Two-Stage Analog Combining in Hybrid Beamforming Systems with Low-Resolution ADCs

Jinseok Choi, Student Member, IEEE, Gilwon Lee, Member, IEEE, and Brian L. Evans, Fellow, IEEE

Abstract—In this paper, we investigate hybrid analog/digital beamforming for multiple-input multiple-output (MIMO) systems with low-resolution analog-to-digital converters (ADCs) for millimeter wave (mmWave) communications. In the receiver, we propose to split the analog combining subsystem into a channel gain aggregation stage followed by a spreading stage. Both stages use phase shifters. Our goal is to design the two-stage analog combiner to optimize mutual information (MI) between the transmitted and quantized signals by effectively managing quantization error. To this end, we formulate an unconstrained MI maximization problem without a constant modulus constraint on analog combiners, and derive a two-stage analog combining solution. The solution achieves the optimal scaling law with respect to the number of radio frequency chains and maximizes the MI for homogeneous singular values of a MIMO channel. We further develop a two-stage analog combining algorithm to implement the derived solution for mmWave channels. By decoupling channel gain aggregation and spreading functions from the derived solution, the proposed algorithm implements the two functions by using array response vectors and a discrete Fourier transform matrix under the constant modulus constraint on each matrix element. Therefore, the proposed algorithm provides a near optimal solution for the unconstrained problem, whereas conventional hybrid approaches offer a near optimal solution only for a constrained problem. The closed-form approximation of the ergodic rate is derived for the algorithm, showing that a practical digital combiner with two-stage analog combining also achieves the optimal scaling law. Simulation results validate the algorithm performance and the derived ergodic rate.

Index Terms—Two-stage analog combining structure, low-resolution ADCs, mutual information, ergodic rate.

I. INTRODUCTION

Millimeter wave communications have emerged as a promising technology for 5G communications [1], [2]. Utilizing multi-gigahertz bandwidth in 30-300 GHz frequency ranges enables cellular networks to achieve an order of magnitude increase in achievable rate [3], and a large number of antennas can be packed into transceivers with very small antenna spacing by leveraging the very small wavelength. Due to the large number of radio frequency (RF) chains and power-demanding high-resolution ADCs coupled with high sampling rates, however, the significant power consumption at the receivers becomes one of the primary challenges to resolve. In this paper, we consider hybrid MIMO receivers with low-resolution ADCs for mmWave communications to address such a challenge by reducing both the number of RF chains and quantization resolution of ADCs. We propose a two-stage analog combining receiver architecture to maximize the mutual information by effectively managing quantization error as shown in Fig. 1.

A. Prior Work

Hybrid beamforming architectures have been widely investigated to reduce the number of RF chains with minimum communication performance degradation. Singular value decomposition (SVD)-based analog combining designs were proposed [4]–[6] as the SVD transceiver maximizes the channel capacity. In [4], hybrid precoder and combiner design methods were developed by extracting the phases of the elements of the singular vectors. Considering correlated channels, the SVD of the MIMO channel covariance matrix was used for analog combiner design to maximize mutual information in [5]. The performance of hybrid precoding systems was analyzed for MIMO downlink communications [7], [8]. It was shown that hybrid beamforming systems with a small number of RF chains can achieve the performance comparable to fully digital beamforming systems. For MIMO uplink communications, the Gram-schmidt based analog combiner design algorithm was developed in [9] to orthogonalize multiuser signals.

For mmWave channels, hybrid beamforming techniques were proposed by exploiting the limited scattering of the channels [10]–[17]. Adopting array response vectors (ARVs) for analog beamformer design, orthogonal matching pursuit (OMP)-based algorithms were developed in [10]–[14]. The proposed OMP-based algorithm in [10] approximates the minimum mean squared error (MMSE) combiner with a fewer number of RF chains than the number of antennas by using ARV-based analog combiners. The OMP-based algorithm in [10] was further improved by combining OMP and local search to reduce the computational complexity [13] and by iteratively updating the phases of the phase shifters [14].
A channel estimation technique was also proposed by using hierarchical multi-resolution codebook-based ARVs with low training overhead in [11]. By leveraging the sparse nature of mmWave channels, the proposed algorithms with ARV-based analog beamformers achieved the comparable performance with greatly reduced cost and power consumption compared to fully digital systems.

While the previous studies [4]–[17] considered infinite-resolution ADCs in hybrid MIMO systems, hybrid beamforming systems with low-resolution ADCs were investigated in [13]–[23] to take advantage of both the hybrid beamforming and low-resolution ADC architectures. The proposed algorithm in [13] attempted to design an analog combiner by minimizing the MSE including the quantization error. The analog combiner, however, is not constrained with a constant modulus, and the entire combining matrix needs to be designed for each transmitted symbol separately. Without considering the coarse quantization effect in combiner design, bit allocation techniques [19] and user scheduling methods [20] were developed for a given ARV-based analog combiner. In [21], [22], an alternating projection method was adopted to implement SVD-based analog combiners. The performance analysis of hybrid MIMO systems with low-resolution ADCs in [21] showed the superior tradeoff between performance and power consumption compared to fully digital systems and hybrid systems with infinite-resolution ADCs. In [23], a subarray antenna structure was considered, and an ARV-based combining algorithm was used to select the ARV that maximizes the aggregated channel gain. Although the analysis in [21]–[23] provided useful insights for the hybrid architecture with low-resolution ADCs such as the achievable rate and power tradeoff, the quantization error was not explicitly taken into account in the hybrid beamformer design. Consequently, considering the coarse quantization effect in the analog combiner design is still an open question.

B. Contributions

In this paper, we derive a near optimal analog combining solution for an unconstrained MI maximization problem in hybrid MIMO systems with low-resolution ADCs. We, then, propose a two-stage analog combining architecture to properly implement the derived solution under a constant modulus constraint on each phase shifter. Splitting the solution into a channel gain aggregation stage by using ARVs and a gain spreading stage by using a discrete Fourier transform (DFT) matrix, the two-stage analog combining structure realizes the derived near optimal combining solution with phase shifter-based analog combiners for mmWave communications. The contributions of this paper can be summarized as follows:

- We further develop an ARV-based two-stage analog combining algorithm to implement the derived solution for mmWave channels under the constant modulus constraint on each phase shifter. Decoupling the channel gain aggregation and spreading functions from the solution, the algorithm implements the aggregation and spreading functions by using ARVs and a DFT matrix without losing the optimality of the solution in the large antenna array regime. Therefore, the two-stage analog combiner obtained from the proposed algorithm under the constant modulus constraint also provides a near optimal solution for the unconstrained MI maximization problem, whereas conventional hybrid approaches offer a near optimal solution only for a constrained problem. Since the DFT matrix is independent of channels, only passive phase shifters need to be appended to a conventional hybrid MIMO architecture with marginal complexity and cost increase, while achieving a large MI gain.

- We derive a closed-form approximation of the ergodic rate with a maximum ratio combining (MRC) digital combiner for the proposed algorithm. The derived rate characterizes the ergodic rate performance of the proposed two-stage analog combining architecture in terms of the system parameters including quantization resolution. The derived rate reveals that the ergodic rate of the MRC combiner achieves the same optimal scaling law with the proposed two-stage analog combiner by reducing the quantization error as the number of RF chains increases.

Simulation results demonstrate that the proposed two-stage analog combining algorithm outperforms conventional algorithms and validate the derived ergodic rate.

Notation: A is a matrix and a is a column vector. \(A^H\) and \(A^T\) denote conjugate transpose and transpose. \([A]_{i,:}\) and \(a_i\) indicate the \(i\)th row and column vector of A. We denote \(a_{i,j}\) or \([A]_{i,j}\) as the \((i,j)\)th element of A and \(a_i\) as the \(i\)th element of a. \(\lambda_i(A)\) denotes the \(i\)th largest singular value of A. \(CN(\mu, \sigma^2)\) is the complex Gaussian distribution with mean \(\mu\) and variance \(\sigma^2\). \(E[\cdot]\) and \(V[\cdot]\) represent an expectation and variance operators, respectively. The correlation matrix is denoted as \(R_{xy} = E[xy^H]\). The diagonal matrix \(\text{diag}\{\lambda\}\) has \(\{\lambda_i\}\) at its \(i\)th diagonal entry, and \(\text{diag}\{a\}\) or \(\text{diag}\{a^T\}\) has \(\{a_i\}\) at its \(i\)th diagonal entry. \(\text{blkdiag}\{A_1, \ldots, A_N\}\) is a block diagonal matrix with diagonal entries \(A_1, \ldots, A_N\). I denotes the identity matrix with a proper dimension and we indicate the dimension \(N\) by \(I_N\) if necessary. \(\hat{0}\) denotes a matrix that has all zeros in its elements with a proper dimension. \(\|A\|\) represents \(L_2\) norm. \(|\cdot|\) indicates an absolute value, cardinality, and determinant for a scalar value \(a\), a set \(\mathcal{A}\), and a matrix \(A\), respectively. \(\text{Tr}\{\cdot\}\) is a trace operator and \(x(N) \sim y(N)\) indicates \(\lim_{N \to \infty} \frac{x}{y} = 1\).
II. SYSTEM MODEL

We consider single-cell uplink wireless communications in which the BS is equipped with $N_r$ receive antennas and $N_{RF}$ RF chains with $N_{RF} < N_r$. The antennas are uniform linear arrays (ULA), and each RF chain is followed by a pair of low-resolution ADCs. We assume that the BS serves $N_u$ users each with a single transmit antenna with $N_u \leq N_{RF}$.

A. Channel Model

The channel $h_{\gamma,k}$ of user $k$ is assumed to be the sum of the contributions of scatterers that contribute $L_k$ propagation paths to the channel $h_{\gamma,k}$ \cite{24}. For mmWave channels, the number of channel paths $L_k$ is expected to be small due to the limited scattering \cite{24}. The discrete-time narrowband channel of user $k$ can be modeled as

$$h_{\gamma,k} = \frac{1}{\sqrt{L_k}} h_k = \sqrt{\frac{N_r}{\gamma_k}} \sum_{\ell=1}^{L_k} g_{\ell,k} a(\phi_{\ell,k})$$

where $\gamma_k$ denotes the pathloss of user $k$, $g_{\ell,k}$ is the complex gain of the $\ell$th propagation path of user $k$, and $a(\phi_{\ell,k})$ is the ARV of the receive antennas corresponding to the azimuth AoA of the $\ell$th path of the $k$th user $\phi_{\ell,k} \in [-\pi/2, \pi/2]$. The complex channel gain $g_{\ell,k}$ follows an independent and identically distributed (i.i.d.) complex Gaussian distribution, $g_{\ell,k} \sim C N(0,1)$. The ARV $a(\theta)$ for the ULA antennas of BS is given as

$$a(\theta) = \frac{1}{\sqrt{N_r}} \left[ 1, e^{-j\pi \theta}, e^{-j2\pi \theta}, \ldots, e^{-j(N_r-1)\pi \theta} \right]^T$$

where the spatial angle $\theta = \frac{2d \sin(\theta)}{\lambda}$ is related to the physical AoA $\theta$, $d$ is the distance between antennas, and $\lambda$ is the signal wavelength. We use $\phi$ and $\theta$ to denote the physical AoAs of a user channel and physical angles of analog combiners, respectively. We also use $\varphi$ and $\vartheta$ to denote the spatial angles for $\phi$ and $\theta$, respectively, where $\varphi, \vartheta \in [-1, 1]$.

B. Signal and Quantization Model

For simplicity, we consider a homogeneous long-term received SNR network \cite{25} where a conventional uplink power control compensates for the pathloss and shadowing effect to achieve the same long-term received SNR target for all users in the cell \cite{25, 26}. Let $x = Ps$ be the transmitted user signals where $P = \text{diag}\{\sqrt{\rho_1}, \ldots, \sqrt{\rho_{N_u}}\}$ is the transmit power matrix and $s$ is the $N_u \times 1$ transmitted symbol vector from $N_u$ users. Further, let $H = HB$ represent the $N_r \times N_u$ channel matrix where $B = \text{diag}\{\sqrt{1/\gamma_1}, \ldots, \sqrt{1/\gamma_{N_u}}\}$. The received baseband analog signal vector is given as

$$r = Hs + n = HBP\bar{s} + n = \sqrt{\rho}Hs + n$$

where $n$ indicates the $N_r \times 1$ additive white noise vector. We assume zero mean and unit variance for the user symbols $s$ and noise $n$. The noise follows the complex Gaussian distribution $n \sim C N(0, I_{N_u})$ and thus, we consider $\rho$ to be the SNR.

After the BS receives the signals from users, the signals are combined via two analog combiners as shown in Fig. 1. Then, the received baseband analog signal vector becomes

$$y = \sqrt{\rho}W_{RF}^H W_{RF}^H Hs + W_{RF}^H W_{RF}^H n$$

where $W_{RF} = W_{RF}^H W_{RF}^H$ denotes the two-stage analog combiner, $W_{RF1} \in C^{N_r \times N_{RF}}$ is the first analog combiner, and $W_{RF2} \in C^{N_{RF} \times N_{RF}}$ is the second analog combiner. Each real and imaginary part of the combined signal are quantized at ADCs with $b$ quantization bits. Assuming a MMSE scalar quantizer and Gaussian signaling, $s \sim C N(0, I_{N_u})$, we adopt an additive quantization noise model (AQNM) \cite{27} which shows reasonable accuracy in the low to medium SNR ranges \cite{28}. The AQNM approximates the quantization process in linear form, which is equivalent to the approximation with Bussgang decomposition for low-resolution ADCs \cite{29}. The quantized signal vector is expressed as \cite{27, 29}

$$y_q = \mathcal{Q}(y) = \alpha_b \sqrt{\rho}W_{RF}^H Hs + \alpha_b W_{RF}^H n + q$$

where $\mathcal{Q}(\cdot)$ is the element-wise quantizer, the scalar quantization gain is $\alpha_b = 1 - \beta_b$ where $\beta_b = \mathbb{E}[|y - y_q|^2]/\mathbb{E}[|y|^2]$, and $q$ denotes the quantization noise vector. For $b > 5$ quantization bits, $\beta_b$ is approximated as $\beta_b \approx \frac{\sqrt{2}}{2} 2^{-b}$. For $b \leq 5$, the values of $\beta_b$ are listed in Table 1 in \cite{30}. The quantization noise vector $q$ is uncorrelated to the quantization input $y$ and follows the complex Gaussian distribution $q \sim C N(0, R_{qq})$, where the covariance matrix is given as \cite{27}

$$R_{qq} = \alpha_b^2 \rho \text{diag}\{\rho W_{RF}^H H H^H W_{RF}^H W_{RF}^H W_{RF}^H W_{RF}^H \}$$

Then, a digital combiner $W_{BB} \in C^{N_{RF} \times N_{RF}}$ is applied to the quantized signal in (3) as

$$z = \alpha_b \sqrt{\rho} W_{BB} W_{RF}^H Hs + \alpha_b W_{BB} W_{RF}^H n + W_{BB} W_{RF}^H q$$

III. OPTIMALITY OF TWO-STAGE ANALOG COMBINING

In this section, we provide a near optimal structure for the first and second analog combiners $W_{RF1}, W_{RF2}$ in low-resolution ADC systems for a general channel. To this end, we first formulate an unconstrained MI maximization problem without a constant modulus condition on the analog combiner $W_{RF}$. Then, we derive a near optimal solution for the unconstrained problem, which can be split into two different functions corresponding to the two-stage analog combiner.

We consider the MI between the transmit symbols $s$ and quantized signals $y_q$ under the AQNM model as a measure to maximize. The MI is given as

$$\mathcal{C}(W_{RF})$$

$$= \log_2 |I_{N_{RF}} + \rho \mathbb{E}(s^H W_{RF}^H W_{RF} + R_{qq})^{-1} W_{RF}^H H H^H W_{RF}|$$

Using (6), we formulate the maximum MI problem by only assuming a semi-unitary constraint on the analog combiner

$$\mathcal{P}1 : W_{RF}^\text{opt}_{RF} = \text{argmax } \mathcal{C}(W_{RF}), \text{s.t. } W_{RF}^H W_{RF} = I$$

We remark that the derived analysis in this paper can also be applicable to a heterogeneous long-term received SNR network with minor modification.
Under the perfect quantization system where the number of quantization bits is assumed to be infinite, the optimal analog combiner for the problem $P_1$ is given as the matrix $U_{1:N_{RF}}$ that consists of the first $N_{RF}$ left singular vectors of $H$. The optimal solution $W_{RF}^{\text{opt}}$ of the problem $P_1$ with a finite number of quantization bits, however, is still not known. We first derive an optimal scaling law with respect to the number of RF chains $N_{RF}$, and provide a solution that achieves the scaling law.

**Theorem 1** (Optimal scaling law). For fixed $N_{RF}/N_{c} = \kappa$ with $\kappa \in (0, 1)$, the MI with the optimal combiner $W_{RF}^{\text{opt}}$ for the problem $P_1$ scales with $N_{RF}$ as

$$C(W_{RF}^{\text{opt}}) \sim N_c \log_2 N_{RF}$$

and this optimal scaling law can be achieved by using $W_{RF}^* = W_{RF1}^*W_{RF2}^*$ such that:

(i) $W_{RF1}^* = \lfloor U_{1:N_u}U_{\perp} \rfloor$, and

(ii) $W_{RF2}^*$ is any $N_{RF} \times N_{RF}$ unitary matrix that satisfies the constant modulus condition on its elements,

where $U_{1:N_u}$ is the matrix of the left-singular vectors corresponding to the first $N_u$ largest singular values of $H$ and $U_{\perp}$ denotes the matrix of any orthonormal vectors whose column space is orthogonal to that of $U_{1:N_u}$.

Proof. Since the optimal solution for $P_1$ is not known, we first derive an upper bound of $C(W_{RF})$ and its scaling law with respect to $N_{RF}$. We then, show that adopting $W_{RF} = W_{RF1}^*W_{RF2}^*$, which satisfies the conditions $(i)$ and $(ii)$ in Theorem 1, achieves the same scaling law of the upper bound.

An arbitrary semi-unitary analog combiner $W_{RF}$ can be decomposed into

$$W_{RF} = [U_{\|} U_{\perp}]W_{RF},$$

where $U_{\|}$ is an $N_{c} \times m$ matrix composed of $m$ orthonormal basis vectors whose column space is in the subspace of $\text{Span}(u_1, \ldots, u_{N_u})$ with $1 \leq m \leq N_u$, $U_{\perp}$ is an $N_{c} \times (N_{RF} - m)$ matrix composed of $(N_{RF} - m)$ orthonormal basis vectors whose column space is in the subspace of $\text{Span}^{\perp}(u_1, \ldots, u_{N_u})$, and $W_{RF}$ is an $N_{RF} \times N_{RF}$ unitary matrix. Here, $u_i$ is the $i$-th left-singular vector of $H$. Using (2), the term $W_{RF}^{H}HH^{H}W_{RF}$ in (3) can be re-written as

$$W_{RF}^{H}HH^{H}W_{RF} = W_{RF}^{H}[U_{\|} U_{\perp}]U^{H}U_{\|}U^{H}W_{RF} = W_{RF}^{H}\left[U_{\|}U_{1:N_{u}}A_{N_{u}}U_{1:N_{u}}^{H}U_{\perp}\right]W_{RF}$$

$$\triangleq Q$$

where $A = \text{diag} \{\lambda_1, \ldots, \lambda_{N_u}, 0, \ldots, 0\} \in C^{N_{c} \times N_{c}}$, $A_{N_u} = \text{diag} \{\lambda_1, \ldots, \lambda_{N_u}\}$, $\lambda_i$ is the $i$th largest singular value of $HH^{H}$, and $U_{1:N_{u}} = \{u_1, \ldots, u_{N_u}\}$. The matrix $Q$ has $m$ ranks and can be decomposed into $Q = U_{Q}A_{Q}U_{Q}^{H}$, where $U_{Q}$ is the $N_{RF} \times N_{RF}$ matrix consisting of $N_{RF}$ singular vectors of $Q$, and $A = \text{diag} \{\lambda_1, \ldots, \lambda_m, 0, \ldots, 0\} \in C^{N_{RF} \times N_{RF}}$. Here, $\bar{\lambda}_i$ is the $i$th largest singular value of $Q$. Since $U_{Q}$ is unitary, $W_{RF}$ can be re-expressed as

$$W_{RF} = U_{Q}W_{RF}.$$

and $W_{RF}$ is still unitary. Substituting (11) into (10), we have

$$W_{RF}^{H}HH^{H}W_{RF} = W_{RF}^{H}A_{Q}W_{RF}$$

and the MI in (6) becomes

$$C(W_{RF}) = \log_2 \left| I + \frac{\alpha_b}{\beta_{b}} \text{diag}^{-1}\left\{ W_{RF}^{H}A_{Q}W_{RF} + \frac{1}{\beta_{b}\rho} I\right\} W_{RF}^{H}A_{Q}W_{RF} \right|$$

Let $G = W_{RF}^{H}A_{Q}^{1/2} = [G_{\text{sub}} \ 0]$, where $G_{\text{sub}}$ is the $N_{RF} \times m$ submatrix of $G$. Then, the MI can be upper bounded as

$$C(W_{RF}) = \log_2 \left| I_{N_{RF}} + \frac{\alpha_b}{\beta_{b}} G^{H} \text{diag}^{-1} \left\{ \left\| G_{\text{sub}} \right\|_{2}^{2} + \frac{1}{\beta_{b}\rho} \right\} G \right|$$

and

$$= \log_2 \left| I_{m} + \frac{\alpha_b}{\beta_{b}} G_{\text{sub}}^{H} \text{diag}^{-1} \left\{ \left\| G_{\text{sub}} \right\|_{2}^{2} + \frac{1}{\beta_{b}\rho} \right\} G_{\text{sub}} \right|$$

(a)

Let $G_{\text{sub}} = \left\{ \sum_{i=1}^{m} \frac{1}{\beta_{b}\rho} \sum_{i=1}^{m} \left\| G_{\text{sub}} \right\|_{2}^{2} + 1 \right\}$

(b)

$$= \log_2 \left| I_{m} + \frac{\alpha_b}{\beta_{b}} \sum_{i=1}^{m} \left\| G_{\text{sub}} \right\|_{2}^{2} + \frac{1}{\beta_{b}\rho} \right|$$

(c)

$$= \log_2 \left| I_{m} + \frac{\alpha_b}{\beta_{b}} \sum_{i=1}^{m} \left\| G_{\text{sub}} \right\|_{2}^{2} + \frac{1}{\beta_{b}\rho} \right|$$

where (a) follows by letting $G_{\text{sub}}$ be the matrix whose each row $i$ is given as $i$-th row of $G_{\text{sub}}$ normalized by $\left( \left\| G_{\text{sub}} \right\|_{2}^{2} + 1 \right)^{1/2}$; (b) comes from Jensen’s inequality and the concavity of $\log_2 (1 + x)$ for $x > 0$; and (c) is from

$$\sum_{i=1}^{m} \lambda_i \left( \bar{G}_{\text{sub}}^{H} \bar{G}_{\text{sub}} \right) = \text{Tr} \left( \bar{G}_{\text{sub}}^{H} \bar{G}_{\text{sub}} \right) = \sum_{i=1}^{N_{RF}} \left\| G_{\text{sub}} \right\|_{i}^{2} + \frac{1}{\beta_{b}\rho}.$$
Here, (a) is from that all diagonal entries of $W_{RF}^* A_{N_{RF}} W_{RF}^*$ are the same as $d_\ell = \frac{\sum_{i=1}^{L_\ell} |g_{\ell,i}|^2}{N_\ell}$, for $j = 1, \ldots, N_{RF}$, because of the constant modulus property of $W_{RF}^*$; (b) follows from the fact that as $N_{RF} \to \infty$, i.e., as $N_u \to \infty$, we have $\frac{1}{N_u} H H^* \to \text{diag}\{\frac{1}{L_\ell} \sum_{i=1}^{L_\ell} |g_{\ell,i}|^2, \ldots, \frac{1}{N_\ell} \sum_{\ell=1}^{N_\ell} |g_{\ell,i}|^2\}$ by the channel model in (1) without the pathloss component and the law of large numbers, which implies

$$\lambda_i \to \frac{1}{L_\ell} \sum_{i=1}^{L_\ell} |g_{\ell,i}|^2 < \infty, \quad \text{for } i = 1, \ldots, N_u.$$  

This completes the proof of Theorem 1.\hfill \blacksquare

We note from (14) that $W_{RF}^*$ of the two-stage analog combining solution $W_{RF}^*$ aggregates all channel gains into the smaller dimension and provides $(N_{RF} - N_u)$ extra dimensions. Then, as observed in (15), $W_{RF}^*$ spreads the aggregated channels gains over all $N_{RF}$ dimensions, which reduces the quantization error by exploiting the extra dimensions. Accordingly, as the number of RF chains $N_{RF}$ increases, the proposed solution $W_{RF}^* = W_{RF1}^* W_{RF2}^*$ achieves the optimal scaling law (3) by reducing the quantization error.

**Corollary 1.** The conventional optimal solution $W_{RF}^{cv} = [U_{1:N_u}, U_{\perp}]$ for perfect quantization systems cannot achieve the optimal scaling law (3) in coarse quantization systems, and it is upper bounded by

$${C(W_{RF}^{cv})} < C_{\text{sub}} = N_u \log_2 \left(1 + \frac{\alpha_b}{1 - \alpha_b}\right). \quad (17)$$

**Proof.** From (14), we have the following MI by setting $W_{RF2} = I$:

$$C(W_{RF}^{cv}) = \log_2 \left| 1 + \frac{\alpha_b}{\beta_b} \text{diag}\{-A_{N_{RF}} + \frac{1}{\beta_b} I\} A_{N_{RF}}\right|$$

$$= \sum_{i=1}^{N_u} \log_2 \left(1 + \frac{\alpha_b \lambda_i}{\beta_b \lambda_i + 1/\rho}\right)^{(a)} < N_u \log_2 \left(1 + \frac{\alpha_b}{\beta_b}\right),$$

where (a) comes from $\rho > 0$.\hfill \blacksquare

Corollary 1 shows that the conventional optimal analog combiner $W_{RF}^{cv}$ can capture all channel gains but the MI does not scale as that of $W_{RF}^* = W_{RF1}^* W_{RF2}^*$. Since all channel gains after processed through $W_{RF}^*$ are concentrated on only $N_u$ RF chains out of $N_{RF}$ RF chains, using $W_{RF}^{cv}$ results in severe quantization errors at each of the $N_u$ RF chains. Although the channel gains $\{\lambda_i\}$ increase as $N_u$ increases, the quantization errors also increase in proportion to the channel gains for $C(W_{RF}^{cv})$, yielding only the bounded MI in (17).

Again, unlike the conventional solution, the additional second stage analog combiner $W_{RF2}^*$ proposed in Theorem 1 spreads the channel gains captured by the first stage combiner $W_{RF1}^*$, to all $N_{RF}$ RF chains evenly, leading to achieving the optimal scaling law by greatly alleviating quantization errors. Intuitively, adopting the second combiner $W_{RF2}^*$ results in distributing the burden of ADCs confined in few RF chains over all available ADCs of the total RF chains. Later, we show that such performance gain from adopting the two-stage analog combining structure can be significant even with a reasonable number of RF chains.

**Theorem 2.** For the case of homogeneous singular values of $H^* H$ where all singular values $\{\lambda_i\}$ are equal, the two-stage analog combining solution $W_{RF}^* = W_{RF1}^* W_{RF2}^*$ in Theorem 1 maximizes the MI in (7) with finite $N_{RF}$, i.e.,

$$W_{RF}^* = \text{arg max}_{W_{RF}} C(W_{RF})$$

s.t. $W_{RF}^* H = H_{N_u}$ and $\lambda_1 = \cdots = \lambda_{N_u} = \lambda$, and the corresponding optimal MI is given as

$$C_{\text{opt}} \triangleq C(W_{RF}^*) = N_u \log_2 \left(1 + \frac{\alpha_b \lambda \lambda_{N_u}}{\lambda_{N_u} (1 - \alpha_b) + N_{RF} / \rho}\right). \quad (18)$$

**Proof.** Recall $G = W_{RF}^* H 1/2 = [G_{sub}]$ in the proof of Theorem 1 where $G_{sub}$ is the $N_{RF} \times m$ submatrix of $G$ and $A = \text{diag}\{\lambda_1, \ldots, \lambda_m, 0, \ldots, 0\}$ is the diagonal matrix composed of the singular values of $Q$, defined in (10). From the assumption of $\lambda_1 = \cdots = \lambda_{N_u} = \lambda$, we have

$$\max_{x \in \mathbb{C}^m : \|x\| = 1} x^H Q x = \max_{y \in \mathbb{C}^m : \|y\| = 1} \lambda \|U_1 : \lambda_{N_u} y\|^2$$

$$(a) \leq \lambda \|U_1 : \lambda_{N_u} y\|^2 \|y\|^2 \leq \lambda,$$

where (a) comes from the sub-multiplicativity of the norm, and the last equality holds by $\|U_1 : \lambda_{N_u} y\| = 1$ and $\|U_1 : y\| = 1$. This implies the singular values of $Q$ are bounded as $\lambda_i \leq \lambda$ for $i = 1, \ldots, m$. Hence, $\|G_{sub} y\|_2$ is maximized for any given $W_{RF}$ when $\lambda_i$ achieves $\lambda$ for all $i = 1, \ldots, m$.

We consider the upper bound of $C(W_{RF}^*)$ in (13) and define

$$G_{sub}^* = W_{RF}^* \left[\sqrt{\lambda_{m}} I_m \cdots 0\right].$$

Then, (13) is further upper bounded as

$$C(W_{RF}^*) \leq m \log_2 \left(1 + \frac{\alpha_b \lambda \lambda_{N_u}}{\lambda_{N_u} (1 - \alpha_b) + N_{RF} / \rho}\right)$$

$$\leq m \log_2 \left(1 + \frac{\alpha_b \lambda \lambda_{N_u}}{\lambda_{N_u} (1 - \alpha_b) + N_{RF} / \rho}\right)$$

$$= m \log_2 \left(1 + \frac{\alpha_b \lambda \lambda_{N_u}}{\lambda_{N_u} (1 - \alpha_b) + N_{RF} / \rho}\right), \quad (19)$$

where (a) holds by Jensen’s inequality and the concavity of $\frac{x^2}{x+1}$ for $x > 0$; and (b) comes from $\sum_{i=1}^{N_{RF}} \|G_{sub}^{*, i}\|^2 = \|G_{sub}^*\|^2 = \lambda m$. Note that (19) is maximized when $m = N_u$ since the derivative of (19) with respect to $m$ is positive for $m > 0$ for any given $\alpha_b, \lambda, \rho, N_{RF} > 0$. By substituting $\lambda_1 = \cdots = \lambda_{N_u} = \lambda$ into (19), it can be shown that the upper bound of $C(W_{RF}^*)$ in (19) with $m = N_u$ can be achieved by adopting $W_{RF}^* = W_{RF1}^* W_{RF2}^*$. This completes the proof of Theorem 2.\hfill \blacksquare

Theorem 2 shows the optimality of the proposed two-stage analog combining solution $W_{RF}^* = W_{RF1}^* W_{RF2}^*$ in maximizing the MI for any number of RF chains $N_{RF} \geq N_u$ with homogeneous singular values. We note that such optimality of $W_{RF}^*$ can be nearly achieved for a fixed number of users in large-scale MIMO systems as shown in Remark 1.
Remark 1. From Theorem 2 the two-stage analog combining solution $W_{RF}^* = W_{RF1}^* W_{RF2}^*$ in Theorem 1 maximizes the MI for $\mathcal{P}1$ as well as achieves the optimal scaling law [3] in homogeneous massive MIMO networks with a large number of antennas $N_r$, where each channel element $h_{ij} \sim \mathcal{CN}(0,1)$. This is because as the number of receive antennas $N_r$ increases, $\frac{1}{N_r}H^*H \to I_{N_u}$, i.e., $\frac{1}{N_r}\lambda_i \to 1, \forall i$ [2].

Figure 2 shows the simulation results of the MI of the proposed two-stage analog combiner $W_{RF}^* = W_{RF1}^* W_{RF2}^*$ in Theorem 1 and the conventional analog combiner $W_{CV}^*$ in Corollary 1 which is optimal for infinite-resolution ADC systems. Here, we use $W_{RF1}^* = W_{RF}^* = U_{1:N_{RF}}$ and $W_{RF2}^* = W_{DFT}$, where $W_{DFT}$ is an $N_{RF} \times N_{RF}$ normalized DFT matrix, and consider Rayleigh MIMO channels described in Remark 1. As shown in Fig. 2(a), the MI of the proposed two-stage analog combiner almost achieves the optimal MI in Theorem 1 and the conventional analog combiner almost achieves the optimal MI as well as achieves the optimal scaling law [8].

IV. TWO-STAGE ANALOG COMBINING ALGORITHM

In the previous section, we derived the analog combining solution for the unconstrained problem $\mathcal{P}1$. However, the constant modulus constraint on each matrix element should be taken into account in designing analog combiners since it is implemented using phase shifters. We further consider a predefined set of phases with a finite cardinality for phase shifters. Considering channels known at the receiver, we propose a codebook-based two-stage analog combining algorithm for mmWave communications.

A. Proposed Two-Stage Analog Combining Algorithm

Theorem 1 provides a practical analog combiner structure that is implementable with a two-stage analog combiner $W_{RF} = W_{RF1} W_{RF2}$: the first analog combiner and the second analog combiner can be considered as a channel gain aggregation matrix and spreading matrix, respectively. Leveraging such insight and the finding in the following Corollary, we propose an ARV-based two-stage analog combining (ARV-TSAC) algorithm for mmWave channels.

Corollary 2. When the number of channel paths $L_k$ is limited, the optimal scaling in [8] can be achieved by using $W_{RF}^* = W_{AOA} W_{RF2}^*$ as $N_r \to \infty$ for fixed $\kappa \in (0,1)$, where $W_{AOA} = [A_{AOA}, A_{AOA}^\perp]$, $A_{AOA} = [a(\phi_{1,1}), a(\phi_{2,1}), \cdots, a(\phi_{L_{N_u, N_u}})]$, and $A_{AOA}^\perp$ is an $N_r \times (N_{RF} - \sum_{k=1}^{N_{RF}} L_k)$ matrix composed of orthonormal basis vectors whose column space is in $\text{Span}^\perp(A_{AOA})$.

Proof. See Appendix A.

According to Corollary 2 using ARVs provides a fair trade-off between practicality in implementation and performance. To design the first analog combiner $W_{RF1}$, we adopt an ARV-codebook based maximum channel gain aggregation approach to collect most channel gains into the lower signal dimension by exploiting the sparse nature of mmWave channels. We set the codebook of the evenly spaced spatial angles $V = \{\vartheta_1, \ldots, \vartheta_{|V|}\}$. Since selecting $N_{RF}$ ARVs out of the total $|V|$ ARVs in the codebook requires $\binom{|V|}{N_{RF}}$ search complexity for the exhaustive method, we propose a greedy-based algorithm to find the best $N_{RF}$ ARVs with greatly reduced complexity.
Algorithm 1: ARV-based TSAC

1. **Initialization**: set $W_{RF_1} = \text{empty matrix}$, $H_{tm} = H$, and $V = \{\vartheta_1, \ldots, \vartheta_{|V|}\}$ where $\vartheta_n = \frac{2\pi}{|V|} - 1$

2. **for** $i = 1 : N_{RF}$ **do**

   3. Maximum channel gain aggregation
      
      \[ a(\vartheta^*) = \text{argmax}_{\vartheta \in V} \|a(\vartheta)^H H_{tm}\|^2 \]
      
      \[ W_{RF_1} = \left[ W_{RF_1} \right] \quad \text{where} \quad \left[ a(\vartheta^*) \right] \]
      
      \[ H_{tm} = P_{a(\vartheta^*)} H_{tm}, \quad \text{where} \quad P_{a(\vartheta^*)} = I - a(\vartheta)a(\vartheta)^H \]
      
      \[ V = \{\vartheta^*\} \]

4. **end**

5. Set $W_{RF_2} = W_{DFT}$ where $W_{DFT}$ is a normalized $N_{RF} \times N_{RF}$ DFT matrix.

6. **return** $W_{RF_1}$ and $W_{RF_2}$.

Algorithm 1 describes the proposed ARV-TSAC method. In Step (a), the ARV with the spatial angle $\vartheta^*$ which captures the largest channel gain in the remaining channel dimensions $H_{tm}$ is selected and it composes a column of the first analog combiner in Step (b). In Step (c), the channel matrix on the remaining dimensions $H_{tm}$ is projected onto the subspace of Span$(a(\vartheta^*))$ to remove the channel gain on the space of the selected ARV. Algorithm 1 repeats these steps until $N_{RF}$ ARVs are selected from the codebook $V$.

**Remark 2.** We can implement the second-stage analog combiner that satisfies the condition (ii) of Theorem 7 by adopting a normalized $N_{RF} \times N_{RF}$ DFT matrix, i.e., $W_{RF_1} = W_{DFT}$.

Employing the DFT matrix for the second analog combiner $W_{RF_2} = W_{DFT}$ (or any unitary matrix with constant modulus) offers benefits in reducing implementation complexity and power consumption since $W_{DFT}$ does not depend on the channel $H$ and can be constructed by using passive (or fixed) analog phase shifters. Accordingly, although the additional $N_{RF}^2$ fully-connected passive phase shifters for the second analog combiner add to the complexity of the proposed architecture in physical area and power consumption, it can be implemented with very low complexity and power consumption in the practical system. Furthermore, if $N_{RF}$ is a power of two, the fast Fourier transform version of the DFT calculation can be implemented, which reduces the number of additional passive phase shifters to $N_{RF} \log_2 N_{RF}$.

### B. Performance Analysis

In this subsection, we analyze the ergodic sum rate of the ARV-TSAC algorithm with an MRC baseband combiner. Once we derive the closed-form ergodic rate, we compare the rate with the one without the second analog combiner $W_{RF_2}$ to quantify the ergodic rate gain from employing $W_{RF_2}$. To this end, we adopt a virtual channel representation [33] for analytic tractability which captures the sparse property of mmWave channels [17], [34]. Under the virtual channel representation, the channel vector $h_k$ in (4) can be modeled as

\[ h_k = \sqrt{\frac{N_u}{L_u}} g_\rho \tilde{A} g_k = \tilde{A} h_{b,k} \]

where $\tilde{A}$ is the $L_u$-sparse beamspace channel of user $k$, i.e., $g_k$ has $L_u$ nonzero entries, $\tilde{A} \in \mathbb{C}^{N_u \times L_u}$, and $\tilde{A} = \{a(\varphi_1), \ldots, a(\varphi_{N_u})\}$ with uniformly spaced spatial angles $\varphi_i$.

Under this representation, we consider the case where the codebook size of Algorithm 1 is equal to the number of antennas $|V| = N_r$. Accordingly, the first analog combiner is the $N_c \times N_{RF}$ submatrix of $A$ which captures the first channel gain, $W_{RF_1} = A_{sub}$. We assume that $W_{RF_1}$ captures all channel propagation paths from $N_u$ users (19), (33), i.e., $L_u$ channels paths for each user fall within $N_{RF}$ RF chains. For simplicity, we further assume $L_u = L_u$, $k$, in the analysis. Thus, after combining with $W_{RF_1} = A_{sub}$, the channel becomes $h_b = W_{RF_1}^H h$, and the channel vector of user $k$ with the reduced dimension $h_{b,k} \in \mathbb{C}^{N_{RF}}$ is

\[ h_{b,k} = \sqrt{\frac{N_u}{L_u}} g_k. \]  

(21)

We consider $L_u$ nonzero channel gains to be uniformly distributed within each user channel $h_{b,k}$ and use an indicator function $I_{(i \in A)}$ to characterize the channel sparsity where $I_{(i \in A)} = 1$ if $i \in A$, and $I_{(i \notin A)} = 0$ otherwise. Utilizing $I_{(i)}$, we model the $i$th complex path gain of user $k$ as

\[ g_{\ell,k} = \xi_{\ell,k} I_{(\ell \in \mathcal{P}_k)}, \quad \ell = 1, \ldots, N_{RF}, \quad k = 1, \ldots, N_u \]

where $\xi_{\ell,k} \sim \mathcal{CN}(0,1)$, $\forall \ell, k$ and $\mathcal{P}_k = \{ i : g_{i,k} \neq 0, i = 1, \ldots, N_{RF} \}$ is the nonzero index set.

We consider the MRC combiner $W_{BB} = \bar{H}_b$ where $\bar{H}_b = W_{RF_2}^H h$, and the received signal $k$ in (5) becomes

\[ z_k = \alpha_b \sqrt{\rho} \bar{h}_{b,k}^H \bar{b}_k s_k + \alpha_b \sqrt{\rho} \sum_{i \neq k} \bar{h}_{b,i}^H \bar{b}_i s_i + \alpha_b \bar{h}_{b,k}^H W_{RF_2}^H n + \bar{h}_{b,k}^H q. \]  

(22)

From (22), the achievable rate of the proposed system for the MRC combiner with simplification is given as

\[ r_{k}^{mrc} = \log_2 \left( 1 + \frac{\rho \alpha_b \| \bar{h}_{b,k} \|^2}{\rho \alpha_b \sum_{i \neq k} \| \bar{h}_{b,k}^H \bar{b}_i s_i \|^2 + \| \bar{h}_{b,k} \|^2 + \rho \beta_k \Psi_k} \right) \]  

(23)

where $\Psi_k = \bar{h}_{b,k}^H \text{diag}(\{ \bar{h}_{b,i}^H \bar{h}_{b,j} \}) h_{b,k}$, and the ergodic rate is

\[ r_{k}^{mrc} = E \left[ r_{k}^{mrc} \right] \]

\[ = E \left[ \log_2 \left( 1 + \frac{\rho \alpha_b \| \bar{h}_{b,k} \|^2 \| \bar{h}_{b,k} \|^2}{\rho \alpha_b \sum_{i \neq k} \| \bar{h}_{b,k}^H \bar{b}_i s_i \|^2 + \| \bar{h}_{b,k} \|^2 + \rho \beta_k \Psi_k} \right) \right]. \]

(24)

Since $W_{RF_2} = W_{DFT}$ is unitary, we have $\| \bar{h}_{b,k}^H \bar{h}_{b,j} \| = \| \bar{h}_{b,k} \| \| \bar{h}_{b,j} \|$, $\forall k, j$. We approximate the ergodic rate (24) as

\[ r_{k}^{mrc} = E \left[ \log_2 \left( 1 + \frac{\rho \alpha_b \| \bar{h}_{b,k} \|^2}{\rho \alpha_b \sum_{i \neq k} \| \bar{h}_{b,k}^H \bar{b}_i s_i \|^2 + \| \bar{h}_{b,k} \|^2 + \rho \beta_k \Psi_k} \right) \right] \]

\[ \approx \log_2 \left( 1 + \frac{\rho \alpha_b \| \bar{h}_{b,k} \|^2}{\rho \alpha_b \sum_{i \neq k} \| \bar{h}_{b,k}^H \bar{b}_i s_i \|^2 + \| \bar{h}_{b,k} \|^2 + \rho \beta_k \Psi_k} \right) \]  

(25)

The similar results can be derived with minor changes for general $L_u$.\footnote{The similar results can be derived with minor changes for general $L_u$.}
Lemma 2. For the considered mmWave channel, the cross quantization error more as seen Appendix B.

Proof. Noting $\Psi = h^H \mathbf{W}_{\text{DFT}} \text{diag} \{ W_{\text{DFT}}^H \mathbf{H}_b W_{\text{DFT}} \} W_{\text{DFT}}^H \mathbf{h}_b$, we decompose $\Psi$ as $\Psi = \Psi^{\text{auto}} + \Psi^{\text{cross}}$, and define the auto quantization noise and cross quantization noise variances as

$$ E[\Psi^{\text{auto}}] = E[\Psi^{\text{cross}}] = \frac{2N_{\text{RF}}^2}{N_{\text{RF}}} = E[\Psi^{\text{cross}}] = \frac{N_{\text{RF}}^2(N_{\text{RF}} - 1)}{N_{\text{RF}}}. $$

Proof. See Appendix [B] [C]

Remark 3. Let $\kappa = N_{\text{RF}}/N_r$ where $\kappa \in (0, 1)$ is a constant value. Then, (30) can reduce to

$$ \tilde{R}_{\text{one}} \approx N_{\text{RF}} \log_2 \left( 1 + \frac{\rho \alpha_b N_{\text{RF}} (1 + 1/L)}{\kappa + \rho (N_{\text{RF}} - 1) + 2 \rho (1 - \alpha_b) N_{\text{RF}} / L} \right). $$

The ergodic sum rate in (31) achieves the optimal scaling law $\sim N_{\text{RF}} \log_2 N_{\text{RF}}$ with respect to $N_{\text{RF}}$ as in [5].

Remark 3 shows that the optimal scaling law can be achieved by the proposed two-stage analog combining algorithm even with the practical baseband combiner. This result verifies that the two-stage analog combining architecture is effective to enhance the achievable rate in mmWave hybrid MIMO systems with low-resolution ADCs. To specify the effect of employing the second analog combiner $\mathbf{W}_{\text{RF}}$, we also derive the ergodic rate (24) without using $\mathbf{W}_{\text{RF}}$.

Corollary 3. For the considered mmWave channel with low-resolution ADCs, the MRC ergodic rate of the ARV- TSAC without the second analog combiner is approximated as

$$ \tilde{R}_{\text{one}} \approx N_{\text{RF}} \log_2 \left( 1 + \frac{\rho \alpha_b N_r N_{\text{RF}} (1 + 1/L)}{N_{\text{RF}} + \rho N_r (N_{\text{RF}} - 1) + 2 \rho (1 - \alpha_b) N_{\text{RF}} / L} \right). $$

Remark 4. Let $\kappa = N_{\text{RF}}/N_r$ where $\kappa \in (0, 1)$ is a constant value. Then, (32) can reduce to

$$ \tilde{R}_{\text{one}} \approx N_{\text{RF}} \log_2 \left( 1 + \frac{\rho \alpha_b N_r N_{\text{RF}} (1 + 1/L)}{\kappa + \rho (N_{\text{RF}} - 1) + 2 \rho (1 - \alpha_b) N_{\text{RF}} / L} \right). $$

Note that the derived ergodic rate of the two-stage analog combining $\tilde{R}_{\text{one}}$ in (33) cannot achieve the optimal scaling law with respect to the number of RF chains $N_{\text{RF}}$.

V. Simulation Results

In this section, we evaluate the performance of the proposed two-stage analog combining algorithm in the MI and ergodic sum rate. In the simulations, we set the codebook size to be $|\mathcal{V}| = N_r$, which guarantees $\mathbf{W}_{\text{RF}}^H \mathbf{W}_{\text{RF}} = \mathbf{I}_{N_{\text{RF}}}$ Consequently, analog combiners used in the simulations are semi-unitary. To provide a reference performance of a conventional one-stage analog combining approach, we simulate a greedy-based MI maximization method which solves the following problem for the given ARV codebook in a greedy way:

$$ \mathbf{W}^{\text{opt.c}} = \arg \max_{\mathbf{W}_{\text{RF}}} C(\mathbf{W}_{\text{RF}}) $$

subject to

$$ \mathbf{W}_{\text{RF}}^H \mathbf{W}_{\text{RF}} = \mathbf{I}, \quad |(\mathbf{W}_{\text{RF}})_{i,j}| = \frac{1}{\sqrt{N_r}}, \forall i, j. $$

At each iteration, the greedy method searches for a single ARV from the codebook $\mathcal{V}$ which maximizes the MI with
respect to the SNR $\rho$ and AR V cases show the MI gap from the AR V-TSAC. The gap decreases as MI and AR V cases provide similar MI to each other but show the best MI over the most SNR values. The Greedy-MI achieves a similar MI as does the SVD+DFT case, and they guarantee the optimality condition for the two-stage analog combining solution in Theorem 2. As more RF chains are used, however, the MI gap between AR V-TSAC/SVD+DFT and Greedy-MI/AR V becomes larger and the performance reversal would happen in even the higher SNR regime. This is because the proposed two-stage analog combining can exploit more RF chains to further reduce quantization errors. The SVD and Greedy-MI/AR V are infeasible in practice due to violating the constant modulus constraint, and SVD+DFT provides a tight upper bound on MI for a homogeneous singular value case from Theorem 2. Here, we adopt $L_k = \max\{1, \text{Poisson}(\lambda_L)\}$ unless mentioned otherwise, where $\lambda_L$ is considered as the average number of channel paths.

### A. Mutual Information

Fig. 3 shows the MI simulation results for $N_r = 128$, $N_u = 8$, $\lambda_L = 3$ average channel paths, $b = 2$ quantization bits, and $N_{RF} \in \{43, 64\}$ RF chains that are $[N_r/3]$ and $[N_r/2]$, respectively.

![Figure 3](image1)

Figure 3. The MI simulation results for $N_r = 128$ receive antennas, $N_u = 8$ users, $\lambda_L = 3$ average channel paths, $b = 2$ quantization bits, and $N_{RF} \in \{43, 64\}$ RF chains that are $[N_r/3]$ and $[N_r/2]$, respectively.

In the simulations, we evaluate the following cases:

1. AR V-TSAC: proposed two-stage analog combining.
2. AR V: one-stage analog combining with $W_{RF} = W_{RF1}$ selected from the AR V-TSAC.
3. SVD+DFT: two-stage analog combining with $W_{RF1} = U_{1: N_{RF}}$ and $W_{RF2} = W_{DFT}$ based on Theorem 1.
4. SVD: one-stage analog combining $W_{RF} = U_{1: N_{RF}}$.
5. Greedy-MI: one-stage analog combining with greedy-based MI maximization.

The SVD+DFT and SVD cases are infeasible in practice due to violating the constant modulus constraint, and SVD+DFT provides a tight upper bound on MI for a homogenous singular value case from Theorem 2. Here, we adopt $L_k = \max\{1, \text{Poisson}(\lambda_L)\}$ unless mentioned otherwise, where $\lambda_L$ is considered as the average number of channel paths.

![Figure 4](image2)

Figure 4. The MI simulation results with $N_u = 8$ users, $\lambda_L = 4$ average channel paths, $b = 2$ quantization bits, and $\rho = 0$ dB SNR for (a) $N_r = 256$ receive antennas and (b) $\kappa = N_{RF}/N_r = 1/3$.

![Figure 5](image3)

### B. Ergodic Sum Rate

Now, we evaluate the ergodic rate for linear digital combiners $W_{BB}$ such as MRC, zero-forcing (ZF), and MMSE. Let $H_{eq} = W_{RF}^H H$. The MRC, ZF, and MMSE combiners provide a tight upper bound on the theoretic upper bound $C_{\text{SD}}$ due to the quantization error.

Fig. 4 shows the MI simulation results with $N_u = 8$, $\lambda_L = 4$, $b = 2$, and $\rho = 0$ dB in terms of $N_{RF}$. In Fig. 4(a), $N_r$ is fixed to be $N_r = 256$. The two-stage combining cases, i.e., SVD+DFT and AR V-TSAC, show that the MI increases logarithmically with $N_{RF}$, and this corresponds to the scaling law derived in Theorem 1. The one-stage combining cases such as the Greedy-MI, AR V, and SVD cases, however, show a marginal increase of the MI as $N_{RF}$ increases. In Fig. 4(b), the ratio between $N_r$ and $N_{RF}$ is fixed to be $\kappa = 1/3$. Here, the Greedy-MI and AR V cases also increase more slowly compared to the SVD+DFT and AR V-TSAC cases. This is because more channel gains can be collected as $N_r$ increases for all cases, but the two-stage combining can reduce more quantization error as $N_{RF}$ increases. Accordingly, the MI gap between the two-stage combining and one-stage combining cases increases as $N_{RF}$ increases.
Figure 5. Simulation results of the ergodic sum rate with \( N_r = 128 \) receive antennas, \( N_{RF} = 43 \) RF chains, \( N_u = 8 \) users, \( \lambda_L = 3 \) average channel paths, and \( b = 2 \) quantization bits for (a) maximum ratio combining (MRC), (b) zero-forcing (ZF), and (c) minimum mean squared error (MMSE) digital combiners.

are given as: \( W_{BB, \text{mrc}} = H_{\text{eq}} \), \( W_{BB, \text{af}} = H_{\text{eq}} (H_{\text{eq}}^H H_{\text{eq}})^{-1} \), and \( W_{BB, \text{mse}} = R_{\text{eq}}^{-1} R_{\text{eq}} \), where \( R_{\text{eq}} = \alpha \sigma^2 H_{\text{eq}} \) and \( R_{\text{eq}} = \alpha^2 \sigma^2 H_{\text{eq}}^H H_{\text{eq}} + W_{RF}^H W_{RF} + R_{\text{eq}} \). For the given analog and digital combiners (\( W_{RF}, W_{BB} \)) with \( W_{RF}^H W_{RF} = I_{N_{RF}} \), the ergodic rate of user \( k \) is expressed as

\[
\hat{r}_k (W_{RF}, W_{BB}) = \mathbb{E} \left[ \log_2 (1 + \frac{\sigma^2}{\rho} W_{BB,k}^H \eta_{BB,k} W_{BB,k}) \right]
\]

where \( \eta_{BB,k} = \sigma^2 \rho^2 \sum_{u \neq k} W_{BB,k}^H w_{BB,k}^u \). In this regard, the ARV-TSAC algorithm achieves the ergodic rate comparable to that of SVD+DFT and shows a noticeable improvement compared to the Greedy-MI, ARV, and SVD cases. As \( b \) increases, the ergodic rates of the ARV-TSAC, Greedy-MI, and ARV algorithms converge to each other with a small gap from the SVD+DFT case. The ergodic rate of the SVD case, however, converges to that of SVD+DFT without any gap because the SVD combining is optimal in maximizing the MI of infinite-resolution ADC systems. The simulation results validate the effectiveness of the proposed two-stage combining in low-resolution ADC systems.

Finally, we validate the derived ergodic rates in Theorem 5 and Corollary 5. We consider \( N_r = 128 \) receive antennas, \( N_{RF} = 43 \) RF chains, \( N_u = 8 \) users each with \( L = 8 \) channel paths for the virtual channels, and \( b = 2 \) quantization bits. In Fig. 7, the theoretical ergodic rates tightly align with the simulation results in the medium to high SNR regime, and show similar trend as the simulation results do. Thus, the derived ergodic rates can characterize the ergodic rate performance of the proposed algorithm for the two-stage analog combining system in terms of the system parameters including quantization resolution.

Overall, the two-stage analog combining structure with the ARV-TSAC algorithm almost achieves the performance of SVD+DFT that is a near optimal solution for the unconstrained problem \( P1 \), while the greedy-MI and ARV algorithms provide a near optimal solution only for the constrained problem \( P2 \). Since \( P1 \) has a larger feasible set than \( P2 \) to find an optimal solution for the same objective function, this leads to \( C(W_{RF}^{\text{opt}}) \geq C(W_{RF}^{\text{opt,c}}) \). In this regard, the ARV-TSAC algorithm achieves the higher performance than that of the Greedy-MI and ARV algorithms in most cases. This shows that the proposed two-stage analog combining architecture with the ARV-TSAC is a practical solution suitable for the mmWave hybrid MIMO systems with low-resolution ADCs.
Conventional hybrid algorithms offer a near optimal solution for the unconstrained problem whereas proposed two-stage analog combining also provides a near optimal value case. To implement the derived solution, we proposed the mutual information for a homogeneous channel singular value case. Then, it can be shown [31] that as \( N_r \to \infty \),

\[
W_{AoA}^H W_{AoA} \to I_{N_{RF}}, \quad \frac{1}{\sqrt{N_r}} W_{AoA}^H H \to \frac{1}{\sqrt{N_r}} \begin{bmatrix} H_V \\ 0 \end{bmatrix}.
\]

Let \( \tilde{H}_V = [H_V, 0]^T \) and \( C_{AoA} = W_{RF_2}^H \tilde{H}_V \tilde{H}_V^H W_{RF_2}^\star \). Using (34), we show \( C(W_{RF}) \) in (12) with \( W_{RF} = W_{RF_2}^\star \) converges as \( N_r \to \infty \) to

\[
\left( C(W_{RF}^\star) \log_2 \left| I + \frac{\alpha_h}{\rho_b} \text{diag}^{-1} \{ C_{AoA} + \rho_b I \} C_{AoA} \right| \right) \to 0.
\]

Note that each diagonal of \( W_{RF_2}^H \tilde{H}_V \tilde{H}_V^H W_{RF_2}^\star \) cannot exceed \( \frac{1}{\rho_b} \sum_{k=1}^{N_u} (\sum_{\ell=1}^{L_k} |g_{l,k}|)^2 = \epsilon_1 < \infty \). Let \( C_\infty(W_{RF}^\star) \) denote the second term in (35). Then, \( C_\infty(W_{RF}^\star) \) can be lower bounded as

\[
C_\infty(W_{RF}^\star) \geq \log_2 \left| I_{N_{RF}} + \frac{\alpha_h}{c_1 \rho_b} W_{RF_2}^H \tilde{H}_V \tilde{H}_V^H W_{RF_2}^\star \right| \quad (a)
\]

where \( (a) \) follows from the same reason of \( (b) \) below (16). This implies that \( C(W_{RF}^\star) \) follows the optimal scaling law.

**APPENDIX B**

**PROOF OF LEMMA**

The auto quantization noise variance term in (26) can be expressed as

\[
E[\Psi_k^{\text{auto}}] = \mathbb{E} \left[ \sum_{i=1}^{N_{RF}} |h_{b,k}^H w_i|^4 \right] = \left( \frac{N_r}{L} \right)^2 \sum_{i=1}^{N_{RF}} \mathbb{E} \left[ |g_{b,k}^H w_i|^4 \right]
\]

\[
= \left( \frac{N_r}{L} \right)^2 \sum_{i=1}^{N_{RF}} \left( \mathbb{E} \left[ |g_{b,k}^H w_i|^2 \right] + \left( \mathbb{E} \left[ |g_{b,k}^H w_i|^2 \right] \right)^2 \right)
\]

\[
(37)
\]
where \( w_i \) is the \( i \)th column of \( W_{\text{DFT}} \). The expectation term \( \mathbb{E}[|g_k^H w_i|^2] \) in (37) is computed as

\[
\mathbb{E}[|g_k^H w_i|^2] = \frac{1}{N_{\text{RF}}} \mathbb{E}\left[ \sum_{\ell=1}^{N_{\text{RF}}} |g_{\ell,k}|^2 \right] = \frac{L}{N_{\text{RF}}}. \tag{38}
\]

Now, let \( \tilde{w}_i = \sqrt{N_{\text{RF}}} w_i \). Then, we can compute the variance term \( \mathbb{V}[|g_k^H w_i|^2] \) in (37) as

\[
\begin{align*}
\mathbb{V}[|g_k^H w_i|^2] &= \frac{1}{N_{\text{RF}}} \mathbb{V}\left[ \sum_{\ell=1}^{N_{\text{RF}}} |g_{\ell,k}|^2 + \sum_{\ell_1 \neq \ell_2} g_{\ell_1,k}^* g_{\ell_2,k} \tilde{w}_{\ell_1,i}^* \tilde{w}_{\ell_2,i} \right] \\
&= \frac{1}{N_{\text{RF}}} \left[ \mathbb{V}\left[ |g_{\ell,k}|^2 \right] + \mathbb{V}\left[ \sum_{\ell_1 \neq \ell_2} g_{\ell_1,k} g_{\ell_2,k} \tilde{w}_{\ell_1,i}^* \tilde{w}_{\ell_2,i} \right] \right] \\
&= \frac{1}{N_{\text{RF}}} \left( \mathbb{V}[\|g_k\|^2] + \mathbb{V}[\|\tilde{w}_i\|^2] \right) \tag{39}
\end{align*}
\]

where \((a)\) and \((b)\) hold as the associated terms are uncorrelated, which can be shown from straightforward mathematics, and \(|\tilde{w}_{\ell,i}| = 1\), \(\forall \ell, i\). Since \(|g_k|^2 \sim \chi^2_{2L} \), which is a chi-square distribution with 2L degrees of freedom, we have \(\mathbb{V}[|g_k|^2] = L\), and \(\mathbb{V}[\|\tilde{w}_i\|^2] \) is computed as

\[
\mathbb{V}[\|\tilde{w}_i\|^2] = \mathbb{V}\left[ \sum_{\ell_1, \ell_2 \neq \ell} g_{\ell_1,k}^* g_{\ell_2,k} \tilde{w}_{\ell_1,i}^* \tilde{w}_{\ell_2,i} \right] = \mathbb{V}[\|g_k\|^2] = L. \tag{40}
\]

Putting (38) and (40) into (37), the auto quantization noise variance \( \mathbb{E}[\Psi_k^{\text{auto}}] \) becomes (28).

**APPENDIX C**

**PROOF OF Lemma 2**

We derive the cross quantization noise variance in (27) as

\[
\mathbb{E}[\Psi_k^{\text{cross}}] = \mathbb{E}\left[ \sum_{i=1}^{N_u} \sum_{u \neq i} h_{b,k}^u w_i^H h_{b,u}^u w_i^H h_{b,k} \right] = \frac{N_r^2}{L} \mathbb{E}_g \sum_{i=1}^{N_u} \sum_{u \neq i} g_{\ell,k}^u g_{\ell,u}^u w_i^H w_i^H g_k \]

where \((a)\) follows from \( \mathbb{E}[|g_k^H w_i|^2] = \frac{L}{N_{\text{RF}}} \) in (38). ■

**APPENDIX D**

**PROOF OF Theorem 3**

To compute (25), we first derive \( \mathbb{E}[\|h_{b,k}\|^2] \) as

\[
\mathbb{E}[\|h_{b,k}\|^2] = \frac{N_r}{L} \mathbb{E}[\|g_k\|^2] \tag{41}
\]

where \((a)\) follows from \( \mathbb{E}[|g_k|^2] \sim \chi^2_{2L} \). Next, we compute \( \mathbb{E}[\|h_{b,k}\|^4] \) as

\[
\begin{align*}
\mathbb{E}[\|h_{b,k}\|^4] &= \mathbb{E}[\|h_{b,k}\|^2 + \mathbb{E}[\|g_k\|^2] + \mathbb{E}[\|g_k\|^4] \\
&= \left( \frac{N_r}{L} \right)^2 \left( \mathbb{E}[\|g_k\|^2] + \mathbb{E}[\|g_k\|^4] \right) \\
&= \frac{N_r^2}{L} \left( 1 + L \right). \tag{42}
\end{align*}
\]

The inter-user interference term \( \mathbb{E}[\|h_{b,k}^u h_{b,i}\|^2] \) is computed as

\[
\begin{align*}
\mathbb{E}[\|h_{b,k}^u h_{b,i}\|^2] &= \frac{N_r}{L} \mathbb{E}[\|g_k^u g_k^i\|^2] = \left( \frac{N_r}{L} \right)^2 \sum_{\ell=1}^{N_{\text{RF}}} \mathbb{E}[\|g_{\ell,k} g_{\ell,i}\|^2] \\
&= \frac{N_r^2}{L} \sum_{\ell=1}^{N_{\text{RF}}} \mathbb{E}[\|g_k\|^2] = \frac{N_r^2}{N_{\text{RF}}}. \tag{43}
\end{align*}
\]

Finally, we compute the quantization variance term \( \mathbb{E}[\Psi_k] \) as

\[
\mathbb{E}[\Psi_k] = \mathbb{E}[\Psi_k^{\text{auto}}] + \mathbb{E}[\Psi_k^{\text{cross}}] = 2N_r^2 + \frac{N_r^2 (N_u - 1)}{N_{\text{RF}}}, \tag{44}
\]

where \( \mathbb{E}[\Psi_k^{\text{auto}}] \) and \( \mathbb{E}[\Psi_k^{\text{cross}}] \) are in (26) and (27), respectively, and \((a)\) follows from Lemma 1 and Lemma 2.

Putting (41), (42), (43), and (44) into (25), we derive the approximated ergodic rate of (25) in closed form. The ergodic rate is equivalent to \( N_u \) users, which leads to the ergodic sum rate in (30). This completes the proof of Theorem 3. ■

**APPENDIX E**

**PROOF OF Corollary 3**

Without the second analog combiner \( W_{\text{RF}} \), the approximated ergodic rate of user \( k \) can be computed as (25) by substituting the average quantization noise variance for the two-stage analog combining \( \mathbb{E}[\Psi_k] \) with the following average quantization noise variance:

\[
\begin{align*}
\mathbb{E}[\hat{\Psi}_k] &= \mathbb{E}\left[ h_{b,k}^u \text{diag}\{ H_k H_k^H \} h_{b,k} \right] \\
&= \mathbb{E}\left[ \left( \frac{N_r}{L} \right)^2 \sum_{\ell=1}^{N_{\text{RF}}} |g_{\ell,k}|^2 \sum_{u=1}^{N_u} |g_{\ell,u}|^2 \right] \\
&= \frac{N_r^2}{L} \sum_{\ell=1}^{N_{\text{RF}}} \mathbb{E}[|g_{\ell,k}|^4] + \sum_{u=1}^{N_u} \mathbb{E}[|g_{\ell,u}|^4] \tag{45}
\end{align*}
\]

where \( \mathbb{E}[|g_k|^2] = \frac{L}{N_{\text{RF}}} \) in (38). ■
Here, $E[|g_{\ell,k}|^4]$ in (45) is computed as
\[
E[|g_{\ell,k}|^4] = \mathbb{E}\left[\prod_{\{\ell \in \mathcal{P}\}} |\xi_{\ell,k}|^4\right] = \frac{L}{N_{RF}} \left( \mathbb{V}[|\xi_{\ell,k}|^2] + \mathbb{E}[|\xi_{\ell,k}|^2]^2 \right) = \frac{2L}{N_{RF}},
\]
and the second expectation term $E[|g_{\ell,k}|^2|g_{\ell,u}|^2]$ is derived as
\[
E[|g_{\ell,k}|^2|g_{\ell,u}|^2] = \mathbb{E}\left[\prod_{\{\ell \in \mathcal{P}\}} |\xi_{\ell,k}|^2 |\xi_{\ell,u}|^2\right] = \left( \frac{L}{N_{RF}} \right)^2.
\]
Putting (46) and (47) into (45), we derive the average quantization noise variance for the one-stage analog combining as
\[
\mathbb{E}[\tilde{\Psi}_k] = \frac{N_0}{L} \left( \frac{2}{L} + \frac{N_0 - 1}{N_{RF}} \right).
\]
This completes the proof of Corollary 3.

**REFERENCES**

[1] P. Z. Pi and F. Khan, “An introduction to millimeter-wave mobile broadband systems,” IEEE Commun. Mag., vol. 49, no. 6, pp. 101–107, Jun. 2011.

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

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, “What will 5G be?” IEEE Journal Sel. Areas Commun., vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[4] X. Zhang, A. F. Molisch, and S.-Y. Kung, “Variable-phase-shifted RF-baseband codesign for MIMO antenna selection,” IEEE Trans. Signal Process., vol. 51, no. 11, pp. 4091–4103, 2005.

[5] P. Sudarshan, N. B. Mehta, A. F. Molisch, and J. Zhang, “Channel statistics-based RF pre-processing with antenna selection,” IEEE Trans. Wireless Commun., vol. 5, no. 12, 2006.

[6] F. Gholam, J. Via, and I. Santamaria, “Beamforming design for simplified analog antenna combining architectures,” IEEE Trans. Veh. Technol., vol. 60, no. 5, pp. 2373–2378, 2011.

[7] D. Ying, F. W. Vook, T. A. Thomas, and D. J. Love, “Hybrid structure in massive MIMO: Achieving large sum rate with fewer RF chains,” in IEEE Int. Conf. Commun., 2015, pp. 2344–2349.

[8] F. Sohrabi and W. Yu, “Hybrid digital and analog beamforming design for large-scale antenna arrays,” arXiv preprint arXiv:1601.06814, 2016.

[9] J. Li, L. Xiao, X. Xu, and Z. Zhou, “Robust and low complexity hybrid beamforming for uplink multiuser mmWave MIMO systems,” IEEE Commun. Lett., vol. 20, no. 6, pp. 1140–1143, 2016.

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

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

[12] T. E. Bogale and L. B. Le, “Beamforming for multiuser massive MIMO systems: Digital versus hybrid analog-digital,” IEEE Global Commun. Conf., 2014.

[13] C. Rusu, R. Méndez-Rial, N. González-Prelcy, and R. W. Heath, “Low complexity hybrid sparse precoding and combining in millimeter wave MIMO systems,” in IEEE Int. Conf. Commun., 2015, pp. 1340–1345.

[14] C.-E. Chen, “An iterative hybrid transceiver design algorithm for millimeter wave MIMO systems,” IEEE Wireless Commun. Lett., vol. 4, no. 3, pp. 285–288, 2015.

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