Dataset Reduction for Neural Network Based Digital Predistorters under Strong Nonlinearities

David López-Bueno  
PHYCOM Department, CTD Division  
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA)  
Castelldefels (Barcelona), Spain  
david.lopez@cttc.es

Pere L. Gilabert  
Department of Signal Theory and Communications  
Universitat Politècnica de Catalunya (UPC)  
Castelldefels (Barcelona), Spain  
p.lgilabert@tsc.upc.edu

Gabriel Montoro  
Department of Signal Theory and Communications  
Universitat Politècnica de Catalunya (UPC)  
Castelldefels (Barcelona), Spain  
gabriel.montoro@upc.edu

Abstract— The artificial neural networks (ANN) are gaining momentum in the digital predistorters (DPD) thanks to their inherently good approximation capabilities. Under strong or complex power amplifier nonlinearities, the size of the ANN can increase and lead to long training periods which are unaffordable in fast-changing waveform scenarios like those proposed for 5G or 6G. In this work we combine the orthogonal matching pursuit technique together with dataset length reduction methods, to significantly shorten the ANN DPD coefficients update time.

Keywords— power amplifiers, predistortion, iterative learning control, machine learning algorithms, artificial neural networks

I. INTRODUCTION

Thanks to the digital predistorter (DPD) the power amplifier (PA) can be operated more efficiently when handling wideband high peak-to-average power ratio (PAPR), and meet the error vector magnitude or the adjacent channel power ratio (ACPR) requirements. Adaptive DPDs are nowadays needed to deal with changing environmental conditions and waveform scenarios and their complexity is increasing with the deployment of massive antenna transmitters. When modeling strongly nonlinear systems, the ANN DPD can offer global approximation capabilities in contrast to the inherent local approximating properties of the polynomial-based DPD. However, when the complexity of the ANN DPD is increased to model complex nonlinear phenomena, and requires large input datasets, the training time is severely increased. There is not a universal procedure to set up the best ANN for a given application but several works found in literature provide indications on the suitable architecture, activation functions and backpropagation algorithms or hyperparameter configurations [1]. Such works mainly refer to the indirect learning (IL) architecture shown in Fig.1 (left) that models the inverse response of the PA during the NN training and replicates the coefficients calculated in the forward path (NN inference). In this work, we target the input dataset reduction in the direct learning architecture (DL) shown in Fig. 1 (right), which models the counteracting distortion signal to be added to the original input signal to compensate for the PA distortion.

II. NEURAL NETWORK ADAPTIVE DIGITAL PREDISTORTER

In this work, we consider the real-valued time delayed feedforward NN in Fig. 2. According to the DL scheme, the input IQ pair is defined as the real and imaginary part of the NN input vector $\mathbf{u}[n]$, and the output IQ pair will be an estimation of the real part and imaginary part of the NN expected output $\hat{\mathbf{y}}[n]$. The input IQ pair is further processed by a group of data functions (DF in Fig. 2) to provide richer basis to the NN and improve the nonlinear modeling. The ANN DPD can benefit from the injection of envelope dependent terms (i.e. $|I_{in}| + jQ_{in}k$ with $k \in \mathbb{N}$) [2], angle dependent terms (i.e. $(\text{atan2}(Q_{in}/I_{in}))k$ with $k \in \mathbb{N}$), or long term memory effects modeling components like those described in [3]. To enable dynamic nonlinear system identification, a linear time invariant (LTI) system built with tapped delay lines is added. In Fig. 2, we have $N$ input dataset basis functions resulting from applying the LTI system to the IQ input signals and to the DF outputs. The input dataset length relies on the length of the IQ data being considered for modeling. The NN features one input layer (with $N$ inputs), two hidden layers with $F$ and $S$ neurons, and an output layer with $2$ neurons. The neurons in the hidden layers employ the $\theta^1(\cdot)$ and $\theta^2(\cdot)$ nonlinear activation functions and a pure linear activation function is used for the output layer. This NN architecture has $NF + FS + 2S$ weights and $F + S + 2$ biases. In Fig. 2, $w_{ij}^k$ is the $j$th weight belonging to the $i$th neuron in the $k$th layer. Similarly, $\theta^k_i$ is the bias belonging to the $i$th neuron in the $k$th layer. In the forward pass we have that

$$I_{out}[n] = \sum_{k=1}^S w_{1k}^3 \theta^2 \left( \sum_{j=1}^F w_{kj}^2\theta^1 \left( \sum_{i=0}^N w_{ij}^1 \varphi_i[n] + \theta_1^1 \right) + \theta_2^1 \right) + \theta_3^1 \right) + \theta_4^1 \right) + \theta_5^1 \right) + \theta_6^1 \right)$$(1)

$$Q_{out}[n] = \sum_{k=1}^S w_{2k}^3 \theta^2 \left( \sum_{j=1}^F w_{kj}^2\theta^1 \left( \sum_{i=0}^N w_{ij}^1 \varphi_i[n] + \theta_1^1 \right) + \theta_2^1 \right) + \theta_3^1 \right) + \theta_4^1 \right) + \theta_5^1 \right) + \theta_6^1 \right)$$

© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. DOI: 10.1109/PAWR51852.2021.9375506
The Levenberg-Marquardt (LM) backpropagation algorithm is used in our ANN to calculate the NN coefficients by minimizing the cost function $E$, where

$$
E = \frac{1}{2K} \sum_{n=1}^{K} \left( (I_{out}[n] - \hat{I}_{out}[n])^2 + (Q_{out}[n] - \hat{Q}_{out}[n])^2 \right)
$$

for every input data batch of length $K$. This forward-backward process is repeated until the desired modeling performance is achieved or the NN fails in generalization.

When going backward, the LM modification to the Gauss-Newton method [4] minimizes $E$ with respect to a parameter $c$ with the weights and biases, which is updated as

$$
c^{n+1} = c^n - (J^T J + \beta I)^{-1} J^T e
$$

where

$$
c = [w_{11} \ldots w_{FN} \theta_1^1 \ldots \theta_2^1 \ldots w_{12}^3 \ldots w_{2S}^3 \theta_1^2 \theta_2^2],
$$

where $I$ is the identity matrix, $\beta$ is a learning rate parameter and $J$ is the Jacobian matrix calculated over the error vector $e$ with respect to $c$ as

$$
J = \begin{bmatrix}
\frac{\partial e[1]}{\partial w_{11}} & \frac{\partial e[1]}{\partial w_{12}} & \ldots & \frac{\partial e[1]}{\partial w_{FN}} & \frac{\partial e[1]}{\partial \theta_1^2} & \frac{\partial e[1]}{\partial \theta_2^2} \\
\frac{\partial e[2]}{\partial w_{11}} & \frac{\partial e[2]}{\partial w_{12}} & \ldots & \frac{\partial e[2]}{\partial w_{FN}} & \frac{\partial e[2]}{\partial \theta_1^2} & \frac{\partial e[2]}{\partial \theta_2^2} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
\frac{\partial e[K]}{\partial w_{11}} & \frac{\partial e[K]}{\partial w_{12}} & \ldots & \frac{\partial e[K]}{\partial w_{FN}} & \frac{\partial e[K]}{\partial \theta_1^2} & \frac{\partial e[K]}{\partial \theta_2^2}
\end{bmatrix}
$$

III. DATASET REDUCTION

In our ANN, we define the $L \times N$ input dataset matrix $U = (\varphi[1], \ldots, \varphi[n], \ldots, \varphi[L])$ with $n = 1, \ldots, L$ and $\varphi^T[n] = (\varphi_1[n], \ldots, \varphi_i[n], \ldots, \varphi_N[n])$ with $i = 1, \ldots, N$. We will reduce the two $U$ matrix dimensions. By reducing the number of basis ($N$ in Fig. 2) both the number of coefficients at the first hidden layer and the training time are reduced. The latter will be further impacted by applying dataset length reduction techniques.

A. Dataset basis reduction

The Orthogonal matching pursuit (OMP) technique can be used as an a priori offline study to determine the best basis functions to linearize the PA. This process is applied once and the basis selection will be kept for a given scenario (i.e. fixed PA output power, waveform and bandwidth). The support vector containing the indices of the basis functions in $U$ is defined as $P^{(m)}$ where $m_{max} = N$ is the maximum number of basis under study. As shown in the procedure of Algorithm 1,

Algorithm 1 Orthogonal matching pursuit for DL ANN-AFPD

1: procedure OMP($x', u$)
2: \hspace{1em} initialization:
3: \hspace{2em} $x'^{(0)} = x' - \hat{x}'^{(0)}$; with $x'^{(0)} = 0$
4: \hspace{2em} $P^{(0)} = []$
5: \hspace{2em} for $m = 1$ to $m_{max}$ do
6: \hspace{3em} $l^{(m)} = \arg \min_{1 \leq i \leq N} \|u[i] - U[j]w_i^{(m)}\|^2$ \hspace{0.5em} $w_i^{(m)} = U[j][m]$
7: \hspace{3em} $p^{(m)} = p^{(m-1)} \cup \{ \hat{p}^{(m-1)} \}$
8: \hspace{3em} $w_{\hat{p}^{(m)}} = (U_{\hat{p}^{(m)}})^{-1} U_{\hat{p}^{(m)}} x'$
9: \hspace{3em} $x'^{(m)} = x' - \hat{x}'^{(m)}$
10: \hspace{3em} end for
11: \hspace{1em} return $P^{(m_{max})}$
12: end procedure
at every iteration of the OMP search, $P^{(m)}$ is fed with the indices that better contribute to minimize the residual modeling error and are therefore sorted according to their relevance. A Bayesian information criterion (BIC) can be used to determine the optimum number of basis and the support set can be truncated for a given pruning factor (i.e., $U_{P}^{(1:mmax/p(factor)})$). In [5], the authors propose the doubly OMP (DOMP) scheme that adds a Gram-Schmidt orthogonalization step into the basis selection procedure which contributes to better normalized mean square error (NMSE) performance.

B. Dataset length reduction

Several sample selection methods (SSM) have been reported in literature [6] to reduce the computational complexity of DPDs. In this work, we introduce a data batch selection (BS) method. The BS method takes the relevant waveform data batches in $u$ depending on the characteristics of the PA output $y$. Two variants of the BS method are presented: consecutive BS (CBS) and sparse BS (SBS). The procedure has the following steps: 1) $u$ and $y$ are divided into a number of data batches $N_b$, which is an integer multiple number of the waveform length reduction factor $R$. 2) With CBS, a sliding window of consecutive samples with length $L_s = N_b/L$ is shifted over $u$ and $y$, with $L/N_b$ shifting step, to calculate both the NMSE, the adjacent channel error power ratio (ACEPR) and the $u$ signal mean power. When using SBS, all the metrics are calculated individually for each of the $N_b$ batches (there is not a wider sliding window). 3) The BS algorithm is in charge of sorting the expanded batch indices (CBS) or the sparse batch indices (SBS) in order of importance according to featuring simultaneously the worst NMSE and ACEPR values, and the highest mean power. When using CBS, all the metrics are calculated individually for each of the $N_b$ batches (there is not a wider sliding window). In [6], the authors propose the doubly OMP (DOMP) scheme that adds a Gram-Schmidt orthogonalization step into the basis selection procedure which contributes to better normalized mean square error (NMSE) performance.

IV. EXPERIMENTAL TEST SETUP AND RESULTS

The dataset reduction techniques have been benchmarked and validated with the Matlab controlled hardware test bench shown in Fig. 4. A GaN HEMT class J PA which is operated at 875 MHz RF frequency and 28 dBm mean output power with 80 MHz bandwidth signals featuring 13-14 dB PAPR. The DPD neural network has been implemented in Matlab. The dataset has 105 inputs/basis each with 737280 samples. These 105 inputs belong to the I and Q components and the augmented products until the sixth power, considering up to 12 consecutive delays, p and involving 8 DPD iterations, and better performance than only using the best BS dataset length reduction method (i.e. see SBS160 vs CBS40+DOMP4 in Table I). These results indicate that by using the input dataset reduction techniques the NN training time can be dramatically reduced and potentially be in the order of magnitude of the classical polynomial DPD adaptation time.

![Fig. 3. MeS, CBS and SBS input dataset length reduction methods (i.e. for 20-reduction factor).](image)

![Fig. 4. Remotely accessible digital linearization test setup.](image)
The direct learning ANN DPD can outperform the classical DPD schemes at the cost of a potentially unaffordable training time for adaptive scenarios. OMP combined with dataset length reduction strategies is effective for selection of NN inputs, NN coefficient reduction, and to greatly cut the NN training time for DPD adaptation while fulfilling signal quality requirements. Application to MIMO DPD schemes and further work on joint design of dataset length reduction methods and feature selection techniques needs to be conducted to reach the best trade-off between NN performance and the DPD update time.

ACKNOWLEDGMENT

The authors would like to thank Dr J.A. Garcia, Dr M.N. Ruiz, Dr J.R. Pérez Cisneros, and D. Vegas from the Universidad de Cantabria for the donation of the class-J PA. This work was supported in part by the Spanish Government (Ministerio de Ciencia, Innovación y Universidades) and the Fondo Europeo de Desarrollo Regional (FEDER) under grants RTI2018-099841-B-I00 and TEC2017-83343-C4-2-R, and in part by the Generalitat de Catalunya under grants 2017 SGR 891 and 2017 SGR 813.

REFERENCES