X. Zhang, L. Xiang, J. Wang, P. Zhu, D. W. K. Ng, and X. Gao, “Hybrid Precoding for mmWave Massive MIMO with Finite Blocklength,” in *IEEE Transactions on Wireless Communications*, vol. 73, no. 8, pp. 6379-6395, Aug. 2025, DoI: 10.1109/TCOMM.2025.3529244.

Hybrid Precoding for mmWave Massive MIMO with Finite Blocklength

Xuzhong Zhang, *Graduate Student Member, IEEE*, Lin Xiang, *Member, IEEE*, Jiaheng Wang, *Senior Member, IEEE*, Pengcheng Zhu, *Member, IEEE*, Derrick Wing Kwan Ng, *Fellow, IEEE*, and Xiqi Gao, *Fellow, IEEE*

***Abstract*—Hybrid digital-analog precoding is essential for balancing communication performance, energy efficiency, and hardware costs in millimeter wave (mmWave) massive multiple-input multiple-output (MIMO) systems. However, most existing designs rely on the Shannon capacity and assume infinite blocklengths, which are impractical for emerging applications, such as massive machine-type communications, operating with finite blocklength (FBL). To address this gap, this paper pioneers a novel hybrid precoding design for mmWave massive MIMO in the FBL regime. We meticulously optimize hybrid precoding based on both the weighted sum-rate (WSR) and the max-min fairness (MMF) criteria, while fulfilling the transmit power budget and users' minimum rate requirements. Both continuous and discrete phase shifters are considered for analog precoding. The formulated optimization problems are highly challenging to solve due to the nonconvex objective functions and nonconvex constraints. These challenges are further intensified by the nonconcave FBL rate function and the intricate coupling between analog and digital precoders. By proposing novel problem transformation and decomposition techniques, we reformulate the original complex problems into forms solvable with the penalty dual decomposition (PDD) method. We then develop two efficient iterative algorithms with parallel, and even closed-form variable updates, and guaranteed convergence to solve the WSR and MMF optimization problems, applicable to both continuous and discrete phase shifters. Simulation results show that our proposed hybrid precoding designs significantly outperform several baseline schemes, especially those adopting the Shannon capacity and infinite blocklength. Additionally, our proposed optimization algorithms enable hybrid precoding exploiting discrete phase shifters with limited quantization resolution (e.g., 3-bit) to closely match the performance of fully digital precoding in FBL scenarios.**

***Index Terms*—Finite blocklength transmission, massive multiple-input multiple-output (MIMO), hybrid precoding, max-min fairness, weighted sum rate.**

I. Introduction

Millimeter wave (mmWave) massive multiple-input multiple-output (MIMO) offers a promising solution to satisfy the escalating demand for ultra-wide frequency bandwidth in future communication systems [1]–[5]. By exploiting large antenna arrays with compact sizes at the base station (BS), mmWave massive MIMO can substantially mitigate multi-user interference and enhance both spectral and energy efficiency. However, the conventional fully digital precoding architecture, which involves connecting each antenna to a dedicated radio frequency (RF) chain, is rendered impractical for mmWave massive MIMO due to the associated prohibitive hardware costs and substantial RF power consumption. To strike an effective balance among system performance, hardware cost, and energy consumption, hybrid digital-analog precoding has emerged as a viable solution for realizing mmWave massive MIMO systems [1], [2], [5]–[19]. In this innovative architecture, digitally precoded baseband signals are up-converted to the carrier frequency through a limited number of RF chains and subsequently processed through a linear network of adjustable phase shifters, constituting the analog precoder.

To enhance the performance of mmWave massive MIMO systems, hybrid precoding optimization is of paramount importance and has been extensively explored in the literature [6]–[19]. For instance, the authors in [6]–[13] proposed several hybrid precoding designs to maximize spectral efficiency for different practical scenarios. Also, in [14], a hybrid precoding scheme was tailored to maximize the minimum user rate in a cache-enabled mmWave system. The energy efficiency of hybrid precoding was evaluated in [15], [16]. Additionally, hybrid precoding was optimized for transmit power minimization while ensuring users' signal-to-interference-plus-noise ratio (SINR) requirements in [17]–[19].

Despite the extensive efforts devoted, most of the aforementioned works [7], [8], [10], [14], [16], [18], [19] assume using continuous phase shifters with infinite resolution in the analog precoder for tractability. However, such phase shifters are prohibitively expensive. In contrast, practical and more affordable phase shifters typically feature limited resolution, specified by a discrete set, which significantly complicates the hybrid precoding optimizations. Specifically, to handle discrete phase shifters, a commonly adopted method involves first configuring the hybrid precoder for infinite resolution, followed by quantizing the analog precoder into discrete values. Unfortu-

This work was supported in part by the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu under Grants BK20212001 and BK20222001, the Key Technologies R&D Program of Jiangsu (Prospective and Key Technologies for Industry) under Grants BE2022068-3 and BE2022067-5, the National Natural Science Foundation of China under Grants U22B2006 and 62331024, the Jiangsu Province Major Science and Technology Project under Grant SBG2024000080, the Science and Technology Major Project of Nanjing under Grant 202405020, the Jiangsu Provincial Scientific Research Center of Applied Mathematics under Grant BK20233002, the Fundamental Research Funds for the Central Universities under Grants 2242023K5003 and 2242022K60002. The work of L. Xiang has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center under grant LOEWE/1/12/519/03/ 05.001(0016)/72 and the Federal Ministry of Education and Research (BMBF) project Open6GHub under grant 16KISKO14. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gui Zhou. *(Corresponding authors: Jiaheng Wang; Xiqi Gao.)*

Xuzhong Zhang, Jiaheng Wang, Pengcheng Zhu, and Xiqi Gao are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China, and also with the Purple Mountain Laboratories, Nanjing 210023, China (e-mail: xzzhang@seu.edu.cn, jhwang@seu.edu.cn, p.zhu@seu.edu.cn, and xqgao@seu.edu.cn).

Lin Xiang is with the Communications Engineering Lab, Technische Universität Darmstadt, 64283 Darmstadt, Germany (e-mail: l.xiang@nt.tu-darmstadt.de).

Derrick Wing Kwan Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: w.k.ng@unsw.edu.au).

nately, this heuristic approach often leads to poor performance or infeasible solutions to the design optimization problem, especially when the phase shifters are of low resolution [9]. To overcome this limitation, hybrid precoding designs tailored to discrete phase shifters are crucial and have been investigated in [6], [9], [15], [17].

Meanwhile, most existing hybrid precoding designs, such as those in [6]–[16], assume an infinite blocklength (IBL) and evaluate performance based on the Shannon capacity. However, practical systems often operate under finite or even limited blocklengths, where the transmissions are generally not error-free and cannot achieve the Shannon capacity [20]. For example, mmWave massive MIMO has recently been increasingly adopted in latency-sensitive communications to enable timely status updates and transmission of control commands in Internet-of-Thing (IoT) networks, massive machine-type communications (mMTC), and tactile internet [21]–[23]. These applications typically rely on short packets to reduce transmission latency, significantly deviating from IBL assumptions. Consequently, hybrid precoding designs based on IBL would lead to poor performance in these practical scenarios. In addition, most FBL services impose stringent minimum rate requirements [24], [25], which are often neglected in current hybrid precoding research. Therefore, it is imperative to rethink the optimal hybrid precoding design for mmWave massive MIMO that accommodates users' quality-of-service (QoS) requirements and supports phase shifters of varying resolutions in the FBL regime.

Recently, the finite blocklength (FBL) regime has emerged as a critical area for precoding design and optimization [26]–[36]. For instance, path-following algorithms were employed in [26], [27] to optimize precoders for maximizing the users' minimum FBL rate. Also, resource allocation for orthogonal frequency division access (OFDMA) downlink multiple-input single-output (MISO) with FBL was studied in [28]. Furthermore, addressing users' QoS requirements, the authors in [29] explored several FBL precoding optimizations to maximize the weighted sum rate (WSR), minimum user rate, and energy efficiency, which were solved utilizing the uplink-downlink duality. Moreover, low-complexity precoding design for massive MIMO multi-group multicasting in the FBL regime was examined in [30]. Additionally, FBL precoding designs for rate-splitting multiple access (RSMA) were investigated in [31]–[33], and RIS-aided FBL systems were further explored in [33]–[35]. However, most existing FBL precoding designs, such as [26]–[35], focused exclusively on the fully-digital architecture and their results cannot be directly applied to hybrid precoding for mmWave massive MIMO due to the complex coupling of digital and analog precoders. While recent research [36] examined hybrid precoding in the FBL regime, its approach is heuristic and does not address users' QoS requirements. To the best of our knowledge, no prior work has yet explored hybrid precoding optimizations with users' QoS constraints in the FBL regime, for either continuous or discrete phase shifters.

To fill in this gap, this paper aims to introduce a novel hybrid precoding design for mmWave massive MIMO in the FBL regime. Given that WSR and max-min fairness (MMF) have been widely considered in hybrid precoding designs in the IBL regime [6], [10]–[14], we also adopt them as our optimization objectives in the FBL regime. Specifically, WSR-based designs aim to maximize the (weighted) sum-rate by prioritizing users with better channel conditions (after applying appropriate weights), making them ideal for applications emphasizing system throughput, such as video streaming and augmented reality [24], [37]. In contrast, MMF-based designs maximize the minimum data rate among users, thus prioritizing users with poorer channel conditions. Although this approach may reduce overall system throughput, it is crucial for fairness-oriented applications such as autonomous vehicle networks and industrial IoT applications [23]. Therefore, it is essential to consider both metrics for balancing system throughput and user fairness, as well as addressing the diverse requirements of next-generation wireless applications.

However, due to the nonconvex nature of the FBL rate function, hybrid precoding optimization problems under the FBL WSR and MMF criteria are significantly more challenging to solve than their IBL counterparts. Existing works [6]–[8], [15] suggested a heuristic hybrid precoding design using the commonly adopted matrix approximation (MAP) method [6], [7], which avoids directly addressing the complex hybrid precoding optimization problem by instead minimizing the Euclidean distance between the hybrid precoder and the fully digital precoder obtained with the solutions in [29], [30]. However, this approach often fails to satisfy the users' QoS requirements and exhibits poor performance, especially with low-resolution phase shifters, as will be shown in Section V. In contrast, this paper presents the first algorithms that directly solve the complex hybrid precoding optimization problems in the FBL regime with guaranteed convergence. Additionally, our hybrid precoder designs support both continuous and discrete phase shifters. We are interested in answering the following fundamental question: *how many bits of quantization are required by hybrid precoding with discrete phase shifters, to achieve a performance comparable to that of fully-digital precoding in FBL systems?* To address these research goals, we have made the following technical contributions in this paper:

- We formulate hybrid precoding optimization problems for mmWave massive MIMO in the FBL regime under the WSR and MMF criteria, while considering the BS's power budget, users' QoS constraints, and both continous and discrete phase shifters. The formulated problems are nonconvex and rather challenging to solve. These complexities are exacerbated by the complex FBL rate function and the intricate coupling between digital and analog precoders.
- Through innovative problem transformation and decomposition techniques, we reformulate the original complex WSR problem into a more tractable form that facilitates the implementation of the penalty dual decomposition (PDD) method [38]. Building upon this foundation, we propose a computationally-efficient algorithm with guaranteed convergence to address the nonconvex WSR problem tailored for mmWave massive MIMO, accommodat-

ing both continuous and discrete phase shifters.

- The WSR solution cannot be directly applied to the nonconex, nonsmooth MMF problem due to its more intricate structure. As a remedy, we further extend the problem transformation and decomposition techniques to effectively reformulate the MMF problem and propose another efficient algorithm employing the PDD method [38], which is specially designed for mmWave massive MIMO and compatible with both continuous and discrete phase shifters.
- Simulation results demonstrate that our proposed hybrid precoding designs achieve the best performance among several considered benchmarks, especially those employing Shannon capacity as the performance metric. Furthermore, even with the minimal number of RF chains needed to support multi-user communication, our optimization algorithms enable hybrid precoding with discrete phase shifters (e.g., 3-bit resolution) to achieve FBL performance closely approaching that of fully digital precoding.

In the remainder of this paper, we introduce in Section II the adopted system model and formulate the WSR and MMF hybrid precoding design optimization problems in the FBL regime. In Sections III and IV, we propose low-complexity hybrid precoding designs for the WSR and MMF problems, respectively. Simulation results are presented in Section V, and finally, conclusions are drawn in Section VI.

Notations: Throughout this paper, vectors and matrices are denoted in bold lower-case and capital letters, respectively. $\mathbb{C}^{N\times 1}$, and $\mathbb{C}^{N\times M}$ denote the sets of complex vectors of length N, and complex matrices of size $N\times M$, respectively. $\Re\{x\}$ denotes the real part of complex number x. $\jmath=\sqrt{-1}$ denotes the imaginary unit. $[\mathbf{A}]_{i,j}$ denotes the (i,j)-th entry of matrix $\mathbf{A}$. $\mathbf{I}_N$ and $\mathbf{1}_N$ denote an $N\times N$ identity matrix and the all-one column vector of length N, respectively. $(\cdot)^T$, $(\cdot)^\star$, $(\cdot)^H$, and $(\cdot)^\dagger$ denote transpose, complex conjugate, Hermitian transpose, and pseudo-inverse of a matrix, respectively. $|\cdot|$, $\|\cdot\|_2$, $\|\cdot\|_\infty$, and $\|\cdot\|_{\rm F}$ denote the absolute value of a complex scalar, Euclidean norm of a vector, infinity norm of a vector, and Frobenius norm of a matrix, respectively. vec$(\cdot)$ denotes the vectorization function. $Q(x)$ is the Q-function defined as $Q(x)=\frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-t^2/2)dt$. $Q^{-1}(\cdot)$ is the inverse Q-function, i.e., $Q(Q^{-1}(x))=x$. $\mathbf{x}\sim\mathcal{CN}(\mathbf{a},\mathbf{R})$ means that $\mathbf{x}$ is a circular symmetric complex Gaussian random vector with mean $\mathbf{a}$ and covariance matrix $\mathbf{R}\succeq\mathbf{0}$.

II. System Model And Problem Formulation

We consider the downlink (DL) of a single-cell mmWave multi-user massive MIMO system, where a BS equipped with N_t transmit antennas and $N_r\le N_t$ RF chains serves K single-antenna users ($K\le N_r$).[1] The users are indexed by set $\mathcal{K}\triangleq\{1,\ldots,K\}$. The BS adopts a fully connected hybrid analog and digital precoding architecture [6], [15] to facilitate efficient FBL transmission and reduce the hardware complexity and energy consumption. Thereby, the data streams intended for the users are precoded at the baseband exploiting digital precoders $\mathbf{w}_k\in\mathbb{C}^{N_r\times 1}$, $k\in\mathcal{K}$, before being up-converted to the carrier frequency in the RF chains and further processed by an analog precoder $\mathbf{F}\in\mathbb{C}^{N_t\times N_r}$. In this paper, we consider that the analog precoder is implemented exploiting one of the following types of phase shifters:

- Continuous phase shifters, which have an arbitrary bit resolution such that
$$[\mathbf{F}]_{i,j}\in\mathcal{F}_{\rm C}\triangleq\left\{\exp\{\jmath\theta\}\,|\,\theta\in[0,2\pi]\right\},\forall i,j; \quad (1)$$
- Discrete phase shifters, which are uniformly quantized with κ-bit resolution and thus
$$[\mathbf{F}]_{i,j}\in\mathcal{F}_{\rm D}\triangleq\left\{\exp\left(\jmath 2^{1-\kappa}m\pi\right)\big|m=0,\ldots,2^\kappa-1\right\}. \quad (2)$$

Note that $\mathcal{F}_{\rm C}$ and $\mathcal{F}_{\rm D}$ define a unit-circle manifold and its discretized set, respectively, both belonging to nonconvex sets.

Let $s_k\sim\mathcal{CN}(0,1)$ be the data symbols intended for user k and $\mathbb{E}\{s_k^\star s_j\}=0,\forall j\neq k$. Moreover, let $\mathbf{h}_k\in\mathbb{C}^{N_t\times 1}$ be the channel vector between the BS and user k. By employing hybrid precoding at the BS, the received signal of user k is given by

$$y_k=\mathbf{h}_k^H\mathbf{F}\mathbf{w}_k s_k+\sum\nolimits_{i\neq k}\mathbf{h}_k^H\mathbf{F}\mathbf{w}_i s_i+n_k, \quad (3)$$

where $n_k\sim\mathcal{CN}(0,\sigma_k^2)$ is the additive white Gaussian noise (AWGN) at user k with zero mean and variance σ_k^2. We assume that the BS possesses full and perfect knowledge of the channel state information (CSI) [9], [10], [14], [16]–[19], [26], [27], [29], [31], [32].[2] Then, the received SINR of user k is

$$\gamma_k=\frac{|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_k|^2}{\sum_{i\neq k}|\mathbf{h}_k^H\mathbf{F}\mathbf{w}_i|^2+\sigma_k^2}. \quad (4)$$

To lower the communication latency, the BS adopts an FBL N for signal transmission, as in typical mMTC and IoT applications [21]–[23]. Consequently, the users cannot decode the messages in an error-free manner [20]. Assume that user k requires a specific block error rate (BLER) $\epsilon_k>0$ for reliable communication. Under this requirement, the achievable data rate of user k in nats/s/Hz is approximately given by [20]

$$R(\gamma_k,\vartheta_k)=\ln(1+\gamma_k)-\vartheta_k\sqrt{V(\gamma_k)}, \quad (5)$$

where $\vartheta_k=Q^{-1}(\epsilon_k)/\sqrt{N}>0$, $V(\gamma_k)=1-(1+\gamma_k)^{-2}$ is the channel dispersion. The second term in (5) denotes a reduction in the achievable data rate to ensure the required BLER ϵ_k, whose value decreases with N. As $N\to\infty$, the second term vanishes such that the FBL rate (5) approaches the Shannon capacity $\ln(1+\gamma_k)$.

In wireless networks, the WSR and the MMF are two key performance metrics having been widely employed in the literature for hybrid precoding design with IBL [6], [10]–[14].

[1] A recent study [34] investigated joint transceiver and RIS design in a multi-cell MIMO system, where each multi-antenna BS transmits multiple streams to each multi-antenna user. Exploring the joint design of hybrid transmit precoding and receiver combining in mmWave massive MIMO systems in the FBL regime is an intriguing direction for future research.

[2] By assuming perfect CSI, we aim to explore the theoretical performance limits of mmWave massive MIMO hybrid precoding in the FBL regime. In time-division duplex (TDD) systems, the BS can acquire DL CSI through uplink (UL) training by leveraging the reciprocity between UL and DL channels. However, with the hybrid precoding architecture, the BS cannot directly access the outputs from individual antenna elements; instead, it can only access the channels through a linear network of phase shifters, which effectively compresses the received signal. Therefore, to estimate UL CSI, the BS can employ compressed-sensing (CS) techniques, such as orthogonal matching pursuit (OMP) [39], [40], approximate message passing (AMP) algorithm [41], and sparse Bayesian leaning (SBL) [42].

To facilitate hybrid precoding design in the FBL regime, in this paper, we adopt both metrics but modify them leveraging the FBL rate (5). For given BLER $\{\epsilon_k\}_{k=1}^K$ and blocklength N, the resulting WSR and MMF hybrid precoding optimization problems are formulated as:

$$\mathbb{P}_1: \max_{\mathbf{W},\mathbf{F}} \ \sum\nolimits_{k=1}^{K} \omega_k R(\gamma_k, \vartheta_k) \tag{6a}$$

$$\text{s.t.} \ \sum\nolimits_{k=1}^{K} \|\mathbf{F}\mathbf{w}_k\|_2^2 \le P, \tag{6b}$$

$$[\mathbf{F}]_{i,j} \in \mathcal{F}, \forall i,j, \tag{6c}$$

$$R(\gamma_k, \vartheta_k) \ge \bar{R}_k, \ \forall k \in \mathcal{K}, \tag{6d}$$

and

$$\mathbb{P}_2: \max_{\mathbf{W},\mathbf{F}} \ \min_{k\in\mathcal{K}} R(\gamma_k, \vartheta_k) \tag{7}$$

$$\text{s.t. (6b), (6c), (6d)},$$

respectively, where $\mathbf{W} \triangleq [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{C}^{N_r \times K}$ is the digital precoder matrix and $\omega_k > 0$ is the weight assigned to user k. In problems $\mathbb{P}_1$ and $\mathbb{P}_2$, we jointly optimize both digital and analog precoders to maximize the WSR and the minimum rate achievable among the users under the given BLER $\{\epsilon_k\}_{k=1}^K$ and blocklength N, respectively, while satisfying the same set of constraints. In particular, constraint (6b) limits the transmit power of the BS by a maximum budget P. Also, (6c) defines the set of allowed phase shifts for continuous and discrete phase shifters by $\mathcal{F} \in \{\mathcal{F}_{\mathrm{C}}, \mathcal{F}_{\mathrm{D}}\}$, respectively.[3] Finally, (6d) requires a minimum FBL rate $\bar{R}_k > 0$ for each user k.[4]

Note that the objective functions of problems $\mathbb{P}_1$ and $\mathbb{P}_2$ are nonconcave, and even the complex FBL rate (5) by itself is nonconcave with respect to (w.r.t.) the SINR γ_k and the precoders $\{\mathbf{W}, \mathbf{F}\}$ [29]. Besides, the objective function of $\mathbb{P}_2$ is nonsmooth. Meanwhile, constraints (6b)-(6d) are nonconvex sets. Furthermore, the analog precoder $\mathbf{F}$ is coupled with the digital precoder $\mathbf{W}$ in the objective functions and constraints (6b) and (6d). As such, it is extremely challenging to optimally solve $\mathbb{P}_1$ and $\mathbb{P}_2$. To the best of our knowledge, the complexities in handling the FBL rate have prevented $\mathbb{P}_1$ and $\mathbb{P}_2$ from being investigated in existing literature.

In fact, the penalty dual decomposition (PDD) method [38] provides an appealing approach to address nonconvex nonsmooth problems with coupling constraints. Based on the PDD method, we develop two computationally efficient algorithms to solve $\mathbb{P}_1$ and $\mathbb{P}_2$, which are applicable for handling both continuous and discrete phase shifters. This enables to assess the performance gaps between them, cf. Section V, and characterize the quantization granularity needed for practical implementation of analog precoder in the FBL regime, to balance performance and implementation costs. Moreover, the design variables can be optimized in a BCD manner with efficient and even closed-form solutions to lower the computational complexity while providing theoretical guarantees for achieving the KKT solutions of $\mathbb{P}_1$ and $\mathbb{P}_2$ under mild conditions.

In the following, we start with tackling the nonconvex smooth problem $\mathbb{P}_1$ in Section III, exploiting the proposed algorithm based on PDD. We then extend the PDD method to address the nonconvex nonsmooth problem $\mathbb{P}_2$ in Section IV.

III. WSR Hybrid Precoding Design

In this section, we address the hybrid precoding optimization problem $\mathbb{P}_1$ for maximizing the WSR. To achieve this, we first reformulate $\mathbb{P}_1$ into a particular form where the optimization variables are coupled solely in the equality constraints. We then formulate its AL problem and handle $\mathbb{P}_1$ with the PDD approach.

A. Transformation of Problem $\mathbb{P}_1$

We first tackle the nonconvexity in constraint (6d). Note that for given $\vartheta_k > 0$, the FBL rate $R(\gamma_k, \vartheta_k)$ is a monotonically increasing function of the received SINR γ_k in the effective SINR regime where $R(\gamma_k, \vartheta_k) > 0$ [29], [30]. Therefore, there exists a unique solution $\bar{\gamma}_k > 0$ satisfying $R(\bar{\gamma}_k, \vartheta_k) = \bar{R}_k$, which can be obtained adopting a typical bisection search. Then constraint (6d) can be equivalently rewritten as

$$\gamma_k \ge \bar{\gamma}_k, \forall k \in \mathcal{K}. \tag{8}$$

Although (8) is still nonconvex, as will be shown below, it is much more convenient to address than (6d) due to the elimination of the complex FBL rate function from (6d).

Next, we reformulate the coupling inequality constraints (6b) and (8). Note that these coupling constraints prevent solving problem $\mathbb{P}_1$ directly with BCD-type algorithms, such as the block successive upper-bound minimization (BSUM) method [43]. Such algorithms are prone to becoming trapped in inefficient solution points, deteriorating the system performance. To resolve the coupling between analog and digital precoders in constraints (6b) and (8), we introduce the following optimization auxiliary variables

$$q_{i,j} = \mathbf{h}_i^H \mathbf{F} \mathbf{w}_j, \forall i,j \in \mathcal{K}, \tag{9}$$

$$\mathbf{d}_k = \mathbf{F}\mathbf{w}_k, \forall k \in \mathcal{K}. \tag{10}$$

Then constraints (6b) and (8) can be equivalently rewritten as

$$\|\mathbf{D}\|_{\mathrm{F}}^2 \le P, \tag{11}$$

$$\Omega_k(\mathbf{q}_k) \triangleq \bar{\gamma}_k\big(\sum\nolimits_{i\neq k} \big|q_{k,i}\big|^2 + \sigma_k^2\big) - \big|q_{k,k}\big|^2 \le 0, \forall k \in \mathcal{K}, \tag{12}$$

respectively, where $\mathbf{D} \triangleq [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{C}^{N_t \times K}$, $\mathbf{q}_k \triangleq (q_{k,1}, \ldots, q_{k,K})^T \in \mathbb{C}^{K\times 1}$, and $\mathbf{Q} \triangleq [\mathbf{q}_1, \ldots, \mathbf{q}_K]^T \in \mathbb{C}^{K\times K}$.

Capitalizing on the above transformation techniques, problem $\mathbb{P}_1$ can be equivalently reformulated as

$$\mathbb{P}_3: \max_{\mathbf{W},\mathbf{F},\mathbf{Q},\mathbf{D}} \ \sum\nolimits_{k=1}^{K} \omega_k R\big(\hat{\gamma}_k(\mathbf{q}_k), \vartheta_k\big) \tag{13}$$

$$\text{s.t.} \ \ (6c), (9), (10), (11), (12),$$

where $\hat{\gamma}_k(\mathbf{q}_k) \triangleq \big|q_{k,k}\big|^2 / \alpha_k(\mathbf{q}_k)$ and $\alpha_k(\mathbf{q}_k) \triangleq \sum_{i\neq k} \big|q_{k,i}\big|^2 + \sigma_k^2$. Note that $\mathbb{P}_3$ is a nonconvex optimization problem with coupled optimization variables only in the equality constraints (9) and (10). In the remainder of this section, we apply the PDD method [38] to address problem $\mathbb{P}_3$.

[3] In this work, we focus primarily on hybrid precoding design in the FBL regime with continuous and discrete phase shifters. Exploring additional features of mmWave communications, such as low-resolution analog-to-digital (ADC) and digital-to-analog converter (DAC), non-ideal power amplifiers (PA), and wideband systems [4], [5], in hybrid precoding optimization with FBL is a promising direction for future research.

[4] In problems $\mathbb{P}_1$ and $\mathbb{P}_2$, the minimum rate $\bar{R}_k$ may vary among users, as they may have heterogeneous service demands in practice. In such scenarios, eliminating the QoS constraint (6d) from problem $\mathbb{P}_2$ would generally lead to a different optimal solution than the original problem. However, if all users require an identical minimum rate constraint, (6d) can be omitted in problem $\mathbb{P}_2$ without altering the optimal solution, provided that $\mathbb{P}_2$ is feasible.

To this end, let $z_{i,j}$ and $\mathbf{u}_k$ be the Lagrangian multipliers associated with constraints (9) and (10), respectively. The AL problem of $\mathbb{P}_3$ for a penalty parameter $\rho > 0$ is formulated as

$$\mathbb{P}_4: \max_{\mathbf{W},\mathbf{F},\mathbf{Q},\mathbf{D}} \ \mathcal{L}_{\mathrm{W}}(\mathbf{W},\mathbf{F},\mathbf{Q},\mathbf{D};\boldsymbol{Z},\boldsymbol{U}) \quad \text{s.t.} \quad (6c),(11),(12), \tag{14}$$

with

$$\mathcal{L}_{\mathrm{W}}(\mathbf{W},\mathbf{F},\mathbf{Q},\mathbf{D};\boldsymbol{Z},\boldsymbol{U}) \triangleq \sum\nolimits_{k=1}^{K} \omega_k R\big(\hat{\gamma}_k\big(\mathbf{q}_k\big),\vartheta_k\big) - \frac{1}{2\rho}\left\|\mathbf{Q}-\mathbf{H}^H\mathbf{F}\mathbf{W}+\rho\boldsymbol{Z}\right\|_{\mathrm{F}}^2 - \frac{1}{2\rho}\left\|\mathbf{D}-\mathbf{F}\mathbf{W}+\rho\boldsymbol{U}\right\|_{\mathrm{F}}^2, \tag{15}$$

where $\mathbf{H} \triangleq [\mathbf{h}_1,\ldots,\mathbf{h}_K]$ is the channel matrix. Moreover, $\{\mathbf{W},\mathbf{F},\mathbf{Q},\mathbf{D}\}$ and $\{\boldsymbol{Z},\boldsymbol{U}\}$ define the sets of primal and dual optimization variables, respectively, with $[\boldsymbol{Z}]_{i,j} \triangleq z_{i,j}$ and $\boldsymbol{U} \triangleq [\mathbf{u}_1,\ldots,\mathbf{u}_K]$. Here, the dual variables are distinguished from the primal variables using italicized fonts.

The PDD method [38] tackles problem $\mathbb{P}_3$ via a nested loop structure. Thereby, the inner loop focuses on optimizing the primal variables by solving the AL problem $\mathbb{P}_4$, while the outer loop updates the dual variables and the penalty factor. Notably, the primal variables in $\mathbb{P}_4$ can be further divided into three blocks, i.e., the auxiliary variables $\{\mathbf{Q},\mathbf{D}\}$, the digital precoding matrix $\mathbf{W}$, and the analog precoder $\mathbf{F}$, where different blocks are not coupled in the constraints. However, due to the presence of nonconvex constraints (6c) and (12), the subproblems involving $\mathbf{Q}$ and $\mathbf{F}$ are still nonconvex. This prevents solving $\mathbb{P}_4$ adopting the classical BCD method.

To overcome this challenge, we solve $\mathbb{P}_4$ using the BSUM approach [43] below, which iteratively updates a block of variables using majorization-minimization (MM) or successive convex approximation (SCA) [44] while keeping the other variables fixed at each iteration. By leveraging the decomposable structure of $\mathbb{P}_4$, we derive a low-complexity solution that can tackle the nonconvex constraints. Details on solving $\mathbb{P}_4$ are provided in Section III-B, while the comprehensive solution for $\mathbb{P}_3$ or $\mathbb{P}_1$ is postponed to Section III-D. For clarity, we adopt t and n to indicate the indices for the outer and inner loop iterations, respectively. Additionally, $\rho^{(t)}$, $\boldsymbol{U}^{(t)}$, and $\boldsymbol{Z}^{(t)}$ denote the values of ρ, $\boldsymbol{U}$, and $\boldsymbol{Z}$ in the t-th outer iteration, and $\mathbf{W}^{(t,n)},\mathbf{F}^{(t,n)},\mathbf{Q}^{(t,n)}$, and $\mathbf{D}^{(t,n)}$ denote the values of $\mathbf{W},\mathbf{F},\mathbf{Q}$, and $\mathbf{D}$ in the n-th inner iteration within the t-th outer iteration.

B. Solving the AL Problem $\mathbb{P}_4$

In this subsection, we derive a BSUM algorithm for solving the AL problem $\mathbb{P}_4$ in the n-th inner iteration within the t-th outer iteration. By capitalizing on the decomposable structure inherent in problem $\mathbb{P}_4$, we derive efficient, and in several instances, closed-form solutions to optimize the variables, facilitating scalable hybrid precoding for mmWave massive MIMO systems.

1) Optimization of Digital Precoder $\mathbf{W}$

For given $\mathbf{F}^{(t,n)}$, $\mathbf{Q}^{(t,n)}$, and $\mathbf{D}^{(t,n)}$, optimizing $\mathbf{W}$ at the inner iteration n reduces to solving the following unconstrained convex optimization problem

$$\min_{\mathbf{W}} \ \|\mathbf{D}^{(t,n)} - \mathbf{F}^{(t,n)}\mathbf{W} + \rho^{(t)}\boldsymbol{U}^{(t)}\|_{\mathrm{F}}^2 + \|\mathbf{Q}^{(t,n)} - \mathbf{H}^H\mathbf{F}^{(t,n)}\mathbf{W} + \rho^{(t)}\boldsymbol{Z}^{(t)}\|_{\mathrm{F}}^2. \tag{16}$$

By setting the first-order derivative of the objective function to zero, the optimal solution of (16) is given by

$$\mathbf{W}^* = \big(\mathbf{F}^{(t,n)}\big)^H\big(\mathbf{I}_{N_t} + \mathbf{H}\mathbf{H}^H\big)\mathbf{F}^{(t,n)}\big)^{\dagger}\big(\mathbf{F}^{(t,n)}\big)^H \cdot \Big\{\mathbf{D}^{(t,n)} + \rho^{(t)}\boldsymbol{U}^{(t)} + \mathbf{H}(\mathbf{Q}^{(t,n)} + \rho^{(t)}\boldsymbol{Z}^{(t)})\Big\}. \tag{17}$$

2) Optimization of Analog Precoder $\mathbf{F}$

The optimization of $\mathbf{F}$ at the inner iteration n, given $\mathbf{W}^{(t,n+1)}$, $\mathbf{Q}^{(t,n)}$, and $\mathbf{D}^{(t,n)}$, can be expressed as

$$\min_{\mathbf{F}} \ \|\mathbf{D}^{(t,n)} - \mathbf{F}\mathbf{W}^{(t.n+1)} + \rho^{(t)}\boldsymbol{U}^{(t)}\|_{\mathrm{F}}^2 + \|\mathbf{Q}^{(t,n)} - \mathbf{H}^H\mathbf{F}\mathbf{W}^{(t,n+1)} + \rho^{(t)}\boldsymbol{Z}^{(t)}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad (6c). \tag{18}$$

Notably, constraint (6c) is nonconvex for both continuous and discrete phase shifters, rendering the solution of problem (18) challenging. To address this obstacle, we first reformulate problem (18) as

$$\min_{\mathbf{F}} \quad F_{\mathrm{W},\mathbf{F}}(\mathbf{F}) \triangleq \mathrm{Tr}\left\{\mathbf{F}^H\mathbf{A}\mathbf{F}\mathbf{C}\right\} - 2\Re\left\{\mathrm{Tr}\big(\mathbf{F}^H\mathbf{B}\big)\right\} \quad \text{s.t.} \quad (6c), \tag{19}$$

where $\mathbf{A} = \mathbf{I}_{N_t} + \mathbf{H}\mathbf{H}^H$,

$$\mathbf{B} = \big(\mathbf{D}^{(t,n)} + \rho^{(t)}\boldsymbol{U}^{(t)} + \mathbf{H}\big(\mathbf{Q}^{(t,n)} + \rho^{(t)}\boldsymbol{Z}^{(t)}\big)\big)\big(\mathbf{W}^{(t,n+1)}\big)^H, \tag{20}$$

and $\mathbf{C} = \mathbf{W}^{(t,n+1)}\big(\mathbf{W}^{(t,n+1)}\big)^H$. The matrices $\mathbf{A}$ and $\mathbf{C}$ are all positive semidefinite. Then, problem (19) can be solved using a BCD algorithm as in [6]. Particularly, let $\tilde{\mathbf{F}}$ be the initial value of the analog precoder $\mathbf{F}$. We then update $[\mathbf{F}]_{i,j}$ based on $\tilde{\mathbf{F}}$ via the following quadratic programming

$$\min_{[\mathbf{F}]_{i,j}} \ \tilde{a}\big|[\mathbf{F}]_{i,j}\big|^2 - 2\Re\big\{\tilde{b}^{\star}\,[\mathbf{F}]_{i,j}\big\} \quad \text{s.t.} \quad [\mathbf{F}]_{i,j} \in \mathcal{F}. \tag{21}$$

In (21), the first term in the objective function can be ignored, because $\tilde{a}$ is a constant whose value depends on $\tilde{\mathbf{F}}$ and $\big|[\mathbf{F}]_{i,j}\big|^2 = 1$ for both continuous and discrete phase shifters, cf. (1) and (2). Besides,

$$\tilde{b} = [\mathbf{A}]_{i,j}\big[\tilde{\mathbf{F}}\big]_{i,j}[\mathbf{C}]_{i,j} - \big[\mathbf{A}\tilde{\mathbf{F}}\mathbf{C}\big]_{i,j} + [\mathbf{B}]_{i,j}. \tag{22}$$

By employing the BCD algorithm, each subproblem (21) can be solved with a closed-form solution, which is given by

$$[\mathbf{F}]_{i,j}^* = \exp\big(\jmath\arg\big(\tilde{b}\big)\big), \tag{23}$$

and

$$[\mathbf{F}]_{i,j}^* = \arg\max_{[\mathbf{F}]_{i,j}\in\mathcal{F}_{\mathrm{D}}} \Re\big\{\tilde{b}^{\star}\,[\mathbf{F}]_{i,j}\big\}, \tag{24}$$

for the continuous and discrete phase shifters, respectively. The procedure for solving problem (18) is outlined in Algorithm 1. For continuous phase shifters, Algorithm 1 is guaranteed to converge to a stationary point of the problem (18) [6], [43]. On the other hand, in the case of discrete phase shifters, the objective function of problem (18) is monotonically decreasing in Algorithm 1. Moreover, as the objective function of problem (18) is bounded below by zero, the convergence of Algorithm 1 is guaranteed.

3) Optimization of Auxiliary Variables $\{\mathbf{Q},\mathbf{D}\}$

Given $\mathbf{W}^{(t,n+1)}$ and $\mathbf{F}^{(t,n+1)}$, $\{\mathbf{Q},\mathbf{D}\}$ can be optimized separately and solved in parallel. In particular, the optimization of $\mathbf{D}$ at inner iteration n can be written as

$$\min_{\mathbf{D}} \ \|\mathbf{D} - \mathbf{F}^{(t,n+1)}\mathbf{W}^{(t,n+1)} + \rho^{(t)}\boldsymbol{U}^{(t)}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \|\mathbf{D}\|_F^2 \le P, \tag{25}$$

Algorithm 1 Proposed Algorithm for Solving (18)

1: Initialize $m=1$, $\mathbf{F}^{(m)}$, $\mathbf{\Theta}=\mathbf{A}\mathbf{F}^{(m)}\mathbf{C}$, set tolerance ε_1 and the maximum number of iterations $N_1^{\max}$.
2: **repeat**
3: $\quad\mathbf{F}=\mathbf{F}^{(m)}$,
4: $\quad$**for** $(i,j)\in\{1,\ldots,N_t\}\times\{1,\ldots,N_r\}$ **do**
5: $\quad\quad$Compute $\tilde{b}=[\mathbf{A}]_{i,j}[\mathbf{F}]_{i,j}[\mathbf{C}]_{i,j}-[\mathbf{AFC}]_{i,j}+[\mathbf{B}]_{i,j}$,
6: $\quad\quad$Compute $x=\exp\big(\jmath\arg(\tilde{b})\big)$ and $x=\arg\max_{[\mathbf{F}]_{i,j}\in\mathcal{F}_{\mathrm{D}}}\Re\{\tilde{b}^{\star}[\mathbf{F}]_{i,j}\}$ for the continuous and discrete phase shifters, respectively,
7: $\quad\quad\mathbf{\Theta}=\mathbf{\Theta}+\Big(x-[\mathbf{F}]_{i,j}\Big)[\mathbf{A}]_{:,i}[\mathbf{C}]_{j,:}$,
8: $\quad\quad[\mathbf{F}]_{i,j}=x$,
9: $\quad$**end for**
10: $\quad\mathbf{F}^{(m+1)}=\mathbf{F}$, $m=m+1$,
11: **until** $\big|F_{\mathrm{W},\mathbf{F}}(\mathbf{F}^{(m)})-F_{\mathrm{W},\mathbf{F}}(\mathbf{F}^{(m-1)})\big|/\big|F_{\mathrm{W},\mathbf{F}}(\mathbf{F}^{(m-1)})\big|\leq\varepsilon_1$ or $m\geq N_1^{\max}$.
12: **return** $\mathbf{F}^{(m)}$.

Algorithm 2 Proposed Algorithm for Solving Problem (28)

1: Initialize $m=0$, $\mathbf{q}_k^{(m)}=\mathbf{q}_k^{(t,n)}$, set tolerance ε_2 and the maximum number of iterations $N_2^{\max}$.
2: **repeat**
3: $\quad$Update $\mathbf{q}_k^{(m+1)}$ according to Theorem 1,
4: $\quad m=m+1$,
5: **until** $\big|F_{\mathrm{W},\mathbf{q}_k}\big(\mathbf{q}_k^{(m)}\big)-F_{\mathrm{W},\mathbf{q}_k}\big(\mathbf{q}_k^{(m-1)}\big)\big|\leq\varepsilon_2$ or $m\geq N_2^{\max}$.
6: **return** $\mathbf{q}_k^{(m)}$.

whose optimal solution is given by [45]

$$\mathbf{D}^*=\min\left\{\frac{\sqrt{P}}{\|\mathbf{F}^{(t,n+1)}\mathbf{W}^{(t,n+1)}-\rho^{(t)}\boldsymbol{U}^{(t)}\|_{\mathrm{F}}},1\right\}\cdot\big(\mathbf{F}^{(t,n+1)}\mathbf{W}^{(t,n+1)}-\rho^{(t)}\boldsymbol{U}^{(t)}\big).\tag{26}$$

It remains to optimize $\mathbf{Q}$ at inner iteration n via solving the following problem

$$\max_{\mathbf{Q}}\ \sum\nolimits_{k=1}^{K}\omega_k R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big)-\frac{1}{2\rho^{(t)}}\|\mathbf{Q}-\mathbf{H}^H\mathbf{F}^{(t,n+1)}\mathbf{W}^{(t,n+1)}+\rho^{(t)}\boldsymbol{Z}^{(t)}\|_{\mathrm{F}}^2\tag{27}$$
s.t. (12).

Problem (27) can naturally divide into K independent subproblems, where subproblem k is expressed as

$$\max_{\mathbf{q}_k}\ F_{\mathrm{W},\mathbf{q}_k}(\mathbf{q}_k)\triangleq\omega_k R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big)-\frac{1}{2\rho^{(t)}}\sum_{k\in\mathcal{K}}\big|q_{k,i}-\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}+\rho^{(t)}z_{k,i}^{(t)}\big|^2\tag{28a}$$
s.t. $\Omega_k(\mathbf{q}_k)\leq 0.$ (28b)

Due to the nonconvex FBL rate function and nonconvex constraint (28b), solving the above problem is prohibitively challenging. In the following, we derive a majorization-minimization (MM)-type algorithm [44] to obtain an effective solution of problem (28). First, we establish a concave surrogate function to approximate $R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big)$ while keeping the nonconvex constraint intact.

Lemma 1. *For any feasible* $\mathbf{q}_k^{(m)}$ *satisfying constraint* (28b)*, a lower bound of* $R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big)$ *is given by:*

$$R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big)\geq\underline{R}_k\big(\mathbf{q}_k,\mathbf{q}_k^{(m)}\big)\triangleq a_k^{(m)}+\sum\nolimits_{i=1}^{K}\Re\Big\{\big(b_{k,i}^{(m)}\big)^{\star}q_{k,i}\Big\}-c_k^{(m)}\sum\nolimits_{i=1}^{K}|q_{k,i}|^2,\tag{29}$$

where $\underline{R}_k\big(\mathbf{q}_k,\mathbf{q}_k^{(m)}\big)$ *is a concave function satisfying*

$$\left.\frac{\partial\underline{R}_k\big(\mathbf{q}_k,\mathbf{q}_k^{(m)}\big)}{\partial q_{k,i}}\right|_{\mathbf{q}_k=\mathbf{q}_k^{(m)}}=\left.\frac{\partial R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big)}{\partial q_{k,i}}\right|_{\mathbf{q}_k=\mathbf{q}_k^{(m)}},\tag{30}$$

and $R\big(\hat{\gamma}_k(\mathbf{q}_k^{(m)}),\vartheta_k\big)=\underline{R}_k\big(\mathbf{q}_k^{(m)},\mathbf{q}_k^{(m)}\big)$*. In (29),* $a_k^{(m)}$ *is a constant whose value depends on* $\mathbf{q}_k^{(m)}$ *according to (31) at the top of next page, where* $\beta_k(\mathbf{q}_k^{(m)})=\alpha_k\big(\mathbf{q}_k^{(m)}\big)+\big|q_{k,k}^{(m)}\big|^2$*,*

$$b_{k,i}^{(m)}=\begin{cases}\dfrac{2q_{k,k}^{(m)}}{\alpha_k\left(\mathbf{q}_k^{(m)}\right)}, & \text{if } i=k,\\ \dfrac{2\vartheta_k\alpha_k\left(\mathbf{q}_k^{(m)}\right)q_{k,i}^{(m)}}{\beta_k^2\left(\mathbf{q}_k^{(m)}\right)\sqrt{V\left(\hat{\gamma}_k\left(\mathbf{q}_k^{(m)}\right)\right)}}, & \text{otherwise},\end{cases}\tag{32}$$

$$c_k^{(m)}=\frac{\big|q_{k,k}^{(m)}\big|^2}{\alpha_k\left(\mathbf{q}_k^{(m)}\right)\beta_k\left(\mathbf{q}_k^{(m)}\right)}+\frac{\vartheta_k}{\sqrt{V\left(\hat{\gamma}_k\left(\mathbf{q}_k^{(m)}\right)\right)}}\frac{\alpha_k^2\left(\mathbf{q}_k^{(m)}\right)}{\beta_k^3\left(\mathbf{q}_k^{(m)}\right)}>0,\tag{33}$$

Proof. Please refer to [30], [33], [35]. □

By replacing $R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big)$ with $\underline{R}_k\big(\mathbf{q}_k,\mathbf{q}_k^{(m)}\big)$ in the objective function, problem (28) naturally separates into a sequence of subproblems given as follows:

$$\min_{\mathbf{q}_k}\ \omega_k c_k^{(m)}\sum_{i=1}^{K}|q_{k,i}|^2-\omega_k\sum_{i=1}^{K}\Re\big\{\big(b_{k,i}^{(m)}\big)^{\star}q_{k,i}\big\}+\frac{1}{2\rho^{(t)}}\sum_{i=1}^{K}\big|q_{k,i}-\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}+\rho^{(t)}z_{k,i}^{(t)}\big|^2\tag{34}$$
s.t. (28b).

Although (34) is still nonconvex due to nonconvex constraint (28b), we derive its optimal solution in closed form as follows.

Theorem 1. *When* $\omega_k b_{k,k}^{(m)}+(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_k^{(t,n+1)}=z_{k,k}^{(t)}$*, the optimal solution of problem* (34) *is given by* (35) *at the top of next page, where* $\theta_{W,k}\in[0,2\pi]$ *is arbitrary. On the other hand, when* $(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_k^{(t,n+1)}\neq z_{k,k}^{(t)}-\omega_k b_{k,k}^{(m)}$*, the optimal solution of problem* (34) *is given by*

$$q_{k,i}^*=\frac{\omega_k b_{k,i}^{(m)}+(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-z_{k,i}^{(t)}}{2\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}+2\tau^*\varsigma_{W,k}\bar{\gamma}_k},\tag{36}$$

where $\varsigma_{W,k}=-1$ *if* $i=k$*, and* $\varsigma_{W,k}=\bar{\gamma}_k$ *otherwise. In* (36)*,* $\tau^*=0$ *if* $\Upsilon(0)<0$*,* $\Upsilon(\tau^*)=\Omega(\mathbf{q}_k^*)$*; otherwise,* $0<\tau^*<\omega_k c_k^{(m)}+1/(2\rho^{(t)})$ *is the unique solution of* $\Upsilon(\tau^*)=0$*, which can be solved in a closed form.*

Proof. Please refer to Appendix A. □

The proposed iterative solution for problem (28) is summarized in Algorithm 2, which is guaranteed to converge to a locally optimal solution of (28) [44]. Note that $\{\mathbf{q}_k\}$ can be optimized in parallel.

Remark 1. To solve problem $\mathbb{P}_1$, one can alternatively introduce an auxiliary variable $\tilde{\mathbf{D}}_k=\left(\tilde{\mathbf{d}}_1,\ldots,\tilde{\mathbf{d}}_K\right)=\mathbf{F}\mathbf{W}$ for

$$a_k^{(m)} = \ln\big(1+\hat{\gamma}_k\big(\mathbf{q}_k^{(m)}\big)\big) - \hat{\gamma}_k\big(\mathbf{q}_k^{(m)}\big) - \frac{\sigma_k^2\big|q_{k,k}^{(m)}\big|^2}{\alpha_k\big(\mathbf{q}_k^{(m)}\big)\beta_k\big(\mathbf{q}_k^{(m)}\big)} - \frac{\vartheta_k\sqrt{V\big(\hat{\gamma}_k\big(\mathbf{q}_k^{(m)}\big)\big)}}{2}\left(1+\frac{1}{V\big(\hat{\gamma}_k\big(\mathbf{q}_k^{(m)}\big)\big)}\right) - \frac{\vartheta_k}{2\sqrt{V\big(\hat{\gamma}_k\big(\mathbf{q}_k^{(m)}\big)\big)}}\left(\frac{\alpha_k\big(\mathbf{q}_k^{(m)}\big)}{\beta_k\big(\mathbf{q}_k^{(m)}\big)}\right)^2 + \frac{\vartheta_k\sigma_k^2\alpha_k\big(\mathbf{q}_k^{(m)}\big)}{\beta_k\big(\mathbf{q}_k^{(m)}\big)\sqrt{V\big(\hat{\gamma}_k\big(\mathbf{q}_k^{(m)}\big)\big)}}\left(\frac{2}{\beta_k\big(\mathbf{q}_k^{(m)}\big)} - \frac{\alpha_k\big(\mathbf{q}_k^{(m)}\big)}{\beta_k^2\big(\mathbf{q}_k^{(m)}\big)}\right). \tag{31}$$

$$q_{k,i}^* = \begin{cases} \sqrt{\bar{\gamma}_k}\mathrm{e}^{\jmath\theta_{\mathrm{W},k}}\left(\frac{\sum_{i\neq k}\left|\omega_k b_{k,i}^{(m)}+(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-z_{k,i}^{(t)}\right|^2}{\left|2\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}\right|^2\left|1+\bar{\gamma}_k\right|^2}+\sigma_k^2\right)^{-1/2}, & \text{if } i=k, \\ \frac{\omega_k b_{k,i}^{(m)}+(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-z_{k,i}^{(t)}}{\left(2\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}\right)(1+\bar{\gamma}_k)}, & \text{otherwise.} \end{cases} \tag{35}$$

$$\Delta_{\mathrm{W}}^{(t,n)} = \frac{\left|\mathcal{L}_{\mathrm{W}}(\mathbf{W}^{(t,n)},\mathbf{F}^{(t,n)},\mathbf{Q}^{(t,n)},\mathbf{D}^{(t,n)},\boldsymbol{Z}^{(t)},\boldsymbol{U}^{(t)}) - \mathcal{L}_{\mathrm{W}}(\mathbf{W}^{(t,n-1)},\mathbf{F}^{(t,n-1)},\mathbf{Q}^{(t,n-1)},\mathbf{D}^{(t,n-1)},\boldsymbol{Z}^{(t)},\boldsymbol{U}^{(t)})\right|}{\left|\mathcal{L}_{\mathrm{W}}(\mathbf{W}^{(t,n-1)},\mathbf{F}^{(t,n-1)},\mathbf{Q}^{(t,n-1)},\mathbf{D}^{(t,n-1)},\boldsymbol{Z}^{(t)},\boldsymbol{U}^{(t)})\right|}, \tag{39}$$

$$\Lambda_{\mathrm{W}}^{(t)} = \max\left\{\|\mathrm{vec}\big(\mathbf{Q}^{(t)}-\mathbf{H}^H\mathbf{F}^{(t)}\mathbf{W}^{(t)}\big)\|_\infty^2, \|\mathrm{vec}\big(\mathbf{D}^{(t)}-\mathbf{F}^{(t)}\mathbf{W}^{(t)}\big)\|_\infty^2\right\}. \tag{40}$$

each quadratic QoS constraint (8) as in [6] and transform them into

$$|\mathbf{h}_k^H\tilde{\mathbf{d}}_k|^2 \geq \bar{\gamma}_k\left(\sum\nolimits_{i\neq k}|\mathbf{h}_k^H\tilde{\mathbf{d}}_i|^2+\sigma_k^2\right), \forall k\in\mathcal{K}. \tag{37}$$

However, this approach introduces $(K+1)N_tK$ auxiliary variables, significant increasing computational complexity. In contrast, our proposed solution requires only N_tK+K^2 auxiliary variables, which can be easily updated in closed form. This reduction in dimensionality is particularly beneficial for massive MIMO systems, where computational efficiency is an essential requirement.

C. Feasibility Analysis and Initialization

Due to the limited transmit power budget in (6b) and stringent users' QoS requirements (8), problem $\mathbb{P}_1$ may be infeasible for some channel realizations, e.g., when the users' channel vectors are highly correlated. Meanwhile, our proposed iterative algorithm for solving $\mathbb{P}_1$ requires a feasible initial solution. To this end, we consider to solve the following power minimization problem:

$$\mathbb{P}_5: \min_{\mathbf{W},\mathbf{F}} \ \|\mathbf{FW}\|_F^2 \quad \text{s.t. (6c), (8).} \tag{38}$$

In problem $\mathbb{P}_5$, the digital and analog precoders are jointly optimized to minimize the transmit power subject to the users' QoS constraints. Problem $\mathbb{P}_1$ is feasible if the optimal objective function value of $\mathbb{P}_5$ is less than P; otherwise, it is infeasible. Problem $\mathbb{P}_5$ can be solved using the methods in [17], [19].

D. Overall Solution of Problem $\mathbb{P}_1$

The proposed algorithm to solve $\mathbb{P}_1$ is described in Algorithm 3, where the relative objective progress (ROP) $\Delta_{\mathrm{W}}^{(t,n)}$ and the constraint violation $\Lambda_{\mathrm{W}}^{(t)}$ are defined in (39) and (40) at the top of this page, respectively. In the inner iteration, the variables $\{\mathbf{W},\mathbf{F},\mathbf{Q},\mathbf{D}\}$ are optimized using the BSUM method [43] to address the AL problem $\mathbb{P}_4$. In the outer loop, the dual variables $\{\boldsymbol{U},\boldsymbol{Z}\}$ are updated using the subgradient method when the constraint violation $\Lambda_{\mathrm{W}}^{(t)}$ is relatively small (steps 13 and 14); otherwise, we decrease the penalty parameter (step 19) to gradually reduce the constraint violation. Algorithm 3 adaptively alternates between the AL approach and the penalty method until the constraints (9) and (10) are approximately satisfied, i.e., the constraint violation $\Lambda_{\mathrm{W}}^{(t)}$ is small sufficiently. According to [38], Algorithm 3 is guaranteed to converge to the KKT solutions of problem $\mathbb{P}_1$ for the continuous phase shifters under mild conditions. Note that this convergence result generally does not apply to the discrete phase shifters, as Algorithm 1 may not converge to a stationary point. Despite of this, our numerical results in Section V show good convergence performance of Algorithm 3 for discrete phase shifters.

E. Complexity Analysis

In Algorithm 1, Step 7 is the most computationally demanding, which has a computational complexity $O\left(N_tN_r\right)$. Therefore, the complexity of Algorithm 1 is $O\left(I_FN_t^2N_r^2\right)$, where I_F is the number of iterations required for convergence. Updating $\mathbf{D}$ and $\mathbf{W}$ require complexities of $O\left(N_tN_rK\right)$ and $O\left(N_t^2N_r\right)$, respectively. Moreover, assuming that Algorithm 2 converges after I_Q iterations, its complexity is $O\left(I_QKN_tN_r\right)$. Consequently, the overall computational complexity of Algorithm 3 is $O\left(I_{\mathrm{outer}}I_{\mathrm{inner}}\left(N_t^2N_r+I_FN_t^2N_r^2+I_QKN_tN_r\right)\right)$, where I_{inner} and I_{outer} are the total numbers of iterations of the inner and the outer loops in Algorithm 3, respectively.[5]

IV. MMF Hybrid Precoding Design

In Section III, we have developed Algorithm 3 based on the PDD method to handle the WSR problem $\mathbb{P}_1$. This section extends the PDD method to solve the MMF problem $\mathbb{P}_2$. Unfortunately, the decomposition technique in Section III and Algorithm 3 cannot be directly applied to problem $\mathbb{P}_2$ due to its nonsmooth objective function and complex structure. To

[5]Deep learning (DL) has the potential to facilitate real-time hybrid precoding design by learning the mapping function from wireless channels to corresponding solutions through training a neural network [46]. The DL-based hybrid precoding design in the FBL regime presents a valuable direction for future research.

Algorithm 3 Proposed Algorithm for Solving Problem $\mathbb{P}_1$

Initialize t=0, $\mathbf{F}^{(t)}, \mathbf{W}^{(t)}$, $\mathbf{Q}^{(t)}, \mathbf{D}^{(t)}$, $\boldsymbol{U}^{(t)}, \boldsymbol{Z}^{(t)}$, $\rho^{(t)}$, set tolerance η, penalty update factor ξ, tolerance ε_3 and the maximum number of iterations $N_3^{\max}$.
repeat
  Initialize $n=0, \mathbf{F}^{(t,n)}=\mathbf{F}^{(t)}, \mathbf{W}^{(t,n)}=\mathbf{W}^{(t)}$, $\mathbf{Q}^{(t,n)}=\mathbf{Q}^{(t)}$, set tolerance ε_4 and the maximum number of iterations $N_4^{\max}$,
  repeat
    Update $\mathbf{W}^{(t,n+1)}$ using (17),
    Update $\mathbf{F}^{(t,n+1)}$ using Algorithm 1,
    Update $\mathbf{D}^{(t,n+1)}$ using (26),
    Update $\mathbf{q}_k^{(t,n+1)}, k=1,\ldots,K$, using Algorithm 2,
    $n=n+1$,
  until $\Delta_{\mathrm{W}}^{(t,n)} \le \varepsilon_4$ or $n \ge N_4^{\max}$.
  $\mathbf{F}^{(t+1)}=\mathbf{F}^{(t,n)}, \mathbf{W}^{(t+1)}=\mathbf{W}^{(t,n)}$, $\mathbf{Q}^{(t+1)}=\mathbf{Q}^{(t,n)}$, $\mathbf{D}^{(t+1)}=\mathbf{D}^{(t,n)}$,
  if $\Lambda_{\mathrm{W}}^{(t+1)} \le \eta$ **then**
    $\boldsymbol{Z}^{(t+1)}=\boldsymbol{Z}^{(t)}+(\rho^{(t)})^{-1}\big(\mathbf{Q}^{(t+1)}-\mathbf{H}^H\mathbf{F}^{(t+1)}\mathbf{W}^{(t+1)}\big)$,
    $\boldsymbol{U}^{(t+1)}=\boldsymbol{U}^{(t)}+(\rho^{(t)})^{-1}\big(\mathbf{D}^{(t+1)}-\mathbf{F}^{(t+1)}\mathbf{W}^{(t+1)}\big)$,
    $\rho^{(t+1)}=\rho^{(t)}$,
  else
    $\boldsymbol{Z}^{(t+1)}=\boldsymbol{Z}^{(t)}$,
    $\boldsymbol{U}^{(t+1)}=\boldsymbol{U}^{(t)}$,
    $\rho^{(t+1)}=\xi\rho^{(t)}$,
  end if
  $t=t+1$,
until $\Lambda_{\mathrm{W}}^{(t)} \le \varepsilon_3$ or $t \ge N_3^{\max}$.
return $\{\mathbf{F}=\mathbf{F}^{(t)}, \mathbf{W}=\mathbf{W}^{(t)}\}$.

overcome this issue and solve $\mathbb{P}_2$, we have to combine the PDD method with novel problem reformulation and decomposition techniques to be detailed below.

A. Transformation of Problem $\mathbb{P}_2$

Following (8), we introduce an auxiliary variable r and reformulate the MMF optimization problem $\mathbb{P}_2$ into the following equivalent smooth problem

$$\begin{aligned}\mathbb{P}_6: \max_{\mathbf{W},\mathbf{F},r}\ & r \\ \text{s.t.}\ & \text{(6b)}, \text{(6c)}, (8), \\ & R(\gamma_k,\vartheta_k) \ge r, \forall k \in \mathcal{K}. \end{aligned} \tag{41}$$

Problem $\mathbb{P}_6$ is intractable due to the coupling between analog and digital precoders in constraints (6b), (8), and (41). To tackle this challenge, we further introduce the auxiliary variables $\mathbf{Q}$ and $\mathbf{D}$, cf. (9), (10), and

$$p_{i,j} = \mathbf{h}_i^H\mathbf{F}\mathbf{w}_j, \forall i,j \in \mathcal{K}, \tag{42}$$

$$\bar{r}_k = r, \forall k \in \mathcal{K}. \tag{43}$$

Then constraints (6b), (8), and (41) can be equivalently transformed into (11),

$$\tilde{\Omega}_k(\mathbf{p}_k) = \bar{\gamma}_k\big(\textstyle\sum_{i\ne k}\big|p_{k,i}\big|^2+\sigma_k^2\big)-\big|p_{k,k}\big|^2 \le 0, \forall k \in \mathcal{K}, \tag{44}$$

$$\bar{r}_k - R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big) \le 0, \forall k \in \mathcal{K}, \tag{45}$$

respectively, where $\mathbf{p}_k = (p_{k,1},\ldots,p_{k,K})^T$. Finally, $\mathbb{P}_6$ can be equivalently reformulated as

$$\begin{aligned}\mathbb{P}_7: \max_{\mathbf{W},\mathbf{F},r,\mathbf{P},\mathbf{D},\mathbf{Q},\bar{\mathbf{r}}}\ & r \\ \text{s.t.}\ & \text{(6c)}, (9), (10), (11), (42), (43), (44), (45), \end{aligned} \tag{46}$$

where $\mathbf{P} \triangleq [\mathbf{p}_1,\ldots,\mathbf{p}_K]^T \in \mathbb{C}^{K\times K}$ and $\bar{\mathbf{r}} = (\bar{r}_1,\ldots,\bar{r}_K)^T \in \mathbb{R}^{K\times 1}$. Similar to problem $\mathbb{P}_3$, the variables of problem $\mathbb{P}_7$ are only coupled in equality constraints (9), (10), (42), and (43), Then, we utilize the PDD method to solve problem $\mathbb{P}_7$, whose AL problem is expressed as

$$\begin{aligned}\mathbb{P}_8: \max_{\mathbf{W},\mathbf{F},r,\mathbf{P},\mathbf{D},\mathbf{Q},\bar{\mathbf{r}}}\ & \mathcal{L}_{\mathrm{M}}(\mathbf{W},\mathbf{F},r,\mathbf{P},\mathbf{D},\mathbf{Q},\bar{\mathbf{r}};\boldsymbol{E},\boldsymbol{Z},\boldsymbol{U},\boldsymbol{v}) \\ \text{s.t.}\ & \text{(6c)}, (11), (44), (45), \end{aligned} \tag{47}$$

where

$$\begin{aligned}&\mathcal{L}_{\mathrm{M}}(\mathbf{W},\mathbf{F},r,\mathbf{P},\mathbf{D},\mathbf{Q},\bar{\mathbf{r}};\boldsymbol{E},\boldsymbol{Z},\boldsymbol{U},\boldsymbol{v}) \\ &= r-\frac{1}{2\rho}\left\|\mathbf{D}-\mathbf{F}\mathbf{W}+\rho\boldsymbol{U}\right\|_{\mathrm{F}}^2-\frac{1}{2\rho}\left\|\mathbf{Q}-\mathbf{H}^H\mathbf{F}\mathbf{W}+\rho\boldsymbol{Z}\right\|_{\mathrm{F}}^2 \\ &-\frac{1}{2\rho}\left\|\bar{\mathbf{r}}-r\mathbf{1}_K+\rho\boldsymbol{v}\right\|_2^2-\frac{1}{2\rho}\left\|\mathbf{P}-\mathbf{H}^H\mathbf{F}\mathbf{W}+\rho\boldsymbol{E}\right\|_{\mathrm{F}}^2, \end{aligned} \tag{48}$$

$\boldsymbol{Z}$, $\boldsymbol{U}$, $\boldsymbol{E} = [\mathbf{e}_1,\ldots,\mathbf{e}_K]^T$, and $\boldsymbol{v} = (v_1,\ldots,v_K)^T$ are the dual variables for constraints (9), (10), (42), and (43), respectively. In problem $\mathbb{P}_8$, the variables naturally divide into three blocks, i.e., the auxiliary variables $\{\mathbf{Q},\mathbf{D},\mathbf{P},\bar{\mathbf{r}}\}$, $\{\mathbf{W},r\}$, and the analog precoder $\mathbf{F}$, where different blocks are not coupled in constraints. Thanks to this decomposable structure, problem $\mathbb{P}_8$ can be solved leveraging the BSUM approach. We denote the value of ρ, $\boldsymbol{v}$, $\boldsymbol{U}$, $\boldsymbol{Z}$, and $\boldsymbol{E}$ in the t-th outer iteration as $\rho^{(t)}$, $\boldsymbol{v}^{(t)}$, $\boldsymbol{U}^{(t)}$, $\boldsymbol{Z}^{(t)}$, and $\boldsymbol{E}^{(t)}$, and let $\mathbf{W}^{(t,n)}$, $\mathbf{F}^{(t,n)}$, $r^{(t,n)}$, $\mathbf{P}^{(t,n)}$, $\mathbf{D}^{(t,n)}$, $\mathbf{Q}^{(t,n)}$, and $\bar{\mathbf{r}}^{(t,n)}$ denote the value of $\mathbf{W}$, $\mathbf{F}$, r, $\mathbf{P}$, $\mathbf{D}$, $\mathbf{Q}$, and $\bar{\mathbf{r}}$ in the n-th inner iteration within the t-th outer iteration. The BSUM algorithm to solve the AL problem $\mathbb{P}_8$ within the t-th outer iteration with given $\rho^{(t)}$, $\boldsymbol{v}^{(t)}$, $\boldsymbol{U}^{(t)}$, $\boldsymbol{Z}^{(t)}$, and $\boldsymbol{E}^{(t)}$ is elaborated in the following.

B. Solving the AL Problem $\mathbb{P}_8$

The problem transformation ($\mathbb{P}_7$) and the PDD method enable us to derive an efficient BCD-type algorithm to solve problem $\mathbb{P}_8$, where the variables are optimized with efficient and even closed-form solutions.

1) Optimization of $\{\mathbf{W}, r\}$

With given $\mathbf{F}^{(t,n)}$, $\mathbf{P}^{(t,n)}$, $\mathbf{D}^{(t,n)}$, $\mathbf{Q}^{(t,n)}$, and $\bar{\mathbf{r}}^{(t,n)}$, the optimization of $\{\mathbf{W}, r\}$ can be written as the following unconstrainted convex optimization problem

$$\begin{aligned}\min_{\mathbf{W},r}\ & \mathcal{L}_{\mathrm{M}}\big(\mathbf{W},r,\mathbf{F}^{(t,n)},\mathbf{P}^{(t,n)},\mathbf{D}^{(t,n)},\mathbf{Q}^{(t,n)} \\ & ,\bar{\mathbf{r}}^{(t,n)};\boldsymbol{E}^{(t)},\boldsymbol{Z}^{(t)},\boldsymbol{U}^{(t)},\boldsymbol{v}^{(t)}\big). \end{aligned} \tag{49}$$

By setting the first-order derivative to zero, the optimal solution is given by

$$r^* = \big(\textstyle\sum_{i=1}^K\big(\bar{r}_i^{(t,n)}+\rho^{(t)}v_i^{(t)}\big)+\rho^{(t)}\big)/K, \tag{50}$$

and (51) at the top of next page.

$$\mathbf{W}^*=\Big(\big(\mathbf{F}^{(t,n)}\big)^H\mathbf{F}^{(t,n)}+2\big(\mathbf{F}^{(t,n)}\big)^H\mathbf{H}\mathbf{H}^H\mathbf{F}^{(t,n)}\Big)^\dagger\big(\mathbf{F}^{(t,n)}\big)^H\Big\{\mathbf{D}^{(t,n)}+\rho^{(t)}\boldsymbol{U}^{(t)}+\mathbf{H}\Big(\mathbf{Q}^{(t,n)}+\mathbf{P}^{(t,n)}+\rho^{(t)}\boldsymbol{Z}^{(t)}+\rho\boldsymbol{E}^{(t)}\Big)\Big\}. \quad (51)$$

$$p_{k,i}^*=\begin{cases}(1+\bar{\gamma}_k)^{-1}(\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n)}-\rho^{(t)}e_{k,i}^{(t)}), & \text{if } i\neq k,\\ \sqrt{\bar{\gamma}_k\Big|\sum_{i\neq k}\Big(1+\bar{\gamma}_k\Big)^{-2}\Big|\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-\rho^{(t)}e_{k,i}^{(t)}\Big|^2+\sigma_k^2\Big|}e^{\jmath\theta_{\mathrm{M},k}}. & \text{otherwise,}\end{cases} \quad (55)$$

2) Optimization of $\mathbf{F}$

With given $\mathbf{W}^{(t,n+1)}$, $\mathbf{P}^{(t,n)}$, $\mathbf{D}^{(t,n)}$, and $\mathbf{Q}^{(t,n)}$, the optimization of the analog precoder $\mathbf{F}$ can be written as

$$\begin{aligned}\min_{\mathbf{F}}\ &\mathrm{Tr}\big(\mathbf{F}^H\big(\mathbf{I}_{N_t}+2\mathbf{H}\mathbf{H}^H\big)\mathbf{F}\mathbf{W}^{(t,n+1)}\big(\mathbf{W}^{(t,n+1)}\big)^H\big)\\ &-2\Re\Big\{\mathrm{Tr}\Big\{\mathbf{F}^H\Big(\mathbf{D}^{(t,n)}+\rho^{(t)}\boldsymbol{U}^{(t)}+\mathbf{H}\Big(\mathbf{Q}^{(t,n)}\\ &+\mathbf{P}^{(t,n)}+\rho^{(t)}\boldsymbol{Z}^{(t)}+\rho^{(t)}\boldsymbol{E}^{(t)}\Big)\Big)\big(\mathbf{W}^{(t,n+1)}\big)^H\Big\}\Big\} \quad (52)\\ \text{s.t. }&(6c),\end{aligned}$$

which admit an identical form as problem (19) and can be efficiently solved exploiting Algorithm 1 for both the continuous and discrete phase shifters.

3) Optimizing $\{\mathbf{P},\mathbf{D},\mathbf{Q},\mathbf{r}\}$

With given $\mathbf{W}^{(t,n+1)}$, $\mathbf{F}^{(t,n+1)}$, and $r^{(t,n+1)}$, the optimization of $\{\mathbf{P},\mathbf{D},\mathbf{Q},\bar{\mathbf{r}}\}$ at inner iteration n divides into three independent optimization subproblems w.r.t. $\mathbf{P}$, $\mathbf{D}$, and $\{\mathbf{Q},\bar{\mathbf{r}}\}$, respectively. The optimization of $\mathbf{D}$ is essentially the same as problem (25), whose optimal solution is given by (26).

The optimization of $\mathbf{P}$ can be written as

$$\begin{aligned}\min_{\mathbf{P}}\ &\Big\|\mathbf{P}-\mathbf{H}^H\mathbf{F}^{(t,n+1)}\mathbf{W}^{(t,n+1)}+\rho^{(t)}\boldsymbol{E}^{(t)}\Big\|_{\mathrm{F}}^2 \quad (53)\\ \text{s.t. }&(44).\end{aligned}$$

Problem (53) is further decomposed into K independent subproblems and subproblem $k\in\mathcal{K}$ can be written as

$$\begin{aligned}\min_{\mathbf{p}_k}\ &F_{\mathrm{M},\mathbf{p}_k}(\mathbf{p}_k)\triangleq\sum\nolimits_{i=1}^K\big|p_{k,i}-\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}+\rho^{(t)}e_{k,i}^{(t)}\big|^2\\ \text{s.t. }&\tilde{\Omega}_k(\mathbf{p}_k)\leq 0. \quad (54)\end{aligned}$$

Problem (54) is nonconvex due to its nonconvex constraint. Note that problem (54) would have the same structure as (28) provided the first term of the latter objective function vanishes. Owing to this slight difference, we show below that, unlike (28), problem (54) can be optimally solved in a closed form.

Theorem 2. *If* $\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_k^{(t,n+1)}-\rho^{(t)}e_{k,k}^{(t)}=0$*, the optimal solution of problem* (54) *is given by* (55) *at the top of this page, where* $\theta_{M,k}\in[0,2\pi]$ *is arbitrary. When* $\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_k^{(t,n+1)}-\rho^{(t)}e_{k,k}^{(t)}\neq 0$*, the optimal solution of* (54) *is given by*

$$p_{k,i}^*=\frac{\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-\rho^{(t)}e_{k,i}^{(t)}}{1+\varsigma_{M,k}\varpi^*}, \quad (56)$$

where $\varsigma_{M,k}=-1$ *if* $i=k$ *and* $\varsigma_{M,k}=\bar{\gamma}_k$ *otherwise. In* (56)*, the optimal dual variable* $\varpi^*=0$ *when* $\Phi(0)\leq 0$*,* $\Phi(\varpi^*)=\tilde{\Omega}_k(\mathbf{p}_k^*)$*; otherwise, there exists the unique solution* $0\leq\varpi^*<1$ *satisfying* $\Phi(\varpi^*)=0$*, which can be solved in a closed form.*

Proof. Theorem 2 can be obtained by replacing $q_{k,i}^*$ and $z_{k,i}$ by $p_{k,i}^*$ and $e_{k,i}$ and setting $\omega_k=0$ in Theorem 1, respectively. □

The optimization of $\{\mathbf{Q},\bar{\mathbf{r}}\}$ can be expressed as

$$\begin{aligned}\min_{\mathbf{Q},\bar{\mathbf{r}}}\ &\sum\nolimits_{k=1}^K\Big|\bar{r}_k-r^{(t,n+1)}+\rho^{(t)}v_k^{(t)}\Big|^2\\ &+\Big\|\mathbf{Q}-\mathbf{H}^H\mathbf{F}^{(t,n+1)}\mathbf{W}^{(t,n+1)}+\rho^{(t)}\boldsymbol{Z}^{(t)}\Big\|_{\mathrm{F}}^2 \quad (57)\\ \text{s.t. }&(45).\end{aligned}$$

Problem (57) can be further divided into K independent subproblems indexed by $k\in\mathcal{K}$, with subproblem k given as

$$\begin{aligned}\min_{\mathbf{q}_k,\bar{r}_k}\ &F_{\mathrm{M},\mathbf{q}_k,\bar{r}_k}(\mathbf{q}_k,\bar{r}_k)\triangleq\Big|\bar{r}_k-r^{(t,n+1)}+\rho^{(t)}v_k^{(t)}\Big|^2\\ &+\sum_{i=1}^K\Big|q_{k,i}-\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}+\rho^{(t)}z_{k,i}^{(t)}\Big|^2 \quad (58)\\ \text{s.t. }&\bar{r}_k-R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big)\leq 0.\end{aligned}$$

Problem (58) is nonconvex due to its nonconvex constraint, which cannot be reduced to (8) and tackled similar to the WSR problem due to the additional variable $\bar{r}_k$. Therefore, although problem (58) has a similar form as (54), it is much more challenging to solve. Particularly, unlike (54), problem (58) does not admit closed-form optimal solutions, due to the nonconvex FBL rate function $R\big(\hat{\gamma}_k(\mathbf{q}_k),\vartheta_k\big)$ in the constraint. Instead, we approximate the FBL rate function by (29), which enables to obtain a locally optimal solution of problem (58) by solving a sequence of subproblems [44]. At the m-th iteration, the subproblem is expressed as

$$\begin{aligned}\min_{\mathbf{q}_k,\bar{r}_k}\ &F_{\mathrm{M},\mathbf{q}_k,\bar{r}_k}(\mathbf{q}_k,\bar{r}_k) \quad (59a)\\ \text{s.t. }&\Upsilon_k(\bar{r}_k,\mathbf{q}_k,\mathbf{q}_k^{(m)})=\bar{r}_k-\underline{R}_k(\mathbf{q}_k,\mathbf{q}_k^{(m)})\leq 0, \quad (59b)\end{aligned}$$

where $\mathbf{q}_k^{(m)}$ is a feasible solution in the m-th iteration. Problem (59) is a convex problem and can be optimally solved in a closed form as shown in the following Theorem.

Theorem 3. *The optimal solution of problem (59) is given by*

$$\bar{r}_k^*=r^{(t,n+1)}-\rho^{(t)}v_k^{(t)}-\mu^*/2, \quad (60a)$$

$$q_{k,i}^*=\frac{\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-\rho^{(t)}z_{k,i}^{(t)}+\mu^*b_{k,i}^{(m)}/2}{1+\mu^*c_k^{(m)}}, \quad (60b)$$

where $\mu^*=0$ *if* $\phi(0)<0$*,* $\phi(\mu^*)=\Upsilon_k(\bar{r}_k^*,\mathbf{q}_k^*)$*; otherwise,* $\mu^*>0$ *is the unique root of* $\phi(\mu^*)=0$ *and can be obtained in a closed form.*

Proof. First, the strong duality holds for problem (59) as it satisfies the Slater's condition. Then the optimal solution (60) is derived from the KKT conditions of problem (59) similar to Theorem 1. The detailed proof is omitted here due to the limited page space. □

The MM-type algorithm to problem (58) is summarized in

Algorithm 4 Proposed Algorithm for Solving Problem (58)

1: Initialize $m = 0$, $\mathbf{q}_k^{(m)} = \mathbf{q}_k^{(t,n)}$, $r_k^{(m)} = r_k^{(t,n)}$, set tolerance ε_5 and the maximum number of iteration $N_5^{\max}$.
2: **repeat**
3: Update $\mathbf{q}_k^{(m+1)}$ and $r_k^{(m+1)}$ according to Theorem 3,
4: $m = m + 1$,
5: **until** $\Delta_{\text{QR}}^{(m)} \leq \varepsilon_5$ or $m \geq N_5^{\max}$.
6: **return** $\mathbf{q}_k^{(m)}$ and $r_k^{(m)}$.

Algorithm 4, where

$$\Delta_{\text{QR}}^{(m)} = \frac{\left| F_{\text{M},\mathbf{q}_k,\bar{r}_k}\left(\mathbf{q}_k^{(m)}, \bar{r}_k^{(m)}\right) - F_{\text{M},\mathbf{q}_k,\bar{r}_k}\left(\mathbf{q}_k^{(m-1)}, \bar{r}_k^{(m-1)}\right) \right|}{\left| F_{\text{M},\mathbf{q}_k,\bar{r}_k}\left(\mathbf{q}_k^{(m-1)}, \bar{r}_k^{(m-1)}\right) \right|}. \tag{61}$$

According to [44], Algorithm 4 is guaranteed to converge to a locally optimal solution of problem (58). Note that $\{\bar{r}_k, \mathbf{q}_k\}$ can be optimized in parallel.

Remark 2. By introducing the auxiliary variables $\bar{r}_k$, $\forall k \in \mathcal{K}$, in (43), r admits a closed-form solution given in (50), and the optimization of $\mathbf{Q}$ also decomposes into small subproblems that can be easily solved in parallel. Both of these strategies facilitate rapid computation, which is attractive for massive MIMO. Without such transformation, the optimization of $\mathbf{Q}$ and r would become a complex nonconvex problem, rendering its solution highly challenging and computationally expensive.

C. Overall Solution of Problem $\mathbb{P}_2$

Algorithm 5 summarizes the procedure for solving problem $\mathbb{P}_2$, where we define the ROP $\Delta_{\text{M}}^{(t,n)}$ and the constraint violation $\Lambda_{\text{M}}^{(t)}$ in (62) and (63) at the top of next page, respectively. Similar to the solution of problem $\mathbb{P}_1$, we propose to first solve problem $\mathbb{P}_5$, which enables a feasibility check of problem $\mathbb{P}_2$ and provides an initial solution to start Algorithm 5. In Algorithm 5, we update the primal variables $\{\mathbf{W}, \mathbf{F}, r, \mathbf{P}, \mathbf{D}, \mathbf{Q}, \bar{\mathbf{r}}\}$ in the inner loop and adaptively update the dual variables and the penalty parameter in the outer loop until convergence. For the continuous phase shifters, Algorithm 5 is guaranteed to converge to a set of KKT solutions of problem $\mathbb{P}_2$ under mild conditions [38]. Though this convergence result does not apply to discrete phase shifters, our numerical results in Section V demonstrate that Algorithm 5 achieves good convergence even in this case.

D. Complexity Analysis

Updating $\mathbf{D}$, $\mathbf{P}$, and $\{\mathbf{W}, r\}$ require complexities of $O(N_t N_r K)$, $O(N_t N_r K)$, and $O\left(N_t^2 N_r\right)$, respectively. Moreover, assuming that Algorithm 4 converges after $I_{Q,r}$ iterations, its complexity is $O(I_{Q,r} K N_t N_r)$. Therefore, the overall computational complexity of Algorithm 5 is $O\left(\tilde{I}_{\text{outer}} \tilde{I}_{\text{Mnner}} \left(N_t^2 N_r + I_F N_t^2 N_r^2 + I_{Q,r} K N_t N_r\right)\right)$, where $\tilde{I}_{\text{Minner}}$ and $\tilde{I}_{\text{Mouter}}$ are the total numbers of iterations of the inner and the outer loops in Algorithm 5, respectively.

V. Simulation Results

In this section, we evaluate the performance of the proposed algorithms by simulation. The BS is equipped with a uniform linear array comprising $N_t = 64$ antennas and serves $K = 12$ single antenna users over a transmission blocklength of

Algorithm 5 Proposed Algorithm for Solving $\mathbb{P}_2$

1: Initialize $t = 0$, $\mathbf{F}^{(t)}, \mathbf{W}^{(t)}$, $\mathbf{Q}^{(t)}, \mathbf{P}^{(t)}$, $r^{(t)}$, $\mathbf{r}^{(t)}$,$\boldsymbol{U}^{(t)}, \boldsymbol{Z}^{(t)}$, $\boldsymbol{E}^{(t)}$, $\boldsymbol{v}^{(t)}$, $\rho^{(t)}$, set tolerance η, penalty update factor ξ, tolerance ε_6 and the maximum number of iterations $N_6^{\max}$.
2: **repeat**
3: Initialize $n = 0$, $\mathbf{F}^{(t,n)} = \mathbf{F}^{(t)}, \mathbf{W}^{(t,n)} = \mathbf{W}^{(t)}$, $\mathbf{Q}^{(t,n)} = \mathbf{Q}^{(t)}$, $\mathbf{P}^{(t,n)} = \mathbf{P}^{(t)}$, set tolerance ε_7 and the maximum number of iterations $N_7^{\max}$,
4: **repeat**
5: Update $r^{(t,n+1)}$ and $\mathbf{W}^{(t,n+1)}$ using (50) and (51), respectively,
6: Update $\mathbf{F}^{(t,n+1)}$ using Algorithm 1,
7: Update $\mathbf{D}^{(t,n+1)}$ using (26),
8: Update $\mathbf{p}_k^{(t,n+1)}, k = 1, \ldots, K$, based on Theorem 2,
9: Update $\mathbf{q}_k^{(t,n+1)}$ and $r_k^{(t,n+1)}$, $k = 1, \ldots, K$, using Algorithm 4,
10: $n = n + 1$,
11: **until** $\Delta_{\text{M}}^{(t,n)} \leq \varepsilon_7$ or $n \geq N_7^{\max}$.
12: $\mathbf{F}^{(t+1)} = \mathbf{F}^{(n)}, \mathbf{W}^{(t+1)} = \mathbf{W}^{(n)}$, $\mathbf{Q}^{(t+1)} = \mathbf{Q}^{(n)}$, $\mathbf{P}^{(t+1)} = \mathbf{P}^{(n)}$, $\mathbf{D}^{(t+1)} = \mathbf{D}^{(n)}$,
13: **if** $\Lambda_{\text{M}}^{(t+1)} \leq \eta$ **then**
14: $\boldsymbol{Z}^{(t+1)} = \boldsymbol{Z}^{(t)} + (\rho^{(t)})^{-1}\left(\mathbf{Q}^{(t+1)} - \mathbf{H}^H \mathbf{F}^{(t+1)} \mathbf{W}^{(t+1)}\right)$,
15: $\boldsymbol{E}^{(t+1)} = \boldsymbol{E}^{(t)} + (\rho^{(t)})^{-1}\left(\mathbf{P}^{(t+1)} - \mathbf{H}^H \mathbf{F}^{(t+1)} \mathbf{W}^{(t+1)}\right)$,
16: $\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)} + (\rho^{(t)})^{-1}\left(\mathbf{r}^{(t+1)} - r^{(t+1)} \mathbf{1}_K\right)$,
17: $\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)} + (\rho^{(t)})^{-1}\left(\mathbf{D}^{(t+1)} - \mathbf{F}^{(t+1)} \mathbf{W}^{(t+1)}\right)$,
18: $\rho^{(t+1)} = \rho^{(t)}$,
19: **else**
20: $\boldsymbol{Z}^{(t+1)} = \boldsymbol{Z}^{(t)}$,
21: $\boldsymbol{E}^{(t+1)} = \boldsymbol{E}^{(t)}$,
22: $\boldsymbol{v}^{(t+1)} = \boldsymbol{v}^{(t)}$,
23: $\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t)}$,
24: $\rho^{(t+1)} = \xi \rho^{(t)}$,
25: **end if**
26: $t = t + 1$,
27: **until** $\Lambda_{\text{M}}^{(t)} \leq \varepsilon_6$ or $t \geq N_6^{\max}$.
28: **return** $\{\mathbf{F} = \mathbf{F}^{(t)}, \mathbf{W} = \mathbf{W}^{(t)}\}$.

N. We adopt a geometric mmWave channel model of $L = 15$ paths with isotropic scatterers [6], [9], where the channel vector between the BS and user k can be expressed as

$$\mathbf{h}_k = \sqrt{\frac{1}{\Gamma_k L}} \sum\nolimits_{l=1}^{L} \zeta_{k,l} \mathbf{a}_t\left(\psi_{k,l}\right), \; k \in \mathcal{K}, \tag{64}$$

$\Gamma_k = 10^{0.366 + 4.14 \log_{10}(d_k) + 2.43 \log_{10}(f_c)}$ is the path-loss of user k experienced at a distance d_k from the BS, $f_c = 28$ GHz is the carrier frequency, and $\zeta_{k,l} \sim \mathcal{CN}(0,1)$ is the complex gain of the lth path between the BS and user k [47]. The users are uniformly and randomly located in a disk with an inner radius of 50 meters and an outer radius of 200 meters. The BS antennas are separated by a spacing of half wavelength. Consequently, the antenna array response at the transmitter $\mathbf{a}_t(\psi_{k,l})$ w.r.t. the azimuth angle $\psi_{k,l} \in [0, 2\pi]$ is given by

$$\mathbf{a}_t(\psi_{k,l}) = \left[1, e^{j\pi \sin(\psi_{k,l})}, \ldots, e^{j\pi(N_t - 1)\sin(\psi_{k,l})}\right]^T, \tag{65}$$

where the azimuth angles $\{\psi_{k,l}\}$ are uniformly distributed within the interval $[0, 2\pi]$. Unless otherwise specified, we set the BLER $\epsilon_k = 10^{-\iota(k)}$ for FBL transmission according

$$\Delta_{\mathrm{M}}^{(t,n)}=\Big|\mathcal{L}_{\mathrm{M}}\Big(\mathbf{W}^{(t,n)},r^{(t,n)},\mathbf{F}^{(t,n)},\mathbf{P}^{(t,n)},\mathbf{D}^{(t,n)},\mathbf{Q}^{(t,n)},\bar{\mathbf{r}}^{(t,n)};\boldsymbol{E}^{(t)},\boldsymbol{Z}^{(t)},\boldsymbol{U}^{(t)},\boldsymbol{v}^{(t)}\Big)$$
$$-\mathcal{L}_{\mathrm{M}}\Big(\mathbf{W}^{(t,n-1)},r^{(t,n-1)},\mathbf{F}^{(t,n-1)},\mathbf{P}^{(t,n-1)},\mathbf{D}^{(t,n-1)},\mathbf{Q}^{(t,n-1)},\bar{\mathbf{r}}^{(t,n-1)};\boldsymbol{E}^{(t)},\boldsymbol{Z}^{(t)},\boldsymbol{U}^{(t)},\boldsymbol{v}^{(t)}\Big)\Big|$$
$$/\Big|\mathcal{L}_{\mathrm{M}}\Big(\mathbf{W}^{(t,n-1)},r^{(t,n-1)},\mathbf{F}^{(t,n-1)},\mathbf{P}^{(t,n-1)},\mathbf{D}^{(t,n-1)},\mathbf{Q}^{(t,n-1)},\bar{\mathbf{r}}^{(t,n-1)};\boldsymbol{E}^{(t)},\boldsymbol{Z}^{(t)},\boldsymbol{U}^{(t)},\boldsymbol{v}^{(t)}\Big)\Big|, \quad (62)$$

$$\Lambda_{\mathrm{M}}^{(t)}=\max\Big\{\Big\|\mathrm{vec}\big(\mathbf{Q}^{(t)}-\mathbf{H}^H\mathbf{F}^{(t)}\mathbf{W}^{(t)}\big)\Big\|_\infty^2,\Big\|\mathrm{vec}\big(\mathbf{D}^{(t)}-\mathbf{F}^{(t)}\mathbf{W}^{(t)}\big)\Big\|_\infty^2,\Big\|\mathrm{vec}\big(\mathbf{P}^{(t)}-\mathbf{H}^H\mathbf{F}^{(t)}\mathbf{W}^{(t)}\big)\Big\|_\infty^2,\Big\|\bar{\mathbf{r}}^{(t)}-r^{(t)}\mathbf{I}_{K\times 1}\Big\|_\infty^2\Big\}. \quad (63)$$

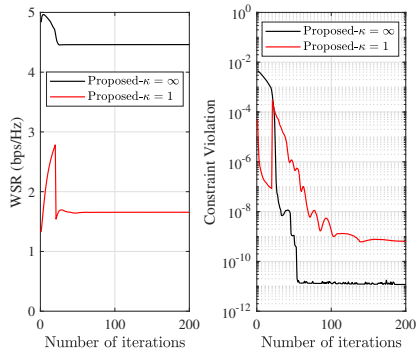

Figure 1. The convergence performance of Algorithm 3 for WSR based hybrid precoder design ($N_t = 64$, $K = 12$, $N = 100$, and $P = 2$ dBm).

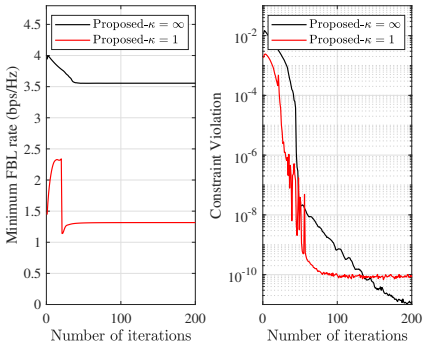

Figure 2. The convergence performance of Algorithm 5 for MMF based hybrid precoder design ($N_t = 64$, $K = 12$, $N = 100$, and SNR= –7 dB).

to $\iota(k) = \min\{5 + (\lceil k/2\rceil - 1) \times \lceil 5/(K-1)\rceil, 10\}$, $\omega_k = 1/K$, $\sigma_k^2 = \sigma^2 = 10^{-11.4}$ mW, and the minimum rate requirements $\bar{R}_k = \bar{R} = 1$ bits/s/Hz. The results are averaged over 200 random independent channel realizations. Similar to [28], the value of the objective function of problems $\mathbb{P}_1$ and $\mathbb{P}_2$ are set to zero if the obtained solution violates any constraints to account for the penalty. Additionally, when using our proposed Algorithms 3 and 5 to optimize the hybrid precoding for discrete phase shifters, their first 20 iterations are performed assuming continuous phase shifters, which can effectively avoid inefficient suboptimal solutions according to our observations.

A. Convergence Performance

In this subsection, we validate the convergence of the proposed algorithms. Unless otherwise specified, we set the relevant algorithmic hyperparameters as $\xi = 0.95$, $\rho^{(0)} = 0.3$, $\eta = 1e^{-3}$, $\epsilon_1 = \epsilon_2 = 10^{-4}$, $\epsilon_3 = \epsilon_6 = 10^{-9}$, $\epsilon_4 = \epsilon_5 = \epsilon_7 = 10^{-4}$, $N_1^{\max} = N_2^{\max} = 20$, and $N_3^{\max} = N_4^{\max} = 100$. Figs. 1 and 2 show the convergence of our proposed Algorithms 3 and 5 by plotting the objective values and the constraint violations averaged over 200 simulations for phase shifters with infinite, i.e., $\kappa = \infty$, and 1-bit resolutions, i.e., $\kappa = 1$, respectively. We observe from Figs. 1 and 2 that the proposed algorithms converge rapidly for both continuous and discrete phase shifters. However, the hybrid precoding design based on the MMF criteria requires on average more iterations to converge than that based on the WSR due to the more complex problem structure in the former. Note that for the hybrid precoder designs with discrete phase shifters, ideal continuous phase shifters are used in the first 20 iterations, while discrete phase shifters are employed in the subsequent iterations. As a result, a noticeable discontinuity is observed in the convergence curves of our proposed Algorithms 3 and 5 in Figs. 1 and 2, reflecting the switch from continuous to discrete phase shifter modeling.

B. WSR Hybrid Precoding

In this subsection, we evaluate the WSR performance of the following schemes: (i) "Proposed", namely the proposed Algorithm 3; (ii) "FD-FBL", which optimizes the WSR for fully digital precoding in the FBL regime subject to the power and QoS constraints (6b) and (6d) as considered in [29], [30]; (iii) "FD-IBL", which optimizes the WSR for fully digital precoding under the power and QoS constraints but adopting the Shannon capacity formula as the performance metric; (iv) "FD-Conventional", which evaluates the FBL rate when employing the fully digital precoder obtained in (iii); (v) "MAP", which designs the hybrid precoder using the matrix approximation (MAP) method [6], [7] by minimizing the Euclidean distance between the hybrid precoder and the fully digital precoder obtained from (ii); (vi) "Con-MAP", which designs the hybrid precoder based on the fully digital precoder obtained from (iii) by using the MAP method. Note that the FBL rate is evaluated for all the considered schemes except (iii).

Fig. 3 illustrates the WSR of the considered fully digital and hybrid precoding designs with continuous phase shifters versus the transmit power budget P. We observe that "FD-IBL" provides a performance upper bound for all considered schemes. This is expected as it employs the IBL and a fully digital precoding architecture. Meanwhile, the fully digital precoding scheme "FD-FBL" achieves the best performance in the FBL regime. Compared with "FD-FBL", the performance of "FD-Conventional" degrades significantly in the low SNR regime, as the impact of blocklength is ignored and the users' QoS constraints (6d) may be violated. This result

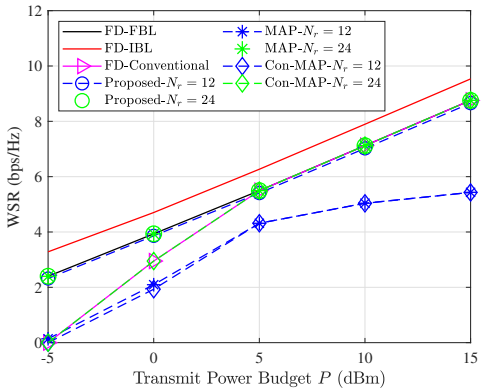

Figure 3. WSR vs. transmit power budget P for continuous phase shifters with different number of RF chains ($N_t = 64$, $K = 12$, and $N = 100$).

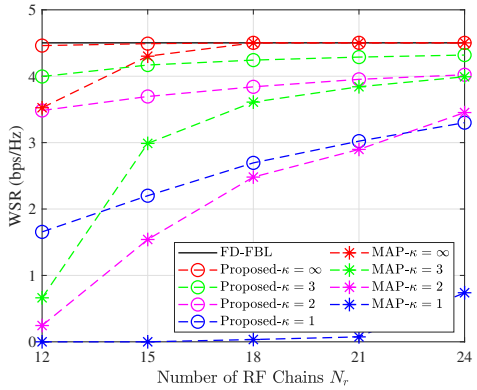

Figure 4. WSR vs. number of RF chains for phase shifters with different resolutions ($N_t = 64$, $K = 12$, $N = 100$, and $P = 2$ dBm).

highlights the importance of tailored precoding design for FBL applications, particularly in the low SNR regime. In contrast, the performance gap between "FD-Conventional" and "FD-FBL" vanishes in the high SNR regime. This is because the users' SINR are sufficiently large such that the channel dispersion $V_k(\gamma_k) \approx 1$ in (5) and its penalty on the data rate in the objective function of problem $\mathbb{P}_1$ becomes negligible. However, "FD-IBL", "FD-FBL", and "FD-Conventional" with the fully digitial precoding architecture would incur high hardware costs for mmWave massive MIMO. When considering hybrid precoding architecture with low hardware costs, Fig. 3 shows that our proposed algorithm always outperforms "MAP" and "Con-MAP" for different number of RF chains. Interestingly, when $N_r = 2K$ RF chains are available, both our proposed hybrid precoding design and "MAP" achieve the same performance as "FD–FBL", while "Con-MAP" and "FD-Conventional" achieve similar performance, which is consistent with the results in [9].

Fig. 4 shows the WSR of the fully digital precoding scheme "FD-FBL" and hybrid precoding methods "Proposed" and "MAP" versus the number of RF chains for phase shifters with different quantization resolutions. We observe that our proposed algorithm always outperforms "MAP" for all considered parameter settings. Particularly, pronounced performance gains are achieved even when the number of

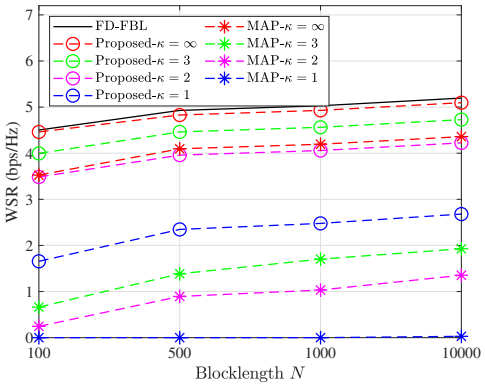

Figure 5. WSR vs. blocklength N for phase shifters with different resolutions ($N_t = 64$, $K = 12$, $N_r = 12$, and $P = 2$ dBm).

RF chains are small or the resolution of phase shifters is low. This is because the "MAP" method is heuristic and lacks theoretical performance guarantees in solving problem $\mathbb{P}_1$, failing to effectively tackling the nonconvex objective functions and constraints of problem $\mathbb{P}_1$. As a result, it leads to poor performance, especially when the level of quantization reduces, since the degree of nonconvexity in $\mathbb{P}_1$ increases. This result highlights the advantages of adopting our proposed hybrid precoding design in practical FBL systems. Moreover, the "MAP" solutions would violate the QoS constraints with a high probability, which is undesirable for FBL applications. Additionally, we observe from Fig. 4 that even when the BS is equipped with $N_r = K$ RF chains, which corresponds to the minimum number of RF chains required for supporting K-user communications, our proposed hybrid precoding adopting discrete phase shifters with several bits (e.g., 3-bit) resolution already achieves a performance that closely match the fully digital precoding counterpart.

In Fig. 5, we further evaluate the WSR of the fully digital precoding scheme "FD-FBL" and hybrid precoding designs "Proposed" and "MAP" versus the blocklength N for phase shifters with different resolutions. We observe that the WSRs of "FD-FBL", "Proposed", and "MAP" increase monotonically with the blocklength N. Meanwhile, our hybrid precoding design always outperforms "MAP", especially for short block-length N. For example, our proposed hybrid precoding scheme achieves performance gains of 3 to 14 times over the "MAP" when the BS adopts the discrete phase shifters of 2-bit resolution. For the discrete phase shifters of 1-bit resolution, the "MAP" scheme can even hardly satisfy the constraints and lead to poor performance. This result fully demonstrates the advantages of our hybrid precoding scheme in the FBL regime.

C. MMF Hybrid Precoding

We further evaluate the minimum rate achieved by the following schemes: (i) "Proposed", namely the proposed Algorithm 3; (ii) "FD-FBL", which maximizes the minimum FBL rate for fully digital precoding subject to the power and QoS constraints (6b) and (6d) using the solutions proposed in [29], [30]; (iii) "FD-IBL", which maximizes the minimum rate specified by the Shannon capacity for fully digital precoding under the power and QoS constraints (6b) and (6d); (iv) "FD-

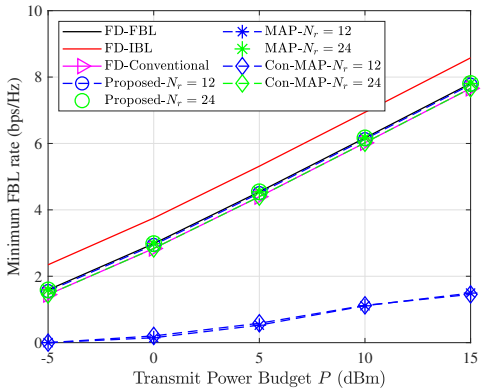

Figure 6. Minimum FBL rate vs. transmit power budget P for continuous phase shifters with different number of RF chains ($N_t = 64$, $K = 12$, and $N = 100$).

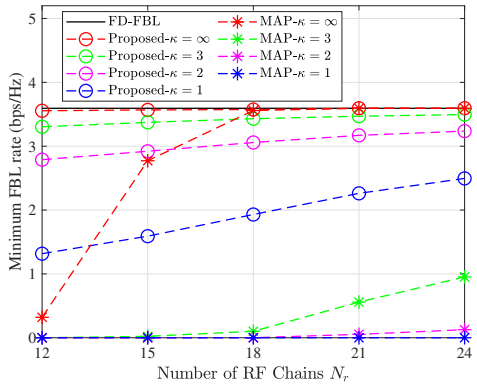

Figure 7. Minimum FBL rate vs. number of RF chains for phase shifters with different resolutions ($N_t = 64$, $K = 12$, $N = 100$, and $P = 2$ dBm).

IBL", which evaluates the FBL rate of the fully digital precoder obtained from (iii); (v) "MAP", which obtains the hybrid precoder using the MAP method [6], [7] based on the fully digital precoder from (ii); and (vi) "Con-MAP", which obtains the hybrid precoder using the MAP method based on the fully digital precoder from (iii).

Fig. 6 presents the minimum rate of the considered fully digital and hybrid precoding schemes with continuous phase shifters versus the transmit power budget P. As expected, the fully digital precoders "FD-IBL" and "FD-FBL" achieve the best IBL and FBL performance, respectively. However, unlike the WSR case, a non-negligible performance gap between "FD-FBL" and "FD-Conventional" always exists, since the second term in the FBL rate (5) cannot be ignored in the objective function of the MMF problem. We observe that our proposed hybrid precoder always outperforms the heuristic MAP-type hybrid precoding designs, as the latter often violates the users' QoS requirements. In contrast, our proposed solution is always feasible, which demonstrates the necessity of considering our solution in practical FBL systems. Similarly, when $N_r = 2K$, both our proposed hybrid precoding algorithm and "MAP" achieve similar performance as "FD–FBL", while "Con-MAP" and "FD-Conventional" perform close to each other.

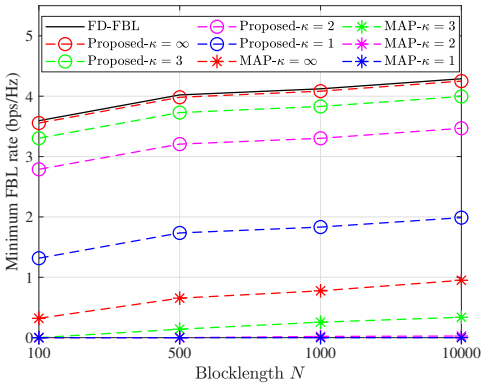

Figure 8. Minimum FBL rate vs. blocklength N for phase shifters with different resolutions ($N_t = 64$, $K = 12$, $N_r = 12$, and $P = 2$ dBm).

In addition, in Fig. 7, we evaluate the minimum FBL rate of the fully digital precoding scheme "FD-FBL" and hybrid precoding methods 'Proposed" and "MAP" versus the number of RF chains for phase shifters with different resolutions. It can be observed that our proposed hybrid precoding design always performs better than "MAP". For example, when the BS employs $N_r = K$ RF chains and the ideal continuous phase shifters, our proposed hybrid precoding scheme achieves approximately 11 times higher performance than that of the "MAP" benchmark. Additionally, under low-resolution phase shifters, such as 1-bit or 2-bit quantization, or when the number of RF chains is small, the heuristic "MAP" design fails to meet QoS requirements and may easily get trapped into bad and infeasible points, particularly for 1-bit resolution phase shifters. Therefore, adopting our hybrid precoder is crucial for latency-sensitive applications. Furthermore, as illustrated in Fig. 7, our proposed optimization algorithm enables hybrid precoding with just a few bits (e.g., 3-bit) quantization to closely approach the performance of fully digital precoding, even when the BS is equipped with $N_r = K$ RF chains, i.e., the minimum number of RF chains required to support multi-user communication.

Finally, Fig. 8 shows the minimum FBL rate of the fully digital precoding scheme "FD-FBL" and hybrid precoding designs "Proposed" and "MAP" versus the blocklength N for phase shifters with different resolutions. We observe that our hybrid precoding solution always achieves the best performance. Meanwhile, for the heuristic "MAP" scheme of low-resolution, e.g., 1-bit and 2-bits, phase shifters, increasing the blocklength N can hardly improve the performance. This again demonstrates the superiority of our method.

VI. Conclusions

In this paper, we explored hybrid precoding designs for massive MIMO systems in the FBL regime under the WSR and MMF criteria. Considering the users' minimum rate requirements, maximum transmit power budget at the BS, and various implementations of phase shifters, we formulated hybrid precoding optimization problems to maximize the WSR and the minimum users' rate. In the formulated problems, the digital and analog precoders are coupled in both the nonconvex objective functions and nonconvex constraints, rendering

their solutions challenging. To address these issues, we first proposed novel problem transformation and decomposition methods to reformulate the original complex problems into specific forms, whose AL problems exhibit decomposable structures and can be solved in a BCD manner. Then we proposed two efficient PDD algorithms to solve the WSR and the MMF hybrid precoding optimization problems, respectively. Our proposed BCD-type solutions are applicable to both continuous and discrete phase shifters. Simulation results demonstrated that our proposed hybrid precoding schemes outperform several considered benchmarks. Furthermore, the results showed that hybrid precoding with several bits (e.g., 3-bit) quantization phase shifters can approach the performance of the fully digital precoding scheme.

APPENDIX A
PROOF OF THEOREM 1

Problem (34) is a quadratically constrained quadratic programming (QCQP) with only one constraint (QCQP-1), and also satisfies the Slater condition. Therefore, strong duality holds for (34) [45], whose optimal solution can be obtained via solving its dual problem given as

$$\max_{\tau\geq 0}\ \Psi(\tau)=\min_{\mathbf{q}_k}\ \mathcal{L}_\tau \tag{66}$$

Here, $\mathcal{L}_\tau$ is the Lagrangian of problem (34), i.e.,

$$\begin{aligned}\mathcal{L}_\tau &= \omega_k c_k^{(m)}\sum\nolimits_{i=1}^{K}|q_{k,i}|^2-\omega_k\sum\nolimits_{i=1}^{K}\Re\big\{(b_{k,i}^{(m)})^\star q_{k,i}\big\},\\ &+\frac{1}{2\rho^{(t)}}\sum_{i=1}^{K}|q_{k,i}-\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}+\rho^{(t)}z_{k,i}^{(t)}|^2+\tau\Omega_k(\mathbf{q}_k)\\ &=\sum_{i=1}^{K}\Re\Big\{\Big(z_{k,i}^{(t)}-\frac{1}{\rho^{(t)}}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-\omega_k b_{k,i}^{(m)}\Big)^\star q_{k,i}\Big\}\\ &+\sum\nolimits_{i\neq k}\big(\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2+\tau\bar{\gamma}_k\big)\,|q_{k,i}|^2\\ &+\big(\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2-\tau\big)\,|q_{k,k}|^2\\ &+\frac{1}{2\rho^{(t)}}\sum\nolimits_{i=1}^{K}|\rho^{(t)}z_{k,i}^{(t)}-\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}|^2+\tau\bar{\gamma}_k\sigma_k^2,\end{aligned} \tag{67}$$

and $\tau\geq 0$ is the dual variable for constraint (28b). Note that $\Psi(\tau)=-\infty$ if $\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2-\tau<0$, we require $\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2\geq\tau$. Additionally, the optimal solution of problem (34) must satisfy the following KKT conditions:

$$\begin{aligned}\frac{\partial\mathcal{L}_\tau}{\partial q_{k,i}} =&2\big(\omega_k c_k^{(m)}+\frac{1}{2\rho^{(t)}}+\varsigma_{\mathrm{W},k}\tau\big)q_{k,i}-\omega_k b_{k,i}^{(m)}\\ &-(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}+z_{k,i}^{(t)}=0,\end{aligned} \tag{68}$$

$$\tau\geq 0, \tag{69}$$

$$\tau\Omega_k(\mathbf{q}_k)=0, \tag{70}$$

$$\Omega_k(\mathbf{q}_k)\leq 0, \tag{71}$$

where $\varsigma_{\mathrm{W},k}=-1$ if $i=k$ and $\varsigma_{\mathrm{W},k}=\bar{\gamma}_k$ otherwise.

Now assume $\omega_k b_{k,k}^{(m)}+(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_k^{(t,n+1)}=z_{k,k}^{(t)}$. As $q_{k,k}\geq\bar{\gamma}_k\sigma_k^2>0$ in constraint (28b), it follows from condition (68) that the optimal dual variable $\tau^*=\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2$. Thus, we have

$$q_{k,i}^*=\frac{\omega_k b_{k,i}^{(m)}+\frac{1}{\rho^{(t)}}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-z_{k,i}^{(t)}}{\Big(2\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}\Big)\,(1+\bar{\gamma}_k)},i\neq k. \tag{72}$$

According to the complementary slackness condition in (70), we have

$$\Omega(\mathbf{q}_k^*)=\bar{\gamma}_k\left(\sum\nolimits_{i\neq k}\left|q_{k,i}^*\right|^2+\sigma_k^2\right)-\left|q_{k,k}^*\right|^2=0. \tag{73}$$

Note that the phase of $q_{k,k}$ cannot be uniquely determined by (73), since the constraint and the objective function only depend on the magnitude of $q_{k,k}$. Thus, the optimal solution $q_{k,k}^*$ is given by

$$q_{k,k}^*=\sqrt{\bar{\gamma}_k}\left(\sum\nolimits_{i\neq k}|q_{k,i}^*|^2+\sigma_k^2\right)^{-1/2}\mathrm{e}^{\jmath\theta_{\mathrm{W},k}}, \tag{74}$$

where $\theta_{\mathrm{W},k}\in[0,2\pi]$ is arbitrary.

Next, assume $\omega_k b_{k,k}^{(m)}+(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_k^{(t,n+1)}\neq z_{k,k}^{(t)}$. Then, the optimal solution (36) can be obtained from (68). Substituting (36) into $\Omega_k(\mathbf{q}_k^*)$, we have

$$\begin{aligned}&\Upsilon(\tau)=\Omega_k(\mathbf{q}_k^*)\\ =&\bar{\gamma}_k\frac{\sum_{i\neq k}\left|\omega_k b_{k,i}^{(m)}+(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-z_{k,i}^{(t)}\right|^2}{\Big(2\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}+2\tau\bar{\gamma}_k\Big)^2}\\ &+\bar{\gamma}_k\sigma_k^2-\frac{\left|\omega_k b_{k,k}^{(m)}+(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_k^{(t,n+1)}-z_{k,k}^{(t)}\right|^2}{(2\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}-2\tau)^2}.\end{aligned} \tag{75}$$

From the complementary slackness condition (70), the optimal dual variable τ^* satisfies

$$\tau^*\Upsilon(\tau^*)=0. \tag{76}$$

Note that $\Upsilon(\tau)$ is monotonically decreasing w.r.t. $0\leq\tau<\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2$ and $\Upsilon\left(\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2\right)=-\infty$. Therefore, we have $\tau^*=0$ when $\Upsilon(0)<0$; otherwise, there must exist a unique root $0<\tau^*<\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2$ satisfying $\Upsilon(\tau^*)=0$. Define

$$\widehat{a}=\bar{\gamma}_k\sum_{i\neq k}\left|\omega_k b_{k,i}^{(m)}+\frac{1}{\rho^{(t)}}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_i^{(t,n+1)}-z_{k,i}^{(t)}\right|^2, \tag{77}$$

$$\widehat{b}=\left|\omega_k b_{k,k}^{(m)}+(\rho^{(t)})^{-1}\mathbf{h}_k^H\mathbf{F}^{(t,n+1)}\mathbf{w}_k^{(t,n+1)}-z_{k,k}^{(t)}\right|^2, \tag{78}$$

$$\widehat{c}=\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2. \tag{79}$$

Then solving $\Upsilon(\tau^*)=0$ is equivalent to solving the following quartic equation

$$\begin{aligned}&\tilde{\Upsilon}(\tau)=(4\bar{\gamma}_k^3\sigma_k^2)\tau^4+\left(8\widehat{c}\bar{\gamma}_k^2\sigma_k^2-8\widehat{c}\bar{\gamma}_k^3\sigma_k^2\right)\tau^3\\ &+\left(\widehat{a}-\widehat{b}\bar{\gamma}_k^2+4\widehat{c}^2\bar{\gamma}_k^3\sigma_k^2+4\bar{\gamma}_k\sigma_k^2\widehat{c}^2-16\widehat{c}^2\bar{\gamma}_k^2\sigma_k^2\right)\tau^2\\ &-\left(2\widehat{a}\widehat{c}\bar{\gamma}_k-8\widehat{c}^3\bar{\gamma}_k^2\sigma_k^2+8\widehat{c}^3\bar{\gamma}_k\sigma_k^2+2\widehat{b}\widehat{c}\bar{\gamma}_k\right)\tau\\ &+\widehat{a}\widehat{c}^2+4\bar{\gamma}_k\sigma_k^2\widehat{c}^4-\widehat{b}\widehat{c}^2=0,\end{aligned} \tag{80}$$

whose solution can be obtained in closed form and the existence of a unique real root τ^* satisfying $0<\tau^*<\omega_k c_k^{(m)}+(\rho^{(t)})^{-1}/2$ is guaranteed.

REFERENCES

[1] J. Wang, X. Zhang, X. Shi, and J. Song, "Higher spectral efficiency for mmWave MIMO: Enabling techniques and precoder designs," *IEEE Commun. Mag.*, vol. 59, no. 4, pp. 116–122, Apr. 2021.

[2] E. Bjornson, L. Van Der Perre, S. Buzzi, and E. G. Larsson, "Massive MIMO in sub-6 GHz and mmWave: Physical, practical, and use-case differences," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 100–108, Apr. 2019.

[3] E. Shi, J. Zhang, H. Du, B. Ai *et al.*, "RIS-aided cell-free massive MIMO

systems for 6G: Fundamentals, system design, and applications," *Proc. IEEE*, vol. 112, no. 4, pp. 331–364, Jun. 2024.

[4] J. Zhang, M. Matthaiou, and H. Yang, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, Jun. 2020.

[5] S. He, Y. Zhang, J. Wang, J. Zhang *et al.*, "A survey of millimeter-wave communication: Physical-layer technology specifications and enabling transmission technologies," *Proc. IEEE*, vol. 109, no. 10, pp. 1666–1705, Oct. 2021.

[6] 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.

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

[8] X. Yu, J.-C. 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.

[9] 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.

[10] C. Feng, W. Shen, J. An, and L. Hanzo, "Weighted sum rate maximization of the mmWave cell-free MIMO downlink relying on hybrid precoding," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2547–2560, Apr. 2022.

[11] J. Zhang, Y. Huang, J. Wang, and L. Yang, "Hybrid precoding for wideband millimeter-wave systems with finite resolution phase shifters," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11 285–11 290, Nov. 2018.

[12] J. Zhang, Y. Huang, T. Yu, J. Wang, and M. Xiao, "Hybrid precoding for multi-subarray millimeter-wave communication systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 440–443, Jun. 2018.

[13] J. Zhang, Y. Huang, J. Wang, X. You, and C. Masouros, "Intelligent interactive beam training for millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2034–2048, Mar. 2021.

[14] S. He, Y. Wu, J. Ren, Y. Huang *et al.*, "Hybrid precoder design for cache-enabled millimeter-wave radio access networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1707–1722, Mar. 2019.

[15] L. You, X. Qiang, K.-X. Li, C. G. Tsinos *et al.*, "Hybrid analog/digital precoding for downlink massive MIMO LEO satellite communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 5962–5976, Aug. 2022.

[16] X. Xue, Y. Wang, L. Yang, J. Shi *et al.*, "Energy-efficient hybrid precoding for massive MIMO mmWave systems with a fully-adaptive-connected structure," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3521–3535, Jun. 2020.

[17] L. Wen, H. Qian, M. Li, X. Luo *et al.*, "QoS-guaranteed hybrid beamforming design for multi-user systems with finite-resolution phase shifters," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 4, pp. 1678–1691, Dec. 2023.

[18] S. Malla and G. Abreu, "Transmit power minimization in multi-user millimeter wave systems," in *Proc. IEEE Int. Symp. Wireless Commun. Syst. (ISWCS)*, Poznan, Poland, Sep. 2016, pp. 409–413.

[19] X. He and J. Wang, "QCQP with extra constant modulus constraints: Theory and application to SINR constrained mmWave hybrid beamforming," *IEEE Trans. Signal Process.*, vol. 70, pp. 5237–5250, 2022.

[20] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[21] Z. Yuan, F. Liu, Q. Guo, X. Yuan *et al.*, "Blind grant-free random access with message-passing-based matrix factorization in mmWave MIMO mMTC," *IEEE Internet Things J.*, vol. 11, no. 3, pp. 4815–4825, Feb. 2024.

[22] M. H. Mazaheri, S. Ameli, A. Abedi, and O. Abari, "A millimeter wave network for billions of things," in *Proc. ACM Special Interest Group Data Commun.*, Beijing China, Aug. 2019, pp. 174–186.

[23] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni *et al.*, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, 3rd Quart., 2019.

[24] K. S. Kim, D. K. Kim, C.-B. Chae, S. Choi *et al.*, "Ultrareliable and low-latency communication techniques for tactile internet services," *Proc. IEEE*, vol. 107, no. 2, pp. 376–393, Feb. 2019.

[25] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park *et al.*, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131 796–131 813, 2020.

[26] A. A. Nasir, H. D. Tuan, H. H. Nguyen, M. Debbah *et al.*, "Resource allocation and beamforming design in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1321–1335, Feb. 2021.

[27] E. Shi, J. Zhang, J. Zhang, D. W. K. Ng *et al.*, "Decentralized coordinated precoding design in cell-free massive MIMO systems for URLLC," *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2638–2642, Feb. 2023.

[28] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink MISO OFDMA-URLLC systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7184–7200, Nov. 2020.

[29] S. He, Z. An, J. Zhu, J. Zhang *et al.*, "Beamforming design for multiuser URLLC with finite blocklength transmission," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8096–8109, Dec. 2021.

[30] X. Zhang, L. Xiang, J. Wang, and X. Gao, "Massive MIMO multicasting with finite blocklength," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 15 018–15 034, 2024.

[31] Y. Wang, V. W. S. Wong, and J. Wang, "Flexible rate-splitting multiple access with finite blocklength," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1398–1412, May 2023.

[32] Y. Xu, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access with finite blocklength for short-packet and low-latency downlink communications," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 12 333–12 337, Nov. 2022.

[33] M. Soleymani, I. Santamaria, E. A. Jorswieck, and B. Clerckx, "Optimization of rate-splitting multiple access in beyond diagonal RIS-assisted URLLC systems," *IEEE Trans. Wireless Comm.*, vol. 23, no. 5, May 2024.

[34] M. Soleymani, I. Santamaria, E. Jorswieck, R. Schober *et al.*, "Optimization of the downlink spectral- and energy-efficiency of RIS-aided multi-user URLLC MIMO systems," 2024, arXiv:2402.16434.

[35] M. Soleymani, I. Santamaria, and E. A. Jorswieck, "Spectral and energy efficiency maximization of MISO STAR-RIS-assisted URLLC systems," *IEEE Access*, vol. 11, pp. 70 833–70 852, Jul. 2023.

[36] S. Liu, Z. Sheng, P. Zhu, D. Wang *et al.*, "Hybrid precoding with low-resolution PSs for URLLC users in cell-free mmWave MIMO systems," in *IEEE Proc. Int. Conf. on Wireless Commun. Signal Process (WCSP)*, Hangzhou, China, Nov. 2023, pp. 1168–1172.

[37] H. Lee and Y.-C. Ko, "Physical layer enhancements for ultra-reliable low-latency communications in 5G new radio systems," *IEEE Commun. Standards Mag.*, vol. 5, no. 4, pp. 112–122, Dec. 2021.

[38] 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.

[39] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[40] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, Jun. 2016.

[41] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *Proc. IEEE Inform. Theory Workshop (ITW)*, Cairo, Egypt, Jan. 2010, pp. 1–5.

[42] D. Wipf and B. Rao, "Sparse bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[43] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, Jan. 2013.

[44] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[45] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[46] K. M. Attiah, F. Sohrabi, and W. Yu, "Deep learning for channel sensing and hybrid precoding in TDD massive MIMO OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10 839–10 853, Dec. 2022.

[47] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch *et al.*, "Overview of millimeter wave communications for fifth-generation (5G) wireless network-with a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.