# Deep Unfolding in Multicell MU-MIMO

Lukas Schynol, Marius Pesavento
Communication Systems Group, Technische Universität Darmstadt, Darmstadt, Germany
Emails: {lschynol, pesavento}@nt.tu-darmstadt.de

***Abstract*—The weighted sum-rate maximization in coordinated multicell MIMO networks with intra- and intercell interference and local channel state at the base stations is considered. Based on the concept of unrolling applied to the classical weighted minimum mean squared error (WMMSE) algorithm and ideas from graph signal processing, we present the GCN-WMMSE deep network architecture for transceiver design in multicell MU-MIMO interference channels with local channel state information. Similar to the original WMMSE algorithm it facilitates a distributed implementation in multicell networks. However, GCN-WMMSE significantly accelerates the convergence and consequently alleviates the communication overhead in a distributed deployment. Additionally, the architecture is agnostic to different wireless network topologies while exhibiting a low number of trainable parameters and high efficiency w.r.t. training data.**

## I. Introduction

The design of downlink (DL) transmit and receive beamformers in multicell multi-user (MU) multiple-input multiple-output (MIMO) is considered. In the presence of power constraints the weighted sum-rate (WSR) maximization utility is commonly leveraged to obtain optimal beamformers. However, in the case of practical linear transceivers, the resulting optimization problem is generally NP-hard [1], and globally optimal algorithms [2] have an exponential runtime. The weighted minimum mean squared error (WMMSE) iterative algorithm [3] is only a locally optimal solution but is widely regarded as a benchmark for WSR maximization since its updates have a simple form while achieving a high WSR. Furthermore, it enables the distributed optimization of beamformers in multicell systems while relying on local channel state information (CSI) and only a limited communication overhead per iteration. Nevertheless, the WMMSE algorithm, and iterative optimization-based algorithms in general, require a large number of iterations to converge, making them difficult to apply in practice [4].

Recently, the concept of deep algorithm unrolling/unfolding gained significant interest in the signal processing community [5], [6]. Here, iterations of problem-specific algorithms are interpreted as layers of machine learning models and then infused with approximations and trainable parameters. The resulting models are trained with data to recover or improve over the performance of the original algorithm with only a limited number of layers, thereby reducing the computational cost. Algorithm unrolling straightforwardly combines expert knowledge with machine learning concepts, enabling a better generalization performance, transferability and interpretability compared to conventional neural network architectures which is key for future robust wireless communication networks.

A number of recent works applied algorithm unrolling in the context of DL beamforming. In [7] efficient solutions to the WSR maximization in multiple-input single-output (MISO) networks consisting of transmitter-receiver pairs are found. This was accomplished by unrolling the inexact cyclic coordinate descent method applied to the WSR maximization problem. The WMMSE algorithm is unfolded in [8] by approximating one of its updates by truncated projected gradient descent (PGD), the step sizes of which are learned using data. This facilitates a trade-off between complexity and performance, however, the discussion is limited to single-cell MISO networks. In [9], the updates of the WMMSE algorithm are replaced by trainable approximations of the original matrix operations. Although limited to MU-MIMO, the architecture achieves performance similar to the WMMSE algorithm with significantly reduced computational complexity. In [10], interfering pairwise single-input single-output (SISO) links are considered. By incorporating graph convolutional networks (GCNs) [11] into the WMMSE algorithm structure, the number of required iterations is significantly reduced.

Nevertheless, to our best knowledge, beamforming algorithms based on deep unrolling have not yet been considered in general multicell network scenarios with MU-MIMO and inter- and intracell interference. The contributions of this paper are, therefore, as follows: 1) Utilizing GCNs, we propose a completely general architecture, which we denote as GCN-WMMSE, based on unfolding the WMMSE algorithm for WSR maximization in multicell MU-MIMO networks that is agnostic to the wireless scenario configuration. 2) We show that GCN-WMMSE significantly reduces the communication overhead over the WMMSE algorithm by reducing the number of required iterations/layers while exhibiting a similar complexity per iteration. 3) We demonstrate the excellent generalization capability and training data efficiency of the proposed GCN-WMMSE architecture in simulations on Rayleigh fading channel models.

The rest of the paper is structured as follows: Section II defines the system model, resource allocation problem and reviews the WMMSE algorithm. The GCN-WMMSE architecture is proposed in Section III. Section IV presents the simulations results. Section V concludes this work.

## II. System Model and WMMSE Algorithm

### A. System Model and Downlink Problem Formulation

Consider a wireless cellular system consisting of $K$ cells with $K$ base stations (BSs), each serving one of $K$ disjoint subsets $\{\mathcal{I}_k\}_{k=1}^{K}$ of user equipments (UEs), for a total of $I = \sum_{k=1}^{K} |\mathcal{I}_k|$ UEs. UE $i$ is equipped with an array of size $N_i$ and BS $k$ is equipped with an antenna array of size $M_k$. We assume complex frequency-flat channels $\mathbf{H}_{ik} \in \mathbb{C}^{N_i \times M_k}$ from BS $k$ to UE $i$, linear DL beamforming from BS $k$ to UE $i \in \mathcal{I}_k$ of a symbol vector $\check{\mathbf{s}}_i \in \mathbb{C}^{N_i}$ with the DL beamformer matrix $\mathbf{V}_i \in \mathbb{C}^{M_k \times N_i}$, and linear receive beamforming with a matrix $\mathbf{U}_i \in \mathbb{C}^{N_i \times N_i}$ at UE $i$ under additive complex white

This work was financially supported by the Federal Ministry of Education and Research of Germany in the project "Open6GHub" (grant no. 16KISK014). The computations for this research paper were conducted on the Lichtenberg high performance computer of the TU Darmstadt.

Gaussian noise $\mathbf{n}_i \sim \mathcal{CN}\left(0, \sigma_i^2 \mathbf{I}\right)$ with power $\sigma_i^2$. Let $\mathbf{Q}_{ij} = \mathbf{H}_{im}\mathbf{V}_j\mathbf{V}_j^{\mathrm{H}}\mathbf{H}_{im}^{\mathrm{H}}$ be the signal covariance at UE $i$ due to the signal for UE $j$ originating at BS $m$ with $j \in \mathcal{I}_m$, then the constrained maximization of the achievable WSR $\mathcal{R}_\Sigma$ becomes

$$\begin{aligned} \max_{\{\mathbf{V}_i\}_{i=1}^I} \quad & \sum_{i=1}^{I} \alpha_i \operatorname{logdet}\left(\mathbf{I} + \mathbf{Q}_{ii}\Bigg(\sum_{m=1}^{K}\sum_{j\in\mathcal{I}_m\setminus\{i\}} \mathbf{Q}_{ij} + \sigma_i^2\mathbf{I}\Bigg)^{-1}\right) \\ \text{s. t.} \quad & \sum_{i\in\mathcal{I}_k} \operatorname{Tr}\left\{\mathbf{V}_i\mathbf{V}_i^{\mathrm{H}}\right\} \le P_k,\ \forall k \in \{1,\dots,K\} \end{aligned} \tag{1}$$

where $\alpha_i > 0$ is a predefined weight and $P_k$ is the power budget of BS $k$. The bandwidth is $1$ without loss of generality.

### B. WMMSE Algorithm

The authors of [3] reformulate the problem in (1) into an equivalent one by introducing a positive semidefinite (PSD) weight matrix $\mathbf{W}_i \succeq \mathbf{0}$ per UE $i$. Let $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{V}$ denote the set of receive beamformers $\{\mathbf{U}_i\}_{i=1}^I$, weight matrices $\{\mathbf{W}_i\}_{i=1}^I$ and DL beamformers $\{\mathbf{V}_i\}_{i=1}^I$. By leveraging the blockwise convexity of the reformulation and the block coordinate descent [12] framework, the convergent WMMSE algorithm [3] is obtained with sequential updates

$$(\mathbf{U}\text{-Step}) \quad \forall i: \mathbf{U}_i^{(\ell)} = (\mathbf{J}_i^{(\ell)})^{-1}\mathbf{H}_{ik}\mathbf{V}_i^{(\ell-1)} \tag{2a}$$

$$(\mathbf{W}\text{-Step}) \quad \forall i: \mathbf{W}_i^{(\ell)} = \left(\mathbf{I} - (\mathbf{V}_i^{(\ell-1)})^{\mathrm{H}}\mathbf{H}_{ik}^{\mathrm{H}}\mathbf{U}_i^{(\ell)}\right)^{-1}, \tag{2b}$$

$$\begin{aligned} (\boldsymbol{\mu}\text{-Step}) \quad \forall k: \mu_k^{(\ell)} &= \operatorname*{argmax}_{\mu_k} \mu_k \\ \text{subject to} \quad 0 &= f_{\mathrm{CS},k}^{(l)}(\mu_k), \quad \mu_k \ge 0, \end{aligned} \tag{2c}$$

$$(\mathbf{V}\text{-Step}) \quad \forall i: \mathbf{V}_i^{(\ell)} = \left(\mathbf{R}_k^{(\ell)} + \mu_k^{(\ell)}\mathbf{I}\right)^{-1}\tilde{\mathbf{V}}_i^{(\ell)} \tag{2d}$$

for iteration $\ell$. Note that the indices $k$ and $i$ are chosen such that $i \in \mathcal{I}_k$. We defined $\mathbf{J}_i^{(\ell)} = \sum_{m=1}^K \sum_{j\in\mathcal{I}_m} \mathbf{Q}_{ij}^{(l-1)} + \sigma_i^2\mathbf{I}$ as the receive signal covariance matrix at UE $i$,

$$\mathbf{R}_k^{(\ell)} = \sum\nolimits_{j=1}^{I} \alpha_j \mathbf{H}_{jk}^{\mathrm{H}}\mathbf{U}_j^{(\ell)}\mathbf{W}_j^{(\ell)}(\mathbf{U}_j^{(\ell)})^{\mathrm{H}}\mathbf{H}_{jk} \tag{3}$$

as the weighted uplink covariance matrix at BS $k$ and

$$\tilde{\mathbf{V}}_i^{(\ell)} = \alpha_i \mathbf{H}_{ik}^{\mathrm{H}}\mathbf{U}_i^{(\ell)}\mathbf{W}_i^{(\ell)} \tag{4}$$

as the candidate beamformer matrix of UE $i \in \mathcal{I}_k$ at BS $k$ respectively. Equations (2c) and (2d) result from applying Karush-Kuhn-Tucker conditions to solve for $\mathbf{V}_i^{(\ell)}$ under the constraint $\sum_{i\in\mathcal{I}_k} \operatorname{Tr}\left\{\mathbf{V}_i^{(\ell)}(\mathbf{V}_i^{(\ell)})^{\mathrm{H}}\right\} \le P_k$, thereby introducing dual variables $\mu_k^{(l)}$ as well as the complementary slackness conditions

$$f_{\mathrm{CS},k}^{(\ell)}(\mu_k) = \left(\sum_{m=1}^{M_k} \frac{\varphi_{km}^{(\ell)}}{(\lambda_{km}^{(\ell)} + \mu_k)^2} - 1\right)\mu_k. \tag{5}$$

The quantities $\lambda_{km}^{(\ell)} = \left[\boldsymbol{\Lambda}_k^{(\ell)}\right]_{mm}$ and $\varphi_{km}^{(\ell)} = \left[\boldsymbol{\Phi}_k^{(\ell)}\right]_{mm}$ are obtained from the eigendecomposition $\mathbf{R}_k^{(\ell)} = \mathbf{D}_k^{(\ell)}\boldsymbol{\Lambda}_k^{(\ell)}(\mathbf{D}_k^{(\ell)})^{\mathrm{H}}$ and $\boldsymbol{\Phi}_k^{(\ell)} = \frac{1}{P_k}(\mathbf{D}_k^{(\ell)})^{\mathrm{H}}\left(\sum_{i\in\mathcal{I}_k} \tilde{\mathbf{V}}_i^{(\ell)}(\tilde{\mathbf{V}}_i^{(\ell)})^{\mathrm{H}}\right)\mathbf{D}_k^{(\ell)}$, respectively.

Importantly, the algorithm can be implemented in a distributed fashion by computing the $\mathbf{U}$-steps and $\mathbf{W}$-steps at the respective UEs, and BS $k$ computes all $\mathbf{V}_i$ where $i \in \mathcal{I}_k$. UE $i$ can locally estimate $\mathbf{J}_i$ and only requires information about its assigned precoding matrix as well as the channel toward its assigned BS. BS $k$ on the other hand only requires local CSI, the beamformer candidate matrices $\tilde{\mathbf{V}}_i$ for $i \in \mathcal{I}_k$ as well as the matrices $\mathbf{U}_i\mathbf{W}_i\mathbf{U}_i^{\mathrm{H}}$ for each UE within its cell radius.

## III. Proposed GCN-WMMSE Architecture

In this section, the proposed **G**raph-**C**onvolutional-**N**etwork-WMMSE architecture, or GCN-WMMSE in short, is introduced by unrolling the WMMSE algorithm and modifying it. The architecture and training procedure is explained in Section III-A1 and III-B, respectively, followed by a brief discussion in Section III-C.

### A. Architecture

*1) WMMSE Algorithm Unfolding:* We unfold $L$ iterations of the WMMSE algorithm by interpreting each iteration $\ell$ with input $\mathbf{V}^{(\ell-1)}$ and output $\mathbf{V}^{(\ell)}$ as a neural network layer. Figure 1 visualizes the forward pass of the resulting network. Each layer consists of the update blocks $\mathcal{F}_{\mathbf{W}}(\mathbf{V};\mathcal{S})$, $\mathcal{F}_{\mathrm{U}}(\mathbf{V},\mathbf{W};\mathcal{S})$ and $\mathcal{F}_{\mathbf{V}}(\mathbf{U},\mathbf{W},\boldsymbol{\mu};\mathcal{S})$, which are the update mappings (2a), (2b) and (2d) for every UE $i$, and $\mathcal{F}_{\boldsymbol{\mu}}(\mathbf{U},\mathbf{W};\mathcal{S})$, which is (2c) for every BS $k$. Compared to previous works [8]–[10], [13], which are restricted to less general wireless scenarios in order to simplify the WMMSE algorithm, the general WMMSE algorithm is considered in this work. In this case, the mapping $\mathcal{F}_{\boldsymbol{\mu}}$, i.e., problem (2c), generally does not possess a closed-form solution, hence, the authors of [3] propose the iterative bisection search to obtain the optimal value $\mu_k^{(\ell,\mathrm{opt})}$. However, since it requires many subiterations $P$ to approach a sufficient accuracy, we instead propose to utilize a method based on rational function approximations as in [14] to accelerate the forward pass of the network. Compared to the iterative bisection algorithm, it yields a highly accurate result after only a low number of subiterations.

Next, the $\mathcal{F}_{\mathbf{W}}$-blocks and $\mathcal{F}_{\mathbf{V}}$-blocks are modified using concepts from graph convolutional filters (GCFs) and GCNs [11] to significantly reduce the number of required layers while maintaining performance.

*2) Weight Matrix Graph Filter:* The idea of GCFs is applied to the $\mathcal{F}_{\mathbf{W}}$-block. To this end, we consider the original weight matrix update in (2b) $\hat{\mathbf{W}}^{(\ell)} = \left(\mathbf{I} - (\mathbf{V}_i^{(\ell-1)})^{\mathrm{H}}\mathbf{H}_{ik}^{\mathrm{H}}\mathbf{J}_i^{(l)}\mathbf{H}_{ik}\mathbf{V}_i^{(\ell-1)}\right)^{-1}$, where $i \in \mathcal{I}_k$, as a graph shift matrix and introduce the weight matrix GCF

$$\mathbf{W}_i^{(\ell)} = a_{\mathrm{W},\ell 0}\mathbf{I} + \sum_{g=1}^{G} \frac{a_{\mathrm{W},\ell g}}{\left(\operatorname{Tr}\left\{\hat{\mathbf{W}}_i^{(\ell)}\right\}/N_i\right)^{g-1}}\left(\hat{\mathbf{W}}_i^{(\ell)}\right)^g \tag{6}$$

of order $G$, where $a_{\mathrm{W},\ell g}$ for $g = 0,\dots,G$ and $\ell = 1,\dots,L$, are a learnable filter taps. Thus, (6) replaces the standard WMMSE weight update (2b). To ensure that the resulting $\mathbf{W}_i^{(\ell)}$ is PSD, the filter taps $a_{\mathrm{W},\ell g}$ are restricted to be non-negative ($a_{\mathrm{W},\ell g} \ge 0$). Additionally, the mean of the eigenvalues $\operatorname{Tr}\left\{\hat{\mathbf{W}}_i^{(\ell)}\right\}/N_i$ normalizes the filter taps for $g \ge 2$ to prevent numerical issues in case of a high signal-to-noise ratio (SNR).

*3) DL Graph Convolutional Neural Network:* The $\mathbf{V}$-step in (2d) can be interpreted as a GCF acting on the candidate beamformer $\tilde{\mathbf{V}}_i^{(\ell)}$ in (4) with the inverse of the modified weighted uplink covariance matrix $\tilde{\mathbf{R}}_k^{(\ell)} = \mathbf{R}_k^{(\ell)} + \mu_k^{(\ell)}\mathbf{I}$ in (3) as the graph shift matrix. The filter is then extended into a complex-valued GCN layer with $F$ features, leading to the modified $\mathcal{F}_{\mathbf{V}}$-update

$$\begin{aligned} \hat{\mathbf{v}}_{id}^{(\ell)} &= \operatorname{modReLU}\left(\tilde{\mathbf{P}}_{id}^{(\ell)}, \frac{1}{b_{\mathrm{S}}}\sqrt{\frac{P_k}{|\mathcal{I}_k|}}\mathbf{1}\mathbf{b}_\ell^{\mathrm{T}}\right)\mathbf{c}_\ell \\ \text{where} \quad \tilde{\mathbf{P}}_{id}^{(\ell)} &= (\tilde{\mathbf{R}}_k^{(\ell)})^{-1}\tilde{\mathbf{v}}_{id}^{(\ell)}\mathbf{a}_{\mathrm{V},\ell 1}^{\mathrm{T}} + \tilde{\mathbf{v}}_{id}^{(\ell)}\mathbf{a}_{\mathrm{V},\ell 0}^{\mathrm{T}} \end{aligned} \tag{7}$$

for every $i \in \mathcal{I}_k$ for all $k$, which obtains the unscaled beamformer $\hat{\mathbf{V}}_i = \left[\hat{\mathbf{v}}_{i1}^{(\ell)},\dots,\hat{\mathbf{v}}_{iN_i}^{(\ell)}\right]$ where $\hat{\mathbf{v}}_{id}^{(\ell)}$ for $d = 1,\dots,N_i$ belong to the individual streams. Similarly $\tilde{\mathbf{V}}_i^{(\ell)} = \left[\tilde{\mathbf{v}}_{i1}^{(\ell)},\dots,\tilde{\mathbf{v}}_{iN_i}^{(\ell)}\right]$. The GCN layer corresponds to a filter of polynomial degree $1$ with $\mathbf{a}_{\mathrm{V},\ell 1} \in \mathbb{C}^F$ and $\mathbf{a}_{\mathrm{V},\ell 0} \in \mathbb{C}^F$ being

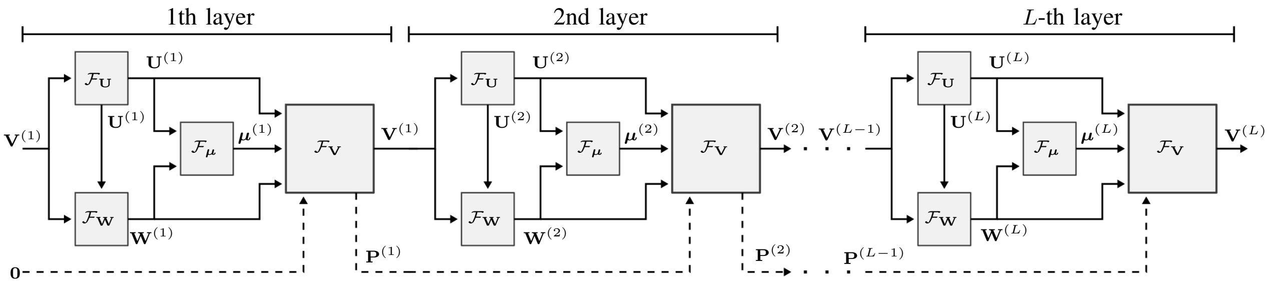

Figure 1. Deep network of the WMMSE algorithm obtained by unrolling $L$ iterations. The blocks $\mathcal{F}_{\mathbf{U}}$, $\mathcal{F}_{\mathbf{W}}$ and $\mathcal{F}_{\mathbf{V}}$ represent the updates (2a), (2b) and (2d) for every $i$. The $\mathcal{F}_{\boldsymbol{\mu}}$-block contains $P$ subiterations for every $k$. The GCN-WMMSE architecture modifies the $\mathcal{F}_{\mathbf{W}}$- and $\mathcal{F}_{\mathbf{V}}$-blocks of the original algorithm. Dashed arrows represent the skip connections of Section III-A4.

trainable complex filter taps, $\mathbf{b}_\ell \in \mathbb{R}^F$ being a trainable bias, and $\mathbf{c}_\ell \in \mathbb{C}^F$ being a trainable vector which recombines the $F$ features contained in $\tilde{\mathbf{P}}_{id}^{(\ell)}$. The generalization performance is enhanced by scaling the bias by the power budget and the number of assigned UEs. The auxiliary scaling parameter $b_{\mathrm{S}}$ stabilizes the training in case of applying an optimizer such as ADAMW [15] which normalizes the gradients. The elementwise complex-valued nonlinearity $\mathrm{modReLU}$ is defined in [16] and preserves the phase. To ensure feasibility w.r.t. the power budgets, the power projection step

$$\mathbf{V}_i^{(\ell)} = \sqrt{P_k}\hat{\mathbf{V}}_i^{(\ell)} \Big/ \sqrt{\max\left\{\sum_{i\in\mathcal{I}_k} \mathrm{Tr}\left\{\hat{\mathbf{V}}_i^{(\ell)}(\hat{\mathbf{V}}_i^{(\ell)})^{\mathrm{H}}\right\}, P_k\right\}} \tag{8}$$

for all $i$ is introduced at the output of each $\mathcal{F}_{\mathbf{V}}$-block.

*4) Skip Connections:* Lastly, we integrate additional connections between layers [17] by replacing the input to the $\mathrm{modReLU}$ nonlinearity $\tilde{\mathbf{P}}_{id}^{(\ell)}$ with $\mathbf{P}_{id}^{(\ell)} = \tilde{\mathbf{P}}_{id}^{(\ell)} + \mathbf{P}_{id}^{(\ell-1)}\mathbf{D}_\ell$, where the trainable matrix $\mathbf{D}_\ell \in \mathbb{C}^{F\times F}$ allows for learned linear combinations of the features of layer $\ell-1$. The resulting direct path enables an additional exchange of gradient information between layers, bypassing operations such as matrix inversions which sometimes lead to noisy gradients. Fig. 1 visualizes the connections with dashed arrows.

### B. Model Training

Optimizing the trainable parameter set $\mathbf{\Gamma}$ of a GCN-WMMSE network $\mathcal{M}(\mathcal{S};\mathbf{\Gamma})$ with $L$ layers, which contains $\{\mathbf{a}_{\mathrm{V},\ell 1}\}_{\ell=1}^L$, $\{\mathbf{a}_{\mathrm{V},\ell 0}\}_{\ell=1}^L$, $\{\mathbf{b}_\ell\}_{\ell=1}^L$, $\{\mathbf{c}_\ell\}_{\ell=1}^L$, $\{\mathbf{D}_\ell\}_{\ell=2}^L$ and $\{\{a_{\mathrm{W},\ell g}\}_{g=0}^G\}_{\ell=1}^L$, is achieved by the maximization of the expected rate $\mathbb{E}_{\mathcal{S}}\left[\mathcal{R}_\Sigma\right]$ over the distribution of scenario realizations $p(\mathcal{S})$, which in practice is approximated by a finite data set $\mathcal{T}$. In this work, we perform stochastic gradient descent (SGD) for $T$ steps by descending along the gradient $\nabla_{\mathbf{\Gamma}} J(\mathcal{T}_t;\mathbf{\Gamma},\mathcal{L})$ of the sample-normalized loss function

$$J(\mathcal{T}_t;\mathbf{\Gamma}) = \frac{1}{|\mathcal{T}_t|}\sum_{\mathcal{S}_n\in\mathcal{T}_t} \frac{J_{\mathrm{WSR}}^{(L)}(\mathcal{S}_n;\mathbf{\Gamma})}{r_{n,\mathbf{\Gamma}}^{(\ell)}}, \tag{9}$$

where $\mathcal{T}_t \subset \mathcal{T}$ is a minibatch of scenario realizations at training step $t$. The partial loss $J_{\mathrm{WSR}}^{(L)}$ is the negative WSR corresponding to (1) achieved on the realization $\mathcal{S}_n$ with the output DL beamformer set $\mathbf{V}^{(L)}$, i.e., $J_{\mathrm{WSR}}^{(L)}(\mathcal{S}_n;\mathbf{\Gamma}) = -\mathcal{R}_\Sigma(\mathcal{M}(\mathcal{S}_n;\mathbf{\Gamma});\mathcal{S}_n)$. The scalar $r_{n,\mathbf{\Gamma}}^{(L)}$ equalizes the impact of all realizations regardless of the achieved WSR and is determined as $\left|J_{\mathrm{WSR}}^{(L)}(\mathcal{S}_n;\mathbf{\Gamma})\right|$ in the forward pass [18]. The scaling parameter $b_{\mathrm{S}}$ in (7) is obtained during training as the empirical expectation $\mathbb{E}_{\mathcal{S}\sim\mathcal{T}}\left[\sqrt{P_k/\left|\mathcal{I}_k\right|}\right]$ dependent on the power budget of scenarios in the training set $\mathcal{T}$.

To perform SGD, the complex gradient $\nabla_{\mathbf{\Gamma}} J(\mathcal{T}_t;\mathbf{\Gamma},\mathcal{L})$ must be available. In deep networks, it is usually obtained by backpropagation which requires the Jacobian of each update block. This is straightforward for $\mathcal{F}_{\mathbf{U}}$, $\mathcal{F}_{\mathbf{W}}$ and $\mathcal{F}_{\mathbf{V}}$. However, the problem represented by the $\mathcal{F}_{\boldsymbol{\mu}}$-block must generally be solved iteratively. Empirically, backpropagation through these subiterations is prone to numerical issues as it involves divisions by small numbers. As a remedy, we apply the notion of derivatives of implicit functions [19, Thm. 8.2] to the complementary slackness condition $f_{\mathrm{CS},k}^{(\ell)}(\mu_k) = 0$, which yields closed-form relations of the local gradient. A detailed treatment of the $\mathcal{F}_{\boldsymbol{\mu}}$-block can be found in our companion work [20].

### C. Discussion

GCN-WMMSE makes use of the permutation equivariance of GCFs and GCNs [11], leading to the achieved WSR being unaffected by relabeling of the transceiver antennas which is a natural property of the wireless system. Simultaneously, complete transferability of the network to any wireless scenario configuration is enabled. Furthermore, the filters are globally optimized in comparison to the local optimality of the block coordinate descent (BCD) operations, which allows for performance improvements over the original algorithm. Lastly, the number of trainable parameters in the set $\mathbf{\Gamma}$ is low with $(L-1)F^2 + L(4F+G+1)$, reducing the risk of overfitting.

## IV. Experimental Evaluation

### A. Simulation Setup

The proposed GCN-WMMSE architecture is implemented[1] using PyTorch and we leverage the AdamW optimizer [15]. The GCN-WMMSE network models are evaluated in a scenario of 3 BSs positioned at the corners of an equilateral triangle of side length $d_{\mathrm{BS}}$. For each scenario realization, all UEs are placed uniformly at random inside the sextant centered on their assigned BS with radius $d_{\mathrm{BS}}/\sqrt{3}$. The picocell model [21] is used to calculate the large-scale path loss $PL_{ik}$ between BS $k$ and UE $i$. We assume equal BS antenna dimensions BS $M_k = M$, BS power budget $P_k = P_{\mathrm{BS}}$, UE antennas dimensions $N_i = N$, UE noise variances $\sigma_i^2 = \sigma_{\mathrm{UE}}^2$ and sum-rate weights $\alpha_i = 1$. Each BS serves the same number of UEs. Assuming rich scattering and Rayleigh fading, the channel matrix coefficients $\left[\mathbf{H}_{ik}\right]_{nm}$ are sampled from $\mathcal{CN}\left(0, 10^{\frac{PL_{ik}}{10\,\mathrm{dB}}}\right)$.

All experiments are conducted using the network and training hyperparameters and scenario configuration parameters

[1] To promote reproducible research, the code is publicly available at https://github.com/lsky96/gcnwmmse.

Table I. Base parameter set for GCN-WMMSE networks, training parameters and base scenario configuration.

| **Base Model Hyperparameters** | |
|---|---|
| Number of Layers $L$ | 7 |
| Polynomial Degree $G$ | 2 |
| Number of Filters $F$ | 4 |
| **Base Training Hyperparameters** | |
| Loss Function | Eq. (9) |
| ADAMW $(\beta_1, \beta_2, \lambda)$ | $(0.9, 0.99, 10^{-3})$ |
| Learning Steps $T$ | $10^4$ |
| Learning Rate $\eta$ | 0.01, /10 after every 2500 steps |
| Minibatch Size $\lvert\mathcal{T}_t\rvert$ | 100 |
| Gradient Clipping Value | 1 |
| **Base Scenario Configuration** | |
| BS Distance $d_{\mathrm{BS}}$ | $200\,\mathrm{m}$ |
| BS Antenna Dimension $M$ | 12 |
| BS Tx Power $P_{\mathrm{BS}}$ | $30\,\mathrm{dBm}$ |
| Num. of UEs $I$ | 12, equally assigned |
| UE Antenna Dimension $N$ | 2 |
| UE Noise Power $\sigma^2_{\mathrm{UE}}$ | $-100\,\mathrm{dBm}$ |

Table II. Absolute and relative WSR of GCN-WMMSE (proposed) for different numbers of layers $L$.

| **# Layers** $L$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| Abs. Rate $\mathcal{R}_\Sigma$ $(\frac{\mathrm{nat}}{\mathrm{Hz}})$ | 82.21 | 87.84 | 89.98 | 91.67 | 92.56 | 93.46 | 93.89 |
| Rel. Rate $\mathcal{R}_\Sigma^{\mathrm{rel.}}$ (%) | 83.72 | 89.45 | 91.63 | 93.35 | 94.26 | 95.17 | 95.61 |

summarized in Tab. I, unless specified otherwise. The filter taps of the DL beamformer GCNs are initialized according to [16], the biases with $\mathbf{0}$ and the taps of the weight matrix GCFs $a_{\mathrm{W},\ell g}$ by $1/(G+1)$. The input beamformers are given by normalized maximum-ratio combining (MRC) beamformers $\mathbf{V}_i^{(0)} \propto \mathbf{H}_{ik}^{\mathrm{H}}$ for $i \in \mathcal{I}_k$. The Lagrangian variable $\mu_k^{(\ell)}$ is initialized as $10^{-12}$ and updated for $P = 8$ subiterations. The validation sets contain $10^3$ scenario realizations, the training sets contain $10^7$ realizations except for the experiments concerning Tab. III. We compare to the classical WMMSE algorithm running for 100 iterations. WMMSE RI denotes the WSR achieved by the WMMSE algorithm averaged over 50 random initializations per scenario realization. WMMSE50 serves as a benchmark and denotes the best WSR achieved over these 50 initializations, however, it has a high computational cost. The acronym TR indicates the WSR the WMMSE algorithm achieves after $L$ iterations, thereby having the same communication overhead as the GCN-WMMSE networks. The average achievable WSR is denoted by $\mathcal{R}_\Sigma$, the WSR relative to the result of WMMSE50 is denoted as $\mathcal{R}_\Sigma^{\mathrm{rel.}}$. Note that comparisons with previous architectures based on unrolling the WMMSE are relegated to the companion work [20].

### *B. Simulation Results*

When varying the number of layers of the proposed GCN-WMMSE network models, Tab. II shows that the WSR increases consistently from 3 to 9 layers up to 95.61% relative WSR, whereas WMMSE RI achieves 95.20% relative WSR only after 100 iterations; WMMSE RI achieves a value comparable to $L = 7$ of the GCN-WMMSE only after 64 iterations. This demonstrates a significant reduction in iterations, therefore, computational cost and communication overhead. The training set size $\lvert\mathcal{T}\rvert$ is considered in Tab. III. For $\lvert\mathcal{T}\rvert = 100$ samples the achieved relative WSR is already 93.26%, which is close to the value for $10^7$ training samples. This remarkable data efficiency is enabled by the equivariance properties of the graph filter structures and the preservation of the WMMSE algorithm structure, which lead to a low number of trainable parameters (here 229 scalars) as well.

Table III. Absolute and relative WSR of GCN-WMMSE (proposed) for a finite set of training samples.

| Training Set Size $\lvert\mathcal{T}\rvert$ | 100 | 300 | 500 | 700 | 1000 |
|---|---|---|---|---|---|
| Abs. Rate $\mathcal{R}_\Sigma$ $(\frac{\mathrm{nat}}{\mathrm{Hz}})$ | 91.58 | 91.73 | 92.17 | 92.09 | 92.16 |
| Rel. Rate $\mathcal{R}_\Sigma^{\mathrm{rel.}}$ (%) | 93.26 | 93.41 | 93.85 | 93.78 | 93.84 |

In the following, the considerable generalization capabilities of the proposed GCN-WMMSE architecture are experimentally validated. We vary individual scenario parameters and study GCN-WMMSE networks (i) trained on samples with matching scenario configuration to the validation data, denoted by MT, and (ii) networks that are trained on samples at a defined pivot scenario configuration, denoted by PT.

*1) BS Power and UE Noise Power:* In Fig. 2, we sweep the BS power budget $P_{\mathrm{BS}}$ (leftmost) and UE noise power (center left) respectively. GCN-WMMSE MT achieves at least 93.64% relative WSR over the entire power budget range, outpeforming WMMSE RI TR substantially, particularly for high $P_{\mathrm{BS}}$, and achieving WSRs close to WMMSE RI while only requiring 7 layers. GCN-WMMSE PT generalizes well to scenarios which have a higher BS power budget compared to the scenarios in its training data. However, GCN-WMMSE PT loses performance when transferring to lower $P_{\mathrm{BS}}$ and achieves a lower WSR than WMMSE RI TR for a offset of $-15\,\mathrm{dB}$. When varying the receiver noise levels $\sigma^2_{\mathrm{UE}}$, both GCN-WMMSE MT and GCN-WMMSE PT achieve a WSR close to WMMSE RI and outperform WMMSE RI TR over the experimental range. GCN-WMMSE MT outperforms WMMSE RI in case of high SNR.

*2) Network Density:* In Fig. 2 (center right) the generalization w.r.t. to different BS distance $d_{\mathrm{BS}}$ is illustrated. Both the GCN-WMMSE MT and PT networks closely follow the WSR of WMMSE RI and achievea WSR above 93.76% relative to WMMSE50 for distances from $100\,\mathrm{m}$ to $300\,\mathrm{m}$. WMMSE RI TR is substantially outperformed.

*3) Number of UEs:* Changing numbers of UEs $I$ are considered in Fig. 2 (rightmost). GCN-WMMSE MT generalizes well and even significantly outperforms WMMSE50 for single user cells. Note that, as an exception, initializing the WMMSE algorithm with the MRC outperforms WMMSE RI in this case with 93.01%. GCN-WMMSE PT outperforms WMMSE RI TR by 13% for $I = 18$ and generalizes well to lower $I$, achieving $\mathcal{R}_\Sigma^{\mathrm{rel.}} = 96.09\%$. Thus, it is advantageous to train with the maximum number of UEs to maximize the transferability, disregarding the increased complexity of training the network.

*4) Array Dimensions:* Varying BS antenna dimension are studied for $I = 6$ UEs in Fig. 3 (left). GCN-WMMSE MT follows WMMSE RI for $M < 12$ and substantially outperforms WMMSE RI for $M = 12$. In this case, the classical WMMSE algorithm tends to find suboptimal beamformers with significant differences between individual UEs rates per scenario realization while GCN-WMMSE favors solutions with more equally distributed rates. However, GCN-WMMSE PT outperforms GCN-WMMSE MT for $M = 16$.

In Fig. 3 (right) the UE antenna array size is swept for $I = 9$ UEs. GCN-WMMSE PT and GCN-WMMSE MT achieve a significantly higher WSR than WMMSE RI TR with at least

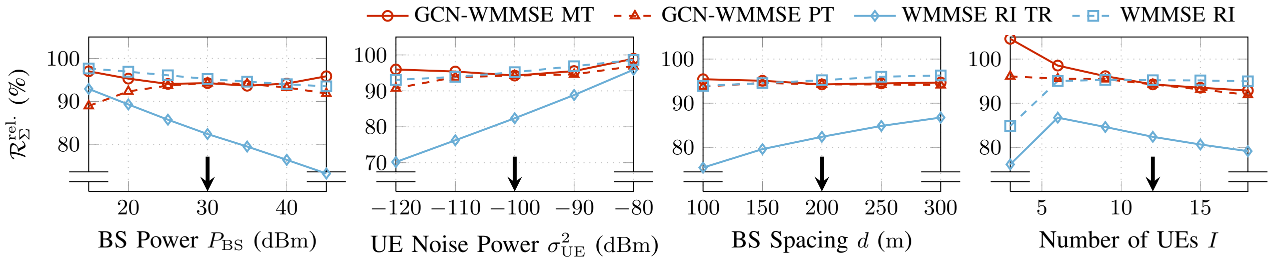

Figure 2. Generalization of GCN-WMMSE (proposed) w.r.t. different BS power budgets $P_k$ (leftmost), UE noise power $\sigma^2_{\text{UE}}$ (center left), BS separation distance $d_{\text{BS}}$ (center right) and number of UEs (rightmost). The arrows indicate the configuration for the training of GCN-WMMSE PT networks.

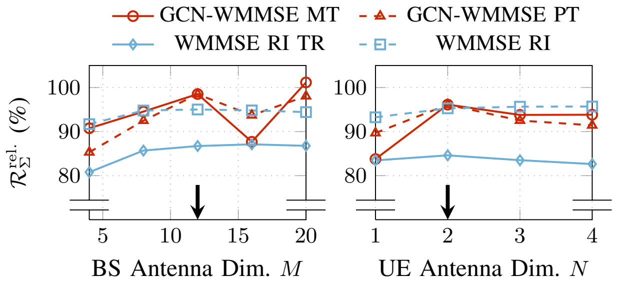

Figure 3. Impact of different numbers of antenna elements at the BSs (left) and at the UEs (right) on GCN-WMMSE (proposed). The arrows indicate the configuration for the training of GCN-WMMSE PT networks.

91.43% relative to WMMSE50 with exception of the MISO link case.

The observation that GCN-WMMSE MT is outperformed by GCN-WMMSE PT for scenario configurations with high degrees of freedom, i.e., when a single BS antenna array exceeds the total number of UE antennas, can be attributed to (2c) and (2d) attempting to solve the *near hard case* of a quadratically constrained quadratic program [22]. To avoid utilization of costly specialized iterative solvers, the ill-conditioned scenario samples can be removed in the backward pass [20]. The most effective solution, however, is to leverage the demonstrated generalization capabilities of GCN-WMMSE by training with scenarios with lower degrees of freedom.

## V. Conclusion

The deep network architecture GCN-WMMSE, which is based on unrolling the classical WMMSE algorithm, for beamforming in multicell MU-MIMO wireless networks is proposed. By reducing the number of required layers/iterations, GCN-WMMSE reduces the computational cost and communication overhead for distributed deployments compared to the WMMSE algorithm. At the same time, GCN-WMMSE exhibits excellent transferability and generalization performance across changing scenario configurations in most instances. Future investigations could address the more practical case of ergodic instead of instantaneous capacities.

## References

[1] 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, 2008. DOI: 10.1109/JSTSP.2007.914876.

[2] E. Björnson, G. Zheng, M. Bengtsson, and B. E. Ottersten, “Robust monotonic optimization framework for multicell MISO systems,” *CoRR*, vol. abs/1104.5240, 2011. arXiv: 1104.5240. [Online]. Available: http://arxiv.org/abs/1104.5240.

[3] Q. Shi, M. Razaviyayn, Z. 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. DOI: 10.1109/TSP.2011.2147784.

[4] 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, 2018. DOI: 10.1109/TSP.2018.2866382.

[5] V. Monga, Y. Li, and Y. C. Eldar, *Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing*, 2020. arXiv: 1912.10557 [eess.IV].

[6] K. Gregor and Y. LeCun, “Learning fast approximations of sparse coding,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML’10, Haifa, Israel: Omnipress, 2010, pp. 399–406, ISBN: 9781605589077.

[7] M. Zhu, T. Chang, and M. Hong, “Learning to beamform in heterogeneous massive MIMO networks,” *CoRR*, vol. abs/ 2011.03971, 2020. arXiv: 2011.03971. [Online]. Available: https://arxiv.org/abs/2011.03971.

[8] L. Pellaco, M. Bengtsson, and J. Jaldén, *Deep unfolding of the weighted MMSE beamforming algorithm*, 2020. arXiv: 2006.08448 [eess.SP].

[9] 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, 2021. DOI: 10.1109/TWC.2020.3033334.

[10] A. Chowdhury, G. Verma, C. Rao, A. Swami, and S. Segarra, *Unfolding WMMSE using graph neural networks for efficient power allocation*, 2021. arXiv: 2009.10812 [eess.SP].

[11] F. Gama, A. G. Marques, G. Leus, and A. Ribeiro, “Convolutional neural network architectures for signals supported on graphs,” *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1034–1049, 2019. DOI: 10.1109/TSP.2018.2887403.

[12] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[13] A. Chowdhury, G. Verma, C. Rao, A. Swami, and S. Segarra, *ML-aided power allocation for tactical MIMO*, 2021. arXiv: 2109.06992 [cs.IT].

[14] T. Liu, A. M. Tillmann, Y. Yang, Y. C. Eldar, and M. Pesavento, *Successive convex approximation for phase retrieval with dictionary learning*, 2021. arXiv: 2109.05646 [eess.SP].

[15] I. Loshchilov and F. Hutter, “Fixing weight decay regularization in adam,” *CoRR*, vol. abs/1711.05101, 2017. arXiv: 1711.05101. [Online]. Available: http://arxiv.org/abs/1711.05101.

[16] C. Trabelsi, O. Bilaniuk, D. Serdyuk, *et al.*, “Deep complex networks,” *CoRR*, vol. abs/1705.09792, 2017. arXiv: 1705.09792. [Online]. Available: http://arxiv.org/abs/1705.09792.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385.

[18] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, “Deep learning coordinated beamforming for highly-mobile millimeter wave systems,” *IEEE Access*, vol. 6, pp. 37 328–37 348, 2018. DOI: 10.1109/ACCESS.2018.2850226.

[19] H. Amann and J. Escher, *Analysis II* (Grundstudium Mathematik). Birkhäuser Basel, 2008, ISBN: 9783764371050. [Online]. Available: https://books.google.de/books?id=izgYzhnyacIC.

[20] L. Schynol and M. Pesavento, *Coordinated sum-rate maximization in multicell MU-MIMO with deep unrolling*, 2022. arXiv: 2202.10371 [eess.SP].

[21] D. Lopez-Perez, I. Guvenc, and X. Chu, “Mobility management challenges in 3GPP heterogeneous networks,” *IEEE Communications Magazine*, vol. 50, no. 12, pp. 70–78, 2012. DOI: 10.1109/MCOM.2012.6384454.

[22] M. Rojas, S. A. Santos, and D. C. Sorensen, “A New Matrix-Free Algorithm for the Large-Scale Trust-Region Subproblem,” *SIAM Journal on Optimization*, vol. 11, no. 3, pp. 611–646, Jan. 2001, ISSN: 1052-6234, 1095-7189. DOI: 10.1137/S105262349928887X.