Lattice-Based mmWave Hybrid Beamforming

Shanxiang Lyu, Zheng Wang, Zhen Gao, Hongliang He, *Member, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*— Conventional hybrid precoding and combining based transceivers require a large number of high-resolution radio frequency (RF) phase shifters (PSs), which impose prohibitive hardware costs and power consumption. To address the above issue, both partially connected RF PSs and low-resolution PSs have been proposed. However, the performance limits of these low-cost designs have not been investigated theoretically. Furthermore, there is room for improvement in their spectral efficiency. To fill this knowledge gap, we derive the mean square error performance discrepancy between an optimal precoder/combiner and the hybrid analog-digital precoder/combiner under the constraint of 1-bit PSs relying on lattice theory. Then, by observing that this performance gap can be reduced by deactivating parts of the PSs whilst improving both the spectral and energy efficiency, we develop an adaptive RF PS connection network. To resolve the associated hybrid precoding and combining problems, we appropriately adapt Babai's algorithm from the lattice decoding literature. Our simulation results demonstrate the superiority of the proposed scheme both in terms of its spectral and energy efficiency.**

***Index Terms*— Lattices, massive MIMO, hybrid beamforming, Babai's algorithm.**

I. Introduction

TO MEET the escalating increase in data traffic predicted for the next generation networks, millimeter-wave (mmWave) multiple-input multiple-output (MIMO) antenna arrays constitute a promising solution. Hence they have drawn extensive attention both from academia and industry [1]–[4], since that they have several GHz unallocated bandwidth which is much wider than the 80 MHz bandwidth of the fifth generation (5G) networks. At such short wavelengths, packing a large number of antennas into a compact space also becomes possible. Hence flexible transmit precoding techniques can be utilized for compensating the high path loss of mmWave carrier frequencies [5]–[7].

To realize mmWave MIMOs in real-world systems, a huge obstacle is the excessive power consumption of large antenna arrays, since each antenna of a fully digital MIMO system requires a dedicated radio-frequency (RF) chain (e.g., digital-to-analog converters, mixers, etc.) [8]. While analog beamforming seems to be a potential alternative as it only employs a single RF chain to reduce the hardware complexity and power consumption, it fails to support multi-stream communications [9], [10]. To mitigate this drawback, hybrid precoding [11] has been put forward as a cost-effective transceiver solution, which utilizes a limited number of RF chains to connect the digital baseband precoder/combiner and the analog RF precoder/combiner. Hence numerous authors have investigated efficient analog and digital schemes conceived for hybrid precoding [11]–[18]. The pioneering technique of Ayach *et al.* [11] formulates the associated entropy maximization problem as a sparse signal reconstruction problem. Later in [12], an alternating minimization algorithm combined with manifold optimization and semi-definite relaxation was proposed by Yu *et al.* to attain a near-optimal performance. However, the phase shifters (PSs) of these schemes usually require infinite resolutions, and the number of PSs is large. These facts suggest that the schemes still suffer from relatively high hardware cost and energy consumption.

A pair of amendments have been proposed for further reducing the hardware cost and power consumption. I) The first is to employ either finite or low-resolution PSs [19]–[21]. Within this setting, the straightforward way of designing beamformers with finite resolution PSs is to design them assuming infinite resolution first and then to quantize the value of each PS to a finite set. However, this approach is inefficient for systems having very low resolution PSs and specific algorithms should be designed. Dedicated algorithms include the classic conjugate descent method of Chen [19], the rank-2 approximation based adaptive design of Wang *et al.* [20], and the improved orthogonal matching pursuit algorithm of Uwaechia *et al.* [21] that selects multiple column indices per iteration. Additionally, algorithms for multi-user settings was considered in [22]–[25]. II) The second is to replace the fully-connected network of RF PSs by a partially-connected one, where an RF chain only

Manuscript received October 11, 2020; revised January 19, 2021 and March 23, 2021; accepted April 19, 2021. Date of publication April 23, 2021; date of current version July 15, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 61902149, 61801216, 62071044, 62032009 and 61932010, the Natural Science Foundation of Guangdong Province under Grant 2020A1515010393, the Fundamental Research Funds for the Central Universities under Grants 21620438 and 21620350, the Major Program of Guangdong Basic and Applied Research under Grant 2019B030302008, the Natural Science Foundation of Jiangsu Province under Grant BK20180420, the Beijing Municipal Natural Science Foundation under Grant L182024, the Young Elite Scientists Sponsorship Program by CAST under No. YESS20180270, the Talent Innovation Project of BIT, the Engineering and Physical Sciences Research Council projects EP/P034284/1 and EP/P003990/1 (COALESCE) as well as the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028). The associate editor coordinating the review of this article and approving it for publication was C. Masouros. *(Corresponding authors: Hongliang He; Lajos Hanzo.)*

Shanxiang Lyu and Hongliang He are with the College of Cyber Security, Jinan University, Guangzhou 510632, China (e-mail: shanxianglyu@gmail.com; hehongliang@stu.xjtu.edu.cn).

Zheng Wang is with the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: z.wang@ieee.org).

Zhen Gao is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: gaozhen16@bit.edu.cn).

Lajos Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2021.3075248.

Digital Object Identifier 10.1109/TCOMM.2021.3075248

TABLE I

A BRIEF COMPARISON OF THE RELATED LITERATURE

Contents \ Literature		[11]–[15]	[33]	[19]–[21]	[9], [12], [26]–[30]	This work
Resolution of PSs	Infinite	✓	✓			
	Finite			✓	✓	✓
Connection of PSs	Full	✓	✓	✓		
	Not full				✓	✓
Theoretical bounds			✓			✓
Exploiting lattice decoding						✓

connects a subset of transmitter antennas (TAs), for which either switches or switches-inverters are deployed [9], [12], [26]–[30]. We refer motivated readers to [31] for a comprehensive study on using the switching mechanism. Designs based on a partially-connected architecture often aim for maximizing the energy efficiency [12], [26]–[28], [32], and recent references have also shown that this optimization target can be readily achieved by either using dynamic RF chain selection [9], or joint bit allocation [29], or alternatively low resolution digital-to-analog converters [30], [32], etc. It is however noteworthy that while the partially-connected network can reduce both the hardware cost and energy consumption significantly, it often trades a significant amount of spectral efficiency for energy efficiency. To sum up, we boldly and explicitly contrast our new contribution to the related literature in Table I.

Although a large body of literature exists in this area, several fundamental questions remain unanswered: (i) *Can low-resolution PS-based schemes be as good as fully digital precoding/combing?* A closely related result in [33] for infinite-resolution PSs stated that provided the number of RF chains is twice the total number of data streams, then the hybrid beamforming structure can indeed match the fully digital beamformer's performance. However, in a quest for answering this question, we more broadly aim for showing that there exists a non-negligible performance gap for a 1-bit-PS-based hybrid precoder. The precoding problem in this specific setting exhibits a lattice structure, where the associated lattices are discrete additive subgroups of $\mathbb{R}^n$. Lattices have been widely used both in coding theory for constructing near-capacity codes [34] and in communications to design integer-forcing based linear receivers for achieving the maximally attainable degrees-of-freedom [35]. (ii) *Can we further improve the spectral efficiency of low-resolution PSs based schemes?* By observing that the objective in low-resolution PS based hybrid transmit precoding (TPC) is reminiscent of the closest vector problem (CVP) of lattices, we construct a low-complexity lattice decoder for solving the CVP. To address the related issues, the novel contributions of this paper are as follows.

1) We derive the performance upper bound of 1-bit resolution PS-based hybrid precoding. The underlying cost function resembles a quantization problem whose quantization accuracy is controlled by the resolution of PSs. The averaged quantization error is non-vanishing for 1-bit resolution PSs, since it entails a quantization process within the lattice structure. In our proof, we exploit the random distributions of query points and the second moment of the fundamental region of lattices.
2) We propose to randomly deactivate some PSs for the sake of improving the achievable spectral efficiency. This idea is reminiscent of the one used in spatial modulation [36], [37] that incorporates a silent/dis-connection point 0 in the constellation designed. Based on this principle, the feasible range in a finite accuracy PS becomes the union of a cyclotomic set and 0. We demonstrate that both the spectral efficiency and energy efficiency are improve by this configuration.
3) In contrast to popular belief, we show that efficient low complexity algorithms exist for solving the combinatorial decoding problem of low resolution hybrid precoding architectures. In particular, both the analog and digital precoders exhibit near-orthogonal properties, hence Babai's low complexity lattice decoder [38] may be adapted. Simulation results confirm that the proposed method converges faster than the major benchmarks, and it improves the spectral efficiency.

The rest of this paper is organized as follows. Background of hybrid TPCs and the associated optimization problems are reviewed in Section II. The performance gap of 1-bit PSs w.r.t. their infinite resolution counterparts is subsequently studied in Section III. Based on the proposed adaptively-connected network of RF PSs, Babai's algorithm is introduced in Section IV. Section V presents our simulation results, followed by our conclusions in the last section.

Notation: Matrices and column vectors are denoted by uppercase and lowercase boldface letters. The (i, j)th element, ith row and jth column of a matrix are respectively denoted as $[\cdot]_{i,j}$, $[\cdot]_{i,:}$ and $[\cdot]_{:,j}$. $\mathbf{I}_n$ and $\mathbf{0}_n$ respectively denote the $n \times n$ identity matrix and $n \times 1$ zero vector. Operations $(\cdot)^\top$, $(\cdot)^H$, $(\cdot)^\dagger$ and $\det(\cdot)$ respectively denote matrix transposition, Hermitian transposition, pseudoinverse, and determinant. $\mathcal{Q}_S(x)$ denotes quantizing x to the nearest element in a set S. $\|\mathbf{x}\|$ denotes the Euclidean norm of vector $\mathbf{x}$, while $\|\mathbf{X}\|_F$ denotes the Frobenius norm of matrix $\mathbf{X}$. V_N refers to the volume of an N-dimensional unit ball.

II. PRELIMINARIES

In this section, we will present the system model of the mmWave MIMO system considered, and then review the classical alternating minimization-based technique of solving the hybrid precoding optimization problem.

A. System Model

We consider a point-to-point mmWave MIMO system using a hybrid TPC and receiver combiner using low-resolution

PSs. The base station (BS) is equipped with N_t TAs and N_{RF}^t RF chains for transmitting N_s data streams to a user relying on N_r receive antennas (RAs) and N_{RF}^r RF chains. To enable the simultaneous transmission of multiple data streams, we assume $N_s \geq \max(N_{RF}^t, N_{RF}^r)$. This configuration makes hybrid precoding particularly desirable. The signal vector $\mathbf{s} \in \mathbb{C}^{N_s}$ first goes through a base-band TPC matrix $\mathbf{F}_{BB} \in \mathbb{C}^{N_{RF}^t \times N_s}$, followed by N_{RF}^t RF chains, and an RF TPC $\mathbf{F}_{RF} \in \mathbb{C}^{N_t \times N_{RF}^t}$. In this paper, we assume the widely used Gaussian transmit signals with normalized signal power of $\mathbb{E}\left(\mathbf{s}\mathbf{s}^H\right) = \frac{1}{N_s}\mathbf{I}_{N_s}$. The transmitted signal vector after the hybrid TPC is then written as

$$\mathbf{x} = \mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s}. \tag{1}$$

The normalized transmit power constraint is given by $\left\|\mathbf{F}_{RF}\mathbf{F}_{BB}\right\|_F^2 = N_s$ [12]. Then, the signal vector $\mathbf{y} \in \mathbb{C}^{N_r}$ received by the user in a narrowband system is given by

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} + \mathbf{n}, \tag{2}$$

where $\mathbf{n} \in \mathbb{C}^{N_r}$ is the additive white Gaussian noise (AWGN) vector with its elements obeying the independent and identical distribution (i.i.d.) $\mathcal{CN}\left(0, \sigma_n^2\right)$, ρ denotes the averaged received power (alternatively, we can remove ρ in Eq. (2) by setting $\left\|\mathbf{F}_{RF}\mathbf{F}_{BB}\right\|_F^2$ as the maximum allocated power controlled by ρ, as used in [9]), and $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ is the MIMO channel matrix such that $\mathbb{E}\left(\left\|\mathbf{H}\right\|_F^2\right) = N_t N_r$. Full channel state information (CSI) is assumed to be available at the transceiver.

Due to its high free-space path loss, the mmWave propagation environment is accurately characterized by a clustered channel model, where the channel matrix $\mathbf{H}$ is assumed to be a sum of the contributions of N_{cl} scattering clusters, each of which contribute N_{ray} propagation paths. Then the discrete-time narrowband channel matrix $\mathbf{H}$ [11], [12] is given by

$$\begin{aligned}\mathbf{H} = \gamma \sum_{i=1}^{N_{cl}} \sum_{l=1}^{N_{ray}} \alpha_{il} \Lambda_r\left(\phi_{il}^r, \theta_{il}^r\right) \Lambda_t\left(\phi_{il}^t, \theta_{il}^t\right) \\ \times \mathbf{a}_r\left(\phi_{il}^r, \theta_{il}^r\right) \mathbf{a}_t^H\left(\phi_{il}^t, \theta_{il}^t\right),\end{aligned} \tag{3}$$

where $\gamma = \sqrt{\frac{N_t N_r}{N_{cl} N_{ray}}}$ denotes a normalization factor, $\alpha_{il} \in \mathbb{C}$ denotes the complex gain of the lth ray and ith cluster, while $\phi_{il}^r/\theta_{il}^r$ and $\phi_{il}^t/\theta_{il}^t$ are its azimuth/elevation angles of arrival and departure (AoAs, AoDs), respectively. The functions $\Lambda_r\left(\phi_{il}^r, \theta_{il}^r\right)$ and $\Lambda_t\left(\phi_{il}^t, \theta_{il}^t\right)$ denote the transmit and receive antenna array gain at a specific AoA and AoD, respectively. Lastly, $\mathbf{a}_r\left(\phi_{il}^r, \theta_{il}^r\right)$ and $\mathbf{a}_t^H\left(\phi_{il}^t, \theta_{il}^t\right)$ are the normalized receive and transmit antenna array response vectors. We adopt the uniform square planar array structure [12] for the array response vectors.

B. Problem Formulation

The RF combiner $\mathbf{W}_{RF} \in \mathbb{C}^{N_r \times N_{RF}^r}$ and the digital baseband combiner $\mathbf{W}_{BB} \in \mathbb{C}^{N_{RF}^r \times N_s}$ process the received signal vector $\mathbf{y}$, yielding:

$$\tilde{\mathbf{y}} = \sqrt{\rho}\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s} + \mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{n}. \tag{4}$$

Algorithm 1 Alternating Minimization Technique for Solving the Hybrid Precoding Problem

```
Input: $\mathbf{F}_{\mathrm{opt}}$.
Output: $\mathbf{F}_{\mathrm{RF}}$,$\mathbf{F}_{\mathrm{BB}}$.
1 Initialize a random RF precoder $\mathbf{F}_{\mathrm{RF}}$ ;
2 while A stopping criterion has not been met do
3     $\mathbf{F}_{\mathrm{BB}} \leftarrow \mathbf{F}_{\mathrm{RF}}^{\dagger}\mathbf{F}_{\mathrm{opt}}$;
4     Obtain $\mathbf{F}_{\mathrm{RF}}$ by solving problem $\mathcal{P}3$;
5 $\mathbf{F}_{\mathrm{BB}} \leftarrow \frac{\sqrt{N_s}}{\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F}\mathbf{F}_{\mathrm{BB}}$.
```

The spectral efficiency based on this equation is therefore written as

$$\begin{aligned}&\mathcal{I}\left(\mathbf{F}_{RF}, \mathbf{F}_{BB}, \mathbf{W}_{RF}, \mathbf{W}_{BB}\right) \\ &\quad = \log_2 \det(\mathbf{I}_{N_s} + \frac{\rho}{\sigma_s^2 N_s}\left(\mathbf{W}_{RF}\mathbf{W}_{BB}\right)^{\dagger}\mathbf{H} \\ &\qquad \times \mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{F}_{BB}^H\mathbf{F}_{RF}^H\mathbf{H}^H\mathbf{W}_{RF}\mathbf{W}_{BB}).\end{aligned} \tag{5}$$

For the sake of maximizing the spectral efficiency, we arrive at the constrained optimization problem of:

$$\mathcal{P}1: \max_{\mathbf{F}_{RF}, \mathbf{F}_{BB}, \mathbf{W}_{RF}, \mathbf{W}_{BB}} \mathcal{I}\left(\mathbf{F}_{RF}, \mathbf{F}_{BB}, \mathbf{W}_{RF}, \mathbf{W}_{BB}\right), \tag{6a}$$

$$s.t. \quad \left\|\mathbf{F}_{RF}\mathbf{F}_{BB}\right\|_F^2 = N_s, \tag{6b}$$

$$\left[\mathbf{F}_{RF}\right]_{m,n} \in \mathcal{F}, \quad \forall m, n, \tag{6c}$$

$$\left[\mathbf{W}_{RF}\right]_{m,n} \in \mathcal{F}, \quad \forall m, n. \tag{6d}$$

The constraint $\mathcal{F}$ is defined by the low-resolution PSs. Given B bits of resolution and an amplitude of $1/\sqrt{N_t}$, the feasible set of phases is

$$\mathcal{F} = 1/\sqrt{N_t}\left\{e^{j\times 0}, e^{j\times\left(2\pi/2^B\right)}, \ldots, e^{j\times\left(2\pi/2^B\right)\times\left(2^B-1\right)}\right\}.$$

The non-convex constraints in $\mathcal{P}1$ imply that no simple closed-form solutions exist.

To simplify the formulation, the popular solution [11], [12] is to separate problem $\mathcal{P}1$ into a hybrid TPC & receiver combiner problem. They have similar mathematical formulations, except that there exists an extra power constraint in the former. For this reason, we focus our attention on the TPC design in the remaining part of this paper, noting that the algorithms proposed in this work can be readily applied for the combiner. Following [11], [12], to design a hybrid TPC we have to solve

$$\begin{aligned}\mathcal{P}2: \min_{\mathbf{F}_{RF}, \mathbf{F}_{BB}} & \left\|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\right\|_F^2, \\ s.t. \;& (6b),\ (6c),\end{aligned} \tag{7}$$

where $\mathbf{F}_{opt}$ is the optimal fully digital TPC that consists of the first N_s columns of the $\mathbf{V}$ matrix in the singular value decomposition $\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^H$, and the diagonal elements of Σ are sorted in descending order.

The popular alternating minimization technique [12], [19], whose convergence has been treated in [39], [40], can decouple the joint optimization into the individual subproblems of $\mathbf{F}_{RF}$ and $\mathbf{F}_{BB}$. The alternating minimization procedure is listed in Algorithm 1. As seen in Step 3 of Algorithm 1, the problem of minimizing $\left\|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\right\|_F^2$ with given $\mathbf{F}_{RF}$ can be solved in a closed-form, while the solution relies on the pseudo-inverse operation. On the other hand, with the known $\mathbf{F}_{BB}$, Step 4 of Algorithm 1 entails a combinatorial search that admits certain statistical constraint:

$$\mathcal{P}3: \min_{\mathbf{F}_{RF}} \|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2, s.t.\ [\mathbf{F}_{RF}]_{m,n} \in \mathcal{F}, \forall m, n. \tag{8}$$

Here the term (6b) is temporarily omitted, since the energy constraint can be met by normalizing $\mathbf{F}_{BB}$, while retaining the fitness of the solutions [12]. However, the combinatorial nature of $\mathcal{P}3$ constitutes a major hindrance when relying on alternating minimization, because an exhaustive search would require the enumeration of $|\mathcal{F}|^{N_t \times N_{RF}^t}$ points, which is computationally prohibitive.

Remark 1: Based on the alternating minimization technique, the optimization of receiver combiner is similar:

$$\begin{aligned} &\min_{\mathbf{W}_{RF}} \|\mathbf{W}_{opt} - \mathbf{W}_{RF}\mathbf{W}_{BB}\|_F^2, \\ &s.t.\ [\mathbf{W}_{RF}]_{m,n} \in \mathcal{F}, \quad \forall m, n. \end{aligned}$$

Here $\mathbf{W}_{opt}$ consists of the first N_s columns of the $\mathbf{U}$ matrix in the above mentioned singular value decomposition.

III. The 1-Bit Resolution Scheme

In this section, we analyze the lower bound of the target function defined in $\mathcal{P}3$, in which a binary constraint is considered for $\mathcal{F}$, and there are in total $N_t \times N_{RF}^t$ PSs for the case of the fully-connected network. Upon harnessing lattice-theoretic techniques [41], [42], the analysis relies on bounding the expected second moment of a convex body with fixed volume.

By inspecting $\mathcal{P}3$ from the perspective of quantization, increasing the number of quantization level in the PSs would certainly reduce the average value of the target function. This can be achieved either by using PSs with higher resolution, or, alternatively, by adding switches to each PS. The latter approach seems more promising, since having a large number of inactive PSs reduces the power consumption.

A. Lattices and Convex Bodies

Let $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]$ consist of n linearly independent vectors. A full-rank lattice of dimension n is defined by

$$\Lambda = \mathcal{L}(\mathbf{B}) = \left\{ \sum_{i=1}^{n} \mathbf{b}_i x_i \mid x_i \in \mathbb{Z} \right\},$$

where $\mathbf{B}$ represents a so-called lattice basis.

In this work, we are mostly concerned with the fundamental parallelotope and the Voronoi region of lattice lattices. Readers are referred to the text [43] for other basic properties. Any lattice basis describes a fundamental parallelotope according to

$$\mathcal{P}(\mathbf{B}) = \left\{ \mathbf{x} \mid \mathbf{x} = \sum_{l=1}^{m} \theta_l \mathbf{b}_l, 0 \leq \theta_l < 1 \right\},$$

where $\mathcal{P}(\mathbf{B})$ is referred to a fundamental region, i.e., a region that completely covers the span of $\mathbf{B}$ when shifted to all points of the lattice. Clearly, different bases lead to different fundamental parallelotopes.

Another important fundamental region is the Voronoi region, defined as the set of points in $\mathbb{R}^n$ that are closer to the origin than to any other lattice point, which is formulated as:

$$\mathcal{V}(\Lambda) = \{\mathbf{x} \mid \|\mathbf{x}\| \leq \|\mathbf{x} - \mathbf{y}\| \,\forall \mathbf{y} \in \Lambda\}. \tag{9}$$

The Voronoi regions can be associated to all other lattice points by a simple translation of $\mathcal{V}(\Lambda)$. In contrast to the fundamental parallelotope $\mathcal{P}(\mathbf{B})$, the Voronoi region $\mathcal{V}(\Lambda)$ is a lattice invariant, i.e., it is independent of the specific choice of a lattice basis.

The closest vector problem is, given a vector $\mathbf{y} \in \mathbb{R}^n$ and a lattice Λ, to find a vector $\mathbf{v} \in \Lambda$ such that

$$\|\mathbf{y} - \mathbf{v}\|^2 \leq \|\mathbf{y} - \mathbf{w}\|^2, \ \forall \mathbf{w} \in \Lambda.$$

B. The Lower Bound of the Target Function

Due to the fact that

$$\begin{aligned} &\|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2 && (10)\\ &\quad = \left\|\mathbf{F}_{opt}^\top - \mathbf{F}_{BB}^\top \mathbf{F}_{RF}^\top\right\|_F^2 && (11)\\ &\quad = \sum_{n=1}^{N_t} \left\| \left[\mathbf{F}_{opt}^\top\right]_{1:N_s,n} - \mathbf{F}_{BB}^\top \left[\mathbf{F}_{RF}^\top\right]_{1:N_s,n} \right\|^2, && (12) \end{aligned}$$

we rewrite $\mathcal{P}3$ as N_t independent optimization problems. The n-th instance is given by

$$\mathcal{P}4: \min_{\mathbf{x} \in \mathcal{F}^{N_{RF}^t}} \|\bar{\mathbf{y}} - \mathbf{B}\mathbf{x}\|^2, \tag{13}$$

in which

$$\mathbf{B} \triangleq \begin{bmatrix} \mathcal{R}(\mathbf{F}_{BB}^\top) \\ \mathcal{I}(\mathbf{F}_{BB}^\top) \end{bmatrix}, \quad \bar{\mathbf{y}} \triangleq \begin{bmatrix} \mathcal{R}(\left[\mathbf{F}_{opt}^\top\right]_{1:N_s,n}) \\ \mathcal{I}(\left[\mathbf{F}_{opt}^\top\right]_{1:N_s,n}) \end{bmatrix},$$

with $\mathcal{R}(\cdot), \mathcal{I}(\cdot)$ defining the real and imaginary parts of the input, respectively. The optimized $\mathbf{x}$ is used to construct $\left[\mathbf{F}_{RF}^\top\right]_{1:N_{RF}^t,n}$.

While we can expect $\mathbf{F}_{RF}\mathbf{F}_{BB}$ to approximate $\mathbf{F}_{opt}$ quite tightly with the aid of high-resolution PSs, it is still unknown whether the worst case configuration, i.e., PSs having 1-bit accuracy, can provide a good approximation of $\mathbf{F}_{opt}$. Here Proposition 1 shows that the performance gap of 1-bit PSs is non-vanishing.

Proposition 1: The expected value of $\min_{\mathbf{F}_{RF} \in \mathcal{F}^{N_t \times N_{RF}^t}} \|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2$ with the constraint $\mathcal{F} = 1/\sqrt{N_t}\{-1, 1\}$ satisfies

$$\begin{aligned} &\mathbb{E}\left(\|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2\right) \\ &\qquad \geq \frac{|\det(\mathbf{F}_{BB})|^{\frac{2}{N_{RF}^t}} \{\Gamma(1 + N_{RF}^t/2)\}^{\frac{2}{N_{RF}^t}}}{4\Gamma(3/2)^2 (N_{RF}^t + 2)}. \end{aligned} \tag{14}$$

Proof: By observing (13), we find that solving $\mathbf{x}$ given $\bar{\mathbf{y}}$ and $\mathbf{B}$ can be formulated as a closest vector problem within the lattice $\mathcal{L}(\mathbf{B})$. More specifically, to analyze $\mathbb{E}\left(\min_{\mathbf{x}} \|\bar{\mathbf{y}} - \mathbf{B}\mathbf{x}\|^2\right)$, $\bar{\mathbf{y}}$ and $\mathbf{B}$ can be respectively regarded as a query point and a lattice basis. The average residual distance between $\bar{\mathbf{y}}$ and the lattice point $\mathbf{B}\mathbf{x}$ can be analyzed.

Firstly, since $\{-1,1\}$ is a finite subset of the integer set $\mathbb{Z}$, the correct decoding region of the problem associated with $[\mathbf{x}]_m \in \{-1,1\}$ is strictly larger than that of $[\mathbf{x}]_m \in \mathbb{Z}$. Secondly, the second moment of the decision region of any decoder is no smaller than the Voronoi region $\mathcal{V}$ of lattice $\mathcal{L}(\mathbf{B}) = \left(\mathbf{B}\mathbf{x} \mid \mathbf{x} \in \mathbb{Z}^{N_{RF}^t}\right)$, where $\mathcal{V}(\Lambda)$ was defined in Eq. (9); it is a convex polytope whose second moment is defined as $\omega^2(\mathcal{V}) = \frac{1}{|\mathcal{V}|}\int_{\mathcal{V}} \|\mathbf{u}\|^2 \, \mathrm{d}\mathbf{u}$.

Let us now find the lower bound of $\omega^2(\mathcal{V})$. Recall that the isoperimetric inequality is a fundamental result in Euclidean geometry, which states that among all n-gons having a fixed perimeter, the one with the largest area is the regular n-gon. Based on the isoperimetric inequality, a ball $\mathcal{B}(r)$ with radius r has the smallest second moment per dimension out of all sets in $\mathbb{R}^{N_{RF}^t}$ with volume $V_N r^N$, where $V_N = 2^N \frac{\Gamma(3/2)^N}{\Gamma(1+N/2)}$ refers to the volume of an N-dimensional unit ball. Then the second moment of $\mathcal{B}(r)$ is given by

$$\begin{aligned}\omega^2(\mathcal{B}(r)) &= \frac{1}{N_{RF}^t V_{N_{RF}^t} r^{N_{RF}^t}} \int_{\mathbf{u}\in\mathcal{B}(r)} \|\mathbf{u}\|^2 \, \mathrm{d}\mathbf{u} \\ &= \frac{1}{N_{RF}^t V_{N_{RF}^t} r^{N_{RF}^t}} \int_0^r \bar{r}^2 \mathrm{d}\left(V_{N_{RF}^t} \bar{r}^{N_{RF}^t}\right) \\ &= \frac{r^2}{N_{RF}^t + 2}. \end{aligned} \tag{15}$$

Thus we have

$$\omega^2(\mathcal{V}) \geq \omega^2(\mathcal{B}(r)) = \frac{r^2}{N_{RF}^t + 2}, \tag{16}$$

and r is obtained by solving

$$\mathrm{Vol}(\mathcal{V}) \triangleq \sqrt{\det\left(\mathbf{B}^H \mathbf{B}\right)} = V_{N_{RF}^t} r^{N_{RF}^t}. \tag{17}$$

Lastly, we multiply $\omega^2(\mathcal{B}(r))$ with a factor of N_t (i.e., the number of instances in $\mathcal{P}4$) to get

$$\begin{aligned}&\mathbb{E}\left(\|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2\right) \\ &\geq \left(\frac{\left(1/\sqrt{N_t}\right)^{N_{RF}^t} |\det(\mathbf{F}_{BB})|}{V_{N_{RF}^t}}\right)^{\frac{2}{N_{RF}^t}} \frac{N_t}{N_{RF}^t + 2} \\ &= \frac{|\det(\mathbf{F}_{BB})|^{\frac{2}{N_{RF}^t}}}{\left(N_{RF}^t + 2\right) V_{N_{RF}^t}^{\frac{2}{N_{RF}^t}}} \\ &= \frac{|\det(\mathbf{F}_{BB})|^{\frac{2}{N_{RF}^t}}}{\left(N_{RF}^t + 2\right)\{2^{N_{RF}^t}\frac{\Gamma(3/2)^{N_{RF}^t}}{\Gamma(1+N_{RF}^t/2)}\}^{\frac{2}{N_{RF}^t}}} \\ &= \frac{|\det(\mathbf{F}_{BB})|^{\frac{2}{N_{RF}^t}} \{\Gamma(1+N_{RF}^t/2)\}^{\frac{2}{N_{RF}^t}}}{4\Gamma(3/2)^2\left(N_{RF}^t + 2\right)}. \end{aligned} \tag{18}$$

□

To show the accuracy of the lower bound in Proposition 1, we will conduct Monte Carlo simulations to compare the actual mean square error $\mathbb{E}\left(\|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2\right)$ and the lower bound $\frac{|\det(\mathbf{F}_{BB})|^{\frac{2}{N_{RF}^t}} \{\Gamma(1+N_{RF}^t/2)\}^{\frac{2}{N_{RF}^t}}}{4\Gamma(3/2)^2\left(N_{RF}^t+2\right)}$. For demonstration, we choose the elements of $\mathbf{F}_{opt}$ and $\mathbf{F}_{BB}$ from a Gaussian distribution $\mathcal{N}(0,1)$. Let $N_t = N_{RF}^t = N_s$, and draw $\mathbb{E}\left(\|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2\right)$ (noted as "simulated value") and $\frac{|\det(\mathbf{F}_{BB})|^{\frac{2}{N_{RF}^t}} \{\Gamma(1+N_{RF}^t/2)\}^{\frac{2}{N_{RF}^t}}}{4\Gamma(3/2)^2\left(N_{RF}^t+2\right)}$ (noted as "theoretical lower bound") for a range of dimensions in Fig. 2. We observe from the figure that our lower bound is loose, but this should not be a surprise, because we have resorted to a relaxation of the quantization set from three points $\{-1,0,1\}$ to the whole integer set $\mathbb{Z}$. Since this step is needed in the proof, how to bypass it to reach a tighter lower bound is left for our future work.

C. The Insight: Deactivate Part of the PSs for Enhanced Spectral Efficiency

Hereby we discuss the configuration of PSs, which in essence controls the domain of elements in $\mathbf{F}_{RF}$. Notice the set $\mathcal{F} = 1/\sqrt{N_t}\left\{e^{j\times 0}, e^{j\times\left(2\pi/2^B\right)}, \ldots, e^{j\times\left(2\pi/2^B\right)\times\left(2^B-1\right)}\right\}$ does not contain the zero point, but adding this point to the quantization set obviously reduces the quantization error. For completeness, we prove the following proposition for the case of 1-bit RF PSs.

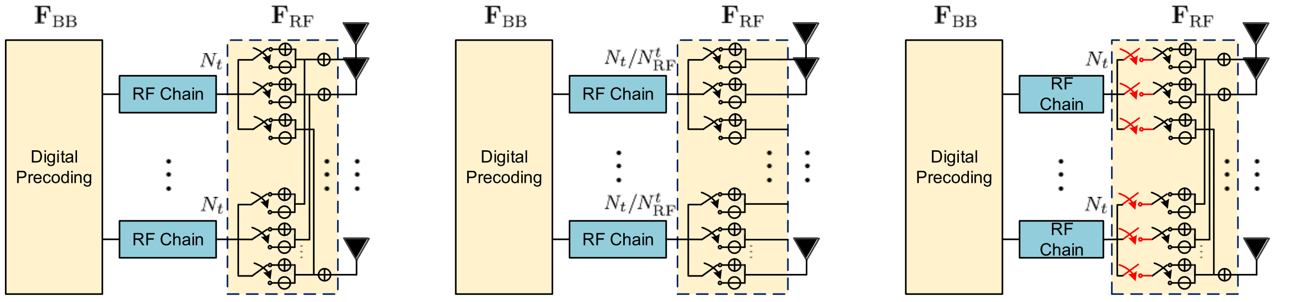
Proposition 2: The expected value of $\min_{\mathbf{F}_{RF}\in\mathcal{F}^{N_t\times N_{RF}^t}} \|\mathbf{F}_{opt} - \mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2$ *with the constraint* $\mathcal{F} = 1/\sqrt{N_t}\{-1,1\}$ *is larger than that with* $1/\sqrt{N_t}\{-1,0,1\}$.

Proof: In one dimension, the comparison is made between the quantization of a random number y with $c \times x$, where c is a multiplicative factor and $x \in \{-1,1\}$ or $\{-1,0,1\}$. For $y \leq -c/2$ or $y \geq c/2$, their quantization errors are the same. However, for $-c/2 \leq y \leq c/2$, obviously making quantization with respect to $\{-1,0,1\}$ has smaller quantization error (i.e. $\leq c/2$). For dimensions higher than 2, partitions over the whole Euclidean space can be similarly made, and the error set associated with $\{0\}$ enjoys a smaller quantization error. □

Based on the aforementioned reason, we propose to add $N_t \times N_{RF}^t$ switches to all the PSs, so as to reach the best possible performance in finite-accuracy TPCs. For high accuracy PSs, it is not essential to add switches, because the improvement of spectral efficiency would remain modest. While the idea of making the PS network sparse is not new [15], [28], [36], [37], these impressive contributions use a small fixed number of PSs. In a quest for fully understanding the best performance of hybrid TPC using low resolution PSs, especially 1-bit PSs, all PSs are followed by switches hereby. We conceive three types of hybrid TPC architectures in Fig. 1. The one-bit accuracy scenario of $\mathcal{F} = 1/\sqrt{N_t}\{-1,1\}$ is drawn in Fig. 1-(a). The partially-connected architecture [28] where each RF chain is connected to N_t/N_{RF}^t antennas is shown in Fig. 1-(b). Lastly, as shown in Fig. 1-(c), the configuration proposed in this work is to deactivate a fraction of $1-\varepsilon$ of the PSs with $0 < \varepsilon < 1$, in which the actual value of ε is determined by the algorithm. We term this an adaptive RF PS connection architecture, and the new constraint is formulated as

$$\mathcal{S} = \mathcal{F} \cup 0.$$

The special case of the proposed pattern with $\mathcal{S} = 1/\sqrt{N_t}\{-1,0,1\}$ is drawn in Fig 1-(c).



(a) Fully-connected network for RF PSs . (b) Sub-connected network for RF PSs. (c) Adaptively connected network for RF PSs.

Fig. 1. Three types of hybrid precoding architectures based on different connections.

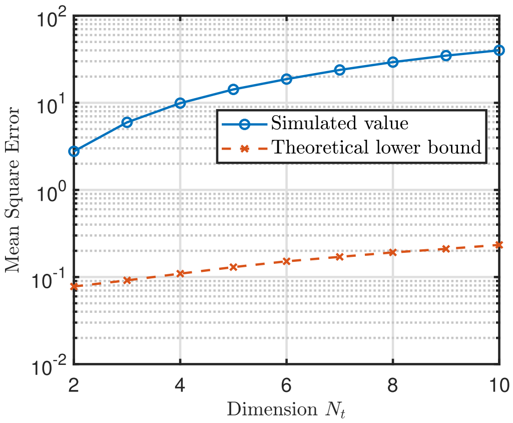


Fig. 2. Demonstration of the lower bound.

The difference between the conventional partially-connected architecture [28] of Fig. 1-(b) and the proposed adaptive RF PS connection architecture can be explicitly seen in the matrix formulation of the analog beamformer $\mathbf{F}_{RF}$. Specifically, $\mathbf{F}_{RF}$ in the sub-connected architecture is in a block diagonal form:

$$\mathbf{F}_{RF} = \begin{bmatrix} \bar{\mathbf{f}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{f}}_2 & & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{f}}_{N^t_{RF}} \end{bmatrix}_{N_t \times N^t_{RF}}, \tag{19}$$

where each RF chain is only connected to a sub-antenna array with $M \triangleq N_t/N^t_{RF}$ and each $\bar{\mathbf{f}}_k$ is an $M \times 1$ vector with elements in $1/\sqrt{N_t}\,\{-1, 1\}$. Given a total of N_t nonzero elements, the sparsity ratio defined by Eq. (19) is therefore $1 - 1/N^t_{RF}$. Regarding the proposed adaptive connection network of RF PSs, *we do not specify the sparsity in* $\mathbf{F}_{RF}$ *and let the positions of zero elements be determined by the hybrid precoding algorithm that optimizes the spectral efficiency.* Since the positions of nonzero elements are random, the $\mathbf{F}_{RF}$ in the proposed network obeys:

$$\mathbf{F}_{RF} = \begin{bmatrix} \mathbf{0} & \bar{\mathbf{f}}_1 & \cdots & \mathbf{0} \\ \bar{\mathbf{f}}_2 & \mathbf{0} & & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{f}}_{N^t_{RF}} \end{bmatrix}_{N_t \times N^t_{RF}}. \tag{20}$$

IV. The Algorithm Proposed for Hybrid Precoding/Combining

While $\mathcal{P}4$ admits non-trivial constraints, the hardness of solving it varies depending on the orthogonality property of $\mathbf{B}$ (i.e., $\mathbf{F}_{BB}$). In this section we propose efficient solvers for $\mathcal{P}4$ so as to accelerate the convergence of alternating minimization while meeting the constraint. Although the open literature [19], [28], [33] has stated the NP-hard nature of solving combinatorial problems, the decoding scheme that we conceived in this paper, represents the very first attempt to design low-complexity lattice decoding [41], [44] rather than using convex optimization.

We focus our attention on the configuration of $N_s \geq \max(N^t_{RF}, N^r_{RF})$ in this work, since QR factorization cannot be applied to under-complete matrices. For the scenario of $N_s < \max(N^t_{RF}, N^r_{RF})$, we refer to [45] for solving combinatorial problems with the aid of approximate message passing.

A. Babai's Algorithm for Alternating Minimization

Hereby we introduce Babai's algorithm [38] from the lattice decoding literature (also known as Babai's nearest plane algorithm [38], and as the successive interference cancellation (SIC) in MIMO detection [46], [47]) to address $\mathcal{P}4$ subject to a uniform prior probability constraint. Based on QR factorization, we have $\mathbf{F}^\top_{BB} = \mathbf{QR}$ where $\mathbf{Q} \in \mathbb{C}^{N_s \times N^t_{RF}}$ has orthogonal columns and $\mathbf{R} \in \mathbb{C}^{N^t_{RF} \times N^t_{RF}}$ is an upper triangular matrix. Without loss of generality, we assume that $[\mathbf{R}]_{m,m} \in \mathbb{R}$, $[\mathbf{R}]_{m,m} > 0$ for $m = 1, \ldots, N^t_{RF}$ throughout the paper.

As for the residual vector $\mathbf{v} \triangleq \bar{\mathbf{y}} - \mathbf{Bx}$, the probability distributions of its entries are unknown. Let us furthermore define $\tilde{\mathbf{y}} = \mathbf{Q}^H \bar{\mathbf{y}}$ and $\tilde{\mathbf{v}} = \mathbf{Q}^H \mathbf{v}$. By left multiplying $(\bar{\mathbf{y}} - \mathbf{Bx})$ with $\mathbf{Q}^H$ we have

$$\min_{\mathbf{x}} \|\bar{\mathbf{y}} - \mathbf{Bx}\|^2 = \min_{\mathbf{x}} \|\tilde{\mathbf{y}} - \mathbf{Rx}\|^2, \tag{21}$$

where the structure within the r.h.s. can be expressed as $\tilde{\mathbf{y}} =$

$$\begin{bmatrix} [\mathbf{R}]_{1,1} & [\mathbf{R}]_{1,2} & \cdots & [\mathbf{R}]_{1,N^t_{RF}} \\ 0 & [\mathbf{R}]_{2,2} & \cdots & [\mathbf{R}]_{2,N^t_{RF}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & [\mathbf{R}]_{N^t_{RF},N^t_{RF}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{N^t_{RF}} \end{bmatrix} + \begin{bmatrix} \tilde{v}_1 \\ \tilde{v}_2 \\ \vdots \\ \tilde{v}_{N^t_{RF}} \end{bmatrix}.$$

The general idea behind Babai's nearest plane algorithm is to process the received vector $\tilde{\mathbf{y}}$ to estimate each component of the transmitted signal $\mathbf{x}$ in a recursive way, thus canceling the effect of the already decoded symbols and nulling those yet unknown. If a symbol $\hat{x}_m$ has been estimated, the decoder will exploit this decision to further estimate the remaining symbols $\hat{x}_{m-1}, \ldots, \hat{x}_1$, forming a nonlinear decoding structure. If the non-convex constraint on $\mathbf{x}$ is temporarily removed, we can obtain analytic solutions for $\hat{x}_{m-1}, \ldots, \hat{x}_1$, as specified by the following proposition.

Proposition 3: When given $\left[x_1, \ldots, x_{N_{RF}^t - 1}\right]$ *or* $\left[x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_{N_{RF}^t}\right]$, *the solutions of*

$$\min_{x_{N_{RF}^t} \in \mathbb{C}} \left| \tilde{y}_{N_{RF}^t} - [\mathbf{R}]_{2,N_{RF}^t} x_{N_{RF}^t} \right|^2, \tag{22}$$

$$\vdots \tag{23}$$

$$\min_{x_1 \in \mathbb{C}} \left| \tilde{y}_1 - \sum_{m=1}^{N_{RF}^t} [\mathbf{R}]_{1,m} x_m \right|^2, \tag{24}$$

are respectively given by

$$\hat{x}_{N_{RF}^t} = \frac{\tilde{y}_{N_{RF}^t}}{[\mathbf{R}]_{N_{RF}^t, N_{RF}^t}}, \tag{25}$$

$$\hat{x}_m = \frac{\tilde{y}_m - \sum_{k=m+1}^{N_{RF}^t} [\mathbf{R}]_{m,k} \hat{x}_k}{[\mathbf{R}]_{m,m}}, \tag{26}$$

for $m = N_{RF}^t - 1, \ldots, 1$.

Proof: The objective functions $\left| \tilde{y}_{N_{RF}^t} - [\mathbf{R}]_{2,N_{RF}^t} x_{N_{RF}^t} \right|^2, \ldots, \left| \tilde{y}_1 - \sum_{m=1}^{N_{RF}^t} [\mathbf{R}]_{1,m} x_m \right|^2$ in the proposition are all convex. By setting

$$\frac{\partial \left| \tilde{y}_{N_{RF}^t} - [\mathbf{R}]_{2,N_{RF}^t} x_{N_{RF}^t} \right|^2}{\partial x_{N_{RF}^t}} = 0,$$

$$\vdots$$

$$\frac{\partial \left| \tilde{y}_1 - \sum_{m=1}^{N_{RF}^t} [\mathbf{R}]_{1,m} x_m \right|^2}{\partial x_1} = 0,$$

the solutions are reached. □

Based on the above, we can further quantize the arguments from $\mathbb{C}$ to $\mathcal{S}$. The estimated symbols can therefore be written as

$$\hat{x}_{N_{RF}^t} = \mathcal{Q}_{\mathcal{S}} \left(\frac{\tilde{y}_{N_{RF}^t}}{[\mathbf{R}]_{N_{RF}^t, N_{RF}^t}} \right), \tag{27}$$

$$\hat{x}_m = \mathcal{Q}_{\mathcal{S}} \left(\frac{\tilde{y}_m - \sum_{k=m+1}^{N_{RF}^t} [\mathbf{R}]_{m,k} \hat{x}_k}{[\mathbf{R}]_{m,m}} \right) \tag{28}$$

for $m = N_{RF}^t - 1, \ldots, 1$. Compared to conventional SIC detectors used in MIMO detection, the quantization function $\mathcal{Q}_{\mathcal{S}}$ here is taken with respect to the discretized unit-power constraint $\mathcal{S}$, and not a quadrature amplitude modulation (QAM) constellation. The whole procedure is summarized in Algorithm 2.

Algorithm 2 Babai's Algorithm for Solving $\mathcal{P}_3$

```
Input: $\mathbf{F}_{\mathrm{opt}}$, $\mathbf{F}_{\mathrm{BB}}$.
Output: $\mathbf{F}_{\mathrm{RF}}$.
Perform QR factorization on $\mathbf{F}_{BB}^\top$ to obtain $\mathbf{F}_{BB}^\top = \mathbf{QR}$ ;
for $n = 1, \ldots N_t$ do
    $\tilde{\mathbf{y}} = \mathbf{Q}^H \left[\mathbf{F}_{opt}^\top\right]_{1:N_s,n}$;
    $\hat{x}_{N_{RF}^t} = \mathcal{Q}_{\mathcal{S}} \left( \frac{\tilde{y}_{N_{RF}^t}}{[\mathbf{R}]_{N_{RF}^t, N_{RF}^t}} \right)$;
    for $m = N_{RF}^t - 1, \ldots, 1$ do
        $\hat{x}_m = \mathcal{Q}_{\mathcal{S}} \left( \frac{\tilde{y}_m - \sum_{k=m+1}^{N_{RF}^t} [\mathbf{R}]_{m,k} \hat{x}_k}{[\mathbf{R}]_{m,m}} \right)$
    $[\mathbf{F}_{RF}]_{n,1:N_{RF}^t} = \hat{\mathbf{x}}^\top$.
```

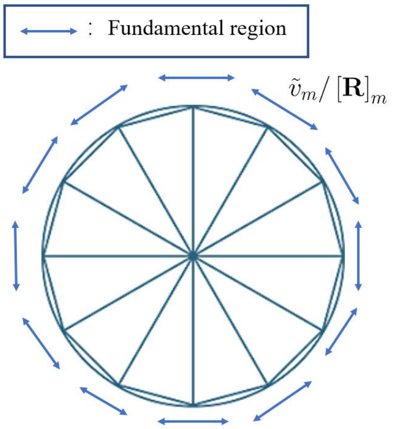


Fig. 3. The quantization region of a 4-bit PS.

Since the effective noises in the quantization process are

$$\begin{cases} \tilde{v}_{N_{RF}^t} / [\mathbf{R}]_{N_{RF}^t, N_{RF}^t}, \\ (\tilde{v}_m + \sum_{k=m+1}^{N_{RF}^t} [\mathbf{R}]_{m,k} (x_k - \hat{x}_k)) / [\mathbf{R}]_{m,m}, \\ \cdots \\ (\tilde{v}_1 + \sum_{k=2}^{N_{RF}^t} [\mathbf{R}]_{1,k} (x_k - \hat{x}_k)) / [\mathbf{R}]_{1,1}, \end{cases}$$

Babai's low-complexity algorithm enjoys the optimal performance if the noise power in each layer falls into the fundamental quantization region. As shown in Fig. 3, where a 4-bit PS with 16 fixed phases divides the unit circle, a fundamental region refers to points in the complex domain, where the quantization with respect to the 16 fixed phases has no ambiguity. All $\tilde{v}_m / [\mathbf{R}]_m$ should fall within the fundamental region to make Babai's algorithm optimal, while optimal is defined in the sense that the error term $||\tilde{\mathbf{y}} - \mathbf{R}\hat{\mathbf{x}}||$ corresponds to that of an exhaustive search.

B. Power Consumption and Energy Efficiency

With reference to the energy consumption model in [28], [48], the total power consumption of a MIMO system using a hybrid TPC is given by

$$P = P_{common} + N_{RF}^t P_{RF} + N_t P_{PA} + N_{PS}(P_{PS,k} + P_{SW}), \tag{29}$$

where P_{common}, P_{RF}, P_{PA}, P_{IN}, and $P_{PS,k}$ respectively denote the common power of the transmitter, the power of each RF chain, the power of each power amplifier, the power consumption of a switch, and the power of a k-bit PS. The power consumed by the PSs depends on the type [1] and on the resolution of the quantized phases. Their approximate values can be chosen as $P_{RF} = 100\,\mathrm{mW}$, $P_{PA} = 100\,\mathrm{mW}$, $P_{SW} = 2.5\,\mathrm{mW}$, $P_{PS,1} = 2.5\,\mathrm{mW}$, $P_{PS,2} = 10\,\mathrm{mW}$, and $P_{PS,4} = 45\,\mathrm{mW}$ [9], [12], [28], [49]. For comparison purposes, we set the power $P_{PS,inf}$ of an infinite resolution PS to $P_{PS,inf} \approx P_{PS,5} = 78\,\mathrm{mW}$ based on [49]. To consume less power, we recommend the usage of 1-bit/2-bit resolution PSs, since we found no substantial spectral efficiency improvements by increasing the number of bits in PSs.

The number of PSs N_{PS} for different connection schemes are given as [26], [27]

$$N_{PS} = \begin{cases} N_t N_{RF}^t & \mathrm{fully-connected,\ Fig.\ 1(a)} \\ N_t & \mathrm{sub-connected,\ Fig1(b)} \\ (1-\varepsilon) N_t N_{RF}^t & \mathrm{adaptively-connected,\ Fig.1(c).} \end{cases}$$

Recall that ε refers to the ratio of sparsity in the configuration of PSs. The power consumption of the proposed scheme varies between that of the fully connected and partially connected schemes when $1/N_{RF}^t \leq 1 - \varepsilon \leq 1$.

Finally, given the spectral efficiency R, the energy efficiency of the TPC is defined as

$$\eta = R/P\,(\mathrm{bps/Hz/W}).$$

As expected, low spectral efficiency schemes may exhibit a high energy efficiency.

C. Analysis of Computational Complexity

Here we only examine the computational complexity of the hybrid TPC in terms of the number of complex floating point operations, since the analysis of the hybrid receiver combiner is similar (e.g., change N_{RF}^t to N_{RF}^r, N_t to N_r). Since $N_s \geq \max(N_{RF}^t, N_{RF}^r)$, the computational complexity of QR factorization is on the order of $O[N_s(N_{RF}^t)^2]$, and the computational complexity of completing N_{RF}^t quantization using Babai's algorithm is no higher than $O[N_{RF}^t\,(N_{RF}^t + |\mathcal{S}|)]$. Since the size of $\mathcal{S}$ is not increasing like the number of data streams or of antennas, its complexity can be omitted. Letting N_{iter}^o refer to the number of iterations in the alternating minimization algorithm, and $O[N_t(N_{RF}^t)^2]$ denote the computational complexity of step 3 in Algorithm 1, the overall computational complexity of our algorithm is

$$\begin{aligned} N_{iter}^o (N_t &\left(O[N_s(N_{RF}^t)^2] + O[N_{RF}^t \left(N_{RF}^t + |\mathcal{S}|\right)]\right) \\ &+ O[N_t(N_{RF}^t)^2]) \\ &= O[N_{iter}^o N_t N_s (N_{RF}^t)^2]. \end{aligned} \tag{30}$$

For comparison, we list the complexity of our algorithm and other state-of-the-art hybrid TPC algorithms based on quantized PSs in TABLE II. Generally, our algorithm enjoys much lower complexity than the CDM-AltMin of [19] and the WLLS of [20], since N_t is often large.

TABLE II

COMPLEXITY COMPARISON

Algorithms	Computational Complexity
Babai-AltMin (Ours)	$O[N_{iter}^o N_t N_s (N_{RF}^t)^2]$
CDM-AltMin [19]	$O[N_{iter}^o N_{iter}^i N_t^3 N_s (N_{RF}^t)^2]$
WLLS ($B > 1$) [20]	$O[N_{iter}^o N_{iter}^i N_t^3 N_{RF}^t]$
WLLS ($B = 1$) [20]	$O[N_{iter}^o N_t^3 N_{RF}^t]$

V. SIMULATION RESULTS

In this section, we perform simulations for a point-to-point mmWave massive MIMO system, whose transmitter and receiver are equipped with a uniform square planar array of half wavelength spacing. As for the channel model of (3), we adopt the setting of $N_{cl} = 5$ clusters and $N_{ray} = 10$ rays per cluster, where the average power of each cluster is 1, while the azimuth and elevation AODs as well as AOAs follow the Laplacian distribution with uniformly distributed mean angles and angular spread of 10°. The simulation results are averaged over 2×10^3 Monte-Carlo Runs. The number of loops in all algorithms are fixed as 10, unless specified otherwise. Babai's algorithm used for our alternation minimization based hybrid TPC is noted as "Babai-AltMin". Our benchmark algorithms include[2]:

- the optimal fully digital TPC that serves as the performance upper bound, denoted as "Fully digital";
- the algorithm from [20] denoted as "WLLS";
- the conjugate descent method of [19] denoted as "CDM-AltMin".
- the algorithm from [27] denoted as "SIC".

Moreover, the different types of configurations shown in Fig. (1) are respectively marked as "full", "sub", and "adap". To foster reproducible research, our programs used in the simulations are of open source and freely available at GitHub.[3]

A. Spectral Efficiency

1) Impact of SNR: In Fig. 4, we plot the spectral efficiency versus SNR based on several state-of-the-art algorithms, where subfigure (a) uses $N_s = 4$ and subfigure (b) uses $N_s = 12$.

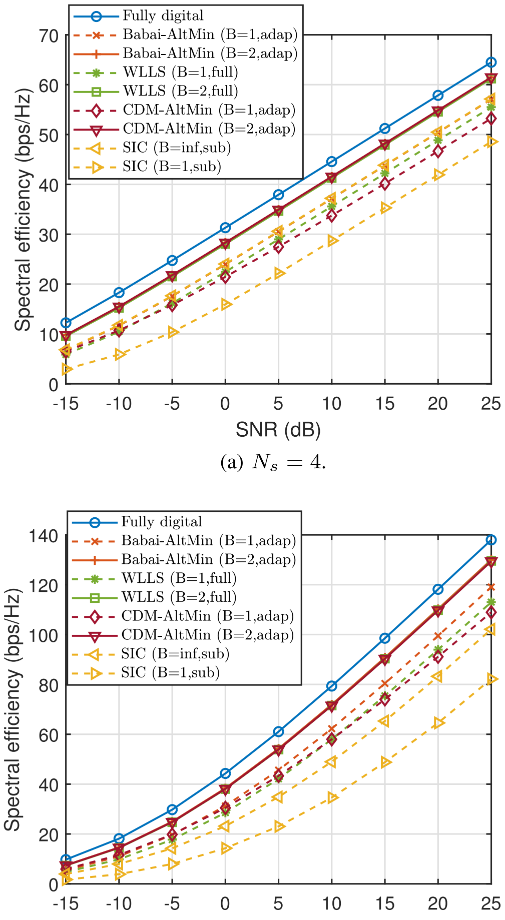
A few observations can be made from the figure:

- CDM-AltMin ($B = 1$,adap) falls behind Babai-AltMin ($B = 1$,adap) in both subfigures, e.g., losing 2.5dB in subfigure (a), even though they both belong to the alternating minimization approach.
- WLLS ($B = 1$) falls behind Babai-AltMin ($B = 1$,adap) by about 1dB in both subfigures.
- In subfigure (a), SIC ($B = inf$,sub) exhibits the same rates as the proposed Babai-AltMin ($B = 1$,adap), but that algorithm employs infinite-resolution PSs, and this result is the best that a sub-connected architecture can achieve. By contrast, other algorithms can still enjoy higher rates by increasing the resolution of PSs. In subfigure (b), SIC ($B = inf$,sub) falls 5dB behind

[1] There exist at least six types of phase shifters at mmWave frequencies: reflective type, loaded line, switched delay, Cartesian vector modulator, LO-path phase shifter, and phase over sampling vector modulator [49].

[2] Since the spectral efficiency of analog beamforming is significantly worse than that of hybrid beamforming [9], [10], it is excluded from our benchmarkers.

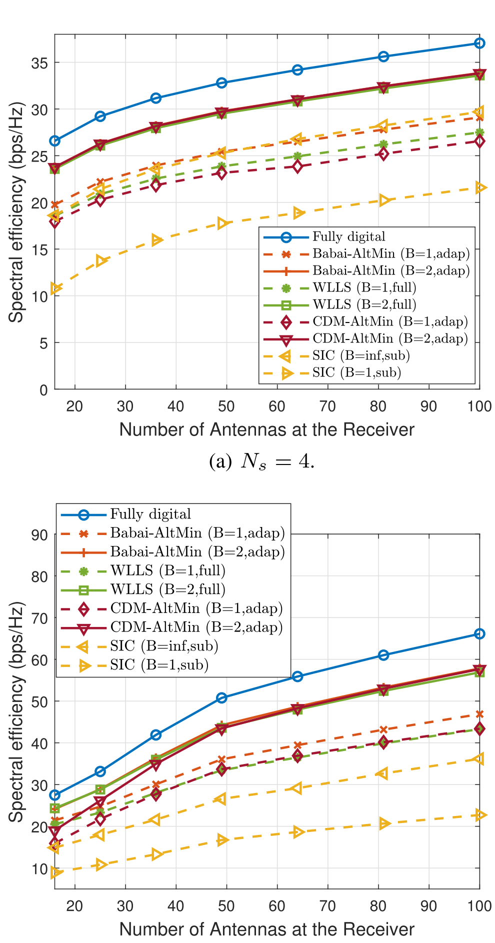
[3] https://github.com/shx-lyu/hybrid-precoding



(a) $N_s = 4$.

(b) $N_s = 12$.

Fig. 4. Spectral efficiency versus SNR ($N_t = 144, N_r = 36, N_{RF}^t = N_{RF}^r = N_s$).



(a) $N_s = 4$.

(b) $N_s = 12$.

Fig. 5. Spectral efficiency versus number of antennas N_r (SNR = 0dB, $N_t = 144, N_{RF}^t = N_{RF}^r = N_s$).

Babai-AltMin ($B = 1$,adap). Additionally, the SIC ($B =$ 1,sub) algorithm performs poorly.

- Once the number of bits B increases to 2, the AltMin based algorithms and WLLS exhibit similar spectral efficiency.

2) Impact of the Number of Antennas: Letting the number of RAs N_r vary from 16 to 100, at SNR = 0dB, and keeping other settings the same as above, we further illustrate the spectral efficiency versus the number of RAs in Fig. 5, respectively using $N_s = 4$ and $N_s = 12$. Similarly to the aforementioned conclusions, the Babai-AltMin algorithm outperform the others in general. It is worth mentioning that SIC (B = inf,sub) surpasses Babai-AltMin ($B = 1$,adap) in subfigure (a) when $N_r \geq 60$.

3) Impact of the Number of Streams: Letting the number of streams N_s vary from 12 to 16, at SNR = 10dB, we plot the spectral efficiency versus the number of streams in Fig. 6. Note that the WLLS algorithm fails to support N_s values higher than the number of RF chains. As revealed in the figure, the spectral efficiencies of the hybrid beamforming schemes decrease slightly when N_s is increased, while our scheme is still the best option among the benchmarkers.

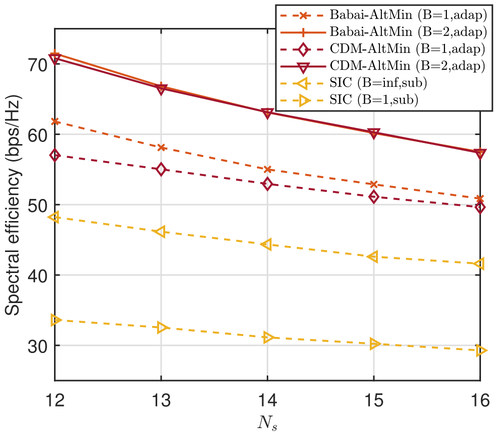
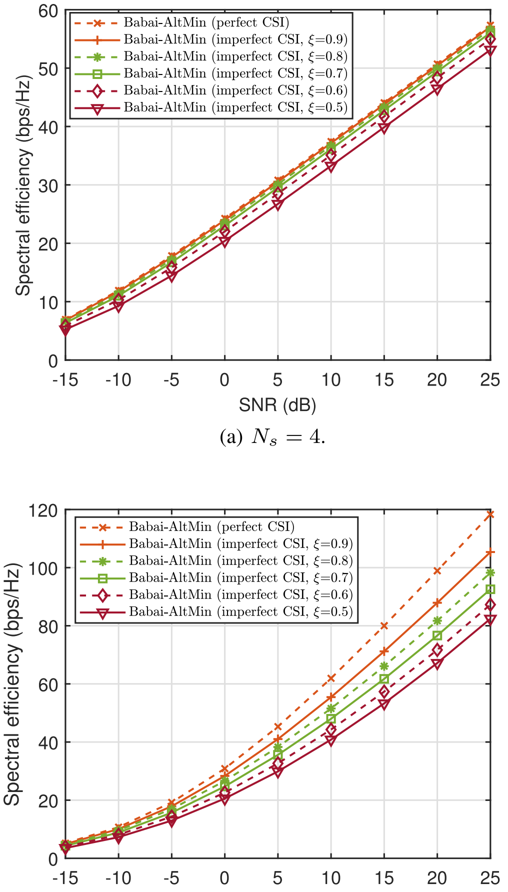


Fig. 6. Spectral efficiency versus the number of streams N_s (SNR = 10dB, $N_t = 144, N_r = 36, N_{RF}^t = N_{RF}^r = 12$).

4) Impact of Imperfect CSI: We further evaluate the impact of imperfect CSI on performance of the proposed algorithm. The estimated channel matrix $\hat{\mathbf{H}}$ is formulated as

$$\hat{\mathbf{H}} = \xi \mathbf{H} + \sqrt{1 - \xi^2} \mathbf{E}, \tag{31}$$



(a) $N_s = 4$.

(b) $N_s = 12$.

Fig. 7. Spectral efficiency versus SNR with imperfect CSI ($N_t = 144$, $N_r = 36$, $N_{RF}^t = N_{RF}^r = N_s, B = 1$).

where $\mathbf{H}$ is the actual channel matrix, $\xi \in [0, 1]$ is a parameter that controls the accuracy of CSI, and $\mathbf{E}$ is the error matrix with entries following the standard complex Gaussian distribution. Fig. 7 shows our achievable rate comparison for mmWave MIMO systems, where both perfect and imperfect CSI have been considered for the 1-bit resolution PSs. We observe that the proposed Babai-AltMin algorithm is not sensitive to the accuracy of CSI. In particular, as shown in the case of $N_s = 4$, the achievable rate of the Babai-AltMin algorithm based on imperfect CSI associated with $\xi = 0.7$ is extremely close (within 0.5dB) to that in the perfect CSI scenario. Even $\xi = 0.5$ can achieve more than 90% of the rate attained in the perfect CSI case. When the number of data streams becomes higher ($N_s = 12$), Fig. 7-(b) reveals that, as the SNR increases, we have at most 3dB SNR gap between the spectral efficiencies of perfect CSI and imperfect CSI of $\xi = 0.9$. Similar conclusions can also be drawn for other ξ values.

B. Complexity

We have shown that the overall complexity of the scheme is $O[N_{iter}^o N_t \left(N_s^3 + N_s|\mathcal{S}|\right)]$. Hereby we also plot the actual running time of the above algorithms. As shown in Fig. 8, Babai's algorithm is faster than both CDM (about 10 times) and WLLS (about 5 times), where these algorithms exhibit an iteration-based structure. We also observe that SIC is significantly faster than all other algorithms, although its structure is quite different from the others.

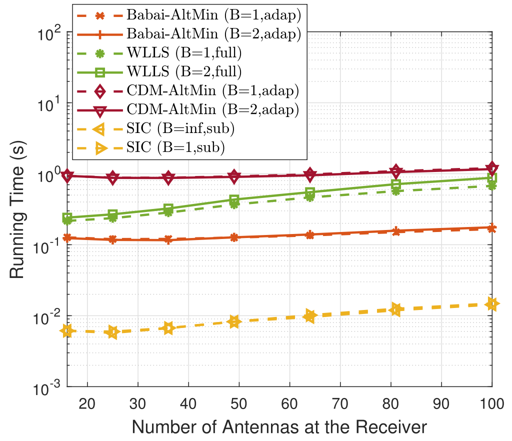


Fig. 8. The running time of different algorithms.

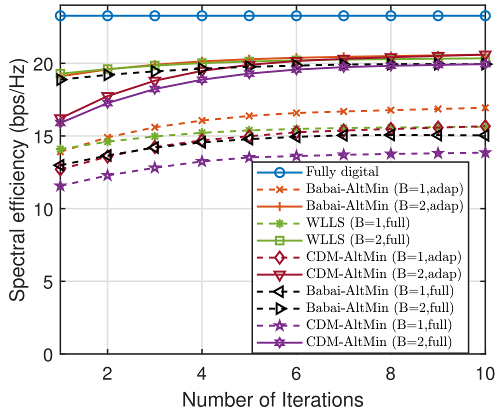


Fig. 9. Spectral efficiency versus number of iterations (SNR = 0dB, $N_t = 144$, $N_r = 36$, $N_s = N_{RF}^t = N_{RF}^r = 4$).

In the second example, we plot the spectral efficiency versus the number of iterations N_{iter}^o in Fig. 9, where all the algorithms compared rely on an iteration-based structure. The figure shows: i) The spectral efficiencies saturate at their best values after around 6 iterations.; ii) Similarly to WLLS, Babai-AltMin converges faster than CDM-AltMin, although they both belong to the AltMin-based category. iii) the advantage of using Babai-AltMin is more evident when $B = 1$. iv) The merits of employing a sparse configuration within the AltMin-based methods become obvious. By using $B = 1$ and $\mathcal{S} = \mathcal{F} \cup 0$, Babai-AltMin AND CDM-AltMin enjoy around 1bps/Hz gain in spectral efficiency over their counterparts configured with $\mathcal{F}$. This gain drops to around 0.5bps/Hz when $B = 2$.

The fast convergence of Babai's algorithm actually hinges on how close the basis $\mathbf{B}$ is to being orthogonal. We introduce a metric referred to as the condition number C, which is defined as the ratio of the largest to smallest singular value in the singular value decomposition of the matrix. The degree of orthogonality is quantified by the amount of distortion of the unit sphere under the transformation by the matrix.

In Fig. (10), we plot the condition number versus the number of iterations within Babai-AltMin algorithm. By increasing

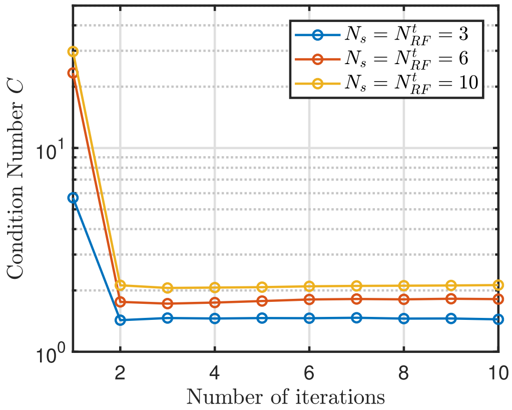


Fig. 10. The condition number of the basis $\mathbf{B}$.

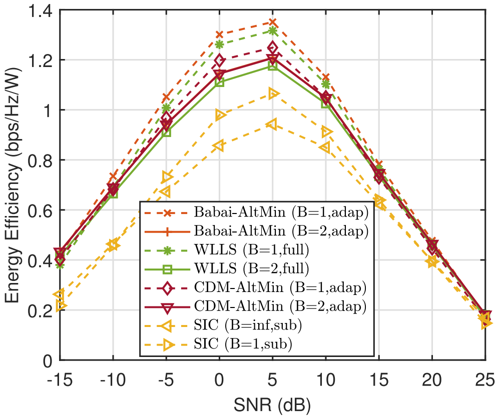
the dimensions of $\mathbf{B}$ from 3×3, to 6×6 and 10×10, Fig. (10) shows that all the condition numbers converge within as few as two iterations, which means that the algorithm only has to use the feedback from Babai's algorithm once and then the solution gets close to the optimal. The implication of this figure is that the alternating minimization method along with Babai's kernel we specified results in a rapidly converging solution. Most importantly, since the quality of the basis $\mathbf{B}$ is good ($C \approx 1$), Babai's algorithm performs well and hence we may dispence with exponential-complexity solutions.

C. Energy Efficiency

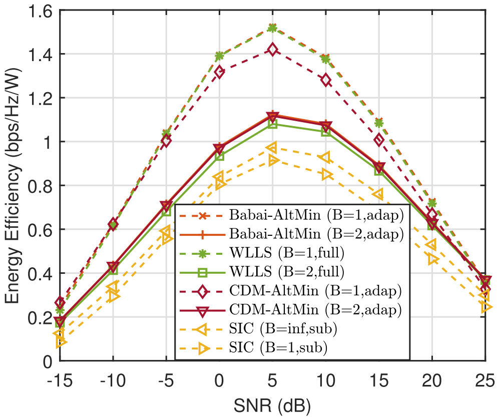
In this subsection we examine the energy efficiency of the related algorithms based on the formulas and typical power consumption values given in Section IV-B.

1) Impact of SNR: Whilst relying on the benchmarks used before, Fig. 11 plots the energy efficiency versus SNR. We can observe that, in Fig. 11-(a), the Babai-AltMin (B = 1,adap) solution outperforms both others with a notable margin when SNR is in the range of $-10 \sim 20$dB, while other solutions using $B = 1$ generally perform better than their high-resolution counterparts (as $P_{PS,1} = 2.5\,\text{mW}$ is significantly lower than $P_{PS,inf} \approx P_{PS,5} = 78\,\text{mW}^4$). In the second subfigure of $N_s = 12$, the major differences with Fig. 11-(a) are that WLLS ($B = 1$, full) gets very close to Babai's algorithm, and SIC ($B = inf$, sub) offers slightly higher energy efficiency than SIC ($B = 1$, sub). Although not shown in the figure, it is noteworthy that high-resolution PS based schemes (such as the Babai-AltMin scheme using many bits) would yield very low energy efficiency, because the corresponding power consumption is excessive.

We further explain why the resolution and the number of PSs exhibit more evident benefits when SNR is not too large. In the expression of the energy efficiency $\eta = R/P$, the numerator only increases logarithmically with the power of the transmitter P_{common}, while the denominator increases linearly. Hence the denominator increases faster and as a result the energy efficiency decreases asymptotically, and the energy efficiency discrepancies of different algorithms vanish.



(a) $N_s = 4$.



(b) $N_s = 12$.

Fig. 11. Energy efficiency versus SNR ($N_t = 144, N_r = 36, N^t_{RF} = N^r_{RF} = N_s$).

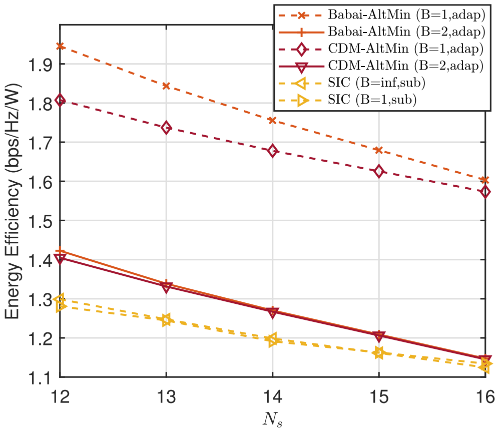
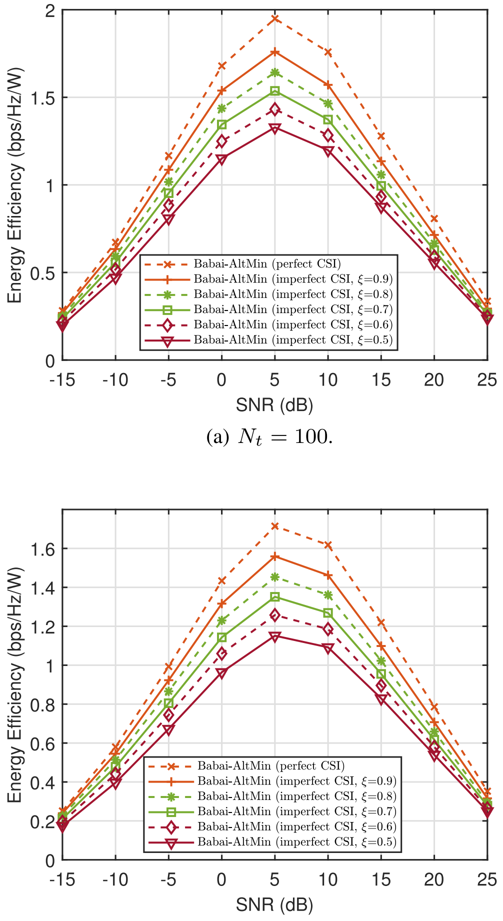


Fig. 12. Energy efficiency versus the number of streams N_s (SNR = 10dB, $N_t = 144, N_r = 36, N^t_{RF} = N^r_{RF} = 12$).

2) Impact of Other Factors: The impact of other factors on the energy efficiency is briefly compared hereby, where the simulation results support the superiority of Babai-AltMin (B = 1,adap) over others. Firstly, as shown in Fig. 12 concerning the relationship of energy efficiency versus the number of streams, Babai-AltMin (B = 1,adap) and CDM-AltMin (B = 1,adap) perform much better than the other algorithms. Secondly, we evaluate how the variations of antennas

[4] Note that [27] has set $P_{PS,inf} = 1\,\text{mW}$, but we argue that our chosen values are more consistent with those in [12], [28], [49].



(a) $N_t = 100$.

(b) $N_t = 144$.

Fig. 13. Energy efficiency versus SNR using imperfect CSI ($N_r = 36, N^t_{RF} = N^r_{RF} = N_s = 12, B = 1$).

and CSIs affect the energy efficiency in Fig. 13. Explicitly, the figure shows that the proposed Babai-AltMin is robust - its energy efficiency degradation is proportional to the CSI accuracy. Additionally, comparing the two subfigures reveals that the achievable energy efficiency is slightly reduced as N_t increases. This justifies the effect of N_t in the power consumption model of Eq. (29).

VI. Conclusion

While it has been recognized that using low resolution PSs leads to a certain loss in spectral efficiency, we have explicitly quantified the spectral penalty of using 1-bit resolution PSs in hybrid TPCs. To improve the performance of the system when given a fixed PS network, an adaptively connected RF PS network has been proposed. For the sake of solving the TPC optimization problem efficiently, we have adapted Babai's lattice decoding algorithm. Our simulation results have shown that our method achieves around 1dB SNR gain in terms of spectral efficiency over the state-of-the-art in a 1-bit PS scenario.

Appendix A
Discussion on Controlling Sparsity

The designated objective function has the constraint of $\mathcal{S} = \mathcal{F} \cup 0$, but it does not reveal the position and the number of the 0s. These two parameters are determined upon solving $\mathcal{P}4$. It is readily understood that tuning the sparsity cannot improve the spectral efficiency, since such feasible solutions only constitute a subset of $\mathcal{S}$. Nevertheless, we may wonder whether a sharp reduction in power consumption is capable of compensating for the loss in spectral efficiency, so as to actually arrive at an improved energy efficiency.

For completeness, we briefly highlight how a Bayesian algorithm is designed to control the sparsity of the RF TPC matrices. Specifically, if we tune the sparsity ratio in both $\mathbf{F}_{RF}$ and $\mathbf{W}_{RF}$, then we have two more constraints imposed on $\mathcal{P}1$, yielding

$$\mathcal{P}5: \max_{\mathbf{F}_{RF},\mathbf{F}_{BB},\mathbf{W}_{RF},\mathbf{W}_{BB}} \mathcal{I}\left(\mathbf{F}_{RF},\mathbf{F}_{BB},\mathbf{W}_{RF},\mathbf{W}_{BB}\right), \tag{32a}$$

$$s.t. \quad \|\mathbf{F}_{RF}\mathbf{F}_{BB}\|_F^2 = N_s, \tag{32b}$$

$$[\mathbf{F}_{RF}]_{m,n} \in \mathcal{S}, \quad \forall m,n, \tag{32c}$$

$$[\mathbf{W}_{RF}]_{m,n} \in \mathcal{S}, \quad \forall m,n, \tag{32d}$$

$$[\mathbf{F}_{RF}]_{m,n} \sim p_{\varepsilon'}, \quad \forall m,n, \tag{32e}$$

$$[\mathbf{W}_{RF}]_{m,n} \sim p_{\varepsilon'}, \quad \forall m,n, \tag{32f}$$

where $p_{\varepsilon'}$ is a probability distribution with the non-zero ratio ε':

$$p_{\varepsilon'}(x) = (1-\varepsilon')\delta(x) + \varepsilon'/|\mathcal{F}| \sum_{y\in\mathcal{F}} \delta(x-y), \tag{33}$$

and $\delta(\cdot)$ is the Dirac delta function. The last two constraints in (32e) and (32f) make this a more complex optimization task than those simply having finite-accuracy constraints [19], [26]. Then we can use the Gaussian approximation of $\tilde{\mathbf{v}} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$, where σ^2 denotes the effective noise variance to be specified. Based on the above, the non-informative likelihood function of $\mathbf{x}$ can be written as

$$p_L(\mathbf{x}) \sim \mathcal{CN}\left(\mathbf{R}^\dagger\tilde{\mathbf{y}}, \sigma^2\left(\mathbf{R}^H\mathbf{R}\right)^{-1}\right).$$

The a priori distribution of $\mathbf{x}$ that describes the inactive PSs is $\prod_{m=1}^{N^t_{RF}} p_{\varepsilon'}(x_m)$. The joint Maximum-a-Posteriori (MAP) function is therefore written as

$$\mathcal{CN}\left(\mathbf{R}^\dagger\tilde{\mathbf{y}}, \sigma^2\left(\mathbf{R}^H\mathbf{R}\right)^{-1}\right) \prod_{m=1}^{N^t_{RF}} p_{\varepsilon'}(x_m). \tag{34}$$

There exists many efficient algorithms for solving the above optimization problem, such as approximate message passing [45], for example. Unfortunately, our simulations show that this type of configurations and algorithms fail to convey non-negligible energy efficiency. For this reason, our low-complexity Babai-AltMin algorithm is eminently suitable for mmWave hybrid TPC.

References

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

[2] I. A. Hemadeh, M. El-Hajjar, S. Won, and L. Hanzo, "Multiuser steered multiset space-time shift keying for millimeter-wave communications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5491–5495, Jun. 2017.

[3] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99878–99888, 2019.

[4] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[5] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[6] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[7] R. Zi, X. Ge, J. Thompson, C.-X. Wang, H. Wang, and T. Han, "Energy efficiency optimization of 5G radio frequency chain systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 758–771, Apr. 2016.

[8] Z. Gao, L. Dai, D. Mi, Z. Wang, M. A. Imran, and M. Z. Shakir, "MmWave massive-MIMO-based wireless backhaul for the 5G ultra-dense network," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 13–21, Oct. 2015.

[9] A. Kaushik, J. Thompson, E. Vlachos, C. Tsinos, and S. Chatzinotas, "Dynamic RF chain selection for energy efficient and low complexity hybrid beamforming in millimeter wave MIMO systems," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 886–900, Dec. 2019.

[10] L. Jiang and H. Jafarkhani, "Multi-user analog beamforming in millimeter wave MIMO systems based on path angle information," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 608–619, Jan. 2019.

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

[12] 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.

[13] M. Xiao *et al.*, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.

[14] X. Gao, L. Dai, and A. M. Sayeed, "Low RF-complexity technologies to enable millimeter-wave MIMO with large antenna array for 5G wireless communications," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 211–217, Apr. 2018.

[15] X. Yu, J. Zhang, and K. B. Letaief, "A hardware-efficient analog network structure for hybrid precoding in millimeter wave systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 282–297, May 2018.

[16] D. Zhang, A. Li, M. Shirvanimoghaddam, P. Cheng, Y. Li, and B. Vucetic, "Codebook-based training beam sequence design for millimeter-wave tracking systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5333–5349, Nov. 2019.

[17] Y. Sun *et al.*, "Principal component analysis-based broadband hybrid precoding for millimeter-wave massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6331–6346, Oct. 2020.

[18] J. Mao, Z. Gao, Y. Wu, and M.-S. Alouini, "Over-sampling codebook-based hybrid minimum sum-mean-square-error precoding for millimeter-wave 3D-MIMO," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 938–941, Dec. 2018.

[19] J.-C. Chen, "Hybrid beamforming with discrete phase shifters for millimeter-wave massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7604–7608, Aug. 2017.

[20] Z. Wang, M. Li, Q. Liu, and A. L. Swindlehurst, "Hybrid precoder and combiner design with low-resolution phase shifters in mmWave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 256–269, May 2018.

[21] A. N. Uwaechia, N. M. Mahyuddin, M. F. Ain, N. M. Abdul Latiff, and N. F. Za'bah, "On the spectral-efficiency of low-complexity and resolution hybrid precoding and combining transceivers for mmWave MIMO systems," *IEEE Access*, vol. 7, pp. 109259–109277, 2019.

[22] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.

[23] A. Li and C. Masouros, "Hybrid analog-digital millimeter-wave MU-MIMO transmission with virtual path selection," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 438–441, Feb. 2017.

[24] A. Li and C. Masouros, "Hybrid precoding and combining design for millimeter-wave multi-user MIMO based on SVD," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[25] F. Dong, W. Wang, and Z. Wei, "Low-complexity hybrid precoding for multi-user mmWave systems with low-resolution phase shifters," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 9774–9784, Oct. 2019.

[26] L. Dai, X. Gao, J. Quan, S. Han, and C.-L. I, "Near-optimal hybrid analog and digital precoding for downlink mmWave massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 1334–1339.

[27] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, "Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.

[28] X. Gao, L. Dai, Y. Sun, S. Han, and I. Chih-Lin, "Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[29] A. Kaushik, E. Vlachos, C. Tsinos, J. Thompson, and S. Chatzinotas, "Joint bit allocation and hybrid beamforming optimization for energy efficient millimeter wave MIMO systems," 2019, *arXiv:1910.01479*. [Online]. Available: http://arxiv.org/abs/1910.01479

[30] A. Kaushik, E. Vlachos, and J. Thompson, "Energy efficiency maximization of millimeter wave hybrid MIMO systems with low resolution DACs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6.

[31] R. Mendez-Rial, C. Rusu, N. Gonzalez-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.

[32] E. Vlachos, A. Kaushik, and J. Thompson, "Energy efficient transmitter with low resolution DACs for massive MIMO with partially connected hybrid architecture," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Porto, Portugal, Jun. 2018, pp. 1–5.

[33] 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.

[34] U. Erez and R. Zamir, "Achieving 1/2 log (1+SNR) on the AWGN channel with lattice encoding and decoding," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2293–2314, Oct. 2004.

[35] J. Zhan, B. Nazer, U. Erez, and M. Gastpar, "Integer-forcing linear receivers," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7661–7685, Dec. 2014.

[36] R. Y. Mesleh, H. Haas, S. Sinanovic, C. W. Ahn, and S. Yun, "Spatial modulation," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2228–2241, Jul. 2008.

[37] J. Choi, Y. Nam, and N. Lee, "Spatial lattice modulation for MIMO systems," *IEEE Trans. Signal Process.*, vol. 66, no. 12, pp. 3185–3198, Jun. 2018.

[38] L. Babai, "On Lovász' lattice reduction and the nearest lattice point problem," *Combinatorica*, vol. 6, no. 1, pp. 1–13, Mar. 1986.

[39] J. Wiktor Both, "On the rate of convergence of alternating minimization for non-smooth non-strongly convex optimization in banach spaces," 2019, *arXiv:1911.00404*. [Online]. Available: http://arxiv.org/abs/1911.00404

[40] A. Beck, "On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes," *SIAM J. Optim.*, vol. 25, no. 1, pp. 185–209, Jan. 2015.

[41] D. Micciancio and S. Goldwasser, *Complexity of lattice problems*. Boston, MA, USA: Springer, 2002.

[42] O. Ordentlich and U. Erez, "A simple proof for the existence of 'good' pairs of nested lattices," *IEEE Trans. Inf. Theory*, vol. 62, no. 8, pp. 4439–4453, Aug. 2016.

[43] R. Zamir, *Lattice Coding for Signals and Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[44] S. Lyu, J. Wen, J. Weng, and C. Ling, "On low-complexity lattice reduction algorithms for large-scale MIMO detection: The blessing of sequential reduction," *IEEE Trans. Signal Process.*, vol. 68, pp. 257–269, 2020.

[45] S. Lyu and C. Ling, "Hybrid vector perturbation precoding: The blessing of approximate message passing," *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 178–193, Jan. 2019.

[46] Z. Wang, S. Liu, and C. Ling, "Decoding by sampling—Part II: Derandomization and soft-output decoding," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4630–4639, Nov. 2013.

[47] C. Ling, "On the proximity factors of lattice reduction-aided decoding," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2795–2808, Jun. 2011.

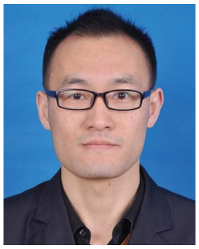
[48] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Computationally efficient vector perturbation precoding using thresholded optimization," *IEEE Trans. Commun.*, vol. 61, no. 5, pp. 1880–1890, May 2013.

[49] R. Méndez-Rial, C. Rusu, A. Alkhateeb, N. González-Prelcic, and R. W. Heath, “Channel estimation and hybrid combining for mmWave: Phase shifters or switches?” in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, 2015, pp. 90–97.

Shanxiang Lyu received the B.S. and M.S. degrees in electronic and information engineering from the South China University of Technology, Guangzhou, China, in 2011 and 2014, respectively, and the Ph.D. degree from the Electrical and Electronic Engineering Department, Imperial College London, U.K., in 2018. He is currently an Associate Professor with the College of Cyber Security, Jinan University. His research interests are in lattice theory, algebraic number theory, and their applications. He received the Superstar Supervisor Award of the National Crypto-Math Challenge of China in 2020.

Zheng Wang received the B.S. degree in electronic and information engineering from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2009, the M.S. degree in communications from the Department of Electrical and Electronic Engineering, The University of Manchester, Manchester, U.K., in 2010, and the Ph.D. degree in communication engineering from Imperial College London, U.K., in 2015. From 2015 to 2016, he served as a Research Associate for Imperial College London. From 2016 to 2017, he was a Senior Engineer with the Radio Access Network Research and Development Division, Huawei Technologies Company. From 2017 to 2020, he was an Associate Professor with the College of Electronic and Information Engineering, NUAA. His current research interests include MIMO systems, machine learning and data analytics over wireless networks, and lattice theory for wireless communications.

Zhen Gao received the B.S. degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2011, and the Ph.D. degree in communication and signal processing from the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China, in 2016. He is currently an Assistant Professor with the Beijing Institute of Technology. His research interest includes wireless communications, with a focus on multi-carrier modulations, multiple antenna systems, and sparse signal processing. He was a recipient of the IEEE Broadcast Technology Society 2016 Scott Helt Memorial Award (Best Paper), the Exemplary Reviewer of IEEE COMMUNICATION LETTERS in 2016, the *IET Electronics Letters* Premium Award (Best Paper) in 2016, and the Young Elite Scientists Sponsorship Program for the period of 2018–2021 from the China Association for Science and Technology.



Hongliang He (Member, IEEE) received the B.E. degree from Harbin Engineering University in 2013, and the Ph.D. degree in information and communication engineering from Xi’an Jiaotong University in 2019. He is currently an Assistant Professor with the College of Cyber Security, Jinan University, Guangzhou, China. His research interests include physical-layer security and cooperative communications.

Lajos Hanzo (Fellow, IEEE) received the master’s and Ph.D. degrees from the Technical University (TU) of Budapest in 1976 and 1983, respectively, and the D.Sc. degree from the University of Southampton in 2004. He has published more than 1900 contributions at IEEE Xplore and 19 Wiley-IEEE Press books. He has helped the fast-track career of 123 Ph.D. students and more than 40 of them are Professors at various stages of their careers in academia and many of them are leading scientists in the wireless industry. He is currently a Foreign Member of the Hungarian Academy of Sciences and a fellow of the Royal Academy of Engineering (REng.), IET, and EURASIP. He received the Honorary Doctorates from the TU of Budapest in 2009 and the University of Edinburgh in 2015. He has served several terms as the Governor for IEEE ComSoc and VTS. He was the former Editor-in-Chief of IEEE Press.