Distributed Precoder based on weighted MMSE with Low Complexity for Massive MIMO Systems

Ningxin Zhou, *Student Member, IEEE,* Zheng Wang, *Senior Member, IEEE,*
Cong Ma, Yongming Huang, *Senior Member, IEEE,* Qingjiang Shi, *Member, IEEE*

***Abstract*—The weighted sum mean squared error minimization (WMMSE) algorithm has gained widespread adoption owing to its superior performance. In this paper, we propose a novel low-complexity distributed WMMSE (LCD-WMMSE) algorithm, which is implemented over a decentralized architecture based on ring topology. Although each distributed unit (DU) in LCD-WMMSE works only with the local channel state information (CSI), LCD-WMMSE algorithm is still able to approach the performance of traditional WMMSE. Moreover, we show that LCD-WMMSE is also scalable since its required interconnection bandwidth is independent of the number of transmitter antennas, making it promising to various scenarios of massive MIMO. Simulation results validate that the proposed LCD-WMMSE algorithm not only achieves the low complexity cost in a distributed manner but also exhibits negligible performance loss.**

***Index Terms*—Distributed precoding, massive MIMO, decentralized architecture, maximization sum-rate, WMMSE**

I. Introduction

Recently, distributed signal processing has emerged as a promising research focus in the field of massive MIMO while a number of distributed downlink precoding schemes have been proposed for alleviating the computational pressure of center computing units [1]. Specifically, a coordinate descent-based precoding algorithm has been proposed based on a fully-decentralized architecture [2]. Meanwhile, a gradient descent-based following the daisy chain topology is given in [3] without sharing the local CSI and the involvement of a central node. As for cell-free massive MIMO systems, a distributed precoding scheme is presented in [4] using only channel statistics of the access points. A distributed precoder based on the virtual weighted sum-rate (WSR) is given by [5], which eliminates the need for additional exchanges throughout the iterative process.

Among numerous downlink precoding schemes, the celebrated WMMSE has been widely employed due to its ability to return the stationary point of the maximum WSR problem [6], [7]. In order to implement it in a distributed way, some distributed WMMSE schemes have been given. Typically, a distributed reduced WMMSE (DR-WMMSE) algorithm is proposed in [8], which works on a decentralized architecture based on star topology. The DR-WMMSE is lossless in performance, but it heavily relies on the central unit (CU). Moreover, a distributed WMMSE algorithm is designed in [9] to serve cell-free MIMO systems. However, since each AP works only with the local CSI, the performance loss becomes inevitable. Different from it, the distributed WMMSE algorithm in [10] applies the over-the-air (OTA) signalling mechanism for the required CSI exchange.

In this paper, to realize the distributed WMMSE for single-cell multi-user massive MIMO systems, a novel low-complexity distributed WMMSE algorithm named as LCD-WMMSE is proposed. In contrast to the distributed WMMSE schemes in [9], [10] that require the information exchange between receivers and transmitters, LCD-WMMSE algorithm only takes the computational information of DUs on the transmit side into account but with the negligible performance loss. On the other hand, because the amount of interconnection data in LCD-WMMSE is independent of the number of transmitter antennas, LCD-WMMSE is also scalable as well to facilitate the demand of higher dimensional systems.

II. WMMSE for downlink massive MIMO systems

Considering a single-cell multi-user massive MIMO system with M antennas base station (BS) and K users, where each user has N receiving antennas. Let $\mathbf{s}_k \in \mathbb{C}^d$ denote the signal vector transmitting to user $k \in \mathcal{K} = \{1, ..., K\}$, where d represents the number of data streams. Meanwhile, it is also assumed that the signal vectors transmitted to different users are independent with $\mathbb{E}\left[\mathbf{s}_k\mathbf{s}_k^H\right] = \mathbf{I}$ and zero mean. Then denoting $\mathbf{V}_k \in \mathbb{C}^{M\times d}$ as the precoding matrix used by BS for sending the signal vector $\mathbf{s}_k$ to user k, the received signal vector $\mathbf{y}_k \in \mathbb{C}^N$ at user k can be expressed as

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{V}_k\mathbf{s}_k + \sum_{m=1/m\neq k}^{K} \mathbf{H}_k\mathbf{V}_m\mathbf{s}_m + \mathbf{n}_k. \tag{1}$$

Here, $\mathbf{H}_k \in \mathbb{C}^{N\times M}$ stands for the channel matrix between BS and user k, and $\mathbf{n}_k \in \mathbb{C}^{N\times 1}$ represents the additive white Gaussian noise (AWGN) with $\mathcal{CN}(\mathbf{0}, \sigma_k^2\mathbf{I})$. Therefore, given the system model in (1), the signal to interference plus noise ratio (SINR) of the k-th user is given by

$$SINR_k = \mathbf{H}_k\mathbf{V}_k\mathbf{V}_k^H\mathbf{H}_k^H\left(\sum_{m\neq k}^{\mathcal{K}} \mathbf{H}_k\mathbf{V}_m\mathbf{V}_m^H\mathbf{H}_k^H + \sigma_k^2\mathbf{I}\right)^{-1}, \tag{2}$$

This work was supported in part by National Key Research and Development Program of China under Grant 2023YFC2205501, in part by the National Natural Science Foundation of China under Grant 62371124, 62231019, 62225107, 61720106003, in part by the Fundamental Research Funds for the Central Universities under Grant 2242022k60002. *(Corresponding author: Zheng Wang.)*

Ningxin Zhou, Zheng Wang, and Yongming Huang are with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (email: ningxin_zhou@seu.edu.cn; wznuaa@gmail.com; huangym@seu.edu.cn); C. Ma is with ZTE Corporation, Shenzhen 518057, China (e-mail:ma.cong@zte.com.cn); Qingjiang Shi is with the School of Software Engineering, Tongji University, Shanghai 201804, China(e-mail:shiqj@tongji.edu.cn).

Algorithm 1 The Classic WMMSE Algorithm [7]

```
Input: $\mathbf{H}_k$, $\varepsilon$, $\{\sigma_k\}_{k\in\mathcal{K}}$, $P$, $\forall k$.
Output: $\mathbf{U}_k$,$\mathbf{W}_k$,$\mathbf{V}_k$, $\forall k$.
Initialize: Set $\mathbf{V}_k$ to satisfy $\mathrm{Tr}(\mathbf{V}_k\mathbf{V}_k^H) = P/K, \forall k$.
repeat
    $\hat{\mathbf{W}}_k = \mathbf{W}_k$
    Update $\mathbf{U}_k$, via (7), $\forall k$.
    Update $\mathbf{W}_k$, via (8), $\forall k$.
    Update $\mathbf{V}_k$, via (9), $\forall k$.
until $|\sum_{k=1}^{K}\log\det(\mathbf{W}_k) - \sum_{k=1}^{K}\log\det(\hat{\mathbf{W}}_k)| < \varepsilon$
```

which leads to the downlink achievable rate of user k as

$$R_k = \log\det(\mathbf{I} + SINR_k). \tag{3}$$

Theoretically, based on R_k, precoding aims to maximize the weighted sum rate $\sum_{k=1}^{K}\alpha_k R_k$ with the power constraint, so that the optimization problem can be formulated as

$$\begin{aligned} &\max_{\{\mathbf{V}_k\}_{k\in\mathcal{K}}} \sum_{k=1}^{K}\alpha_k R_k \\ &\text{s.t.} \quad \mathrm{Tr}\left(\sum_{k=1}^{K}\mathbf{V}_k\mathbf{V}_k^H\right) \le P, \end{aligned} \tag{4}$$

where α_k is the weight of user k and P denotes the total transmit power budget at BS.However, since such a problem is non-convex and NP-hard, finding its optimal solution turns out to be unaffordable. In order to solve it, an equivalent problem is established as follows [7], [8]

$$\min_{\{\mathbf{U}_k,\mathbf{W}_k,\mathbf{V}_k\}_{k\in\mathcal{K}}} \sum_{k=1}^{K}\alpha_k\Big(\mathrm{Tr}\left(\mathbf{W}_k\mathbf{E}_k\right) - \log\det\left(\mathbf{W}_k\right)\Big) \tag{5a}$$

$$\text{s.t.} \quad \mathrm{Tr}\left(\sum_{k=1}^{K}\mathbf{V}_k\mathbf{V}_k^H\right) \le P, \tag{5b}$$

with the mean squared error (MSE) matrix

$$\begin{aligned}\mathbf{E}_k = &\left(\mathbf{I} - \mathbf{U}_k^H\mathbf{H}_k\mathbf{V}_k\right)\left(\mathbf{I} - \mathbf{U}_k^H\mathbf{H}_k\mathbf{V}_k\right)^H \\ &+ \sum_{m\neq k}^{K}\mathbf{U}_k^H\mathbf{H}_k\mathbf{V}_m\mathbf{V}_m^H\mathbf{H}_k^H\mathbf{U}_k + \sigma_k^2\mathbf{U}_k^H\mathbf{U}_k,\end{aligned} \tag{6}$$

where $\mathbf{U}_k \in \mathbb{C}^{N\times d}$ and $\mathbf{W}_k \in \mathbb{C}^{d\times d}$ are two introduced auxiliary matrices.

Although the problem in (5) is also non-convex, the problem of optimizing any single variable of $\mathbf{U}_k$, $\mathbf{W}_k$ and $\mathbf{V}_k$ in (5) is convex. To this end, WMMSE employs the block coordinate descent (BCD) method to converge towards the stationary point of the problem in (5), where the expressions of updating $\mathbf{U}_k$, $\mathbf{W}_k$ and $\mathbf{V}_k$ during each BCD iteration are [7]

$$\mathbf{U}_k = \left(\sum_{m=1}^{K}\mathbf{H}_k\mathbf{V}_m\mathbf{V}_m^H\mathbf{H}_k^H + \sigma_k^2\mathbf{I}\right)^{-1}\mathbf{H}_k\mathbf{V}_k, \tag{7}$$

$$\mathbf{W}_k = \left(\mathbf{I} - {\mathbf{U}_k}^H\mathbf{H}_k\mathbf{V}_k\right)^{-1}, \tag{8}$$

$$\mathbf{V}_k = \alpha_k\left(\sum_{m=1}^{K}\alpha_m\mathbf{H}_m^H\mathbf{U}_m\mathbf{W}_m\mathbf{U}_m^H\mathbf{H}_m + \mu\mathbf{I}\right)^{-1}\mathbf{H}_k^H\mathbf{U}_k\mathbf{W}_k. \tag{9}$$

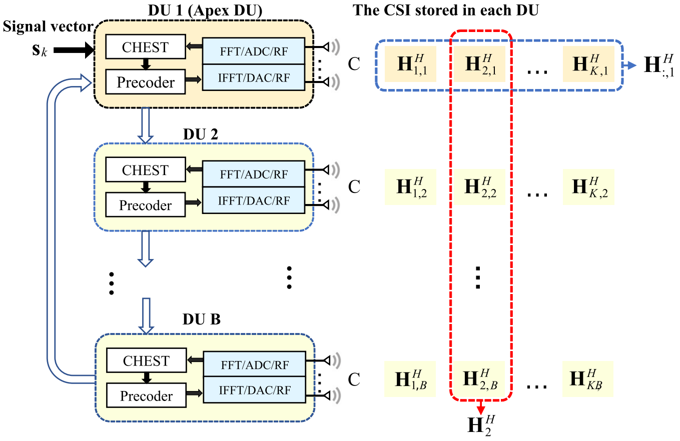

Fig. 1. Decentralized architecture based on ring topology with apex DU 1, where CHEST represents channel estimation.

III. Low-complexity Distributed WMMSE Algorithm

In particular, assume that the M antennas at BS can be partitioned into B ($1 \le B \le M$) distributed units (DUs), where each DU has C transmit antennas (i.e. $BC = M$) with its own dedicated RF circuitry and baseband signal processing units. Consequently, the local channel matrix between DU b and user k is $\mathbf{H}_{k,b} \in \mathbb{C}^{N\times C}$, and $\mathbf{H}_{:,b} = [\mathbf{H}_{1,b}^H, \cdots, \mathbf{H}_{K,b}^H]^H \in \mathbb{C}^{KN\times C}$ represents the local channel state information (CSI) stored in the b-th DU. Similarly, the local precoding matrix in the b-th DU is $\mathbf{V}_{:,b} = [\mathbf{V}_{1,b}, \cdots, \mathbf{V}_{K,b}] \in \mathbb{C}^{C\times Kd}$, where $\mathbf{V}_{k,b} \in \mathbb{C}^{C\times d}$ denotes the precoding matrix for user k. Moreover, the relationship between the local channel matrix $\mathbf{H}_{k,b}$ and the global channel matrix $\mathbf{H}_k$ is $\mathbf{H}_k = [\mathbf{H}_{k,1}, \cdots, \mathbf{H}_{k,B}]$, and similarly we have $\mathbf{V}_k^H = [\mathbf{V}_{k,1}^H, \cdots, \mathbf{V}_{k,B}^H]^H$.

In order to solve the problem in (5) based on these DUs, a decentralized architecture designed for the proposed LCD-WMMSE algorithm is displayed in Fig. 1. Intuitively, every DU is connected to another two adjacent DUs with unidirectional interconnections, forming a ring structure. Besides, the first DU called apex DU is additionally connected to CPU for receiving the signal vector $\mathbf{s}_k$, and the apex DU needs to undertake additional computational load compared with other DUs. In other words, CPU only sends the information $\mathbf{s}_k$ to the apex DU without participating in the following calculations. Clearly, each DU should independently calculate its local precoding matrix $\mathbf{V}_{:,b}$ only with its local CSI $\mathbf{H}_{:,b}$ and the information received from the last DU.

Based on such a decentralized architecture, the proposed LCD-WMMSE algorithm has a dual-layer structure. Specifically, the outer loop indexed with t is used to update $\mathbb{U}^t = \{\mathbf{U}_k^t\}_{k\in\mathcal{K}}$ and $\mathbb{W}^t = \{\mathbf{W}_k^t\}_{k\in\mathcal{K}}$ and contains the inner loop with L iterations for locally updating $\mathbf{V}_{:,b}$ at each DU, which is shown as follows,

$$\underbrace{\mathbb{U}^t \to \mathbb{W}^t \to \overbrace{\mathbf{V}_{:,1}^{t,1} \to \ldots \to \overbrace{\mathbf{V}_{:,1}^{t,l} \to \ldots \to \mathbf{V}_{:,B}^{t,l}}^{l\text{-th Inner Loop}} \to \ldots \to \mathbf{V}_{:,B}^{t,L}}^{L\text{ Inner Loops}}}_{t\text{-th Outer Loop}}.$$

A. Distributed Computation of $\mathbf{V}_k$ in the Inner Loop

During the inner loop of LCD-WMMSE, the block non-linear Gauss-Seidel (BGS) method is used to compute each precoding submatrix $\mathbf{V}_{:,b}$ locally. Technically speaking, the block non-linear GS method is an effective method to solve the minimization problem with constraint in a distributed way [11]. In particular, consider a constrained convex optimization problem as follows

$$\begin{aligned} \min_{\mathbf{x}} \quad & F(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n, \end{aligned} \tag{10}$$

where the constraint set $\mathcal{X}$ can be written as the *Cartesian* product of I convex compact sets, i.e $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_I$ with $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}, i \in \{1, ..., I\}$ and $\sum_{i=1}^{I} n_i = n$. Then, according to the structure of constraint $\mathcal{X}$, the optimized variable $\mathbf{x}$ can be expressed as $\mathbf{x}^T = [\mathbf{x}_1^T \quad \mathbf{x}_2^T \cdots \mathbf{x}_I^T]$ with $\mathbf{x}_i \in \mathbb{R}^{n_i}$. To solve the problem in (10), BGS iteratively updates the component vector $\mathbf{x}_i$ via the following subproblem,

$$\mathbf{x}_i^l = \arg\min_{\mathbf{x}_i \in \mathcal{X}_i} F(\mathbf{x}_1^l, ..., \mathbf{x}_{i-1}^l, \mathbf{x}_i, \mathbf{x}_{i+1}^{l-1}, ..., \mathbf{x}_I^{l-1}), \tag{11}$$

where the superscript l represents the l-th iteration. It has been shown in [11] that BGS is able to descend to the minimum objective function on each component $\mathbf{x}_i$ under its respective constraint $\mathcal{X}_i$ in turn, so as the global minimum of (10).

However, different from the problem in (10), the optimization problem in (5) has the coupled constraint. Therefore, to apply BGS, the constraint in (5b) need to be decoupled into the local and independent ones. And then the optimization problem of $\mathbf{V}_{:,b}$ for the b-th DU at the l-th inner iteration and t-th outer iteration can be formulated as follows:

$$\min_{\mathbf{V}_{:,b}} \quad f(\mathbb{U}^t, \mathbb{W}^t, \cdots, \mathbf{V}_{:,b-1}^{t,l}, \mathbf{V}_{:,b}, \mathbf{V}_{:,b+1}^{t,l-1}, \cdots, \mathbf{V}_{:,B}^{t,l-1}) \tag{12a}$$

$$\text{s.t.} \mathbf{V}_{:,b} \in \mathcal{V}_b = \left\{ \mathbf{V}_{:,b} \middle| \text{Tr}(\mathbf{V}_{:,b}\mathbf{V}_{:,b}^H) \leq P/B \right\}, \forall b, \tag{12b}$$

where $f(\cdot)$ represents the objective function of problem in (5). According to the *Lagrange* multipliers method, the closed-form solution of the convex problem (12) can be derived by

$$\begin{aligned} \mathbf{V}_{:,b}^{t,l} = & \left(\mathbf{H}_{:,b}^H \mathbf{A}^t \mathbf{H}_{:,b} + \mu_b^{t,l}\mathbf{I}\right)^{-1} \Big(\mathbf{H}_{:,b}^H \hat{\mathbf{U}}^t \hat{\mathbf{W}}^t \\ & - \mathbf{A}^t \big(\sum_{\bar{b}=1}^{b-1} \mathbf{H}_{:,\bar{b}} \mathbf{V}_{:,\bar{b}}^{t,l} + \sum_{\bar{b}=b+1}^{B} \mathbf{H}_{:,\bar{b}} \mathbf{V}_{:,\bar{b}}^{t,l-1}\big)\Big). \end{aligned} \tag{13}$$

Here, $\hat{\mathbf{U}}^t$ =blkdiag($\mathbf{U}_1^t, \cdots, \mathbf{U}_K^t$) and $\hat{\mathbf{W}}^t$ =blkdiag($\alpha_1 \mathbf{W}_1^t, \cdots, \alpha_K \mathbf{W}_K^t$) with $\mathbf{A}^t = \hat{\mathbf{U}}^t \hat{\mathbf{W}}^t \hat{\mathbf{U}}^{tH}$, where blkdiag($\mathbf{X}_1, \mathbf{X}_2$) represents a block diagonal matrix with $\mathbf{X}_1, \mathbf{X}_2$ as the diagonal blocks. Meanwhile, the *Lagrange* multiplier $\mu_b^{t,l}$ is found by the classic bisection method to satisfy the local power constraint in (12b).

Motivated by BGS, we compute (13) sequentially from the first DU to B-th DU to obtain $\mathbf{V}_{:,b}$ locally, which corresponds to one iteration of BGS and iterates L times to form an inner loop for seeking $\{\mathbf{V}_k\}_{k \in \mathcal{K}}$.

However, as shown in (13), only with the updated results of all previous DUs (i.e. $\mathbf{V}_{:,\bar{b}}^{t,l}$ and $\mathbf{V}_{:,\bar{b}}^{t,l-1}$) and the CSI of other DUs (i.e. $\mathbf{H}_{:,\bar{b}}$) can each DU locally computes $\mathbf{V}_{:,b}^{t,l}$. Therefore, to accumulate the information from the previous computations, we define $\mathbf{\Phi}_b^{t,l}$ as follows

$$\mathbf{\Phi}_b^{t,l} = \sum_{\bar{b}=1}^{b-1} \mathbf{H}_{:,\bar{b}} \mathbf{V}_{:,\bar{b}}^{t,l} + \sum_{\bar{b}=b}^{B} \mathbf{H}_{:,\bar{b}} \mathbf{V}_{:,\bar{b}}^{t,l-1} \in \mathbb{C}^{KN \times Kd}. \tag{14}$$

Actually, it represents the interconnection data transmitted from $(b-1)$-th DU to b-th DU at the l-th inner iteration and t-th outer iteration, enabling b-th DU to compute its precoding matrix with the local CSI. Intuitively, because of the ring topology, it is clear to see $\mathbf{\Phi}_1^{t,l+1} = \mathbf{\Phi}_{B+1}^{t,l}, \forall l \in \{1, 2, ..., L-1\}$ and $\mathbf{\Phi}_1^{t,1} = \mathbf{\Phi}_{B+1}^{t-1,L}$.

Consequently, with the interconnection data $\mathbf{\Phi}_b^{t,l}$, the update for local precoding matrix $\mathbf{V}_{:,b}$ in (13) becomes

$$\begin{aligned} \mathbf{V}_{:,b}^{t,l} = & \left(\mathbf{H}_{:,b}^H \mathbf{A}^t \mathbf{H}_{:,b} + \mu_b^{t,l}\mathbf{I}\right)^{-1} \Big(\mathbf{H}_{:,b}^H \hat{\mathbf{U}}^t \hat{\mathbf{W}}^t \\ & - \mathbf{A}^t (\mathbf{\Phi}_b^{t,l} - \mathbf{H}_{:,b} \mathbf{V}_{:,b}^{t,l-1})\Big), \end{aligned} \tag{15}$$

Clearly, after receiving $\mathbf{\Phi}_b^{t,l}$ from $(b-1)$-th DU, b-th DU can compute its local precoding matrices $\mathbf{V}_{:,b}^{t,l}$ only with the local CSI $\mathbf{H}_{:,b}$. Therefore, after calculation of (15), b-th DU updates the interconnection data, i.e.

$$\mathbf{\Phi}_{b+1}^{t,l} = \mathbf{\Phi}_b^{t,l} - \mathbf{H}_{:,b} \mathbf{V}_{:,b}^{t,l-1} + \mathbf{H}_{:,b} \mathbf{V}_{:,b}^{t,l} \tag{16}$$

to support the upcoming local calculation of $(b+1)$-th DU.

B. Performance Analysis of the Inner Loop

Assuming that $\mathbb{V}^* = \{\mathbf{V}_k^*\}_{k \in \mathcal{K}}$ is obtained by (9) with $f^* = f(\mathbb{U}^t, \mathbb{W}^t, \mathbb{V}^*)$, it represents the optimal solution of the following problem

$$\min_{\{\mathbf{V}_k\}_{k \in \mathcal{K}} \in \mathcal{V}} f(\mathbb{U}^t, \mathbb{W}^t, \{\mathbf{V}_k\}_{k \in \mathcal{K}}) \tag{17}$$

with $\mathcal{V} = \{\{\mathbf{V}_k\}_{k \in \mathcal{K}} | \text{Tr}(\sum_{k=1}^{K} \mathbf{V}_k \mathbf{V}_k^H) \leq P\}$.

To better analyze performance, considering $\hat{f}$ as the optimal value of following problem

$$\begin{aligned} \hat{f} = \min_{\{\mathbf{V}_k\}_{k \in \mathcal{K}}} \quad & f(\mathbb{U}^t, \mathbb{W}^t, \{\mathbf{V}_k\}_{k \in \mathcal{K}}) \\ \text{s.t.} \quad & \{\mathbf{V}_k\}_{k \in \mathcal{K}} \in \bar{\mathcal{V}}(C, B) = \mathcal{V}_1 \times \mathcal{V}_2 \cdots \times \mathcal{V}_B, \end{aligned} \tag{18}$$

which has different constraint compared to problem in (17) with $\bar{\mathcal{V}}(C, B) \subseteq \mathcal{V}$. Then, the performance loss Δ, i.e. the gap between f_B^L obtained by the L inner iterations of LCD-WMMSE and f^*, can be described by two losses Δ_1 and Δ_2 shown below

$$\Delta = f_B^L - f^* = \underbrace{(f_B^L - \hat{f})}_{\Delta_1} + \underbrace{(\hat{f} - f^*)}_{\Delta_2}. \tag{19}$$

Lemma 1. *With increasing the number of inner iterations L, the loss Δ_1 monotonously decreases to 0, i.e $\lim_{L \to \infty} \Delta_1 = 0$.*

Proof: As for f_B^L, it is obtained by iterative sequential solving problem in (12), which is the subproblem of question in (18). Because of GS update mechanism and the convexity of problem (18), it follows

$$f_1^1 \geq ... \geq f_b^l \geq f_{b+1}^l \geq ... \geq f_b^{l+1} \geq ... \geq f_B^L \geq \hat{f}, \tag{20}$$

This article has been accepted for publication in IEEE Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2025.3526155

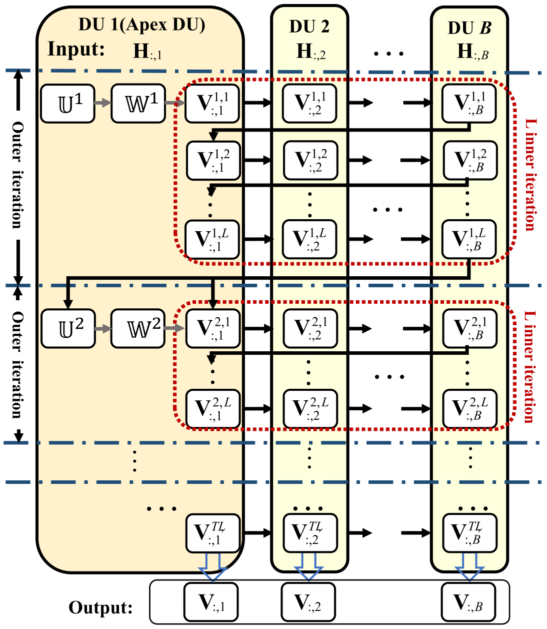

Fig. 2. The illustrations of LCD-WMMSE, where the update of interconnection data $\mathbf{\Phi}^{t,l}_{b+1}$ following with $\mathbf{V}^{t,l}_{:,b}$ is omitted for simplicity.

ensuring the monotonic decreasing of Δ_1 with increasing L. Meanwhile, guaranteed by the convergence of BGS [11], f^L_B will return to $\hat{f}$ with an infinite L for any given B, so as to

$$\lim_{L\to\infty} \Delta_1 = \lim_{L\to\infty} f^L_B - \hat{f} = 0. \tag{21}$$

■

On the other hand, we analyze the impact of B upon Δ_1 and Δ_2. When L and M are fixed, a smaller B leads to a higher level of centralization, leading to a faster convergence and a smaller Δ_1. Moreover, if $B_1 > B_2$, we have $\bar{\mathcal{V}}(M/B_1, B_1) \subset \bar{\mathcal{V}}(M/B_2, B_2) \subset \mathcal{V}$, so that $\Delta_2(B_1) \geq \Delta_2(B_2) \geq 0$ due to the convexity of objective function $f(\cdot)$. In summary, the larger L and the smaller B account for a smaller performance loss Δ but with higher level of centralization and more complexity cost. Hence, suitable choices of L and B are highly demanded to balance the complexity and performance in a better way.

C. Distributed Computation of $\mathbf{U}_k$ and $\mathbf{W}_k$ at Outer Loop

Before executing the inner loop to update $\mathbf{V}_k$ locally by (15), the apex DU should be specified to undertake the calculation of $\mathbf{U}_k$ in (7) and $\mathbf{W}_k$ in (8) at outer loop.

However, updating $\mathbf{U}_k$ and $\mathbf{W}_k$ requires the global CSI $\mathbf{H}_k$ and the precoding matrix $\mathbf{V}_k$, which are unavailable for the apex DU. Fortunately, thanks to the introduced interconnection data $\mathbf{\Phi}^{t,1}_1 = \mathbf{\Phi}^{t-1,L}_{B+1}$ obtained by apex DU, the updates of $\mathbf{U}^t_k$ and $\mathbf{W}^t_k$ can be carried out by

$$\mathbf{U}^t_k = \left(\boldsymbol{\xi}^t_{k:}{\boldsymbol{\xi}^t_{k:}}^H + \sigma^2_k\mathbf{I}\right)^{-1}\boldsymbol{\xi}^t_{kk}, \tag{22}$$

$$\mathbf{W}^t_k = \left(\mathbf{I} - {\mathbf{U}^t_k}^H\boldsymbol{\xi}^t_{kk}\right)^{-1}, \tag{23}$$

where $\boldsymbol{\xi}^t_{k,m} = \mathbf{\Phi}^{t,1}_1[(k-1)N+1 : kN, (m-1)d+1 : md] \in \mathbb{C}^{N\times d}$ and $\boldsymbol{\xi}^t_{k,:} = \mathbf{\Phi}^{t,1}_1[(k-1)N+1 : kN, :] \in \mathbb{C}^{N\times Kd}$ are different submatrixes of $\mathbf{\Phi}^{t,1}_1$. By doing this, the updating $\mathbf{U}^t_k$ and $\mathbf{W}^t_k$ only occurs at the transmitter without global CSI, which avoids the information exchange between transmitters and users in other distributed WMMSE schemes [9], [10]. Then, $\mathbf{U}^t_k$ and $\mathbf{W}^t_k$ will be transmitted together with $\mathbf{\Phi}^{t,l}_b$ along the ring link, serving for the calculations at each DU.

Algorithm 2 LCD-WMMSE Algorithm

```
Input: $\mathbf{H}_{k,b}$, $\varepsilon$, $\{\sigma_k\}_{k\in\mathcal{K}}$, $P$, $\forall b, k$, $L$.
Output: $\mathbf{U}_k$, $\mathbf{W}_k$, $\mathbf{V}_{k,b}$, $\forall b, k$.
Initialize: Set $\mathbf{V}^{0,0}_{k,b}$ to satisfy $\mathrm{Tr}(\mathbf{V}^{0,0}_{k,b}\mathbf{V}^{0,0H}_{k,b}) = P/(BK)$. Initialize $\mathbf{\Phi}^{1,1}_1 = \sum_{b=1}^{B}\mathbf{H}_{:,b}\mathbf{V}^{0,0}_{:,b}$ and $t = 0$.
repeat(Outer loop)
    $t = t + 1$
    Apex DU update $\mathbf{U}^t_k$ via (22) and $\mathbf{W}^t_k$ via (23), $\forall k$.
    for $l = 1$ to $L$ do (Inner loop)
        for $b = 1$ to $B$ do
            $b$-th DU updates $\mathbf{V}^{t,l}_{k,b}$ via (15),$\forall k$.
            Update $\mathbf{\Phi}^{t,l}_{b+1}$ with (16).
        end for
    end for
until $|\sum_{k=1}^{K}\log\det(\mathbf{W}^t_k) - \sum_{k=1}^{K}\log\det(\mathbf{W}^{t-1}_k)| < \varepsilon$
```

As it can be seen in Fig.2, the operations that follow the double-layer structure in LCD-WMMSE are shown. During a single outer iteration, apex DU first updates $\mathbb{U}^t$ and $\mathbb{W}^t$ with (22) and (23), and then these results are used to execute the inner loop. In each inner iteration, each DU updates $\mathbf{V}_{:,b}$ and the interconnection data in order. Note that given the initial precoding matrix $\mathbf{V}^{0,0}_{k,b}$, the interconnection data $\mathbf{\Phi}^{1,1}_1$ needs to be initialized through the sequential accumulation along DUs, i.e. $\mathbf{\Phi}^{1,1}_1 = \sum_{b=1}^{B}\mathbf{H}_{:,b}\mathbf{V}^{0,0}_{:,b}$.

IV. Complexity and Scalability Analysis

Assuming T as the iteration numbers of outer loop respectively, the total complexity of updating $\mathbb{U}^t$ and $\mathbb{W}^t$ in LCD-WMMSE is $\mathcal{O}(TKN^3 + TK^2N^2)$. For each DU during single inner iteration, the complexity of calculating (15) is $\mathcal{O}(C^3 + K^2CN^2 + KC^2N)$, and the update of interconnection data $\mathbf{\Phi}^{t,l}_{b+1}$ costs $2K^2CNd$ complex multiplications. Therefore, the total complexity of LCD-WMMSE is $\mathcal{O}(TLMC^2 + TLK^2MN^2 + TLKMCN + TKN^3 + TK^2N^2)$.

As for the traditional WMMSE, its complexity is $\mathcal{O}(TM^3 + TKN^3 + TKM^2 + TKMN^2 + TK^2MN)$ with the same T iterations. Since $M \gg K, N, C, L$, the computational complexity of two algorithms are dominated by the order of M, and LCD-WMMSE reduces the complexity from the cubic of M to the linear of M. Table I shows the number of complex multiplications required for different algorithms with $T = 20$ at different system models, where the significant complexity reduction can be confirmed. Although the computational complexity of the LCD-WMMSE algorithm is still higher than that of the up-to-date R-WMMSE proposed in [12], LCD-WMMSE can implemented on a decentralized architecture, while R-WMMSE can not.

On the other hand, since the data transmitted among DUs are $\mathbb{U}^t$, $\mathbb{W}^t$ and $\mathbf{\Phi}^{t,l}_b$, the amount of interconnection data is $TLBK^2Nd + TB(KNd + Kd^2)$. Clearly, the amount of data transmitted among DUs is independent of the number of transmitted antennas M, leading to the desired scalability.

TABLE I
THE COMPLEXITY OF UPDATING PRECODING MATRIX IN DIFFERENT $M \times N$ WITH $L = 3, K = 16, T = 20$ AND $d = 2$

Algorithm		64×2	128×2	256×4
WMMSE [7]		$5.51 \cdot 10^7$	$4.09 \cdot 10^8$	$3.22 \cdot 10^9$
	$C = 8$	$2.06 \cdot 10^7$	$4.11 \cdot 10^7$	$1.54 \cdot 10^8$
LCD-WMMSE	$C = 4$	$1.80 \cdot 10^7$	$3.59 \cdot 10^7$	$1.42 \cdot 10^8$
	$C = 1$	$1.69 \cdot 10^7$	$3.38 \cdot 10^7$	$1.38 \cdot 10^8$
R-WMMSE [12]		$3.78 \cdot 10^6$	$3.92 \cdot 10^6$	$1.26 \cdot 10^7$

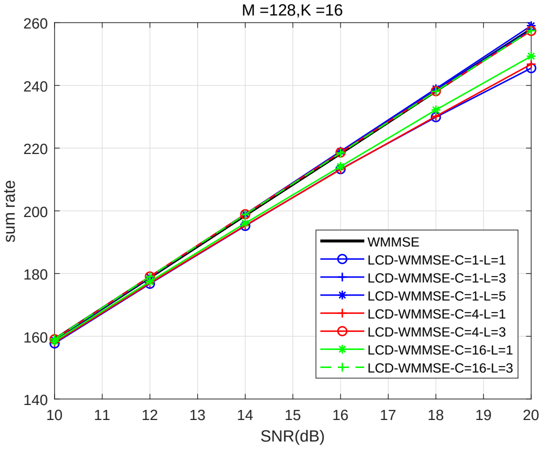

Fig. 3. WSR comparison under different C and L in a massive MU-MIMO system with $M = 128, K = 16, N = 2, d = 2$ and $\alpha_k = 1, \forall k$.

V. SIMULATION

Fig. 3 presents a comparison of the sum rates achieved by different DU antennas C and the iteration number of inner loop L. To compare the efficiency in a fair way, the iteration number of WMMSE and the outer iteration number of LCD-WMMSE T are both limited to 20. As expected, LCD-WMMSE achieves negligible performance loss compared to the traditional WMMSE. In addition, the performance improves as L increases with fixed C, which is consistent with the conclusion drawn before. Meanwhile, with the same L and M, increasing antennas number C at each DU also improves performance. When $L = 3$ and $C = 1$, the performance of two schemes can be close, but the complexity of LCD-WMMSE maybe less than $1/10$ of WMMSE, as shown in Table I.

The convergence behaviour of LCD-WMMSE is shown by Fig. 4. Here, all algorithms start with a uniform initial value, and type1 and type2 are instances where the antennas number of different DUs is different. Intuitively, the performance of LCD-WMMSE improves as the number of outer iterations increases, and both the increasing of L and concentration level displayed by C are beneficial to the convergence. Nevertheless, the number of inner iterations L has a greater impact on the convergence than C. Overall, the simulation results in Fig. 3 and Fig. 4 show that the proposed LCD-WMMSE algorithm can achieve near-WMMSE performance while maintaining much lower computational complexity.

VI. CONCLUSION

In this paper, a distributed low-complexity WMMSE algorithm running in a decentralized architecture based on ring topology is proposed. Analysis shows that LCD-WMMSE has extremely low complexity, only consuming computing resources at the transmitting end, and has scalability due to the independence between the amount of interconnection data and the number of transmitter antennas. The simulation results show that LCD-WMMSE algorithm has negligible performance loss with lower computational complexity.

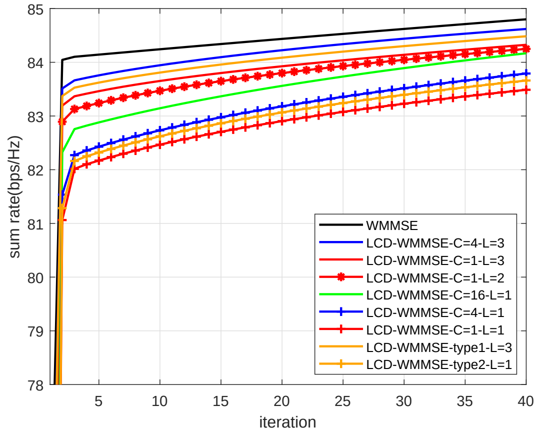

Fig. 4. Iterative performance comparison under different C and L in a massive MU-MIMO system with $M = 128, K = 10, N = 2, d = 1, \alpha_k = 1, \forall k$ and $\mathrm{SNR} = 15\mathrm{dB}$ (type1: 10 4-antenna DUs, 20 2-antenna DUs and 48 single-antenna DUs;type2: 50 2-antenna DUs and 28 single-antenna DUs).

REFERENCES

[1] L. Wengang, X. Yaqin, Z. Chenmeng, T. Yiheng, L. Mohan, and H. Jun, "Multi-Frequency-Ranging positioning algorithm for 5G OFDM communication systems," *Chinese Journal of Electronics*, vol. 32, no. 4, pp. 773–784, 2023.

[2] K. Li, O. Castaeda, C. Jeon, J. R. Cavallaro, and C. Studer, "Decentralized Coordinate-Descent data detection and precoding for massive MU-MIMO," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2019, pp. 1–5.

[3] J. Rodrguez Snchez, F. Rusek, O. Edfors, M. Sarajli, and L. Liu, "Decentralized massive MIMO processing exploring daisy-chain architecture and recursive algorithms," *IEEE Transactions on Signal Processing*, vol. 68, pp. 687–700, 2020.

[4] Z. Wang, J. Zhang, H. Q. Ngo, B. Ai, and M. Debbah, "Iteratively weighted MMSE uplink precoding for cell-free massive MIMO," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 231–236.

[5] W.-J. Zhu, C. Sun, and X. Gao, "Distributed precoding for network massive MIMO systems without data sharing," *IEEE Systems Journal*, vol. 17, no. 4, pp. 6057–6068, 2023.

[6] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 4792–4799, 2008.

[7] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331–4340, 2011.

[8] X. Zhao, M. Li, Y. Liu, T.-H. Chang, and Q. Shi, "Communication-efficient decentralized linear precoding for massive MU-MIMO systems," *IEEE Transactions on Signal Processing*, vol. 71, pp. 4045–4059, 2023.

[9] Y. Zhong, Y. Cao, and T. Lv, "Low-complexity distributed precoding in user-centric cell-free mmWave MIMO systems," in *2022 Wireless Telecommunications Symposium (WTS)*, 2022, pp. 1–5.

[10] I. Atzeni, B. Gouda, and A. Tlli, "Distributed precoding design via over-the-air signaling for cell-free massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1201–1216, 2021.

[11] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss–Seidel method under convex constraints," *Operations research letters*, vol. 26, no. 3, pp. 127–136, 2000.

[12] X. Zhao, S. Lu, Q. Shi, and Z.-Q. Luo, "Rethinking WMMSE: Can its complexity scale linearly with the number of BS antennas?" *IEEE Transactions on Signal Processing*, vol. 71, pp. 433–446, 2023.