Energy Efficiency Maximization Precoding for Quantized Massive MIMO Systems

Jinseok Choi, *Member, IEEE*, Jeonghun Park, *Member, IEEE*, and Namyoon Lee, *Senior Member, IEEE*

***Abstract*— The use of low-resolution digital-to-analog and analog-to-digital converters (DACs and ADCs) significantly benefits energy efficiency (EE) at the cost of high quantization noise for massive multiple-input multiple-output (MIMO) systems. This paper considers a precoding optimization problem for maximizing EE in quantized downlink massive MIMO systems. To this end, we jointly optimize an active antenna set, precoding vectors, and allocated power; yet acquiring such joint optimal solution is challenging. To resolve this challenge, we decompose the problem into precoding direction and power optimization problems. For precoding direction, we characterize the first-order optimality condition, which entails the effects of quantization distortion and antenna selection. We cast the derived condition as a functional eigenvalue problem, wherein finding the principal eigenvector attains the best local optimal point. To this end, we propose generalized power iteration based algorithm. To optimize precoding power for given precoding direction, we adopt a gradient descent algorithm for the EE maximization. Alternating these two methods, our algorithm identifies a joint solution of the active antenna set, the precoding direction, and allocated power. In simulations, the proposed methods provide considerable performance gains. Our results suggest that a few-bit DACs are sufficient for achieving high EE in massive MIMO systems.**

***Index Terms*— Low-resolution analog-to-digital converter (ADC)/digital-to-analog converter (DAC), energy efficiency (EE), precoding, antenna selection, massive multiple-input multiple-output (MIMO), eigenvalue problem.**

I. Introduction

MASSIVE multiple-input multiple-output (MIMO) [1] is a key enabler for future cellular systems because it can provide substantial gains in both spectral efficiency (SE) and coverage by employing large-scale antenna arrays at a base station (BS). In principle, massive MIMO increases the SE gain by scaling antennas at the BS under ideal conditions. Unfortunately, in practice, the use of very large-antenna elements makes the BS hardware and radio frequency (RF) circuit architectures intricate and also gives rise to excessive energy consumption in the BS. Since the energy consumption of quantizers (digital-to-analog converters (DACs) and analog-to-digital converters (ADCs)) exponentially increases with its resolution bits, using low-resolution quantizers can alleviate the excessive energy consumption in massive MIMO systems [2]. Motivated by this, implementing massive MIMO with low-resolution DACs and ADCs are rapidly gaining momentum [3]–[5].

The use of the low-resolution quantizers in transmission and reception causes severe quantization error. For example, in the downlink MIMO transmission using low-resolution DACs, the transmitted signals are distorted by the low-resolution DACs, and the quantization error incurs a significant amount of inter-user interference. This fundamentally limits the SE gains in massive MIMO. As a result, it is crucial to incorporate the quantization effects in designing a downlink transmission strategy to maximize communication performance such as SE and energy efficiency (EE) in the massive MIMO using low-resolution quantizers. In particular, design for energy efficient communications is critical in realizing massive MIMO systems. However, finding a precoding solution for maximizing the EE under low-resolution quantizers constraints is highly challenging. The challenge involves the non-convexity of the EE function, defined as the sum SE normalized by the total transmission power. Further, since the total transmission power is a function of an active antenna set, it entails a non-smooth function. This non-smooth part makes the optimization problem more difficult to solve. In this paper, we make progress toward finding a sub-optimal solution that jointly identifies a set of active antenna elements and the corresponding linear precoding vectors to maximize EE in downlink quantized massive MIMO systems.

A. Prior Works

In the literature, linear precoding methods for maximizing EE have been widely studied in high-resolution quantizer setups. This includes a study of the EE of massive MIMO uplink systems with traditional linear receive beamforming such as zero-forcing (ZF) and minimum mean-square-error (MMSE) [6], the capacity limit in massive MIMO with non-ideal hardware impacts [7], joint antenna selection and

Manuscript received 6 August 2021; revised 23 December 2021 and 8 February 2022; accepted 9 February 2022. Date of publication 28 February 2022; date of current version 12 September 2022. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government, Ministry of Science and ICT (MSIT), South Korea, under Grant 2019R1G1A1094703, Grant 2021R1C1C1004438, and Grant 2020R1C1C1013381; and in part by MSIT under the Information Technology Research Center (ITRC) Support Program, supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2021-2017-0-01635. The associate editor coordinating the review of this article and approving it for publication was C.-K. Wen. *(Corresponding authors: Jeonghun Park; Namyoon Lee.)*

Jinseok Choi is with the Department of Electrical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, South Korea (e-mail: jinseokchoi@unist.ac.kr).

Jeonghun Park is with the School of Electronics Engineering, College of IT Engineering, Kyungpook National University, Daegu 41566, South Korea (e-mail: jeonghun.park@knu.ac.kr).

Namyoon Lee is with the School of Electrical Engineering, Korea University, Seoul 02841, South Korea (e-mail: namyoon@korea.ac.kr).

Digital Object Identifier 10.1109/TWC.2022.3152491

precoding design using semidefinite programming (SDP) with successive convex approximation (SCA) [8]. A comprehensive survey on energy-efficient design is found in [9].

Despite the abundant previous studies, the aforementioned prior work is not applicable when low-resolution quantizers are used. This is because low-resolution quantizers induce non-negligible non-linear quantization distortion, which may render the performance characterization totally different from a conventional high-resolution system. Motivated by this, there exist several prior works that performed performance analysis for low-resolution quantizers. Especially, thanks to the analytical tractability of 1-bit quantizers, there have been rich studies regarding design and analysis of 1-bit quantizers. In [10]–[12], the achievable rate of 1-bit ADCs was studied. In [13], it was revealed that higher-rank transmit covariance matrices can improve the sum spectral efficiency even with 1-bit quantizer. In [14], [15], a MIMO system with one-bit Sigma-Delta ($\Sigma\Delta$) quantizers is presented. In [16], [17], constant-envelope precoding methods were developed by harnessing the fact that a 1-bit quantizer only can extract the signal phase information.

More relevant to this work, several prior works studied precoding methods in downlink systems with low-resolution quantizers. A main obstacle in developing such precoders is that non-linear quantization distortion is not tractable. To handle this in a tractable way, in [18], the non-linearity of the quantization distortion was resolved by adopting Bussgang theorem [19], then by using the linearization, a closed-form expression of a MMSE precoder was derived. In [4], a linear precoder for 3 to 4-bit DACs and a non-linear precoder for 1-bit DACs were proposed. Especially, in designing the linear precoder, it adopted a conventional linear precoder such as MMSE or ZF, and quantized the adopted linear precoder to use with low-resolution DACs. A key finding of [4] is that using 3 to 4-bit DACs offers comparable performance with high-resolution DACs, provided that proper design of precoding is used. This finding is supported by the SE and bit-error-rate (BER) analyses in [5]. A similar approach was also used in [20], where conventional ZF precoding is applied, thereafter 1-bit quantization is conducted to the precoded signal. The approach used in [4] was further improved by using alternating minimization in [21]. For precoding design in more general bits quantizers, an additive quantization noise model (AQNM) was also used in several works [22]–[25], providing a linear approximation of a quantized signal. Employing the AQNM, [22] considered hybrid precoding architecture with low-resolution DACs in a point-to-point MIMO channel and performed extensive performance evaluation by using a conventional precoder. In [24], [25], an algorithm that selects an active antenna set was proposed for a given precoder. Without linear modeling, the alternating direction method of multipliers (ADMM) was used to solve an inter-user interference minimization problem in [26]. Based on this result, a general precoder was designed.

As explained above, abounding studies provided crucial insights on low-resolution quantizer systems. The existing precoding methods, however, are mostly limited to variants of traditional linear precoding methods such as ZF or MMSE. Specifically, the traditional linear precoders are firstly adopted, then quantization distortion effects are reflected. Nevertheless, the traditional precoders have limited SE performance due to the quantization distortion, and do not care to save energy consumption. These two features lead to mediocre EE performance.

Beyond traditional linear precoders, prior work such as [25] developed an antenna selection strategy to maximize the EE. Nonetheless, the approach in [25] is limited in that only antenna selection is carefully designed while fixing precoders. Since precoder design significantly affects the energy consumption of the BS, it is inefficient compared to the joint design of antenna selection and precoders. Besides downlink, the prior works [22], [27] tackled EE maximization in point-to-point massive MIMO systems. Notwithstanding, their design principles are mostly based on singular-value decomposition (SVD); therefore, this cannot be extended to downlink MIMO systems. In summary, no prior work rigorously investigated a joint design approach of antenna selection and precoders for maximizing the EE in downlink quantized massive MIMO, which motivates to develop a novel precoding method with joint antenna selection.

B. Contributions

We investigate a EE maximization problem with regard to precoders in a downlink multiuser massive MIMO system, where low-resolution DACs and low-resolution ADCs are employed at the BS and users, respectively. Our main contributions are summarized as follows:

- In the first phase of this paper, we put forth a precoding strategy for maximizing the EE in quantized massive MIMO systems by reformulating the problem. To accomplish the precoding optimization, by adopting the AQNM [28], which is a linear approximation technique of the non-linear quantizer function with additional quantization noise, we define the EE of the quantized massive MIMO system. The EE function is the sum SE function normalized by the total transmit power, a fractional programming form. Therefore, it is critical to find an optimal transmit power level while maximizing the SE. Unfortunately, the EE maximization problem in the downlink quantized massive MIMO system is NP-hard, similar to the case using infinite-resolution DACs and ADCs. Therefore, finding a global optimum solution is infeasible using polynomial-time complexity algorithms. To overcome this challenge, we take the Dinkelbach method [29] to relax a fractional programming, then we decompose the optimization variables, i.e., the precoding vectors, into two parts: 1) scaling and 2) directional components.
- Leveraging this decomposition, we propose an alternating optimization framework for the EE maximization called Q-GPI-EEM. To be specific, for a fixed Dinkelbach variable and a directional component, we obtain the optimum power scaling component by using a gradient descent method. Since this sub-optimization problem is convex, gradient descent is sufficient to find the optimum point. Subsequently, we find the directional component with the obtained power scaling component; we derive

the first-order optimality condition for the non-convex precoding direction optimization problem. The derived condition is cast as a functional eigenvalue problem, and it insinuates that a local optimal point that has zero gradient is obtained by finding the principal eigenvector of the functional eigenvalue problem. To this end, modifying the algorithm in [30], we present a precoding algorithm called quantized generalized power iteration-based algorithm for direction optimization (Q-GPI-DO) that iteratively identifies the principal eigenvector with a few numbers of iterations. The algorithm iterates by updating the Dinkelbach variable until it converges. Unlike other existing algorithms for the EE maximization in the quantized massive MIMO systems, the most prominent feature of Q-GPI-EEM is to jointly identify a set of active antennas and the corresponding precoding direction by considering the effect of RF circuit power consumption for the active transmit antennas.

- In the second phase of the paper, as a byproduct, we also present a precoding method for maximizing the SE in quantized massive MIMO systems, called Q-GPI-SEM. We show that the SE maximization method is a special case of the EE maximization method, reduced by simplifying some parameters of the EE maximization setups. Q-GPI-SEM ensures to find a local optimal point of the SE maximization for any important system parameters, including the number of antennas, the number of downlink users, and the number of DAC and ADS resolution bits. Besides, our algorithm generalizes, and the prior GPI-based algorithm [30] by incorporating the quantization error effects caused by DACs and ADCs.
- Simulation results demonstrate that the proposed algorithms, Q-GPI-EEM and Q-GPI-SEM considerably outperform conventional algorithms in both EE and SE. Our EE maximization algorithm provides robustness to the maximum transmit power constraint in terms of EE. For example, the EE performances of the other methods eventually degrade as the transmit power increases, while our algorithm offers the monotonically increasing EE by adjusting the actual transmit power in the algorithm. In terms of SE, the SE is saturated as the transmit power increases because of the quantization distortion effects of low-resolution quantizers. Using the proposed algorithm, we can pull up this saturation level more than $2\times$ compared to the conventional methods.
- In addition to the SE and EE improvement, we also elucidate a system design guideline for quantized massive MIMO systems. Exploiting the proposed joint method, we observe that using large-scale antenna elements, each with low-resolution quantizers, provides considerable benefits in both the SE and EE. We note that this was also found in several prior works (e.g., [4] or [25]), yet our finding is obtained by exploiting the joint design of antenna selection and precoders. Thanks to this, our method achieves higher EE compared to the baseline methods at the same number of antennas. More detailed system design insights are provided as follows: (i) Regarding the EE, there exists the optimal number of the DAC bits that maximize the EE. For instance, using 4-bit DACs achieves the maximum EE if the BS has 8 antennas, while 3-bit DACs achieves the maximum EE if the BS has 32 antennas. (ii) Regarding the SE, our method is highly efficient if the number of antennas is enough. In particular, the proposed method achieves a similar level of SE of the high-resolution DACs, e.g., $9 \sim 11$ bits even with low-resolution DACs, e.g., $3 \sim 5$ bits. (iii) Under the constraint of the total number of DAC bits, using the homogeneous DACs at the BS is beneficial for the SE and the EE. In addition, the proposed algorithm shows higher robustness for DAC configuration, and thus, it can provide more flexibility in the system design. Overall, the proposed algorithm provides high SE and EE performance in the massive MIMO regime, allowing the BS to employ low-resolution DACs and offering high system design flexibility.

Notation: $\mathbf{A}$ is a matrix and $\mathbf{a}$ is a column vector. Superscripts $(\cdot)^*$, $(\cdot)^{\mathsf{T}}$, $(\cdot)^{\mathsf{H}}$, and $(\cdot)^{-1}$ denote conjugate, transpose, Hermitian, and matrix inversion, respectively. $\mathbf{I}_N$ is an identity matrix of size $N \times N$ and $\mathbf{0}_{M\times N}$ is a zero matrix of size $M \times N$. $\mathcal{CN}(\mu, \sigma^2)$ is a complex Gaussian distribution with mean μ and variance σ^2. $\mathrm{Unif}[a, b]$ denotes a discrete uniform distribution from a to b. A diagonal matrix $\mathrm{diag}(\mathbf{A})$ has the diagonal entries of $\mathbf{A}$ at its diagonal entries. Assuming that $\mathbf{A}_1, \ldots, \mathbf{A}_N \in \mathbb{C}^{K\times K}$, $\mathrm{blkdiag}\left(\mathbf{A}_1, \ldots, \mathbf{A}_N\right)$ is a block diagonal matrix of size $KN \times KN$ whose nth block diagonal entry is $\mathbf{A}_n$. $\|\mathbf{A}\|$ represents L2 norm, $\|\mathbf{A}\|_F$ represents Frobenius norm, $\mathbb{E}[\cdot]$ represents an expectation operator, $\mathrm{tr}(\cdot)$ denotes a trace operator, $\mathrm{vec}(\cdot)$ indicates a vectorization operator, and $\otimes$ is Kroncker product.

II. System Model

We consider a downlink multiuser massive MIMO system where the base station (BS) is equipped with $N \gg 1$ antennas and each user is equipped with a single antenna, and there are K users to be served. We further assume that the BS employs low-resolution DACs and the users employ low-resolution ADCs. We also consider a general case for low-resolution quantizers where each DAC and ADC can have any bit configuration. We remark that such an assumption covers the case of high-resolution quantizers as well as low-resolution quantizers.[1] At the BS, a precoded digital baseband transmit signal vector $\mathbf{x} \in \mathbb{C}^N$ is expressed as

$$\mathbf{x} = \sqrt{P}\mathbf{W}\mathbf{s}, \tag{1}$$

where $\mathbf{s} \sim \mathcal{CN}(\mathbf{0}_{K\times 1}, \mathbf{I}_K)$ is a symbol vector, $\mathbf{W} \in \mathbb{C}^{N\times K}$ represents a precoding matrix, and P is the maximum transmit power.

The digital baseband signal $\mathbf{x}$ in (1) is further quantized at the DACs prior to transmission. To characterize the quantized signal, we adopt the AQNM [28], that approximates the quantization process in a linear form. Then, the analog baseband transmit signal after quantization becomes

$$\mathcal{Q}(\mathbf{x}) \approx \mathbf{x}_{\mathsf{q}} = \sqrt{P}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{W}\mathbf{s} + \mathbf{q}_{\mathsf{bs}},$$

[1] An example of a low-resolution quantizer at the users: in internet-of-things, the users can be energy-hungry even with a single antenna, which requires to use low-resolution ADCs.

where $\mathcal{Q}(\cdot)$ is an element-wise quantizer that applies for each real and imaginary part, $\mathbf{\Phi}_{\alpha_{\mathsf{bs}}} = \mathrm{diag}(\alpha_{\mathsf{bs},1}, \ldots, \alpha_{\mathsf{bs},N}) \in \mathbb{R}^{N\times N}$ denotes a diagonal matrix of quantization loss, and $\mathbf{q}_{\mathsf{bs}} \in \mathbb{C}^N$ is a quantization noise vector. The quantization loss of the nth DAC $\alpha_{\mathsf{bs},n} \in (0,1)$ is defined as $\alpha_{\mathsf{bs},n} = 1-\beta_{\mathsf{bs},n}$, where $\beta_{\mathsf{bs},n}$ is a normalized mean squared quantization error with $\beta_{\mathsf{bs},n} = \frac{\mathbb{E}[|x-\mathcal{Q}_n(x)|^2]}{\mathbb{E}[|x|^2]}$ [28], [31]. The values of $\beta_{\mathsf{bs},n}$ are characterized depending on the number of quantization bits at the nth BS antenna $b_{\mathsf{DAC},n}$. Specifically, for $b_{\mathsf{DAC},n} \leq 5$, $\beta_{\mathsf{bs},n}$ is quantified in Table 1 in [32]. For $b_{\mathsf{DAC},n} > 5$, $\beta_{\mathsf{bs},n}$ can be approximated as $\beta_{\mathsf{bs},n} \approx \frac{\pi\sqrt{3}}{2} 2^{-2b_{\mathsf{DAC},n}}$ [33]. The quantization noise $\mathbf{q}_{\mathsf{bs}}$ is uncorrelated with $\mathbf{x}$ and follows $\mathbf{q}_{\mathsf{bs}} \sim \mathcal{CN}(\mathbf{0}_{N\times 1}, \mathbf{R}_{\mathbf{q}_{\mathsf{bs}}\mathbf{q}_{\mathsf{bs}}})$. The covariance $\mathbf{R}_{\mathbf{q}_{\mathsf{bs}}\mathbf{q}_{\mathsf{bs}}}$ is computed as [28], [31]

$$\begin{aligned}\mathbf{R}_{\mathbf{q}_{\mathsf{bs}}\mathbf{q}_{\mathsf{bs}}} &= \mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathsf{H}}\right]\right)\\ &= \mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(P\mathbf{W}\mathbf{W}^{\mathsf{H}}\right).\end{aligned}$$

The quantized signal $\mathbf{x}_{\mathsf{q}}$ is amplified by a power amplifier under a power constraint at the BS. Since the maximum transmit power is defined as P, we have the following power constraint [34]

$$\mathrm{tr}\left(\mathbb{E}\left[\mathbf{x}_{\mathsf{q}}\mathbf{x}_{\mathsf{q}}^{\mathsf{H}}\right]\right) \leq P.$$

Now, we represent the received analog baseband signals at all the K users as

$$\mathbf{y} = \mathbf{H}^{\mathsf{H}}\mathbf{x}_{\mathsf{q}} + \mathbf{n}, \tag{2}$$

where $\mathbf{H} \in \mathbb{C}^{N\times K}$ is a channel matrix and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}_{K\times 1}, \sigma^2\mathbf{I}_K)$ indicates an $K\times 1$ additive white Gaussian noise vector with zero mean and variance of σ^2. Each column of the channel matrix $\mathbf{h}_k$ represents the channel between user k and the BS. We assume a block fading model, where $\mathbf{h}_k$ is invariant within one transmission block and changes independently over transmission blocks. We focus on a single transmission block and assume that the CSI is perfectly known at the BS.

The received analog signals in (2) are quantized at the ADCs of the users. Accordingly, the received digital baseband signals are given as [28], [35]

$$\begin{aligned}\mathcal{Q}(\mathbf{y}) \approx \mathbf{y}_{\mathsf{q}} &= \mathbf{\Phi}_{\alpha}\mathbf{y} + \mathbf{q}\\ &= \mathbf{\Phi}_{\alpha}\mathbf{H}^{\mathsf{H}}\mathbf{x}_{\mathsf{q}} + \mathbf{\Phi}_{\alpha}\mathbf{n} + \mathbf{q}\\ &= \sqrt{P}\mathbf{\Phi}_{\alpha}\mathbf{H}^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{W}\mathbf{s} + \mathbf{\Phi}_{\alpha}\mathbf{H}^{\mathsf{H}}\mathbf{q}_{\mathsf{bs}} + \mathbf{\Phi}_{\alpha}\mathbf{n} + \mathbf{q},\end{aligned}$$

where $\mathbf{\Phi}_{\alpha} = \mathrm{diag}(\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^{K\times K}$ is a diagonal matrix of ADC quantization loss defined as $\alpha_k = 1-\beta_k$, and $\mathbf{q} \in \mathbb{C}^K$ is a quantization noise vector, which is uncorrelated with $\mathbf{y}$. Here, α_k and β_k are similarly defined as $\alpha_{\mathsf{bs},n}$ and $\beta_{\mathsf{bs},n}$, respectively. The quantization noise $\mathbf{q}$ has zero mean and follows a complex Gaussian distribution. Consequently, the digital baseband signal at user k is given as

$$\begin{aligned}y_{\mathsf{q},k} = \sqrt{P}\alpha_k\mathbf{h}_k^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{w}_k s_k + \sqrt{P}\sum_{\ell\neq k}\alpha_k\mathbf{h}_k^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{w}_\ell s_\ell\\ +\alpha_k\mathbf{h}_k^{\mathsf{H}}\mathbf{q}_{\mathsf{bs}} + \alpha_k n_k + q_k,\end{aligned} \tag{3}$$

where $\mathbf{w}_k \in \mathbb{C}^N$ denotes the kth column vector of $\mathbf{W}$, and s_k, n_k, and q_k represent kth element of $\mathbf{s}$, $\mathbf{n}$, and $\mathbf{q}$, respectively. The variance of q_k is computed as [28]

$$\begin{aligned}r_{\mathsf{q}_k\mathsf{q}_k} &= \alpha_k\beta_k\mathbb{E}\left[y_k y_k^{\mathsf{H}}\right]\\ &= \alpha_k\beta_k\left(\mathbf{h}_k^{\mathsf{H}}\mathbb{E}\left[\mathbf{x}_{\mathsf{q}}\mathbf{x}_{\mathsf{q}}^{\mathsf{H}}\right]\mathbf{h}_k^{\mathsf{H}} + \sigma^2\right)\\ &= \alpha_k\beta_k\left(\mathbf{h}_k^{\mathsf{H}}\left(P\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{W}\mathbf{W}^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{\mathsf{H}} + \mathbf{R}_{\mathbf{q}_{\mathsf{bs}}\mathbf{q}_{\mathsf{bs}}}\right)\mathbf{h}_k + \sigma^2\right).\end{aligned} \tag{4}$$

The following sections formulate an EE maximization problem, propose an algorithm, and validate the performance via simulations accordingly.

Remark 1 (Comparison Between AQNM and Bussgang Decomposition): In the literature, two approaches are known to model the non-linear quantization distortion effect: AQNM and Bussgang decomposition. In some studies, these two approaches were described as two different methods. AQNM and Bussgang decomposition are actually equivalent. More specifically, AQNM is a special case of Bussgang decomposition when MMSE tailors a quantization function. We refer [36] for more details.

III. Energy Efficiency Maximization Problem Formulation

A. Performance Metric

We characterize the EE for the considered system as our main performance metric. Using (3), we define the downlink SINR of user k as

$$\Gamma_k = \frac{P\alpha_k^2|\mathbf{h}_k^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{w}_k|^2}{P\alpha_k^2\sum_{\ell\neq k}|\mathbf{h}_k^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{w}_\ell|^2 + \alpha_k^2\mathbf{h}_k^{\mathsf{H}}\mathbf{R}_{\mathbf{q}_{\mathsf{bs}}\mathbf{q}_{\mathsf{bs}}}\mathbf{h}_k + \alpha_k^2\sigma^2 + r_{\mathsf{q}_k\mathsf{q}_k}}. \tag{5}$$

Unlike a perfect quantization system where the resolution of DACs and ADCs is infinite, the SINR in (5) includes the quantization noise power and the quantization loss induced by both low-resolution DACs and ADCs. Accordingly, the SE for user k is expressed as

$$R_k = \log_2\left(1+\Gamma_k\right). \tag{6}$$

Now, we define the EE based on (6). Let P_{LP}, P_{M}, P_{LO}, P_{H}, P_{PA}, P_{RF}, P_{DAC}, P_{cir}, and P_{BS} denote the power consumption of low-pass filter, mixer, local oscillator, 90° hybrid with buffer, power amplifier (PA), radio frequency (RF) chain, DAC, analog circuits, and BS, respectively. The DAC power consumption P_{DAC} (in Watt) is defined as [22], [37]

$$P_{\mathsf{DAC}}(b_{\mathsf{DAC}}, f_s) = 1.5\times 10^{-5}\cdot 2^{b_{\mathsf{DAC}}} + 9\times 10^{-12}\cdot f_s\cdot b_{\mathsf{DAC}},$$

where f_s is the sampling rate. If antenna n is active, then the corresponding DAC pair and RF chain consume the circuit power of $2\,P_{\mathsf{DAC}}(b_{\mathsf{DAC},n}, f_s) + P_{\mathsf{RF}}$ where $P_{\mathsf{RF}} = 2P_{\mathsf{LP}} + 2P_{\mathsf{M}} + P_{\mathsf{H}}$. On the contrary, if the antenna is inactive, no power is consumed in the corresponding DAC and RF chain. Considering such behaviour, the circuit power consumption P_{cir} is formulated as [22]

$$P_{\mathsf{cir}} = P_{\mathsf{LO}} + \sum_{n=1}^{N}\mathbb{1}_{\{n\in\mathcal{A}\}}\left(2\,P_{\mathsf{DAC}}(b_{\mathsf{DAC},n}, f_s) + P_{\mathsf{RF}}\right), \tag{7}$$

TABLE I
ABBREVIATION AND DEFINITION OF SYSTEM PARAMETERS

Abbreviation	Definition	Value
P_{LP}	power consumption of low-pass filter	14 mW
P_{M}	power consumption of mixer	0.3 mW
P_{LO}	power consumption of local oscillator	22.5 mW
P_{H}	power consumption of 90° hybrid with buffer	3 mW
P_{PA}	power consumption of power amplifier	P_{tx}/κ
P_{RF}	power consumption of radio frequency chain	$2P_{\mathsf{LP}}+2P_{\mathsf{M}}+P_{\mathsf{H}}$
P_{DAC}	power consumption of DAC	$1.5\times10^{-5}\cdot 2^{b_{\mathsf{DAC}}}+9\times10^{-12}\cdot f_s\cdot b_{\mathsf{DAC}}$
P_{cir}	power consumption of analog circuits	$P_{\mathsf{LO}}+\sum_{n=1}^{N}\mathbb{1}_{\{n\in\mathcal{A}\}}\left(2\,P_{\mathsf{DAC}}(b_{\mathsf{DAC},n},f_s)+P_{\mathsf{RF}}\right)$
P_{BS}	power consumption of BS	$P_{\mathsf{cir}}+\kappa^{-1}P_{\mathsf{tx}}$
P_{tx}	transmit power	$\mathrm{tr}(\mathbb{E}[\mathbf{x}_{\mathsf{q}}\mathbf{x}_{\mathsf{q}}^{\mathsf{H}}])$
κ	power amplifier efficiency	$P_{\mathsf{tx}}/P_{\mathsf{PA}}=0.27$
$b_{\mathsf{DAC},n}$	DAC resolution bit of the n-th antenna	$b_{\mathsf{DAC},n}\in[2,12]$
f_s	sampling rate	10^8 Hz

where $\mathbb{1}_{\{a\}}$ is the indicator function which is $\mathbb{1}_{\{a\}}=1$ only when a is true and $\mathbb{1}_{\{a\}}=0$ otherwise, and $\mathcal{A}$ is an index set of active antennas. Then the BS power consumption P_{BS} is given as [22]

$$P_{\mathsf{BS}}(N,\mathbf{b}_{\mathsf{DAC}},f_s,\mathbf{W})=P_{\mathsf{cir}}+\kappa^{-1}P_{\mathsf{tx}},$$

where P_{tx} is the transmit power $P_{\mathsf{tx}}=\mathrm{tr}(\mathbb{E}[\mathbf{x}_{\mathsf{q}}\mathbf{x}_{\mathsf{q}}^{\mathsf{H}}])$, κ is a PA efficiency, i.e., $\kappa=P_{\mathsf{tx}}/P_{\mathsf{PA}}$, and $\mathbf{b}_{\mathsf{DAC}}=[b_{\mathsf{DAC},1},\ldots,b_{\mathsf{DAC},N}]^{\mathsf{T}}$. Finally, the EE of the considered system is defined as

$$\eta=\frac{\Omega\sum_{k=1}^{K}R_k}{P_{\mathsf{BS}}(N,\mathbf{b}_{\mathsf{DAC}},f_s,\mathbf{W})},\tag{8}$$

where Ω denotes transmission bandwidth. We summarize the used system parameters, their definitions, and values in Table I.

B. Formulated Problem

We now formulate an EE maximization problem with respect to a precoder. We first normalize the EE η in (8) by the bandwidth Ω since it is irrelevant to precoder design. Throughout the paper, we use the EE and normalized EE interchangeably as they do not change the problem. Then the EE maximization problem is formulated as

$$\underset{\mathbf{W}}{\text{maximize}}\ \frac{\sum_{k=1}^{K}R_k}{P_{\mathsf{BS}}(N,\mathbf{b}_{\mathsf{DAC}},f_s,\mathbf{W})}\tag{9}$$

$$\text{subject to}\ \mathrm{tr}\left(\mathbb{E}\left[\mathbf{x}_{\mathsf{q}}\mathbf{x}_{\mathsf{q}}^{\mathsf{H}}\right]\right)\le P,\tag{10}$$

We remark that the BS power consumption includes the circuit power consumption, which is a function of active antennas, and thus, the EE maximization problem needs to be solved by designing $\mathbf{W}$ with incorporating the impact of active and inactive antenna sets. In this regard, we propose an algorithm that jointly designs a precoder and performs an antenna selection in Section IV.

IV. JOINT PRECODING AND ANTENNA SELECTION ALGORITHM

A direct solution for the formulated problem in (9) is not available since it is non-smooth and non-convex. We first reformulate the problem and then propose an algorithm that provides the best sub-optimal solution to resolve these challenges.

A. Problem Reformulation

We reformulate (9) by using the Dinkelbach method [29] as

$$\underset{\mathbf{W},\mu}{\text{maximize}}\ \sum_{k=1}^{K}R_k-\mu P_{\mathsf{BS}}(N,\mathbf{b}_{\mathsf{DAC}},f_s,\mathbf{W})\tag{11}$$

$$\text{subject to}\ \mathrm{tr}\left(\mathbb{E}\left[\mathbf{x}_{\mathsf{q}}\mathbf{x}_{\mathsf{q}}^{\mathsf{H}}\right]\right)\le P\tag{12}$$

$$\mu>0,\tag{13}$$

where μ is an auxiliary variable. To cast the problem (11) into a tractable form, we first rewrite the DAC quantization noise covariance term coupled with a user channel in (5) as

$$\begin{aligned}\mathbf{h}_k^{\mathsf{H}}\mathbf{R}_{\mathbf{q}_{\mathsf{bs}}\mathbf{q}_{\mathsf{bs}}}\mathbf{h}_k&=\mathbf{h}_k^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(P\mathbf{W}\mathbf{W}^{\mathsf{H}}\right)\mathbf{h}_k\\&=P\sum_{n=1}^{N}|h_{n,k}|^2\alpha_{\mathsf{bs},n}\beta_{\mathsf{bs},n}\sum_{\ell=1}^{K}|w_{n,\ell}|^2\\&=P\sum_{\ell=1}^{K}\sum_{n=1}^{N}w_{n,\ell}^{*}\alpha_{\mathsf{bs},n}\beta_{\mathsf{bs},n}h_{n,k}h_{n,k}^{*}w_{n,\ell}\\&=P\sum_{\ell=1}^{K}\mathbf{w}_\ell^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\right)\mathbf{w}_\ell.\end{aligned}\tag{14}$$

We also rewrite the ADC quantization noise variance $r_{\mathsf{q}_k\mathsf{q}_k}$ in (4) as

$$\begin{aligned}r_{\mathsf{q}_k\mathsf{q}_k}&\overset{(a)}{=}\alpha_k\beta_k\Bigg(P\mathbf{h}_k^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{W}\mathbf{W}^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{\mathsf{H}}\mathbf{h}_k\\&\qquad+P\sum_{\ell=1}^{K}\mathbf{w}_\ell^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\right)\mathbf{w}_\ell+\sigma^2\Bigg)\\&=\alpha_k\beta_k\Bigg(P\sum_{\ell=1}^{K}\mathbf{w}_\ell^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{\mathsf{H}}\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{w}_\ell\\&\qquad+P\sum_{\ell=1}^{K}\mathbf{w}_\ell^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\right)\mathbf{w}_\ell+\sigma^2\Bigg)\\&=P\alpha_k\beta_k\sum_{\ell=1}^{K}\mathbf{w}_\ell^{\mathsf{H}}\\&\qquad\times\left(\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{\mathsf{H}}\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}+\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\right)\right)\\&\qquad\times\mathbf{w}_\ell+\alpha_k\beta_k\sigma^2,\end{aligned}\tag{15}$$

where (a) comes from (14). Applying (14), (15), and $\beta_k = 1 - \alpha_k$ to the SINR in (5), we re-organize the SINR term to represent the SINR in a more compact form as shown in (16), shown at the bottom of the page. Now, we simplify the transmit power constraint in (12) as

$$\begin{aligned}&\mathrm{tr}\left(\mathbb{E}\left[\mathbf{x_q}\mathbf{x_q}^{\mathsf{H}}\right]\right)\\&= \mathrm{tr}\left(P\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{W}\mathbf{W}^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{\mathsf{H}} + P\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\boldsymbol{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(\mathbf{W}\mathbf{W}^{\mathsf{H}}\right)\right)\\&\overset{(a)}{=} \mathrm{tr}\left(P\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{W}\mathbf{W}^{\mathsf{H}}\right) \le P,\end{aligned} \tag{17}$$

where (a) comes from $\boldsymbol{\Phi}_{\beta_{\mathsf{bs}}} = \mathbf{I}_N - \boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}$, $\mathrm{tr}\big(\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{W}\mathbf{W}^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\big) = \mathrm{tr}\big(\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^2\mathbf{W}\mathbf{W}^{\mathsf{H}}\big)$, and the definition of the trace operation.

Regarding the SE point of view, using the maximum transmit power, i.e., $P_{\mathrm{tx}} = P$, maximizes the SE for a given precoder. For the EE point of view, however, using the maximum transmit power may decrease the EE because it increases the SE with a logarithmic scale while it increases the power consumption with an (approximately) linear scale. For this reason, using the maximum power without adjusting a power level, the EE converges to zero eventually, as shown in [25]. To prevent this phenomenon, an optimal power scaling solution needs to be found to maximize the EE. To this end, we introduce a scalar weight $\tau : 0 < \tau \le 1$ in the power constraint as

$$\mathrm{tr}\left(\mathbb{E}\left[\mathbf{x_q}\mathbf{x_q}^{\mathsf{H}}\right]\right) = \tau P. \tag{18}$$

Then applying (18) to (17), the power constraint reduces to

$$\mathrm{tr}\left(\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{W}\mathbf{W}^{\mathsf{H}}\right) = \tau. \tag{19}$$

To incorporate the power scaling in the precoder, we define a weighted and normalized precoding matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]$ where

$$\mathbf{v}_k = \frac{1}{\sqrt{\tau}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2}\mathbf{w}_k \in \mathbb{C}^N. \tag{20}$$

Then, applying (20) to (16), the SE of user k is re-written as a function of τ and $\mathbf{V}$, which is (21), shown at the bottom of the page. Using (19), (20), and (21), the problem in (11) becomes

$$\underset{\mathbf{V},\tau,\mu}{\text{maximize}} \sum_{k=1}^{K} \bar{R}_k(\mathbf{V},\tau) - \mu P_{\mathsf{BS}}(N, \mathbf{b}_{\mathsf{DAC}}, f_s, \mathbf{V}, \tau) \tag{22}$$

$$\begin{aligned}\text{subject to } &\mathrm{tr}\left(\mathbf{V}\mathbf{V}^{\mathsf{H}}\right) = 1\\&0 < \tau \le 1\\&\mu > 0.\end{aligned} \tag{23}$$

Note that the power constraint is equivalent to $\|\bar{\mathbf{v}}\|^2 = 1$ where $\bar{\mathbf{v}} = \mathrm{vec}(\mathbf{V})$, which is interpreted as a directional component of the precoder. Consequently, the EE maximization problem is now a problem of jointly finding an optimal power level τ and precoding direction $\bar{\mathbf{v}}$. In this regard, we decompose (22) into two subsequent problems: find $\mathbf{V}$ for given τ and find τ for a given $\mathbf{V}$. With the obtained τ and $\mathbf{V}$, we iteratively update the Dinkelbach variable μ.

B. Proposed Algorithm

1) Optimal Direction $\mathbf{V}^\star$: We solve the problem in (22) regarding $\mathbf{V}$ for given τ and μ. Again, this phase is mainly related to designing an optimal direction of precoding since the power constraint becomes $\|\bar{\mathbf{v}}\|^2 = 1$. Let $\mathbf{G}_k = (\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2})^{\mathsf{H}}\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2} + \boldsymbol{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\right)$. Then leveraging the fact that $\|\bar{\mathbf{v}}\|^2 = 1$, we cast the SE $\bar{R}_k$ in (21) into a Rayleigh quotient form as

$$\bar{R}_k = \log_2\left(\frac{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{A}_k\bar{\mathbf{v}}}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{B}_k\bar{\mathbf{v}}}\right) \tag{24}$$

where

$$\begin{aligned}\mathbf{A}_k &= \mathrm{blkdiag}\big(\mathbf{G}_k, \cdots, \mathbf{G}_k\big) + \frac{\sigma^2}{\tau P}\mathbf{I}_{NK},\\\mathbf{B}_k &= \mathbf{A}_k\\&\quad -\mathrm{blkdiag}\left(\mathbf{0}_{N\times N}, \cdots, \alpha_k(\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2})^{\mathsf{H}}\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2}, \cdots, \mathbf{0}_{N\times N}\right).\end{aligned}$$

Consequently, we have the following problem for given τ and μ:

$$\begin{aligned}\underset{\bar{\mathbf{v}}}{\text{maximize}} \ &\sum_{k=1}^{K}\log_2\left(\frac{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{A}_k\bar{\mathbf{v}}}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{B}_k\bar{\mathbf{v}}}\right)\\&-\mu\sum_{n=1}^{N}\left(2\,P_{\mathsf{DAC}}(b_{\mathsf{DAC},n}, f_s) + P_{\mathsf{RF}}\right)\mathbb{1}_{\{n\in\mathcal{A}\}}.\end{aligned} \tag{25}$$

Note that the constraint in (23) is ignored at this step. In the proposed algorithm, however, $\bar{\mathbf{v}}$ will be normalized, which indeed satisfies the constraint.

Now, the major challenge in solving the problem in (25) is to handle the indicator function $\mathbb{1}_{\{n\in\mathcal{A}\}}$. It is, however, highly difficult due to the lack of smoothness. To resolve this challenge, we transform the indicator function into a more favorable form as follows: let $\tilde{\mathbf{w}}_n$ be the nth row of the precoder $\mathbf{W}$. Then we clarify that antenna n is active if and only if $\|\tilde{\mathbf{w}}_n\|^2 > 0$. Equivalently, we consider $\mathbb{1}_{\{n\in\mathcal{A}\}} = \mathbb{1}_{\{\|\frac{1}{\sqrt{\alpha_{\mathsf{bs},n}}}\tilde{\mathbf{v}}_n\|^2>0\}}$ because $\mathbf{W} \propto \boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{-1/2}\mathbf{V}$ from the definition in (20), where $\tilde{\mathbf{v}}_n$ is the nth row vector of $\mathbf{V}$. We remark that

$$\Gamma_k = \frac{\alpha_k|\mathbf{h}_k^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{w}_k|^2}{\sum_{\ell=1}^{K}|\mathbf{h}_k^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{w}_\ell|^2 - \alpha_k|\mathbf{h}_k^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{w}_k|^2 + \sum_{\ell=1}^{K}\mathbf{w}_\ell^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\boldsymbol{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\right)\mathbf{w}_\ell + \sigma^2/P}. \tag{16}$$

$$\bar{R}_k(\mathbf{V},\tau) = \log_2\left(1 + \frac{\tau\alpha_k|\mathbf{h}_k^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2}\mathbf{v}_k|^2}{\tau\sum_{\ell=1}^{K}|\mathbf{h}_k^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2}\mathbf{v}_\ell|^2 - \tau\alpha_k|\mathbf{h}_k^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2}\mathbf{v}_k|^2 + \tau\sum_{\ell=1}^{K}\mathbf{v}_\ell^{\mathsf{H}}\boldsymbol{\Phi}_{\beta_{\mathsf{bs}}}\mathrm{diag}\left(\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\right)\mathbf{v}_\ell + \sigma^2/P}\right). \tag{21}$$

we omit the effect of τ since it applies identically to all the antennas, but the quantization loss term is included so that the quantization effect can be applied differently across antennas. Subsequently, we approximate the indicator function by using the following approximation [38]:

$$\mathbb{1}_{\{|x|^2>0\}} \approx \frac{\log_2(1+|x|^2/\rho)}{\log_2(1+1/\rho)}, \tag{26}$$

where the approximation becomes tight as $\rho \to 0$. Using (26), the indicator function in (7) is represented as

$$\begin{aligned}\mathbb{1}_{\{n\in\mathcal{A}\}} &= \mathbb{1}_{\{\|\tilde{\mathbf{v}}_n/\sqrt{\alpha_{\mathsf{bs},n}}\|^2>0\}} \\ &\approx \log_2\left(1+\rho^{-1}\left\|\frac{1}{\sqrt{\alpha_{\mathsf{bs},n}}}\tilde{\mathbf{v}}_n\right\|^2\right)^{\omega_\rho},\end{aligned} \tag{27}$$

where $\omega_\rho = 1/\log_2(1+\rho^{-1})$ and $\rho > 0$ is a small enough value. Further denoting that $P_{\mathsf{ant},n} = 2\,P_{\mathsf{DAC}}(b_{\mathsf{DAC},n}, f_s) + P_{\mathsf{RF}}$, we have

$$\begin{aligned}&\sum_{n=1}^{N}\left(2\,P_{\mathsf{DAC}}(b_{\mathsf{DAC},n},f_s)+P_{\mathsf{RF}}\right)\mathbb{1}_{\{n\in\mathcal{A}\}} \\ &\quad\approx \sum_{n=1}^{N}\log_2\left(1+\rho^{-1}\left\|\frac{1}{\sqrt{\alpha_{\mathsf{bs},n}}}\tilde{\mathbf{v}}_n\right\|^2\right)^{\omega_\rho P_{\mathsf{ant},n}}.\end{aligned} \tag{28}$$

We note that the approximation technique (27) makes it possible to incorporate the indicator function into the problem formulation without introducing explicit discrete variables. The next step is to express (28) in terms of the $\bar{\mathbf{v}}$. To this end, we let $\mathbf{e}_n$ be the N dimensional nth canonical basis vector with a single 1 at its nth coordinate and zeros elsewhere. Then we can write $\frac{1}{\sqrt{\alpha_{\mathsf{bs},n}}}\tilde{\mathbf{v}}_n$ as $\tilde{\mathbf{v}}_n = \mathbf{e}_n^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{-1/2}\mathbf{V}$. Subsequently, we rewrite $\|\frac{1}{\sqrt{\alpha_{\mathsf{bs},n}}}\tilde{\mathbf{v}}_n\|^2$ as

$$\begin{aligned}\left\|\frac{1}{\sqrt{\alpha_{\mathsf{bs},n}}}\tilde{\mathbf{v}}_n\right\|^2 &= \mathbf{e}_n^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{-1/2}\mathbf{V}\mathbf{V}^{\mathsf{H}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{-1/2}\mathbf{e}_n \\ &= \mathrm{vec}(\tilde{\mathbf{e}}_n^{\mathsf{H}}\mathbf{V}\mathbf{V}^{\mathsf{H}}\tilde{\mathbf{e}}_n) \\ &\overset{(a)}{=} \left(\left(\tilde{\mathbf{e}}_n^{\mathsf{T}}\mathbf{V}^*\right)\otimes\tilde{\mathbf{e}}_n^{\mathsf{H}}\right)\mathrm{vec}(\mathbf{V}) \\ &\overset{(b)}{=} \left(\left(\left(\mathbf{I}_K\otimes\tilde{\mathbf{e}}_n^{\mathsf{T}}\right)\mathrm{vec}(\mathbf{V}^*)\right)^{\mathsf{T}}\otimes\tilde{\mathbf{e}}_n^{\mathsf{H}}\right)\mathrm{vec}(\mathbf{V}) \\ &\overset{(c)}{=} \bar{\mathbf{v}}^{\mathsf{H}}\left(\mathbf{I}_K\otimes\tilde{\mathbf{e}}_n\otimes\tilde{\mathbf{e}}_n^{\mathsf{H}}\right)\bar{\mathbf{v}},\end{aligned} \tag{29}$$

where $\tilde{\mathbf{e}}_n = \boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{-1/2}\mathbf{e}_n$, (a) and (b) follow from $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathsf{T}}\otimes\mathbf{A})\mathrm{vec}(\mathbf{B})$, and (c) comes from $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$ and the definition of $\bar{\mathbf{v}} = \mathrm{vec}(\mathbf{V})$. Using (28) and (29), the objective function in (25) can be represented as

$$L_v(\bar{\mathbf{v}}) = \log_2\lambda_v(\bar{\mathbf{v}}), \tag{30}$$

where

$$\lambda_v(\bar{\mathbf{v}}) = \prod_{k=1}^{K}\left(\frac{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{A}_k\bar{\mathbf{v}}}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{B}_k\bar{\mathbf{v}}}\right)\prod_{n=1}^{N}\left(\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{E}_n\bar{\mathbf{v}}\right)^{-\mu\omega_\rho P_{\mathsf{ant},n}}$$

and $\mathbf{E}_n = \mathbf{I}_{NK} + \rho^{-1}\mathbf{I}_K\otimes\tilde{\mathbf{e}}_n\otimes\tilde{\mathbf{e}}_n^{\mathsf{H}}$ as $\|\bar{\mathbf{v}}\|^2 = 1$. With the reformulated objective function in (30), we derive a condition for a local-optimal stationary point and propose an algorithm to find such a local optimal point.

Lemma 1: The first-order optimality condition of the optimization problem (25) *with the approximated objective function in* (30) *for given* τ *and* μ *is satisfied if the following holds*:

$$\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}})\bar{\mathbf{v}} = \lambda_v(\bar{\mathbf{v}})\mathbf{B}_{\mathsf{KKT}}(\bar{\mathbf{v}})\bar{\mathbf{v}}, \tag{31}$$

where

$$\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}}) = \sum_{k=1}^{K}\frac{\mathbf{A}_k}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{A}_k\bar{\mathbf{v}}}\prod_{\ell=1}^{K}\left(\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{A}_\ell\bar{\mathbf{v}}\right), \tag{32}$$

$$\begin{aligned}\mathbf{B}_{\mathsf{KKT}}(\bar{\mathbf{v}}) = &\left(\sum_{k=1}^{K}\frac{\mathbf{B}_k}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{B}_k\bar{\mathbf{v}}}+\mu\omega_\rho\sum_{n=1}^{N}P_{\mathsf{ant},n}\frac{\mathbf{E}_n}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{E}_n\bar{\mathbf{v}}}\right) \\ &\times\prod_{\ell=1}^{K}\left(\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{B}_\ell\bar{\mathbf{v}}\right)\prod_{m=1}^{N}\left(\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{E}_m\bar{\mathbf{v}}\right)^{\mu\omega_\rho P_{\mathsf{ant},m}}.\end{aligned} \tag{33}$$

Proof: We compute the first derivative of (30) as

$$\frac{\partial L_v(\bar{\mathbf{v}})}{\partial\bar{\mathbf{v}}^{\mathsf{H}}} = \frac{1}{\lambda_v(\bar{\mathbf{v}})\ln 2}\frac{\partial\lambda_v(\bar{\mathbf{v}})}{\partial\bar{\mathbf{v}}^{\mathsf{H}}}$$

and subsequently, we derive $\partial\lambda_v(\bar{\mathbf{v}})/\partial\bar{\mathbf{v}}^{\mathsf{H}}$ and set it to zero

$$\begin{aligned}\frac{\partial\lambda_v(\bar{\mathbf{v}})}{\partial\bar{\mathbf{v}}^{\mathsf{H}}} &= 2\lambda_v(\bar{\mathbf{v}})\left(\sum_{k=1}^{K}\left(\frac{\mathbf{A}_k\bar{\mathbf{v}}}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{A}_k\bar{\mathbf{v}}}-\frac{\mathbf{B}_k\bar{\mathbf{v}}}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{B}_k\bar{\mathbf{v}}}\right)\right. \\ &\qquad\left.-\mu\omega_\rho\sum_{n=1}^{N}P_{\mathsf{ant},n}\frac{\mathbf{E}_n\bar{\mathbf{v}}}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{E}_n\bar{\mathbf{v}}}\right) \\ &= 0.\end{aligned}$$

Then, the first-order optimality condition can be derived as

$$\begin{aligned}&\sum_{k=1}^{K}\frac{\mathbf{A}_k}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{A}_k\bar{\mathbf{v}}}\lambda_{v,\mathsf{num}}(\bar{\mathbf{v}})\bar{\mathbf{v}} \\ &\quad=\lambda_v(\bar{\mathbf{v}})\left(\sum_{k=1}^{K}\frac{\mathbf{B}_k}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{B}_k\bar{\mathbf{v}}}+\mu\omega_\rho\sum_{n=1}^{N}P_{\mathsf{ant},n}\frac{\mathbf{E}_n}{\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{E}_n\bar{\mathbf{v}}}\right) \\ &\qquad\lambda_{v,\mathsf{denom}}(\bar{\mathbf{v}})\bar{\mathbf{v}},\end{aligned}$$

where $\lambda_{v,\mathsf{num}}(\bar{\mathbf{v}}) = \prod_{\ell=1}^{K}\left(\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{A}_\ell\bar{\mathbf{v}}\right)$ and $\lambda_{v,\mathsf{denom}}(\bar{\mathbf{v}}) = \prod_{\ell=1}^{K}\left(\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{B}_\ell\bar{\mathbf{v}}\right)\prod_{m=1}^{N}\left(\bar{\mathbf{v}}^{\mathsf{H}}\mathbf{E}_m\bar{\mathbf{v}}\right)^{\mu\omega_\rho P_{\mathsf{ant},m}}$. This completes the proof. ■

We interpret the derived condition (31) as a functional eigenvalue problem.

Remark 2: The derived first-order optimality condition in (31) can be transformed to a functional eigenvalue problem regarding the matrix $\mathbf{B}_{\mathsf{KKT}}^{-1}(\bar{\mathbf{v}})\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}})$ as

$$\mathbf{B}_{\mathsf{KKT}}^{-1}(\bar{\mathbf{v}})\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}})\bar{\mathbf{v}} = \lambda(\bar{\mathbf{v}})\bar{\mathbf{v}}. \tag{34}$$

From Remark 2, the first-order optimality condition in (34) is cast as a functional eigenvalue problem. More specifically, (34) is included in a class of nonlinear eigenvector dependent eigenvalue problem (NEPv) [39], where the matrix itself is a function of eigenvectors. Consequently, treating $\bar{\mathbf{v}}$ as an eigenvector of $\mathbf{B}_{\mathsf{KKT}}^{-1}(\bar{\mathbf{v}})\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}})$, $\lambda(\bar{\mathbf{v}})$ is interpreted as a corresponding eigenvalue of $\mathbf{B}_{\mathsf{KKT}}^{-1}(\bar{\mathbf{v}})\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}})$. In this regard, we need to find the principal eigenvector of $\mathbf{B}_{\mathsf{KKT}}^{-1}(\bar{\mathbf{v}})\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}})$ that leads $\lambda(\bar{\mathbf{v}})$ to be a maximum eigenvalue of $\mathbf{B}_{\mathsf{KKT}}^{-1}(\bar{\mathbf{v}})\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}})$ and also satisfies the first-order

Algorithm 1 Q-GPI-DO

initialize: $\bar{\mathbf{v}}^{(0)}$
Set the iteration count $t = 1$
while $\left\|\bar{\mathbf{v}}^{(t)} - \bar{\mathbf{v}}^{(t-1)}\right\| > \epsilon$ & $t \leq t_{\max}$ **do**
  Build matrix $\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}}^{(t-1)})$ in (32)
  Build matrix $\mathbf{B}_{\mathsf{KKT}}(\bar{\mathbf{v}}^{(t-1)})$ in (33)
  Compute $\bar{\mathbf{v}}^{(t)} = \mathbf{B}_{\mathsf{KKT}}^{-1}(\bar{\mathbf{v}}^{(t-1)})\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}}^{(t-1)})\bar{\mathbf{v}}^{(t-1)}$
  Normalize $\bar{\mathbf{v}}^{(t)} = \bar{\mathbf{v}}^{(t)} / \left\|\bar{\mathbf{v}}^{(t)}\right\|$
  $t \leftarrow t + 1$
$\bar{\mathbf{v}}^{\star} \leftarrow \bar{\mathbf{v}}^{(t)}$
return $\bar{\mathbf{v}}^{\star}$.

optimality condition in (31). To find such $\bar{\mathbf{v}}$, we propose a quantization-aware generalized power iteration-based direction optimization algorithm (Q-GPI-DO).

Algorithm 1 describes the proposed algorithm. The key idea used in Q-GPI-DO is to modify a power iteration method to be appliable for the corresponding functional eigenvalue problem (34). Specifically, Algorithm 1 first initializes the stacked precoding vector $\bar{\mathbf{v}}^{(0)}$. Then, the precoding vector $\bar{\mathbf{v}}^{(0)}$ is updated iteratively: at iteration t, the matrices $\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}}^{(t-1)})$ and $\mathbf{B}_{\mathsf{KKT}}(\bar{\mathbf{v}}^{(t-1)})$ are computed according to (32) and (33). Subsequently, the precoder $\bar{\mathbf{v}}^{(t)}$ is re-computed as $\bar{\mathbf{v}}^{(t)} = \mathbf{B}_{\mathsf{KKT}}^{-1}(\bar{\mathbf{v}}^{(t-1)})\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}}^{(t-1)})\bar{\mathbf{v}}^{(t-1)}$ and normalized as $\bar{\mathbf{v}}^{(t)} = \bar{\mathbf{v}}^{(t)} / \left\|\bar{\mathbf{v}}^{(t)}\right\|$. The iteration stops when one of the stopping criteria are met: either converges, i.e., $\|\bar{\mathbf{v}}^{(t)} - \bar{\mathbf{v}}^{(t-1)}\| \leq \epsilon$ where $\epsilon > 0$ denotes a tolerance level or reaches a maximum iteration count $t_{\max}$ which may differ depending on a system requirement.

One remarkable benefit of the proposed Q-GPI-DO is that it is not required to exploit any off-the-shelf optimization solver such as CVX. Distinguished from other convex-relaxation based approaches, the only computational load of the proposed method is caused when calculating $\bar{\mathbf{v}}^{(t)} = \mathbf{B}_{\mathsf{KKT}}^{-1}(\bar{\mathbf{v}}^{(t-1)})\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}}^{(t-1)})\bar{\mathbf{v}}^{(t-1)}$. As implementing CVX in practical hardware is nearly infeasible due to the high complexity, Q-GPI-DO has a substantial advantage from an implementation perspective.

2) Optimal Power Scaling $\tau^{\star}$: the problem in (22) for given $\mathbf{V}$ and μ reduces to

$$\begin{aligned} \underset{\tau}{\text{maximize}} \ & \sum_{k=1}^{K} \bar{R}_k - \frac{\mu P}{\kappa}\tau \\ \text{subject to } & 0 < \tau \leq 1 \\ & \mu > 0. \end{aligned} \tag{35}$$

The objective function in (35) is now concave with respect to τ. In this regard, the optimal τ can be derived by using a gradient descent algorithm for given $\bar{\mathbf{v}}$ and μ. Let the objective function in (35) be $L_{\tau}(\tau, \bar{\mathbf{v}})$. Then the gradient update is given as

$$\tau^{(t+1)} = \tau^{(t)} + \delta_{\mathsf{GD}} \frac{\partial L_{\tau}(\tau^{(t)}, \bar{\mathbf{v}})}{\partial \tau^{(t)}}, \tag{36}$$

where δ_{GD} denotes a step size and the partial derivative $\partial L_{\tau}(\tau^{(t)}, \bar{\mathbf{v}})/\partial \tau^{(t)}$ is computed as

$$\begin{aligned} &\frac{\partial L_{\tau}(\tau, \bar{\mathbf{v}})}{\partial \tau} \\ &= \frac{1}{\ln 2} \sum_{k=1}^{K} \left(\frac{\Xi_k + \Psi_k}{\sigma^2/P + (\Xi_k + \Psi_k)\tau} - \frac{\Psi_k}{\sigma^2/P + \Psi_k \tau} \right) - \frac{\mu P}{\kappa}, \end{aligned} \tag{37}$$

where

$$\begin{aligned} \Xi_k &= \alpha_k |\mathbf{h}_k^{\mathsf{H}} \mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2} \mathbf{v}_k|^2, \\ \Psi_k &= \sum_{\ell=1}^{K} |\mathbf{h}_k^{\mathsf{H}} \mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2} \mathbf{v}_{\ell}|^2 - \alpha_k |\mathbf{h}_k^{\mathsf{H}} \mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2} \mathbf{v}_k|^2 \\ &\quad + \sum_{\ell=1}^{K} \mathbf{v}_{\ell}^{\mathsf{H}} \mathbf{\Phi}_{\beta_{\mathsf{bs}}} \mathrm{diag}\left(\mathbf{h}_k \mathbf{h}_k^{\mathsf{H}}\right) \mathbf{v}_{\ell}. \end{aligned}$$

To decide δ_{GD}, we use a backtracking line search method [40] in simulations.

3) Thresholding of τ: Recall that we have the constraint of $0 < \tau \leq 1$. From (36), however, it is not guaranteed to satisfy the constraint. Accordingly, for each update, we perform thresholding of τ; if $\tau > 1$, then set $\tau = 1$, and if $\tau < 0$, then set $\tau = 0$.

4) Dinkelbach Update μ: Once the precoder $\mathbf{V}$ and scaling factor τ are derived, we update μ for the next outer iteration as

$$\mu = \frac{\sum_{k=1}^{K} \bar{R}_k(\mathbf{V}, \tau)}{P_{\text{approx}}}, \tag{38}$$

where

$$\begin{aligned} P_{\text{approx}} = P_{\mathsf{LO}} + \sum_{m=1}^{N} \log_2 \left(1 + \rho^{-1} \left\| \frac{1}{\sqrt{\alpha_{\mathsf{bs},m}}} \tilde{\mathbf{v}}_m \right\|^2 \right)^{\omega_{\rho}} \\ \times \left(2\, P_{\mathsf{DAC}}(b_{\mathsf{DAC},m}, f_s) + P_{\mathsf{RF}} \right) + \tau P/\kappa. \end{aligned}$$

Algorithm 2 describes the proposed quantization-aware GPI-based EE maximization algorithm (Q-GPI-EEM). With initialization, the algorithm computes τ by using the gradient descent method for given $\bar{\mathbf{v}}$ and μ and also finds $\bar{\mathbf{v}}$ by using the Q-GPI-DO algorithm for given τ and μ. Then $\mathbf{W}$ is computed from the derived τ and $\bar{\mathbf{v}}$. The algorithm repeats these steps until $\mathbf{W}$ converges. Once converged, the outer loop computes μ by using the derived $\mathbf{W}$ and repeats the previous steps until μ converges. Once μ converged, we check whether $\|\frac{1}{\sqrt{\alpha_{\mathsf{bs},n}}} \tilde{\mathbf{v}}_n\|^2 \geq \epsilon_{\mathsf{as}}$ to select antennas which have effective gain where $\epsilon_{\mathsf{as}} > 0$ is a small enough value. This selection approach is effective as the rows of $\mathbf{V}$ are jointly designed with each other, and thus, the designed $\mathbf{V}$ incorporates relative gains across the antennas. The norm of $\tilde{\mathbf{v}}_n$, however, highly depends on the number of antennas because the norm of $\bar{\mathbf{v}}$ is limited as $\|\bar{\mathbf{v}}\| = 1$. In this regard, the threshold ϵ may vary with N. To avoid this issue, we normalize $\tilde{\mathbf{v}}_n$ by $\max_i \|\frac{1}{\sqrt{\alpha_{\mathsf{bs},i}}} \tilde{\mathbf{v}}_i\|$, i.e., we perform

$$\hat{\mathbf{v}}_n = \tilde{\mathbf{v}}_n / \max_i \left\| \frac{1}{\sqrt{\alpha_{\mathrm{bs},i}}} \tilde{\mathbf{v}}_i \right\|$$

Algorithm 2 Q-GPI-EEM

initialize: $\bar{\mathbf{v}}^{(0)}$, $\tau^{(0)}$, and $\mu^{(0)}$
Set the iteration count $n_0 = 1$
while $|\mu^{(t_0)} - \mu^{(t_0-1)}|/|\mu^{(t_0)}| > \epsilon_0$ & $t_0 \leq t_{\max}$ **do**
  Set the iteration count $t_1 = 1$
  while $\left\|\mathbf{W}^{(t_1)} - \mathbf{W}^{(t_1-1)}\right\|_F / \left\|\mathbf{W}^{(t_1)}\right\|_F > \epsilon_1$ & $t_1 \leq t_{\max}$ **do**
    Set the iteration count $n_2 = 1$
    while $|\tau^{(t_2)} - \tau^{(t_2-1)}|/|\tau^{(t_2)}| > \epsilon_2$ & $t_2 \leq t_{\max}$ **do**
      Set $\partial L_\tau(\tau^{(t_2)}, \bar{\mathbf{v}})/\partial \tau^{(t_2)}$ according to (37)
      Update $\tau^{(t_2)}$ according to (36)
      Thresholding τ
      $t_2 \leftarrow t_2 + 1$
    $\bar{\mathbf{v}} = \text{Q-GPI-DO}(\mathbf{A}_{\mathsf{KKT}}(\bar{\mathbf{v}}, \tau^{(t_2)}), \mathbf{B}_{\mathsf{KKT}}(\bar{\mathbf{v}}, \tau^{(t_2)}))$
    Compute $\mathbf{W}^{(t_1)} = \sqrt{\tau^{(t_2)}}\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{-1/2}[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K]$
    $t_1 \leftarrow t_1 + 1$
  Update $\mu^{(t_0)}$ according to (38)
  $t_0 \leftarrow t_0 + 1$
Set $\tilde{\mathbf{w}}_n^{(t_1)} = \mathbf{0}_{1\times K}$ if $\|\frac{1}{\sqrt{\alpha_{\mathrm{bs},n}}}\hat{\mathbf{v}}_n\|^2 < \epsilon_{\mathsf{as}}$, $n = 1, \ldots, N$
return $\mathbf{W}^{(t_1)}$

and set $\tilde{\mathbf{w}}_n = \mathbf{0}_{1\times K}$ if $\|\frac{1}{\sqrt{\alpha_{\mathrm{bs},n}}}\hat{\mathbf{v}}_n\|^2 < \epsilon_{\mathsf{as}}$ $\forall n$, where $\tilde{\mathbf{w}}_n$ represents the nth row of $\tilde{\mathbf{W}}$. Finally, the proposed method returns $\mathbf{W}$.

Remark 3 (Complexity): The computational complexity for each step in the proposed Q-GPI-EEM is analyzed as follows. In the Q-GPI-DO step, the main computational load is caused by calculating the inversion of $\mathbf{B}_{\mathsf{KKT}}(\bar{\mathbf{f}})$. Since $\mathbf{B}_{\mathsf{KKT}}(\bar{\mathbf{f}})$ is comprised of $NK \times NK$ block-diagonal matrices, the inversion can be obtained by calculating the inversion of each sub-matrix. This results in that the complexity of Q-GPI-DO with T_{GPI} iterations is characterized as $\mathcal{O}(T_{\mathrm{GPI}}KN^3)$. The backtracking iteration has the complexity of $\mathcal{O}(\max(K^2N,\ T_{\mathrm{BT}}K))$, where T_{BT} is the number of iterations of the backtracking algorithm. Considering the iterations of the gradient descent T_{GD}, the complexity of the gradient descent with the backtracking algorithm is

$$\mathcal{O}(T_{\mathrm{GD}} \cdot \max(K^2N,\ T_{\mathrm{BT}}K)). \tag{39}$$

Note that there is some abuse of notation in (39): T_{BT} denotes the number of the backtracking iteration per gradient descent step, which may not be the same for every step. But here, we consider T_{BT} to be constant, assuming that T_{BT} will be similar for every gradient descent step for ease of analysis. This complexity analysis will apply a similar assumption, which does not change the final result. In addition to this, we have the following outer loops for both the gradient descent and Q-GPI-DO: $\mathbf{W}$ optimization loop T_W, and μ optimization loop T_μ. Therefore the overall complexity of Algorithm 2 is given as

$$\mathcal{O}\big(T_\mu T_W \cdot \max\left(T_{\mathrm{GPI}}KN^3,\ T_{\mathrm{GD}} \cdot \max(K^2N,\ T_{\mathrm{BT}}K)\right)\big). \tag{40}$$

Then, in the massive MIMO regime, it is reasonable to consider $KN > T_{\mathrm{BT}}$ and $T_{\mathrm{GPI}}KN^3 > T_{\mathrm{GD}}K^2N$. Therefore, considering T as the total iteration count of the GPI method, i.e., $T = T_{\mathrm{GPI}}T_W T_\mu$, the *overall complexity* in (40) becomes $\mathcal{O}(TKN^3)$.

Remark 4 (Convergence): It is challenging to prove the convergence of the proposed method rigorously. The main obstacle is Q-GPI-DO. As explained above, through the lens of a functional eigenvalue problem, Q-GPI-DO is interpreted as a power iteration type algorithm to find the principal eigenvector of the NEPv (31). Conventionally, the self-consistent field iteration (SCF) can be used to solve the NEPv [39], whose main idea is to perform the eigenvector decomposition iteratively based on the previously obtained eigenvectors. Extending this to the power iteration type algorithm is a key for the convergence proof of Q-GPI-DO. We shall leave this as future work. In the later section, we empirically show that μ converges very well in general system environments.

V. Special Case: Sum Spectral Efficiency Maximization

We can show that a sum SE maximization problem in the considered massive MIMO system with low-resolution DACs and ADCs is regarded as a special case of the EE maximization. Accordingly, we provide a brief description of finding precoders that maximize the sum SE by exploiting the results derived in the EE. The sum SE maximization problem is formulated by setting μ in (11) to zero as

$$\underset{\mathbf{W}}{\text{maximize}} \ \sum_{k=1}^{K} R_k \tag{41}$$

$$\text{subject to } \mathrm{tr}\left(\mathbb{E}\left[\mathbf{x}_{\mathsf{q}}\mathbf{x}_{\mathsf{q}}^{\mathsf{H}}\right]\right) \leq P. \tag{42}$$

Since the SINR in (16) increases with the transmit power, the SE is maximized when we meet

$$\mathrm{tr}\left(\boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}\mathbf{W}\mathbf{W}^{\mathsf{H}}\right) = 1, \tag{43}$$

i.e., the BS transmits signals use the maximum transmit power $\mathrm{tr}\left(\mathbb{E}\left[\mathbf{x}_{\mathsf{q}}\mathbf{x}_{\mathsf{q}}^{\mathsf{H}}\right]\right) = P$. Since this is equivalent to having $\tau = 1$, the results derived in the EE problem can be directly used by setting $\tau = 1$ and $\mu = 0$. Now let us define a weighted precoding vector as

$$\mathbf{f}_k = \boldsymbol{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2}\mathbf{w}_k \tag{44}$$

which is equivalent to $\mathbf{v}_k$ in (20) with $\tau = 1$. Then we have the following optimality condition:

Lemma 2: The first-order optimality condition of the optimization problem (41) *is satisfied if the following holds*:

$$\mathbf{C}_{\mathsf{KKT}}(\bar{\mathbf{f}})\bar{\mathbf{f}} = \lambda(\bar{\mathbf{f}})\mathbf{D}_{\mathsf{KKT}}(\bar{\mathbf{f}})\bar{\mathbf{f}}, \tag{45}$$

where

$$\mathbf{C}_{\mathsf{KKT}}(\bar{\mathbf{f}}) = \sum_{k=1}^{K} \frac{\mathbf{C}_k}{\bar{\mathbf{f}}^{\mathsf{H}}\mathbf{C}_k\bar{\mathbf{f}}} \prod_{\ell=1}^{K}\left(\bar{\mathbf{f}}^{\mathsf{H}}\mathbf{C}_\ell\bar{\mathbf{f}}\right), \tag{46}$$

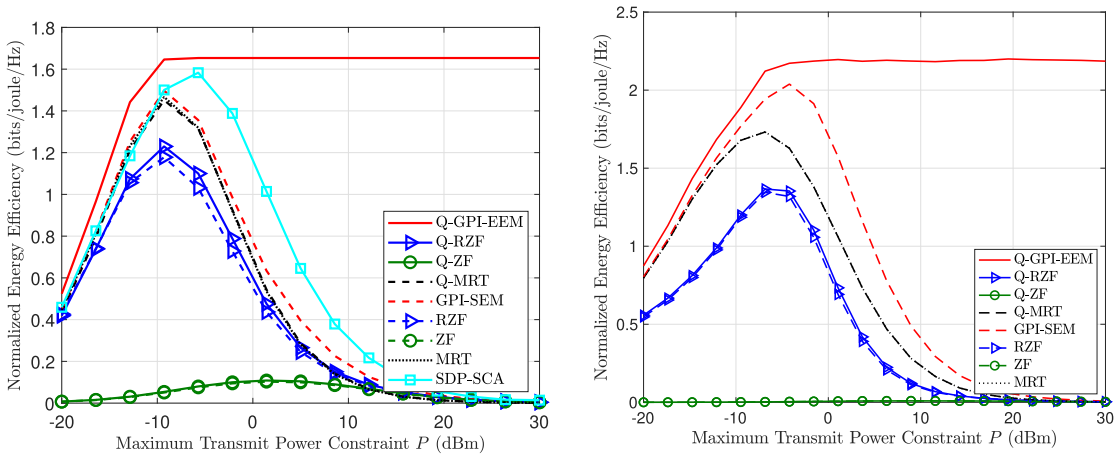


Fig. 1. The energy efficiency versus maximum transmit power constraint P results for (a) $N = 16$ BS antennas, $K = 4$ users, $b_{\mathsf{DAC},n} \sim \mathrm{Unif}[2,5]$, and $b_{\mathsf{ADC},k} \sim \mathrm{Unif}[2,5]$, and (b) $N = 32$ BS antennas, $K = 8$ users, $b_{\mathsf{DAC},n} \sim \mathrm{Unif}[2,12]$, and $b_{\mathsf{ADC},k} \sim \mathrm{Unif}[2,6]$.

$$\mathbf{D}_{\mathsf{KKT}}(\bar{\mathbf{f}}) = \sum_{k=1}^{K} \frac{\mathbf{D}_k}{\bar{\mathbf{f}}^{\mathsf{H}}\mathbf{D}_k\bar{\mathbf{f}}} \prod_{\ell=1}^{K} \left(\bar{\mathbf{f}}^{\mathsf{H}}\mathbf{D}_\ell\bar{\mathbf{f}}\right), \tag{47}$$

$$\mathbf{C}_k = \mathrm{blkdiag}\big(\mathbf{G}_k, \cdots, \mathbf{G}_k\big) + \frac{\sigma^2}{P}\mathbf{I}_{NK},$$

$$\mathbf{D}_k = \mathbf{C}_k - \mathrm{blkdiag}\Big(\mathbf{0}_{N\times N}, \cdots, \alpha_k(\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2})^{\mathsf{H}}\mathbf{h}_k\mathbf{h}_k^{\mathsf{H}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2}, \cdots, \mathbf{0}_{N\times N}\Big),$$

$$\lambda(\bar{\mathbf{f}}) = \prod_{k=1}^{K}\left(\frac{\bar{\mathbf{f}}^{\mathsf{H}}\mathbf{C}_k\bar{\mathbf{f}}}{\bar{\mathbf{f}}^{\mathsf{H}}\mathbf{D}_k\bar{\mathbf{f}}}\right).$$

Proof: We derive the results by setting $\mu = 0$ and $\tau = 1$ from Lemma 1. ■

We note that we can solve (45) by leveraging Algorithm 1 with $\mu = 0$ and $\tau = 1$; replacing (32), (33), and $\mathbf{v}_k$ with (46), (47), and $\mathbf{f}_k$, respectively and computing $\mathbf{W}$ based on (44). The algorithm provides the best precoder that maximizes the sum SE among all the stationary points. We call the quantization-aware GPI-based SE maximization algorithm as Q-GPI-SEM.

VI. Numerical Results

This section evaluates the proposed algorithms to validate the performance and draw key system design insights. We also evaluate benchmark algorithms for comparison. The following cases are included in the simulations: (1) the proposed algorithms, (2) quantization-ignorant conventional GPI-based SE maximization (GPI-SEM) [30], (3) quantization-aware linear precoders such as regularized zero-forcing (Q-RZF), zero-forcing (Q-ZF), and maximum ratio transmission (Q-MRT), and (4) conventional linear precoders such as RZF, ZF, and MRT. The quantization-aware linear precoders are derived based on the AQNM system model.

A. Simulation Environments

We adopt a one-ring model [41] to generate the channel vector $\mathbf{h}_k = \sqrt{\rho_k}\mathbf{g}_k$. To generate pathloss ρ_k, we adopt the log-distance pathloss model in [42]; cell radius is 1 km, the minimum distance between the BS and users is 100 m, pathloss exponent is 4, and 2.4 GHz carrier frequency with 100 MHz bandwidth (passband), 8.7 dB lognormal shadowing variance, and 5 dB noise figure are considered.

For the BS power consumption, we set $P_{\mathsf{LP}} = 14$ mW, $P_{\mathsf{M}} = 0.3$ mW, $P_{\mathsf{LO}} = 22.5$ mW, $P_{\mathsf{H}} = 3$ mW, and $\kappa = 0.27$ [22]. For the parameters used in the proposed algorithms, we set $t_{\mathsf{max}} = 10$, $\rho = 10^{-8}$, $\epsilon_{\mathsf{as}} = 0.05$ $\epsilon_0 = 0.001$, and ϵ_{gpi}, ϵ_1, ϵ_2 to be 0.1 unless mentioned otherwise, and set $\delta_{\mathsf{GD}} = 1$ for an initial step size which will be updated according to the backtracking line search algorithm [40]. We also initialize $\mathbf{W}^{(0)} = \mathbf{H}$, $\tau^{(0)} = 1$, and $\mu^{(0)} = 0$. Accordingly, based on the definition, $\bar{\mathbf{v}}^{(0)}$ is initialized as $\bar{\mathbf{v}}^{(0)} = \mathrm{vec}(\frac{1}{\sqrt{\tau^{(0)}}}\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2}\mathbf{W}^{(0)}) = \mathrm{vec}(\mathbf{\Phi}_{\alpha_{\mathsf{bs}}}^{1/2}\mathbf{H})$.

B. Evaluation

1) EE Comparison: We compare the proposed algorithm and other baselines in terms of the EE performance. We depict the comparison results in Fig. 1, wherein Fig. 1(a) assumes $N = 16$ and Fig. 1(b) assumes $N = 32$. The caption of Fig. 1 includes the detailed simulation setups. Particularly, in Fig. 1(a), we compare our method and the SDP based on SCA (SDP-SCA) proposed in [8], where we adjust the assumptions in [8] to fit our setup. Over the considered maximum transmit power constraint, the proposed Q-GPI-EEM algorithm achieves the highest EE performance while maintaining the highest EE once it attains the highest EE. This is because the Q-GPI-EEM suitably optimizes antenna selection and precoders to maximize the EE, by incorporating the quantization distortion effects of low-resolution quantizers. We note that although the SDP-SCA provides the second-best EE in Fig. 1(a), it is hard to use the SDP-SCA in massive MIMO systems due to its enormous computational complexity. We also note that the proposed Q-GPI-EEM is shown to be robust to the maximum transmit power constraint by maintaining its highest EE and thus, provides high EE performance regardless of the power constraint. This is because we identify the actual transmit power level τ $(0 < \tau \leq 1)$ to maximize

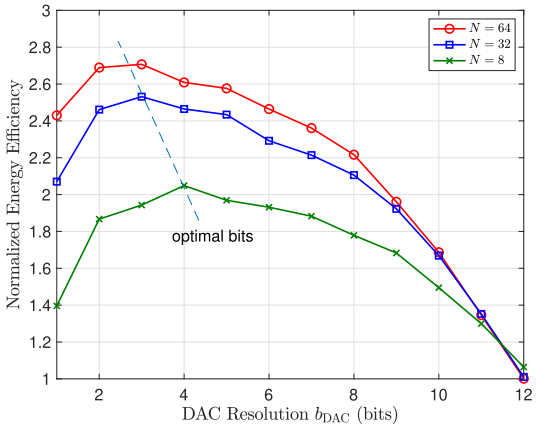


Fig. 2. The energy efficiency versus DAC resolutions b_{DAC} for $N \in \{8, 32, 64\}$ BS antennas, $K = 8$ users, $P = 30$ dBm maximum transmit power constraint, and $b_{\mathsf{ADC},k} = 10$ ADC bits.

the EE in the proposed method (the optimal power scaling $\tau^\star$ step).

2) EE by DAC Bits: The EE with respect to the number of DAC bits of the proposed Q-GPI-EEM is evaluated for $N \in \{8, 32, 64\}$ BS antennas, $K = 8$ users, $P = 30$ dBm maximum transmit power constraint, and $b_{\mathsf{ADC},k} = 10$ ADC bits in Fig. 2. In this case, the resolution of all DACs is the same, and the selection threshold is relaxed to $\epsilon_{\mathsf{as}} = 0.2$ when $N = 64$. We note that the EEs of all the cases can be maximized when 3 to 4 bits are used per DAC. In addition, the optimal number of bits tends to become smaller, and the EE increases with the number of antennas. In this regard, using more antennas with the proposed algorithm provides gains in both the SE and EE, and also it allows the BS to use coarser quantizers, thereby saving more power and simplifying each RF chain.

3) Antenna Selection Ratio by DAC Bits: Now, we present the DAC selection ratios for a different number of bits. In the simulation, we set $N = 42$ BS antennas, $K = 8$ users, $b_{\mathsf{ADC},k} \sim \mathrm{Unif}[2, 6]$, and $P = 30$ dBm maximum transmit power constraint. Each DAC resolution is assigned to three BS antennas. Accordingly, the selection ratio indicates the ratio of selected antennas within each resolution on average. For example, the selection ratio is about 36% for 13 bits, which means that only one antenna out of the three antennas with 13 bits is selected at each transmission on average. It is shown in Fig. 3(a) that about 90% of the antennas are selected for the low-to-medium resolution DACs. However, the selection ratio rapidly decreases with the number of DAC bits in the high-resolution regime since such antennas with high-resolution DACs consume unnecessarily high power, which corresponds to our intuition. Consequently, less than one antenna out of the three antennas with 14 bits is selected at each transmission on average, and no antenna with 15 bits is selected in most cases.

4) Number of Selected Antennas: In Fig. 3(b), we simulate the number of selected antennas over different parameters such as the maximum transmit power constraint, the number of DAC bits, and the number of users. For the case of the number of DAC bits, we use a homogeneous DAC bit distribution, i.e., all DACs have 3 bits. For the case of the number of users,

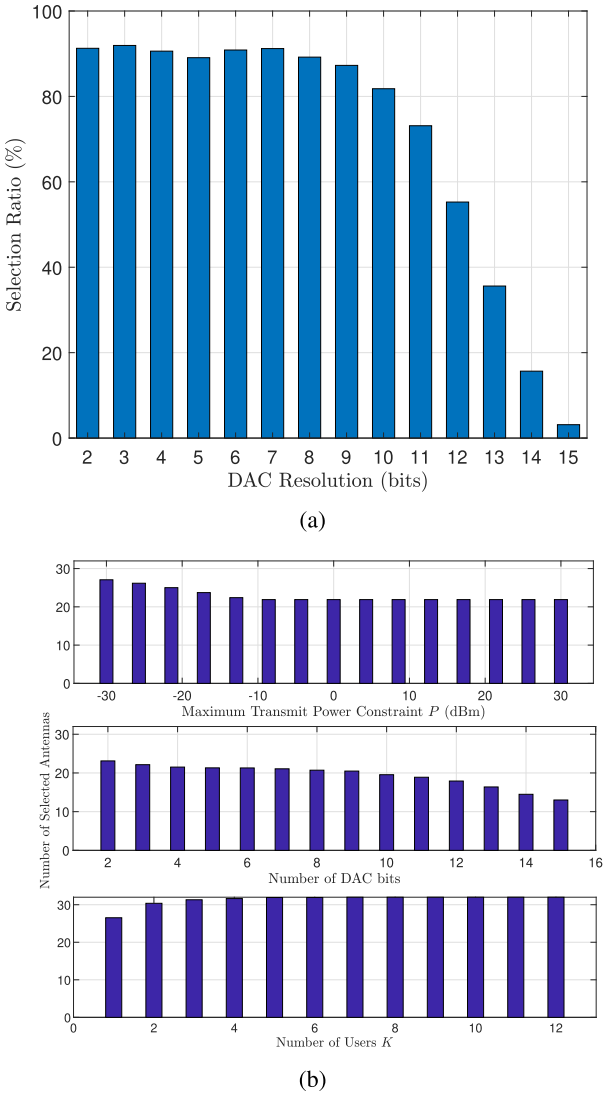


Fig. 3. (a) Selection ratio of each DAC resolution for $N = 42$ BS antennas, $K = 8$ users, $b_{\mathsf{ADC},k} \sim \mathrm{Unif}[2, 6]$, and $P = 30$ dBm maximum transmit power. Each DAC resolution is assigned to three BS antennas. (b) The number of selected antennas versus the power constraint P, number of DAC bits b_{DAC}, and number of users K for $N = 32$ BS antennas, 3 bits for DACs and ADCs, and $P = 20$ dBm power constraint unless mentioned otherwise.

we force the pathloss of all users to be -100 dB for a fair comparison. The trends of the antenna selection correspond to the general intuition: 1) when the power constraint becomes larger, the BS can use transmit power more effectively, thereby turning off more antennas to save RF power. 2) When the number of DAC bits increases, more antennas tend to be turned off to save RF power. 3) When the number of users increases, using more antennas can bring greater SE improvement, leading to higher EE, thereby selecting more BS antennas. We note that the proposed EE method does not increase the transmit power once it reaches the highest EE. Accordingly, the number of selected antennas remains the same beyond a certain P.

5) Convergence: To exam the convergence of the proposed Q-GPI-EEM algorithm, we provide the numerical convergence μ. In Fig. 4, we verify the convergence of μ under the assumptions that $N = 32$ BS antennas, $K = 8$

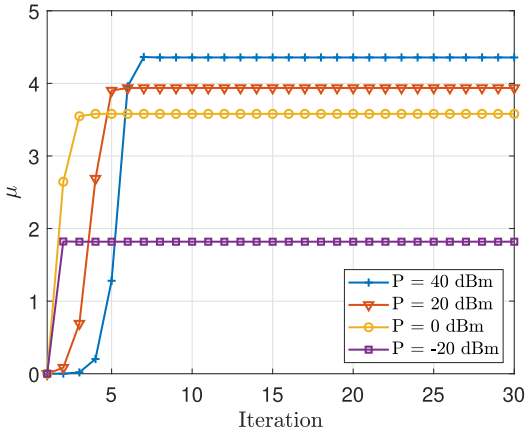


Fig. 4. Convergence of μ for $N = 32$ BS antennas, $K = 8$ users, $b_{\mathrm{DAC},n} \sim \mathrm{Unif}[2, 12]$, $b_{\mathrm{ADC},k} \sim \mathrm{Unif}[2, 6]$, and $P \in \{-20, 0, 20, 40\}$ dBm maximum transmit power constraint.

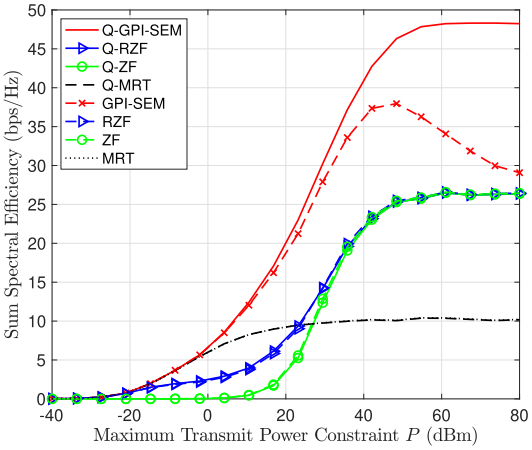
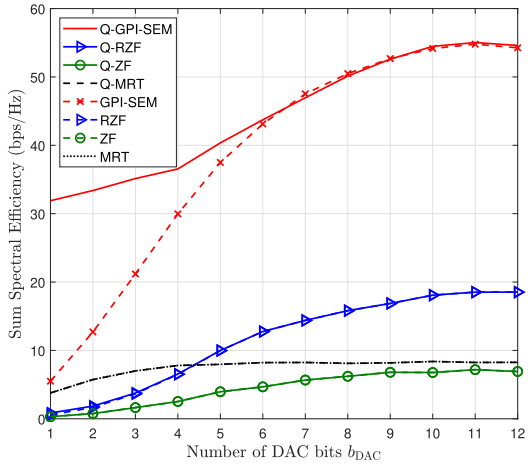
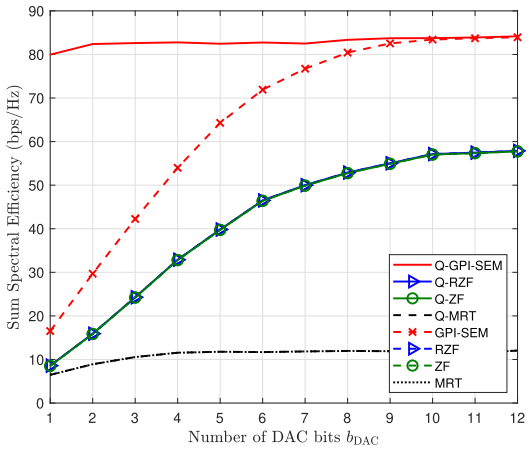


Fig. 5. The sum spectral efficiency versus maximum transmit power constraint P results for $N = 32$ BS antennas, $K = 8$ users, $b_{\mathrm{DAC},n} \sim \mathrm{Unif}[2, 12]$, and $b_{\mathrm{ADC},k} \sim \mathrm{Unif}[2, 6]$.



(a) $N = 8$ BS antennas



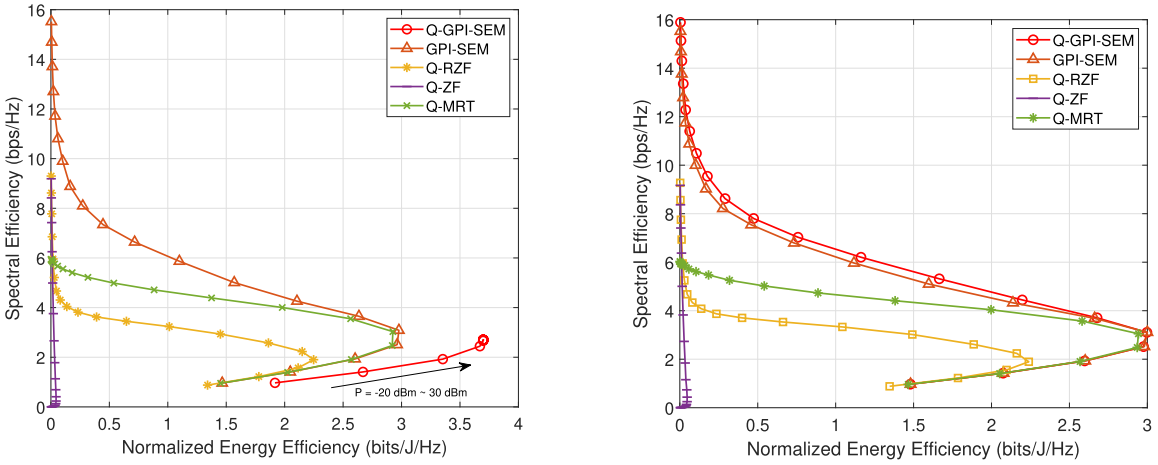
(b) $N = 32$ BS antennas

Fig. 6. The sum spectral efficiency versus the number of DAC bits b_{DAC} for $K = 8$ users, $b_{\mathrm{ADC},k} = 10$ ADC bits for all k, $P = 50$ dBm maximum transmit power constraint, and $N \in \{8, 32\}$ BS antennas.

users, $b_{\mathrm{DAC},n} \sim \mathrm{Unif}[2, 12]$, $b_{\mathrm{ADC},k} \sim \mathrm{Unif}[2, 6]$, and $P \in \{-20, 0, 20, 40\}$ dBm maximum transmit power constraint. As we observe in Fig. 4, μ converges fast for all the considered cases. Although the convergence takes longer with higher P, the results in Fig. 4 still show that μ converges within 8 iterations, which can guarantee fast convergence of Q-GPI-EEM in the practical transmit power regime. We note that the convergence of μ means the convergence of the EE since we reformulate the fractional programming by using the Dinkelbach approach as in (11).

6) SE Comparison: Fig. 5 shows the sum SE versus maximum transmit power constraint P results for $N = 32$ BS antennas, $K = 8$ users, $b_{\mathrm{DAC},n} \sim \mathrm{Unif}[2, 12]$, and $b_{\mathrm{ADC},k} \sim \mathrm{Unif}[2, 6]$. As shown in Fig. 5, Q-GPI-SEM achieves the highest sum SE over the most range of P. As the maximum transmit power constraint increases, the SE is saturated because the quantization distortion also increases. The proposed algorithm can pull up this saturation level more than 2 times that of the conventional methods. We note that GPI-SEM provides a higher sum SE than the RZF, ZF, and MRT-based precoders. GPI-SEM, however, shows a huge gap from Q-GPI-SEM as P increases, i.e., quantization noise also increases. Moreover, its SE even decreases in the very high transmit power regime because the interference from the quantization error, which cannot be fully treated with GPI-SEM dominates the SE performance in the regime. In the Q-RZF/RZF, Q-ZF/ZF, and Q-MRT/MRT cases, the sum SE performance is not comparable with that of Q-GPI-SEM except in the very low transmit power regime where the quantization error is buried in the thermal noise. Therefore, Fig. 5 validates the sum SE performance of the proposed method over the practical transmit power regime.

7) SE by DAC Bits: Fig. 6 shows the sum SE versus the number of DAC bits b_{DAC} for $K = 8$ users, $b_{\mathrm{ADC},k} = 10$ ADC bits for all k, $P = 50$ dBm maximum transmit power constraint, and $N \in \{8, 32\}$ BS antennas. In this case, the entire DACs have the same resolution. We first note that Q-GPI-SEM provides the highest sum SE, and the sum SE of GPI-SEM converges to that of Q-GPI-SEM as the number of DAC bits increases since both the DACs and ADCs have high resolutions. More importantly, the performance variation of Q-GPI-SEM over DAC resolutions becomes marginal in



(a) EE maximization (b) SE maximization

Fig. 7. The spectral efficiency versus the normalized energy efficiency results for $N = 16$ BS antennas, $K = 4$ users, $b_{\mathsf{DAC},k} = 3$ DAC bits, and $b_{\mathsf{ADC},k} = 3$ ADC bits.

$N = 32$ case compared to $N = 8$ case, whereas the other algorithms still show high-performance variations. Therefore, in the massive MIMO, the proposed algorithm is highly robust to quantization noise at the BS by achieving the SE of high-resolution DACs with low-resolution DACs ($2 \sim 5$ bits).

Overall, the proposed Q-GPI-SEM algorithm outperforms the conventional precoding algorithms, and it is indeed an efficient method in the massive MIMO communication systems with low-resolution DACs or ADCs, showing robustness to quantization error.

8) SE-EE Tradeoff: We also provide the SE-EE trade-off results with the proposed EE maximization method and SE maximization method, including other benchmarks. In Fig. 7, we plot the SE-EE tradeoffs over $P = -20$ dBm to $P = 30$ dBm maximum transmit power constraint. Fig. 7(a) shows the tradeoff result with the proposed EE maximization algorithm. In the beginning, both the EE and SE increase with P. Then the EE of the benchmarks decreases beyond certain P, whereas the EE of the proposed algorithm never decreases. Fig. 7(b) shows the tradeoff result with the proposed SE maximization algorithm. The EE of the proposed SE maximization method also decreases beyond certain P while providing the best SE-EE tradeoff among all. We note that the EE maximization algorithm achieves the highest EE performance with the limited SE performance by not fully using P. The achieved SE, in our algorithm, can be adjusted to a desirable level by introducing a control parameter $\nu > 0$ in the BS power consumption of the EE optimization problem as $\nu P_{\mathsf{BS}}(N, \mathbf{b}_{\mathsf{DAC}}, f_s, \mathbf{W})$.

9) Symbol Error Rate (SER): In Fig. 8, we present SER results. We use 4-QAM modulation with $N = 16$ BS antennas, $K = 8$ users, 3 bits for all DACs and ADCs, and MMSE scalar quantization. As shown in Fig. 8, the proposed method achieves the best SER, followed by GPI-SEM. Similar to the SE results presented in the manuscript, the GPI-SEM also shows the inverse trend in the high SNR regime due to the quantization effect. The Q-RZF shows the worst SER results with significant gaps from the others. Because the user channel

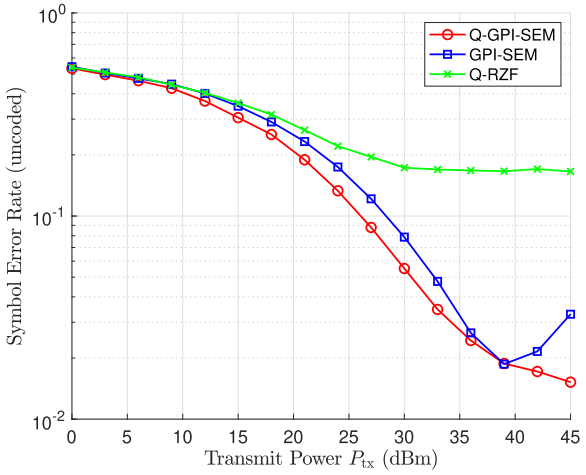
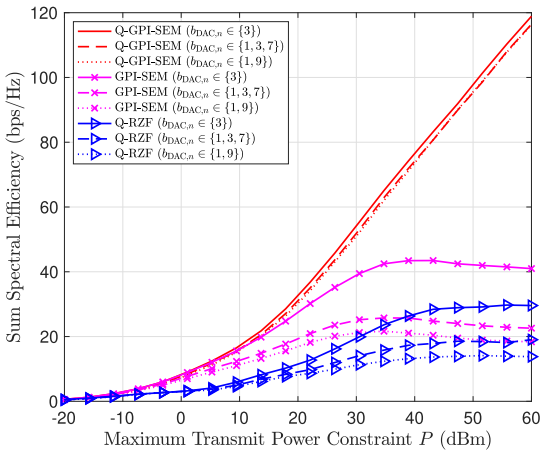


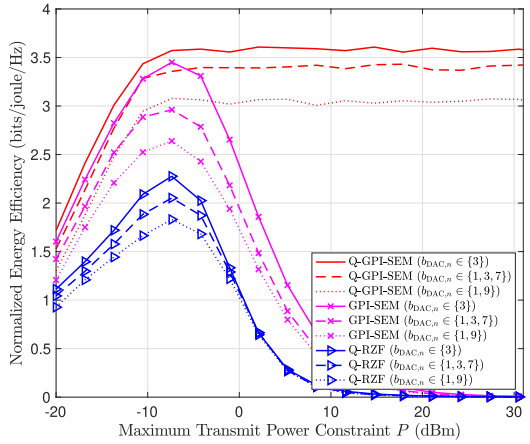
Fig. 8. Symbol error rates of 4-QAM modulation with $N = 16$ BS antennas, $K = 12$ users, 3 bits for all DACs and ADCs, and MMSE scalar quantization.

gains are highly heterogeneous, the Q-RZF that involves a matrix inversion suffers in the high SNR and results in poor SER for users with severe pathloss.

10) SE by DAC Configuration: Finally, to provide system design insights, we compare various DAC configurations such as (i) $b_{\mathsf{DAC},n} = 3$, (ii) $b_{\mathsf{DAC},n} \in \{1, 3, 7\}$, and (iii) $b_{\mathsf{DAC},n} \in \{1, 9\}$ under the constraint of the total number of DAC resolution bits (96 bits total) for $K = 8$, $b_{\mathsf{ADC},k} = 10$ for all k, and $N = 32$. Fig. 9 reveals that the homogeneous DAC configuration where all DACs have the same resolution achieves the highest sum SE and EE. In particular, the homogeneous DAC configuration provides a noticeable improvement in the EE since medium- and high-resolution DACs in the other configurations cause inefficiency in power consumption. In addition, the proposed algorithm shows a relatively small variation in the sum SE across different DAC configurations, whereas the other algorithms reveal a noticeable performance gap across configurations. Therefore, the result demonstrates that the proposed algorithm is more



(a) Spectral Efficiency



(b) Energy Efficiency

Fig. 9. The spectral efficiency and energy efficiency versus maximum transmit power constraint P for $K = 8$ users, $b_{\mathrm{ADC},k} = 10$ ADC bits for all k, and $N = 32$ BS antennas with different DAC configurations.

robust to DAC configurations, providing more system design flexibility.

VII. Conclusion

In this paper, we solved a precoding problem for EE maximization in downlink multiuser massive MIMO systems with low-resolution DACs and ADCs. To take into account the effects of RF circuit power consumption, we incorporated an antenna selection feature into the EE maximization problem. Managing the quantization errors, we reformulated the SINR and adopted the Dinkelbach method. Subsequently, we decomposed the problem into precoding direction and power scaling problems and proposed the joint precoding and antenna selection algorithm. As a special case, we showed that the proposed algorithm can reduce to the SE maximization algorithm by leveraging the product of Rayleigh quotients form of the reformulated SINR. The simulation results demonstrated that the proposed algorithms improve both EE and SE compared to conventional methods. In particular, the EE maximization algorithm presented robustness to the maximum transmit power constraint with fast convergence, while other methods suffer from EE degradation as the maximum transmit power increases. In addition, it was shown that the proposed methods achieve high enough EE and SE even with low-resolution DACs in the massive MIMO regime, which means that the performance degradation caused by low-resolution quantizers can be compensated by using our method with large-scale arrays. As a result, the proposed algorithms can provide considerable benefits in the future massive MIMO systems by offering high flexibility on quantizer configuration and improving the SE and EE performance. Considering wideband systems with low-resolution quantizers for maximizing the EE would be a desirable future research direction.

References

[1] 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.
[2] J. Zhang, L. Dai, X. Li, Y. Liu, and L. Hanzo, "On low-resolution ADCs in practical 5G millimeter-wave massive MIMO systems," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 205–211, Jul. 2018.
[3] J. Choi, G. Lee, A. Alkhateeb, A. Gatherer, N. Al-Dhahir, and B. L. Evans, "Advanced receiver architectures for millimeter-wave communications with low-resolution ADCs," *IEEE Commun. Mag.*, vol. 58, no. 8, pp. 42–48, Aug. 2020.
[4] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.
[5] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Linear precoding with low-resolution DACs for massive MU-MIMO-OFDM downlink," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1595–1609, Mar. 2019.
[6] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.
[7] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.
[8] S. He, Y. Huang, J. Wang, L. Yang, and W. Hong, "Joint antenna selection and energy-efficient beamforming design," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1165–1169, Sep. 2016.
[9] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Energy efficiency in massive MIMO-based 5G networks: Opportunities and challenges," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 86–94, Jun. 2017.
[10] J. Singh, O. Dabeer, and U. Madhow, "On the limits of communication with low-precision analog-to-digital conversion at the receiver," *IEEE Trans. Commun.*, vol. 57, no. 12, pp. 3629–3639, Dec. 2009.
[11] A. Mezghani and J. A. Nossek, "Analysis of Rayleigh-fading channels with 1-bit quantized output," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 260–264.
[12] J. Mo and R. W. Heath, Jr., "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5498–5512, Oct. 2015.
[13] O. De Candido, H. Jedda, A. Mezghani, A. L. Swindlehurst, and J. A. Nossek, "Reconsidering linear transmit signal processing in 1-bit quantized multi-user MISO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 254–267, Jan. 2019.
[14] H. Pirzadeh, G. Seco-Granados, S. Rao, and A. L. Swindlehurst, "Spectral efficiency of one-bit sigma-delta massive MIMO," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2215–2226, Sep. 2020.
[15] M. Shao, W.-K. Ma, Q. Li, and A. L. Swindlehurst, "One-bit sigma-delta MIMO precoding," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 1046–1061, Sep. 2019.
[16] M. Shao, Q. Li, W.-K. Ma, and A. M.-C. So, "A framework for one-bit and constant-envelope precoding over multiuser massive MISO channels," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5309–5324, Oct. 2019.
[17] H. Jedda, A. Mezghani, A. L. Swindlehurst, and J. A. Nossek, "Quantized constant envelope precoding with PSK and QAM signaling," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8022–8034, Dec. 2018.
[18] A. Mezghani, R. Ghiat, and J. A. Nossek, "Transmit processing with low resolution D/A-converters," in *Proc. 16th IEEE Int. Conf. Electron., Circuits Syst. (ICECS)*, Dec. 2009, pp. 683–686.

[19] J. J. Bussgang, "Crosscorrelation functions of amplitude-distorted Gaussian signals," Res. Lab. Electron., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. 216, 1952.
[20] A. K. Saxena, I. Fijalkow, and A. L. Swindlehurst, "Analysis of one-bit quantized precoding for the multiuser massive MIMO downlink," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4624–4634, Sep. 2017.
[21] J.-C. Chen, "Alternating minimization algorithms for one-bit precoding in massive multiuser MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7394–7406, Aug. 2018.
[22] L. N. Ribeiro, S. Schwarz, M. Rupp, and A. L. F. de Almeida, "Energy efficiency of mmWave massive MIMO precoding with low-resolution DACs," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 298–312, May 2018.
[23] J. Dai, J. Liu, J. Wang, J. Zhao, C. Cheng, and J.-Y. Wang, "Achievable rates for full-duplex massive MIMO systems with low-resolution ADCs/DACs," *IEEE Access*, vol. 7, pp. 24343–24353, 2019.
[24] J. Choi, J. Sung, N. Prasad, X.-F. Qi, B. L. Evans, and A. Gatherer, "Base station antenna selection for low-resolution ADC systems," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1951–1965, Mar. 2020.
[25] E. Vlachos and J. Thompson, "Energy-efficiency maximization of hybrid massive MIMO precoding with random-resolution DACs via RF selection," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1093–1104, Feb. 2021.
[26] C.-J. Wang, C.-K. Wen, S. Jin, and S.-H. Tsai, "Finite-alphabet precoding for massive MU-MIMO with low-resolution DACs," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4706–4720, Jul. 2018.
[27] Q. Ding, Y. Deng, and X. Gao, "Spectral and energy efficiency of hybrid precoding for mmWave massive MIMO with low-resolution ADCs/DACs," *IEEE Access*, vol. 7, pp. 186529–186537, 2019.
[28] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Robust predictive quantization: Analysis and design via convex optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 618–632, Dec. 2007.
[29] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.
[30] J. Choi, N. Lee, S.-N. Hong, and G. Caire, "Joint user selection, power allocation, and precoding design with imperfect CSIT for multi-cell MU-MIMO downlink systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 162–176, Jan. 2020.
[31] J. Zhang, L. Dai, Z. He, B. Ai, and O. A. Dobre, "Mixed-ADC/DAC multipair massive MIMO relaying systems: Performance analysis and power optimization," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 140–153, Jan. 2019.
[32] L. Fan, S. Jin, C. K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2186–2189, Oct. 2015.
[33] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, vol. 159. Cham, Switzerland: Springer, 2012.
[34] J. Choi, Y. Cho, and B. L. Evans, "Quantized massive MIMO systems with multicell coordinated beamforming and power control," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 946–961, Feb. 2021.
[35] 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.
[36] O. T. Demir and E. Bjornson, "The Bussgang decomposition of nonlinear systems: Basic theory and MIMO extensions [lecture notes]," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 131–136, Jan. 2021.
[37] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2349–2360, Sep. 2005.
[38] B. K. Sriperumbudur, D. A. Torres, and G. R. G. Lanckriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Mach. Learn.*, vol. 85, nos. 1–2, pp. 3–39, Oct. 2011.
[39] Y. Cai, L.-H. Zhang, Z. Bai, and R.-C. Li, "On an eigenvector-dependent nonlinear eigenvalue problem," *SIAM J. Matrix Anal. Appl.*, vol. 39, no. 3, pp. 1360–1382, Jan. 2018.
[40] L. Armiju, "Minimization of functions having Lipschitz continuous first partial derivatives," *Pacific J. Math.*, vol. 16, no. 1, pp. 1–3, 1996.
[41] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.
[42] V. Erceg *et al.*, "An empirically based path loss model for wireless channels in suburban environments," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 7, pp. 1205–1211, Jul. 1999.

Jinseok Choi (Member, IEEE) received the B.S. degree from the Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2014, and the M.S. and Ph.D. degrees in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2016 and 2019, respectively. He was a Student Member of the Wireless Networking and Communications Group (WNCG) and the Embedded Signal Processing Laboratory, under supervision of Prof. Brian L. Evans. He was also a Senior System Engineer at the Qualcomm Wireless Research and Development, San Diego, CA, USA. He is currently an Assistant Professor at the Ulsan National Institute of Science and Technology (UNIST). His primary research interest is to develop and analyze future wireless communication systems and to develop algorithms for intelligent devices that require ultra-high speed, high-reliability, and low-latency communications.

Jeonghun Park (Member, IEEE) received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2010 and 2012, respectively, and the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2017. He is currently working as an Assistant Professor with the School of Electronics Engineering, Kyungpook National University (KNU), Daegu, South Korea. Prior to joining KNU, he worked at Qualcomm Wireless Research and Development, San Diego, CA, USA. His main research interests include developing and analyzing future wireless communication systems using tools of optimization, information theory, and machine learning.

Namyoon Lee (Senior Member, IEEE) received the Ph.D. degree from The University of Texas at Austin, Austin, TX, USA, in 2014. He was with the Communications and Network Research Group, Samsung Advanced Institute of Technology, South Korea, from 2008 to 2011, and Wireless Communications Research, Intel Labs, Santa Clara, CA, USA, from 2015 to 2016. He is currently an Associate Professor at Korea University. His main research interests include communications, sensing, and machine learning. He was a recipient of the 2016 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award, the 2020 IEEE Best YP Award (Outstanding Nominee), the 2021 IEEE-IEIE Joint Award for Young Engineer and Scientist, and the 2021 Heading Young Researcher Award. Since 2021, he has been an Associate Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.