End-to-End Learning of Transmitter and Receiver Filters in Bandwidth Limited Fiber Optic Communication Systems

Søren Føns Nielsen, Francesco Da Ros, Mikkel N. Schmidt and Darko Zibar

Abstract—This paper investigates the application of end-to-end (E2E) learning for joint optimization of pulse-shaper and receiver filter to reduce intersymbol interference (ISI) in bandwidth-limited communication systems. We investigate this in two numerical simulation models: 1) an additive white Gaussian noise (AWGN) channel with bandwidth limitation and 2) an intensity modulated direct detection (IM/DD) link employing an electro-absorption modulator. For both simulation models, we implement a wavelength division multiplexing (WDM) scheme to ensure that the learned filters adhere to the bandwidth constraints of the WDM channels. Our findings reveal that E2E learning greatly surpasses traditional single-sided transmitter pulse-shaper or receiver filter optimization methods, achieving significant performance gains in terms of symbol error rate with shorter filter lengths. These results suggest that E2E learning can decrease the complexity and enhance the performance of future high-speed optical communication systems.

Index Terms—End-to-end learning, intersymbol interference, digital communication, intensity modulation

I. INTRODUCTION

AS we continue to integrate digital technology into all aspects of society, the traffic demands are estimated to grow at an exponential rate towards 2030 [1]. This in turn requires hyperscale datacenters (DC) located all around the world with high-speed optical interconnects, where intensity modulation with direct detection (IM/DD) is the most widely used technology.

Future DC solutions will likely employ higher order PAM signaling and baud rates reaching 100 GBd or beyond to achieve 800 Gb/s and 1.6 Tb/s [2]. This will require ultra-high bandwidths of optical and electrical components, which may be challenging to realize in practice. It is therefore expected that the next generation data center fiber-optic communication systems will need to deal with strong inter-symbol-interference (ISI), while at the same time keeping the complexity, and thereby power consumption, down [3]. The focus of the next generation of transmitter- and receiver-side digital signal processing (DSP) techniques should therefore focus on low complexity solutions for ISI mitigation.

As we move towards 100 GBd or higher data rates, the bandwidth limitations of the transmitter and the receiver front-end components (digital-to-analog converter, optical modulator, photodiodes and analog-to-digital converter), along with chromatic dispersion in the fiber channel, will lead to strong ISI.

Realizing a zero ISI system is in practice a difficult problem, as it requires that the total transfer function—including pre-distortion, pulse-shaper, transmitter front-end, channel, receiver front-end, receiver (matched) filter and equalizer—must fulfill the Nyquist criterion [4]. Typically, the components that can be adjusted to meet the Nyquist criterion are the pre-distortion, the pulse-shaper, the receiver filter and the equalizer. On top of that, it is desirable to find a parsimonious solution promoting low complexity for many practical applications.

Current solutions, for mitigating ISI are predominantly focused on receiver-side *equalization*, i.e. the process of removing the distortion introduced by the system, which can be done using an adaptive linear filter, denoted a feed-forward equaliser (FFE) [4]. Feedback filter structures are often used in conjunction with the FFE, either after the FFE at the receiver, denoted decision feedback equalizer (DFE) [5], or at the transmitter, denoted Tomlinson-Harashima Precoding (THP) [6]. In systems with non-linear distortion due to hardware imperfections and fiber dispersion, non-linear compensation can be employed. Examples include non-linear adaptive filters such as the Volterra equalizer [7] and neural networks (see [8] for an overview). Some effort has also gone into mitigating distortion from the transmitter side, either through *pulse-shaping* [9], i.e. designing spectrally efficient filters to minimize ISI, or predistortion [10], [11] and constellation shaping [12] to linearize the transmitter response. Typically these compensation mechanisms are designed or optimized independently of each other, which may lead to suboptimal performance.

End-to-end learning (E2E) [13], [14] has emerged as a framework for jointly optimizing all adaptable DSP blocks in the system on both the transmitter and receiver side. In [14] this was shown for an IM/DD system, where both transmitter and receiver were modeled by a feed-forward neural network using an autoencoder-style structure. The system was trained offline (through numerical simulation) using the backpropagation algorithm given a known pilot sequence, and evaluated experimentally with a fine-tuning step for the receiver. This enabled information rates above 42 Gb/s

beyond 40 km of transmission distance without amplification. E2E has since been used in the context of geometric constellation shaping [15], joint geometric and probabilistic constellation shaping [16], wavelength division multiplexing systems [17], learning parameters in directly modulated laser transmission [18] and learning the pulse shape and receiver demapper in a 6G wireless application [19], to mention a few examples.

However, to the best of our knowledge, no systematic analysis of E2E learning has been performed for the joint optimization of pulse-shaper and receiver filter, to reach the zero ISI criterion for AWGN and IM/DD channels under strong ISI.

In this paper we focus on the setting where the pulse-shaper and receiver filter are implemented as FIR filters, which is of great importance due to low complexity of implementation. We numerically investigate the use of E2E for jointly optimizing the pulse-shaper and receiver filter to combat ISI due to bandwidth limitations of the digital-to-analogue (DAC) and analogue-to-digital (ADC) converter, as well as chromatic dispersion, in a 100 GBd four pulse amplitude modulation (4-PAM) communication link. We first show the benefits of joint pulse-shaper and receiver filter optimization in a simple additive white Gaussian noise (AWGN) system with bandwidth limitation, where zero ISI can be reached for fairly short filter lengths. Next, we show the E2E framework's benefits in a wavelength division multiplexing (WDM) scheme, where the learned filters must adhere to the bandwidth restrictions imposed by the presence of interfering channels. We then show the benefits of joint optimization for a 1 km WDM IM/DD link with an electro-absorption modulator (EAM) that exhibits some degree of nonlinearity. In the non-linear case, we additionally compare our proposed joint optimization to a flexible Volterra equalizer with a large number of parameters to investigate the theoretical limits of this type of optimization. Finally, we evaluate the robustness of the learned filters to fiber non-linearities using the split-step Fourier method (SSFM).

The rest of the paper is structured as follows. In section II, we introduce the learning framework and the two communication links (AWGN and IM/DD) by their system model. Furthermore, we describe how WDM is implemented for both links. Next, in section III, we present the results of our simulations, including symbol-error rate curves for different scenarios. Finally, in section IV, we present our conclusions and perspectives on our findings.

II. NUMERICAL SETUP

We begin by introducing the learning framework which is based on stochastic gradient descent optimization. In the following sections, the employed end-to-end learning frameworks for AWGN and IM/DD are described.

A. End-to-end Learning Using Stochastic Gradient Descent

In the communication systems we investigate here, we are interested in obtaining pulse-shaper and receiver filters that jointly reduce ISI as much as possible. We employ a gradient based learning of the filter coefficients. Let the *loss* function be

denoted $\mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{x}}, \boldsymbol{\theta})$, in which $\boldsymbol{x}$ is the true symbol sequence, $\hat{\boldsymbol{x}}$ is the estimated symbol sequence before demapping and $\boldsymbol{\theta} = \{\boldsymbol{h}_\mathrm{p}, \boldsymbol{h}_\mathrm{r}\}$, is the collection of all parameters we want to optimize, in this case the pulse-shaper $\boldsymbol{h}_\mathrm{p}$ and the receiver filter $\boldsymbol{h}_\mathrm{r}$. In each iteration of the algorithm, we use gradient descent to minimize $\mathcal{L}$. This requires calculating the gradient of $\mathcal{L}(\cdot)$ with respect to $\boldsymbol{\theta}$, $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$. Furthermore, to avoid having to observe the entire sequence of symbols before updating the model, we use a stochastic gradient descent approach, processing only a batch of sybmols of size N_batch, denoted $\boldsymbol{x}^{(b)}$. In all simulations, we use the mean-square error as our loss function, which for one batch, can be written as,

$$\mathcal{L}(\hat{\boldsymbol{x}}^{(b)}, \boldsymbol{x}^{(b)}, \boldsymbol{\theta}) = \frac{1}{N_\mathrm{batch}} \sum_{n=1}^{N_\mathrm{batch}} (\hat{x}_n^{(b)} - x_n^{(b)})^2 \qquad (1)$$

Note here that the parameters $\boldsymbol{\theta}$ are implicitly used to produce $\hat{\boldsymbol{x}}$ but are not directly appearing in the right-hand side of (1).

All the systems described below are implemented in PyTorch [20] which provides automatic differentiation for gradient computations using the backpropagation algorithm, as well as routines for stochastic gradient optimization. We use the Adam optimizer [21], which combines momentum and adaptive learning rates to achieve fast convergence and stable training. At each gradient update, we use gradient norm clipping [22] and we normalize the filters to have unit L2-norm. We use a learning rate scheduling called OneCLR [23]. A high-level pseudo-code of our optimization routine can be seen in Alogrithm 1, and our implementation is available online[1].

**Algorithm 1** Optimization of FIR Filters

---

1: **Input:** Initial filters $\boldsymbol{h}_\mathrm{p}$, $\boldsymbol{h}_\mathrm{r}$, learning rate schedule η_b, training data $\boldsymbol{x}$ and batch size N_batch.
2: **for** $b = 1$ to $N_\mathrm{symbols}/N_\mathrm{batch}$ **do**
3:     Evaluate system on $\boldsymbol{x}^{(b)}$ to get predictions $\hat{\boldsymbol{x}}^{(b)}$.
4:     Calculate loss $\mathcal{L}_b = \mathcal{L}(\boldsymbol{x}^{(b)}, \hat{\boldsymbol{x}}^{(b)}, \boldsymbol{\theta})$.
5:     Compute the gradients: $\boldsymbol{g}_\mathrm{p} = \frac{\partial \mathcal{L}_b}{\partial \boldsymbol{h}_\mathrm{p}}$, $\boldsymbol{g}_\mathrm{r} = \frac{\partial \mathcal{L}_b}{\partial \boldsymbol{h}_\mathrm{r}}$,
        and apply norm clipping.
6:     Update parameters:
        $\boldsymbol{h}_\mathrm{p} := \boldsymbol{h}_\mathrm{p} - \eta_b \Delta_b(\boldsymbol{g}_\mathrm{p})$, $\boldsymbol{h}_\mathrm{r} := \boldsymbol{h}_\mathrm{r} - \eta_b \Delta_b(\boldsymbol{g}_\mathrm{r})$,
        where Δ_b is the Adam [21] update.
7: **end for**
8: **Output:** Optimized filters $\boldsymbol{h}_\mathrm{p}$, $\boldsymbol{h}_\mathrm{r}$.


B. Additive White Gaussian Noise Channel with Bandwidth Limitation

We first investigate end-to-end learning in a simple additive white Gaussian noise (AWGN) channel with bandwidth limitation, which is depicted in Figure 1. We assume that the bandwidth limitation originates from the DAC and ADC. The pulse-shaper and receiver filter are implemented as FIR filters with adjustable coefficients that are optimized using E2E learning.

[1]Link to our Github repository:
https://github.com/sfvnielsen/marble-end-to-end
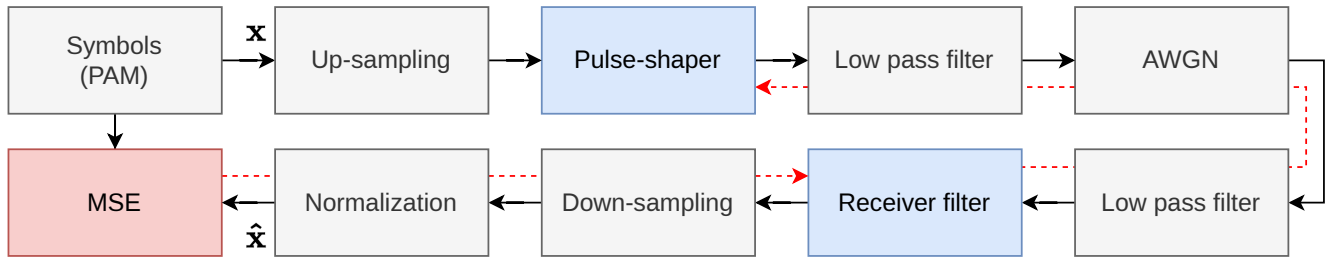
(a) Training mode. Pulse-shaper and receiver filters are updated using backpropagation, visualized with the dashed red line.
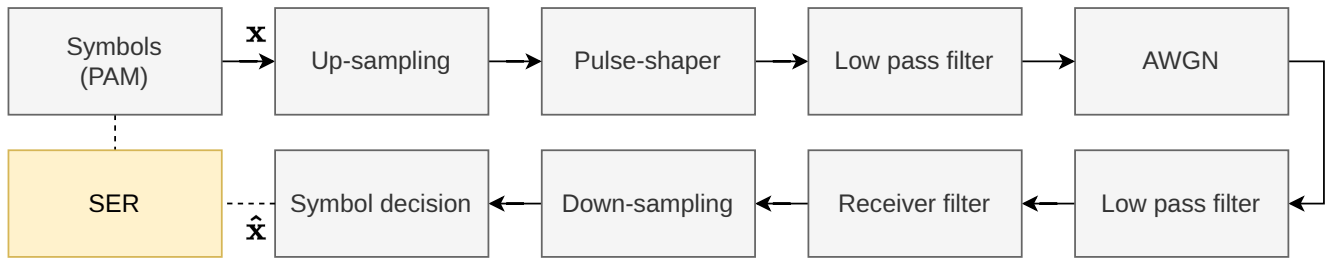
(b) Evaluation mode. Pulse-shaper and receiver filters are fixed to the values from last iteration of the training phase.

Fig. 1: Blockdiagram of the system model for the AWGN channel with bandwidth limitation. When filters are optimized (training) they are done so using the mean square error (MSE) between the output of the system and the symbol sequence, depicted in Figure 1a. During evaluation (depicted in Figure 1b), a new symbol sequence is generated, filters are fixed and the symbol error rate (SER) is calculated.

Pulse-amplitude modulation (PAM) symbols with constellation order 4 are drawn from a uniform distribution after which they are upsampled by a factor of 4. The up-sampled symbol sequence is passed through a finite impulse response (FIR) pulse-shaping filter such that at the output we have a symbol rate of $R_s = 100$ GBd. To emulate the bandwidth limitation of the DAC, the signal is then low-pass filtered by a 5th order Bessel filter with 3dB cutoff $f_{3dB} = 45$ GHz. Next, AWGN is added such that the resulting signal's SNR can be varied. After the noise is added, another Bessel low-pass filter, emulating the bandwidth limitation of the ADC, is applied with same cutoff as the first one. A receiver FIR (matched) filter is applied and the signal is downsampled to a single sample per symbol (sps). Finally, we apply a normalization that rescales the signal to match the average power of the constellation. In all simulations, the pulse-shaper and receiver filters are initialized to a root-raised cosine (RRC) with a rolloff of $\rho = 0.01$ and we vary the length of the filters as part of the simulations. With the aforementioned initialization, it is apparent that due to the bandwidth limitations of the DAC and ADC, the total impulse response will not satisfy the Nyquist criterion. Therefore, the pulse-shaper and the matched filter coefficients need to be adapted.

During training (see Figure 1a), the output of the receiver filter, $\hat{x}$, is compared to the true symbol sequence x using MSE, as described in the previous section. The $\hat{x}$ and x are coarsely aligned before calculating the cost function, by a priori calculating the average group-delay of the two Bessel

filters in the passband and compensating for that in the MSE calculation.

During evaluation (see Figure 1b), a new sequence of symbols is generated and propagated through the system with the learned filters fixed to their value from the last iteration of the training phase. The output of the downsampler is mapped to the closest constellation point in the Euclidean distance to obtain the symbol estimate, $\hat{x}$. Finally the symbol error rate (SER) between the sequences x and $\hat{x}$ is calculated. In the case of the AWGN, we train at one (high) SNR and evaluate over a range of SNRs to get performance curves. This can be done due to the linearity of the system (which is not true for the IM/DD system).

C. Multilevel Intensity Modulated and Direct Detected System

Next, we investigate the end-to-end learning framework for an intensity modulated direct detection (IM/DD) link that employs an EAM and fiber transmission link. We would like to investigate the benefits of joint pulse-shaper and receiver filter optimization in the presence of EAM nonlinearity and chromatic dispersion. Again, to keep the complexity low, the pulse-shaper and the receiver filter are implemented as FIR filters.

The block diagram for the system can be seen in Figure 2. We generate a PAM-4 symbol sequence, that after up-sampling at $sps = 8$ and pulse-shaping, has an $R_s = 100$ GBd. The DAC is comprised of a normalization step and a low-pass filter and outputs a voltage signal. We ignore quantization

noise during training in the DAC. The normalization step is a multiplication with a gain, g_{dac}, after which a clipping function is applied with range $[-1/2, 1/2]$. A voltage peak-to-peak factor, V_{pp}, is now multiplied onto the signal such that it now has range $[-V_{\text{pp}}/2, V_{\text{pp}}/2]$. The low-pass filter is applied with same parameters as for the AWGN system (5th order Bessel, $f_{\text{3dB}} = 45$ GHz).

The resulting voltage signal, $V(t)$, is combined with a bias voltage V_{b} and applied to the modulator, i.e. $V_{\text{m}}(t) = V_{\text{b}} + V(t)$. We study two modulators, an ideal linear modulator and an electro-absorption modulator (EAM) [12], [24]. The output of the ideal modulator has the form,

$$E_{\text{ideal}}(t) = \sqrt{P_{\text{in}}^{(\text{ideal})} \cdot V_{\text{m}}(t)}, \qquad (2)$$

in which P_{in} is the power of the laser and $E_{\text{ideal}}(t)$ is the resulting signal in the optical domain.

In the case of the EAM, the amplitude of the signal at the output of the modulator is written as,

$$A_{\text{EAM}}(t) = \sqrt{P_{\text{in}}^{(\text{EAM})} \cdot 10^{-\alpha(V_{\text{m}}(t))/10}}, \qquad (3)$$

in which $\alpha(V_{\text{m}}(t))$ is the voltage-to-absorption function. We based $\alpha(V(t))$ on the one reported in [12] using a cubic spline fit (cf. Figure 3a). To model the signal-intensity induced chirp, we follow the model from [12], yielding following expression for the signal after the EAM,

$$E_{\text{EAM}}(t) = A_{\text{EAM}}(t) \cdot e^{j\left(\frac{\alpha}{2} \ln A_{\text{EAM}}^2(t)\right)}, \qquad (4)$$

in which α is the linewidth enhancement factor that controls the amount of chirp. The voltage-to-optical amplitude relationship and the absorption curve for the EAM used in the simulations can be seen in Figure 3. For both types of modulators, we keep the laser power fixed at a given level (cf. Table I) and only change V_{pp} in our simulations. Chromatic dispersion and fiber attenuation, associated with the SMF, is modeled using a standard frequency domain filter [25]. We use the laser and fiber parameters from [12] (cf. Table I). The signal is then detected using a square-law detection, with thermal noise and shot noise both modeled as white Gaussian random variables with variances $\sigma_{\text{thermal}}^2$ and σ_{shot}^2, respectively. The noise variances are parameterized as,

$$\sigma_{\text{thermal}}^2 = \frac{4 \cdot k \cdot T \cdot F_{\text{s}}}{B \cdot Z}, \qquad (5)$$

$$\sigma_{\text{shot}}^2 = \frac{2e_{\text{c}}\left(R \cdot \mathbb{E}[|y|^2] + I_{\text{d}}\right) F_{\text{s}}}{B}, \qquad (6)$$

in which k is the Boltzmann constant, T is the temperature of the photodiode in Kelvin, F_{s} is the sampling frequency, B is the bandwidth of the photodiode, Z is the impedance load, e_{c} is the electron charge, R is the responsivity of the photodiode, $\mathbb{E}[|y|^2]$ is the empirical average power received and I_{d} is the dark current. At the receiver, the signal is converted to digital domain using an ADC with the same low-pass filter characteristics as the DAC. The receiver filter is then applied and the signal is downsampled to 1 sps and the resulting signal is normalized to match the average power of the constellation. An overview of the parameters and the values used in our experiments can be seen in Table I.

The training and evaluation stages follow the same structure as in the previous section, namely, during training the MSE is calculated between the symbol sequence of interest and the output of the normalization block. The filters are updated using gradient descent as described in Algorithm 1. During evaluation, we add a quantization step, with 2^5 levels, to the DAC and ADC. The downsampled output of the receiver filter is mapped to its closest symbol in the constellation and the SER is calculated.

Component	Parameter Name	Value
System	Samples per symbol (sps)	8
	Symbol rate (R_s)	100 GBd
	Constellation	PAM-4
	Bit rate (with 5.8% FEC overhead)	189 GBit/s
DAC	DAC gain (g_{dac})	-
	3dB cutoff frequency (f_{3dB})	45 GHz
	Bias voltage (V_{b})	-
	Voltage peak-to-peak (V_{pp})	-
Modulator	Laser power ($P_{\text{in}}^{(\text{ideal})}$)	-13.0 dBm
	Laser power ($P_{\text{in}}^{(\text{EAM})}$)	-6.0 dBm
	Linewidth enhancement (α)	1.0
	Laser wavelength (λ)	1270 nm
Fiber	Dispersion slope (S_0)	0.092 ps/(nm²·km)
	Zero-dispersion wavelength (λ_0)	1310 nm
	Attenuation (α_{smf})	0.2 dB/km
	Dispersion parameter (D)	-15.43 ps/(nm·km)
Photodiode	Boltzmann constant (k)	1.38×10^{-23} J/K
	Temperature (T)	293 K
	Bandwidth (B)	45 GHz
	Impedance load (Z)	50 Ω
	Electron charge (e_{c})	1.6×10^{-19} C
	Responsivity (R)	1 A/W
	Dark current (I_{d})	1×10^{-8} A
ADC	3dB cutoff frequency (f_{3dB})	45 GHz

TABLE I: Parameters used in the IM/DD system. Parameter values marked with '-' are either learned or varied during simulations.

D. Wavelength division multiplexing

The minimization of the MSE, which is used for learning the pulse-shaper and receiver filter coefficients, does not put any constraints on the bandwidth of the learned filters directly. It is thereby expected that filters will occupy parts of the frequency spectrum outside the information bandwidth [17] (see also Section III-A). This will in turn make this framework much less desirable for wavelength division multiplexing (WDM) applications.

To investigate this, we implemented a WDM scheme with three channels—one channel of interest, on which the pulse-shaper and receiver filter coefficients are learned and two interfering channels (see Figure 4). This scheme was employed both during training and evaluation to make the learned filters robust towards neighboring channels. Three independent symbol sequences are generated and processed by the transmitter. Note that the same, learned pulse-shaping filter, is applied to all three channels. The three independently modulated optical signals are then multiplexed with channel spacing f_0 after which the fiber model (either AWGN or IM/DD) is applied.
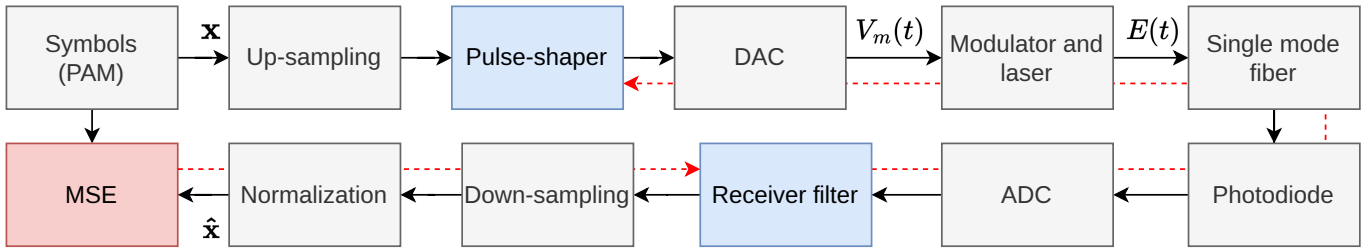
(a) Training mode. Pulse-shaper and receiver filters are updated using backpropagation, visualized with the dashed red line.
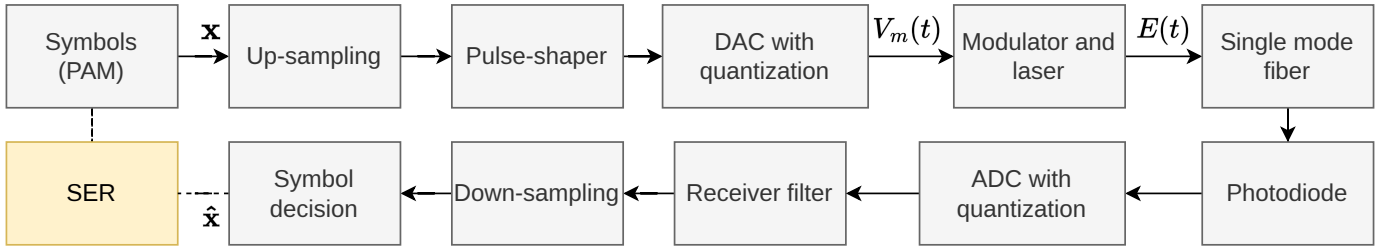

(b) Evaluation mode. Pulse-shaper and receiver filters are fixed to the values from last iteration of the training phase.

Fig. 2: System model for the simulated IM/DD link. When filters are optimized (training) they are done so using the mean square error (MSE) between the output of the system and the symbol sequence, depicted in Figure 2a. During evaluation (depicted in Figure 2b), a new symbol sequence is generated, filters are fixed and the sybmol error rate (SER) is calculated. Additionally, quantization in the ADC and DAC is applied in the evaluation step.
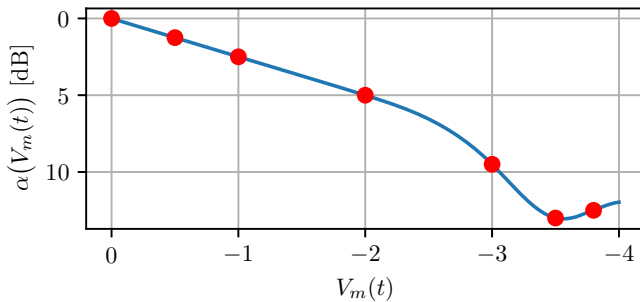
(a) Absorption curve $\alpha\left(V(t)\right)$. Red dots indicate the knee-points used to make the interpolation.
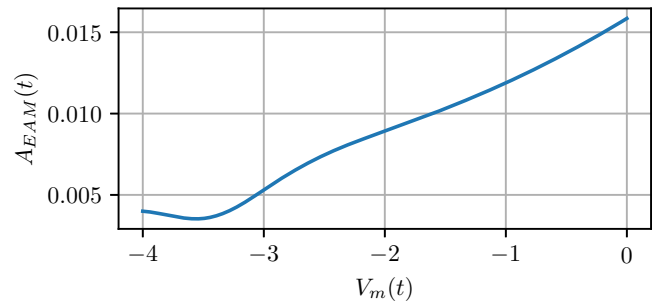
(b) Relation between voltage and optical amplitude given by equation (3).

Fig. 3: Electro-absorption modulator model used in the IM/DD system. Absorption curve was derived from [12] using a cubic spline fit.

After the fiber, we isolate the middle channel using a 5th order Gaussian frequency domain filter with a given bandwidth (see section III). Receiver side processing is then applied, after which we calculate the SER (for the middle channel only).

III. RESULTS

We now investigate the performance of jointly optimizing the pulse-shaper and the receiver filter within the two simulation models described. To make a fair comparison, we consider the following four configuration variants:

PS where the pulse-shaping filter is learned and the receiver filter is held fixed at RRC.

RxF where the pulse-shaping filter is held fixed at RRC and the receiver filter is learned.

PS & RxF where both pulse-shaping filter and receiver filter are learned.

RRC & FFE where both pulse-shaping filter and receiver filter are held fixed at RRC and a linear feed-forward equaliser (FFE) is learned after the receiver filter but before downsampling.

During simulations, we vary the length of the filters but always keep the pulse-shaping filter and receiver filter at the same number of taps N. In the case where an equaliser is used (RRC & FFE), we run the equaliser in the oversampled domain (i.e.
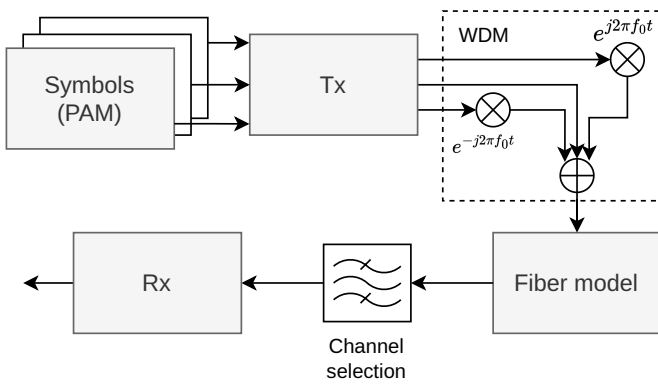
Fig. 4: Wavelength division multiplexing (WDM) applied to both the AWGN and IM/DD systems during training and evaluation. The same pulse-shaping filter is applied to all channels in the Tx. We only decode the channel centered at 0 Hz.

samples per symbol is equal to the oversampling rate) and use N (additional) taps.

A. Additive White Gaussian Noise Channel with Bandwidth Limitation

We compare the four learning methods on the AWGN channel described in section II-B. We vary the SNR and plot the SER for different filter lengths, which can be seen in Figure 5. Note that all filter optimization is carried out at an SNR of 12 dB. The choice of SNR does not influence the learned filters as the channel transfer function is linear. The theoretical curve is based on an AWGN channel without bandwidth limitation. The joint optimization framework (PS & RF) comes extremely close to the theoretical limit even for very short filter lengths. The penalty to the theory almost vanishes with a filter length of $N = 25$. All other methods incur substantial penalties irrespective of the filter length. The baseline (RRC & FFE) has a slight advantage over the other single-sided methods, which is attributed to the extra taps in the FFE. However, this advantage reduces as we increase the filter lengths.

To illustrate what the different methods learn, we show the pulse-shaping amplitude response and the Nyquist zero ISI criterion [4] for the total system response (i.e. pulse-shaper, DAC, receiver filter and ADC) in Figure 6. Given the total frequency response of the system, $H(f)$, then we are interested in analyzing the following term,

$$B(f) = \sum_{m=-\infty}^{\infty} H\left(f + \frac{m}{T}\right). \tag{7}$$

in which T is the sampling interval. In an ISI-free system the quantity $B(f)$ is constant over all frequencies [4]. We see in Figure 6b that the joint optimization scheme achieves the flattest curve, which explains the near-optimal symbol-error performance in the AWGN channel (see Figure 5). Furthermore, we see in Figure 6a that the methods that learn the pulse-shaper (PS and PS & RxF) utilize a much wider

range of frequencies compared to the RRC filter. This is due to the unconstrained nature of our cost function in the optimization problem [17].

B. Additive White Gaussian Noise Channel with Bandwidth Limitation and Wavelength Division Multiplexing

We now study what happens if we include WDM in both training and evaluation and vary the channel spacing f_0. To accommodate for three channels, each with bandwidth of 100 GHz, we increase the oversampling rate from 4 to 8. We set the bandwidth of the Gaussian channel selection filter to be $f_{3dB} = 65$ GHz. We tried different values for f_{3dB} between 55 and 65 GHz and the relative performance between methods did not significantly change. The performance of the systems can be seen in Figure 7. For a fixed channel spacing of 150 GHz (Figure 7a), we see that the relative performance of the methods is similar to the previous section. Again, the join-optimization scheme is consistently better and can close the gap to the theoretical limit with a filter length of 65. If we look at the performance when varying the channel spacing (Figure 7b), we naturally see a degradation as the channels are spaced more closely together. The joint optimization PS & RxF is still consistently better compared to the other configurations across the board. However, the benefit of optimizing both the PS and RxF filters jointly is only marginal in the low channel spacing regime (< 125 GHz).

C. Intensity Modulated and Direct Detected System

We now turn to IM/DD system with WDM. During simulations we vary the peak-to-peak voltage, V_{pp}, in the DAC while keeping laser power constant. We test the system in a back-to-back (B2B) scenario and for a fiber length of 1.0 km. In all simulations, we learn the normalization constant g_{dac} as part of the optimization such that all methods utilize the full voltage range. For the IM/DD system, we compare the performance of our methods to the pre-forward error correction threshold KP4-FEC, which we in symbol-error rate approximate by $KP4_{SER} \approx \log_2(4) \cdot 2.4 \cdot 10^{-4}$. The results of the numerical simulations with an ideal modulator can be seen in Figure 8 and 9. For the ideal modulator we fix the bias in the DAC such that the voltages have range $[0, V_{pp}]$. We set the bandwidth of the Gaussian channel selection filter to be $f_{3dB} = 55$ GHz. Due to the ideal modulator, we observe a linear improvement in the SER as a function of V_{pp} for all the methods, with the PS & RxF method having the best performance. As in the AWGN simulation, when varying the channel spacing we need at least a channel spacing of 125 GHz to see a meaningful relative improvement for the joint optimization method compared to the others. We note that the receiver-side methods (RxF and RRC & FFE) slightly *increase* in SER for higher filter lengths. We attribute this to the clipping constraint in the DAC, which will penalize filters with higher peak-to-average power ratio (PAPR). Because we increase the filter length of the pulse-shaper and receiver filter jointly, this will create pulse-shaper filters with higher and higher PAPR (due to the very low RRC rolloff $\rho = 0.01$). The methods that optimize the pulse-shaper have the flexibility to learn filters with lower PAPR that avoid
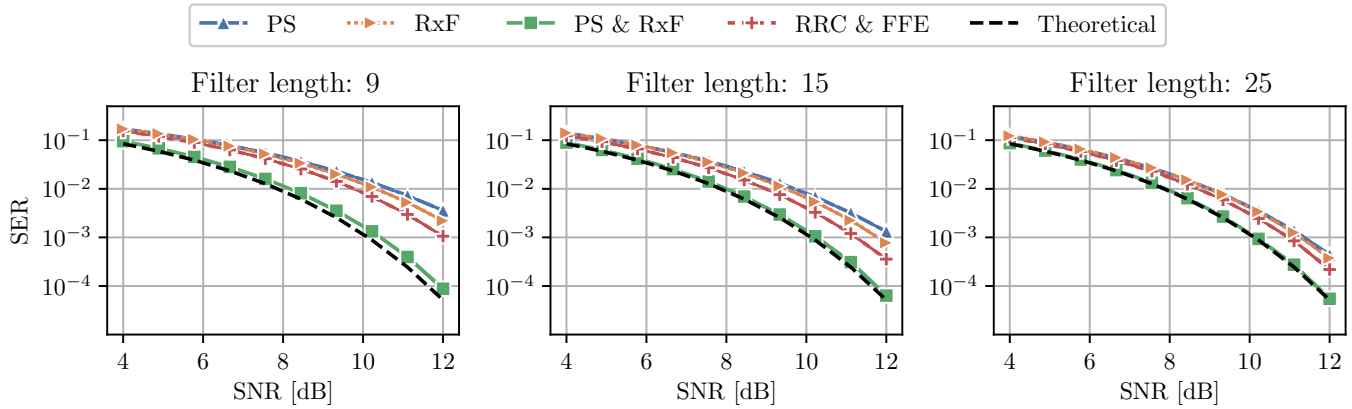
Fig. 5: SER a function SNR for different filter lengths in the AWGN channel with bandwidth limitation. Filters were trained at SNR = 12 dB and evaluated in the range from 4 to 12 dB

(a) Pulse-shaping filter amplitude response

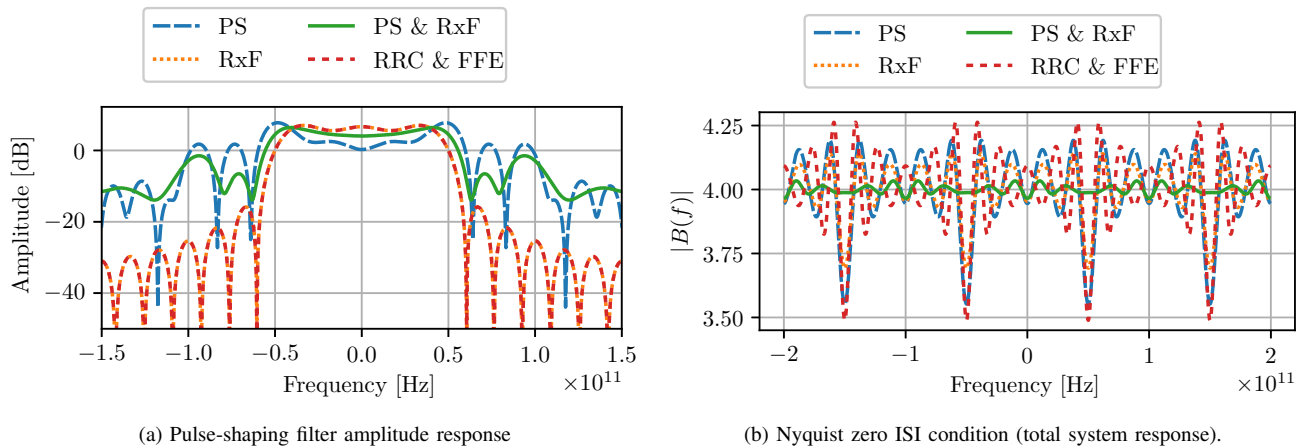(b) Nyquist zero ISI condition (total system response).

Fig. 6: AWGN: We show the amplitude spectrum of the pulse-shaper in 6a for the different configurations. In 6b, we show the Nyquist zero ISI criterion. In this simulation, filter lengths were set to 25 for both pulse-shaper and receiver filter.

clipping and thus do not suffer an SER penalty at higher filter lengths.

Finally, we investigate the learning framework with a non-linear modulator, namely the EAM (see input-output characteristic in Figure 3b). In all simulations with the EAM, in addition to the filters, we optimize g_{dac} and the bias, V_{b}, in the DAC such that each method has the flexibility to utilize the modulator range in the best way possible. We initialized the bias in modulator to $V_{\text{b}} = -1.0$ for all simulations as we found that to yield consistently good results. Due to the non-linear modulator characteristic, we additionally compare the filter optimization methods with a second order Volterra equalizer. We set the number of taps in the first order filter to 101 and 45 taps in the second order, yielding 1136 free parameters. We use the non-linear equalizer in two configurations — 1) together with a fixed RRC filter on the Tx side and a fixed RRC filter on the Rx side denoted RRC & Volterra* and 2) together with a learnable pulse-shaper and fixed RRC filter on the Rx side denoted PS & Volterra*. The two Volterra* methods can viewed as a best-case comparison, where we have not taken into account computational complexity constraints. The

results of the numerical simulations can be seen in Figure 10 and 11. In the B2B condition in Figure 10, due to the non-linear characteristic of the EAM, we see that the SER curve flattens as we increase the peak-to-peak voltage, except for the Volterra* methods that continue to improve. The relative performance between the methods is consistent with previous results, namely that the joint-learning framework outperforms the "single-side" frameworks. One might expect that the Volterra equalizer alone (RRC & Volterra*) should be enough to reach the optimum performance, with a clear advantage in the number of free parameters over the other standard filtering methods (PS, RxF and PS & RxF). We attribute this to the ISI of the system stemming from bandwidth limitation being the limiting factor, and *not* the non-linearity of the system. In a channel where the non-linearities are more dominant we would expect the Volterra alone to yield better results, but this was not the focus of the paper. If we additionally optimize the pulse-shaper, as represented by the PS & Volterra method, we observe further gains in terms of performance compared to the linear approach (PS & RxF). When we increase the fiber length to 1 km, we see a dramatic drop in performance and more

(a) SER as a function SNR with channel spacing $f_0 = 150$ GHz.

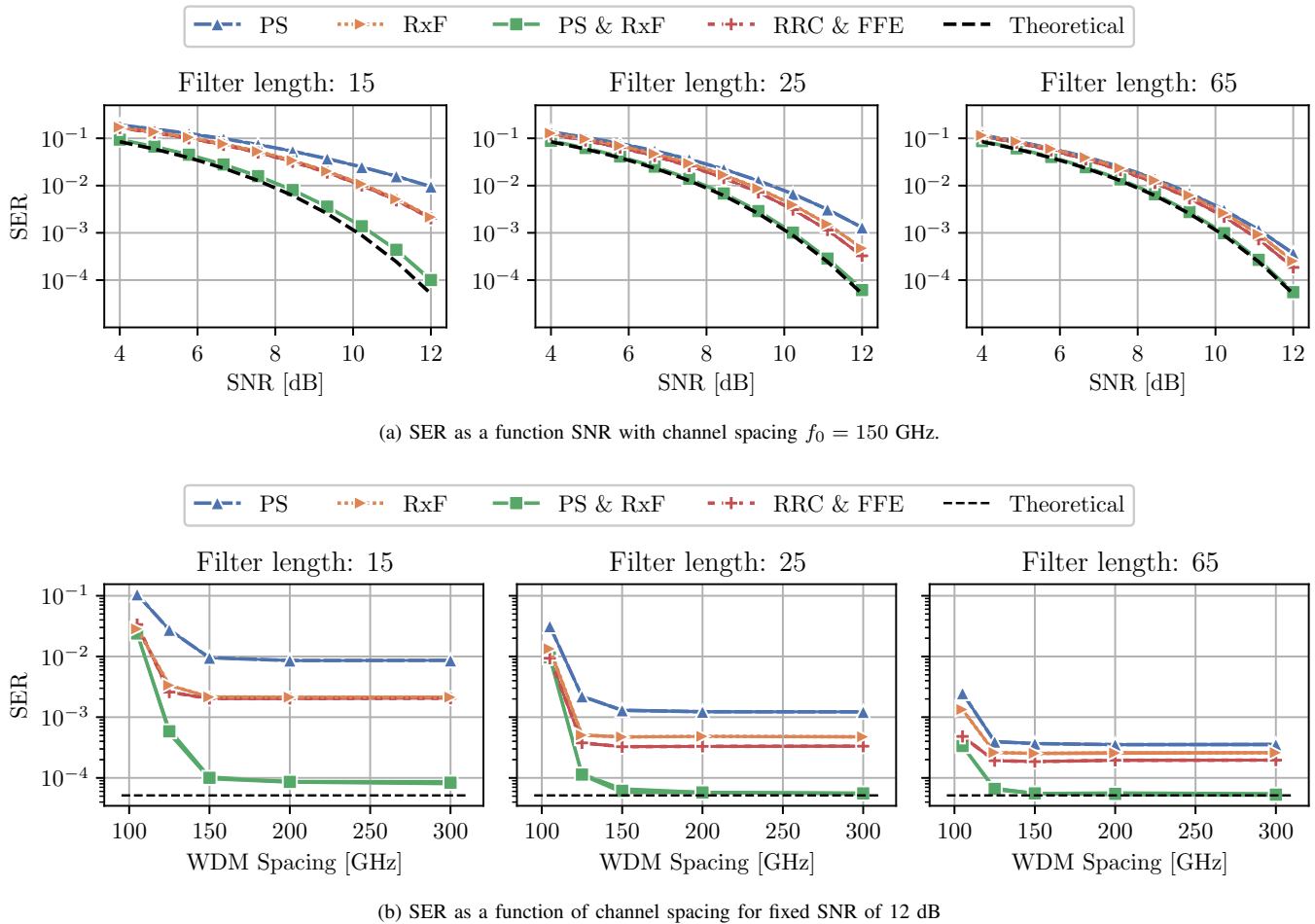(b) SER as a function of channel spacing for fixed SNR of 12 dB

Fig. 7: Performance plots for the AWGN channel with bandwidth limitation and wavelength division multiplexing. In Figure 7a, we have varied the SNR for a fixed channel spacing of $f_0 = 150$ GHz and in Figure 7b we fixed the SNR to 12 dB and varied the channel spacing.

linear behavior as a function of V_{pp}. We primarily attribute this to the phase noise introduced by the chirp in the modulator that significantly impacts the performance of all methods. Similar to the other simulations, the channel spacing in the WDM needs to be at least 125 GHz for the end-to-end learning framework to have an advantage over the other methods (see Figure 11). However, we note that the for a fiber length of 1 km all methods struggle to reach good SER performance and the difference over channel spacing is relatively low (especially for short filters). As in the simulations with the ideal modulator, the "receiver-side only" methods (RxF, RRC & FFE and RRC & Volterra*) all incur a penalty when increasing the filter length from 25 to 65. As already discussed in the previous simulation, we attribute this to the clipping constraint in the DAC and the high PAPR of the RRC filter at the Tx. An example of eyediagrams after receiver filtering in the 1 km case can be seen in Figure 12. The joint optimization (PS & RxF) achieves the clearest eye-opening compared to the other methods.

1) Filter Robustness Evaluation with Non-linear Fiber Model: Robustness of end-to-end frameworks will always be of interest as deviations from the simulation model will lead

to performance degradation, unless specifically accounted for during training [14], [15]. To investigate the learned pulse-shaper and receiver filters' robustness and generalizability, we devised a small simulation in which the channel conditions are altered during evaluation.

The E2E systems under test, are trained on the IM/DD WDM system described above with a 1 km SMF with parameters described in Table I. We used 3 WDM channels with a channel spacing of 150 GHz. During evaluation, we introduce a non-linear fiber with the split-step Fourier method (SSFM) with a non-linearity coefficient of $\gamma = 1.3$ [1/W/km]. The step size was set to 0.25 km and we did not use any amplification. We introduce 2 additional interferer channels, bringing the total number of channels up to 5, and sweep the launch power that enters the fiber. The results of this simulation for PS, RxF and PS & RxF can be seen in Figure 13. Overall, all the methods follow the same trend, namely that the SER decreases until the non-linear distortion from the fiber becomes dominant and we start to see the SER increase. The joint optimization method is across the board the best method, even in the highly non-linear regime, however, it also sees the biggest increase in SER compared relative to the optimum performance.
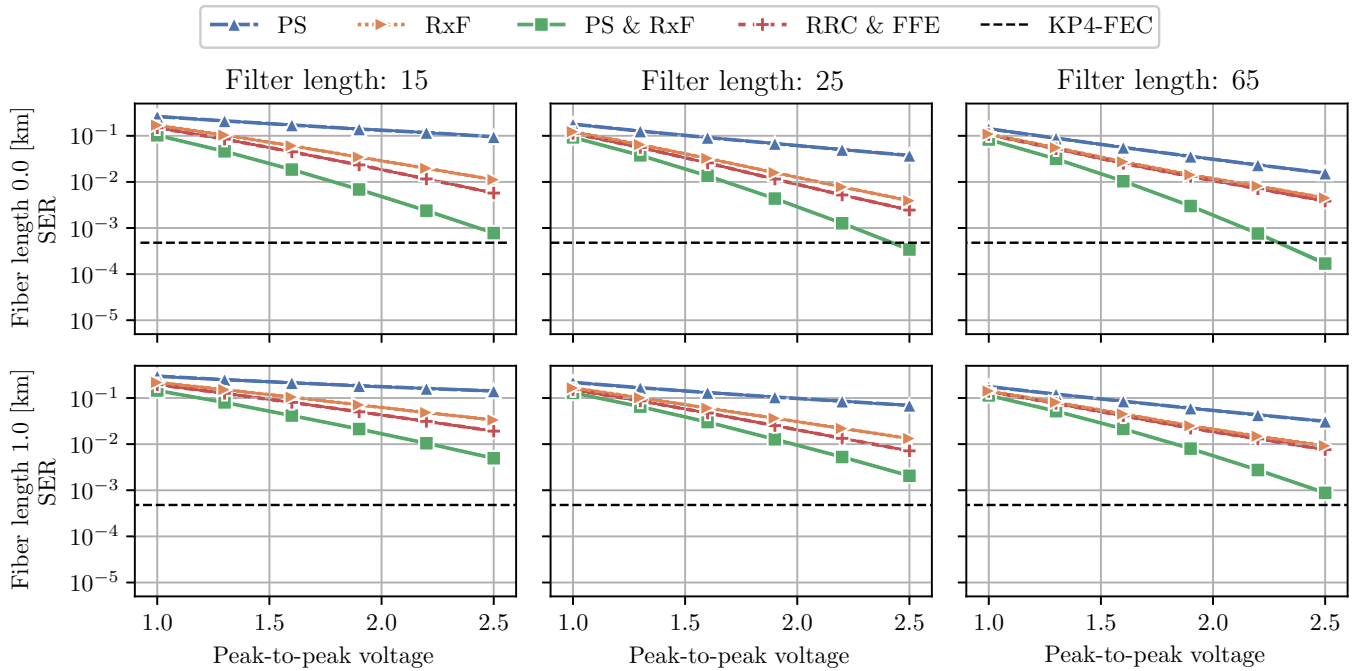
Fig. 8: IM/DD system with an ideal linear modulator and channel spacing 200 GHz: SER vs V_{pp} in DAC for different fiber and filter lengths.
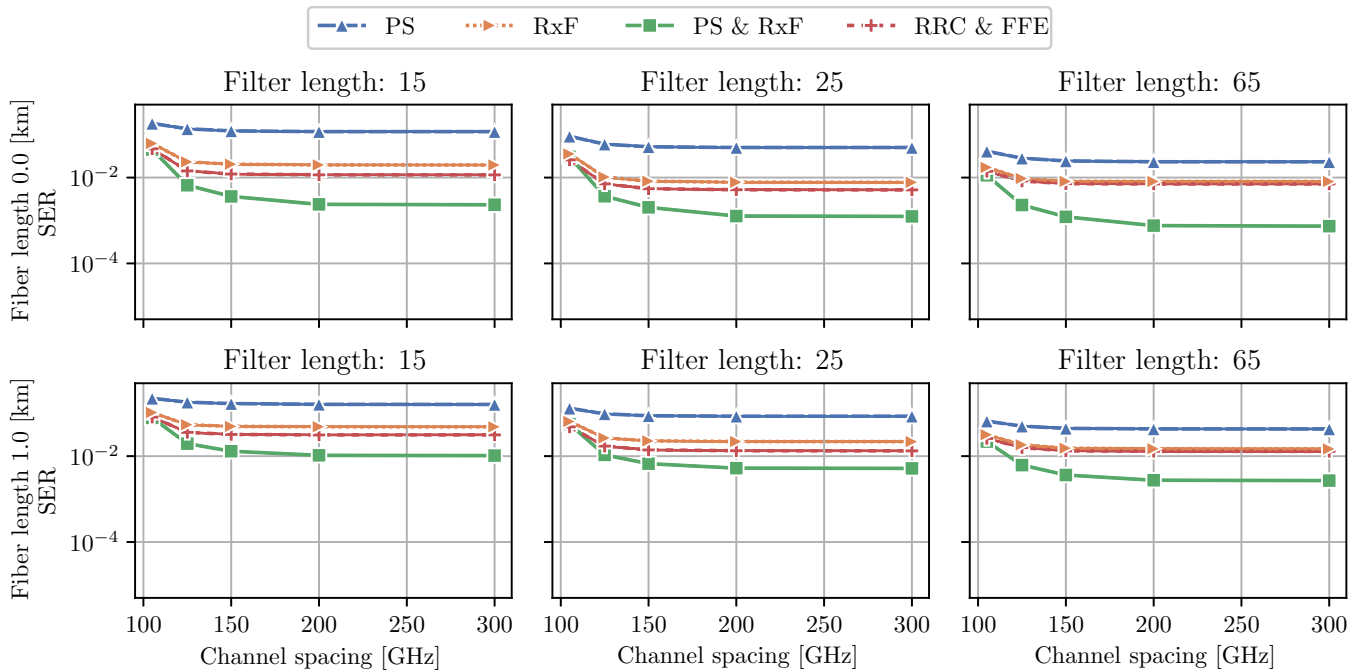

Fig. 9: IM/DD system with ideal modulator: SER vs WDM channel spacing for different fiber and filter lengths with $V_{pp} = 2.2$ in DAC.
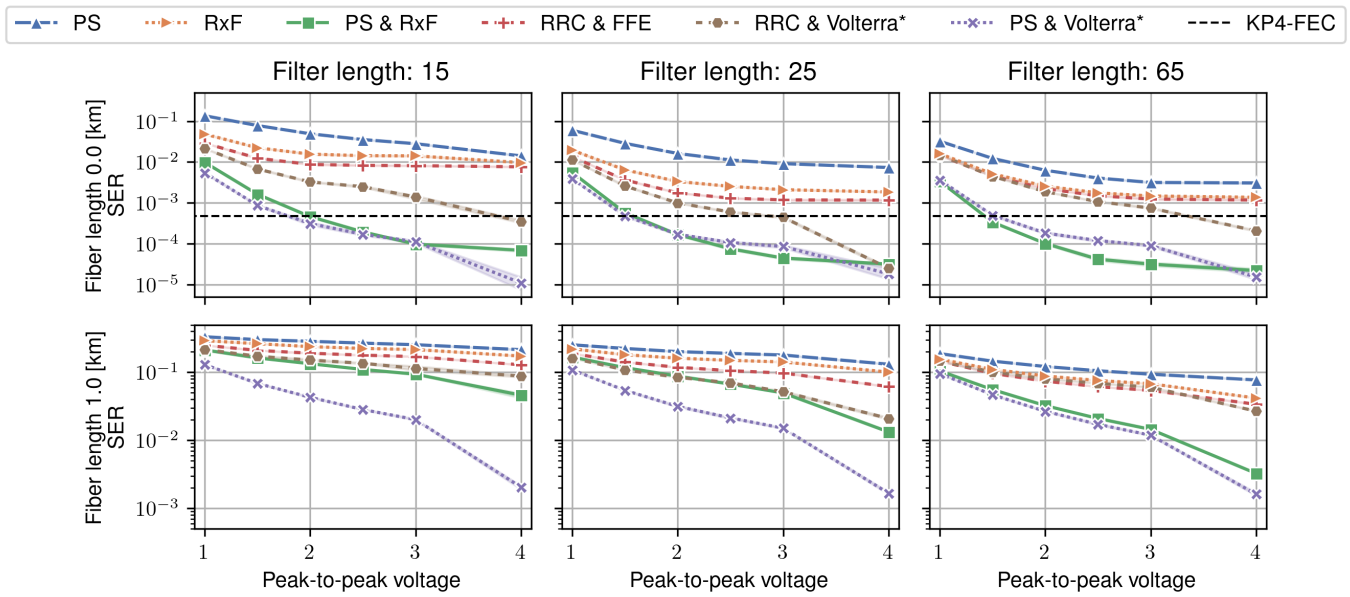
Fig. 10: IM/DD system with EAM and channel spacing 200 GHz: SER vs V_{pp} in DAC for different fiber and filter lengths. Note, the Volterra* methods has many more free parameters than the other methods and are only meant as a theoretical comparisons. The shaded area represents a 95% confidence interval estimated over 5 random restarts of the simulation.

Fig. 11: IM/DD system with EAM modulator: SER vs WDM channel spacing for different fiber and filter lengths with $V_{\text{pp}} = 4.0$ in DAC. Note, the Volterra* methods has many more free parameters than the other methods and are only meant as a theoretical comparisons. The shaded area represents a 95% confidence interval estimated over 5 random restarts of the simulation.
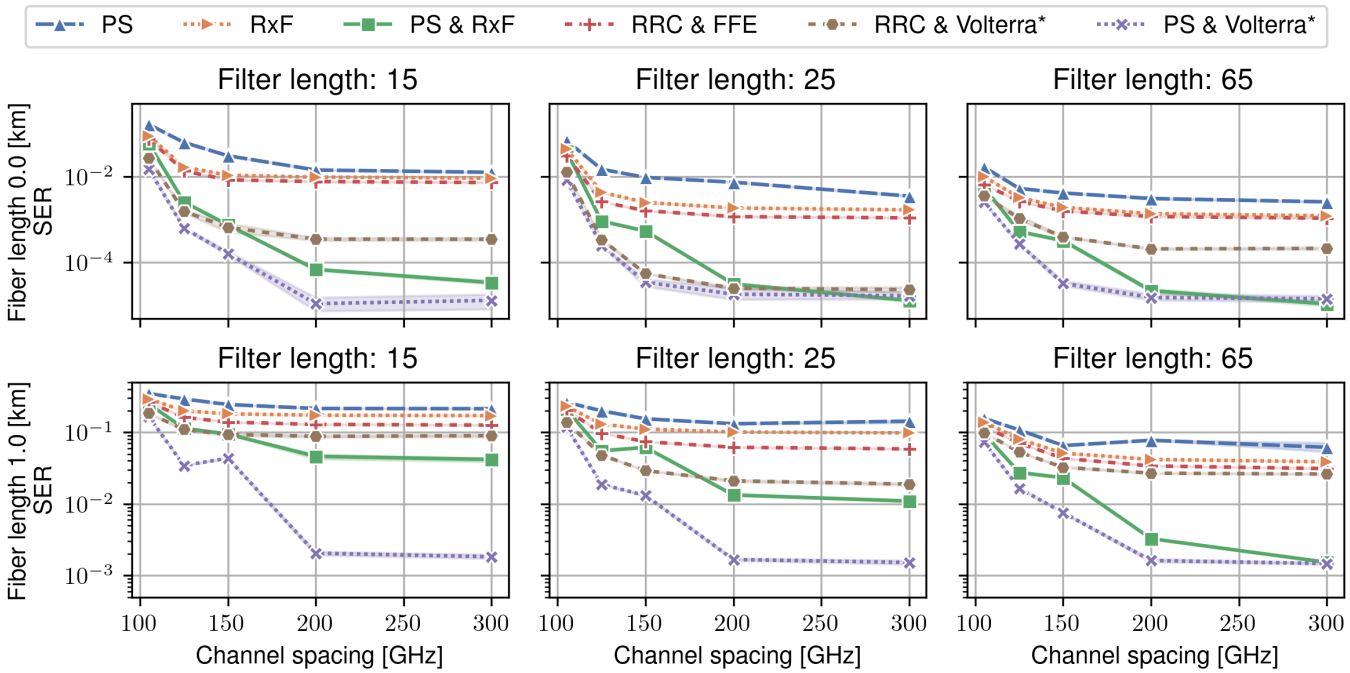
(a) Optimized pulse-shaper

(b) Optimized receiver filter

(c) Joint optimization of pulse-shaper and receiver filter
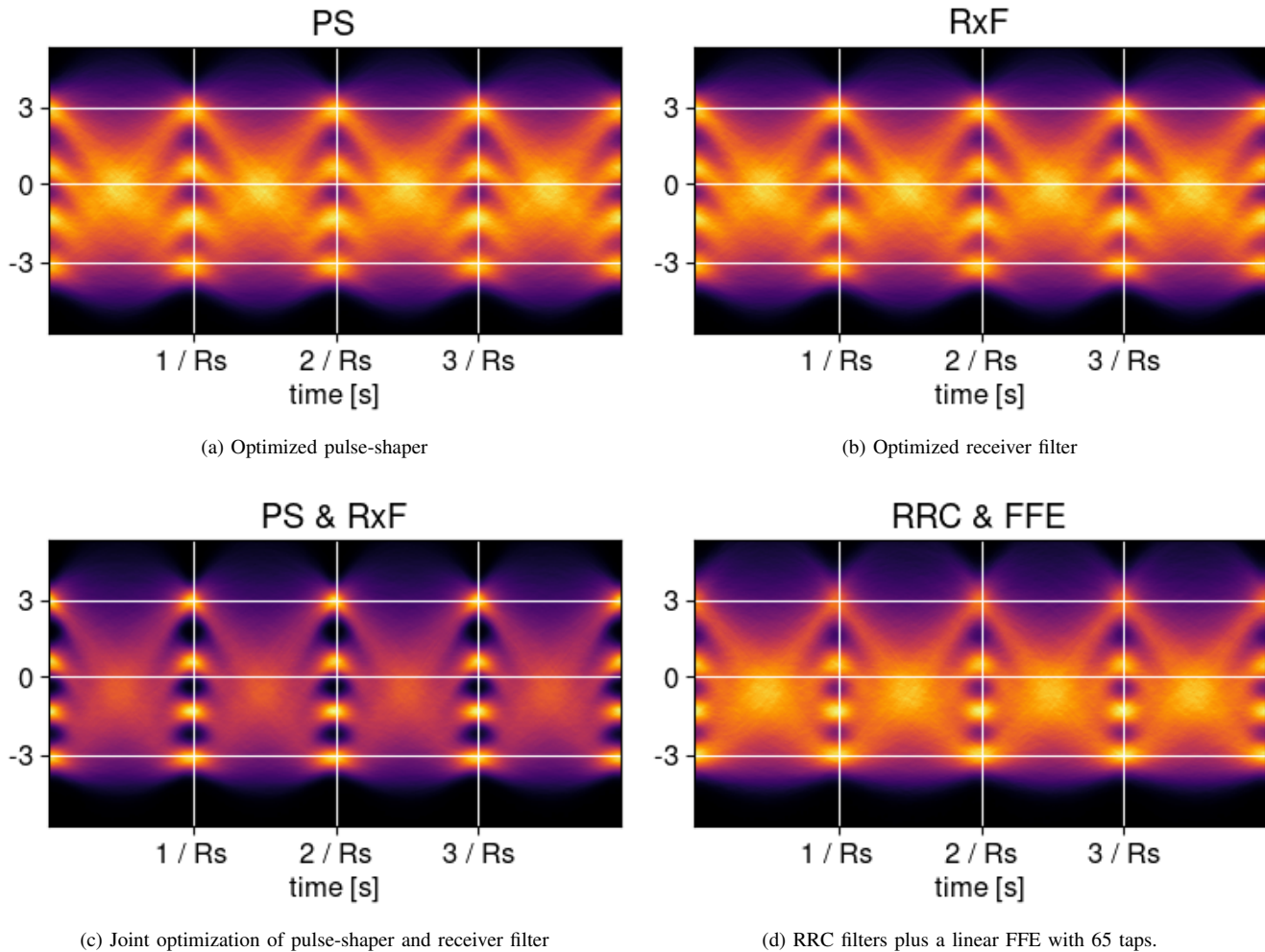
(d) RRC filters plus a linear FFE with 65 taps.

Fig. 12: Eyediagrams of the signal just before the downsampling in the IM/DD system. Fiber length was 1.0 km, $V_{pp} = 4.0$ and filter lengths of 65 (with an additional 65 equalizer taps in the FFE case).

IV. CONCLUSION

We investigated the use of end-to-end learning for joint optimization of pulse-shaping and reciver-side FIR filters to mitigate ISI in bandwidth-limited communication channels. By implementing the channels in PyTorch, with the use of automatic differentiation we numerically investigated the performance gains in two types of systems; an AWGN channel with bandwidth-limitation and an IM/DD short-reach link with a non-linear modulator. We compared joint optimization to their respective single-sided counterpart, fixing one side to a RRC filter and learning the other side. We consistently found that joint optimization outperformed the other methods across both channels and SNRs. In the IM/DD system, we found support for short filter lengths ($N_{taps} = 15$) being able to respect the KP4 forward-error correction threshold which none of the other linear methods came close to. We hypothesize that joint optimizations' superiority comes from the additional degrees of freedom at the Tx, which results in the channel response, including hardware components, being more efficiently invertible. This points toward joint-optimization methods being able to deliver considerable savings in terms of

operating IM/DD links, at the (minor) cost of some calibration and optimization.

Our framework does not directly translate to an experimental setup, as it requires differentiating through the channel. However, we believe that a combination of good characterization of electrical and optical components together with approximate optimization routines will be able to realize the reported performance gains.

REFERENCES

[1] A. S. Andrae, "New perspectives on internet electricity use in 2030," *Engineering and Applied Science Letters*, 2020.

[2] E. Berikaa, M. S. Alam, S. Bernal, R. Gutierrez-Castrejon, W. Li, Y. Hu, B. Krueger, F. Pittala, and D. V. Plant, "Next-generation o-band coherent transmission for 1.6 tbps 10 km intra-datacenter interconnects," *Journal of Lightwave Technology*, vol. 42, no. 3, pp. 1126–1135, 2024.

[3] D. Che and X. Chen, "Modulation Format and Digital Signal Processing for IM-DD Optics at Post-200G Era," *Journal of Lightwave Technology*, vol. 42, no. 2, pp. 588–605, 2023.

[4] J. Proakis and M. Salehi, *Digital Communications*. McGraw Hill,, 2008.

[5] Q. Yu, "On the Decision-Feedback Equalizer in Optically Amplified Direct-Detection Systems," *Journal of Lightwave Technology*, vol. 25, pp. 2090–2097, Aug. 2007.
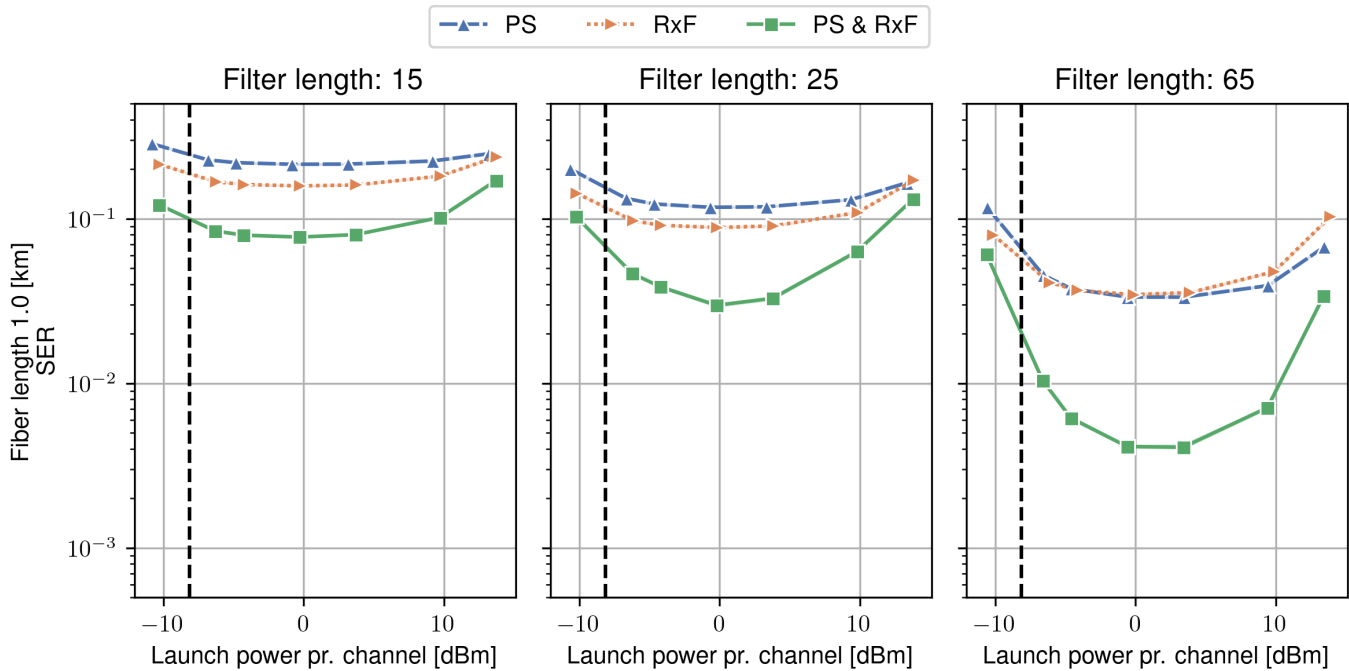
Fig. 13: Robustness simulation: The plots show the evaluation performance of the IM/DD WDM (5 channel) systems with a non-linear fiber ($\gamma = 1.3$ [1/W/km]) as a function of laser input power ($P_{\text{in}}^{(\text{EAM})}$). The systems were trained with 3 channels on a standard SMF at $P_{\text{in}}^{(\text{EAM})} = -6.0$ dBm, which in launch power is indicated by the dashed black vertical line. During both training and evaluation the channel spacing was set to 150 GHz.

[6] R. Rath, D. Clausen, S. Ohlendorf, S. Pachnicke, and W. Rosenkranz, "Tomlinson–Harashima Precoding For Dispersion Uncompensated PAM-4 Transmission With Direct-Detection," *Journal of Lightwave Technology*, vol. 35, pp. 3909–3917, Sept. 2017.

[7] N. Stojanovic, F. Karinou, Z. Qiang, and C. Prodaniuc, "Volterra and Wiener Equalizers for Short-Reach 100G PAM-4 Applications," *Journal of Lightwave Technology*, vol. 35, pp. 4583–4594, Nov. 2017.

[8] B. Karanov, P. Bayvel, and L. Schmalen, "Machine learning for short reach optical fiber systems," in *Machine Learning for Future Fiber-Optic Communication Systems*, pp. 65–89, Academic Press, Jan. 2022.

[9] C. B. Czegledi, M. R. Khanzadi, and E. Agrell, "Bandlimited Power-Efficient Signaling and Pulse Design for Intensity Modulation," *IEEE Transactions on Communications*, vol. 62, pp. 3274–3284, Sept. 2014.

[10] X. Wu, J. Zhang, A. P. T. Lau, and C. Lu, "C-band 120-Gb/s PAM-4 transmissions over a 100-km dispersion-uncompensated SSMF using joint combined pulse shaping and low-complexity nonlinear equalization," *Optics Letters*, vol. 47, pp. 5144–5147, Oct. 2022.

[11] V. Bajaj, F. Buchali, M. Chagnon, S. Wahls, and V. Aref, "Deep Neural Network-Based Digital Pre-Distortion for High Baudrate Optical Coherent Transmission," *Journal of Lightwave Technology*, vol. 40, pp. 597–606, Feb. 2022.

[12] E. Liang and J. Kahn, "Geometric Shaping for Distortion-Limited Intensity Modulation/Direct Detection Data Center Links," *IEEE Photonics Journal*, vol. 15, no. 6, pp. 1–17, 2023.

[13] T. O'Shea and J. Hoydis, "An Introduction to Deep Learning for the Physical Layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, pp. 563–575, Dec. 2017.

[14] B. Karanov, M. Chagnon, F. Thouin, T. A. Eriksson, H. Bülow, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-End Deep Learning of Optical Fiber Communications," *Journal of Lightwave Technology*, vol. 36, pp. 4843–4855, Oct. 2018.

[15] 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, pp. 3316–3324, May 2022.

[16] V. Neskorniuk, A. Carnio, D. Marsella, S. K. Turitsyn, J. E. Prilepsky, and V. Aref, "Memory-aware end-to-end learning of channel distortions in optical coherent communications," *Optics Express*, vol. 31, pp. 1–20, Jan. 2023.

[17] J. Song, C. Hager, J. Schroder, A. G. I. Amat, and H. Wymeersch, "Model-Based End-to-End Learning for WDM Systems With Transceiver Hardware Impairments," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 28, pp. 1–14, July 2022.

[18] S. Hernandez, C. Peucheret, F. D. Ros, and D. Zibar, "End-to-end optimization of optical communication systems based on directly modulated lasers," *Journal of Optical Communications and Networking*, vol. 16, pp. D29–D43, Aug. 2024.

[19] D. Marasinghe, L. H. Nguyen, J. Mohammadi, Y. Chen, T. Wild, and N. Rajatheva, "Waveform Learning under Phase Noise Impairment for Sub-THz Communications," *IEEE Transactions on Communications*, pp. 1–1, 2024.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[22] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), pp. 1310–1318, PMLR, June 2013.

[23] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, pp. 369–386, SPIE, May 2019.

[24] A. D. Gallant and J. C. Cartledge, "Characterization of the Dynamic Absorption of Electroabsorption Modulators With Application to OTDM Demultiplexing," *Journal of Lightwave Technology*, vol. 26, pp. 1835–1839, July 2008.

[25] G. P. Agrawal, *Fiber-Optic Communication Systems*. Wiley-Interscience, a John Wiley &Sons, 3rd edition ed., 2002.