Efficient Multi-User Hybrid Precoding and Combining for mmWave Massive MIMO Systems

Kaijie Zeng, Yuanli Ma, Bin Yan, Zheng Wang, Yili Xia
School of Information Science and Engineering
Southeast University, Nanjing, China
Email: kaijie_zeng@seu.edu.cn, myl@seu.edu.cn, bin_yan@seu.edu.cn, wznuaa@gmail.com, yili_xia@seu.edu.cn

***Abstract*—For downlink millimeter wave (mmWave) massive multiple-input multiple-output (MIMO) systems, this paper proposes an efficient multi-user hybrid precoding and combining (EMU-HPC) scheme to maximize the sum-rate with low feedback overhead. Specifically, we adopt a non-iterative generalized low rank approximation of matrices (NI-GLRAM) framework in the analog stage to harvest the array gain with minimal feedback overhead. Meanwhile, for each mobile station (MS), a low-complexity truncated singular value decomposition (TSVD) is implemented for efficient analog combining. To further enhance the sum-rate, we propose a joint analog-digital interference suppression strategy, in which the orthogonal matrix projection is applied in analog precoding, followed by the regularized channel diagonalization (RCD) in the digital stage. Numerical results confirm that the proposed EMU-HPC method achieves superior sum-rate performance with low feedback overhead and computational complexity.**

***Index Terms*—Massive MIMO, multi-user MIMO, hybrid precoding, sum-rate maximization.**

I. Introduction

In 5G systems and beyond, the integration of mmWave and massive MIMO compensates for severe path loss and enables high sum-rate via precoding [1]. However, fully digital precoding is impractical due to its prohibitive hardware cost and power consumption from radio frequency (RF) chains [2]. To address this issue, hybrid precoding has emerged as a solution, which connects a digital baseband precoder to an analog RF precoder with far fewer RF chains [3]. In multi-user MIMO (MU-MIMO), the problem of hybrid precoding and combining design is normally divided into analog and digital stages for sum-rate optimization [4]–[7]. Yet achieving higher sum-rate without increasing feedback overhead remains challenging.

In particular, the work in [4] employs exhaustive codebook search and equal gain transmission in the analog stage to obtain array gains, followed by block diagonalization in the digital stage for interference elimination. With the same digital design, [5] exploits low rank approximations to enhance array gains, while [6] maximizes the capacity of baseband channels. Although these methods reduce feedback overhead by separately designing analog precoder and combiners, the interference is managed solely in the digital stage, which limits the sum-rate. To solve this problem, [7] incorporates interference management in the analog stage using orthogonal vector projections and employs alternating optimization (AO) for array gains, but the increased feedback overhead from successive AOs may hinder its deployment.

To achieve high sum-rate with low feedback overhead, an efficient multi-user hybrid precoding and combining scheme, named EMU-HPC, is proposed for mmWave massive MIMO systems. Specifically, to maximize the array gain, we employ the NI-GLRAM framework in the analog stage with low feedback cost, while the TSVD in analog combining is realized efficiently via the variance-reduced principal component analysis (VR-PCA) in the LazySVD framework. For sum-rate improvement, we propose a joint analog-digital interference management scheme, with the orthogonal matrix projection first applied in analog precoding, followed by RCD in the digital stage. Numerical results demonstrate that EMU-HPC attains superior sum-rate with minimal feedback overhead and reduced computational complexity.

II. System Model

We consider a single-cell massive MU-MIMO downlink system. The base station (BS) equipped with N_t transmit antennas and M_t RF chains serves K MSs, and each MS has N_r receive antennas and M_r RF chains to multiplex N_s data streams. For effective communications, the constraints $N_\mathrm{s} = M_\mathrm{r} \ll N_\mathrm{r}$ and $KN_\mathrm{s} = M_\mathrm{t} \ll N_\mathrm{t}$ are assumed. The narrowband mmWave channel for the k-th MS (MS-k) is modeled by the Saleh–Valenzuela model [4] as

$$\mathbf{H}_k = \sqrt{\frac{\beta_k N_\mathrm{t} N_\mathrm{r}}{N_\mathrm{c} N_\mathrm{ray}}} \sum_{i=1}^{N_\mathrm{c}} \sum_{l=1}^{N_\mathrm{ray}} \alpha_{il}^k \mathbf{a}_\mathrm{r}(\phi_{il}^k) \mathbf{a}_\mathrm{t}^H(\theta_{il}^k), \tag{1}$$

which consists of N_c clusters, each propagating N_ray paths, with β_k denoting the large-scale fading factor. Uniform linear arrays (ULAs) are employed at the BS and all MSs. For the (i,l)-th path in $\mathbf{H}_k$, $\alpha_{il}^k \sim \mathcal{CN}(0,1)$ is the path gain, ϕ_{il}^k and θ_{il}^k are azimuth angles of arrival (AoA) and departure (AoD). $\mathbf{a}_\mathrm{r}(\phi_{il}^k)$ and $\mathbf{a}_\mathrm{t}(\theta_{il}^k)$ represent the receive and transmit array response vectors, respectively, of the form $\mathbf{a}(\psi) = \frac{1}{\sqrt{N}}[1, e^{\mathrm{j}\frac{2\pi}{\lambda} d \sin(\psi)}, \ldots, e^{\mathrm{j}(N-1)\frac{2\pi}{\lambda} d \sin(\psi)}]^T$, where N is the number of antennas and ψ can be the AoA or AoD, λ is the signal wavelength and $d = \frac{\lambda}{2}$ is the antenna spacing without loss of generality.

At the BS, the aggregated data streams $\mathbf{s} = [\mathbf{s}_1^T, \ldots, \mathbf{s}_K^T]^T$, where $\mathbf{s}_k \in \mathbb{C}^{N_\mathrm{s} \times 1}$ contains the data streams for MS-k,

Corresponding author: Zheng Wang (email: wznuaa@gmail.com) is processed by a digital precoder $\mathbf{F}_{\mathrm{D}} \in \mathbb{C}^{M_{\mathrm{t}} \times KN_{\mathrm{s}}}$ and an analog precoder $\mathbf{F}_{\mathrm{A}} \in \mathbb{C}^{N_{\mathrm{t}} \times M_{\mathrm{t}}}$. Thus, the signal transmitted to the mmWave channels is written as $\mathbf{x} = \mathbf{F}_{\mathrm{A}}\mathbf{F}_{\mathrm{D}}\mathbf{s}$. The total power budget P_{t} is allocated as $\mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \frac{P_{\mathrm{t}}}{KN_{\mathrm{s}}}\mathbf{I}_{KN_{\mathrm{s}}}$, where $\mathbb{E}\{\cdot\}$ denotes expectation. Thus, to meet the transmit power constraint $\mathbb{E}\{\|\mathbf{x}\|_2^2\} = P_{\mathrm{t}}$, the hybrid precoder should be normalized such that $\|\mathbf{F}_{\mathrm{A}}\mathbf{F}_{\mathrm{D}}\|_F^2 = KN_{\mathrm{s}}$.

At MS-k, the received signal is processed by an analog combiner $\mathbf{W}_{\mathrm{A},k} \in \mathbb{C}^{N_{\mathrm{r}} \times M_{\mathrm{r}}}$ and a digital combiner $\mathbf{W}_{\mathrm{D},k} \in \mathbb{C}^{M_{\mathrm{r}} \times N_{\mathrm{s}}}$. In particular, since the analog precoder and combiner are implemented by phase shifters, their entries are subject to constant modulus (CM) constraints, i.e., $|\mathbf{F}_{\mathrm{A}}(i,j)| = \frac{1}{\sqrt{N_{\mathrm{t}}}}, \forall i,j$ and $|\mathbf{W}_{\mathrm{A},k}(i,j)| = \frac{1}{\sqrt{N_{\mathrm{r}}}}, \forall i,j,k$. After the analog precoding and combining, the baseband channel for MS-k is defined as $\widetilde{\mathbf{H}}_k = \mathbf{W}_{\mathrm{A},k}^H \mathbf{H}_k \mathbf{F}_{\mathrm{A}}$.

After digital combining, the signal-to-interference-plus-noise ratio (SINR) of the i-th data stream in MS-k is

$$\mathrm{SINR}_{k_i} = \frac{S_{k_i}}{I_{k_i} + N_{k_i}}, \begin{cases} S_{k_i} = P_{k_i}|\mathbf{w}_{k_i}^H \widetilde{\mathbf{H}}_k \mathbf{f}_{k_i}|^2, \\ I_{k_i} = \sum_{\substack{l=1 \\ l \neq i}}^{N_{\mathrm{s}}} P_{k_l}|\mathbf{w}_{k_i}^H \widetilde{\mathbf{H}}_k \mathbf{f}_{k_l}|^2 \\ \quad + \sum_{\substack{m=1 \\ m \neq k}}^{K} \sum_{j=1}^{N_{\mathrm{s}}} P_{m_j}|\mathbf{w}_{k_i}^H \widetilde{\mathbf{H}}_k \mathbf{f}_{m_j}|^2, \\ N_{k_i} = \sigma_{\mathrm{n}}^2 \|\mathbf{w}_{k_i}^H \mathbf{W}_{\mathrm{A},k}^H\|_2^2, \end{cases} \tag{2}$$

where P_{k_i} is the power allocated to the i-th data stream of $\mathbf{s}_k$. The first term of I_{k_i} is the intra-user interference and the second term is the multi-user interference (MUI). σ_{n}^2 is the power of additive white Gaussian noise, $\mathbf{w}_{k_i}$ is the i-th column of $\mathbf{W}_{\mathrm{D},k}$, and $\mathbf{f}_{k_i}$ is the $[(k-1)N_{\mathrm{s}} + i]$-th column of $\mathbf{F}_{\mathrm{D}}$.

Assuming Gaussian data symbols, the design problem of hybrid precoding and combining to maximize the sum-rate is formulated as

$$\max_{\substack{\mathbf{F}_{\mathrm{A}}, \mathbf{F}_{\mathrm{D}}, \\ \{\mathbf{W}_{\mathrm{A},k}, \mathbf{W}_{\mathrm{D},k}\}_{k=1}^K}} \quad R = \sum_{k=1}^{K} \sum_{i=1}^{N_{\mathrm{s}}} \mathbb{E}\{\log_2(1 + \mathrm{SINR}_{k_i})\}, \tag{3a}$$

$$\text{s.t.} \quad \|\mathbf{F}_{\mathrm{A}}\mathbf{F}_{\mathrm{D}}\|_F^2 = KN_{\mathrm{s}}, \tag{3b}$$

$$\mathbf{F}_{\mathrm{A}} \in \mathcal{F}_{\mathrm{A}}, \tag{3c}$$

$$\mathbf{W}_{\mathrm{A},k} \in \mathcal{W}_{\mathrm{A},k}, \ \forall k, \tag{3d}$$

where $\mathcal{F}_{\mathrm{A}}$ and $\mathcal{W}_{\mathrm{A},k}, \forall k$, are the feasible sets for the analog precoder and combiners, respectively, satisfying the CM constraints. Unfortunately, the optimization problem (3) is non-convex, rendering its global optimum intractable. Consequently, existing methods [5], [7] adopt the two-stage design pattern as follows:

a) ***Analog Stage Design:*** The analog precoder $\mathbf{F}_{\mathrm{A}}$ and combiners $\{\mathbf{W}_{\mathrm{A},k}\}_{k=1}^K$ are obtained at this stage such that the aggregated baseband channel $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{H}}_1^T, \dots, \widetilde{\mathbf{H}}_K^T]^T$ is a diagonal low-rank approximation of $\mathbf{H} = [\mathbf{H}_1^T, \dots, \mathbf{H}_K^T]^T$. In this way, array gains are maximized and MUI is sufficiently suppressed. Assuming $\{\mathbf{W}_{\mathrm{A},k}\}_{k=1}^K$ and $\mathbf{F}_{\mathrm{A}}$ to be semi-unitary, the analog stage design problem is given as [5]

$$\max_{\mathbf{F}_{\mathrm{A}}, \{\mathbf{W}_{\mathrm{A},k}\}_{k=1}^K} \sum_{k=1}^{K} \|\mathbf{W}_{\mathrm{A},k}^H \mathbf{H}_k \mathbf{F}_{\mathrm{A}}\|_F^2, \tag{4a}$$

$$\text{s.t.} \quad \mathbf{F}_{\mathrm{A}}^H \mathbf{F}_{\mathrm{A}} = \mathbf{I}_{M_{\mathrm{t}}}, \ \mathbf{W}_{\mathrm{A},k}^H \mathbf{W}_{\mathrm{A},k} = \mathbf{I}_{M_{\mathrm{r}}}, \ \forall k, \tag{4b}$$

$$\mathbf{F}_{\mathrm{A}} \in \mathcal{F}_{\mathrm{A}}, \ \mathbf{W}_{\mathrm{A},k} \in \mathcal{W}_{\mathrm{A},k}, \ \forall k. \tag{4c}$$

After relaxing the CM constraints, non-convex problem (4) can be approximately solved by the generalized low rank approximations of matrices [5], which involves AO between $\mathbf{F}_{\mathrm{A}}$ and $\{\mathbf{W}_{\mathrm{A},k}\}_{k=1}^K$, incurring prohibitive feedback overhead.

b) ***Digital Stage Design:*** Based on the baseband channel $\{\widetilde{\mathbf{H}}_k\}_{k=1}^K$ formed by $\mathbf{F}_{\mathrm{A}}$ and $\{\mathbf{W}_{\mathrm{A},k}\}_{k=1}^K$, the digital precoder $\mathbf{F}_{\mathrm{D}}$ and combiners $\{\mathbf{W}_{\mathrm{D},k}\}_{k=1}^K$ are derived for interference mitigation.

III. Hybrid Precoding and Combining Design

In this section, we elaborate and analyze the proposed EMU-HPC method. In the analog stage, to maximize array gains without AO, we adopt the NI-GLRAM framework [8], which designs the combiners and precoder non-iteratively. Moreover, the analog precoding achieves a balance between array gains and MUI suppression. In the digital stage, RCD [9] is applied to mitigate the residual interference and noise, followed by an analysis of complexity and feedback overhead.

A. Analog Combining Design

The analog combiner for MS-k is designed to maximize the array gain. Specifically, following NI-GLRAM, we decouple the analog combining design from (4) as

$$\max_{\{\mathbf{W}_{\mathrm{A},k}\}_{k=1}^K} \sum_{k=1}^{K} \|\mathbf{W}_{\mathrm{A},k}^H \mathbf{H}_k\|_F^2, \tag{5a}$$

$$\text{s.t.} \quad \mathbf{W}_{\mathrm{A},k}^H \mathbf{W}_{\mathrm{A},k} = \mathbf{I}_{M_{\mathrm{r}}}, \mathbf{W}_{\mathrm{A},k} \in \mathcal{W}_{\mathrm{A},k}, \ \forall k. \tag{5b}$$

Since (5) remains non-convex, we relax CM constraints and obtain a semi-unitary solution $\widehat{\mathbf{W}}_{\mathrm{A},k}$ for MS-$k, \forall k$, given by the first M_{r} left singular vectors of $\mathbf{H}_k$, i.e., $\widehat{\mathbf{W}}_{\mathrm{A},k} = \mathbf{U}_k(:,1:M_{\mathrm{r}})$, with the singular value decomposition (SVD) $\mathbf{H}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^H$ [8]. However, direct TSVD at each MS may be prohibitive under limited computational resources.

To reduce this computational burden, we implement the VR-PCA method [10] within the LazySVD framework [11] to approximate the required TSVD at each MS, which is based on stochastic gradient ascent with guaranteed convergence [10]. Specifically, to estimate the m-th left singular vector of $\mathbf{H}_k$ via VR-PCA, the t-th iteration within epoch s updates as follows:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma[\mathbf{h}_{i_t}^{(m)}(\mathbf{h}_{i_t}^{(m)^H}\mathbf{w}_{t-1} - \mathbf{h}_{i_t}^{(m)^H}\widetilde{\mathbf{w}}_{s-1}) + \boldsymbol{\mu}_s], \tag{6}$$

where γ is the step size, $\mathbf{w}_t$ and $\widetilde{\mathbf{w}}_s$ are the per-iteration and per-epoch estimates with unit ℓ_2 norm. $\boldsymbol{\mu}_s$ is the gradient at $\widetilde{\mathbf{w}}_{s-1}$, and $\mathbf{h}_{i_t}^{(m)}$ is the i_t-th column of the deflated channel matrix $\mathbf{H}_k^{(m)}$ with i_t randomly selected. The maximum number of epochs is N_{epoch}, and the number of iterations per epoch is N_{t} [10]. VR-PCA converges if $\|\mathbf{w}_{N_{\mathrm{t}}} - \widetilde{\mathbf{w}}_{s-1}\|_2 \le \varepsilon$ is satisfied at the end of an iteration, where ε is the error tolerance.

In the LazySVD framework, the m-th left singular vector $\mathbf{u}_m^{'}$ obtained from VR-PCA is orthonormalized with all previous vectors $\mathbf{U}^{(m-1)} = [\mathbf{u}_1, \dots, \mathbf{u}_{m-1}]$ as

$$\mathbf{u}_m = \frac{(\mathbf{I}_{N_{\mathrm{r}}} - \mathbf{U}^{(m-1)}\mathbf{U}^{(m-1)^H})\mathbf{u}_m^{'}}{\left\|(\mathbf{I}_{N_{\mathrm{r}}} - \mathbf{U}^{(m-1)}\mathbf{U}^{(m-1)^H})\mathbf{u}_m^{'}\right\|_2}. \tag{7}$$

Algorithm 1 Analog Combining Scheme for MS-k

```
Input: $\mathbf{H}_k$, $N_{\text{epoch}}$, $\gamma$, $\varepsilon$
Output: $\mathbf{W}_{\text{A},k}$
Initialize $\mathbf{U}^{(0)} = \emptyset$, $\mathbf{H}_k^{(1)} = \mathbf{H}_k$
for $m = 1$ to $M_{\text{r}}$ do
    Initialize $\widetilde{\mathbf{w}}_0 \sim \mathcal{CN}(\mathbf{0}_{N_{\text{r}}\times 1}, \mathbf{I}_{N_{\text{r}}})$, $\widetilde{\mathbf{w}}_0 \leftarrow \frac{\widetilde{\mathbf{w}}_0}{\|\widetilde{\mathbf{w}}_0\|_2}$
    Initialize the error vector $\mathbf{e} = \widetilde{\mathbf{w}}_0$
    s = 1
    while $s \le N_{\text{epoch}}$ and $\|\mathbf{e}\|_2 > \varepsilon$ do
        $\boldsymbol{\mu}_s = \frac{1}{N_{\text{t}}}\sum_{i=1}^{N_{\text{t}}} \mathbf{h}_i^{(m)} \left(\mathbf{h}_i^{(m)^H} \widetilde{\mathbf{w}}_{s-1}\right)$
        $\mathbf{w}_0 = \widetilde{\mathbf{w}}_{s-1}$
        for $t = 1$ to $N_{\text{t}}$ do
            Sample $i_t$ randomly from $\{1, 2, \ldots, N_{\text{t}}\}$
            Get $\mathbf{h}_{i_t}^{(m)}$ and update $\mathbf{w}_t$ by (6)
            Normalize $\mathbf{w}_t$ as $\mathbf{w}_t \leftarrow \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|_2}$
        end for
        Obtain the error vector as $\mathbf{e} = \mathbf{w}_{N_{\text{t}}} - \widetilde{\mathbf{w}}_{s-1}$
        $\widetilde{\mathbf{w}}_s = \mathbf{w}_{N_{\text{t}}}$
        $s \leftarrow s + 1$
    end while
    Obtain $\mathbf{u}_m^{'}$ as $\mathbf{u}_m^{'} = \widetilde{\mathbf{w}}_{s-1}$ and obtain $\mathbf{u}_m$ by (7)
    Obtain $\mathbf{H}_k^{(m+1)}$ by (8)
    Enlarge $\mathbf{U}^{(m)}$ by $\mathbf{U}^{(m)} = [\mathbf{U}^{(m-1)}, \mathbf{u}_m]$
end for
Given $\widehat{\mathbf{W}}_{\text{A},k} = \mathbf{U}^{(M_{\text{r}})}$, obtain $\mathbf{W}_{\text{A},k}$ by (9)
```

Based on this, the deflated channel matrix $\mathbf{H}_k^{(m)}$ is updated by projecting it onto the orthogonal complement of $\mathbf{u}_m$ as

$$\mathbf{H}_k^{(m+1)} = (\mathbf{I}_{N_{\text{r}}} - \mathbf{u}_m \mathbf{u}_m^H)\mathbf{H}_k^{(m)}, \tag{8}$$

which is the input for the next round of VR-PCA. The resulting M_{r} vectors form the semi-unitary analog combiner $\widehat{\mathbf{W}}_{\text{A},k}$. Under the CM constraint, the analog combiner $\mathbf{W}_{\text{A},k}$ approximates $\widehat{\mathbf{W}}_{\text{A},k}$ via phase extraction, given as

$$\mathbf{W}_{\text{A},k} = \frac{1}{\sqrt{N_{\text{r}}}} \exp\left[\text{j}\angle\left(\widehat{\mathbf{W}}_{\text{A},k}\right)\right], \tag{9}$$

where $\angle(\cdot)$ represents the element-wise phase extraction for a matrix. After that, MS-k computes $\bar{\mathbf{H}}_k = \mathbf{W}_{\text{A},k}^H \mathbf{H}_k$ and feeds it back to the BS for analog precoding. The analog combining scheme for MS-k is summarized in Algorithm 1.

B. Analog Precoding Design

The analog precoding is designed to balance array gain enhancement and MUI suppression based on $\{\bar{\mathbf{H}}_k\}_{k=1}^K$ from all MSs. Following NI-GLRAM, the analog precoding design based on (4) is given as

$$\max_{\mathbf{F}_{\text{A}}} \sum_{k=1}^{K} \|\bar{\mathbf{H}}_k \mathbf{F}_{\text{A}}\|_F^2, \tag{10a}$$

$$\text{s.t.} \quad \mathbf{F}_{\text{A}}^H \mathbf{F}_{\text{A}} = \mathbf{I}_{M_{\text{t}}},\ \mathbf{F}_{\text{A}} \in \mathcal{F}_{\text{A}}. \tag{10b}$$

However, for each MS, (10) fails to distinguish its desired array gain from the MUI it causes to the other MSs, resulting in maximizing both. This leads to an imbalance between array gain harvesting and MUI suppression. To address this issue, we design a dedicated analog precoder $\mathbf{F}_{\text{A},k} \in \mathbb{C}^{N_{\text{t}} \times M_{\text{r}}}$ for each MS-k and form the overall analog precoder as $\mathbf{F}_{\text{A}} = [\mathbf{F}_{\text{A},1}, \ldots, \mathbf{F}_{\text{A},K}]$. Based on this, the desired array gain for MS-k is $\|\bar{\mathbf{H}}_k \mathbf{F}_{\text{A},k}\|_F^2$, while the MUI from MS-$k$ to the other MSs is $\sum_{p \neq k} \|\bar{\mathbf{H}}_p \mathbf{F}_{\text{A},k}\|_F^2$. However, jointly maximizing the former and minimizing the latter is difficult since the row space of $\bar{\mathbf{H}}_k$ misaligns with the common null space of $\bar{\mathbf{H}}_p$'s, $\forall p \neq k$.

To resolve this, we relax the CM constraint and design the column space of semi-unitary analog precoder $\widehat{\mathbf{F}}_{\text{A},k}$ to align with the row space of $\bar{\mathbf{H}}_k$ after projecting $\bar{\mathbf{H}}_k$ onto the joint orthogonal complement of $\widehat{\mathbf{F}}_{\text{A},1}, \ldots, \widehat{\mathbf{F}}_{\text{A},k-1}$, resulting in the residual channel matrix $\bar{\mathbf{H}}_{\text{res},k}$, given by

$$\bar{\mathbf{H}}_{\text{res},k} = \begin{cases} \bar{\mathbf{H}}_1, & k = 1 \\ \bar{\mathbf{H}}_k \prod_{i=1}^{k-1} \left(\mathbf{I}_{N_{\text{t}}} - \widehat{\mathbf{F}}_{\text{A},i} \widehat{\mathbf{F}}_{\text{A},i}^H\right), & k > 1 \end{cases} \tag{11}$$

if the previously obtained $\widehat{\mathbf{F}}_{\text{A},i}$, $\forall i < k$, are semi-unitary for $k = 2$, or mutually orthogonal for $k \ge 3$. Then, we modify the analog precoding problem for MS-k from (10) to a MUI-aware array gain maximization problem as follows:

$$\max_{\mathbf{F}_{\text{A},k}} \|\bar{\mathbf{H}}_{\text{res},k} \mathbf{F}_{\text{A},k}\|_F^2, \tag{12a}$$

$$\text{s.t.} \quad \mathbf{F}_{\text{A},k}^H \mathbf{F}_{\text{A},k} = \mathbf{I}_{M_{\text{r}}},\ \mathbf{F}_{\text{A},k} \in \mathcal{F}_{\text{A},k}, \tag{12b}$$

where $\mathcal{F}_{\text{A},k}$ is the feasible set for $\mathbf{F}_{\text{A},k}$. Its semi-unitary solution is $\widehat{\mathbf{F}}_{\text{A},k} = \bar{\mathbf{V}}_k(:, 1 : M_{\text{r}})$ obtained via SVD $\bar{\mathbf{H}}_{\text{res},k} = \bar{\mathbf{U}}_k \bar{\mathbf{D}}_k \bar{\mathbf{V}}_k^H$. Thus, by designing MS-1 to MS-K recursively, each $\widehat{\mathbf{F}}_{\text{A},k}$ satisfies the semi-unitary condition for $k = 1$ and the mutual orthogonality condition for $k \ge 2$, i.e., $\widehat{\mathbf{F}}_{\text{A},p}^H \widehat{\mathbf{F}}_{\text{A},q} = \mathbf{0}_{M_{\text{r}} \times M_{\text{r}}}, \forall p, q \le k, p \neq q$, thereby validating the computation of $\bar{\mathbf{H}}_{\text{res},k+1}$ via (11).

The elements of $\widehat{\mathbf{F}}_{\text{A},k}$ are then phase-extracted to meet the CM constraint. Therefore, $\mathbf{F}_{\text{A},k}$ is given by

$$\mathbf{F}_{\text{A},k} = \frac{1}{\sqrt{N_{\text{t}}}} \exp\left[\text{j}\angle\left(\widehat{\mathbf{F}}_{\text{A},k}\right)\right], \tag{13}$$

with the following MUI suppression property.

Proposition 1. *The analog precoder in* (13) *minimizes the upper bound of the prior MUI term* $\|\bar{\mathbf{H}}_p \mathbf{F}_{\text{A},k}\|_F^2$, $\forall p < k$.

Proof. From the mutual orthogonality condition and (11), it follows that the approximate prior MUI term satisfies $\bar{\mathbf{H}}_p \widehat{\mathbf{F}}_{\text{A},k} = \mathbf{0}_{M_{\text{r}} \times M_{\text{r}}}, \forall p < k$. Leveraging this property, the actual prior MUI term is upper-bounded as

$$\begin{aligned} \left\|\bar{\mathbf{H}}_p \mathbf{F}_{\text{A},k}\right\|_F^2 &= \left\|\bar{\mathbf{H}}_p \mathbf{F}_{\text{A},k} - \bar{\mathbf{H}}_p \widehat{\mathbf{F}}_{\text{A},k}\right\|_F^2 \\ &\le \left\|\bar{\mathbf{H}}_p\right\|_2^2 \left\|\widehat{\mathbf{F}}_{\text{A},k} - \mathbf{F}_{\text{A},k}\right\|_F^2. \end{aligned} \tag{14}$$

Since $\|\bar{\mathbf{H}}_p\|_2$ is constant, minimizing the upper bound is equivalent to minimizing $\|\widehat{\mathbf{F}}_{\text{A},k} - \mathbf{F}_{\text{A},k}\|_F^2$ to obtain $\mathbf{F}_{\text{A},k}$ under the CM constraint. As shown in [12], the optimal analog precoder achieving this minimization is given by (13), concluding the proof. □

In summary, for each MS, the MUI generated to prior MSs is effectively suppressed, while the MUI to later MSs persists due to the balance with array gain harvesting.

C. Digital Precoding and Combining Design

In the digital stage, RCD is applied to the baseband channel $\widetilde{\mathbf{H}}_k, \forall k$, to suppress the remaining interference and noise.

The RCD method first addresses the MUI plus noise under minimum mean square error criterion, where the digital precoder for MS-k is $\mathbf{T}_k^{(\mathrm{a})}$. We define $\bar{\bar{\mathbf{H}}}_k = [\widetilde{\mathbf{H}}_1^T, \dots, \widetilde{\mathbf{H}}_{k-1}^T, \widetilde{\mathbf{H}}_{k+1}^T, \dots, \widetilde{\mathbf{H}}_K^T]^T$, so that $\|\bar{\bar{\mathbf{H}}}_k \mathbf{T}_k^{(\mathrm{a})}\|_F^2$ quantifies the MUI generated by MS-k to the other MSs, and the design problem of $\{\mathbf{T}_k^{(\mathrm{a})}\}_{k=1}^K$ is given by [7]

$$\min_{\{\mathbf{T}_k^{(\mathrm{a})}\}_{k=1}^K} \sum_{k=1}^{K} \left(\left\| \bar{\bar{\mathbf{H}}}_k \mathbf{T}_k^{(\mathrm{a})} \right\|_F^2 + \frac{K N_{\mathrm{s}} \sigma_{\mathrm{n}}^2}{P_{\mathrm{t}}} \|\mathbf{T}_k^{(\mathrm{a})}\|_F^2 \right). \tag{15}$$

One of its non-trivial solutions is given by

$$\mathbf{T}_k^{(\mathrm{a})} = \left(\bar{\bar{\mathbf{H}}}_k^H \bar{\bar{\mathbf{H}}}_k + \frac{K N_{\mathrm{s}} \sigma_{\mathrm{n}}^2}{P_{\mathrm{t}}} \mathbf{I}_{M_{\mathrm{t}}} \right)^{-1}, \tag{16}$$

which reduces the MUI to negligible levels. This enables each MS to optimize the data stream transmission in its effective subchannel $\widetilde{\mathbf{H}}_k \mathbf{T}_k^{(\mathrm{a})}$, with SVD $\widetilde{\mathbf{H}}_k \mathbf{T}_k^{(\mathrm{a})} = \mathbf{U}_{\mathrm{b},k} \mathbf{D}_{\mathrm{b},k} \mathbf{V}_{\mathrm{b},k}^H$. For the k-th subchannel, its optimal digital precoder $\mathbf{T}_k^{(\mathrm{b})}$ and the digital combiner $\mathbf{W}_{\mathrm{D},k}$, which jointly eliminate the intra-user interference are designed as

$$\mathbf{T}_k^{(\mathrm{b})} = \mathbf{V}_{\mathrm{b},k}(:, 1 : N_{\mathrm{s}}), \quad \mathbf{W}_{\mathrm{D},k} = \mathbf{U}_{\mathrm{b},k}(:, 1 : N_{\mathrm{s}}). \tag{17}$$

Finally, under the total transmit power constraint, the normalized digital precoder for MS-k is given as

$$\mathbf{F}_{\mathrm{D},k} = \sqrt{\frac{K N_{\mathrm{s}}}{\sum_{i=1}^{K} \|\mathbf{F}_{\mathrm{A}} \mathbf{T}_i^{(\mathrm{a})} \mathbf{T}_i^{(\mathrm{b})}\|_F^2}} \mathbf{T}_k^{(\mathrm{a})} \mathbf{T}_k^{(\mathrm{b})}, \tag{18}$$

and the overall digital precoder is $\mathbf{F}_{\mathrm{D}} = [\mathbf{F}_{\mathrm{D},1}, \dots, \mathbf{F}_{\mathrm{D},K}]$.

In summary, the proposed EMU-HPC method is shown in Algorithm 2 and its computational complexity is further examined. Specifically, the analog combining has a complexity order of $O(K N_{\text{epoch}} N_{\mathrm{s}} N_{\mathrm{t}} N_{\mathrm{r}})$, dominated by the computation of the $\boldsymbol{\mu}_s$ and $\mathbf{w}_t$ in VR-PCAs. In analog precoding, the complexity order is $O(K^2 N_{\mathrm{s}}^2 N_{\mathrm{t}})$, primarily from computing the residual channel matrices. In the digital stage, RCD incurs a complexity of $O(K^4 N_{\mathrm{s}}^3)$. Overall, EMU-HPC has a total complexity order of $O\big(K N_{\mathrm{s}} N_{\mathrm{t}} N_{\mathrm{r}} N_{\text{epoch}} + K^2 N_{\mathrm{s}}^2 N_{\mathrm{t}} + K^4 N_{\mathrm{s}}^3\big)$.

For comparison, Table I summarizes the complexity orders of different hybrid precoding and combining algorithms. It shows that the complexity of EMU-HPC is significantly lower than HyEB [6] and HRCD [9], and is no higher than HyBD [4], HySBD [12], and SROA [7], as verified by simulations.

Additionally, the feedback overhead of EMU-HPC in the analog stage is $K N_{\mathrm{s}} N_{\mathrm{t}}$, since $\{\bar{\mathbf{H}}_k\}_{k=1}^K$ is fed back only once. This is identical to HyBD, HySBD, and HyEB, but much lower than $K N_{\mathrm{s}} N_{\text{iter}}(N_{\mathrm{t}} + N_{\mathrm{r}})$ required by SROA, where N_{iter} is the maximum number of data-feedback rounds among all AOs. These AOs may cause extra communication delays and further hinder practical deployment. Besides, all algorithms have identical feedback overhead in the digital stage.

Algorithm 2 The Proposed Efficient Multi-User Hybrid Precoding and Combining (EMU-HPC) Algorithm

```
Input: $\{\mathbf{H}_k\}_{k=1}^K$, $N_{\text{epoch}}$, $\gamma$, $\varepsilon$
Output: $\{\mathbf{W}_{\mathrm{A},k}, \mathbf{W}_{\mathrm{D},k}\}_{k=1}^K$, $\mathbf{F}_{\mathrm{A}}$, $\mathbf{F}_{\mathrm{D}}$
for $k = 1$ to $K$ do
    Obtain $\mathbf{W}_{\mathrm{A},k}$ by Algorithm 1
    Compute $\bar{\mathbf{H}}_k$ by $\bar{\mathbf{H}}_k = \mathbf{W}_{\mathrm{A},k}^H \mathbf{H}_k$
end for
Initialize $\mathbf{F}_{\mathrm{A}} = \emptyset$ and $\bar{\mathbf{H}}_{\mathrm{res},k} = \bar{\mathbf{H}}_k, \forall k$
for $k = 1$ to $K$ do
    Perform SVD of $\bar{\mathbf{H}}_{\mathrm{res},k}$ as $\bar{\mathbf{H}}_{\mathrm{res},k} = \bar{\mathbf{U}}_k \bar{\mathbf{D}}_k \bar{\mathbf{V}}_k^H$
    Obtain $\widehat{\mathbf{F}}_{\mathrm{A},k}$ by $\widehat{\mathbf{F}}_{\mathrm{A},k} = \bar{\mathbf{V}}_k(:, 1 : M_{\mathrm{r}})$
    Obtain $\mathbf{F}_{\mathrm{A},k}$ by (13)
    for $p = k + 1$ to $K$ do
        $\bar{\mathbf{H}}_{\mathrm{res},p} \leftarrow \bar{\mathbf{H}}_{\mathrm{res},p}(\mathbf{I}_{N_{\mathrm{t}}} - \widehat{\mathbf{F}}_{\mathrm{A},k} \widehat{\mathbf{F}}_{\mathrm{A},k}^H)$
    end for
    $\mathbf{F}_{\mathrm{A}} \leftarrow [\mathbf{F}_{\mathrm{A}}, \mathbf{F}_{\mathrm{A},k}]$
end for
for $k = 1$ to $K$ do
    Obtain $\mathbf{T}_k^{(\mathrm{a})}$ by (16)
    Obtain $\widetilde{\mathbf{H}}_k \mathbf{T}_k^{(\mathrm{a})}$ with SVD $\widetilde{\mathbf{H}}_k \mathbf{T}_k^{(\mathrm{a})} = \mathbf{U}_{\mathrm{b},k} \mathbf{D}_{\mathrm{b},k} \mathbf{V}_{\mathrm{b},k}^H$
    Obtain $\mathbf{T}_k^{(\mathrm{b})}$ and $\mathbf{W}_{\mathrm{D},k}$ by (17)
end for
Obtain the digital precoders by (18), $\forall k$, and form $\mathbf{F}_{\mathrm{D}}$
```

TABLE I
COMPARISON OF COMPUTATIONAL COMPLEXITY ORDERS

Algorithm	**Computational Complexity Order**
EMU-HPC	$O\left[K N_{\mathrm{t}} N_{\mathrm{s}} (N_{\text{epoch}} N_{\mathrm{r}} + K N_{\mathrm{s}}) + K^4 N_{\mathrm{s}}^3\right]$
HyBD	$O\left[K N_{\mathrm{t}} \left(N_{\mathrm{r}}^2 + K N_{\mathrm{s}}^2\right) + K^4 N_{\mathrm{s}}^3\right]$
HySBD	$O\left[K N_{\mathrm{t}} \left(N_{\mathrm{r}}^2 + K N_{\mathrm{s}}^2\right) + K^4 N_{\mathrm{s}}^3\right]$
HyEB	$O[K N_{\mathrm{s}}^2 N_{\text{iter}} (N_{\mathrm{t}}^2 + N_{\mathrm{r}}^2) + K^2 N_{\mathrm{s}}^4 N_{\mathrm{t}} + K^4 N_{\mathrm{s}}^3]$
HRCD	$O\left[K N_{\mathrm{t}} \left(N_{\mathrm{r}}^2 + K N_{\mathrm{s}}^2\right) + K^4 N_{\mathrm{s}}^3 + K N_{\mathrm{s}} N_{\mathrm{c}} N_{\text{ray}}\right]$
SROA	$O\left[K N_{\mathrm{t}} N_{\mathrm{s}} (N_{\text{iter}} N_{\mathrm{r}} + K N_{\mathrm{r}}) + K^4 N_{\mathrm{s}}^3\right]$

IV. Simulation Results

Apart from setting the large-scale fading factor to be $\beta_k = 1, \forall k$, the mmWave channel configuration follows that in [4]. The signal-to-noise ratio (SNR) is defined as $P_{\mathrm{t}}/\sigma_{\mathrm{n}}^2$. The EMU-HPC algorithm is compared with HyBD [4], HySBD [12], MGLRAM [5], HRCD [9], and SROA [7]. The capacity-achieving iterative waterfilling-based dirty paper coding (IWDPC) [13] is used as a benchmark for performance upper bound. For EMU-HPC, the step size of VR-PCA is $\gamma = 0.01$, the maximum number of epochs is 10 and the error tolerance is $\varepsilon = 10^{-3}$ (same as SROA for fairness).

Fig. 1 illustrates the sum-rate performance over SNR in the mmWave channel, showing that EMU-HPC outperforms other hybrid precoding and combining methods, except at very low SNR. This improvement is mainly achieved through joint interference management. Compared with SROA, EMU-HPC requires far less feedback overhead to achieve a superior sum-rate, rendering it more practical for deployment. Fig. 2 further

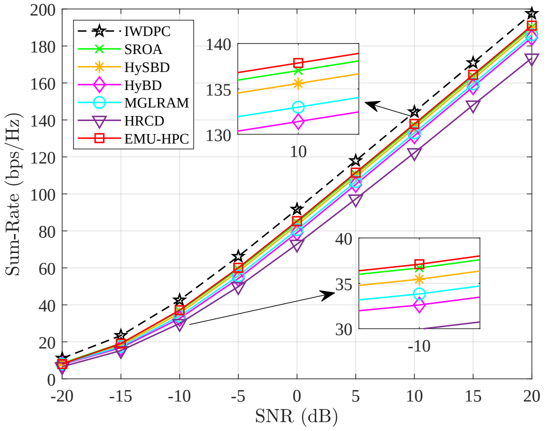

Fig. 1. Sum-rate vs. SNR in range of [-20,20] dB for 256 × 16 8-user mmWave massive MIMO systems with $N_\mathrm{s} = 2$.

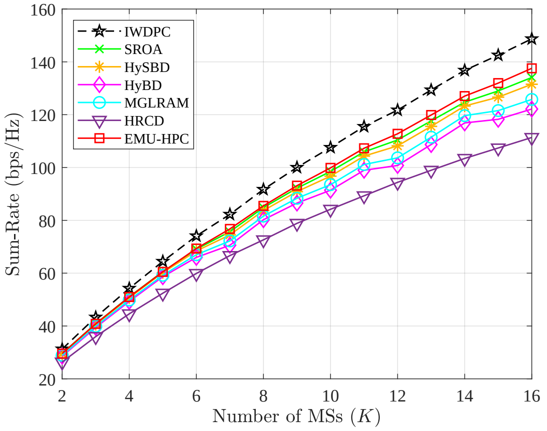

Fig. 2. Sum-rate vs. the number of MSs in the range of $[2, 16]$ for 256×16 mmWave massive MIMO systems with $N_\mathrm{s} = 2$ (SNR=0dB).

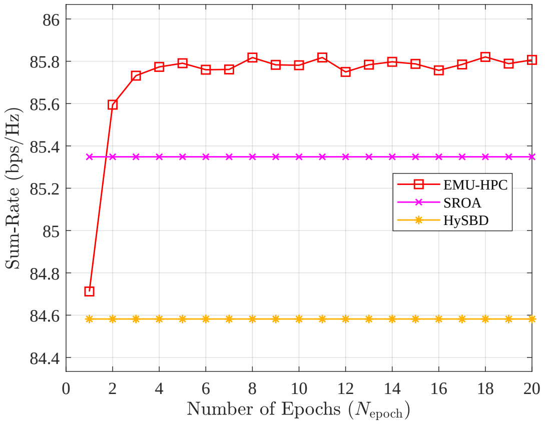

Fig. 3. Sum-rate vs. the maximum number of epochs in the range of $[1, 20]$ for 256×16 8-user mmWave massive MIMO systems with $N_\mathrm{s} = 2$ (SNR=0dB).

presents the sum-rate performance versus the increasing numbers of MSs, where EMU-HPC achieves higher sum-rate than other hybrid precoding and combining schemes at medium and high K, and remains competitive at low K. Compared with MGLRAM, which maximizes both MUI and array gains via (10), EMU-HPC attains better performance by balancing array gains and MUI suppression during analog precoding.

Fig. 3 illustrates the sum-rate performance as N_epoch increases, indicating that EMU-HPC requires only $N_\mathrm{epoch} = 2$ to surpass SROA, thereby satisfying $N_\mathrm{epoch} N_\mathrm{r} + K N_\mathrm{s} < N_\mathrm{iter} N_\mathrm{r} + K N_\mathrm{r}$ with N_iter averaging $\bar{N}_\mathrm{iter} = 13.96$. Moreover, $N_\mathrm{epoch} = 1$ is sufficient to outperform HySBD, resulting in $N_\mathrm{s} N_\mathrm{epoch} N_\mathrm{r} \ll N_\mathrm{r}^2$. Hence, EMU-HPC achieves the lowest computational complexity among all hybrid precoding and combining methods in Table I, demonstrating its efficiency.

V. Conclusion

In this paper, we propose the EMU-HPC method for mmWave massive MU-MIMO to improve the sum-rate with low feedback overhead. NI-GLRAM is employed in the analog stage to maximize the array gain with minimal feedback, while joint analog–digital interference suppression is realized through the orthogonal matrix projection in analog precoding and RCD in the digital stage. Numerical results demonstrate that EMU-HPC achieves better performance compared with existing hybrid precoding and combining algorithms.

Acknowledgment

This work was supported in part by the National Key R&D Program of China under Grants No.2023YFC2205501, and in part by National Natural Science Foundation of China under Grants No.62371124.

References

[1] N. Li and P. Fan, "Distributed cell-free massive MIMO versus cellular massive MIMO under UE hardware impairments," *Chinese Journal of Electronics*, vol. 33, no. 5, pp. 1274–1285, 2024.

[2] X. Zhan, Z. Sun, F. Shu, Y. Chen, X. Cheng, Y. Wu, Q. Zhang, Y. Li, and P. Zhang, "Rapid phase ambiguity elimination methods for DOA estimator via hybrid massive MIMO receive array," *Chinese Journal of Electronics*, vol. 33, no. 1, pp. 175–184, 2024.

[3] Q. Zhang and B. Wu, "Characteristic mode analysis for pattern diversity and beamforming: A survey," *Chinese Journal of Electronics*, vol. 33, no. 5, pp. 1117–1126, 2024.

[4] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201–211, Jan. 2016.

[5] N. Song, H. Sun, and T. Yang, "Coordinated hybrid beamforming for millimeter wave multi-user massive MIMO systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2016, pp. 1–6.

[6] C. Hu, J. Liu, X. Liao, Y. Liu, and J. Wang, "A novel equivalent baseband channel of hybrid beamforming in massive multiuser MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 764–767, Apr. 2018.

[7] G. M. Zilli and W. P. Zhu, "Constrained channel decomposition-based hybrid beamforming for mmwave massive MIMO systems," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1707–1720, 2020.

[8] J. Liu and S. Chen, "Non-iterative generalized low rank approximation of matrices," *Pattern Recognit. Lett.*, vol. 27, no. 9, pp. 1002–1008, 2006.

[9] F. Khalid, "Hybrid beamforming for millimeter wave massive multiuser MIMO systems using regularized channel diagonalization," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 705–708, Jun. 2019.

[10] O. Shamir, "A stochastic PCA and SVD algorithm with an exponential convergence rate," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 144–152.

[11] Z. Allen-Zhu and Y. Li, "LazySVD: Even faster SVD decomposition yet without agonizing pain," in *Proc. 30th Int. Conf. on Neural Inf. Process. Syst.*, 2016, pp. 982–990.

[12] X. Wu, D. Liu, and F. Yin, "Hybrid beamforming for multi-user massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3879–3891, Sep. 2018.

[13] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.