Energy-Efficient Hybrid Symbol-Level Precoding via Phase Shifter Selection in mmWave MU-MIMO Systems

Alireza Haqiqatnejad[1], Farbod Kayhan[1,2], and Björn Ottersten[1]

[1]Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg
Email: {alireza.haqiqatnejad,bjorn.ottersten}@uni.lu
[2]Institute for Communication Systems, Department of Electrical and Electronic Engineering, University of Surrey
Email: f.kayhan@surrey.ac.uk

Abstract—We address the symbol-level precoding design problem for the downlink of a multiuser millimeter wave (mmWave) multiple-input multiple-output wireless system. We consider a hybrid analog-digital architecture with phase shifter selection where a small-sized baseband precoder is followed by two successive networks of analog on-off switches and variable phase shifters according to a fully-connected structure. We jointly optimize the digital baseband precoder and the states of the switching network on a symbol-level basis, i.e., by exploiting both the channel state information (CSI) and the instantaneous data symbols, while the phase-shifting network is designed only based on the CSI. Our approach to this joint optimization is to minimize the Euclidean distance between the optimal fully-digital and the hybrid symbol-level precoders. It is shown via numerical results that using the proposed approach, up to 50 percent of the phase shifters can be switched off on average, allowing for reductions in the power consumption of the phase-shifting network. Adopting appropriate power consumption models for the analog precoder, our energy efficiency analysis further shows that this power reduction can substantially improve the energy efficiency of the hybrid precoding compared to the fully-digital and the state-of-the-art schemes.

Index Terms—Energy efficiency, hybrid symbol-level precoding, multiuser mmWave massive MIMO, phase shifter selection.

I. Introduction

Millimeter wave (mmWave) communication has been widely accepted as a prime technology for the emerging outdoor/indoor wireless communication deployments, enabling multi-gigabit-per-second (Gbps) data rates thanks to the enormously available spectrum resources within 30-300 GHz frequency band [1], [2]. Communication in the mmWave band, however, suffers from an order-of-magnitude increase in the free-space path loss, higher shadow fading and more severe penetration losses compared to the legacy lower-frequency systems [3]. On the other hand, the shorter wavelength of mmWave signals makes it possible to pack a larger number of antenna elements in the same physical dimension, allowing for large-scale spatial multiplexing and highly directional beamforming. Employing large-scale antenna arrays, which is commonly known as massive multiple-input multiple-output (MIMO), further provides considerable beamforming gain to combat severe propagation losses at mmWave frequencies [4], and thus is essential to achieve high quality communication links in mmWave systems.

In traditional MIMO systems, the convention is to perform baseband precoding fully in the digital domain, which enables modification of both the amplitudes and phases of complex signals. This fully-digital signal processing, however, requires one dedicated radio frequency (RF) chain per antenna element, which is impractical to be implemented in systems with large antenna arrays due to the prohibitive cost and high power consumption of mixed-signal components, especially when operating at mmWave frequencies [5]. Hybrid analog-digital precoding is a cost-effective alternative to enable both multi-stream transmission and large beamforming gains via splitting the signal processing operation between the digital and analog domains. In hybrid architectures, a small-sized digital precoder is followed by a high-dimensional analog precoder which is usually implemented using RF phase shifters and/or switches [6]. Such a setup allows for using fewer RF chains, scaling with the number of data streams rather than the number of antennas.

For large-scale multiuser mmWave systems, design of block-level hybrid schemes where the precoding solution solely relies on the channel state information (CSI), has been extensively addressed. However, symbol-level approaches to hybrid precoding is not yet well studied. In the latter design approach, the data information is further exploited in optimizing the precoder such that the inter-user interference constructively contributes to the received signal by each user. Symbol-level hybrid precoding design under mmWave hardware constraints has been addressed in some recent work [7]–[9].

Hybrid architectures mostly employ phase shifters, switches or even a combination of both where the phase-shifting network is controlled by a preceding network of switches [6]. Employing a combination of phase-shifting and switching networks in the analog precoder comes with a two-fold advantage. On the one hand, it can provide additional degrees-of-freedom (brought by the switching network) when designing the analog precoder, and on the other hand, it allows for potential power-savings through switching some of the phase shifters off. Hybrid implementations with the so-called phase shifter selection, where a two-state on-off switch precedes each phase shifter, have been studied for conventional block-level precoding. For example, in [10], it has been shown that significant power reductions are possible without sacrificing the spectral efficiency even with up to 50% of the phase shifters being turned off. To the best of authors' knowledge, such an approach has not been investigated

so far for hybrid symbol-level precoding.

In this paper, we consider a hybrid analog-digital architecture with phase shifter selection for the symbol-level precoder where the analog precoder is implemented using a network of variable phase shifters preceded by an on-off switching network of the same dimension according to a fully-connected structure. As for the analog precoder, the phase states of the phase-shifting network are designed only based on the instantaneous CSI, i.e., they remain unchanged within the duration of one channel coherence block. However, in our design, the on-off states of the switches as well as the baseband digital precoder are jointly optimized on a symbol-level basis. Our design approach is then to minimize the ℓ_2-norm distance between the hybrid scheme and the optimal fully-digital symbol-level precoder. The phase shifter selection approach allows for power-savings in the RF precoder by switching off some of the phase shifters according to the optimized states of the switches. By adopting appropriate power consumption models, we will show through simulation results that this power reduction can lead to noticeable improvements in the energy efficiency of precoding.

Notations: We use bold-faced uppercase and lowercase letters to represent matrices and vectors, respectively. For matrices and vectors, $\|\cdot\|$ stands for the spectral norm and the ℓ_2 norm, respectively, and $[\cdot]^H$ and $[\cdot]^T$ denote conjugate transpose and transpose. For vectors, $\preceq$ and $\succeq$ denote elementwise inequality. Operators $\mathrm{diag}(\cdot)$ and $\mathrm{blkdiag}(\cdot)$ represent diagonal and block-diagonal matrices, and $\mathrm{vec}(\cdot)$ denotes vectorization operation. We use $\mathbf{I}$ and $\mathbf{1}$ to represent, respectively, the identity matrix and the all-ones vector of appropriate dimension. Operators $\otimes$ and $\circ$ stand for the Kronecker product and the Hadamard product, respectively. The probability function and the statistical expectation are denoted by $\mathbb{P}\{\cdot\}$ and $\mathbb{E}\{\cdot\}$.

II. SYSTEM AND CHANNEL MODEL

We consider a narrowband single-cell mmWave multiuser MIMO system. The base station (BS), which is equipped with a large-scale antenna array of N_t elements and a (typically) much smaller number of transmit RF chains, denoted by N_l, simultaneously communicates independent data streams to N_u single-antenna users, each supporting single-stream transmission. The maximum number of transmitted data streams is limited by the number of available RF chains at the BS, which leads to the assumption $N_\mathrm{u} \leq N_\mathrm{l} < N_\mathrm{t}$. We assume a hybrid digital-analog architecture where the digital baseband precoder is followed by the RF chains and an analog RF precoder, as shown in Fig. 1.

We consider a (non-linear) symbol-level baseband precoder that calculates the digital outputs specifically for every set of input symbols. Accordingly, the discrete-time $N_\mathrm{u} \times 1$ complex modulated symbol vector $\mathbf{s} = [s_1, s_2, ..., s_{N_\mathrm{u}}]^T$, where $\mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \mathbf{I}_{N_\mathrm{u}}$, is preprocessed in the digital domain using the symbol-level precoder, resulting in the output baseband signal $\mathbf{u}_{\mathrm{BB}} \in \mathbb{C}^{N_\mathrm{l} \times 1}$. In contrast to linear precoding schemes, the nonlinear-precoded signal $\mathbf{u}_{\mathrm{BB}}$ is directly designed and thus may not be explicitly decomposable as a linear combination of the users' precoding vectors.
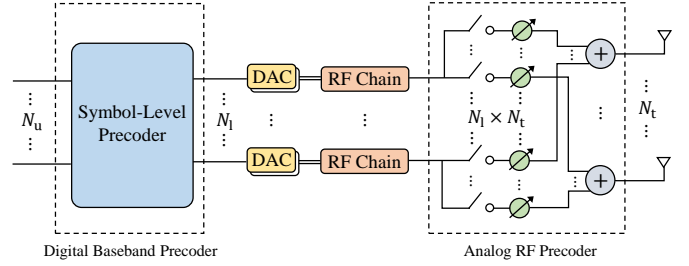

Fig. 1. The considered hybrid symbol-level precoding architecture with fully-connected switching and phase-shifting networks.

We assume that the analog precoder is implemented following a fully-connected architecture where two successive $N_\mathrm{t} \times N_\mathrm{l}$ networks of switches and phase shifters map N_l digital outputs to N_t analog signals feeding the antennas [6], [11]. Let $\mathbf{F} \in \mathbb{C}^{N_\mathrm{t} \times N_\mathrm{l}}$ and $\mathbf{T} \in \mathbb{B}$ represent the phase-shifting network and the on/off states of the switching network, respectively, where $\mathbb{B} \triangleq \{\mathbf{X} \in \{0,1\}^{N_\mathrm{t} \times N_\mathrm{l}} : \mathbf{X}\mathbf{1}_{N_\mathrm{l} \times 1} \succeq \mathbf{1}_{N_\mathrm{t} \times 1}, \mathbf{X}^T\mathbf{1}_{N_\mathrm{t} \times 1} \succeq \mathbf{1}_{N_\mathrm{l} \times 1}\}$. Then, the entire RF precoder can be represented by $\mathbf{F} \circ \mathbf{T}$. Note that the set $\mathbb{B}$ is defined such that the selection matrix $\mathbf{T}$ has no all-zero row and column, where the former case corresponds to an antenna selection scheme and the latter case excludes an RF chain from the transmitter's analog circuitry, but neither is the focus of this paper. Under the described system model, the vector collecting the baseband received signals for all N_u users is given by

$$\mathbf{y} = \sqrt{\rho}\,\mathbf{H}\left(\mathbf{F} \circ \mathbf{T}\right)\mathbf{u}_{\mathrm{BB}} + \mathbf{z}, \tag{1}$$

where $\mathbf{y} \in \mathbb{C}^{N_\mathrm{u} \times 1}$ is the received signal vector, ρ is the instantaneous transmit power, $\mathbf{H} \in \mathbb{C}^{N_\mathrm{u} \times N_\mathrm{t}}$ represents the mmWave multiuser channel, and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma})$ is a circularly symmetric complex Gaussian noise vector with $\mathbf{\Sigma} \triangleq \mathrm{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_{N_\mathrm{u}}^2)$ where σ_j^2 denotes the noise variance at the receiver of the jth user. The total instantaneous transmit power is constrained by ρ through enforcing $\|(\mathbf{F} \circ \mathbf{T})\mathbf{u}_{\mathrm{BB}}\|^2 = 1$. It is further assumed that the BS has perfect knowledge of the instantaneous channel matrix $\mathbf{H}$; see, e.g., [12], [13] for efficient mmWave channel estimation techniques. At the receiver side, we assume that each user performs optimal single-user decoding of the received signal via, e.g., maximum likelihood (ML) detector.

The mmWave propagation environment is known to feature limited multipath components. To capture this sparse scattering nature, the narrowband clustered channel modeling based on the Saleh-Valenzuela model is commonly used [14]. Under this model, the channel vector corresponding to a single user is a summation over the contributions of N_c scattering clusters, with each cluster contributing N_p propagation paths between the BS and the user. Accordingly, the narrowband mmWave channel vector for the jth user can be expressed as

$$\mathbf{h}_j^H = \sqrt{\frac{N_\mathrm{t}}{N_\mathrm{c}N_\mathrm{p}}} \sum_{i=1}^{N_\mathrm{c}} \sum_{l=1}^{N_\mathrm{p}} \alpha_{j,i,l}\,\mathbf{a}^H(\phi_{j,i,l}, \theta_{j,i,l}), \tag{2}$$

where $\mathbf{h}_j \in \mathbb{C}^{N_\mathrm{t} \times 1}$ such that $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{N_\mathrm{u}}]^H$. For the

lth path in the ith scattering cluster seen by the jth user, $\alpha_{j,i,l} \sim \mathcal{CN}(0, \sigma_\alpha^2)$ denotes the circularly symmetric complex Gaussian gain of the path (i.e., the small-scale fading component), $\phi_{j,i,l}$ and $\theta_{j,i,l}$ are respectively the azimuth and elevation angles of departure (AoD), and $\mathbf{a}(\phi_{j,i,l}, \theta_{j,i,l})$ represents the normalized transmit array response vector at azimuth and elevation angles $\phi_{j,i,l}$ and $\theta_{j,i,l}$. The array response vector further depends on the array geometry. For uniform linear arrays (ULA), where the antenna elements are linearly and equally spaced, the array response vector is independent of the elevation angles $\theta_{j,i,l}$ and follows the Vandermonde structure given by

$$\mathbf{a}(\phi_{j,i,l}) = \frac{1}{\sqrt{N_\mathrm{t}}} \left[1, e^{\mathrm{j}\frac{2\pi}{\lambda} d \sin(\phi_{j,i,l})}, ..., e^{\mathrm{j}(N_\mathrm{t}-1)\frac{2\pi}{\lambda} d \sin(\phi_{j,i,l})} \right]^T, \tag{3}$$

where λ and d respectively denote the signal wavelength and the inter-element antenna spacing. The path gain variances $\alpha_{j,i,l}$ and the normalization constant $\sqrt{N_\mathrm{t}/(N_\mathrm{c} N_\mathrm{p})}$ are set such that $\mathbb{E}\{\|\mathbf{H}\|_\mathrm{F}\} = \sqrt{N_\mathrm{t} N_\mathrm{u}}$, where $\|\cdot\|_\mathrm{F}$ denotes the Frobenius norm.

III. Hybrid Symbol-Level Precoder Design

We start off by designing the analog phase-shifting network. The matrix $\mathbf{F}$ representing the phase shifters' angles is usually considered to be dependent only on the aggregate channel $\mathbf{H}$. Here, similar to [8], we adopt an analog design based on the singular value decomposition of the channel matrix $\mathbf{H}$ given by $\mathbf{H} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H$, where $\boldsymbol{\Sigma}$ is an $N_\mathrm{u} \times N_\mathrm{t}$ rectangular diagonal matrix with the singular values on the diagonal in a descending order, and $\mathbf{U}$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{N_\mathrm{t}}]$ are respectively $N_\mathrm{u} \times N_\mathrm{u}$ and $N_\mathrm{t} \times N_\mathrm{t}$ unitary matrices with the columns representing the left and the right singular vectors. Accordingly, the angles of the phase shifters are aligned to those of the first N_l right singular vectors of $\mathbf{H}$, with an element-wise normalization due to constant modulus constraint of the phase shifters. Denoting by $f_{n,k}$ the nth element of the kth column of $\mathbf{F}$, we have

$$f_{n,k} = \frac{1}{\sqrt{N_\mathrm{t}}} e^{\mathrm{j}\varphi_{n,k}}, \quad \forall n \in \{1, 2, ..., N_\mathrm{t}\}, \forall k \in \{1, 2, ..., N_\mathrm{l}\}, \tag{4}$$

where $\varphi_{n,k}$ denotes the phase of the nth entry in $\mathbf{v}_k$. Although infinite-precision phase shifters are required for an accurate implementation of this approach, in practice, the use of finite-precision phase shifters is preferred due to practical constraints of variable phase shifters. In a more realistic implementation, phase shifters with discrete phase states are used and each entry of $\mathbf{F}$ is quantized up to (typically) low bits of precision, i.e.,

$$\hat{\varphi}_{n,k} = \frac{2\hat{m}\pi}{2^{b_\mathrm{PS}}}, \quad \hat{m} = \underset{m \in \{0,1,...,2^{b_\mathrm{PS}}-1\}}{\mathrm{argmin}} \left| \varphi_{n,k} - \frac{2m\pi}{2^{b_\mathrm{PS}}} \right|, \tag{5}$$

with b_PS denoting the number of phase shifter's resolution bits. Note, however, that our design process is independent of the precision of the entries of $\mathbf{F}$.

Accordingly, for given symbol vector $\mathbf{s}$, the channel matrix $\mathbf{H}$ and the phase-shifting network matrix $\mathbf{F}$, our design objective is to jointly and instantaneously (i.e., on a symbol-level basis) optimize the digitally precoded signal $\mathbf{u}$ as well as the states of the switching network, represented by $\mathbf{T}$. To this end, we first

study the symbol-level precoding solution corresponding to a fully-digital architecture which will be of essential use in our subsequent elaboration of the hybrid precoding approach.

A. Optimal Fully-Digital Precoder

Let us consider a fully-digital architecture for the symbol-level precoder where each antenna element is driven by a dedicated RF chain, i.e., $N_\mathrm{u} \leq N_\mathrm{l} = N_\mathrm{t}$. We adopt a power-constrained max-min SNR design criterion subject to constructive interference (CI) constraint for each user. In order to exploit the CI at the receiver side, we adopt the so-called distance-preserving CI regions, as presented in [15], in which any two points belonging to two distinct CI regions are distanced by at least the distance between the corresponding constellation points. In general, the distance-preserving CI region associated with any constellation symbol is a subset of its decision region, and therefore, the entire noise-free received signal can be treated as desired signal component. Therefore, the convention of using SINR as the performance measure is no longer applicable and one needs to alternatively consider the users' received SNRs; see, e.g., [16]. Denoting by $\mathbf{u}_\mathrm{FD} \in \mathbb{C}^{N_\mathrm{t} \times 1}$ and $\mathbb{D}_j$, respectively, the fully-digital precoded signal and the distance-preserving CI region for symbol s_j, the corresponding optimization problem can be expressed as

$$\max_{\mathbf{u}_\mathrm{FD}} \quad \min_j \quad |\mathbf{h}_j^H \mathbf{u}_\mathrm{FD}|^2 / \sigma_j^2$$
$$\text{s.t.} \quad \sqrt{\rho}\, \mathbf{h}_j^H \mathbf{u}_\mathrm{FD} \in \mathbb{D}_j, \ j = 1, ..., N_\mathrm{u}, \quad \|\mathbf{u}_\mathrm{FD}\|^2 \leq 1. \tag{6}$$

In formulating (6), we attempt to exploit the data information (i.e., the symbol vector $\mathbf{s}$) in order to instantaneously align the multiuser interference so that the noise-free signal received by each user locates within the distance-preserving CI region that corresponds to its target symbol. It has been shown in [17] that problem (6) can be expressed in a compact form as

$$\max_{\bar{\mathbf{u}}_\mathrm{FD}, \mathbf{d} \succeq \mathbf{0}} \quad \min(\mathbf{d})$$
$$\text{s.t.} \quad \sqrt{\rho}\, \bar{\mathbf{H}}\bar{\mathbf{u}}_\mathrm{FD} = \boldsymbol{\Sigma}\bar{\mathbf{s}} + \mathbf{A}^{-1}\mathbf{W}\mathbf{d}, \quad \|\bar{\mathbf{u}}_\mathrm{FD}\|^2 \leq 1, \tag{7}$$

where $\min(\cdot)$ denotes elementwise minimum. Furthermore, the following real-valued notations are used: $\bar{\mathbf{u}}_\mathrm{FD} \triangleq [\mathrm{Re}(\mathbf{u}_\mathrm{FD})^T, \mathrm{Im}(\mathbf{u}_\mathrm{FD})^T]^T$, $\bar{\mathbf{H}} \triangleq [\bar{\mathbf{H}}_1^T, \bar{\mathbf{H}}_2^T, ..., \bar{\mathbf{H}}_{N_\mathrm{u}}^T]^T$ with

$$\bar{\mathbf{H}}_j \triangleq \begin{bmatrix} \mathrm{Re}(\mathbf{h}_j^T) & -\mathrm{Im}(\mathbf{h}_j^T) \\ \mathrm{Im}(\mathbf{h}_j^T) & \mathrm{Re}(\mathbf{h}_j^T) \end{bmatrix}, \quad j = 1, 2, ..., N_\mathrm{u},$$

and $\mathbf{A} \triangleq \mathrm{blkdiag}(\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_{N_\mathrm{u}})$ with $\mathbf{A}_j = [\mathbf{a}_{j,1}, \mathbf{a}_{j,2}]^T$ denoting an invertible 2×2 matrix containing the normal vectors of the decision region boundaries associated with s_j; $\boldsymbol{\Sigma} \triangleq \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_{N_\mathrm{u}}) \otimes \mathbf{I}_2$; $\bar{\mathbf{s}} \triangleq [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{N_\mathrm{u}}]^T$ with $\mathbf{s}_j \triangleq [\mathrm{Re}(s_j), \mathrm{Im}(s_j)]^T$; and $\mathbf{W} \triangleq \mathrm{diag}(w_1, w_2, ..., w_{N_\mathrm{u}}) \otimes \mathbf{I}_2$ is a $2N_\mathrm{u} \times 2N_\mathrm{u}$ diagonal matrix with $w_j = 1$ if the intended symbol for the jth user is an outer constellation symbol, and $w_j = 0$ otherwise. Further, the $2N_\mathrm{u} \times 1$ non-negative slack vector $\mathbf{d}$ collects the orthogonal distances between the received symbols and the boundaries of their corresponding CI regions.

B. Hybrid Precoder with Phase Shifter Selection

We use the optimal transmit signal corresponding to the fully-digital formulation in order to design the hybrid symbol-level precoder. More specifically, denoting by $\mathbf{u}_{\mathrm{FD}}^\star$ the optimal solution to (7), we aim to find the digitally precoded signal $\mathbf{u}_{\mathrm{BB}}$ and the selection matrix $\mathbf{T}$ such that the output of the hybrid precoder, i.e., $(\mathbf{F} \circ \mathbf{T})\mathbf{u}_{\mathrm{BB}}$ is as close as possible to $\mathbf{u}_{\mathrm{FD}}^\star$. The corresponding optimization problem is therefore given by

$$\min_{\mathbf{u}_{\mathrm{BB}}, \mathbf{T} \in \mathbb{B}} \quad \left\| (\mathbf{F} \circ \mathbf{T})\mathbf{u}_{\mathrm{BB}} - \mathbf{u}_{\mathrm{FD}}^\star \right\|^2 \quad \text{s.t.} \quad \left\| (\mathbf{F} \circ \mathbf{T})\mathbf{u}_{\mathrm{BB}} \right\|^2 = 1. \tag{8}$$

To proceed, by denoting $\mathbf{g} \triangleq \mathrm{vec}(\mathbf{G})$ where $\mathbf{G} \triangleq 2\mathbf{T} - \mathbf{1}_{N_t \times N_l}$, and redefining the set $\mathbb{B}$, denoted by $\bar{\mathbb{B}}$, according to the latter denotation, we recast problem (8) as

$$\min_{\mathbf{u}_{\mathrm{BB}}, \mathbf{g}} \quad \left\| (\mathbf{u}_{\mathrm{BB}}^T \otimes \mathbf{I}_{N_t}) \mathrm{diag}(\mathrm{vec}(\mathbf{F})) \, \mathbf{g} + \mathbf{F}\mathbf{u}_{\mathrm{BB}} - 2 \, \mathbf{u}_{\mathrm{FD}}^\star \right\|^2$$
$$\text{s.t.} \quad \frac{1}{4} \| (\mathbf{u}_{\mathrm{BB}}^T \otimes \mathbf{I}_{N_t}) \mathrm{diag}(\mathrm{vec}(\mathbf{F})) \, \mathbf{g} + \mathbf{F}\mathbf{u}_{\mathrm{BB}} \|^2 = 1,$$
$$\mathbf{g} \in \left\{ \{-1, +1\}^{N_t N_l} \cap \bar{\mathbb{B}} \right\}. \tag{9}$$

Problem (9) belongs to the class of minimization of quadratic forms over binary vectors (i.e., the binary constraints on the elements of $\mathbf{g}$), which is known to be NP-hard in general [18]. To tackle this difficulty, inspired by [18], we use an equivalent implication of the binary constraints given in a biconvex form. Accordingly, let $\mathbf{e}$ be a real-valued slack vector of length $N_t N_l$. Provided that $-\mathbf{1} \preceq \mathbf{g} \preceq \mathbf{1}$ and $\mathbf{e}^T \mathbf{e} \leq N_t N_l$, the constraint $\mathbf{g}^T \mathbf{e} = N_t N_l$ implies that $\mathbf{g} \in \{-1, +1\}^{N_t N_l}$. Therefore, we can rewrite problem (9) in an equivalent form where all the optimization variables are taken from continuous domains, i.e.,

$$\min_{\mathbf{u}_{\mathrm{BB}}, -\mathbf{1} \preceq \mathbf{g} \preceq \mathbf{1}, \mathbf{e}} \quad \left\| (\mathbf{u}_{\mathrm{BB}}^T \otimes \mathbf{I}_{N_t}) \mathrm{diag}(\mathrm{vec}(\mathbf{F})) \, \mathbf{g} + \mathbf{F}\mathbf{u}_{\mathrm{BB}} - 2 \, \mathbf{u}_{\mathrm{FD}}^\star \right\|^2$$
$$\text{s.t.} \quad \frac{1}{4} \| (\mathbf{u}_{\mathrm{BB}}^T \otimes \mathbf{I}_{N_t}) \mathrm{diag}(\mathrm{vec}(\mathbf{F})) \, \mathbf{g} + \mathbf{F}\mathbf{u}_{\mathrm{BB}} \|^2 = 1,$$
$$\mathbf{g}^T \mathbf{e} = N_t N_l, \quad \mathbf{e}^T \mathbf{e} \leq N_t N_l, \quad \mathbf{g} \in \bar{\mathbb{B}}, \tag{10}$$

which is still a non-convex formulation due to the biconvex constraint $\mathbf{g}^T \mathbf{e} = N_t N_l$. This constraint can be handled by adding the difference $N_t N_l - \mathbf{g}^T \mathbf{e}$ multiplied by $\mu > 0$ as a penalizing term to the objective function. Thus, denoting the objective function of (10) by $g(\mathbf{u}_{\mathrm{BB}}, \mathbf{g})$, we can write

$$\min_{\mathbf{u}_{\mathrm{BB}}, -\mathbf{1} \preceq \mathbf{g} \preceq \mathbf{1}, \mathbf{e}} \quad g(\mathbf{u}_{\mathrm{BB}}, \mathbf{g}) + \mu \left(N_t N_l - \mathbf{g}^T \mathbf{e} \right)$$
$$\text{s.t.} \quad \frac{1}{4} \left\| (\mathbf{u}_{\mathrm{BB}}^T \otimes \mathbf{I}_{N_t}) \mathrm{diag}(\mathrm{vec}(\mathbf{F})) \, \mathbf{g} + \mathbf{F}\mathbf{u}_{\mathrm{BB}} \right\|^2 = 1,$$
$$\mathbf{e}^T \mathbf{e} \leq N_t N_l, \quad \mathbf{g} \in \bar{\mathbb{B}}. \tag{11}$$

Given $\mathbf{u}_{\mathrm{BB}}$, it is straightforward to show that function $g(\mathbf{u}_{\mathrm{BB}}, \mathbf{g})$ is a Lipschitz continuous convex function on $-\mathbf{1} \preceq \mathbf{g} \preceq \mathbf{1}$ with Lipschitz constant

$$L = 2\sqrt{N_t N_l} \left\| (\mathbf{u}_{\mathrm{BB}}^T \otimes \mathbf{I}_{N_t}) \mathrm{diag}(\mathrm{vec}(\mathbf{F})) \right\|^2$$
$$+ 2 \left\| \left((\mathbf{u}_{\mathrm{BB}}^T \otimes \mathbf{I}_{N_t}) \mathrm{diag}(\mathrm{vec}(\mathbf{F})) \right)^H \left(\mathbf{F}\mathbf{u}_{\mathrm{BB}} - 2 \, \mathbf{u}_{\mathrm{FD}}^\star \right) \right\|. \tag{12}$$

Algorithm 1: BCD algorithm solving (11)

> **input:** $\mathbf{F}, \mathbf{u}_{\mathrm{FD}}^\star$
> **output:** $\mathbf{u}_{\mathrm{BB}}, \mathbf{g}$
> **initialize:** $\mathbf{g}^{(0)} = \mathbf{e}^{(0)} = \mathbf{1}_{N_t N_l \times 1}, \mathbf{u}_{\mathrm{BB}}^{(0)} = \mathbf{0}_{N_l \times 1}, \mu^{(0)}, t = 0$
> **set:** $\vartheta > 1$
> 1 **do**
> 2 $\quad$ $t \leftarrow t + 1$
> 3 $\quad$ *compute $\mathbf{g}^{(t)}$ by solving* (13)
> 4 $\quad$ $\mathbf{e}^{(t)} \leftarrow \sqrt{N_t N_l} \, \mathbf{g}^{(t)} / \|\mathbf{g}^{(t)}\|$
> 5 $\quad$ *compute $\mathbf{u}_{\mathrm{BB}}^{(t)}$ using* (15)
> 6 $\quad$ *obtain $L^{(t)}$ from* (12)
> 7 $\quad$ $\mu^{(t)} \leftarrow \min\{2L^{(t)}, \vartheta \, \mu^{(t-1)}\}$
> 8 **until** *the terminating condition is met*;

As a result, the local and global minima of (11) are identical to those of the original problem (10) for $\mu \geq 2L$; see [18, Th. 1]. Note, further, that the objective function of the minimization problem (11), i.e., $g(\mathbf{u}_{\mathrm{BB}}, \mathbf{g}) + \mu \left(N_t N_l - \mathbf{g}^T \mathbf{e} \right)$ is a biconvex quadratic function in $\mathbf{g}$ and $\mathbf{e}$, i.e., fixing either $\mathbf{g}$ or $\mathbf{e}$ gives a convex function in the other variable. It follows that finding at least a locally optimal solution to problem (10) is equivalent to solving (11) via, e.g., a standard block coordinate descent (BCD) algorithm, with a coordinate block referring to either of the vector-valued variables $\mathbf{u}_{\mathrm{BB}}, \mathbf{g}$ or $\mathbf{e}$. The penalty multiplier μ can be cyclically updated until the threshold $2L$ is reached. In each cycle, the two main sub-problems are as follows. Given $\mathbf{e}$ and $\mathbf{u}_{\mathrm{BB}}$, the value of $\mathbf{g}$ can be computed by solving

$$\mathbf{g} = \operatorname*{argmin}_{-\mathbf{1} \preceq \mathbf{g} \preceq \mathbf{1}, \mathbf{g} \in \bar{\mathbb{B}}} g(\mathbf{u}_{\mathrm{BB}}, \mathbf{g}) + \mu \left(N_t N_l - \mathbf{g}^T \mathbf{e} \right), \tag{13}$$

which is a linearly-constrained quadratic program. On the other hand, given $\mathbf{g}$ and $\mathbf{e}$, the minimization (11) is equivalent to

$$\min_{\mathbf{u}_{\mathrm{BB}}} \quad g(\mathbf{u}_{\mathrm{BB}}, \mathbf{g})$$
$$\text{s.t.} \quad \frac{1}{4} \left\| (\mathbf{u}_{\mathrm{BB}}^T \otimes \mathbf{I}_{N_t}) \mathrm{diag}(\mathrm{vec}(\mathbf{F})) \, \mathbf{g} + \mathbf{F}\mathbf{u}_{\mathrm{BB}} \right\|^2 = 1. \tag{14}$$

Using the method of Lagrange multipliers, one can obtain the solution to (14), which is given by

$$\mathbf{u}_{\mathrm{BB}} = \frac{2 \left((\mathbf{F} \circ \mathbf{G}) + \mathbf{F} \right)^\dagger \mathbf{u}_{\mathrm{FD}}^\star}{\left\| \left((\mathbf{F} \circ \mathbf{G}) + \mathbf{F} \right) \left((\mathbf{F} \circ \mathbf{G}) + \mathbf{F} \right)^\dagger \mathbf{u}_{\mathrm{FD}}^\star \right\|}, \tag{15}$$

where $(\cdot)^\dagger$ stands for the matrix Moore-Penrose inverse. The pseudocode of the BCD algorithm to solve the design problem (11) is presented in Algorithm 1.

IV. ENERGY EFFICIENCY ANALYSIS

Hybrid implementations may differ from one another in their complexity and power consumption. For the sake of comparing different hybrid architectures and also to assess their efficiency versus the fully-digital counterpart, one needs to incorporate both performance and complexity/power consumption aspects into a single figure of merit. A well known choice is energy efficiency which can simply be expressed as the ratio between spectral efficiency and power consumption. Here, due to the assumption of finite-alphabet signaling, we measure the spectral

efficiency in bits per symbol. Thereby, the energy efficiency of the precoding scheme, in bits per Joule, is defined as the ratio between goodput and power consumption, i.e.,

$$\eta \triangleq \frac{R(1 - P_e)}{P}, \qquad (16)$$

where $P_e \triangleq 1 - (1/N_u) \sum_{j=1}^{N_u} P_{e,j}$ is the average symbol error probability across all N_u users with $P_{e,j}$ denoting the symbol error probability for the jth user, and the spectral efficiency R and the power consumption P are defined as follows.

A. Spectral Efficiency

Using an uncoded transmission scheme with finite-alphabet signaling, the communication rate towards the jth user can be evaluated, in terms of bits per symbol per unit bandwidth, through calculating the average mutual information between the target symbol s_j and the received signal y_j, i.e.,

$$I(s_j; y_j) = \mathbb{E}_{s_j, y_j, \mathbf{H}} \left\{ \log_2 \frac{\mathbb{P}_{y_j | s_j, \mathbf{H}}(y_j | s_j, \mathbf{H})}{\mathbb{P}_{y_j | \mathbf{H}}(y_j | \mathbf{H})} \right\}. \qquad (17)$$

Assuming transmission with Nyquist rate over a double-sided bandwidth of W Hz, the maximum allowable symbol rate is W symbols per second, which results in a bit rate of $W \times I(s_j; y_j)$ for user j. Putting this together for all N_u users, the sum-rate over the downlink channel is given by

$$R = W \sum_{j=1}^{N_u} I(s_j; y_j). \qquad (18)$$

Notice that no closed-form expressions exist for the conditional probability mass functions in (17). Alternatively, one can obtain empirical probability distributions over sufficiently many independent channel and symbol realizations to approximate the mutual information $I(s_j; y_j)$ for each user $j \in \{1, 2, ..., N_u\}$.

B. Power Consumption

We specifically tailor a power consumption model according to each precoding architecture, namely, fully-digital and hybrid (with/without phase shifter selection). The transmit RF front-end of a multi-antenna system is commonly composed of one baseband processor, several RF chains, each preceded by a pair of DACs (i.e., one DAC for each I/Q channel), and power amplifiers (PA). The use of analog components such as switches and/or phase shifters are limited to hybrid architectures.

Let P_{BB}, P_{PA}, P_{PS} and P_{SW} respectively denote the power consumption of the baseband processor, a single PA, a single phase shifter and a single analog switch. The fully-digital BS architecture requires $2N_t$ DACs, and N_t RF chains and PAs, and therefore its power consumption can be modeled as

$$P_{FD} = P_{BB} + N_t(2P_{DAC} + P_{RF} + P_{PA}), \qquad (19)$$

where P_{DAC} is the power consumption of binary-weighted current-steering type, which is given in [19] by

$$P_{DAC} = \frac{3}{2} \left(2^{b_{DAC}} - 1 \right) \times 10^{-5} + \frac{9}{2} b_{DAC} F_s \times 10^{-12}, \qquad (20)$$

with b_{DAC} and F_s respectively denoting the number of precision bits and the sampling frequency. On the other hand, the hybrid architecture with fully-connected phase-shifting network can be implemented using $2N_l$ DACs, N_l RF chains, N_t PAs, and $N_t N_l$ phase shifters. The resulting power dissipation is

$$P_H = P_{BB} + N_l(2P_{DAC} + P_{RF}) + N_t N_l P_{PS} + N_t P_{PA}. \qquad (21)$$

To calculate the power consumption of the hybrid architecture with phase shifter selection, we assume the associated RF processes are turned off while a phase shifter is deactivated, and further, the phase shifter has negligible static power dissipation. Denoting the average ratio of deactivated phase shifters by β, the power consumed by the entire phase-shifting network is equal to $(1 - \beta)N_t N_l P_{PS}$. According to Fig. 1, the phase shifter selection process is implemented through an $N_t \times N_l$ network of switches, and therefore the corresponding power consumption can be obtained as

$$\begin{aligned} P_{HPSS} = P_{BB} + N_l(2P_{DAC} + P_{RF}) \\ + N_t N_l \left((1 - \beta)P_{PS} + P_{SW} \right) + N_t P_{PA}. \end{aligned} \qquad (22)$$

Using the above power consumption models with appropriate parameter selection, we will compare the energy efficiencies of different fully-digital and hybrid architectures in Section V.

V. SIMULATION RESULTS

The simulation setup is as follows. We consider the hybrid analog-digital precoding architecture depicted in Fig. 1 for a downlink mmWave massive multiuser MIMO system, performing an uncoded transmission with QPSK signaling over a carrier frequency of 60 GHz with a bandwidth of 1 GHz. We assume unit noise variances at the receivers of all the users, i.e., $\sigma_j^2 = 1, \forall j = 1, 2, ..., N_u$. As described in Section II, we adopt a geometric model for the mmWave propagation environment with $N_c = 1$ clusters and $N_p = 12$ scatterers between the BS and each user. For all the paths, the variances of the complex gains are set to be $\sigma_\alpha^2 = 1$, and the azimuth angles of departure $\phi_{j,i,l}$ are drawn independently from a uniform distribution over $[0, 2\pi)$. To calculate power consumptions according to the model introduced in Section IV, we consider $P_{RF} = 40$ mW, $P_{PA} = 20$ mW, $P_{PS} = 30$ mW, $P_{BB} = P_{DAC}$ [11], $P_{SW} = 1$ mW [20], and $F_s = 1$ GHz. For the proposed cyclic algorithm, we set $\vartheta = 1.1$ to avoid overshooting, and initialize the penalty parameter as $\mu^{(0)} = 10^{-4}$ to have a reasonable starting point. We consider the optimal fully-digital symbol-level precoding (SLP) as our performance benchmark, and further, provide comparisons with the hybrid precoders in [8] and [9]. The method in [8] performs a symbol-based optimization of the digital baseband precoder subject to one-bit DACs, and adopts a CSI-only design for the phase-shifting network. On the other hand, the hybrid scheme in [9] jointly optimizes both the digital baseband precoder and the phase-shifting network on a symbol-level basis. Based on the adopted hybrid architecture and the precoder design approach, we refer to the methods in [8] and [9] and the proposed scheme in this paper respectively as "Hybrid BB SLP", "Hybrid BB+PS SLP" and "Hybrid BB+SW SLP".
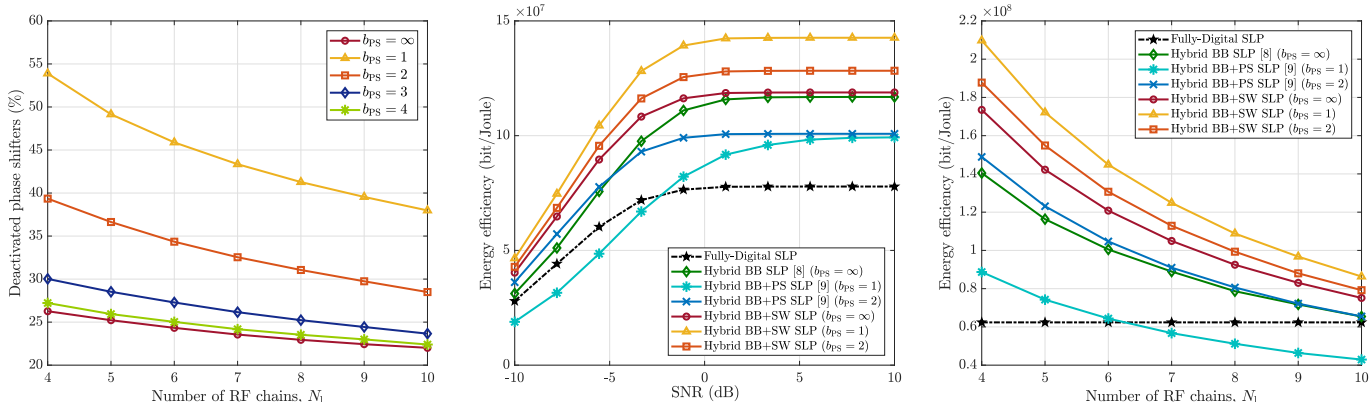
Fig. 2. Performance comparison of different precoding schemes: (a) Average percentage of deactivated phase shifters versus N_l with $N_t = 64$ and $N_u = 4$; (b) Energy efficiency versus SNR with $(N_t, N_l, N_u) = (64, 8, 4)$; (c) Energy efficiency as a function of N_l with $N_t = 64$ and $N_u = 4$ at SNR $= -5$ dB.

For those architectures with high-resolution DACs, we assume $b_{DAC} = 12$ to calculate the power consumption using (20). For a scenario with $N_t = 64$ and $N_u = 4$, the percentages of deactivated phase shifters for different values of b_{PS} are shown versus the number of RF chains in Fig. 2 (a). It follows form the results that by employing low-resolution phase shifters, a higher percentage of the phase shifters can be switched off, and hence more power-savings are possible. In particular, in the case with $N_l = 4$, up to 55% of the phase shifters with $b_{PS} = 1$ can be turned off, which can be roughly translated to a power reduction of $\alpha N_t N_l P_{PS} \approx 4200$ mW in the phase-shifting network. It can further be inferred from Fig. 2 (a) that the percentage of deactivated phase shifters decreases with increasing N_l.

We compare in Fig. 2 (b) the energy efficiencies of different fully-digital and hybrid symbol-level precoders as a function of the transmit SNR for a system with $(N_t, N_l, N_u) = (64, 8, 4)$. As it can be seen, almost all of the the hybrid approaches are more energy-efficient than the fully-digital SLP. The proposed Hybrid BB+SW SLP with phase shifter selection outperforms the other schemes either with infinite or low resolution phase shifters. The most energy-efficient scheme is Hybrid BB+SW SLP with $b_{PS} = 1$, using which energy efficiency gains of up to 75 Mbps/Joule can be achieved with respect to the fully-digital SLP. In contrast to the Hybrid BB+SW SLP scheme, employing phase shifters with lower precision bits improves the energy-efficiency of Hybrid BB+SW SLP. This is a because more phase shifter can be switched off using low-precision phase shifters, which results in larger reductions of the power consumption. We are also interested in the variations of energy efficiency as a function of N_l, which is plotted in Fig. 2 (c) at a fixed SNR of -5 dB. A common trend among all the hybrid schemes is that their energy efficiency become lower as N_l increases. We can also see that the differences in energy efficiency of Hybrid BB+SW SLP schemes resulted from using phase shifters with different precision bits b_{PS} reduces with increasing N_l.

VI. CONCLUSIONS

We proposed a hybrid precoding scheme for multiuser mmWave downlink systems, where successive fully-connected networks of RF switches and phase shifters are assumed to enable phase shifter selection. We adopted a CSI-only design approach for the phase-shifting network with either infinite or finite precision phase shifters, whereas the digital precoder and the switching network are optimized in a symbol-level basis. We formulated our design problem so as to minimize the Euclidean distance between the hybrid symbol-level precoder and its optimal fully-digital counterpart, where a max-min SINR design criterion subject to constructive interference constraints was adopted. To evaluate the energy efficiency, we adopted a power consumption model for the RF precoder. Our results indicated that using the phase shifter selection approach, up to half of the phase shifters can be turned off, allowing for power reductions of multi-Watts in the analog circuitry. This power reduction translates to significant improvements in the energy efficiency of the precoding, where simulation results showed that the proposed approach outperforms both the fully-digital and the existing hybrid symbol-level precoders.

REFERENCES

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

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

[3] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5g wireless communication networks," *IEEE communications magazine*, vol. 52, no. 2, pp. 122–130, 2014.

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

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

[6] S. S. Ioushua and Y. C. Eldar, "A family of hybrid analog–digital beamforming methods for massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3243–3257, June 2019.

[7] G. Hegde, C. Masouros, and M. Pesavento, "Analog beamformer design for interference exploitation based hybrid beamforming," in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2018, pp. 109–113.

[8] A. Li, C. Masouros, and F. Liu, "Hybrid analog-digital precoding for interference exploitation," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 812–816.

[9] R. Liu, H. Li, and M. Li, "Symbol-level hybrid precoding in mmWave multiuser MISO systems," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1636–1639, Sep. 2019.

[10] S. Payami, M. Ghoraishi, and M. Dianati, "Hybrid beamforming for large antenna arrays with phase shifter selection," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 11, pp. 7258–7271, Nov. 2016.

[11] R. Méndez-Rial, C. Rusu, N. G. Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.

[12] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics in Signal Process.*, vol. 8, no. 5, pp. 831–846, 2014.

[13] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, 2010.

[14] T. S. Rappaport, R. W. Heath Jr, R. C. Daniels, and J. N. Murdock, *Millimeter wave wireless communications*. Pearson Education, 2015.

[15] A. Haqiqatnejad, F. Kayhan, and B. Ottersten, "Constructive interference for generic constellations," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 586–590, Apr. 2018.

[16] ——, "Symbol-level precoding design based on distance preserving constructive interference regions," *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5817–5832, Nov. 2018.

[17] A. Haqiqatnejad, F. Kayhan, and B. Ottersten, "An approximate solution for symbol-level multiuser precoding using support recovery," in *2019 IEEE 20th Int. Workshop on Signal Process. Advances in Wirel. Commun. (SPAWC)*, Jul. 2019, pp. 1–5.

[18] G. Yuan and B. Ghanem, "Binary optimization via mathematical programming with equilibrium constraints," *arXiv preprint:1608.04425*, 2016.

[19] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wirel. Commun.*, vol. 4, no. 5, pp. 2349–2360, 2005.

[20] R. L. Schmid, P. Song, C. T. Coen, A. Ç. Ulusoy, and J. D. Cressler, "On the analysis and design of low-loss single-pole double-throw W-band switches utilizing saturated SiGe HBTs," *IEEE Trans. Microwave Theory and Techniques*, vol. 62, no. 11, pp. 2755–2767, 2014.