Optimization of RIS-aided MIMO Systems via the Cutoff Rate

Nemanja Stefan Perović, Le-Nam Tran, *Senior Member, IEEE,* Marco Di Renzo, *Fellow, IEEE,* and Mark F. Flanagan, *Senior Member, IEEE*

***Abstract*—The main difficulty concerning optimizing the mutual information (MI) in reconfigurable intelligent surface (RIS)-aided communication systems with discrete signaling is the inability to formulate this optimization problem in an analytically tractable manner. Therefore, we propose to use the cutoff rate (CR) as a more tractable metric for optimizing the MI and introduce two optimization methods to maximize the CR. The first method is based on the projected gradient method (PGM), while the second method is derived from the principles of successive convex approximation (SCA). Simulation results show that the proposed optimization methods significantly enhance the CR and the corresponding MI.**

***Index Terms*—Channel cutoff rate (CR), mutual information (MI), multiple-input multiple-output (MIMO), optimization, reconfigurable intelligent surfaces (RISs).**

I. Introduction

The recently developed reconfigurable intelligent surfaces (RISs) have the potential to shape and control the radio wave propagation in wireless networks, which makes them a promising candidate for future beyond-5G communication systems. RISs consist of a large number of small, low-cost, and nearly-passive elements each of which can reflect the incident signal with an adjustable phase shift, thereby modifying the wavefront of the scattered wave [1]. Changing the wavefront of the reflected signals enables us to shape how the radio waves propagate through the channel, and thus improve key system performance metrics such as the achievable rate. The main body of research work in this area concentrates on the achievable rate optimization for single-user [2] and multi-user [3] multiple-input single-output (MISO) communication systems. Another significant body of research work has focused on the achievable rate optimization for multiple-input multiple-output (MIMO) systems equipped with RISs in single-user [4], [5] and multi-user [6], [7] communications. In the aforementioned papers, the transmitted symbols are distributed according to a circularly-symmetric complex Gaussian distribution, which is a capacity achieving distribution. However, in practice the transmitted symbols are usually chosen from a discrete signal constellation and thus the present solutions cannot be used to establish realistic bounds on the achievable data rate in practical RIS-aided communication systems.

The main difficulty concerning the *practical* achievable rate, i.e., the analysis and optimization of the mutual information (MI) in RIS-aided communication systems with discrete signaling, is the inherent difficulty of formulating this optimization problem in an analytically tractable manner. Hence, MI optimization and analysis were considered in only a few publications. In [8], the authors considered the mutual information optimization for an RIS-aided system, where the transmit information is encoded into in-phase and quadrature (IQ) symbols transmitted from a single transmit antenna and also into the (discrete) RIS phase shifts. However, the proposed optimization approach in [8] is not directly implementable to MIMO systems that transmit multiple IQ symbols in parallel. The mutual information analysis for a MISO communication system, where the transmit information is encoded into the IQ symbol and a subset of active RIS elements, is presented in [9]. The receiver device is equipped with a single receive antenna in [9], so the proposed optimization method is not applicable to multi-stream MIMO systems.

Motivated by this, we propose to use the channel cutoff rate (CR) as a surrogate metric for optimizing the MI of RIS-aided multi-stream MIMO communication systems. The idea comes from the fact that the channel CR serves as a practical upper limit on the information rate for reliable communications [10]. In mathematical terms, the relation between the channel CR R_0 and the codeword error probability P_e can be formulated as $P_e \leq e^{-n(R_0-R)}$, where n is the codeword length and R is the information rate in bits per channel use (bpcu). In other words, for very long coded sequences (i.e., $n \to \infty$), P_e can be made arbitrarily small as long as $R < R_0$. Since the channel capacity is the theoretical upper limit of the information rate for reliable communications, the CR is usually employed as a practical lower bound on the channel capacity. More specifically, it can be shown that a lower bound on the MI is determined by the CR as [11, Eq. (36)]

$$\mathrm{MI} \geq N_r(1-\log_2 e) + R_0^h$$

where R_0^h corresponds to the CR evaluated at half the noise power. In practical terms, the use of the channel CR facilitates the otherwise intractable optimization of the channel capacity (i.e., MI) and provides us with a tool to optimize modulation techniques for communication systems.

Against this background, the contributions of this paper are listed as follows:

1) Instead of optimizing the MI directly, which results in an intractable problem, we propose to use the CR as a surrogate metric for the MI optimization. We show that the CR can lead to a deterministic optimization problem for which efficient numerical algorithms can be derived.
2) To maximize the CR for the considered system, we formulate a joint optimization problem of the precoding matrix and the RIS elements' phase shifts. Since the

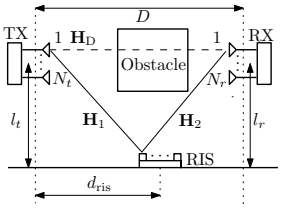


Fig. 1. Aerial view of the considered communication system.

problem of interest is nonconvex and our numerical experiments indicates that it contains many local optima (see Section V), we propose two local optimization methods for solving this problem. The first method is a projected gradient method (PGM) which admits closed-form expressions in each iteration. The second method is derived from the principles of successive convex approximation (SCA). In this way, one of the two proposed methods can avoid being trapped in an unsuitable local optimum, and thus will provide a near-optimal solution for a given channel realization.

3) We present simulation results which show that the proposed optimization methods, by maximizing the CR, can substantially increase the MI.

II. System Model

A. System Model

An aerial view of the considered system is shown in Fig. 1. It contains a transmitter equipped with N_t antennas and a receiver equipped with N_r antennas, where we assume $N_t \geq N_r$. The separations between the adjacent antennas in the transmit and the receive antenna array are s_t and s_r, respectively. These antenna arrays are placed on parallel vertical walls, which are at a distance D from each other[1]. To mitigate the blockage (i.e., attenuation) of the direct link, an RIS, whose midpoint is at a distance d_ris from the plane containing the transmit antenna array, is installed. The RIS is placed on a vertical wall that is perpendicular to both the transmit and the receive antenna array. We assume that the RIS, the transmit antenna array and the receive antenna array are approximately at the same height. The RIS consists of N_ris reflecting elements, which are placed in a rectangular formation such that the separation between the centers of adjacent RIS elements in both dimensions is $s_\mathrm{ris} = \frac{\lambda}{2}$, where λ denotes the wavelength. The distance between the midpoint of the transmit (respectively, receive) antenna array and the plane containing the RIS is l_t (respectively, l_r).

The signal vector at the receive antenna array is given by

$$\mathbf{y} = \mathbf{H}\mathbf{P}\mathbf{x}_i + \mathbf{n}, \tag{1}$$

where $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix, $\mathbf{P} \in \mathbb{C}^{N_t \times N_r}$ is the transmit precoding matrix and $\mathbf{x}_i \in \mathbb{C}^{N_r \times 1}$ is the transmit symbol vector. The elements of $\mathbf{x}_i$ are chosen from a discrete symbol alphabet of size M with unit average symbol energy. The number of different transmit symbol vectors is $N_s = M^{N_r}$. We assume that the precoding matrix $\mathbf{P}$ preserves the average power of the transmit signal (i.e., $\mathrm{Tr}(\mathbf{P}\mathbf{P}^H) = N_r$). The noise vector $\mathbf{n} \in \mathbb{C}^{N_r \times 1}$ consists of independent and identically distributed (i.i.d.) elements that are distributed according to $\mathcal{CN}(0, \sigma^2)$, where σ^2 denotes the noise variance.

Since an RIS is present in this system, the channel matrix can be expressed as $\mathbf{H} = \sqrt{\beta_\mathrm{DIR}^{-1}}\mathbf{H}_\mathrm{D} + \sqrt{\beta_\mathrm{INDIR}^{-1}}\mathbf{H}_2\mathbf{F}(\boldsymbol{\theta})\mathbf{H}_1$, where $\mathbf{H}_\mathrm{D} \in \mathbb{C}^{N_r \times N_t}$ represents the *direct* link between the transmitter and the receiver, $\mathbf{H}_1 \in \mathbb{C}^{N_\mathrm{ris} \times N_t}$ represents the link between the transmitter and the RIS, and $\mathbf{H}_2 \in \mathbb{C}^{N_r \times N_\mathrm{ris}}$ represents the link between the RIS and the receiver. The distance-dependent path loss for the direct link is β_DIR^{-1} and the free space path loss (FSPL) for the indirect link is $\beta_\mathrm{INDIR}^{-1}$. Signal reflection from the RIS is modeled by $\mathbf{F}(\boldsymbol{\theta}) = \mathrm{diag}(\boldsymbol{\theta}) \in \mathbb{C}^{N_\mathrm{ris} \times N_\mathrm{ris}}$, where $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_{N_\mathrm{ris}}]^T \in \mathbb{C}^{N_\mathrm{ris} \times 1}$. Since we assume that the reflection is without any power loss, we may write $\theta_l = e^{j\phi_l}$ (i.e., $|\theta_l| = 1$) for $l = 1, 2, \ldots, N_\mathrm{ris}$, where ϕ_l is the phase shift induced by the l-th RIS element. In this letter, we assume perfect knowledge of the channel state information (CSI).

III. Problem Formulation

Since the considered system is a discrete-input continuous-output memoryless channel (DCMC), the MI is given by [12, Eq. (7.2.1)]

$$\mathrm{MI} = \sum\nolimits_{i=1}^{N_s} \int_\mathbf{y} p(\mathbf{y}, \mathbf{x}_i) \log_2\Big[\frac{p(\mathbf{y}|\mathbf{x}_i)}{\sum_{j=1}^{N_s} p(\mathbf{y}, \mathbf{x}_j)}\Big] d\mathbf{y}, \tag{2}$$

where the conditional probability density function is

$$p(\mathbf{y}|\mathbf{x}_i) = \Big(1/\left(\pi\sigma^2\right)^{N_r}\Big)\exp\Big(-\frac{\|\mathbf{y} - \mathbf{H}\mathbf{P}\mathbf{x}_i\|^2}{\sigma^2}\Big). \tag{3}$$

Assuming that all the transmit symbol vectors are equally probable $p(\mathbf{x}_i) = 1/N_s$ $(i = 1, \ldots, N_s)$, we obtain

$$\mathrm{MI} = \log_2 N_s - \frac{1}{N_s}\sum\nolimits_{i=1}^{N_s} \mathbb{E}_\mathbf{n}\Big\{\log_2 \sum\nolimits_{j=1}^{N_s} \exp(\psi_{i,j})\Big\}, \tag{4}$$

where $\psi_{i,j} = (-\|\mathbf{H}\mathbf{P}(\mathbf{x}_i - \mathbf{x}_j) + \mathbf{n}\|^2 + \|\mathbf{n}\|^2)/\sigma^2$.

We remark that the expression for the MI in (4) leads to an intractable stochastic optimization problem, if one wishes to maximize the MI directly. The fact that the MI expression in (4) is neither convex nor concave with respect to $\mathbf{P}$ and $\boldsymbol{\theta}$ also adds to the difficulty. Also, the feasible set for θ_l, which satisfies $|\theta_l| = 1$, is nonconvex. To overcome these issues, we propose the use of the CR as an auxiliary metric whose optimization is well-aligned with that of the MI.

A. Derivation of the CR

The CR of the considered system for equiprobable symbol vectors $\mathbf{x}_i$ is given by [10]

$$R_0 = -\log_2\Big[\frac{1}{N_s^2}\int_\mathbf{y} \sum\nolimits_{i,j=1}^{N_s} \sqrt{p(\mathbf{y}|\mathbf{x}_i)p(\mathbf{y}|\mathbf{x}_j)}d\mathbf{y}\Big]. \tag{5}$$

By inserting (3) into (5) and after some algebraic manipulations, the CR of the considered system can be written as

$$R_0 = -\log_2\Big[\frac{1}{N_s^2}\sum\nolimits_{i,j=1}^{N_s} \exp\Big(-\frac{\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P})}{4\sigma^2}\Big)\Big], \tag{6}$$

where

$$\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}) = \|\mathbf{H}\mathbf{P}\left(\mathbf{x}_i - \mathbf{x}_j\right)\|^2 = \|\mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\|^2 \tag{7}$$

and $\mathbf{e}_{i,j} = \mathbf{x}_i - \mathbf{x}_j$.

[1]This system geometry described in the paper is adopted for ease of exposition, and to provide the reader with a concrete use case. However, the optimization approach proposed in this letter is applicable to any system geometry.

B. Optimization Problem

Upon close inspection of (6), it is easy to see that the CR optimization problem can be formulated as

$$\underset{\boldsymbol{\theta},\mathbf{P}}{\text{minimize}}\ f(\boldsymbol{\theta},\mathbf{P})=\sum\nolimits_{i,j=1}^{N_s}\exp\Big(-\frac{\Phi_{i,j}(\boldsymbol{\theta},\mathbf{P})}{4\sigma^2}\Big) \tag{8a}$$

$$\text{subject to } \mathrm{Tr}(\mathbf{P}\mathbf{P}^H)=N_r \tag{8b}$$

$$\big|\theta_l\big|=1, l=1,2,\ldots,N_{\mathrm{ris}}. \tag{8c}$$

Since the objective of (8) is nonconvex, finding a globally optimal solution is very difficult. Hence, in the next section, we present two local optimization algorithms to solve (8).

IV. Proposed Optimization Methods

A. Projected Gradient Method (PGM)

The first proposed method is based on the PGM [13], which consists of the following iterations:

$$\boldsymbol{\theta}_{n+1}=P_{\Theta}(\boldsymbol{\theta}_n-\mu_1\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta}_n,\mathbf{P}_n)), \tag{9a}$$

$$\mathbf{P}_{n+1}=P_{\mathcal{P}}(\mathbf{P}_n-\mu_2\nabla_{\mathbf{P}}f(\boldsymbol{\theta}_{n+1},\mathbf{P}_n)), \tag{9b}$$

where $\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta},\mathbf{P})$ and $\nabla_{\mathbf{P}}f(\boldsymbol{\theta},\mathbf{P})$ are the gradients of $f(\boldsymbol{\theta},\mathbf{P})$ with respect to $\boldsymbol{\theta}^*$ and $\mathbf{P}^*$, respectively[2], and μ_1 and μ_2 are the corresponding step sizes. Also, $P_{\Theta}(\cdot)$ and $P_{\mathcal{P}}(\cdot)$ denote the projection onto Θ and $\mathcal{P}$, respectively, which are detailed in the sequel. Note that we use the complex-valued gradients [14], the explicit forms of which are provided in the next lemma.

Lemma 1. *The gradients of $f(\boldsymbol{\theta},\mathbf{P})$ with respect to $\boldsymbol{\theta}^*$ and $\mathbf{P}^*$ are given by*

$$\nabla_{\boldsymbol{\theta}}f(\boldsymbol{\theta},\mathbf{P})=-\frac{1}{4\sigma^2}\sum_{i,j=1}^{N_s}e^{-\frac{\Phi_{i,j}(\boldsymbol{\theta},\mathbf{P})}{4\sigma^2}} \times\mathrm{vec}_d\left(\mathbf{H}_2^H\mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H\mathbf{P}^H\bar{\mathbf{H}}_1^H\right), \tag{10a}$$

$$\nabla_{\mathbf{P}}f(\boldsymbol{\theta},\mathbf{P})=-\frac{1}{4\sigma^2}\mathbf{H}^H\mathbf{H}\mathbf{P}\sum_{i,j=1}^{N_s}e^{-\frac{\Phi_{i,j}(\boldsymbol{\theta},\mathbf{P})}{4\sigma^2}}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H, \tag{10b}$$

where $\bar{\mathbf{H}}_1=\sqrt{\beta_{\mathrm{INDIR}}^{-1}}\mathbf{H}_1$.

Proof: See Appendix A. ∎

1) Projection Operations: Next we show that the projection operations in (9a) and (9b) can be calculated in closed form. Note that the constraint $\big|\theta_l\big|=1$ states that θ_l lies on the unit circle in the complex plane. Thus, for a given point $\boldsymbol{\theta}\in\mathbb{C}^{N_{\mathrm{ris}}\times 1}$, $P_{\Theta}(\boldsymbol{\theta})$ is given by

$$\bar{\theta}_l=\begin{cases}\frac{\theta_l}{|\theta_l|} & \theta_l\neq 0\\ e^{j\phi},\phi\in[0,2\pi] & \theta_l=0\end{cases},\ l=1,\ldots,N_{\mathrm{ris}}. \tag{11}$$

In particular, $\bar{\theta}_l$ can be any point on the unit circle if $\theta_l=0$, and thus $P_{\Theta}(\boldsymbol{\theta})$ is not unique.

Similarly, the constraint $\mathrm{Tr}(\mathbf{P}\mathbf{P}^H)=N_r$ implies that the projection of the precoding matrix $P_{\mathcal{P}}(\mathbf{P})$ is given by

$$\bar{\mathbf{P}}=\mathbf{P}\sqrt{N_r/\mathrm{Tr}(\mathbf{P}\mathbf{P}^H)}. \tag{12}$$

[2]Here x^* denotes the complex conjugate of x.

2) Backtracking Line Search: Appropriate choices of the step sizes in (9a) and (9b) are instrumental for ensuring the convergence of the PGM. Ideally, each step size should be inversely proportional to the Lipschitz constant of the corresponding gradient. However, the optimal Lipschitz constants are difficult to find for our considered problem. Therefore, we utilize the Armijo-Goldstein backtracking line search to determine the step sizes μ_1 and μ_2 at each iteration.

Let $L_0>0$, $\delta>0$ be a small constant, and $\rho\in(0,1)$. The step size μ_1 in (9a) is found as $L_o\rho^{k_n}$, where k_n is the smallest nonnegative integer such that

$$f(\boldsymbol{\theta}_{n+1},\mathbf{P}_n)\leq f(\boldsymbol{\theta}_n,\mathbf{P}_n)-\delta||\boldsymbol{\theta}_{n+1}-\boldsymbol{\theta}_n||^2. \tag{13}$$

The step size μ_2 is found similarly. The proposed line search procedure ensures that the objective sequence strictly decreases after each iteration. Thus, the PGM is guaranteed to converge to a stationary point of (8), which is, however, not necessarily a globally optimal solution.

B. Successive Convex Approximation (SCA)

Since (8) is a nonconvex problem, the proposed PGM can become trapped in an unsuitable local optimum. Thus, it is desirable to check the obtained solution with another local optimization method that is developed by applying a different optimization paradigm. Our expectation is that the two proposed local optimization methods can complement each other and one of them can escape unsuitable local optima. In particular, the second proposed method is derived from the SCA, which has been shown to be a powerful tool for a range of nonconvex optimization problems.

First, we relax the equality constraints in (8), leading to the following program:

$$\underset{\boldsymbol{\theta},\mathbf{P}}{\text{minimize}}\ f(\boldsymbol{\theta},\mathbf{P})=\sum\nolimits_{i,j=1}^{N_s}\exp\left(-\frac{\Phi_{i,j}(\boldsymbol{\theta},\mathbf{P})}{4\sigma^2}\right) \tag{14a}$$

$$\text{subject to } \mathrm{Tr}(\mathbf{P}\mathbf{P}^H)\leq N_r \tag{14b}$$

$$|\theta_l|\leq 1, l=1,2,\ldots,N_{\mathrm{ris}}. \tag{14c}$$

It is easy to see that $\Phi_{i,j}(\boldsymbol{\theta},\alpha\mathbf{P})>\Phi_{i,j}(\boldsymbol{\theta},\mathbf{P})$ for any $\alpha>1$. Thus, the constraint in (14b) must hold with equality at the optimum. Otherwise, we can always scale up $\mathbf{P}$ and achieve a strictly smaller objective. Similarly, it can be shown that (14c) holds with equality at the optimum (see Appendix B). These two observations imply that the set of optimal solutions of (14) is the same as that of (8). Also, the aforementioned relaxation ensures the convexity of the feasible sets for $\boldsymbol{\theta}$ and $\mathbf{P}$, while the objective function $f(\boldsymbol{\theta},\mathbf{P})$ remains nonconvex.

Let us suppose that we have already obtained a feasible point $(\mathbf{P}_n,\boldsymbol{\theta}_n)$. To find $\mathbf{P}_{n+1}$ based on the SCA method, we need to find a convex upper bound on $f(\boldsymbol{\theta}_n,\mathbf{P})$. Since $\Phi_{i,j}(\boldsymbol{\theta}_n,\mathbf{P})$ is convex with respect to $\mathbf{P}$, we have

$$\Phi_{i,j}(\boldsymbol{\theta}_n,\mathbf{P})\geq\Phi_{i,j}(\boldsymbol{\theta}_n,\mathbf{P}_n)+\left\langle\nabla_{\mathbf{P}}\Phi_{i,j}(\boldsymbol{\theta}_n,\mathbf{P}_n),\mathbf{P}-\mathbf{P}_n\right\rangle \triangleq\tilde{\Phi}_{i,j}(\boldsymbol{\theta}_n,\mathbf{P}_n;\mathbf{P}) \tag{15}$$

where $\left\langle\mathbf{X},\mathbf{Y}\right\rangle=2\Re(\mathrm{Tr}(\mathbf{X}^H\mathbf{Y}))$. In fact, $\tilde{\Phi}_{i,j}(\boldsymbol{\theta}_n,\mathbf{P}_n;\mathbf{P})$ is an affine approximation of $\Phi_{i,j}(\boldsymbol{\theta}_n,\mathbf{P})$ around $\mathbf{P}_n$. Next,

$\mathbf{P}_{n+1}$ is a solution to the following convex problem:

$$\underset{\mathbf{P}}{\text{minimize}} \ \sum\nolimits_{i,j=1}^{N_s} \exp\Big(-\frac{\tilde{\Phi}_{i,j}(\boldsymbol{\theta}_n, \mathbf{P}_n; \mathbf{P})}{4\sigma^2}\Big) \tag{16a}$$

$$\text{subject to } \mathrm{Tr}(\mathbf{P}\mathbf{P}^H) \leq N_r. \tag{16b}$$

To find $\boldsymbol{\theta}_{n+1}$, we again need to derive a convex upper bound on $f(\boldsymbol{\theta}, \mathbf{P}_{n+1})$. It is easy to see that $\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}_{n+1})$ is convex with respect to $\boldsymbol{\theta}$ and thus we have

$$\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}_{n+1}) \geq \Phi_{i,j}(\boldsymbol{\theta}_n, \mathbf{P}_{n+1}) + \Big\langle \nabla_{\boldsymbol{\theta}} \Phi_{i,j}(\boldsymbol{\theta}_n, \mathbf{P}_{n+1}), \boldsymbol{\theta} - \boldsymbol{\theta}_n \Big\rangle \triangleq \bar{\Phi}_{i,j}(\boldsymbol{\theta}_n, \mathbf{P}_{n+1}; \boldsymbol{\theta}). \tag{17}$$

Next, $\boldsymbol{\theta}_{n+1}$ is a solution of the following convex problem:

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \ \sum\nolimits_{i,j=1}^{N_s} \exp\Big(-\frac{\bar{\Phi}_{i,j}(\boldsymbol{\theta}_n, \mathbf{P}_{n+1}; \boldsymbol{\theta})}{4\sigma^2}\Big) \tag{18a}$$

$$\text{subject to } |\theta_l| \leq 1, l = 1, 2, \ldots, N_{\text{ris}}. \tag{18b}$$

Both convex optimization problems in (16) and (18) can be solved by off-the-shelf convex solvers. By minimizing an upper bound in each iteration, the SCA-based method generates a decreasing objective sequence. It can shown that the SCA-based method converges to a stationary point of (14).

C. Complexity Comparison

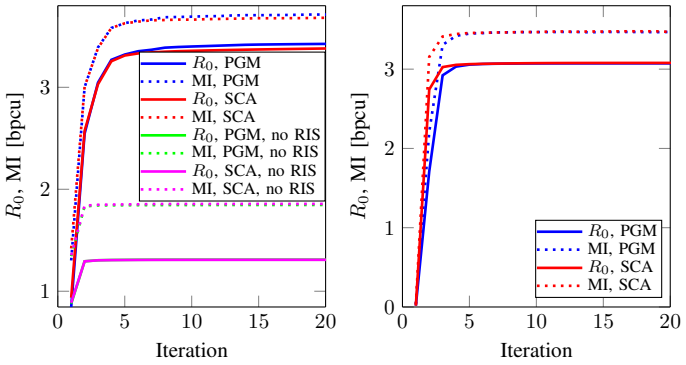
We next provide a brief analysis of the complexity of the proposed methods based on the required number of complex multiplications. The per-iteration complexity of the PGM method is $\mathcal{O}(N_s^2 N_{\text{ris}}(N_t + N_r))$. To solve subproblems (16) and (18), we first need to compute (15) and (17), and this requires approximately the same per-iteration complexity as the PGM. By treating (16) and (18) as generic convex problems, the worst-case complexity for solving (16) and (18) is $\mathcal{O}(N_t^4 N_r^4)$ and $\mathcal{O}(N_{\text{ris}}^4)$, respectively.

The SCA-based method generally requires higher computational complexity to return a solution than the PGM method. The reason is that convex solvers commonly use an interior point method to solve problems (16) and (18) which involve exponential cones. The memory requirement and the computational complexity increase very quickly with the problem size. Consequently, the SCA-based method is not suitable for large-scale scenarios, i.e., when N_s or N_{ris} is large. For such cases, the PGM is the only viable option. For problems of moderate size, the SCA-based method is a good choice since it is a descent method without a line search. On the other hand, the PGM method requires a line search to ensure a good convergence rate.

V. Simulation Results

In this section, we evaluate the CR and the MI of the proposed optimization algorithms with the aid of Monte Carlo simulations. More precisely, we utilize $\boldsymbol{\theta}$ and $\mathbf{P}$ obtained by optimizing the CR to calculate the MI according to the expression in (4), where the expectation is evaluated by generating random noise vectors. Also, we compare the MI optimization results in the case of discrete and Gaussian signaling.

In the following simulations, all of the channel matrices are modeled according to the Rician fading channel model with Rician factor K, as specified in [5]. Also, we assume no spatial correlation[3] exists among the elements of matrices $\mathbf{H}_1$ and $\mathbf{H}_2$. The distance-dependent path loss for the direct link is $\beta_{\text{DIR}} = (4\pi/\lambda)^2 d_0^{\alpha_{\text{DIR}}}$, where d_0 is the distance between the transmit array midpoint and the receive array midpoint, and the path loss exponent of the direct link is denoted by α_{DIR}. The far-field FSPL for the indirect link $\beta_{\text{INDIR}}^{-1}$ can be computed according to [16] as $\beta_{\text{INDIR}}^{-1} = \lambda^4 (l_t/d_1 + l_r/d_2)^2/(256\pi^2 d_1^2 d_2^2)$, where d_1 is the distance between the transmit antenna array midpoint and the RIS center, and d_2 is the distance between the RIS center and the receive antenna array midpoint.



(a) The direct link is present. (b) The direct link is blocked.

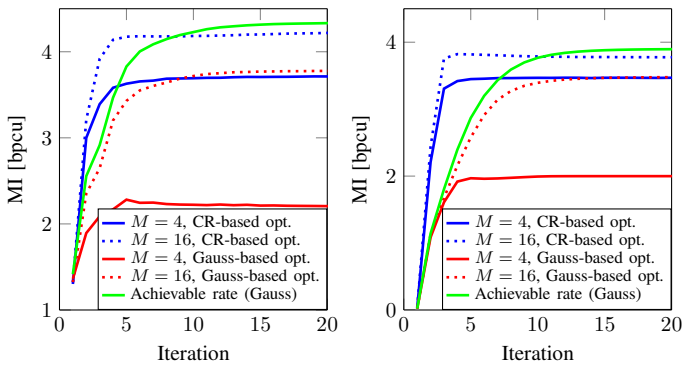
Fig. 2. The CR (R_0) and the MI of the proposed PGM and SCA methods.

In the following simulation setup, the parameters are $f = 2\,\text{GHz}$ (i.e., $\lambda = 15\,\text{cm}$), $s_t = s_r = s_{\text{ris}} = \lambda/2 = 7.5\,\text{cm}$, $D = 500\,\text{m}$, $l_t = l_r = 20\,\text{m}$, $d_{\text{ris}} = 30\,\text{m}$, $N_t = 8$, $N_r = 2$, $\alpha_{\text{DIR}} = 3$, $K = 1$, $M = 4$, and $\sigma^2 = -110\,\text{dB}$. The RIS consists of $N_{\text{ris}} = 225$ elements placed in a 15×15 square formation. The line search procedure for the PGM utilizes the parameters $L_0 = 10^3$, $\delta = 10^{-3}$ and $\rho = 1/2$. For the SCA-based method, we use the CVX tool with MOSEK as the internal software package to solve (16) and (18). The initial values of $\boldsymbol{\theta}$ and $\mathbf{P}$ are randomly chosen for both methods. All results are averaged over 30 independent channel realizations.

In Fig. 2, we show the CR and the MI results of the proposed PGM and SCA for scenarios in which the direct link is present and in which the the direct link is blocked. As benchmark schemes, we consider the presented system without the RIS, whose precoding matrix is optimized by using the proposed methods. The two proposed optimization methods achieve approximately the same results and need only a few iterations to reach the optimum. In general, the MI shows the same behavioral trend as the CR, but the MI is always larger than the CR. For both metrics, the proposed algorithms achieve a similar gain through the parameter optimization, which is around 2 bpcu and 3 bpcu if the direct link is present and blocked, respectively. These results justify our initial claim that the CR presents a suitable metric to optimize the MI. In addition, removing the RIS from the considered system causes a significant reduction of the CR and the MI.

In Fig. 3, we compare the MI when discrete and Gaussian signaling are considered. Specifically, we plot the MI where

[3]Spatial correlation models for RIS channel matrices have recently become available in the literature (e.g., in [15]).



(a) The direct link is present. (b) The direct link is blocked.

Fig. 3. MI optimization results for discrete and Gaussian signaling.

the precoding matrix and the RIS coefficients are obtained by implementing the CR optimization (referred to as CR-based optimization), and the MI where the precoding matrix and the RIS coefficients are obtained based on the achievable rate optimization for Gaussian signaling (referred to as Gaussian signaling based optimization) [5]. Moreover, we present the achievable rate for Gaussian signaling [5]. As can be seen, the CR-based optimization yields a higher MI than the Gaussian signaling based optimization, particularly for a small size of symbol alphabet. As expected, the achievable rate for Gaussian signaling is always higher than the MI. However, for $M = 16$ the gap between the MI obtained by using the discrete optimization and the achievable rate is less than 0.2 bpcu.

VI. Conclusion

In this letter, we have investigated the use of the CR as a simple and meaningful metric for optimizing the MI in RIS-aided MIMO communication systems. Since the maximization of the CR is a nonconvex optimization problem, we proposed two local optimization methods. The first method is based on the PGM which uses closed-form expressions in each iteration. The second method is derived from the principles of SCA. Simulation results show that both optimization methods produce very similar performance results and that the CR is indeed a suitable substitute metric for optimizing the MI in RIS-aided MIMO communication systems.

Appendix A
Complex-valued Gradient of $f(\boldsymbol{\theta}, \mathbf{Q})$

The complex gradient of $f(\boldsymbol{\theta}, \mathbf{Q})$ with respect to $\boldsymbol{\theta}^*$ is

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mathbf{P}) = -\frac{1}{4\sigma^2} \sum_{i,j=1}^{N_s} \exp\left(-\frac{\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P})}{4\sigma^2}\right) \nabla_{\boldsymbol{\theta}} \Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}). \tag{19}$$

Also, the complex differential of $\Phi_{i,j}$ with respect to $\mathbf{F}(\boldsymbol{\theta}) = \mathrm{diag}(\boldsymbol{\theta})$ and $\mathbf{F}^*(\boldsymbol{\theta})$ can be expressed as

$$d\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}) = d(\|\mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\|^2) = \mathbf{e}_{i,j}^H \mathbf{P}^H d(\mathbf{H}^H\mathbf{H})\mathbf{P}\mathbf{e}_{i,j}. \tag{20}$$

With the aid of a few algebraic steps, we obtain

$$d\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}) = \mathrm{vec}^T\left(\mathbf{H}_2^H \mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H \mathbf{P}^H \bar{\mathbf{H}}_1^H\right) \mathrm{vec}(d(\mathbf{F}^*)) + \mathrm{vec}^T\left(\left(\bar{\mathbf{H}}_1^H \mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H \mathbf{P}^H \mathbf{H}^H \mathbf{H}_2\right)^T\right) \mathrm{vec}(d(\mathbf{F})) \tag{21}$$

where we used the identity $\mathrm{Tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{vec}^T(\mathbf{A})\,\mathrm{vec}(\mathbf{B})$, where $\mathrm{vec}(\mathbf{A})$ denotes the vector obtained by vertical stacking of the columns of $\mathbf{A}$. Let $\mathbf{L}_d$ be the matrix used to place the diagonal elements of a square matrix $\mathbf{A}$ on $\mathrm{vec}(\mathbf{A})$, i.e., $\mathrm{vec}(\mathbf{A}) = \mathbf{L}_d\,\mathrm{vec}_d(\mathbf{A})$ [14, Def. 2.12]. Then, we have

$$d\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}) = \mathrm{vec}^T\left(\mathbf{H}_2^H \mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H \mathbf{P}^H \bar{\mathbf{H}}_1^H\right) \mathbf{L}_d\,\mathrm{vec}(d\boldsymbol{\theta}^*) + \mathrm{vec}^T\left(\left(\bar{\mathbf{H}}_1^H \mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H \mathbf{P}^H \mathbf{H}^H \mathbf{H}_2\right)^T\right) \mathbf{L}_d\,\mathrm{vec}(d\boldsymbol{\theta}). \tag{22}$$

Using [14, Table 3.2] and [14, Eqn. (2.140)], we obtain

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}) &= \mathbf{L}_d^T\,\mathrm{vec}\left(\mathbf{H}_2^H \mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H \mathbf{P}^H \bar{\mathbf{H}}_1^H\right) \\ &= \mathrm{vec}_d\left(\mathbf{H}_2^H \mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H \mathbf{P}^H \bar{\mathbf{H}}_1^H\right)\end{aligned} \tag{23}$$

where $\mathrm{vec}_d(\mathbf{A})$ denotes the vector comprised of the diagonal elements of $\mathbf{A}$. Substituting (23) into (19), we obtain (10a).

In a similar way, the gradient of $f(\boldsymbol{\theta}, \mathbf{P})$ with respect to $\mathbf{P}^*$ is equal to

$$\nabla_{\mathbf{P}} f(\boldsymbol{\theta}, \mathbf{P}) = -\frac{1}{4\sigma^2} \sum_{i,j=1}^{N_s} \exp\left(-\frac{\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P})}{4\sigma^2}\right) \nabla_{\mathbf{P}} \Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}). \tag{24}$$

The complex differential of $\Phi_{i,j}$ with respect to $\mathbf{P}$ and $\mathbf{P}^*$is

$$\begin{aligned} d\Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}) &= d(\mathbf{e}_{i,j}^H \mathbf{P}^H \mathbf{H}^H \mathbf{H}\mathbf{P}\mathbf{e}_{i,j}) = d(\mathrm{Tr}\{\mathbf{P}^H \mathbf{H}^H \mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H\}) \\ &\overset{(a)}{=} \mathrm{Tr}\{(\mathbf{H}^H \mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H)^T d\mathbf{P}^* + \mathbf{e}_{i,j}\mathbf{e}_{i,j}^H \mathbf{P}^H \mathbf{H}^H \mathbf{H} d\mathbf{P}\}\end{aligned} \tag{25}$$

where (a) is obtained from [14, Table 4.3]. Similarly, we have

$$\nabla_{\mathbf{P}} \Phi_{i,j}(\boldsymbol{\theta}, \mathbf{P}) = \mathbf{H}^H \mathbf{H}\mathbf{P}\mathbf{e}_{i,j}\mathbf{e}_{i,j}^H. \tag{26}$$

Substituting (26) into (24), we obtain (10b).

Appendix B
Connection Between (8c) and (14c)

Let $\boldsymbol{\theta}^*$ an the optimal solution to (14) and suppose that there is some l such that $|\theta_l^*| < 1$. Note that we can write $\Phi_{i,j}(\boldsymbol{\theta}^*, \mathbf{P}) = \mathrm{Tr}\Big(\left(\mathbf{H}_{\mathrm{DIR}}^H + \mathbf{H}_1^H \mathbf{F}^H(\boldsymbol{\theta}^*)\mathbf{H}_2^H\right)\left(\mathbf{H}_{\mathrm{DIR}} + \mathbf{H}_2\mathbf{F}(\boldsymbol{\theta}^*)\mathbf{H}_1\right)\mathbf{P}\mathbf{e}_{ij}\mathbf{e}_{ij}^H\mathbf{P}^H\Big)$ as

$$\Phi_{i,j}(\boldsymbol{\theta}^*, \mathbf{P}) = a_{i,j}\,|\theta_l^*|^2 + 2\Re\big(b_{i,j}\theta_l^*\big) + c_{i,j}$$

where $a_{i,j} \geq 0$, $b_{i,j}$ and $c_{i,j}$ are the resulting constants obtained by rearranging $\Phi_{i,j}(\boldsymbol{\theta}^*, \mathbf{P})$ with respect to θ_l^*. Now consider the point $\tilde{\theta}_l = -\theta_l^*$ which is also feasible to (14c). Also, we define $\tilde{\boldsymbol{\theta}} = [\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_{N_{\mathrm{ris}}}]^T \in \mathbb{C}^{N_{\mathrm{ris}} \times 1}$ such that $\tilde{\theta}_k = \theta_k^*$ for $k = 1, \ldots, N_{\mathrm{ris}}$ and $k \neq l$, and $\tilde{\theta}_l = -\theta_l^*$. It is easy to see that $\Phi_{i,j}(\tilde{\boldsymbol{\theta}}, \mathbf{P}) = a_{i,j}\,|\theta_l^*|^2 - 2\Re\big(b_{i,j}\theta_l^*\big) + c_{i,j}$. Thus we can conclude that $2\Re\big(b_{i,j}\theta_l^*\big) \geq 0$, otherwise $\Phi_{i,j}(\tilde{\boldsymbol{\theta}}, \mathbf{P}) > \Phi_{i,j}(\boldsymbol{\theta}^*, \mathbf{P})$, and thus, a strictly smaller objective can be obtained. Next, since $2\Re\big(b_{i,j}\theta_l^*\big) \geq 0$ we immediately have that $\Phi_{i,j}(\alpha\theta_l^*, \mathbf{P}) = \alpha^2 a_{i,j}\,|\theta_l^*|^2 + \alpha 2\Re\big(b_{i,j}\theta_l^*\big) + c_{i,j} \geq a_{i,j}\,|\theta_l^*|^2 + 2\Re\big(b_{i,j}\theta_l^*\big) + c_{i,j} = \Phi_{i,j}(\theta_l^*, \mathbf{P})$ for any $\alpha > 1$. That means, if $|\theta_l^*| < 1$ then we can always scale θ_l^* up to achieve a better objective.

References

[1] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[2] 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.

[3] Q.-U.-A. Nadeem *et al.*, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7748–7764, Dec. 2020.

[4] N. S. Perović *et al.*, "Channel capacity optimization using reconfigurable intelligent surfaces in indoor mmWave environments," in *Proc. IEEE Int. Conf. on Communications (ICC)*, 2020, pp. 1–7.

[5] ——, "Achievable rate optimization for MIMO systems with reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, 2021, Early Access.

[6] C. Pan *et al.*, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5218–5233, Aug. 2020.

[7] ——, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1719–1734, Aug. 2020.

[8] R. Karasik *et al.*, "Adaptive coding and channel shaping through reconfigurable intelligent surfaces: An information-theoretic analysis," *arXiv preprint arXiv:2012.00407*, 2020.

[9] S. Lin *et al.*, "Reconfigurable intelligent surfaces with reflection pattern modulation: Beamforming design and performance analysis," *IEEE Trans. Wireless Commun.*, 2020, Early Access.

[10] J. L. Massey, "Coding and modulation in digital communications," in *Proc. of Int. Zurich Seminar*, 1974, pp. E2.1–E2.4.

[11] N. S. Perović *et al.*, "Optimization of the cut-off rate of generalized spatial modulation with transmit precoding," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4578–4595, Oct. 2018.

[12] R. G. Gallager, *Information theory and reliable communication*. New York: Springer, 1968, vol. 2.

[13] S. Boyd *et al.*, *Convex optimization*. Cambridge university press, 2004.

[14] A. Hjörungnes, *Complex-Valued Matrix Derivatives With Applications in Signal Processing and Communications*. Cambridge University Press, 2011.

[15] E. Björnson and L. Sanguinetti, "Rayleigh fading modeling and channel hardening for reconfigurable intelligent surfaces," *IEEE Wireless Commun. Lett.*, 2020, Early Access.

[16] F. H. Danufane *et al.*, "On the path-loss of reconfigurable intelligent surfaces: An approach based on Green's theorem applied to vector fields," *arXiv preprint arXiv:2007.13158*, 2020.