Machine Learning Inspired Energy-Efficient Hybrid Precoding for MmWave Massive MIMO Systems

Xinyu Gao*, Linglong Dai*, Ying Sun*, Shuangfeng Han†, and Chih-Lin I†
*Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Electronic Engineering, Tsinghua University, Beijing, China
†Green Communication Research Center, China Mobile Research Institute, Beijing 100053, China

***Abstract*—Hybrid precoding is a promising technique for mmWave massive MIMO systems, as it can considerably reduce the number of required radio-frequency (RF) chains without obvious performance loss. However, most of the existing hybrid precoding schemes require a complicated phase shifter network, which still involves high energy consumption. In this paper, we propose an energy-efficient hybrid precoding architecture, where the analog part is realized by a small number of switches and inverters instead of a large number of high-resolution phase shifters. Our analysis proves that the performance gap between the proposed hybrid precoding architecture and the traditional one is small and keeps constant when the number of antennas goes to infinity. Then, inspired by the cross-entropy (CE) optimization developed in machine learning, we propose an adaptive CE (ACE)-based hybrid precoding scheme for this new architecture. It aims to adaptively update the probability distributions of the elements in hybrid precoder by minimizing the CE, which can generate a solution close to the optimal one with a sufficiently high probability. Simulation results verify that our scheme can achieve the near-optimal sum-rate performance and much higher energy efficiency than traditional schemes.**

I. Introduction

Millimeter-wave (mmWave) massive multiple-input multiple-output (MIMO) has been considered as a promising technology for future 5G wireless communications [1], since it can provide wider bandwidth [2] and achieve higher spectral efficiency [3]. However, in MIMO systems, each antenna usually requires a dedicated radio-frequency (RF) chain (including high-resolution digital-to-analog converter, mixer, etc.) to realize the fully digital signal processing (e.g., precoding) [4]. For mmWave massive MIMO, this will result in unaffordable hardware complexity and energy consumption, as the number of antennas becomes huge and the energy consumption of RF chain is high [5]. To reduce the number of required RF chains, hybrid precoding has been recently proposed [6]. Its key idea is to decompose the fully digital precoder into a large-size analog beamformer (realized by the analog circuit) and a small-size digital precoder (requiring a small number of RF chains). Thanks to the low-rank characteristics of mmWave channels [2], a small-size digital precoder can achieve the spatial multiplexing gains, making hybrid precoding enjoy the near-optimal performance [5].

Nevertheless, most of the existing hybrid precoding schemes require a complicated phase shifter network, where each RF chain is connected to all antennas with high-resolution phase shifters [6], [7]. Although this architecture can provide high design freedom to achieve the near-optimal performance, it requires hundreds or even thousands of high-resolution phase shifters with high hardware cost and energy consumption [5]. To solve this problem, two categories of schemes have been proposed very recently. The first category is to directly employ finite-resolution phase shifters instead of high-resolution phase shifters [8], [9]. It can reduce the energy consumption of phase shifter network without obvious performance loss, but it still requires a large number of phase shifters with considerable energy consumption. The second category is to utilize the switch network to replace the phase shifter network [10]–[12]. It can significantly reduce the hardware cost and energy consumption, but it suffers from an obvious performance loss.

In this paper, we propose a switch and inverter (SI)-based hybrid precoding architecture with considerably reduced hardware cost and energy consumption. Instead of using phase shifters, the analog part of the proposed architecture is realized by a small number of energy-efficient switches and inverters. Then, we provide the performance analysis to quantify the performance gap between the proposed hybrid precoding architecture and the traditional ones. After that, inspired by the cross-entropy (CE) optimization developed in machine learning [13], we propose an adaptive CE (ACE)-based hybrid precoding scheme for this new architecture. Specifically, according to the probability distributions of the elements in hybrid precoder, this scheme first randomly generates several candidate hybrid precoders. Then, it adaptively weights these candidate hybrid precoders according to their achievable sum-rates, and refines the probability distributions of elements in hybrid precoder by minimizing the CE. Repeating such procedure, we can finally generate a hybrid precoder close to the optimal one with a sufficiently high probability. Simulation results verify that our scheme can achieve the near-optimal sum-rate performance and much higher energy efficiency than traditional schemes.

Notation: Lower- and upper-case boldface letters denote vectors and matrices, respectively; $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^{-1}$, $\mathrm{tr}\,(\cdot)$, and $\|\cdot\|_F$ denote the transpose, conjugate transpose, inversion, trace, and Frobenius norm of a matrix, respectively; $|\cdot|$ denotes the absolute operator; $\mathbb{E}(\cdot)$ denotes the expectation; $\otimes$ denotes the kronecker product; $\mathbf{I}_N$ is the $N \times N$ identity matrix.

II. System Model

In this paper, we consider a typical mmWave massive MIMO system, where the base station (BS) employs N antennas

and N_{RF} RF chains to simultaneously serve K single-antenna users (the extension to users with multiple-antennas is also possible as will be explained later). To fully achieve the multiplexing gains, we assume $N_{\mathrm{RF}} = K$ [9]. For hybrid precoding systems as shown in Fig. 1, the $K \times 1$ received signal vector $\mathbf{y}$ for all K users can be presented by

$$\mathbf{y} = \mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_K]^H$ is the channel matrix with $\mathbf{h}_k$ presenting the $N \times 1$ channel vector between the BS and the kth user, $\mathbf{s}$ is the $K \times 1$ transmitted signal vector for all K users satisfying $\mathbb{E}\left(\mathbf{s}\mathbf{s}^H\right) = \mathbf{I}_K$, $\mathbf{F}_{\mathrm{RF}}$ of size $N \times N_{\mathrm{RF}}$ is the analog beamformer realized by analog circuit (different architectures incur different hardware constraints as will be discussed later), $\mathbf{F}_{\mathrm{BB}}$ is the baseband digital precoder of size $N_{\mathrm{RF}} \times K$ satisfying the total transmit power constraint as $\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\|_F^2 = \rho$, where ρ is total transmit power. Finally, $\mathbf{n} \sim \mathcal{CN}\left(0, \sigma^2\mathbf{I}_K\right)$ of size $K \times 1$ is the additive white Gaussian noise (AWGN) vector, where σ^2 presents the noise power.

For the channel vector $\mathbf{h}_k$ of the kth user, we adopt the geometric channel model to capture the characteristics of mmWave massive MIMO channels as [5]

$$\mathbf{h}_k = \sqrt{\frac{N}{L_k}} \sum_{l=1}^{L_k} \alpha_k^{(l)} \mathbf{a}\left(\varphi_k^{(l)}, \theta_k^{(l)}\right), \tag{2}$$

where L_k denotes the number of paths for user k, $\alpha_k^{(l)}$ and $\varphi_k^{(l)}$ ($\theta_k^{(l)}$) for $1 \le l \le L_k$ are the complex gain and azimuth (elevation) angle of departure (AoD) of the path l for user k, $\mathbf{a}(\varphi, \theta)$ presents the $N \times 1$ array steering vector. For the typical uniform planar array (UPA) with N_1 elements in horizon and N_2 elements in vertical ($N = N_1 N_2$), we have [6]

$$\mathbf{a}(\varphi, \theta) = \mathbf{a}_{\mathrm{az}}(\varphi) \otimes \mathbf{a}_{\mathrm{el}}(\theta), \tag{3}$$

where $\mathbf{a}_{\mathrm{az}}(\varphi) = \frac{1}{\sqrt{N_1}}\left[e^{j2\pi i(d_1/\lambda)\sin\varphi}\right]^T$ for $i \in \mathcal{I}(N_1)$, $\mathbf{a}_{\mathrm{el}}(\theta) = \frac{1}{\sqrt{N_2}}\left[e^{j2\pi j(d_2/\lambda)\sin\theta}\right]^T$ for $j \in \mathcal{I}(N_2)$, $\mathcal{I}(n) = \{0, 1, \cdots, n-1\}$, λ is the signal wavelength, and d_1 (d_2) is the horizontal (vertical) antenna spacing. At mmWave frequencies, we usually have $d_1 = d_2 = \lambda/2$ [14].

III. Energy Efficient Hybrid Precoding

In this section, we first describe the proposed SI-based hybrid precoding architecture. Then, we propose an ACE-based hybrid precoding scheme for this new architecture. Finally, the complexity analysis is provided.

A. The proposed SI-based hybrid precoding architecture

Fig. 1 (a) and (b) show the traditional precoding architectures, i.e., the one with finite-resolution phase shifters (PS-based architecture) [9] and the one with switches (SW-based architecture) [11], respectively.

As shown in Fig. 1 (a), the traditional PS-based architecture requires a complicated phase shifter network, and the corresponding energy consumption can be presented as

$$P_{\mathrm{PS-architecture}} = \rho + N_{\mathrm{RF}}P_{\mathrm{RF}} + NN_{\mathrm{RF}}P_{\mathrm{PS}} + P_{\mathrm{BB}}, \tag{4}$$

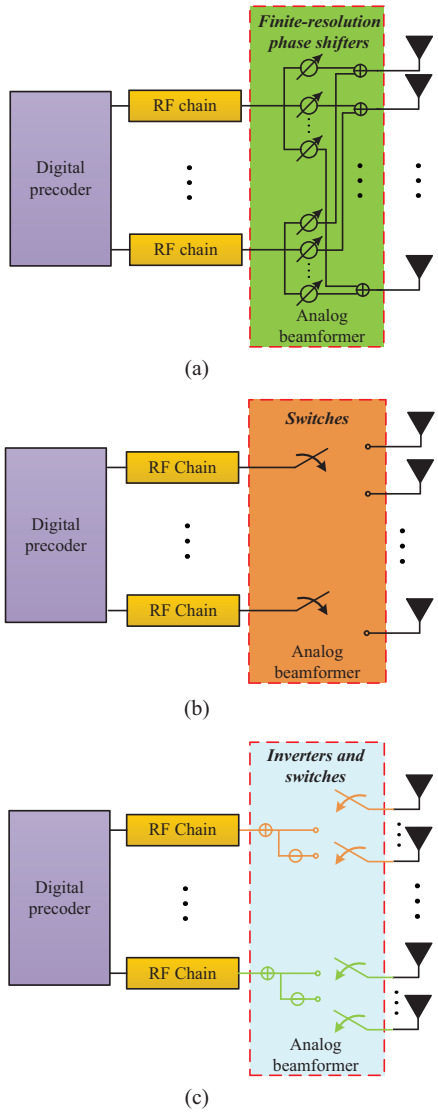

Fig. 1. Hybrid precoding: (a) traditional PS-based architecture; (b) traditional SW-based architecture; (c) proposed SI-based architecture.

where P_{RF}, P_{PS}, and P_{BB} are the energy consumption of RF chain, finite-resolution phase shifter, and baseband, respectively. Note that although the PS-based architecture enjoys high design freedom to achieve the near-optimal performance [9], it requires a large number (e.g., $N \times N_{\mathrm{RF}} = 64 \times 16 = 1024$ [9]) of phase shifters. Moreover, the energy consumption of finite phase shifter is also considerable (e.g., $P_{\mathrm{PS}} = 40\mathrm{mW}$ for 4-bit phase shifter [11]), These make the traditional PS-based architecture still suffer from high energy consumption [14].

By contrast, as shown in Fig. 1 (b), SW-based architecture can efficiently relieve such problem by employing a small number (N_{RF} instead of $N \times N_{\mathrm{RF}}$) of energy-efficient switches. The energy consumption of SW-based architecture can be presented as

$$P_{\mathrm{SW-architecture}} = \rho + N_{\mathrm{RF}}P_{\mathrm{RF}} + N_{\mathrm{RF}}P_{\mathrm{SW}} + P_{\mathrm{BB}}, \tag{5}$$

where P_{SW} is the energy consumption of switch, which is much lower than P_{PS} (e.g., $P_{\mathrm{SW}} = 5\mathrm{mW}$ [11]). Nevertheless,

since only N_{RF} antennas are active simultaneously, SW-based architecture cannot fully achieve the array gains of mmWave massive MIMO, leading to an obvious performance loss [15].

To overcome the problems faced by traditional architectures, we propose the SI-based architecture as shown in Fig. 1 (c), which can be considered as a better trade-off between the near-optimal PS-based architecture and the energy-efficient SW-based architecture. Specifically, in the proposed SI-based architecture, each RF chain is only connected to a sub antenna array with $M = N/N_{\mathrm{RF}}$ (assumed to be an integer) antennas instead of all N antennas [16]. Moreover, each RF chain is connected to the sub antenna array via only one inverter and M switches instead of N phase shifters. The energy consumption of SI-based architecture can be presented by

$$P_{\mathrm{SI-architecture}} = \rho + N_{\mathrm{RF}} P_{\mathrm{RF}} + N_{\mathrm{RF}} P_{\mathrm{IN}} + N P_{\mathrm{SW}} + P_{\mathrm{BB}}, \quad (6)$$

where P_{IN} is the energy consumption of inverter. It worth pointing out that the inverters can be realized by the digital chip with the energy consumption similar to switches (i.e., $P_{\mathrm{IN}} \approx P_{\mathrm{SW}}$) [15]. As a result, by comparing (4)-(6), we can conclude that the energy consumption of the proposed SI-based architecture is much lower than that of PS-based one. Furthermore, as all antennas are used, SI-based architecture can also achieve the potential array gains of mmWave massive MIMO, which will be further proved as follows.

To do this, we need to first explain the hardware constraints induced by the proposed SI-based architecture, which are different from those of the traditional ones. The first constraint is that the analog beamformer $\mathbf{F}_{\mathrm{RF}}$ should be a block diagonal matrix instead of a full matrix as

$$\mathbf{F}_{\mathrm{RF}} = \begin{bmatrix} \overline{\mathbf{f}}_1^{\mathrm{RF}} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{f}}_2^{\mathrm{RF}} & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \overline{\mathbf{f}}_{N_{\mathrm{RF}}}^{\mathrm{RF}} \end{bmatrix}_{N \times N_{\mathrm{RF}}}, \quad (7)$$

where $\overline{\mathbf{f}}_n^{\mathrm{RF}}$ is the $M \times 1$ analog beamformer on the nth sub antenna array. The second constraint is that since only inverters and switches are used, all the N nonzero elements of $\mathbf{F}_{\mathrm{RF}}$ should belong to

$$\frac{1}{\sqrt{N}} \{-1, +1\}. \quad (8)$$

Based on these constraints, we have the following **Lemma 1**.

Lemma 1. *Assume that the channel $\mathbf{h}_k$ of user k only has single path, i.e., $L_k = 1$ [12]. When $N \to \infty$ and $N/M = N_{\mathrm{RF}}$, the ratio ζ between the array gains achieved by SI-based architecture and that achieved by PS-based architecture with sufficiently high-resolution phase shifters can be presented by*

$$\lim_{N\to\infty, \frac{N}{M}=N_{\mathrm{RF}}} \zeta = \frac{4}{N_{\mathrm{RF}}\pi^2}. \quad (9)$$

Proof: For the traditional PS-based architecture with sufficiently high-resolution phase shifters, the phases of the elements in the analog beamformer can be arbitrarily adjusted to capture the power of $\mathbf{h}_k$. Therefore, the array gains achieved by PS-based architecture is $\left|\alpha_k^{(1)}\right|^2$. By contrast, the array gains achieved by SI-based architecture can be presented by

$$\begin{aligned} \left|\mathbf{h}_k^H \mathbf{f}_k^{\mathrm{RF}}\right|^2 &= N \left|\alpha_k^{(1)}\right|^2 \left|\mathbf{a}^H(\varphi_k)\, \mathbf{f}_k^{\mathrm{RF}}\right|^2 \quad (10) \\ &= \frac{1}{N}\left|\alpha_k^{(1)}\right|^2 \left|\sum_{m=1}^{M} e^{j\bar{\phi}_m}\right|^2 \\ &= \frac{1}{N}\left|\alpha_k^{(1)}\right|^2 \left(\left|\sum_{m=1}^{M} \cos\left(\bar{\phi}_m\right)\right|^2 + \left|\sum_{m=1}^{M} \sin\left(\bar{\phi}_m\right)\right|^2\right), \end{aligned}$$

where $\mathbf{f}_k^{\mathrm{RF}}$ is the kth column of $\mathbf{F}_{\mathrm{RF}}$ including the zeros and $\bar{\phi}_m$ denotes the phase quantization error. Since the nonzero elements in $\mathbf{f}_k^{\mathrm{RF}}$ belong to $\frac{1}{\sqrt{N}}\{-1, +1\}$, $\bar{\phi}_m$ can be assumed to follow the uniform distribution $\mathcal{U}(-\pi/2, \pi/2)$ for $1 \le m \le M$ [10]. Then, we have

$$\begin{aligned} \lim_{\substack{N\to\infty \\ N/M=N_{\mathrm{RF}}}} \zeta &= \frac{1}{N_{\mathrm{RF}} M}\left\{\left|\sum_{m=1}^{M} \cos\left(\bar{\phi}_m\right)\right|^2 + \left|\sum_{m=1}^{M} \sin\left(\bar{\phi}_m\right)\right|^2\right\} \\ &= \frac{1}{N_{\mathrm{RF}}}\left(\mathbb{E}\left[\cos\left(\bar{\phi}_m\right)\right]\right)^2 + \left(\mathbb{E}\left[\sin\left(\bar{\phi}_m\right)\right]\right)^2 \\ &= \frac{4}{N_{\mathrm{RF}}\pi^2}, \quad (11) \end{aligned}$$

which verifies the conclusion in **Lemma 1**. ■

Lemma 1 indicates that although the proposed SI-based architecture suffers from some loss of array gains compared to the near-optimal PS-based architecture, the performance loss keeps constant and limited, which does not grow as the number of BS antennas goes to infinity. Recalling the low energy consumption of SI-based architecture, we can conclude that our scheme is a better trade-off between the traditional architectures, which will be also verified by simulation. Next, we will design a near-optimal hybrid precoding scheme for SI-based architecture with quite different hardware constraints.

B. ACE-based hybrid precoding scheme

We aim to design the analog beamformer $\mathbf{F}_{\mathrm{RF}}$ and the digital precoder $\mathbf{F}_{\mathrm{BB}}$ to maximize the achievable sum-rate R, which can be presented as

$$\begin{aligned} \left(\mathbf{F}_{\mathrm{RF}}^{\mathrm{opt}}, \mathbf{F}_{\mathrm{BB}}^{\mathrm{opt}}\right) &= \arg\max_{\mathbf{F}_{\mathrm{RF}}, \mathbf{F}_{\mathrm{BB}}} R, \\ &\text{s.t.} \quad \mathbf{F}_{\mathrm{RF}} \in \mathcal{F}, \\ &\quad \left\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\right\|_F^2 = \rho, \end{aligned} \quad (12)$$

where $\mathcal{F}$ denotes the set with all possible analog beamformers satisfying the two constraints (7) and (8) described above, and the achievable sum-rate R can be presented by

$$R = \sum_{k=1}^{K} \log_2(1 + \gamma_k), \quad (13)$$

where γ_k presents the signal-to-interference-plus-noise ratio (SINR) of the kth user as

$$\gamma_k = \frac{\left|\mathbf{h}_k^H \mathbf{F}_{\mathrm{RF}} \mathbf{f}_k^{\mathrm{BB}}\right|^2}{\sum_{k' \neq k}^{K} \left|\mathbf{h}_k^H \mathbf{F}_{\mathrm{RF}} \mathbf{f}_{k'}^{\mathrm{BB}}\right|^2 + \sigma^2}, \quad (14)$$

where $\mathbf{f}_k^{\mathrm{BB}}$ is the kth column of $\mathbf{F}_{\mathrm{BB}}$.

It is worth pointing out that the constraints (7) and (8) on the analog beamformer $\mathbf{F}_{\mathrm{RF}}$ are non-convex. This makes (12) very difficult to be solved. Fortunately, as all the N nonzero elements of $\mathbf{F}_{\mathrm{RF}}$ belong to the set $\frac{1}{\sqrt{N}}\{-1,+1\}$, the number of possible $\mathbf{F}_{\mathrm{RF}}$ is finite. Therefore, (12) can be regarded as a non-coherent combining problem [13]. To solve it, we can first select a candidate $\mathbf{F}_{\mathrm{RF}}$, and compute the optimal $\mathbf{F}_{\mathrm{BB}}$ according to the effective channel matrix $\mathbf{HF}_{\mathrm{RF}}$ without non-convex constraints. After all possible $\mathbf{F}_{\mathrm{RF}}$'s have been searched, we can obtain the optimal analog beamformer $\mathbf{F}_{\mathrm{RF}}^{\mathrm{opt}}$ and digital precoder $\mathbf{F}_{\mathrm{BB}}^{\mathrm{opt}}$. However, such exhaustive search scheme requires to search 2^N possible $\mathbf{F}_{\mathrm{RF}}$'s and $\mathbf{F}_{\mathrm{BB}}$'s, which involves unaffordable complexity as N is usually large in mmWave massive MIMO systems (e.g., $N=64$, $2^{64}\approx 1.84\times 10^{19}$). To solve this problem, we propose an ACE algorithm, which can be considered as an improved version of the CE algorithm developed from machine learning [13].

At first, we would like to briefly introduce the conventional CE algorithm, which is a probabilistic model-based algorithm to solve the combining problem by an iterative procedure. In each iteration, the CE algorithm first generates S candidates (e.g., possible hybrid precoders in our problem) according to a probability distribution. Then, it computes the objective value (e.g., achievable sum-rate in our problem) of each candidate, and selects S_{elite} best candidates as "elite" [13]. Finally, based on the selected elites, the probability distribution is updated by minimizing the CE. Repeating such procedure, the probability distribution will be refined to generate a solution close to the optimal one with a sufficiently high probability. However, although the CE algorithm has been widely used in machine learning [13], it still has some disadvantages. One is that the contributions of all elites are treated as the same. Intuitively, the elite with better objective value should be more important when we update the probability distribution. Therefore, if we can adaptively weight the elites according to their objective values, better performance can be expected. Following this idea, we propose an ACE algorithm to solve (12).

The pseudo-code of the proposed ACE-based hybrid precoding scheme[1] is summarized in **Algorithm 1**, which can be explained as follows. At the beginning, we formulate the nonzero elements in $\mathbf{F}_{\mathrm{RF}}$ as $N\times 1$ vector $\mathbf{f}=\left[\left(\bar{\mathbf{f}}_1^{\mathrm{RF}}\right)^T,\left(\bar{\mathbf{f}}_2^{\mathrm{RF}}\right)^T,\cdots\left(\bar{\mathbf{f}}_{N_{\mathrm{RF}}}^{\mathrm{RF}}\right)^T\right]^T$, and set the probability parameter $\mathbf{u}=[u_1,u_2,\cdots,u_N]^T$ as an $N\times 1$ vector, where $0\le u_n\le 1$ presents the probability that $f_n=1/\sqrt{N}$, f_n is the nth element of $\mathbf{f}$. Then, by initializing $\mathbf{u}^{(0)}=\frac{1}{2}\times\mathbf{1}_{N\times 1}$ ($\mathbf{1}$ is the all-one vector), we assume that all the N nonzero elements of $\mathbf{F}_{\mathrm{RF}}$ belong to $\frac{1}{\sqrt{N}}\{-1,+1\}$ with equal probability, since no priori information is available. During the ith iteration, in step 1, we first generate S candidate analog beamformers $\{\mathbf{F}_{\mathrm{RF}}^s\}_{s=1}^S$ based on the probability distribution $\Xi\left(\mathcal{F};\mathbf{u}^{(i)}\right)$ (i.e., generate $\{\mathbf{f}^s\}_{s=1}^S$ according to $\mathbf{u}^{(i)}$, and reshape them as matrices belong to $\mathcal{F}$). Then, in step 2, we calculate the corresponding digital precoder $\mathbf{F}_{\mathrm{BB}}^s$ according to the effective channel $\mathbf{H}_{\mathrm{eq}}^s=\mathbf{HF}_{\mathrm{RF}}^s$ for $1\le s\le S$. Note that there are lots of advanced digital precoder schemes [3]. In this paper, we adopt the classical ZF digital precoder with the near-optimal performance and low complexity as the example [3], and $\mathbf{F}_{\mathrm{BB}}^s$ can be computed as

$$\mathbf{G}^s=\left(\mathbf{H}_{\mathrm{eq}}^s\right)^H\left(\mathbf{H}_{\mathrm{eq}}^s\left(\mathbf{H}_{\mathrm{eq}}^s\right)^H\right)^{-1},\quad \mathbf{F}_{\mathrm{BB}}^s=\beta^s\mathbf{G}^s, \tag{15}$$

where $\beta^s=\sqrt{\rho}/\|\mathbf{F}_{\mathrm{RF}}^s\mathbf{G}^s\|_F$ is power normalized factor.

After that, in step 3, the achievable sum-rate $\{R(\mathbf{F}_{\mathrm{RF}}^s)\}_{s=1}^S$ are calculated by substituting $\mathbf{F}_{\mathrm{RF}}^s$ and $\mathbf{F}_{\mathrm{BB}}^s$ (also a function of $\mathbf{F}_{\mathrm{RF}}^s$) into (13). We sort $\{R(\mathbf{F}_{\mathrm{RF}}^s)\}_{s=1}^S$ in a descend order in step 4. Then, the elites can be obtained in step 5. In the conventional CE algorithm, the next step is to using elites to update $\mathbf{u}^{(i+1)}$ by minimizing CE, which is equivalent to [13]

$$\mathbf{u}^{(i+1)}=\arg\max_{\mathbf{u}^{(i)}}\frac{1}{S}\sum_{s=1}^{S_{\mathrm{elite}}}\ln\Xi\left(\mathbf{F}_{\mathrm{RF}}^{[s]};\mathbf{u}^{(i)}\right), \tag{16}$$

where $\Xi\left(\mathbf{F}_{\mathrm{RF}}^{[s]};\mathbf{u}^{(i)}\right)$ denotes the probability to generate $\mathbf{F}_{\mathrm{RF}}^{[s]}$. As mentioned above, in (16), the contributions of all elites are treated as the same, leading to performance degradation. To solve this problem, we propose to weight each elite adaptively based on its achievable sum-rate. Specifically, we first define an auxiliary parameter T presenting the average achievable sum-rate of all elites as

$$T=\frac{1}{S_{\mathrm{elite}}}\sum\nolimits_{s=1}^{S_{\mathrm{elite}}}R\left(\mathbf{F}_{\mathrm{RF}}^{[s]}\right). \tag{17}$$

Then, the weight w_s of the elite $\mathbf{F}_{\mathrm{RF}}^{[s]}$ can be calculated in step 6 as $w_s=R\left(\mathbf{F}_{\mathrm{RF}}^{[s]}\right)/T$. Based on $\{w_s\}_{s=1}^{S_{\mathrm{elite}}}$, (16) can be modified as

$$\mathbf{u}^{(i+1)}=\arg\max_{\mathbf{u}^{(i)}}\frac{1}{S}\sum_{s=1}^{S_{\mathrm{elite}}}w_s\ln\Xi\left(\mathbf{F}_{\mathrm{RF}}^{[s]};\mathbf{u}^{(i)}\right). \tag{18}$$

Input: Channel matrix $\mathbf{H}$; Number of iterations I; Number of candidates S; Number of elites S_{elite}.
Initialization: $i=0$; $\mathbf{u}^{(0)}=\frac{1}{2}\times\mathbf{1}_{N\times 1}$.
for $0\le i\le I$
1. Randomly generate S candidate analog beamformers $\{\mathbf{F}_{\mathrm{RF}}^s\}_{s=1}^S$ based on $\Xi\left(\mathcal{F};\mathbf{u}^{(i)}\right)$;
2. Compute S corresponding digital precoders $\{\mathbf{F}_{\mathrm{BB}}^s\}_{s=1}^S$ based on the effective channel $\mathbf{H}_{\mathrm{eq}}^s=\mathbf{HF}_{\mathrm{RF}}^s$;
3. Calculate the achievable sum-rate $\{R(\mathbf{F}_{\mathrm{RF}}^s)\}_{s=1}^S$ (13);
4. Sort $\{R(\mathbf{F}_{\mathrm{RF}}^s)\}_{s=1}^S$ in a descend order as $R\left(\mathbf{F}_{\mathrm{RF}}^{[1]}\right)\ge R\left(\mathbf{F}_{\mathrm{RF}}^{[2]}\right)\ge\cdots\ge R\left(\mathbf{F}_{\mathrm{RF}}^{[S]}\right)$;
5. Select elites as $\mathbf{F}_{\mathrm{RF}}^{[1]},\mathbf{F}_{\mathrm{RF}}^{[2]},\cdots,\mathbf{F}_{\mathrm{RF}}^{[S_{\mathrm{elite}}]}$;
6. Calculate weight w_s for each elite $\mathbf{F}_{\mathrm{RF}}^{[s]}$, $1\le s\le S_{\mathrm{elite}}$;
7. Update $\mathbf{u}^{(i+1)}$ according to $\{w_s\}_{s=1}^{S_{\mathrm{elite}}}$ and $\left\{\mathbf{F}_{\mathrm{RF}}^{[s]}\right\}_{s=1}^{S_{\mathrm{elite}}}$;
8. $i=i+1$;
end for
Output: Analog beamformer $\mathbf{F}_{\mathrm{RF}}^{[1]}$; Digital precoder $\mathbf{F}_{\mathrm{BB}}^{[1]}$.

Algorithm 1: The proposed ACE-based hybrid precoding

[1]Note that the convergence of the proposed ACE-based hybrid precoding scheme can be proved by extending the Theorem 1 in [17].

Note that $\Xi\left(\mathbf{F}_{\mathrm{RF}}^{[s]};\mathbf{u}^{(i)}\right)=\Xi\left(\mathbf{f}^{[s]};\mathbf{u}^{(i)}\right)$, and the nth element $f_n^{[s]}$ of $\mathbf{f}^{[s]}$ is a Bernoulli random variable, where $f_n^{[s]}=1/\sqrt{N}$ with probability $u_n^{(i)}$ and $f_n^{[s]}=-1/\sqrt{N}$ with probability $1-u_n^{(i)}$. Therefore, we have

$$\Xi\left(\mathbf{F}_{\mathrm{RF}}^{[s]};\mathbf{u}^{(i)}\right)=\prod_{n=1}^{N}\left(u_n^{(i)}\right)^{\frac{1}{2}\left(1+\sqrt{N}f_n^{[s]}\right)}\left(1-u_n^{(i)}\right)^{\frac{1}{2}\left(1-\sqrt{N}f_n^{[s]}\right)}. \tag{19}$$

After substituting (19) into (18), the first derivative of the target in (18) with respect to $u_n^{(i)}$ can be derived as

$$\frac{1}{S}\sum_{s=1}^{S_{\mathrm{elite}}}w_s\left(\frac{1+\sqrt{N}f_n^{[s]}}{2u_n^{(i)}}-\frac{1-\sqrt{N}f_n^{[s]}}{2\left(1-u_n^{(i)}\right)}\right). \tag{20}$$

Setting (20) to zero, $\mathbf{u}^{(i+1)}$ can be updated in step 7 as

$$u_n^{(i+1)}=\frac{\sum_{s=1}^{S_{\mathrm{elite}}}w_s\left(\sqrt{N}f_n^{[s]}+1\right)}{2\sum_{s=1}^{S_{\mathrm{elite}}}w_s}. \tag{21}$$

Such procedure above will be repeated ($i=i+1$ in step 8) until the maximum number of iterations I is reached, and the analog beamformer and digital precoder will be selected as $\mathbf{F}_{\mathrm{RF}}^{[1]}$ and $\mathbf{F}_{\mathrm{BB}}^{[1]}$, respectively. Finally, it is worth pointing out that the proposed ACE-based hybrid precoding scheme can be also extended to the scenario where users employ multiple antennas. In this case, the analog beamformers at the BS and users should be jointly searched by the ACE algorithm.

C. Computational complexity analysis

In this subsection, the computational complexity of the proposed ACE-based hybrid precoding scheme is discussed.

From **Algorithm 1**, we can observe that the complexity of the ACE-based hybrid precoding scheme mainly comes from steps 2, 3, 6, and 7. In step 2, we need to compute S effective channel matrices $\left\{\mathbf{H}_{\mathrm{eq}}^{s}\right\}_{s=1}^{S}$ and digital precoders $\left\{\mathbf{F}_{\mathrm{BB}}^{s}\right\}_{s=1}^{S}$ according to (15). Therefore, the complexity of this part is $\mathcal{O}\left(SNK^2\right)$. In step 3, the achievable sum-rate of each candidate is computed. Since we employ the digital ZF precoder, the SINR γ_k^s of the kth user for the sth candidate is simplified to $\gamma_k^s=\left(\beta^s/\sigma\right)^2$. As a result, this part only involves the complexity $\mathcal{O}\left(S\right)$. In step 6, we calculate S_{elite} weights based on (17), which is quite simple with the complexity $\mathcal{O}\left(S_{\mathrm{elite}}\right)$. Finally, in step 7, the probability parameter $\mathbf{u}^{(i+1)}$ is updated according to (21) with the complexity $\mathcal{O}\left(NS_{\mathrm{elite}}\right)$.

In summary, after I iterations, the total computational complexity of the proposed ACE hybrid precoding scheme is $\mathcal{O}\left(ISNK^2\right)$. Since K is usually small, I and S also do not have to be very large as will be verified in the next section, we can conclude that the complexity of the proposed ACE-based hybrid precoding scheme is acceptable, which is comparable with the simple least squares (LS) algorithm.

IV. Simulation Results

In this section, we provide the simulation results in terms of achievable sum-rate and energy-efficiency to evaluate the performance of the proposed ACE-based hybrid precoding scheme. The simulation parameters are described as follows: We assume that the BS employs an UPA with antenna spacing $d_1=d_2=\lambda/2$. For the kth user, we generate the channel $\mathbf{h}_k$ based on (2), where we assume: 1) $L_k=3$; 2) $\alpha_k^{(l)}\sim\mathcal{CN}\left(0,1\right)$ for $1\le l\le L_k$; 3) $\varphi_k^{(l)}$ and $\theta_k^{(l)}$ follow the uniform distribution $\mathcal{U}\left(-\pi,\pi\right)$ for $1\le l\le L_k$ [12]. Finally, the signal-to-noise ratio (SNR) is defined as ρ/σ^2.

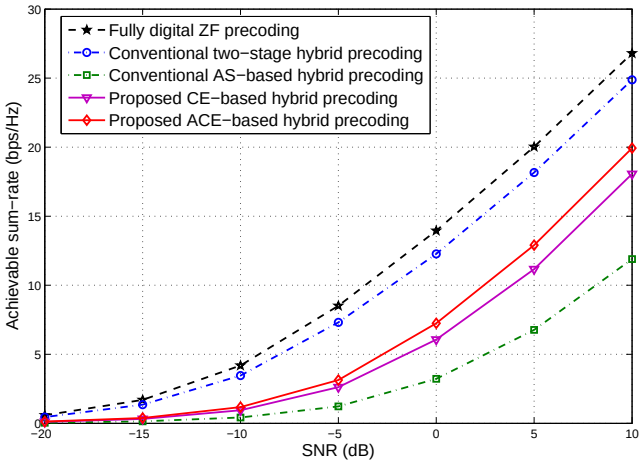

Fig. 2. Achievable sum-rate comparison.

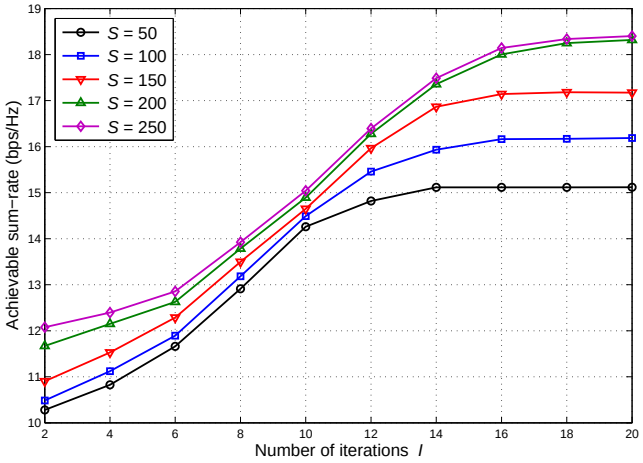

Fig. 3. Achievable sum-rate against S and I.

Fig. 2 shows the achievable sum-rate comparison in a typical mmWave massive MIMO system with $N=64$, $N_{\mathrm{RF}}=K=4$. In Fig. 2, the proposed CE-based (i.e., using the conventional CE algorithm to solve (12)) and ACE-based hybrid precoding schemes are designed for SI-based architecture, where we set $S=200$, $S_{\mathrm{elite}}=40$, and $I=20$ for **Algorithm 1**, the conventional two-stage hybrid precoding scheme is designed for PS-based architecture with 4-bit phase shifters [9], and the conventional antenna selection (AS)-based hybrid precoding scheme is designed for SW-based architecture [11] with switches. Firstly, we can observe from Fig. 2 that the proposed ACE algorithm outperforms the traditional CE algorithm, where the SNR gap is about 1 dB. Note that the ACE algorithm only involves one additional step (i.e., step 6 in **Algorithm 1**) with negligible complexity. Therefore, the proposed ACE algorithm is more efficient. Moreover, Fig. 2 also shows that the proposed ACE-based hybrid precoding can

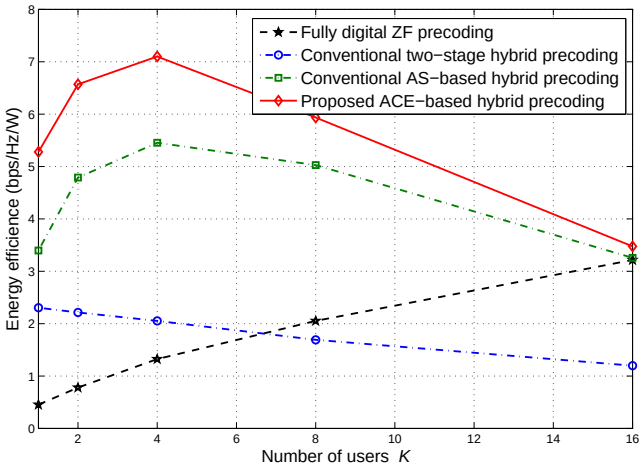

Fig. 4. Energy efficiency comparison.

achieve the sum-rate much higher than the conventional AS-based hybrid precoding, as it can achieve the potential array gains in mmWave massive MIMO systems. Finally, we can observe that the performance gap between ACE-based hybrid precoding and two-stage hybrid precoding is limited and keeps constant, which further verifies the conclusion in **Lemma 1**.

Fig. 3 shows the achievable sum-rate of the proposed ACE-based hybrid precoding against the number of candidates S and the number of iterations I, when $S_{\text{elite}}/S = 0.2$, $N = 64$, $N_{\text{RF}} = K = 4$, and SNR = 10 dB. From Fig. 3, we can observe that when S is small, increasing S will lead to an obvious improvement in the sum-rate performance. However, when S is sufficiently large, such trend is no longer obvious. This indicates that the number of candidates S does not have to be very large, e.g., $S = 200$ is enough. Furthermore, Fig. 3 also indicates that the proposed ACE-based hybrid precoding can converge with a small number of iterations, e.g., $I = 20$. These observations verify the rationality of the parameters for **Algorithm 1** we used in Fig. 2.

Fig. 4 shows the energy efficiency comparison when $N = 64$ is fixed and $N_{\text{RF}} = K$ varies from 1 to 16. The parameters for **Algorithm 1** are the same as Fig. 2. According to [15], [16], the energy efficiency can be defined as the ratio between the achievable sum-rate and the energy consumption, which should be (4), (5), and (6) for two-stage hybrid precoding, AS-based hybrid precoding, and ACE-based hybrid precoding, respectively. In this paper, we adopt the practical values $\rho = 30\text{mW}$ [16], $P_{\text{RF}} = 300\text{mW}$ [16], $P_{\text{BB}} = 200\text{mW}$ [15], $P_{\text{PS}} = 40\text{mW}$ (4-bit phase shifter) [15], and $P_{\text{SW}} = P_{\text{IN}} = 5\text{mW}$ [15]. From Fig. 4, we can observe that the proposed ACE-based hybrid precoding with SI-based architecture can achieve much higher energy efficiency than the others, especially when K is not very large (e.g., $K \leq 12$). Furthermore, it is interesting to observe that when $K \geq 8$, the energy efficiency of the two-stage hybrid precoding with PS-based architecture is even lower than that of the fully digital ZF precoding. This is due to the fact that as K grows, the number of phase shifters in PS-based architecture increases rapidly. As a result, the energy consumption of the phase shifter network will be huge, even higher than that of RF chains.

V. Conclusions

In this paper, we propose an energy-efficient SI-based hybrid precoding architecture, where the analog part is realized by a small number of switches and inverters. The performance analysis proves that the performance gap between the proposed SI-based architecture and the traditional near-optimal one keeps constant and limited. Then, by employing the idea of CE optimization in machine learning, we further propose an ACE-based hybrid precoding scheme with low complexity for SI-based architecture. Simulation results verify that our scheme can achieve the satisfying sum-rate performance and much higher energy efficiency than traditional schemes.

References

[1] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: The next wireless revolution?" *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 56–62, Sep. 2014.

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

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[4] H. Xie, F. Gao, S. Zhang, and S. Jin, "A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model," to appear in *IEEE Trans. Veh. Technol.*, 2017.

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

[6] O. El 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.

[7] H. Xie, F. Gao, and S. Jin, "An overview of low-rank channel estimation for massive MIMO systems," *IEEE Access*, vol. 4, pp. 7313–7321, Nov. 2016.

[8] F. Sohrabi and W. Yu, "Hybrid beamforming with finite-resolution phase shifters for large-scale MIMO systems," in *Proc. IEEE SPAWC Workshops*, Jul. 2015, pp. 136–140.

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

[10] A. Alkhateeb, Y.-H. Nam, J. Zhang, and R. W. Heath, "Massive MIMO combining with switches," *IEEE Wireless Commun. Lett.*, vol. 5, no. 3, pp. 232–235, Jun. 2016.

[11] R. Mendez-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. ITA Workshops*, Feb. 2015, pp. 90–97.

[12] A. Sayeed and J. Brady, "Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 3679–3684.

[13] R. Y. Rubinstein and D. P. Kroese, *The cross-entropy method: A unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media, 2013.

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

[15] R. Méndez-Rial, C. Rusu, N. González-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, Jan. 2016.

[16] 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.

[17] J.-C. Chen, C.-K. Wen, and K.-K. Wong, "An efficient sensor node selection algorithm for sidelobe control in collaborative beamforming," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 5984–5994, Aug. 2016.