End-to-end optimization of coherent optical communications over the split-step Fourier method guided by the nonlinear Fourier transform theory

Gaiarin, Simone; Daros, Francesco; Jones, Rasmus Thomas; Zibar, Darko

Published in:
Journal of Lightwave Technology

Link to article, DOI:
10.1109/JLT.2020.3033624

Publication date:
2021

Document Version
Peer reviewed version

End-to-end optimization of coherent optical communications over the split-step Fourier method guided by the nonlinear Fourier transform theory

Simone Gaiarin, Francesco Da Ros, Rasmus T. Jones, and Darko Zibar

***Abstract*—Optimizing modulation and detection strategies for a given channel is critical to maximizing the throughput of a communication system. Such an optimization can be easily carried out analytically for channels that admit closed-form analytical models. However, this task becomes extremely challenging for nonlinear dispersive channels such as the optical fiber. End-to-end optimization through autoencoders (AEs) can be applied to define symbol-to-waveform (modulation) and waveform-to-symbol (detection) mappings, but so far it has been mainly shown for systems relying on approximate channel models. Here, for the first time, we propose an AE scheme applied to the full optical channel described by the nonlinear Schrödinger equation (NLSE). Transmitter and receiver are jointly optimized through the split-step Fourier method (SSFM) which accurately models an optical fiber. In this first numerical analysis, the detection is performed by a neural network (NN), whereas the symbol-to-waveform mapping is aided by the nonlinear Fourier transform (NFT) theory in order to simplify and guide the optimization on the modulation side. This proof-of-concept AE scheme is thus benchmarked against a manually-optimized NFT-based system and a three-fold increase in achievable distance (from 2000 to 6640 km) is demonstrated.**

***Index Terms*—auto-encoder, modulation, detection, nonlinear frequency division multiplexing, nonlinear Fourier transform**

I. Introduction

OPTICAL communication systems are striving to maximize the achievable information rate distance product. In order to achieve this goal, optimizing the signal constellation (e.g. constellation shaping [1]–[4]) may not be sufficient and modulation and detection strategies need to be jointly improved. Whereas for simple transmission channels, such as additive white Gaussian noise (AWGN) channels, it is possible to analytically derive optimal modulation and detection strategies, such an optimization is particularly challenging for the nonlinear optical channel. As a closed-form expression to describe the signal propagation through an optical fiber is not available, current research relies on approximate analytical or numerical models achieving only a limited degree of accuracy [5]. Even following this direction though, no definitive answer to optimal modulation and detection strategies is known. In order to address these challenges, full end-to-end learning through autoencoders (AEs), which does not require closed-form channel models, has been proposed [6]. The first proposal analyzing an AWGN channel [6] was followed by a number of works focusing on the optical fiber channel, e.g. [4], [7]–[10], however, all relying on approximate channel models. A general AE model of a coherent communication system (complex signal transmission) targets using the full nonlinear dispersive channel model and replacing the transmitter and the receiver with neural networks (NNs). After the training phase, this model can provide two key functions: the optimal encoding of symbols onto time-domain waveforms resilient to the optical channel impairments; the decoding of the received waveforms to recover the transmitted symbols. An ideal AE should therefore both be trained on an accurate channel model, as well as provide symbol-to-waveform and waveform-to-symbol transformation.

In this work, we extend our initial proposal of [11], discussing the first numerical analysis of an AE scheme for coherent communication making use of the split-step Fourier method (SSFM) to accurately model the optical fiber channel. However, in order to restrict the space of all possible modulation strategies, in this work, we demonstrate our proposed method by guiding the optimization through the mathematical theory of the nonlinear Fourier transform (NFT) [12], [13]. Our proposed AE scheme jointly optimizes an NFT-aided transmitter with an NN-based receiver over the channel modeled by solving the nonlinear Schrödinger equation (NLSE) through the SSFM. This choice results in pulse-shapes (solitons) which do not suffer from significant pulse broadening over propagation, therefore enabling a small (low-complexity) and memoryless (1-symbol in input) NN to be used at the receiver. Due to the system similarity with nonlinear frequency division multiplexing (NFDM) systems, the proposed scheme is benchmarked against standard NFDM transmission (NFT transmitter and receiver). This proof-of-concept demonstration shows more than three times extension in transmission reach compared to a manually-optimized NFDM system. The 2000 km achieved by standard NFDM transmission are extended to more than 6000 km through the end-to-end optimization.

The remaining of the paper is organized as follows. In Section II the concept of AE is reviewed, discussing its application to communication systems and positioning this work within the state-of-the-art on this topic. In Section III the numerical setup is presented, including a brief review of the fundamentals of NFT. The specific system model and its implementation are also described in detail. In Section IV the training procedure

Manuscript received April 1st, 2020;

S. Gaiarin, F. Da Ros, and D. Zibar are with the Department of Photonics Engineering, Technical University of Denmark, Kongens Lyngby, 2800 Denmark, e-mail: {simga,fdro,dazi}@fotonik.dtu.dk

R. T. Jones was with the Department of Photonics Engineering, Technical University of Denmark, Kongens Lyngby, 2800 Denmark. He is now with Oticon, Smørum, 2765 Denmark, email: rajo@fotonik.dtu.dk

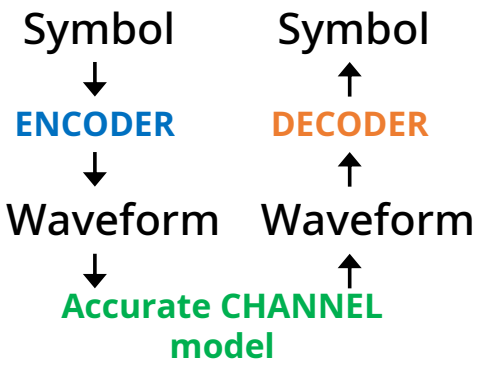


Fig. 1. Ideal autoencoder model for an optical communication system.

used is detailed and key practical trade-offs are identified. The key transmitter and receiver parameters that have been optimized are introduced, and the different optimization cases are presented. In Section V the main results of this work are reported, starting with the results of the optimization and the training performance and following with the performance of the system during the testing phase. The different optimized transmission schemes are compared against each other and benchmarked against a conventional NFDM system. Finally, the conclusions are summarized in Section VI.

II. End-to-end communication systems

A communication system is composed of three main blocks: a transmitter, a channel, and a receiver. The goal of the system is to faithfully reproduce the information entered at its input to its output, i.e. to estimate the transmitted symbol sequence with the lowest error probability. The structure of the communication system resembles that of an AE [6], as shown in Fig. 1. In its general definition, an AE consists of two key transformations: an encoder function that maps the input data to a *code*, i.e. an encoded version of the data, and a decoder function that tries to reconstruct the original data from the code [14]. According to the specific problem and context, different AE architectures can be chosen to generate codes optimized according to different metrics, e.g. lower dimensionality compared to the input data, or robustness to noise and distortion sources. For the case of a communication system, the transmitter, i.e. the encoder, maps the information (symbols) to a time-domain waveform, i.e. the code, that is transmitted over the channel. The channel is the source of noise and distortion. The receiver, i.e. the decoder, is responsible to recover the original data symbols from the corrupted version of the code, which is the waveform after the transmission. The goal for training the transmitter and the receiver of this communication system is to minimize the error probability of the transmitted symbols, and it is achieved by generating transmitter waveforms that are resilient to the noise and distortions introduced by the channel.

The ideal AE for optical coherent communication systems needs to consider the full nonlinear dispersive channel model and can be constructed by replacing the transmitter and the receiver with NNs. Indeed a sufficiently large NN can theoretically approximate any function, including the one that generates this optimal set of waveforms [15]. Nonetheless, to achieve this final goal the correct optimization strategy and NN architecture must be devised, and this is challenging to do in practice, especially for non-trivial channels such as the optical fiber. Therefore, the preliminary demonstrations reported for optical communications have considered a simplified version of the overall structure of Fig. 1. All the works reported consider approximate channel models such as a simplified memoryless nonlinear channel model [7], a dispersive linear fiber channel model [9], [10], and perturbative models of the nonlinear fiber channels [4], [8]. End-to-end learning for the full nonlinear dispersive fiber channel, i.e. not relying on perturbative/approximate models, has yet to be reported. In order to approach the desired system shown in Fig. 1, in this work, the AE considers the SSFM to numerically implement the physical channel. The encoder and decoder are therefore trained over a highly accurate representation of the channel, moving one step closer to the final goal of Fig. 1.

Concerning the encoder, however, of the available literature, only the work from [9], [10] discussed full encoding, i.e. from symbols to waveforms, whereas the other reports focused on mapping from bit/symbol to complex constellations, and assumed conventional pulse shaping. A fully blind AE consisting of two NNs incurs in the challenge of a vast and complex optimization landscape. In [9], [10], the optimization was aided by considering intensity-only modulation and, therefore, halving the dimensionality of the problem. Here, in order to avoid this challenge worsened by the even more complex optimization landscape introduced by the accurate channel model, the receiver/decoder is replaced by an NN whereas the transmitter/encoder is implemented as a conventional NFDM transmitter with trainable parameters. The NFT theory helps in guiding the optimization of the encoder by liming the solutions space of the time-domain waveforms generated by the transmitter. In particular, the NFDM transmitter is constrained to solitonic pulse-shapes, which are exact analytical solutions of the NLSE in absence of loss, and have been shown to provide appropriate pulse shapes also for practical transmissions with losses [16], [17]. Moreover, given that solitons do not disperse with the transmission distance (the pulse broadening can be kept minimal also in the presence of losses [12]), they are not subject to strong inter-symbol interference (ISI) and can be detected with a low-complexity memoryless receiver as the NN considered in this work.

III. Numerical setup

The complete simulation setup used for both the end-to-end optimization of the communication system and the performance evaluation is depicted in Fig. 2 and its parameters are summarized in TABLE I. The same figure also shows the standard NFDM receiver that has been used to compute a benchmark test performance. The setup was implemented in Tensorflow to perform the optimization of the parameters (training phase, Section IV) while the performances of the system using the optimal parameters have been estimated using our existing MATLAB implementation of the setup (test phase, Section V). In the following sub-sections, after

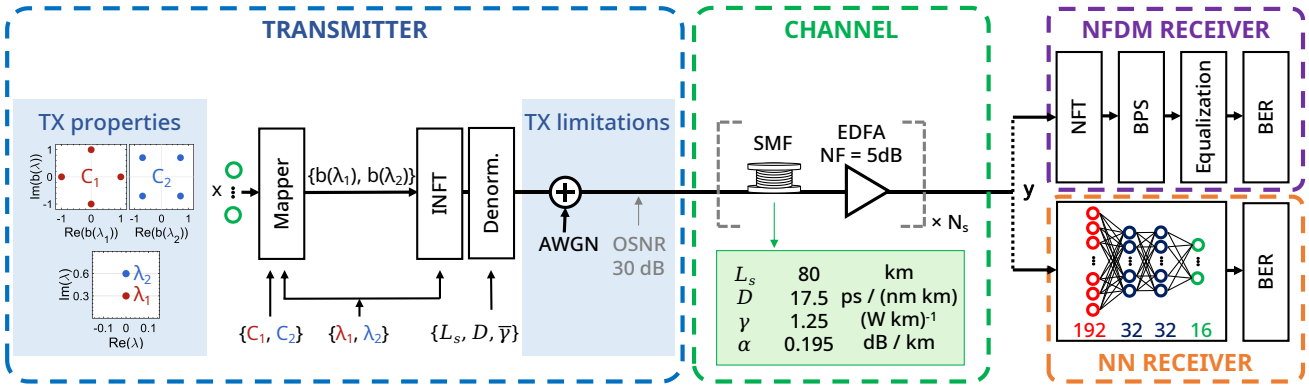


Fig. 2. Setup of the optical communication system used for both the end-to-end optimization of the system parameters and the BER performance evaluation.

a brief introduction to the NFT theory, transmitter, channel, and receivers (both NN-based and conventional for NFDM) of Fig. 2 are described in detail.

A. Nonlinear Fourier transform

The complex field envelope $E = E(\tau, \ell)$ of a single polarization signal propagating in a standard single-mode fiber (SSMF) with losses evolves according to the NLSE

$$\frac{\partial E}{\partial \ell} = -\frac{\alpha}{2}E - j\frac{\beta_2}{2}\frac{\partial^2 E}{\partial \tau^2} + j\gamma|E|^2 E. \tag{1}$$

where τ is the retarded time, ℓ the distance, α the attenuation coefficient, β_2 the group velocity dispersion (GVD), and γ the Kerr nonlinear coefficient of the fiber. The direct and inverse NFT transforms are defined for a lossless and noiseless NLSE [13]. To abstract from the specific channel parameters this NLSE is usually presented in its normalized form that for the anomalous dispersion regime ($\beta_2 < 0$) becomes

$$j\frac{\partial q}{\partial z} = \frac{\partial^2 q}{\partial t^2} + 2|q|^2 q \tag{2}$$

where t is the normalized retarded time, and z the normalized distance. This equation is derived from (1) ignoring the loss term and through the change of variables

$$q(t) = \frac{E(\tau)}{\sqrt{P}}, \qquad t = \frac{\tau}{T_0}, \qquad z = -\frac{\ell}{\mathcal{L}}, \tag{3}$$

with $P = |\beta_2|/(\gamma T_0^2)$, $\mathcal{L} = 2T_0^2/|\beta_2|$, and T_0 a free normalization parameter.

The direct NFT maps a time-domain waveform to a so-called nonlinear spectrum. When the time-domain waveforms are solitons, the nonlinear spectrum is said to be discrete and it is composed of a set of complex eigenvalues (nonlinear frequencies) λ_i with associated complex scattering coefficients $a(\lambda_i)$, $b(\lambda_i)$ with $a(\lambda_i) = 0$ [17].

The nonlinear spectrum can be modulated and used to carry information. A discrete NFDM communication system works as follows: the data is encoded onto the discrete nonlinear spectrum that is then mapped to a solitonic time-waveform through an inverse nonlinear Fourier transform (INFT) operation. The waveform is transmitted through the nonlinear fiber channel. At the receiver, the waveform is mapped back to a nonlinear spectrum (direct NFT) to retrieve the encoded data. Specific details of the NFDM communication system are given in the following sections.

B. Transmitter

The transmitter is a standard single-polarization NFDM transmitter with the same structure as the one reported in [18] and experimentally validated in [19].

In details, a sequence of uniform random symbols $x \in X = \{i \mid i = 1, \ldots, M\}$, with $M = 16$ the size of the chosen symbol alphabet, is generated. Each symbol is mapped onto a discrete nonlinear spectrum consisting of two discrete eigenvalues λ_i and their associated complex scattering coefficients $b(\lambda_i)$, $i = 1, 2$, which are independently modulated using 4-phase-shift keying (PSK) as illustrated in Fig. 2. The set of values $\{\lambda_i, b(\lambda_i)\}$, $i = 1, 2$ constitutes an NFDM symbol that carries a total of four information bits, similar to two parallel 4-PSK channels, at the symbol rate of 1 GBd. Note that the cardinality of the alphabet, as well as the number of eigenvalues, have not been optimized. The values have been chosen mainly to simplify the comparison with existing literature on NFDM systems [17]–[22]. The choice of b-modulation stems from its better noise resilience, as discussed in [23].

An INFT implemented with a Darboux transform (DT) [24] maps each NFDM symbol into a time-domain waveform $q(t)$ with 96 samples-per-symbol and solitonic pulse shape.

TABLE I
SIMULATION PARAMETERS

	Parameter name	Parameter	Value
Transmitter	# Polarizations		1
	Symbol rate		1 GBd
	Oversampling		96
	NFT free normalization factor	T_0	47 ps
	Lossless path-averaged $\bar{\gamma}$	$\bar{\gamma}_{LPA}$	0.34
Channel	Span length	L_s	80 km
	Dispersion parameter	D	17.5 ps / (nm km)
	Nonlinear coefficient	γ	1.25 (W km)$^{-1}$
	Attenuation	α	0.195 dB/km
	EDFA noise figure		5 dB
Receiver NN	# nodes in input layer		192
	Activation functions		SeLu
	# hidden layers		2
	# nodes per hidden layer		32
	# nodes in output layer		16

In order to obtain an optical field, $E(\tau)$, matched to the NLSE in (1), the waveform $q(t)$ needs to be de-normalized (Denorm. block in Fig. 2) using (3) with the parameters reported in TABLE I. This last operation ensures that the optical field, which carries the information, is matched to a lossless optical channel. Given that the fiber loss is not accounted for by the NFT theory, which relies on upon (2), the obtained waveform is not perfectly matched to the actual channel constituted of lossy single-mode fiber (SMF) spans interleaved by erbium-doped fiber amplifiers (EDFAs). It is possible to obtain a better match by performing the de-normalization in (3) replacing the fiber nonlinear coefficient γ with a different value $\bar{\gamma}$. In a standard NFDM communication system this value is usually set to the one provided by the lossless path-averaged (LPA) approximation [12], which for the channel considered in this work is $\bar{\gamma}_{LPA} = 0.34$.

After de-normalization, the field $E(\tau)$ is used to ideally modulate a laser, thus disregarding transmitter impairments such as laser phase noise and distortions introduced by the Mach-Zehnder modulator (MZM). The obtained signal is loaded with AWGN noise to limit its optical signal-to-noise ratio (OSNR) to 30 dB and finally transmitted over the channel.

C. Channel

The channel is composed of N_s spans of $L_s = 80$-km long SSMF interleaved by EDFAs (5-dB noise figure) used to compensate for the fiber loss. The full set of channel parameters are reported in TABLE I.

The signal propagation is numerically simulated using the well-know SSFM [25]. Being the SSFM a composition of basic differentiable mathematical operations, it is possible to numerically compute the gradient propagation from the output of the channel to its input, thus enabling an optimization of the performance of the communication system in terms of BER across the channel. The SSFM in this work is implemented in Tensorflow, and thanks to the automatic differentiation capability of the Tensorflow framework [26], the gradient is automatically computed. One limitation of a Tensorflow implementation of the channel is that it is not possible to use the variations of the SSFM using an adaptive step size. Indeed Tensorflow requires that the computational graph that implements the channel must be created before the computation happens. However, using a predetermined constant step size within the SSFM yields very slow and unpractical computations for the batch size and oversampling rate used in this work, especially for long transmission distances. As a trade-off between computing time, SSFM precision and Tensorflow requirements, the SSFM is implemented with 80 fixed step sizes per span, that increase logarithmically along the fiber length [27]. This choice of step sizes guarantees that at each step of the propagation the signal undergoes a maximum nonlinear phase rotation of 0.01 degrees when the simulation is performed with the reference configuration (see Section IV). Additionally, the simulation bandwidth of 96-GHz IV is sufficient to ensure that the SSFM provides an accurate solution to the NLSE describing the optical signal propagation.

D. Nonlinear Fourier transform receiver

The NFT receiver is a simplified version of the one in [28]. In particular, a bandpass filter with a bandwidth of 20 GHz has been used to filter the out-of-band noise prior to the nonlinear spectrum computation with the NFT. The detected discrete nonlinear spectrum was processed with a blind phase search (BPS) carrier phase estimation algorithm to compensate for the phase rotation introduced by the nonlinear domain transfer function of the channel over the $b(\lambda_i)$ [12], [28]. The compensated spectrum was finally equalized with a linear minimum mean square error (LMMSE) equalizer [28] prior to computing the BER.

E. Neural network receiver

The signal coming from the channel is sliced in non-overlapping blocks $\mathbf{y}$ of 96 samples (each corresponding to a single NFDM symbol) that are sequentially fed to the receiver, which consists of a feed-forward NN for multi-class classification with $M = 16$ output classes corresponding to all the possible transmitted symbols. As the NN is implemented as a real-valued network, the input layer consists of 192 nodes, considering the real and the imaginary parts of each sample as separate inputs. The NN takes as input the sequence $\mathbf{y}$ and provides at the output node i the probability $p(x_i|\mathbf{y})$ of having transmitted the symbol $x_i \in X$ given the received vector $\mathbf{y}$. The activation function of the hidden layers is a scaled exponential linear unit (SELU) [29] while the output layer uses a softmax activation function [14]. The NN weights are initialized using the Glorot algorithm [30], and optimized within the end-to-end training discussed in Section IV. The receiver NN parameters are reported in TABLE I. An argmax operation on the NN output probabilities is used to perform a hard decision that provides the estimated transmitted symbol $\hat{x}$. The detected symbols are finally used to compute the BER. Remark that the use of a memoryless (1-input symbols) receiver is enabled by the choice of an NFT-aided transmitter, which in turn results in transmitted waveforms (solitons) not affected by significant pulse broadening.

IV. End-to-end training

The goal of training the AE is to maximize the probability that the output symbol $\hat{x}$ of the communication system is equal to the input symbol x [6], [31]. The training consists of iteratively varying a set of trainable parameters and evaluating the performance of the system in terms of cross-entropy loss [14]. The details on the trainable parameters and the training process are given in the next two sections and the whole process is highlighted in Fig. 3. Note that, due to the data processing inequality, the joint optimization of transmitter and receiver through an AE approach, will theoretically yield equal or better performance compared to independent optimization of transmitter and receiver. The disadvantage of an independent optimization of the different blocks of a communication system is already introduced in [9].

A. Trainable parameters

The parameters of the transmitter chosen for the optimization are the imaginary part of the purely-imaginary eigenvalues λ_i, and the radius r_i and phase ϕ_i of the two PSK $b(\lambda_i)$-constellations:

$$C_i = r_i \exp\left(j\left(k\frac{\pi}{2} + \phi_i\right)\right), \; k = 0,1,2,3; \; i = 1,2. \quad (4)$$

Unlike a classical coherent modulation where the impulse response of a fixed pulse shaping filter is linearly modulated by the amplitude and phase of the transmitted symbols, the INFT nonlinearly maps the symbols-eigenvalue pairs to a time-domain waveform that is theoretically optimal for transmission over the nonlinear channel modeled by the lossless NLSE. Therefore, by constraining the transmitter pulse shape through the INFT, it is the nonlinear relation between symbols and waveforms which is kept fixed. This is in contrast with previous literature [4], [8] where the pulse shape is kept fixed regardless of the symbols optimization. This choice, whereas not reaching the target AE scheme of Fig. 1, moves one step closer in that direction. The components of the nonlinear spectrum affect the generated waveform $q(t)$ pulse shape in the following way [13]:

a) The imaginary part of the eigenvalues controls the energy E of the waveform according to

$$E = 4\sum_{i=1}^{2} \mathrm{Im}(\lambda_i), \quad (5)$$

and, at the same time, the waveform duration. Indeed the amplitude and duration of soliton pulses are inversely related. The energy of the waveform is only dependent on the λ_is and it is independent on the $b(\lambda_i)$-constellations.

b) The radii r_i of the constellations control the temporal position of the two components constituting the soliton waveform and thus their temporal overlap/separation.

c) The difference $\Delta\phi$ between the constellation phases ϕ_i governs the shape, and thus the bandwidth, of the waveform. The absolute phase of each C_i is not particularly relevant. A constant phase offset of both C_is maps to the same phase offset in the time-domain waveform [32].

To study the individual effects of these transmitter parameters on the performance of the communication system, four different training *configurations* are considered. In each configuration, a subset of these parameters is trained, while the remaining transmitter parameters are kept constant. For each of the configurations considered, the weights and biases (one bias per layer) of the NN that constitutes the receiver, denoted with $\mathbf{w}_{NN}$, are trained. For each configuration, the set of trained parameters θ and the static values used for the parameters that are not trained are reported in TABLE II. The static values are the same as those used in [18]. In the *configuration* 0, none of the transmitter parameters are optimized. This configuration is used as a benchmark to compare the performance of the other configurations. In *configuration* 1, only the constellations, i.e. radii and phases, are trained, in *configuration* 2 only the eigenvalues, and finally, in *configuration* 3 all parameters, constellations, and eigenvalues are fully trainable.

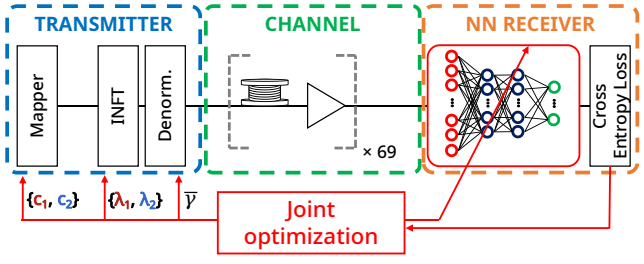


Fig. 3. Setup used for the training phase highlighting the transmitter and receiver parameters jointly optimized by the AE.

TABLE II
TRANSMITTER TRAINABLE PARAMETERS ($\bar{\gamma}$ OPTIMIZED IN $\bar{\gamma}_{E2E}$-SCENARIO, $\bar{\gamma}$ KEPT FIXED TO $\bar{\gamma}_{LPA}$ OTHERWISE.)

Configuration	λ_1, λ_2	ϕ_1, ϕ_2	r_1, r_2	θ
0	$0.3j, 0.6j$	$0, 0.25\pi$	1, 1	$[\bar{\gamma}, \mathbf{w}_{NN}]$
1	$0.3j, 0.6j$	trained	trained	$[C_i, \bar{\gamma}, \mathbf{w}_{NN}]$
2	trained	$0, 0.25\pi$	1, 1	$[\lambda_i, , \bar{\gamma}\mathbf{w}_{NN}]$
3	trained	trained	trained	$[\lambda_i, C_i, \bar{\gamma}, \mathbf{w}_{NN}]$

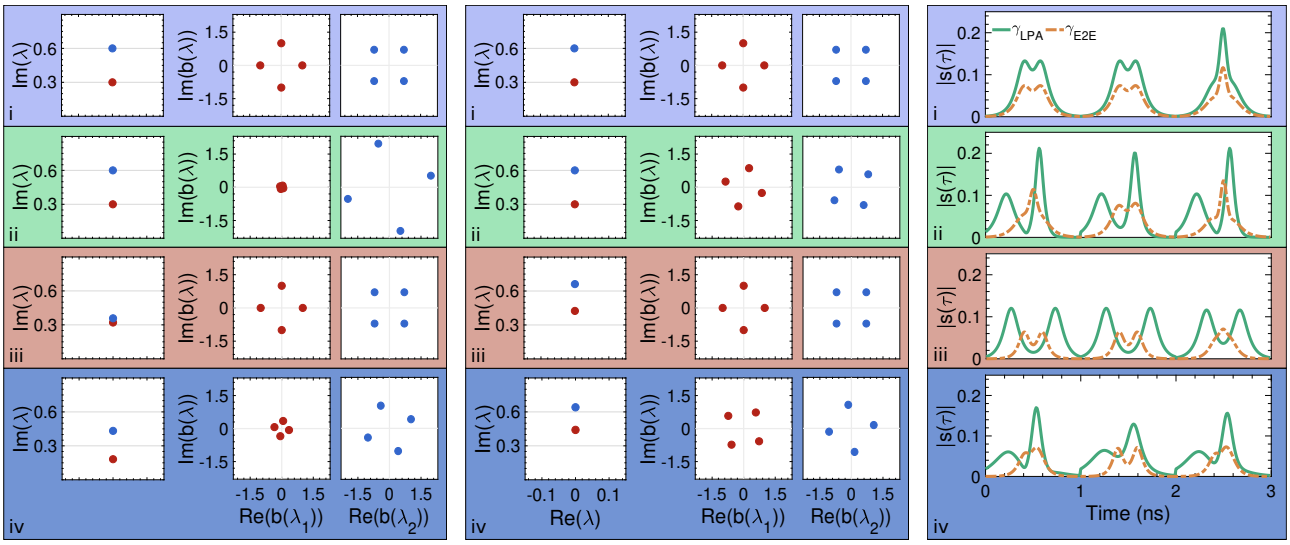
The choice of guiding the encoder optimization through the NFT allows to restrict the solution space that the AE has to search, but it results in enforcing a strict constraint on the waveform amplitude-duration relation [13]. For example, the AE can decrease the imaginary part of the eigenvalues, thus decreasing the amplitude and average power of the waveform while extending its temporal duration. The temporal broadening is however restricted by the NFDM symbol temporal slot of 1 ns. By further decreasing the amplitude, and thus further broadening the waveforms, it will spread beyond the available symbol slot yielding a negative impact on the system performance. This condition can be relaxed by making the variable $\bar{\gamma}$ a trainable parameter.

It should be noted that the parameter $\bar{\gamma}$ affects only the amplitude de-normalization parameter P, thus, given that the symbol period is fixed at 1 ns, changing $\bar{\gamma}$ is equivalent to changing the launch power of the signal $E(\tau, \ell)$.

By training $\bar{\gamma}$, any launch power can be set for a given transmitted waveform shape. Whereas this slightly deviates from the strict NFT theory, it gives the AE one additional degree of freedom to improve the overall system performance. It is therefore interesting to compare the two scenarios to understand how strictly the NFT theory can be applied to a lossy channel while aiming to improve transmission performance. The four different configurations of TABLE II have been optimized for both scenarios: $\bar{\gamma}$ is set equal to the value provided by the LPA approximation explained in Section III-B ($\bar{\gamma} = \bar{\gamma}_{LPA}$) and $\bar{\gamma}$ is optimized jointly with the other trainable variables ($\bar{\gamma} = \bar{\gamma}_{E2E}$).

B. Training

The AE optimization procedure is reported in Algorithm 1. The transmission distance was fixed at 69×80 km. This distance has been chosen so that for all the configurations the accuracy of the AE during the training is not too close to saturation. This, in turn, results in BER values for the testing



(a) Power factor unoptimized ($\bar{\gamma} = \hat{\gamma}_{LPA}$) (b) Power factor optimized ($\bar{\gamma} = \hat{\gamma}_{E2E}$) (c) Transmitted waveforms

Fig. 4. *Trained transmitter parameters* - Eigenvalues and constellations for the four configurations (i) to (iv) and scenario $\bar{\gamma} = \hat{\gamma}_{LPA}$ (a) and $\bar{\gamma} = \hat{\gamma}_{E2E}$ (b). Transmitted waveforms for $\bar{\gamma} = \hat{\gamma}_{LPA}$ (solid) and $\bar{\gamma} = \hat{\gamma}_{E2E}$ (dashed) for the four configurations (i) to (iv).

Algorithm 1 End-to-end training procedure

Inputs: Trainable parameters $\theta^{(0)}$
Outputs: Optimized trainable parameters $\theta^{(N_{iter})}$

Initialize trainable parameters
$[C_i, \lambda_i]\, i = 1, 2 \leftarrow$ configuration 0 in TABLE II
$\bar{\gamma} \leftarrow \gamma$
$\mathbf{w}_{NN} \leftarrow$ *Glorot* uniform initialization
Initialize feature vector $\theta^{(0)}$ as in Table II
for $n \leftarrow 1, \ldots, N_{iter}$ **do**
  Generate B random symbols $\{x\}_B$
  Map $\{x\}_B$ to $\{b(\lambda_i)\}_B$ using $C_i^{(n)}, i = 1, 2$
  Generate B waveforms $E(\tau) = \mathrm{INFT}(\lambda_i, b(\lambda_i))$
  Construct channel input $s(\tau)$ by serializing $\{E(\tau)\}_B$
  Propagate $s(\tau)$ in the channel using SSFM to get $r(\tau)$
  Slice $r(\tau)$ into B waveforms $\{\mathbf{y}\}_B$
  Detect $\{\mathbf{y}\}_B$ using rx NN with $\mathbf{w}_{NN}$
  Rx NN outputs B symbols $\{\hat{x}\}_B$
  Compute cross-entropy $L(\theta) = \mathrm{xentr}(\{x\}_B, \{\hat{x}\}_B)$
  Backpropagate to compute gradient $\nabla_\theta \tilde{L}(\theta^{(n)})$
  Compute $\theta^{(n+1)} \leftarrow \theta^{(n)} - \eta\ \mathrm{Nadam}(\nabla_\theta \tilde{L}(\theta^{(n)}))$
  Update trainable parameters from $\theta^{(n+1)}$
end for

phase which are neither too low (and thus challenging to count accurately), nor too high, overall allowing for a non-trivial comparison of the performance of the different configurations.

The AE is trained using the Adam optimizer with Nesterov gradient [33]. The optimization is run for a total of $N_{iter} = 6400$ iterations, a fixed value that was observed to ensure the convergence of all the 2 × four training configurations considered.

In each iteration, a batch of B = 64 symbols $x \in X$ is generated. This batch size is the maximum size for which the entire AE network (including the SSFM) fits into the available memory of our graphics processing unit (GPU) using the memory saving gradient technique [34], [35]. During the forward propagation of one training iteration, the results of each operation throughout the whole communication model - including the results of each step of the SSFM - need to be saved, to be used later during the backpropagation that computes the gradient of the loss function. This implies large memory requirements that limit the training batch size, which, in turn, causes a noisy estimation of the gradient at each training iteration. Being able to increase the batch size by increasing the available computing memory, may lead to improved performance or shorter training time [6]. A systematic complexity analysis of the proposed approach is beyond the scope of this work. Nevertheless, the current implementation is particularly time-consuming within the training phase whereas compares favorably against our implementation of the NFT receiver in the testing phase.

The learning rate was tuned according to a step decay schedule that sets it to $\eta = 0.01$ in the first 1600 iterations, $\eta = 0.003$ for the following 2400 iterations, and $\eta = 0.001$ for the remaining 2400 iterations. This scheduling allowed a quick convergence in the initial part of the training and a fine-tuning of the performance on the final part.

The architecture and hyper-parameters of the receiver NN have been kept fixed to provide a fair comparison among the different transmitter configurations. The chosen NN architecture yielded a sufficiently small network size while still guaranteeing good detection performances. A fine optimization of the receiver hyper-parameters for the optimal transmitter configuration would potentially improve the communication system performance, but doing it is beyond the scope of this work.

As the EDFAs noise in the channel is randomly generated at each training iteration, no identical batch is seen more than once by the AE during the training, even though the

transmitted symbols may be the same due to the finite batch size (online learning). This prevents overfitting problems that may instead arise when using a training dataset of fixed and limited size that is reused multiple times during the training process (batch learning). This choice allows for avoiding a separate cross-validation loss analysis to monitor eventual overfitting.

V. Simulation Results

In this section, the results of the AE training and testing are discussed. In Sections V-A and V-B, the trained transmitter parameters and the training performance are reported, whereas Sections V-C and V-D show the performance under testing for an NFT- and NN-based receiver, respectively.

A. AE-trained transmitter parameters

The values of the transmitter trainable parameters found by the end-to-end optimization of the communication system are reported in TABLE III for $\bar{\gamma} = \bar{\gamma}_{LPA}$, and in TABLE IV for $\bar{\gamma} = \bar{\gamma}_{E2E}$. The tables also report the average power of the transmitted waveforms generated using those parameters. These optimized transmitter parameters (eigenvalues and $b(\lambda_i)$-constellations) and the resulting time-domain waveforms are shown in Fig. 4 for (a) $\bar{\gamma} = \bar{\gamma}_{LPA}$ and (b) $\bar{\gamma} = \bar{\gamma}_{E2E}$.

Starting from the ($\bar{\gamma} = \bar{\gamma}_{LPA}$) scenario, a qualitative analysis of the optimized constellations (Fig. 4a) indicates that the AE optimizes the eigenvalues (configurations 2 and 3) by decreasing their imaginary part. As the transmitted power in this scenario is only determined by the imaginary part of the eigenvalues, decreasing the eigenvalues is equivalent to reducing the average power of the waveforms and extending their temporal duration. Observing the corresponding waveforms (dashed curves in Fig. 4c-(iii, iv)), the imaginary parts of the eigenvalues are decreased up to the point where the waveforms would not fit any more within the symbol slot of 1 ns. For the case where only the $b(\lambda_i)$-constellations can be optimized (configuration 1), the constellations found have a relative phase difference of 0.25π. More importantly, the radius of the first constellation (r_1) is decreased during the optimization whereas the radius of the second constellation is increased (r_2). These radius values cause the time waveform to have the first (second) solitonic component respectively delayed (anticipated) with respect to the center of the time window, as can be seen in Fig. 4c-(iii). Interestingly, the AE optimization (both in terms of phase difference and radii) converges to a set of constellations very similar to that reported in [17] where the constellations were manually optimized to maximize performance by minimizing the peak-to-average power ratio (PAPR) of the generated waveforms, and further investigated in [19]. Note that the transmission system in [17], [19] was equivalent, i.e. 80-km spans with lumped EDFA amplification. In the following, this system will therefore be considered as the reference NFDM system for benchmarking Finally, in *configuration* 3 the radii of the constellations are tuned by the AE similarly to the previous case, but given that now the imaginary part of the eigenvalues is reduced, enlarging the temporal duration of the soliton, the value of the radii of the constellations are closer to 1 compared to the *configuration* 1. A deviation from unitary radii would quite rapidly result in waveforms extending beyond the allowed symbol slot.

TABLE III
Optimized parameters ($\bar{\gamma} = \bar{\gamma}_{LPA}$).
The highlighted values are the results of the training.

Configuration	λ_1, λ_2	$\Delta\phi$	r_1, r_2	$\bar{\gamma}_{LPA}$	Power (dBm)
0	$0.3j, 0.6j$	0.25π	1, 1	0.34	7.03
1	$0.3j, 0.6j$	0.25π	0.08, 2.03	0.34	7.03
2	$0.33j, 0.37j$	0.25π	1, 1	0.34	5.97
3	$0.18j, 0.43j$	0.32π	0.35, 1.10	0.34	5.34

TABLE IV
Optimized parameters ($\bar{\gamma} = \bar{\gamma}_{E2E}$)
The highlighted values are the results of the training.

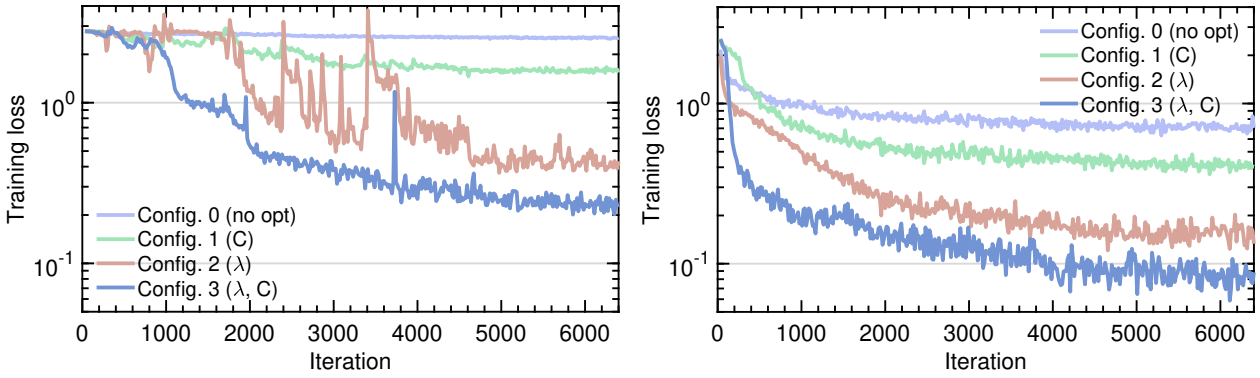
Configuration	λ_1, λ_2	$\Delta\phi$	r_1, r_2	$\bar{\gamma}_{E2E}$	Power (dBm)
0	$0.3j, 0.6j$	0.25π	1, 1	1.09	1.96
1	$0.3j, 0.6j$	0.29π	0.90, 0.99	0.97	2.46
2	$0.42j, 0.66j$	0.25π	1, 1	2.41	-0.67
3	$0.44j, 0.64j$	0.26π	0.93, 1.07	2.04	0.04

When the constraint of the amplitude-duration of the soliton is removed ($\bar{\gamma} = \bar{\gamma}_{E2E}$), the strategy of the AE to reduce the average power of the generated waveforms is even more evident. As shown by comparing TABLE IV to TABLE III, $\bar{\gamma}$ is optimized in order to significantly reduce the power launched into the channel. As the power can be reduced through $\bar{\gamma}$, for this scenario, the eigenvalues are not varied as heavily as for the previous scenario, and their imaginary part is rather slightly increased instead. Additionally, the radii of the constellations are less affected and they keep values close to unity even for configurations 1 and 3, leading to a lower separation in time between the two solitonic components. Finally, the relative phase rotations, instead, converge to similar values as for the scenario of $\bar{\gamma} = \bar{\gamma}_{LPA}$ as can be inferred by comparing the values reported in TABLE IV with those in TABLE III.

B. Autoencoder training performance

Looking at the cross-entropy loss curves as a function of the end-to-end training iteration in Fig. 5, it is possible to observe that in the scenario when $\bar{\gamma}$ is not trained ($\bar{\gamma}_{LPA}$, Fig. 5a), the loss curve does not converge to values as low as those for the scenario when $\bar{\gamma}$ is optimized (scenario $\bar{\gamma}_{E2E}$, Fig. 5b). Moreover, in the first scenario the loss curves do not decrease monotonically nor as quickly and smoothly as in the second scenario. In order to reduce the instability in the convergence, the learning rate of the AE was optimized. Nevertheless, the curves are still unstable and the performance reported in Fig. 5a shows the lowest losses achieved. The unstable behavior and the poorer performance obtained suggest that the loss function for the $\bar{\gamma} = \bar{\gamma}_{LPA}$ scenario is not smooth and contains multiple local minima that slow down and hinder the overall optimization.

Focusing on *configuration* 1 for both scenarios though, the optimization brings improvement in terms of loss compared to the reference case (*configuration* 0), but not as great as



(a) Power factor unoptimized ($\bar{\gamma} = \hat{\gamma}_{LPA}$) (b) Power factor optimized ($\bar{\gamma} = \hat{\gamma}_{E2E}$)

Fig. 5. *Training performance* Cross-entropy loss (training loss) during the training process for the four configurations within the two scenarios.

optimizing the imaginary part of the eigenvalues (*configuration* 2). By optimizing both degrees of freedom (*configuration* 3) it is possible to further reduce the cross-entropy loss but only minimally compared to *configuration* 2, i.e. eigenvalue-only optimization. This shows that the optimization of the eigenvalues is critical for the performance of the system. These general trends during the training phase are well aligned with the results achieved during the testing phase and shown in Section V-D.

C. Nonlinear Fourier transform receiver BER performance

Although the transmitter parameters are optimized for an NN receiver, and thus are not necessarily optimal for an NFT receiver, the performance of a conventional NFDM system (NFT transmitter + NFT receiver) is presented here. The counted BERs as a function of the transmission distance for the four configurations are shown in Fig. 6. Only the scenario where $\bar{\gamma} = \bar{\gamma}_{LPA}$ has been considered, as once the power of the waveform is not matched to the channel according to the NFT theory, a conventional NFDM receiver fails to correctly demodulate the received signal. The BER of the NFT receiver has been computed using 5×10^5 symbols.

For *configuration* 0 (unoptimized transmitter), the BER degrades rapidly and reaches values above the hard-decision forward error correction (HD-FEC) threshold (BER = 3.8×10^{-3}) already after a 3×80 km transmission. This is consistent with the results from [18], and it is due to the presence of fiber loss and long fiber spans. As the power varies significantly over the fiber length, the channel strongly deviates from the lossless fiber channel over which the NFT is defined, even considering the LPA approximation. When only the eigenvalues are optimized (*configuration* 2) the receiver cannot decode the data for any of the transmission distances. This is expected as the two eigenvalues are almost overlapped making it extremely challenging for the receiver to discriminate between them. When the $b(\lambda_i)$-constellations are optimized (*configurations* 1 and 3) the BER performance improves drastically compared to the previous configurations, and it is possible to reach a transmission distance of 24×80 km spans and 25×80 km, respectively, still considering an HD-FEC BER target.

In the future, to fully test the limits of the NFT system, the end-to-end training can be performed using the NFT receiver. This, however, requires to implement the NFT transform in Tensorflow using only non-adaptive algorithms. This limitation poses a challenge because it excludes the possibility to use some commonly used algorithms for locating the eigenvalues such as the search methods [36].

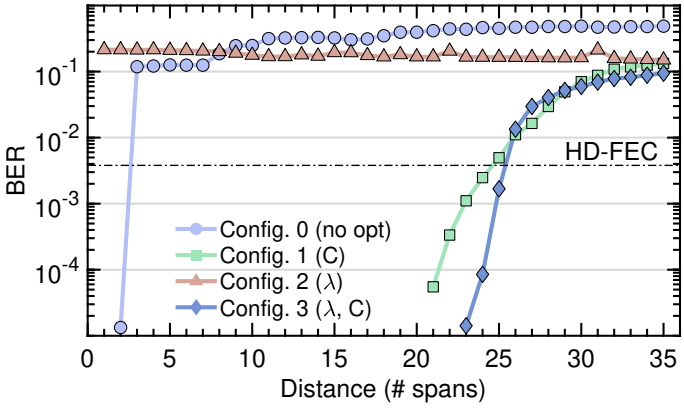
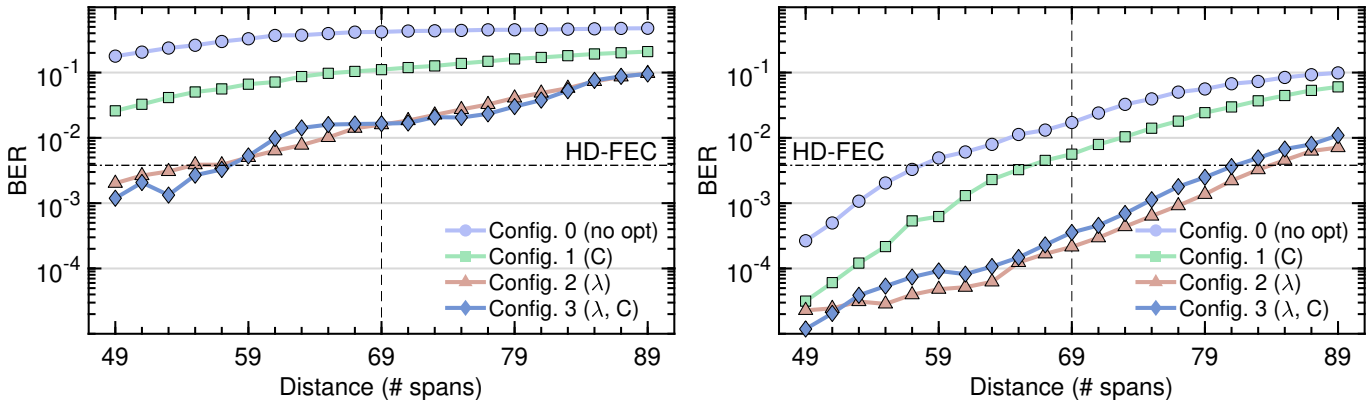


Fig. 6. *Test performance NFT receiver* - BER as a function of the transmission distance using the AE-optimized transmitter with $\bar{\gamma} = \bar{\gamma}_{LPA}$ and the NFT receiver.

D. Neural network receiver BER performance

The performances of the communication system, which was trained at a distance of 69×80 km, have been evaluated over a range of transmission distances from 49×80 km to 89×80 km in order to verify the robustness of the optimized transmitter to a transmission distance mismatch. The transmitter of the system uses the parameters optimized by the AE (eigenvalues, $b(\lambda_i)$-constellations, and $\bar{\gamma}$). The receiver NN used for the performance evaluation has the same topology as the one described in Section III-E but it was re-trained independently for each of the transmission distances considered. The re-training was performed using a training dataset of 500×10^3 symbols, a batch size of 1000 symbols, and 1000 training iterations. The re-training is necessary for the different transmission lengths, as the different amounts of accumulated dispersion



(a) Power factor unoptimized ($\bar{\gamma} = \hat{\gamma}_{LPA}$)

(b) Power factor optimized ($\bar{\gamma} = \hat{\gamma}_{E2E}$)

Fig. 7. *Test performance NN receiver* - BER as a function of the transmission distance. The vertical dashed line marks the optimization distance.

and nonlinear phase shift lead to significant variations in the time-domain waveforms seen by the receiver. The re- training was also performed for 69×80 km. As already shown in [9], re-training the receiver NN once the transmitter parameters have been optimized by the AE improves the performance. An intuitive explanation is that during the training of the AE, both the transmitter and the receiver NN parameters are changed at each training iteration so that the final NN is effectively trained for the specific transmitter parameters only for a single batch. The small batch size may not allow converging to the optimal set of weights. The re-training is particularly useful in this work given the very limited size of the batch used during the training of the AE. Larger batch sizes may provide improved performance potentially removing the need for re-training at 69×80 km [6]. Note that, alternatively to re-training for each transmission distance, the receiver NN could be trained with a dataset containing samples from several transmission distances. As shown in [9], this may lead to stronger robustness to variation in the target link length, but overall sub-optimal performance compared to selectively training for each distance independently. After re-training, the BER has been computed using a testing dataset of 5×10^5 symbols.

Fig. 7 shows the BER performance as a function of the transmission distance for (a) the transmitter power unoptimized scenario ($\bar{\gamma} = \bar{\gamma}_{LPA}$) and (b) the power-optimized scenario ($\bar{\gamma} = \bar{\gamma}_{E2E}$).

In Fig. 7 (a) ($\bar{\gamma} = \bar{\gamma}_{LPA}$), the worst performance is shown by *configuration* 0, i.e. when none of the transmitter parameters are optimized. The BER is close to 0.5 at the trained distance and drops only slightly for shorter transmission distances. When the $b(\lambda_i)$-constellations are optimized (*configuration* 1), the performance is improved compared to *configuration* 0, but the BER is still above 1×10^{-2} over the full range of distances considered. Separating the solitonic components in time provides only a slight improvement. When the AE optimizes the imaginary part of the eigenvalues (*configuration* 2), the BER at optimization distance is further reduced by almost one order of magnitude. Finally when both the eigenvalues and the $b(\lambda_i)$-constellations are optimized (*configuration* 3) the BER is similar to that *configuration* 2 for all the distances considered, consistently with the training performance. The optimization of the imaginary part of the eigenvalue, which controls the transmitted power, plays a key role in the system performance. The fact that a reduction of the signal power reduces the BER hints that the system performance at the optimization distance might be more limited by nonlinear effects and not by the OSNR so that reducing the transmitted power improves the performance.

In Fig. 7 (b) the curves for the second scenario ($\bar{\gamma} = \bar{\gamma}_{E2E}$) are shown. At the optimization distance, all configurations show an improved BER, with values of 1.72×10^{-2}, 5.63×10^{-3}, 2.14×10^{-4} and 3.56×10^{-4} for *configurations* 0 to 3, respectively. The relative performance between the four configurations follows the discussion reported for the $\bar{\gamma} = \bar{\gamma}_{LPA}$ scenario, but the additional degree of freedom provides a significant improvement for all the configurations. This proves that the average launch power (through the parameter $\bar{\gamma}$) is a critical optimization parameter. The BER obtained in the best case (*configuration* 2, $\bar{\gamma} = \bar{\gamma}_{E2E}$) is three orders of magnitudes lower than the reference case (*configuration* 0, $\bar{\gamma} = \bar{\gamma}_{LPA}$) where none of the transmitter parameters are optimized. This best configuration among those considered allows reaching a transmission distance more than three times what is achievable with a manually-optimized NFT receiver (approx. 83 spans vs. the 25 spans of Fig. 6).

We can observe in Fig. 7 (b) that the *configuration* 3 performs slightly worse than the *configuration* 2, despite having more optimization degrees of freedom. This is likely due to the presence of local optima in the cost function from where the optimization procedure was not able to escape. This result further justifies our choice of guiding the encoder rather than performing a fully blind optimization. The re-training of the receiver can only partially improve on the negative impact of local optima as the transmitter parameters are not improved and they have a clear impact on the overall performance. In the future, strategies to move towards this latter goal without suffering from local optima in the optimization landscape need to be devised [37]. Overall, the comparison between *configuration* 1 and *configuration* 2, under both optimization conditions ($\bar{\gamma} = \bar{\gamma}_{LPA}$ and $\bar{\gamma} = \bar{\gamma}_{E2E}$), hints that the optimization of

the eigenvalues is more critical than the optimization of the spectral amplitudes.

For all the eight different configurations tested we can observe that the performance gain is preserved across all the transmission distances between 49 and 89×80 km, even though the optimization was performed at 69 spans, showing the robustness of the transmitter parameters optimization to the transmission distance.

VI. Conclusion

In this work, we proposed for the first time an AE scheme for coherent fiber-optic communications considering the accurate model of a nonlinear dispersive fiber channel. The system uses an NFDM transmitter that performs the symbol-to-waveform encoding and an NN-based receiver that performs the waveform-to-symbol decoding. The chosen transmitter constraints the solutions space of the generated time-domain waveforms to solitonic pulse-shapes, facilitating the optimization of the AE. Moreover the minimal dispersion of the solitons, even in the presence of losses, allows using a memoryless receiver implemented with a low-complexity NN. The full optical nonlinear channel model has been implemented using the SSFM. Given that this method is a composition of basic differentiable operations, it was possible to use the automatic differentiation capabilities of the Tensorflow library to perform the training of the AE across this channel model.

The AE has been trained using $2 \times$ four different configurations of the transmitter trainable parameters (eigenvalues, $b(\lambda_i)$-constellations, and $\bar{\gamma}$). The numerical results of the performance testing of the system demonstrated that the best configuration of these parameters allows a reduction of three orders of magnitude in BER at the optimization distance compared to the un-optimized transmitter configuration. In particular, it was shown that the system performance is particularly sensitive to the imaginary part of the eigenvalues. Moreover, it was shown that the best results are obtained when the AE is also free to optimize the waveform launch power through the optimization of $\bar{\gamma}$. Compared to a manually-optimized NFDM communication system used as a benchmark, the proposed proof-of-concept system allows extending the transmission reach from 2000 to 6640 km at the HD-FEC threshold.

We believe that this work moves the research on the end-to-end optimization of communication systems a step closer to the final goal of realizing a general AE communication scheme employing NNs for both the transmitter and the receiver and defined over a nonlinear dispersive fiber-optic channel.

Acknowledgment

This work is supported by the European Research Council through the ERC-CoG FRECOM project (grant agreement no. 771878) and by the Villum Foundation through the Villum Young Investigator fellowship OPTIC-AI (grant no. 29344).

References

[1] M. P. Yankov, F. Da Ros, E. P. da Silva, S. Forchhammer, K. J. Larsen, L. K. Oxenløwe, M. Galili, and D. Zibar, "Constellation shaping for wdm systems using 256QAM/1024QAM with probabilistic optimization," *Journal of Lightwave Technology*, vol. 34, no. 22, pp. 5146–5156, 2016.

[2] F. Buchali, F. Steiner, G. Böcherer, L. Schmalen, P. Schulte, and W. Idler, "Rate adaptation and reach increase by probabilistically shaped 64-QAM: An experimental demonstration," *Journal of Lightwave Technology*, vol. 34, no. 7, pp. 1599–1609, 2016.

[3] J. Renner, T. Fehenberger, M. P. Yankov, F. Da Ros, S. Forchhammer, G. Böcherer, and N. Hanik, "Experimental comparison of probabilistic shaping methods for unrepeated fiber transmission," *Journal of Lightwave Technology*, vol. 35, no. 22, pp. 4871–4879, 2017.

[4] R. T. Jones, T. A. Eriksson, M. P. Yankov, and D. Zibar, "Deep Learning of Geometric Constellation Shaping Including Fiber Nonlinearities," *European Conference on Optical Communication (ECOC)*, pp. 1–3, 2018.

[5] V. Oliari, E. Agrell, and A. Alvarado, "Regular perturbation on the group-velocity dispersion parameter for nonlinear fibre-optical communications," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[6] 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, 2017.

[7] S. Li, C. Hager, N. Garcia, and H. Wymeersch, "Achievable Information Rates for Nonlinear Fiber Communication via End-to-end Autoencoder Learning," in *European Conference on Optical Communication (ECOC 2018)*, 2018, pp. 1–3.

[8] R. T. Jones, T. A. Eriksson, M. P. Yankov, B. J. Puttnam, G. Rademacher, R. S. Luis, and D. Zibar, "Geometric Constellation Shaping for Fiber Optic Communication Systems via End-to-end Learning," pp. 1–9, 2018. arXiv: 1810.00774.

[9] B. Karanov, M. Chagnon, F. Thouin, T. A. Eriksson, H. Bulow, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-End Deep Learning of Optical Fiber Communications," *Journal of Lightwave Technology*, vol. 36, no. 20, pp. 4843–4855, 2018.

[10] B. Karanov, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-end optimized transmission over dispersive intensity-modulated channels using bidirectional recurrent neural networks," *Optics Express*, vol. 27, no. 14, p. 19 650, 2019.

[11] S. Gaiarin, R. Jones, F. Da Ros, and D. Zibar, "End-to-end optimized nonlinear fourier transform-basedcoherent communications," in *Conference on Lasers and Electro-Optics (CLEO)*, Optical Society of America, 2020, SF2L.4.

[12] A. Hasegawa and Y. Kodama, *Solitons in optical communications*, 7. Oxford University Press, USA, 1995.

[13] M. J. Ablowitz, B. Prinari, and A. D. Trubatch, "Integrable Nonlinear Schrödinger Systems and their Soliton Dynamics," *Dynamics of PDE*, vol. 1, no. 3, pp. 239–299, 2004.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[15] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, and M. Tornatore, "An overview on application of machine learning techniques in optical networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1383–1408, 2018.

[16] S. K. Turitsyn, J. E. Prilepsky, S. T. Le, S. Wahls, L. L. Frumin, M. Kamalian, and S. A. Derevyanko, "Nonlinear fourier transform for optical data processing and transmission: Advances and perspectives," *Optica*, vol. 4, no. 3, pp. 307–322, 2017.

[17] S. Gaiarin, A. M. Perego, E. P. da Silva, F. Da Ros, and D. Zibar, "Dual-polarization nonlinear Fourier transform-based optical communication system," *Optica*, vol. 5, no. 3, p. 263, 2018.

[18] R. T. Jones, S. Gaiarin, M. P. Yankov, and D. Zibar, "Time-Domain Neural Network Receiver for Nonlinear Frequency Division Multiplexed Systems," *IEEE Photonics Technology Letters*, vol. 30, no. 12, pp. 1079–1082, 2018.

[19] S. Gaiarin, F. DaRos, N. De Renzis, R. T. Jones, and D. Zibar, "Experimental demonstration of nonlinear frequency division multiplexing transmission with neural network receiver," *Journal of Lightwave Technology*, 2020.

[20] F. Da Ros, S. Civelli, S. Gaiarin, E. P. da Silva, N. De Renzis, M. Secondini, and D. Zibar, "Dual-polarization nfdm transmission with continuous and discrete spectral modulation," *Journal of Lightwave Technology*, vol. 37, no. 10, pp. 2335–2343, 2019.

[21] S. T. Le, V. Aref, and H. Buelow, "Nonlinear signal multiplexing for communication beyond the kerr nonlinearity limit," *Nature Photonics*, vol. 11, no. 9, pp. 570–576, 2017.

[22] Z. Dong, S. Hari, T. Gui, K. Zhong, M. I. Yousefi, C. Lu, P. A. Wai, F. R. Kschischang, and A. P. T. Lau, "Nonlinear frequency division multiplexed transmissions based on nft," *IEEE Photonics Technology Letters*, vol. 27, no. 15, pp. 1621–1623, 2015.

[23] S. Wahls, "Generation of time-limited signals in the nonlinear fourier domain via b-modulation," in *2017 European Conference on Optical Communication (ECOC)*, IEEE, 2017, pp. 1–3.

[24] V. B. Matveev and M. A. Salle, *Darboux transformations and solitons*. Springer-Verlag, 1991.

[25] G. P. Agrawal, *Nonlinear Fiber Optics*, 5th ed. Academic Press, 2013.

[26] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic differentiation in machine learning: A survey," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 5595–5637, 2017.

[27] G. Bosco, A. Carena, V. Curri, R. Gaudino, P. Poggiolini, and S. Benedetto, "Suppression of spurious tones induced by the split-step method in fiber systems simulation," *IEEE Photonics Technology Letters*, vol. 12, no. 5, pp. 489–491, 2000.

[28] S. Gaiarin, F. Da Ros, N. De Renzis, E. P. Da Silva, and D. Zibar, "Dual-Polarization NFDM Transmission Using Distributed Raman Amplification and NFT-Domain Equalization," *IEEE Photonics Technology Letters*, vol. 30, no. 22, pp. 1983–1986, 2018.

[29] G. Klambauer, T. Unterthiner, and A. Mayr, "Self-Normalizing Neural Networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 971–980.

[30] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[31] F. N. Khan, Q. Fan, C. Lu, and A. P. T. Lau, "An optical communication's perspective on machine learning and its applications," *Journal of Lightwave Technology*, vol. 37, no. 2, pp. 493–516, 2019.

[32] M. I. Yousefi and F. R. Kschischang, "Information transmission using the nonlinear fourier transform, part I: Mathematical tools," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4312–4328, 2014.

[33] T. Dozat, "Incorporating Nesterov momentum into Adam," in *International Conference on Learning Representations (ICLRW)*, 2016, pp. 1–6.

[34] T. Chen, B. Xu, C. Zhang, and C. Guestrin, "Training Deep Nets with Sublinear Memory Cost," pp. 1–12, 2016.

[35] *Gradient checkpoint*, https://github.com/cybertronai/gradient-checkpointing.

[36] M. I. Yousefi and F. R. Kschischang, "Information Transmission Using the Nonlinear Fourier Transform, Part II: Numerical Methods," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4329–4345, 2014.

[37] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Advances in Neural Information Processing Systems*, 2018, pp. 6389–6399.