|\boldsymbol{s}_{[i]}}$ *is the covariance matrix of* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *conditioned on* $\boldsymbol{s}$ *given by*

$$\boldsymbol{\mathcal{C}}_{\boldsymbol{xy}|\boldsymbol{s}_{[i]}} = \boldsymbol{Q}(\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}_{[i]}\boldsymbol{G}_1+\boldsymbol{H}_0)^{\mathrm{H}}, \tag{21}$$

and $\boldsymbol{\mathcal{C}}_{\boldsymbol{yy}|\boldsymbol{s}_{[i]}}$ *is the covariance matrix of* $\boldsymbol{y}$ *conditioned on* $\boldsymbol{s}$ *given by*

$$\begin{aligned}\boldsymbol{\mathcal{C}}_{\boldsymbol{yy}|\boldsymbol{s}_{[i]}}=&\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}_{[i]}\boldsymbol{G}_1\boldsymbol{Q}\boldsymbol{G}_1^{\mathrm{H}}\boldsymbol{S}_{[i]}^{\mathrm{H}}\boldsymbol{\Theta}^{\mathrm{H}}\boldsymbol{G}_2^{\mathrm{H}}+\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}_{[i]}\boldsymbol{G}_1\boldsymbol{Q}\boldsymbol{H}_0^{\mathrm{H}}\\&+\left(\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}_{[i]}\boldsymbol{G}_1\boldsymbol{Q}\boldsymbol{H}_0^{\mathrm{H}}\right)^{\mathrm{H}}+\boldsymbol{H}_0\boldsymbol{Q}\boldsymbol{H}_0^{\mathrm{H}}+\sigma_w^2\boldsymbol{I}.\end{aligned} \tag{22}$$

Proof: First, for an arbitrarily given $\boldsymbol{\Sigma}$, the problem of $\min_{\boldsymbol{\Phi},\boldsymbol{\Sigma}\succeq\mathbf{0}}\mathcal{J}_{[i]}(\boldsymbol{\Phi},\boldsymbol{\Sigma}),\forall i\in\mathcal{I}_{\ell_s}$ is converted to

$$\begin{aligned}\min_{\boldsymbol{\Phi}}\ \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}_{[i]}}\Big(&\boldsymbol{y}^{\mathrm{H}}\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\boldsymbol{y}-\boldsymbol{x}^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\boldsymbol{y}\\&-\boldsymbol{y}^{\mathrm{H}}\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\Big),\quad\forall i\in\mathcal{I}_{\ell_s}.\end{aligned} \tag{23}$$

Problem (23) is quadratic and convex in $\boldsymbol{\Phi}$, and so the optimal solution can be found by taking the first-order derivative over $\boldsymbol{\Phi}$ as

$$\begin{aligned}\frac{\partial\mathcal{J}_{[i]}(\boldsymbol{\Phi},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Phi}} &= -2\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}_{[i]}}\left((\boldsymbol{x}-\boldsymbol{\Phi}\boldsymbol{y}_{[i]})\boldsymbol{y}_{[i]}^{\mathrm{H}}\right)\\&= -2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mathcal{C}}_{\boldsymbol{xy}|\boldsymbol{s}_{[i]}}-\boldsymbol{\Phi}\boldsymbol{\mathcal{C}}_{\boldsymbol{yy}|\boldsymbol{s}_{[i]}}).\end{aligned} \tag{24}$$

Setting $\frac{\partial\mathcal{J}_{[i]}(\boldsymbol{\Phi},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Phi}}=\mathbf{0}$, we obtain the optimal $\boldsymbol{\Phi}$ as in (19). We see that the optimal $\boldsymbol{\Phi}$ is irrelevant to $\boldsymbol{\Sigma}$. Then, by substituting $\boldsymbol{\Phi}=\boldsymbol{\Phi}_{[i]}$, the optimization problem of $\min_{\boldsymbol{\Phi},\boldsymbol{\Sigma}\succeq\mathbf{0}}\mathcal{J}_{[i]}(\boldsymbol{\Phi},\boldsymbol{\Sigma}),\forall i\in\mathcal{I}_{\ell_s}$ reduces to

$$\max_{\boldsymbol{\Omega}\succeq\mathbf{0}}\ \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}_{[i]}}\Big(-\log\det(\boldsymbol{\Omega})+\boldsymbol{\Omega}\left\|\left(\boldsymbol{x}-\boldsymbol{\Phi}_{[i]}\boldsymbol{y}_{[i]}\right)\right\|_2^2\Big), \tag{25}$$

where $\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}$ is a semidefinite precision matrix. From the concavity of the log-determinant function, (25) is convex in $\boldsymbol{\Omega}$ and the optimal solution can be found by setting the first-order derivative of (25) over $\boldsymbol{\Omega}$ to zero. As a result, the optimal $\boldsymbol{\Sigma}$ has the form in (20). ∎

C. Overall Iterative Algorithm

The overall iterative algorithm is summarized as the pseudocode in Algorithm 2. We refer to this algorithm as the SAA-based P-BF algorithm. The convergence of the SAA-based P-BF algorithm is guaranteed since the objective function of (9) decreases monotonically in each step. We now prove that this algorithm converges to a local minimum of the original problem (9). The optimization problem in the first step is non-convex and is approximately solved by the MM algorithm. The converged solution of the MM algorithm satisfies the KKT optimality conditions of the subproblem (11) [20]. The optimization problem in the second step is convex and its close-form solution satisfies the KKT optimality conditions of the subproblem (9c). As a result, the converged solution of the overall iteration algorithm satisfies the KKT conditions of both the two subproblems and thus satisfies the KKT conditions of the original problem (9). Therefore, we conclude that the converged solution of the SAA-based P-BF algorithm is at least a local minimum of the original problem (9). However, we observe from numerical experiments that the convergence of this algorithm is relatively slow, which incurs a high computational cost. This inspire us to develop a simplified P-BF method to strike a balance between complexity and performance, as detailed in the next section.

IV. Simplified Beamforming Design

In this section, we present the simplified P-BF method by slightly modifying the objective function of problem (8). Specifically, by exchanging the order of the minimization over $(\boldsymbol{\Phi},\boldsymbol{\Sigma})$ and the expectation over $\boldsymbol{s}$, we can recast the P-BF design problem in (8) as

$$\min_{\boldsymbol{\Phi},\boldsymbol{\Sigma}\succeq\mathbf{0},\boldsymbol{\Theta}}\ \mathbb{E}_{\boldsymbol{y},\boldsymbol{x},\boldsymbol{s}}\left(\log\det(\boldsymbol{\Sigma})+\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\Phi}\boldsymbol{y})\right\|_2^2\right) \tag{26a}$$

$$\text{s.t. } |\theta_n|=1,\quad\forall n\in\mathcal{I}_N. \tag{26b}$$

It is not difficult to see that the expectation over $\boldsymbol{y},\boldsymbol{x},\boldsymbol{s}$ in (26a) can be solved explicitly, and so the problem formulated in (26) is computationally more friendly to handle than the one in (9). Yet, the problem in (26) is still non-convex, and it is in general difficult to find a global minimum to (26). Similarly to the approach described in the preceding section, we propose an alternating optimization method that iteratively optimize $\boldsymbol{\Theta}$ and $(\boldsymbol{\Phi},\boldsymbol{\Sigma})$ to find a suboptimal solution of (26). The details of the algorithm are described below.

A. Optimization of $\boldsymbol{\Theta}$ for Fixed $(\boldsymbol{\Phi},\boldsymbol{\Sigma})$

For fixed $(\boldsymbol{\Phi},\boldsymbol{\Sigma})$, the optimization problem in (26) reduces to

$$\min_{\boldsymbol{\Theta}}\ \mathbb{E}_{\boldsymbol{y},\boldsymbol{x},\boldsymbol{s}}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\Phi}\boldsymbol{y})\right\|_2^2 \tag{27a}$$

$$\text{s.t. } |\theta_n|=1,\forall n\in\mathcal{I}_N. \tag{27b}$$

Problem (27) can be equivalently rewritten as

$$\min_{\boldsymbol{\Theta}}\ \boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\Lambda}'\boldsymbol{\theta}+(\boldsymbol{\alpha}')^{\mathrm{H}}\boldsymbol{\theta}+\boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\alpha}' \tag{28a}$$

$$\text{s.t. } |\theta_n|=1,\quad\forall n\in\mathcal{I}_N \tag{28b}$$

where

$$\boldsymbol{\Lambda}' = \sum_{k=1}^{K} \rho^2 p_k \text{diag}\{\boldsymbol{g}_{1,k}\}^{\text{H}} \boldsymbol{G}_2^{\text{H}} \boldsymbol{\Phi}^{\text{H}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \text{diag}\{\boldsymbol{g}_{1,k}\} + \rho(1-\rho)\text{diag}\left\{\boldsymbol{V}^{\text{H}} \boldsymbol{G}_2^{\text{H}} \boldsymbol{\Phi}^{\text{H}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \boldsymbol{V}\right\}, \quad (29)$$

$$\boldsymbol{\alpha}' = \rho \left(\text{diag}\left\{\boldsymbol{G}_1 \boldsymbol{Q} (\boldsymbol{H}_0^{\text{H}} \boldsymbol{\Phi}^{\text{H}} - \boldsymbol{I}) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2\right\}\right)^*. \quad (30)$$

The detailed derivation of the equivalence between (27) and (28) can be found in Appendix B. Similarly to (18), the non-convex QCQP problem in (28) can be approximately solved by the MM algorithm as described in Section III-A. The details are omitted for brevity.

B. Optimization of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ for Fixed $\boldsymbol{\Theta}$

For a fixed $\boldsymbol{\Theta}$, the optimization problem in (26) reduces to

$$\min_{\boldsymbol{\Phi}, \boldsymbol{\Sigma} \succeq \boldsymbol{0}} \ \mathcal{J}(\boldsymbol{\Phi}, \boldsymbol{\Sigma}) = \mathbb{E}_{\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{s}} \Big(\log\det(\boldsymbol{\Sigma}) + \left\| \boldsymbol{\Sigma}^{-\frac{1}{2}} (\boldsymbol{x} - \boldsymbol{\Phi} \boldsymbol{y}) \right\|_2^2 \Big) \quad (31)$$

The close-form solution of $\min_{\boldsymbol{\Phi}, \boldsymbol{\Sigma} \succeq \boldsymbol{0}} \mathcal{J}(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ is given by

$$\boldsymbol{\Phi}^{opt} = \mathbb{E}_{\boldsymbol{s}}(\boldsymbol{\mathcal{C}}_{\boldsymbol{xy}|\boldsymbol{s}}) \left[\mathbb{E}_{\boldsymbol{s}}(\boldsymbol{\mathcal{C}}_{\boldsymbol{yy}|\boldsymbol{s}})\right]^{-1}, \quad (32)$$

$$\boldsymbol{\Sigma}^{opt} = \boldsymbol{\mathcal{C}}_{\boldsymbol{xx}} - \mathbb{E}_{\boldsymbol{s}}(\boldsymbol{\mathcal{C}}_{\boldsymbol{xy}|\boldsymbol{s}}) \mathbb{E}_{\boldsymbol{s}}(\boldsymbol{\mathcal{C}}_{\boldsymbol{yy}|\boldsymbol{s}})^{-1} \mathbb{E}_{\boldsymbol{s}}(\boldsymbol{\mathcal{C}}_{\boldsymbol{yx}|\boldsymbol{s}}). \quad (33)$$

where

$$\mathbb{E}_{\boldsymbol{s}}(\boldsymbol{\mathcal{C}}_{\boldsymbol{xy}|\boldsymbol{s}}) = \boldsymbol{Q}(\rho \boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{G}_1 + \boldsymbol{H}_0)^{\text{H}}, \quad (34)$$

$$\mathbb{E}_{\boldsymbol{s}}\left(\boldsymbol{\mathcal{C}}_{\boldsymbol{yy}|\boldsymbol{s}}\right) = \rho^2 \boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{G}_1^{\text{H}} \boldsymbol{\Theta}^{\text{H}} \boldsymbol{G}_2^{\text{H}} + \rho(1-\rho) \boldsymbol{G}_2 \boldsymbol{\Theta} \times \left[\boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{G}_1^{\text{H}}\right]_{\text{diag}} \boldsymbol{\Theta}^{\text{H}} \boldsymbol{G}_2^{\text{H}} \rho \boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{H}_0^{\text{H}} + \rho \left(\boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{H}_0^{\text{H}}\right)^{\text{H}} + \boldsymbol{H}_0 \boldsymbol{Q} \boldsymbol{H}_0^{\text{H}} + \sigma_w^2 \boldsymbol{I}. \quad (35)$$

The proof of the above conclusion strictly follows that of Proposition 1. We omit the details for brevity.

C. Overall Iterative Algorithm

The overall iterative algorithm is summarized as the pseudocode in Algorithm 3. The convergence of Algorithm 3 is guaranteed since the objective function of (26) decreases monotonically in each update of $\boldsymbol{\Theta}$ and $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$. Furthermore, following the argument in Section III-C, the simplified P-BF algorithm converges to a local minimum of the original problem (26). We now briefly compare the computational complexities of the simplified P-BF algorithm (i.e., Algorithm 3) and the SAA-base P-BF algorithm (i.e., Algorithm 2). By inspection, the differences between Algorithm 3 and Algorithm 2 are: Lines 3 and 4 of Algorithm 2 are replaced by Line 3 of Algorithm 3; and Lines 5 and 6 of Algorithm 2 are replaced by Line 4 of Algorithm 3. With these replacements, the complexity of the related part is reduced by $\ell_{\boldsymbol{s}}$ times since no sample averaging is involved in Line 3 of Algorithm 3. Furthermore, it can be seen from the numerical results presented later in Section VII that the convergence speed of the Algorithm 3 is order-of-magnitude faster than that of Algorithm 2. This is expected since deterministic optimization algorithms usually converge much faster than stochastic optimization algorithms.

Algorithm 3 Simplified P-BF Algorithm

1: Initialize $\boldsymbol{\Theta}$ randomly
2: **repeat**
3: Compute $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ according to (32)
4: Compute $\boldsymbol{\Lambda}'$ and $\boldsymbol{\alpha}'$ based on (29) and (30)
5: Update $\boldsymbol{\theta}$ by the MM algorithm given in Algorithm 1
6: **until** the objective function in (26a) is reduced by less than ϵ_{bf2}, or the iteration index reaches $iter_{bf2}$

V. Receiver Design

A. Problem Formulation

The receiver at the BS aims to reliably recover both the information from the users and the RIS (i.e., $\boldsymbol{X}$ and $\boldsymbol{s}$). This recovery problem can be formulated by using the maximum *a posteriori* principle as

$$\left(\hat{\boldsymbol{X}}, \hat{\boldsymbol{s}}\right) = \arg\max_{\boldsymbol{X}, \boldsymbol{s}} p_{\boldsymbol{X}, \boldsymbol{s}|\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{s}|\boldsymbol{Y}). \quad (36)$$

Exactly solving (36) is difficult by noting the bilinear model of $\boldsymbol{X}$ and $\boldsymbol{s}$ in (2). Message passing is a powerful technique to yield near-optimal low-complexity solutions to sophisticated inference problems. Particularly, the PBiGAMP algorithm [22] can possibly handle the bilinear inference problem in (36). To see this, we rewrite the model in (2) as

$$\begin{aligned} \boldsymbol{Y} &= (\boldsymbol{G}_2 \boldsymbol{\Theta} \boldsymbol{S} \boldsymbol{G}_1 + \boldsymbol{H}_0) \boldsymbol{X} + \boldsymbol{W} \\ &= \left(\sum_{n=1}^{N} s_n \theta_n \boldsymbol{g}_{2,n} \boldsymbol{g}_{1,n}^{\text{T}} + \boldsymbol{H}_0\right) \boldsymbol{X} + \boldsymbol{W} \\ &= \sum_{n=0}^{N} s_n \boldsymbol{H}_n \boldsymbol{X} + \boldsymbol{W}, \end{aligned} \quad (37)$$

where $\boldsymbol{H}_n \triangleq \theta_n \boldsymbol{g}_{2,n} \boldsymbol{g}_{1,n}^{\text{T}} \in \mathbb{C}^{M \times K}$ is a rank-1 matrix $\forall n \in \mathcal{I}_N$ and $s_0 = 1$ is a constant. Since (37) is special case of the model [22, eq. 2], we see that the PBiGAMP algorithm is applicable to problem (36).

However, there are two issues with the PBiGAMP algorithm when it is applied to solve (36). First, the PBiGAMP algorithm requires that the measurement matrixes consist of i.i.d. Gaussian elements. In (37), the measurement matrices $\{\boldsymbol{H}_1, \ldots, \boldsymbol{H}_N\}$ are rank-1, which violates the i.i.d. assumption in PBiGAMP. Second, the computational complexity of PBiGAMP is $\mathcal{O}(MNKT^2)$, which is unaffordable for a large T.

To avoid the above two issues, we develop a turbo message passing algorithm as follows. We start with the factor graph representation of the probability model involved in (36). From the Bayes' rule, we obtain

$$p_{\boldsymbol{X}, \boldsymbol{s}|\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{s}|\boldsymbol{Y}) \propto p_{\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{s}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{s}) p_{\boldsymbol{X}}(\boldsymbol{X}) p_{\boldsymbol{s}}(\boldsymbol{s}) \quad (38a)$$

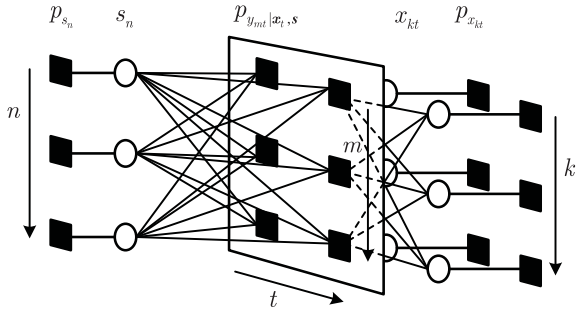


Fig. 2. The factor graph representation for the joint probability in (38) with $M = N = K = 3$ and $T = 2$.

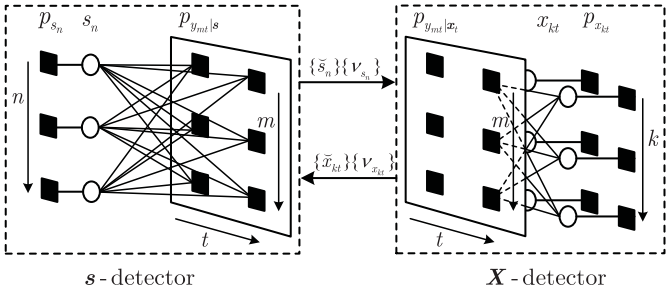


Fig. 3. The factor graph representation for turbo message passing with $M = N = K = 3$ and $T = 2$.

$$= \left[\prod_{m=1}^{M}\prod_{t=1}^{T} p_{y_{mt}|\boldsymbol{x}_t,\boldsymbol{s}}\left(y_{mt}|\sum_{n=1}^{N}\sum_{k=1}^{K} s_n h_{n,mk} x_{k,t}\right)\right] \times \left[\prod_{k=1}^{K}\prod_{t=1}^{T} p_{x_{kt}}(x_{kt})\right]\left[\prod_{n=0}^{N} p_{s_n}(s_n)\right], \tag{38b}$$

where the notation $\propto$ in (38a) means equality up to a constant scaling factor; (38b) is from the fact that $\{x_{kt}\}$ and $\{s_n\}$ are independent of each other. The factorized posterior distribution in (38b) can be represented by a factor graph, as depicted in Fig. 2. Note that in Fig. 2, we use a hollow circle to represent a "variable node" and a solid square to represent a "factor node".

We are now ready to present the turbo message passing framework by dividing the whole factor graph in Fig. 2 into two modules, namely the $\boldsymbol{X}$-detector and the $\boldsymbol{s}$-detector, as depicted in Fig. 3. We iteratively estimate $\boldsymbol{X}$ and $\boldsymbol{s}$ by performing message passing in each module. The messages are exchanged between the two modules until convergence or the maximum iteration number is reached. The $\boldsymbol{X}$-detector estimates $\boldsymbol{X}$ based on the observed signal $\boldsymbol{Y}$ and the messages from the $\boldsymbol{s}$-detector. The outputs of the $\boldsymbol{X}$-detector are the means and variances of $\{x_{kt}\}$ denoted by $\{\breve{x}_{kt}\}$ and $\{\upsilon_{x_{kt}}\}$. Similarly, the $\boldsymbol{s}$-detector estimates $\boldsymbol{s}$ based on the observed signal $\boldsymbol{Y}$ and the messages from the $\boldsymbol{X}$-detector, and outputs the means and variances of $\{s_n\}$ denoted by $\{\breve{s}_n\}$ and $\{\upsilon_{s_n}\}$. In the following subsections, we give a detailed description of the turbo message passing algorithm.

B. Design of $\boldsymbol{X}$-Detector

We first describe the design of the $\boldsymbol{X}$-detector. From Fig. 3, the input means and variances of $\boldsymbol{s}$ are respectively given by $\breve{\boldsymbol{s}}$ and $\{\upsilon_{s_n}\}$. Let $\tilde{\boldsymbol{s}} = \boldsymbol{s} - \breve{\boldsymbol{s}}$ be the residual interference of $\boldsymbol{s}$ with mean zero and variances $\{\upsilon_{s_n}\}$. Then the observed signal $\boldsymbol{Y}$ can be expressed as

$$\boldsymbol{Y} = \sum_{n=0}^{N}(\breve{s}_n + \tilde{s}_n)\boldsymbol{H}_n\boldsymbol{X} + \boldsymbol{W} = \breve{\boldsymbol{H}}\boldsymbol{X} + \tilde{\boldsymbol{W}}_{\boldsymbol{X}}, \tag{39}$$

where $\breve{\boldsymbol{H}} = \sum_{n=0}^{N}\breve{s}_n\boldsymbol{H}_n$ is the linear transform matrix known by the $\boldsymbol{X}$-detector and $\tilde{\boldsymbol{W}}_{\boldsymbol{X}} \triangleq \sum_{n=0}^{N}\tilde{s}_n\boldsymbol{H}_n\boldsymbol{X} + \boldsymbol{W}$ is the equivalent noise. Assume that the elements of $\tilde{\boldsymbol{W}}_{\boldsymbol{X}}$ are independent of each other and obey a CSCG distribution. Denote by $\tilde{w}_{\boldsymbol{X},mt}$ the (m,t)th element of $\tilde{\boldsymbol{W}}_{\boldsymbol{X}}$. The variance of each $\tilde{w}_{\boldsymbol{X},mt}$ is

$$\upsilon_{\tilde{w}_{\boldsymbol{X},mt}} = \sum_{i=0}^{N}\sum_{k=1}^{K}\upsilon_{s_i}|h_{i,mk}|^2 p_k + \sigma_w^2, \forall m,t, \tag{40}$$

where p_k is the power of user k. The derivation of (40) is given in Appendix C-A. Thus, $\forall m,t$, we have

$$p_{y_{mt}|\boldsymbol{x}_t}(y_{mt}|\boldsymbol{x}_t) = \mathcal{CN}\left(y_{mt}; \sum_{n=1}^{N}\sum_{k=1}^{K}\breve{s}_n h_{n,mk}x_{k,t}, \upsilon_{\tilde{w}_{\boldsymbol{X},mt}}\right). \tag{41}$$

Given the model in (39), the recovery of $\boldsymbol{X}$ from the observation $\boldsymbol{Y}$ is a linear estimation problem. The existing algorithm GGAMP-SBL [27] is designed to handle such a linear estimation problem with an arbitrary measurement matrix, and thus can be used here for the recovery of $\boldsymbol{X}$.

For completeness, the details of the GGAMP-SBL algorithm are summarized in Algorithm 4. The algorithm uses a Gaussian distribution $\mathcal{CN}(\cdot;\breve{x}_{kt},\upsilon_{x_{kt}}), \forall k,t$ to approximate the prior distribution of each x_{kt} with mean $\breve{x}_{kt}$ and variance $\upsilon_{x_{kt}}$. The variances $\{\upsilon_{x_{kt}}\}$ are updated in the M-step of each expectation maximization (EM) iteration by step 18. The E-step estimates the marginal posterior distributions of $\{x_{kt}\}$ by the GGAMP algorithm, as given in steps 4-16. In specific, steps 5 and 6 update the estimates of $\{z_{mt}\}$ with the variances $\{\nu_{p_{mt}}^{(l)}\}$ and means $\{\hat{p}_{mt}^{(l)}\}$ by accumulating the messages from variable nodes $\{x_{kt}\}$ to check nodes $\{p_{y_{mt}}\}$. Steps 7 and 8 give the estimates of the marginal posterior variances $\{\nu_{z_{nt}}^{(l)}\}$ and means $\{\hat{z}_{mt}^{(l)}\}$ of $\{z_{mt}\}$. Steps 9 and 10 calculate the inverse-residual-variances $\{\nu_{u_{n,t}}^{(l)}\}$ and the scaled residuals $\{\hat{u}_{nt}^{(l)}\}$. Steps 11 and 12 update the estimates of $\{x_{kt}\}$ with variances $\{\nu_{r_{kt}}^{(l)}\}$ and the means $\{\hat{r}_{kt}^{(l)}\}$ by accumulating the messages from check nodes $\{p_{y_{mt}}\}$ to variable nodes $\{x_{kt}\}$. Steps 13 and 14 updates the marginal posterior variances $\{\nu_{x_{kt}}^{(l)}\}$ and means $\{\hat{x}_{kt}^{(l)}\}$ of $\{x_{kt}\}$. Steps 15 and 19 define the termination conditions of the GGAMP iteration and the EM iteration, respectively, where ϵ_{gamp} and ϵ_{em} are the corresponding tolerance parameters. ℓ_{max}^{x} and I_{max}^{x} are the maximum numbers of the GGAMP iteration and the EM iteration, respectively. The outputs of the GGAMP-SBL algorithm are the marginal posterior means $\{\breve{x}_{kt}^{(i)}\}$ and variances $\{\breve{\nu}_{x_{kt}}^{(i)}\}$ of the last iteration.

The damping strategy is used in the steps 10 and 14 to ensure the convergence of the GAMP algorithm, where $\xi_u^x \in [0,1]$ and $\xi_x^x \in [0,1]$ are the damping factors of $\hat{u}_{mt}$

Algorithm 4 The GGAMP-SBL Algorithm

Input: $\boldsymbol{Y}$, $\breve{\boldsymbol{H}}$, $\{\upsilon_{\tilde{w}_{\boldsymbol{X},mt}}\}$, and $\mathcal{CN}(\cdot;\breve{x}_{kt},\upsilon_{x_{kt}}),\forall\, k,t$

1: Initialization: $\forall m,k,t:\ \breve{x}_{kt}^{(0)}=\breve{x}_{kt}$, $\upsilon_{x_{kt}}^{(0)}=\upsilon_{x_{kt}}$, $\breve{\nu}_{x_{kt}}^{(0)}=\upsilon_{x_{kt}}$, $\breve{u}_{mt}^{(0)}=0$
2: **for** $i=1,2,\ldots,I_{max}^{x}$ % EM iteration
3: $\forall m,k,t:\ \hat{x}_{k,t}^{(0)}=\breve{x}_{kt}^{(i-1)}$, $\nu_{x_{k,t}}^{(0)}=\breve{\nu}_{x_{kt}}^{(i-1)}$, $\hat{u}_{mt}^{0}=\breve{u}_{mt}^{(i-1)}$
4: **for** $\ell=1,2,\ldots,\ell_{max}^{x}$ % GGAMP for E-Step
5: $\forall m,t:\ \nu_{p_{mt}}^{(\ell)}=\sum_{k=1}^{k}|\breve{h}_{mk}|^2\nu_{x_{k,t}}^{(\ell-1)}$
6: $\forall m,t:\ \hat{p}_{mt}^{(\ell)}=\sum_{k=1}^{k}\breve{h}_{mk}\hat{x}_{(kt)}^{(\ell-1)}-\nu_{p_{mt}}^{(\ell)}\hat{u}_{mt}^{(\ell-1)}$
7: $\forall m,t:\ \nu_{z_{mt}}^{(\ell)}=\mathbb{V}\mathtt{ar}(z_{mt}|y_{mt};\hat{p}_{m}^{(\ell)},\nu_{p_{mt}}^{(\ell)})$
8: $\forall m,t:\ \hat{z}_{mt}^{(\ell)}=\mathbb{E}\big(z_{mt}|y_{mt};\hat{p}_{mt}^{(\ell)},\nu_{p_{mt}}^{(\ell)}\big)$
9: $\forall m,t:\ \nu_{u_{mt}}^{(\ell)}=\left(1-\nu_{z_{mt}}^{(\ell)}/\nu_{p_{mt}}^{(\ell)}\right)/\nu_{p_{mt}}^{(\ell)}$
10: $\forall m,t:\ \hat{u}_{mt}^{(\ell)}=(1-\xi_u^x)\hat{u}_{mt}^{(\ell-1)}+\xi_u^x\big(\hat{z}_{mt}^{(\ell)}-\hat{p}_{mt}^{(\ell)}\big)\nu_{p_{mt}}^{(\ell)}$
11: $\forall k,t:\ \nu_{r_{kt}}^{(\ell)}=\left(\sum_{m=1}^{M}|\breve{h}_{mk}|^2\nu_{u_{mt}}^{(\ell)}\right)^{-1}$
12: $\forall k,t:\ \hat{r}_{kt}^{(\ell)}=\hat{x}_{kt}^{(\ell-1)}+\nu_{r_{kt}}^{(\ell)}\sum_{m=1}^{M}\breve{h}_{mk}\hat{u}_{mt}^{(\ell)}$
13: $\forall k,t:\ \nu_{x_{kt}}^{(\ell)}=\mathbb{V}\mathtt{ar}\big(x_{kt}|\boldsymbol{y}_t;\hat{r}_{kt}^{(\ell)},\nu_{r_{kt}}^{(\ell)}\big)$
14: $\forall m,t:\ \hat{x}_{kt}^{(\ell)}=(1-\xi_x^x)\hat{x}_{kt}^{(\ell-1)}+\xi_x^x\mathbb{E}\big(x_{kt}|\boldsymbol{Y};\hat{r}_{kt}^{(\ell)},\nu_{r_{kt}}^{(\ell)}\big)$
15: **if** $\|\hat{x}_{kt}^{(\ell)}-\hat{x}_{kt}^{(\ell-1)}\|^2/\|\hat{x}_{kt}^{(\ell)}\|^2\le\epsilon_{\text{gamp}}^{x}$, **break**
16: **end for**
17: $\forall m,k,t:\ \breve{x}_{kt}^{(i)}=\hat{x}_{kt}^{(\ell)}$, $\breve{\nu}_{x_{k,t}}^{(i)}=\nu_{x_{kt}}^{(\ell)}$, $\breve{u}_{mt}^{(i)}=\hat{u}_{mt}^{(\ell)}$
18: $\forall k,t:\ \upsilon_{x_{kt}}^{(i)}=|\breve{x}_{kt}^{(i)}|^2+\breve{\nu}_{x_{k,t}}^{(i)}$ % M-Step
19: **if** $\|\hat{x}_{kt}^{(i)}-\hat{x}_{kt}^{(i-1)}\|^2/\|\hat{x}_{kt}^{(i)}\|^2\le\epsilon_{\text{em}}^{x}$, **break**
20: **end for**

Output: : $\big\{\breve{x}_{kt}^{(i)}\big\}$ and $\big\{\breve{\nu}_{x_{kt}}^{(i)}\big\}$.

and $\hat{x}_{kt}$, respectively. The variance and mean in steps 7 and 8 are taken over the probability distribution $p(z_{mt}|y_{mt};\hat{p}_m,\nu_{p_{mt}})=\frac{1}{C}p(y_{mt}|\boldsymbol{x}_t)\mathcal{CN}(z_{mt}|\hat{p}_m,\nu_{p_{mt}})$, where C is a normalization factor. The variance and mean in steps 12 and 13 are taken over $p(x_{kt};\hat{r}_{kt},\nu_{r_{kt}})=\frac{1}{C}\mathcal{CN}(\cdot;\breve{x}_{kt},\upsilon_{x_{kt}})\mathcal{CN}(x_{kt}|\hat{r}_{kt},\nu_{r_{kt}})$.

C. Design of $\boldsymbol{s}$-Detector

We now describe the design of the $\boldsymbol{s}$-detector. Recall from Fig. 3 that the input mean and variances of $\boldsymbol{X}$ are respectively given by $\breve{\boldsymbol{X}}$ and $\{\upsilon_{x_{kt}}\}$. Let $\tilde{\boldsymbol{X}}=\boldsymbol{X}-\breve{\boldsymbol{X}}$ be the residual interference of $\boldsymbol{X}$ with mean zero and variances $\{\upsilon_{x_{kt}}\}$. Then the observed signal $\boldsymbol{Y}$ can be expressed as

$$\boldsymbol{Y}=\sum_{n=0}^{N}s_n\boldsymbol{H}_n(\breve{\boldsymbol{X}}+\tilde{\boldsymbol{X}})+\boldsymbol{W}=\sum_{n=0}^{N}s_n\boldsymbol{H}_n\breve{\boldsymbol{X}}+\tilde{\boldsymbol{W}}_{\boldsymbol{s}},\tag{42}$$

where $\tilde{\boldsymbol{W}}_{\boldsymbol{s}}\triangleq\sum_{n=0}^{N}s_n\boldsymbol{H}_n\tilde{\boldsymbol{X}}+\boldsymbol{W}$ is the equivalent noise for the $\boldsymbol{s}$-detector. Denoted $\breve{\boldsymbol{A}}_n=\boldsymbol{H}_n\breve{\boldsymbol{X}},\forall n\in\{0,1,\ldots,N\}$ and $\breve{\boldsymbol{A}}=[\text{vec}(\breve{\boldsymbol{A}}_0),\ldots,\text{vec}(\breve{\boldsymbol{A}}_N)]$. Then, we have

$$\text{vec}(\boldsymbol{Y})=\breve{\boldsymbol{A}}\boldsymbol{s}+\text{vec}(\tilde{\boldsymbol{W}}_{\boldsymbol{s}}).\tag{43}$$

Similarly, assume that the elements of $\tilde{\boldsymbol{W}}_{\boldsymbol{s}}$ are independent of each other and obey a CSCG distribution. Denote by $\tilde{w}_{\boldsymbol{s},mt}$ the (m,t)th element of $\tilde{\boldsymbol{W}}_{\boldsymbol{s}}$. The variance of each $\tilde{w}_{\boldsymbol{s},mt}$ is given by

$$\upsilon_{\tilde{w}_{\boldsymbol{s},mt}}=\rho^2\sum_{k=1}^{K}\upsilon_{x_{kt}}\left|\sum_{n=1}^{N}h_{n,mk}\right|^2+\rho(1-\rho)\times\sum_{k'=1}^{K}\sum_{n'=1}^{N}\upsilon_{x_{k't}}|h_{n',mk}|^2+\sigma_w^2,\quad\forall m,t.\tag{44}$$

The derivation of (44) is given in Appendix C-B. Thus, we have

$$p_{y_{mt}|\boldsymbol{s}}(y_{mt}|\boldsymbol{s})=\mathcal{CN}\left(y_{mt};\sum_{n=1}^{N}\sum_{k=1}^{K}s_nh_{n,mk}\breve{x}_{k,t},\upsilon_{\tilde{w}_{\boldsymbol{s},mt}}\right),\quad\forall m,t.\tag{45}$$

Given the model in (43), the recovery of $\boldsymbol{s}$ from $\boldsymbol{Y}$ can also be solved by using the GGAMP-SBL algorithm. The GGAMP-SBL algorithm described in Section V-B can be directly used to detect $\boldsymbol{s}$ after changing the measurement matrix form $\breve{\boldsymbol{H}}$ to $\breve{\boldsymbol{A}}$, the likelihood function from $p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X})$ to $p_{\boldsymbol{Y}|\boldsymbol{s}}(\boldsymbol{Y}|\boldsymbol{s})$, and the *priori* from $\mathcal{CN}(\cdot;\breve{x}_{kt},\upsilon_{x_{kt}}),\forall k,t$ to $\mathcal{CN}(\cdot;\breve{s}_n,\upsilon_{s_n}),\forall n$. Moreover, the parameters I_{max}^{x}, ℓ_{max}^{x}, ξ_u^x, ξ_x^x, $\epsilon_{\text{gamp}}^{x}$ and ϵ_{em}^{x} in the $\boldsymbol{X}$-detector are correspondingly replaced by I_{max}^{s}, ℓ_{max}^{s}, ξ_u^s, ξ_x^s, $\epsilon_{\text{gamp}}^{s}$ and ϵ_{em}^{s}.

Algorithm 5 Turbo Message Passing Algorithm

Input: $\boldsymbol{Y}$, $p_{\boldsymbol{s}}(\boldsymbol{s})$, $p_{\boldsymbol{X}}(\boldsymbol{X})$, and σ_w^2

1: Initialization: $\forall m,n,k,t:\ \breve{s}_n^{(0)}=\int_{s_n}s_np_{s_n}(s_n)$, $\breve{x}_{kt}^{(0)}=\int_{x_{kt}}x_{kt}p_{x_{kt}}(x_{kt})$, $\upsilon_{s_n}^{(0)}=\int_{s_n}\left|s_n-\breve{s}_n^{(0)}\right|^2\ p_{s_n}(s_n)$, $\upsilon_{x_{kt}}^{(0)}=\int_{x_{kt}}\left|x_{kt}-\breve{x}_{kt}^{(0)}\right|^2\ p_{x_{kt}}(x_{kt})$
2: **for** $\tau=1,2,\ldots,\tau_{max}$ % Turbo iteration
3: $\breve{\boldsymbol{H}}^{(\tau)}=\sum_{n=0}^{N}\breve{s}_n^{(\tau-1)}\boldsymbol{H}_n$ % $\boldsymbol{X}$-detector
4: $\forall m,t:\ \nu_{\tilde{w}_{\boldsymbol{X},mt}}^{(\tau)}=\sum_{n=0}^{N}\sum_{k=1}^{K}\upsilon_{s_n}^{(\tau-1)}|h_{n,mk}|^2\ p_k+\sigma_w^2$
5: $\forall k,t:\ \breve{x}_{kt}=\breve{x}_{kt}^{(\tau-1)}$, $\upsilon_{x_{kt}}=\upsilon_{x_{kt}}^{(\tau-1)}$
6: Perform GGAMP-SBL to estimate $\boldsymbol{X}$ by invoking Algorithm 4, and output $\{\breve{x}_{kt}^{(\tau)}\}$ and $\{\upsilon_{x_{kt}}^{(\tau)}\}$
7: $\forall k,t:\ \breve{x}_{kt}^{(\tau)}=\arg\min_{c\in\mathcal{C}}|c-\breve{x}_{kt}^{(\tau)}|^2$
8: $\forall n:\ \breve{\boldsymbol{A}}_n^{(\tau)}=\boldsymbol{H}_n\breve{\boldsymbol{X}}^{(\tau)}$, $\breve{\boldsymbol{A}}^{(\tau)}=[\text{vec}(\hat{\boldsymbol{A}}_0)^{(\tau)},\ldots\ \text{vec}(\hat{\boldsymbol{A}}_N)^{(\tau)}]$ % $\boldsymbol{s}$-detector
9: $\forall m,t:\ \nu_{\tilde{w}_{\boldsymbol{s},mt}}^{(\tau)}=\rho^2\sum_{k=1}^{K}\upsilon_{x_{kt}}^{(\tau)}\left|\sum_{n=1}^{N}h_{n,mk}\right|^2+\rho(1-\rho)\sum_{k'=1}^{K}\sum_{n'=1}^{N}\upsilon_{x_{k't}}^{(\tau)}|h_{n',mk}|^2+\sigma_w^2$
10: $\forall n:\ \breve{s}_n=\breve{s}_n^{(\tau-1)}$, $\upsilon_{s_n}=\upsilon_{s_n}^{(\tau-1)}$
11: Perform GGAMP-SBL to estimate $\boldsymbol{s}$, and output $\{\breve{s}_n^{(\tau)}\}$ and $\{\upsilon_{s_n}^{(\tau)}\}$
12: $\forall n:\ \breve{s}_n^{(\tau)}=\arg\min_{c\in\{0,1\}}|c-\breve{s}_n^{(\tau)}|^2$
13: **if** $\left|\left\|\boldsymbol{Y}-\sum_{n=0}^{N}\breve{s}_n^{(\tau)}\boldsymbol{H}_n\breve{\boldsymbol{X}}^{(\tau)}\right\|_{\text{F}}^2/M/T-\sigma_w^2\right|^2\le\epsilon_{\text{td}}$, **break**
14: **end for**

Output: : $\{\breve{x}_{kt}^{(\tau)}\}$ and $\{\breve{s}_n^{(\tau)}\}$

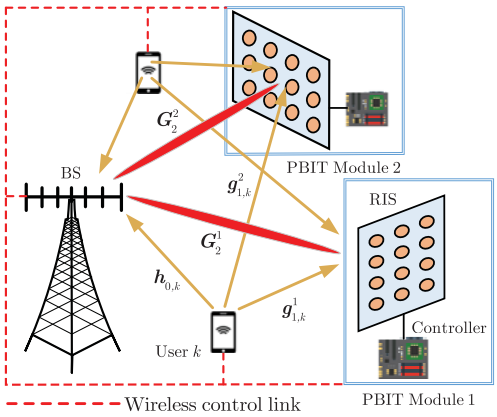


Fig. 4. A PBIT-enhanced uplink MIMO system with multi-RIS.

D. Algorithm Summary

The turbo message passing algorithm is summarized in Algorithm 5. In specific, step 3 updates $\breve{\boldsymbol{H}}$. Step 4 updates the variances $\{\nu_{\tilde{w}_{\boldsymbol{X},mt}}\}$. Step 5 initializes the means $\{\breve{x}_{kt}\}$ and variances $\{v_{x_{kt}}\}$, Step 6 performs the GGAMP-SBL algorithm to update the estimate of $\boldsymbol{X}$. Step 7 maps each $\breve{x}_{kt}^{(\tau)}$ to the nearest constellation point. Step 8 updates $\breve{\boldsymbol{A}}$. Step 9 updates the variances $\{\nu_{\tilde{w}_{\boldsymbol{s},mt}}\}$. Step 10 initializes the means $\{\breve{s}_n\}$ and variances $\{v_{s_n}\}$, Step 11 performs the GGAMP-SBL algorithm to update the estimate of $\boldsymbol{s}$. Step 12 makes a hard decision on each $\breve{s}_n^{(\tau)}$. Step 13 defines the termination condition of the turbo iteration, where ϵ_{td} is the tolerance parameter, and τ_{max} is the maximum number of the turbo iteration. The outputs of the turbo message passing algorithm are the marginal posterior means $\{\breve{x}_{kt}^{(\tau)}\}$ and $\{\breve{s}_n^{(\tau)}\}$ of the last iteration.

We now analyse the computational complexity of the turbo message passing algorithm. The complexity in step 3 is $\mathcal{O}(MNK)$. The complexity in step 4 is $\mathcal{O}(MNKT)$. In the $\boldsymbol{X}$-detector, the complexity of the GGAMP-SBL algorithm is $\mathcal{O}(MKT)$. The complexity in step 8 is $\mathcal{O}(MKT)$. The complexity in step 9 is $\mathcal{O}(MNKT)$. In the $\boldsymbol{s}$-detector, the complexity of the GGAMP-SBL algorithm is $\mathcal{O}(MNT)$. Thus, the computational complexity of the turbo message passing algorithm is dominated by the steps 4 and 9, and is given by $\mathcal{O}(MNKT)$.

VI. Extension on Multi-RIS

We now extend the discussions from the single-RIS aided Mu-MIMO system to the multi-RIS case. As illustrated in Fig. 4, we consider an L-RIS aided wireless system, where each RIS belongs to a PBIT module equipped with an individual controller. The phase shifts of the reflecting elements in all the RISs are jointly designed to enhance the user-BS communication.

The channel model of the multi-RIS aided Mu-MIMO system is described as follows. The channel of the user-BS link is still denoted by $\boldsymbol{H}_0$, where $\boldsymbol{H}_0 \in \mathbb{C}^{M\times K}$ is defined in the same way as in the single-RIS system in (1). Denote by RIS^l the RIS in the lth PBIT module. The number of the reflecting elements in RIS^l is $N_l, \forall l \in \mathcal{I}_L$. Denote by $\boldsymbol{G}_1^l = [\boldsymbol{g}_{1,1}^l, \ldots, \boldsymbol{g}_{1,N_l}^l]^{\mathrm{T}} \in \mathbb{C}^{N_l\times K}$ the baseband equivalent channel of the user-RIS^l link, where $\boldsymbol{g}_{1,n}^l \in \mathbb{C}^{K\times 1}$ represents the channel coefficient vector between all the users and the nth reflecting element of RIS^l. $\boldsymbol{S}^l = \text{diag}\{\boldsymbol{s}^l\}$ is the diagonal state matrix of RIS^l, where $\boldsymbol{s}^l = [s_1^l, \ldots, s_{N_l}^l]^{\mathrm{T}}$ carries the information of the lth PBIT module. $\boldsymbol{\Theta}^l = \text{diag}\{\boldsymbol{\theta}^l\}$ is the diagonal phase-shift matrix of RIS^l, with $\boldsymbol{\theta}^l = [\theta_1^l, \ldots, \theta_{N_l}^l]^{\mathrm{T}}$. Denote by $\boldsymbol{G}_2^l = [\boldsymbol{g}_{2,1}^l, \ldots, \boldsymbol{g}_{2,N_l}^l] \in \mathbb{C}^{M\times N_l}$ the baseband equivalent channel of the RIS^l-BS link, where $\boldsymbol{g}_{2,n}^l \in \mathbb{C}^{M\times 1}$ is the channel coefficient vector between the nth reflecting element of RIS^l and the BS. Then, the observed signal matrix in the multi-RIS system over a transmission block can be expressed as

$$\boldsymbol{Y} = \left(\sum_{l=1}^{L} \boldsymbol{G}_2^l \boldsymbol{\Theta}^l \boldsymbol{S}^l \boldsymbol{G}_1^l + \boldsymbol{H}_0\right)\boldsymbol{X} + \boldsymbol{W}, \tag{46}$$

where $\boldsymbol{X}$ and $\boldsymbol{W}$ are defined in the same way as in the single-RIS system.

Denoted by $N = \sum_{l=1}^{L} N_l$ the total number of the reflecting elements in all the L RISs. $\boldsymbol{G}_1 \triangleq \left[\left(\boldsymbol{G}_1^1\right)^{\mathrm{T}}, \ldots, \left(\boldsymbol{G}_1^L\right)^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{N\times K}$ represents the baseband equivalent channel of the user-RIS link. $\boldsymbol{S} = \text{diag}\{\boldsymbol{s}\}$ with $\boldsymbol{s} = \left[\left(\boldsymbol{s}^1\right)^{\mathrm{T}}, \ldots, \left(\boldsymbol{s}^L\right)^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{N\times 1}$ represents the diagonal state matrix of all the RISs. $\boldsymbol{\Theta} = \text{diag}\{\boldsymbol{\theta}\}$ with $\boldsymbol{\theta} = \left[\left(\boldsymbol{\theta}^1\right)^{\mathrm{T}}, \ldots, \left(\boldsymbol{\theta}^L\right)^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{N\times 1}$ represents the diagonal phase-shift matrix of all the RISs. $\boldsymbol{G}_2 \triangleq [\boldsymbol{G}_2^1, \ldots, \boldsymbol{G}_2^L] \in \mathbb{C}^{M\times N}$ represents the baseband equivalent channel of the RIS-BS link. Then, the system model in the multi-RIS system can still be expressed by (2) (which is the same as the single-RIS system).

Denote by ρ_l the probability that $s_n^l, \forall n \in \mathcal{I}_{N_l}, \forall l \in \mathcal{I}_L$ takes the value of 1, where s_n^l is the state of the nth passive element in RIS^l. Then, the probability of $\boldsymbol{s}$ in the multi-RIS system can be written as

$$p(\boldsymbol{s}) = \prod_{l=1}^{L}\left(\prod_{n=1}^{N_l} p(s_n)\right) = \prod_{l=1}^{L}\Bigg(\prod_{n=1}^{N_l}\Big((1-\rho_l)\delta(s_n^l), + \rho_l\delta(s_n^l - 1)\Big)\Bigg). \tag{47}$$

Then, we have

$$\mathbb{E}_{\boldsymbol{s}}\left(\boldsymbol{S}\right) = \text{diag}\{\mathcal{E}_s\} \tag{48}$$

$$\mathbb{E}_{\boldsymbol{s}}\left(\boldsymbol{S}\boldsymbol{X}\boldsymbol{S}^{\mathrm{H}}\right) = \left(\mathcal{E}_s\mathcal{E}_s^{\mathrm{H}}\right)\odot\boldsymbol{X} + \text{diag}\{\mathcal{E}_s(\boldsymbol{1}_N - \mathcal{E}_s)\}\left[\boldsymbol{X}\right]_{\text{diag}}, \tag{49}$$

where $\mathcal{E}_s \triangleq [\rho_1\boldsymbol{1}_{N_1}^{\mathrm{T}}, \ldots, \rho_L\boldsymbol{1}_{N_L}^{\mathrm{T}}]^{\mathrm{T}}$.

We now discuss the P-BF design and the receiver design for the multi-RIS aided Mu-MIMO system. Note that the system model of the multi-RIS system shares the same expression (i.e., (2)) as that of the single-RIS system. The only difference is the PDF of $\boldsymbol{s}$ in (47) for the multi-RIS case as compared to (5) for the single-RIS case. This difference induces the following modifications when extending the P-BF and receiver design of the single-RIS case to the multi-RIS case:

- *SAA-based P-BF algorithm:* The independently generated samples $\{\boldsymbol{s}_{[1]}, \ldots, \boldsymbol{s}_{[\ell_{\boldsymbol{s}}]}\}$ are based on the PDF in (47) rather than in (5).
- *Simplified P-BF algorithm:* We replace (29) and (30) respectively by

$$\begin{aligned}\boldsymbol{\Lambda} = &\sum_{k=1}^{K} p_k \mathrm{diag}\{\mathcal{E}_s \odot \boldsymbol{g}_{1,k}\}^{\mathrm{H}} \boldsymbol{G}_2^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \\ &\times \mathrm{diag}\{\mathcal{E}_s \odot \boldsymbol{g}_{1,k}\} + \mathrm{diag}\Big\{ \boldsymbol{G}_2^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \\ &\times \mathrm{diag}\{\mathcal{E}_s(\mathbf{1}_N - \mathcal{E}_s)\} \left[\boldsymbol{G}_1 \boldsymbol{Q} \boldsymbol{G}_1^{\mathrm{H}}\right]_{\mathrm{diag}} \Big\}, \end{aligned} \tag{50}$$

$$\boldsymbol{\alpha} = \rho \left(\mathrm{diag}\left\{ \boldsymbol{G}_1 \boldsymbol{Q} (\boldsymbol{H}_0^{\mathrm{H}} \boldsymbol{\Phi}^{\mathrm{H}} - \boldsymbol{I}) \boldsymbol{\Sigma}^{-1} \boldsymbol{\Phi} \boldsymbol{G}_2 \right\} \right)^{*}. \tag{51}$$

- *Turbo message passing algorithm:* First, the initialization of $\{\check{s}_n\}$ and $\{\upsilon_{s_n}\}$ in Algorithm 5 is based on the PDF in (47) rather than in (5). Second, the variance of each $\tilde{w}_{\boldsymbol{s},mt}$ in (67) is changed to

$$\begin{aligned}\upsilon_{\tilde{w}_{\boldsymbol{s},mt}} = &\sum_{k=1}^{K} \upsilon_{x_{kt}} \left| \sum_{l=1}^{L} \left(\sum_{n=1}^{N_l} \rho_l h_{n,mk} \right) \right|^2 + \sum_{k'=1}^{K} \sum_{l'=1}^{L} \\ &\times \left(\sum_{n'=1}^{N_l} \rho_l (1-\rho_l) \upsilon_{x_{k't}} |h_{n',mk}|^2 \right) + \sigma_w^2. \end{aligned} \tag{52}$$

By the above modifications, we obtain the corresponding designs for the multi-RIS system.

VII. Numerical Results

A. Generation Model of the Channel

We consider a three-dimensional (3D) Cartesian coordinate system. The location of the reference antennas/elements at the users is $(0, 0, h_{\mathrm{BS}})$, that of the lth RIS is $(x_{\mathrm{R},l}, y_{\mathrm{R},l}, h_{\mathrm{R},l})$, and that of the kth user is $(x_{\mathrm{U},k}, y_{\mathrm{U},k}, h_{\mathrm{U},k}), k \in \mathcal{I}_K$. The distance-dependent larger-scale fading factor for all channel links[5] is modeled as $\beta = \beta_0 \left(\frac{d}{d_0}\right)^{-\alpha}$, where β_0 is the path fading at the reference distance d_0, d is the link distance, and α is the path fading exponent. In simulation, the specific parameter settings are: $h_{\mathrm{BS}} = 50$; for single-RIS case, $x_{\mathrm{R}} = 0$, $y_{\mathrm{R}} = 30$, $h_{\mathrm{R}} = 30$, $x_{\mathrm{U},k} \in [0, 30]$, $y_{\mathrm{U},k} \in [0, 50]$, and $h_{\mathrm{U},k} = 0$, $\forall k \in \mathcal{I}_K$; for multi-RIS case, $x_{\mathrm{R},l} \in [0, 50]$, $y_{\mathrm{R},l} \in [-30, 30]$, $h_{\mathrm{R},l} = 30$, $x_{\mathrm{U},k} \in [0, 80]$, $y_{\mathrm{U},k} \in [-50, 50]$, and $y_{\mathrm{U},k} = 0$, $\forall l \in \mathcal{I}_L$, $\forall k \in \mathcal{I}_K$, where the unit is meter, and all the location parameters are uniformly random over their specific ranges, expect for those with fixed values; $\beta_0 = -30$ dB, $d_0 = 1$, $\alpha = 3.5$ for the user-BS channel link, and $\alpha = 2.2$ for the user-RIS and RIS-BS channel links.

The line of sight (LoS) linkage of $\boldsymbol{H}_0$ may be blocked while there usually exist a plenty of scatterers in the wireless channel. Thus, we model the small-scale fading component of $\boldsymbol{H}_0$ as Rayleigh fading with the elements independently taking over $\mathcal{CN}(0, 1)$. The RIS is installed to ensure that the LoS linkages from the user to the RIS and from the RIS to the BS exist in the practical scenarios. Thus, for the small-scale fading components of $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$, we adopt a Rician fading model to account for both the LoS and non-LoS (NLoS) effects. The specific formula of the Rician fading model is given by

$$\boldsymbol{H} = \left(\sqrt{\frac{\kappa}{1+\kappa}} \bar{\boldsymbol{H}} + \sqrt{\frac{1}{1+\kappa}} \tilde{\boldsymbol{H}} \right), \tag{53}$$

where κ is the Rician factor, $\bar{\boldsymbol{H}}$ denotes the LoS component that remains fixed within a transmission block, and $\tilde{\boldsymbol{H}}$ denotes the NLoS component with its elements independently taken from $\mathcal{CN}(0, 1)$. In simulations, κ is set to 3 dB for the user-RIS channel link and 10 dB for the RIS-BS channel link.

The LoS component $\bar{\boldsymbol{H}}$ of the Rician fading model can be further expressed as follows. Assume that both the BS and the RIS are equipped with a 2D uniform rectangular array. The 2D array steering vector $\mathbf{a}(\vartheta, \psi)$ is given by

$$\mathbf{a}(\vartheta, \psi) = \mathbf{a}_{\mathsf{az}}(\vartheta, \psi) \otimes \mathbf{a}_{\mathsf{el}}(\vartheta, \psi), \tag{54}$$

where ϑ and ψ are the azimuth and elevation angles, respectively, and $\mathbf{a}_{\mathsf{az}}(\vartheta, \psi) \in \mathbb{C}^{M_1 \times 1}$ and $\mathbf{a}_{\mathsf{el}}(\vartheta, \psi) \in \mathbb{C}^{M_2 \times 1}$ are the uniform linear array (ULA) steering vector given by

$$\left[\mathbf{a}_{\mathsf{az}}(\vartheta, \psi)\right]_n = e^{-j\frac{2\pi d(n-1)}{\varrho} \sin(\vartheta)\cos(\psi)}, \quad \forall n \in \mathcal{I}_{N_1} \tag{55}$$

$$\left[\mathbf{a}_{\mathsf{el}}(\vartheta, \psi)\right]_n = e^{j\frac{2\pi d(n-1)}{\varrho} \sin(\vartheta)\cos(\psi)}, \quad \forall n \in \mathcal{I}_{N_2} \tag{56}$$

with ϱ being the wavelength of propagation and d being the distance of any two adjacent antennas. Denote by $\mathbf{a}(\vartheta_{l,k}^r, \psi_{l,k}^r)$, $\mathbf{a}(\vartheta_l^t, \psi_l^t)$, and $\mathbf{a}(\vartheta_l^r, \psi_l^r)$ the arrival steering vector of RIS^l from user $k, \forall k \in \mathcal{I}_K$, the departure steering vector of RIS^l to the BS, $\forall l \in \mathcal{I}_L$, and the arrival steering vector of the BS from RIS^l, $\forall l \in \mathcal{I}_L$, respectively. $\{\vartheta_{l,k}^r\}$, $\{\psi_{l,k}^r\}$, $\{\vartheta_l^t\}$, $\{\psi_l^t\}$, $\{\vartheta_l^r\}$, and $\{\psi_l^r\}$ are determined by the locations of the users, the RISs, and the BS. For example, $\vartheta_{l,k}^r = \arctan \frac{x_{\mathrm{RIS}^l} - x_{\mathrm{U}_k}}{y_{\mathrm{RIS}^l} - y_{\mathrm{U}_k}}$ and $\psi_{l,k}^r = \arctan \frac{-h_{\mathrm{RIS}^l}}{\sqrt{(x_{\mathrm{U}_k} - x_{\mathrm{RIS}^l})^2 + (y_{\mathrm{U}_k} - y_{\mathrm{RIS}^l})^2}}$. The expressions of the other angles can be obtained similarly and are omitted for brevity. The other parameters are set as: $\frac{d}{\varrho} = \frac{1}{2}$; the dimension of the 2D array of the RIS^l is $16 \times \frac{N_l}{16}$, $\forall l \in \mathcal{I}_l$; and the dimension of the 2D array of the BS is $8 \times \frac{M}{8}$.

B. Simulations for P-BF Design

In this subsection, we present numerical results to validate the efficiency of the proposed P-BF algorithms. The parameter settings for the P-BF design are: $M = 32$, $K = 4$, $p_k = 1, \forall k$, $\ell_{\boldsymbol{s}} = 100$, $\ell_{\boldsymbol{\theta}} = 10$, $\epsilon_{MM} = 10^{-4}$, $iter_{MM} = 100$, $\epsilon_{bf1} = 10^{-6}$, $iter_{bf1} = 5000$, $\epsilon_{bf1} = 10^{-2}$, and $iter_{bf2} = 50$.

[5]There are two competing models on the large-scale fading of the reflect channel links (i.e., user-RIS link and RIS-BS link) in the existing literature. One model assumes that the RIS can work as a specular reflector and thus the distance-dependent large-scale fading factor $\beta \propto \frac{1}{(d_1+d_2)^2}$, where d_1 and d_2 are the link distances of user-to-RIS and RIS-to-BS, respectively [4], [5]. The other model claims that $\beta \propto \frac{1}{(d_1 d_2)^2}$ under a practical communication scenario [28]. In fact, the choice of a model to characterize the path-loss conditions depends on the ratio of the front-wave size and the actual reflector size. To achieve specular reflection, each RIS element is required to have the size of at least $10\lambda \times 10\lambda$ under the near-filed assumption [29], where λ is the wavelength of the incident electromagnetic wave. In this paper, we follow [7], [20], [30] to take the large-scale fading model as $\beta \propto \frac{1}{(d_1 d_2)^2}$ in simulations.

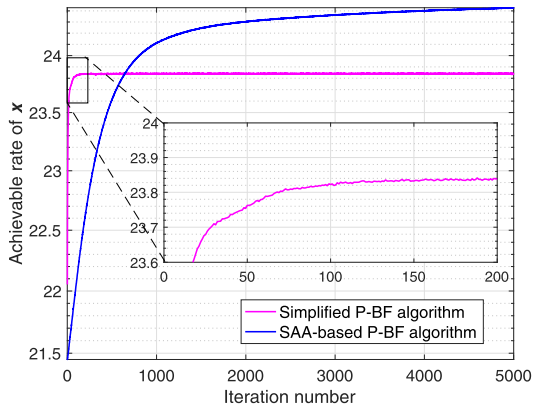


Fig. 5. Comparison of the achievable rate of $\boldsymbol{x}$ versus the iteration number in the single-RIS system with the SAA-based P-BF algorithm and the simplified P-BF algorithm. $N = 32$ and $\rho = 0.5$.

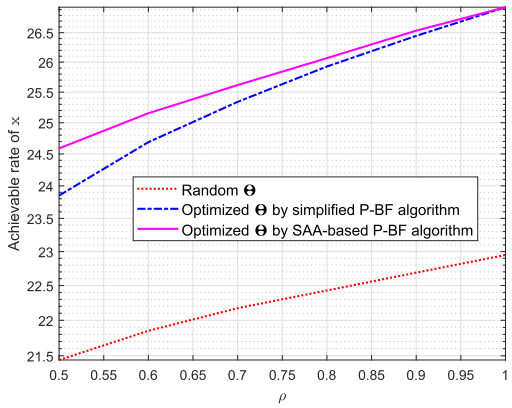


Fig. 6. Achievable rate of $\boldsymbol{x}$ versus ρ in the single-RIS system under different $\boldsymbol{\Theta}$ with $N = 32$.

We consider a system bandwidth with 10 MHz. The noise power spectrum density is -160 dBm/Hz, and the signal power is 0 dBm, yielding $\sigma^2 = -90$ dBm and SNR $= 90$ dB. The simulation results are obtained by taking average over 100 random realizations. The achievable rate of $\boldsymbol{x}$ is calculated by using (8a).

Fig. 5 compares the achievable rate of $\boldsymbol{x}$ versus the iteration number in the single-PBIT system with the SAA-based P-BF algorithm and the simplified P-BF algorithm. We see that the SAA-based P-BF algorithm outperforms the simplified P-BF algorithm by about 0.6 bit per channel use with $N = 32$, and $\rho = 0.5$ at the iteration number $= 5000$. We also see that the simplified P-BF algorithm achieves convergence at the iteration number ≈ 100 and is at least 50 times faster than the SAA-based P-BF algorithm. This demonstrates a good tradeoff between the performance and the computational cost.

Fig. 6 compares the achievable rate of $\boldsymbol{x}$ versus ρ in the single-RIS system with random $\boldsymbol{\Theta}$, optimized $\boldsymbol{\Theta}$ by the simplified P-BF algorithm, and optimized $\boldsymbol{\Theta}$ by the SAA-based P-BF algorithm. The number of the RIS elements is set to $N = 32$. We see that the optimization gains of the simplified P-BF algorithm and the SAA-based P-BF algorithm are at least 2.5 bits per channel use throughout the range of ρ under consideration. We also see that the performance of the simplified P-BF algorithm becomes closer to that of the SAA-based P-BF algorithm with the increase of ρ, and reaches the SAA-based P-BF curve at $\rho = 1$. Fig. 6 shows a good tradeoff between the achievable rate of $\boldsymbol{x}$ and the information rate of the RIS. We see that with ρ increasing from 0.5 to 1, the achievable rate of $\boldsymbol{x}$ increases about 2.5 to 3 bit per channel use for all the three cases of $\boldsymbol{\Theta}$, and the RIS information rate decreases from $\frac{32}{50}$ bit per channel use to zero. In particular, for $\rho = 0.9$ (corresponding to the RIS information rate of $\frac{N}{T}H(\rho) \approx 0.3$ bit per channel use), the achievable rate of $\boldsymbol{x}$ is 26.5 bits per channel use for the optimized $\boldsymbol{\Theta}$ by the SAA-based P-BF algorithm. Compared to the case of $\rho = 1$, the rate loss of $\boldsymbol{x}$ is only about 0.4 bit per channel use. Therefore, a favorable balance between the user information transfer and the RIS information transfer can be achieved by adjusting the value of ρ.

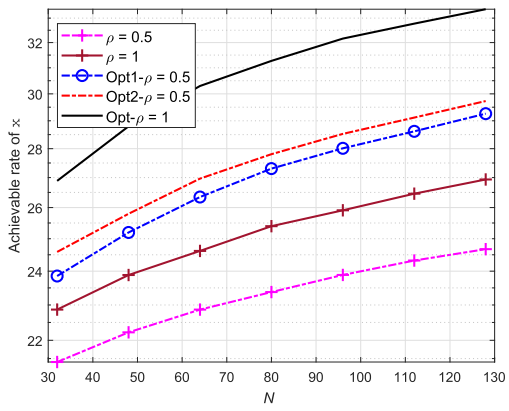


Fig. 7. Achievable rate of $\boldsymbol{x}$ versus N in the single-RIS system.

Fig. 7 studies the achievable rate of $\boldsymbol{x}$ in the single-RIS system with N varying from 32 to 128. We consider the following five scenarios: 1) $\rho = 0.5$: random $\boldsymbol{\Theta}$ (i.e., $\boldsymbol{\Theta}$ is randomly taken from a uniform distribution $[0, 2\pi)$) with $\rho = 0.5$; 2) $\rho = 1$: random $\boldsymbol{\Theta}$ with $\rho = 1$; 3) Opt1-$\rho = 0.5$: $\boldsymbol{\Theta}$ is optimized by the simplified P-BF algorithm with $\rho = 0.5$; 4) Opt2-$\rho = 0.5$: $\boldsymbol{\Theta}$ is optimized by the SAA-based P-BF algorithm with $\rho = 0.5$; 5) Opt-$\rho = 1$: $\boldsymbol{\Theta}$ is optimized by the SAA-based P-BF algorithm with $\rho = 1$. (Note that the above two P-BF algorithms achieve the same performance at $\rho = 1$.) We see that for a random $\boldsymbol{\Theta}$, the rate loss of $\rho = 0.5$ as compared to the case of $\rho = 1$ is 2 bits per channel use throughout the range of N under consideration. For optimized $\boldsymbol{\Theta}$, the corresponding rate loss is increased to 4 bits per channel use. Note that these losses are the cost for the RIS information transfer at the rate of $\frac{N}{T}H(\rho = 0.5) = \frac{N}{50}$ bits per channel use. We also see that the rate gap between the simplified P-BF algorithm and the SAA-based P-BF algorithm is about 0.5 bit per channel use throughout $N \in [32, 128]$. Since the computational cost of the SAA-based P-BF algorithm quickly becomes unaffordable as N increases, in the remaining figures, we only present the simulation results of the simplified P-BF algorithm.

Fig. 8 compares the achievable rate of $\boldsymbol{x}$ in the multi-RIS system with the number of RISs L varying from 1 to 5.

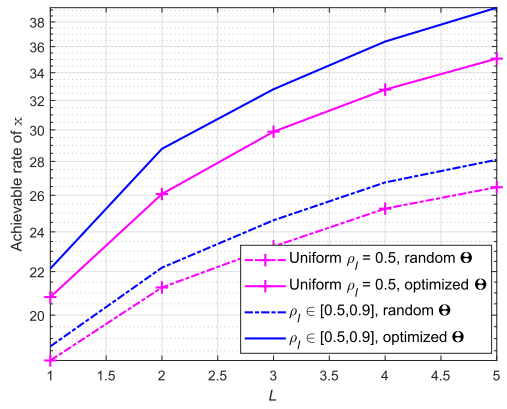


Fig. 8. Achievable rate of $\boldsymbol{x}$ versus L in the multi-RIS system.

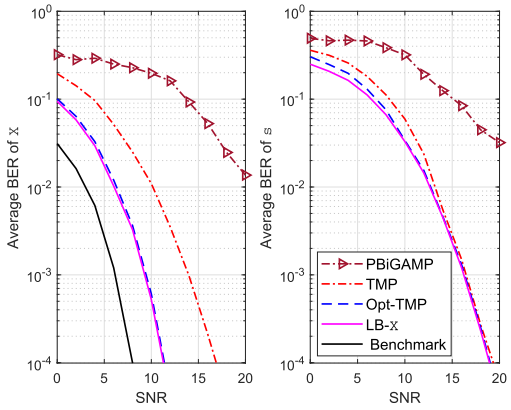


Fig. 9. The BERs of $\boldsymbol{X}$ and $\boldsymbol{s}$ versus the SNR in the single-RIS system with $N = 128$, $M = 32$, and $K = 4$.

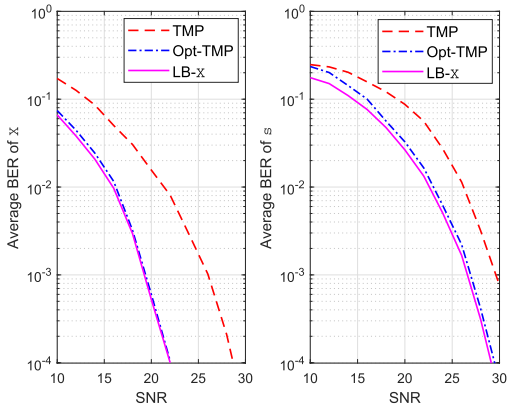


Fig. 10. The BERs of $\boldsymbol{X}$ and $\boldsymbol{s}$ versus the SNR in the multi-RIS system with $\rho \in [0.5, 0.9]$, $L = 3$, $M = 32$, and $K = 4$.

We consider the following four scenarios: 1) Uniform $\rho_l = 0.5, \forall l \in \mathcal{I}_L$ (uniform-ρ_l) with random $\boldsymbol{\Theta}$; 2) Uniform-ρ_l with optimized $\boldsymbol{\Theta}$; 3) $\rho_l, \forall l \in \mathcal{I}_L$ is randomly taken from a uniform distribution over $[0.5, 0.9]$ (random-ρ_l) with random $\boldsymbol{\Theta}$; 4) random-ρ_l with optimized $\boldsymbol{\Theta}$. We see that for uniform-ρ_l case, the optimization gain increases from about 3 bits per channel use to about 8 bits per channel use with L increasing from 1 to 5. For random-ρ_l case, the optimization gain increases from about 3 bits per channel use to about 10 bits per channel use with L increasing from 1 to 5. We also see that the random-ρ_l case outperforms the uniform-ρ_l case for both random $\boldsymbol{\Theta}$ and optimized $\boldsymbol{\Theta}$. In addition, the rate gap increases with L varying from 1 to 5.

C. Simulations for Detector Design

In this subsection, we present numerical results to validate the efficiency of the proposed the TMP algorithm. In simulations, the elements of $\boldsymbol{X}$ are randomly taken from the quadrature phase shift keying modulation with Gray-mapping. We set $N_1 = \ldots = N_L = 32$ in the multi-RIS cases. The parameter settings for the $\boldsymbol{X}$-detector are: $I^x_{max} = 10$, $\ell^x_{max} = 200$, $\xi^x_x = 0.6$, $\xi^x_u = \frac{2[(2-\xi^x_x)K+\xi^x_x M]}{1.1\xi^x_x MK\|\boldsymbol{H}\|^2_2/\|\boldsymbol{H}\|^2_{\rm F}}$, $\epsilon^x_{\rm gamp} = 10^{-10}$, and $\epsilon^x_{\rm em} = 10^{-10}$. The parameter settings for the $\boldsymbol{s}$-detector are: $I^s_{max} = 10$, $\ell^s_{max} = 1000$, $\xi^s_x = 0.2$, $\xi^s_u = \frac{2[(2-\xi^s_x)N+\xi^s_x MT]}{1.1\xi^s_x MTN\|\boldsymbol{A}\|^2_2/\|\boldsymbol{A}\|^2_{\rm F}}$, $\epsilon^s_{\rm gamp} = 10^{-10}$ and $\epsilon^s_{\rm em} = 10^{-10}$. The other settings are $\tau_{max} = 20$, $\epsilon_{td} = 10^{-8}$. The presented simulation results are obtained by taking average over 100 random realizations. The receiver SNR is based on the case of random $\boldsymbol{\Theta}$ with $\rho = 0.5$.

In simulations, we consider the following four scenarios: 1) PBiGAMP: the receiver detects $\boldsymbol{X}$ and $\boldsymbol{s}$ by the PBiGAMP algorithm [22] with optimized $\boldsymbol{\Theta}$; 2) TMP: the receiver detects $\boldsymbol{X}$ and $\boldsymbol{s}$ by the turbo message passing algorithm proposed in this paper with random $\boldsymbol{\Theta}$; 3) Opt-TMP: the receiver detects $\boldsymbol{X}$ and $\boldsymbol{s}$ by the turbo message passing algorithm proposed in this paper with $\boldsymbol{\Theta}$ optimized by the simplified P-BF algorithm; 4) Lower bound for detecting $\boldsymbol{X}/\boldsymbol{s}$ (LB-$\boldsymbol{X}/\boldsymbol{s}$): the receiver detects $\boldsymbol{X}$ by the GGAMP-SBL algorithm [27] with perfect knowledge of $\boldsymbol{s}$ and $\boldsymbol{\Theta}$ optimized by the simplified P-BF algorithm; 5) Lower bound for detecting $\boldsymbol{s}$ (LB-$\boldsymbol{s}$) by the GGAMP-SBL algorithm [27] with perfect knowledge of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$ optimized by the simplified P-BF algorithm.

Fig. 9 compares the BERs of $\boldsymbol{X}$ and $\boldsymbol{s}$ of the single-RIS system versus the SNR with $N = 128$, $M = 32$, and $K = 4$. We consider the PBiGAMP, TMP, Opt-TMP, LB-$\boldsymbol{X}$, and LB-$\boldsymbol{s}$ schemes. We see that the turbo message passing algorithm significantly outperforms the PBiGAMP algorithm and tightly approaches the lower bound for both detecting $\boldsymbol{X}$ and $\boldsymbol{s}$. We also see that, by optimization, the BER of $\boldsymbol{X}$ is improved by about 6 dB at the average BER $= 10^{-4}$.

Fig. 10 compares the BERs of $\boldsymbol{X}$ and $\boldsymbol{s}$ of the multi-PBIT system versus the SNR with $\rho_l \in [0.5, 0.9], \forall l \in \mathcal{I}_L$. The other settings are $L = 3$, $M = 32$, and $K = 4$. We consider the TMP, Opt-TMP, LB-$\boldsymbol{X}$, and LB-$\boldsymbol{s}$ schemes. We see that the turbo message passing algorithm tightly approaches the lower bounds for both detecting $\boldsymbol{X}$ and $\boldsymbol{s}$. We also see that, by optimization, the system achieves about 8 dB performance improvement for detecting $\boldsymbol{X}$ at the average BER $= 10^{-4}$ and about 3 dB performance improvement for detecting $\boldsymbol{s}$ at the average BER $= 10^{-4}$.

VIII. Conclusions

In this paper, we studied the design of P-BF and information transfer for the RIS-aided Mu-MIMO system. For the P-BF design, we formulated the problem as a stochastic program

in which the achievable user sum rate of the RIS-aided Mu-MIMO channel is used as the metric for optimization, and developed the SAA-based P-BF algorithm to solve the problem. To reduce the high computational complexity of the SAA-based P-BF algorithm, we further developed a simplified P-BF algorithm by approximating the stochastic program as a deterministic alternating optimization problem. For the receiver design, we proposed a turbo message passing algorithm to iteratively estimate the user signals and the RIS states. Furthermore, we extended all the designs from the single-RIS case to the multi-RIS case. Numerical results were provided to demonstrate the superior performance of our proposed designs.

The information transfer is an essential function of the RIS in its practical implementation. In this regard, there are many issues need to be studied beyond the work in this paper. For example, the P-BF design in this paper aimed to maximize the user sum rate without considering user fairness, i.e., some remote users may experience very poor quality of service in achieving the maximum system throughput. To impose user fairness, we need to conduct the P-BF optimization under the constraint of a minimum required data rate of each user. Besides, instead of user sum rate, other design criteria, such as energy efficiency and physical-layer security need to be considered in the P-BF design. Moreover, the PBIT scheme developed in this paper achieves RIS information transfer at the cost of compromising the P-BF capability of the RIS. As an alternative, the RIS information can possibly be transferred by using a dedicated transmitter. Clearly, there are pros and cons of the PBIT scheme as compared to the dedicated transmitter strategy for RIS information transfer in terms of cost, complexity, and efficiency. A comprehensive comparison study of these issues desires our future research effort.

APPENDIX A: EQUIVALENCE BETWEEN (7) AND (8)

From the Blahut-Arimoto algorithm in [23], $I(\boldsymbol{x};\boldsymbol{y}|\boldsymbol{s})$ can be written as

$$\begin{aligned} I(\boldsymbol{x};\boldsymbol{y}|\boldsymbol{s}) &= \mathbb{E}_{\boldsymbol{s}}\left(\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}\log\frac{p_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})}{p_{\boldsymbol{x}}(\boldsymbol{x})}\right) \\ &= \mathbb{E}_{\boldsymbol{s}}\left(\max_{q(\cdot|\cdot)}\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}\log\frac{q_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})}{p_{\boldsymbol{x}}(\boldsymbol{x})}\right), \end{aligned} \tag{57}$$

where $q(\cdot|\cdot)$ is an arbitrary distribution of $\boldsymbol{x}$ conditioned on $\boldsymbol{y}$ and $\boldsymbol{s}$. Note that (57) is based on the fact that

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}\log\frac{p_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})}{p_{\boldsymbol{x}}(\boldsymbol{x})} - \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}\log\frac{q_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})}{p_{\boldsymbol{x}}(\boldsymbol{x})} \\ &= \int_{\boldsymbol{y},\boldsymbol{x}} p_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})\log\frac{p_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})}{q_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})} \\ &= \mathrm{KL}\left(p_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})\|q_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s})\right) \geq 0, \end{aligned} \tag{58}$$

Define an auxiliary variable $\boldsymbol{z} = (\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}\boldsymbol{G}_1 + \boldsymbol{H}_0)\boldsymbol{x}$ and recall that $\boldsymbol{w}$ is an AWGN with the elements independently drawn from $\mathcal{CN}(0,\sigma_w^2)$. Then, we have

$$p_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{s}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{s}) = \frac{1}{(\pi\sigma_w^2)^M}\exp\left(-\frac{1}{\sigma_w^2}(\boldsymbol{y}-\boldsymbol{z})^{\mathrm{H}}(\boldsymbol{y}-\boldsymbol{z})\right). \tag{59}$$

Since $I(\boldsymbol{x};\boldsymbol{y}|\boldsymbol{s})$ is maximized when $\boldsymbol{x}$ is Gaussian, we only need to consider the Gaussian input:

$$p_{\boldsymbol{x}}(\boldsymbol{x}) = \frac{1}{\pi^K\det(\boldsymbol{Q})}\exp(-\boldsymbol{x}^{\mathrm{H}}\boldsymbol{Q}^{-1}\boldsymbol{x}), \tag{60}$$

where $\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}^{\mathrm{H}}\right)$ is the covariance matrix of $\boldsymbol{x}$. Note that $\boldsymbol{Q}$ is a diagonal matrix with the kth diagonal element given by $P_k, \forall k \in \mathcal{I}_K$. The probability distribution of $\boldsymbol{s}$ is given in (5). The conditional PDF $p_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s})$ is given by

$$\begin{aligned} p_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}) = &\frac{1}{(\pi\sigma_w^2)^M}\exp\left(-\frac{1}{\sigma_w^2}(\boldsymbol{y}-\boldsymbol{z})^{\mathrm{H}}(\boldsymbol{y}-\boldsymbol{z})\right) \\ &\times\frac{1}{\pi^K\det(\boldsymbol{Q})}\exp(-\boldsymbol{x}^{\mathrm{H}}\boldsymbol{Q}^{-1}\boldsymbol{x}). \end{aligned} \tag{61}$$

From [31], the optimal choice of $q(\cdot|\cdot)$ follows the Gaussian distribution and can be written as

$$\begin{aligned} &q_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{s}) \\ &\quad= \frac{1}{\pi^K\det(\boldsymbol{\Sigma})}\exp\left(-(\boldsymbol{x}-\boldsymbol{\Phi}\boldsymbol{y})^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\Phi}\boldsymbol{y})\right), \end{aligned} \tag{62}$$

where $\boldsymbol{\Phi}\boldsymbol{y}$ and $\boldsymbol{\Sigma} \succeq \boldsymbol{0}$ are the mean and variance of the Gaussian distribution. Plugging (60) and (62) into (6), we obtain

$$\begin{aligned} \max_{q(\cdot|\cdot)} I(\boldsymbol{x};\boldsymbol{y}|\boldsymbol{s}) = \mathbb{E}_{\boldsymbol{s}}\Big[\max_{\boldsymbol{\Phi},\boldsymbol{\Sigma}\succeq\boldsymbol{0}}\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}\Big(&\log\det(\boldsymbol{Q}) - \log\det(\boldsymbol{\Sigma}) \\ &- (\boldsymbol{x}-\boldsymbol{\Phi}\boldsymbol{y})^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\Phi}\boldsymbol{y}) + \boldsymbol{x}^{\mathrm{H}}\boldsymbol{Q}^{-1}\boldsymbol{x}\Big)\Big]. \end{aligned} \tag{63}$$

Then, we obtain (8) by omiting the terms $\mathbb{E}_{\boldsymbol{s}}\left[\max_{\boldsymbol{\Phi},\boldsymbol{\Sigma}}\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}\log\det(\boldsymbol{Q})\right]$ and $\mathbb{E}_{\boldsymbol{s}}\left[\max_{\boldsymbol{\Phi},\boldsymbol{\Sigma}}\mathbb{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{s}}\boldsymbol{x}^{\mathrm{H}}\boldsymbol{Q}^{-1}\boldsymbol{x}\right]$ since they are irrelevant to $\boldsymbol{\Phi}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Theta}$.

APPENDIX B

From the model in (1), we obtain

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{y},\boldsymbol{x},\boldsymbol{s}}\left\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\Phi}\boldsymbol{y})\right\|_2^2 \\ &= \mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1}\Big[\boldsymbol{Q} - \boldsymbol{\Phi}\boldsymbol{H}_0\boldsymbol{Q} - (\boldsymbol{\Phi}\boldsymbol{H}_0\boldsymbol{Q})^{\mathrm{H}} + \boldsymbol{\Phi}\boldsymbol{H}_0\boldsymbol{Q}\boldsymbol{H}_0^{\mathrm{H}}\boldsymbol{\Phi}^{\mathrm{H}} \\ &\quad + \sigma_w^2\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{H}} + \mathbb{E}_{\boldsymbol{s}}\Big(\boldsymbol{\Phi}\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}\boldsymbol{G}_1\boldsymbol{Q}\boldsymbol{G}_1^{\mathrm{H}}\boldsymbol{S}^{\mathrm{H}}\boldsymbol{\Theta}^{\mathrm{H}}\boldsymbol{G}_2^{\mathrm{H}}\boldsymbol{\Phi}^{\mathrm{H}} \\ &\quad + \boldsymbol{\Phi}\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}\boldsymbol{G}_1\boldsymbol{Q}\boldsymbol{H}_0^{\mathrm{H}}\boldsymbol{\Phi}^{\mathrm{H}} + (\boldsymbol{\Phi}\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}\boldsymbol{G}_1\boldsymbol{Q}\boldsymbol{H}_0^{\mathrm{H}}\boldsymbol{\Phi}^{\mathrm{H}})^{\mathrm{H}} \\ &\quad - \boldsymbol{\Phi}\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}\boldsymbol{G}_1\boldsymbol{Q} - (\boldsymbol{\Phi}\boldsymbol{G}_2\boldsymbol{\Theta}\boldsymbol{S}\boldsymbol{G}_1\boldsymbol{Q})^{\mathrm{H}}\Big)\Big]\Big\}. \end{aligned} \tag{64}$$

Based on the probability distribution of $\boldsymbol{s}$ in (5), we obtain

$$\mathbb{E}_{\boldsymbol{s}}(\boldsymbol{S}) = \rho\boldsymbol{I},\ \mathbb{E}_{\boldsymbol{s}}\left(\boldsymbol{S}\boldsymbol{X}\boldsymbol{S}^{\mathrm{H}}\right) = \rho^2\boldsymbol{X} + \rho(1-\rho)\left[\boldsymbol{X}\right]_{\mathrm{diag}}. \tag{65}$$

Plugging (13) and (65) into (64), we obtain (28).

APPENDIX C

A. Derivation of (40)

Denote by $\boldsymbol{h}_{n,m}^{\mathrm{T}}$ the mth row of $\boldsymbol{H}_n$ and let $\boldsymbol{B}_m \triangleq [\boldsymbol{h}_{0,m},\ldots,\boldsymbol{h}_{N,m}]^{\mathrm{T}} \in \mathbb{C}^{(N+1)\times K}, \forall m \in \mathcal{I}_M$. Then, each

element of $\tilde{\boldsymbol{W}}_{\boldsymbol{X}}$ can be written as $\tilde{w}_{\boldsymbol{X},mt} = (\tilde{\boldsymbol{s}})^{\mathrm{T}} \boldsymbol{B}_m \boldsymbol{x}_t + w_{mt}, \forall m, t$. The variance of $\tilde{w}_{\boldsymbol{X},mt}$ is

$$\begin{aligned}
\upsilon_{\tilde{w}_{\boldsymbol{X},mt}} &= \mathbb{E}\left[\left(\tilde{\boldsymbol{s}}^{\mathrm{T}}\boldsymbol{B}_m\boldsymbol{x}_t + w_{mt}\right)\left(\tilde{\boldsymbol{s}}^{\mathrm{T}}\boldsymbol{B}_m\boldsymbol{x}_t + w_{mt}\right)^{\mathrm{H}}\right] \\
&= \mathbb{E}\left(\tilde{\boldsymbol{s}}^{\mathrm{T}}\boldsymbol{B}_m\boldsymbol{Q}\boldsymbol{B}_m^{\mathrm{H}}\tilde{\boldsymbol{s}}\right) + \sigma_w^2 \\
&= \sum_{i=0}^{N}\sum_{j=0}^{N}\mathbb{E}\left(\tilde{s}_i\tilde{s}_j\right)\left(\boldsymbol{h}_{i,m}\boldsymbol{Q}\boldsymbol{h}_{j,m}^{\mathrm{H}}\right) + \sigma_w^2 \\
&= \sum_{i=0}^{N}\sum_{k=1}^{K}\upsilon_{s_i}|h_{i,mk}|^2 p_k + \sigma_w^2, \quad \forall m, t, \qquad (66)
\end{aligned}$$

where $\mathbb{E}\left(\tilde{s}_i\tilde{s}_j\right) = \upsilon_{s_i}$ for $i = j \in \mathcal{I}_N$ and $\mathbb{E}\left(\tilde{s}_i\tilde{s}_j\right) = 0$ otherwise.

B. Derivation of (44)

$$\begin{aligned}
\upsilon_{\tilde{\boldsymbol{w}}_{\boldsymbol{s},mt}} &= \mathbb{E}\left[\left(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{B}_m\tilde{\boldsymbol{x}}_t + w_{mt}\right)^{\mathrm{H}}\left(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{B}_m\tilde{\boldsymbol{x}}_t + w_{mt}\right)\right] \\
&= \mathbb{E}\left[\tilde{\boldsymbol{x}}_t^{\mathrm{H}}\boldsymbol{B}_m^{\mathrm{H}}\left(\rho^2\mathbf{1}_N\cdot\mathbf{1}_N^{\mathrm{T}} + \rho(1-\rho)\mathbf{I}\right)\boldsymbol{B}_m\tilde{\boldsymbol{x}}_t\right] + \sigma_w^2 \\
&= \rho^2\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left(\tilde{x}_{it}^*\tilde{x}_{jt}\right)\left(\mathbf{1}_N^{\mathrm{T}}\boldsymbol{B}_m\right)_i^*\left(\mathbf{1}_N^{\mathrm{T}}\boldsymbol{B}_m\right)_j \\
&\quad + \rho(1-\rho)\sum_{i'=1}^{K}\sum_{j'=1}^{K}\mathbb{E}\left(\tilde{x}_{i't}^*\tilde{x}_{j't}\right)\boldsymbol{b}_{m,i'}^{\mathrm{H}}\boldsymbol{b}_{m,j'} + \sigma_w^2 \\
&= \rho^2\sum_{k=1}^{K}\upsilon_{x_{kt}}\left|\sum_{n=1}^{N}h_{n,mk}\right|^2 + \rho(1-\rho) \\
&\quad \times\sum_{k'=1}^{K}\sum_{n'=1}^{N}\upsilon_{x_{k't}}|h_{n',mk}|^2 + \sigma_w^2, \qquad (67)
\end{aligned}$$

where $\mathbb{E}\left[\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}}\right] = \rho^2\mathbf{1}_N\cdot\mathbf{1}_N^{\mathrm{T}} + \rho(1-\rho)\mathbf{I}$, $\mathbb{E}\left(\tilde{x}_{it}^*\tilde{x}_{jt}\right) = \upsilon_{x_{it}}$ for $i = j \in \mathcal{I}_K$ and $\mathbb{E}\left(\tilde{x}_{it}^*\tilde{x}_{jt}\right) = 0$ otherwise, and $\boldsymbol{b}_{m,k}$ is the kth column of $\boldsymbol{B}_m$.

REFERENCES

[1] W. Yan, X. Yuan, Z.-Q. He, and X. Kuai, "Large intelligent surface aided multiuser MIMO: Passive beamforming and information transfer," in *Proc. IEEE Int. Conf. Commun. (ICC)*, to be published.

[2] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] S. V. Hum and J. Perruisseau-Carrier, "Reconfigurable reflectarrays and array lenses for dynamic antenna beam control: A review," *IEEE Trans. Antennas Propag.*, vol. 62, no. 1, pp. 183–198, Jan. 2014.

[4] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," *J. Wireless Commun. Netw.*, vol. 2019, p. 129, May 2019, doi: 10.1186/s13638-019-1438-9.

[5] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, Aug. 2019.

[6] Q.-U.-U. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," 2019, *arXiv:1903.08127*. [Online]. Available: http://arxiv.org/abs/1903.08127

[7] 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.

[8] 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.

[9] L. Subrt and P. Pechac, "Intelligent walls as autonomous parts of smart indoor environments," *IET Commun.*, vol. 6, no. 8, pp. 1004–1010, May 2012.

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

[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] X. Yuan, T. Yang, and I. B. Collings, "Multiple-input multiple-output two-way relaying: A space-division approach," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6421–6440, Oct. 2013.

[13] S. Hu, F. Rusek, and O. Edfors, "Beyond massive MIMO: The potential of data transmission with large intelligent surfaces," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2746–2758, May 2018.

[14] Q. Zhang, H. Guo, Y.-C. Liang, and X. Yuan, "Constellation learning-based signal detection for ambient backscatter communication systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 2, pp. 452–463, Feb. 2019.

[15] 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

[16] 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.

[17] X. Yuan, Y.-J. Zhang, Y. Shi, W. Yan, and H. Liu, "Reconfigurable-Intelligent-Surface empowered 6G wireless communications: Challenges and opportunities," 2020, *arXiv:2001.00364*. [Online]. Available: http://arxiv.org/abs/2001.00364

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

[19] S. Zhang and R. Zhang, "Capacity characterization for intelligent reflecting surface aided MIMO communication," 2019, *arXiv:1910.01573*. [Online]. Available: http://arxiv.org/abs/1910.01573

[20] C. Pan *et al.*, "Multicell MIMO communications relying on intelligent reflecting surface," 2019, *arXiv:1907.10864*. [Online]. Available: http://arxiv.org/abs/1907.10864

[21] W. Gropp *et al.*, *Using MPI: Portable Parallel Programming With the Message-Passing Interface*. Cambridge, MA, USA: MIT Press, 1999.

[22] J. T. Parker and P. Schniter, "Parametric bilinear generalized approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 795–808, Jun. 2016.

[23] T. M. Cover and J. A. Thomas, *Elements Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[24] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM J. Optim.*, vol. 12, no. 2, pp. 479–502, Jan. 2002.

[25] L. Dai, C. H. Chen, and J. R. Birge, "Convergence properties of two-stage stochastic programming," *J. Optim. Theory Appl.*, vol. 106, no. 3, pp. 489–509, Sep. 2000.

[26] A. Shapiro and T. Homem-de-Mello, "On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs," *SIAM J. Optim.*, vol. 11, no. 1, pp. 70–86, Jan. 2000.

[27] M. Al-Shoukairi, P. Schniter, and B. D. Rao, "A GAMP-based low complexity sparse Bayesian learning algorithm," *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 294–308, Jan. 2018.

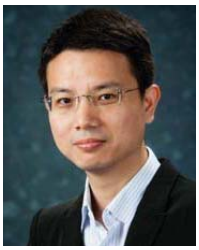
[28] Ö. Özdogan, E. Björnson, and E. G. Larsson, "Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling," 2019, *arXiv:1911.03359*. [Online]. Available: http://arxiv.org/abs/1911.03359

[29] W. Tang *et al.*, "Wireless communications with reconfigurable intelligent surface: Path loss modeling and experimental measurement," 2019, *arXiv:1911.05326*. [Online]. Available: http://arxiv.org/abs/1911.05326

[30] E. Björnson, Ö. Özdogan, and E. G. Larsson, "Intelligent reflecting surface vs. Decode-and-forward: how large surfaces are needed to beat relaying?" 2019, *arXiv:1906.03949*. [Online]. Available: http://arxiv.org/abs/1906.03949

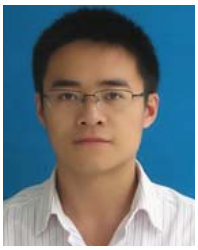
[31] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

Wenjing Yan (Graduate Student Member, IEEE) received the B.S. degree in electronic and information engineering from Chongqing University, Chongqing, China, in 2018. She is currently pursuing the M.S. degree in electronic and communication engineering with the University of Electronic Science and Technology of China, Chengdu, China.



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 Hawai'i at Mānoa, in Spring and Summer of 2009 and in the same period of 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, supported by the Thousand Youth Talents Plan in China. His research interests include broad range of signal processing, machine learning, and wireless communications, including but not limited to multiantenna and cooperative communications, sparse and structured signal recovery, Bayesian approximate inference, network coding, and so on. He has published over 160 peer-reviewed research papers in the leading international journals and conferences in the related areas. He has served on a number of technical programs for international conferences. He was a co-recipient of the Best Paper Award of the IEEE International Conference on Communications (ICC) 2014 and the Best Journal Paper Award of the IEEE Technical Committee on Green Communications and Computing (TCGCC) 2017. He has been an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS since 2017 and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2018.



Zhen-Qing He (Member, IEEE) received the Ph.D. degree in communication and information system from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2017. From 2015 to 2016, he was a Visiting Ph.D. Student with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA. Since 2018, he has been a Post-Doctoral Researcher with the National Key Laboratory of Science and Technology on Communications, UESTC. His main research interests include statistical signal processing, wireless communications, and machine learning.

Xiaoyan Kuai (Member, IEEE) received the Ph.D. degree in communication engineering from Xiamen University in 2017. From 2014 to 2016, she was a Visiting Student with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. From 2017 to 2019, she was a Post-Doctoral Researcher with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China. Her general research interests include the areas of signal processing and communications and machine learning.