On the Optimality of Reconfigurable Intelligent Surfaces (RISs): Passive Beamforming, Modulation, and Resource Allocation

Minchae Jung, Member, IEEE, Walid Saad, Fellow, IEEE, Mrouane Debbah, Fellow, IEEE, and Choong Seon Hong, Senior Member, IEEE

Abstract

Reconfigurable intelligent surfaces (RISs) have recently emerged as a promising technology that can achieve high spectrum and energy efficiency for future wireless networks by integrating a massive number of low-cost and passive reflecting elements. An RIS can manipulate the properties of an incident wave, such as the frequency, amplitude, and phase, and, then, reflect this manipulated wave to a desired destination, without the need for complex signal processing. In this paper, the asymptotic optimality of achievable rate in a downlink RIS system is analyzed under a practical RIS environment with its associated limitations. In particular, a passive beamformer that can achieve the asymptotic optimal performance by controlling the incident wave properties is designed, under a limited RIS control link and practical reflection coefficients. In order to increase the achievable system sum-rate, a modulation scheme that can be used in an RIS without interfering with existing users is proposed and its average symbol error ratio is asymptotically derived. Moreover, a new resource allocation algorithm that jointly considers user scheduling and power control is designed, under consideration of the proposed passive beamforming and modulation schemes. Simulation results show that the proposed schemes are in close agreement with their upper bounds in presence of a large number of RIS reflecting elements thereby verifying that the achievable rate in practical RISs satisfies the asymptotic optimality.

1In fact, an optimal performance can be achieved by an ideal RIS which has infinite and continuous phase shifts and lossless reflection coefficients and is higher than that of a massive multiple-input and multiple-output system as proved in [1].
Index Terms

large intelligent surface (LIS), metasurface, passive beamforming, resource allocation, reconfigurable intelligent surface (RIS).

I. INTRODUCTION

The concept of a metasurface is rapidly emerging as a key solution to support the demand for massive connectivity, mainly driven by upcoming Internet of Things (IoT) and 6G applications [1]–[16]. A metasurface relies on a massive integration of artificial meta-atoms that are commonly made of metal structures of low-cost and passive elements. Each meta-atom can manipulate the incident electromagnetic (EM) wave impinging on it, in terms of frequency, amplitude, and phase, and reflect it to a desired destination, without additional signal processing. A metasurface can potentially provide reliable and pervasive wireless connectivity given that man-made structures, such as buildings, walls, and roads, can be equipped with metasurfaces in the near future and used for wireless transmission [4]–[6]. Moreover, a tunable metasurface can significantly enhance the signal quality at a receiver by allowing a dynamic manipulation of the incident EM wave. Tunable metasurfaces are mainly controlled by electrical, optical, mechanical, and fluid operations [7] that can be programmed in software using a field programmable gate array (FPGA) [8]. The concept of a reconfigurable intelligent surface (RIS) is essentially an electronically operated metasurface controlled by programmable software, as introduced in [7] and [8]. In wireless communication systems, a base station (BS) can send control signals to an RIS controller (i.e., FPGA) via a dedicated control link and controls the properties of the incident wave to enhance the signal quality at the receiver. In principle, the electrical size of the unit reflecting elements (i.e., meta-atoms) deployed on RIS is between $\lambda/8$ and $\lambda/4$, where $\lambda$ is a wavelength of radio frequency (RF) signal [7]. Note that conventional large antenna-array systems, such as a massive multiple-input and multiple-output (MIMO) and MIMO relay system, typically require antenna spacing of greater than $\lambda/2$ [5]. Therefore, an RIS can provide more reliable and space-intensive communications compared to conventional antenna-array systems as clearly explained in [4]–[6]. Moreover, a large number of reflecting elements can be arranged on each RIS thus offering precise control of the reflection wave and allowing it to coherently align with the desired channel.

A. Related Works

Owing to these advantages, the use of an RIS in wireless communication systems has recently received significant attention as in [1] and [8]–[14]. An RIS is typically used for two main
wireless communication purposes: a) RIS as an RF chain-free transmitter and b) RIS as a passive beamformer that amplifies the incident waveform (received from a BS) and reflects it to the desired user. In [8], the authors analyzed the error rate performance of a phase-shift keying (PSK) signaling and proved that an RIS transmitter equipped with a large number of reflecting elements can convey information with high reliability. The works in [9] and [10] proposed RF chain-free transmitter architectures enabled by an RIS that can support PSK and quadrature amplitude modulation (QAM). Meanwhile, the works in [1] and [11] designed joint active and passive beamformer that minimize the transmit power at the BS, under discrete and continuous phase shifts, respectively. Also, in [12] and [13], the authors designed a passive beamformer that maximizes the ergodic data rate and the energy efficiency, respectively. Moreover, the work in [8] theoretically analyzed the average symbol error rate (SER) resulting from an ideal passive beamformer and proved that the SER decays exponentially as the number of reflecting elements on RIS increases. In [14], we provided a first insight on a passive beamformer that can achieve, asymptotically, an ideal RIS performance. However, these previous studies in [1] and [8]–[14] have not considered practical RIS environments and their limitations, such as practical reflection coefficients and the limited capacity of the RIS control link. In fact, an RIS can manipulate the properties of an incident wave based on the resonant frequency of the tunable reflecting circuit. Then, the incident EM power is partially consumed at the resistance of the reflecting circuit according to the difference between the incident wave frequency and the resonant frequency. This results in the amplitude of the reflection coefficients less than or equal to one depending on the phase shifts of the incident wave. However, the works in [1], [8], [11] and [12] assumed an ideal RIS whose amplitude of the reflection coefficients are always equal to one which is impractical for an RIS. Moreover, in [8] and [11]–[14], the authors assumed a continuous phase shift at each reflecting element. However, this continuous phase shift requires infinite bits to control each reflecting element and the RIS control link between a BS and an RIS cannot support those infinite control bits. Finally, the signals from the RIS transmitters proposed in [9] and [10] can be undesired interference for existing cellular network, given that those RISs operate as an underlay coexistence with cellular networks. Therefore, there is a need for new analysis of practical RISs when dealing with a limited RIS control link capacity and practical reflection coefficients that can verify the asymptotic optimality of realistic RISs.
B. Contributions

The main contribution of this paper is a rigorous optimality analysis of the data rates that can be achieved by an RIS under consideration of practical reflection coefficients with a limited RIS control link capacity. In this regard, we first design a passive beamformer that achieves asymptotic signal-to-noise ratio (SNR) optimality, regardless of the reflection power loss and the number of RIS control bits. In particular, the proposed passive beamformer with one bit RIS control can achieve the asymptotic SNR of an ideal RIS with infinite control bits, lead to a much simpler operation at the BS especially for a large number of reflecting elements on an RIS. We then propose a new modulation scheme that can be used in an RIS to achieve sum-rate higher than the one achieved by a conventional network without RIS. In the proposed modulation scheme, each RIS utilizes an ambient RF signal, convert it into desired signal by controlling the properties of incident wave, and transmit it to the desired user, without interfering with existing users. We also prove that the achievable SNR from the proposed modulation converges to the asymptotic SNR resulting from a conventional massive MIMO or MIMO relay system, as the number of reflecting elements on an RIS increases. Given the aforementioned passive beamformer and modulation scheme, we finally develop a novel resource allocation algorithm whose goal is to maximize the average sum-rate under the minimum rate requirements at each user. We then study, analytically, the potential of an RIS by showing that a practical RIS can achieve the asymptotic performance of an ideal RIS, as the number of RIS reflecting elements increases without bound. Our simulations show that the proposed schemes can asymptotically achieve the performance resulting from an ideal RIS and its upper bound.

The rest of this paper is organized as follows. Section II describes the fundamentals of RIS and the system model. Section III describes the optimality of achievable rate in downlink RIS system. Simulation results are provided in Section IV to support and verify the analyses, and Section V concludes the paper.

Notations: Throughout this paper, boldface upper- and lower-case symbols represent matrices and vectors respectively, and $I_M$ is a size-$M$ identity matrix. The conjugate, transpose, and Hermitian transpose operators are $(\cdot)^*$, $(\cdot)^T$, and $(\cdot)^H$, respectively. The norm of a vector $a$ is $\|a\|$, the amplitude and phase of a complex number $a$ are denoted by $|a|$ and $\angle a$, respectively. $E[\cdot]$ and $\text{Var}[\cdot]$ denote expectation and variance operators, respectively. $\mathcal{O}(\cdot)$ denotes the big O notation and $\mathcal{CN}(m, \sigma^2)$ denotes a complex Gaussian distribution with mean $m$ and variance $\sigma^2$. 
II. SYSTEM MODEL

In this section, we first describe the fundamental operating principle of RISs and we also discuss practical RIS properties in terms of reflection phase, reflection amplitude, and their relationship. These fundamentals are then used to develop our system model.

A. Fundamentals of Reconfigurable Intelligent Surface

We consider an RIS that consists of multiple two-dimensional layers and operates based on the varactor-tuned resonator as shown in Fig. 1. A metasurface is located on the top of a dielectric substrate and a large number of metallic patches are periodically printed on it. In a reflecting element (i.e., a meta-atom), two separate metallic patches are connected to each other through a common varactor diode and each of which is linked to a conductive cylindrical post. This results in a parallel connection between the varactor and the cylindrical post. Those cylindrical posts are connected to a metal layer through a dielectric substrate and a control layer, and the metal layer performs the role of a ground plane. Therefore, the length of the cylindrical post and the thickness of the substrate determine the inductance of the circuit. In the control layer, a bias direct-current (DC) voltage is controlled by an FPGA-based controller and the capacitance of each varactor changes according to separately controlled bias voltages. As discussed in [1] and [8], the FPGA-based controller takes the role of a gateway which can receive and decode signals from other networks (e.g., cellular network) and control the bias voltages based on this signal. Electromagnetic leakage from the metal layer is eliminated by an insulation layer. When an RF signal arrives on the metallic patch, the current flows through the metallic patch and is divided into two different directions of the cylindrical post and the varactor, according to their
impedance. Then, the separated current converges on the other side of the metallic patch and reflects RF signal through this metallic patch. The equivalent circuit of this RIS is shown on the right side of Fig. 1. In this equivalent circuit, $L_0$ represents the inductance resulting from two conductive cylindrical posts, and $L_v$, $R_v$, and $C_n$ are passive elements belonging to the varactor which can be determined according to the varactor model. Hereinafter, we use the SMV1231-079 varactor which has $L_v = 0.7 \, \text{nH}$, $R_v = 2.5 \, \Omega$, and $0.466 \, \text{pF} \leq C_n \leq 2.35 \, \text{pF}$, and assume that $L_0 = 2.5 \, \text{nH}$ as in [16] and [17]. By controlling $C_n$ from 0.466 pF to 2.35 pF, the impedance seen by the incident wave will change which yields a phase shift of the incident wave for entire range of $[-\pi, \pi]$ at each reflecting element.

In conventional RF antennas, the antenna acts as an impedance transformer (i.e. impedance matching network) between the feedline and free space as shown in the left side of Fig. 2. This matching network enables the feedline impedance of RF antenna (i.e., $Z_L = 50 \, \Omega$) to match, perfectly, the impedance of free space (i.e., $Z_0 = 377 \, \Omega$) at RF carrier frequency $f_c$. Given a perfect matching network, a resonant frequency, $f_r$, is generated through the parallel connection between $L_m$ and $C_m$ yielding $f_r = f_c = \frac{1}{2\pi\sqrt{L_mC_m}}$. Then, only the incident wave at frequency $f_c$ flows into the receiver and the power of this incident wave is consumed, thoroughly, at the receiver $Z_L$. Therefore, a maximum data rate can be achieved and there will be no reflection wave at $f_c$, as shown in the left side of Fig. 2. On the other hand, a reflecting RIS does not need to match the impedance at $f_c$. Instead, it needs to reflect the entire power of the incident wave to achieve the maximum data rate. Hence, a reflecting RIS does not require an RF antenna or impedance matching network unlike conventional RF systems. As shown in the right side of Fig. 2, the considered reflecting element $n$ generates the resonance resulting from the parallel connection between $L_0$ and the varactor $n$, as follows: $f_r = \frac{1}{2\pi\sqrt{(L_v+L_0)C_n}}$, where $f_r$ changes.
according to $C_n$ which is controlled by the RIS controller. Since the impedance between the reflecting element and free space will not be matched, the incident wave at $f_i$ flows through the varactor and its power is consumed, partially, at $R_v$. Here, the incident wave at $f_c$ can be fully reflected when $f_c \neq f_i$. However, as $f_i$ gradually approaches to $f_c$, the phase shift decreases from $\pi$ to zero or increases from $-\pi$ to zero \[15\]. Hence, in order to cover the phase shifts of the incident wave for entire range of $[-\pi, \pi]$, power loss is inevitable and it is necessary to analyze the optimal phase shifts, under this practical limitation of an RIS. In \[17\], the relation between a reflection amplitude and its phase is approximated under this partial power loss, as follows:

$$|\Gamma_n| \approx A(\angle \Gamma_n) = 0.8 \left( \frac{\sin (\angle \Gamma_n - 0.43\pi) + 1}{2} \right)^{1.6} + |\Gamma|_{\min}, \quad (1)$$

where $\Gamma_n$, $|\Gamma_n|$, and $\angle \Gamma_n$ are the reflection coefficient, amplitude, and phase, respectively, and $|\Gamma|_{\min} = 0.2$ is the minimum reflection amplitude when $f_c = 2.4$ GHz. Given the approximated reflection amplitude $A_n(\angle \Gamma_n)$, as seen later in Section III, we will derive asymptotic optimal reflection phases at each reflecting element for the maximum performance as the number of reflecting elements increases to infinity.

**B. System Model**

Consider a single BS multiple-input single-output (MISO) system that consists of a set $\mathcal{K}$ of $K$ single-antenna user equipments (UEs) and multiple RISs each of which having $N$ reflecting elements, as shown in Fig. 3. The BS is equipped with $M$ antennas and serves one UE at each time slot based on a time-division multiple access (TDMA) \[18\]. Also, the BS transmits the downlink signal to scheduled UE through the transmit beamforming. In our system model, we consider two types of UEs: a) UEs directly connected to the BS (called DUEs) and b) UEs connected to the BS via an RIS (called RUEs). Each UE can measure the downlink channel quality information (CQI) and transmit this information to the BS as done in existing cellular systems \[19\]. For UEs whose CQI exceeds a pre-determined threshold, the BS will directly transmit downlink signals to these UEs (which are now DUEs) without using the RIS. When the CQI is below a pre-determined threshold (i.e., the direct BS-UE channel is poor), the BS will have to allocate, respectively, suitable RISs to those UEs (that become RUEs) experiencing this poor CQI and, then, send a control signal to each RIS controller via a dedicated control link. Given the received BS control signal, the RIS controller determines $N$ bias DC voltages for
all reflecting elements and then, the varactor capacitance can be controlled, resulting in phase shifts of the reflection wave. Note that an RIS cannot coherently align, simultaneously, with the desired channels of all RUEs which, in turn, limits system performance [4]–[6]. For densely located RISs, we assume that each RUE is connected to different RISs (i.e., one RUE per one RIS) depending on the location of each RUE. Also, given a practical range of mobility speed and carrier frequency, we consider that all channels are generated from quasi-static block fading whose coherence time covers the downlink transmission period [20], as shown in Fig. 4. In accordance with 3GPP LTE specification [21], we consider two types of reference signals (RSs): Channel state information-RS (CSI-RS) and demodulation RS (DMRS). A CSI-RS is used to estimate the CSI and report CQI back to the BS, and a DMRS is a beamformed RS used to estimate an effective CSI for demodulation [21]. In order to estimate accurate CSI, an RIS will not operate during the CSI-RS period and, simultaneously, the BS can send a control signal to the RIS controller via a dedicated control link during this period. Then, the RIS operates based on this control signal, and reflects the DMRS and data signal with controlled phase shifts. The RUE receives the phase-shifted DMRS and estimates the effective CSI, and eventually, the downlink signal can be decoded. Hereinafter, we assume that the channel state of each wireless link follows a stationary stochastic process under a perfect CSI at the BS and, hence, our analysis will result in a performance bound of practical channel estimation scenarios.

We divide the UE set into two sets, such that $\mathcal{K} = \mathcal{D} \cup \mathcal{R}$ where $\mathcal{D}$ is the set of DUEs and
\( \mathcal{R} \) is the set of RUEs. Then, the received signal at UE \( k \) is obtained as

\[
y_k = \begin{cases} 
\sqrt{P} h_k^H w_k x_d^k + n_d^k, & \text{if } k \in \mathcal{D}, \\
\sqrt{P} f_k^H \Phi_k G_k w_k x_r^k + n_r^k, & \text{if } k \in \mathcal{R},
\end{cases}
\]

(2)

where \( P \) is the BS transmit power and \( h_k \in \mathbb{C}^{M \times 1} \), \( G_k \in \mathbb{C}^{N \times M} \), and \( f_k \in \mathbb{C}^{N \times 1} \) are, respectively, the fading channels between the BS and DUE \( k \), between the BS and RIS \( k \), and between RIS \( k \) and RUE \( k \). Also, \( w_k \in \mathbb{C}^{N \times 1} \) is the transmit beamforming vector and \( x_d^k \) and \( x_r^k \) are downlink transmit symbols for DUE \( k \) and RUE \( k \), respectively, with noise terms \( n_d^k \sim \mathcal{CN}(0, N_0) \) and \( n_r^k \sim \mathcal{CN}(0, N_0) \). In (2), \( \Phi_k \in \mathbb{C}^{N \times N} \) is a reflection matrix (i.e., passive beamformer) that includes reflection amplitudes and phases resulting from \( N \) reflecting elements. This reflection matrix is controlled by the RIS control signal from the BS and then, \( \Phi_k \) can be obtained by using (1) as follows:

\[
\Phi_k = \text{diag}\left( A(\angle \Gamma_1) e^{j\angle \Gamma_1}, A(\angle \Gamma_2) e^{j\angle \Gamma_2}, \ldots, A(\angle \Gamma_N) e^{j\angle \Gamma_N} \right).
\]

(3)

Hence, the instantaneous SNR at UE \( k \) can be obtained as follows:

\[
\gamma_k = \begin{cases} 
P E_{k}^d |h_k^H w_k|^2 / N_0, & \text{if } k \in \mathcal{D}, \\
P E_{k}^r |f_k^H \Phi_k G_k w_k|^2 / N_0, & \text{if } k \in \mathcal{R},
\end{cases}
\]

(4)

where \( E_{k}^d \) and \( E_{k}^r \) are the average energy per symbol for DUE \( k \) and RUE \( k \), respectively. Given this practical RIS model, our goal is to maximize (4) and eventually achieve (asymptotically) the SNR of an ideal RIS as \( N \to \infty \). In most prior studies such as [11] and [11]–[13], the properties of the reflection wave, such as the frequency, amplitude, and phase, are assumed to be independently controlled, however, these properties are closely related to each other as discussed in Section II. A. Hence, their relationship should be considered in the system model to accurately verify the potential of practical RISs. Note that the SNR of an ideal RIS system increases with \( \mathcal{O}(N^2) \) as \( N \to \infty \) [11]. Since the diversity order of a conventional antenna array system is linearly proportional to the number of transmit antennas [22], an ideal RIS
can achieve a squared diversity order of a conventional array system equipped with \( N \) transmit antennas. This squared diversity gain can be obtained from an ideal RIS assumption in which the incident wave is reflected by an RIS without power loss and the BS sends infinite bits to the RIS controller via unlimited RIS control link capacity. Given a practical RIS model in [4], we will propose a novel passive beamformer that can achieve \( O(N^2) \), asymptotically, even with one bit control for each reflecting element. Moreover, we will propose a new modulation scheme and an effective resource allocation algorithm that can be used in our RIS system to increase an achievable sum-rate under the aforementioned practical RIS considerations.

III. Optimality of the Achievable Downlink Rate in an RIS

We analyze the optimality of the achievable rate using practical RISs under consideration of the limited capacity of the RIS control link and practical reflection coefficients, as \( N \) increases to infinity. As proved in [11], given an ideal RIS that reflects the incident wave without power loss under unlimited control link capacity, the downlink SNR of the RIS achieves, asymptotically, the order of \( O(N^2) \), as \( N \) increases to infinity. However, the downlink SNRs of a conventional massive MIMO or MIMO relay system, each of which equipped with \( N \) antennas, equally increase with \( O(N) \) as proved in [22]. This squared SNR gain of RIS will analytically result in twice as much performance as conventional array systems in terms of achievable rate, without additional radio resources. In order to prove the optimality of the achievable rate using practical RISs under the aforementioned limitations, we first design a passive beamformer that achieves the SNR order of \( O(N^2) \) asymptotically. We then design a modulation scheme which can be used in an RIS that uses ambient RF signals to transmit data without additional radio resource and achieves the asymptotic SNR in order of \( O(N) \) like a conventional massive MIMO (or MIMO relay) system. We finally propose a resource allocation algorithm to maximize the sum-rate of the considered RIS-based MISO system.

A. Passive Beamformer Design

The maximum instantaneous SNR at DUE \( k \) can be achieved by using a maximum ratio transmission (MRT) where \( \mathbf{w}_k = \mathbf{h}_k / \| \mathbf{h}_k \| \), which yields an SNR \( \gamma_k = PE_k \| \mathbf{h}_k \|^2 / N_0 \). We then formulate an optimization problem whose goal is to maximize instantaneous SNR at RUE \( k \) with respect to \( \Phi_k \) and \( \mathbf{w}_k \), as follows:

\[
\max_{\Phi_k, \mathbf{w}_k} \frac{PE_k | f_k^H \Phi_k G_k \mathbf{w}_k |^2}{N_0},
\]

(5)
Algorithm 1 Reflection Phase Selection Algorithm

1: Initialization: Select \( \hat{m} = \arg \max_{1 \leq m \leq M} \| g_m \| \) and set \( s_0 = 0 \) and \( i = 1 \).
2: Reflection phase selection: \( \hat{\theta} = \arg \max_{\theta \in \mathcal{P}} |s_{i-1} + a_{i0} \phi(\theta)| \).
3: Update reference vector: \( s_i = s_{i-1} + a_{i0} \phi(\hat{\theta}) \).
4: Select \( \phi_i = \phi(\hat{\theta}) \).
5: Set \( i \leftarrow i + 1 \) and go to Step 2 until \( i = N + 1 \).
6: Return \( \Phi_k = \text{diag}(\phi_1, \phi_2, \ldots, \phi_N) \).

s.t. \( |\Gamma|_{\min} \leq A(\angle \Gamma_n) \leq 1, \forall n, \) \hspace{1cm} (5a)

\[ -\pi \leq \angle \Gamma_n \leq \pi, \forall n. \] \hspace{1cm} (5b)

For any given \( \Phi_k \), it is well known that the MRT precoder is the optimal solution to problem \( (5) \) such that \( w_k = \frac{G_k^H \Phi_k^H f_k}{\| G_k^H \Phi_k^H f_k \|} \) \cite{23}. Then, we formulate an optimization problem with respect to \( \Phi_k \) as follows:

\[ \max_{\Phi_k} \frac{P E_n^r \| f_k^H \Phi_k G_k \|}{N_0} , \] \hspace{1cm} (6)

s.t. \( (5a), (5b) \).

This problem is non-convex since \( \| f_k^H \Phi_k G_k \| \) is not concave with respect to \( \Phi_k \). Let \( f_k = [f_1, \cdots, f_N]^T \) and \( G_k = [g_1, \cdots, g_M] \), where \( g_m \in \mathbb{C}^{N \times 1} = [g_{1m}, \cdots, g_{Nm}]^T \) is the channel between BS antenna \( m \) and RIS \( k \). Then, \( \| f_k^H \Phi_k G_k \| \) is obtained by

\[ \| f_k^H \Phi_k G_k \| = \sum_{m=1}^{M} \sum_{n=1}^{N} |f_n| |g_{nm}| A(\angle \Gamma_n) e^{j(\angle \Gamma_n + \angle f_n + \angle g_{nm})} \] \hspace{1cm} (7)

\[ = \sum_{m=1}^{M} \sum_{n=1}^{N} a_{nm} \phi(\angle \Gamma_n) \] \hspace{1cm} (8)

where \( a_{nm} = |f_n| |g_{nm}| e^{j(\angle f_n + \angle g_{nm})} \) and \( \phi(\angle \Gamma_n) = A(\angle \Gamma_n) e^{j\angle \Gamma_n} \). Since \( |f_n| |g_{nm}| A(\angle \Gamma_n) \) is always greater than zero given that \( |\Gamma|_{\min} > 0 \), \( (8) \) can readily achieve \( O(N^2) \) when \( \angle \Gamma_n = -\angle f_n - \angle g_{mn0} \) for \( 1 \leq m_0 \leq M \) and \( \forall n \). However, using continuous reflection phases is impractical when we have a limited RIS control link capacity and practical RIS hardware. Given a discrete reflection phase set \( \mathcal{P} = \{-\pi, -\pi + \Delta \phi, \cdots, -\pi + \Delta \phi(2^b - 1)\} \) where \( \Delta \phi = 2\pi/2^b \) and \( b \) is the number of RIS control bits at each reflecting element, we design a suboptimal reflection
matrix $\hat{\Phi}_k$ which can achieve $O(N^2)$ as $N \to \infty$, asymptotically, as shown in Algorithm 1. In this algorithm, we first select antenna $\hat{m}$ which has the largest channel gain among $M$ channels between the BS and the RIS. Since a reflection amplitude is always 1 when its phase equals to $-\pi$, $\phi(-\pi)$ is selected as a reflection phase $\phi_1$ and we determine $a_{i\hat{m}}\phi(-\pi)$ as a reference vector, $s_1$, in the first round. Note that $\left\| f_k^H \Phi_k G_k \right\|^2$ is calculated based on the sum of $N$ vectors such as $\sum_{n=1}^{N} a_{nm}\phi(\angle \Gamma_n)$, as shown in (8). Therefore, when $i > 1$, we compare the Euclidean norm of vector additions between $s_{i-1}$ and $\theta$ shifted candidates, i.e., $\left| s_{i-1} + a_{i\hat{m}}\phi(\theta) \right|$, $\forall \theta \in \mathcal{P}$, and select $\hat{\theta}$ with the maximum Euclidean norm. Therefore, we can derive the suboptimal solution such that $\phi_i = \phi(\hat{\theta})$ for each reflecting element $i$. Algorithm 1 results in a suboptimal solution and will not achieve the optimal performance that can be obtained by the exhaustive search method with $O(2^{bN})$ complexity. However, we can prove the following result related to the asymptotic optimality of Algorithm 1.

**Proposition 1.** Algorithm 1 can achieve an instantaneous SNR in order of $O(N^2)$ regardless of the number of RIS control bits, $b \geq 1$, as $N \to \infty$.

**Proof:** In order to analyze the impact of $b$ on the instantaneous SNR resulting from Algorithm 1 we first consider the case of $b = 1$ with $\mathcal{P} = \{-\pi, 0\}$. In Step 2 of Algorithm 1 the Euclidean norm of vector addition is obtained as follows:

$$|s_{i-1} + a_{i\hat{m}}\phi(\theta)| = \sqrt{|s_{i-1}|^2 + |a_{i\hat{m}}\phi(\theta)|^2 + 2|s_{i-1}||a_{i\hat{m}}\phi(\theta)| \cos \delta}$$

$$(9) \quad \geq \sqrt{|s_{i-1}|^2 + |f_i|^2|g_{i\hat{m}}|^2|\Gamma|_{\min}^2 + 2|s_{i-1}||f_i||g_{i\hat{m}}||\Gamma|_{\min}\cos \delta},$$

$$(10)$$

where $\delta = \angle s_{i-1} - \angle a_{i\hat{m}} - \theta$ and (a) results from the worst case scenario where $|\phi(\theta)| = |\Gamma|_{\min}$. Given that $\theta \in \{-\pi, 0\}$ and $|\Gamma|_{\min} > 0$, we can always select $\theta$ that satisfies $\cos \delta \geq 0$ and then, $|s_i|$ will increase as $i$ increases until $i = N$. Therefore, $|s_N|$ will increase with $O(N)$ as $N \to \infty$. Since $\left| \sum_{n=1}^{N} a_{m\hat{m}} \phi(\angle \Gamma_n) \right|^2$ in (8) equals to $|s_N|^2$ in Algorithm 1 $\left\| f_k^H \hat{\Phi}_k G_k \right\|^2$ increases with $O(N^2)$ as $N \to \infty$, which completes the proof.

Proposition 1 shows that the instantaneous SNR resulting from Algorithm 1 can be in order of $O(N^2)$ even with one bit control for each reflecting element. Moreover, Algorithm 1 requires a complexity of $O(N)$ resulting in simpler RIS control compared to the various existing works on RIS in [1], [11], and [17].

Next, we analyze the average SNR of a downlink RIS system, under consideration of the RIS reflection matrix derived from Algorithm 1. We first consider an ideal RIS control link that can
use an infinite and continuous reflection phases at the RIS, and assume that \( f_k \sim \mathcal{CN}(0, I_N) \)
and \( g_m \sim \mathcal{CN}(0, I_M) \) considering Rayleigh fading channels. From (7), the instantaneous SNR
at RUE \( k \) is obtained by
\[
\gamma_k = \frac{P E_k^t}{N_0} \| \Phi_k G_k \|^2 = \frac{P E_k^t}{N_0} \sum_{m=1}^{N} \left| f_n \right| \left| g_{nm} \right| A(\theta_n) \left| e^{j(\angle \Gamma_n + \angle f_n^* + \angle g_{nm})} \right|^2.
\]  
(11)

By selecting \( \angle \Gamma_n = \theta_n = -\angle f_n - \angle g_{nm} \) for \( 0 \leq m_0 \leq M \) and \( \forall n \), we have the following:
\[
\frac{P E_k^t}{N_0} \left\| \Phi_k G_k \right\|^2 \geq \frac{P E_k^t}{N_0} \left( \sum_{n=1}^{N} \left| f_n \right| \left| g_{nm} \right| A(\theta_n) \left| e^{j\Delta g_{nm} \theta_n} \right| \right)^2 \geq \left( \sum_{n=1}^{N} \left| f_n \right| \left| g_{nm} \right| A(\theta_n) \left| e^{j\Delta g_{nm} \theta_n} \right| \right)^2.
\]  
(12)

where \( \Delta g_{nm} = \angle g_{nm} - \angle g_{nm} \) and \( \hat{\Phi}_k \) and (b) result from Algorithm 1 with infinite \( b \) and the minimum reflection amplitude, i.e., \( A(\theta_n) \geq |\Gamma|_{\text{min}}, \forall n \), respectively. We refer to (12) as an instantaneous SNR lower bound \( \gamma_k \). For notational convenience, we define \( \gamma_1 = \sum_{n=1}^{N} \left| f_n \right| \left| g_{nm} \right| \) and \( \gamma_{r,m} = \sum_{n=1}^{N} \left| f_n \right| \left| g_{nm} \right| e^{j\Delta g_{nm} \theta_n} \) in (12). Then, the random variable \( \gamma_k \) follows Lemma 1.

**Lemma 1.** Based on the central limit theorem (CLT), as \( N \to \infty \), \( \gamma_k \) follows a non-central chi-square distribution with one degree of freedom and its mean and variance are asymptotically converge to
\[
E[\gamma_k] = \frac{N^2 P E_k^t \pi^2}{16 N_0} |\Gamma|_{\text{min}}^2, \quad \text{Var}[\gamma_k] = \left( 1 - \frac{\pi^2}{16} \right) \frac{N^3 P^2 (E_k^t)^2 \pi^2 |\Gamma|_{\text{min}}^4}{4 N_0^2}.
\]  
(13)

**Proof:** Since \( |f_n| \) and \( |g_{nm}| \) are independent, \( \gamma_1 \) converges to a Gaussian distribution based on the CLT: \( \gamma_1 \sim \mathcal{N}\left( \frac{N \pi^2}{4}, N \left( 1 - \frac{\pi^2}{16} \right) \right) \), as \( N \to \infty \). Then, \( |\gamma_1|^2 \) follows a non-central chi-square distribution with one degree of freedom with mean of \( N \left( 1 + \frac{\pi^2}{16} (N - 1) \right) \) and variance of \( N^2 \left( 1 - \frac{\pi^2}{16} \right) \left( 2 - \frac{\pi^2}{8} + \frac{N \pi^2}{4} \right) \). Similarly, \( \gamma_{r,m} \) converges to a complex Gaussian distribution as \( \gamma_{r,m} \sim \mathcal{CN}(0, N) \), and \( |\gamma_{r,m}|^2 \) follows a central chi-square distribution with two degrees of freedom with mean \( N \) and variance \( N^2 \), as \( N \to \infty \). Since \( |\gamma_{r,m}|^2 \) are independent random variables for different \( m \) and also independent with \( |\gamma_1|^2 \), we have the following mean and variance of \( \gamma_k \), respectively:
\[
E[\gamma_k] = \frac{N P E_k^t |\Gamma|_{\text{min}}^2}{N_0} \left( M + \frac{\pi^2 (N - 1)}{16} \right),
\]  
(14)
\[
\text{Var}[\gamma_k] = \frac{N^2 P^2 (E_k^t)^2 |\Gamma|_{\text{min}}^4}{N_0^2} \left\{ \left( 1 - \frac{\pi^2}{16} \right) \left( 2 - \frac{\pi^2}{8} + \frac{N \pi^2}{4} \right) + M - 1 \right\}.
\]  
(15)
As $N \to \infty$, (14) and (15) converge to (13), which completes the proof.

Lemma 1 shows that the lower bound of the average SNR increases with $O(N^2)$ and the average SNR gains from $M-1$ antennas become negligible compared to those from $m_0$, as $N \to \infty$. Moreover, we can observe from (13) that this lower bound is equal to the single antenna case in [8]. Since $O(N^2)$ can also be achieved by the instantaneous SNR resulting from Algorithm I with limited $b$, as proved in Proposition 1, the average SNR gains from $M-1$ antennas are also negligible compared to those from $m_0$ in the limited RIS control link capacity, as $N \to \infty$. Therefore, the average SNR resulting from Algorithm I with limited $b$ will also converge to that of the single antenna case. Given this convergence of the average SNR, we can use $w_k$ as an antenna selection that can achieve full multi-antenna diversity with a low-cost and low-complexity instead of a MRT [24]. By selecting the BS antenna whose channel gain has the maximum value such as in Step 1 of Algorithm I, we can determine the transmit precoding vector, $w_k = [w_1, \ldots, w_M]^T$, as follows:

$$
\begin{align*}
  w_{m} &= 1, \quad \text{if } m = \hat{m}, \\
  w_{m} &= 0, \quad \text{if } m \neq \hat{m},
\end{align*}
$$

(16)

where $\hat{m} = \arg \max_{1 \leq m \leq M} \|g_m\|$. Although an MRT precoder achieves the optimal performance for a single-user MISO system, it requires multiple RF chains associated with multiple antennas resulting in higher cost and hardware complexity compared to the transmit antenna selection scheme. Moreover, since the average SNR will converge to the single antenna system as $N$ increases, a large $N$ results in a performance convergence between MRT and transmit antenna selection. Lemma 1 also shows that the variance of the instantaneous SNR increases with $O(N^3)$, asymptotically, and this will result in scheduling diversity. To achieve this scheduling diversity for a large $N$, we will develop a new resource allocation algorithm that can achieve the maximum scheduling diversity in Section IV. C. Moreover, we can observe from Lemma 1 that (14) with $M = 1$ is equal to the average SNR derived in [8], showing that the SER resulting from Algorithm I also decays exponentially as a function of $N$.

B. Modulation for Unscheduled RUE

Next, we develop a modulation scheme that can be used to increase the achievable RIS sum-rate. In our downlink TDMA scheme, the BS can transmit the downlink signal to only one scheduled UE at each time slot. However, by reflecting the ambient RF signals generated
from the BS, each RIS also can send the downlink signal to its unscheduled RUE at each time slot, as done in ambient backscatter communications [25]. Different from ambient backscatter communications, an RIS can convey information by reflecting the ambient RF signal without interfering with existing scheduled UEs. In addition to the data rates obtained from the BS’s downlink DUEs, we can obtain additional data rates at unscheduled RUEs without additional radio resources, resulting in higher achievable sum-rate than a conventional network without RIS. For convenience, we refer to each unscheduled RUE and its connected RIS as uRUE and uRIS, respectively. As shown in [8] and [9], an RIS can transmit PSK signals to serving user by controlling its reflection phase. As such, we consider that each uRIS controls its reflection phase to send the downlink signal to each corresponding uRUE by reflecting the incident wave from the BS, whenever the DUE is scheduled. In order to avoid undesired interference from uRISs to the scheduled DUE, we consider the following procedure. First, the BS sends the RIS control signals, which are related to the data symbols for each uRUE, to uRISs during the CSI-RS period (see Fig. 4). Each uRIS controls $N$ bias DC voltages for all reflecting elements depending on the control signal and reflects the incident wave from the BS with the same reflection phase during the DMRS and data transmission periods. Meanwhile, each scheduled DUE (sDUE) receives the DMRS from the BS and estimates the effective CSI which includes the transmit precoding at the BS and the phase shifts from the uRISs. Since all uRISs keep using the same reflection phase from the beginning of the DMRS to the end of the data transmission, this effective CSI will not change during the downlink data transmission and this results in zero interference. Similarly, each uRUE receives the CSI-RS from the BS and estimates the CSI between the BS and each uRUE without controlled phase shifts. Note that the BS broadcasts the information about the transmit precoder and the modulation of sDUE by using the downlink control indicator (DCI), based on the LTE specification in [26] and [27]. Given the estimated CSI and the broadcast DCI, the uRUE can calculate the effective CSI and eventually decode the downlink signal transmitted from the uRIS. By using this procedure, each uRIS needs to transmit only one symbol during the channel coherence time, however, the uRUEs can achieve the asymptotic SNR in order of $\mathcal{O}(N)$ as will be proved in Theorem 1. Moreover, since the uRUE receives the same symbols from the uRIS during the data transmission period, the uRUE will achieve a diversity gain proportional to the length of the data transmission period. Furthermore, since the BS can send the RIS control signal related to one symbol for each uRIS during the entire channel coherence time, we can
reduce the link burden of the RIS control link. Given this procedure, the received downlink signal at uRUE $i$ can be obtained by:

$$y_i = \sqrt{P} f_i^H \bar{\Phi}_i G_i w_k x_k^d + n_i^r,$$

(17)

where $i \in R$, $k \in D$, and $w_k$ and $x_k^d$ are the transmit precoder and the downlink signal for sDUE $k$, respectively. To modulate the downlink data bits into the reflection matrix, we propose the following modulated reflection matrix for the uRUE $i$:

$$\bar{\Phi}_i = A(\omega_i) e^{j\omega_i} I_N, \quad \omega_i \in \mathcal{M},$$

(18)

where $e^{j\omega_i}$ is the proposed $M_o$-PSK modulation symbol for uRUE $i$ and $\mathcal{M} = [\mu_1, \ldots, \mu_{2M_o}]$ is a set of the corresponding $M_o$-PSK modulation. For two examples of QPSK signaling as shown in Fig. 5, $\mathcal{M} = \{0, \pi/4, \pi/2, 3\pi/4\}$ when $x_k^d$ is BPSK symbol and $\mathcal{M} = \{0, \pi/8, \pi/4, 3\pi/8\}$ when $x_k^d$ is QPSK symbol. When $x_k^d$ is a QAM symbol, we can also obtain $\mathcal{M}$ according to the minimum angle between adjacent QAM symbols. Hence, using (17) and (18), we obtain:

$$y_i = \sqrt{P} f_i^H G_i w_k A(\omega_i) e^{j\omega_i} x_k^d + n_i^r.$$

(19)

Note that $f_i^H G_i$ can be estimated from the CSI-RS, and the modulation scheme of $x_k^d$ and $w_k$ are known at the uRUE from the broadcasted DCI. Assuming that all RISs are equipped with the same passive elements resulting in identical reflection coefficient models as in (1), the uRUE can calculate $A(\omega_i)$ for $|\mathcal{M}|$ symbols and eventually estimate $\omega_i$ resulting in additional data rate in the considered RIS system. From (19), we can prove the following result related to the average SER at the uRUE.
Theorem 1. The uRUE achieves an average SNR in order of $O(N)$ as $N \to \infty$ and the average SER with the proposed $M_o$-PSK signaling can be approximated by

$$P_e = \frac{1}{2^{M_o}} \sum_{p=1}^{2^{M_o}} \int_{0}^{\pi} \frac{1}{1 - \frac{NN_sE_s(\theta)A(\mu_p)^2}{N_0}} d\theta,$$

where $t(\theta) = \frac{-\sin^2(\Delta \mu/2)}{\sin^2(\theta)}$, $\Delta \mu$ is the angular spacing of the proposed $M_o$-PSK symbols, and $N_s$ is the number of transmitted symbols during the downlink data transmission period.

Proof: See the appendix.

Theorem 1 shows that the uRUE can achieve an asymptotic SNR in order of $O(N)$ by using the proposed modulation scheme resulting in several implications. First, an RIS can provide the same asymptotic SNR as a conventional massive MIMO or MIMO relay system for all unscheduled RUEs, simultaneously. For the scheduled RUEs, an RIS also provides an asymptotic SNR in order of $O(N^2)$, which is much higher than that of conventional MIMO systems, as proved in Proposition 1. Hence, an RIS can support the demand for massive connectivity and high data traffic, without additional radio resources. Moreover, Theorem 1 shows that the average SER of uRUE can be obtained based on deterministic values and we can evaluate the reliability of the considered RIS system without extensive simulations. In particular, we can observe from (20) that the average SER decreases as $N$ increases and it eventually reaches zero as $N \to \infty$, resulting in reliable communication regardless of $E_s/N_0$ even at the uRUEs.

C. Resource Allocation Algorithm

Next, we develop a new resource allocation algorithm for RIS based on the approaches of Section IV. We propose a resource allocation algorithm that includes joint transmission power control and user scheduling for the maximum average sum-rate. Given our downlink TDMA system, we allow the channel states to vary over time slot, $t$, in which each time slot has duration of the channel coherence time. In accordance with the LTE specification, the minimum scheduling period is equal to one transmission time interval (TTI) and lasts for 1 ms duration [28]. For analytical simplicity, we assume that the channel coherence time is equal to one TTI and consider a scheduling indicator $q^t_k$ that $q^t_k = 1$ when UE $k$ is scheduled in time slot $t$, otherwise $q^t_k = 0$, resulting in $\sum_{k \in K} q^t_k = 1$. Given the proposed phase selection algorithm and the modulation scheme for uRUE, the instantaneous achievable rate at UE $k$ in time slot $t$ is
given as follows:

\[
R^t_k = \begin{cases} 
q_k^t \log \left( 1 + \frac{P^t ||h_k^t||^2}{N_0} \right), & \text{if } k \in D, \\
q_k^t \log \left( 1 + \frac{P^t (f^t_k)^H \tilde{\Phi}^t_k \hat{g}_k^t)^2}{N_0} \right) + \sum_{i \in D} q_i^t \log \left( 1 + \gamma_{ak}^t \right), & \text{if } k \in R,
\end{cases}
\]

(21)

where we use superscript \( t \) for all time-varying variables and we assume that \( E^t_k = E^d_k = 1 \), \( \forall k \in K \). Also, \( \hat{g}_k^t \) is the selected channel from \( G_k^t \) according to (16) and \( \gamma_{ak}^t \) is the received signal-to-interference-plus-noise ratio (SINR) at uRUE \( k \) in time slot \( t \), which corresponds to the additional rate achieved at the uRUE, as given by:

\[
\tilde{\gamma}_{ak}^t = \frac{P^t \alpha_k^t \|f^t_k\|^2}{N_0 \|h_k^t\|^2} + \sum_{j \neq k, j \in R} P^t \|f^t_{jk}\|^2 \|\tilde{\Phi}_{jk}^t \hat{g}_j^t h_i^t\|^2.
\]

(22)

where \( f^t_{jk} \) is the interference channel between uRIS \( j \) and uRUE \( k \), and \( \alpha_k^t \) denotes the SNR loss at uRUE resulting from the modulation order at sDUE \( i \). As the modulation order at the sDUE increases, \( \Delta_\mu \) in (20) decreases resulting in a throughput loss at the uRUE and \( \alpha_k^t \) captures this throughput loss. Since the CQI of DUE \( i \) is higher than the pre-determined threshold, we assume that \( \alpha_k^t \ll 1 \), \( \forall i \in D \), resulting from a high order modulation at sDUE \( i \). Given that we consider a TDMA system and uRISs will not interfere with scheduled UEs, we do not consider the interference from uRISs to the scheduled UEs as shown in (21). However, since uRUEs will experience interference resulting from the undesired reflection wave generated by neighboring uRISs, we consider the interference from other uRISs to uRUE as in (22). Then, the instantaneous sum-rate in time slot \( t \) can be obtained as follows:

\[
R^t = \sum_{i \in D} q_i^t \left\{ \log \left( 1 + \gamma_{di}^t \right) + \sum_{k \in R} \log \left( 1 + \gamma_{ak}^t \right) \right\} + \sum_{k \in R} q_k^t \log \left( 1 + \gamma_{ik}^t \right),
\]

(23)

where \( \gamma_{di}^t = \frac{P^t ||h_i^t||^2}{N_0} \) and \( \gamma_{ik}^t = \frac{P^t (f^t_k)^H \tilde{\Phi}^t_k \hat{g}_k^t)^2}{N_0} \). Hence, the average sum-rate of the considered RIS-based MISO system will be: \( R = \sum_{s_t \in S} \pi_s R^t \), where the set \( S \) consists of \( S \) system states and each system state represents one of the possible channel states for all links [29]. Also, \( s_t \) is the system state at time slot \( t \) and \( \pi_s \) is the probability that the actual system state is in state \( s \), \( \forall s \in S \). In accordance with the LTE specification [19], a CSI is the quantized information which consists of the limited feedback bits and discrete system states, \( S \), is known at the BS which include all possible quantized CSI (i.e., \( s_t \in S \)). Moreover, the number of those system states
will increase as the number of feedback bits and the number of wireless links increase. Since we consider a large $N$, we will analyze the average sum-rate of the considered RIS system as $S \to \infty$.

Next, we formulate an optimization problem whose goal is to maximize the average sum-rate with respect to the scheduling indicator (i.e., $q = [q_1, \cdots, q^S]$ where $q^t = [q^t_1, \cdots, q^t_K]$) and the transmission power at the BS (i.e., $p = [P^1, \cdots, P^S]$) as follows:

$$\max_{q, p} \sum_{s_t \in S} \pi_t R^t,$$

s.t. $\sum_{s_t \in S} \pi_t R^t_k \geq \bar{R}, \forall k \in \mathcal{K}$,  
$\sum_{s_t \in S} \pi_t P^t \leq \bar{P}$,  
$0 \leq P^t \leq P_{\text{max}}, \forall s_t \in S$,  
$q^t_k \in \{0, 1\}, \forall k \in \mathcal{K}, \forall s_t \in S$,  
$\sum_{k \in \mathcal{K}} q^t_k = 1, \forall s_t \in S$,

where $\bar{R}$ is the minimum average rate requirement and $\bar{P}$ and $P_{\text{max}}$ are the maximum average and instantaneous transmission power at the BS, respectively. Note that (24) is a mixed integer optimization problem which is generally hard to solve. Hence, we relax $q^t_k$ into continuous value such that $0 \leq q^t_k \leq 1, \forall k \in \mathcal{K}, \forall s_t \in S$. Given that the objective function and constraints in (24) are continuous after relaxation, we can easily show that the duality gap of (24) becomes zero as $N \to \infty$ by using [29, Proposition 1]. To formulate a dual problem, we define a Lagrangian function corresponding to the optimization problem in (24) to solve its dual problem as follows:

$$L(q, p, \lambda, \mu) = \sum_{s_t \in S} \pi_t R^t + \sum_{k \in \mathcal{K}} \lambda_k \left( \sum_{s_t \in S} \pi_t R^t_k - \bar{R} \right) + \mu \left( \bar{P} - \sum_{s_t \in S} \pi_t P^t \right)$$

$$= \sum_{s_t \in S} \pi_t L^t (q^t, P^t, \lambda, \mu) - \sum_{k \in \mathcal{K}} \lambda_k \bar{R} + \mu \bar{P},$$

where $L^t (q^t, P^t, \lambda, \mu) = R^t + \sum_{k \in \mathcal{K}} \lambda_k R^t_k - \mu P^t$, and $\lambda = [\lambda_1, \cdots, \lambda_K]$ and $\mu$ are the Lagrangian multipliers. Then, we have the following dual problem:

$$\min_{\lambda \geq 0, \mu \geq 0} F(\lambda, \mu),$$

where

$$F(\lambda, \mu) = \max_{q, p} L(q, p, \lambda, \mu),$$
s.t. \( 0 \leq P^t \leq P_{\text{max}}, \forall s_t \in S \), \hspace{1cm} (28a)

\[ q^t \in Q^t, \forall s_t \in S, \] \hspace{1cm} (28b)

and \( Q^t = \{ q^t | 0 \leq q_k^t \leq 1, \forall k \in K, \text{ and } \sum_{k \in K} q_k^t = 1 \} \). Given Lagrangian multipliers \( \lambda \) and \( \mu \), we can solve (28) by maximizing \( L^t(q^t, P^t, \lambda, \mu) \) for each slot \( t \):

\[
\max_{q^t, P^t} \sum_{k \in K} (1 + \lambda_k) R_k^t - \mu P^t,
\]

s.t. \( 0 \leq P^t \leq P_{\text{max}} \), \hspace{1cm} (29a)

\[ q^t \in Q^t. \] \hspace{1cm} (29b)

(29) shows that the optimal \( q^t \) and \( P^t \) can be obtained without knowledge of the underlying probability \( \pi_t \). For notational simplicity, we define \( \bar{R}^t = \sum_{k \in K} (1 + \lambda_k) R_k^t \) which, using (23), can be given by:

\[
\bar{R}^t = \sum_{k \in K} q_k^t \bar{R}_k^t,
\]

where

\[
\bar{R}_k^t = \begin{cases} 
\sum_k (1 + \lambda_k) \left( \log (1 + \bar{\gamma}_d^t) + \sum_{i \in R} \log (1 + \bar{\gamma}_a^t) \right), & \text{if } k \in D, \\
\sum_k (1 + \lambda_k) \log (1 + \bar{\gamma}_r^t), & \text{if } k \in R.
\end{cases}
\]

Since (29) is still nonconvex optimization problem, we first determine the optimal \( P^t \) for given \( q^t \) and then, obtain the optimal \( q^t \) under the pre-determined \( P^t \). For the fixed \( P^t \), we can observe from (30) that \( \bar{R}_k^t \) is an affine function with respect to \( q_k^t \) and thus, the optimal \( q_k^t \) will be one of the boundary conditions (i.e., \( q_k^t \in \{0, 1\} \)). Hence, the gap between the original problem in (24) and the problem after \( q_k^t \) relaxation will be zero. Given an integer solution \( q_k^t \), we have the following result related to the optimal \( q^t \) and \( P^t \).

**Proposition 2.** The optimal \( q^t \) and \( P^t \) that solve the optimization problem in (29) are obtained, respectively, as follows:

\[
\hat{q}_k^t = \begin{cases} 
1, & \text{if } k = \arg \max_i \left( \bar{R}_i^t | p_i^t = \hat{P}_i - \mu \hat{P}_i \right), \\
0, & \text{otherwise},
\end{cases}
\]

\[
\hat{P}_k^t = \begin{cases} 
\hat{P}_k, & \text{if } \hat{q}_k^t = 1, \\
0, & \text{otherwise},
\end{cases}
\]
for all \( k \in \mathcal{K} \), where \( \hat{P}_t^i = \arg \max_{P^t} (\bar{R}_t^i - \mu P^t) \) for all \( i \in \mathcal{K} \).

**Proof:** The integer solution \( q_k^t \) always has a value of 0 or 1 for the entire range of \( P^t \). If we can calculate the optimal \( P^t \) that maximizes \( \bar{R}_k^t \) for each \( q_k^t = 1 \), \( k \in \mathcal{K} \) (i.e., \( q_k^t = 0 \), \( \forall i \neq k \in \mathcal{K} \)), we can jointly find the optimal \( P^t \) and \( q_k^t \) by comparing those maximum \( \bar{R}_k^t \) values. Since \( \bar{\gamma}_{dk}^t \), \( \bar{\gamma}_{ak}^t \), and \( \bar{\gamma}_{rk}^t \) in (31) are strictly concave functions with respect to \( P^t \), \( \bar{R}_k^t - \mu P^t \) is also a concave function with respect to \( P^t \). For \( k \in \mathcal{R} \), we can readily find the maximizer \( \hat{P}_k^t \) by letting the first derivative be zero under consideration of the power constraint in (29a):

\[
\hat{P}_k^t = \left[ 1 + \lambda_k - \frac{N_0}{\left( f_k^t \right)^H \Phi_k^t \tilde{g}_k^t} \right] \frac{P_{\max}}{2}, \forall k \in \mathcal{R}, \tag{34}
\]

where \( [x]^b_a = a \) if \( x < a \), \( [x]^b_a = b \) if \( x > b \), and otherwise, \( [x]^b_a = x \). For \( k \in \mathcal{D} \), we can also obtain the globally optimal \( \hat{P}_k^t \) using a simple gradient method under the constraint in (29a). From those results of \( \hat{P}_k^t \), \( \forall k \in \mathcal{K} \), we can find the optimal \( q_k^t \) as in (32) and the corresponding \( \hat{P}_k^t \) is the optimal transmission power, which completes the proof.

Proposition 2 shows that, for given \( \lambda \) and \( \mu \), we can obtain the optimal solution by solving \( \hat{q}_k^t (\lambda, \mu) = [q_k^1, \ldots, q_k^K] \) and \( \hat{P}_k^t (\lambda, \mu) = \hat{P}_k^t \). In fact, \( \bar{\gamma}_{dk}^t \) and \( \bar{\gamma}_{rk}^t \) should be calculated based on Algorithm 1 and (18), respectively, to obtain the optimal \( \hat{P}_k^t \) for all \( k \in \mathcal{K} \) and this will result in high complexity. In case of \( \bar{\gamma}_{dk}^t \), \( \tilde{g}_k^t \) is used for the proposed PSK modulation and is the pre-determined reflection matrix regardless of its channel state. However, in case of \( \bar{\gamma}_{rk}^t \), \( \Phi_k^t \) is designed based on Algorithm 1 and should be updated at each time slot \( t \) for all \( k \in \mathcal{K} \),

---

**Algorithm 2 Resource Allocation Algorithm**

1: **Initialization:** \( \lambda^0 = 0, \mu^0 = 0, \) and \( t = 0 \).
2: **For each time slot** \( t \): \( \lambda = \lambda^t \) and \( \mu = \mu^t \).
3: Calculate \( \hat{P}_t^i = \arg \max_{P^t} (\bar{R}_t^i - \mu P^t), \forall i \in \mathcal{K} \).
4: Select \( \forall k \in \mathcal{K} \),
   \[
   \hat{q}_k^t = \begin{cases} 
   1, & \text{if } k = \arg \max_i (\bar{R}_t^i | P^t = \hat{P}_t^i), \\
   0, & \text{otherwise},
   \end{cases}
   \]

   \[
   \hat{P}_t^k = \begin{cases} 
   \hat{P}_t^i, & \text{if } \hat{q}_k^t = 1, \\
   0, & \text{otherwise},
   \end{cases}
   \]

5: **Scheduling and power control:** \( \hat{q}^t (\lambda, \mu) = [\hat{q}_1^t, \ldots, \hat{q}_K^t] \) and \( \hat{P}^t (\lambda, \mu) = \hat{P}^t \).
6: **Update** \( \lambda^{t+1} = [\lambda^t - \Delta^t \lambda^t]^{+}, \mu^{t+1} = [\mu^t - \Delta^t \mu^t]^{+} \), and \( t \leftarrow t + 1 \).
7: **End for**
resulting in high computational complexity. Note that Algorithm 1 achieves $O(N^2)$ in terms of instantaneous SNR regardless of the number of RIS control bits as $N \to \infty$, as proved in Proposition 1. Consider an ideal RIS that can achieve an optimal instantaneous SNR using a continuous reflection phases with lossless reflection amplitudes, we can obtain the upper bound of the instantaneous SNR at RUE $k$ as follows:

$$\bar{\gamma}_{tk} \leq \tilde{\gamma}_{tk} = \frac{P_t}{N_0} \sum_{m=1}^{M} \sum_{n=1}^{N} \left| f_{tn} \right| \left| g_{nm}^t \right|^2.$$  \hspace{1cm} (35)

In (35), we can observe that $\tilde{\gamma}_{tk}$ also increases with $O(N^2)$ as $N \to \infty$, verifying the SNR optimality of Algorithm 1. Therefore, we can use $\tilde{\gamma}_{tk}$ instead of $\bar{\gamma}_{tk}$ that can significantly reduce the computational complexity.

Next, we solve the dual problem whose goal is to minimize $F(\lambda, \mu)$ in (27). Since $F(\lambda, \mu)$ is a convex function with respect to $\lambda$ and $\mu$, we use the stochastic subgradient algorithm as the following iterative updates in [30] and [31]:

$$\lambda^{t+1} = \left[ \lambda^t - \Delta^t r^t \right]^+, \quad \mu^{t+1} = \left[ \mu^t - \Delta^t p^t \right]^+,$$  \hspace{1cm} (36)

where $[x]^+ = \max\{0, x\}$ and $\Delta^t$ is the step size at time slot $t$. Also, $r^t$ and $p^t$ are the stochastic subgradients of $F(\lambda, \mu)$ which can be obtained by using Danskin’s theorem in [32]:

$$r^t = [r_1^t, r_2^t, \ldots, r_K^t], \quad p^t = \bar{P} - \hat{P}^t(\lambda^t, \mu^t),$$  \hspace{1cm} (37)

where $r_k = R_k^t \hat{q} = q^t(\lambda^t, \mu^t), P^t = \hat{P}^t(\lambda^t, \mu^t) - \bar{R}$. The stochastic subgradient algorithm in (36) always converges when we consider the step size of the subgradient algorithm as $\Delta^t = 1/t$, as proved.
in [30] and [31]. Since the duality gap goes to zero and the optimality of $\hat{\gamma}_r$ is satisfied as $N \to \infty$, the proposed dual optimal solution asymptotically achieves the maximum average sum-rate under the constraints. The detailed procedure of the proposed algorithm is provided in Algorithm 2.

![Performance comparison](image)

**Fig. 7.** Performance comparison of the average SERs resulting from the proposed modulation with different $N$ values.

**IV. SIMULATION RESULTS AND ANALYSIS**

We run extensive simulations to assess the downlink performance, in terms of the ergodic rate and SER, under a practical-sized RIS environment with finite $N$. We assume that all channels are generated by Rayleigh fading, resulting in $f_k \sim \mathcal{CN}(0, I_N)$ and $g_m \sim \mathcal{CN}(0, I_M)$, $\forall k, m$, where $M = 2$. Note that all numerical results are obtained from Monte Carlo simulations that are statistically averaged over a large number of independent runs.

Fig 6 shows the ratio between the ergodic rate at the RUE, $R_k^e$, resulting from Algorithm 1 and theoretical upper bound. The theoretical upper bound is derived from (35) as follows:

$$\hat{R}_k = \mathbb{E} \left[ \log \left( 1 + \frac{P}{N_0} \sum_{m=1}^{M} \sum_{n=1}^{N} |f_n| |g_{nm}|^2 \right) \right].$$

(38)

We compare the results with the AO algorithm proposed in [17] which is shown up to $N = 500$ due to its computational complexity. As shown in Fig. 6, the ergodic rate ratios resulting from the proposed scheme increase toward 1 as $N$ increases, verifying the optimality of Algorithm 1 as proved in Proposition 1. Although the AO algorithm can also achieve the upper bound
performance, asymptotically, it requires very high complexity. In the AO algorithm, a complexity in order of $O(N^2)$ is required for each iteration and it continues until convergence [17]. However, the proposed Algorithm \[1\] requires $O(N)$ resulting in a much simpler operation at the BS especially for a large $N$. Moreover, the AO algorithm uses MRT which requires multiple RF chains resulting in higher cost and hardware complexity compared to the proposed algorithm.

In Fig. [7] Theorem 1 is verified in the following scenario. The BS transmits BPSK signals to the sDUE via a wireless channel, $h_k$, and also sends the RIS control signals related to the data symbols for the uRIS via a dedicated RIS control link. Data symbols for the uRIS are modulated based on the proposed modulation technique assuming the BPSK signaling (i.e., $M_0 = 2$). As shown in Fig. [7] the asymptotic SERs derived from Theorem 1 are close to the results of our simulations. Moreover, the SERs linearly decrease as $N$ increases given that the SNR difference is always equal to 3 dB when $N$ is doubled. For instance, when the target SER is $2 \cdot 10^{-2}$, the corresponding SNRs are 5, 2, −1, −4, −7, −10 dB for $N = 16, 32, 64, 128, 256, 512$, respectively. This result shows that the SER can be reduced by increasing $N$ and eventually it converges to zero as $N \to \infty$.

In Figs. [8–11] Algorithm \[2\] is verified in the following scenario. The average transmission power at the BS and the noise power are set to $−50$ dBm/Hz and $−174$ dBm/Hz, respectively, and 10 UEs (5 DUE and 5 RUE) are located around the BS. For experimental simplicity, we consider fixed locations for the RISs and UEs such that the distances from the BS to DUEs,
from the BS to RISs, and from the RISs to their serving RUEs are set equally to 5 m, and the distances between the interfering RISs and RUEs are set equally to 10 m. The minimum rate requirement is set to 3 bps/Hz per UE and we consider a path loss model as $11 + 20 \log_{10} d [\text{m}]$ where $d [\text{m}]$ is distance in meters [23].

Fig 8 shows the convergence of the average sum-rate resulting from Algorithm 2. This convergence shows that Algorithm 2 always satisfies the constraints of the transmission power and the rate requirements in (24) regardless of $N$, verifying the convergence of Algorithm 2. Also, the convergence is satisfied within seconds after initial access (e.g., $1 \text{ms} \times 1600 \text{TTIs} = 1.6 \text{s}$) regardless of $N$, showing that the optimal performance can always be achieved after a few seconds.

Figs. 9 and 10 compare the average data rates resulting from Algorithm 2 with and without the minimum rate requirements. As proved in Proposition 1, the received SNRs at the RUEs increase with $\mathcal{O}(N^2)$ while those at the DUEs approximately keep constant as $N$ increases. For a large $N$ without considering the rate requirements, Algorithm 2 always selects RUEs for better sum-rate resulting in unfairness of individual rates, as shown in Fig. 9. However, in this case, we can achieve better performance in terms of the average sum-rate compared to the case in which the UEs have the requirements on the minimum rates, as shown in Fig. 10. On the other hand, when we consider the minimum rate requirements, all individual rates satisfy those requirements while the average sum-rate decreases, as shown in Figs. 9 and 10. Hence, Algorithm 2 can...
control the tradeoff between fairness and maximum performance.

Fig. 11. The average sum-rates with the minimum rate requirements when we do not consider the interference at uRUEs.

In Fig. 11 we assume that the numbers of DUEs and RUEs are set equally to 10, 20, and 50 for $K = 20, 40, 100$, respectively, and the upper bound is derived from (38) as follows: $\hat{R} = \max_{k \in \mathcal{R}} \hat{R}_k$. In the considered downlink RIS system, the uRUEs will experience interference from neighboring uRISs resulting in the performance degradation as given in (22). To compensate for this performance degradation, the BS can send the same RIS control signal to neighboring uRISs and prevent interference, as done in coordinated multipoint transmission [33]. Then, the interference from the neighboring uRISs will be eliminated and the uRUE can rather achieve the diversity gain proportional to the number of those uRISs. In this regard, Fig. 11 shows the average sum-rates resulting from Algorithm 2 without considering the interference from the neighboring uRISs. Since the BS serves a single UE at each time slot, the user selection gain resulting from $\hat{R} = \max_{k \in \mathcal{R}} \hat{R}_k$ will be limited at high SNR region. Therefore, as shown in Fig. 11, the upper bounds for different $K$ (e.g., $K = 20, 40, 100$) converge equally to the same value as $N$ increases. However, since the additional sum-rate is calculated by the sum of $|\mathcal{R}|$ additional rates from all uRUEs, the additional sum-rate linearly increases as $K$ increases as shown in Fig. 11. This additional sum-rate gradually increases its effects on the average sum-rate as $N$ and $K$ increase. Moreover, when $N \geq 850$ and $K \geq 100$, the average sum-rate resulting from the proposed scheme achieves the performance higher than that of single-user upper bound.
In this paper, we have asymptotically analyzed the optimality of the achievable rate using practical RISs in presence of limitations such as practical reflection coefficients and limited RIS control link capacity. In particular, we have designed a passive beamformer that can achieve the asymptotic optimal SNR under discrete reflection phases with a practical reflection power loss, and shown that it achieves a SNR optimality even with one bit RIS control. We have also proposed a modulation scheme that can be used in a downlink RIS system resulting in higher achievable sum-rate than a conventional network without RIS. Moreover, we have derived the approximated SER of the proposed modulation scheme, showing that it achieves an asymptotic SNR of a conventional massive array systems such as a massive MIMO or MIMO relay system. Furthermore, we have proposed the resource allocation algorithm under consideration of the aforementioned passive beamforming and the modulation schemes that achieves the asymptotic optimal sum-rate. We have shown that the proposed algorithms can analytically achieve the performance of an ideal RIS. Simulation results have shown that the results of our algorithms converge to the asymptotic upper bound as $N \rightarrow \infty$. In particular, we have observed that the approximated SER is in close agreement with the result from our simulations and the proposed resource allocation algorithm asymptotically achieves the optimal performance satisfying the minimum rate requirements at each UE. Moreover, our results have shown that the proposed resource algorithm can control the tradeoff between fairness of individual performance and maximum system performance. We finally have shown that the performance of our algorithm becomes higher than the single-user upper bound as the number of UEs increases. Therefore, we expect that our algorithm will be invaluable solution for future wireless networks supporting massive connectivity.

**APPENDIX**

**PROOF OF THEOREM 1**

Given that the maximum instantaneous SNR at sDUE $k$ can be achieved by using the MRT such as $w_k = h_k / \|h_k\|$, the instantaneous SNR at uRUE $i$ can be derived by using (19), as follows:

$$
\gamma_i^{uR} = \frac{PA(\omega_i)^2N_sE_k|f_i^H G_i h_k|^2}{N_0|\|h_k\|^2}.
$$

(39)
Since the uRIS transmits the same symbols during the data transmission period, the uRUE achieves the diversity gain proportional to $N_s$ resulting in $N_s$-fold of the desired signal power as shown in (39). In order to analyze the impact on (39) of the number of BS antennas, we consider two extreme cases: $M = 1$ and $M \to \infty$. We first analyze (39) when $M = 1$. Then, (39) is obtained by

$$\gamma_{ui} = \frac{PA(\omega_i)^2 N_s E_k^d \left| \sum_{n=1}^{N} f_n^* g_n \right|^2}{N_0}.$$  

By using the CLT, $\sum_{n=1}^{N} f_n^* g_n$ converges to a complex Gaussian distributed random variable with zero mean and variance of $N$, as $N \to \infty$. Then, we have

$$\lim_{N \to \infty} \gamma_{ui} \overset{d}{\longrightarrow} \frac{PA(\omega_i)^2 N N_s E_k^d}{2 N_0} \chi_2,$$  

where “$\overset{d}{\longrightarrow}$” denotes the convergence in distribution and $\chi_2$ denotes the chi-square distribution with $2$ degrees of freedom. We consider a random variable $Y_1$ as $\gamma_{ui}$ when $M = 1$. Then, the mean and the probability density function (PDF) of $Y_1$ can be obtained, respectively, as follows:

$$E[Y_1] = \frac{PA(\omega_i)^2 N N_s E_k^d}{N_0},$$  

$$f_{Y_1}(y) = \frac{N_0}{PA(\omega_i)^2 N N_s E_k^d} e^{-\frac{N_0 y}{PA(\omega_i)^2 N N_s E_k^d}}.$$  

Next, we analyze (39) as $M \to \infty$. Then, (39) can be obtained by:

$$\gamma_{ui} = \lim_{M \to \infty} \frac{PA(\omega_i)^2 N_s E_k^d \sum_{m=1}^{M} h_m \sum_{n=1}^{N} f_n^* g_{nm}^2}{N_0 \sum_{m=1}^{M} \left| h_m \right|^2},$$  

where $h_k = [h_1, \ldots, h_M]^T$. On the basis of the CLT for a large $M$, $\sum_{m=1}^{M} h_m \sum_{n=1}^{N} f_n^* g_{nm}$ in (43) converges to a complex Gaussian distributed random variable with zero mean and variance of $NM$. Hence, we have

$$\left| \sum_{m=1}^{M} h_m \sum_{n=1}^{N} f_n^* g_{nm} \right|^2 \overset{d}{\longrightarrow} \frac{NM}{2} \chi_2^2,$$  

Also, the denominator in (43) follows the chi-square distribution: $N_0 \sum_{m=1}^{M} \left| h_m \right|^2 \sim \frac{N_0}{2} \chi_2^2$. Note that the distribution of the ratio of two chi-square distributed random variables is the beta prime distribution [34]. Considering a random variable $Y_M$ as $\gamma_{ui}$ with a large $M$, the mean and the PDF of $Y_M$ can be obtained based on the beta prime distribution, respectively:

$$E[Y_M] = \frac{PA(\omega_i)^2 N M N_s E_k^d}{N_0 (M - 1)},$$  

$$f_{Y_M}(y) = \frac{N_0}{PA(\omega_i)^2 N N_s E_k^d} \left( 1 + \frac{N_0 y}{PA(\omega_i)^2 N M N_s E_k^d} \right)^{-M - 1}.$$  

(46)
As $M \to \infty$, (45) and (46) are obtained, respectively, as follows:
\[
E[Y_\infty] = \lim_{M \to \infty} E[Y_M] = PA(\omega_i)^2 N N_s E_k^d / N_0,
\]
\[
f_{Y_\infty}(y) = \lim_{M \to \infty} f_{Y_M}(y) = \frac{N_0}{PA(\omega_i)^2 N N_s E_k^d} e^{-\frac{N_0 y^2}{2 PA(\omega_i)^2 N N_s E_k^d}},
\]
where $c$ is obtained from the exponential function definition $e^x = \lim_{n \to \infty} (1 + x/n)^n$. Hence, we can obtain the following equalities: $E[Y_1] = E[Y_\infty]$ and $f_{Y_1}(y) = f_{Y_\infty}(y)$. Since the transmit precoder at the BS is designed to match $h_k$ independent with $f_H^i G_i$, the moments of instantaneous SNR at the uRUE will be a monotonic function with respect to $M$. Therefore, $\gamma_{uR_i}$ can be approximated as the chi-square distributed random variable given in (40) regardless of $M$. By using the characteristic of a chi-square distribution, $\gamma_{uR_i}$ has the following moment generating function (MGF) [35]: $M_{\gamma_{uR_i}}(\mu_p, t) = \frac{1}{1 - \frac{N N_s E_k^d A(\mu_p)^2}{N_0 t}}$. Based on this MGF, we can obtain the average SER of the proposed $M_0$-PSK signal at the uRUE as follows [36]:
\[
P_e = \frac{1}{2 M_0} \sum_{p=1}^{2 M_0} \int_{0}^{\pi} -\Delta \mu_p \frac{\sin^2(\Delta \mu_p/2)}{\sin^2\theta} M_{\gamma_{uR_i}}(\mu_p, -\sin^2(\Delta \mu_p/2)) d\theta.
\]

REFERENCES

[1] Q. Wu and R. Zhang, “Beamforming optimization for intelligent reflecting surface with discrete phase shifts,” available online: arxiv.org/abs/1810.10718, Oct. 2018.
[2] W. Saad, M. Bennis, and M. Chen, “A vision of 6G wireless systems: Applications, trends, technologies, and open research problems,” IEEE Network, to appear, 2019.
[3] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “Mobile unmanned aerial vehicles (uavs) for energy-efficient internet of things communications,” IEEE Transactions on Wireless Communications, vol. 16, no. 11, pp. 7574–7589, Nov 2017.
[4] M. Jung, W. Saad, Y. Jang, G. Kong, and S. Choi, “Reliability analysis of large intelligent surfaces (LISs): Rate distribution and outage probability,” IEEE Wireless Communications Letters, to appear, 2019.
[5] M. Jung, W. Saad, Y. Jang, G. Kong, and S. Choi, “Performance analysis of large intelligent surfaces (LISs): Asymptotic data rate and channel hardening effects,” available online: arxiv.org/abs/1810.05667, Oct. 2018.
[6] M. Jung, W. Saad, and G. Kong, “Large intelligent surfaces (LISs): Uplink spectral efficiency and pilot training,” available online: arxiv.org/abs/1904.00453, Mar. 2019.
[7] T. J. Cui, S. Liu, and L. Zhang, “Information metamaterials and metasurfaces,” Journal of Materials Chemistry C, vol. 5, pp. 3644–3668, 2017.
[8] E. Basar, M. D. Renzo, J. de Rosny, M. Debbah, M. S. Aloui, and R. Zhang, “Wireless communications through reconfigurable intelligent surfaces,” available online: arxiv.org/abs/1906.09490, Jul. 2019.
[9] W. Tang, J. Y. Dai, M. Chen, X. Li, Q. Cheng, S. Jin, K. Wong, and T. J. Cui, “Programmable metasurface-based RF chain-free 8PSK wireless transmitter,” Electronics Letters, vol. 55, no. 7, pp. 417–420, 2019.
[10] W. Tang, X. Li, J. Y. Dai, S. Jin, Y. Zeng, Q. Cheng, and T. J. Cui, “Wireless communications with programmable metasurface: Transceiver design and experimental results,” available online: arxiv.org/abs/1811.08119, Nov. 2018.
[11] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design,” available online: arxiv.org/abs/1809.01423, Sep. 2018.

[12] Y. Han, W. Tang, S. Jin, C. Wen, and X. Ma, “Large intelligent surface-assisted wireless communication exploiting statistical CSI,” available online: arxiv.org/abs/1812.05429, Dec. 2018.

[13] C. Huang, G. Alexandropoulos, A. Zaponne, M. Debbah, and C. Yuen, “Energy efficient multi-user MISO communication using low resolution large intelligent surfaces,” in IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, UAE, Dec. 2018.

[14] M. Jung, W. Saad, M. Debbah, and C. S. Hong, “Asymptotic optimality of reconfigurable intelligent surfaces: Passive beamforming and achievable rate,” submitted to IEEE International Conference on Communications (ICC), Jun. 2020.

[15] D. F. Sievenpiper, J. H. Schaffner, H. J. Song, R. Y. Loo, and G. Tagonan, “Two-dimensional beam steering using an electrically tunable impedance surface,” IEEE Transactions on Antennas and Propagation, vol. 51, no. 10, pp. 2713–2722, Oct. 2003.

[16] B. O. Zhu, J. Zhao, and Y. Feng, “Active impedance metasurface with full 360 reflection phase tuning,” Scientific Reports, vol. 3, pp. 3059–3064, Oct. 2013.

[17] S. Abeywickrama, R. Zhang, and C. Yuen, “Intelligent reflecting surface: Practical phase shift model and beamforming optimization,” available online: arxiv.org/abs/1907.06002, Aug. 2019.

[18] S. Santivanez and I. Stavrakakis, “Study of various TDMA schemes for wireless networks in the presence of deadlines and overhead,” IEEE Journal on Selected Areas in Communications, vol. 17, no. 7, pp. 1284–1304, July 1999.

[19] 3rd Generation Partnership Project, “Technical Specification Group Radio Access Network; User Equipment (UE) conformance specification; Radio transmission and reception; Part 1: Conformance testing,” TS 36.521-1, V14.3.0, Aug. 2017.

[20] E. Biglieri, G. Caire, and G. Taricco, “Limiting performance of block-fading channels with multiple antennas,” IEEE Transactions on Information Theory, vol. 47, no. 4, pp. 1273–1289, May 2001.

[21] 3rd Generation Partnership Project, “Technical Specification Group Radio Access Network; Physical layer procedures for data,” TS 38.214, V15.2.0, Jul. 2018.

[22] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, “Energy and spectral efficiency of very large multiuser MIMO systems,” IEEE Transactions on Communications, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[23] D. Tse and P. Viswanath, Fundamentals of Wireless Communication. Cambridge Univ. Press, 2005.

[24] S. Sanayei and A. Nosratinia, “Antenna selection in MIMO systems,” IEEE Communications Magazine, vol. 42, no. 10, pp. 68–73, Oct. 2004.

[25] V. C. Liu, A. N. Parks, V. Talla, S. Gollakota, D. Wetherall, and J. S. R. Smith, “Ambient backscatter: wireless communication out of thin air,” in ACM Special Interest Group on Data Communication (SIGCOMM), 2013.

[26] 3rd Generation Partnership Project, “Technical Specification Group Radio Access Network; Physical layer procedures,” TS 36.213, V14.6.0, Apr. 2018.

[27] ——, “Technical Specification Group Radio Access Network; Multiplexing and channel coding,” TS 36.212, V14.2.0, Apr. 2017.

[28] ——, “Technical Specification Group Radio Access Network; Physical channels and modulation,” TR 36.211, V15.0.0, Dec. 2017.

[29] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, “Opportunistic power scheduling for dynamic multi-server wireless systems,” IEEE Transactions on Wireless Communications, vol. 5, no. 6, pp. 1506–1515, Jun. 2006.
[30] Y. Ermoliev, “Stochastic quasigradient methods and their application to system optimization,” *Stochastics*, vol. 9.

[31] P. Kall and S. W. Wallace, *Stochastic Programming*. Wiley, 1994.

[32] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[33] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. Mayer, L. Thiele, and V. Jungnickel, “Coordinated multipoint: Concepts, performance, and field trial results,” *IEEE Communications Magazine*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[34] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*. Wiley, 1995.

[35] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 2008.

[36] M. Simon and M. S. Alouini, *Digital Communications over Fading Channels*. New Jersey: John Wiley & Sons, 2005.