Design of Per-Antenna Constant-Envelope Precoding for Massive MIMO Communications With Beyond Diagonal Reconfigurable Intelligent Surface

Chunxuan Shi, Yongzhe Li, and Ran Tao
School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
Emails: cxshi@bit.edu.cn, lyz@ieee.org/yongzhe.li@bit.edu.cn, rantao@bit.edu.cn

***Abstract*—This paper addresses the design of per-antenna constant-envelope precoding for massive MIMO communications, where a beyond diagonal reconfigurable intelligent surface (BD-RIS) is considered to be involved. The primary goal of the design is to mitigate the multi-user interference for communications, and the reflection-coefficient matrix of the BD-RIS is required to be both unitary and symmetric. Meanwhile, we also impose constant-modulus constraints for the precoding at each antenna. These considerations culminate in a non-convex optimization problem characterized by a min-max type objective and multiple intricate constraints. To solve the design problem, we propose a cyclic alternating framework for finding solutions. For each subproblem arising from the alternating optimization framework, we reformulate the min-max problem into a tractable form, and subsequently devise a fixed-point iteration scheme via proximal splitting. These enable us to obtain a closed-form solution at each iteration. Furthermore, we provide a theoretical analysis that demonstrates convergence to stable points for the developed update procedures. Simulation results confirm that the proposed approach outperforms existing methods across various aspects.**

***Index Terms*—Beyond diagonal reconfigurable intelligent surface (BD-RIS), constant-envelope precoding, proximal splitting.**

I. Introduction

Precoding has been a research field of significant interest for communications during the past several decades [1]–[4]. It has emerged as a leading approach to address various challenges, whose importance is particularly pronounced in massive MIMO communications [5]. For example, the precoding with a large-scale antenna array benefits the communications from tackling issues related to the increasing number of radio-frequency chains [6]. Through proper waveform/code designs, the radio-frequency chains can avoid signal distortion by means of restricting the total transmit power [7], peak-to-average-power ratio (PAPR) [8], and magnitudes of signal elements [9]. On the other hand, the large degrees of freedom enabled by massive MIMO communications help the precoding to obtain potential improvements on symbol error rate (SER) [10], signal-to-noise ratio (SNR) [11], multi-user interference mitigation [12], etc.

Among relevant works on precoding in recent years, the ones that focus on constant-envelope precoding have attracted considerable attention [13], [14]. This type of precoding enforces all the elements of transmit signals to have a same magnitude, which leads to low PAPRs of waveform [14]. However, when it comes to harsh environment with undesired channel responses, the conventional constant-envelope precoding may fail to meet the requirement of high-accuracy communications. To address this issue, precoding with the aid of reconfigurable intelligent surface (RIS) has emerged as a new trend [15], [16]. By means of adjusting the overall channel response, the RIS can support constant-envelope precoding to further explore performance gains in massive MIMO communications.

Technically, the performance of RIS-aided precoding is subject to the physical form of RIS. For example, the early type which controls the elements independently is prone to result in a diagonal reflection-coefficient matrix with limited design flexibility [17]. By contrast, the recently emerged category allows for mutual coupling of elements [18], which enables a beyond diagonal (BD) reflection-coefficient matrix [19]. For the BD-RIS aided constant-envelope precoding, an interesting aspect is to investigate the optimal design for massive MIMO communications, which however, is seldom studied.

In this paper, we study the BD-RIS aided constant-envelope precoding for massive MIMO communications. The goal is to suppress the multi-user interference. To this end, we minimize the maximum difference between the desired and received noise-free symbols among all users by the joint design of transmit signals and reflection coefficients for the BD-RIS. Meanwhile, we also guarantee the inherent constraints associated with the BD-RIS. A non-convex optimization problem with min-max objective is therefore formulated. To tackle it, we exploit a cyclic manner to find its solutions. For each resulted alternating problem, we first reformulate the "min-max" type optimization into a pure minimization form. Then, we devise a fixed-point iteration rule based on proximal splitting [20]. Our major contribution also lies in obtaining a closed-form solution for updating reflection coefficients of the BD-RIS via Takagi factorization. Moreover, we prove the convergence of devised update procedures. Simulation results verify the superiority of our proposed design in terms of different aspects.

Notations: We use notations $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$, $\otimes$, $|\cdot|$, $\|\cdot\|$, $\|\cdot\|_{\mathrm{F}}$, and ∂ to denote the conjugate, transpose, Hermitian, Kronecker product, modulus, Euclidean norm, Frobenius norm, and sub-differential operations, respectively. Moreover, notations $\mathbb{C}$, $\preceq$, $\arg(\cdot)$, and $\lambda_{\max}(\cdot)$ stand for the complex filed, generalized inequality, argument of a complex value, and largest eigenvalue of a matrix, respectively. In addition, $\mathbf{0}_M$ is an $M \times 1$ vector

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants 62271054 and U21A20456.

with all elements equal to 0, and $\mathbf{I}_M$ stands for the $M \times M$ identity matrix.

II. Signal Model and Problem Formulation

Let us consider a downlink communication system serving K single-antenna users with the aid of an N-element BD-RIS. We assume that the base station of the system is equipped with M antenna elements. Moreover, we denote $\mathbf{x} \triangleq [x_1, \ldots, x_M]^{\mathrm{T}} \in \mathbb{C}^{M\times 1}$ as the vector of transmit symbols for precoding at a certain time, each of which has a constant magnitude.

The received signal observed at the k-th user, denoted by y_k, can be expressed as

$$y_k = s_k + ((\mathbf{d}_k + \mathbf{H}\mathbf{\Phi}\mathbf{h}_k)^{\mathrm{H}}\mathbf{x} - s_k) + w_k \tag{1}$$

where s_k is the desired communication symbol, w_k is the zero-mean white Gaussian noise, both associated with the k-th user, $\mathbf{\Phi} \in \mathbb{C}^{N\times N}$ is the reflection-coefficient matrix associated with the BD-RIS, and $\mathbf{d}_k \in \mathbb{C}^{M\times 1}$, $\mathbf{H} \in \mathbb{C}^{M\times N}$, and $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$ are vectors/matrices that store the channel information associated with paths from the base station to the k-th user, base station to BD-RIS, and BD-RIS to the k-th user, respectively. The second sum component in (1), which corresponds to the difference between the received noise-free and desired symbols, denotes the multi-user interference for the k-th user.

Note that the reflection-coefficient matrix $\mathbf{\Phi}$ for the BD-RIS differs from that for the conventional RIS. The latter requires the reflection-coefficient matrix to be diagonal with non-zero elements being unimodular, while the former only requires to be both unitary and symmetric [19]. Correspondingly, the BD-RIS is capable of enabling extra degrees of freedom to assist precoding compared to the conventional RIS, but it is complex for hardware implementation [18].

The primary objective here is to suppress the largest multi-user interference among all users, and meanwhile, to guarantee the aforementioned conditions on transmit symbols and the BD-RIS. Hence, we can formulate the precoding problem with the aid of BD-RIS in the form as follows

$$\min_{\mathbf{x},\mathbf{\Phi}} \max_{k\in\mathcal{K}} \quad |(\mathbf{d}_k + \mathbf{H}\mathbf{\Phi}\mathbf{h}_k)^{\mathrm{H}}\mathbf{x} - s_k|^2 \tag{2a}$$

$$\text{s.t.} \quad |x_m| = c,\ m \in \mathcal{M} \tag{2b}$$

$$\mathbf{\Phi}^{\mathrm{H}}\mathbf{\Phi} = \mathbf{I}_N \tag{2c}$$

$$\mathbf{\Phi} = \mathbf{\Phi}^{\mathrm{T}} \tag{2d}$$

where $\mathcal{K} \triangleq \{1, \ldots, K\}$, $\mathcal{M} \triangleq \{1, \ldots, M\}$, and c denotes the constant magnitude of transmit symbols. Note that the objective function (2a) originates from the second sum term of (1), the constraints (2b) guarantee the constant-modulus property of the transmit symbols in $\mathbf{x}$, and (2c) and (2d) ensure the unitary and symmetric structures of $\mathbf{\Phi}$, respectively. Overall, the problem (2) takes a "min-max" form, which is non-convex and difficult to solve. Therefore, an efficient solution to (2) is to be found.

III. Joint Design of Constant-Envelope Precoding and Reflection-Coefficient Matrix for BD-RIS

In order to solve (2), we exploit a cyclic manner to optimize $\mathbf{x}$ and $\mathbf{\Phi}$ alternately. For each alternating optimization with respect to $\mathbf{x}$ or $\mathbf{\Phi}$, we first convert the "min-max" type problem into a pure minimization form, and then address it by means of proximal splitting techniques.

A. Per-Antenna Constant-Envelope Precoding to Obtain $\mathbf{x}$

Fixing $\mathbf{\Phi}$ and additionally introducing an auxiliary variable t for the problem (2), we can obtain the alternating optimization problem with respect to $\mathbf{x}$ and t as follows

$$\min_{\mathbf{x},t} \quad t^2 \tag{3a}$$

$$\text{s.t.} \quad |(\mathbf{d}_k + \mathbf{H}\mathbf{\Phi}\mathbf{h}_k)^{\mathrm{H}}\mathbf{x} - s_k|^2 \le t^2,\ k \in \mathcal{K} \tag{3b}$$

$$|x_m| = c,\ m \in \mathcal{M}. \tag{3c}$$

To tackle (3), we exploit the idea that transforms it into an unconstrained problem with the support of introducing indicator functions for (3b) and (3c). To this end, we respectively denote the feasible sets for (3b) and (3c) as $\{\mathcal{C}_k\}_{k=1}^K$ and $\widetilde{\mathcal{C}}$, and we also denote $\{I_{\mathcal{C}_k}(\mathbf{x}, t)\}_{k=1}^K$ and $I_{\widetilde{\mathcal{C}}}(\mathbf{x})$ as the corresponding indicator functions, which equal zeros when the argument(s) (jointly) fall within the feasible sets while otherwise are $+\infty$. Hence, the problem (3) can be rewritten as

$$\min_{\mathbf{x},t} \quad t^2 + I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}(\mathbf{x}, t) + I_{\widetilde{\mathcal{C}}}(\mathbf{x}) \tag{4}$$

where the second sum term in the objective equals the sum of all the indicator functions $\{I_{\mathcal{C}_k}(\mathbf{x}, t)\}_{k=1}^K$, i.e., $I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}(\mathbf{x}, t) = \sum_{k\in\mathcal{K}} I_{\mathcal{C}_k}(\mathbf{x}, t)$. For the obtained unconstrained problem (4), we adopt the proximal splitting technique to find its solutions, which are shown in the following Lemma.

Lemma 1. *The local stable points of* (4) *can be obtained via the iterations given below*

$$t := \tfrac{\gamma}{\gamma+2}\mathrm{prox}_{I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}}(t) \tag{5}$$

$$\mathbf{x} := \mathrm{prox}_{I_{\widetilde{\mathcal{C}}}}\big(\tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\mathrm{prox}_{I_{\mathcal{C}_k}}(\mathbf{x})\big) \tag{6}$$

with $\mathrm{prox}_{(\cdot)}(\cdot)$ *being the proximal operator whose generalized definition is given by* $\mathrm{prox}_f(\mathbf{z}) \triangleq \mathrm{argmin}_{\mathbf{y}}\ f(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{z}\|^2$ *[20], and* $\gamma > 0$ *being a parameter of user choice.*

Proof. We denote the objective function of (4) as $\zeta(\mathbf{x}, t)$. For known $\mathbf{x}$ or t, the subdifferential of $\zeta(\mathbf{x}, t)$ with respect to t or $\mathbf{x}$ can be expressed as

$$\partial_t\zeta(\mathbf{x}, t) = 2t + \partial_t I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}(\mathbf{x}, t) \tag{7}$$

$$\partial_{\mathbf{x}}\zeta(\mathbf{x}, t) = \textstyle\sum_{k\in\mathcal{K}}\partial_{\mathbf{x}} I_{\mathcal{C}_k}(\mathbf{x}, t) + \partial_{\mathbf{x}} I_{\widetilde{\mathcal{C}}}(\mathbf{x}) \tag{8}$$

where we use a subscript to mark the variable for subdifferential $\partial_{(\cdot)}(\cdot)$. Here, the subdifferentials obey the generalized definition given by [21], which is not subject to the convexity. Its relevant properties include the sum rule, chain rule, Fermat's rule, and mean-value theorem [21], [22].

According to the Fermat's rule [21], the locally stationary points of (4) can be obtained by enforcing the subdifferentials (7) and (8) to include zeros. Hence, the following tasks are to find any subsets of $\partial_t I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}(\mathbf{x}, t)$, $\{\partial_{\mathbf{x}} I_{\mathcal{C}_k}(\mathbf{x}, t)\}_{k=1}^K$, and $\partial_{\mathbf{x}} I_{\widetilde{\mathcal{C}}}(\mathbf{x})$. To this end, we can use the following results

$$\gamma\big(t - \mathrm{prox}_{I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}}(t)\big) \in \partial_t I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}(\mathbf{x}, t),\ \forall\gamma > 0 \tag{9}$$

$$\mathbf{x} - \operatorname{prox}_{I_{\mathcal{C}_k}}(\mathbf{x}) \in \partial_{\mathbf{x}} I_{\mathcal{C}_k}(\mathbf{x}, t),\ \forall k \in \mathcal{K} \tag{10}$$

$$K\big(\mathbf{x} - \operatorname{prox}_{I_{\widetilde{\mathcal{C}}}}(\mathbf{x})\big) \in \partial_{\mathbf{x}} I_{\widetilde{\mathcal{C}}}\big(\operatorname{prox}_{I_{\widetilde{\mathcal{C}}}}(\mathbf{x})\big) \tag{11}$$

which originate from the definitions of the subdifferential and indicator function whose detailed derivations are omitted due to the space limitation.

Substituting the left side of (9) into (7) as the desired subset for $\partial_t I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}(\mathbf{x}, t)$, we obtain a stable point that satisfies the relation $t = \frac{\gamma}{\gamma+2}\operatorname{prox}_{I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}}(t)$. This leads to the fixed-point iteration for t shown by (5). Likewise, using (8), (10), and (11), after some straightforward derivations, we obtain the fixed-point iteration for $\mathbf{x}$ shown by (6). The proof is complete. □

The remaining task is to derive the explicit expressions of $\operatorname{prox}_{I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}}(t)$ and $\operatorname{prox}_{I_{\widetilde{\mathcal{C}}}}\big(\frac{1}{K}\sum_{k\in\mathcal{K}}\operatorname{prox}_{I_{\mathcal{C}_k}}(\mathbf{x})\big)$ in Lemma 1 to solve (4), which are given based on their definitions as

$$\operatorname{prox}_{I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}}(t) = \max_{k\in\mathcal{K}}\{|(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k)^{\mathrm{H}}\mathbf{x} - s_k|, t\} \tag{12}$$

$$\operatorname{prox}_{I_{\widetilde{\mathcal{C}}}}\big(\tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\operatorname{prox}_{I_{\mathcal{C}_k}}(\mathbf{x})\big) = c\cdot e^{j\cdot\arg(\sum_{k\in\mathcal{K}}\operatorname{prox}_{I_{\mathcal{C}_k}}(\mathbf{x}))} \tag{13}$$

with the vector $\operatorname{prox}_{I_{\mathcal{C}_k}}(\mathbf{x})$ being derived as

$$\operatorname{prox}_{I_{\mathcal{C}_k}}(\mathbf{x}) = \mathbf{x} + \frac{(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k)^{\mathrm{H}}\mathbf{x} - s_k}{|(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k)^{\mathrm{H}}\mathbf{x} - s_k|}(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k) \times \frac{\min\{0, t - |(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k)^{\mathrm{H}}\mathbf{x} - s_k|\}}{\|\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k\|^2} \tag{14}$$

whose detailed derivations have been omitted due to the space limitation. To conclude, we use (5), (6), and (12)–(14) to find the optimal per-antenna constant-envelope precoding.

B. Reflection-Coefficient Matrix Design to Obtain $\boldsymbol{\Phi}$

Recalling the problem (2), fixing $\mathbf{x}$ and additionally introducing an auxiliary variable $\tilde{t}$, we can obtain the alternating optimization problem with respect to $\boldsymbol{\Phi}$ and $\tilde{t}$ as follows

$$\min_{\boldsymbol{\Phi},\tilde{t}} \quad \tilde{t}^{\,2} \tag{15a}$$

$$\text{s.t.} \quad |(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k)^{\mathrm{H}}\mathbf{x} - s_k|^2 \le \tilde{t}^{\,2},\ k \in \mathcal{K} \tag{15b}$$

$$\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\Phi} = \mathbf{I}_N \tag{15c}$$

$$\boldsymbol{\Phi} = \boldsymbol{\Phi}^{\mathrm{T}}. \tag{15d}$$

To solve (15), we first denote $\{\mathcal{S}_k\}_{k=1}^K$ and $\widetilde{\mathcal{S}}$ as feasible sets for its constraints (15b) and (15c)–(15d), respectively. Similarly as in the previous subsection, we introduce indicator functions $\{I_{\mathcal{S}_k}(\boldsymbol{\Phi}, \tilde{t})\}_{k=1}^K$ and $I_{\widetilde{\mathcal{S}}}(\boldsymbol{\Phi})$ to transform (15) into the unconstrained optimization problem given as follows

$$\min_{\boldsymbol{\Phi},\tilde{t}} \quad \tilde{t}^{\,2} + I_{\cap_{k\in\mathcal{K}}\mathcal{S}_k}(\boldsymbol{\Phi}, \tilde{t}) + I_{\widetilde{\mathcal{S}}}(\boldsymbol{\Phi}) \tag{16}$$

which can also be solved by proximal splitting techniques as used in the previous subsection. Mathematically, the solutions to (16) can be derived as the fixed-point iterations given by

$$\tilde{t} := \tfrac{\gamma}{\gamma+2}\max_{k\in\mathcal{K}}\{|(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k)^{\mathrm{H}}\mathbf{x} - s_k|, \tilde{t}\} \tag{17}$$

$$\boldsymbol{\Phi} := \operatorname{prox}_{I_{\widetilde{\mathcal{S}}}}\big(\tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\operatorname{prox}_{I_{\mathcal{S}_k}}(\boldsymbol{\Phi})\big) \tag{18}$$

with the explicit expression of $\operatorname{prox}_{I_{\mathcal{S}_k}}(\boldsymbol{\Phi})$ being expressed as

$$\operatorname{prox}_{I_{\mathcal{S}_k}}(\boldsymbol{\Phi}) = \boldsymbol{\Phi} + \frac{(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k)^{\mathrm{T}}\mathbf{x}^* - s_k^*}{|(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k)^{\mathrm{T}}\mathbf{x}^* - s_k^*|}\mathbf{h}_k^{\mathrm{H}} \otimes \big(\mathbf{H}^{\mathrm{H}}\mathbf{x}\big) \times \frac{\min\{0, \tilde{t} - |(\mathbf{d}_k + \mathbf{H}\boldsymbol{\Phi}\mathbf{h}_k)^{\mathrm{T}}\mathbf{x}^* - s_k^*|\}}{\|\mathbf{h}_k\|^2\|\mathbf{H}^{\mathrm{H}}\mathbf{x}\|^2} \tag{19}$$

whose detailed derivations are also omitted to be shown here.

The following task is to derive the final expression for (18), which can be obtained as follows

$$\boldsymbol{\Phi} := \operatorname*{argmin}_{\mathbf{B}\in\widetilde{\mathcal{S}}} \big\|\mathbf{B} - \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}}\operatorname{prox}_{I_{\mathcal{S}_k}}(\boldsymbol{\Phi})\big\|_{\mathrm{F}}^2 \tag{20}$$

$$= \operatorname*{argmin}_{\mathbf{B}\in\widetilde{\mathcal{S}}} \big\|\mathbf{B} - \tfrac{1}{2K}\textstyle\sum_{k\in\mathcal{K}}\big(\operatorname{prox}_{I_{\mathcal{S}_k}}(\boldsymbol{\Phi}) + \operatorname{prox}_{I_{\mathcal{S}_k}}^{\mathrm{T}}(\boldsymbol{\Phi})\big)\big\|_{\mathrm{F}}^2 \tag{21}$$

where the definition of proximal operator has been used to derive from (18) to the former update, and the condition of $\widetilde{\mathcal{S}}$ has been used to derive the latter. Note that $\mathbf{B}$ is required to be both symmetric and unitary because $\mathbf{B} \in \widetilde{\mathcal{S}}$. Hence, it can be decomposed as a product between a unitary matrix and its transpose via Takagi factorization [23], i.e., $\mathbf{B} = \mathbf{U}_{\mathbf{B}}\mathbf{U}_{\mathbf{B}}^{\mathrm{T}}$. The matrix $\frac{1}{2K}\sum_{k\in\mathcal{K}}\big(\operatorname{prox}_{I_{\mathcal{S}_k}}(\boldsymbol{\Phi}) + \operatorname{prox}_{I_{\mathcal{S}_k}}^{\mathrm{T}}(\boldsymbol{\Phi})\big)$ is also symmetric which can be rewritten as $\mathbf{U}_{\boldsymbol{\Phi}}\boldsymbol{\Sigma}_{\boldsymbol{\Phi}}\mathbf{U}_{\boldsymbol{\Phi}}^{\mathrm{T}}$ via Takagi factorization. Therefore, the optimal $\mathbf{B}$ in (21) can be obtained when the Frobenius norm $\|\mathbf{U}_{\mathbf{B}}\mathbf{U}_{\mathbf{B}}^{\mathrm{T}} - \mathbf{U}_{\boldsymbol{\Phi}}\boldsymbol{\Sigma}_{\boldsymbol{\Phi}}\mathbf{U}_{\boldsymbol{\Phi}}^{\mathrm{T}}\|_{\mathrm{F}}$ is minimized. In order to find a solution to the minimization of this Frobenius norm, we use the fact that it equals the trace of $-\mathbf{U}_{\mathbf{B}}^*\mathbf{U}_{\mathbf{B}}^{\mathrm{H}}\mathbf{U}_{\boldsymbol{\Phi}}\boldsymbol{\Sigma}_{\boldsymbol{\Phi}}\mathbf{U}_{\boldsymbol{\Phi}}^{\mathrm{T}}$. It is straightforward that such a trace is no smaller than the trace of $-\boldsymbol{\Sigma}_{\boldsymbol{\Phi}}$ because both $\mathbf{U}_{\mathbf{B}}$ and $\mathbf{U}_{\boldsymbol{\Phi}}$ are unitary. Based on these derivations, the aforementioned Frobenius norm can achieve the minimum on the condition that $\mathbf{U}_{\mathbf{B}} = \mathbf{U}_{\boldsymbol{\Phi}}$, which finally leads to the update of $\boldsymbol{\Phi}$ given by

$$\boldsymbol{\Phi} := \mathbf{U}_{\boldsymbol{\Phi}}\mathbf{U}_{\boldsymbol{\Phi}}^{\mathrm{T}} \tag{22}$$

We use (17), (19), and (22) to obtain the optimal $\boldsymbol{\Phi}$.

C. Convergence Analysis on Iteration Updates

To conduct convergence analysis for iteration updates (5) and (6), we first denote their expressions on the right side as $p(\mathbf{x}, t)$ and $\mathbf{q}(\mathbf{x}, t)$, respectively. Then, we use them to investigate the residues of optimization variables across iterations. According to the mean-value theorem of subdifferential [21], the residue of $[t, \mathbf{x}^{\mathrm{T}}]^{\mathrm{T}}$ between two neighboring iterations satisfies

$$\begin{bmatrix} t^{(r+1)} \\ \mathbf{x}^{(r+1)} \end{bmatrix} - \begin{bmatrix} t^{(r)} \\ \mathbf{x}^{(r)} \end{bmatrix} = \begin{bmatrix} p(\mathbf{x}^{(r)}, t^{(r)}) \\ \mathbf{q}(\mathbf{x}^{(r)}, t^{(r)}) \end{bmatrix} - \begin{bmatrix} p(\mathbf{x}^{(r-1)}, t^{(r-1)}) \\ \mathbf{q}(\mathbf{x}^{(r-1)}, t^{(r-1)}) \end{bmatrix} \in \begin{bmatrix} \partial_t p(\mathbf{x}, t) & \partial_t \mathbf{q}(\mathbf{x}, t) \\ \partial_{\mathbf{x}} p(\mathbf{x}, t) & \partial_{\mathbf{x}} \mathbf{q}(\mathbf{x}, t) \end{bmatrix}^{\mathrm{H}}\Bigg|_{\substack{t=t_0 \\ \mathbf{x}=\mathbf{x}_0}} \left(\begin{bmatrix} t^{(r)} \\ \mathbf{x}^{(r)} \end{bmatrix} - \begin{bmatrix} t^{(r-1)} \\ \mathbf{x}^{(r-1)} \end{bmatrix}\right) \tag{23}$$

where we use superscripts $(\cdot)^{(r-1)}$, $(\cdot)^{(r)}$, and $(\cdot)^{(r+1)}$ to mark the $(r-1)$-th, r-th, and $(r+1)$-th iterations, respectively, and t_0 and $\mathbf{x}_0$ are points on the lines from $t^{(r)}$ to $t^{(r-1)}$ and $\mathbf{x}^{(r)}$ to $\mathbf{x}^{(r-1)}$, respectively.

Exploiting the explicit expressions of $p(\mathbf{x}, t)$ and $\mathbf{q}(\mathbf{x}, t)$ in (5) and (6), respectively, we can express the subdifferentials

$\partial_t p(\mathbf{x},t)$ and $\partial_{\mathbf{x}}\mathbf{q}(\mathbf{x},t)$ as the following forms

$$\partial_t p(\mathbf{x},t) = \tfrac{\gamma}{\gamma+2}\cdot \partial_t \mathrm{prox}_{I_{\cap_{k\in\mathcal{K}}\mathcal{C}_k}}(t) \tag{24}$$

$$\partial_{\mathbf{x}}\mathbf{q}(\mathbf{x},t) \subseteq \tfrac{1}{K}\textstyle\sum_{k\in\mathcal{K}} \partial_{\mathbf{x}}\mathrm{prox}_{I_{\mathcal{C}_k}}(\mathbf{x})\partial_{\mathbf{x}_\star}\mathrm{prox}_{I_{\bar{\mathcal{C}}}}(\mathbf{x}_\star) \tag{25}$$

where $\mathbf{x}_\star \triangleq \frac{1}{K}\sum_{k\in\mathcal{K}}\mathrm{prox}_{I_{\mathcal{C}_k}}(\mathbf{x})$, and the sum rule and chain rule of the subdifferentials [21], [22] have been used in the derivations to (25). To investigate the properties of $\partial_t p(\mathbf{x},t)$ and $\partial_{\mathbf{x}}\mathbf{q}(\mathbf{x},t)$ respectively given by (24) and (25), we present the following result.

Lemma 2. *For an indicator function $I_{\mathcal{C}}(\mathbf{z})$ with respect to $\mathbf{z}$ and $\mathcal{C}$, all matrices in $\partial_{\mathbf{z}}\mathrm{prox}_{I_{\mathcal{C}}}(\mathbf{z})$ are Hermitian, and their maximum eigenvalues are no larger than* 1.

Proof. Applying subdifferentials twice to the definition of the proximal operator [20] with respect to indicator functions, we can obtain the following result

$$\partial_{\mathbf{z}}\mathrm{prox}_{I_{\mathcal{C}}}(\mathbf{z}) \subseteq \left(\mathbf{I}_L + \partial^2_{\mathrm{prox}_{I_{\mathcal{C}}}(\mathbf{z})} I_{\mathcal{C}}(\mathrm{prox}_{I_{\mathcal{C}}}(\mathbf{z}))\right)^{-1} \tag{26}$$

where $\partial^2_{(\cdot)}(\cdot)$ denotes the second-order differential defined as the set that includes all the subdifferentials of the first-order subgradients in this paper, and the chain rule [22] has been used in the derivations to (26).

Since the indicator function $I_{\mathcal{C}}(\mathbf{z})$ is locally convex within the neighborhood of $\mathrm{prox}_{I_{\mathcal{C}}}(\mathbf{z})$, $\partial^2_{\mathrm{prox}_{I_{\mathcal{C}}}(\mathbf{z})} I_{\mathcal{C}}(\mathrm{prox}_{I_{\mathcal{C}}}(\mathbf{z}))$ only includes positive semi-definite matrices [24], which enables the set $\partial_{\mathbf{z}}\mathrm{prox}_{I_{\mathcal{C}}}(\mathbf{z})$ to include Hermitian matrices with maximum eigenvalues no larger than 1 based on the expression of (26). The proof is complete. □

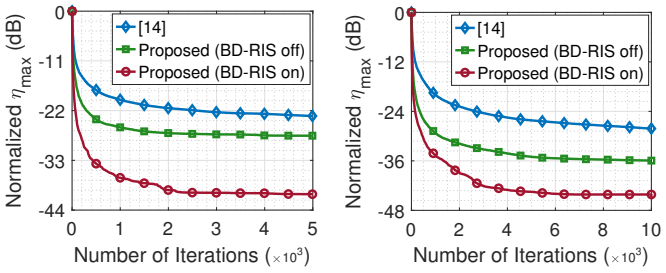
Applying Lemma 2 to (24) and (25), respectively, we can obtain the inequalities $\partial_t p(\mathbf{x},t) \le \frac{\gamma}{\gamma+2} < 1$ and $\partial_{\mathbf{x}}\mathbf{q}(\mathbf{x},t) \preceq \mathbf{I}_N$ using some elementary properties of the eigenvalue. Based on these results together with (23), we can conclude that

$$\begin{aligned}&|t^{(r+1)} - t^{(r)}|^2 + \|\mathbf{x}^{(r+1)} - \mathbf{x}^{(r)}\|^2 \\ &\qquad \le |t^{(r)} - t^{(r-1)}|^2 + \|\mathbf{x}^{(r)} - \mathbf{x}^{(r-1)}\|^2\end{aligned} \tag{27}$$

which means the variable $[t, \mathbf{x}^{\mathrm{T}}]^{\mathrm{T}}$ can monotonically converge to a locally stationary point. Thus, the convergence of (5) and (6) can be guaranteed. The convergence of (17) and (18) can be proved through the same routine.

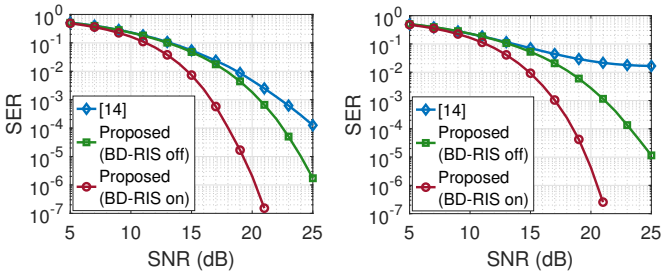
IV. Simulation Results

In this section, we conduct our proposed iteration updates and evaluate their convergence performance, obtained SERs, and the maximum power of multi-user interference (denoted hereafter by $\eta_{\max}$). We also compare our proposed method with the algorithm in [13]. The SQUAREM scheme [25] is used for accelerating our method. The desired communication symbols are generated from the unit-power constellation points of 16-QAM, and the channel parameters are assumed to be Rayleigh distributed. The constant magnitude of transmit symbols is set to be 1, and the stopping criterion of the tested algorithms is chosen as a predefined maximum number of iterations, which is set to be 10^4. Throughout simulations, all the data are averaged over 50 independent trails, and we use the same hardware and software configurations for comparisons.



(a) $M = 32$ and $K = 20$ (b) $M = 64$ and $K = 36$

Fig. 1. Convergence evaluations with $N = 5$.



(a) $M = 100$ and $K = 54$ (b) $M = 200$ and $K = 108$

Fig. 2. SER evaluations with $N = 10$.

Evaluations on convergence: We evaluate the convergence performances of tested algorithms. The presented objective is adopted as $\eta_{\max}$ after normalization by its initial values. Both the scenarios of BD-RIS off and BD-RIS on are studied for our proposed method. The size of BD-RIS is set to be $N = 5$, and two cases of $M = 32$ with $K = 20$ and $M = 64$ with $K = 36$ are investigated. The corresponding results are shown in Fig. 1. It can be seen that our proposed method presents faster convergence speeds and achieves lower $\eta_{\max}$ compared to the method of [13]. For example, after 3000 iterations in the former case, our proposed method reaches normalized $\eta_{\max}$ equalling -27.26 dB and -40.18 dB for the scenarios with BD-RIS off and on, respectively, while the method of [13] only achieves -22.34 dB normalized $\eta_{\max}$. In the latter case, the normalized $\eta_{\max}$ obtained by our proposed method with BD-RIS off and on respectively give 7.81 dB and 16.03 dB improvements on that obtained by the method of [13].

Evaluations on SER: We then evaluate the SER performances of the tested algorithms, wherein both the cases of $M = 100$ with $K = 54$ and $M = 200$ with $K = 108$ are investigated. The size of BD-RIS is set to be $N = 10$, and the corresponding results obtained over 10^5 channel realizations are shown in Fig. 2. It can be seen that our proposed method shows lower SERs than those obtained by the method of [13] in both tested cases, even for the scenario with BD-RIS off. The reason is that our proposed "min-max" optimization gives potential performance gains on SER. For instance, in the latter case, our proposed method achieves low SERs equalling 2.7×10^{-3} and 4.0×10^{-6} at SNR$=$ 20 dB for the scenarios with BD-RIS off and on, respectively, while the method of [13] obtains a high SER equalling 2.48×10^{-2}.

TABLE I
EVALUATIONS ON NORMALIZED MAXIMUM POWER OF INTERFERENCE VERSUS NUMBERS OF TRANSMIT ANTENNAS.

$N = 5$, $K = 24$	$M = 30$	$M = 36$	$M = 42$	$M = 48$	$M = 54$
[13]	-17.64 dB	-20.91 dB	-29.77 dB	-34.93 dB	-41.54 dB
Proposed (RIS-off)	-20.40 dB	-24.89 dB	-34.30 dB	-47.12 dB	-92.64 dB
Proposed (RIS-on)	-22.15 dB	-31.89 dB	-156.81 dB	-212.49 dB	-267.88 dB

TABLE II
EVALUATIONS ON NORMALIZED MAXIMUM POWER OF INTERFERENCE VERSUS NUMBERS OF USERS AND SIZES OF RIS.

$M = 60$	$K = 25$	$K = 30$	$K = 35$	$K = 40$	$K = 45$	$K = 50$
$N = 5$	-268.59 dB	-118.26 dB	-45.87 dB	-29.43 dB	-24.09 dB	-19.74 dB
$N = 10$	-283.22 dB	-267.23 dB	-163.21 dB	-44.37 dB	-29.28 dB	-25.31 dB
$N = 20$	-290.57 dB	-287.95 dB	-280.09 dB	-253.11 dB	-180.27 dB	-48.41 dB

Evaluations on the worst power of multi-user interference: We evaluate the normalized values of $\eta_{\max}$ obtained by tested algorithms. The first case investigates multiple numbers of transmit antennas $M \in \{30, 36, 42, 48, 54\}$ associated with a 5-element BD-RIS and 24 users. The second case studies multiple sizes of BD-RIS $N \in \{5, 10, 20\}$ and also multiple numbers of users $K \in \{25, 30, 35, 40, 45, 50\}$ associated with $M = 60$ transmit antennas. The corresponding results for the two cases are shown in Tables I and II, respectively. It can be seen that for all tested algorithms, the obtained $\eta_{\max}$ generally increases as the ratio K/M becomes large. It can also be seen that our proposed method can reach lower normalized values of $\eta_{\max}$ than those obtained by the method of [13] (e.g., -92.64 dB for BD-RIS off and -267.88 dB for BD-RIS on compared to -41.54 dB when $M = 54$ as shown in Table I). Moreover, a large size of BD-RIS enables a low normalized $\eta_{\max}$ (e.g., -45.87, -163.21, and -280.09 dBs for $N = 5$, 10, and 20, respectively, as shown in Table II when $K = 35$).

V. CONCLUSION

We have studied the BD-RIS aided constant-envelope precoding for massive MIMO communications. Specifically, we have minimized the maximum difference between the desired and received noise-free symbols among all users. By incorporating the inherent constraints of transmit signals and the BD-RIS, we have formulated a new precoding design problem that involves a "min-max" objective and non-convex constraints. To tackle it, we have exploited a cyclic manner that involves a reformulation from the "min-max" type problem to a solvable form in each alternating optimization. Then, we have devised a fixed-point iteration rule via proximal splitting. The convergence guarantee of our proposed iteration rule has been proved. Simulations have verified the superiority of our proposed precoding design over existing methods in terms of different aspects.

REFERENCES

[1] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector perturbation technique for near-capacity multiantenna multiuser communication-Part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[2] M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1711–1721, May 2007.

[3] C. Masouros, "Correlation rotation linear precoding for MIMO broadcast communications," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 252–262, Jan. 2011.

[4] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.

[5] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[6] P. V. Amadori and C. Masouros, "Constant envelope precoding by interference exploitation in phase shift keying-modulated multiuser transmission," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 538–550, Jan. 2017.

[7] C. Masouros and G. Zheng, "Exploiting known interference as green signal power for downlink beamforming optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 14, pp. 3628–3640, Jul. 2015.

[8] C. Studer and E. G. Larsson, "PAR-aware large-scale multi-user MIMO-OFDM downlink," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 303–313, Feb. 2013.

[9] C. Shi, F. Wang, M. Sellathurai, J. Zhou, and S. Salous, "Power minimization-based robust OFDM radar waveform design for radar and communication systems in coexistence," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1316–1330, Mar. 2018.

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

[11] C. Masouros and E. Alsusa, "Dynamic linear precoding for the exploitation of known interference in MIMO broadcast systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1396–1404, Mar. 2009.

[12] M. Alodeh, S. Chatzinotas, and B. Ottersten, "Constructive multiuser interference in symbol level precoding for the MISO downlink channel," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2239–2252, May 2015.

[13] S. K. Mohammed and E. G. Larsson, "Per-antenna constant envelope precoding for large multi-user MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 3, pp. 1059–1071, 2013.

[14] S. Zhang, R. Zhang, and T. J. Lim, "Constant envelope precoding for MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 149–162, Jan. 2018.

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

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

[17] 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.

[18] S. Shen, B. Clerckx, and R. Murch, "Modeling and architecture design of reconfigurable intelligent surfaces using scattering parameter network analysis," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1229–1243, Feb. 2022.

[19] H. Li, S. Shen, and B. Clerckx, "Beyond diagonal reconfigurable intelligent surfaces: From transmitting and reflecting modes to single-, group-, and fully-connected architectures," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2311–2324, Apr. 2023.

[20] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2014.

[21] F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*. London, U.K.: Springer, 2013.

[22] W. Schirotzek, *Nonsmooth Analysis*. Heidelberg, Germany: Springer, 2007.

[23] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[24] N. H. Chieu, T. D. Chuong, J.-C. Yao, and N. D. Yen, "Characterizing convexity of a function by its fréchet and limiting second-order subdifferentials," *Set-Valued Var. Anal.*, vol. 19, pp. 75–96, Mar. 2011.

[25] R. Varadhan and C. Roland, "Simple and globally convergent methods for accelerating the convergence of any EM algorithm," *Scand. J. Statist.*, vol. 35, no. 2, pp. 335–353, 2008.