A Multi-Rate Approach for Nonlinear Pre-Distortion Using End-to-End Deep Learning in IM-DD Systems

A Multi-Rate approach for Nonlinear Pre-distortion using End-to-end Deep Learning in IM-DD systems

Leonardo Minelli, Fabrizio Forghieri *Senior Member, IEEE*, Antonino Nespola, Stefano Straullu and Roberto Gaudino *Senior Member, IEEE*

***Abstract*—Modern intra-data center (IDC) interconnects leverage robust and low-cost intensity modulation (IM) and direct detection (DD) optical links, based on multimode fibers (MMFs) and vertical-cavity surface-emitting lasers (VCSELs). Current solutions, based on on-off keying (OOK) modulations, reach up to 25-50 Gbps per lane over nearly 100 meters. The actual target for IDCs is to increase VCSEL-MMF links capacity up to 100 Gbps, using PAM-4 on the same devices. To counteract the consequent linear and nonlinear distortions affecting the transmitted signals, an effective solution is to exploit digital signal processing (DSP). In this manuscript, we propose a novel method to optimize a nonlinear artificial neural network (ANN) digital pre-distorter (DPD), based on End-to-end (E2E) learning, that, trained jointly with a Feed-Forward Equalizer (FFE), fulfills physical amplitude constraints and handles different ratio between the sampling rates incurring along with an optical IM-DD system. We indeed propose an E2E ANN system operating simultaneously at different sampling frequencies. We moreover propose in our training method a substitution to the time-domain injection of the receiver noise in the system with an additive regularization term in the FFE gradient loss. We experimentally show the advantages of our proposed DPD comparing the bit error rate (BER) performance against the same scenario without DPD. We assess the gain in terms of Gross Bit Rate and Optical Path Loss (OPL), at given BER targets, for different fiber lengths.**

***Index Terms*—IM-DD systems, nonlinear equalization, Artificial Neural Networks, VCSEL, Multi Mode Fibers, Intra Data-Center Interconnection.**

I. Introduction

NEXT-generation telecommunication systems are requested to fulfill an increasing number of data-driven services, that will leverage technologies such as 5G, Internet of Things, and Cloud Computing. In front of the consequent traffic demand growth, data centers need to upgrade their communication links accordingly, to exchange data at ever-higher rates. Modern intra Data Center Inter Connects (DCI) exploit Intensity Modulation-Direct Detection (IM-DD) On-Off Keying (OOK) modulated optical links. Around 50 % of these DCI are still using Multi-Mode Fibers (MMF) [1], together with Vertical-Cavity Surface-Emitting Lasers (VCSEL) due to their low-cost chip manufacturing and high power efficiency. While a long-term goal could be the introduction of coherent technology in DCI, as well as the switch to Single-Mode Fibers (SMF), the current focus is on increasing the capacity of the deployed links beyond 100 Gbps per lane, using the same physical devices. Therefore, moving to multi-level formats such as Pulse Amplitude Modulation (PAM) and exploiting Digital Signaling Processing (DSP) techniques could be an effective solution. Among possible DSP solutions, nonlinear equalization allows counteracting bandwidth limitations and nonlinear distortions, that severely impair the signals when transmitted at the aforementioned data rates. Consequently, in the past years, considerable effort has been spent on designing equalization technologies, especially for PAM-4 modulated signals [2]. In particular, several types of nonlinear equalizers have been recently investigated either at the receiver (RX) side as post-equalizers [3] [4] [5] [6], or at the transmitter (TX) side as Digital Pre-Distorters (DPD) [7] [9] [10]. The latter approach is favored in optical short-reach links since nonlinear DSP algorithms tend to be easier to implement at TX rather than RX. The optimization of a nonlinear DPD to be applied on an optical IM-DD link requires particular attention to several critical factors: indeed, the laser (i.e., VCSEL) nonlinear effects and the severe bandwidth limitations must be addressed jointly to physical constraints such as the limited VCSEL input dynamics and the different ratios between Baud Rate, Digital-to-Analog (DAC) sampling frequency, and Analog-to-Digital (ADC) sampling frequency. Moreover, the DPD must mitigate impairments caused by stochastic disturbances affecting the system which cannot be merely modeled as White Gaussian Noise.

In this article we propose a solution able to effectively apply nonlinear DPD jointly with a RX linear post-equalizer on an experimental VCSEL-MMF IM-DD optical link characterized by these physical constraints. We illustrate a novel optimization method, based on End-to-end (E2E) learning [12] of an optical link, to train a Neural Network based DPD, able to encode PAM-M symbols in a sequence with arbitrary sampling frequency that satisfies the signal amplitude swing constraints imposed by the system in which it is applied. The novelty of our approach resides in designing an E2E ANN-based architecture that supports an online optimization method performing forward and backward propagation at different sampling rates, handling generic non-integer (but rationale) ratios between them. Moreover, we propose as a novelty a procedure to model the noise affecting the experimental transmission system, in order to simulate its effect in the E2E optimization in an analytical fashion. Physically, the main noise sources in our experimental system are the electrical noise generated by the Trans-Impedance Amplifier (TIA) that follows the photodiode, and the Relative Intensity Noise (RIN) of the VCSEL (the relevance of this noise in the

L. Minelli and R. Gaudino are with the Department of Electronics and Telecommunications, Politecnico di Torino, Torino, Italy.

A. Nespola and S. Straullu are with LINKS Foundation, Torino, Italy.

F. Forghieri is with CISCO Photonics Italy, Vimercate, Italy.

system can be seen in Fig. 6). We indeed characterize it using experimental measurements, and we introduced it as an additive regularization term during the FFE coefficients's gradient update. We apply and test our proposed DPD training algorithm on an experimental setup, demonstrating the validity of our optimized DPDs in a wide range of conditions.

The following Sections in this manuscript are organized as follows: in Section II we review the approaches to the nonlinear predistortion, illustrating our method and the constraints imposed by the considered transmission system. In Section III we present our designed multi-rate E2E system, explaining its behavior when forwarding signals and backpropagating the gradients. In Section IV we examine the application of our method in an experimental setup, assessing the performances of the trained DPDs in the considered scenario. Finally, in Section V we discuss the implications of the designed method and the obtained results, drawing some conclusions.

II. Digital Pre-distortion in an optical IM-DD link

A nonlinear Digital Pre-Distorter (DPD) is a device meant to fully or partially pre-compensate for the distortions affecting a signal during the propagation over a communication link. More specifically, in an optical VCSEL-MMF IM-DD system a DPD is supposed to counteract bandwidth limitations arising from several devices at TX and RX when transmitting at very high data rates, impairments caused by chromatic and modal dispersion along with the MMF, and nonlinear distortions, caused mainly by the VCSEL. The pre-compensation at TX can provide advantages to post-compensation at RX: for instance, compensating bandwidth limitations using a post-equalizer leads to an enhancement of the receiver noise, due to the high-pass filter effect synthesized by the device after the noise is added to the signal. Moreover, a DSP nonlinear DPD at TX tends to be easier to implement with respect to a nonlinear post-equalizer at RX (e.g., by leveraging structures resembling Look-Up Tables (LUT)).

A. Approaches to nonlinear predistortion in optical links

Nonlinear DPD for optical links have been developed using several technologies, such as Volterra-equalizer [11] [10], Artificial Neural Networks [13], or LUT [7] [8]. Different solutions have been proposed in the literature for optimizing DPD coefficients.

One of the best known is the Indirect Learning Approach (ILA): this method is based on performing a single optimization step, in which the DPD is trained to learn the inverse model of the transmission system. The DPD thus estimates a post-distortion function that is assumed to be equal to the required pre-distortion function [16].

The other main technique, called Direct Learning Approach (DLA), divides the DPD optimization into two steps. In the first phase, a differentiable architecture (e.g., a Neural Network) is trained to get a direct surrogate model (also called "digital twin") of the transmission system. In the second phase, the obtained channel model's input is attached to the output of the DPD, forming a unique cascaded structure. The latter can be then treated as a unique ANN, whose dependancies along its computational graph can be exploited to optimize through gradient-based methods the DPD coefficients. The cascaded ANN is thus trained without updating the channel model coefficients to optimize the DPD [16].

Recently, End-to-end (E2E) deep learning approaches have been studied for both coherent and IM-DD optical links, to optimize transceivers's DSP blocks [14] [15] [17] [18] [19] [20] [21] [22]. E2E deep learning of a communication system consists of modeling the transmitter (i.e., the DPD in our case), the channel and the receiver (i.e., the post-equalizer) as a unique autoencoder ANN [12] (or "E2E system"). Equivalently to the DLA, the E2E approach exploits a differentiable digital twin of the channel to optimize the DPD while training the autoencoder. In this case, the DPD gets optimized jointly with the post-equalizer, while the channel coefficients are not updated (as in DLA). This approach is supposed to be preferable to optimize a DPD with respect to other block-wise optimization techniques, such as ILA and DLA. Infact, with an E2E approach the DPD is aware while trained that the impairments affecting the signals are also compensated at the receiver side. A predistorter trained in this way has thus the potential to lead to optimal end-to-end performance [14]. One of the crucial points of the E2E learning approach is obtaining a reliable differentiable digital twin of the channel. As indeed optical links are affected by deterministic and stochastic impairments (i.e., the "noise"), these must be properly modeled to get a faithful representation of the system. Several solution were proposed, such as digital twins based physical models of the noise perturbations [14] [15] [22], or channel models based on Generative Adversarial Networks (GAN) [23].

In situations where it is difficult or unfeasible to retrieve a channel differentiable model, alternative model-free optimization methods of the E2E autoencoders were proposed [24] [25], training for instance DPD using Reinforcement Learning [26].

B. The proposed DPD optimization method

In this paper, we investigate the joint use of the DLA and the E2E approaches to optimize an Artificial Neural Network (ANN) based nonlinear DPD. We assume a fully-digital implementation based on DSP, using DAC at Tx and ADC at Rx: therefore, we do not consider analog equalization implementations. In the first step, we directly model the transmission channel through an ANN, which we then exploit as a channel derivable model in an E2E system, through which we then train an ANN-based DPD at TX together with a post-equalizer at Rx. We study the application of this method in a VCSEL-MMF IM-DD optical link, where the DPD is applied on the current directly modulating the laser. Since modern IM-DD systems leverage using Feed-Forward Equalizers (FFE) at RX as reference for applications such as TDECQ test [27], we selected this linear post-equalizer for the E2E optimization.

In the considered transmission scenario, we must take into account two constraints, related to the physical limitations of the transmission system:

- The modulation swing of the pre-distorted signal cannot exceed the smallest among the two following dynamics:
  - The Digital-to-Analog (DAC) converter output dynamic (e.g., the maximum peak-to-peak voltage it can provide)
  - VCSEL's input current dynamics imposed for safety reasons: it cannot be inverted with respect to the laser's bias current, neither it can exceed a certain amplitude.
- The predistorted signal cannot have a sampling rate higher than the Digital-to-Analog Converter (DAC) sampling frequency: for an arbitrary Baud Rate the samples-per-symbol (sps) ratio tends thus to be fractional.

Obtaining a pre-distorted signal that jointly fulfills these limitations is not straightforward. For instance, a conventional 2 sps DPD could indeed satisfy the dynamic constraint by imposing an output layer that saturates the signal to the given modulation bounds. However, if the Baud Rate R_s is not an integer multiple of the DAC sampling frequency f_{DAC}, the resulting 2 sps pre-distorted signal would break this constraint after being resampled to f_{DAC}: this operation indeed produces new samples that are not necessarily bound to the desired dynamics. The solution we propose in this paper is a DPD whose output already matches f_{DAC} for any R_s, having thus an arbitrary sps ratio and natively satisfying the amplitude constraint. In this work, we created such DPD by starting from the design of an E2E autoencoder as an FIR-based Neural Network (FIRNN) [28]. We choose this specific type of ANN in order to properly deal with a transmission system characterized by significant memory effects (i.e., bandwidth limitations) present either in the channel, at the TX or at the RX. The peculiarity of this neural model resides indeed in its "synapses": while in a Feed-Forward Neural Network (FFNN) they represent a static multiplication operator, in the FIRNN this gets extended to an FIR filter. FIRNN can thus be exploited to implement advanced nonlinear cascades of FIR filters, representing the optical link signal processing chain. A more detailed explanation is provided in the Appendix. In this paper, we modeled the E2E system as a FIRNN, where different sampling rates are involved along with its structure. In order to support this, we extended the online optimization algorithm proposed in [28] to back-propagate in parallel gradients related to different sampling frequencies. End-to-end systems involving different sampling frequencies have been already approached in a simplified way, with integer ratios (e.g. 2 or 8) among rates [15]. In this paper we propose a multi-rate End-to-end system that extends and generalizes this feature, by allowing to handle generic non-integer (but rationale) ratios among the involved sampling rates. Moreover, we provide a detailed explanation and comment on the gradient backpropagation when signal resampling occurs, as we will illustrate in the next Section.

III. The Multi-rate end-to-end system

A. Transmitted signal forward propagation

To train DPD through a multi-rate end-to-end learning approach, we implemented a structure whose scheme is shown in Fig. 1.In the notation used in the next pages, we refer to each discrete time-index of a signal propagated in the system as n_{f_s}, where in subscript we have the associated sampling frequency f_s (expressed in Hertz).

During the E2E optimization, a PAM-M sequence of symbols $a[n_{R_s}]$ (where R_s is the Baud rate), gets forwarded through the three main building blocks of the E2E architecture:

1) Multi-rate Transmitter (tx): This device is the nonlinear DPD. It is composed of a polyphase resampler cascaded to a FIRNN with 1 hidden ReLU layer (more info in Section III-B). It processes the transmitted symbols to obtain a pre-distorted sequence $x[n_{f_{DAC}}]$, whose rate f_{DAC} is the sampling frequency of the DAC. The output activation function of the DPD is modeled as a hard-limiter, according to the usual equation:

$$DAC(x) = \begin{cases} -A & x < -A \\ A & x > A \\ x & otherwise \end{cases} \tag{1}$$

where A is the given amplitude constraint. This function plays actually a key role in our algorithm: it allows indeed to natively take into account the DAC+VCSEL amplitude constraint during the backpropagation algorithm.

2) Multi-rate Channel (ch): This device is topologically almost identical to tx (i.e., a resampler plus a ReLU FIRNN without output activation function). It is the differentiable surrogate model of the analog transmission channel (i.e., from the DAC output to the ADC input). It is optimized through direct modeling using noiseless experimental measurements, using successive transmitted samples as training examples and received samples as labels. Trained with this approach, it becomes a digital twin of the channel, that can be exploited to simulate the deterministic impairments affecting the transmitted signal, such as bandwidth limitations and VCSEL nonlinear effects. In our E2E system modelization, the noise affecting the system is assumed to be introduced later at the RX side (details in Sec. III-A3). The Multi-rate channel accepts in input the signal $x[n_{f_{DAC}}]$, producing as output the distorted signal $z[n_{f_{ADC}}]$ whose rate f_{ADC} is the sampling frequency of the Analog-to-Digital Converter (ADC).

3) Multi-rate Receiver (rx): This device is composed of a poly-phase resampler cascaded to a linear Feed Forward Equalizer (FFE), internally running at 2 sps: conceptually, an FFE can be interpreted as a FIRNN with only 1 neuron without any nonlinear activation functions. The device resamples $z[n_{f_{ADC}}]$ obtaining the signal $c[n_{2\cdot R_s}]$, which has the sampling rate of the FFE.

To simulate the stochastic impairments affecting the transmission system, $c[n_{2\cdot R_s}]$ is summed to the *receiver noise*. This disturbance is modeled as a White Gaussian Noise (WGN) n_{in} filtered by a properly designed FIR filter. The latter is modeled such that its frequency response follows the Power Spectral Density (PSD) of the true experimental noise retrieved from the experimental measurements. The resulting noisy signal is finally equalized through the FFE. As we will explain in Section III-C, the noise is not added in the time domain, but semi-analytically during the E2E optimization.

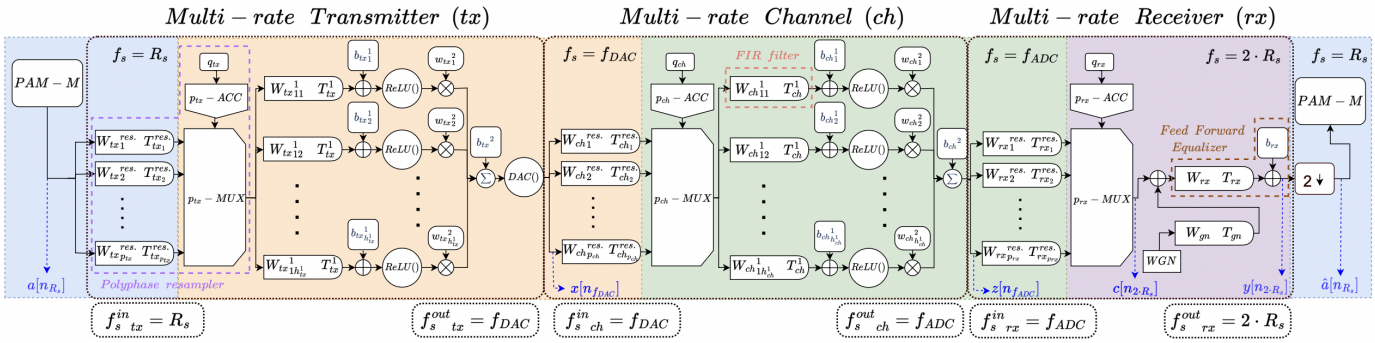

Fig. 1. Scheme representing the Multi-rate end-to-end system. Each main building block contains a polyphase resampler, composed by a bank of p_d FIR filters, a p_d-multiplexer ($p_d - MUX$) and a modulo-p_d accumulator ($p_d - ACC$), where d is the device among tx, ch and rx. The output of each MUX thus is given in input to a FIRNN. Building blocks operate at the sample rate indicated in the upper right corners of the colored regions. Blue: $f_s = R_s$; Orange: $f_s = f_{DAC}$; Green: $f_s = f_{ADC}$; Purple: $f_s = 2 \cdot R_s$.

At the end of the system the FFE outputs the sequence $y[n_{2 \cdot R_s}]$. The latter is decimated by a factor 2 at the decision instant, to get an estimation of the transmitted sequence $\hat{a}[n_{R_s}] = y[n_{R_s}]$.

Each multi-rate device d works in parallel at two operatives sampling frequencies, ${f_s^{in}}_d\ [Hz]$ and ${f_s^{out}}_d\ [Hz]$. The change of rate is performed by the poly-phase resampler, which conceptually performs the following operations:

1) It upsamples the signal by an integer factor p_d, by inserting $p_d - 1$ zeros between two consecutive input samples.
2) It applies a low-pass anti-aliasing FIR filter to the upsampled signal.
3) It decimates the filtered signal samples by an integer factor q_d, thus producing a discrete output signal whose sampling rate is p_d/q_d times the input one.

As illustrated in Fig.1, Steps 1 and 2 are implemented through a bank of p_d FIR filters (i.e., the *polyphase* components), whose taps are defined by the vector:

$$W_{d_k}^{res.} = [w_{d_{(k)}}^{res.}, w_{d_{(p_d+k)}}^{res.}, \cdots, w_{d_{(T_{d_k}^{res.} \cdot p_d + k)}}^{res.}]^\top \quad (2)$$

where $w_{d_i}^{res.}$ is the i-th tap of the resampler's anti-aliasing FIR-filter ($i = k, p_d + k, \cdots, T_{d_k}^{res.} \cdot p_d + k$) and $T_{d_k}^{res.}$ is the number of taps of the k-th polyphase component ($k = 1, 2, \cdots, p_d$). Step 3 is then implemented through the use of a multiplexer with p_d input ports, whose selection is shifted of a factor q_d by a modulo-p_d accumulator. In the E2E system, all the p_d and q_d values satisfy the 2 following conditions:

$$\frac{p_d}{q_d} = \frac{{f_s^{out}}_d}{{f_s^{in}}_d} \quad (3)$$

$$HCF(p_d, q_d) = 1 \quad (4)$$

where $HCF()$ computes the Highest Common Factor. The above equation implies thus that all the ratio between the E2E architecture sampling rates must be a rationale number. For instance, in our experimental setup, transmitting a R_s=50 GBaud signal with f_{DAC}= 92 Gsample/s and f_{ADC}= 200 Gsample/s implies $p_{tx} = 46$, $q_{tx} = 25$, $p_{ch} = 50$, $q_{tx} = 23$, $p_{rx} = 1$ and $q_{tx} = 2$.

B. Loss gradient backward propagation

The Multi-rate end-to-end system previously illustrated, if all the rates were identical and no noise was injected, would be conceptually a cascade of three FIRNN (being thus a FIRNN itself), having the following characteristics:

- They all have input size and output size equal to 1. Moreover, the first layer is composed of only 1 synapse, that is the polyphase resampler anti-aliasing filter.
- The ANN cascaded to the polyphase resampler is a FIRNN characterized by "static" branches (the synapses are single-taps FIR-filters with no delays) after its first layer. This ANN can be seen equivalently as a FFNN whose inputs are entries of a tap delay line.
- Neurons activation functions can be the following types: $DAC()$, $ReLU()$ or $none$.

Therefore, with the previous assumptions, the overall E2E system could be optimized as a generic FIRNN, updating only the DPD and FFE coefficients.

Neglecting in this Section the receiver noise presence for simplicity, we show that the FIRNN back-propagation algorithm (illustrated in Appendix) can be extended to work in a multi-rate scenario. According to the different rates in the E2E system, the loss gradient musts change during its temporal back-propagation its sampling frequency, as an actual signal itself. There are two types of rate changes along with the network:

The first kind is the change from the rate of the decision instants (R_s) to the rate of the FFE ($2 \cdot R_s$). Looking at Fig. 1 as reference (i.e., left-end and right-end sides), we define the per-symbol squared error loss $e[n_{R_s}]^2$ as:

$$e[n_{R_s}]^2 = |a[n_{R_s} - S] - \hat{a}[n_{R_s}]|^2 \quad (5)$$

where S is the delay (in symbols) introduced by the E2E system. It must be noticed that this constitues a stochastic implementation of the Mean Squared Error (MSE), which is the most commonly used loss function for optimizing adaptive FFEs. Its derivative with respect to $\hat{a}[n_{R_s}]$ can be computed as follows:

$$\delta_{\hat{a}}[n_{R_s}] = \frac{\partial e[n_{R_s}]^2}{\partial \hat{a}[n_{R_s}]} = -2e[n_{R_s}] \quad (6)$$

The consequent loss gradient with respect to the fractionally spaced FFE output, defined as $\delta_y[n_{2\cdot R_s}]$, can be computed as follows:

$$\delta_y[n_{2\cdot R_s}] = \frac{\partial e[n_{2\cdot R_s}]^2}{\partial y[n_{2\cdot R_s}]} = \begin{cases} \delta_{\hat{a}}[n_{R_s}] & n_{2\cdot R_s} = 2 \cdot n_{R_s} \\ 0 & otherwise \end{cases} \quad (7)$$

The latter gradient indeed assumes non-null values only for the instants in which a variation in $y[n_{2\cdot R_s}]$ would actually affect $e[n_{R_s}]^2$: when this isn't the case, $\delta_y[n_{2\cdot R_s}]$ is consequently null. According then to [28], the gradient with respect to the output of the *rx* resampler $\delta_c[n_{2\cdot R_s}]$ is computed as follows:

$$\delta_c[n_{2\cdot R_s}] = \Delta_y[n_{2\cdot R_s}]^\top \cdot W_{rx} \quad (8)$$

where :

$$\Delta_y[n_{2\cdot R_s}] = \\ [\delta_y[n_{2\cdot R_s}], \delta_y[(n+1)_{2\cdot R_s}], \cdots, \delta_y[(n+T_{rx}-1)_{2\cdot R_s}]]^\top \quad (9)$$

From a DSP point of view, the signal $\delta_c[n_{2\cdot R_s}]$ can be interpreted as a interpolated version of the signal $\delta_{\hat{a}}[n_{R_s}]$. The latter is indeed upsampled by a factor 2, with the addition of zeros between the samples. Then it passes through of a backward FIR "interpolating" filter, whose taps are the entries of the vector W_{rx}.

The second kind is related to the backpropagation through each resampler. Taking for instance $\delta_c[n_{2\cdot R_s}]$:

1) It gets upsampled by the addition of $q_{rx} - 1$ zeros between two consecutive samples: a variation on the anti-aliasing FIR filter output would affect the resampler output only on the decimated instants.
2) It gets backward filtered through the FIR anti-aliasing FIR filter: the result of this operation is the gradient $\delta_z[n_{p_{rx}\cdot f_{ADC}}]$
3) It gets decimated by a factor p_{rx}. Only 1 sample every p_{rx} of $\delta_z[n_{p_{rx}\cdot f_{ADC}}]$ is indeed actually related to $z[n_{f_{ADC}}]$:

$$\delta_z[n_{p_{rx}\cdot f_{ADC}}] = \begin{cases} \frac{\partial e[n_{2\cdot R_s}]^2}{\partial z[n_{f_{ADC}}]} & n_{p_{rx}\cdot f_{ADC}} = p_{rx} \cdot n_{f_{ADC}} \\ \frac{\partial e[n_{2\cdot R_s}]^2}{\partial \mathbf{0}} & otherwise \end{cases} \quad (10)$$

where $\mathbf{0}$ is any padding zero added for upsampling $z[n_{f_{ADC}}]$ during forward propagation.

It must be noticed that steps 1 and 2 are equivalent to the operations performed to compute $\delta_c[n_{2\cdot R_s}]$. Moreover, we can observe that any gradient backpropagated through a resampler is subject to a sequence of operations that is equivalent to the 3 operations performed during forward propagation. Therefore, a polyphase structure (symmetric to the forward resampler) can be adopted for implementing the resampler backward propagation as well. The adopted resampler can be then viewed as a more complex FIR-based synapse, with a backward structure dual to the forward one, equivalently to what derived in [28]. Conceptually, the E2E multi-rate could be therefore viewed as a particular FIRNN, where some synapses are made by polyphase resamplers rather than FIR filters, still preserving the symmetry between signal forward propagation and gradient backward propagation [28].

In practice, the E2E optimization we developed consists thus of simulating a real-time digital system in which it is applied an *online training algorithm*: an optimization where a predictor (i.e., DPD and FFE) exploits data (i.e., gradients) that become available in a sequential order (due to the causality in the temporal backpropagation) to update its coefficients at each step. In the E2E multi-rate FIRNN system indeed, at each discrete time step n_{f_s} (related to any sampling rate f_s in the system), the forward and backward filters operating at f_s are incremented, and the involved coefficients that must be optimized are subject to a Stochastic Gradient Descent (SGD) update.

C. Analytical introduction of the receiver noise

In the multi-rate end-to-end training process, the DPD and the FFE are jointly optimized through successive SGD updates using an MSE criterion. However, it can be observed that the MSE gradient computed with respect to the DPD coefficients is independent of the receiver noise since this is injected into the E2E system after all the nonlinear stages, as shown in Fig.1. Therefore, the DPD online SGD training can benefit from not injecting directly the noise into the E2E system, to thus backpropagate "clean" realizations of the MSE gradient: in this way the DPD gradients estimations have a reduced variance, leading to a more stable optimization. We can then introduce the noise effect as an additive regularization term [29] in the FFE Loss gradient, for a correct E2E optimization. The Mean Squared Error (MSE) at the output of the receiver is indeed described as follows:

$$MSE_{(a,y)} = \mathbb{E}\left[\left|a[n_{R_s} - S] - y[n_{R_s}]\right|^2\right] \quad (11)$$

By defining $\hat{y}[n_{R_s}]$ as the noiseless output of the receiver, obtained when only $c[n_{R_s}]$ is given in input to the FFE, the MSE can be splitted into two components:

$$MSE_{(a,y)} = MSE_{(a,\hat{y})} + \sigma^2_{n_{in}} \sum_{k=1}^{T_{(gn*rx)}} \left|w_{(gn*rx)_k}\right|^2 \quad (12)$$

where $w_{(gn*rx)}$ are the taps of the discrete linear convolution between the noise filter and the FFE. These terms can be expressed as entries of a vector $W_{(gn*rx)}$ with length $T_{gn*rx} = T_{gn} + T_{rx} - 1$. According to Eq. 12, the MSE gradient with respect to the FFE taps can be thus defined as:

$$\nabla_{W_{rx}} MSE_{(a,y)} = \nabla_{W_{rx}} MSE_{(a,\hat{y})} + 2 \cdot \sigma^2_{n_{in}} W^\top_{(gn*rx)} \cdot \bar{\mathbf{W}}_{gn} \quad (13)$$

where the matrix $\bar{\mathbf{W}}_{gn} \in \mathbb{R}^{T_{(gn*rx)} \times T_{rx}}$ is defined as follows:

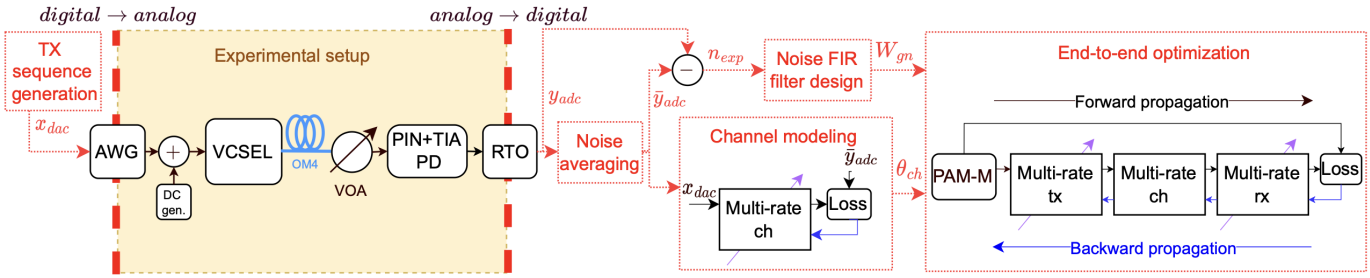

Fig. 2. Scheme illustrating an overview of the overall experimental optimization scheme, starting with signal acquisitions in the experimental setup (left) and proceeding in the several DSP operations such as Noise Averaging, Channel modeling, Noise FIR filter Design (center) and finally the End-to-end optimization (right). DC gen.: bias current generator; PD: PhotoDiode; TIA: Trans-Impedance Amplifier

$$\bar{\mathbf{W}}_{gn} = \begin{bmatrix} W_{gn} & 0 & 0 & \cdots & 0 \\ 0 & W_{gn} & 0 & \cdots & 0 \\ 0 & 0 & W_{gn} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & W_{gn} \end{bmatrix} \tag{14}$$

The E2E optimization of the DPD can be thus performed by backpropagating stochastic realizations of $MSE_{(a,\hat{y})}$. During the SGD update, the receiver will then take into account the noise presence by adding the 2^{nd} component of Eq. 13 to its "clean" gradient. This however requires the knowledge of the noise filter taps W_{gn}: in Section IV we will explain how to retrieve them from experimental measurements.

IV. Experimental Application of the Multi-Rate System

In this Section, we illustrate in detail the experimental implementation of our multi-rate end-to-end optimization. We first illustrate the optimization steps, then we evaluate the performance gain obtained when applying the trained DPD. Specifically, we investigate in this article the application of the multi-rate nonlinear DPD for a PAM-4 signal, transmitted over an experimental setup illustrated in Fig. 2 (left side). The optoelectronic part is a typical VCSEL+MMF IM-DD system. Specifically, we employed an λ=850 nm VCSEL on-chip with a 3-dB Bandwidth B_{3dB}= 20 GHz (hence severely bandlimited when transmitting $>$ 100 Gbps PAM-4 signals), and 2 different OM4 fiber cables: a 2 m cable to perform Back-to-back (B2B) analysis and a 125 m cable.

A. DSP modeling of the experimental optoelectronic channel

As a first step, to fully characterize the optoelectronic channel, we perform a non-pre-distorted acquisition over the experimental setup. A PAM-M sequence with $R_s = 58\ GBaud$, $M = 4$ and a periodicity of 2^{15} symbols is shaped through a Gaussian filter (order 2, $B_{3dB} = 0.75 \cdot R_s$), to get a digital signal $x_{dac}[n_{f_{DAC}}]$. The latter is then converted into the analog domain through an Arbitrary Waveform Generator (AWG), operating at 92 Gsample/s with an output modulation peak-to-peak voltage set to 700 mV: we selected this value since this is the maximum allowed to operate the used VCSEL (leading to approximately 14 mA of modulation current swing), and moreover, it leads to obtaining a VCSEL model in a condition that includes a non-linear regime. The modulated signal is injected into the laser together with a bias current equal to 9 mA (i.e., the value found to maximize performances in non-pre-distorted conditions). The optical light emitted from the laser (avg. TX power $\bar{P}_{TX}$=5 mW) is sent into an OM4 fiber, after which it is attenuated with a Variable Optical Attenuator (VOA) and sent to a PIN+TIA photodiode (B_{3dB}= 22 GHz). Finally, a 200 GSample/s Real-Time Oscilloscope (RTO), whose clock is shared with the AWG, is used for the ADC conversion and sequence acquisition. In this phase, 50 separated measurements are acquired from the experimental setup, to retrieve a reasonably high amount (in the order of 10^3) replicas of the transmitted sequence.

In DSP post-processing, a *Noise averaging* procedure is then applied to the received signal y_{adc}. The operation consists of averaging the sequence over its periodic replicas: as the noise samples are decorrelated, the disturbance gets filtered out, and a noiseless sequence $\bar{y}_{adc}$ is thus retrieved. The Power Spectral Density (PSD) of the sequence before and after denoising is illustrated in Fig.3. It can be noticed how the spectrum between 30 and 60 GHz gets lowered by nearly 30 dB, with a strict relation to the average performed over nearly 10^3 repetitions of the transmitted sequence.

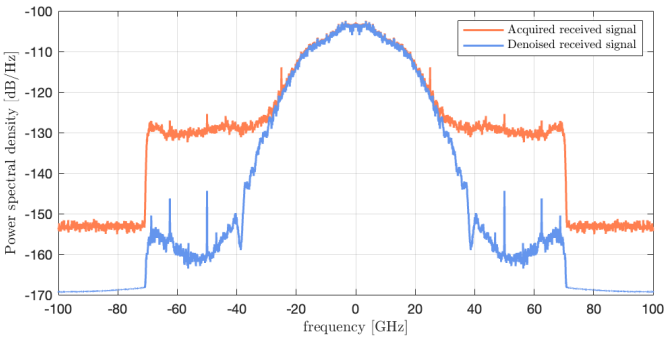

Fig. 3. Welch PSD estimates of an experimental received signal before (orange) and after (blue) denoising. Reference PSD power: $\sigma^2_{y_{adc}}$.

Once the denoised signal $\bar{y}_{adc}$ has been retrieved, it is used to perform the *Channel modeling*: we train an ANN version of the Multi-rate Channel (with $f_s^{in}{}_{ch} = 92\ GHz$ and $f_s^{out}{}_{ch} = 200\ GHz$), to synthesize a digital twin of the considered transmission channel: this has a strict analogy to what was proposed as a first step of the DLA introduced in [16], from which we took inspiration: the training of the ANN is indeed performed under MSE criterion using one period of x_{dac} as input examples and $\bar{y}_{adc}$ as output labels. In Fig. 4 we illustrate qualitatively the effectiveness of the optimized multi-rate channel, tested with a sequence different from the one used during training to avoid overfitting issues (i.e., train sequence: R_s=58 GBaud, test sequence: R_s=50 GBaud). It can be noticed how the digital twin ANN is able to faithfully reproduce the signal retrieved by noise averaging experimental acquisitions, exhibiting the same bandwidth limitations that severely close the eyediagram as well as a time-domain skew (see violet dashed arrows in Fig.4), which is a typical nonlinear effect caused by VCSELs when driven in nonlinear conditions.

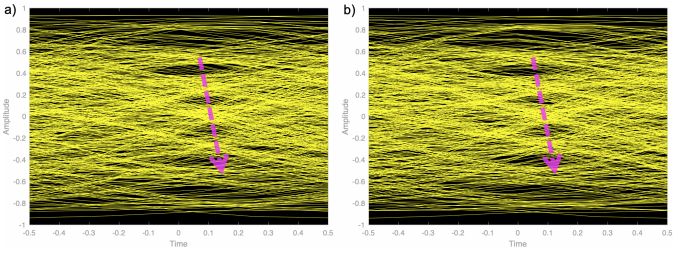

Fig. 4. Eyediagram of the noiseless channel output $\bar{y}_{adc}$ transmitted at 50 GBaud over the transmission system a) obtained by *Noise averaging* acquisitions taken from the experimental setup b) obtained through simulation using the Multi-rate Channel model, previously trained using noiseless sequence $\bar{y}_{adc}$ when transmitted at 58 GBaud in B2B.

In order to quantitatively assess the accuracy of the channel ANN digital twin, we report in Fig. 5 the training loss evolution over the model optimization together with the test loss after convergence. The train and test losses, expressed in terms of MSE normalized with respect to the power (i.e., the signal variance) of the noiseless sequence $\bar{y}_{adc}$, are respectively equal to $2.9 \cdot 10^{-3}$ and $3.4 \cdot 10^{-3}$ (corresponding to -25.4 dB and -24.7 dB). This result can be interpreted as follows: at the digital twin output, the variance of the error signal (i.e. the difference between the experimentally measured signal and the numerical output of the ANN channel model) is about 25 dB below the useful signal's variance. This is quantitative proof that the digital twin achieved very good replication of the physical channel under test.

After optimization, the ANN parameters (i.e., the weights and biases in the ANN layers) optimized during the training, named θ_{ch} (see Fig. 2), are inserted in the Multi-rate Channel of the E2E architecture.

Subsequently, by performing an element-wise subtraction of $\bar{y}_{adc}$ to each replica of y_{adc}, an estimation of the receiver noise signal n_{exp} is extracted to perform the *Noise FIR filter design*: the filter taps W_{gn} illustrated in Sec.III-C are obtained by synthesizing a FIR filter whose magnitude response fits the PSD behavior of n_{exp} (this operation can for instance be easily performed in MATLAB^{TM} exploiting the *designfilt()* function

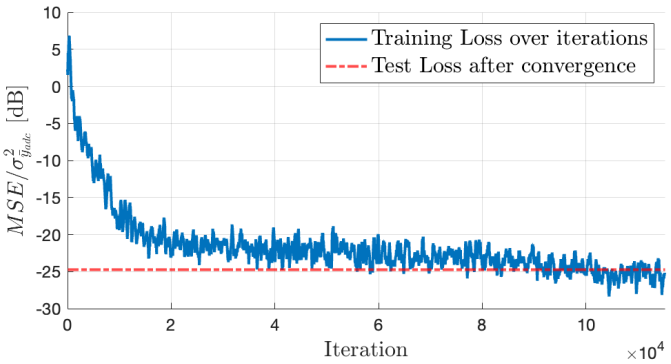

Fig. 5. Evolution of the loss function (i.e., MSE normalized with respect to $\bar{y}_{adc}$ power) over training iterations. Blue curve: training loss over iterations; Red dashed line: test loss after model convergence. Training and test sequences are relative to the qualitative analysis illustrated in Fig. 4

using the "arbmagfir" option). In Fig. 6 we present the PSD of the resampled experimental noise, compared to the noise simulated through our modelization,showing an almost perfect agreement.

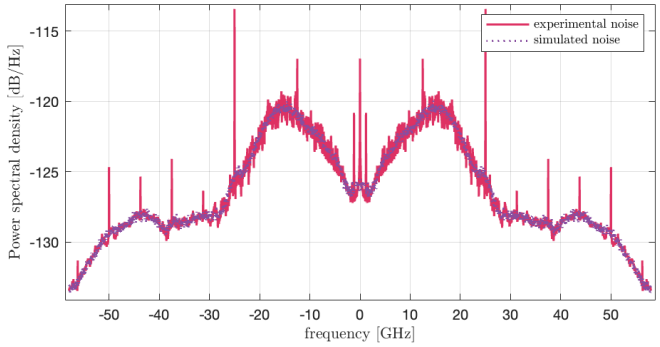

Fig. 6. Welch PSD estimates of the receiver noise extracted from denoising (red), compared to the one simulated exploiting the noise FIR filter (violet). Reference PSD power: $\sigma^2_{\bar{y}_{adc}}$

In fact, it can be clearly observed how the simulated noise spectrum behaves exactly like the experimental one, except for the spurious harmonic disturbances, which are actually introduced by the internal RTO clock spurious signals: as these impairments have however very low power, we verified their effect is thus negligible.

B. End-to-end optimization of the DPD

After obtaining the digital twin ANN parameters θ_{ch} and W_{gn} (Fig. 2) (i.e., that determine the behavior of the Multi-rate Channel and of the RX noise, respectively), the multi-rate *End-to-end optimization* of the DPD together with an FFE is performed. As training sequence, 10^5 PAM-4 random symbols are generated (i.e., using *randi()* MATLAB^{TM} function). During the end-to-end optimization, the Multi-rate Ch coefficients are kept fixed, while the DPD and FFE coefficients are jointly optimized. It must be pointed out that:

- The Baud Rate R_s adopted in the E2E optimization is independent of that used to carry out the experimental acquisition in Sec. IV-A. Since the Multi-rate Channel

is modeled to distort a generic signal with given f_s^{in} and f_s^{out} (in our case equal to 92 Gsample/s and 200 Gsample/s, respectively), once trained as illustrated in Sec. IV-A it can be exploited to obtain DPDs operating at arbitrary values of R_s.

- During the training, only the FIRNN coefficients of the DPD and the FFE are actually optimized, while the resamplers parameters (i.e., the taps in the polyphase components) are kept fixed. Training indeed even these coefficients would only increase the complexity without improving the generalization capacity, since the resampler linear transformation is followed by another linear operation (i.e., the one performed by the first FIRNN synapses) [32].

During the E2E optimization, since the transmission system is severely bandlimited, both the DPD (for its "linear" part) and the FFE tend to synthesize a high pass filtering effect. While at the receiver side this tends to enhance the noise impairments, at the transmitter side instead this causes the presence of time-domain amplitude overshoots in the pre-distorted signal: as these oscillations would exceed the VCSEL+DAC constraint, the $DAC()$ function at the output DPD layer bounds the transmitted signal to the given maximum dynamics. The consequent intrinsic clipping operation is mostly applied on the pre-distorted outer PAM-M levels, causing at the end of the E2E optimization an unbalanced pre-compensation on the TX signal, that leads to an additive nonlinear impairment that cannot be compensated by the linear FFE at the receiver side: this can be observed in Fig. 7.a where we illustrate the noiseless eye-diagram of the PAM-4 signal at the output of the E2E Multi-rate receiver (i.e., once training is complete). The E2E optimization therefore automatically converges to a solution that is MSE optimal according to Eq.11, but with a consequent unequal per-level distortion strongly penalizing the BER performances. To overcome this issue we thus designed a heuristic method that, taking inspiration from Weighted Mean Squared Error approaches [30], gives more weight in the MSE loss to the errors related to the outer PAM-M symbols. Considering for instance a PAM-4 modulation with nominal levels set to $-3, -1, +1, +3$ the error $e[n_{R_s}]^2$ to be backpropagated can be computed as follows:

$$e[n_{R_s}]^2 = \begin{cases} (a[n_{R_s} - S] - \hat{a}[n_{R_s} - S])^2 & if \ \ |a[n_{R_s} - S]| = 1 \\ h \cdot (a[n_{R_s} - S] - \hat{a}[n_{R_s} - S])^2 & if \ \ |a[n_{R_s} - S]| = 3 \end{cases} \tag{15}$$

where h is the value of the *heuristic weight* applied. The results of the application of this heuristic method in the E2E optimization is illustrated in Fig. 7.b When the heuristic is applied, the eyes are all balanced and open: the openness is anyway not full as expected, since the FFE is still considering the presence of the noise.

C. Preliminary simulation results

After the E2E optimization, we can exploit the numerically obtained E2E system as a simulation setup to evaluate preliminarly the performances of the trained DPD. By fixing the TX and RX parameters trained during E2E optimization, and now injecting in time-domain the RX noise, we simulated the BER measurement for DPDs working at different Bit Rates in a fixed OPL condition. We compare the performances of a trained nonlinear DPD (the same used then in Sec. IV-D) with respect to the case in which DPD is not applied. Details on non-DPD signal are provided in Sec. IV-D). Moreover, in order to assess the gain provided by the use of a nonlinear DPD (i.e., the ReLU FIRNN), we evaluated the performance using also a linear DPD (that is, a Multi-rate TX with a FIR filter in place of a FIRNN, but preserving the $DAC()$ function), having the same FIR length as the nonlinear DPD (i.e., 21 taps). In Fig. 8 we illustrate the results observed through simulation by using the E2E system obtained from B2B scenario (L_{OPL}=8.5 dB).

As it can be observed, the linear DPD can only slightly improve the performance at higher Bit Rates, presumably where the DSP shaping of the non-DPD signal starts to lose effectiveness due to the decreasing sps ratio: as the $DAC()$ function would clip any time-domain overshoot, the linear DPD has little room to improve the performances. The nonlinear DPD exhibits instead a noticeable improvement in the performances, with a gain of nearly 6 Gbps for a BER=10^{-3}. Although the linear pre-compensation is severely limited by the $DAC()$ function, in this case the nonlinear DPD is able to pre-distort for the nonlinear effects caused by the VCSEL, thus producing a performance gain.

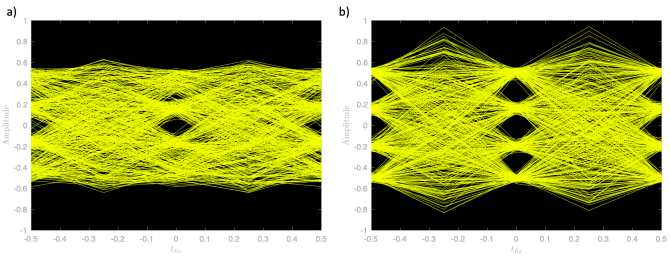

Fig. 7. E2E Rx normalized output eye-diagram after optimization a) without heuristic application; b) with heuristic applied ($h = 50$). E2E system features: B2B experimental setup, L_{OPL}=8.5 dB, R_s=50 GBaud.

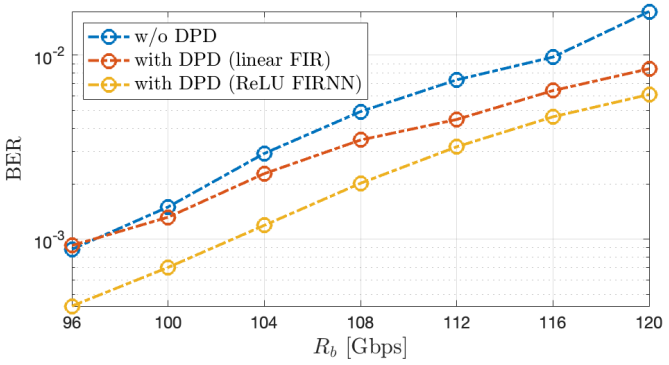

Fig. 8. BER vs Gross Bit Rate (R_b) curves obtained through simulation on the End-to-end system (B2B, L_{OPL}=8.5 dB). a) without DPD applied b) with linear FIR DPD applied c) with ReLU FIRNN DPD applied.

D. Experimental performance evaluation

By following the optimization steps illustrated in Sections IV-A and IV-B, we trained several DPDs using different values of Optical Path Loss (L_{OPL}) and several Bit Rates in the 2 tested conditions (i.e., B2B and using 125 m of OM4 fiber). At any fixed L_{OPL}, we performed one single acquisition for modeling the channel: the retrieved digital twin was then used for several parallel E2E optimizations, differing only in the DPD and FFE operative rates and in the noise FIR filter taps values. In Table I we report the parameters that, independently from the Baud rate (f_{DAC} and f_{ADC} were fixed by the AWG and the RTO, respectively), we kept fixed for all the channel and E2E optimizations.

TABLE I
FIXED PARAMETERS IN THE OPTIMIZATION PROCESSES.

Device	Transmitter	Channel	Receiver
FIR length	$T_{tx} = 21$	$T_{ch} = 61$	$T_{rx} = 31$
Hidden layers	1	1	0 (single FIR)
Activation Functions	ReLU, DAC	ReLU	none (single FIR)
Hidden neurons	$h_{tx} = 21$	$h_{ch} = 61$	0 (single FIR)
Learn rate	0.0001	0.001	0.0001

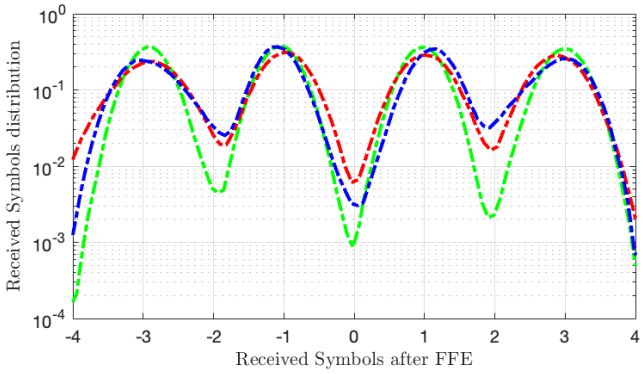

Fig. 9. Experimental received symbols distribution (R_s=50 GBaud, L_{OPL}=8.5 dB) during performance evaluation without DPD (red line, $BER = 2 \cdot 10^{-3}$), with DPD trained without heuristic (blue line, $BER = 5 \cdot 10^{-3}$) and with DPD trained using the heuristic (green line, $BER = 4 \cdot 10^{-4}$).

In addition, the heuristic weight has been fixed to h = 50 (observed value providing the best DPD performance), and the number of taps of the noise FIR filter (i.e., the length of the W_{gn} vector) has been set to $T_{gn} = 61$. Regarding the coefficients of the resamplers in the E2E architecture, for simplicity, they have been taken from the default MATLAB™ *resampler()* implementation: the number of parameters is not reported here since it varies depending on the involved frequency ratios. The choice of all the aforementioned parameters has been done consistently to the application of the DPD in IM-DD systems: therefore, we look for the lowest number of coefficients giving effective performances (for instance, using 1 single hidden layer in the FIRNNs). However, the search for optimal values for FIR filter lengths was beyond the scope of our investigation, and is therefore postponed for future work.

After training the DPDs, we then evaluated the performances in the experimental setup by pre-distorting and transmitting a PAM-4 pseudo-random sequence (PRBS) with a periodicity of 2^{16} symbols. This PRBS is completely different from the one used for the training acquisitions to avoid any overfitting effects. To assess the gain provided by the DPD performance, we transmitted the not pre-distorted sequence shaped with a Gaussian filter (i.e., the one already used in Sec.IV-A) and the same peak-to-peak modulation swing (700 mV): we selected this value since in the considered attenuated conditions we verified the noise is always the dominant impairment. Maximizing thus OMA allows for increasing the SNR at RX, which still gives advantages despite the nonlinear distortions. Therefore, the scenario that we use for comparison is the one giving the best performance without applying the DPD. The selected non-predistorted scenario exhibited an average TX output power $\bar{P}_{TX}$=5 mW, as during the training acquisition. Noticeably, the same $\bar{P}_{TX}$=5 mW was also measured in the case in which the trained DPD was applied: this is mainly due to the use of the same bias current (9 mA), and because of similar modulated signal statistics (the DPD signal has a variance less than 1 dB smaller than the non-DPD case). The channel conditions are thus mainly preserved when applying DPD, as the TX power is one of the main factors determining the VCSEL operating point (mainly related to its temperature).

On the received signals we applied the same FFE used in the E2E optimizations, training it for $2 \cdot 10^5$ symbols. We then evaluated the BER over $5 \cdot 10^5$ symbols using a hard-decisor with thresholds optimized as in [31] with respect to symbols'centroids and variances after FFE. Computed BER was thus further averaged over 3 consecutive acquisitions to obtain a very stable BER estimation, by counting errors over $3 \cdot 10^6$ bits. The distributions of the post-equalized signals with and without DPD (normalized to give a probability density estimate) are illustrated in Fig. 9.

Consistently to what has been observed in Sec.IV-B, after a simple MSE-based E2E optimization (blue line in Fig. 9) the DPD is able to reduce mostly the distortions affecting the inner PAM-4 symbols. However, this compensation penalizes the outer levels, increasing the overall $BER = 2 \cdot 10^{-3}$ achieved without DPD (red line in Fig. 9) to a $BER = 5 \cdot 10^{-3}$. Applying the heuristic instead (green line in Fig. 9) provides a significant decision improvement for all the transmitted symbols with respect to the non-pre-distorted signal, achieving in the considered scenario an overall $BER = 4 \cdot 10^{-4}$.

We then compare the DPD versus non-DPD BER performances in Fig. 10: we present the results as BER contour plots, for target BERs equal to 10^{-2} and 10^{-3}. These plots, which are the main experimental results of our paper, demonstrate that our DPD implementation gives significant improvements in any condition with respect to the best-performing non-pre-distorted scenario.

At BER=10^{-3} the gain in terms of Bit rates is relevant, increasing as the L_{OPL} gets lower: specifically, nearly 8 Gbps for L_{OPL}=8 dB in B2B and around 14 Gbps for L_{OPL}=6 dB using the 125 m OM4. The reason is that the non-DPD scenario exhibits a noticeable performance degradation as the nonlinear distortions overcome the noise disturbance. Evaluating instead the gain in terms of L_{OPL}, for BER=10^{-2} this is considerably high: with $R_b = 110$ Gbps the nonlinear

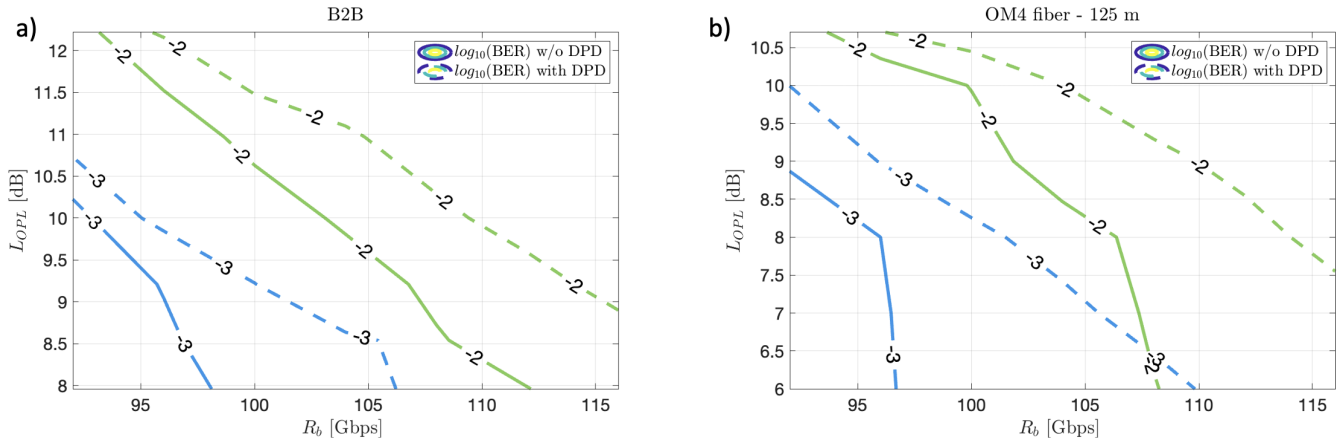

Fig. 10. Optical Path Loss (L_{OPL}) vs Bit Rate (R_b) curves for BER target equal to 10^{-2} and 10^{-3}, without DPD (solid lines) and with DPD applied (dashed lines) a) B2B scenario b) using 125 meters of OM4 fiber.

DPD improves the sensitivity by nearly 1.5 dB in B2B and by at least 3 dB with the OM4-125 m fiber (according to the available results).

V. Discussion and Conclusion

In this article, we proposed a new method to optimize a nonlinear DPD for $>$ 100 Gbps PAM-4 signals transmitted in VCSEL-MMF IM-DD systems. The trained predistorter is able, in a transmission system severely bandlimited (e.g., VCSEL B_{3dB}= 20 GHz), to jointly fulfill the constraints imposed by a TX maximum amplitude and the mismatch between sampling frequency and Baud Rate of the transmitted signal. By exploiting a DLA-based modeling of the channel, a characterization of the experimental noise, and the use of FIRNN to properly address the memory of the transmission system, we implemented a novel FIRNN-based E2E system, whose backpropagation has been extended to cope with multiple sampling rates within its structure. As result, we obtained an optimization algorithm that provides a multi-rate DPD able to produce optimized pre-distorted signals while natively supporting amplitude constraints at a fractional sps ratio. Moreover, the method introduces the noise as an additive regularization term in the FFE loss gradient, leading to an effective SGD optimization.

On the considered transmission system, we observed that this DPD optimization approach was actually the only method providing an effective predistorter (i.e., giving better performance than our reference non-DPD scenario). The two other approaches that we attempted, ILA and DLA respectively, have shown to be problematic in such severely bandlimited system:

- Using **ILA**, the dynamics constraints cannot be fulfilled using techniques such as the one we proposed (i.e., introducing an hard-limiter function at the DPD output during the optimization) . Unless using alternative techniques, the trained DPD is led to produce severe time-domain overshoots to counteract bandwidth limitations, that must necessarily be clipped before DAC conversion. Such nonlinear mitigation leads to an excessive degradation in the signal, that cannot be recovered by a linear FFE at the RX side.
- Using **DLA**, whose implementation in this scenario is equivalent to performing the E2E optimization with the E2E system deprived of the RX equalizer FFE block, we verified that, although the dynamics constraints can be fulfilled, the DPD still attempts to synthesize overshoots, that are consequently clipped by the $DAC()$ function: even applying a strong heuristic weight cannot help in overcoming this issue. According to our study, we thus believe that an E2E optimization in which the DPD is aware that a RX FFE will partially compensate for bandwidth limitations is necessary to correctly perform nonlinear DPD on a VCSEL-MMF optical link as the one considered.

Concerning complexity, we want to point out that the implementation of an online training algorithm requires reduced computational effort and memory requirements with respect to conventional deep-learning offline SGD optimizations [32]: only 1 single non-periodic random sequence is used to train the E2E system, and the temporal backpropagation structure requires a number of computations that is roughly the same as the forward propagation. For instance, DPDs required in our experiments less than 10 minutes to be optimized on a low-cost commercial laptop without using GPU: this could be an acceptable solution for a factory-level DPD optimization. Parallel training of different DPDs can also be implemented, considering that for a given channel condition (i.e., a fixed L_{OPL}), the retrieved Multi-rate Channel can be exploited for E2E optimizations at any Baud rates.

Considering the DPD FIRNN structure, our proposal demonstrated to be effective by performing $\sim$733 multiplications per symbol (mps) with R_s=58 GBaud. Further study is still required to assess if using more complex architectures would increase the performance.

The presented work on nonlinear DPD over IM-DD optical link highlights anyway the actual limitations affecting this kind of transmission system at high rates. Strong bandwidth limitations showed being not fully compensable by a DPD, as outer PAM-M levels have no room for overshoots: the performance gain from pre-compensation in such a constrained scenario seems thus difficult to be further increased. Moreover,

in such constrained optimizations, a simple MSE criterion appears to not be the optimal choice: unbalanced distortions affecting the symbols after the FFE lead to solutions that are not merely optimal for a minimum BER criterion.

Our multi-rate approach, here applied using FIRNN but extendible to other linear and nonlinear structures, could be therefore exploited to train different DPD architectures, with even a post-equalizer different from the FFE. We are therefore confident that the presented approach for nonlinear DPD could inspire further works in this research field.

APPENDIX
FIR-BASED NEURAL NETWORKS

Proposed by Eric A. Wan in [28], FIR-based Neural Networks (FIRNN) are generalized Feed Forward Neural Networks (FFNN) where each network synapse is extended to be a Finite Impulse Response Filter (FIR), as illustrated in Fig. 11.

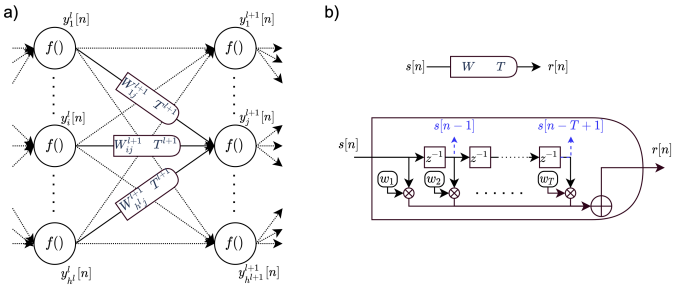

Fig. 11. (a) FIRNN structure from the l-th to the $(l+1)$-th layer. (b) Structure of a FIR filter.

In a FIRNN with L layers, its h^L-dimensional output vector $\boldsymbol{y}^L[n] = [{y^L}_1[n], {y^L}_2[n], \cdots, {y^L}_{h^L}[n]]^\top$ is produced from a h^0-dimensional input vector $\boldsymbol{y}^0[n] = [{y^0}_1[n], {y^0}_2[n], \cdots, {y^0}_{h^0}[n]]^\top$, through the following iterative computation (we omit biases for semplicity):

$$for\ \ l = 0, .., L-1:$$
$${y^{l+1}}_j[n] = f\left(\sum_{i=1}^{h^l} {W_{ij}^{l+1}}^\top Y_i^l[n]\right) \quad j = 1, .., h^{l+1} \tag{16}$$

In Eq.16, $f()$ is the nonlinear activation function, and W_{ij}^{l+1} is the vector containing the taps of the FIR synapse from the i-th node in the l-th layer to the j-th node in the $(l+1)$-th layer, defined as:

$$W_{ij}^l = \left[w_{ij_1}^l, w_{ij_2}^l, \cdots, w_{ij_{T^l}}^l\right]^\top \quad l = 1, .., L \tag{17}$$

Finally, in 16 $Y_i^l[n]$ is the sequence vector whose entries are successive samples of $y_i^l[n]$, as follows:

$$Y_i^l[n] = [y_i^l[n], y_i^l[n-1], \cdots, y_i^l[n-(T^l-1)]]^\top \tag{18}$$

where T^l is the number of taps in all the FIR filters of the $l-th$ layer.

FIRNNs, as well as all neural networks, can be trained to perform a given task by optimizing their coefficients so that a given cost function is minimized. Specifically, by defining the target output vector at the n-th discrete-time index as $\Xi[n] = [\xi_1[n], \xi_2[n], .., \xi_{h^L}[n]]^\top$, the FIRNN can be trained by using a Stochastic Gradient Descent (SGD) online training algorithm, which minimizes the total squared error over all time defined as follows:

$$e^2 = \sum_{n=0}^{\infty} \sum_{i=1}^{h^L} \left(\xi_i[n] - {y^L}_i[n]\right)^2 \tag{19}$$

At each discrete time index n then, the loss derivatives with respect to the FIRNN coefficients can be retrieved by exploiting an online temporal backpropagation algorithm, like the one proposed in [28]. This method extends the conventional recursive gradient computation algorithm used for FFNN, interpreting each FIR filter's tap delay as a "virtual neuron", whose output is a delayed version of its input. According to the equation (35) in [28], the algorithm can be thus summarized as follows:

$$\delta_j^l[n] = \frac{\partial e^2}{\partial y_j^l[n]} = \begin{cases} -2\left(\xi_j[n] - y_j^L[n]\right) \cdot f'(y_j^L) & l = L \\ f'(y_j^l) \cdot \sum_{m=1}^{h^{l+1}} \Delta_m^{l+1}[n]^\top W_{jm}^{l+1} & 1 \le l \le L-1 \end{cases} \tag{20}$$

$$\Delta_m^l[n] = [\delta_m^l[n], \delta_m^l[n+1], .., \delta_m^l[n+(T^l-1)]] \tag{21}$$

$$\nabla_{w_{ij_k}^l} e^2 = \delta_j^l[n] \cdot y_i^{l-1}[n-k+1] \tag{22}$$

where $f'()$ is the derivative of the nonlinear activation function. The coefficients can be then updated as follows:

$$W_{ij}^l \leftarrow W_{ij}^l - \epsilon \cdot \delta_j^l[n] \cdot Y_i^{l-1}[n] \tag{23}$$

where ϵ is the learning rate.

The peculiarity of this online training algorithm is that the gradient recursive computation can be implemented as an FIR-based structure, where the backward propagated δ terms are symmetric to the forward propagated y states. Computation of the δ terms is non-causal: this can be anyway solved in practice by adding a finite number of delay operators into the network, as the structure is FIR-based [28].

In this article, the proposed End-to-end multi-rate system can be viewed as an enhanced FIRNN, able to deal with different sampling rates within its structure.

ACKNOWLEDGMENT

This work was carried out under a research contract with Cisco Photonics. We also acknowledge the PhotoNext initiative at Politecnico di Torino (http://www.photonext.polito.it/) and its laboratory, where all experiments have been performed.

REFERENCES

[1] Jonathan King, "In Support of 200G MMF Ethernet PMDs," *IEEE 802.3 Next-generation 200 Gb/s and 400 Gb/s MMF PHYs Study Group*, https://www.ieee802.org/3/NGMMF/public/Jan18/young_NGMMF_01a_jan18.pdf, last accessed on 5 May 2022.

[2] Honghang Zhou, Yan Li, Yuyang Liu, Lei Yue, Chao Gao, Wei Li, Jifang Qiu, Hongxiang Guo, Xiaobin Hong, Yong Zuo and Jian Wu, "Recent Advances in Equalization Technologies for Short-Reach Optical Links Based on PAM4 Modulation: A Review," *Applied Sciences*, 2019

[3] Y. Yu, T. Bo, Y. Che, D. Kim and H. Kim, "Low-Complexity Equalizer Based on Volterra Series and Piecewise Linear Function for DML-Based IM/DD System," *2020 Optical Fiber Communications Conference and Exhibition (OFC)*, 2020, pp. 1-3.

[4] Anzhong Liang, Chuanchuan Yang, Cheng Zhang, Yue Liu, Fan Zhang, Zhenrong Zhang, Hongbin Li, "Experimental study of support vector machine based nonlinear equalizer for VCSEL based optical interconnect," *Optics Communications*, Volume 427, 2018, Pages 641-647.

[5] Qingyi Zhou, Chuanchuan Yang, Anzhong Liang, Xiaolong Zheng, Zhangyuan Chen,"Low computationally complex recurrent neural network for high speed optical fiber transmission," *Optics Communications*, Volume 441, 2019, Pages 121-126.

[6] Y. Gao et al., "288 Gb/s 850 nm VCSEL-based Interconnect over 100 m MMF based on Feature-enhanced Recurrent Neural Network," *2022 Optical Fiber Communications Conference and Exhibition (OFC)*, 2022, pp. 01-03.

[7] J. Zhang et al., "PAM-8 IM/DD Transmission Based on Modified Lookup Table Nonlinear Predistortion," *IEEE Photonics Journal*, vol. 10, no. 3, pp. 1-9, June 2018, Art no. 7903709, doi: 10.1109/JPHOT.2018.2828869.

[8] Z. He, K. Vijayan, M. Mazur, M. Karlsson, and J. Schröder, "Look-up Table based Pre-distortion for Transmitters Employing High-Spectral-Efficiency Modulation Formats," *2020 European Conference on Optical Communications (ECOC)*, 2020, pp. 1-4, doi: 10.1109/ECOC48923.2020.9333231.

[9] Q. Zhang, Z. Wang, S. Duan, N. Jiang, B. Cao and Y. Wu, "An Improved End-to-end Optical Transmission System Based On Deep Learning," *2021 19th International Conference on Optical Communications and Networks (ICOCN)*, pp. 1-3,2021, doi: 10.1109/ICOCN53177.2021.9563723.

[10] V. Shivashankar, C. Kottke, V. Jungnickel and R. Freund, "Investigation of Linear and Nonlinear Pre-Equalization of VCSEL," *Broadband Coverage in Germany*; 11. ITG-Symposium, 2017, pp. 1-5.

[11] V. Bajaj, M. Chagnon, S. Wahls and V. Aref, "Efficient Training of Volterra Series-Based Pre-distortion Filter Using Neural Networks," *2022 Optical Fiber Communications Conference and Exhibition (OFC)*, 2022, pp. 1-3.

[12] T. O'Shea and J. Hoydis, "An Introduction to Deep Learning for the Physical Layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563-575, Dec. 2017, doi: 10.1109/TCCN.2017.2758370.

[13] V. Bajaj, F. Buchali, M. Chagnon, S. Wahls and V. Aref, "Deep Neural Network-Based Digital Pre-Distortion for High Baudrate Optical Coherent Transmission," *IEEE/OSA Journal of Lightwave Technology*, vol. 40, no. 3, pp. 597-606, 1 Feb.1, 2022, doi: 10.1109/JLT.2021.3122161.

[14] B. Karanov et al., "End-to-End Deep Learning of Optical Fiber Communications," *IEEE/OSA Journal of Lightwave Technology*, vol. 36, no. 20, pp. 4843-4855, 15 Oct.15, 2018, doi: 10.1109/JLT.2018.2865109.

[15] V. Neskorniuk et al., "End-to-End Deep Learning of Long-Haul Coherent Optical Fiber Communications via Regular Perturbation Model," *2021 European Conference on Optical Communication (ECOC)*, 2021, pp. 1-4, doi: 10.1109/ECOC52684.2021.9605928.

[16] G. Paryanti, H. Faig, L. Rokach and D. Sadot, "A Direct Learning Approach for Neural Network Based Pre-Distortion for Coherent Nonlinear Optical Transmitter," *IEEE/OSA Journal of Lightwave Technology*, vol. 38, no. 15, pp. 3883-3896, 1 Aug.1, 2020, doi: 10.1109/JLT.2020.2983229.

[17] V. Aref and M. Chagnon, "End-to-End Learning of Joint Geometric and Probabilistic Constellation Shaping," *2022 Optical Fiber Communications Conference and Exhibition (OFC)*, 2022, pp. 1-3.

[18] J. Song, C. Häger, J. Schröder, A. G. i. Amat and H. Wymeersch, "End-to-end Autoencoder for Superchannel Transceivers with Hardware Impairment," 2021 Optical Fiber Communications Conference and Exhibition (OFC), 2021, pp. 1-3.

[19] B. Karanov, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-end optimized transmission over dispersive intensity- modulated channels using bidirectional recurrent neural networks," Opt. Exp., vol. 27, no. 14, 2019, Art. no. 19650.

[20] S. Gaiarin, F. Da Ros, R. T. Jones, and D. Zibar, "End-to-end optimization of coherent optical communications over the split-step Fourier method guided by the nonlinear Fourier transform theory," J. Lightw. Technol., vol. 39, no. 2, pp. 418–428, 2021.

[21] R. T. Jones, M. P. Yankov, and D. Zibar, "End-to-end learning for GMI optimized geometric constellation shape," in Proc. Eur. Conf.Opt. Commun., 2019, pp. 1–4.

[22] O. Jovanovic, M. P. Yankov, F. Da Ros, and D. Zibar, "End-to-End Learning of a Constellation Shape Robust to Channel Condition Uncertainties," Journal of Lightwave Technology, vol. 40, no. 10, pp. 3316 - 3324, 2022.

[23] B. Karanov, M. Chagnon, V. Aref, D. Lavery, P. Bayvel, and L. Schmalen, "Concept and experimental demonstration of optical IM/DD end-to-end system optimization using a generative model," in Proc. Opt. Fiber Commun. Conf. Exhib., 2020, pp. 1–3.

[24] F. A. Aoudia and J. Hoydis, "Model-Free Training of End-to-End Communication Systems," IEEE Journal on Selected Areas in Communications, vol. 37, no. 11,pp. 2503–2516, 2019.

[25] O. Jovanovic, M. P. Yankov, F. Da Ros, and D. Zibar, "Gradient-Free Training of Autoencoders for Non- Differentiable Communication Channels," Journal of Lightwave Technology, vol. 39, no. 20, pp. 6381–6391, 2021.

[26] J. Song et al., "Over-the-fiber Digital Predistortion Using Reinforcement Learning," 2021 European Conference on Optical Communication (ECOC), 2021, pp. 1-4, doi: 10.1109/ECOC52684.2021.9605972.

[27] S. Echeverri-Chacón et al., "Transmitter and Dispersion Eye Closure Quaternary (TDECQ) and Its Sensitivity to Impairments in PAM4 Waveforms," *Journal of Lightwave Technology*, vol. 37, no. 3, pp. 852-860, 1 Feb.1, 2019, doi: 10.1109/JLT.2018.2881986.

[28] E. A. Wan, "Temporal backpropagation for FIR neural networks," *1990 IJCNN International Joint Conference on Neural Networks*, vol.1, pp. 575-580, June 1990.

[29] C. M. Bishop, "Training with Noise is Equivalent to Tikhonov Regularization," *Neural Computation*, vol. 7, no. 1, pp. 108-116, Jan. 1995, doi: 10.1162/neco.1995.7.1.108.

[30] Y. Sai, R. Jinxia and L. Zhongxia, "Learning of Neural Networks Based on Weighted Mean Squares Error Function," *2009 Second International Symposium on Computational Intelligence and Design*, 2009, pp. 241-244, doi: 10.1109/ISCID.2009.67.

[31] L. Minelli, A. Abdellatif, and R. Gaudino, "Optimization of 50G-PON APD-based receivers," *Italian Conference on Optics and Photonics*, 2022.

[32] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning". *MIT press*, 2016.