

Frequency-domain digital predistortion for Massive MU-MIMO-OFDM Downlink

Yibo Wu^{*†}, Ulf Gustavsson^{*}, Mikko Valkama[‡], Alexandre Graell i Amat[†], and Henk Wymeersch[†]

^{*}Ericsson Research, Gothenburg, Sweden

[†]Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden

[‡]Department of Electrical Engineering, Tampere University, Tampere, Finland

Abstract—Digital predistortion (DPD) is a method commonly used to compensate for the nonlinear effects of power amplifiers (PAs). However, the computational complexity of most DPD algorithms becomes an issue in the downlink of massive multi-user (MU) multiple-input multiple-output (MIMO) orthogonal frequency division multiplexing (OFDM), where potentially up to several hundreds of PAs in the base station (BS) require linearization. In this paper, we propose a convolutional neural network (CNN)-based DPD in the frequency domain, taking place before the precoding, where the dimensionality of the signal space depends on the number of users, instead of the number of BS antennas. Simulation results on generalized memory polynomial (GMP)-based PAs show that the proposed CNN-based DPD can lead to very large complexity savings as the number of BS antenna increases at the expense of a small increase in power to achieve the same symbol error rate (SER).

I. INTRODUCTION

Massive multi-user (MU)-multiple-input multiple-output (MIMO) is one of the prominent technologies in fifth generation (5G) and beyond. With proper precoding techniques, up to several hundreds of antennas at the base station (BS) increases the capacity when serving many tens of user equipments (UEs) [1]. However, such large number of antennas poses a difficult linearization problem for the nonlinear power amplifiers (PAs). To meet the linearization performance of each PA, it is common to use digital predistortion (DPD) for each PA. However, the computational complexity increases linearly with the number of PAs and becomes unacceptably large for massive MU-MIMO [2], [3].

To tackle this complexity problem, recent works have shifted toward reducing the number of DPDs [2]–[5], so that each DPD linearizes more than one single PA. These methods unavoidably degrade the linearization performance as each PA has a different behavior due to variations in component characteristics. Another research direction goes toward frequency domain (FD) DPD [6], [7], where DPD is implemented before the OFDM inverse discrete Fourier transform (IDFT). The complexity of the DPD is reduced as the DPD sampling rate becomes equal to the symbol rate, instead of using oversampling as in conventional DPD methods. While [6] considers hybrid massive MIMO, [7] considers a fully digital MU-MIMO system with a neural network (NN)-based DPD, where the DPD is implemented

prior to the precoder. Thus, the complexity of the DPD is reduced as the dimensionality of the signal increases with the number of UEs instead of the number of BS antennas. However, this method requires several inverse discrete Fourier transforms (IDFTs) and discrete Fourier transforms (DFTs) as the NN-based DPD still operates in the time domain (TD). More problematically, the complexity cost of the method in [7] still increases with the number of BS antennas due to the additional precoding cost of the guard-band subcarriers. To address this complexity problem, an FD DPD model that only processes data subcarriers without IDFTs would be a possible solution. To the best of our knowledge, such a model has not yet been proposed.

In this work, we consider a massive MU-MIMO-OFDM system. We propose a convolutional neural network (CNN)-based FD DPD model that takes place before the OFDM IDFT and the precoder. We refer to this model as FD-CNN. Compared with conventional TD per-antenna DPD, the complexity cost of FD-CNN increases with the symbol rate and the number of UEs instead of with the much higher sampling rate and the number of BS antennas. Furthermore, to avoid high complexity cost caused by the large number of data subcarriers at the DPD input, we consider two-dimensional (2D) convolutional layers that can efficiently extract information from data subcarriers. We focus on the in-band metric SER as the out-of-band metric adjacent channel power ratio (ACPR) requirements have been considerably relaxed in millimeter wave [8]. Simulation results on different GMP-based PA models at each antenna and a frequency-selective Rayleigh fading channel show that the proposed FD-CNN DPD can provide very large complexity savings when the number of BS antennas grows at the expense of a small increase in power to achieve the same SER performance compared with state-of-the-art TD and FD DPDs.

Notation: Lowercase and uppercase boldface letters denote column vectors and matrices such as \mathbf{x} and \mathbf{X} ; \mathbf{x}^H denotes the Hermitian transpose of \mathbf{x} ; \mathbb{R} and \mathbb{C} denote real and complex numbers, respectively; x_n or $[\mathbf{x}]_n$ denote the n -th element of \mathbf{x} , and $\mathbf{x}_{n:n+k}$ denotes a vector consisting of the n -th to $(n+k)$ -th elements of \mathbf{x} ; the $B \times U$ all-zeros matrix and the $U \times U$ identity matrix are denoted by $\mathbf{0}_{B \times U}$ and \mathbf{I}_U , respectively; $\mathbb{E}\{x\}$ denotes the expectation of x .

This work was supported by the Swedish Foundation for Strategic Research (SSF), grant no. ID19-0021. The authors would like to thank Fan Jiang at Chalmers University of Technology for fruitful discussions.

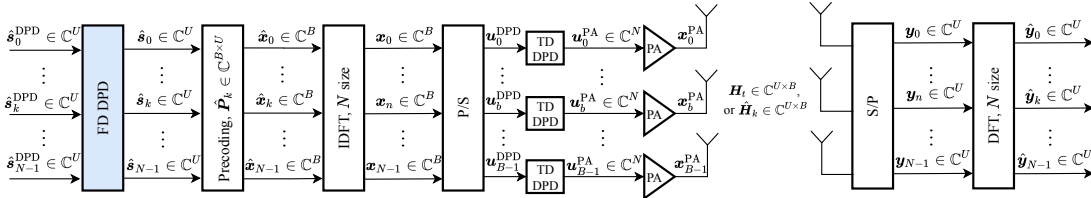


Fig. 1: System model of a massive MU-MIMO-OFDM downlink with nonlinear PAs in each RF chain of the BS. The conventional TD per-antenna DPDs take place before each PA while the proposed FD DPD operates before the precoder.

II. SYSTEM MODEL

A. System Model

We consider a massive MU-MIMO-OFDM downlink system model as shown in Fig. 1. The BS is equipped with B antennas and transmits messages to $U \ll B$ single-antenna UEs. Each OFDM symbol consists of N subcarriers with N_d data subcarriers and $N_g = N - N_d$ guard subcarriers. The subcarrier spacing is denoted by Δf . Accordingly, the sampling and symbol rate are defined as $f_s = N\Delta f$ and $f_d = N_d\Delta f$, respectively. The oversampling rate is $R = f_s/f_d = N/N_d$.

Specifically, let $\hat{\mathbf{s}}_k \in \mathbb{C}^U$ denote the symbol vector for U UEs at subcarrier k in the FD, where all symbols are generated independently from an M -QAM constellation. A precoder maps $\hat{\mathbf{s}}_k$ to a B -dimensional antenna array based on the available channel state information (CSI). We consider linear precoders due to their low complexity and good performance [9]. Given the data vector $\hat{\mathbf{s}}_k$ at subcarrier k , the FD output of a linear precoder, $\hat{\mathbf{x}}_k \in \mathbb{C}^B$, is obtained by [10]

$$\hat{\mathbf{x}}_k = \hat{\mathbf{P}}_k \hat{\mathbf{s}}_k, \quad (1)$$

where $\hat{\mathbf{P}}_k \in \mathbb{C}^{B \times U}$ denotes the FD precoding matrix for subcarrier k . We consider the precoding matrices $\hat{\mathbf{P}}_k = \mathbf{0}_{B \times U}$ for guard subcarriers. The precoded vectors $\hat{\mathbf{x}}_k$ are transformed to TD signals by N -size IDFTs. The TD signal vector at time sample n , $\mathbf{x}_n \in \mathbb{C}^B$, is given by

$$\mathbf{x}_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \hat{\mathbf{x}}_k \exp\left(jk \frac{2\pi}{N} n\right). \quad (2)$$

Without the impact of PA imperfections, the received signal at the U UEs in the TD at time sample n , $\mathbf{y}_n \in \mathbb{C}^U$, is given by [10]

$$\mathbf{y}_n = \sum_{t=0}^{T-1} \mathbf{H}_t \mathbf{x}_{n-t} + \mathbf{w}_n, \quad (3)$$

where $\mathbf{H}_t \in \mathbb{C}^{U \times B}$ denotes the TD channel matrix at the t -th tap with a total of T taps. The elements of the channel matrix are independently generated from a complex Gaussian distribution with zero mean and variance T^{-1} [10]. The channel is assumed to be block-constant with a coherence time of one OFDM symbol. In addition, $\mathbf{w}_n \sim \mathcal{CN}(\mathbf{0}_{U \times 1}, N_0 \mathbf{I}_U)$ denotes the additive white Gaussian noise (AWGN) vector at time sample n with the noise variance N_0 . Note that this yields a spatially white frequency-selective Rayleigh-fading channel with uniform power-delay profile [10].

Equivalently, (3) can be written in the FD as

$$\hat{\mathbf{y}}_k = \hat{\mathbf{H}}_k \hat{\mathbf{x}}_k + \hat{\mathbf{w}}_k, \quad (4)$$

where $\hat{\mathbf{H}}_k \in \mathbb{C}^{U \times B} = \sum_{t=0}^{T-1} \mathbf{H}_t \exp(-jk \frac{2\pi}{N} t)$ is the FD channel matrix at subcarrier k , $\hat{\mathbf{y}}_k \in \mathbb{C}^U = \sum_{n=0}^{N-1} \mathbf{y}_n \exp(-jk \frac{2\pi}{N} n)$ is the FD received signal at subcarrier k , and $\hat{\mathbf{w}}_k \in \mathbb{C}^U = \sum_{n=0}^{N-1} \mathbf{w}_n \exp(-jk \frac{2\pi}{N} n)$ is the FD AWGN noise at subcarrier k . Substituting (1) into (4), $\hat{\mathbf{y}}_k$ with precoder $\hat{\mathbf{P}}_k$ can be expressed as

$$\hat{\mathbf{y}}_k = \hat{\mathbf{H}}_k \hat{\mathbf{P}}_k \hat{\mathbf{s}}_k + \hat{\mathbf{w}}_k. \quad (5)$$

To minimize the MU interference, the pseudo-inverse of the channel matrix is used as the precoding matrix, which is known as the zero-forcing (ZF) precoding. The precoding matrix $\hat{\mathbf{P}}_k$ in (5) can be expressed as [9]

$$\hat{\mathbf{P}}_k = \alpha \hat{\mathbf{H}}_k^H (\hat{\mathbf{H}}_k \hat{\mathbf{H}}_k^H)^{-1}, \quad (6)$$

where α denotes the normalization factor to ensure the power constraint $\mathbb{E}\{\|\hat{\mathbf{x}}_k\|^2\} = P_T$ is met. The expectation is over the symbols of all UEs, and P_T denotes the average transmit power. With an imperfect knowledge of the CSI, the BS has an estimated TD channel matrix $\mathbf{H}_t^{\text{est}} = \sqrt{1-\eta} \mathbf{H}_t + \sqrt{\eta} \mathbf{H}_t^{\text{err}}$, where $\eta \in [0, 1]$ and $\mathbf{H}_t^{\text{err}} \sim \mathcal{CN}(\mathbf{0}_{U \times B}, \mathbf{I}_{U \times B})$ [10]. Thus, the FD channel matrix in (6) is replaced by a channel estimation $\hat{\mathbf{H}}_k^{\text{est}} = \sum_{t=0}^{T-1} \mathbf{H}_t^{\text{est}} \exp(-jk \frac{2\pi}{N} t)$.

B. PA nonlinearity

After the IDFT, the TD OFDM symbols are mapped to B radio frequency (RF) chains and sent to each TD DPD with input at the b -th RF chain, $\mathbf{u}_b^{\text{DPD}} \in \mathbb{C}^N = [[\mathbf{x}_0]_b, [\mathbf{x}_1]_b, \dots, [\mathbf{x}_{N-1}]_b]^T$. The DPD outputs are then converted to analog domain by the digital-to-analog converters (DACs). For simplicity, we assume ideal DACs with infinite-resolution. The PA input signal at the b -th RF chain is $\mathbf{u}_b^{\text{PA}} \in \mathbb{C}^N$.¹ The PA associated with the b -th BS antenna is represented by the nonlinear function $f_b^{\text{PA}}: \mathbb{C}^{L_1+1} \rightarrow \mathbb{C}$ with memory length L_1 , input $[\mathbf{u}_b^{\text{PA}}]_{n:n-L_1} = [[\mathbf{u}_b^{\text{PA}}]_n, \dots, [\mathbf{u}_b^{\text{PA}}]_{n-L_1}]^T \in \mathbb{C}^{L_1+1}$, and output $x_{b,n}^{\text{PA}}$. The input-output relation of the b -th PA at time sample n can be expressed as

$$x_{b,n}^{\text{PA}} = f_b^{\text{PA}}([\mathbf{u}_b^{\text{PA}}]_{n:n-L_1}). \quad (7)$$

Assuming an ideal PA with a linear behavior and with no need of DPD, (7) reduces to $x_{b,n}^{\text{PA}} = [\mathbf{u}_b^{\text{PA}}]_n = [\mathbf{u}_b^{\text{DPD}}]_n = [\mathbf{x}_n]_b$,

¹ In Section II-C, we describe the input-output relation of TD DPD and how the PA input \mathbf{u}_b^{PA} is obtained.

where $[x_n]_b$ is the b -th element of x_n . With the PA gain, the average PA output power is defined as $P_{\text{PA}} = \mathbb{E}\{|x_{b,n}^{\text{PA}}|\}$, where the expectation is over all time steps and BS antennas. In this case, no intercarrier interference (ICI) is introduced, and MU interference is eliminated by the ZF precoding, given perfect CSI. However, the PA nonlinearity creates distortion which in FD breaks the orthogonality of the subcarriers, and in spatial domain breaks the nulls created by the ZF precoding. Thus, the nonlinear PAs cause both ICI and MU interference.

C. Time-domain DPD

DPD is applied to compensate for the nonlinear behavior of the PA. Conventionally, DPD is implemented in TD before the PA and makes use of oversampling in order to cancel out adjacent channel power created by the PA nonlinearity. We represent the TD DPD associated with the b -th BS PA by a parametric model $f_{\theta_b^{\text{DPD}}} : \mathbb{C}^{L_2+1} \rightarrow \mathbb{C}$ with parameters θ_b^{DPD} and input memory length L_2 . Then, the mapped analog TD signal u_b^{DPD} is sent to the corresponding DPD as

$$u_{b,n}^{\text{PA}} = f_{\theta_b^{\text{DPD}}}([u_b^{\text{DPD}}]_{n:n-L_2}), \quad (8)$$

where the DPD output $u_{b,n}^{\text{PA}}$ corresponds to the PA input in (7). In the case of massive MU-MIMO, one can use one DPD for each PA. However, the total computational complexity cost of the DPD increases linearly with the number of PAs and antennas, which makes this approach highly impractical when several hundreds of PAs are deployed [2].

III. COMPLEXITY ANALYSIS OF DPD BENCHMARKS IN MASSIVE MU-MIMO

A. Measure of DPD Complexity

When working with DPD algorithm design, there are two complexity figures to consider, the *running complexity* and *training complexity*. The *running complexity* is the computational complexity needed to continuously run the DPD algorithm, which makes it the major overall contributor to complexity. Unlike the *training complexity*, which is an offline cost, the running complexity directly relates to the number of operations required for the inference of a certain number of samples [11].

There are several ways to measure the running complexity of DPD such as the Bachmann-Landau measure $\mathcal{O}(\cdot)$, running time, and number of parameters. These measures are either approximations or depend on the specific type of platform and thus are implementation-dependent. In contrast, the number of floating point operations (FLOPs) can accurately measure every addition, subtraction, and multiplication operation for both Volterra series-based models [12] and NNs [13], [14] for DPD. We therefore utilize the number of FLOPs to measure the running complexity of DPD, as this is the figure of merit which brings us closest to a fair comparison without going into implementation details. In this paper, we follow the same FLOP calculation as in [11, Table I]. In massive MU-MIMO, for a fair complexity comparison for different DPD schemes in the TD and FD, we calculate the total number of FLOPs required per UE for one OFDM symbol.

B. Time-domain Per-antenna GMP-based DPD

While there exist a large variety of published TD DPD models, we choose the commonly used GMP [15] for comparison because it has been shown to outperform many other models in terms of linearization performance versus complexity [11]. For simplicity, we refer to this TD per-antenna GMP method as TD-GMP.

To linearize every PA in massive MU-MIMO-OFDM, each RF chain deploys a GMP-based DPD with the same memory length L_3 , nonlinear order K , and cross-term length G . Given the TD signal, u_b , at the b -th RF chain as the input, the output of the GMP-based DPD associated with the b -th RF chain at time sample n can be expressed as

$$u_{b,n}^{\text{PA}} = \sum_{k=0}^{K-1} \sum_{l=0}^{L_3} a_{k,l} [u_b^{\text{DPD}}]_{n-l} | [u_b^{\text{DPD}}]_{n-l} |^k + \sum_{k=1}^{K-1} \sum_{l=0}^{L_3} \sum_{g=1}^G \left(b_{k,l,g} [u_b^{\text{DPD}}]_{n-l} | [u_b^{\text{DPD}}]_{n-l-g} |^k + c_{k,l,g} [u_b^{\text{DPD}}]_{n-l} | [u_b^{\text{DPD}}]_{n-l+g} |^k \right), \quad (9)$$

where $a_{k,l}$, $b_{k,l,g}$, and $c_{k,m,g}$ are complex-valued coefficients, which are identified using conventional mean squared error (MSE) estimation.

The number of FLOPs required for the GMP with each input sample is computed as [11]

$$C_{\text{Samp}}^{\text{TD-GMP}} = 8((L_3 + 1)(K + 2KG) - \frac{G(G+1)}{2}(K-1)) + 10 + 2K + 2(K-1)G + 2K \min(G, L_3) \quad (10)$$

Now consider that we use GMP-based DPD for each antenna in a massive MU-MIMO antenna system. The number of FLOPs required per UE for one OFDM symbol is calculated as

$$C^{\text{TD-GMP}} = C_{\text{Samp}}^{\text{TD-GMP}} NB/U. \quad (11)$$

We note that the number of FLOPs for the TD GMP-based DPD grows linearly with the number of BS antennas because every PA requires one DPD. This causes a complexity problem in massive MU-MIMO with hundreds of antennas.

C. FD Neural Network-based DPD

To address the complexity problem of TD DPD in massive MU-MIMO, the authors in [7] proposed a NN-based DPD, which operates in the FD prior to the precoder. In this paper, we refer to this FD NN-based DPD model as FD-NN. To form the input of the FD-NN, each UE stream is first converted to the TD. Specifically, given the symbol vector for the u -th UE in the FD, $\tilde{s}_u^{\text{DPD}} \in \mathbb{C}^N = [[\hat{s}_0^{\text{DPD}}]_u, [\hat{s}_1^{\text{DPD}}]_u, \dots, [\hat{s}_{N-1}^{\text{DPD}}]_u]^T$, the TD symbol vector for the u -th UE, $s_u^{\text{DPD}} \in \mathbb{C}^N$, is given by an IDFT

$$s_u^{\text{DPD}} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} [\tilde{s}_u^{\text{DPD}}]_k \exp(jk \frac{2\pi}{N} n) = \mathbf{F}_N^H \tilde{s}_u^{\text{DPD}}, \quad (12)$$

where \mathbf{F}_N denotes the unitary $N \times N$ DFT matrix. Then, using tapped delay lines with memory length L_4 , the input

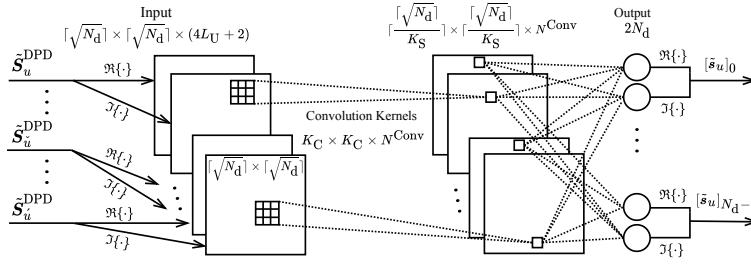


Fig. 2: Structure of the proposed FD-CNN. For the u -th UE, the input of FD-CNN is formed by a selection of L_U UE symbol vectors, $\{\tilde{s}_u^{\text{DPD}}, \tilde{s}_u^{\text{DPD}}, \dots, \tilde{s}_u^{\text{DPD}}\}$. The output is the predistorted symbol vector of u -th UE, \tilde{s}_u .

signal of the first layer consists of each TD UE symbol vector, $[\mathbf{s}_u^{\text{DPD}}]_{n:n-L_4} = [[\mathbf{s}_u^{\text{DPD}}]_n, [\mathbf{s}_u^{\text{DPD}}]_{n-1}, \dots, [\mathbf{s}_u^{\text{DPD}}]_{n-L_4}]^T \in \mathbb{C}^{L_4+1}$. In total, the input signal of the first layer FD-NN is given by

$$\mathbf{s}_{\text{in}}^{\text{FD-NN}} = [[\mathbf{s}_0^{\text{DPD}}]_{n:n-L_4}^T, [\mathbf{s}_1^{\text{DPD}}]_{n:n-L_4}^T, \dots, [\mathbf{s}_{U-1}^{\text{DPD}}]_{n:n-L_4}^T]^T, \quad (13)$$

which is then decomposed into real and imaginary parts connecting with their own neurons. Thus, this yields $2U(L_4+1)$ neurons for the first layer. All layers in the FD-NN are fully connected. There are K^{NN} hidden layers, each with D neurons and a nonlinear activation function ReLU. The number of neurons at the output layer is $2U$. The predistorted TD signal for each UE is then converted back to the FD by DFTs before being sent to the precoder.

In this method, the additional IDFTs and DFTs introduce an extra complexity cost. More importantly, the guard subcarriers of the predistorted signal are not empty anymore due to the DFTs, which leads to extra complexity cost due to the precoding. For each UE, the number of FLOPs required for the FD-NN per OFDM symbol can be computed as

$$C^{\text{FD-NN}} = \frac{1}{U} \left(\underbrace{N(4U(L_3+1)D + 2(K^{\text{NN}}-1)D^2 + 4DU)}_{\text{NN}} + \underbrace{10UN \log_2 N}_{\text{(IDFTs)}} + \underbrace{(6+2)N_g UB}_{\text{Extra precoding}} \right), \quad (14)$$

where the 6 and 2 in the extra precoding part are for complex-number multiplication and addition, respectively. Considering the fast Fourier transform, the complexity of a N -size (ID)FT is approximated by $5N \log_2 N$ FLOPs with $\frac{N}{2} \log_2 N$ complex-number multiplications and $N \log_2 N$ complex-number additions, and there are U IDFTs and U DFTs. We note that due to extra precoding cost, the number of FLOPs for FD-NN still increases with the number of BS antennas, which dominates the complexity when hundreds of antennas are deployed in massive MU-MIMO.

IV. PROPOSED FD CONVOLUTIONAL NEURAL NETWORK

A. Structure of the FD-CNN

The structure of the proposed FD-CNN is shown in Fig. 2. FD-CNN works in the FD with the input of UE symbol vectors. To efficiently extract appropriate information from the FD UE symbol vectors without using IDFTs as [6], we propose to

use 2D convolutional layers as the first layer, which save complexity compared with a fully connected layer. To facilitate the 2D convolution, the FD UE symbol vector $\tilde{s}_u^{\text{DPD}} \in \mathbb{C}^N$ is converted to a matrix $\tilde{\mathbf{S}}_u^{\text{DPD}} \in \mathbb{C}^{\lceil \sqrt{N_d} \rceil \times \lceil \sqrt{N_d} \rceil}$, where only N_d data subcarriers are involved to save complexity. Here $\lceil \cdot \rceil$ denotes the ceiling function and zero-padding is used if $\lceil \sqrt{N_d} \rceil > \sqrt{N_d}$. To offer more complexity saving, for the u -th UE, an arbitrary selection of L_U UE symbol vectors from all U UEs, $\{\tilde{s}_u^{\text{DPD}}, \tilde{s}_u^{\text{DPD}}, \dots, \tilde{s}_u^{\text{DPD}}\}$, including \tilde{s}_u^{DPD} , are converted to matrices. In total, these matrices form the input of the FD-CNN as a tensor, $\tilde{\mathbf{S}}_u \in \mathbb{C}^{\lceil \sqrt{N_d} \rceil \times \lceil \sqrt{N_d} \rceil \times L_U}$,

$$\tilde{\mathbf{S}}_u^{\text{DPD}} = [\tilde{\mathbf{S}}_u^{\text{DPD}}, \dots, \tilde{\mathbf{S}}_u^{\text{DPD}}, \dots, \tilde{\mathbf{S}}_u^{\text{DPD}}]. \quad (15)$$

Then, each complex-valued UE symbol matrix in $\tilde{\mathbf{S}}_u^{\text{DPD}}$ is decomposed into real and imaginary matrices with the same dimension as $\tilde{\mathbf{S}}_u^{\text{DPD}}$. In total, there are $2L_U$ real-valued UE symbol matrices, which are convoluted by N^{Conv} convolutional kernels with the same kernel size K_C and stride length K_S . This yields an output tensor, $\tilde{\mathbf{S}}_u^{\text{Conv}} \in \mathbb{R}^{\lceil \frac{\lceil \sqrt{N_d} \rceil}{K_S} \rceil \times \lceil \frac{\lceil \sqrt{N_d} \rceil}{K_S} \rceil \times N^{\text{Conv}}}$, where zero-padding is also considered if $\lceil \frac{\lceil \sqrt{N_d} \rceil}{K_S} \rceil > \frac{\lceil \sqrt{N_d} \rceil}{K_S}$.

$\tilde{\mathbf{S}}_u^{\text{Conv}}$ is passed through an element-wise activation function $\sigma(\cdot)$ and then flatten into a column vector, $\tilde{s}_u^{\text{Flat}} \in \mathbb{R}^{\lceil \frac{\lceil \sqrt{N_d} \rceil}{K_S} \rceil \lceil \frac{\lceil \sqrt{N_d} \rceil}{K_S} \rceil N^{\text{Conv}}}$. After that $\tilde{s}_u^{\text{Flat}}$ is fully connected with the linear output layer with $2N_d$ neurons, which correspond to the real and imaginary parts of the predistorted symbol vector at the u -th UE, $\tilde{s}_u \in \mathbb{C}^{N_d}$.

B. Complexity Analysis

We split the complexity of the proposed FD-CNN into the convolution layer part and the fully-connected layer part. For the complexity of the convolution layer part, each K_C -size 2D convolution requires $(2K_C^2 - 1)$ FLOPs consisting of K_C^2 real-valued multiplications and $(K_C^2 - 1)$ real-valued additions. This 2D convolution operates on $2L_U$ real-valued UE symbol matrices with stride length K_S , and in total there are N^{Conv} convolutional kernels. Thus, the number of FLOPs for the convolution layer part can be expressed as

$$C_{\text{Conv}}^{\text{FD-CNN}} = 2N^{\text{Conv}} L_U (2K_C^2 - 1) \left(\frac{\lceil \sqrt{N_d} \rceil}{K_S} \right)^2. \quad (16)$$

²The vector-to-matrix conversion follows a contiguous order so that the convolution kernels can efficiently extract information from adjacent subcarriers.

The number of FLOPs for the fully-connected layer part can be expressed as

$$C_{\text{FC}}^{\text{FD-CNN}} = 8N_d L_U \left(\frac{\lceil \sqrt{N_d} \rceil}{K_S} \right)^2. \quad (17)$$

In total, the number of FLOPs required for the FD-CNN with one OFDM symbol per UE is calculated as

$$C^{\text{FD-CNN}} = C_{\text{Conv}}^{\text{FD-CNN}} + C_{\text{FC}}^{\text{FD-CNN}} \quad (18)$$

We note that $C^{\text{FD-CNN}}$ is sensitive to the number of data subcarriers N_d because $C_{\text{FC}}^{\text{FD-CNN}}$ is in the magnitude of N_d^2 . Compared with the complexity for TD-GMP and FD-NN in (11) and (14), we note that the complexity of FD-CNN does not depend on the number of BS antennas, B , which naturally saves complexity when B is large in massive MU-MIMO.

V. NUMERICAL RESULTS

A. Simulation Setup

1) *Parameters:* We consider a set of simulated PAs using the GMP model in (9) with nonlinear order $K = 7$, memory length $M = 3$, and cross-term length $G = 1$. These parameters are estimated using real measurements from the RF WebLab using a 100 MHz OFDM signal [16]. Using the estimated PA parameters as the mean of a Gaussian distribution, the PA parameters of each antenna are drawn with variance 0.01. For instance, given the original GMP-based PA coefficient $\alpha_{k,l}$, the coefficient at b -th antenna $\alpha_{k,l}^b \sim \mathcal{N}(\alpha_{k,l}, 0.01\alpha_{k,l})$. Note the GMP-based PA coefficients are fixed once being generated. The saturation point and measurement noise standard deviation of each PA are 24.02 V (≈ 37.6 dBm with a 50 Ω load impedance) and 0.053 V, respectively.

We select a 50 MHz OFDM setup in the operating band n257 from the 3GPP 5G standardization [17, Table 5.3.2-1 and 5.3.5-1] with subcarrier spacing $\Delta f = 120$ kHz, OFDM symbol length $N = 4096$, number of data subcarriers $N_d = 384$, and $M = 256$ QAM modulation. Each channel realization has $T = 10$ taps. The AWGN noise variance is $N_0 = 1$ Watts/Hz. We assume the BS has imperfect CSI with $\eta = 0.001$, which affects the ZF precoding.

For the benchmark of TD-GMP, we consider the same GMP parameter setup for each DPD as the PA model, namely $K = 7$, $L_3 = 3$, and $G = 1$. For the benchmark of FD-NN, we consider the same memory length $L_4 = 3$. Other parameters are set the same as in [7] with $N_{\text{FD-NN}} = 1$ hidden layer and $D = 15$ neurons. For the proposed FD-CNN, we set $\lceil \sqrt{N_d} \rceil = 20$, $L_U = 1$, $K_C = 3$, and $K_S = 1$, and $N^{\text{Conv}} = 2$. The activation function $\sigma(\cdot)$ in the convolution layer is the ReLU. To preserve the same input and output dimension of the convolutional layer, zero padding is used.

2) *DPD Coefficients Identification:* For the TD-GMP, the coefficients of each GMP-based DPD are estimated using the indirect learning architecture (ILA) at the output of each PA [18], where the least squares algorithm is used to minimize the MSE between the PA input signal and post-distorter output signal. For the FD-NN, since all blocks in the given

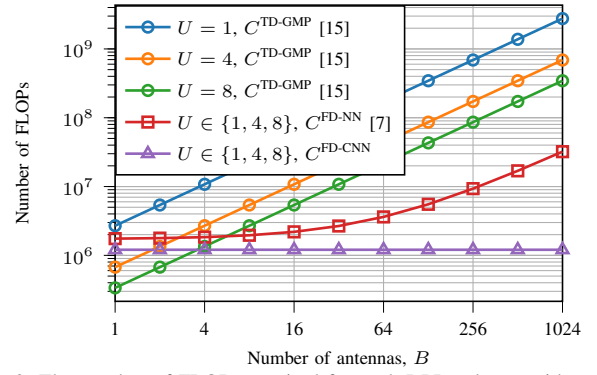


Fig. 3: The number of FLOPs required for each DPD scheme with one UE and one OFDM symbol versus the number of BS antennas, B .

communication system are simulated and differentiable, we simply utilize supervised learning with a loss function at the receiver side to minimize the MSE between received symbols \hat{y}_k and transmitted symbols \hat{s}_k for all data subcarriers. This symbol-based criterion for DPD optimization has been shown to achieve similar SER performance in [19]. Note new channel realization is generated for every training mini-batch. The FD-NN is then trained till convergence using the gradient descent optimizer Adam [20] with a learning rate 0.001.

For the proposed FD-CNN, we choose the same supervised learning, loss function, and optimizer Adam as the training for FD-NN. Each mini-batch consists of 10 OFDM symbols. In practical applications, it is straightforward to use a training method similar to the one in [7]. This training method creates the training data of the NN output by converting back PA output signals through MP-based DPDs and the pseudo-inverse of precoders. This requires dedicated feedback paths to collect each PA output at the transmitter. Alternatively, one can use over-the-air method that utilize an observation receiver for data acquisition as in [19], [21].

B. Simulation results

1) *Complexity Versus Number of BS Antennas:* Fig. 3 shows the number of FLOPs for one OFDM symbol per UE as a function of the number of BS antennas, $B \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$, for DPD schemes of TD-GMP, FD-NN, and the proposed FD-CNN. We also plot results for different number of UEs, $U \in \{1, 4, 8\}$.

We note that, as B increases, the number of FLOPs for TD-GMP grows linearly as each antenna is associated with one TD DPD. While the number of FLOPs for FD-NN grows gently for $B < 16$, it eventually increases linearly with B as the extra cost of guard-band precoding starts to dominate. However, the number of FLOPs for the proposed FD-CNN is completely invariant to B , which would greatly benefit very large antenna systems. Compared with FD-NN and TD-GMP, FD-CNN saves around $2.2\times$ and $8.9\times$ FLOPs for $B = 32$ and $U = 8$, respectively, and these savings can be further increased up to $26\times$ and $286\times$ when B increases to 1024.

We also observe that as U increases, the number of FLOPs per UE decreases accordingly for TD-GMP, while this number

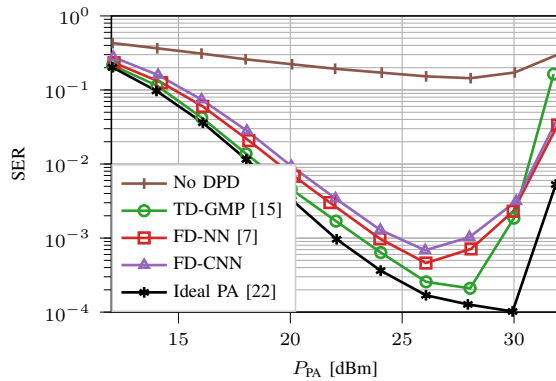


Fig. 4: SER as a function of the average PA output power, P_{PA} .

for FD-NN and FD-CNN is the same.

2) *SER Versus Average PA Output Power*: Now we fix $B = 32$ and $U = 8$ for different DPD schemes from Fig. 3 and plot the SER results as a function of the average PA output power, P_{PA} , in Fig. 4. Specifically, we plot the case of ideal linear-clipping PA [22]. The linear-clipping PA has a linear behavior before the clipping region, which has the minimum distortion that any DPDs can achieve. The number of QAM symbols used to calculate the SER is around 10^6 .

We note that the SER curves exhibit two different behaviors depending on the nonlinear and clipping regions of the PAs. On one hand, as P_{PA} is below 25 dBm, most PAs in the massive MU-MIMO system operate in their linear/nonlinear regions, so the SER of all DPD cases improve dramatically with P_{PA} since the signal-to-noise ratio (SNR) at the UEs are increasing accordingly. The proposed FD-CNN requires $2.2\times$ and $8.9\times$ less complexity than TD-GMP and FD-NN, respectively, at the expense of some increase in the required power to achieve a certain SER. For example, it requires 0.8 dBm and 1.8 dBm more PA output power to achieve an SER 10^{-3} . On the other hand, when $P_{PA} > 25$ dBm, the clipping effect of the PA brings unrecoverable distortions, which lead to a quick SER degradation for all DPD cases. The SER gap between all the DPD cases and the linear-clipping case may come from some residual distortions due to irreversible nonlinearity.

VI. CONCLUSION

We proposed a novel DPD method in the FD based on CNNs to address the rising complexity issue of linearization for large antenna systems in massive MU-MIMO-OFDM. The complexity of the proposed FD-CNN DPD is in the magnitude of the number of UEs and the symbol rate, which naturally avoids high complexity as the number of BS antennas increases. Simulation results on a MU-MIMO-OFDM system with different behavior PAs on each RF chain show that the FD-CNN can save $2.2\times$ and $8.9\times$ number of FLOPs with 32 BS antennas and 8 UEs, compared with FD NN-based and TD per-antenna GMP-based DPDs, respectively. This saving can be further improved as the number of BS antennas increases. Furthermore, the SERs degradation of FD-CNN is minor.

REFERENCES

- [1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
- [2] X. Liu, W. Chen, L. Chen, F. M. Ghannouchi, and Z. Feng, "Linearization for hybrid beamforming array utilizing embedded over-the-air diversity feedbacks," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 12, pp. 5235–5248, Oct. 2019.
- [3] X. Wang, Y. Li, C. Yu, W. Hong, and A. Zhu, "Digital predistortion of 5G massive MIMO wireless transmitters based on indirect identification of power amplifier behavior with OTA tests," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 1, pp. 316–328, Jan. 2020.
- [4] L. Liu, W. Chen, L. Ma, and H. Sun, "Single-PA-feedback digital predistortion for beamforming MIMO transmitter," in *IEEE Int. Conf. Microw. Millim. Wave Technol.*, vol. 2, Jun. 2016, pp. 573–575.
- [5] H. Yan and D. Cabric, "Digital predistortion for hybrid precoding architecture in millimeter-wave massive MIMO systems," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2017, pp. 3479–3483.
- [6] A. Brihuega, L. Anttila, and M. Valkama, "Frequency-domain digital predistortion for OFDM," *IEEE Microw. Wireless Compon. Lett.*, vol. 31, no. 6, pp. 816–818, Mar. 2021.
- [7] C. Tarver, A. Balasoukas-Slimining, C. Studer, and J. R. Cavallaro, "Virtual DPD neural network predistortion for OFDM-based MU-massive MIMO," in *Asilomar Conf. Signals, Syst. Comput.*, Oct. 2021, pp. 376–380.
- [8] A. Brihuega, M. Abdelaziz, L. Anttila, M. Turunen, M. Allén, T. Eriksson, and M. Valkama, "Piecewise digital predistortion for mmwave active antenna arrays: Algorithms and measurements," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 9, pp. 4000–4017, 2020.
- [9] A. Wiesel, Y. C. Eldar, and S. Shamai, "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, Sep. 2008.
- [10] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Linear precoding with low-resolution DACs for massive MU-MIMO-OFDM downlink," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1595–1609, Jan. 2019.
- [11] A. S. Tehrani, H. Cao, S. Afsardoost, T. Eriksson, M. Isaksson, and C. Fager, "A comparative analysis of the complexity/accuracy tradeoff in power amplifier behavioral models," *IEEE Trans. Microw. Theory Techn.*, vol. 58, no. 6, pp. 1510–1520, Jun. 2010.
- [12] J. Moon and B. Kim, "Enhanced hammerstein behavioral model for broadband wireless transmitters," *IEEE Trans. Microw. Theory Techn.*, vol. 59, no. 4, pp. 924–933, Apr. 2011.
- [13] T. Liu, S. Boumaiza, and F. M. Ghannouchi, "Dynamic behavioral modeling of 3G power amplifiers using real-valued time-delay neural networks," *IEEE Trans. Microw. Theory Techn.*, vol. 52, no. 3, pp. 1025–1033, Mar. 2004.
- [14] Y. Wu, U. Gustavsson, A. Graell i Amat, and H. Wymeersch, "Low complexity joint impairment mitigation of IQ modulator and PA using neural networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 54–64, Nov. 2021.
- [15] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt, and J. Pastalan, "A generalized memory polynomial model for digital predistortion of RF power amplifiers," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3852–3860, Oct. 2006.
- [16] P. N. Landin, S. Gustafsson, C. Fager, and T. Eriksson, "WebLab: A web-based setup for PA digital predistortion and characterization [application notes]," *IEEE Microw. Mag.*, vol. 16, no. 1, pp. 138–140, Feb. 2015.
- [17] 3GPP, "NR; User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone," TS 38.101-2, Jul. 2020, version 16.4.0.
- [18] C. Eun and E. J. Powers, "A new Volterra predistorter based on the indirect learning architecture," *IEEE Trans. Signal Process.*, vol. 45, no. 1, pp. 223–227, Jan. 1997.
- [19] Y. Wu, J. Song, C. Häger, U. Gustavsson, A. Graell i Amat, and H. Wymeersch, "Symbol-based over-the-air digital predistortion using reinforcement learning," *arXiv preprint arXiv:2111.11923*, 2021.
- [20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR'15*, May. 2015.
- [21] K. Hausmair, U. Gustavsson, C. Fager, and T. Eriksson, "Modeling and linearization of multi-antenna transmitters using over-the-air measurements," in *Proc. IEEE ISCAS'18*. IEEE, May. 2018, pp. 1–4.
- [22] J. Chani-Cahuana, C. Fager, and T. Eriksson, "Lower bound for the normalized mean square error in power amplifier linearization," *IEEE Microw. Wireless Compon. Lett.*, vol. 28, no. 5, pp. 425–427, May. 2018.