Achievable Rate Optimization for MIMO Systems with Reconfigurable Intelligent Surfaces

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

***Abstract*—Reconfigurable intelligent surfaces (RISs) represent a new technology that can shape the radio wave propagation in wireless networks and offers a great variety of possible performance and implementation gains. Motivated by this, we study the achievable rate optimization for multi-stream multiple-input multiple-output (MIMO) systems equipped with an RIS, and formulate a joint optimization problem of the covariance matrix of the transmitted signal and the RIS elements. To solve this problem, we propose an iterative optimization algorithm that is based on the projected gradient method (PGM). We derive the step size that guarantees the convergence of the proposed algorithm and we define a backtracking line search to improve its convergence rate. Furthermore, we introduce the total free space path loss (FSPL) ratio of the indirect and direct links as a first-order measure of the applicability of RISs in the considered communication system. Simulation results show that the proposed PGM achieves the same achievable rate as a state-of-the-art benchmark scheme, but with a significantly lower computational complexity. In addition, we demonstrate that the RIS application is particularly suitable to increase the achievable rate in indoor environments, as even a small number of RIS elements can provide a substantial achievable rate gain.**

***Index Terms*—Achievable rate, gradient projection, multiple-input multiple-output (MIMO), optimization, reconfigurable intelligent surface (RIS).**

I. Introduction

In recent years, there has been a tremendous, almost exponential, increase in the demands for higher data rates. The main driving forces that constantly increase this demand are the increasing number of mobile devices and the appearance of services that require high data rates (e.g., video streaming and online gaming). Consequently, many technology solutions have been proposed to address this ever-increasing demand, such as massive multiple-input multiple-output (MIMO) and millimeter-wave (mmWave) communications. In spite of providing potentially significant achievable rate gains, these technologies generally incur additional power and hardware costs, so that the total benefit of their implementation has to be independently evaluated for each user scenario. Broadly speaking, these technologies can be seen as novel transmitter and receiver features that enable us to achieve higher data rates. However, they do not have the capability of directly influencing the propagation channel, the stochastic nature of which can sometimes limit the efficiency of these proposed technology solutions.

A possible approach to overcome the aforementioned issue lies in the use of the recently-developed reconfigurable intelligent surfaces (RISs) [1]. The key component to realize the RIS function is a software-defined surface that is reconfigurable in such a way as to adapt itself to changes in the wireless environment. It consists of a large number of small, low-cost, and passive elements, each of which can reflect the incident signal with an adjustable phase shift, thereby modifying the radio waves. Optimization of the wavefront of the reflected signals enables us to shape how the radio waves interact with the surrounding objects, and thus control their scattering and reflection characteristics [2], [3], [4]. Hence, the introduction of RISs fundamentally changes the wave propagation in wireless communication systems and offers a wide variety of possible implementation gains, thus potentially presenting a new milestone in wireless communications.

In recent years, researchers have investigated many important aspects of RIS-assisted wireless communication systems. The problem of estimating the required channel state information (CSI) was considered in [5] by embedding an active sensor in the RIS, and in [6] by estimation of a combined transmitter-RIS-receiver channel. Other emerging body of work studies an accurate modeling of the interactions (considering reflection, refraction, diffraction and polarization) of the incident wave with the RIS, and elucidates the dependence of these interactions on the size of the RIS elements, the distance between the adjacent RIS elements, the angle of incidence and so on [7], [8]. All of these aspects are critical for the practical implementation of RIS-aided wireless communication systems to become feasible.

From a theoretical standpoint, the evaluation and optimization of the achievable rate of an RIS-aided wireless communication system is crucial. This problem is significantly more challenging to solve than in the conventional case without an RIS, since in the case without an RIS the channel capacity can be completely determined in closed-form for deterministic (or fixed) channels. A variety of different optimization methods for enhancing the achievable rate in RIS-

The work of N. S. Perović and M. F. Flanagan was funded by the Irish Research Council under grant number IRCLA/2017/209. The work of L. N. Tran was supported in part by a Grant from Science Foundation Ireland under Grant number 17/CDA/4786. M. Di Renzo's work was supported in part by the European Commission through the H2020 ARIADNE project under grant agreement number 871464 and through the H2020 RISE-6G project under grant agreement number 101017011.

N. S. Perović, L. N. Tran, and M. F. Flanagan are with School of Electrical and Electronic Engineering, University College Dublin, Belfield, Dublin 4, Ireland (Email: nemanja.stefan.perovic@ucd.ie, nam.tran@ucd.ie and mark.flanagan@ieee.org).

M. Di Renzo is with Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, 3 Rue Joliot-Curie, 91192 Gif-sur-Yvette, France (E-mail: marco.di-renzo@universite-paris-saclay.fr).

aided wireless communication systems have been proposed in the literature, which attempt to find a near-optimal solution with a reasonable computational complexity and run time. The vast majority of these methods are particularly tailored for downlink communication with single-antenna receive devices. In [9], the authors introduced an optimization method that increases the receive signal-to-noise ratio (SNR) and consequently enhances the achievable rate in multiple-input single-output (MISO) systems. The proposed solution is based on the alternating optimization (AO) method, which adjusts the transmit beamformer and the RIS element phase shifts in an alternating fashion. The AO technique has also been successfully utilized to increase the data rate for secure communications in environments with multiple RISs and single-antenna users [10]. In contrast to AO, the spectral efficiency optimization for a single-user MISO system in [11] was performed by jointly adjusting the transmit beamformer and the RIS element phase shifts. In [12], the authors employed a gradient-based algorithm to enhance the receive signal-to-interference-plus-noise ratio (SINR), and hence the achievable rate, for single-antenna users that do not have a direct link with the base station. The achievable rate optimization for multi-user downlink communications is specifically considered for mmWave sparsely scattered channels in [13]. An algorithm for energy efficiency optimization in a multi-user downlink communication system was presented in [14]. The sum-rate optimization for multi-user downlink communications using a deep reinforcement learning based algorithm was introduced in [15].

In contrast to the previous papers, the achievable rate optimization in [16] is realized by jointly controlling the phase and the amplitude adjustment of each RIS element. Furthermore, in [17] the authors developed a practical phase shift model that captures the phase-dependent amplitude variation in the RIS element-wise reflection coefficient and utilized it to enhance the achievable rate. The achievable rate optimization for an RIS with discrete phase shifts in multi-user downlink communications was considered in [18]. A system for serving paired power-domain non-orthogonal multiple access (NOMA) users by designing the RIS phase shifts was introduced in [19]. In [20], the authors studied the joint optimization of the RIS reflection coefficients and the orthogonal frequency division multiple access (OFDMA) time-frequency resource block, as well as power allocations, to maximize the users' common (minimum) rate. Energy-efficiency optimization for multi-user uplink MIMO communications was presented in [21], where the users' covariance matrices and the RIS phase shifts are optimized in an alternating fashion based on partial knowledge of the CSI.

Although RIS-aided communication systems with single-antenna devices are well-studied in the literature, there is only a limited number of papers that consider the design and analysis of an RIS-aided *MIMO* communication system. In particular, the achievable rate optimization in those systems remains relatively unknown. It was demonstrated in [22] how an RIS can be implemented and optimized to increase the rank of the channel matrix, leading to substantial achievable rate gains in *multi-stream* MIMO communications. The method proposed in [22] was specifically designed for pure line-of-sight (LOS) channels, neglecting the presence of any non-LOS (NLOS) component. Optimization of the achievable rate for a *single-stream* MIMO system in an indoor mmWave environment with a blocked direct link was analyzed in [23]. Since in indoor mmWave communications NLOS channel components are usually significantly weaker than the LOS component, all communication links were also modeled as pure LOS links. Although the proposed optimization schemes in [23] provide the near-optimal achievable rate, they require very low computational and hardware complexity. In [24], the authors utilized the AO method to enhance the achievable rate of an RIS-aided *multi-stream* MIMO communication system. Although this optimization method is simple to implement, it can require many iterations to converge, especially when the number of RIS elements is very large, which corresponds precisely to the case where the RIS is the most useful.

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

- To maximize the achievable rate of a multi-stream MIMO system equipped with an RIS, we formulate a joint optimization problem of the covariance matrix of the transmitted signal and the RIS elements (i.e., phase shifts). We then propose an iterative projected gradient method (PGM) to solve this nonconvex problem, for which we present exact gradient and projection expressions in closed form. The proposed method is provably convergent to a critical point of the considered problem, which is desirable for a nonconvex program.
- We derive a Lipschitz constant for the proposed PGM, which is then used to determine an appropriate step size which guarantees its convergence. Also, to improve the rate of convergence of the proposed algorithm, we propose a data scaling step and employ a backtracking line search, which increases the convergence rate significantly, and more importantly, outperforms the existing AO approach in terms of convergence rate.
- As a tool to estimate the applicability of an RIS, we introduce the concept of the *total free space path loss* (total FSPL). Since the computation of the total FSPL of the indirect link is an intractable problem in a MIMO system, we instead derive the total FSPLs for a single-input single-output (SISO) system. We then show that the ratio of the total FSPL of the indirect and direct links can be used as an accurate first-order measure of the applicability of an RIS.
- We show through simulations that the proposed PGM provides the same achievable rate as the AO, but with a significantly lower number of iterations. This is particularly visible in the case where the direct link is blocked, as in this case the PGM needs just a few iterations to reach the convergent achievable rate. As a side product, we demonstrate that the total FSPL of the indirect link is primarily determined by the RIS position, while the total length of the indirect link is of relatively minor importance. Also, we show that scaling the number of RIS elements with the operational frequency can compen-

sate the FSPL increase and ensure communication via the indirect link at all frequencies. Furthermore, we study the application of an RIS in an indoor environment and show that a small number of RIS elements is sufficient to enable the indirect link to have a higher achievable rate than the direct link. Last but not least, we demonstrate that the proposed PGM has a significantly lower computational complexity compared to the AO method.

The rest of this paper is organized as follows. In Section II, we introduce the system model and formulate the optimization problem to maximize the achievable rate of a MIMO system equipped with an RIS. In Section III, we propose and derive the PGM algorithm to solve the previous optimization problem. The convergence and the complexity analysis of the proposed optimization algorithm are presented in Section IV. The applicability of an RIS in the considered communication system is discussed in Section V. In Section VI, we illustrate simulation results of the achievable rate for the proposed PGM algorithm, and use these to illustrate its advantages. Finally, Section VII concludes this paper.

Notation: Bold lower and upper case letters represent vectors and matrices, respectively. $\mathbb{C}^{a\times b}$ denotes the space of complex matrices of dimensions $a \times b$. $(\cdot)^T$, $(\cdot)^*$ and $(\cdot)^H$ represent transpose, complex conjugate and Hermitian transpose, respectively. $\ln(x)$ denotes the natural logarithm of x. $\lambda_{\max}(\mathbf{X})$ denotes the largest singular value of matrix $\mathbf{X}$. To simplify the notation we denote by $||\cdot||$ the Euclidean norm if the argument is a vector and the Frobenius norm if the argument is a matrix. $\mathrm{diag}\,(\mathbf{x})$ denotes the square diagonal matrix which has the elements of $\mathbf{x}$ on the main diagonal. $|x|$ is the absolute value of x and $(x)_+$ denotes $\max(0, x)$. $\arg\{x\}$ denotes the argument of x. The l-th entry of vector $\mathbf{x}$ is denoted by x_l. $\mathrm{Tr}(\mathbf{X})$ is the trace of matrix $\mathbf{X}$, and $\mathbb{E}\{\cdot\}$ stands for the expectation operator. $\det(\mathbf{X})$ is the determinant of $\mathbf{X}$. The notation $\mathbf{A} \succeq (\succ)\mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite (definite). $\nabla_{\mathbf{X}} f(\cdot)$ is the gradient of f with respect to $\mathbf{X}^* \in \mathbb{C}^{m\times n}$, which also lies in $\mathbb{C}^{m\times n}$. $\mathrm{vec}_d(\mathbf{X})$ denotes the vector comprised of the diagonal elements of $\mathbf{X}$. $\mathrm{vec}(\mathbf{X})$ denotes the vectorization operator which stacks the columns of $\mathbf{X}$ to create a single long column vector. $A(i, k)$ denotes the k-th element of the i-th row of matrix $\mathbf{A}$.

II. System Model and Problem Formulation

A. System Model

We consider a wireless communication system with N_t transmit and N_r receive antennas, whose aerial view is depicted in Fig. 1. Both the transmit and receive antennas are placed in uniform linear arrays (ULAs) on vertical walls that are parallel to each other. The distance between these walls is denoted by D. For simplicity, both antenna arrays are parallel to the ground and are assumed to be at the same height. The inter-antenna separations of these arrays are denoted by s_t and s_r, respectively. The direct link is attenuated by an obstacle (e.g., a building) which is situated between the two antenna arrays, and for this reason, a rectangular RIS of size $a \times b$ is utilized to improve the system performance. The RIS is installed on a vertical wall that is perpendicular to the antenna arrays and its center is at the same height as the transmit and the receive antenna arrays[1]. It consists of reflection elements placed in an uniform rectangular array (URA) with N_a and N_b elements per dimension respectively (the total number of reflection elements of the RIS then being $N_{\mathrm{ris}} = N_a N_b$). All RIS elements are of size $\frac{\lambda}{2} \times \frac{\lambda}{2}$, where λ denotes the wavelength of operation. The separation between the centers of adjacent RIS elements in both dimensions is $s_{\mathrm{ris}} = \frac{\lambda}{2}$. The distance between the midpoint of the RIS and the plane containing the transmit antenna array is d_{ris}. The distance between the midpoint of the transmit antenna array and the plane containing the RIS is l_t, and the distance between the midpoint of the receive antenna array and the plane containing the RIS is l_r. We assume that the RIS elements are ideal and that each of them can independently influence the phase and the reflection angle of the impinging wave.

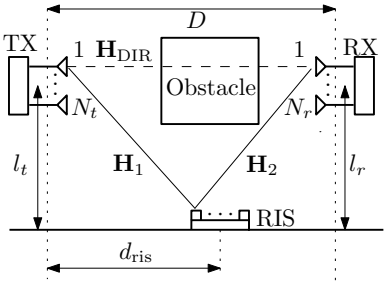


Fig. 1. Aerial view of the considered communication system.

The signal vector at the receive antenna array is given by

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

where $\mathbf{H} \in \mathbb{C}^{N_r\times N_t}$ is the channel matrix, $\mathbf{x} \in \mathbb{C}^{N_t\times 1}$ is the transmit signal vector and $\mathbf{n} \in \mathbb{C}^{N_r\times 1}$ is the noise vector which is distributed according to $\mathcal{CN}(\mathbf{0}, N_0\mathbf{I})$. We assume that the total average transmit power has a maximum value of P_t, i.e., $\mathbb{E}\{\mathbf{x}^H\mathbf{x}\} \le P_t$. Let $\mathbf{Q} \succeq \mathbf{0}$ be the covariance matrix of the transmitted signal, i.e., $\mathbf{Q} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\}$, then the transmit power constraint can be equivalently written as

$$\mathrm{Tr}(\mathbf{Q}) \le P_t. \tag{2}$$

B. Channel Model

Since an RIS is present in this system, the channel matrix can be expressed as

$$\mathbf{H} = \mathbf{H}_{\mathrm{DIR}} + \mathbf{H}_{\mathrm{INDIR}},$$

where $\mathbf{H}_{\mathrm{DIR}} \in \mathbb{C}^{N_r\times N_t}$ represents the *direct* link between the transmitter and the receiver, and $\mathbf{H}_{\mathrm{INDIR}} \in \mathbb{C}^{N_r\times N_t}$ represents the *indirect* link between the transmitter and the receiver (i.e., via the RIS). Adopting the Rician fading channel model, the direct link channel matrix is given by

$$\mathbf{H}_{\mathrm{DIR}} = \frac{\sqrt{\beta_{\mathrm{DIR}}^{-1}}}{\sqrt{K+1}}(\sqrt{K}\mathbf{H}_{\mathrm{D,LOS}} + \mathbf{H}_{\mathrm{D,NLOS}}), \tag{3}$$

[1]For ease of exposition, we assume that the RIS, the transmit and the receive antenna arrays are at the same height. It can be shown by simulations that introducing different heights has a negligible influence on the achievable rate.

where $H_{\mathrm{D,LOS}}(r,t) = e^{-j2\pi d_{r,t}/\lambda}$ and $d_{r,t}$ is the distance between the t-th transmit and the r-th receive antenna. The elements of $\mathbf{H}_{\mathrm{D,NLOS}}$ are independent and identically distributed (i.i.d.) according to $\mathcal{CN}(0,1)$. The FSPL for the direct link is given by $\beta_{\mathrm{DIR}} = (4\pi/\lambda)^2 d_0^{\alpha_{\mathrm{DIR}}}$ [25], where $d_0 = \sqrt{D^2 + (l_t - l_r)^2}$ is the distance between the transmit array midpoint and the receive array midpoint. The path loss exponent of the direct link, whose value is influenced by the obstacle present, is denoted by α_{DIR}. The Rician factor K is chosen from the interval $[0, +\infty)$.

We assume that the far-field model is valid for signal transmission via the RIS (i.e., for the indirect link), and thus $\mathbf{H}_{\mathrm{INDIR}}$ can be written as

$$\mathbf{H}_{\mathrm{INDIR}} = \sqrt{\beta_{\mathrm{INDIR}}^{-1}}\mathbf{H}_2\mathbf{F}(\boldsymbol{\theta})\mathbf{H}_1, \tag{4}$$

where $\mathbf{H}_1 \in \mathbb{C}^{N_{\mathrm{ris}} \times N_t}$ represents the channel between the transmitter and the RIS, $\mathbf{H}_2 \in \mathbb{C}^{N_r \times N_{\mathrm{ris}}}$ represents the channel between the RIS and the receiver, and $\beta_{\mathrm{INDIR}}^{-1}$ represents the overall FSPL for the indirect link. Signal reflection from the RIS is modeled by the matrix $\mathbf{F}(\boldsymbol{\theta}) = \mathrm{diag}(\boldsymbol{\theta}) \in \mathbb{C}^{N_{\mathrm{ris}} \times N_{\mathrm{ris}}}$, where $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_{N_{\mathrm{ris}}}]^T \in \mathbb{C}^{N_{\mathrm{ris}} \times 1}$. In this paper, similar to related works [9], [24], we assume that the signal reflection from any RIS element is ideal, i.e., without any power loss. In other words, we may write $\theta_l = e^{j\phi_l}$ for $l = 1, 2, \ldots, N_{\mathrm{ris}}$, where ϕ_l is the phase shift induced by the l-th RIS element. Equivalently, we may write

$$|\theta_l| = 1, \quad l = 1, 2, \ldots, N_{\mathrm{ris}}. \tag{5}$$

Utilizing the Rician fading channel model, the channel between the transmitter and the RIS $\mathbf{H}_1$ is given by

$$\mathbf{H}_1 = \frac{1}{\sqrt{K+1}}(\sqrt{K}\mathbf{H}_{1,\mathrm{LOS}} + \mathbf{H}_{1,\mathrm{NLOS}}), \tag{6}$$

where $H_{1,\mathrm{LOS}}(l,t) = e^{-j2\pi d_{l,t}/\lambda}$ and $d_{l,t}$ is the distance between the t-th transmit antenna and the l-th RIS element. The elements of $\mathbf{H}_{1,\mathrm{NLOS}}$ are i.i.d. according to $\mathcal{CN}(0,1)$. It is worth noting that the channel matrix expression (6) does not contain any FSPL term.

In a similar way, $\mathbf{H}_2$ can be expressed as

$$\mathbf{H}_2 = \sqrt{\frac{1}{K+1}}(\sqrt{K}\mathbf{H}_{2,\mathrm{LOS}} + \mathbf{H}_{2,\mathrm{NLOS}}) \tag{7}$$

where $\mathbf{H}_{2,\mathrm{LOS}}(r,l) = e^{-j2\pi d_{r,l}/\lambda}$ and $d_{r,l}$ is the distance between the l-th RIS element and the r-th receive antenna. The FSPL for the indirect link can be computed according to [8], [26], [27, Eqn. (18.13.6)] as

$$\beta_{\mathrm{INDIR}}^{-1} = \frac{\lambda^4}{256\pi^2}\frac{(\cos\gamma_1 + \cos\gamma_2)^2}{d_1^2 d_2^2}, \tag{8}$$

where $d_1 = \sqrt{d_{\mathrm{ris}}^2 + l_t^2}$ is the distance between the transmit array midpoint and the RIS center, and $d_2 = \sqrt{(D - d_{\mathrm{ris}})^2 + l_r^2}$ is the distance between the RIS center and the receive array midpoint. Also, γ_1 is the angle between the incident wave direction from the transmit array midpoint to the RIS center and the vector normal to the RIS, and γ_2 is the angle between the vector normal to the RIS and the reflected wave direction from the RIS center to the receive array midpoint. Therefore, we have $\cos\gamma_1 = l_t/d_1$ and $\cos\gamma_2 = l_r/d_2$, which finally gives

$$\beta_{\mathrm{INDIR}}^{-1} = \frac{\lambda^4}{256\pi^2}\frac{(l_t/d_1 + l_r/d_2)^2}{d_1^2 d_2^2}. \tag{9}$$

C. Problem Formulation

In this paper, we are interested in maximizing the achievable rate[2] of the considered RIS-assisted wireless communication system. It is well known that for a MIMO channel, Gaussian signaling provides the maximum achievable rate, and that for a given input covariance matrix $\mathbf{Q}$, when $\mathbf{H}$ is known perfectly at both transmitter and receiver, the following rate is achievable:

$$R = \log_2 \det\Big(\mathbf{I} + \frac{1}{N_0}\mathbf{H}\mathbf{Q}\mathbf{H}^H\Big)(\text{bit/s/Hz}). \tag{10}$$

We note that the channel matrix $\mathbf{H}$ also depends on $\boldsymbol{\theta}$. Thus, for the total power P_t, the problem of the achievable rate optimization for the considered system can be mathematically stated as:

$$\underset{\boldsymbol{\theta},\mathbf{Q}}{\text{maximize}}\ f(\boldsymbol{\theta},\mathbf{Q}) = \ln\det\Big(\mathbf{I} + \mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}\mathbf{Z}^H(\boldsymbol{\theta})\Big) \tag{11a}$$

$$\text{subject to}\ \ \mathrm{Tr}(\mathbf{Q}) \le P_t; \mathbf{Q} \succeq \mathbf{0}; \tag{11b}$$

$$\big|\theta_l\big| = 1, l = 1, 2, \ldots, N_{\mathrm{ris}}. \tag{11c}$$

where

$$\mathbf{Z}(\boldsymbol{\theta}) = \bar{\mathbf{H}}_{\mathrm{DIR}} + \mathbf{H}_2\mathbf{F}(\boldsymbol{\theta})\bar{\mathbf{H}}_1 \tag{12}$$

$$\bar{\mathbf{H}}_{\mathrm{DIR}} = \mathbf{H}_{\mathrm{DIR}}/\sqrt{N_0} \tag{13}$$

$$\bar{\mathbf{H}}_1 = \mathbf{H}_1\sqrt{\beta_{\mathrm{INDIR}}^{-1}/N_0}. \tag{14}$$

III. Solution Approach via Projected Gradient Method

In contrast to the conventional MIMO channel where the water-filling algorithm can be used to efficiently find the maximum achievable rate, problem (11) is nonconvex and thus difficult to solve. Further, we note that the objective is neither convex nor concave in the involved variables.

Previously proposed methods for rate optimization in RIS communication systems were primarily based on the *alternating optimization* (AO) technique [9], [24]. The main idea of this method is that the RIS phase shifts and the covariance matrix are optimized in an alternating fashion, each independently of the other. This method is motivated by the fact that the optimization over one variable can be performed efficiently (i.e., in closed form) while others are kept fixed. Although the AO method is easy to implement, it may require many iterations to converge, especially when the number of RIS elements is very large (which corresponds to the case in which the RIS is the most useful). In other words, the simplicity of an iteration in the AO method does not necessarily translate into low actual run time.

[2]Note that this achievable rate does not correspond to the channel *capacity*, as we do not consider the possibility of encoding the transmitted data into the phase shift values of the RIS. If such encoding is performed, the capacity of the RIS-aided MIMO system may be achieved [28].

Motivated by the above discussion, we propose an optimization method to solve (11), based on the PGM presented in [29]. Our proposed method is motivated by the fact that the projection onto the feasible set (albeit nonconvex with respect to $\boldsymbol{\theta}$) can be performed efficiently.

A. Description of Proposed Algorithm

To describe the proposed algorithm, we define the following two sets:

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{C}^{N_{\rm ris}\times 1} : \left|\theta_l\right| = 1, l = 1, 2, \ldots, N_{\rm ris}\} \tag{15}$$

$$\mathcal{Q} = \{\mathbf{Q} \in \mathbb{C}^{N_t \times N_t} : \mathrm{Tr}(\mathbf{Q}) \leq P_t; \mathbf{Q} \succeq \mathbf{0}\} \tag{16}$$

It is clear that the feasible set of (11) is the Cartesian product of Θ and $\mathcal{Q}$. We denote by $P_{\mathcal{U}}(\mathbf{u})$ the Euclidean projection from a point $\mathbf{u}$ onto a set $\mathcal{U}$, i.e., $P_{\mathcal{U}}(\mathbf{u}) = \arg\min_{\mathbf{x}}\{||\mathbf{x}-\mathbf{u}|| : \mathbf{x} \in \mathcal{U}\}$.

The proposed algorithm is outlined in Algorithm 1 and follows the projected gradient method in order to solve (11). The main idea behind Algorithm 1 is as follows. Starting from an arbitrary point $(\boldsymbol{\theta}_0, \mathbf{Q}_0)$, we move in each iteration in the direction of the gradient of $f(\boldsymbol{\theta}, \mathbf{Q})$. The size of this move is determined by the step size $\mu > 0$ (see Section IV for details regarding the choice of an appropriate step size). As a result of this step, the resulting updated point may lie outside of the feasible set. Therefore, before the next iteration, we project the newly computed points $\boldsymbol{\theta}$ and $\mathbf{Q}$ onto Θ and $\mathcal{Q}$, respectively. As shall be seen shortly, the projection onto Θ or $\mathcal{Q}$ can be determined in closed form. Another important remark concerning Algorithm 1 is also in order. Since (11) involves complex variables, we adopt the complex-valued gradient defined in [30, Eq. (4.37)]. In particular, it is proved that the directions where $f(\boldsymbol{\theta}, \mathbf{Q})$ has maximum rate of change with respect to $\boldsymbol{\theta}$ and $\mathbf{Q}$ are $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mathbf{Q})$ and $\nabla_{\mathbf{Q}} f(\boldsymbol{\theta}, \mathbf{Q})$, respectively [30, Theorem 3.4].

In our method, all optimization variables are updated simultaneously in each iteration. This is in sharp contrast to the AO method, in which each iteration only updates a single variable. As a result, the proposed method converges much faster than the AO method as we demonstrate via extensive numerical results in Section VI.

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

Let $\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q}) = (\mathbf{I} + \mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}\mathbf{Z}^H(\boldsymbol{\theta}))^{-1}$. Then we have the following result.

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

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mathbf{Q}) = \mathrm{vec}_d\left(\mathbf{H}_2^H \mathbf{K}(\boldsymbol{\theta}, \mathbf{Q}) \mathbf{Z}(\boldsymbol{\theta}) \mathbf{Q} \bar{\mathbf{H}}_1^H\right) \tag{17a}$$

$$\nabla_{\mathbf{Q}} f(\boldsymbol{\theta}, \mathbf{Q}) = \mathbf{Z}^H(\boldsymbol{\theta}) \mathbf{K}(\boldsymbol{\theta}, \mathbf{Q}) \mathbf{Z}(\boldsymbol{\theta}). \tag{17b}$$

Proof: See Appendix A. ∎

C. Projection onto Θ and $\mathcal{Q}$

We now show that the projection operations in Algorithm 1 can be carried out very efficiently, and thus Algorithm 1 indeed requires low complexity to implement. Note that the constraint $\left|\theta_l\right| = 1$ means that θ_l should lie on the unit circle in the complex plane. Thus, it is straightforward to see that, for a given point $\mathbf{u} \in \mathbb{C}^{N_{\rm ris}\times 1}$, $P_\Theta(\mathbf{u})$ is the vector $\bar{\mathbf{u}}$ where

$$\bar{u}_l = \begin{cases} \frac{u_l}{|u_l|} & u_l \neq 0 \\ e^{j\phi}, \phi \in [0, 2\pi] & u_l = 0 \end{cases}, l = 1, \ldots, N_{\rm ris}. \tag{18}$$

Note that $\bar{u}_l$ can be any point on the unit circle if $u_l = 0$, and thus the projection onto Θ is not unique. Despite this issue, we are still able to prove the convergence of Algorithm 1, which is shown in the next section. Next we turn our attention to the projection onto $\mathcal{Q}$, which general problem has already been studied previously (e.g., in [31]). For a given $\mathbf{Y} \succeq \mathbf{0}$, the projection of $\mathbf{Y}$ onto $\mathcal{Q}$ is the solution of the following problem:

$$\underset{\mathbf{Q}}{\text{minimize}} \quad \|\mathbf{Q} - \mathbf{Y}\|^2 \tag{19a}$$

$$\text{subject to} \quad \mathrm{Tr}(\mathbf{Q}) \leq P_t; \mathbf{Q} \succeq \mathbf{0} \tag{19b}$$

Let $\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^H$ be the eigenvalue decomposition of $\mathbf{Y}$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{N_t})$. Now, we can write $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}^H$ for some $\mathbf{D} \succeq \mathbf{0}$ and $\mathrm{Tr}(\mathbf{Q}) = \mathrm{Tr}(\mathbf{D})$. Then, we obtain $\|\mathbf{Q} - \mathbf{Y}\|^2 = \|\boldsymbol{\Sigma} - \mathbf{D}\|^2$. Thus, $\mathbf{D}$ must be diagonal to be an optimal solution, i.e., $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_{N_t})$. Therefore, (19) is equivalent to the following program:

$$\underset{\{d_i\}}{\text{minimize}} \quad \sum\nolimits_{i=1}^{N_t} (d_i - \sigma_i)^2 \tag{20a}$$

$$\text{subject to} \quad \sum\nolimits_{i=1}^{N_t} d_i \leq P_t; d_i \geq 0 \tag{20b}$$

The solution to the above problem is achieved by the water-filling algorithm and is given as

$$d_i = (\sigma_i - \gamma)_+, \quad i = 1, \ldots, N_t, \tag{21}$$

where $\gamma \geq 0$ is the water level.

Algorithm 1 Proposed projected gradient method (PGM).

1: $\mathbf{Input}$: $\theta_0, \mathbf{Q}_0, \mu > 0$.
2: **for** $n = 1, 2, \ldots$ **do**
3: $\quad \theta_{n+1} = P_\Theta(\theta_n + \mu \nabla_\theta f(\theta_n, \mathbf{Q}_n))$
4: $\quad \mathbf{Q}_{n+1} = P_{\mathbf{Q}}(\mathbf{Q}_n + \mu \nabla_{\mathbf{Q}} f(\theta_n, \mathbf{Q}_n))$
5: **end for**

D. Improved Convergence Rate by Data Scaling

For any first-order method, exploiting the structure of the optimization problem is key to speed up its convergence. In this regard, we remark that the effective channel via the RIS is $\sqrt{\beta_{\rm INDIR}^{-1}}\mathbf{H}_2\mathbf{F}(\boldsymbol{\theta})\bar{\mathbf{H}}_1$ which can be some orders of magnitude weaker or stronger than the direct link $\bar{\mathbf{H}}_{\rm DIR}$. This unbalanced data in (11) makes Algorithm 1 converge slowly.

To increase the convergence speed of Algorithm 1, we propose a change of variable as follows:

$$\bar{\mathbf{Q}} = k^2 \mathbf{Q} \tag{22a}$$

$$\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}/k \tag{22b}$$

$$\bar{\mathbf{H}}_{\rm DIR} = \mathbf{H}_{\rm DIR}/(k\sqrt{N_0}) \tag{22c}$$

for some $k > 0$. Accordingly, the equivalent optimization problem with respect to the new variables $\bar{\boldsymbol{\theta}}$ and $\bar{\mathbf{Q}}$ reads

$$\underset{\bar{\boldsymbol{\theta}},\bar{\mathbf{Q}}}{\text{maximize}}\ f(\bar{\boldsymbol{\theta}},\bar{\mathbf{Q}}) = \ln\det\left(\mathbf{I}+\mathbf{Z}(\bar{\boldsymbol{\theta}})\bar{\mathbf{Q}}\mathbf{Z}^H(\bar{\boldsymbol{\theta}})\right) \tag{23a}$$

$$\text{subject to}\ \ \mathrm{Tr}(\bar{\mathbf{Q}}) \le \bar{P}_t; \bar{\mathbf{Q}} \succeq \mathbf{0} \tag{23b}$$

$$\left|\bar{\theta}_l\right| = \frac{1}{k}, l = 1, 2, \ldots, N_{\text{ris}} \tag{23c}$$

where $\bar{P}_t = k^2 P_t$. The above change of variable step is equivalent to scaling the gradient of the original objective, which can improve the convergence rate. Now we solve (23) following the iterative procedure in Algorithm 1, but instead of $\mathbf{Q}$ and $\boldsymbol{\theta}$ we use their scaled versions $\bar{\mathbf{Q}}$ and $\bar{\boldsymbol{\theta}}$ which are defined by the expressions (22a) and (22b), respectively. Accordingly, the sets that contain all valid $\bar{\mathbf{Q}}$ and $\bar{\boldsymbol{\theta}}$ are denoted as $\bar{\Theta}$ and $\bar{Q}$, respectively. The projections of computed $\bar{\mathbf{Q}}$ and $\bar{\boldsymbol{\theta}}$ onto $\bar{\Theta}$ and $\bar{Q}$ are performed in the same way as the projections in the previous subsections, and the constraints (2) and (5) are replaced by (23b) and (23c), respectively. Also, it should be pointed out that $\bar{\mathbf{H}}_{\text{DIR}}$ is scaled (i.e., divided by k) in (22c) compared to (13). An appropriate value for k should reflect the difference between the direct and indirect links. When the direct link is absent, k should take into account the difference between the feasible sets of $\mathbf{Q}$ and $\boldsymbol{\theta}$. From our extensive numerical experiments, an appropriate value for k, depending on the presence or absence of the direct link, is given by[3]

$$k = \begin{cases} 10\max\{1, \frac{1}{\sqrt{P_t}}\}\sqrt{\frac{1}{\beta_{\text{INDIR}}^{-1/2}}\frac{\|\mathbf{H}_{\text{DIR}}\|}{\|\mathbf{H}_2\mathbf{H}_1\|}}, & \mathbf{H}_{\text{DIR}} \neq \mathbf{0} \\ 10, & \mathbf{H}_{\text{DIR}} = \mathbf{0}. \end{cases} \tag{24}$$

Based on the previous expressions, we can see that the PGM requires only the knowledge of the *cascaded* channel, and not the individual channels $\mathbf{H}_1$ and $\mathbf{H}_2$, for the indirect link. Estimation of this channel can be performed at the receiver or the transmitter in TDD mode [32], [33].

IV. Convergence and Complexity Analysis

A. Convergence Analysis

In this subsection we prove the convergence of Algorithm 1 for solving (23), following the framework in [29]. To achieve this, we first show that $f(\bar{\boldsymbol{\theta}},\bar{\mathbf{Q}})$ has a Lipschitz continuous gradient with a Lipschitz constant L, and then assert that Algorithm 1 is convergent if the step size satisfies $\mu \le \frac{1}{L}$. For the first part of the proof, recall that a function $f(\mathbf{x})$ is said to be L-Lipschitz continuous (also known as L-smooth) over a set $\mathcal{X}$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ we have

$$||\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})|| \le L||\mathbf{x}-\mathbf{y}||. \tag{26}$$

In our present context, the inequality (26) corresponds to (25), to prove which we may make use of the following lemma.

[3]Since the gradients with respect to $\mathbf{Q}$ and $\boldsymbol{\theta}$ are of different sizes, using two different step sizes for each of them can actually increase the convergence rate of PGM. In order to preserve the single step size, we introduce the scaling factor k which provides the same effect as if two independent step sizes were used. The value of the scaling factor k is obtained in a heuristic manner, by performing numerical experiments.

Lemma 2. *The following inequalities hold for* $\nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}},\bar{\mathbf{Q}})$ *and* $\nabla_{\bar{\mathbf{Q}}} f(\bar{\boldsymbol{\theta}},\bar{\mathbf{Q}})$

$$\begin{aligned}&||\nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1) - \nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_2,\bar{\mathbf{Q}}_2)|| \\ &\le (ab+ab^3\bar{P}_t)||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2|| + (a^2\bar{P}_t + 2a^2b^2\bar{P}_t^2)||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2||\end{aligned} \tag{27}$$

$$\begin{aligned}&||\nabla_{\bar{\mathbf{Q}}} f(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1) - \nabla_{\bar{\mathbf{Q}}} f(\bar{\boldsymbol{\theta}}_2,\bar{\mathbf{Q}}_2)|| \\ &\qquad \le b^4||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2|| + (2ab+2ab^3\bar{P}_t)||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2||\end{aligned} \tag{28}$$

where

$$a = \lambda_{\max}(\bar{\mathbf{H}}_1)\lambda_{\max}(\mathbf{H}_2) \tag{29}$$

$$b = \lambda_{\max}(\bar{\mathbf{H}}_{\text{DIR}}) + k^{-1}\lambda_{\max}(\bar{\mathbf{H}}_1)\lambda_{\max}(\mathbf{H}_2). \tag{30}$$

Proof: See Appendix B. ∎

With the aid of the above lemma, we assert the smoothness of $f(\bar{\boldsymbol{\theta}},\bar{\mathbf{Q}})$ in the next theorem.

Theorem 1. *The objective* $f(\bar{\boldsymbol{\theta}},\bar{\mathbf{Q}})$ *is* L*-smooth with a constant* L *given by*

$$L = \sqrt{\max(L_{\bar{\boldsymbol{\theta}}}^2, L_{\bar{\mathbf{Q}}}^2)}, \tag{31}$$

where

$$\begin{aligned}L_{\bar{\boldsymbol{\theta}}}^2 =&(2ab^5+4a^2b^2) + (a^3b+2ab^7+8a^2b^4)\bar{P}_t \\ &+ (3a^3b^3+a^4+4a^2b^6)\bar{P}_t^2 + (2a^3b^5+4a^4b^2)\bar{P}_t^3 \\ &+ 4a^4b^4\bar{P}_t^4\end{aligned} \tag{32}$$

$$\begin{aligned}L_{\bar{\mathbf{Q}}}^2 =&(a^2b^2+b^8+2ab^5) + (2a^2b^4+a^3b+2ab^7)\bar{P}_t \\ &+ (a^2b^6+3a^3b^3)\bar{P}_t^2 + 2a^3b^5\bar{P}_t^3\end{aligned} \tag{33}$$

Proof: Theorem 1 follows immediately from Lemma 2 and the inequality

$$2||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2|| \times ||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2|| \le ||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2||^2 + ||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2||^2.$$

Specifically, we have

$$\begin{aligned}&||\nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1) - \nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_2,\bar{\mathbf{Q}}_2)||^2 \\ &\qquad + ||\nabla_{\bar{\mathbf{Q}}} f(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1) - \nabla_{\bar{\mathbf{Q}}} f(\bar{\boldsymbol{\theta}}_2,\bar{\mathbf{Q}}_2)||^2 \\ &\qquad \le L_{\bar{\mathbf{Q}}}^2||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2||^2 + L_{\bar{\boldsymbol{\theta}}}^2||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2||^2 \\ &\qquad \le \max(L_{\bar{\boldsymbol{\theta}}}^2, L_{\bar{\mathbf{Q}}}^2)\left(||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2||^2 + ||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2||^2\right).\end{aligned}$$

Taking the square root of both sides of the above inequality, we can see that $\sqrt{\max(L_{\bar{\boldsymbol{\theta}}}^2, L_{\bar{\mathbf{Q}}}^2)}$ is a Lipschitz constant of the gradient of $f(\bar{\boldsymbol{\theta}},\bar{\mathbf{Q}})$; this completes the proof. ∎

The convergence of Algorithm 1 is stated in the following theorem.

Theorem 2. *Assume the step size satisfies* $\mu < \frac{1}{L}$*, where* L *is given in* (31)*. Then the iterates* $(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)$ *generated by Algorithm 1 are bounded. Let* $(\boldsymbol{\theta}^*,\mathbf{Q}^*)$ *be any accumulation point of the set* $\{(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)\}$*, then,* $(\boldsymbol{\theta}^*,\mathbf{Q}^*)$ *is a critical point of* (11)*.*

Proof: See Appendix C. ∎

Before proceeding further, a subtle point regarding the convergence of Algorithm 1 is worth mentioning. Specifically, the proposed method is provably convergent to a critical point

$$\left(\left\|\nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1) - \nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_2, \bar{\mathbf{Q}}_2)\right\|^2 + \left\|\nabla_{\bar{\mathbf{Q}}} f(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1) - \nabla_{\bar{\mathbf{Q}}} f(\bar{\boldsymbol{\theta}}_2, \bar{\mathbf{Q}}_2)\right\|^2\right)^{1/2} \leq L\left(\left\|\bar{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_2\right\|^2 + \left\|\bar{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_2\right\|^2\right)^{1/2} \quad (25)$$

of the considered problem, also known as a stationary solution, which satisfies the *necessary* optimality conditions for (11). However, since (11) is nonconvex, these optimality conditions may not be sufficient in general, and thus the solution obtained from Algorithm 1 may not be globally optimal. However, it is often the case that with a good initialization, a stationary solution is good enough for practical applications. We note that the same comments also apply to the AO method proposed in [24].

B. Complexity Analysis

In this subsection, we analyze the computational complexity of Algorithm 1 (i.e., the PGM).[4] To simplify the analysis while still providing a good approximation to the complexity of Algorithm 1, we concentrate on the number of complex multiplications required per iteration. To this end, we first recall some fundamental results. Specifically, the multiplication of $\mathbf{A} \in \mathbb{C}^{m\times n}$ and $\mathbf{B} \in \mathbb{C}^{n\times p}$ needs mnp complex multiplications when $\mathbf{A}$ and $\mathbf{B}$ are dense matrices.[5] This complexity reduces to mn for the case of a square diagonal matrix $\mathbf{B} \in \mathbb{C}^{n\times n}$. Calculating $\mathrm{vec}_d(\mathbf{ABC})$, $\mathbf{A} \in \mathbb{C}^{m\times n}$, $\mathbf{B} \in \mathbb{C}^{n\times p}$ and $\mathbf{C} \in \mathbb{C}^{p\times m}$, needs $mp(n+1)$ complex multiplications, which is justified as follows. Multiplying a row of $\mathbf{A}$ with $\mathbf{B}$ requires np complex multiplications and multiplying the resulting row vector with the corresponding column of $\mathbf{C}$ requires a further p complex multiplications.

It is obvious that the complexity of the proposed method is determined by Steps 3 and 4 in Algorithm 1. The computation of $\mathbf{Z}(\boldsymbol{\theta})$ is dominated by that of the term $\mathbf{H}_2\mathbf{F}(\boldsymbol{\theta})\bar{\mathbf{H}}_1$, which requires $N_r N_{\mathrm{ris}} + N_r N_t N_{\mathrm{ris}}$ complex multiplications. To compute $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mathbf{Q})$, we also need to compute the term $\mathbf{A} = \mathbf{K}(\boldsymbol{\theta}, \mathbf{Q})\mathbf{Z}(\boldsymbol{\theta}) \in \mathbb{C}^{N_r \times N_t}$. Instead of directly computing $\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q}) = (\mathbf{I} + \mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}\mathbf{Z}^H(\boldsymbol{\theta}))^{-1}$ using matrix inversion and then multiplying $\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q})$ with $\mathbf{Z}(\boldsymbol{\theta})$, we note that $\mathbf{A}$ is in fact the solution to the linear system $\big(\mathbf{I} + \mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}\mathbf{Z}^H(\boldsymbol{\theta})\big)\mathbf{X} = \mathbf{Z}(\boldsymbol{\theta})$. To form $\mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}\mathbf{Z}^H(\boldsymbol{\theta})$, we first need $N_r N_t^2$ multiplications to achieve $\mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}$ and then $(N_r^2 + N_r)N_t/2$ to multiply $\mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}$ with $\mathbf{Z}^H(\boldsymbol{\theta})$. Solving the linear system using Cholesky decomposition, by solving two triangular systems using forward and backward substitution, requires a complexity which is $\mathcal{O}(N_r^3 + N_r^2 N_t)$. In summary, the computation of $\mathbf{A}$ takes $\mathcal{O}(N_t^2 N_r + \frac{3}{2}N_t N_r^2 + N_r^3)$ multiplications. Next, the computation of $\mathbf{AQ}$ requires $N_r N_t^2$ complex multiplications. To calculate $\mathrm{vec}_d(\mathbf{H}_2^H \mathbf{A}\mathbf{Q}\bar{\mathbf{H}}_1^H)$, we need $N_{\mathrm{ris}} N_t (N_r + 1)$ complex multiplications. As $\mathbf{A}$ is also common to (17b), the complexity of computing $\nabla_{\mathbf{Q}} f(\boldsymbol{\theta}, \mathbf{Q})$ is only $N_r N_t^2$. In summary, the computational complexity of $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mathbf{Q})$ and $\nabla_{\mathbf{Q}} f(\boldsymbol{\theta}, \mathbf{Q})$ is $\mathcal{O}\big(2N_{\mathrm{ris}} N_t N_r + 2N_t^2 N_r + \frac{3}{2}N_t N_r^2 + N_r^3 + N_r N_{\mathrm{ris}} + N_t N_{\mathrm{ris}}\big)$. When N_{ris} is much larger than N_t and N_r, then the complexity can be approximated by $\mathcal{O}\big(N_{\mathrm{ris}} N_t N_r\big)$.

Next, multiplying μ with $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mathbf{Q})$ and then projecting the result onto Θ requires $3N_{\mathrm{ris}}$ complex multiplications. Similarly, we need $N_t^2/2$ operations to multiply μ with $\nabla_{\mathbf{Q}} f(\boldsymbol{\theta}_n, \mathbf{Q}_n)$. The projection of $\mathbf{Q}_n + \mu\nabla_{\mathbf{Q}} f(\boldsymbol{\theta}_n, \mathbf{Q}_n)$ onto $\mathcal{Q}$ requires: $\mathcal{O}(N_t^3)$ operations for the eigenvalue decomposition, $\mathcal{O}(N_t^2)$ operations for the water-filling algorithm in (21) and $N_t^2 + (N_t^2 + N_t)N_t/2$ operations for the matrix multiplication $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}^H$. Therefore, the complexity for the update and projection operations in Step 4 is given by $\mathcal{O}(\frac{3}{2}N_t^3)$. Thus, the per-iteration complexity of Algorithm 1 is finally determined as

$$C_{\mathrm{PGM,IT}} = \mathcal{O}(2N_{\mathrm{ris}} N_t N_r + 2N_t^2 N_r + \frac{3}{2}N_t N_r^2 + N_r^3 + N_r N_{\mathrm{ris}} + N_t N_{\mathrm{ris}} + 3N_{\mathrm{ris}} + \frac{3}{2}N_t^3), \quad (34)$$

while the total complexity C_{PGM} also depends from the number of required iterations I_{PGM}. The computational complexity of the proposed PGM and the AO method from [24] are presented in Table I. The complexity of the AO is expressed with respect to the number of outer iterations I_{OI}, where one outer iteration is actually a sequence of $N_{\mathrm{ris}} + 1$ conventional iterations. Further details and discussion on this complexity comparison will be presented in Subsection VI-D.

TABLE I
COMPARISON OF THE COMPUTATIONAL COMPLEXITY REQUIRED BY THE PROPOSED PGM METHOD AND THE AO METHOD TO REACH 95 % OF THE AVERAGE ACHIEVABLE RATE AT THE 500TH ITERATION.

Direct link	N_{ris}	I_{PGM}	$C_{\mathrm{PGM,IT}}$	C_{PGM}	I_{OI}	C_{AO}
Present	100	19	9436	179284	1	394304
	225	6	19311	115866	1	862304
	400	4	33136	132544	1	1517504
	625	3	50911	152733	1	2359904
Blocked	100	2	9436	18872	1	394304
	225	2	19311	38622	1	862304
	400	2	33136	66272	1	1517504
	625	2	50911	101822	1	2359904

C. Improved Convergence by Backtracking Line Search

It often occurs that the Lipschitz constant given in (31) is much larger than the best Lipschitz constant for the gradient of the objective. The corresponding step size required (according to Theorem 2) to guarantee the convergence is then very small, and adopting this step size can lead to a very slow convergence. To speed up the convergence of the proposed PGM, we can employ a backtracking line search to find a possibly larger step size at each iteration. In the

[4]Although the PGM is actually implemented in the scaled-variable form described in Subsection III-D, this scaling does not affect the complexity of the PGM which is equal to the complexity of Algorithm 1. Therefore, we analyze the complexity of Algorithm 1 in the sequel.

[5]Special algorithms can reduce the complexity further, but this is not our focus in this paper.

following, we present a line search procedure, based on the Armijo–Goldstein condition [34], that is numerically shown to be efficient for our considered problem.

Let $L_0 > 0$, $\delta > 0$ be a small constant, and $\rho \in (0,1)$. In Steps 3 and 4 of Algorithm 1, we replace the step size μ by $L_o\rho^{k_n}$ and obtain (35a) and (35b), where k_n is the smallest nonnegative integer that satisfies (35c).

$$\boldsymbol{\theta}_{n+1} = P_{\Theta}(\boldsymbol{\theta}_n + L_o\rho^{k_n}\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_n, \mathbf{Q}_n)) \tag{35a}$$

$$\mathbf{Q}_{n+1} = P_{\mathcal{Q}}(\mathbf{Q}_n + L_o\rho^{k_n}\nabla_{\mathbf{Q}} f(\boldsymbol{\theta}_n, \mathbf{Q}_n)) \tag{35b}$$

$$f(\boldsymbol{\theta}_{n+1}, \mathbf{Q}_{n+1}) \geq f(\boldsymbol{\theta}_n, \mathbf{Q}_n) + \delta\big(||\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n||^2 + ||\mathbf{Q}_{n+1} - \mathbf{Q}_n||^2\big). \tag{35c}$$

The above backtracking line search can be found through an iterative procedure, which is guaranteed to terminate after a finite number of iterations since $f(\bar{\boldsymbol{\theta}}, \bar{\mathbf{Q}})$ is L-smooth. It is easy to see that the convergence of Algorithm 1 (i.e., Theorem 2) still holds when this procedure is used to find the step size. We remark that the line search described above results in increased per-iteration complexity. Suppose that the line search stops after I_{LS} steps, the additional complexity is $\mathcal{O}\big(I_{\mathrm{LS}}(3N_{\mathrm{ris}}+2N_t^3)\big)$. However, this computational cost turns out to be immaterial, since the line search can significantly reduce the required number of iterations and hence the actual overall run time.

V. Total FPSL Ratio - A Metric of RIS Applicability

It can be very useful to have a first-order estimate of the benefit (if any) provided to a wireless communication system by adding an RIS. We can achieve this by considering the total FSPL of the indirect and direct links. Note that when speaking about the "total FSPL" of the indirect link, we require (in contrast to (8)) a definition which takes into account also the RIS phase shift values.

The computation of the total FSPL of the indirect link is an intractable problem in a MIMO system, since the optimal RIS element phase shifts are *a priori* unknown and can only be obtained by implementing an iterative optimization method. This problem was approximately tackled only for the single-stream scenario in [33], and the obtained results were then used in [35] to quantify the performance of RISs in the far field regime (which is also the case considered in this paper), but only for Rayleigh and deterministic LOS channels. To overcome this issue, we consider the total FSPL of the indirect link in a SISO system, which is given by

$$\beta_{\mathrm{INDIR,T}}^{-1} = \beta_{\mathrm{INDIR}}^{-1}\mathbb{E}\left\{|\mathbf{h}_2\mathbf{F}(\boldsymbol{\theta})\mathbf{h}_1|^2\right\}, \tag{36}$$

where $\mathbf{h}_1$ models the channel between the transmit antenna and the RIS, and $\mathbf{h}_2$ models the channel between the RIS and the receive antenna. The optimal RIS element phase shift values in (36) satisfy $\phi_i = -\arg\{h_2(i)h_1(i)\}$ and as a result we have $\mathbf{h}_2\mathbf{F}(\boldsymbol{\theta})\mathbf{h}_1 = \sum_{i=1}^{N_{\mathrm{ris}}}|h_2(i)h_1(i)|$. As all of the terms $|h_2(i)h_1(i)|$ follow the same distribution, we may write

$$\mathbb{E}\{\sum_{i=1}^{N_{\mathrm{ris}}}|h_2(i)h_1(i)|\} = N_{\mathrm{ris}}\mathbb{E}\{|h_2(1)h_1(1)|\}. \tag{37}$$

From Jensen's inequality we obtain

$$\mathbb{E}\left\{|\mathbf{h}_2\mathbf{F}(\boldsymbol{\theta})\mathbf{h}_1|^2\right\} = \mathbb{E}\left\{\left|\sum_{i=1}^{N_{\mathrm{ris}}}|h_2(i)h_1(i)|\right|^2\right\} \geq \left(\mathbb{E}\left\{\sum_{i=1}^{N_{\mathrm{ris}}}|h_2(i)h_1(i)|\right\}\right)^2 = N_{\mathrm{ris}}^2\left(\mathbb{E}\{|h_2(1)h_1(1)|\}\right)^2. \tag{38}$$

Substituting (38) into (36), we finally obtain

$$\beta_{\mathrm{INDIR,T}}^{-1} \geq \beta_{\mathrm{INDIR}}^{-1}N_{\mathrm{ris}}^2\left(\mathbb{E}\{|h_2(1)h_1(1)|\}\right)^2, \tag{39}$$

which constitutes an upper-bound on the total FSPL. The total FSPL of the direct link in a SISO system is given by $\beta_{\mathrm{DIR,T}} = \beta_{\mathrm{DIR}}$, since the direct link does not alter the average signal power. Finally, the ratio between the total FSPL of the indirect and direct links can be expressed as

$$T = \frac{\beta_{\mathrm{INDIR,T}}}{\beta_{\mathrm{DIR,T}}} = \frac{16}{\lambda^2}\frac{(d_1d_2)^2}{d_0^{\alpha_{\mathrm{DIR}}}}\frac{1}{(l_t/d_1 + l_r/d_2)^2N_{\mathrm{ris}}^2E}, \tag{40}$$

where $E = \left(\mathbb{E}\{|h_2(1)h_1(1)|\}\right)^2$. The obtained T serves as a first-order measure of the applicability of an RIS for a given communication scenario[6]. For $T > 1$, the direct link is expected to be always stronger than the indirect link, even if the RIS phase shifts are optimally adjusted. Consequently, the RIS is capable of achieving limited performance gains with respect to the case when only the direct link is utilized for communication. For $T < 1$, the indirect link with the optimized RIS phase shifts is stronger than the direct link and the gains of using the RIS are usually more substantial.

VI. Simulation Results

In this section, we evaluate the achievable rate of the proposed optimization algorithm with the aid of Monte Carlo simulations. First, the study is conducted for a typical outdoor propagation environment in two different scenarios: with the direct link present and with the direct link blocked. For the case study where the direct link is present, we utilize three benchmark schemes. The first benchmark scheme is based on the implementation of the AO method from [24]. The second and third benchmark schemes are based on the use of the PGM in the case where only the indirect link is active and where only the direct link is active, respectively. In the case where the direct link is blocked, we only consider the AO method as the benchmark scheme. Additionally, we show the variation of the achievable rate with the number of RIS elements. Furthermore, we study the suitability of RIS-aided wireless communications (with the proposed optimization method) for implementation in indoor propagation environments. We also present a comparison of the proposed and benchmark schemes in terms of computational complexity and run time. In addition, we analyze the sensitivity and robustness of the proposed PGM. Finally, we evaluate the influence of data scaling and the line search procedure.

[6]Since (40) is derived for a SISO system, T is realistically only a rough measure of the applicability of an RIS in a MIMO system. However, as we shall see in Section V, this metric is quite useful for MIMO scenarios.

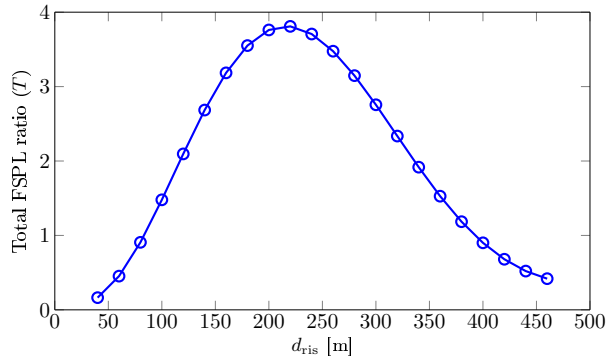
In the following simulation setup, the parameters are $f = 2\,\text{GHz}$ (i.e., $\lambda = 15\,\text{cm}$), $s_t = s_r = \lambda/2 = 7.5\,\text{cm}$, $s_{\text{ris}} = \lambda/2 = 7.5\,\text{cm}$, $D = 500\,\text{m}$, $N_t = 8$, $N_r = 4$, $\alpha_{\text{DIR}} = 3$, $N_{\text{ris}} = 225$, $K = 1$, $P_t = 0\,\text{dB}$ and $N_0 = -120\,\text{dB}$. The RIS elements are placed in a 15×15 square formation so that the area of the RIS is slightly larger than $1\,\text{m}^2$. The line search procedure for the proposed gradient algorithms utilizes the parameters $L_0 = 10^4$, $\delta = 10^{-5}$ and $\rho = 1/2$. Also, the minimum allowed step size value is the largest step size value lower than 10^{-4}. Unless otherwise specified, we assume the initial values $\boldsymbol{\theta} = [1\ 1\ \cdots\ 1]^T$ and $\mathbf{Q} = (P_t/N_t)\mathbf{I}$ for all optimization algorithms. To maintain compatibility with [24], we set the number of random initializations for the AO to $L_{\text{AO}} = 100$. All of the achievable rate results, except those for very large N_{ris} in Figs. 8 and 9, are averaged over 200 independent channel realizations.

A. Achievable Rate in Outdoor Environments

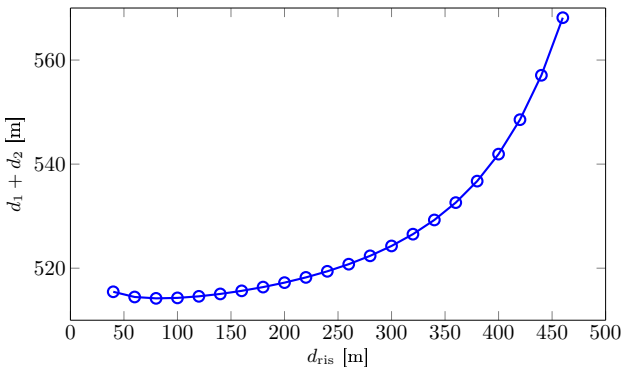
1) Direct link present: In this subsection, we present the achievable rate simulation results when the considered communication system is located in an outdoor environment. To obtain a more complete picture, the positions of the transmitter and the receiver, as well as the position of the RIS, are varied in simulations. In general, we analyze two cases for the transmitter and receiver positions: the transmitter and receiver are at substantially different distances from the plane containing the RIS ($l_t \neq l_r$), and the transmitter and receiver are at the same distance from the plane containing the RIS ($l_t = l_r$). In each of these cases, the position of the RIS is also varied.

In the first case, we assume $l_t = 20\,\text{m}$ and $l_r = 100\,\text{m}$. The variation of the FSPL ratio T given by (40) and the total indirect link length with the RIS position are shown in Fig. 2. We observe that the highest FSPL ratio T is obtained when the RIS is placed close to the center, and the lowest T is obtained when the RIS is placed close to the transmitter or the receiver. In other words, placing the RIS in the vicinity of the transmitter or the receiver ensures the lowest signal attenuation for the indirect communication link. It is interesting to note from Fig. 2b that in contrast to signal propagation principles for conventional communication systems, the total length of the indirect link (d_1+d_2) does not determine the total FSPL of that link, and that the relationship between these variables is not monotonic. For example, the minimum and the maximum of the FSPL ratio T in Fig. 2a are obtained for almost the same total length of the indirect link, as shown in Fig. 2b. Also, the largest indirect link length in Fig. 2b does not coincide with the highest FSPL of the indirect link in Fig. 2a. The reason for this is that the FSPL of the indirect link is determined by the *product* of the distances d_1 and d_2 rather than by their sum. Therefore, finding the optimal position for the RIS is not a straightforward task.

Based on the previous observations, we assume in further simulations that the RIS is placed in the vicinity of the transmitter ($d_{\text{ris}} = 40\,\text{m}$) or in the vicinity of the receiver ($d_{\text{ris}} = D - 40\,\text{m}$). The achievable rate results for the proposed PGM approach and for the benchmark schemes are shown



(a) The total FSPL ratio (T) versus d_{ris}.

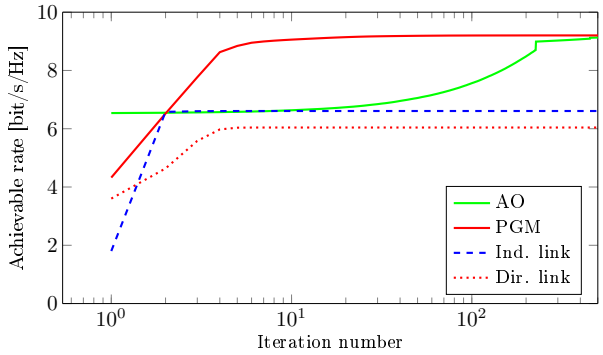


(b) The total indirect link length $(d_1 + d_2)$ versus d_{ris}.

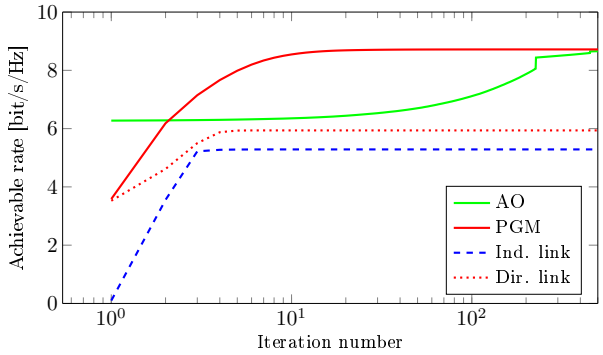
Fig. 2. The total FSPL ratio and the indirect link length for $l_t = 20\,\text{m}$ and $l_r = 100\,\text{m}$.

in Fig. 3. It can be seen that the proposed gradient-based optimization method converges relatively fast to the optimum achievable rate value. On the other hand, the AO requires significantly more iterations (at least one outer iteration, which consists of a sequence of $N_{\text{ris}} + 1$ conventional iterations [24]) to reach its optimum value. It can be also observed that the initial achievable rate for the AO is higher than for the PGM. The reason for this is that for the AO, we select from a large set of randomly generated RIS phase shift realizations and optimized $\mathbf{Q}$ matrices the ones that provide the highest achievable rate and use these as a starting point for the AO (thus, this specifies the initial achievable rate). In contrast to this, the initial achievable rate for the PGM is obtained for the aforementioned initial RIS phase shifts and $\mathbf{Q}$ matrix.

In addition, the RIS is capable of providing a significant enhancement of the achievable rate, which is proportional to the achievable rate of the indirect link. As expected, this gain is higher when the RIS is located in the vicinity of the transmitter, due to the lower FSPL of the indirect link. Finally, we observe that the FPSL ratio T, which is derived for a SISO system, is not entirely trustworthy for predicting the achievable rate in a MIMO system. Although the total FSPL of the indirect link is lower than the total FSPL of the direct link when the RIS is placed in the vicinity of the receiver in Fig. 2a, the direct link will ultimately provide a higher achievable rate in Fig. 3b.



(a) $d_{\text{ris}} = 40\,\text{m}$.



(b) $d_{\text{ris}} = D - 40\,\text{m}$.

Fig. 3. Average achievable rate of the PGM versus the benchmark schemes. Here $l_t = 20\,\text{m}$ and $l_r = 100\,\text{m}$.

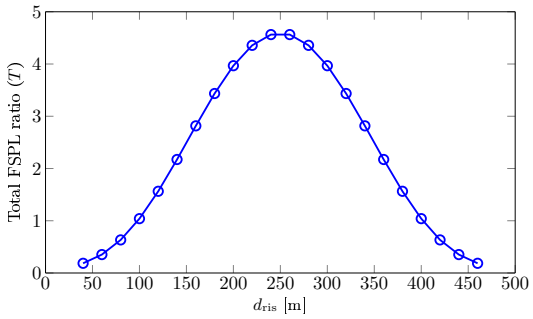
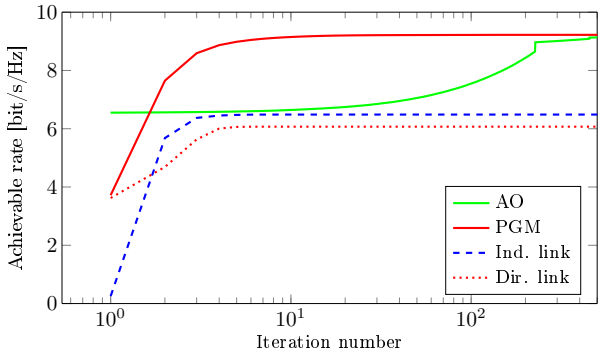


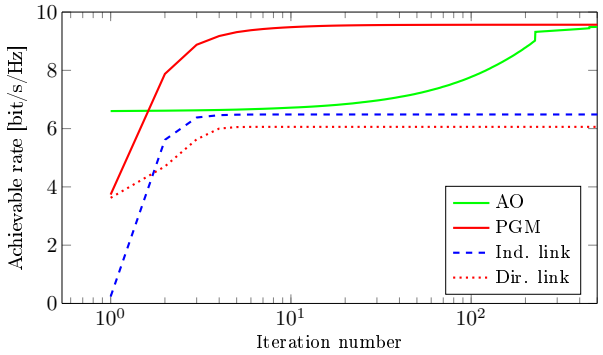
Fig. 4. The total FSPL ratio (T) for the case $l_t = l_r = 50\,\text{m}$.

In the second case, we assume $l_t = 50\,\text{m}$ and $l_r = 50\,\text{m}$. The variation of the total FSPL ratio T with the RIS position is shown in Fig. 4. It can be seen that T is perfectly symmetric due to the equal values of the distances l_t and l_r. The same is true for the total indirect link length, which is not shown for brevity reasons. The achievable rate of the proposed PGM approach versus the benchmark schemes is shown in Fig. 5. As expected, the PGM has a much higher convergence rate than the AO. Also, it can be seen that the achievable rate is slightly higher when the RIS is located in the vicinity of the receiver than in the vicinity of the transmitter. If the communication link are used individually, the indirect link has a slightly higher achievable rate than the direct link.

2) Direct link blocked: If the direct link between the transmitter and the receiver is blocked, the only means of



(a) $d_{\text{ris}} = 40\,\text{m}$.



(b) $d_{\text{ris}} = D - 40\,\text{m}$.

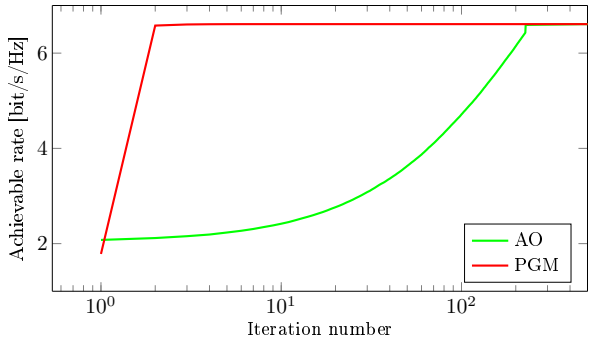
Fig. 5. Average achievable rate of the PGM versus the benchmark schemes. Here $l_t = l_r = 50\,\text{m}$.

signal transmission is via the RIS. It can be easily seen that the main observations made concerning the optimal RIS position in the previous subsection are also applicable here. Therefore, we analyze the achievable rate when the RIS is placed in the vicinity of the transmitter or in the vicinity of the receiver. The achievable rate results of the PGM versus the AO are shown in Fig. 6. In both cases, the optimal achievable[7] rates match the achievable rates of the second benchmark scheme in Fig. 2. The PGM requires only a few iterations to converge to the optimum value. On the other hand, the AO needs approximately $N_{\text{ris}} + 1$ iterations (i.e., one outer iteration) to reach the optimum value. Interestingly, the achievable rate enhancement during the first outer iteration is higher when the direct link is blocked. It seems that the absence of the direct link may have a significant influence on choosing the initial covariance matrix and RIS phase shifts, before starting the AO.

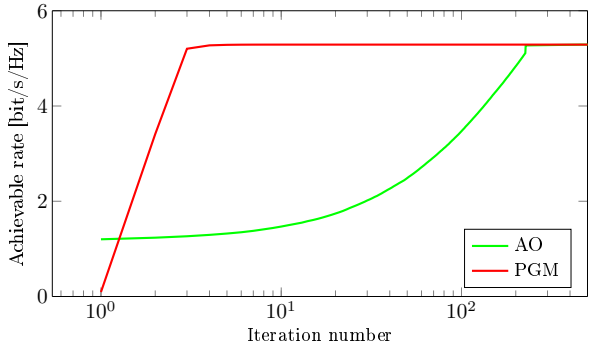
B. Scaling with N_{ris}

This subsection consists of three parts. First, we demonstrate the correctness of the expression (38) in Section V. Then, we show how increasing the number of RIS elements influences the achievable rate of the considered system. Finally, we present the trade-off between the operating frequency and the number of RIS elements.

[7]In our simulations, we take the "optimal achievable rate" to be that which is obtained at the final (i.e., 500th) iteration.



(a) $d_{\text{ris}} = 40\,\text{m}$.



(b) $d_{\text{ris}} = D - 40\,\text{m}$.

Fig. 6. Average achievable rate of the PGM versus the AO. The parameter setup is the same as in Fig. 2.

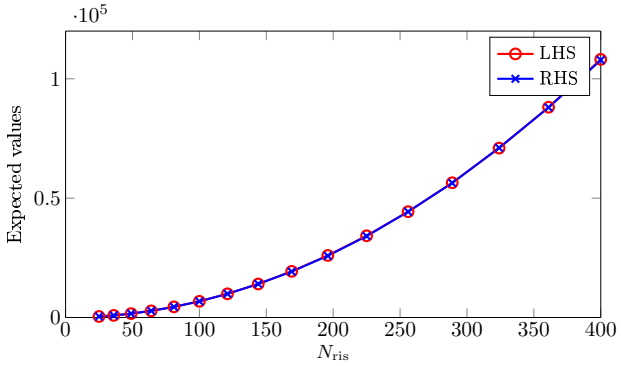


Fig. 7. Comparison of the left-hand side and right-hand side of (38).

To verify the correctness of the upper-bound expression in (38), we compare the values of the left and right hand sides of this expression in Fig. 7. As the expression pertains to single-antenna systems, we assume $N_t = N_r = 1$ in this simulation. For completeness, the presented results are computed and averaged for different positions of the RIS. The graph shows a very good match between the two sides of the aforementioned expression, which means that in practice the total FSPL of the indirect link in a SISO system is very well approximated by (39).

In general, it is not easy to assess the expected achievable rate for some arbitrary value of N_{ris}, when gradient-based optimization methods are applied. Therefore, to obtain a better understanding of the variation of the achievable rate with N_{ris}, we present a numerical evaluation of the achievable rate in Fig. 8. The parameter setup is the same as for Fig. 2 and $d_{\text{ris}} = 40\,\text{m}$. In this case, the physical size of the RIS is actually increasing, while the RIS is always operating in the far field [8]. As a result, we observe that there is an increase in the achievable rate when N_{ris} is doubled[8] and this increase becomes larger as we increase N_{ris}. Also, the slope of the achievable rate curve gradually reduces with N_{ris}, as a consequence of the logarithm function in the achievable rate expression.

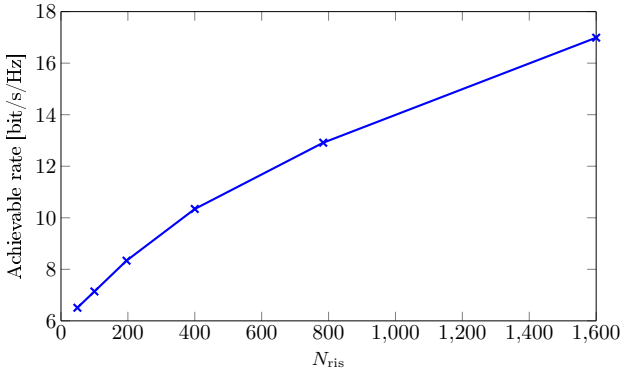


Fig. 8. Average achievable rate versus the number of RIS elements N_{ris}.

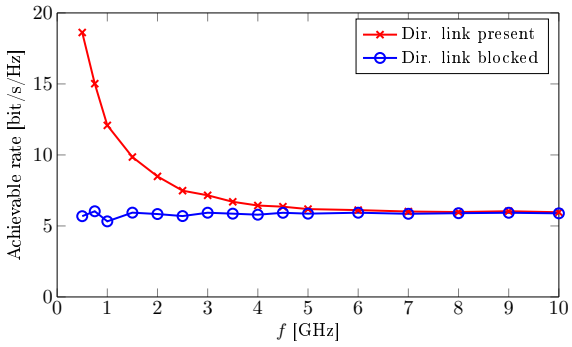


Fig. 9. Average achievable rate versus frequency f. The parameter setup is the same as for Fig. 2.

The number of RIS elements that can be placed on an RIS having a constant physical size, without causing coupling between the neighboring RIS elements, increases with the frequency of operation. Therefore, the RIS may consist of a large number of RIS elements, if it is intended to work at high frequencies. Motivated by this fact, we analyze the trade-off between the operating frequency and the number of RIS elements N_{ris}, and their influence on the achievable rate in the considered system. We assume that the RIS elements are placed in an RIS of size $1\,\text{m} \times 1\,\text{m}$ and $s_{\text{ris}} = \lambda/2$ at all frequencies. The achievable rate of the PGM versus the operating frequency f is shown in Fig. 9. In the frequency

[8]It should be noted that the number of RIS elements is not exactly doubled in simulations, since we aim to have a square RIS. Therefore, the achievable rate is computed and plotted for N_{ris} equal to 49, 196 and 784 instead of 50, 200 and 800, respectively.

range up to 5 GHz, the achievable rate of the considered system decreases primarily because of the FSPL increase of the direct link. At higher frequencies, the achievable rate remains almost constant regardless of whether the direct link is present or blocked. In other words, the FSPL of the direct link is so high in this case that the direct link becomes practically useless for communicating information. On the other hand, the indirect link has approximately the same achievable rate across the entire frequency range. It is because the increase in the FSPL of the indirect link is compensated by the increased number of RIS elements in the considered system. Finally, we conclude that the direct link is only useful at lower frequencies, while the indirect link can be used at all frequencies if a sufficient number of RIS elements is provided.

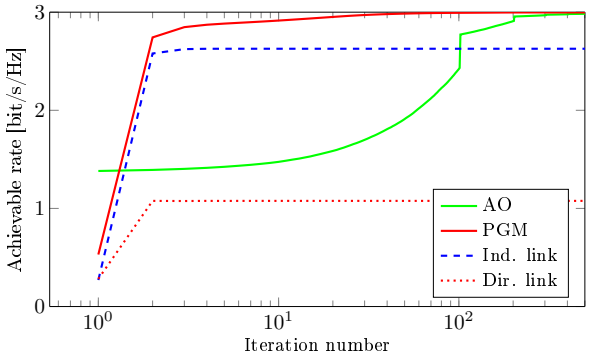
C. Achievable Rate in Indoor Environments

All of the previous simulation results are obtained for a wireless communication system operating in an outdoor environment. To further demonstrate the effectiveness of the proposed gradient-based optimization method, we consider its implementation in an indoor environment. Since the communication distances are now much smaller and the communication bandwidths are usually larger (i.e., typically 20/22 MHz), the following simulation parameters have the following altered values: $D = 30\,\mathrm{m}$, $d_{\mathrm{ris}} = 5\,\mathrm{m}$, $l_t = 3\,\mathrm{m}$, $l_r = 7\,\mathrm{m}$, $N_{\mathrm{ris}} = 100$, $P_t = -30\,\mathrm{dB}$ and $N_0 = -100\,\mathrm{dB}$.

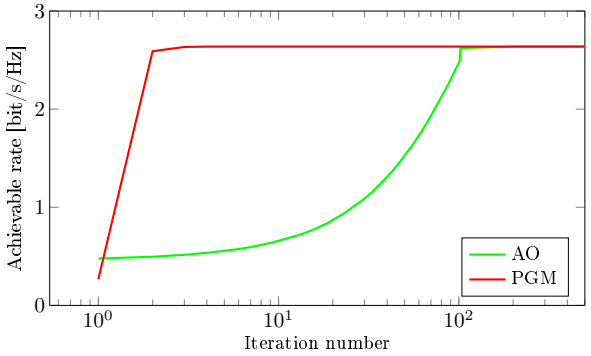
The simulation results of the proposed optimization method versus the benchmark schemes in an indoor environment are presented in Fig. 10. The PGM again requires a lower number of iterations than the AO to converge to the optimal achievable rate. In contrast to the previous simulation results, the achievable rate in indoor environments is almost entirely determined by the indirect link signal transmission, which can explained by the following argument. Reducing the distances in the considered communication system results in the reduction of the total FSPLs, and the total FSPL of the indirect link is particularly affected by this, since it is inversely proportional to the product of distances. Hence, a very small number of RIS elements is sufficient to enable the indirect link to have a lower total FSPL than the direct link, and any further increase of the number of RIS elements will render the direct link comparatively useless for indoor communications.

D. Computational Complexity and Run Time Results

In reality, it is not practical to the wait for an optimization algorithm to reach a critical point, but rather some value that is not too far from it. Hence in this subsection, we consider the computational complexity required for the PGM and the AO to reach an achievable rate that is equal to 95 % of the average achievable rate at the 500th iteration. These complexities are heavily influenced by the number of iterations that are needed to achieve this target achievable rate. For the PGM, the computational complexity per iteration is $C_{\mathrm{PGM,IT}}$ given by (34) and the number of iterations needed to achieve the optimal achievable rate is denoted as I_{PGM}. Their product



(a) Direct link present.



(b) Direct link blocked.

Fig. 10. Average achievable rate of the PGM versus the benchmark schemes in an indoor environment.

determines the total computational complexity[9] C_{PGM} of the PGM. For the AO, the computational complexity is given by[10] C_{AO} and the number of outer iterations needed to achieve the optimal achievable rate is I_{OI} (see Appendix D). To maintain compatibility with [24], the number of randomly generated RIS phase shift realizations at the beginning of the AO is taken to be $L_{\mathrm{AO}} = 100$.

The computational complexity of the PGM and of the AO is shown in Table I. The parameter setup is the same as for Figs. 3b and 6b. In general, the PGM is able to achieve a significantly lower computational complexity than the AO, while at the same time it requires a small number of iterations. If the direct link is present, we observe that I_{PGM} becomes smaller with an increase of the number of RIS elements, or in other words, the convergence of the PGM improves. As a result, the computational complexity of the PGM does not increase proportionally to the number of RIS elements. On the other hand, I_{PGM} remains constant when the direct link is blocked and the computational complexity of the PGM increases if the number of RIS elements is made larger. The AO needs one outer iteration to reach the target achievable

[9]We neglect the number of multiplications needed to compute the achievable rate after every iteration, due to the low number of iterations that PGM needs to reach the target achievable rate.

[10]This complexity is derived under the assumption that the achievable rate is computed at the end of each outer iteration, as was proposed in [24]. Therefore, we are not interested in the number of conventional iterations, but in the number of *outer* iterations needed to achieve a certain rate.

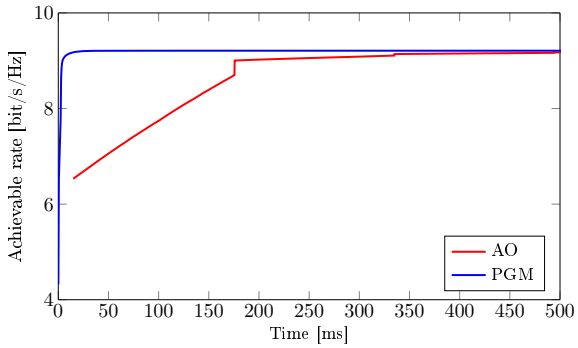


Fig. 11. Average achievable rate of the PGM and the AO versus the run time. The parameter setup is the same as for Fig. 3a.

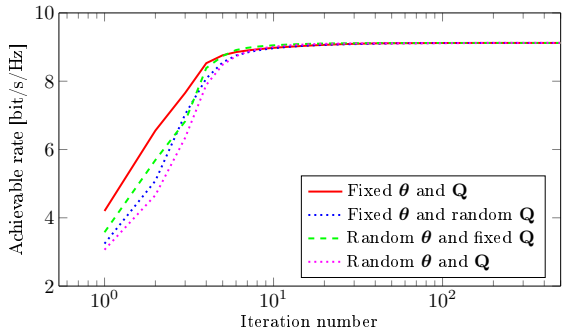


Fig. 12. Average achievable rate results for different initial values of $\boldsymbol{\theta}$ and $\mathbf{Q}$. The parameter setup is the same as for Fig. 3a.

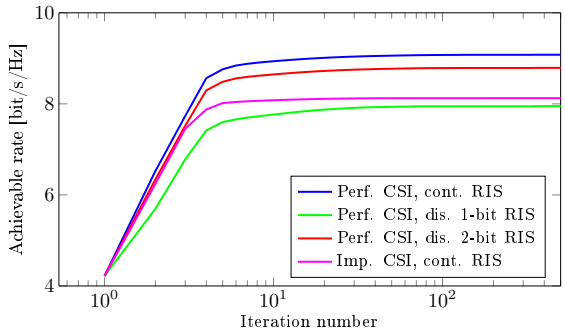


Fig. 13. Average achievable rate for the case of discrete RIS phase shifts and imperfect CSI. The parameter setup is the same as for Fig. 3a.

rate, and its computational complexity increases in proportion to the number of RIS elements.

To make this subsection complete, we also compare in Fig. 11 the achievable rate of the AO and the PGM with respect to the run time of the algorithm's software implementation. The achievable rate for both methods is computed at the end of each iteration. It can be seen that the PGM needs an extremely low run time to converge. Approximately the same time is needed for the AO just to select the optimal initial point. Since the first iteration of the AO is executed after the initial point is chosen, the achievable rate curve for the AO in Fig. 11 starts from about 10 ms. Even after 500 ms the AO is not entirely capable of reaching the same achievable rate.

E. Sensitivity of PGM to Initialization

In this subsection, we study the sensitivity of the PGM to the initial values of $\boldsymbol{\theta}$ and $\mathbf{Q}$. Hence, we consider four cases, where the initial value of $\boldsymbol{\theta}$ is either set to $[1 \ 1 \ \cdots \ 1]^T$ (referred to as "fixed $\boldsymbol{\theta}$") or is randomly generated, and the initial value of $\mathbf{Q}$ is either set to $(P_t/N_t)\mathbf{I}$ (referred to as "fixed $\mathbf{Q}$") or is randomly generated. The achievable rate results for the different initial values of $\boldsymbol{\theta}$ and $\mathbf{Q}$ are shown in Fig. 12. The only visible difference between the considered cases is in the first few iterations, where the PGM with fixed initial $\boldsymbol{\theta}$ and $\mathbf{Q}$ achieves a slightly higher achievable rate than in other cases. In later iterations, the achievable rates in all four cases are approximately equal. Hence, the PGM can always reach the same achievable rate in approximately the same number of iterations, independently of the initial values of $\boldsymbol{\theta}$ and $\mathbf{Q}$.

F. Robustness to System Imperfections

In order to better understand the applicability of the proposed PGM, it is necessary to consider the influence of realistic imperfections in an RIS-aided communication system. Motivated by this, the achievable rate for the case of discrete RIS phase shifts and imperfect CSI is shown in Fig. 13. The achievable rate for the RIS with discrete phase shifts is obtained by discretizing the continuous RIS phase shifts in the final iteration of the PGM and then calculating the achievable rate. The proposed PGM is also directly applicable to discrete phase shifts, since the projection of a given point onto the set of discrete phase shifts is equivalent to finding the minimum distance between the point and all possible phase shifts. It can be seen that utilizing 1-bit and 2-bit discrete RIS phase shifts can reduce the optimal achievable rate by approximately 1.1 bit/s/Hz and 0.2 bit/s/Hz, respectively. Hence, even a very low resolution of discrete RIS phase shifts is sufficient to ensure a limited reduction of the optimal achievable rate.

In the case of imperfect CSI, we assume that the estimated channel matrix can be presented as a sum of the true channel matrix and an estimation error matrix. The estimation error matrix consists of i.i.d. elements that are distributed according to $\mathcal{CN}(0, \sigma^2)$, where $\sigma^2 = 0.2$. Also, it is assumed that the channel matrix FSPLs are not affected by imperfect CSI. From the results, which are plotted in Fig. 13, we can observe that the optimal achievable rate decreases by approximately 1 bit/s/Hz, which is an acceptable level of reduction.

G. Influence of Data Scaling and Line Search

In this subsection, we analyze the influence of data scaling and line search on the PGM. Hence, we compare the proposed PGM with two benchmark schemes. For the first benchmark scheme (i.e., PGM without line search), the PGM is implemented without the line search procedure and we assumed a constant step size equal to 10. For the second benchmark scheme, the PGM is implemented without data scaling. The achievable rate results are shown in Fig. 14. As expected, the proposed PGM has the best achievable rate results among the considered schemes. PGM without line search needs significantly more iterations to reach the optimal

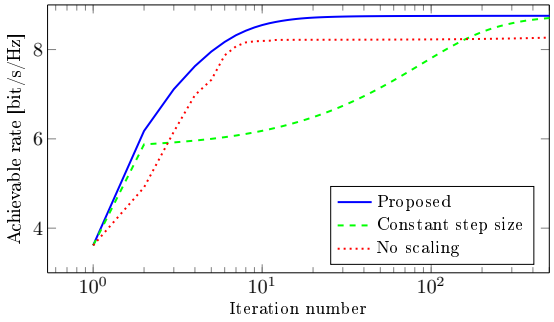


Fig. 14. Average achievable rate of the proposed PGM and the two benchmark schemes (i.e., PGM without line search and PGM without data scaling). The setup of parameters is the same as for Fig. 3b.

achievable rate, for a step size that is multiple times larger than the inverse of the Lipschitz constant (see Theorem 2). Generally, the larger step size enables faster convergence, but the risk of misconvergence is then higher. Furthermore, PGM without data scaling has an achievable rate that is not very significantly worse than the achievable rate of the proposed PGM.

VII. Conclusion

In this paper, we proposed a new PGM algorithm for the achievable rate optimization in multi-stream MIMO system equipped with an RIS. Also, we derived a Lipschitz constant that guarantees the convergence of the PGM. To improve the rate of convergence of the PGM algorithm, we proposed a data scaling step and employed a backtracking line search, which enable the PGM to significantly outperform the existing AO algorithm. In addition, we defined the new metric of total FSPL, and showed that the ratio between the total FSPL of the indirect and direct links can successfully serve as a first-order measure of the applicability of an RIS. Numerical results confirm that the PGM requires a significantly lower number of iterations, and correspondingly a substantially lower computational complexity, than the AO in order to reach a target (near-optimal) achievable rate. Furthermore, we showed that the RIS is particularly convenient for application in an indoor environment, since a small number of RIS elements is sufficient to enable the indirect link to have a higher achievable rate than the direct link.

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

We first note that (17b) is given in [30, Eq. (6.207)] and is relatively well known in the related literature. To derive (17a) we follow the procedure to compute the complex-valued gradient of a general function detailed in [30, Sect. 3.3.1]. Note that in the following, we adopt the notations introduced in [30]: $df(\mathbf{X})$ denotes the complex differential of $f(\mathbf{X})$. To proceed, we recall that the complex differential of $f(\boldsymbol{\theta}, \mathbf{Q})$ with respect to $\mathbf{F}(\boldsymbol{\theta}) = \mathrm{diag}(\boldsymbol{\theta})$ and $\mathbf{F}^*(\boldsymbol{\theta})$ is given by

$$df(\boldsymbol{\theta}, \mathbf{Q}) = \mathrm{Tr}\left\{\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q}) d\left(\mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}\mathbf{Z}^H(\boldsymbol{\theta})\right)\right\} = \mathrm{Tr}\left\{\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q})\left(d(\mathbf{Z}(\boldsymbol{\theta}))\mathbf{Q}\mathbf{Z}^H(\boldsymbol{\theta}) + \mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}d\mathbf{Z}^H(\boldsymbol{\theta})\right)\right\}. \quad (41)$$

After a few algebraic steps we obtain

$$df(\boldsymbol{\theta}, \mathbf{Q}) = \mathrm{vec}^T\left(\left(\bar{\mathbf{H}}_1\mathbf{Q}\mathbf{Z}^H(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q})\mathbf{H}_2\right)^T\right)\mathrm{vec}(d\mathbf{F}(\boldsymbol{\theta})) + \mathrm{vec}^T\left(\left(\bar{\mathbf{H}}_1^*\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{Q}^T\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q})\mathbf{H}_2^*\right)^T\right)\mathrm{vec}\left(d\mathbf{F}^*(\boldsymbol{\theta})\right), \quad (42)$$

where we have used the equality $\mathrm{Tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{vec}^T(\mathbf{A})\,\mathrm{vec}(\mathbf{B})$. 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. $\mathrm{vec}(\mathbf{A}) = \mathbf{L}_d\,\mathrm{vec}_d(\mathbf{A})$ [30, Definition 2.12]. Then we can rewrite $df(\boldsymbol{\theta}, \mathbf{Q})$ as

$$df(\boldsymbol{\theta}, \mathbf{Q}) = \mathrm{vec}^T\left(\left(\bar{\mathbf{H}}_1\mathbf{Q}\mathbf{Z}^H(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q})\mathbf{H}_2\right)^T\right)\mathbf{L}_d\,\mathrm{vec}(d\boldsymbol{\theta}) + \mathrm{vec}^T\left(\left(\bar{\mathbf{H}}_1^*\mathbf{Z}^T(\boldsymbol{\theta})\mathbf{Q}^T\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q})\mathbf{H}_2^*\right)^T\right)\mathbf{L}_d\,\mathrm{vec}(d\boldsymbol{\theta}^*). \quad (43)$$

Using [30, Table 3.2] and [30, Eqn. (2.140)] we obtain

$$\begin{aligned}\nabla_\theta f(\boldsymbol{\theta}, \mathbf{Q}) &= \mathbf{L}_d^T\,\mathrm{vec}\left(\mathbf{H}_2^H\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q})\mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}\bar{\mathbf{H}}_1^H\right)\\ &= \mathrm{vec}_d\left(\mathbf{H}_2^H\mathbf{K}(\boldsymbol{\theta}, \mathbf{Q})\mathbf{Z}(\boldsymbol{\theta})\mathbf{Q}\bar{\mathbf{H}}_1^H\right). \quad (44)\end{aligned}$$

In a similar manner, we can prove the expression for $\nabla_{\mathbf{Q}} f(\boldsymbol{\theta}, \mathbf{Q})$. The details are omitted here due to the page limit.

Appendix B
Proof of Lemma 2

To make the proof easy to follow, we first recall the following inequalities, which are well known or can be proved easily. For the norm of a matrix product it holds that

$$||\mathbf{AB}|| \le \lambda_{\max}(\mathbf{A})||\mathbf{B}|| \quad (45a)$$

$$||\mathbf{ABC}|| \le \lambda_{\max}(\mathbf{A})||\mathbf{B}||\lambda_{\max}(\mathbf{C}) \quad (45b)$$

where $\lambda_{\max}(\mathbf{X})$ denotes the largest singular value of $\lambda_{\max}(\mathbf{X})$. Since $\mathbf{K}(\bar{\boldsymbol{\theta}}, \bar{\mathbf{Q}}) = \left(\mathbf{I} + \mathbf{Z}(\bar{\boldsymbol{\theta}})\bar{\mathbf{Q}}\mathbf{Z}(\bar{\boldsymbol{\theta}})^H\right)^{-1} \preceq \mathbf{I}$, we have

$$\lambda_{\max}\left(\mathbf{K}(\bar{\boldsymbol{\theta}}, \bar{\mathbf{Q}})\right) \le 1. \quad (46)$$

It is easy to check that

$$\lambda_{\max}\left(\mathbf{F}(\bar{\boldsymbol{\theta}})\right) = k^{-1}; \lambda_{\max}\left(\bar{\mathbf{Q}}\right) \le \bar{P}_t. \quad (47)$$

A. Proof of (27)

From (17a) we obtain

$$\begin{aligned}||\nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1) - \nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_2, \bar{\mathbf{Q}}_2)|| &= ||\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{Q}}_1\bar{\mathbf{H}}_1^H - \mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_2, \bar{\mathbf{Q}}_2)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H|| \le ||\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{Q}}_1\bar{\mathbf{H}}_1^H\\ &\quad - \mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H|| + ||\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H\\ &\quad - \mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_2, \bar{\mathbf{Q}}_2)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H||. \quad (48)\end{aligned}$$

The first term on the right-hand side of (48) can be upper-bounded as

$$\begin{aligned}||\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{Q}}_1\bar{\mathbf{H}}_1^H - \mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H||\\ \le a\lambda_{\max}(\bar{\mathbf{H}}_{\mathrm{DIR}})||\bar{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_2||\\ + a\lambda_{\max}(\mathbf{H}_2)||\mathbf{F}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_1 - \mathbf{F}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_2||. \quad (49)\end{aligned}$$

Furthermore, we have

$$||\mathbf{F}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_1-\mathbf{F}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_2|| \leq k^{-1}\lambda_{\max}(\bar{\mathbf{H}}_1)||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2|| + \lambda_{\max}(\bar{\mathbf{H}}_1)\bar{P}_t||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2||. \quad (50)$$

Substituting (50) into (49) gives

$$||\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{Q}}_1\bar{\mathbf{H}}_1^H-\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H|| \leq ab||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2||+a^2\bar{P}_t||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2||. \quad (51)$$

Similarly, the second term on the right-hand side (RHS) of (48) can be upper-bounded as

$$||\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H-\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_2,\bar{\mathbf{Q}}_2)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H|| \leq ab\bar{P}_t||\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)^H-\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{Q}}_1\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H||. \quad (52)$$

Furthermore, we obtain

$$\begin{aligned}&||\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)^H-\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{Q}}_1\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H|| \\ &\leq ||\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2(\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)^H-\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H)|| \\ &+||(\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2-\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H||. \quad (53)\end{aligned}$$

The following inequalities hold for the two norms in the RHS of the above equation:

$$||\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2(\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)^H-\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H)|| = ||(\bar{\mathbf{H}}_{\mathrm{DIR}}+\mathbf{H}_2\mathbf{F}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{H}}_1) \times \bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H(\mathbf{F}(\bar{\boldsymbol{\theta}}_2)-\mathbf{F}(\bar{\boldsymbol{\theta}}_1))^H\mathbf{H}_2^H|| \leq ab\bar{P}_t||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2|| \quad (54)$$

and

$$\begin{aligned}&||(\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2-\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H|| \\ &\leq ||\bar{\mathbf{H}}_{\mathrm{DIR}}(\bar{\mathbf{Q}}_2-\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H|| \\ &+||\left[\mathbf{H}_2\mathbf{F}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_2-\mathbf{H}_2\mathbf{F}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_1\right]\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H||. \quad (55)\end{aligned}$$

To upper-bound the two terms on the RHS of (55), we use

$$\left\|\bar{\mathbf{H}}_{\mathrm{DIR}}(\bar{\mathbf{Q}}_2-\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H\right\| \leq b\lambda_{\max}(\bar{\mathbf{H}}_{\mathrm{DIR}})\left\|\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2\right\| \quad (56)$$

and

$$\begin{aligned}&||\left[\mathbf{H}_2\mathbf{F}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_2-\mathbf{H}_2\mathbf{F}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_1\right]\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H|| \\ &\leq \lambda_{\max}(\mathbf{H}_2)||\mathbf{F}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_2-\mathbf{F}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{H}}_1\bar{\mathbf{Q}}_1|| \\ &\times\left[\lambda_{\max}(\bar{\mathbf{H}}_{\mathrm{DIR}}^H)+\lambda_{\max}(\bar{\mathbf{H}}_1^H)\lambda_{\max}(\mathbf{H}_2^H)\right] \\ &\leq k^{-1}ab||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2||+ab\bar{P}_t||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2||. \quad (57)\end{aligned}$$

Substituting (54), (55), (56) and (57) into (53), we obtain

$$\begin{aligned}&||\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)^H-\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)\bar{\mathbf{Q}}_1\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H|| \\ &\leq b^2||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2||+2ab\bar{P}_t||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2|| \quad (58)\end{aligned}$$

and (52) then implies

$$||\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H-\mathbf{H}_2^H\mathbf{K}(\bar{\boldsymbol{\theta}}_2,\bar{\mathbf{Q}}_2)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)\bar{\mathbf{Q}}_2\bar{\mathbf{H}}_1^H|| \leq ab^3\bar{P}_t||\bar{\mathbf{Q}}_1-\bar{\mathbf{Q}}_2||+2a^2b^2\bar{P}_t^2||\bar{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_2||. \quad (59)$$

Substituting (51) and (59) into (48), we obtain (27).

B. *Proof of* (28)

From (17b) immediately have

$$\begin{aligned}&||\nabla_{\bar{\mathbf{Q}}}f(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1)-\nabla_{\bar{\mathbf{Q}}}f(\bar{\boldsymbol{\theta}}_2,\bar{\mathbf{Q}}_2)|| \\ &=||\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)-\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)^H\mathbf{K}(\bar{\boldsymbol{\theta}}_2,\bar{\mathbf{Q}}_2)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)|| \\ &\leq||\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)-\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)|| \\ &+||\mathbf{Z}(\bar{\boldsymbol{\theta}}_1)^H\mathbf{K}(\bar{\boldsymbol{\theta}}_1,\bar{\mathbf{Q}}_1)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)-\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)^H\mathbf{K}(\bar{\boldsymbol{\theta}}_2,\bar{\mathbf{Q}}_2)\mathbf{Z}(\bar{\boldsymbol{\theta}}_2)||. \quad (60)\end{aligned}$$

Following the same steps used to prove (27) we can further upper bound the two norms in the RHS of the above equation to prove (28). The details are omitted here due to the page limit.

Appendix C
Proof of Theorem 2

We recall the following inequality for any function $f(x)$ which is L-smooth:

$$f(\mathbf{y}) \geq f(\mathbf{x})+\left\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\right\rangle-\frac{L}{2}||\mathbf{y}-\mathbf{x}||^2. \quad (61)$$

The projection of $\bar{\boldsymbol{\theta}}_{n+1}$ onto $\bar{\Theta}$ can be written as

$$\begin{aligned}\bar{\boldsymbol{\theta}}_{n+1}&=\arg\min_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}}\left\|\bar{\boldsymbol{\theta}}-\bar{\boldsymbol{\theta}}_n-\mu\nabla_{\bar{\boldsymbol{\theta}}}f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)\right\|^2 \\ &=\arg\max_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}}\left\langle\nabla_{\bar{\boldsymbol{\theta}}}f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n),\bar{\boldsymbol{\theta}}-\bar{\boldsymbol{\theta}}_n\right\rangle-\frac{1}{2\mu}||\bar{\boldsymbol{\theta}}-\bar{\boldsymbol{\theta}}_n||^2 \quad (62)\end{aligned}$$

where $\left\langle\mathbf{x},\mathbf{y}\right\rangle=\Re(\mathbf{x}^H\mathbf{y})$ and we have used the fact that $||\mathbf{a}-\mathbf{b}||^2=||\mathbf{a}||^2+||\mathbf{b}||^2-2\Re(\mathbf{a}^H\mathbf{b})$. Note that when $\bar{\boldsymbol{\theta}}=\bar{\boldsymbol{\theta}}_n$, the objective in the above problem is equal to 0, and thus we have

$$\left\langle\nabla_{\bar{\boldsymbol{\theta}}}f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n),\bar{\boldsymbol{\theta}}_{n+1}-\bar{\boldsymbol{\theta}}_n\right\rangle-\frac{1}{2\mu}||\bar{\boldsymbol{\theta}}_{n+1}-\bar{\boldsymbol{\theta}}_n||^2\geq 0. \quad (63)$$

An analogous inequality also holds for $\bar{\mathbf{Q}}_{n+1}$, i.e.,

$$\left\langle\nabla_{\bar{\mathbf{Q}}}f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n),\bar{\mathbf{Q}}_{n+1}-\bar{\mathbf{Q}}_n\right\rangle-\frac{1}{2\mu}||\bar{\mathbf{Q}}_{n+1}-\bar{\mathbf{Q}}_n||^2\geq 0. \quad (64)$$

Applying (61) yields

$$\begin{aligned}f(\bar{\boldsymbol{\theta}}_{n+1},\bar{\mathbf{Q}}_{n+1})&\geq f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)+\left\langle\nabla_{\bar{\boldsymbol{\theta}}}f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n),\bar{\boldsymbol{\theta}}_{n+1}-\bar{\boldsymbol{\theta}}_n\right\rangle \\ &+\left\langle\nabla_{\bar{\mathbf{Q}}}f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n),\bar{\mathbf{Q}}_{n+1}-\bar{\mathbf{Q}}_n\right\rangle \\ &-\frac{L}{2}\left\|\bar{\boldsymbol{\theta}}_{n+1}-\bar{\boldsymbol{\theta}}_n\right\|^2-\frac{L}{2}\left\|\bar{\mathbf{Q}}_{n+1}-\bar{\mathbf{Q}}_n\right\|^2 \\ &\geq f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)+\left(\frac{1}{2\mu}-\frac{L}{2}\right)\left(\left\|\bar{\boldsymbol{\theta}}_{n+1}-\bar{\boldsymbol{\theta}}_n\right\|^2\right. \\ &\left.+\left\|\bar{\mathbf{Q}}_{n+1}-\bar{\mathbf{Q}}_n\right\|^2\right). \quad (65)\end{aligned}$$

It is easy to see that $f(\bar{\boldsymbol{\theta}}_{n+1},\bar{\mathbf{Q}}_{n+1})\geq f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)$ if $\mu<\frac{1}{L}$. Since the feasible set of the considered problem is closed and bounded, the iterate $(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)$ is bounded and thus $(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)$ has accumulation points. Since, as shown above, $f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)$ is nondecreasing, f has the same value, denoted by f^*, at all of these accumulation points. From (65) we have

$$\begin{aligned}f(\bar{\boldsymbol{\theta}}_{n+1},\bar{\mathbf{Q}}_{n+1})-f(\bar{\boldsymbol{\theta}}_n,\bar{\mathbf{Q}}_n)&\geq\left(\frac{1}{2\mu}-\frac{L}{2}\right)\left(\left\|\bar{\boldsymbol{\theta}}_{n+1}-\bar{\boldsymbol{\theta}}_n\right\|^2\right. \\ &\left.+\left\|\bar{\mathbf{Q}}_{n+1}-\bar{\mathbf{Q}}_n\right\|^2\right), \quad (66)\end{aligned}$$

which results in

$$\infty > f^* - f(\bar{\boldsymbol{\theta}}_1, \bar{\mathbf{Q}}_1) \geq \sum_{n=1}^{\infty} \big(\frac{1}{2\mu} - \frac{L}{2}\big)\big(\big\|\bar{\boldsymbol{\theta}}_{n+1} - \bar{\boldsymbol{\theta}}_n\big\|^2 + \big\|\bar{\mathbf{Q}}_{n+1} - \bar{\mathbf{Q}}_n\big\|^2\big). \quad (67)$$

Since $\mu < \frac{1}{L}$ we can conclude that

$$\big\|\bar{\boldsymbol{\theta}}_{n+1} - \bar{\boldsymbol{\theta}}_n\big\| \to 0; \big\|\bar{\mathbf{Q}}_{n+1} - \bar{\mathbf{Q}}_n\big\| \to 0. \quad (68)$$

The optimality condition of (62) implies

$$\big\langle \frac{1}{\mu}(\bar{\boldsymbol{\theta}}_{n+1} - \bar{\boldsymbol{\theta}}_n) - \nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_n, \bar{\mathbf{Q}}_n), \bar{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}_{n+1}\big\rangle \leq 0, \ \forall \bar{\boldsymbol{\theta}} \in \bar{\Theta}. \quad (69)$$

Similarly we have

$$\big\langle \frac{1}{\mu}(\bar{\mathbf{Q}}_{n+1} - \bar{\mathbf{Q}}_n) - \nabla_{\bar{\mathbf{Q}}} f(\bar{\boldsymbol{\theta}}_n, \bar{\mathbf{Q}}_n), \bar{\mathbf{Q}} - \bar{\mathbf{Q}}_{n+1}\big\rangle \leq 0, \ \forall \bar{\mathbf{Q}} \in \bar{\mathcal{Q}}. \quad (70)$$

Let $(\boldsymbol{\theta}^*, \mathbf{Q}^*)$ be any accumulation point of $(\bar{\boldsymbol{\theta}}_n, \bar{\mathbf{Q}}_n)$, say $(\bar{\boldsymbol{\theta}}_n, \bar{\mathbf{Q}}_n) \to (\boldsymbol{\theta}^*, \mathbf{Q}^*)$ as $n \to \infty$. We also note that the gradient of $f(\bar{\boldsymbol{\theta}}_n, \bar{\mathbf{Q}}_n)$ is continuous and thus $\nabla_{\bar{\boldsymbol{\theta}}} f(\bar{\boldsymbol{\theta}}_n, \bar{\mathbf{Q}}_n) \to \nabla_{\bar{\boldsymbol{\theta}}} f(\boldsymbol{\theta}^*, \mathbf{Q}^*)$ and $\nabla_{\bar{\mathbf{Q}}} f(\bar{\boldsymbol{\theta}}_n, \bar{\mathbf{Q}}_n) \to \nabla_{\bar{\mathbf{Q}}} f(\boldsymbol{\theta}^*, \mathbf{Q}^*)$. By letting $n \to \infty$ in (69) and (70), we have

$$\big\langle -\nabla_{\bar{\boldsymbol{\theta}}} f(\boldsymbol{\theta}^*, \mathbf{Q}^*), \bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\big\rangle \leq 0, \ \forall \bar{\boldsymbol{\theta}} \in \bar{\Theta} \quad (71)$$

$$\big\langle -\nabla_{\bar{\mathbf{Q}}} f(\boldsymbol{\theta}^*, \mathbf{Q}^*), \bar{\mathbf{Q}} - \mathbf{Q}^*\big\rangle \leq 0, \ \forall \bar{\mathbf{Q}} \in \bar{\mathcal{Q}}, \quad (72)$$

which means that $(\boldsymbol{\theta}^*, \mathbf{Q}^*)$ is indeed a critical point of (11). This completes the proof.

Appendix D
Computational Complexity for Alternating Optimization (AO)

The computational complexity for the AO method, introduced in [24], is derived in this appendix. To make the following derivation more accessible, the mathematical notation in this appendix is the same as in [24].

The channel matrix from the transmitter to the receiver is given by $\tilde{\mathbf{H}} = \mathbf{H} + \mathbf{R}\boldsymbol{\phi}\mathbf{T}$, where $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ presents the *direct* signal transmission between the transmitter and the receiver, $\mathbf{T} \in \mathbb{C}^{N_{\text{ris}} \times N_t}$ presents the signal transmission between the transmitter and the RIS, $\mathbf{R} \in \mathbb{C}^{N_r \times N_{\text{ris}}}$ presents the signal transmission between the RIS and the receiver, and $\boldsymbol{\phi}$ models the RIS response. Let $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_{N_{\text{ris}}}]$, $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_{N_{\text{ris}}}]^H$ and $\boldsymbol{\phi} = \text{diag}[\alpha_1, \ldots, \alpha_{N_{\text{ris}}}]$, so that the channel matrix can be written as $\tilde{\mathbf{H}} = \mathbf{H} + \sum_{i=1}^{N_{\text{ris}}} \alpha_i \mathbf{r}_i \mathbf{t}_i^H$.

In the first step of the AO algorithm, L_{AO} independent realizations of $\{\alpha_m\}_{m=1}^{N_{\text{ris}}}$ are randomly generated and for each of these the optimal covariance matrix $\mathbf{Q}$ is computed. To do this, the channel matrix $\tilde{\mathbf{H}}$ has to be calculated for every $\{\alpha_m\}_{m=1}^{N_{\text{ris}}}$ realization. This calculation starts by computing all $\mathbf{r}_m \mathbf{t}_m^H$ matrices and for this $N_r N_t N_{\text{ris}}$ multiplications are needed. Further, the computation of all $\alpha_m \mathbf{r}_m \mathbf{t}_m^H$ matrices requires $N_r N_t N_{\text{ris}}$ multiplications (per one $\{\alpha_m\}_{m=1}^{N_{\text{ris}}}$ realization). Hence, the complexity of calculating L_{AO} channel matrices $\tilde{\mathbf{H}}$ is $(L_{\text{AO}} + 1) N_r N_t N_{\text{ris}}$.

For each $\tilde{\mathbf{H}}$ it is required to perform the the truncated singular value decomposition $\tilde{\mathbf{H}} = \tilde{\mathbf{U}} \tilde{\boldsymbol{\Lambda}} \tilde{\mathbf{V}}^H$, where $\tilde{\mathbf{V}} \in \mathbb{C}^{N_t \times D}$ and $D = \min(N_t, N_r)$. The complexity of this decomposition is approximately $\mathcal{O}(D^3)$. Next, $\mathbf{Q}$ is computed as $\mathbf{Q} = \tilde{\mathbf{V}} \, \text{diag}\{p_1, \ldots, p_D\} \tilde{\mathbf{V}}^H$, where $\{p_1, \ldots, p_D\}$ are obtained using a water-filling algorithm and $D = \min(N_t, N_r)$. The complexity of the water-filling algorithm is $\mathcal{O}(D^2)$ and the complexity of the matrix multiplication is $N_t D + (N_t^2 + N_t) D/2$. Therefore, the calculation of L_{AO} covariance matrices $\mathbf{Q}$ requires $\mathcal{O}(L_{\text{AO}}(D^3 + \frac{1}{2} N_t^2 D))$ multiplications.

In the sequel, the optimal $\{\alpha_m\}_{m=1}^{N_{\text{ris}}}$ and $\mathbf{Q}$ are iteratively determined. In one conventional iteration, one α_m or $\mathbf{Q}$ is adjusted. A set of $N_{\text{ris}} + 1$ successive conventional iterations constitutes one "outer" iteration, in which *all* α_m and $\mathbf{Q}$ are adjusted. The AO method stops when the convergence criterion at the end of an outer iteration is fulfilled.

At the beginning of each outer iteration, the eigenvalue decomposition $\mathbf{Q} = \mathbf{U}_Q \boldsymbol{\Sigma}_Q \mathbf{U}_Q^H$ is performed, which requires $\mathcal{O}(N_t^3)$ multiplications. The calculation of the matrices $\mathbf{H}' = \mathbf{H}\mathbf{U}_Q \boldsymbol{\Sigma}_Q^{\frac{1}{2}} \in \mathbb{C}^{N_r \times N_t}$ and $\mathbf{T}' = \mathbf{T}\mathbf{U}_Q \boldsymbol{\Sigma}_Q^{\frac{1}{2}} \in \mathbb{C}^{N_{\text{ris}} \times N_t}$ has the complexity $\mathcal{O}(N_t^3 + N_t^2 N_{\text{ris}})$.

To form $\mathbf{S} = \mathbf{H}' + \sum_{i=1}^{N_{\text{ris}}} \alpha_i \mathbf{r}_i \mathbf{t}_i^{'H}$, $N_r N_t N_{\text{ris}}$ multiplications are needed first to obtain all $\mathbf{r}_i \mathbf{t}_i^{'H}$ and $N_r N_t N_{\text{ris}}$ multiplications are needed for all $\alpha_i \mathbf{r}_i \mathbf{t}_i^{'H}$. Hence, the complexity of computing $\mathbf{S}$ is $2 N_r N_t N_{\text{ris}}$.

The optimization of the m-th RIS element requires the computation of the following auxiliary matrices

$$\mathbf{A}_m = \mathbf{I} + \frac{1}{N_0} \mathbf{S}_m \mathbf{S}_m^H + \frac{1}{N_0} \mathbf{r}_m \mathbf{t}_m^{'H} \left(\mathbf{r}_m \mathbf{t}_m^{'H}\right)^H \quad (73)$$

$$\mathbf{B}_m = \frac{1}{N_0} \mathbf{r}_m \mathbf{t}_m^{'H} \mathbf{S}_m^H \quad (74)$$

where $\mathbf{S}_m = \mathbf{H}' + \sum_{i=1, i \neq m}^{N_{\text{ris}}} \alpha_i \mathbf{r}_i \mathbf{t}_i^{'H} = \mathbf{S} - \alpha_m \mathbf{r}_m \mathbf{t}_m^{'H}$. It can be easily shown that the complexities for calculating $\mathbf{A}_m$ and $\mathbf{B}_m$ from the previous expressions are both $\mathcal{O}(N_r^2 N_t)$.

Utilizing the results from Subsection IV-B, the complexity of computing $\mathbf{A}_m^{-1} \mathbf{B}_m$ is $\mathcal{O}(2 N_r^3)$. The subsequent calculation of α_m and update of $\mathbf{S} = \mathbf{S}_m + \alpha_m \mathbf{r}_m \mathbf{t}_m^{'H}$ require a negligible complexity.

After adjusting all RIS elements, the optimization of $\mathbf{Q}$ is performed according to the aforementioned procedure, which requires $\mathcal{O}(D^3 + \frac{1}{2} N_t^2 D)$ multiplications.

If I_{OI} is the number of outer iterations, then the computational complexity of the AO algorithm is given by

$$\begin{aligned} C_{\text{AO}} = \mathcal{O}((L_{\text{AO}} + 1) N_r N_t N_{\text{ris}} + L_{\text{AO}}(D^3 + \frac{1}{2} N_t^2 D) \\ + I_{\text{OI}}[N_t^3 + N_t^2 N_{\text{ris}} + 2 N_r N_t N_{\text{ris}} \\ + (2 N_r^2 N_t + 2 N_r^3) N_{\text{ris}} + D^3 + \frac{1}{2} N_t^2 D]) \end{aligned} \quad (75)$$

where $D = \min(N_t, N_r)$.

References

[1] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," *EURASIP J. Wireless Commun. and Netw.*, vol. 2019, no. 1, pp. 1–20, 2019.

[2] ——, "Reconfigurable intelligent surfaces vs. relaying: Differences, similarities, and performance comparison," *IEEE Open Jour. of the Commun. Society*, vol. 1, pp. 798–807, Jun. 2020.

[3] E. Basar *et al.*, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, 2019.

[4] C. Huang *et al.*, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wirel. Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[5] A. Taha *et al.*, "Enabling large intelligent surfaces with compressive sensing and deep learning," *arXiv preprint arXiv:1904.10136*, 2019.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] X. Yu *et al.*, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2637–2652, Nov. 2020.

[11] ——, "MISO wireless communication systems via intelligent reflecting surfaces," in *Proc. IEEE/CIC Int. Conf. on Commun. in China (ICCC)*. IEEE, Aug 2019.

[12] 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.

[13] P. Wang *et al.*, "Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design," *IEEE Trans. Veh. Technol.*, 2020, Early Access.

[14] C. Huang *et al.*, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[15] ——, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[16] M.-M. Zhao *et al.*, "Exploiting amplitude control in intelligent reflecting surface aided wireless communication with imperfect CSI," *arXiv preprint arXiv:2005.07002*, 2020.

[17] S. Abeywickrama *et al.*, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, Sep. 2020.

[18] B. Di *et al.*, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1809–1822, Aug. 2020.

[19] T. Hou *et al.*, "Reconfigurable intelligent surface aided NOMA networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2575–2588, Nov. 2020.

[20] Y. Yang *et al.*, "IRS-enhanced OFDMA: Joint resource allocation and passive beamforming optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 760–764, Jun. 2020.

[21] J. Xiong *et al.*, "Reconfigurable intelligent surfaces assisted MIMO-MAC with partial CSI," in *Proc. IEEE Int. Conf. on Communications (ICC)*, 2020, pp. 1–6.

[22] Ö. Özdogan *et al.*, "Using intelligent reflecting surfaces for rank improvement in MIMO communications," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020, pp. 9160–9164.

[23] 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.

[24] S. Zhang and R. Zhang, "Capacity characterization for intelligent reflecting surface aided MIMO communication," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1823–1838, Aug. 2020.

[25] T. S. Rappaport *et al.*, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks–With a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.

[26] S. W. Ellingson, "Path loss in reconfigurable intelligent surface-enabled channels," *arXiv preprint arXiv:1912.06759*, 2019.

[27] S. J. Orfanidis, *Electromagnetic waves and antennas*. Rutgers University New Brunswick, NJ, 2002.

[28] R. Karasik *et al.*, "Beyond max-SNR: Joint encoding for reconfigurable intelligent surfaces," in *Proc. International Symposium on Information Theory (ISIT)*, 2020, pp. 2965–2970.

[29] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Advances in neural information processing systems*, 2015, pp. 379–387.

[30] A. Hjörungnes, *Complex-calued matrix derivatives with applications in signal processing and communications*. Cambridge University Press, 2011.

[31] T. M. Pham *et al.*, "Revisiting the MIMO capacity with per-antenna power constraint: Fixed-point iteration and alternating optimization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 388–401, Jan. 2019.

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

[33] A. Zappone *et al.*, "Overhead-aware design of reconfigurable intelligent surfaces in smart radio environments," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 126–141, Jan. 2021.

[34] L. Armijo, "Minimization of functions having Lipschitz continuous first partial derivatives," *Pac. J. Math.*, vol. 16, no. 1, pp. 1–3, 1966.

[35] X. Qian *et al.*, "Beamforming through reconfigurable intelligent surfaces in single-user MIMO systems: SNR distribution and scaling laws in the presence of channel fading and phase noise," *IEEE Wireless Commun. Lett.*, vol. 10, no. 1, pp. 77–81, Jan. 2021.