MIMO-Aided Nonlinear Hybrid Transceiver Design for Multiuser mmWave Systems Relying on Tomlinson-Harashima Precoding

Kaidi Xu, Yunlong Cai, Minjian Zhao, Yong Niu, and Lajos Hanzo

***Abstract*—Hybrid analog-digital (A/D) transceivers designed for millimeter wave (mmWave) systems have received substantial research attention, as a benefit of their lower cost and modest energy consumption compared to their fully-digital counterparts. We further improve their performance by conceiving a Tomlinson-Harashima precoding (THP) based nonlinear joint design for the downlink of multiuser multiple-input multiple-output (MIMO) mmWave systems. Our optimization criterion is that of minimizing the mean square error (MSE) of the system under channel uncertainties subject both to realistic transmit power constraint and to the unit modulus constraint imposed on the elements of the analog beamforming (BF) matrices governing the BF operation in the radio frequency domain. We transform this optimization problem into a more tractable form and develop an efficient block coordinate descent (BCD) based algorithm for solving it. Then, a novel two-timescale nonlinear joint hybrid transceiver design algorithm is developed, which can be viewed as an extension of the BCD-based joint design algorithm for reducing both the channel state information (CSI) signalling overhead and the effects of outdated CSI. Moreover, we determine the near-optimal cancellation order for the THP structure based on the lower bound of the MSE. The proposed algorithms can be guaranteed to converge to a Karush-Kuhn-Tucker (KKT) solution of the original problem. The simulation results demonstrate that our proposed nonlinear joint hybrid transceiver design algorithms significantly outperform the existing linear hybrid transceiver algorithms and approach the performance of the fully-digital transceiver, despite its lower cost and power dissipation.**

***Index Terms*—Nonlinear precoding, hybrid A/D beamforming, hardware-efficient, mmWave, two-timescale.**

The work of Y. Cai was supported in part by the National Natural Science Foundation of China under Grants 61971376 and 61831004, in part by the Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars under Grant LR19F010002, and in part by the State Key Laboratory of Rail Traffic Control and Safety (Contract No. RCS2020K010), Beijing Jiaotong University. L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/P034284/1 and EP/P003990/1 (COALESCE) as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028). (Corresponding author: Yunlong Cai.)

K. Xu and M. Zhao are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: xukaidi13@126.com; mjzhao@zju.edu.cn).

Y. Cai is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and also with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: ylcai@zju.edu.cn).

Y. Niu is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: niuy11@163.com).

L. Hanzo is with the Department of ECS, University of Southampton, U.K. (e-mail: lh@ecs.soton.ac.uk).

I. Introduction

The global spectrum shortage has stimulated considerable interest in the development of millimeter wave (mmWave) communications for the next generation wireless networks [1]–[9]. At a carrier frequency of 30 GHz or 1cm wavelength, numerous antenna elements can be packed into a compact space. This facilitates large-scale spatial multiplexing and high-gain directional beamforming (BF) and thereby significantly increases the system capacity. However, for large-scale multiple-input multiple-output (MIMO) mmWave systems the conventional fully-digital (FD) BF architecture requires numerous radio frequency (RF) chains which results in extremely high fabrication cost and high power consumption. In order to circumvent these drawbacks, hybrid analog-digital (A/D) BF architectures have been proposed, which require less RF chains than the FD BF architecture, when using the same number of antennas [10]–[27].

In [10], the authors analyzed the beam-alignment performance of both exhaustive and hierarchical search techniques, with the time-domain training overhead taken into account. An optimized two-stage search algorithm was proposed in [11] for transmitter and receiver beam alignment. In [12], the authors established that a hybrid A/D BF structure with twice as many RF chains as data streams is capable of realizing any FD BF structure exactly. A series of matrix-decomposition based hybrid BF design algorithms have been proposed in [14]–[16]. By exploiting the sparse nature of the channel matrix, the authors of [17] formulated the hybrid BF design problem as a sparse matrix reconstruction problem and solved it using the modified orthogonal matching pursuit (OMP) algorithm. In [18], the authors addressed the joint optimization of computation and communication power for multi-user massive MIMO systems with partially-connected structures. In [19], the authors investigated a hybrid transceiver design using realistic limited feedback in their multi-user mmWave systems. As a further advance, the authors of [13] have developed an alternating minimization algorithm for their hybrid BF design with the aid of manifold optimization (MO). The authors of [20] considered the uplink of large-scale multiuser MIMO mmWave systems, where the implementation cost of their joint hybrid BF algorithm was reduced with the aid of antenna selection. In order to mitigate the hardware-induced performance erosion, a number of codebook-based hybrid BF algorithms were conceived in [21], [22]. In [23], [24] the unit-modulus constraint and power constraints imposed upon the A/D hybrid BF were mitigated by the penalty dual

decomposition (PDD) [28] based hybrid BF design algorithm, which can be guaranteed to achieve the Karush-Kuhn-Tucker (KKT) solution. In particular, the authors of [23] directly optimized the spectral efficiency of the mmWave downlink in a multiuser multistream MIMO system. Then Cai *et al.* [24] extended the solution advocated in [23] to a mmWave full-duplex MIMO relay-aided system. As another development, both the channel state information (CSI) feedback overhead and the implementation complexity were reduced as part of a series of two-timescale based studies for the design of A/D hybrid BF [25]–[27]. Explicitly, the long-timescale analog BF matrices were optimized based on the channel statistics, while the short-timescale digital precoding matrices were updated according to the near-instantaneous CSI.

In parallel to the low-complexity linear transceiver structures, more sophisticated nonlinear transceivers, such as the Tomlinson-Harashima precoding (THP) have also evolved from the seminal contributions of [29], [30], leading to powerful spatial-domain MIMO solutions [31]. The THP-based nonlinear transceiver algorithms have a remarkable performance gain over their linear counterparts and thus have found numerous applications [32]–[34]. However, determining the optimal cancellation order under the THP structure, which achieves the optimal performance gain is quite a challenge [35]. In [36], the authors proposed a multi-branch (MB) THP scheme, where each branch contains a THP with a predefined ordering strategy, and a selection criterion is applied to choose the best branch to generate the final output. Moreover, the THP-based robust nonlinear transceiver design has also been further developed by taking the CSI errors into account in relay-aided multiuser MIMO systems [37]. This solution has also been extended to a full-duplex relay-aided wireless power transfer system in [38]. Finally, the authors of [39], [40] proposed techniques for reducing the power-loss imposed by the modulo and feedback operations used in the THP.

However, to the best of our knowledge, the aforementioned A/D hybrid transceiver design algorithms are all based on the linear precoding structure, which suffers from the performance degradation caused by the multiuser interference and by the reduced number of available RF chains. Against this background, we propose a THP-based joint A/D hybrid transceiver design algorithm for the downlink of multiuser mmWave MIMO systems for further improving the system performance. Specifically, we jointly optimize the analog BF matrices and the digital processing matrices, i.e., the digital precoding and the receiver as well as feedback matrices of the THP structure. Explicitly, we minimize the system's mean square error (MSE) subject to both the transmit power constraint and the unit modulus constraint imposed on each element of the analog RF BF matrices. The optimization problem formulated is quite challenging to tackle. By efficiently exploiting the particular structure of this problem, we first transform it into a more tractable form. Then we propose an efficient block coordinate descent (BCD) based algorithm for solving the converted problem. Furthermore, we extend the proposed BCD-based joint design algorithm to a novel two-timescale nonlinear joint hybrid transceiver design algorithm in order to reduce both the CSI signalling overhead and the effects of outdated CSI caused by its signalling delay. The proposed algorithms can be guaranteed to obtain a KKT solution of the original problem.

The main contributions of this work are summarized as follows:

1) There is a paucity of literature on optimizing the nonlinear A/D hybrid transceiver matrices by minimizing the MSE, because this problem is very challenging. Hence we first transform this problem into a more tractable form and optimize the matrix variables in a BCD fashion, where the subproblems of each block can be solved in closed form.
2) We develop a novel two-timescale nonlinear hybrid transceiver design algorithm based on two-stage online successive convex approximation (TOSCA). Although the proposed TOSCA-based two-timescale algorithm suffers from a certain performance degradation compared to the proposed BCD-based joint design algorithm in the presence of small delays, both the CSI signalling overhead and the effects of outdated CSI caused by high CSI signalling delays can be substantially reduced. In this scheme, the long-timescale analog BF matrices are optimized based on the channel statistics, while the short-timescale digital processing matrices are designed based on the low-dimensional effective CSI matrices for each time slot.
3) We determine the near-optimal cancellation order for the proposed THP-based hybrid transceiver design based on the lower bound of the MSE. Our simulation results demonstrate that the proposed BCD-based joint nonlinear hybrid transceiver design algorithm significantly outperforms the existing linear hybrid transceiver algorithms and approaches the performance of the fully-digital transceiver. Furthermore, compared to the proposed BCD-based joint design algorithm, the proposed two-timescale joint design algorithm provides better performance in the scenario of severe CSI delays, although it suffers from some performance degradation for small delays.

The rest of this paper is structured as follows. Section II introduces the proposed THP-based mmWave multiuser MIMO system and the optimization problems formulated. In Section III, we first transform the problem into a more tractable form and then propose a BCD-based joint design algorithm to solve it. In Section IV, we propose the TOSCA-based two-timescale joint nonlinear transceiver design algorithm. In Section V, we derive the lower bound of the MSE and determine the near-optimal cancellation order for the proposed THP-based hybrid transceiver design. Our simulation results are presented in Section VI. Finally, Section VII offers our conclusions.

Notations: Scalars, vectors and matrices are respectively denoted by lower case, boldface lower case and boldface upper case letters. $\mathbf{I}$ represents an identity matrix and $\mathbf{0}$ denotes an all-zero matrix. For a matrix $\mathbf{A}$, $\mathbf{A}^T$, $\mathbf{A}^*$, $\mathbf{A}^H$ and $\|\mathbf{A}\|$ denote its transpose, conjugate, conjugate transpose and Frobenius norm, respectively. For a square matrix $\mathbf{A}$, $\mathrm{Tr}(\mathbf{A})$ denotes its trace, $\mathbf{A} \succeq \mathbf{0}$ ($\mathbf{A} \preceq \mathbf{0}$) means that $\mathbf{A}$ is positive (negative) semidefinite. $[\mathbf{A}]_{a:b\,,c:d}$ represents a submatrix of $\mathbf{A}$. For a vector $\mathbf{a}$, $\|\mathbf{a}\|$ represents its Euclidean norm. $\mathbb{E}\{.\}$ denotes the statistical expectation. $\mathrm{Re}(.)$ ($\mathrm{Im}(.)$)

denotes the real (imaginary) part of a variable. The operator vec$(\cdot)$ stacks the elements of a matrix in one long column vector. $|\cdot|$ denotes the absolute value of a complex scalar. The operator $\angle$ takes the phase angles of the elements in a matrix. $\mathbb{C}^{m\times n}$ ($\mathbb{R}^{m\times n}$) denotes the space of $m \times n$ complex (real) matrices. The symbol $\otimes$ denotes the Kronecker product of two vectors/matrices and the symbol $\circ$ denotes the Hadamard product of two vectors/matrices.

II. System model and problem formulation

In this section, we first introduce the system model of mmWave multiuser MIMO systems, and then mathematically formulate the optimization problem of interest.

A. System model

As illustrated in Fig. 1, we consider a mmWave communication system comprising of one BS and M users. The BS equipped with N_s antennas and R_s RF chains, where $N_s \geq R_s$, transmits a signal vector $\mathbf{s} = [\mathbf{s}_1^T, \mathbf{s}_2^T, \ldots, \mathbf{s}_M^T]^T \in \mathbb{C}^{D\times 1}$ to the users, where $\mathbf{s}_m \in \mathbb{C}^{D_m\times 1}$ denotes the signal vector for user m, $m \in \mathcal{M} \triangleq \{1, 2, ..., M\}$, and $D = \sum_{m=1}^{M} D_m$ denotes the total number of transmit data streams. User m is equipped with $N_{d,m}$ antennas and $R_{d,m}$ RF chains. Besides, we assume the necessary condition $R_s \geq D$ for sufficient degree of freedom. Each entry of the transmitted signal vector $\mathbf{s}$ is a Q-ary quadrature amplitude modulation (QAM) signal. Hence the real and imaginary parts of each entry of the signal vector $\mathbf{s}$ are independent and identically distributed (i.i.d.) random variables generated from a Q-ary QAM alphabet $\mathcal{A}$. Specifically, we let $\mathcal{A} = \{\pm\sqrt{\frac{3}{2(Q-1)}}, \pm 3\sqrt{\frac{3}{2(Q-1)}}, \pm 5\sqrt{\frac{3}{2(Q-1)}}, \ldots, \pm(\sqrt{Q}-1)\sqrt{\frac{3}{2(Q-1)}}\}$, $\mathrm{E}\{\mathbf{s}\} = \mathbf{0}$ and $\mathrm{E}\{\mathbf{s}\mathbf{s}^H\} = \mathbf{I}$, where $\mathrm{Re}(s_k) \in \mathcal{A}$, $\forall k$, $\mathrm{Im}(s_k) \in \mathcal{A}$, $\forall k$ and s_k is the kth entry of the vector $\mathbf{s}$.

$\mathrm{MOD}_Q(.)$ in Fig. 1 is a modulo operator used to constrain a value in $(-\sqrt{\tau}, \sqrt{\tau}]$, where $\tau = \sqrt{\frac{3Q}{2(Q-1)}}$. This operator can be formulated as

$$\mathrm{MOD}_Q(x) = x - 2\sqrt{\tau}\lfloor\frac{x+\sqrt{\tau}}{2\sqrt{\tau}}\rfloor = x + e, \tag{1}$$

where e is the residual error.

With the modulo operator in (1), we can generate the transmit symbols x_k successively as

$$x_k = s_k - \sum_{n=1}^{k-1}[\mathbf{C}]_{k,n}x_n + e_k, \tag{2}$$

where $\mathbf{C} \in \mathbb{C}^{D\times D}$ is a strictly lower triangle matrix and $\mathbf{e} = [e_1, e_2, \ldots, e_M]^T$ is the residual error vector generated by the modulo operator. Equation (2) can be rewritten in a matrix form as

$$\mathbf{x} = \mathbf{U}^{-1}\mathbf{v}, \tag{3}$$

where $\mathbf{U} = \mathbf{I} + \mathbf{C}$ is a lower triangle matrix with ones on the main diagonal and $\mathbf{v} = \mathbf{s} + \mathbf{e} = [\mathbf{v}_1^T, \mathbf{v}_2^T, \ldots, \mathbf{v}_M^T]^T$ is the target signal vector[1]. Together with the consideration in [41], we have $\mathrm{E}\{\mathbf{x}\mathbf{x}^H\} = \mathbf{I}$ for a high order Q-ary QAM constellation.

Before transmission, the processed signal $\mathbf{x}$ is passed through a linear digital precoding matrix $\mathbf{W} \in \mathbb{C}^{R_s\times D}$ followed by an analog BF matrix $\mathbf{T} \in \mathbb{C}^{N_s\times R_s}$. The transmit power constraint at the BS is given by

$$\mathrm{E}\{\|\mathbf{T}\mathbf{W}\mathbf{x}\|^2\} = \|\mathbf{T}\mathbf{W}\|^2 \leq P_t, \tag{4}$$

where P_t is the transmit power budget. The signal received at user m is given by

$$\mathbf{y}_m = \mathbf{H}_m\mathbf{T}\mathbf{W}\mathbf{x} + \mathbf{n}_m, \tag{5}$$

where $\mathbf{H}_m \in \mathbb{C}^{N_{d,m}\times N_s}$ denotes the MIMO channel matrix between the BS and user m, while $\mathbf{n}_m$ denotes the complex-valued circular Gaussian noise at user m with zero mean and correlation matrix $\mathrm{E}\{\mathbf{n}_m\mathbf{n}_m^H\} = \sigma_m^2\mathbf{I}$.[2]

At user m, a linear A/D hybrid receiver consisting of an analog BF matrix $\mathbf{F}_m \in \mathbb{C}^{R_{d,m}\times N_{d,m}}$ and a digital receiving matrix $\mathbf{P}_m \in \mathbb{C}^{D_m\times R_{d,m}}$ is employed for detecting symbols. The output of the hybrid receiver is expressed as

$$\hat{\mathbf{v}}_m = \mathbf{P}_m\mathbf{F}_m\mathbf{H}_m\mathbf{T}\mathbf{W}\mathbf{x} + \mathbf{P}_m\mathbf{F}_m\mathbf{n}_m, \tag{6}$$

while the final estimate of the signal vector for user m is given by

$$\hat{\mathbf{s}}_m = \mathrm{MOD}_Q(\hat{\mathbf{v}}_m). \tag{7}$$

In practice, channel estimation errors are inevitable. According to [44], the channel estimation errors can be modelled as

$$\mathbf{H}_m = \bar{\mathbf{H}}_m + \sigma_{e,m}\Delta\mathbf{H}_m \quad \forall m, \tag{8}$$

where $\bar{\mathbf{H}}_m \in \mathbb{C}^{N_{d,m}\times N_s}$ denotes the estimated channel matrix, $\Delta\mathbf{H}_m$ denotes the channel estimation error matrix, and $\sigma_{e,m}$ denotes the estimation error variance. Specifically, $\Delta\mathbf{H}_m$ is i.i.d. with zero-mean and unit-variance circular complex Gaussian distribution.

Furthermore, the ordering scheme for the THP structure is considered as a matrix $\mathbf{L} \in \mathbb{R}^{D\times D}$ whose elements are zeros and ones. The ordering matrix $\mathbf{L}$ follows the constraints $\mathbf{L}\mathbf{1} = \mathbf{1}$, $\mathbf{1}^T\mathbf{L} = \mathbf{1}^T$, that is, in each row and column only one entry is 1 and the others are 0s. Hence the permutation process can be expressed as $\mathbf{s} = \mathbf{L}\tilde{\mathbf{s}}$, where $\tilde{\mathbf{s}}$ is the original transmit data vector and $\mathbf{s}$ is the permutated data vector. Then we have the desired output signal vector of the linear receiver for user m is $\tilde{\mathbf{v}}_m = \mathbf{A}_m\mathbf{L}^T\mathbf{v}$, where $\mathbf{A}_m = [\mathbf{0}_{D_m\times\sum_{i=1}^{m-1}D_i}, \mathbf{I}_{D_m}, \mathbf{0}_{D_m\times\sum_{i=m+1}^{M}D_i}]$ denotes a selection matrix extracting the entries of user m in vector $\mathbf{L}^T\mathbf{v}$.[3]

[1] This is because the users can recover $\mathbf{s}$ from $\mathbf{v}$ via $\mathbf{s} = \mathrm{MOD}_Q(\mathbf{v})$.

[2] The CSI can be obtained via either the traditional least square channel estimation algorithm or some newly proposed iterative channel estimation algorithms [42], [43].

[3] This is because the desired signal vector for user m is $\mathbf{A}_m\mathbf{L}^T\mathbf{s} = \mathbf{A}_m\mathbf{L}^T\mathrm{MOD}_Q(\mathbf{v}) = \mathrm{MOD}_Q(\mathbf{A}_m\mathbf{L}^T\mathbf{v})$.

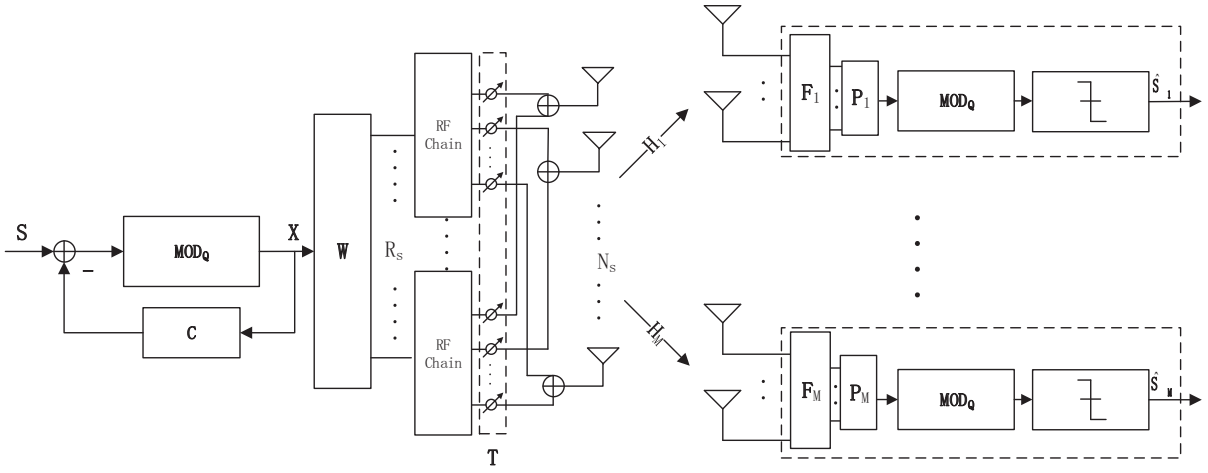

Fig. 1: System model

Finally, the MSE of all users can be expressed as

$$\begin{aligned}
&\text{MSE}(\{\mathbf{P}_m,\mathbf{F}_m\},\mathbf{T},\mathbf{W},\mathbf{U})\\
&=\sum_{m=1}^{M}\text{E}\{\|\hat{\mathbf{v}}_m-\tilde{\mathbf{v}}_m\|^2\}\\
&=\sum_{m=1}^{M}\text{tr}\Big(\mathbf{P}_m\mathbf{F}_m\bar{\mathbf{H}}_m\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H\bar{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\\
&\quad+\sigma_{e,m}^2\text{tr}(\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H)\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H\\
&\quad+\sigma_m^2\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H-\mathbf{P}_m\mathbf{F}_m\bar{\mathbf{H}}_m\mathbf{T}\mathbf{W}\mathbf{U}^H\mathbf{L}\mathbf{A}_m^H\\
&\quad-\mathbf{A}_m\mathbf{L}^T\mathbf{U}\mathbf{W}^H\mathbf{T}^H\bar{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\Big)+\text{tr}(\mathbf{U}\mathbf{U}^H),
\end{aligned}\tag{9}$$

where the expectation here is taken over the random variables $\{\Delta\mathbf{H}_m,\mathbf{n}_m\}$.

B. Problem formulation

1) Joint design problem: With the expression of MSE shown in (9) and the power constraint shown in (4), we are now able to formulate the proposed THP-based hybrid transceiver design problem. We aim to jointly design the digital precoding and feedback matrices in the THP structure and the analog BF matrices to minimize the MSE, hence this problem can be formulated as follows

$$\min_{\{\mathbf{P}_m,\mathbf{F}_m\},\mathbf{T},\mathbf{W},\mathbf{U}}\ \text{MSE}(\{\mathbf{P}_m,\mathbf{F}_m\},\mathbf{T},\mathbf{W},\mathbf{U})\tag{10a}$$

$$s.t.\quad |[\mathbf{F}_m]_{i,j}|=1\quad\forall m,i,j,\tag{10b}$$

$$|[\mathbf{T}]_{i,j}|=1\quad\forall i,j,\tag{10c}$$

$$\|\mathbf{T}\mathbf{W}\|^2\le P_t,\tag{10d}$$

$$\mathbf{U}\in\mathbb{L}.\tag{10e}$$

where the constant modulus constraints given by (10b) and (10c) are due to the fact that the analog beamformer is implemented using low-cost phase shifters. (10e) is the structural constraint of matrix $\mathbf{U}$, where $\mathbb{L}$ is a set of all lower triangle matrices with ones on the main diagonal.

2) Two-timescale joint design problem: In practice, the analog BF matrices can also update over a longer timescale than the digital processing matrices aiming at reducing the CSI signalling overhead needed for the exchange of CSI. Specifically, the long-timescale variables, i.e., the analog BF matrices, are designed based on the slowly varying channel statistics[4] while the short-timescale variables, i.e., the digital processing matrices, are optimized based on the instantaneous effective low-dimensional CSI matrices.

In particular, as illustrated in Fig. 2, the time axis is divided into some super-frames within which the channel statistics remains coherent. Each super-frame consists of T_f frames, each of which is made up of T_s time slots. Within each time slot, the instantaneous effective CSI remains unchanged. During the implementation of our proposed two-timescale algorithm, the long-timescale variables are updated at the end of each frame based on a channel sample, while the short-timescale variables are updated at the beginning of each time slot based on the instantaneous effective CSI. Consequently, we formulate our two-timescale optimization problem as

$$\begin{aligned}
\min_{\{\mathbf{F}_m\},\mathbf{T},\mathbf{\Theta}}\quad & f(\{\mathbf{F}_m\},\mathbf{T},\mathbf{\Theta})\triangleq\mathbb{E}_{\bar{\mathbf{H}}_m}\{\text{MSE}(\{\mathbf{F}_m\},\mathbf{T},\mathbf{\Theta})\}\\
\text{s.t.}\quad & \text{(10b)}-\text{(10e)},
\end{aligned}\tag{11}$$

where $\mathbf{\Theta}\triangleq\{\{\mathbf{P}_m\},\mathbf{W},\mathbf{U}\}$ denotes a collection of the short-timescale variables and the expectation here is taken over the channel samples $\{\bar{\mathbf{H}}_m\}$ within a super-frame.

III. Proposed Hybrid Transceiver Joint Design Algorithm

In this section, we first transform problem (10) into a more tractable form and then propose a novel iterative BCD-based algorithm to efficiently solve the converted problem. Subsequently, we carry out the convergence and computational complexity analyses for the proposed algorithm.

[4]The channel statistics refer to the distribution of channel fading realizations. We only need to obtain a single (potentially outdated) channel sample at each frame, based on which the analog BF matrices can be updated directly.

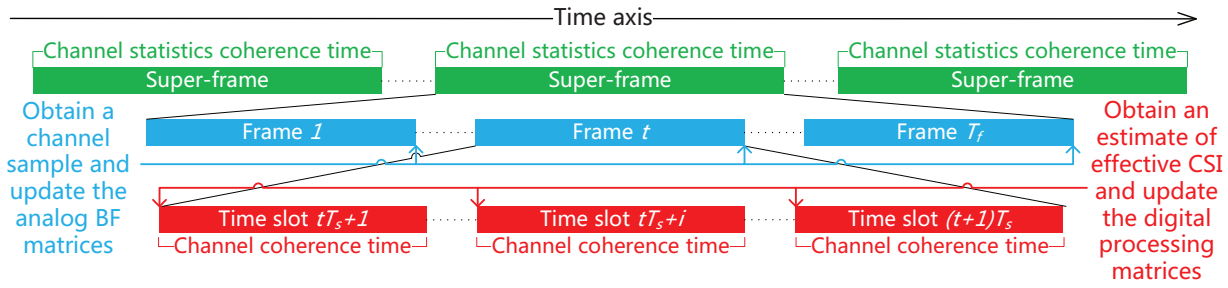

Fig. 2: Decomposition of the time axis into two timescales

A. Problem transformation

Problem (10) is hard to solve due to the highly coupled and nonconvex constraints. Hence, we provide Theorem 1 to simplify problem (10).

Theorem 1. *Let us first define the KKT solution of the following problem* (12) *as* $(\{\mathbf{P}_m^\star, \mathbf{F}_m^\star\}, \mathbf{T}^\star, \mathbf{W}^\star, \mathbf{U}^\star)$,

$$\begin{aligned} \min_{\mathcal{S}} \quad & \mathrm{MSE}_\sigma(\{\mathbf{P}_m, \mathbf{F}_m\}, \mathbf{T}, \mathbf{W}, \mathbf{U}) \\ s.t. \quad & (10b), (10c), (10e), \end{aligned} \tag{12}$$

where $\mathcal{S} \triangleq \{\{\mathbf{P}_m, \mathbf{F}_m\}, \mathbf{T}, \mathbf{W}, \mathbf{U}\}$ *and the objective function is given by*

$$\begin{aligned} &\mathrm{MSE}_\sigma(\{\mathbf{P}_m, \mathbf{F}_m\}, \mathbf{T}, \mathbf{W}, \mathbf{U}) \\ &= \sum_{m=1}^{M} \mathrm{tr}\Big(\mathbf{P}_m\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H \\ &\quad + \hat{\sigma}_{e,m}^2 \mathrm{tr}(\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H)\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H \\ &\quad + \frac{1}{P_t}\mathrm{tr}(\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H)\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H \\ &\quad - \mathbf{P}_m\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T}\mathbf{W}\mathbf{U}^H\mathbf{L}\mathbf{A}_m^H \\ &\quad - \mathbf{A}_m\mathbf{L}^T\mathbf{U}\mathbf{W}^H\mathbf{T}^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\Big) + \mathrm{tr}(\mathbf{U}\mathbf{U}^H), \end{aligned} \tag{13}$$

where $\hat{\mathbf{H}}_m \triangleq \frac{\bar{\mathbf{H}}_m}{\sigma_m}$ *denotes the scaled channel matrix and* $\hat{\sigma}_{e,m} \triangleq \frac{\sigma_{e,m}}{\sigma_m}$ *denotes the scaled channel estimation error variance.*

Then we can obtain the KKT solution of problem (10) *via scaling, i.e.,* $\bar{\mathcal{S}}^\star \triangleq (\{\bar{\mathbf{P}}_m^\star, \mathbf{F}_m^\star\}, \mathbf{T}^\star, \bar{\mathbf{W}}^\star, \mathbf{U}^\star) \triangleq (\{\frac{1}{a}\sigma_m\mathbf{P}_m^\star, \mathbf{F}_m^\star\}, \mathbf{T}^\star, a\mathbf{W}^\star, \mathbf{U}^\star)$, *where* $\bar{\mathcal{S}}^\star$ *is the KKT solution of problem* (10) *and* $a = \frac{\sqrt{P_t}}{\|\mathbf{T}^\star\mathbf{W}^\star\|}$ *denotes the scaling factor.*

Proof: See **Appendix** A. ∎

Remark: Note that in problem (10), the analog beamforming matrix $\mathbf{T}$ and the digital precoding matrix $\mathbf{W}$ are coupled in the transmit power constraint (10d), which makes it difficult to obtain a KKT solution of problem (10) directly. Typically, we can handle this coupled power constraint and the modulus constraints by penalty based algorithms, such as PDD [28] or penalty concave-convex procedure (PCCCP) [45]. However, these penalty based algorithms are double-loop algorithms which usually have a higher complexity. Theorem 1 absorbs the power constraint into the objective function and makes it possible for us to obtain a KKT solution of problem (10) via a single-loop algorithm with a relatively low complexity.

B. Proposed BCD-based joint iterative design

In this subsection, we propose an efficient BCD-based iterative algorithm to solve problem (12). The variables are partitioned into several convenient blocks which are updated sequentially at each iteration. The subproblems in each block can be solved in closed form. At this point, we partition the search variables into five blocks as follows: 1) Update $\{\mathbf{P}_m\}$ in parallel, $\forall m \in \mathcal{M}$, by fixing the variables within the other blocks; 2) Update $\mathbf{U}$ by fixing the other variables; 3) Update $\{[\mathbf{F}_m]_{i,j}\}$, $\forall i, j$, sequentially by fixing other variables, $\forall m \in \mathcal{M}$. The unit modulus constraints (10c) are handled within this block; 4) Update $[\mathbf{T}]_{i,j}$, $\forall i, j$, sequentially by fixing the other variables. The unit modulus constraints (10b) are handled within this block; 5) Update $\mathbf{W}$ by fixing the other variables. The detailed updating procedure is presented as follows.

In **Step 1**, the subproblem for $\forall m$ can be expressed as

$$\min_{\mathbf{P}_m} \quad \hat{f}_m, \tag{14}$$

where the objective function is given by

$$\begin{aligned} \hat{f}_m \triangleq \mathrm{tr}\Big(& \mathbf{P}_m\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H \\ & + \hat{\sigma}_{e,m}^2 \mathrm{tr}(\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H)\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H \\ & + \frac{1}{P_t}\mathrm{tr}(\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H)\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H \\ & - \mathbf{P}_m\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T}\mathbf{W}\mathbf{U}^H\mathbf{L}\mathbf{A}_m^H \\ & - \mathbf{A}_m\mathbf{L}^T\mathbf{U}\mathbf{W}^H\mathbf{T}^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\Big). \end{aligned} \tag{15}$$

This is an unconstrained convex optimization problem with respect to $\mathbf{P}_m$. By checking the first order optimality condition, we obtain the solution of $\mathbf{P}_m$ as

$$\begin{aligned} \mathbf{P}_m^\star = & \mathbf{A}_m\mathbf{L}^T\mathbf{U}(\mathbf{T}\mathbf{W})^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H \\ & \times \Big(\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T}\mathbf{W}(\mathbf{T}\mathbf{W})^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H \\ & + \hat{\sigma}_{e,m}^2\|\mathbf{T}\mathbf{W}\|^2\mathbf{F}_m\mathbf{F}_m^H + \frac{\|\mathbf{T}\mathbf{W}\|^2}{P_t}\mathbf{F}_m\mathbf{F}_m^H\Big)^{-1}. \end{aligned} \tag{16}$$

In **Step 2**, we seek the optimization of matrix $\mathbf{U}$ with the other variables fixed, which minimizes the MSE. This subproblem is given by

$$\begin{aligned} \min_{\mathbf{U}} \quad & \hat{g}, \\ \text{s.t.} \quad & (10e). \end{aligned} \tag{17}$$

where $\hat{g} \triangleq \sum_{m=1}^{M} -2\text{Re}\big(\text{tr}(\mathbf{P}_m\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T}\mathbf{W}\mathbf{U}^H\mathbf{L}\mathbf{A}_m^H)\big) + \text{tr}(\mathbf{U}\mathbf{U}^H)$. We can solve this subproblem by taking the partial derivative of $\hat{g}$ with respect to the elements in the strictly lower triangle area of the matrix $\mathbf{U}$.

Let us define an operator vecLT($\mathbf{X}$), which extracts the elements in the strictly lower triangle area of the square matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ and vectorizes these elements in the form of column, i.e., vecLT($\mathbf{X}$) $= [[\mathbf{X}]_{2,1}, \ldots, [\mathbf{X}]_{n,1}, [\mathbf{X}]_{3,2}, \ldots [\mathbf{X}]_{n,2}, [\mathbf{X}]_{4,3} \ldots, [\mathbf{X}]_{n,n-1}]^T$. It is readily seen that $\frac{\partial \hat{g}}{\partial \text{vecLT}(\mathbf{U})^*} = \text{vecLT}(\frac{\partial \hat{g}}{\partial \mathbf{U}^*})$. Hence, by checking the first order optimality condition, the optimal strictly lower triangle part of $\mathbf{U}$ can be given by

$$\text{vecLT}(\mathbf{U}^\star) = \text{vecLT}(\sum_{m=1}^{M} \mathbf{L}\mathbf{A}_m^H\mathbf{P}_m\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T}\mathbf{W}), \quad (18)$$

where $\mathbf{U}^\star$ denotes the optimal lower triangle matrix.

In **Step 3**, we optimize $\mathbf{F}_m, \forall m \in \mathcal{M}$, in parallel, with the other variables fixed. Specifically, the elements of $\mathbf{F}_m$, i.e., $[\mathbf{F}_m]_{i,j}$, $\forall i, j$, are optimized sequentially. The corresponding subproblem is given by

$$\min_{[\mathbf{F}_m]_{i,j}} \quad \hat{f}_m \quad (19a)$$
$$\text{s.t.} \quad |[\mathbf{F}_m]_{i,j}| = 1. \quad (19b)$$

By appropriate rearrangement, we can rewrite problem (19) as

$$\min_{[\mathbf{F}_m]_{i,j}} \quad \text{tr}(\mathbf{F}_m^H\mathbf{A}_{F_m}\mathbf{F}_m\mathbf{C}_{F_m} - 2\text{Re}(\mathbf{F}_m^H\mathbf{B}_{F_m})) \quad \text{s.t.} \quad (19b), \quad (20)$$

where $\mathbf{A}_{F_m} \triangleq \mathbf{P}_m^H\mathbf{P}_m$, $\mathbf{B}_{F_m} \triangleq \mathbf{P}_m^H\mathbf{A}_m\mathbf{L}^T\mathbf{U}\mathbf{W}^H\mathbf{T}^H\hat{\mathbf{H}}_m^H$ and $\mathbf{C}_{F_m} \triangleq \hat{\mathbf{H}}_m\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H\hat{\mathbf{H}}_m^H + \hat{\sigma}_{e,m}^2\text{tr}(\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H)\mathbf{I} + \frac{1}{P_t}\text{tr}(\mathbf{T}\mathbf{W}\mathbf{W}^H\mathbf{T}^H)\mathbf{I}$. It is readily seen that the objective function of problem (20) is a quadratic function with respect to $[\mathbf{F}_m]_{i,j}$. Thus, by omitting some constant terms in the objective function, problem (20) can be further rewritten as

$$\min_{[\mathbf{F}_m]_{i,j}} \quad \bar{a}_{F,m,i,j}|[\mathbf{F}_m]_{i,j}|^2 - \text{Re}(\bar{b}_{F,m,i,j}^*[\mathbf{F}_m]_{i,j}) \quad \text{s.t.} \quad (19b), \quad (21)$$

where $\bar{a}_{F,m,i,j}$ and $\bar{b}_{F,m,i,j}$ denote some coefficients. The optimal solution of problem (21) is given by $[\mathbf{F}_m]_{i,j}^\star = \frac{\bar{b}_{F,m,i,j}}{|\bar{b}_{F,m,i,j}|}$. Therefore, we only need to know the value of $\bar{b}_{F,m,i,j}$ to update $[\mathbf{F}_m]_{i,j}$. The value of $\bar{b}_{F,m,i,j}$ is given by

$$\bar{b}_{F,m,i,j} = [\mathbf{A}_{F,m}]_{i,i}[\mathbf{F}_m]_{i,j}[\mathbf{C}_{F,m}]_{j,j} - [\mathbf{A}_{F,m}\mathbf{F}_m\mathbf{C}_{F,m}]_{i,j} + [\mathbf{B}_{F,m}]_{i,j}. \quad (22)$$

Besides, in order to reduce the computational complexity of updating $[\mathbf{F}_m]_{i,j}$, we can update $[\mathbf{F}_m]_{i,j}$ sequentially by following similar steps in **Algorithm 3** in [23].

In **Step 4**, we optimize the elements in $\mathbf{T}$, i.e., $[\mathbf{T}]_{i,j}$, $\forall i, j$, by fixing the other variables. The corresponding subproblem is provided as

$$\min_{[\mathbf{T}]_{i,j}} \quad \text{MSE}_\sigma \quad s.t. \quad |[\mathbf{T}]_{i,j}| = 1. \quad (23)$$

This subproblem can be solved by following the same method introduced in **Step 3**. For simplicity, we omit the derivation and provide the solution of problem (23) directly,

$$\mathbf{T}t_{i,j}^\star = \frac{\bar{b}_{T,i,j}}{|\bar{b}_{T,i,j}|}, \quad (24)$$

where $\bar{b}_{T,i,j} = [\mathbf{A}_T]_{i,i}[\mathbf{T}]_{i,j}[\mathbf{C}_T]_{j,j} - [\mathbf{A}_T\mathbf{T}\mathbf{C}_T]_{i,j} + [\mathbf{B}_T]_{i,j}$, $\mathbf{A}_T \triangleq \sum_{m=1}^{M}(\bar{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\mathbf{P}_m\mathbf{F}_m\bar{\mathbf{H}}_m + \sigma_{e,m}^2\text{tr}(\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H)\mathbf{I} + \frac{\sigma_m^2}{P_t}\text{tr}(\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H)\mathbf{I})$, $\mathbf{C}_T \triangleq \mathbf{W}\mathbf{W}^H$ and $\mathbf{B}_T \triangleq \sum_{m=1}^{M}\bar{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\mathbf{A}_m\mathbf{L}^T\mathbf{U}\mathbf{W}^H$.

In **Step 5**, we optimize the variable $\mathbf{W}$ with the other variables fixed. We need to solve the following convex subproblem

$$\min_{\mathbf{W}} \quad \text{MSE}_\sigma. \quad (25)$$

By checking the first order optimality condition, we obtain the optimal solution of this subproblem as

$$\begin{aligned}\mathbf{W}^\star = &\Big(\sum_{m=1}^{M}\mathbf{T}^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\mathbf{P}_m\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T} \\ &+ \text{tr}(\Sigma_{m=1}^{M}\frac{1}{Pt}\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H)\mathbf{T}^H\mathbf{T} \\ &+ \sum_{m=1}^{M}\hat{\sigma}_{e,m}^2\text{tr}(\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H)\mathbf{T}^H\mathbf{T}\Big)^{-1} \\ &\times(\sum_{m=1}^{M}\mathbf{T}^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\mathbf{A}_m\mathbf{L}^T\mathbf{U}).\end{aligned} \quad (26)$$

In each iteration of the proposed BCD-based algorithm, we implement the above five steps to update the optimization variables. The overall procedure of the proposed algorithm is summarized in **Algorithm 1**.[5]

Algorithm 1 Proposed BCD-based algorithm for nonlinear hybrid transceiver design

1. Define the accuracy tolerance δ. Initialize all the variables in $\mathcal{S}$ with a feasible point. Set the iteration number $i = 0$.
2. **Repeat**
   - 2.1 Update $\mathbf{P}_m, \forall m \in \mathcal{M}$, in parallel based on (16).
   - 2.2 Update $\mathbf{U}$ based on (18).
   - 2.3 Update $\mathbf{F}_m, \forall m \in \mathcal{M}$, in parallel. In particular, the elements of $\mathbf{F}_m$ are optimized sequentially based on the method introduced in **Step 3**.
   - 2.4 Update the elements of $\mathbf{T}$ sequentially based on the method introduced in **Step 3**.
   - 2.5 Update $\mathbf{W}$ based on (26).
   - 2.6 Update the iteration number: $i = i + 1$.
3. **Until** the difference between two successive objective value is less than δ.

*C. Convergence and complexity of **Algorithm 1***

In this subsection, we analyze the convergence and the computational complexity of the proposed algorithm for the nonlinear hybrid transceiver joint design.

[5]We can obtain the KKT point of the original problem (10) via scaling: $(\{\bar{\mathbf{P}}_m^\star, \mathbf{F}_m^\star\}, \mathbf{T}^\star, \bar{\mathbf{W}}^\star, \mathbf{U}^\star) = (\{\frac{1}{a}\sigma_m\mathbf{P}_m^\star, \mathbf{F}_m^\star\}, \mathbf{T}^\star, a\mathbf{W}^\star, \mathbf{U}^\star)$.

The following proposition demonstrates the convergence of **Algorithm 1**.

Proposition 1. ***Algorithm 1*** *can be guaranteed to converge to a KKT point of problem* (12)*.*

Proof: From proposition 2.7.1 (Convergence of Block Coordinate Descent) in [46], we know that the BCD algorithm will finally converge to a KKT point if the objective function is continuously differentiable and each subproblem of the blocks can be solved uniquely and globally. It is readily seen that MSE_σ is a continuously differentiable function with respect to all the variables and each subproblem in **Algorithm 1** can be solved uniquely and globally in closed form. Thus our proposed **Algorithm 1** can be guaranteed to converge to a KKT point of problem (12). ■

The complexity of the proposed BCD-based algorithm is dominated by the inversion operations in **Step 1** and **Step 5** and the multiplications in **Step 3** and **Step 4**, of which the complexities are $\mathcal{O}(R_{d,m}^3)$, $\mathcal{O}(R_s^3)$, $\mathcal{O}(N_{d,m}^2 R_{d,m}^2)$ and $\mathcal{O}(N_s^2 R_s^2)$, respectively. Therefore, by omitting the lower order terms, the complexity of our proposed BCD-based algorithm is given by $\mathcal{O}(I(N_s^2 R_s^2))$, where I denotes the maximum iteration number of the proposed BCD-based algorithm.

Remark: Note that for a FD system, the corresponding BCD algorithm has a complexity of $\mathcal{O}(N_s^3)$ due to the inverse operation for calculating the BF matrix at the BS. Thus, when the number of antennas is much larger than that of RF chains at the BS, i.e., $N_s >> R_s$, the computational complexity of our proposed algorithm can be lower than that of the FD system.

IV. Proposed two-timescale Hybrid Transceiver joint design algorithm

In order to reduce the CSI signalling overhead and the effects of outdated CSI caused by the associated delays, in this section we propose a novel two-timescale nonlinear hybrid transceiver design algorithm. In this scheme, the long-timescale analog BF matrices are optimized based on the channel statistics and the short-timescale digital processing matrices are designed based on the instantaneous low-dimensional effective CSI matrices. Based on the TOSCA framework [47], we can see that problem (11) can be decomposed into a long-timescale master problem and a short-timescale subproblem.

A. Short-timescale subproblem

By fixing the long-timescale variables $\{\mathbf{F}_m\}$ and $\mathbf{T}$, the short-timescale subproblem is given by

$$\begin{aligned} \min_{\mathbf{W},\mathbf{U},\{\mathbf{P}_m\}} \quad & \text{MSE}(\{\mathbf{P}_m\},\mathbf{W},\mathbf{U}) \\ \text{s.t.} \quad & \text{(10d)},\text{(10e)}. \end{aligned} \tag{27}$$

Note that Theorem 1 can be also applied to this subproblem similarly. This subproblem can be transformed into the following problem,

$$\begin{aligned} \min_{\mathbf{W},\mathbf{U},\{\mathbf{P}_m\}} \quad & \text{MSE}_\sigma(\{\mathbf{P}_m\},\mathbf{W},\mathbf{U}) \\ \text{s.t.} \quad & \text{(10e)}. \end{aligned} \tag{28}$$

We can solve this converted short-timescale problem based on a BCD algorithm which is similar to **Algorithm 1** (without Step 2.3 and Step 2.4). Then, we obtain the solution of the short-timescale subproblem (27) via scaling: $(\{\bar{\mathbf{P}}_m^\star\},\bar{\mathbf{W}}^\star,\mathbf{U}^\star) = (\{\frac{1}{a}\sigma_m \mathbf{P}_m^\star\}, a\mathbf{W}^\star, \mathbf{U}^\star)$.

B. Long-timescale master problem

By fixing the short-timescale variables, the long-timescale master problem is given by

$$\min_{\boldsymbol{\theta}_T,\{\boldsymbol{\theta}_{F_m}\}} \quad \tilde{f}(\boldsymbol{\theta}_T,\{\boldsymbol{\theta}_{F_m}\},\boldsymbol{\Theta}^\star) = \mathbb{E}_{\bar{\mathbf{H}}_m}\{g(\boldsymbol{\theta}_T,\{\boldsymbol{\theta}_{F_m}\},\boldsymbol{\Theta}^\star)\} \tag{29}$$

where $\boldsymbol{\theta}_{F_m} \triangleq \angle\mathbf{F}_m, \forall m$, $\boldsymbol{\theta}_T \triangleq \angle\mathbf{T}$, $\boldsymbol{\Theta}^\star \triangleq \{\{\bar{\mathbf{P}}_m\}^\star, \bar{\mathbf{W}}^\star, \mathbf{U}^\star\}$ denotes the solution of problem (27) and

$$g(\boldsymbol{\theta}_T,\{\boldsymbol{\theta}_{F_m}\},\boldsymbol{\Theta}^\star) \triangleq \text{MSE}(\{\mathbf{P}_m,\mathbf{F}_m\},\mathbf{T},\mathbf{W},\mathbf{U}). \tag{30}$$

To solve problem (29), the long-timescale variables are updated at the end of each frame by solving an approximation problem obtained via replacing the objective function of problem (29) with a quadratic surrogate function. Thus, we introduce the following quadratic surrogate function to approximate the objective function for frame t:

$$\begin{aligned} &\bar{f}^t(\boldsymbol{\theta}_T,\{\boldsymbol{\theta}_{F_m}\},\boldsymbol{\Theta}^{\star,t}) \\ =&f^t + (\mathbf{f}_T^t)^T(\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^t) + \sum_{m=1}^M (\mathbf{f}_{F_m}^t)^T(\boldsymbol{\theta}_{F_m} - \boldsymbol{\theta}_{F_m}^t) \\ &+ \tau\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^t\|^2 + \sum_{m=1}^M \tau\|\boldsymbol{\theta}_{F_m} - \boldsymbol{\theta}_{F_m}^t\|^2, \end{aligned} \tag{31}$$

where $\boldsymbol{\Theta}^{\star,t}$ denotes the solution of solving problem (27) with given $\{\bar{\mathbf{H}}_m^t\}$, $\boldsymbol{\theta}_T^t$ and $\boldsymbol{\theta}_{F_m}^t$. $\tau > 0$ is a constant. f^t, $\mathbf{f}_T^t$ and $\mathbf{f}_{F_m}^t$ denote the approximations of objective function $\tilde{f}$, the partial derivatives $\frac{\partial \tilde{f}}{\partial \boldsymbol{\theta}_T}$ and $\frac{\partial \tilde{f}}{\partial \boldsymbol{\theta}_{F_m}}$, respectively, based on the current channel sample $\{\bar{\mathbf{H}}_m^t\}$ and $\boldsymbol{\Theta}^{\star,t}$. The quantities can be updated based on the following expressions:

$$f^t = (1-\rho^t)f^{t-1} + \rho^t g(\boldsymbol{\theta}_T^t,\{\boldsymbol{\theta}_{F_m}^t\},\boldsymbol{\Theta}^{\star,t}), \tag{32}$$

$$\mathbf{f}_T^t = (1-\rho^t)\mathbf{f}_T^{t-1} + \rho^t \frac{\partial g}{\partial \boldsymbol{\theta}_T}|_{(\boldsymbol{\theta}_T^t,\{\boldsymbol{\theta}_{F_m}^t\},\boldsymbol{\Theta}^{\star,t})}, \tag{33}$$

and

$$\mathbf{f}_{F_m}^t = (1-\rho^t)\mathbf{f}_{F_m}^{t-1} + \rho^t \frac{\partial g}{\partial \boldsymbol{\theta}_{F_m}}|_{(\boldsymbol{\theta}_T^t,\{\boldsymbol{\theta}_{F_m}^t\},\boldsymbol{\Theta}^{\star,t})}. \tag{34}$$

The details of the derivatives are given in **Appendix B**. Here $\{\rho^t\}$ is a sequence of parameters satisfying condition (38).

Subsequently, let us solve the approximated problem for (29), which is given by

$$\min_{\boldsymbol{\theta}_T,\{\boldsymbol{\theta}_{F_m}\}} \bar{f}^t(\boldsymbol{\theta}_T,\{\boldsymbol{\theta}_{F_m}\},\boldsymbol{\Theta}^{\star,t}). \tag{35}$$

It is readily seen that (35) can be solved as follows

$$\bar{\boldsymbol{\theta}}_T = \boldsymbol{\theta}^t - \frac{\mathbf{f}_T^t}{2\tau}, \quad \bar{\boldsymbol{\theta}}_{F_m} = \boldsymbol{\theta}_{F_m}^t - \frac{\mathbf{f}_{F_m}^t}{2\tau}, \forall m, \tag{36}$$

where $\bar{\boldsymbol{\theta}}_T$ and $\{\bar{\boldsymbol{\theta}}_{F_m}\}$ are the optimal solution of the quadratic approximation problem (35).

Then, the long-timescale variables are updated as

$$\begin{aligned} \boldsymbol{\theta}_T^{t+1} &= (1-\gamma^t)\boldsymbol{\theta}_T^t + \gamma^t\bar{\boldsymbol{\theta}}_T, \\ \boldsymbol{\theta}_{F_m}^{t+1} &= (1-\gamma^t)\boldsymbol{\theta}_{F_m}^t + \gamma^t\bar{\boldsymbol{\theta}}_{F_m}, \forall m, \end{aligned} \tag{37}$$

where $\{\gamma^t\}$ denotes a sequence of parameters satisfying condition (38). The proposed two-timescale joint design algorithm is summarized in **Algorithm 2**.

Algorithm 2 Proposed TOSCA-based two-timescale joint design algorithm

1. A super-frame starts. Initialize the long-timescale variables $\{\boldsymbol{\theta}_T^0, \{\boldsymbol{\theta}_{F_m}^0\}\}$ and the short-timescale variables $\{\mathbf{W}^0, \mathbf{U}^0, \{\mathbf{P}_m^0\}\}$ to a feasible point. Set the frame index $t = 0$ and the time slot index $i = 0$.
2. **Repeat**
   2.1 Obtain the effective CSI matrices $\{\tilde{\mathbf{H}}_m^i\}$ for time slot i.
   2.2 Solve problem (27) and obtain the solution $\{\bar{\mathbf{W}}^i, \mathbf{U}^i, \{\bar{\mathbf{P}}_m^i\}\}$.
   2.3 Update the time slot index: $i = i + 1$.
   Until the frame ends, i.e. $i = (t+1)T_s$.
3. Obtain a CSI sample $\{\bar{\mathbf{H}}_m^t\}$ at the end of frame t.
4. Update the surrogate function (31) using $\boldsymbol{\Theta}^{\star,t}$, $\{\boldsymbol{\theta}_T^t, \{\boldsymbol{\theta}_{F_m}^t\}\}$ and $\{\bar{\mathbf{H}}_m^t\}$.
5. Solve (35) to obtain $\{\bar{\boldsymbol{\theta}}_T, \bar{\boldsymbol{\theta}}_{F_m}\}$.
6. Update $\{\boldsymbol{\theta}_T^{t+1}, \{\boldsymbol{\theta}_{F_m}^{t+1}\}\}$ according to (37).
7. Set $t = t + 1$ and return to Step 2.

According to [47], if we choose the sequences of the parameters $\{\rho^t, \gamma^t\}$ so that they satisfy the following condition

$$\begin{aligned}&\rho^t \to 0, \frac{1}{\rho^t} \le O(t^\beta) \text{ for some } \beta \in (0,1), \sum_t (\rho^t)^2 < \infty,\\ &\gamma^t \to 0, \sum_t \gamma^t = \infty, \sum_t (\gamma^t)^2 < \infty, \lim_{t\to\infty} \frac{\gamma^t}{\rho^t} = 0.\end{aligned} \tag{38}$$

then our proposed two-timescale algorithm can be guaranteed to converge to a KKT solution of problem (11). The overall computational complexity of **Algorithm 2** is dominated by the updating of the short-timescale variables, which is given by $\mathcal{O}(T_s I(R_s^3))$, where I denotes the maximum iteration number of the proposed BCD-based algorithm for the short-timescale subproblem.

Remark: Note that the design of the short-timescale digital processing matrices only requires the effective CSI matrices $\tilde{\mathbf{H}}_m$, which can be obtained by pre-multiplying and post-multiplying $\hat{\mathbf{H}}_m^i$ with $\mathbf{F}_m^t$ and $\mathbf{T}^t$, respectively, i.e., $\mathbf{F}_m^t \hat{\mathbf{H}}_m^i \mathbf{T}^t = \tilde{\mathbf{H}}_m^i$. The effective CSI matrices $\{\tilde{\mathbf{H}}_m\}$ have much lower dimension than the instantaneous estimated CSI matrices $\{\hat{\mathbf{H}}_m\}$ and only a small number of CSI samples are needed for updating the long-timescale variables, thus the overhead of estimating CSI can be significantly reduced.

V. Design of cancellation ordering matrix

The cancellation order of the proposed THP-based nonlinear hybrid transceiver design affects the system performance. In order to further increase the performance, in this section we seek to design the cancellation ordering matrix $\mathbf{L}$.[6]

[6]In this work, we mainly focus on the cancellation order among the users rather than that among the data streams, since the channel conditions related to different antennas per user are quite similar.

However, it is very difficult to design the optimal cancellation ordering matrix $\mathbf{L}$ due to the MSE expression with unknown optimization variables. Thus, we seek to develop a low-complexity approach to find the near-optimal cancellation order based on a comparable lower bound of MSE, which does not contain the coupled terms of $\mathbf{L}$ and other variables that would affect the objective value. To this end, we derive a lower bound for MSE_σ and then design the matrix $\mathbf{L}$ based on the lower bound.

The following lemma 1 gives the expression of the desired lower bound of MSE_σ, based on which we can decide the cancellation order for the THP structure in a low-complexity manner.

Lemma 1. *The objective function MSE_σ has the following lower bound,*

$$MSE_\sigma(\mathcal{P}, \mathbf{U}; \mathbf{L}) \ge g(\mathcal{P}; \mathbf{L}) - C_1^2 (\sum_{i=1}^{M} i \|\hat{\mathbf{H}}_{(i)}\|^2), \tag{39}$$

where the subscript (i) of $\hat{\mathbf{H}}_{(i)}$ indicates the permutated cancellation order of the users and is generated by using $\mathbf{L}$ to permutate the cancellation order of the users at the THP structure, $\mathcal{P} \triangleq \{\{\mathbf{P}_m, \mathbf{F}_m\}, \mathbf{T}, \tilde{\mathbf{W}}\}$ denotes a set of variables, C_1 is a positive constant and $g(\mathcal{P}; \mathbf{L})$ is expressed as

$$g(\mathcal{P}; \mathbf{L}) \triangleq \|\mathbf{L}\bar{\mathbf{A}}\mathbf{T}\tilde{\mathbf{W}}\mathbf{L}^T\|^2 + \tilde{g}(\mathcal{P}) = \|\bar{\mathbf{A}}\mathbf{T}\tilde{\mathbf{W}}\|^2 + \tilde{g}(\mathcal{P}). \tag{40}$$

Here, we have $\tilde{\mathbf{W}} = \mathbf{W}\mathbf{L}$, $\bar{\mathbf{A}} \triangleq [(\mathbf{P}_1\mathbf{F}_1\hat{\mathbf{H}}_1)^T, (\mathbf{P}_2\mathbf{F}_2\hat{\mathbf{H}}_2)^T, \ldots, (\mathbf{P}_M\mathbf{F}_M\hat{\mathbf{H}}_M)^T]^T$ and

$$\begin{aligned}\tilde{g}(\mathcal{P}) \triangleq \sum_{m=1}^{M} tr\Big(&\hat{\sigma}_{e,m}^2 tr(\mathbf{T}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^H\mathbf{T}^H)\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H\\ &+ tr(\mathbf{T}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^H\mathbf{T}^H)\mathbf{P}_m\mathbf{F}_m\mathbf{F}_m^H\mathbf{P}_m^H\\ &- \mathbf{P}_m\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T}\tilde{\mathbf{W}}\mathbf{A}_m^H\\ &- \mathbf{A}_m\tilde{\mathbf{W}}^H\mathbf{T}^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\Big) + D.\end{aligned} \tag{41}$$

Proof: Let us rewrite the expression of the original MSE_σ as follows

$$MSE_\sigma(\mathcal{P}, \mathbf{U}; \mathbf{L}) = g(\mathcal{P}; \mathbf{L}) + f(\mathcal{P}, \mathbf{C}; \mathbf{L}) \tag{42}$$

where

$$\begin{aligned}f(\mathcal{P}, \mathbf{C}; \mathbf{L}) \triangleq \sum_{m=1}^{M} \text{tr}\Big(&- \mathbf{P}_m\mathbf{F}_m\hat{\mathbf{H}}_m\mathbf{T}\tilde{\mathbf{W}}\mathbf{L}^T\mathbf{C}^H\mathbf{L}\mathbf{A}_m^H\\ &- \mathbf{A}_m\mathbf{L}^T\mathbf{C}\mathbf{L}\tilde{\mathbf{W}}^H\mathbf{T}^H\hat{\mathbf{H}}_m^H\mathbf{F}_m^H\mathbf{P}_m^H\Big)\\ &+ \text{tr}(\mathbf{C}\mathbf{C}^H).\end{aligned} \tag{43}$$

By recalling the solution of $\mathbf{U}$ in (18) and substituting $\mathbf{C}^\star = \mathbf{U}^\star - \mathbf{I}$ into $f(\mathcal{P}, \mathbf{U}; \mathbf{L})$, we obtain

$$f(\mathcal{P}, \mathbf{C}^\star; \mathbf{L}) = -\|\Delta(\mathbf{L}\bar{\mathbf{A}}\mathbf{T}\tilde{\mathbf{W}}\mathbf{L}^T)\|^2, \tag{44}$$

where the operation $\Delta(.)$ is defined as that $\Delta(\mathbf{X})$ extracts the elements in the strictly lower triangle area of the square matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ and forms a strictly lower triangle matrix, i.e.

$$[\Delta(\mathbf{X})]_{i,j} = \begin{cases}[\mathbf{X}]_{i,j}, & \text{if } i > j,\\ 0, & \text{otherwise.}\end{cases} \tag{45}$$

In the following, we derive the comparable lower bound of $f(\mathcal{P}, \mathbf{C}^\star; \mathbf{L})$. First, let us define $\bar{\mathbf{B}} \triangleq \mathbf{TW} = [\bar{\mathbf{B}}_1, \bar{\mathbf{B}}_2, ..., \bar{\mathbf{B}}_M] \in \mathbb{C}^{N_s \times D}$, where $\bar{\mathbf{B}}_m \in \mathbb{C}^{N_s \times D_m}$ denotes a submatrix of $\bar{\mathbf{B}}$, which is formulated from the $(\sum_{i=1}^{m-1} D_i + 1)$th column vector to the $(\sum_{i=1}^{m} D_i)$th column vector of matrix $\bar{\mathbf{B}}$, then we have $f(\mathcal{P}, \mathbf{C}^\star; \mathbf{L}) = -\|\Delta(\mathbf{L}\bar{\mathbf{A}}\bar{\mathbf{B}})\|^2 = -\|\Delta(\hat{\mathbf{A}}\bar{\mathbf{B}})\|^2 = -\|\Delta(\mathbf{\Omega})\|^2$, where $\hat{\mathbf{A}} \triangleq \mathbf{L}\bar{\mathbf{A}} = [(\mathbf{P}_{(1)}^\star \mathbf{F}_{(1)}^\star \hat{\mathbf{H}}_{(1)})^T, (\mathbf{P}_{(2)}^\star \mathbf{F}_{(2)}^\star \hat{\mathbf{H}}_{(2)})^T, \ldots, (\mathbf{P}_{(M)}^\star \mathbf{F}_{(M)}^\star \hat{\mathbf{H}}_{(M)})^T]^T \in \mathbb{C}^{D \times N_s}$ denotes a matrix obtained by permutating the rows of $\bar{\mathbf{A}}$ with the ordering matrix $\mathbf{L}$. $\mathbf{\Omega} = \hat{\mathbf{A}}\bar{\mathbf{B}}$ can be structured as

$$\mathbf{\Omega} \triangleq \begin{bmatrix} \mathbf{\Omega}_{(1),1} & \mathbf{\Omega}_{(1),2} & \ldots & \mathbf{\Omega}_{(1),M} \\ \mathbf{\Omega}_{(2),1} & \mathbf{\Omega}_{(2),2} & \ldots & \mathbf{\Omega}_{(2),M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Omega}_{(M),1} & \mathbf{\Omega}_{(M),2} & \ldots & \mathbf{\Omega}_{(M),M} \end{bmatrix}, \quad (46)$$

where $\mathbf{\Omega}_{(i),k} \triangleq \mathbf{P}_{(i)}^\star \mathbf{F}_{(i)}^\star \hat{\mathbf{H}}_{(i)} \bar{\mathbf{B}}_k \in \mathbb{C}^{D_{(i)} \times D_k}, \forall i, k$.

Then we can rewrite $f(\mathcal{P}, \mathbf{C}^\star; \mathbf{L})$ as follows

$$f(\mathcal{P}, \mathbf{C}^\star; \mathbf{L}) = -\sum_{i>k} \|\mathbf{\Omega}_{(i),k}\|^2 - \sum_{i=1}^{M} \|\Delta(\mathbf{\Omega}_{(i),i})\|^2. \quad (47)$$

Let us define C_1 as the upper bound of $\|\mathbf{P}_m \mathbf{F}_m\| \|\bar{\mathbf{B}}_k\|$, i.e., $\|\mathbf{P}_m \mathbf{F}_m\| \|\bar{\mathbf{B}}_k\| \leq C_1, \forall m, k$. Then, we have the upper bound for $\|\mathbf{\Omega}_{(i),k}\|^2$ as

$$\|\mathbf{\Omega}_{(i),k}\|^2 \leq (\|\mathbf{P}_{(i)}^\star \mathbf{F}_{(i)}^\star\| \|\hat{\mathbf{H}}_{(i)}\| \|\bar{\mathbf{B}}_k\|)^2 \leq C_1^2 \|\hat{\mathbf{H}}_{(i)}\|^2. \quad (48)$$

Based on (47) and (48), we obtain the lower bound for $f(\mathcal{P}, \mathbf{C}^\star; \mathbf{L})$ as

$$\begin{aligned} f(\mathcal{P}, \mathbf{C}^\star; \mathbf{L}) &\geq -\sum_{i>k} \|\mathbf{\Omega}_{(i),k}\|^2 - \sum_{i} \|\mathbf{\Omega}_{(i),i}\|^2 \\ &\geq -C_1^2 (\sum_{i=1}^{M} i \|\hat{\mathbf{H}}_{(i)}\|^2). \end{aligned} \quad (49)$$

Finally, by substituting (49) into (42), we have (39). ■

Remark: From the proof above, we can see how the THP structure gains in terms of MSE. Based on (44), (40) and (42), we can see that $g(\mathcal{P}; \mathbf{L})$ is the MSE of a linear hybrid transceiver, and it is slightly affected by the cancellation ordering matrix $\mathbf{L}$. The MSE gain of the THP comes from $f(\mathcal{P}, \mathbf{C}^\star; \mathbf{L})$. Thus, the MSE performance of the THP-based hybrid transceiver always outperforms that of its linear counterpart no matter what the cancellation ordering matrix $\mathbf{L}$ is. Note that the matrix $\Delta(\mathbf{L}\bar{\mathbf{A}}\mathbf{T}\tilde{\mathbf{W}}\mathbf{L}^T)$ is a strictly lower triangle matrix and the elements in its upper triangle area are forced to zero. The permutation matrix $\mathbf{L}$ can change the positions of the elements to reduce the value of $MSE_\sigma(\mathcal{P}, \mathbf{U}^\star; \mathbf{L})$.

Since the permutation matrix $\mathbf{L}$ does not change the minimum value of $g(\mathcal{P}; \mathbf{L})$, it is readily seen that the cancellation order can be generated from the smallest value to the largest value based on the sequence $\|\hat{\mathbf{H}}_1\|^2, \|\hat{\mathbf{H}}_2\|^2, \ldots, \|\hat{\mathbf{H}}_M\|^2$ aiming at minimizing the lower bound of MSE (39). Thus, the cancellation ordering matrix $\mathbf{L}$ can be straightforwardly formulated based on this order.

VI. Simulation results

In this section, we evaluate the performance of the proposed THP-based hybrid transceiver joint design algorithms. We consider the widely used narrow-band mmWave channel model with the uniform linear antenna array configuration [4]. The channel matrix between user m and the BS is given by

$$\begin{aligned} \mathbf{H}_m = &\sum_{n_{cl}}^{N_{cl}} \sum_{p}^{N_p} \Gamma_{n_{cl}} \mathbf{a}(\theta_{n_{cl}}^t + \psi_{n_{cl},n_p}^t) \mathbf{a}^H(\theta_{n_{cl}}^r + \psi_{n_{cl},n_p}^r) \\ &\times \exp(j2\pi f_d \tau \cos(\theta_{n_{cl}}^r + \psi_{n_{cl},n_p}^r)), \end{aligned} \quad (50)$$

where N_{cl} and N_p are the number of aggregated clusters and the number of rays within the cluster p, respectively, and $\Gamma_{n_{cl}} \sim \mathcal{CN}(0, 1)$ represents the complex channel gain of the cluster n_{cl}. $\theta_{n_{cl}}^t$ and ψ_{n_{cl},n_p}^t are the n_{cl}-th cluster's central angle of departure and bias angles of departure due to the angle spread, correspondingly, while $\theta_{n_{cl}}^r$ and ψ_{n_{cl},n_p}^r are the n_{cl}-th cluster's counterparts of the angles of arrival. f_d is the maximum Doppler shift, τ is the delay, and $\mathbf{a}(\theta)$ is the array response vector whose generic expression can be given by

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{N}} [1, e^{jk_o d_a \pi \sin(\theta)}, ..., e^{jk_o d_a \pi (N-1) \sin(\theta)}]^T, \quad (51)$$

where $k_o = 2\pi/\lambda_o$, λ_o is the wavelength at the operating frequency and d_a is the antenna spacing. We assume that there are 3 clusters and 5 rays within each cluster, i.e., totally 15 rays as in [23]. Besides, we limit $\boldsymbol{\theta}_{n_{cl}}^t$ and $\boldsymbol{\theta}_{n_{cl}}^r$ in a range of $(-\frac{\pi}{8}, \frac{\pi}{8})$ in order to simulate the scenario where the users are close to each other and set, unless specified, $\sigma_{e,m} = \sigma_e = 0.1$, $\forall m$.

We assume that there are $M = 4$ users and each user is equipped with $N_{d,m} = 8$ antennas and $R_{d,m} = 2$ RF chains, while the BS has $N_s = 32$ antennas and $R_s = 8$ RF chains. We employ the 16-QAM modulation. The number of data streams for each user is set to $D_m = 2, \forall m$, hence the number of data streams at the BS is $D = MD_m = 8$. The level of noise variance is normalized to $\sigma_m = 1, \forall m$. The signal-to-noise ratio (SNR) is defined as SNR$= 10\log_{10} \frac{P_t}{\sigma_m^2}$dB. We consider the following algorithms for comparison:

- **Nonlinear joint**: The proposed BCD-based algorithm (**Algorithm 1**) for the THP-based joint hybrid transceiver design.
- **Nonlinear separate**: The analog BF matrices are first obtained by using the channel matching approach as in [24]. The THP-based digital processing matrices are optimized jointly.
- **FD**: The proposed THP-based fully digital transceiver design algorithm.
- **Linear joint**: The joint linear hybrid transceiver design algorithm proposed in [23].
- **Linear separate**: The analog BF matrices are first obtained by using the channel matching approach as in [24]. The linear transceiver matrices are optimized jointly.
- **ZF**: The analog BF matrices are first obtained by using the channel matching approach as in [24]. The digital transceiver matrices are designed based on the conventional zero-forcing (ZF) BF.

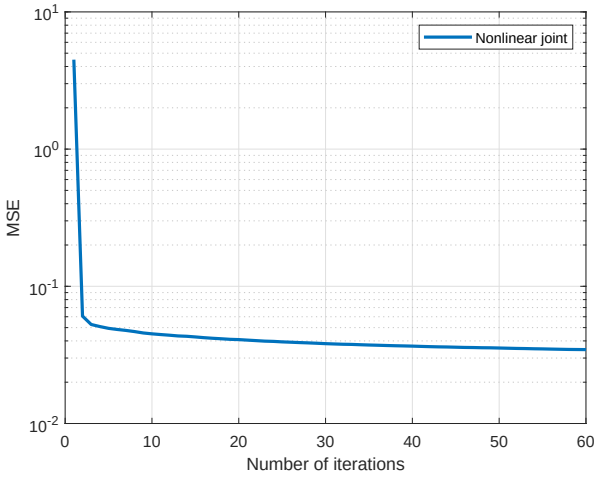

Fig. 3: Convergence performance of the proposed algorithm (SNR= 20dB).

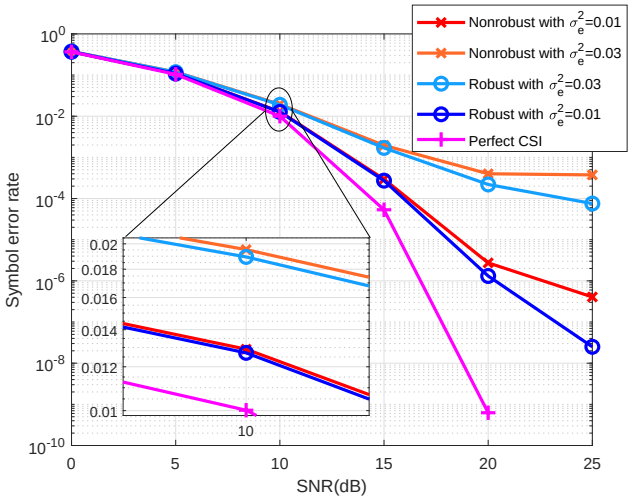

Fig. 5: SER performance comparison of the proposed nonlinear hybrid transceiver design algorithm with different variances of CSI errors.

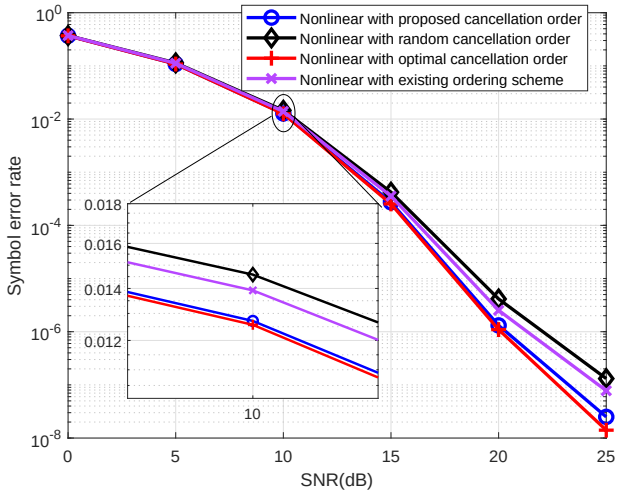

Fig. 4: SER performance comparison of the proposed nonlinear hybrid transceiver design algorithm with different cancellation ordering schemes.

- **Two-timescale joint**: The proposed TOSCA-based two-timescale joint design algorithm.

Remark: Except the TOSCA-based two-timescale joint design algorithm, all the analyzed designs are single-timescale algorithms.

A. Single-timescale joint design algorithm

We first study the convergence performance of this proposed **Algorithm 1**. Fig. 3 shows the MSE performance versus the number of iterations, where the SNR is set to 20dB. It can be observed that the objective value of the optimization problem nearly converges within less than 50 iterations, which indicates the convergence behaviour of the proposed BCD-based joint design algorithm. Moreover, the proposed algorithm provides relatively low complexity due to the closed-form solutions in each block.

Then, we investigate the effect of different cancellation ordering schemes on the symbol error rate (SER) performance. Four different cancellation ordering schemes are taken into consideration, including, the proposed ordering scheme, the random ordering scheme, the optimal ordering scheme and the ordering scheme proposed in [48][7]. Fig. 4 shows the SER performance of the analyzed algorithms. The results indicate that our proposed cancellation ordering scheme approaches the performance of the optimal scheme. Besides, the proposed ordering scheme renders significant SER gain over the existing ordering scheme and the random ordering scheme for the high SNR region. The existing ordering scheme is affected by its ZF precoding setting. The results verify the effectiveness of the proposed ordering scheme.

Next, we show the SER performance of the proposed BCD-based joint design algorithm in the presence of different variances of CSI errors. As shown in Fig. 5, a smaller variance of CSI errors leads to better SER performance for the proposed robust nonlinear transceiver design algorithm and its nonrobust counterpart.[8] With the same channel estimation error variance, the SER performance of the proposed robust design algorithm is always better than that of the nonrobust design algorithm. Furthermore, the gap of the SER performance between the robust and nonrobust design algorithms increases as the increasing of SNR. The results verify the robustness of the proposed nonlinear transceiver design algorithm.

Fig. 6 shows the SER performance versus SNR for different transceiver design algorithms, including the ZF algorithm, linear joint design algorithm, proposed nonlinear joint design algorithm, linear separate design algorithm and nonlinear separate design algorithm. We observe that, as expected, the nonlinear transceiver design algorithms provide better SER

[7]For the random ordering scheme, the cancellation order is generated randomly and for the optimal ordering scheme, the order is obtained via exhaustive search.

[8]The nonrobust transceiver design algorithm updates the optimization variables only based on the estimated CSI matrices $\{\bar{\mathbf{H}}_m\}$ without considering the channel estimation errors.

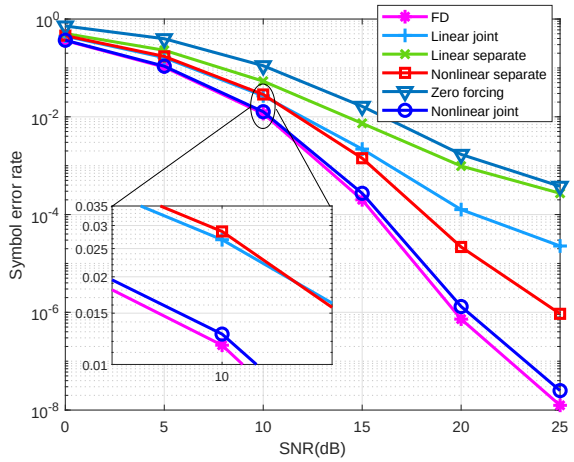

Fig. 6: SER performance comparison for different transceiver design algorithms.

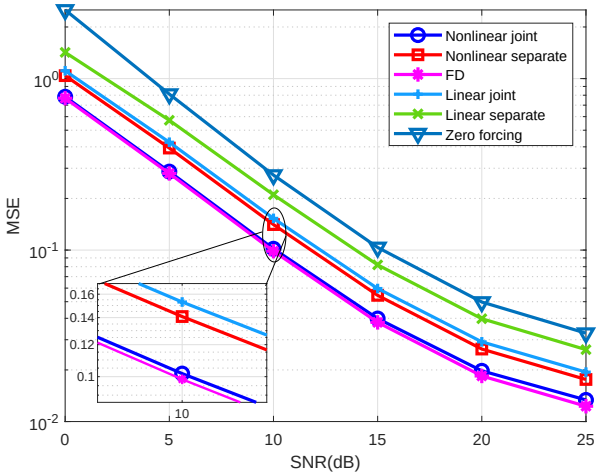

Fig. 7: MSE performance comparison for different transceiver design algorithms.

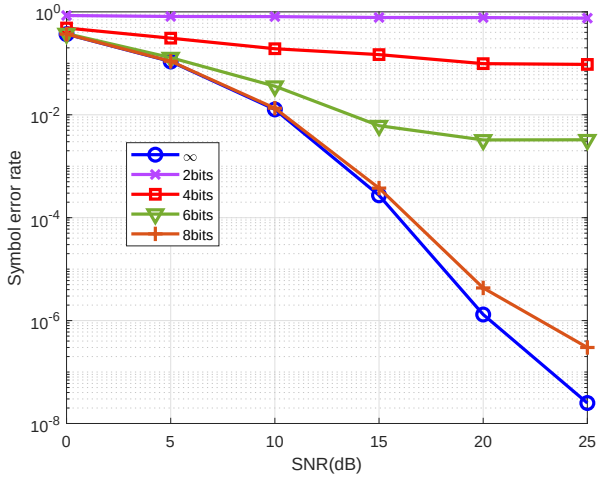

Fig. 8: SER performance comparison of the proposed nonlinear hybrid transceiver design algorithm with different numbers of quantization bits of phase shifters.

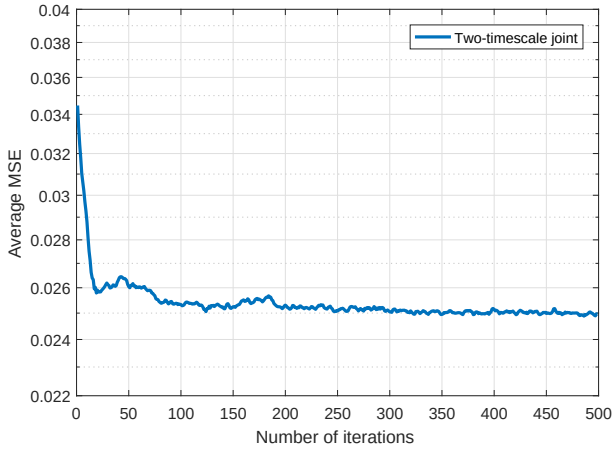

Fig. 9: Convergence performance of the proposed two-timescale joint design algorithm (SNR= 20dB).

performance compared to the linear transceiver design algorithms all the time due to the successive interference suppression based preprocessing. Besides, the linear and nonlinear joint design algorithms significantly outperform the linear and nonlinear separate design algorithms, respectively, due to the joint optimization techniques. Among the hybrid transceiver design algorithms, the best performance is achieved by the proposed nonlinear transceiver joint design algorithm followed by the nonlinear transceiver separate design algorithm, the linear transceiver joint design algorithm, the linear transceiver separate design algorithm and the ZF algorithm. The performance of the FD nonlinear transceiver design algorithm is provided as a reference. We can see that the proposed hybrid algorithm can approach the performance of the performance of the FD transceiver design algorithm. The corresponding MSE performance is shown in Fig. 7 which coincides with the results in Fig. 6.

Finally, we examine the SER performance or our proposed algorithm with the number of quantization bits of phase shifters. Fig. 8 shows the influence of number of quantization bits on our proposed algorithm. As expected, it is readily seen that the performance of our proposed algorithm improves with the number of quantization bits and when the number of quantization bits reaches 8, its performance can approach that with infinite resolution phase shifters.

B. Two-timescale joint design algorithm

We illustrate the performance of the proposed two-timescale joint design algorithm (**Algorithm 2**). We assume that the CSI delay of the single-timescale algorithm is proportional to the number of antennas while the effective CSI delay of the two-timescale algorithm is proportional to the number of RF chains.[9] Thus, we have $\tau/\tau_{TTS} = 4$, where τ is the full CSI delay of the single-timescale algorithm and τ_{TTS} is the effective

[9]Here we mainly consider the CSI delay caused by the pilots transmission. Typically the number of pilots is proportional to the number of transmit antennas (or RF chains if estimate the effective channel).

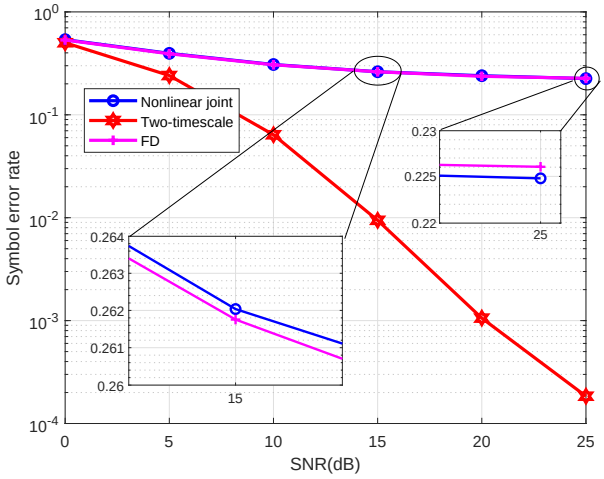

Fig. 10: SER performance comparison for different analyzed algorithms under the CSI delay $\tau = 5$ms.

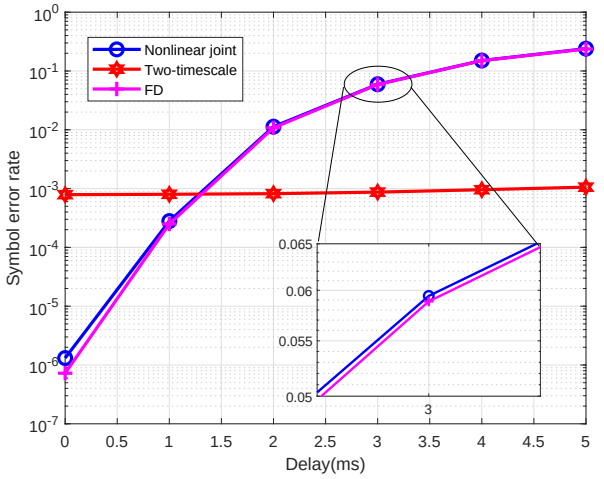

Fig. 11: SER performance of different analyzed algorithms versus τ (SNR$=$ 20dB).

CSI delay of the two-timescale algorithm. We first study the convergence performance of the proposed two-timescale joint design algorithm under the setting SNR$=$ 20dB, $\sigma_e^2 = 0.01$ and $\tau = 1$ms. τ_{TTS} can be calculated as $\tau_{TTS} = \tau/4 = 0.25$ms. For simplicity, we omit the computation of τ_{TTS} in the following experiments. Fig. 9 illustrates the average MSE versus the number of iterations. We can see that the proposed two-timescale algorithm converges within 300 iterations.

Fig. 10 shows the SER performance for different analyzed transceiver design algorithms under the CSI delay $\tau = 5$ms, including the proposed single-timescale joint design algorithm, the proposed two-timescale joint design algorithm and the FD algorithm. The two-timescale joint design algorithm provides the best SER performance while the other algorithms provide almost the same SER performance. With the increasing of SNR, the performance gap between the two-timescale algorithm and the single-timescale algorithm becomes larger.

In Fig. 11, we show the SER performance of the proposed single-timescale joint design algorithm, the proposed two-timescale joint design algorithm and the FD algorithm versus the CSI delay τ. It can be observed from Fig. 11 that with the increasing of τ, the performance of the single-timescale algorithms degrades dramatically while the performance of the two-timescale algorithm varies slightly. The two-timescale algorithm starts to outperform the single-timescale algorithms at the delay of 2ms. This is mainly because the two-timescale algorithm has much lower CSI estimation overhead and therefore it creates much smaller CSI delay. Moreover, the proposed single-timescale algorithm provides better performance compared to the proposed two-timescale algorithm in the presence of smaller delays. The results verify the robustness of our proposed TOSCA-based two-timescale joint design algorithm. This verifies the effectiveness of the proposed two-timescale algorithm against the CSI mismatch caused by the delays.

VII. Conclusion

In this work, we conceived the THP-based joint nonlinear hybrid A/D transceiver design algorithms for the downlink multiuser MIMO mmWave systems, where we considered the minimization of MSE subject to the transmit power constraint and the unit modulus constraint on each element of the RF analog BF matrices. Due to the highly coupled constraints, this optimization problem is hard to tackle. We first transformed it into a simpler form and then developed an innovative BCD-based algorithm to solve it. Besides, we proposed a novel TOSCA-based two-timescale joint design algorithm to further reduce the CSI signalling overhead and the effects of outdated CSI caused by the severe delays. These proposed algorithms can be guaranteed to obtain the KKT solution of the original problem. Moreover, with the aid of the lower bound of the MSE, we also determined the near-optimal cancellation order for the THP structure. Our simulation results demonstrated that the proposed THP-based hybrid transceiver design algorithm can significantly outperform the existing linear hybrid transceiver design algorithms and that the two-timescale joint design algorithm has stronger robustness against the CSI delay than the single-timescale algorithms. Hence, the proposed BCD-based joint design algorithm should be employed for the scenario of small CSI delays, while the extended two-timescale joint design algorithm should be applied for the case of severe CSI delays.

Appendix A
Proof of Theorem 1

It is readily seen that the KKT solution of problem (10) always makes the power constraint meet equality. By checking the first order conditions for problem (12) and problem (10), it is obvious that the scaled solution $\bar{\mathcal{S}}^\star$ is a KKT solution of problem (12) and we can obtain a series of equations (52) at the top of the next page via comparing the derivatives of MSE and MSE$_\sigma$ at the point $\bar{\mathcal{S}}^\star$.

The series of equations (52) indicates that $\mathcal{S}^\star$ is a KKT solution of problem (10) with the Lagrange multiplier attached to the power constraint being $\lambda_{P,T} = \sum_{m=1}^{M} \frac{1}{P_t} \|\bar{\mathbf{P}}_m^\star \mathbf{F}_m^\star\|^2$. This completes the proof.

$$\frac{1}{\sigma_m}\frac{\partial \text{MSE}}{\partial \mathbf{P}_m^*}|_{\bar{\mathcal{S}}^\star} = \frac{\partial \text{MSE}_\sigma}{\partial \mathbf{P}_m^*}|_{\bar{\mathcal{S}}^\star} = \mathbf{0}, \forall m, \tag{52a}$$

$$\frac{\partial \text{MSE}}{\partial \mathbf{F}_m^*}|_{\bar{\mathcal{S}}^\star} + \boldsymbol{\lambda}_{F,m} \circ \mathbf{F}_m^\star = \frac{\partial \text{MSE}_\sigma}{\partial \mathbf{F}_m^*}|_{\bar{\mathcal{S}}^\star} + \boldsymbol{\lambda}_{F,m} \circ \mathbf{F}_m^\star = \mathbf{0}, \forall m, \tag{52b}$$

$$\frac{\partial \text{MSE}}{\partial \mathbf{T}^*}|_{\bar{\mathcal{S}}^\star} + \boldsymbol{\lambda}_T \circ \mathbf{T}^\star + \sum_{m=1}^{M} \frac{1}{P_t} \|\bar{\mathbf{P}}_m^\star \mathbf{F}_m^\star\|^2 \mathbf{T}^\star \bar{\mathbf{W}}^\star (\bar{\mathbf{W}}^\star)^H = \frac{\partial \text{MSE}_\sigma}{\partial \mathbf{T}^*}|_{\bar{\mathcal{S}}^\star} + \boldsymbol{\lambda}_T \circ \mathbf{T}^\star = \mathbf{0}, \tag{52c}$$

$$\frac{\partial \text{MSE}}{\partial \mathbf{W}^*}|_{\bar{\mathcal{S}}^\star} + \sum_{m=1}^{M} \frac{1}{P_t} \|\bar{\mathbf{P}}_m^\star \mathbf{F}_m^\star\|^2 (\mathbf{T}^\star)^H \mathbf{T}^\star \bar{\mathbf{W}}^\star = \frac{\partial \text{MSE}_\sigma}{\partial \mathbf{W}^*}|_{\bar{\mathcal{S}}^\star} = \mathbf{0}, \tag{52d}$$

$$\frac{\partial \text{MSE}}{\partial \mathbf{U}^*}|_{\bar{\mathcal{S}}^\star} = \frac{\partial \text{MSE}_\sigma}{\partial \mathbf{U}^*}|_{\bar{\mathcal{S}}^\star} = \mathbf{0}, \tag{52e}$$

$$|[\mathbf{F}_m]_{i,j}^\star|^2 = 1 \quad \forall m, i, j, \tag{52f}$$

$$|[\mathbf{T}]_{i,j}^\star|^2 = 1 \quad \forall i, j, \tag{52g}$$

$$\|\bar{\mathbf{T}}\bar{\mathbf{W}}\|^2 = P_t. \tag{52h}$$

APPENDIX B
DERIVATION OF GRADIENTS

The partial derivatives with respect to the phase matrices can be associated with the partial derivatives with respect to the analog BF matrices by the following equations

$$\frac{\partial g}{\partial \boldsymbol{\theta}_T} = \frac{\partial g}{\partial \mathbf{T}} \circ 1j\mathbf{T} - \frac{\partial g}{\partial \mathbf{T}^*} \circ 1j\mathbf{T}^*, \tag{53}$$

and

$$\frac{\partial g}{\partial \boldsymbol{\theta}_{F_m}} = \frac{\partial g}{\partial \mathbf{F}_m} \circ 1j\mathbf{F}_m - \frac{\partial g}{\partial \mathbf{F}_m^*} \circ 1j\mathbf{F}_m^*. \tag{54}$$

Besides we have the expressions of the partial derivatives with respect to the analog BF matrices

$$\begin{aligned}\frac{\partial g}{\partial \mathbf{T}^*} =& \sum_{m=1}^{M} \bar{\mathbf{H}}_m^H \mathbf{F}_m^H \mathbf{P}_m^H \mathbf{P}_m \mathbf{F}_m \bar{\mathbf{H}}_m \mathbf{T}\mathbf{W}\mathbf{W}^H \\ &+ \sum_{m=1}^{M} \sigma_{e,m}^2 \text{tr}(\mathbf{P}_m \mathbf{F}_m \mathbf{F}_m^H \mathbf{P}_m^H) \mathbf{T}\mathbf{W}\mathbf{W}^H \\ &- \sum_{m=1}^{M} \bar{\mathbf{H}}_m^H \mathbf{F}_m^H \mathbf{P}_m^H \mathbf{A}_m \mathbf{L}^T \mathbf{U}\mathbf{W}^H,\end{aligned} \tag{55}$$

and

$$\begin{aligned}\frac{\partial g}{\partial \mathbf{F}_m^*} =& \mathbf{P}_m^H \mathbf{P}_m \mathbf{F}_m \bar{\mathbf{H}}_m \mathbf{T}\mathbf{W}\mathbf{W}^H \mathbf{T}^H \bar{\mathbf{H}}_m^H \\ &+ \sigma_{e,m}^2 \text{tr}(\mathbf{T}\mathbf{W}\mathbf{W}^H \mathbf{T}^H) \mathbf{P}_m^H \mathbf{P}_m \mathbf{F}_m \\ &+ \sigma_m^2 \mathbf{P}_m^H \mathbf{P}_m \mathbf{F}_m - \mathbf{P}_m^H \mathbf{A}_m \mathbf{L}^T \mathbf{U}\mathbf{W}^H \mathbf{T}^H \bar{\mathbf{H}}_m^H.\end{aligned} \tag{56}$$

Finally, we can obtain the partial derivatives $\frac{\partial g}{\partial \boldsymbol{\theta}_T}$ and $\frac{\partial g}{\partial \boldsymbol{\theta}_{F_m}}$ by substituting (55) and (56) into (53) and (54), respectively.

REFERENCES

[1] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[2] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[3] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[4] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surv. Tut.*, vol. 20, no. 2, pp. 870–913, second quarter 2018.

[5] R. He, B. Ai, G. Wang, Z. Zhong, C. Schneider, D. A. Dupleich, R. S. Thomae, M. Boban, J. Luo, and Y. Zhang, "Propagation channels of 5G millimeter-wave vehicle-to-vehicle communications: Recent advances and future challenges," *IEEE Veh. Technol. Mag.*, vol. 15, no. 1, pp. 16–26, 2020.

[6] R. He, B. Ai, G. L. Stüber, G. Wang, and Z. Zhong, "Geometrical-based modeling for millimeter-wave MIMO mobile-to-mobile channels," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 2848–2863, 2018.

[7] Y. Chen, D. Chen, T. Jiang, and L. Hanzo, "Millimeter-wave massive MIMO systems relying on generalized sub-array-connected hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8940-8950, Sept. 2019.

[8] K. Satyanarayana, M. El-Hajjar, P. Kuo, A. Mourad, and L. Hanzo, "Hybrid beamforming design for full-duplex millimeter wave communication," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1394-1404, Feb. 2019.

[9] K. Satyanarayana, M. El-Hajjar, A. A. M. Mourad, and L. Hanzo, "Deep learning aided fingerprint-based beam alignment for mmWave vehicular communication," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 10858-10871, Nov. 2019.

[10] C. Liu, M. Li, S. V. Hanly, I. B. Collings, and P. Whiting, "Millimeter wave beam alignment: Large deviations analysis and design insights," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1619–1631, 2017.

[11] M. Li, C. Liu, S. V. Hanly, I. B. Collings, and P. Whiting, "Explore and eliminate: Optimized two-stage search for millimeter-wave beam alignment," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4379–4393, 2019.

[12] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.

[13] X. Yu, J. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[14] W. Ni, X. Dong, and W. Lu, "Near-optimal hybrid processing for massive MIMO systems via matrix decomposition," *IEEE Trans. Signal Process*, vol. 65, no. 15, pp. 3922–3933, Aug. 2017.

[15] C. Chen, "An iterative hybrid transceiver design algorithm for millimeter wave MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 3, pp. 285–288, Jun. 2015.

[16] R. Méndez-Rial, C. Rusu, N. González-Prelcic, and R. W. Heath, "Dictionary-free hybrid precoders and combiners for mmWave MIMO systems," *Proc. 2015 IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun.*, pp. 151–155, Stockholm, Jun. 2015.

[17] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[18] X. Ge, Y. Sun, H. Gharavi, and J. Thompson, "Joint optimization of computation and communication power in multi-user massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4051-4063, Jun. 2018.

[19] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

[20] X. Zhai, Y. Cai, Q. Shi, M. Zhao, G. Y. Li, and B. Champagne, "Joint transceiver design with antenna selection for large-scale MU-MIMO mmWave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2085–2096, Sep. 2017.

[21] S. He, J. Wang, Y. Huang, B. Ottersten, and W. Hong, "Codebook-based hybrid precoding for millimeter wave multiuser systems," *IEEE Trans. Signal Process.*, vol. 65, no. 20, pp. 5289–5304, Oct. 2017.

[22] C. Lin, G. Y. Li, and L. Wang, "Subarray-based coordinated beamforming training for mmWave and sub-THz communications," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2115–2126, Sep. 2017.

[23] Q. Shi and M. Hong, "Spectral efficiency optimization for millimeter wave multiuser MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 455–468, Jun. 2018.

[24] Y. Cai, Y. Xu, Q. Shi, B. Champagne, and L. Hanzo, "Robust joint hybrid transceiver design for millimeter wave full-duplex MIMO relay systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1199–1215, Feb. 2019.

[25] A. Liu, V. K. N. Lau, and B. Kananian, "Stochastic successive convex approximation for non-convex constrained stochastic optimization," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4189–4203, 2019.

[26] X. Chen, A. Liu, Y. Cai, V. K. N. Lau, and M. Zhao, "Randomized two-timescale hybrid precoding for downlink multicell massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4152–4167, 2019.

[27] R. Mai, T. Le-Ngoc, and D. H. N. Nguyen, "Two-timescale hybrid RF-baseband precoding with MMSE-VP for multi-user massive MIMO broadcast channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4462–4476, 2018.

[28] Q. Shi and M. Hong, "Penalty dual decomposition method for non-smooth nonconvex optimization–part I: Algorithms and convergence analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 4108-4122, 2020.

[29] M. Tomlinson, "New automatic equaliser employing modulo arithmetic," *Electron. Lett.*, vol. 7, no. 5, pp. 138–139, Mar. 1971.

[30] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *IEEE Trans. Commun.*, vol. 20, no. 4, pp. 774–780, Aug. 1972.

[31] R. F. Fischer, C. Windpassinger, A. Lampe, and J. B. Huber, "Space-time transmission using Tomlinson-Harashima precoding," in *Proc. 2002 ITG Conf. on Source and Channel Coding*, pp. 139–148.

[32] X. Geng, B. An, F. Liu, and F. Cao, "Robust THP transceiver design for MIMO interference channel," *IEEE Commun. Lett.*, vol. 19, no. 9, pp. 1640–1643, Sep. 2015.

[33] S. Zarei, W. H. Gerstacker, R. Weigel, M. Vossiek, and R. Schober, "Robust MSE-balancing hierarchical linear/Tomlinson-Harashima Precoding for downlink massive MU-MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7309–7324, Nov. 2018.

[34] F. Tseng, M. Chang, and W. Wu, "Joint Tomlinson-Harashima source and linear relay precoder design in amplify-and-forward MIMO relay systems via MMSE criterion," *IEEE Trans. Veh. Technol.*, vol. 60, no. 4, pp. 1687–1698, May 2011.

[35] J. Liu and W. A. Kizymien, "Improved Tomlinson-Harashima precoding for the downlink of multi-user MIMO systems," *Can. J. Elect. Comput. Eng.*, vol. 32, no. 3, pp. 133–144, Summer 2007.

[36] K. Zu, R. C. de Lamare, and M. Haardt, "Multi-branch Tomlinson-Harashima precoding design for MU-MIMO systems: Theory and algorithms," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 939–951, Mar. 2014.

[37] L. Zhang, Y. Cai, R. C. de Lamare, and M. Zhao, "Robust multibranch Tomlinson-Harashima precoding design in amplify-and-forward MIMO relay systems," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3476–3490, Oct. 2014.

[38] L. Zhang, Y. Cai, M. Zhao, B. Champagne, and L. Hanzo, "Nonlinear MIMO transceivers improve wireless-powered and self-interference-aided relaying," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6953–6966, Oct. 2017.

[39] H. Gao, Y. Fu, C. Jiang, and Y. Zhao, "Optimization of Tomlinson-Harashima precoding for joint suppression of interference and power loss in MU-MIMO downlinks," in *2018 IEEE 3rd Int. Conf. on Communication and Information Systems (ICCIS)*, Singapore, Dec. 2018, pp. 137–140.

[40] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Interference optimization for transmit power reduction in Tomlinson-Harashima precoded MIMO downlinks," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2470–2481, May 2012.

[41] M. B. Shenouda and T. N. Davidson, "A framework for designing MIMO systems with decision feedback equalization or Tomlinson-Harashima precoding," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 401–411, Feb. 2008.

[42] C. Huang, L. Liu, C. Yuen and S. Sun, "Iterative channel estimation using LSE and sparse message passing for mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 67, no. 1, pp. 245-259, Jan. 2019.

[43] L. Chen, A. Liu, and X. Yuan, "Structured turbo compressed sensing for massive MIMO channel estimation using a Markov prior," *IEEE Trans. Vehi. Technol.*, vol. 67, no. 5, pp. 4635-4639, May 2018.

[44] N. Lee, O. Simeone, and J. Kang, "The effect of imperfect channel knowledge on a MIMO system with interference," *IEEE Trans. Commun.*, vol. 60, no. 8, pp. 2221–2229, 2012.

[45] Y. Cai, Q. Shi, B. Champagne, and Y. G. Li, "Joint transceiver design for secure downlink communications over an amplify-and-forward MIMO relay," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3691–3704, Sep. 2017.

[46] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., 1999.

[47] A. Liu, V. K. N. Lau, and M. Zhao, "Online successive convex approximation for two-stage stochastic nonconvex optimization," *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5941–5955, 2018.

[48] M. Joham, J. Brehmer, and W. Utschick, "MMSE approaches to multiuser spatio-temporal Tomlinson-Harashima precoding," *Proc. ITG SCC' 04,* Erlangen, Germany, Jan. 2004, pp. 387–394.