MM-Net: Recurrent Unfolding with Adaptive Majorization for Weighted Sum-Rate Beamforming

Zhexian Yang, *Student Member, IEEE*, Zepeng Zhang, *Student Member, IEEE*, and Ziping Zhao, *Member, IEEE*

***Abstract*—Weighted sum-rate (WSR) maximization for beamforming is a fundamental problem in communication systems. Classical approaches, such as the weighted sum-minimum mean-square error (WMMSE) algorithm, often suffer from high computational complexity and slow convergence. In this letter, we propose MM-Net for WSR beamforming, an unfolding-based framework derived from a recently developed minorization–maximization (MM) algorithm [2], [3] that avoids explicit matrix inversion in WMMSE. MM-Net learns an adaptive majorization parameter controlling surrogate tightness, thereby accelerating convergence while preserving the convergence guarantees of the underlying MM algorithm. By sharing parameters across unfolded iterations, MM-Net forms a recurrent architecture, where a compact multilayer perceptron predicts the adaptive majorization parameter. The inference complexity of the learned component scales linearly with the number of transmit antennas and is independent of the number of users, data streams, and receive antennas. Numerical results demonstrate that MM-Net achieves faster convergence than conventional optimization methods and state-of-the-art learning-based schemes.**

***Index Terms*—Weighted sum-rate maximization, minorization–maximization, adaptive majorization, algorithm unfolding, RNN.**

I. Introduction

Weighted sum-rate (WSR) maximization is a fundamental task in communication system design, particularly for beamformer design in interference-limited systems [4]. The problem is generally NP-hard [5], and practical optimization algorithms therefore aim to compute stationary solutions of the resulting nonconvex formulation. A widely adopted approach is the weighted sum-minimum mean-square error (WMMSE) algorithm [6], [7], which leverages the equivalence between WSR maximization and mean-square error minimization. Under the commonly imposed total power constraint, each WMMSE iteration requires solving a quadratic constrained quadratic program (QCQP), whose closed-form solution involves a scalar dual variable obtained via line search. As the number of transmit antennas increases, this step becomes computationally expensive, creating a major bottleneck in massive multiple-input multiple-output (MIMO) systems [8]. To alleviate this issue, [2], [3], [9] reinterpret WMMSE within the minorization–maximization (MM) framework [10], [11] and develop a matrix-inversion-free MM algorithm. Moreover, the MM algorithm naturally accommodates more general beamforming constraints beyond the total power constraint, such as per-antenna power constraints [12].

Part of the paper was preliminarily presented at the 33rd European Signal Processing Conference (EUSIPCO), Palermo, Italy, September 8-12, 2025 [1].

Beyond optimization-based approaches, deep learning has recently been explored to improve the computational efficiency of WSR maximization. In [13], a multilayer perceptron (MLP) is trained end-to-end to approximate the input–output mapping of the WMMSE algorithm. Numerical results demonstrate that the MLP can effectively mimic WMMSE in single-input single-output (SISO) interference channels, but the approach is limited to power control. Extending this framework to MIMO systems requires substantially larger networks due to increased input–output dimensions, leading to significant training challenges. To address this issue, [14] considers WSR maximization in MISO broadcasting channels and proposes a convolutional neural network (CNN) for beamforming design, where the beamforming structure [15] is exploited to reduce output dimensionality. Notably, the CNN degenerates to the MLP architecture in [13]. However, both approaches involve a large number of parameters, particularly in massive MIMO scenarios. An alternative direction is deep unfolding [16]. In [17], an iterative algorithm induced deep-unfolding neural network (IAIDNN) is constructed by unfolding the WMMSE algorithm for MIMO broadcasting channels, reducing inference complexity by replacing matrix inversions with nonlinear approximations. Similarly, [18] proposes a matrix-inversion-free WMMSE deep unfolding network (WMMSE-Net) for MISO broadcasting channels, where the QCQP subproblems in each iteration are approximated using a fixed number of projected gradient (PG) steps. This idea is further extended to MIMO broadcasting channels in [19] by replacing matrix inversions with a single Schultz iteration step. In [20], the PG method is unfolded into a neural network architecture, referred to as BLN-PGP, for MISO interference channels. Other deep learning approaches for WSR beamforming design largely follow similar network design principles, as in [21]–[23].

In this paper, building on the recent developments in [2], [3], we propose a deep unfolding network derived from the MM algorithm, termed MM-Net, for WSR beamforming in MIMO broadcasting channels. Unlike existing unfolding networks [13], [14], [17]–[23], which typically adopt feedforward architectures with layer-wise trainable parameters and therefore result in structured MLPs whose depth equals the number of unfolded iterations, MM-Net resembles a recurrent neural network (RNN) with shared parameters across layers, substantially reducing the number of trainable parameters. Moreover, the learnable parameters of MM-Net scale linearly with the number of transmit antennas and are independent of the number of users, data streams, and receive antennas, which enhances its practicality for large-scale systems. MM-Net also accommodates beamforming constraints beyond the

total power constraint. Compared with the conventional MM algorithm, MM-Net can be interpreted as a data-driven realization of adaptive majorization within the MM framework. While classical MM methods choose the majorization parameter based on conservative spectral bounds that ensure monotonicity but often lead to slow convergence, MM-Net learns this parameter directly from data and implicitly captures curvature information that is difficult to exploit analytically. As a result, MM-Net balances surrogate tightness and descent aggressiveness in an instance-dependent manner while remaining embedded in a provably convergent MM structure. Furthermore, thanks to parameter sharing across unfolded iterations, MM-Net exhibits strong cross-configuration generalization: a single trained model can be applied to different system settings without retraining. This capability can be further enhanced via a mixed-training paradigm, where the network is trained simultaneously on samples from multiple system configurations, enabling robust performance across a wider range of user numbers, data streams, and antenna setups. We establish its convergence properties and analyze its computational complexity, and extensive numerical experiments demonstrate that MM-Net outperforms conventional optimization algorithms such as WMMSE and MM, as well as state-of-the-art learning-based methods.

II. Problem Formulation and MM Algorithm

Consider a MIMO broadcast channel where a base station equipped with M antennas serves K users. User k is equipped with N_k antennas, $k = 1, \ldots, K$. Let $\mathbf{W}_k \in \mathbb{C}^{M \times D_k}$ denote the beamforming matrix for user k, and $\mathbf{s}_k \in \mathbb{C}^{D_k}$ the corresponding transmit symbol vector, where D_k represents the number of data streams intended for user k. The channel matrix from the base station to user k is denoted by $\mathbf{H}_k \in \mathbb{C}^{N_k \times M}$, and σ_k^2 denotes the noise variance at user k. Let $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_K] \in \mathbb{C}^{M \times \sum_{k=1}^{K} D_k}$. The WSR beamforming problem is formulated as

$$\max_{\mathbf{W} \in \mathcal{C}} \quad \sum_{k=1}^{K} \omega_k \log\det\left(\mathbf{I} + \mathbf{W}_k^{\mathsf{H}} \mathbf{H}_k^{\mathsf{H}} \mathbf{F}_k^{-1} \mathbf{H}_k \mathbf{W}_k\right), \quad (1)$$

where ω_k is a predefined weight and $\mathbf{F}_k = \sum_{j \neq k}^{K} \mathbf{H}_k \mathbf{W}_j \mathbf{W}_j^{\mathsf{H}} \mathbf{H}_k^{\mathsf{H}} + \sigma_k^2 \mathbf{I}_{N_k}$ for $k = 1, \ldots, K$. We assume that $\mathcal{C}$ is a convex and bounded beamforming set. For example, under the total power constraint, $\mathcal{C} = \left\{\mathbf{W} \,\middle|\, \|\mathbf{W}\|_F^2 \leq P\right\}$, where P denotes the total transmit power budget. Under the per-antenna power constraint, $\mathcal{C} = \left\{\mathbf{W} \,\middle|\, \|\mathbf{W}(n,:)\|_2^2 \leq P_n,\ n = 1, \ldots, M\right\}$, where $\mathbf{W}(n,:)$ denotes the n-th row of $\mathbf{W}$ and P_n is the power budget of the n-th transmit antenna.

We briefly review the MM algorithm for WSR maximization developed in [2], [3]. Let $\mathrm{WSR}(\mathbf{W})$ denote the objective function in (1). The central step of the MM algorithm is to construct a tight lower bound of the objective such that each resulting subproblem admits a computationally efficient solution. Based on [3, Proposition 11, 16], a surrogate function of $\mathrm{WSR}(\mathbf{W})$ at the iterate $\mathbf{W}^{(t)}$ is given by

$$\begin{aligned} \underline{\mathrm{WSR}}\left(\mathbf{W}; \mathbf{W}^{(t)}\right) &= -\frac{1}{\alpha^{(t)}} \left\|\mathbf{W} - \mathbf{W}^{(t)}\right\|_F^2 \\ &+ 2\Re\left[\mathrm{tr}\left(\left(\mathbf{G}^{(t)}\right)^{\mathsf{H}} \left(\mathbf{W} - \mathbf{W}^{(t)}\right)\right)\right] + \mathrm{WSR}\left(\mathbf{W}^{(t)}\right) \end{aligned} \quad (2)$$

where $\alpha^{(t)} > 0$ is chosen such that $\frac{1}{\alpha^{(t)}} \mathbf{I} \succeq \mathbf{A}^{(t)}$, with

$$\begin{aligned} \mathbf{A}^{(t)} = \sum_{j=1}^{K} \omega_j \mathbf{H}_j^{\mathsf{H}} \Bigg(&\left(\mathbf{F}_j^{(t)}\right)^{-1} - \\ &\left(\mathbf{F}_j^{(t)} + \mathbf{H}_j \mathbf{W}_j^{(t)} \left(\mathbf{W}_j^{(t)}\right)^{\mathsf{H}} \mathbf{H}_j^{\mathsf{H}}\right)^{-1} \Bigg) \mathbf{H}_j, \end{aligned} \quad (3)$$

$$\mathbf{G}^{(t)} = \left[\mathbf{G}_1^{(t)}, \ldots, \mathbf{G}_K^{(t)}\right], \quad (4)$$

where $\mathbf{G}_k^{(t)} = \left(\omega_k \mathbf{H}_k^{\mathsf{H}} \left(\mathbf{F}_k^{(t)}\right)^{-1} \mathbf{H}_k - \mathbf{A}^{(t)}\right) \mathbf{W}_k^{(t)}$, $k = 1, \ldots, K$. The tightest choice of $1/\alpha^{(t)}$ is given by the largest eigenvalue of $\mathbf{A}^{(t)}$, namely, $\frac{1}{\alpha^{(t)}} = \lambda_{\max}\left(\mathbf{A}^{(t)}\right)$. In each iteration of the MM algorithm, the beamforming matrix $\mathbf{W}^{(t+1)}$ is obtained by minimizing the surrogate function $\underline{\mathrm{WSR}}\left(\mathbf{W}; \mathbf{W}^{(t)}\right)$. Specifically, the update rule is given by

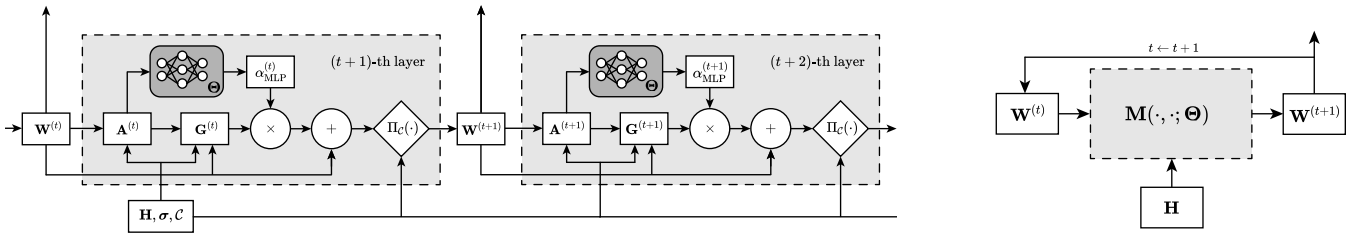
$$\begin{aligned} \mathbf{W}^{(t+1)} &= \arg\min_{\mathbf{W} \in \mathcal{C}} \quad \left\|\mathbf{W} - \left(\mathbf{W}^{(t)} + \alpha^{(t)} \mathbf{G}^{(t)}\right)\right\|_F^2, \\ &= \Pi_{\mathcal{C}}\left(\mathbf{W}^{(t)} + \alpha^{(t)} \cdot \mathbf{G}^{(t)}\right), \end{aligned} \quad (5)$$

where $\Pi_{\mathcal{C}}(\cdot)$ denotes the Euclidean projection onto the feasible set $\mathcal{C}$. Define $\mathbf{V}^{(t)} = \mathbf{W}^{(t)} + \alpha^{(t)} \cdot \mathbf{G}^{(t)}$. Under the total power constraint, the projection admits the closed-form solution $\mathbf{W}^{(t+1)} = \min\left\{\frac{\sqrt{P}}{\left\|\mathbf{V}^{(t)}\right\|_F}, 1\right\} \mathbf{V}^{(t)}$. Under the per-antenna power constraint, the projection is performed row-wise as $\mathbf{W}^{(t+1)}(n,:) = \min\left\{\frac{\sqrt{P_n}}{\left\|\mathbf{V}^{(t)}(n,:)\right\|_2}, 1\right\} \mathbf{V}^{(t)}(n,:)$, $n = 1, \ldots, M$. In fact, this projection can be extended to any convex beamformer constraint set $\mathcal{C}$ that admits an efficient projection operator. A key challenge in the MM algorithm lies in the computation of the constant $\alpha^{(t)}$ in (5) at each iteration, which serves as a stepsize analogous to that in the PG method [2], [3]. Its computation can be particularly demanding in large-scale antenna systems. To mitigate this issue, [2], [3] propose a lightweight approximation by setting $\alpha^{(t)} = \|\mathbf{A}^{(t)}\|_F^{-1}$. Although this choice significantly reduces the computational burden, it generally yields a looser surrogate function for the subproblem, which may consequently result in slow numerical convergence.

III. MM-Net: Deep Unfolding of MM

In this section, we adopt the algorithm unfolding paradigm to develop a deep unfolding network based on the MM algorithm in Section II, termed MM-Net. The proposed MM-Net inherits the interpretability of the underlying MM algorithm while leveraging the flexibility of data-driven learning. This design aims to improve the computational efficiency of the original MM procedure, and a properly trained network can significantly accelerate convergence. The overall architecture of MM-Net is illustrated in Fig. 1(a).

This article has been accepted for publication in IEEE Signal Processing 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/LSP.2026.3710384



(a) The MM-Net architecture. (b) An RNN interpretation.

Fig. 1: The MM-Net architecture for WSR beamforming.

MM-Net employs an MLP with trainable parameters denoted by $\boldsymbol{\Theta}$. Taking $\mathbf{A}^{(t)}$ as the input, the MLP outputs a learned stepsize, denoted by $\alpha_{\mathrm{MLP}}^{(t)}$. The update rule of MM-Net is then given by $\mathbf{W}^{(t+1)} = \Pi_{\mathcal{C}}\left(\mathbf{W}^{(t)} + \alpha_{\mathrm{MLP}}^{(t)} \cdot \mathbf{G}^{(t)}\right)$. For the MLP module, using the entire matrix $\mathbf{A}^{(t)}$ as input would incur an inference complexity of $\mathcal{O}(M^2)$. To reduce the computational burden, we instead adopt $\mathrm{diag}(\mathbf{A}^{(t)}) \in \mathbb{R}^M$ as the input feature. Since $\mathrm{tr}(\mathbf{A}^{(t)}) \geq \lambda_{\max}(\mathbf{A}^{(t)})$, the diagonal entries provide an upper bound on the maximum eigenvalue, which governs the allowable step size, while lowering complexity to $\mathcal{O}(M)$. The hidden layers of the MLP use ReLU activations, while the output layer adopts the softplus function, $\log(1+e^x)$, to ensure nonnegativity of the output and to mitigate vanishing gradient issues during training [24].

MM-Net distinguishes itself from existing feedforward architectures such as IAIDNN [17] and WMMSE-Net [19] by employing parameter sharing across layers, resulting in an RNN structure. Denote the iterative mapping inside the dashed gray box in Fig. 1(a) by $\mathbf{W}^{(t+1)} = \mathbf{M}\left(\mathbf{H}, \mathbf{W}^{(t)}; \boldsymbol{\Theta}\right)$. This parameter-sharing mechanism leads to a compact representation where the number of trainable parameters is independent of the number of unfolded layers, as illustrated in Fig. 1(b), significantly reducing the overall parameter count. Consequently, unlike feedforward unfolding networks where inference depth is tied to the training depth, MM-Net can be trained with a fixed number of layers yet deployed with an arbitrary number of iterations during inference. In contrast, the parameter sizes of MLP [13], CNN [14], [21], IAIDNN [17], WMMSE-Net [18], [19], and BLN-PGP [20] grow linearly with the number of unfolding layers, and their inference depth is strictly limited to the number of layers used during training.

The parameter counts of MLP [13], CNN [14], [21], IAIDNN [17], WMMSE-Net [18], [19], and BLN-PGP [20] grow linearly with the number of users. In MLP and IAIDNN, they also scale with the numbers of transmit/receive antennas, and IAIDNN scales quadratically with the number of data streams. Consequently, these networks typically require retraining when the beamforming scenario changes. MM-Net, by contrast, is highly parameter-efficient: its only trainable component is an MLP whose size depends solely on the number of transmit antennas, independent of user count or receive antennas. This allows the same trained network to be deployed across varying system configurations without retraining, making it particularly practical for real-world deployment.

Given an initial state $\mathbf{W}^{(0)}$, the following theorem establishes the convergence property of MM-Net.

Theorem 1. *Let $\{\mathbf{W}^{(t)}\}_{t\geq 0}$ be the beamformer sequence generated by MM-Net. Assume that there exists a constant $C > 0$ such that $\frac{2}{\alpha_{\mathrm{MLP}}^{(t)}} - \lambda_{\max}(\mathbf{A}^{(t)}) \geq C, \quad \forall t \geq 0$. Then all limit points of $\{\mathbf{W}^{(t)}\}_{t\geq 0}$ are stationary points of problem* (1). *Moreover,*

$$\min_{t=0,\ldots,T-1} \|\mathbf{W}^{(t+1)} - \mathbf{W}^{(t)}\|_F \leq \sqrt{\frac{\mathrm{WSR}^\star - \mathrm{WSR}(\mathbf{W}^{(0)})}{CT}},$$

where $\mathrm{WSR}^\star$ *denotes the optimal value of problem* (1).

Proof. The proof is given in Section **??** of the Appendix. □

This result establishes a sublinear convergence rate of order $\mathcal{O}(1/\sqrt{T})$ to a stationary point of problem (1).

IV. Complexity Comparisons

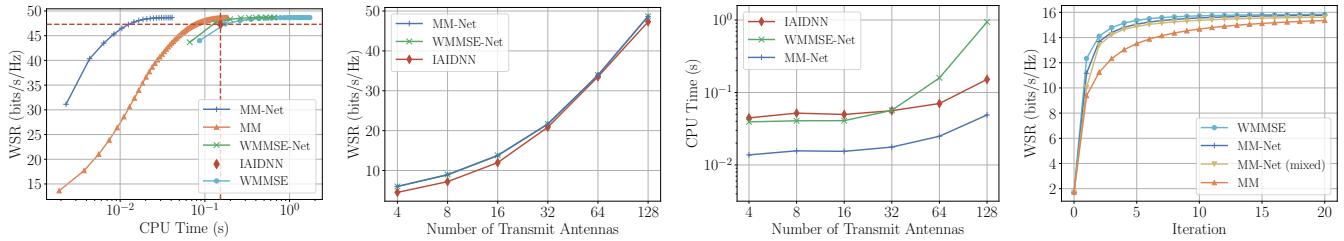
We next present a formal computational-complexity comparison of MM-Net with its optimization-based counterpart, MM, the WMMSE algorithm, and two deep-unfolding MIMO beamforming architectures: IAIDNN [17] and WMMSE-Net [19]. For clarity, we assume $N_k = N$ and $D_k = D$ for $k = 1, \ldots, K$, with $M \geq N = D$. The per-iteration computational complexity of the MM algorithm, with $\alpha^{(t)}$ computed as $\left\|\mathbf{A}^{(t)}\right\|_F$, is $\mathcal{O}\left(K^2M^2N\right)$. For MM-Net, assuming an MLP with L hidden layers of width I, the per-layer complexity is $\mathcal{O}\left(K^2M^2N + MI + (L-1)I^2\right)$. The per-iteration complexity of the WMMSE algorithm is $\mathcal{O}\left(K^2M^2N + KM^3\right)$. For IAIDNN, the per-layer complexity is $\mathcal{O}(K^2(MN^2 + M^2N) + K(M^3+N^3)) = \mathcal{O}(K^2M^2N + KM^3)$, where the cubic term can be reduced using fast matrix multiplication. For WMMSE-Net, the per-layer complexity is $\mathcal{O}(J_uK^2(MND + N^2D + M^2N + MN^2) + K\left(D^3 + MD^2\right) + J_wKD^3 + J_v(K^2MND + KMD^2)) = \mathcal{O}(J_uK^2M^2N + J_wKD^3 + J_vK^2MN^2)$, where J_u, J_w, and J_v denote the numbers of update steps for variables $\mathbf{U}_k$, $\mathbf{W}_k$, and $\mathbf{V}_k$, respectively (see [19] for details). Considering the massive MIMO regime [8], where $M \gg KN$, the computational complexities of MM-Net and MM are comparable. WMMSE-Net exhibits higher complexity than MM-Net and MM, while IAIDNN is more complex than WMMSE-Net but remains less than WMMSE. Hence, the per-iteration (or per-layer) complexities satisfy

$$\text{MM-Net} \approx \text{MM} < \text{WMMSE-Net} < \text{IAIDNN} < \text{WMMSE}.$$

Importantly, while retaining the same order complexity as MM, MM-Net achieves faster convergence in practice through learned adaptive majorization, resulting in improved overall computational efficiency without increasing asymptotic cost.

This article has been accepted for publication in IEEE Signal Processing 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/LSP.2026.3710384

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Theta}} = \mathbb{E}_{\mathbf{H},\mathbf{W}^{(0)}} \left\{ -\frac{1}{T} \sum_{t=0}^{T-1} \frac{\partial \mathrm{WSR}\left(\mathbf{W}^{(t+1)}\right)}{\partial \boldsymbol{\Theta}} \right\} = \mathbb{E}_{\mathbf{H},\mathbf{W}^{(0)}} \left\{ -\frac{1}{T} \sum_{t=0}^{T-1} 2\Re \left[\frac{\partial \mathrm{WSR}\left(\mathbf{W}^{(t+1)}\right)}{\partial \operatorname{vec}\left(\mathbf{W}^{(t+1)}\right)} \cdot \frac{\partial \operatorname{vec}\left(\mathbf{W}^{(t+1)}\right)}{\partial \boldsymbol{\Theta}} \right] \right\}$$
$$= \mathbb{E}_{\mathbf{H},\mathbf{W}^{(0)}} \left\{ -\frac{1}{T} \sum_{t=0}^{T-1} 2\Re \left[\operatorname{vec}\left(\mathbf{G}^{(t+1)}\right)^{\mathsf{H}} \left(\frac{\partial \operatorname{vec}\left(\mathbf{M}\right)}{\partial \boldsymbol{\Theta}} + \cancel{\frac{\partial \operatorname{vec}\left(\mathbf{M}\right)}{\partial \operatorname{vec}\left(\mathbf{W}^{(t)}\right)} \cdot \frac{\partial \operatorname{vec}\left(\mathbf{W}^{(t)}\right)}{\partial \boldsymbol{\Theta}}} \right) \right] \right\} \tag{6}$$



(a) $M = 128$ (b) WSR versus M (c) CPU time versus M (d) $M = 16$

Fig. 2: Comparison of the performance of optimization-based and learning-based methods.

V. Numerical Simulations

In this section, we conduct numerical experiments on a desktop equipped with two Intel Xeon Gold 6226R CPUs. We consider a MIMO broadcast channel under a total power constraint. The channel coefficients are generated independently from $\mathcal{CN}(0,1)$. The noise variances are identical for all users, i.e., $\sigma_k^2 = \sigma^2$. The signal-to-noise ratio is set to $\mathrm{SNR} = 10\log_{10}(P/\sigma^2) = 0\,\mathrm{dB}$, with $\sigma^2 = 1$. The methods under comparison include WMMSE [7], MM [3], IAIDNN [17], WMMSE-Net [19], and the proposed MM-Net. The training and testing of IAIDNN, WMMSE-Net, and MM-Net are all implemented on CPUs. All results are averaged over 1000 Monte Carlo simulations. MM-Net is implemented in PyTorch 2.5, while IAIDNN and WMMSE-Net follow the implementations described in their respective papers. For MM-Net, the MLP consists of two hidden layers, each with width M. The training loss is defined as $\mathcal{L}\left(\boldsymbol{\Theta}\right) = \mathbb{E}_{\mathbf{H},\mathbf{W}^{(0)}}\left[-\frac{1}{T}\sum_{t=0}^{T-1}\mathrm{WSR}\left(\mathbf{M}\left(\mathbf{H},\mathbf{W}^{(t)};\boldsymbol{\Theta}\right)\right)\right]$. This formulation enables MM-Net to operate in a fully unsupervised manner. Using backpropagation through time [25], we derive the gradient expression in (6). To reduce computational cost, we adopt the truncation method [26] by stopping gradient propagation at the t-th layer, i.e., neglecting the second term in $\frac{\partial \operatorname{vec}(\mathbf{W}^{(t+1)})}{\partial \boldsymbol{\Theta}}$. The network parameters are trained using random realizations of channels $\mathbf{H}$ and initializations $\mathbf{W}^{(0)}$. The training set contains 10^3 samples with a batch size of 10. We use the Adam optimizer with step size 10^{-3}. The code is available at: https://github.com/yangzhexian/MM-Net.

We compare the performance of MM-Net with both optimization-based and learning-based methods. The number of layers in MM-Net is set to $T = 20$. For the learning-based baselines, WMMSE-Net and IAIDNN use 10 layers; in WMMSE-Net, the update steps for $\mathbf{U}_k$, $\mathbf{W}_k$, and $\mathbf{V}_k$ are set to $J_u = 2$, $J_w = 2$, and $J_v = 4$, respectively.

We first consider $K = 4$ users, each with $D_k = 4$ data streams and $N_k = 4$ antennas for $k = 1,\ldots,K$. Fig. 2(a) shows runtime convergence for $M = 128$. MM-Net converges much faster than WMMSE and MM, consistent with its design that inherits the MM structure while introducing learnable components for acceleration. Compared with WMMSE-Net, MM-Net achieves the same WSR with less runtime, demonstrating superior efficiency. MM-Net also achieves better WSR than IAIDNN, for which only the final WSR is reported as intermediate outputs violate the total power constraint. We further examine performance as the number of transmit antennas M increases, particularly in the massive MIMO regime. Fig. 2(b) plots WSR versus M, while Fig. 2(c) reports runtimes. As M grows, MM-Net slightly outperforms IAIDNN and WMMSE-Net. Meanwhile, its runtime advantage over these learning-based baselines becomes increasingly pronounced. This aligns with Section IV, where MM-Net preserves the same order of complexity as MM while avoiding heavier computations in WMMSE-based unfolding. Finally, we examine generalization in mixed-training settings, as shown in Fig. 2(d). For $M = 16$ and $D_k = N_k$, a mixed training set combines samples from $(K, D_k) = (4,4),(4,8),(8,4),(8,8)$, each with 250 samples, to train MM-Net (mixed). Under $(8,4)$, MM-Net (mixed) is compared with MM-Net trained solely on $(8,4)$. Results show that MM-Net (mixed) generalizes well, and although its convergence in WSR is slightly slower than that of the configuration-specific MM-Net, it still surpasses the conventional MM algorithm.

VI. Conclusion

In this paper, we have proposed an MM-based deep unfolding network, termed MM-Net, for WSR beamforming. By sharing parameters across layers, MM-Net forms an RNN structure that improves the efficiency of the MM algorithm while remaining highly parameter-efficient. We have established its convergence properties and have analyzed its computational complexity. Simulation results have shown that MM-Net achieves the same WSR performance while converging faster than both traditional optimization methods and existing learning-based unfolding networks. Finally, we note that the proposed MM-based unfolding framework naturally leads to an RNN structure, suggesting a general approach for designing efficient learning architectures from MM-type iterative algorithms, which may be extended to a broader class of optimization problems [27].

REFERENCES

[1] Z. Yang, Z. Zhang, and Z. Zhao, "Weighted Sum-Rate Maximization for Beamforming Design Using Minorization-Maximization: Convergence Rate and Deep Unfolding," in *2025 33rd European Signal Processing Conference (EUSIPCO)*, 2025, pp. 2457–2461.

[2] Z. Zhang, Z. Zhao, and K. Shen, "Enhancing the Efficiency of WMMSE and FP for Beamforming by Minorization-Maximization," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[3] Z. Zhang, Z. Zhao, K. Shen, D. P. Palomar, and W. Yu, "Discerning and Enhancing the Weighted Sum-Rate Maximization Algorithms in Communications," *Available at arXiv:2311.04546*, 2023.

[4] M. Chiang, P. Hande, T. Lan, C. W. Tan *et al.*, "Power Control in Wireless Cellular Networks," *Foundations and Trends® in Networking*, vol. 2, no. 4, pp. 381–533, 2008.

[5] Z.-Q. Luo and S. Zhang, "Dynamic Spectrum Management: Complexity and Duality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[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, Sep. 2011.

[8] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a Reality - What is Next? Five Promising Research Directions for Antenna Arrays," *Digital Signal Processing*, vol. 94, pp. 3–20, 2019.

[9] Z. Zhang and Z. Zhao, "Weighted Sum-Rate Maximization for Multi-Hop RIS-Aided Multi-User Communications: A Minorization-Maximization Approach," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2021, pp. 106–110.

[10] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[11] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[12] W. Yu and T. Lan, "Transmitter Optimization for the Multi-Antenna Downlink With Per-Antenna Power Constraints," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2646–2660, Jun. 2007.

[13] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to Optimize: Training Deep Neural Networks for Interference Management," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.

[14] W. Xia, G. Zheng, Y. Zhu, J. Zhang, J. Wang, and A. P. Petropulu, "A Deep Learning Framework for Optimization of MISO Downlink Beamforming," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1866–1880, Mar. 2020.

[15] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal Multiuser Transmit Beamforming: A Difficult Problem with a Simple Solution Structure," *IEEE Signal Processing Magazine*, vol. 31, no. 4, pp. 142–148, Jul. 2014.

[16] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing," *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[17] Q. Hu, Y. Cai, Q. Shi, K. Xu, G. Yu, and Z. Ding, "Iterative Algorithm Induced Deep-Unfolding Neural Networks: Precoding Design for Multiuser MIMO Systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1394–1410, Feb. 2021.

[18] L. Pellaco, M. Bengtsson, and J. Jaldén, "Matrix-Inverse-Free Deep Unfolding of the Weighted MMSE Beamforming Algorithm," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 65–81, 2022.

[19] L. Pellaco and J. Jaldén, "A Matrix-Inverse-Free Implementation of the MU-MIMO WMMSE Beamforming Algorithm," *IEEE Transactions on Signal Processing*, vol. 70, pp. 6360–6375, 2022.

[20] M. Zhu, T.-H. Chang, and M. Hong, "Learning to Beamform in Heterogeneous Massive MIMO Networks," *IEEE Transactions on Wireless Communications*, vol. 22, no. 7, pp. 4901–4915, Jul. 2023.

[21] H. Hojatian, J. Nadal, J.-F. Frigon, and F. Leduc-Primeau, "Unsupervised Deep Learning for Massive MIMO Hybrid Beamforming," *IEEE Transactions on Wireless Communications*, vol. 20, no. 11, pp. 7086–7099, Nov. 2021.

[22] A. Chowdhury, G. Verma, A. Swami, and S. Segarra, "Deep Graph Unfolding for Beamforming in MU-MIMO Interference Networks," *IEEE Transactions on Wireless Communications*, vol. 23, no. 5, pp. 4889–4903, May 2024.

[23] C. Xu, Y. Jia, S. He, Y. Huang, and D. Niyato, "Joint User Scheduling, Base Station Clustering, and Beamforming Design Based on Deep Unfolding Technique," *IEEE Transactions on Communications*, vol. 71, no. 10, pp. 5831–5845, Oct. 2023.

[24] K. Zhang, L. Van Gool, and R. Timofte, "Deep Unfolding Network for Image Super-Resolution," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 3214–3223.

[25] P. Werbos, "Backpropagation Through Time: What It Does and How to Do It," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[26] H. Jaeger, *Tutorial on Training Recurrent Neural Networks, Covering BPPT, RTRL, EKF and the "echo state network" Approach*. GMD-Forschungszentrum Informationstechnik Bonn, 2002.

[27] Z. Yang and Z. Zhao, "From WMMSE to XMMSE: An Old Melody Sung in a Fast New Tune," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 21 151–21 155.