# Pilot-Domain NOMA for RIS-Assisted Communications

Mehdi Karbalayghareh, *Member, IEEE,* and Aria Nosratinia, *Fellow, IEEE*

***Abstract*—Due to variations in node mobility or differences in the scattering environment, wireless links in multi-user systems often experience non-identical coherence intervals; this is true also in systems assisted by reconfigurable intelligent surfaces (RIS). This paper studies RIS-assisted multi-user downlink systems under link coherence disparity. Since the RIS channel model has many parameters to be estimated, controlling the training overhead under coherence disparity has a strong impact on the practical operation of RIS. Thus motivated, we propose a novel pilot-domain non-orthogonal multiple access (NOMA) technique for RIS-assisted downlink systems. The transmit beamforming and RIS reflection coefficient vectors are jointly designed to maximize the achieved sum-rate. We analyze the resulting reduction in training overhead and the corresponding rate improvements. We investigate efficient pilot placement strategies for multi-user scenarios with arbitrary coherence intervals. Numerical results illustrate the effectiveness of the proposed technique.**

***Index Terms*—Reconfigurable intelligent surfaces, non-orthogonal multiple access, product superposition, coherence disparity, multi-user downlink, channel state feedback.**

## I. Introduction

Reconfigurable intelligent surfaces (RIS) are widely considered to be a promising direction for the future of wireless communication systems. The effectiveness of RIS-assisted communication systems is critically dependent on the availability of channel state information (CSI). However, the number of channel parameters in the RIS-assisted channel path grows linearly with not only the number of transmit and receive antennas, but also the number of RIS reflectors [1]–[14]. As a result, the channel parameters in the presence of a typical RIS are often much more numerous than otherwise. Therefore, channel training and estimation in RIS systems is challenging, and the management of training overhead has been an acknowledged area of important research [15]. Several methods have been proposed for reducing channel training overhead in RIS, including element grouping [16]–[18], infrequent updates [19]–[21], and opportunistic methods [22], [23], among others. In this paper, we explore a new method for reducing pilot overhead in RIS-aided downlink transmission when the different links have non-identical coherence intervals. This is a condition that often occurs in practice due to different mobile velocities or different reflectors in their vicinity, therefore the proposed technique is highly relevant to the practice of mobile communication.

For downlink channel estimation, all receivers are served by the same pilots, thus pilot time slots and pilot power are identical for all users [24], [25]. This remains true even though the coherence time of different links may be non-identical. Since the channel state for some links varies more rapidly (shorter coherence times) than for some other links, the pilot sequence that is geared toward some links may be either inadequate or excessive for other links. Efficiency can be restored if the users employ different pilot duty cycles, but then the temporal orthogonality of pilots and data must be relinquished. We propose a pilot-domain non-orthogonal multiple access (NOMA) strategy, which allows simultaneous pilot and data transmission for different users during pilot slots. The method leverages *product superposition* [26]–[34], a technique initially developed for MIMO downlink systems in the absence of RIS. It enables faster users to be trained *as often as needed*, while at the same time the slower users can be trained *no more than they need*, even though all users are being trained using the same downlink pilots. The slower users reuse some pilot slots for data transmission via a non-orthogonal scheme, *with little or no contamination of the pilots for the fast users*.

In the literature, integrating RIS and NOMA primarily refers to conventional NOMA, i.e., power-domain NOMA, which is typically implemented using transmit-side superposition coding and receiver-side successive interference cancellation [35]–[51]. While this integration can yield gains, particularly in spectral efficiency, it does not address the unequal need for pilots among different users in the downlink, a common occurrence in practical wireless networks. The proposed pilot-domain NOMA in this work is not an *alternative* to conventional power-domain NOMA, as the two operate on parallel tracks: one over pilot slots and the other over data slots. They do not replace each other; they can be combined, and their benefits are cumulative.

Focusing on pilot/data non-orthogonal signaling, there are few available alternatives *in the absence of RIS*, where pilots and data are superimposed additively. The concept of co-timing pilots and data has been used in the context of synchronization [52]–[55]. There are some works on channel estimation for orthogonal time frequency space (OTFS) systems [56]–[59], and for uplink massive MIMO [60], [61], where users transmit pilot symbols at reduced power alongside the data throughout the coherence block. Our work avoids two major weaknesses of the previous models/approaches in additive superposition data/pilot non-orthogonal transmission. First, the data and pilot were previously superimposed without attention to the unequal need of users for pilots. Without

M. Karbalayghareh and A. Nosratinia are with the Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75083-0688, USA (e-mails: mehdi.karbalayghareh@utdallas.edu; aria@utdallas.edu).

A preliminary version of this work appeared in ICC 2024.

This work was supported in part by NSF grants 1956213 and 2148211.

this differentiation, the broader mixing of data and pilots in the previous examples lacked the solid foundational basis that underlies our work. Second, under additive superposition as used in previous works, data will act as noise on the pilots. In comparison, under our method, *for the purposes of channel estimation via the pilot,* any superimposed data will "disappear" by merging into an equivalent/virtual channel link gain, and will not act as noise for the pilot estimation process. This is one of the key aspects that separates the proposed scheme from other techniques and makes it more promising.

RIS-induced channel coefficients need to be estimated at the receiver for coherent communication and shared with the transmitter to design beamforming and RIS reflection coefficient vectors. We investigate the channel estimation at the receivers using either ordinary or superposition pilots passively reflected by the RIS elements, and then return the channel estimates to the transmitter via an imperfect feedback link. In our proposed signaling, some pilot slots dedicated to fast-fading links also transmit data for users experiencing slower links. Each user must receive and process at least one pilot (per transmit antenna) per fading coherence interval. We introduce an efficient pilot placement strategy that further determines which pilot slots are reused for the data of a given user. We design a joint beamforming and RIS precoding scheme that maximizes the achievable sum-rate through the proposed transmission scheme under imperfect CSI feedback.

The main contributions of this work are as follows:

- We propose a non-orthogonal pilot/data transmission scheme for RIS-assisted downlink SISO (single-input single-output) systems under *coherence disparity*, reducing the training overhead and achieving gain in both degrees of freedom and rate. This scheme introduces an efficient pilot placement strategy and establishes a foundation for utilizing product superposition in RIS systems. The allocated powers for the pilot and data are optimized under this setting.
- We extend the proposed pilot-domain NOMA to multi-antenna transmissions in multi-user RIS systems under general coherence disparities. The technical novelty lies in reconciling the distinct requirements of beamforming (from both the transmitter and RIS) with the proposed non-orthogonal transmission scheme (via product superposition) under *unequal* coherence conditions and imperfect channel state feedback. This is an important outcome because exploiting the efficiencies outlined in this paper also requires the handling of a mismatch in CSI between the transmitter and receiver.
- We jointly optimize transmit beamforming and the RIS reflection coefficients to maximize the achievable sum-rate via the proposed transmission scheme.

A preliminary version of this work [62] was limited to single-antenna transmitters.

## II. Notation and System Model

Matrices and vectors are denoted by bold capital letters and bold small letters, respectively. For a matrix $\mathbf{A}$, the transpose is denoted by $\mathbf{A}^T$, the Hermitian by $\mathbf{A}^H$, the conjugate by $\mathbf{A}^*$. Statistical expectation is denoted by $\mathbb{E}(\cdot)$. $\operatorname{diag}(\mathbf{a})$ denotes a diagonal matrix whose entries are the elements of the vector $\mathbf{a}$. $\operatorname{vec}(\cdot)$ concatenates the columns of a $p \times q$ matrix into a vector of size $pq$. $\operatorname{Re}\{\cdot\}$ denotes the real part of the argument. $\otimes$ denotes the Kronecker product, $\odot$ denotes the element-wise product, and the least common multiple of integers is denoted by $\operatorname{lcm}(\cdot,\cdot)$. The base of the logarithm throughout the paper is 2.

We consider an $M$-antenna transmitter serving $L$ single-antenna receivers. This communication is assisted by a RIS with $N$ passive elements (see Fig. 1). $\mathbf{H} \in \mathbb{C}^{N\times M}$ is the channel matrix between the transmitter and the RIS; $\mathbf{g}_\ell \in \mathbb{C}^{N\times 1}$ and $\mathbf{h}_{d,\ell} \in \mathbb{C}^{M\times 1}$ are the channel vectors from the RIS and the transmitter to User $\ell$, respectively. All channel matrices and vectors have independent identically distributed (i.i.d.) entries $\mathcal{CN}(0,1)$. The system operates under block-fading, where $\mathbf{H}$, $\mathbf{g}_\ell$ and $\mathbf{h}_{d,\ell}$ remain constant over $T$, $T_\ell$ and $T_{d,\ell}$ symbols, respectively, and change independently across blocks. Let $\mathbf{\Theta} = \operatorname{diag}(\theta_1, \theta_2, \cdots, \theta_N)$ denote the complex-valued RIS reflection coefficient matrix with $|\theta_i| \le 1$. The received signal at User $\ell$ is

$$y_\ell = \big(\mathbf{h}_{d,\ell}^T + \mathbf{g}_\ell^T \mathbf{\Theta}\mathbf{H}\big)\mathbf{x} + n_\ell, \qquad \ell = 1, \cdots, L, \qquad (1)$$

where $n_\ell$ is additive noise distributed $\mathcal{CN}(0,\sigma_0^2)$, and $\mathbf{x}$ is the transmit signal which is subject to the average power constraint $\mathbb{E}[\|\mathbf{x}\|^2] \le \rho$ at each time slot.

The transmit signal is

$$\mathbf{x} = \sum_{\ell=1}^{L} \mathbf{w}_\ell s_\ell,$$

where $\mathbf{w}_\ell \in \mathbb{C}^{M\times 1}$ is the beamforming vector for User $\ell$ that is determined at the transmitter through CSI feedback. $s_\ell$ is the data symbol intended for User $\ell$ satisfying $\mathbb{E}[|s_\ell|^2] = 1$.

Let $\mathbf{H}_{c,\ell} \triangleq \mathbf{H}^T \operatorname{diag}(\mathbf{g}_\ell)$ and $\boldsymbol{\theta} \triangleq [\theta_1 \cdots \theta_N]^T$ respectively denote the cascaded channel[1] for User $\ell$ and the $N \times 1$ RIS reflection coefficient vector. The received signal in (1) can be re-written as

$$\begin{aligned} y_\ell &= \big(\mathbf{h}_{d,\ell}^T + \boldsymbol{\theta}^T \mathbf{H}_{c,\ell}^T\big)\mathbf{x} + n_\ell \\ &= \mathbf{x}^T\big(\mathbf{h}_{d,\ell} + \mathbf{H}_{c,\ell}\boldsymbol{\theta}\big) + n_\ell. \end{aligned} \qquad (2)$$

For compact notation, we define the end-to-end channels:

$$\mathbf{H}_\ell \triangleq \big[\mathbf{h}_{d,\ell}\ \mathbf{H}_{c,\ell}\big]$$

and further, for convenience, we produce a vectorized version $\mathbf{h}_\ell \triangleq \operatorname{vec}(\mathbf{H}_\ell)$ and $\widetilde{\boldsymbol{\theta}} \triangleq [1\ \theta_1 \cdots \theta_N]^T$. Then, the received signal in (2) is alternatively expressed as

$$\begin{aligned} y_\ell &= \mathbf{x}^T \mathbf{H}_\ell \widetilde{\boldsymbol{\theta}} + n_\ell \\ &= \Big(\widetilde{\boldsymbol{\theta}}^T \otimes \mathbf{x}^T\Big)\mathbf{h}_\ell + n_\ell \\ &\triangleq \mathbf{u}^T \mathbf{h}_\ell + n_\ell, \end{aligned} \qquad (3)$$

wherein $\mathbf{u} \triangleq \widetilde{\boldsymbol{\theta}} \otimes \mathbf{x}$ represents a vector that contains the transmit vector as well as the RIS reflection coefficients.

The true values of the channel gains $\mathbf{H}_\ell$ are not known a-priori at either the receivers or the transmitter. The values of

[1] Also known as product channel

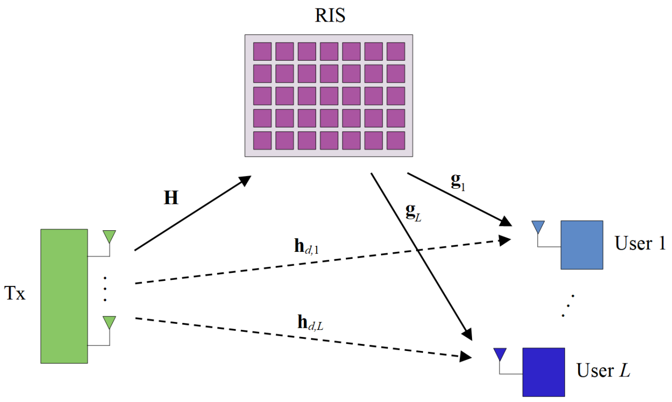

Fig. 1. Channel model

channel gains *estimated* at the receiver are $\overline{\mathbf{H}}_\ell$, and the channel gains estimated at the transmitter (through noisy feedback from the receiver) are represented with $\widehat{\mathbf{H}}_\ell$. The transmitter has the estimated values of the end-to-end channels for both users and uses them to design the transmit beamforming and RIS reflection coefficient vectors. We assume that all the design computations occur at the transmitter according to the channel estimates available at the transmitter, after which the transmitter shares with the RIS controller the computed RIS reflection vector.

## III. Developing the Foundations: Two-User SISO System

We begin by developing the main techniques used in this paper in the simplest possible context, for clarity and ease of exposition. We consider a single-antenna transmitter and an RIS serving two single-antenna receivers, i.e., $M = 1$ and $L = 2$. In this section, we assume for simplicity that no direct propagation path exists between the transmitter and receivers.

The links have unequal coherence times. The link between the transmitter and RIS is stationary, i.e., $T = \infty$, and $T_1 = KT_2$, $K \in \mathbb{Z}$. The integer ratio simplifies the initial presentation of ideas and will be relaxed in Section IV. The Transmitter-RIS link is common for both users. The end-to-end channel $\mathbf{h}_1$ must be updated every $T_1$ time slots, and channel $\mathbf{h}_2$ every $T_2$ time slots. The signaling for both users is designed within a length-$T_1$ block, which as mentioned a moment ago is also an integer multiple of $T_2$.

### A. Transmission Scheme and Channel Estimation

As mentioned earlier, we concentrate on a single time period of length-$T_1$, with all operations repeating every $T_1$ time slots. User 2 has the smaller coherence interval $T_2$, thus it experiences $\frac{T_1}{T_2} = K$ different channel realizations within $T_1$ time slots and needs as many pilot transmissions and channel estimations. During the same $T_1$ time slots, User 1 has a channel that remains unchanged, therefore it only requires one pilot interval. Within $T_1$ time slots, signaling occurs over $K$ sub-blocks, each with length $T_2$. Please refer to the top part of Figure 2 for a depiction of channel states.

Each length-$T_2$ interval consists of a *training phase* that consumes $T_p < T_2$ channel uses and a *data phase* with length $T_2 - T_p$. In Fig. 2, training phases are shown in either gray or light blue, and data phases are shown in green. Each of the channel vectors $\mathbf{h}_1, \mathbf{h}_2$ have $N$ unknowns to be estimated. Therefore, the MMSE estimation requires at least $T_p = N$ pilot slots.[2] The transmitter sends pilot $x_\tau = 1$ over $N$ training slots, during which the RIS states $[\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N]$ are the columns of an $N \times N$ DFT matrix, denoted by $\mathbf{D}_N$:

$$\mathbf{D}_N = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega & \cdots & \omega^{(N-1)} \\ 1 & \omega^2 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \cdots & \omega^{(N-1)(N-1)} \end{bmatrix}. \tag{4}$$

The pre-determined RIS states during the pilot transmission are known as *training states,* and the choice of DFT matrix during this time is called *DFT training.* DFT training can accommodate any RIS size $N$ and is power-efficient but needs accurate implementation of reflection coefficients. The pros and cons of different RIS training sequences are described in [13].

Our scheme admits two types of pilot/training phase. The first kind is employed when both links experience a channel transition, and neither knows its channel. This occurs once every $T_1$ time slots. In this case, the transmitter emits pilots ($x_\tau = 1$) intended for both receivers (shown in gray in Fig. 2). The corresponding received signal at User $\ell$ is

$$\mathbf{y}_\ell = \sqrt{\rho_\tau}\,\mathbf{D}_N^H \mathbf{h}_\ell + \mathbf{n}_\ell, \qquad \ell = 1, 2, \tag{5}$$

where $\rho_\tau$ is the transmit power at each pilot slot and $\mathbf{n}_\ell \in \mathbb{C}^N$ is the additive noise at the receiver with i.i.d. entries $\mathcal{CN}(0, \sigma_0^2)$. The MMSE estimate of $\mathbf{h}_\ell$ is obtained as

$$\begin{aligned} \overline{\mathbf{h}}_\ell &= \mathbb{E}\big[\mathbf{h}_\ell \mathbf{y}_\ell^H\big]\mathbb{E}\big[\mathbf{y}_\ell \mathbf{y}_\ell^H\big]^{-1}\mathbf{y}_\ell \\ &= \sqrt{\rho_\tau}\,\mathbb{E}\big[\mathbf{h}_\ell \mathbf{h}_\ell^H\big]\mathbf{D}_N \Big[\rho_\tau\,\mathbf{D}_N^H \mathbb{E}\big[\mathbf{h}_\ell \mathbf{h}_\ell^H\big]\mathbf{D}_N + \sigma_0^2 \mathbf{I}_N\Big]^{-1}\mathbf{y}_\ell \\ &= \frac{\sqrt{\rho_\tau}\,\mathbf{D}_N \mathbf{y}_\ell}{N\rho_\tau + \sigma_0^2}, \end{aligned} \tag{6}$$

where the facts $\mathbb{E}[\mathbf{h}_\ell \mathbf{h}_\ell^H] = \mathbf{I}_N$ and $\mathbf{D}_N^H \mathbf{D}_N = N\mathbf{I}_N$ are used. The estimation error is denoted by $\mathbf{e}_\ell \triangleq \mathbf{h}_\ell - \overline{\mathbf{h}}_\ell$ which is Gaussian and uncorrelated with the channel estimate $\overline{\mathbf{h}}_\ell$. The covariance of $\mathbf{e}_\ell$ is calculated as

$$\begin{aligned} \mathbb{E}\big[\mathbf{e}_\ell\,\mathbf{e}_\ell^H\big] =& \mathbb{E}\Bigg[\Big(\frac{\sigma_0^2}{N\rho_\tau + \sigma_0^2}\mathbf{h}_\ell - \frac{\sqrt{\rho_\tau}\,\mathbf{D}_N}{N\rho_\tau + \sigma_0^2}\mathbf{n}_\ell\Big) \\ & \Big(\frac{\sigma_0^2}{N\rho_\tau + \sigma_0^2}\mathbf{h}_\ell - \frac{\sqrt{\rho_\tau}\,\mathbf{D}_N}{N\rho_\tau + \sigma_0^2}\mathbf{n}_\ell\Big)^H\Bigg] \\ =& \frac{\sigma_0^2}{N\rho_\tau + \sigma_0^2}\mathbf{I}_N. \end{aligned} \tag{7}$$

The end-to-end channel estimates are fed back to the transmitter to design the RIS reflection coefficient vectors. The system operates under frequency-division duplexing (FDD) mode. We consider an analog feedback scheme [63] in which

[2] Recall that at this point, we are analyzing a system with $M = 1$ transmit antennas and $N$ RIS coefficients.

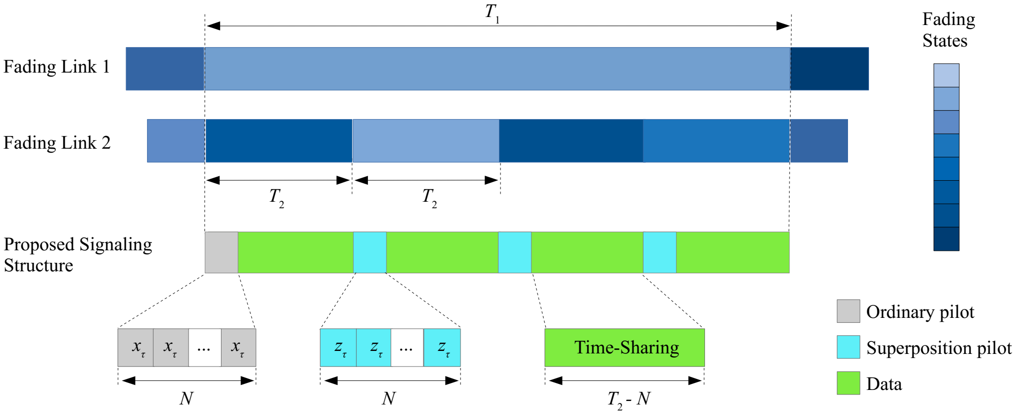

Fig. 2. Link coherence times and the proposed signaling structure. Achievable rates via this scheme are represented by Eqs. (17) and (18), and its full scope is highlighted in Remark 3.

each user transmits on the feedback channel a scaled version of its downlink observation in (5). The received signal at the transmitter is given by

$$\begin{aligned}\mathbf{r}_\ell &= \frac{\sqrt{\rho_\tau}}{\sqrt{\rho_\tau+\sigma_0^2}}\mathbf{y}_\ell + \widetilde{\mathbf{n}}\\ &= \frac{\rho_\tau \mathbf{D}_N^H}{\sqrt{\rho_\tau+\sigma_0^2}}\mathbf{h}_\ell + \frac{\sqrt{\rho_\tau}}{\sqrt{\rho_\tau+\sigma_0^2}}\mathbf{n}_\ell + \widetilde{\mathbf{n}}\\ &= \frac{\rho_\tau \mathbf{D}_N^H}{\sqrt{\rho_\tau+\sigma_0^2}}\mathbf{h}_\ell + \widehat{\mathbf{n}}, \qquad (8)\end{aligned}$$

where $\widetilde{\mathbf{n}} \in \mathbb{C}^N$ denotes the Gaussian noise at the transmitter with i.i.d. entries $\mathcal{CN}(0,\sigma_0^2)$, and is independent of $\mathbf{n}_\ell$. In Equation (8), $\widehat{\mathbf{n}} \triangleq \frac{\sqrt{\rho_\tau}}{\sqrt{\rho_\tau+\sigma_0^2}}\mathbf{n}_\ell + \widetilde{\mathbf{n}}$ denotes the total additive noise at the transmitter, which is Gaussian and its covariance is calculated as

$$\mathbb{E}\big[\widehat{\mathbf{n}}\,\widehat{\mathbf{n}}^H\big] = \sigma_0^2\Big(1+\frac{\rho_\tau}{\rho_\tau+\sigma_0^2}\Big)\mathbf{I}_N. \qquad (9)$$

The transmitter observes $\mathbf{r}_\ell$ and computes the MMSE estimate of $\mathbf{h}_\ell$ as

$$\begin{aligned}\widehat{\mathbf{h}}_\ell &= \mathbb{E}\big[\mathbf{h}_\ell\mathbf{r}_\ell^H\big]\mathbb{E}\big[\mathbf{r}_\ell\mathbf{r}_\ell^H\big]^{-1}\mathbf{r}_\ell\\ &= \frac{\rho_\tau\sqrt{\rho_\tau+\sigma_0^2}\,\mathbf{D}_N}{N\rho_\tau^2+\sigma_0^2(2\rho_\tau+\sigma_0^2)}\mathbf{r}_\ell. \qquad (10)\end{aligned}$$

The estimation error is defined as $\widehat{\mathbf{e}}_\ell \triangleq \mathbf{h}_\ell - \widehat{\mathbf{h}}_\ell$ with covariance

$$\mathbb{E}\big[\widehat{\mathbf{e}}_\ell\,\widehat{\mathbf{e}}_\ell^H\big] = \frac{\sigma_0^2(2\rho_\tau+\sigma_0^2)}{N\rho_\tau^2+\sigma_0^2(2\rho_\tau+\sigma_0^2)}\mathbf{I}_N. \qquad (11)$$

A different kind of pilot/training phase occurs when User 2 experiences a channel transition while User 1 does not (see Figure 2). This happens in exactly $K-1$ instances during each $T_1$ time slots. In these instances, User 2 needs channel estimation, while User 1 does not. Therefore, the corresponding pilot slots (shown in light blue in Figure 2) are reused to carry data for User 1. We propose an efficient non-orthogonal (superposition) transmission of pilot and data over these training phases, each with length $N$. In each of these pilot slot, the transmit signal is

$$z_\tau = \sqrt{\rho_\tau}\,x_1 x_\tau, \qquad (12)$$

where $x_\tau$ is the pilot and $x_1$ is the intended symbol for User 1, such that $\mathbb{E}[|x_1|^2] = 1$. This signaling is repeated $N$ times, covering the entire training phase. The corresponding received signal at User $\ell$ is

$$\begin{aligned}\mathbf{y}'_\ell &= z_\tau\mathbf{D}_N^H\mathbf{h}_\ell + \mathbf{n}_\ell\\ &= \sqrt{\rho_\tau}\,\mathbf{D}_N^H\mathbf{h}_\ell\,x_1x_\tau + \mathbf{n}_\ell, \qquad \ell = 1,2. \qquad (13)\end{aligned}$$

Clearly, the main difference of Equation (13) compared with Equation (5) is in $z_\tau$, a signal component designed to carry data for User 1 and act as a pilot for User 2. User 1 receives $\mathbf{y}'_1$ over the pilot slots and since its channel is remaining unchanged since the previous pilot slot, it attempts to decode $x_1$, resulting in gains in rate and degrees of freedom. User 2 has a channel that is transitioned to a new value since its last channel estimation. Therefore, it estimates its virtual channel $\mathbf{f}_2 \triangleq \mathbf{h}_2 x_1$ and feeds it back to the transmitter. The MMSE estimate of the virtual channel at the receiver is

$$\begin{aligned}\overline{\mathbf{f}}_2 &= \mathbb{E}\big[\mathbf{f}_2\,\mathbf{y}_2'^H\big]\mathbb{E}\big[\mathbf{y}'_2\,\mathbf{y}_2'^H\big]^{-1}\mathbf{y}'_2\\ &= \frac{\sqrt{\rho_\tau}\,\mathbf{D}_N\mathbf{y}'_2}{N\rho_\tau+\sigma_0^2}. \qquad (14)\end{aligned}$$

Channel state feedback follows a process similar to Equation (8), except the feedback value is the virtual channel $\mathbf{f}_2$ and not the physical link gain $\mathbf{h}_2$. The transmitter calculates an estimate of the true channel gain $\widehat{\mathbf{h}}_2$ from the noisy feedback version of the virtual channel. This is possible because the transmitter knows $x_1$. The estimate $\widehat{\mathbf{h}}_2$ is calculated via:

$$\widehat{\mathbf{h}}_2 = \frac{\rho_\tau\sqrt{\rho_\tau+\sigma_0^2}\,\mathbf{D}_N}{N\rho_\tau^2+\sigma_0^2(2\rho_\tau+\sigma_0^2)}\Big(\frac{\rho_\tau\mathbf{D}_N^H}{\sqrt{\rho_\tau+\sigma_0^2}}\mathbf{h}_2 + \widehat{\mathbf{n}}\,x_1^{-1}\Big). \qquad (15)$$

Let $\widehat{\mathbf{e}}_2 \triangleq \mathbf{h}_2 - \widehat{\mathbf{h}}_2$ denote the channel estimation error. The covariance of $\widehat{\mathbf{e}}_2$ is calculated as

$$\mathbb{E}\big[\widehat{\mathbf{e}}_2\,\widehat{\mathbf{e}}_2^H\big] = \frac{(\sigma_0^2(2\rho_\tau+\sigma_0^2))^2+\rho_\tau^3+\rho_\tau^2\sigma_0^2}{(N\rho_\tau^2+\sigma_0^2(2\rho_\tau+\sigma_0^2))^2}\mathbf{I}_N. \qquad (16)$$

*Remark 1:* Channel knowledge for User 1 in some sub-blocks arises from the differences in coherence times of the two links. This happens because the same sub-block size is used for analyzing both users, but one of them has a longer coherence interval (User 1). As long as the next channel transition has not occurred for User 1, it maintains the same channel gain. In other words, after every channel gain transition, there are $K-1$ sub-blocks in which the channel gain knowledge for User 1 is inherited from the past.

*Remark 2:* In the superposition pilot slots given in Equation (12), when the fast-fading User 2 is estimating and using the channel, the slow-fading User 1 repeats its data transmission. Thus, the value of $\mathbf{f}_2 \triangleq \mathbf{h}_2 x_1$ remains unchanged from the time it is estimated to the time it is used. Consequently, the product in Equation (12) does not change User 2's coherence interval and pilot duty cycle. User 2's coherence interval is $T_2$, requiring pilots every $T_2$ time slots.

### B. Achievable Rates

We outline an accounting of time slots needed for rate calculations over $T_1$ time slots (see Fig. 2). Recall that $\frac{T_1}{T_2} = K$ therefore we have $K$ sub-blocks. During the interval of length $T_1$, we need $KN$ pilot slots. $N$ pilots are needed at the beginning of the $T_1$-length interval to estimate the channel gains of both users. The remaining $(K-1)N$ pilot slots are only needed for estimating the channel of User 2, therefore, they will also carry data to User 1. The users employ time-sharing with ratios $\eta_\ell \in [0,1]$ during the data phases, each with length $T_2 - N$. During each data phase, passive beamforming from RIS to all users occurs. With this, the following rates per channel use are achieved:

$$R_1 = \Big(1 - \frac{1}{K}\Big)\frac{N}{T_2}\log(1+\gamma_1') + \eta_1\Big(1 - \frac{N}{T_2}\Big)\log(1+\gamma_1), \tag{17}$$

$$R_2 = \eta_2\Big(1 - \frac{N}{T_2}\Big)\log(1+\gamma_2), \tag{18}$$

where $\gamma_1'$ denotes the signal-to-noise ratio (SNR) at User 1 over the superposition pilot slots. $\gamma_1$ and $\gamma_2$ respectively denote the SNR at User 1 and User 2 over the data slots. The first term in Equation (17) is the achieved rate for User 1 over $K-1$ training phases, each with length $N$. The second term in Equation (17) is the achieved rate for User 1 over $K$ data phases, each with length $T_2 - N$. The rate for User 2 in Equation (18) has one term, because it only receives data over $K(T_2 - N)$ data slots.

*Remark 3:* The rate expressions in (17) and (18) essentially represent a corner point within the rate region for a fixed $\eta_\ell$. The system can also operate in single-user mode, in which pilots are only necessary if the channel state of the active user requires updating. The overall rate region consists of the convex hull of the rates under superposition (17), (18), and the single-user rates.

*Remark 4:* In our rate derivations, the time-sharing variables $\eta_\ell$ are "knobs" that allow the boundary of the rate-region to be traversed. These parameters allow one user's rate to be larger or smaller at the expense of other users. The setting of time-sharing variables depends on the rates requested by different users in the system and corresponds to multi-user multiplexing. The optimization of these factors has been extensively discussed in the literature, employing various techniques. In particular, it has been shown that in time-sharing signaling, the maximum sum-rate is achieved by allocating transmission time to the user with the largest channel gain [64]. In the interest of brevity, these well-studied issues are not repeated in this paper.

*Remark 5:* The achievable rates in Eqs. (17) and (18) are subject to the optimization of RIS coefficients. This part is omitted in the present section because in a two-user SISO scenario, the RIS coefficient optimization reduces to the single-user RIS scenario, which is adequately covered in the literature [13]. The joint Transmitter-RIS beamforming optimization is thoroughly addressed in Section IV-C.

### C. Power Allocation

Regardless of whether ordinary or superposition pilots are transmitted, the transmit power at each time slot during the pilot phases is $\rho_\tau$. Let $\rho_d$ denote the transmit power at each time slot during the data phases. Given that $\mathbb{E}[|x_1|^2] = 1$ and $\eta_1 + \eta_2 = 1$, in all $K$ sub-blocks, each with length $T_2$, the power constraint $\rho$ is satisfied, with:

$$\rho_\tau N + \rho_d(T_2 - N) = \rho T_2. \tag{19}$$

The effective SNR for User 1 and User 2 depends on $\rho_\tau$ and $\rho_d$. The proposed transmission scheme employs pilot/data superposition over $K-1$ pilot phases for carrying the data of User 1. The transmit signal in these pilot phases is given by Equation (12), with the corresponding received values expressed in Equation (13). From Equation (17), $\gamma_1'$ is the average SNR for User 1 at each superposition pilot slot. To calculate $\gamma_1'$, we re-express Equation (13) for User 1 in terms of the channel estimate and estimation error:

$$\begin{aligned}\mathbf{y}_1' &= \sqrt{\rho_\tau}\,\mathbf{D}_N^H\,\widehat{\mathbf{h}}_1 x_1 x_\tau + \sqrt{\rho_\tau}\,\mathbf{D}_N^H\,\widehat{\mathbf{e}}_1 x_1 x_\tau + \mathbf{n}_1 \\ &= \sqrt{\rho_\tau}\,\mathbf{D}_N^H\,\widehat{\mathbf{h}}_1 x_1 x_\tau + \mathbf{n}_1',\end{aligned} \tag{20}$$

where $\mathbf{n}_1' \triangleq \sqrt{\rho_\tau}\,\mathbf{D}_N^H\,\widehat{\mathbf{e}}_1 x_1 x_\tau + \mathbf{n}_1$ denotes the sum of additive noise and the residual channel estimation error. Substituting (10) in (20) and utilizing the covariance of channel estimation error in (11), the effective SNR $\gamma_1'$ is obtained:

$$\gamma_1' = \frac{N^2\rho_\tau^3}{\sigma_0^2(2\rho_\tau + \sigma_0^2)}. \tag{21}$$

As the proposed signaling offers User 1 (slow-fading link) the opportunity to receive additional data over the pilot phases, we focus on $\rho_\tau$ and $\rho_d$ that maximize the effective SNR of User 2 (fast-fading link) over the data phases. The received signal at User 2 in each time slot of the data phase is

$$\begin{aligned}y_2' &= \sqrt{\rho_d}\,\widehat{\mathbf{h}}_2^H\boldsymbol{\theta}x + \sqrt{\rho_d}\,\widehat{\mathbf{e}}_2^H\boldsymbol{\theta}x + n_2 \\ &= \sqrt{\rho_d}\,\widehat{\mathbf{h}}_2^H\boldsymbol{\theta}x + n_2',\end{aligned} \tag{22}$$

where $n_2' \triangleq \sqrt{\rho_d}\,\widehat{\mathbf{e}}_2^H\boldsymbol{\theta}x + n_2$ denotes the sum of receiver noise and residual channel estimation error. The choice of RIS states that maximizes the desired signal power in Equation (22), is $|\theta_i| = 1$ and $\angle\theta_i = -\angle h(i)$, $i = 1, \cdots, N$. Utilizing this and the channel estimate in Equation (15), the power of desired signal in Equation (22) is calculated, denoted as $\sigma_p^2$:

$$\begin{aligned}\sigma_p^2 &= \rho_d\,\mathbb{E}\big[\boldsymbol{\theta}^H \widehat{\mathbf{h}}_2 \widehat{\mathbf{h}}_2^H \boldsymbol{\theta}\big] \\ &= N\rho_d\bigg(\Big(\frac{N\rho_\tau^2}{N\rho_\tau^2+\sigma_0^2(2\rho_\tau+\sigma_0^2)}\Big)^2 + \\ &\qquad \Big(\frac{N\rho_\tau^2\mathbb{E}[|1/x|^2]}{(N\rho_\tau^2+\sigma_0^2(2\rho_\tau+\sigma_0^2))^2}\Big)\bigg). \qquad (23)\end{aligned}$$

Using (16), we calculate

$$\begin{aligned}\mathbb{E}\big[n_2' n_2'^H\big] &= \rho_d\,\boldsymbol{\theta}^H \mathbb{E}\big[\widehat{\mathbf{e}}_2\widehat{\mathbf{e}}_2^H\big]\boldsymbol{\theta} + 1 \\ &= N\rho_d\frac{(\sigma_0^2(2\rho_\tau+\sigma_0^2))^2+\rho_\tau^3+\rho_\tau^2\sigma_0^2}{(N\rho_\tau^2+\sigma_0^2(2\rho_\tau+\sigma_0^2))^2}+1. \qquad (24)\end{aligned}$$

The effective SNR $\gamma_2$ is then obtained as Equation (25), shown at the bottom of this page.

Let $\delta$ denote the fraction of total power allocated for data. That is, $\rho_d(T_2 - N) = \delta\rho T_2$ and $\rho_\tau N = (1-\delta)\rho T_2$. We replace $\rho_d = \delta\rho\frac{T_2}{T_2-N}$ and $\rho_\tau = (1-\delta)\rho\frac{T_2}{N}$ in (25) and set the derivative of $\gamma_2$ with respect to $\delta$ to zero. This results in the optimal value of $\delta$:

$$\delta^* = \frac{\sqrt{(\sigma_0^2+\frac{\rho T_2}{N})(\sigma_0^2+\frac{N\rho T_2}{T_2-N})}-(\sigma_0^2+\frac{\rho T_2}{N})}{\sigma_0^2(\frac{N\rho T_2}{T_2-N}-\frac{\rho T_2}{N})}. \qquad (26)$$

The optimal value, denoted as $\delta^*$, for pilot and data power allocation in the first length-$T_2$ sub-block, where ordinary pilots are transmitted, follows the same analysis and hence is omitted here for brevity. The only difference lies in the fact that it utilizes the estimate of the true channel instead of the estimate of the virtual channel.

*Remark 6:* The power allocations discussed in this section were obtained for a single-antenna transmitter configuration. In the context of a multi-antenna setup, we will later develop a joint transmit beamforming and RIS reflection coefficient scheme, which will have a different power allocation outcome.

## IV. Multi-User MIMO RIS-Assisted Pilot-Domain NOMA

We now leverage the key insights obtained in Section III to develop a comprehensive approach for pilot domain NOMA transmission in a multi-user MIMO RIS-assisted system. We begin by characterizing the signaling and NOMA structure in a multi-user scenario, resulting in achievable rate expressions as a function of transmitter and RIS beamforming vectors. Then, we use a sum-rate criterion to find the optimal transmitter and RIS beamforming. Our optimization approach is easily generalizable to any (convex) weighted sum of rates, therefore the entire multi-user achievable rate region is attainable using this technique.

### A. Multi-User MIMO Link and Signaling Schemes

Consider a multi-user downlink system with $L$ receivers, assisted by an RIS, as outlined in Section II. Each user receives signals through both direct and cascaded (RIS) paths. Our focus is on the *coherence disparity* condition, where the coherence times for RIS-Receiver and Transmitter-Receiver links vary among users. Specifically, this section analyzes the interesting case where for some User(s) $\ell \in \{1, \cdots, L\}$, at least two of the three quantities $T$, $T_\ell$, and $T_{d,\ell}$ are non-identical. Due to variations in node mobility and scattering environment, such a broad condition can easily occur. We define $T_\ell' \triangleq \min\{T_\ell, T_{d,\ell}\}$. Since the Transmitter-RIS link is shared across all end-to-end channels, when $T \le \min\{T_1', \cdots, T_L'\}$, the common Transmitter-RIS link creates a channel estimation bottleneck, and no additional degrees of freedom can be achieved through the *coherence diversity*. However, when $T > \min\{T_1', \cdots, T_L'\}$, the required pilot duty cycle for different users is not identical, creating opportunities for pilot-domain NOMA.

Without loss of generality, the link coherence times $T_\ell'$ are indexed in descending order:

$$T_1' > \cdots > T_L'.$$

This leads to $T > T_L'$, indicating that User $L$ needs to estimate its channel $\mathbf{h}_L$ every $T_L'$ time slots. To have a general notation for all end-to-end link coherence times, we define $T_\ell^* \triangleq \min\{T, T_\ell'\}$. Thus, any User $\ell$ requires refreshing its channel estimates every $T_\ell^*$ time slots. Since the links have unequal coherence intervals, we design the transmission scheme over a super interval whose length is the least common multiple (LCM) of the coherence times for all end-to-end links (see Fig. 3).

$$T_c \triangleq \mathrm{lcm}(T_1^*, \cdots, T_L^*)$$

We concentrate on a single time period of length $T_c$ for signaling, with all operations repeating every $T_c$ time slots. User $L$ has the fastest link (smallest coherence time $T_L^*$), therefore at least $\frac{T_c}{T_L^*}$ pilot intervals are required within $T_c$. User 1, the slowest user, needs $\frac{T_c}{T_1^*}$ pilot intervals during $T_c$. Other users fall somewhere in between. The signaling is designed over $\frac{T_c}{T_L^*}$ sub-blocks, each with length $T_L^*$. Every length-$T_L^*$ sub-block consists of *training phase* with $T_p \le T_L^*$ time slots and *data phase* with $T_L^* - T_p$ time slots. The channel vector $\mathbf{h}_\ell$ has $M(N+1)$ unknowns, whose estimation requires $T_p = M(N+1)$ observations in each training phase (see Fig. 3). Each RIS training state $\boldsymbol{\theta}$ is maintained for $M$ time intervals during which the transmitter emits linearly independent orthogonal pilots. This is repeated $N+1$ times with different RIS training states. Each group of $M$ pilot transmissions is called a *training sub-phase*; the groups are indexed with the variable $\tau = 1, \ldots, N+1$.

Let $\mathbf{P}_\tau \in \mathbb{C}^{M\times M}$ denote the pilot matrix over the training sub-phase $\tau$, which is *unitary* and known to both users. The

$$\gamma_2 = \frac{N^2\rho_\tau^2\rho_d(N\rho_\tau^2+\mathbb{E}[|1/x|^2])}{N\rho_d((\sigma_0^2(2\rho_\tau+\sigma_0^2))^2+\rho_\tau^3+\rho_\tau^2\sigma_0^2)+(N\rho_\tau^2+\sigma_0^2(2\rho_\tau+\sigma_0^2))^2} \qquad (25)$$

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3524943

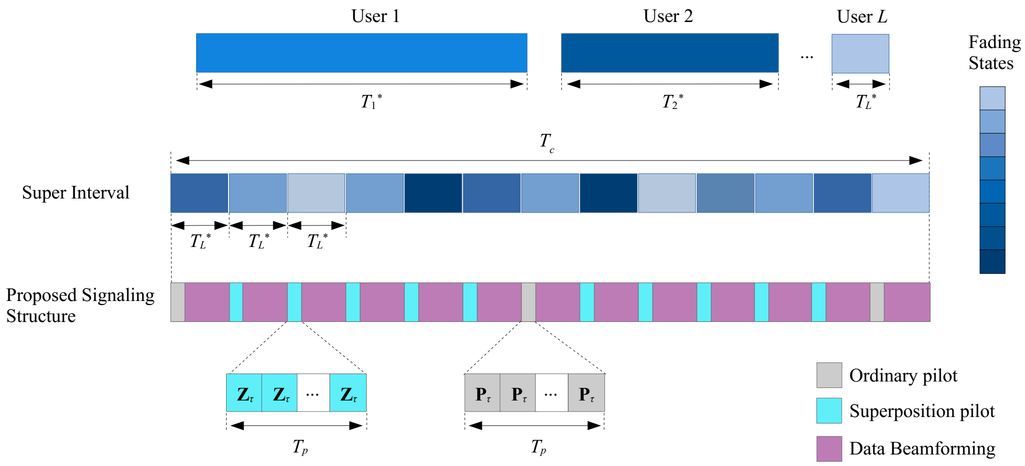

Fig. 3. Link coherence times, super interval, and the proposed signaling structure within the super interval.

corresponding received signal at each user during $M$ time slots is

$$\mathbf{y}_{\ell,\tau} = \sqrt{\rho}\,\mathbf{U}_\tau\,\mathbf{h}_\ell + \mathbf{n}_\ell, \tag{27}$$

where $\mathbf{U}_\tau \triangleq \widetilde{\boldsymbol{\theta}}_\tau^T \otimes \mathbf{P}_\tau^T$ represents the combined action of pilots and RIS training states on the channel coefficients $\mathbf{h}_\ell$ during training sub-phase $\tau$. $\mathbf{n}_\ell \in \mathbb{C}^M$ is the additive Gaussian noise with i.i.d. entries $\mathcal{CN}(0,\sigma_0^2)$. This process is repeated $N+1$ times to estimate all unknown channels. Let

$$\widetilde{\mathbf{y}}_\ell \triangleq \begin{bmatrix} \mathbf{y}_{\ell,1} \\ \vdots \\ \mathbf{y}_{\ell,N+1} \end{bmatrix} \text{ and } \widetilde{\mathbf{U}} \triangleq \begin{bmatrix} \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_{N+1} \end{bmatrix}.$$

The MMSE estimate of $\mathbf{h}_\ell$ is denoted by $\bar{\mathbf{h}}_\ell$:

$$\bar{\mathbf{h}}_\ell = \sqrt{\rho}\,\mathbb{E}\big[\mathbf{h}_\ell\mathbf{h}_\ell^H\big]\widetilde{\mathbf{U}}^H\Big[\rho\,\widetilde{\mathbf{U}}\mathbb{E}\big[\mathbf{h}_\ell\mathbf{h}_\ell^H\big]\widetilde{\mathbf{U}}^H + \mathbf{I}\Big]^{-1}\widetilde{\mathbf{y}}_\ell. \tag{28}$$

The estimated channel depends on the RIS training states. We consider the DFT training as discussed in Equation (6). This leads to

$$\bar{\mathbf{h}}_\ell = \frac{\sqrt{\rho}}{1+\rho T_p}\widetilde{\mathbf{U}}^H\widetilde{\mathbf{y}}_\ell. \tag{29}$$

The channel estimates are fed back to the transmitter to calculate the transmit beamforming and RIS reflection coefficients (see Fig. 4). We consider an analog feedback scheme [63] over an AWGN feedback channel with power constraint $\rho$.[3] Each user feeds back its downlink observation from Eq. (27), scaled appropriately to satisfy the feedback channel power constraint. The output of the feedback channel is:

$$\mathbf{r}_\ell = \frac{\sqrt{\rho}}{\sqrt{\rho+\sigma_0^2}}\widetilde{\mathbf{y}}_\ell + \widetilde{\mathbf{n}}, \tag{30}$$

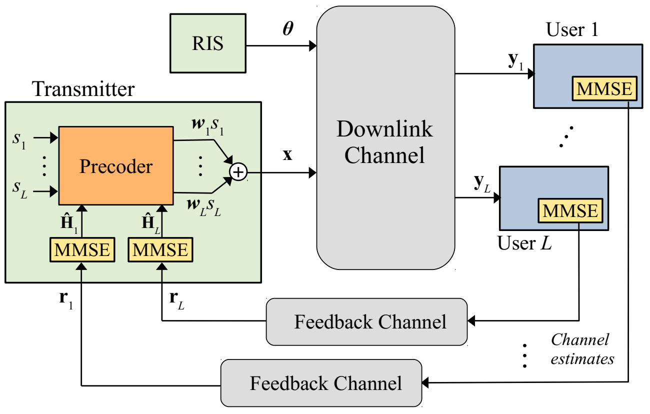

Fig. 4. Channel state feedback

where $\widetilde{\mathbf{n}} \in \mathbb{C}^{T_p}$ is AWGN with power $\sigma_0^2$ per entry. The transmitter observes $\mathbf{r}_\ell$ and computes the MMSE estimate of the channel $\mathbf{h}_\ell$ as

$$\widehat{\mathbf{h}}_\ell = \frac{\rho}{(\rho+\sigma_0^2)\sqrt{\rho+\sigma_0^2}}\mathbf{r}_\ell. \tag{31}$$

In the remaining $\frac{T_c}{T_L^*} - \frac{T_c}{T_1^*}$ sub-blocks, each of length $T_L^*$, User 1 does not require further pilots for channel estimation. However, any other User $\ell \neq 1$ might need to refresh its channel estimate. Therefore, the pilot slots in the following sub-blocks may be used to transmit data for users whose channel states remain unchanged. To this end, we propose non-orthogonal pilot/data transmissions during all such training phases.

During a training sub-phase $\tau$, assume a User $i$ does not need the pilot because its channel has not changed since the last channel estimate was formed, and is eligible to receive data represented by the signal matrix $\mathbf{X}_i \in \mathbb{C}^{M\times M}$. The transmitter will emit:

$$\mathbf{Z}_\tau = \sqrt{\rho}\,\mathbf{X}_i\mathbf{P}_\tau, \tag{32}$$

[3] As in prior works [63], [65], in our model the feedback link for each user operates through a "side channel" that does not interfere or interact with the feedback link of other users, or with the forward link.

$\mathbf{P}_\tau \in \mathbb{C}^{M\times M}$ is the pilot matrix. The corresponding received signal at all users will be

$$\begin{aligned}\check{\mathbf{y}}_{\ell,\tau} &= \sqrt{\rho}\,\mathbf{Z}_\tau^T\,\mathbf{H}_\ell\widetilde{\boldsymbol{\theta}}_\tau + \mathbf{n}_\ell \\ &= \sqrt{\rho}\,\mathbf{P}_\tau^T\,\mathbf{X}_i^T\,\mathbf{H}_\ell\widetilde{\boldsymbol{\theta}}_\tau + \mathbf{n}_\ell, \qquad \ell = 1,\cdots,L. \end{aligned} \tag{33}$$

Recall that User $\ell = i$ already knows $\mathbf{H}_i, \mathbf{P}_\tau, \widetilde{\boldsymbol{\theta}}_\tau$, therefore it can coherently decode $\mathbf{X}_i$. Any other User $\ell \neq i$ *whose channel has experienced a transition*[4] will try to estimate the virtual channel $\mathbf{F}_\ell \triangleq \mathbf{X}_i^T\,\mathbf{H}_\ell$ and feed it back to the transmitter. The transmitter knows $\mathbf{X}_i$, therefore it can estimate the true channel $\mathbf{H}_\ell$ and use it for data beamforming.

*Remark 7:* In the multi-user superposition pilot slots described in Equation (32), when User $\ell \neq i$, with a varying channel, estimates and utilizes the channel, User $i$ continues its data transmission $\mathbf{X}_i$. As a result, $\mathbf{F}_\ell$ remains constant for User $\ell$ from estimation to usage. Consequently, the product in Equation (32) does not impact User $\ell$'s coherence interval or pilot duty cycle. The end-to-end coherence interval for User $\ell$ remains $T_\ell^*$, requiring pilot transmissions every $T_\ell^*$ time slots.

Let $\mathbf{f}_\ell \triangleq \mathrm{vec}(\mathbf{F}_\ell)$. Then, Equation (33) is alternatively expressed as

$$\begin{aligned}\check{\mathbf{y}}_{\ell,\tau} &= \sqrt{\rho}\left(\widetilde{\boldsymbol{\theta}}_\tau^T \otimes \mathbf{P}_\tau^T\right)\mathbf{f}_\ell + \mathbf{n}_\ell \\ &= \sqrt{\rho}\,\mathbf{U}_\tau\,\mathbf{f}_\ell + \mathbf{n}_\ell, \end{aligned} \tag{34}$$

where $\mathbf{U}_\tau$ is the overall training matrix at the training sub-phase $\tau$ and is defined in Equation (27). The transmitter sends $\mathbf{Z}_\tau$ in $N+1$ sub-phases with the corresponding receives values $\check{\mathbf{y}}_{\ell,\tau}$ at each sub-phase. Let

$$\check{\mathbf{y}}_\ell \triangleq \begin{bmatrix}\check{\mathbf{y}}_{\ell,1} \\ \vdots \\ \check{\mathbf{y}}_{\ell,N+1}\end{bmatrix}.$$

Through the same manners used in Equations (28) and (29), the MMSE estimate of $\mathbf{f}_\ell$ is obtained as

$$\bar{\mathbf{f}}_\ell = \frac{\sqrt{\rho}}{1+\rho T_p}\widetilde{\mathbf{U}}^H\check{\mathbf{y}}_\ell. \tag{35}$$

Channel state feedback follows similar concepts as that of Equation (30) with the only difference being that the virtual channel is returned to the transmitter, not the true one. The transmitter first estimates the virtual channel $\mathbf{f}_\ell$ through the same manner as (31) and then computes the estimate of the true channel $\mathbf{h}_\ell$ noting that it has the full knowledge of $\mathbf{X}_\ell$. Let $\widehat{\mathbf{f}}_\ell \triangleq \mathrm{vec}(\widehat{\mathbf{F}}_\ell)$ denote the estimate of $\mathbf{f}_i$ at the transmitter. Then, the estimate of the true channel $\mathbf{H}_\ell$ is denoted by $\widehat{\mathbf{H}}_\ell = (\mathbf{X}_i^T)^{-1}\widehat{\mathbf{F}}_\ell$.

### B. Achievable Rates

We now outline an accounting of time slots needed for rate calculations over $T_c$ time slots. To estimate all the channels within the super interval, we need $\frac{T_c}{T_1^*}M(N+1)$ pilot slots. The proposed scheme allocates $\frac{T_c}{T_{\ell+1}^*} - \frac{T_c}{T_\ell^*}$ length-$T_p$ pilot intervals to transmit the message of User $\ell \neq L$ over the pilot phases. User $L$ with the most rapidly varying link requires all available $\frac{T_c}{T_L^*}$ pilot intervals for channel estimation. Beamforming to all users occurs during the $\frac{T_c}{T_L^*}$ data phases, each with length $T_L^* - M(N+1)$. With this, the following rates per channel use are achieved:

$$\begin{aligned} R_\ell = &\Big(\frac{1}{T_L^*} - \frac{1}{T_1^*}\Big)M(N+1)\log(1+\gamma_\ell{}') \\ &+ \Big(1 - \frac{M}{T_L^*}(N+1)\Big)\log(1+\gamma_\ell) \quad \ell = 1,\cdots,L-1, \end{aligned} \tag{36}$$

$$R_L = \Big(1 - \frac{M}{T_L^*}(N+1)\Big)\log(1+\gamma_L), \tag{37}$$

where $\gamma_\ell'$ denotes the SNR at User $\ell$ over the reused pilot slots. $\gamma_\ell$ and $\gamma_L$ respectively denote the signal-to-interference-plus-noise ratio (SINR) at User $\ell$ and User $L$ over the data slots. The first term in (36) is the achieved rate for User $\ell$ over $\frac{T_c}{T_L^*} - \frac{T_c}{T_1^*}$ training phases, each with length $M(N+1)$. The second term in (36) is the achieved rate for User $\ell$ over $\frac{T_c}{T_L^*}$ data phases, each with length $T_L^* - M(N+1)$. The rate for User $L$ in (37) has only one term, because it only receives data over $\frac{T_c}{T_L^*}(T_L^* - M(N+1))$ data slots.

*Remark 8:* Equations (36) and (37) together describe individual rates for all users. By varying the allocated powers subject to the overall power constraint $\rho$, the Equations (36) and (37) yield the overall rate region.

*Remark 9:* The proposed method can be generalized to the case where there are more users than transmit antennas. When $L > M$, we can choose groups of $M$ users, each, so that the users within each group will not interfere with each other. These groups will be scheduled via a scheduling algorithm through time-sharing, or will be able to utilize a pilot slot for data transmission, whenever the group does not need the channel state information provided by that pilot. The details are straightforward but tedious to enumerate case-by-case, and therefore are omitted for brevity.

*Remark 10:* Unlike other NOMA techniques, such as power-domain NOMA, the performance of the proposed pilot-domain NOMA is not directly dependent on the number of users but rather on the level of coherence disparities among them. Different coherence disparity regimes yield varying performance gains. Since pilot-domain NOMA inherently relies on the coherence lengths of all users, performance cannot be evaluated independently of these coherence conditions. Consequently, our analysis differs from that typically found in conventional NOMA literature. This distinction stems from fundamental differences between the two methods, their performance gains, and the specific conditions under which these gains are achieved. Therefore, while it is straightforward to present rate performance as a function of user count for power-domain NOMA, this approach is less applicable to pilot-domain NOMA.

*Remark 11:* The pilot-domain NOMA for RIS-assisted systems presented in this paper exploits a very different source of gains compared with conventional (power-domain) NOMA whose gain often comes with the requirement of interference-

[4] There may be some users $\ell \neq i$ whose channel has *not* experienced a transition since their last channel estimate. Any such users will ignore this training interval.

canceling (peeling) receivers. Pilot-domain NOMA sends data during pilots that are unnecessary for some users. Because of the different nature of the gains in pilot-domain NOMA, and the neat structure of product superposition, this gain can be harvested without the need to explicitly perform interference cancellation. These pilot-domain and power-domain NOMA gains are *not* mutually exclusive; in fact, they are independent of each other. Since power-domain NOMA has been explored in numerous other works, and its combination with pilot-domain NOMA does not produce new points of interest, its discussion is omitted in this paper.

### *C. Beamforming Optimization*

The joint design of beamforming for the transmitter and RIS occurs centrally (presumably at the base station) and is communicated with the RIS through a backhaul link. The objective is to maximize the achievable sum-rate.[5] We describe the optimization algorithm in the blocks where superposition pilots are employed. The optimization in non-superposition blocks is essentially similar with a change of variable, and is therefore omitted for brevity.

The channel estimates $\widehat{\mathbf{H}}_\ell$ are available at the transmitter. The transmitter calculates $\gamma_\ell$ as

$$\gamma_\ell = \frac{|\widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_\ell|^2}{|\widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_j|^2_{j\neq\ell} + \rho\sigma_{e,\ell}^2 + \sigma_0^2}, \tag{38}$$

where $\sigma_{e,\ell}^2$ denotes the power of channel estimation error that can be obtained by calculating the covariance of the channel estimation error $\mathbf{e}_\ell = \mathbf{h}_\ell - \widehat{\mathbf{h}}_\ell$ [66]:

$$\mathbb{E}\big[\mathbf{e}_\ell\, \mathbf{e}_\ell^H\big] = \sigma_{e,\ell}^2 \mathbf{I}.$$

The original sum-rate maximization problem is

$$\max_{\mathbf{W},\boldsymbol{\theta}} \quad \sum_{\ell=1}^{L} R_\ell \tag{39a}$$

$$\text{s.t.} \quad \sum_{\ell=1}^{L} \|\mathbf{w}_\ell\|^2 \leq \rho \tag{39b}$$

$$|\theta_i| \leq 1,\ \forall i = 1, \cdots, N, \tag{39c}$$

where $R_\ell$ is given in Equations (36) and (37), and $\mathbf{W} \triangleq [\mathbf{w}_1, \cdots, \mathbf{w}_L]$ denotes the beamforming matrix.

The beamforming for all users is performed during the data phase of the block. In the sum-rate $\sum_{\ell=1}^{L} R_\ell$, the term $\log(1+\gamma_\ell')$ refers to the data rate transmitted in superposition during the RIS *training states*, therefore it is independent of the subsequent *beamforming states* $\mathbf{W}, \boldsymbol{\theta}$, and does not participate in their optimization. The remaining terms in the sum-rate are $\big(1 - \frac{M}{T_L^*}(N+1)\big)\sum_{\ell=1}^{L}\log(1+\gamma_\ell)$, in which the multiplying constant upfront involves variables that are also independent of optimization variables $\mathbf{W}, \boldsymbol{\theta}$. Therefore, the sum-rate optimization is reduced to:

$$\max_{\mathbf{W},\boldsymbol{\theta}} \quad \sum_{\ell=1}^{L} \log\bigg(1 + \frac{|\widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_\ell|^2}{|\widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_j|^2_{j\neq\ell} + \rho\sigma_{e,\ell}^2 + \sigma_0^2}\bigg) \tag{40a}$$

$$\text{s.t.} \quad \sum_{\ell=1}^{L} \|\mathbf{w}_\ell\|^2 \leq \rho \tag{40b}$$

$$|\theta_i| \leq 1,\ \forall i = 1, \cdots, N. \tag{40c}$$

To solve this non-convex problem, we utilize fractional programming [67], [68]. Incorporating both Lagrangian and Quadratic transforms, this method transforms the cost function, which is a sum of logarithms of ratios, into a sum of logarithmic and linear functions. This process introduces two sets of auxiliary variables: $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_L]$ and $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_L]$, corresponding to the Lagrangian dual transform and the Quadratic transform, respectively. The optimization in (40) is equivalently reformulated as:

$$\begin{aligned}\max_{\mathbf{W},\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \sum_{\ell=1}^{L} \log(1+\alpha_\ell) - \alpha_\ell \\ & + \sum_{\ell=1}^{L} 2\sqrt{1+\alpha_\ell}\,\mathrm{Re}\{\beta_\ell^*\ \widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_\ell\} \\ & - \sum_{\ell=1}^{L} |\beta_\ell|^2 \Big(\sum_{j=1}^{L} |\widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_j|^2 + \rho\sigma_{e,\ell}^2 + \sigma_0^2\Big)\end{aligned} \tag{41a}$$

$$\text{s.t.} \quad \sum_{\ell=1}^{L} \|\mathbf{w}_\ell\|^2 \leq \rho \tag{41b}$$

$$|\theta_i| \leq 1,\ \forall i = 1, \cdots, N \tag{41c}$$

$$\alpha_\ell \in \mathbb{R}_+, \qquad \beta_\ell \in \mathbb{C}. \tag{41d}$$

Details of the transformation are available in [67], [68].

The four variables $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{W}$, and $\boldsymbol{\theta}$, are updated iteratively. During each iteration, they are updated one-by-one, while keeping the others fixed with the results of last iteration. The update equation for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is obtained by finding the root of the gradient of (41a) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively:

$$\begin{aligned}\alpha_\ell = \frac{1}{2}\big(\mathrm{Re}\{\beta_\ell^* \widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_\ell\}\big)^2 \times \\ \Big(1 + \sqrt{1 + 4(\mathrm{Re}\{\beta_\ell^* \widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_\ell\})^{-2}}\Big),\end{aligned} \tag{42}$$

$$\beta_\ell = \frac{\sqrt{(1+\alpha_\ell)}\,\widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_\ell}{\sum_{j=1}^{L} |\widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_j|^2 + \rho\sigma_{e,\ell}^2 + \sigma_0^2}. \tag{43}$$

The update for $\mathbf{W}$ solves the following:

$$\begin{aligned}\max_{\mathbf{W}} \quad & \sum_{\ell=1}^{L} 2\sqrt{1+\alpha_\ell}\,\mathrm{Re}\{\beta_\ell^*\ \widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_\ell\} \\ & - \sum_{\ell=1}^{L} |\beta_\ell|^2 \Big(\sum_{j=1}^{L} |\widetilde{\boldsymbol{\theta}}^H \widehat{\mathbf{H}}_\ell^H \mathbf{w}_j|^2\Big)\end{aligned} \tag{44a}$$

$$\text{s.t.} \quad \sum_{\ell=1}^{L} \|\mathbf{w}_\ell\|^2 \leq \rho, \tag{44b}$$

which is equivalent to (41) when $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are fixed. The problem (44) is a standard quadratically constrained quadratic

[5] Any convex combination of rates may equally well be optimized in the same manner.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3524943

programming, whose solution via Lagrange multipliers [69] is:

$$\mathbf{w}_\ell = \beta_\ell\sqrt{1+\alpha_\ell}\Big(\lambda\mathbf{I}_M + \sum_{j=1}^{L}|\beta_j|^2\widehat{\mathbf{H}}_j\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^H\widehat{\mathbf{H}}_j^H\Big)^{-1}\widehat{\mathbf{H}}_\ell\widetilde{\boldsymbol{\theta}}, \tag{45}$$

where $\lambda$ is a Lagrange multiplier, whose optimal value can be obtained via a grid search [69]. The update for $\boldsymbol{\theta}$ solves the problem:

$$\max_{\boldsymbol{\theta}} \quad \mathrm{Re}\{2\boldsymbol{\theta}^H\boldsymbol{\kappa}\} - \boldsymbol{\theta}^H\boldsymbol{\Phi}\boldsymbol{\theta} \tag{46a}$$

$$\text{s.t.} \quad |\theta_i| \le 1, \ \forall i = 1, \cdots, N, \tag{46b}$$

where $\boldsymbol{\kappa}$ and $\boldsymbol{\Phi}$ are functions of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\mathbf{W}$ that act as constants in the update of $\boldsymbol{\theta}$, as follows:

$$\boldsymbol{\kappa} \triangleq \sum_{\ell=1}^{L}\sqrt{(1+\alpha_\ell)}\mathrm{diag}(\beta_\ell^*\mathbf{g}_\ell)\mathbf{H}\mathbf{w}_\ell - \sum_{\ell=1}^{L}|\beta_\ell|^2\mathrm{diag}(\mathbf{g}_\ell)\mathbf{H}\sum_{j=1}^{L}\mathbf{w}_j\mathbf{w}_j^H\mathbf{h}_{d,\ell},$$

$$\boldsymbol{\Phi} \triangleq \sum_{\ell=1}^{L}|\beta_\ell|^2\sum_{j=1}^{L}\mathrm{diag}(\mathbf{g}_\ell)\mathbf{H}\mathbf{w}_j\mathbf{w}_j^H\mathbf{H}^H\mathrm{diag}(\mathbf{g}_\ell).$$

This is a quadratic programming whose optimal solution via Lagrange multipliers is:

$$\boldsymbol{\theta} = (\boldsymbol{\Phi} + \mu\mathbf{I}_N)^{-1}\boldsymbol{\kappa}, \tag{47}$$

where $\mu$ is the Lagrange multiplier whose optimal value at the solution can be obtained via a grid search [69].

### D. Complexity

Unlike most non-orthogonal schemes that require interference-canceling receivers, product superposition in our pilot-domain NOMA offers a structured approach where gains are achievable without explicitly performing interference cancellation. This is because the interfering signal is multiplied by the true channel, creating a 'virtual channel' for some users, which is easily estimated at the receiver side. This provides unique advantages in terms of computational complexity, compared with other NOMA techniques (e.g., power-domain NOMA).

Compared with a baseline orthogonal transmission technique, the main additional computation required by the proposed technique lies in the transmitter performing superposition in each training sub-phase. It is important to note, though, that the scope of this computational burden is relatively limited, because this only applies to pilot slots in which data is transmitted in parallel. In other words, the amount of added computational complexity in our technique is linearly proportional to the gains in rate. Unlike some others, our method is unburdened by any overhead unrelated to rate gains.

The total complexity of the proposed scheme over the sub-blocks with superposition pilots is $\mathcal{O}(M^2(N+1)+\ell' M^2(N+1)+2\ell' M^3(N+1)^3+LM^3(N+1)^6N^2)$, where $\ell'$ denotes the number of users that need to update their channels. The first term corresponds to a $M \times M$ data matrix that is transmitted over the pilot slots. The second term represents the computational complexity to estimate the true channel at the transmitter, where the inverse of a $M \times M$ matrix should be calculated for $\ell'$ users. The third term corresponds to the sum of computational complexities for MMSE estimation of the virtual channel at both the receiver and the transmitter. The last term represents the complexity of the beamforming (from the transmitter and RIS) for $L$ users over the data phase.

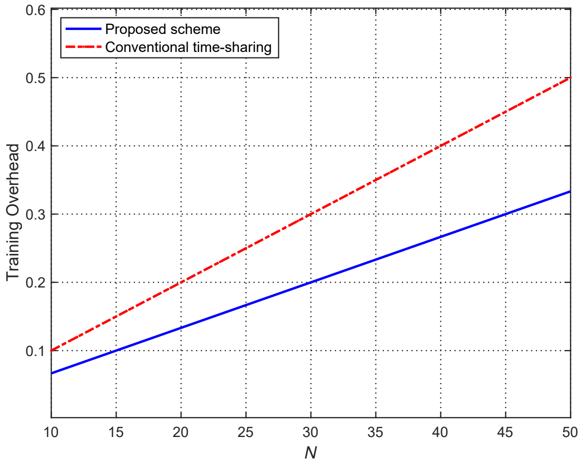

Fig. 5. Training overhead as a function of RIS elements

## V. Numerical Results

Unless stated otherwise, in this section $T = \infty$, $M = L$, $\sigma_0^2 = 1$ W, $\rho = 10$ dB, and power allocation is considered for pilot and data slots ($\rho_\tau$ and $\rho_d$).

Fig. 5 illustrates the required training overhead with respect to the number of RIS elements for two users. Here, $L = 2$, $M = 1$, $T_1 = 250$ and $T_2 = 150$. For comparison, we also provide the resulting training overhead via the conventional transmission scheme, where ordinary pilots are sent every $T_2$ time slot (i.e., transmission without pilot reuse). This shows that the proposed scheme is more economical in training overhead than conventional techniques.

Fig. 6 compares the rate regions achieved via the proposed transmission scheme and the conventional time-sharing when $L = 2$, $M = 1$, $N = 32$, $T_1 = 500$, and $T_2 = 100$. The proposed transmission scheme provides a significantly better achievable rate region compared with conventional signaling, under similar conditions.

Fig. 7 shows the individual rates and sum-rate achieved through the proposed transmission scheme. Here, $L = 3$, $M = 1$, $N = 64$, $T = 500$, $T_1 = 600$, $T_2 = 250$, and $T_3 = 150$. For comparison, we also provide the achieved sum-rate results via benchmark schemes: conventional time-sharing (TDMA with equal time-sharing factors), optimized TDMA (allocating transmission time to the user with the largest channel gain [64]), and conventional power-domain NOMA (without pilot reuse and with equal power allocation). It is observed that the proposed scheme can achieve significant gains over conventional schemes. For example, at the

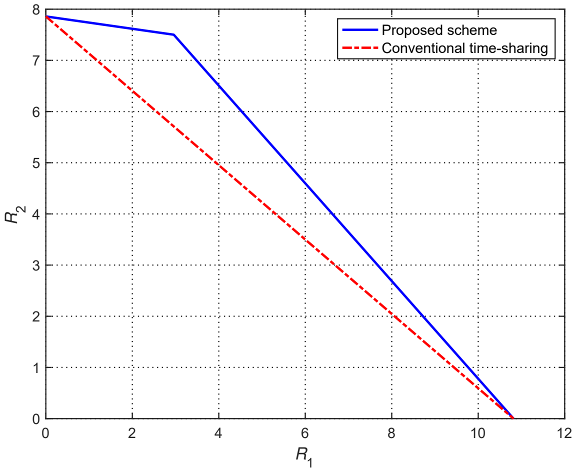

Fig. 6. Comparison of rate regions achieved through the proposed transmission scheme and the conventional time-sharing

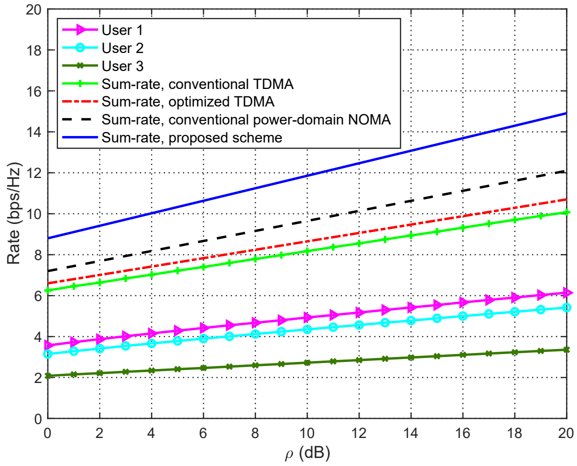

Fig. 7. Comparison of achievable sum-rate for downlink SISO

transmit power $\rho = 10$ dB, the proposed scheme achieves approximately 3.5 bps/Hz and 2.4 bps/Hz gain compared with optimized TDMA and conventional power-domain NOMA, respectively.

Although we compare the proposed scheme with the conventional power-domain NOMA, it is important to note that our scheme is not in competition with power-domain NOMA, as the two operate on different tracks: one over pilot slots and the other over data slots. As stated in the Introduction and Conclusion sections, these two techniques can be combined, and their benefits are cumulative, to achieve even more gain.

Fig. 8 shows the impact of feedback channel power on achievable sum-rates. Here, $L = 2$, $M = 1$, $N = 32$, $T = 500$, $T_1 = 250$, and $T_2 = 100$. For comparison, we also provide the achieved result via the perfect (error-free) feedback link. This shows that as the power of the feedback channel increases, the achievable sum-rates increase, converging to the results achieved under a perfect feedback link.

Fig. 9 shows the impact of the shortest coherence time on achievable rates. Here, $L = M = 2$, $N = 16$, and $T_1 = 150$. We vary $T_2$ from 40 to 150. This shows that as $T_2$ increases, the rate for User 2 improves since its training overhead is reduced. In contrast, the rate for User 1 decreases as $T_2$ is increased, because there are fewer reused pilot slots it can exploit. When $T_1 = T_2$, there is no opportunity for pilot reuse strategy and thus both users achieve the same rates. We notice that when $T_2 \leq 45$, the RIS produces no gains, because in smaller coherence intervals, there are insufficient samples to amortize the training cost of RIS. The limits incurred by training overhead are broadly an issue for all RIS-assisted systems.

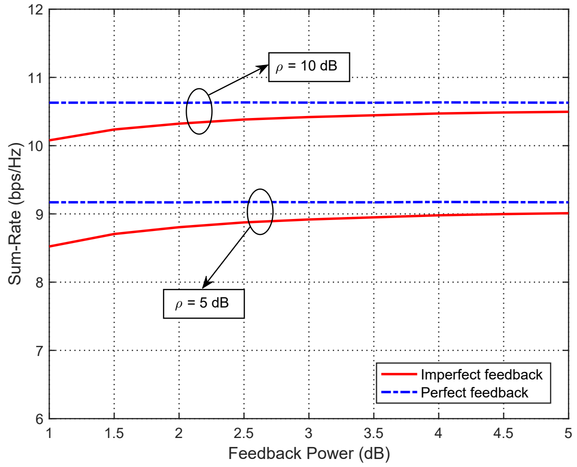

Fig. 8. Impact of analog feedback power on achievable rates

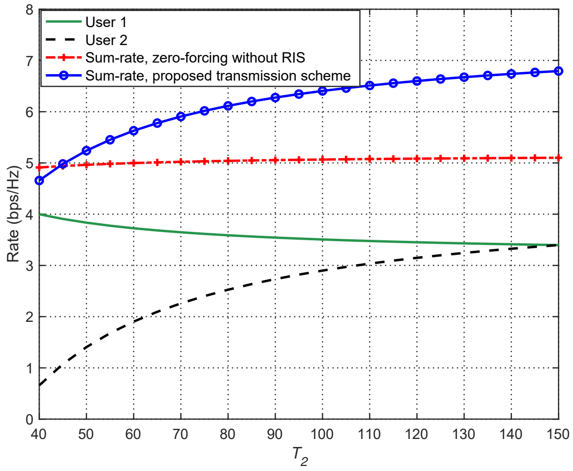

Fig. 9. Achievable rates as a function of block length $T_2$

Fig. 10 shows the achievable sum-rate via the proposed scheme with joint beamforming and RIS reflection coefficient optimization. Here, $L = M = 3$, $N = 64$, $T = 500$, $T_1 = 600$, $T_2 = 250$, and $T_3 = 150$. For reference, we also show the effect of randomized RIS coefficients, as well as no RIS. This figure shows that beamforming is important at both the RIS and the transmitter for maximizing gains. At the target sum-rate of 10 bps/Hz, a gain of $\sim 10$ dB is achieved over the

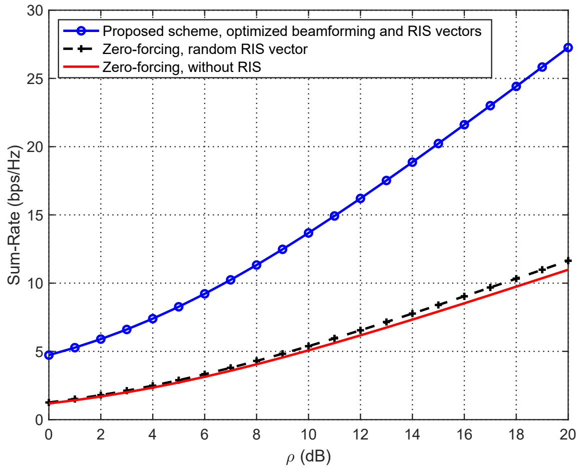

Fig. 10. Achievable sum-rate via the proposed and conventional transmission schemes

conventional zero-forcing (without pilot reuse) with a random RIS vector.

## VI. Conclusion

This paper proposes a new and efficient pilot-domain non-orthogonal multiple access (NOMA) signaling scheme for a multi-user downlink channel with reconfigurable intelligent surfaces (RIS) whose links experience unequal coherence times. The channel state feedback is available imperfectly. The outcomes of this paper include a significant reduction of training overhead and corresponding rate gains. The technical contribution of this work includes a harmonious combination of product superposition and beamforming from both transmitter and RIS under imperfect channel state information, in a manner that maximizes sum-rates.

The proposed pilot-domain NOMA in this paper exploits a different source of gains compared with conventional data-domain (a.k.a. power-domain) NOMA whose gain often comes with the requirement of interference-canceling (peeling) receivers. Because of the different nature of the gains and the neat structure of product superposition, the gain of the proposed scheme can be harvested without the need to explicitly perform interference cancellation. The gains of these two NOMA schemes are independent of each other and both of them can be achieved at the same time within the same system. In other words, the pilot-domain NOMA gains of the present paper can be combined with power-domain NOMA gains obtained in previous works.

## References

[1] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Select. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, 2020.

[2] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, 2019.

[3] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3313–3351, 2021.

[4] E. Bjornson, O. Ozdogan, and E. G. Larsson, "Reconfigurable intelligent surfaces: Three myths and two critical questions," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 90–96, 2020.

[5] X. Wei, D. Shen, and L. Dai, "Channel estimation for RIS assisted wireless communications—part I: Fundamentals, solutions, and future opportunities," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1398–1402, 2021.

[6] Z. Zhou, N. Ge, Z. Wang, and L. Hanzo, "Joint transmit precoding and reconfigurable intelligent surface phase adjustment: A decomposition-aided channel estimation approach," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1228–1243, 2021.

[7] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Matrix-calibration-based cascaded channel estimation for reconfigurable intelligent surface assisted multiuser MIMO," *IEEE J. Select. Areas Commun.*, vol. 38, no. 11, pp. 2621–2636, 2020.

[8] Q.-U.-A. Nadeem, H. Alwazani, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Intelligent reflecting surface-assisted multi-user MISO communication: Channel estimation and beamforming design," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 661–680, 2020.

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

[10] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Communications Surveys and Tutorials*, vol. 23, no. 3, pp. 1546–1577, 2021.

[11] C. Pan, H. Ren, K. Wang, J. F. Kolb, M. Elkashlan, M. Chen, M. Di Renzo, Y. Hao, J. Wang, A. L. Swindlehurst, X. You, and L. Hanzo, "Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 14–20, 2021.

[12] S. Noh, J. Lee, G. Lee, K. Seo, Y. Sung, and H. Yu, "Channel estimation techniques for RIS-assisted communication: Millimeter-wave and sub-thz systems," *IEEE Veh. Technol. Mag.*, vol. 17, no. 2, pp. 64–73, 2022.

[13] B. Shamasundar, N. Daryanavardan, and A. Nosratinia, "Channel training and estimation for reconfigurable intelligent surfaces: Exposition of principles, approaches, and open problems," *IEEE Access*, vol. 11, pp. 6717–6734, 2023.

[14] Z. Ding, L. Lv, F. Fang, O. A. Dobre, G. K. Karagiannidis, N. Al-Dhahir, R. Schober, and H. V. Poor, "A state-of-the-art survey on reconfigurable intelligent surface-assisted non-orthogonal multiple access networks," *Proceedings of the IEEE*, vol. 110, no. 9, pp. 1358–1379, 2022.

[15] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6607–6620, 2020.

[16] Z. Mao, W. Wang, Q. Xia, C. Zhong, X. Pan, and Z. Ye, "Element-grouping intelligent reflecting surface: Electromagnetic-compliant model and geometry-based optimization," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5362–5376, 2022.

[17] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, "Intelligent reflecting surface meets OFDM: Protocol design and rate maximization," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4522–4535, 2020.

[18] Y. Gao, C. Yong, Z. Xiong, J. Zhao, Y. Xiao, and D. Niyato, "Reflection resource management for intelligent reflecting surface aided wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6971–6986, 2021.

[19] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8238–8242, 2019.

[20] M.-M. Zhao, Q. Wu, M.-J. Zhao, and R. Zhang, "Intelligent reflecting surface enhanced wireless networks: Two-timescale beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 2–17, 2021.

[21] K. Zhi, C. Pan, H. Ren, and K. Wang, "Power scaling law analysis and phase shift optimization of RIS-aided massive MIMO systems with statistical CSI," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3558–3574, 2022.

[22] J. An and L. Gan, "The low-complexity design and optimal training overhead for IRS-assisted MISO systems," *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1820–1824, 2021.

[23] J. An, Q. Wu, and C. Yuen, "Scalable channel estimation and reflection optimization for reconfigurable intelligent surface-enhanced OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 796–800, 2022.

[24] L. Tong, B. Sadler, and M. Dong, "Pilot-assisted wireless transmissions: general model, design criteria, and signal processing," *IEEE Signal Processing Mag.*, vol. 21, no. 6, pp. 12–25, 2004.

[25] A. Lozano and N. Jindal, "Optimum pilot overhead in wireless communication: A unified treatment of continuous and block-fading channels," in *2010 European Wireless Conference (EW)*, 2010, pp. 725–732.

[26] Y. Li and A. Nosratinia, "Coherent product superposition for downlink multiuser MIMO," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1746–1754, 2015.

[27] M. Fadel and A. Nosratinia, "Coherence disparity in broadcast and multiple access channels," *IEEE Trans. Inform. Theory*, vol. 62, no. 12, pp. 7383–7401, 2016.

[28] M. Fadel and A. Nosratinia, "Frequency-selective multiuser downlink channels under mismatched coherence conditions," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2393–2404, 2019.

[29] M. Fadel Shady and A. Nosratinia, "MISO broadcast channel under unequal link coherence times and channel state information," *Entropy*, vol. 22, no. 9, p. 976, 2020.

[30] F. Zhang and A. Nosratinia, "The impact of coherence diversity on MIMO relays," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 6906–6919, 2022.

[31] M. Fadel and A. Nosratinia, "Broadcast channel under unequal coherence intervals," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 275–279.

[32] ——, "Coherence diversity in time and frequency," in *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6.

[33] M. Karbalayghareh and A. Nosratinia, "Interaction of pilot reuse and channel state feedback under coherence disparity," in *2022 IEEE Information Theory Workshop (ITW)*, 2022, pp. 190–195.

[34] ——, "RIS-assisted downlink transmission under unequal coherence intervals and CSI feedback," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2023, pp. 1–6.

[35] Y. Li, M. Jiang, Q. Zhang, and J. Qin, "Joint beamforming design in multi-cluster MISO NOMA reconfigurable intelligent surface-aided downlink communication networks," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 664–674, 2021.

[36] T. Hou, Y. Liu, Z. Song, X. Sun, Y. Chen, and L. Hanzo, "Reconfigurable intelligent surface aided NOMA networks," *IEEE J. Select. Areas Commun.*, vol. 38, no. 11, pp. 2575–2588, 2020.

[37] Z. Ding and H. Vincent Poor, "A simple design of IRS-NOMA transmission," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1119–1123, 2020.

[38] Y. Liu, X. Mu, X. Liu, M. Di Renzo, Z. Ding, and R. Schober, "Reconfigurable intelligent surface-aided multi-user networks: Interplay between NOMA and RIS," *IEEE Wireless Communications*, vol. 29, no. 2, pp. 169–176, 2022.

[39] J. Zuo, Y. Liu, Z. Qin, and N. Al-Dhahir, "Resource allocation in intelligent reflecting surface assisted NOMA systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7170–7183, 2020.

[40] Y. Guo, Z. Qin, Y. Liu, and N. Al-Dhahir, "Intelligent reflecting surface aided multiple access over fading channels," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 2015–2027, 2021.

[41] J. Zuo, Y. Liu, Z. Ding, L. Song, and H. V. Poor, "Joint design for simultaneously transmitting and reflecting (STAR) RIS assisted NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 611–626, 2023.

[42] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting intelligent reflecting surfaces in NOMA networks: Joint beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6884–6898, 2020.

[43] T.-H. Vu, T.-V. Nguyen, D. B. d. Costa, and S. Kim, "Intelligent reflecting surface-aided short-packet non-orthogonal multiple access systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4500–4505, 2022.

[44] W. Wang, X. Liu, J. Tang, N. Zhao, Y. Chen, Z. Ding, and X. Wang, "Beamforming and jamming optimization for IRS-aided secure NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1557–1569, 2022.

[45] X. Yu, D. Xu, and R. Schober, "MISO wireless communication systems via intelligent reflecting surfaces : (invited paper)," in *2019 IEEE/CIC International Conference on Communications in China (ICCC)*, 2019, pp. 735–740.

[46] M. Fu, Y. Zhou, and Y. Shi, "Intelligent reflecting surface for downlink non-orthogonal multiple access networks," in *2019 IEEE Globecom Workshops (GC Wkshps)*, 2019, pp. 1–6.

[47] G. Yang, X. Xu, and Y.-C. Liang, "Intelligent reflecting surface assisted non-orthogonal multiple access," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1–6.

[48] B. Zheng, Q. Wu, and R. Zhang, "Intelligent reflecting surface-assisted multiple access with user pairing: NOMA or OMA?" *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 753–757, 2020.

[49] C. Wu, C. You, Y. Liu, S. Han, and M. D. Renzo, "Two-timescale design for STAR-RIS-aided NOMA systems," *IEEE Trans. Commun.*, vol. 72, no. 1, pp. 585–600, 2024.

[50] N. Zhang, Y. Liu, X. Mu, W. Wang, and A. Huang, "Queue-aware STAR-RIS assisted NOMA communication systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 5, pp. 4786–4801, 2024.

[51] Q. Gao, Y. Liu, X. Mu, M. Jia, D. Li, and L. Hanzo, "Joint location and beamforming design for STAR-RIS assisted NOMA systems," *IEEE Trans. Commun.*, vol. 71, no. 4, pp. 2532–2546, 2023.

[52] D. Makrakis and K. Feher, "A novel pilot insertion-extraction method based on spread spectrum techniques," in *Miami Technicon*, 1987, pp. 129–132.

[53] T. Holden and K. Feher, "A spread spectrum based system technique for synchronization of digital mobile communication systems," in *IEEE Vehicular Technology Conference (VTC)*, 1989, pp. 780–787.

[54] A. Steingass, A. van Wijngaarden, and W. Teich, "Frame synchronization using superimposed sequences," in *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, 1997, pp. 489–.

[55] F. Tufvesson, M. Faulkner, P. Hoeher, and O. Edfors, "OFDM time and frequency synchronization by spread spectrum pilot technique," in *1999 IEEE Communications Theory Mini-Conference*, 1999, pp. 115–119.

[56] W. Yuan, S. Li, Z. Wei, J. Yuan, and D. W. K. Ng, "Data-aided channel estimation for OTFS systems with a superimposed pilot and data transmission scheme," *IEEE Wireless Commun. Lett.*, vol. 10, no. 9, pp. 1954–1958, 2021.

[57] H. B. Mishra, P. Singh, A. K. Prasad, and R. Budhiraja, "OTFS channel estimation and data detection designs with superimposed pilots," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2258–2274, 2022.

[58] C. Yang, J. Wang, Z. Pan, and S. Shimamoto, "Delay-doppler frequency domain-aided superimposing pilot OTFS channel estimation based on deep learning," in *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, 2022, pp. 1–6.

[59] Y. Liu, Y. L. Guan, and D. González G., "BEM OTFS receiver with superimposed pilots over channels with doppler and delay spread," in *2022 IEEE International Conference on Communications (ICC)*, 2022, pp. 2411–2416.

[60] J. Ma, C. Liang, C. Xu, and L. Ping, "On orthogonal and superimposed pilot schemes in massive MIMO NOMA systems," *IEEE J. Select. Areas Commun.*, vol. 35, no. 12, pp. 2696–2707, 2017.

[61] H. Zhang, S. Gao, D. Li, H. Chen, and L. Yang, "On superimposed pilot for channel estimation in multicell multiuser MIMO uplink: Large system analysis," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1492–1505, 2016.

[62] M. Karbalayghareh and A. Nosratinia, "Pilot-domain NOMA for multi-user RIS-assisted communications," in *2024 IEEE International Conference on Communications (ICC)*, 2024.

[63] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inform. Theory*, vol. 56, no. 6, pp. 2845–2866, 2010.

[64] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Select. Areas Commun.*, vol. 24, no. 3, pp. 528–541, 2006.

[65] M. Kobayashi and G. Caire, "Joint beamforming and scheduling for a multi-antenna downlink with imperfect transmitter channel knowledge," *IEEE J. Select. Areas Commun.*, vol. 25, no. 7, pp. 1468–1477, 2007.

[66] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 951–963, 2003.

[67] K. Shen and W. Yu, "Fractional programming for communication systems—part I: Power control and beamforming," *IEEE Trans. Signal Processing*, vol. 66, no. 10, pp. 2616–2630, 2018.

[68] ——, "Fractional programming for communication systems—part II: Uplink scheduling via matching," *IEEE Trans. Signal Processing*, vol. 66, no. 10, pp. 2631–2644, 2018.

[69] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, ser. Foundations and Trends in Machine learning. NOW Publishers, 2011.

**Mehdi Karbalayghareh** (Member, IEEE) received the B.Sc. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 2015, and the M.Sc. degree in electrical engineering from Ozyegin University, Istanbul, Turkey, in 2019. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Texas at Dallas, Richardson, TX, USA. His research interests include wireless communications, information theory, and machine learning. He was the recipient of the 2023 Excellence in Education Doctoral Fellowship, and the 2024 Research Excellence Award at the University of Texas at Dallas. He was also awarded the IEEE ComSoc Travel Grant for IEEE ICC 2024.

**Aria Nosratinia** (S'87, M'97, SM'04, F'10) is Erik Jonsson Distinguished Professor and associate head of the electrical engineering department at the University of Texas at Dallas. He received his Ph.D. in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign in 1996. He had visiting appointments at Princeton University, Rice University, and UCLA. His interests lie in the broad area of information theory and signal processing, with applications in wireless communications, data security and privacy. Dr. Nosratinia is a fellow of IEEE for contributions to multimedia and wireless communications. He has served as editor and area editor for the IEEE Transactions on Wireless Communications, and editor for the IEEE Transactions on Information Theory, IEEE Transactions on Image Processing, IEEE Signal Processing Letters, IEEE Wireless Communications (Magazine), and Journal of Circuits, Systems, and Computers. He has received the National Science Foundation career award, and the Outstanding Service award from the IEEE Signal Processing Society, Dallas Chapter. Dr. Nosratinia is a registered professional engineer in the state of Texas and a Clarivate Analytics highly cited researcher.