Research Article Vol. 31, No. 1 / 2 Jan 2023 / Optics Express 1
Optics EXPRESS

Memory-aware end-to-end learning of channel distortions in optical coherent communications

VLADISLAV NESKORNIUK,[1,2,*] ANDREA CARNIO,[3] DOMENICO MARSELLA,[3] SERGEI K. TURITSYN,[2] JAROSLAW E. PRILEPSKY,[2] AND VAHID AREF[1]

[1]*Nokia, 70469 Stuttgart, Germany*
[2]*Aston Institute of Photonic Technologies, Aston University, B4 7ET Birmingham, UK*
[3]*Nokia, Vimercate 20871, Italy*
[*]*v.neskorniuk@gmail.com*

Abstract: We implement a new variant of the end-to-end learning approach for the performance improvement of an optical coherent-detection communication system. The proposed solution enables learning the joint probabilistic and geometric shaping of symbol sequences by using auxiliary channel model based on the perturbation theory and the refined symbol probabilities training procedure. Due to its structure, the auxiliary channel model based on the first order perturbation theory expansions allows us performing an efficient parallelizable model application, while, simultaneously, producing a remarkably accurate channel approximation. The learnt multi-symbol joint probabilistic and geometric shaping demonstrates a considerable bit-wise mutual information gain of 0.47 bits/2D-symbol over the conventional Maxwell-Boltzmann shaping for a single-channel 64 GBd transmission through the 170 km single-mode fiber link.

1. Introduction

When dealing with optical communication systems, there is still the lack of the general communication theory describing nonlinear fiber channels (where the dispersive effects intertwine with the fiber nonlinearity), in contrast to the classical additive white Gaussian noise (AWGN) communication channel. Thus, many fundamental and practically important questions related to the optimal coding, modulation, pulse-shaping, and channel equalization for the nonlinear optical transmission systems, remain either partially solved or even yet to be answered [1,2]. In particular, the optimal signal statistics (the Gaussian distribution for AWGN [3]) is not known for the transmission affected by the simultaneous action of fiber dispersion and nonlinearity. Addressing this problem numerically, in general, is not practically feasible due to the high computational cost of modelling high-speed transmission via dispersive nonlinear channels. Thus, there remains the research challenge related to developing practical transmitters and receivers with signal format and modulation inherently adjusted to nonlinear transmission. We note that currently-used transceivers are suboptimal, which brings about a cap on the achievable data rates and transmission distances. Recently, machine learning (ML) methods and, in particular, artificial neural networks (NNs) have been applied to the design of optical communication systems, see, e.g., recent Refs. [4–10] and numerous literature sources therein. Albeit it is rather difficult (if possible at all) to obtain the general conclusions on the optimal signal shaping and modulation in the realistic nonlinear fiber systems, it is possible to obtain some sub-optimal results by using specific ML techniques tailored to deal with complex nonlinear problems.

It is important to point out that the properties of optical fiber-channels strongly depend on signal parameters: the power is linked to the nonlinear effects, and the carrier pulse width defines the dispersive signal broadening and respective memory effects. Then, the transmission

#470154 https://doi.org/10.1364/OE.470154
Journal © 2023 Received 25 Jul 2022; revised 27 Oct 2022; accepted 30 Oct 2022; published 19 Dec 2022

Optics EXPRESS

distance defines how essential the noise-induced corruptions are, the strength of nonlinear signal distortions, and the dispersive effects. Therefore, different nonlinear fibre channels can have rather distinct optimal signal modulations and coding.

End-to-end (E2E) learning approach, proposed in [11], is a machine learning technique offering a way to automatically tailor signal modulation and coding for an arbitrary communication link. The basic idea of E2E learning lies in representing the link from messages-in to messages-out as a single NN, differentiable with respect to the link parameters, to simultaneously optimize these parameters by using, e.g., the efficient gradient-descent-based optimization. Many applications of the E2E learning in different communication links then followed [12,13].

The first applications of E2E learning to optical communication systems were done for intensity modulation / direct detection (IM/DD) systems. In Refs. [14–17], the E2E learning of geometric constellation shaping (GS), i.e., optimal symbol locations, for IM/DD optical communication systems was proposed. Further, in [18] the E2E learning of waveforms was considered for the specific case of a link based on nonlinear frequency division multiplexing.

The following works dealt with the E2E learning of the optimal constellation points locations, i.e. on the GS, and the pre-distortion techniques in coherent systems. In [19–24], the E2E learning of single-symbol GS was considered for a coherent communication system. While [19–23] considered the distortions generated only by the optical channel, in [24] a more realistic link model that included the local oscillator laser noise, was studied. In [13,25–30], the E2E learning of GS, signal waveform, and nonlinear pre-distortion resistant to transmitter distortions was considered. However, the true complexity of nonlinear fiber-optic channel distortions was neglected in these works, namely, the distortions were either completely neglected, or modelled via a simplified Gaussian noise model. Conversely, in [31] the authors considered the joint E2E learning of GS and linear pre-distorter mitigating the fiber channel distortion. Nonetheless, the pre-distorter learnt in that work was rather trivial: it combined the Nyquist pulse shaper and chromatic dispersion compensator, and, hence, it actually did not contribute to the nonlinearity mitigation.

In our current paper, we propose the machine learning algorithm for E2E learning of the constellation shaping that takes into account the nonlinearities and memory present in optical channel distortions. With this algorithm, we jointly optimized the symbol locations in the constellation diagram, the symbol probabilities, and the nonlinear pre-distortion. The learnt transmitted signal distribution chooses the transmitted symbol based not only on the message sent in the corresponding time slot, as in the conventional constellation shaping, but also on the messages sent in the neighbouring time slots. Therefore, we refer to the resulting signal distribution as the multi-symbol constellation shaping (MSCS).

In the proposed E2E learning algorithm, we implement two key new features. The first one is the utilization of the auxiliary channel model based on perturbation theory [32], which allows us to reduce the computational efforts/resources needed to model the complex mixture of nonlinear and linear distortions taking place in fiber-optic links. Second, we implemented the training procedure for the simultaneous learning of symbol probabilities and locations according to [33], in which the accuracy of the procedure was demonstrated for the AWGN channel. These innovations let the algorithm learn the signal distribution outperforming a conventional Maxwell-Boltzmann (MB) one, where the latter is optimal for the AWGN channel [34]. In more details, we applied the proposed algorithm to a single-channel dual-prolarized 64 GBd transmission over 170 km standard single mode fiber (SMF) link, where the expected gains by nonlinear shaping are particularly high [35]. The learnt MSCS led to the bit-wise mutual information (BMI) gain of 0.47 bits/2D-symbol/polarization over the conventional MB shaping. Furthermore, we show that the proposed E2E learning is applicable for the cases when, because of hardware- or complexity-related limitations, we cannot use multi-symbol constellation shaping. For the same test case, we learnt a single-symbol joint probabilistic and geometric shaping

Optics EXPRESS

showing 0.074 bits/2D-symbol/pol. BMI gain over the reference MB shaping. Moreover, for the case when geometric shaping is not an option, we learnt the single-symbol probabilistic shaping which outperforms the MB shaping by 0.043 bits/2D-symbol/pol. in terms of BMI. Note that the proposed method is rather generic and is applicable to improving the quality of transmission in an arbitrary coherent fiber-optic communication link. In [36] we successfully applied the similar E2E learning technique to a 64 GBd transmission over long-haul 30x80 km (2400 km) SMF link.

This work extends our previous approaches from Refs. [36,37]. The remainder of the paper is organized as follows: Section 2 describes in the proposed E2E learning algorithm; Section 3 describes the considered testcase and the results achieved by the proposed E2E learning; and Section 4 concludes the paper.

2. Models and the algorithm description

2.1. End-to-end learning

In the end-to-end learning approach, the whole system is implemented as a single NN from bits-in to bits-out, including transmitter, channel, and receiver. This enables the joint training of transmitter and receiver. The idea of the approach was first suggested in [11]. The scheme of end-to-end platform used in this paper is given in Fig. 1.

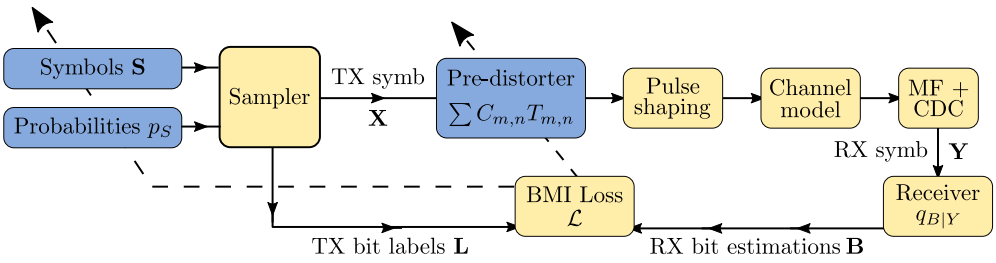


Fig. 1. Principal scheme of the end-to-end learning algorithm implemented in this paper. Blue denotes trainable blocks, dashed lines denote feedback from loss function. MF stands for matched filtering, CDC stands for chromatic dispersion compensation.

2.2. Transmitter design, constellation shaping, and pre-distortion

We consider the two-stage transmitter, consisting of a sampler followed by a nonlinear pre-distorter. Note that the transmitter's separability into autonomous stages, with each stage having its well-defined purpose, improves its interpretability and cost-efficiency [13].

In the sampler, the input data is mapped to some complex-valued constellation points $s_m \in \mathbf{S} \in \mathbb{C}, |\mathbf{S}| = M$ and then transmitted over a channel. A fixed bit label $\mathbf{b}_m = [b_{m,1}, \ldots, b_{m,J}]$, $b_{m,j} \in \{0, 1\}$, $J = \lceil \log_2(M) \rceil$ is set for each symbol in the alphabet $\mathbf{b}_m \Longleftrightarrow s_m$, $\forall s_m \in \mathbf{S}$. The input data is generated in a way that symbols s_m are sampled according to the discrete probabilistic distribution $p_S(s_m)\, \forall s_m \in \mathbf{S}$. In this work, we assume that for an arbitrary distribution p_S there exists a mapper which maps the original input data stream in such a manner that the resulting symbols are independent and identically distributed according to p_S. Such an idealistic mapping can be achieved by a distribution matcher, e.g., the constant composition distribution matching (CCDM), with a long enough sequence.

Because of the imperfections of the communication channel, the information rate of the system depends on the locations of constellation points $\mathbf{S}$ and their probability distribution p_S. Therefore, it is possible to maximize the system information rate by optimizing symbol locations $\mathbf{S}$ and occurrence probabilities p_S. Optimizing $\mathbf{S}$ is called a geometric constellation shaping (GS), and optimizing p_S is called a probabilistic constellation shaping (PS) [38]. Notably, for linear channels

Optics EXPRESS

with AWGN, the family of Maxwell-Boltzmann (MB) distributions results in information rates very close to the optimal PS [34]

$$p_S(s_m) \approx \exp\left(-\theta|s_m|^2\right), \tag{1}$$

where θ is a distribution hyper-parameter referred to as the MB shaping parameter. Nevertheless, finding the optimal shaping for nonlinear-dispersive optical fiber channels is a relatively subtle problem [35].

In this work, we consider a separate optimization of PS, GS, along with the joint optimization of both $\mathbf{S}$ and p_S; the joint optimization method is referred to as the joint shaping (JS). The sampler generated the dual-polarized signal $X = \{X_h, X_v\}$ with both polarization components X_h, X_v sampled from the same alphabet $x_{i,h/v} \in \mathbf{S}$ according to the same distribution p_S.

The sampler was followed by a trainable nonlinear pre-distorter. The goal of the latter is to pre-compensate the nonlinear channel distortions through pre-processing the symbol sequence $X_{h/v}$ generated by the sampler, before the sequence is sent into the optical channel. The pre-compensation is made in such a way that the symbol sequence at the channel output, Y, approximates the sampled input symbol sequence. We consider a perturbation-based pre-distortion (PPD) [39], which adds to each transmitted symbol $x_{i,h/v}$ an additive correction $\Delta_{\text{PPD}}x_{i,h/v}$, depending on the characters transmitted in the neighboring slots in both polarizations $x_{i,h}, x_{i,v}$. The pre-distortion Δ_{PPD} is defined as a linear combination of cubic polynomials $T_{m,n}$ calculated from the symbols co-propagated with the pre-processed symbol at the neighboring time slots. In more detail, for any symbol $x_{i,h/v}$ transmitted at the i-th time slot in the H- or V-polarization, the pre-distortion takes the form:

$$\Delta_{\text{PPD}}(x_{i,h/v}) = \sum_{m,n} C_{m,n} \cdot T_{m,n};\ T_{m,n} = x_{i+m,h/v} \cdot \left(x^*_{i+n+m,h/v}x_{i+n,h/v} + x^*_{i+n+m,v/h}x_{i+n,v/h}\right), \tag{2}$$

where $x_{i+m,h/v}$ is the symbol in H-/V-polarization shifted by m time slots from the target symbol $x_{i,h/v}$, and $C_{m,n}$ are the trainable weights.

The performance-to-complexity ratio of PPD is determined by the range of polynomials $T_{m,n}$ taken into consideration in a particular algorithm. Since the PPD is a linear regression over $T_{m,n}$ terms, the importance of each term in the trained PPD can be assessed by the absolute value of the coefficient $|C_{m,n}|$ corresponding to it. Hence, one can reach a cost-effective PPD by training an excessively complex one and then pruning it, i.e. removing terms $T_{m,n}$ with the absolute value smaller than the chosen cut-off value $T_{m,n} : |C_{m,n}| < C_{\text{cutoff}}$. This makes the cut-off value C_{cutoff} an important hyper-parameter of PPD learning, defining the performance-to-complexity ratio of the resulting algorithm. We consider its optimization as part of E2E learning in Section 3.4.

2.3. RP-based channel model

The second autoencoder block, the *auxiliary channel model*, maps the signal generated by the transmitter X to a sequence of symbols Y collected by the receiver. The model includes both deterministic and stochastic distortions, expressed via the probability of receiving Y given X was transmitted, $p_{Y|X}$. The approximate channel model must be computationally simple and easy to differentiate, to allow for fast transmitter learning; but, simultaneously, it has to describe the distortion introduced by the channel accurately enough, so that the learned transmitter and receiver could emulate well a real-world communication link.

The nonlinear dispersive channel is typically modelled by the Manakov equations, see Eq. (3) below, which are simulated by a serial cascade of alternating convolutional and pointwise nonlinear operators; this solution scheme is referred to as the split-step Fourier method (SSFM) [40]. The SSFM can be represented by the convolutional NN consisting of many layers [31,41], and the complexity of such a convolutional NN makes the calculation of gradients of the model

Optics EXPRESS

outputs over its inputs (in the back-propagation learning [42]) very slow and rather challenging. First, the learning process implies that all the intermediate states are stored, and it makes the process memory hungry. Second, the back-propagation through many layers often results in the numerical errors, leading to the infamous uncontrolled growth or vanishing of gradients [9,43]. One way to circumvent these problems is to consider a simpler auxiliary channel model. This approach, while typically offering the lesser degree of channel approximation precision than SSFM and, hence, the smaller shaping gain of E2E learned constellation shaping, can greatly reduce the numerical complexity of E2E learning. Among others, the following models were proposed for this purpose: the E2E learning using dispersion-free nonlinear channel model [44]; a nonlinear interference noise (NLIN) model [20,37,45], which considers nonlinear distortion as a AWGN process with a constellation-dependent variance [46,47]; or neglecting the optical channel nonlinearity at all by modelling its distortion as a fixed AWGN [13]. Nonetheless, all these models neglect the channel memory, i.e., inter-symbol interactions. On top of it, NLIN and AWGN-based channel models erase the information about the determinism of nonlinear distortions, modelling it as a stochastic noise. These two factors prevent the E2E learning of any memory-aware constellation shaping or nonlinearity pre-distorter, both of which rely on the information about the inter-symbol behavior of nonlinear distortion.

At the same time, NLIN-based auxiliary channel model can be useful for the limited task of E2E learning the single-symbol constellation shaping. In general, we expect the E2E learning via memory-aware auxiliary channel models to obtain higher shaping gain than NLIN, because of these models being free from the assumption of Gaussianity of the injected channel distortion. Nonetheless, the simplicity of NLIN model may lead to faster and cheaper training procedure. Therefore, we consider comparing NLIN with memory-aware channel models to be a fruitful, yet still open research question.

Recently, the data-driven approaches, learning the channel model properties from analysing the propagated data, were suggested [48,49], with [23] proposing the E2E GS learning in a coherent communication link through a data-driven auxiliary channel model. On one side, compared to the aforementioned approaches, the data-driven models do not require precise information about the optical channel properties. Nonetheless, the data-driven models are often less cost-effective since they require additional computational resources to learn the link parameters. In this paper, we assume the common case of the channel parameters being known in advance, which, therefore, renders the data-driven approach excessively complex.

Following our previous work [36], we propose here the application of first-order regular perturbation model (RP model) as an auxiliary channel model. Particularly, we suggest using the RP model for calculating both loss value and gradients, referred to as the forward and backward passes, correspondingly. In the following, we describe the RP model and the benefits of its application in more detail.

Consider the Manakov equations describing the evolution of the waveform of a dual-polarized optical signal $\mathbf{E}(z,t) = \mathbf{u}(z,t)\sqrt{f(z)}$ during its propagation over a fiber-optic link with lumped optical amplifiers (OAs) [50]:

$$\frac{\partial \mathbf{u}}{\partial z} = -i\frac{\beta_2}{2}\frac{\partial^2 \mathbf{u}}{\partial t^2} + i\frac{8}{9}\gamma f(z)\|\mathbf{u}\|^2\mathbf{u} + \eta(z,t), \quad f(z) = \exp\left(-\frac{\alpha}{2}\left(z - L_{\mathrm{sp}}\lfloor z/L_{\mathrm{sp}}\rfloor\right)\right), \tag{3}$$

where $f(z)$ models the optical losses (with α being the attenuation coefficient) and amplification accumulated till point z, L_{sp} denotes the fiber span length, β_2 and γ are the chromatic dispersion and Kerr nonlinearity coefficients; $\eta(z,t)$ denotes the amplified spontaneous emission noise (ASE) injected by OAs.

The first-order regular perturbation (RP) [32,50,51] is an elaborate method to approximate $\mathbf{u}(z,t)$ in a weakly nonlinear regime. The principal scheme of RP model is given in Fig. 3. The

Optics EXPRESS

channel output $\mathbf{u}(z,t)$ is approximated using the perturbations according to expressions:

$$\begin{aligned}\mathbf{u}(z,t) &= \mathbf{u}_{\mathrm{L}}(z,t) + \mathbf{u}_{\mathrm{NL}}(z,t) + O(\gamma^2),\\ \mathbf{u}_{\mathrm{L}}(z,t) &= \mathcal{D}_z\left[\mathbf{u}(0,t) + \int_0^z dx\,\eta(x,t)\right],\\ \mathbf{u}_{\mathrm{NL}}(z,t) &= \sum_{m=1}^{N_{\mathrm{st}}} \mathcal{D}_{z-(m-0.5)\delta}\left[\mathcal{K}_{\delta,m}[\mathbf{u}_{\mathrm{L}}((m-0.5)\delta,t)]\right],\end{aligned} \tag{4}$$

with

$$\begin{aligned}\mathcal{D}_z[\cdot] &= \mathcal{F}^{-1}\left[\exp(i\beta_2 z\omega^2/2)\mathcal{F}[\cdot]\right],\\ \mathcal{K}_{\delta,m}[\mathbf{u}(t)] &= i\tfrac{8}{9}\gamma L_{\mathrm{eff}} f\left((m-1)\delta\right)\|\mathbf{u}(t)\|^2\mathbf{u}(t),\\ f(z) &= \exp\left(-\tfrac{\alpha}{2}\left(z - L_{\mathrm{sp}}\lfloor z/L_{\mathrm{sp}}\rfloor\right)\right), \quad L_{\mathrm{eff}} = \tfrac{1-e^{-\alpha\delta}}{\alpha}.\end{aligned}$$

Here $\delta = z/N_{\mathrm{st}}$ is the algorithm's spatial step size, $\|\cdot\|$ is the Euclidean vector norm, and $\mathcal{F}$ denotes the Fourier transform. $\mathcal{D}_z[\cdot]$ is the operator introducing the chromatic dispersion accumulated by our signal over the distance z, $\mathcal{K}_{\delta,m}[\cdot]$ introduces the Kerr nonlinear phase shift accumulated over the fiber span of length δ centered around the point $(m-0.5)\delta$, L_{eff} is effective step length, and $f(z)$ was introduced in Eq. (3). Since the linear distortion $\mathbf{u}_{\mathrm{L}}$ and every term of sum in $\mathbf{u}_{\mathrm{NL}}$ can be calculated independently, we refer to their calculation routines as to "branches". The number of branches $N_{\mathrm{br}} = N_{\mathrm{st}} + 1$ is a main RP-model hyper-parameter, defining both its precision of approximation and its complexity. Note that the amplifier noise should be realized in the model according to Eq. (4), where the integral $\int_0^z dx\,\eta(x,t)$ stands for all the ASE noise injected into the link on the coordinate range $x \in [0,z]$. In more detail, when calculating $\mathbf{u}_{\mathrm{L}}(z,t)$, prior to applying the $\mathcal{D}_z$ operator, one has to add to the input signal $\mathbf{u}(0,t)$ all the noises injected by the OAs situated in the range $x \in [0,z]$ including points $x = 0$ and $x = z$. For the single-span link with post-amplification, considered in this work, the ASE noise was effectively just added as an AWGN to $\mathbf{u}(z,t)$ calculated noiselessly. The case of multi-span transmission was considered in [36].

The RP model can be better understood through its comparison with the split-step Fourier method (SSFM) applied for solving the Manakov equations [32], see Figs. 2, 3. SSFM is formulated as a sequence of alternating steps: the linear steps introducing the dispersive broadening $D_\delta[\cdot]$ and accounting for linear losses, and the full nonlinear one, $\exp(K_{\delta,m})$, introducing Kerr nonlinear phase shift applied at the m-th step centered around point $z = (m-0.5)\delta$:

$$\begin{aligned}\mathbf{u}(z,t) &= \underbrace{\mathcal{D}_{\delta/2}\,\exp\left(\mathcal{K}_{\delta,m}\right)\,\mathcal{D}_{\delta/2}}_{\text{repeat } z/\delta \text{ times}}[\mathbf{u}(0,t)],\\ \exp\left(\mathcal{K}_{\delta,m}\right)[\mathbf{u}(t)] &= \exp\left(i\tfrac{8}{9}\gamma L_{\mathrm{eff}} f\left((m-1)\delta\right)\|\mathbf{u}(t)\|^2\right)\mathbf{u}(t).\end{aligned} \tag{5}$$

So we can think of the RP model as of the simplified SSFM model Eq. (5). In more detail, the RP model is the first order approximation of SSFM in the nonlinear parameter γ, meaning that it disregards the effect of all the nonlinear steps $\exp\left(\mathcal{K}_{\delta,m}\right)$ on each other, and reduces the exponent in the nonlinear step of SSFM in Eq. (5) to a linear approximation $\exp\left(\mathcal{K}_{\delta,m}\right) \approx 1 + \mathcal{K}_{\delta,m}$ via the expansion $\exp(x) \approx 1 + x + O(x^2)$.

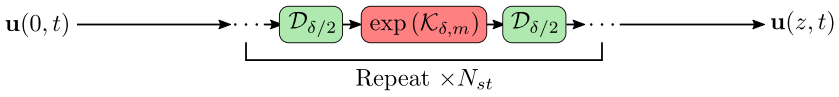


Fig. 2. Principal scheme of the split-step Fourier method, Eq. (5). The scheme is given to illustrate the derivation of RP model, Eq. (4).

Optics EXPRESS

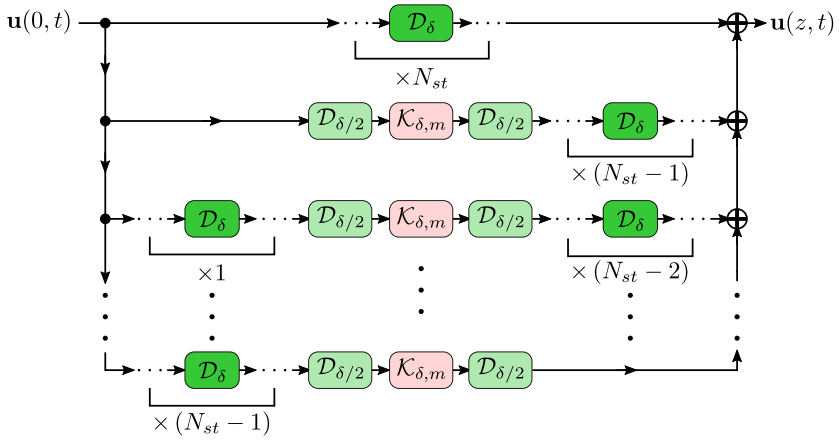


Fig. 3. Principal scheme of the first-order regular perturbation (RP) model [32] introduced in Eq. (4).

Compared to the "conventional" channel modelling by applying the full SSFM, the RP-model application for channel modelling has two main advantages in the context of end-to-end learning. First, the key advantage of RP is its parallel structure, i.e. its separability into the branches, where each particular branch is calculated via nearly the same routine. This RP property allows speeding up the computation of signal evolution and the computation of gradients via the "back-propagation" stage. Hence, the overall E2E learning is parallelizable in a straightforward way, which allows considerably speed-up its calculation by using modern multi-thread computational hardware, e.g. multicore processors and graphics processing units (GPUs). Second, the parallel structure of the RP model can increase the numerical stability of end-to-end learning. As noted in [52], the SSFM structure mimics the one of a typical convolutional NN made up of interchanging convolutions and point-wise nonlinearities. These complex multi-layered structures are prone to numerical errors in gradient estimation, referred to as infamous exploding/vanishing problem [9,43]. The simpler structure of RP model allows us to circumvent the gradient estimation problems by having a single nonlinear step per branch.

Notably, in addition to all the aforementioned benefits, the RP model offers a rather good approximation for the precise SSFM model in the practically important weakly nonliear regime [53] and around the optimal launch power level, where the link performance is the highest. At the same time, as expected, RP model poorly approximates optical channel at higher power levels. We measure how well does RP model approximate a precise SSFM one later in Sec. 3.2.

2.4. Receiver and performance metrics

At the receiver, the goal is to estimate the transmitted symbol x_n from each received symbol $y_n \in \mathbf{Y}$. The optimal way is to decide according to the posterior probability $P_{X|Y}(x_n|y_n)$. This is equivalent to computing the posterior probability $P(\mathbf{b}_n|y_n)$ based on the bit labels $\mathbf{b}_n = [b_{n,1}, \ldots, b_{n,K}], b_{n,j} \in \{0,1\}$. The alternative sub-optimal approach is to estimate the posterior probabilities $P(b_{n,j}|y_n)$ and decide on each $b_{n,j}$ individually. This approach corresponds to many practical receivers using bit-interleaved coded modulation with soft-decision binary forward error correction (FEC) codes. In such systems, the information rate determines (and is limited) by the *bit-wise mutual information (BMI)* Eq. (10) [38]. We consider here such a practical receiver.

We applied a commonly used mismatched Gaussian receiver (MGR) to approximate the bit posterior probabilities vector $q(b_{n,j}|y_n)$ [54,55]. In this approach, the conditional probability

Optics EXPRESS

linking the channel input and output, $q_{Y|X}$, is assumed to obey a Gaussian distribution:

$$q_{Y|X}(y_n|x_n) = \frac{1}{\pi\sigma_G^2}\exp\left(-\frac{\|y_n - x_n\|^2}{\sigma_G^2}\right), \tag{6}$$

with its variance σ_G estimated as the mean squared error between the received $\mathbf{Y}$ and the transmitted symbol sequence $\mathbf{X}$ recorded before the pre-distorter: $\sigma_G^2 = \mathbb{E}\left[\|Y - X\|^2\right]$. Here $\mathbb{E}[\cdot]$ is mathematical expectation. While more complex trainable decoders, notably, the artificial-neural-network based ones, can better approximate the exact channel distribution $p_{Y|X}$, MGR offers a practical cost-effective solution.

Applying the Bayes rule, we can write,

$$q_{X|Y}(x_n = s_m|y_n) = \frac{q_{Y|X}(y_n|x_n = s_m)p_S(s_m)}{\sum_{\tilde{m}=1}^{M} q_{Y|X}(y_n|x_n = s_{\tilde{m}})p_S(s_{\tilde{m}})}. \tag{7}$$

Let us define $\mathbf{S}_{0/1}^j := \{s_m \in \mathbf{S} : b_{m,j} = 0/1\}$ as the subset of transmission alphabet for which the j-th bit in the corresponding bit labels is set as '0' or '1', respectively. The bit posterior probability $q(b_{n,j}|y_n)$ is then is obtained by

$$q(b_{n,j} = 1|y_n) = \sum_{s_m \in \mathbf{S}} p(b_{n,j} = 1|x_n = s_m)\, q_{X|Y}(x_n = s_m|y_n) = \sum_{s_m \in \mathbf{S}_1^j} q_{X|Y}(x_n = s_m|y_n), \tag{8}$$

since $p(b_{n,j} = 1|x_n = s_m) = \{1, \text{ if } s_m \in \mathbf{S}_1^j;\ 0, \text{ if } s_m \notin \mathbf{S}_1^j\}$. Similarly,

$$q(b_{n,j} = 0|y_n) = \sum_{s_m \in \mathbf{S}_0^j} q_{X|Y}(x_n = s_m|y_n). \tag{9}$$

Note that, clearly, $q(b_{n,j} = 0|y_n) + q(b_{n,j} = 1|y_n) \equiv 1$.

2.5. Optimization and training procedure

During the training, we optimize the symbol locations, probabilities, and pre-distorter parameters via a batch gradient descent procedure, i.e., by a repeated generation of the training symbols' batches of fixed size and updating the trainable parameters using the numerical gradients of our objective function (averaged over the batches). The most challenging part here is the computation of gradient over individual symbol probabilities. We followed the training procedure proposed in [33]. This approach relies on a key consideration that the order of differentiation and expectation may not be exchangeable and this can affect the accuracy of numerical gradient calculation if the back-propagation algorithm is applied from some off-the-shelf toolboxes, like Tensorflow or PyTorch, without some necessary changes. To convey this point better, we reproduce here the algorithm presented in [33] in more detail.

Let $\bar{s}$ be a random constellation point generated by the transmitter according to symbol probability $P_S = \{p_S(s_1), p_S(s_2), \ldots, p_S(s_M)\}$. Since each symbol has a fixed bit labeling, we denote the corresponding random bit labeling by $\bar{\mathbf{b}} = [\bar{b}_1, \ldots, \bar{b}_K]$. Accordingly, the transmitter sends $\bar{x} = \bar{s}/\sqrt{\sum_{m=1}^{M} p_S(s_m)|s_m|^2}$. The power normalization is required to constrain the average power of the transmitter. Whatever the actual transmitter power is, it affects the channel distribution $p_{Y|X}(\bar{y}|\bar{x})$. The receiver receives symbol $\bar{y}$ according to $p_{Y|X}(\bar{y}|\bar{x})$ and computes the likelihood of each bit labels $\bar{b}_j$ according to the receiver's trained posterior distribution $q(\bar{b}_j|\bar{y})$. The objective is to train $q(\bar{b}_j|\bar{y})$ for all j at receiver as well as P_S and the set of constellation points

Optics EXPRESS

S to maximize the end-to-end BMI given by [38],

$$\mathcal{I} := H(\bar{s}) - \sum_{j=1}^{K} \mathbb{E}_{\bar{y}} \left[\mathcal{X} \left(p(\bar{b}_j|\bar{y}), q(\bar{b}_j|\bar{y}) \right) \right], \tag{10}$$

where $H(\bar{s})$ is the transmitter entropy, equal to

$$H(\bar{s}) := -\sum_{m=1}^{M} p_S(s_m) \log_2(p_S(s_m)) \tag{11}$$

and $\mathcal{X}(p(\bar{b}_j|\bar{y}), q(\bar{b}_j|\bar{y}))$ is the cross-entropy between the actual posterior distribution $p(\bar{b}_j|\bar{y})$ and its approximation $q(\bar{b}_j|\bar{y})$ from our receiver for each bit-label b_j. Explicitly, for each b_j

$$\mathbb{X}(j) := \mathbb{E}_{\bar{y}} \left[\mathcal{X} \left(p(\bar{b}_j|\bar{y}), q(\bar{b}_j|\bar{y}) \right) \right] = -\sum_{m=1}^{M} p_S(s_m) \int p_{Y|X}(y|x_m) \log_2 \left(q(\bar{b}_j|y) \right) \mathrm{d}y \tag{12}$$

where $x_m = s_m / \sqrt{\sum_{\omega=1}^{M} p_S(s_\omega)|s_\omega|^2}$. It is known that for a given P_S and constellation points S, Eq. (12) is minimized when $q(\bar{b}_j = 1|\bar{y})$ becomes equal to $p(\bar{b}_j = 1|\bar{y})$ for "almost every" y values (e.g. see [38]). Therefore, the learning task of receiver is to output $q(\bar{b}_j = 1|\bar{y})$ as close as possible to $p(\bar{b}_j = 1|\bar{y})$. To remind $q(\bar{b}_j = 0|\bar{y})$ then can be estimated as $q(\bar{b}_j = 0|\bar{y}) = 1 - q(\bar{b}_j = 0|\bar{y})$.

In our case of using MGR, no receiver training is needed. Nonetheless, if one instead employs another trainable model as a receiver, e.g. an artificial neural network, the receiver learning becomes simply equivalent to training of K independent binary classifiers $q(\bar{b}_j|y)\, \forall j \in \{1, \ldots, K\}$. This optimization is a classical logistic regression, so we skip it here and refer to [56], where the authors detailed the training of such receivers.

The learning task of transmitter is more challenging. We should optimize P_S and S such that the aggregate BMI $\mathcal{I}$ is maximized. Since P_S is constraint by $P_S(s_m) \geq 0$ and $\sum P_S(s_m) = 1$, it is more recommended to train its unconstrained logarithm ℓ_m such that

$$P_S(s_m) = \frac{\exp(\ell_m)}{\sum_{\omega=1}^{M} \exp(\ell_\omega)}. \tag{13}$$

Therefore, the training requires the partial derivatives of $\mathcal{I}$ in terms of s_m, ℓ_m and pre-distortion parameters $C_{m,n}$. Over symbol's location s_i,

$$\frac{\partial \mathcal{I}}{\partial s_i} = \sum_{j=1}^{K} \sum_{m=1}^{M} p_S(s_m) \int p_{Y|X}(y|x_m) \frac{\partial \log_2 \left(q(\bar{b}_j|y) \right)}{\partial s_i} \mathrm{d}y. \tag{14}$$

The above expression is obtained because probabilities $p_S(s_m)$ are another optimization parameters independent of the considered location s_i. Moreover, the dependence of channel conditional distribution $p_{Y|X}(y|x_m)$ on x_m and, accordingly on s_m can be neglected. Recall that the role of the channel is to draw random samples y independently according to $p_{Y|X}(y|x_m)$. In typical communication systems, the channel is mainly governed by independent random noise sources, e.g. AWGN, and thus its dependency on x_m is a secondary effect and it can be ignored. The dependency of $q(\bar{b}_j|y)$ on s_m is expressed from Eqs. (8, 9). Note that Eq. (14) is the basis of geometric shaping optimization in [20,45,56] where the gradient is computed not directly, but using back-propagation algorithm thorough the entire differential auto-encoder blocks in Fig. 1 from a training batch of input-output samples. We explain the numerical optimization later.

Optics EXPRESS

Differentiation over pre-distortion parameters $C_{m,n}$ is followed similarly, i.e.

$$\frac{\partial \mathcal{I}}{\partial C_{m,n}} = \sum_{j=1}^{K}\sum_{i=1}^{M} p_S(s_i)\int p_{Y|X}(y|x_i)\frac{\partial \log_2\left(q(\bar{b}_j|y)\right)}{\partial C_{m,n}}\mathrm{d}y, \tag{15}$$

and training is similarly done using batch gradient ascent algorithm and applying back-propagation algorithm.

Training over ℓ_m is slightly different as several terms in objective function $\mathcal{I}$ depend on ℓ_m. Let us first express the differentiation explicitly and then explain how one computes it using back-propagation algorithm. Differentiation of entropy $H(\bar{s})$ results

$$\frac{\partial H(\bar{s})}{\partial \ell_m} = -\log_2(e)p_S(s_m)\left(\ell_m - \sum_{\omega=1}^{M} p_S(s_\omega)\ell_\omega\right), \tag{16}$$

where $e = \exp(1)$ and $P_S(s_m)$ in terms of ℓ_m is defined in Eq. (13). Differentiation of $\mathbb{X}(j)$ over ℓ_m results in the sum of two terms

$$\begin{aligned}\frac{\partial \mathbb{X}(j)}{\partial \ell_m} = -p_S(s_m)\left(\mathbb{X}(j) + \int p_{Y|X}(y|x_m)\log_2\left(q(\bar{b}_j|y)\right)\mathrm{d}y\right)\\ - \sum_{i=1}^{M} p_S(s_i)\int p_{Y|X}(y|x_i)\frac{\partial \log_2\left(q(\bar{b}_j|y)\right)}{\partial \ell_m}\mathrm{d}y.\end{aligned} \tag{17}$$

Summing all terms together, we have the differentiation of $\mathcal{I}$ as

$$\frac{\partial \mathcal{I}}{\partial \ell_m} = \frac{\partial H(\bar{s})}{\partial \ell_m} - \sum_{j=1}^{K}\frac{\partial \mathbb{X}(j)}{\partial \ell_m}. \tag{18}$$

We explain now how to train the auto-encoder via batch gradient ascent algorithm to numerically maximize $\mathcal{I}$. Recall that the considered transmitter consists of two stages: sampler and pre-distorter. To generate a batch of size N, the sampler randomly draws with replacement N indices from the current symbol distribution $P_S = \{p_S(s_1), p_S(s_2), \ldots, p_S(s_M)\}$, $s_m \in \mathbf{S}$. Each drawn index, let say s_n, is mapped to a corresponding transmitted normalized symbol $x_n = s_n/\sqrt{\sum_{\omega=1}^{M} p_S(s_\omega)|s_\omega|^2}$ with a given, fixed bit label $\mathbf{b}_n$, e.g. Gray bit labeling. The sampled normalized symbols x_n and its corresponding bit-labels $\mathbf{b_n}$ are stacked, respectively, in the input symbol vector $\mathbf{X} = [x_1, x_2, \ldots, x_N]$ and the bit vector $\mathbf{B} = [\mathbf{b}_1, \ldots \mathbf{b}_N]$. The symbol power normalization is required to keep the batch power constant during the training.

The generated input batch is processed by a cascade of pre-distorter, followed by the channel model. The cascade maps input symbols $\mathbf{X}$ to the vector of channel output symbols $\mathbf{Y} = [y_1, y_2, \ldots, y_N]$ according to the cascade's conditional probability density function $P_{Y|X}$. The vectors of input symbols $\mathbf{X}$, input bit-labels $\mathbf{B}$, and received symbols $\mathbf{Y}$ constitute the training dataset. The differentiability of both pre-distorter and channel model allows calculating the gradients of the output symbols over the input symbols and the pre-distorter parameters. The next step is to compute the gradient numerically and apply batch gradient ascent algorithm [42, Sec. 8.1.3]. Notably, gradient ascent, not descent as usually done in ML, should be applied there since we are interested in increasing the optimized metrics BMI. For instance, Eq. (14) is numerically estimated by,

$$\frac{\partial \mathcal{I}}{\partial s_i} \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{K}\frac{\partial \log_2\left(q(b_{n,j}|y_n)\right)}{\partial s_i}. \tag{19}$$

Similarly, the gradient of pre-distortion parameters $C_{m,n}$ can be expressed numerically in terms of training batch. The usual remedy is to apply back-propagation algorithm [42, Sec. 6.5] to

Optics EXPRESS

compute these gradients efficiently. The back-propagation is in fact based on the chain rule of derivatives, deriving the gradients at the input of a layer from the gradients at the output of the layer. This way, the gradients "propagate backward" from the final output variables $z_{n,j} = q(b_{n,j}|y_n)$ to the input parameters $\{s_i\}$ or the pre-distortion parameters $\{C_{m,n}\}$.

Numerical computation of gradient for ℓ_m is slightly different. The derivatives of $\frac{\partial H(\bar{s})}{\partial \ell_m}$ are expressed in terms of optimized parameters in Eq. (16) and, hence, do not need to be computed via training data. The derivatives of $\frac{\partial \mathbb{X}(j)}{\partial \ell_m}$ must be however computed numerically. Let $\mathbb{Y}_m$ denote the subset of outputs y_n in the batch for which the same constellation point s_m produced their corresponding input symbol, i.e., $x_n = s_m/\sqrt{\sum_{\omega=1}^{M} p_S(s_\omega)|s_\omega|^2}$. Let N_m denote the number of output samples in $\mathbb{Y}_m$. Then, Eq. (17) can be partitioned into three terms

$$\frac{\partial \mathbb{X}(j)}{\partial \ell_m} = D_1(j,m) + D_2(j,m) + D_3(j,m), \tag{20}$$

where

$$D_1(j,m) \approx \frac{p_S(s_m)}{N} \sum_{n=1}^{N} \log_2\left(q(b_{n,j}|y_n)\right) \tag{21}$$

$$D_2(j,m) \approx -\frac{1}{N} \sum_{y_n \in \mathbb{Y}_m} \log_2\left(q(b_{n,j}|y_n)\right) \text{ (remark : } s_m \text{ is the input for those } y_n) \tag{22}$$

$$D_3(j,m) \approx -\frac{1}{N} \sum_{n=1}^{N} \frac{\partial \log_2\left(q(b_{n,j}|y_n)\right)}{\partial \ell_m}. \tag{23}$$

As a result, the numerical estimate of Eq. (18) is expressed by,

$$\frac{\partial \mathcal{I}}{\partial \ell_m} \approx \frac{\partial H(\bar{s})}{\partial \ell_m} - \sum_{j=1}^{K} \left(D_1(j,m) + D_2(j,m) + D_3(j,m)\right). \tag{24}$$

Note that $D_3(j,m)$ has a similar form to Eq. (19). Thus, this term is computed via back-propagation, as explained before. The other terms $D_1(j,m)$ and $D_2(j,m)$ are available as they are part of the cost function $\mathcal{I}$ (indeed, $D_2(j,m)$ is the sum of a subset of terms participating in the sum in $D_1(j,m)$). These terms and $\frac{\partial H(\bar{s})}{\partial \ell_m}$ Eq. (16) are available and no back-propagation is required to compute them. To summarize, one needs to compute $D_3(j,m)$ terms via back-propagation from output variables $z_{n,j} = q(b_{n,j}|y_n)$ to the design parameters ℓ_m. And then its results must be summed up with other terms, i.e. $D_1(j,m)$, $D_2(j,m)$ and $\frac{\partial H(\bar{s})}{\partial \ell_m}$.

Finally, all the computed gradients were then used by the Adam optimizer [57] to train the respective communication system parameters. This completes the training process of our E2E system.

Remark: if one computes the derivatives from the numerical estimate of $\mathcal{I}$, i.e.

$$\mathcal{I}_{\text{num.}} \approx H(\bar{s}) + \frac{1}{N} \sum_{j=1}^{K} \sum_{n=1}^{N} \log_2\left(q(b_{n,j}|y_n)\right) \tag{25}$$

then one would find that $\partial \mathcal{I}_{\text{num.}}/\partial s_i$ is the same as Eq. (19). Similarly, $\partial \mathcal{I}_{\text{num.}}/\partial C_{m,n}$ is equal to our approximation, but this is not true for $\partial \mathcal{I}_{\text{num.}}/\partial \ell_m$ which excludes $D_1(j,m)$ and $D_2(j,m)$ terms. The reason isthat a change in statistics of sampler, brought by updating symbol probabilities $p_S(s_m)$ cannot be precisely expressed via the already generated samples.

Optics EXPRESS

3. Results

3.1. Testcase

Having proposed the end-to-end learning algorithm for the general coherent fiber-optic communication link, we illustrate its benefits on the particular case of a single-span link. More specifically, we numerically consider the dual-polarized (DP) 64 GBd transmission of 256-symbol constellations over the 1x170 km single mode fiber (SMF) link. The SMF parameters were taken as: chromatic dispersion coefficient D = 16.8 ps/(nm*km), effective nonlinear coefficient $\gamma = 1.14$ (W*km)$^{-1}$, loss coefficient $\alpha = 0.21$ dB/km. The span is followed by a lumped optical amplifier (OA) with the noise figure 4.5 dB. No transceiver phase noise was considered.

During the E2E system training, we used the RP model, Eq. (5), with $N_{br} = 100$ branches, as an auxiliary channel model. At the same time, the performance of the learnt constellations was estimated using the "precise" SSFM, i.e. via Eq. (4) channel model.

In our current work, we focus on a single-span link, because in such a case the nonlinearity-aware constellation shaping is expected to produce a considerable gain [35]. At the same time, we emphasize that our method is equally applicable for other fiber-optic communication systems, where, of course, the ultimate gain figures can be different. First, the training procedure, described in Section 2.5, was successfully applied in [33] to E2E learning the constellation shaping for the AWGN channel. The obtained results agreed with the analytical predictions for AWGN channel [58]. Particularly, the learned PS matched the performance of the "optimal" MB shaping Eq. (1), while the learned combination of PS and GS slightly outperformed MB, thanks to better ability to approach the truly optimal continuous Gaussian input symbol distribution. Second, in [36] the similar E2E learning technique was successfully applied to train the geometric constellation shaping and the nonlinear pre-distorter for the 64 GBd transmission over the long-haul 30x80 km (2400 km) SMF link.

Nonetheless, we expect the proposed E2E learning algorithm to reap smaller performance gains over MB shaping Eq. (1) in the links with bigger numbers of spans or co-propagating channels. We attribute it to the weakening of the overall gain obtainable by a nonlinearity-aware constellation shaping with the increase in link length [35] or number of co-propagated channels [59]. The channel distribution $p_{Y|X}$ in these testcases becomes more a Gaussian-like [47], for which MB shaping is optimal.

3.2. RP model channel approximation precision

We start from showing that RP model, despite being a simplified one, offers a decent approximation of the precise SSFM model. For the testcase, considered in this paper, we compared how do RP and SSFM models simulate the propagation of unshaped 256-QAM signal. In Fig. 4, first, we plot the signal-to-noise ratio (SNR) of the received signals after the chromatic dispersion compensation (CDC) was applied. One can see that the SNR is nearly the same for both the SSFM and RP models in the weakly nonlinear regime (up to 10 dBm), and, most notably, the correspondence is still good around the optimal launch power level of $P_{\text{opt}} \approx 10$ dBm. To demonstrate that this SNR-value-similarity comes directly from RP model correctly approximating the deterministic nonlinear distortions introduced by SSFM, we compared the outputs of the models in the noiseless scenario: we fed the same input signal to both models, while they injected no ASE noise, i.e. only the deterministic effects were considered. We quantified the difference between the outputs of RP $\mathbf{Y}_{\text{RP}}$ and SSFM $\mathbf{Y}_{\text{SSFM}}$ in terms of signal-to-distortion ratio (SDR), defined as $-20\log_{10}(\|\mathbf{Y}_{\text{SSFM}}\|/\|\mathbf{Y}_{\text{RP}} - \mathbf{Y}_{\text{SSFM}}\|)$, also added to Fig. 4(b). We see there that up to the launch power of 10.5 dBm, SDR is at least 20 dB larger than the SSFM-modelled received SNR. This indicates that the approximation error of the RP model is much smaller than the total distortion in the link. To sum up, from Fig. 4 we readily see that the RP model renders a very

Optics EXPRESS

good channel approximation for the case considered, and the deterministic mismatch between the models cannot noticeably affect the E2E system training.

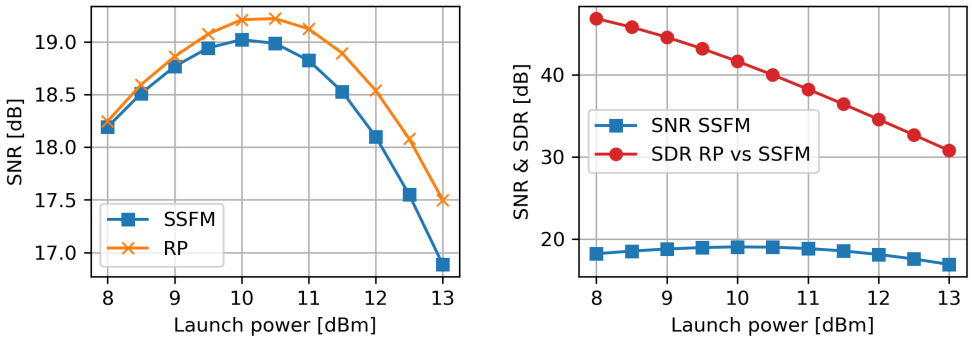


Fig. 4. Comparison between channel models based on first-order regular perturbation (RP) and split-step Fourier method (SSFM) approximating the 64 GBd single channel dual-polarised transmission of unshaped 256QAM signal over 1x170 km SMF link.

3.3. Learning the constellation shaping without pre-distorter

Once we have tested the accuracy of the RP model, we turn to the results of the E2E learning the constellation shaping.

In the first case, we consider learning the constellation shaping in a link without a pre-distorter, i.e. it was disabled and we learnt only PS and GS of separate symbols. This option is well suited for the use cases when there is no opportunity to implement a separate pre-distorter at the transmitter, typically, because of the limited complexity and/or power budget.

Before training, we initialized the encoder with a MB-shaped 256QAM constellation, defined in Eq. (1). To find the MB shaping parameter θ we, first, initialized the constellation as an unshaped 256QAM one and recorded the SNR produced by this constellation between the input and output of the auxiliary channel model. The MB shaping parameter θ was chosen as to have the optimal value for the AWGN channel, producing the same SNR level. The MB constellation with the θ parameter value found in this way was used as a starting seed for all the following learning of constellation shapings and its performance was used as a reference one.

We did the two types of constellation shaping learning. In the first experiment, we learnt just the PS. This is a preferable case for the links, where introducing the GS is not favored. The benefit of PS is that, while providing the performance gain, it keeps the square grid of QAM constellation intact (see Fig. 6(b)). Having the square grid of the transmitted constellation allows one to use the cost-effective blind DSP algorithms to recover the signal at the receiver [60] and lower the required precision of a digital-to-analog converter at the transmitter by limiting the range of signal powers along I- and Q- component present in the transmitted signal. In the second experiment, we jointly learnt the PS and GS to highlight the full single-symbol joint shaping (JS) gains reachable by the E2E learning strategy. For the PS, the training started from the reference MB shaped constellation. Then, the PS, learnt at the first stage, was used as an initial seed for the JS E2E learning.

The whole training procedure: constellation initialization as the MB-shaped one, learning the PS, and learning the JS, was done separately for a range of launch power levels. Figure 5 showcases the performance achieved by these shaped constellations in a validation SSFM simulation. The learnt PS (E2E-PS-256QAM) resulted in the 0.043 bits/2D-symbol/pol BMI gain on top of MB-shaping, and led to $\approx$ 0.25 dBm increase in optimal power level. The learnt JS

Optics EXPRESS

(E2E-JS-256) including both PS and GS, led to nearly doubling the performance gain: it resulted in 0.074 bits/2D-symbol/pol BMI gain on top of MB-shaping.

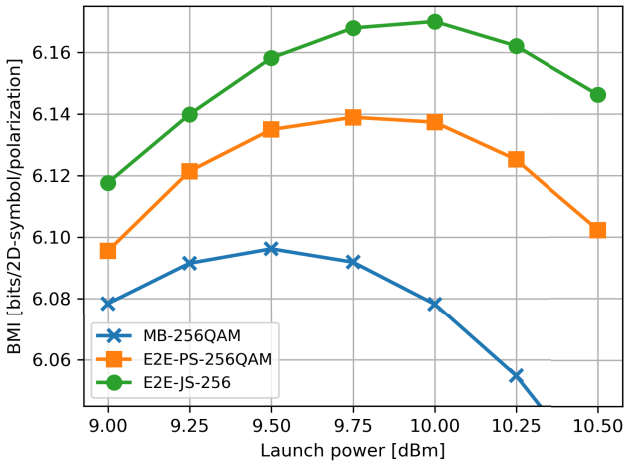
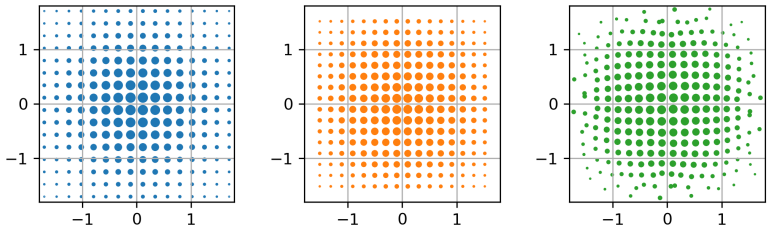


Fig. 5. The performance of the constellation shaping without pre-distortion: the reference Maxwell-Boltzmann (MB-256QAM) shaping, learnt probabilistic shaping (E2E-PS-256QAM), and the learnt joint probabilistic and geometric constellation shaping (E2E-JS-256).



(a) MB-256QAM (b) E2E-PS-256QAM (c) E2E-JS-256

Fig. 6. Constellations applied at the optimal power level for the case of the link without nonlinear pre-distortion.

The better performance of learned constellations can be partly explained by the fact that they produce a higher effective SNR in the link. Figure 7(a) shows that the effective SNR of the learned JS is higher than that of the reference MB shaping. Furthermore, we can quantify the difference between the constellations, leading to different SNR values. The extended Gaussian noise (EGN) model [47] suggests that the SNR of a signal propagated over a non-linear channel depends on the constellation and is inversely proportional to its 4th μ_4 and 6th μ_6 standard moments, where the k-th standard moment μ_k of the input symbol sequence $\mathbf{X}$ is defined as

$$\mu_k[\mathbf{X}] = \frac{\mathbb{E}\left[|\mathbf{X} - \mathbb{E}[\mathbf{X}]|^k\right]}{\left(\mathbb{E}\left[|\mathbf{X} - \mathbb{E}[\mathbf{X}]|^2\right]\right)^{k/2}} \tag{26}$$

and $\mathbb{E}[\cdot]$ stands for the expectation value. Indeed, Figs. 7(b), 7(c) show that μ_4, μ_6 moments of the E2E learnt JS are lower than the ones for the MB constellation. Furthermore, the difference in SNR, and μ_4, μ_6 moments, increases with the rise of the launch power, when the nonlinear distortions strengthen.

Optics EXPRESS

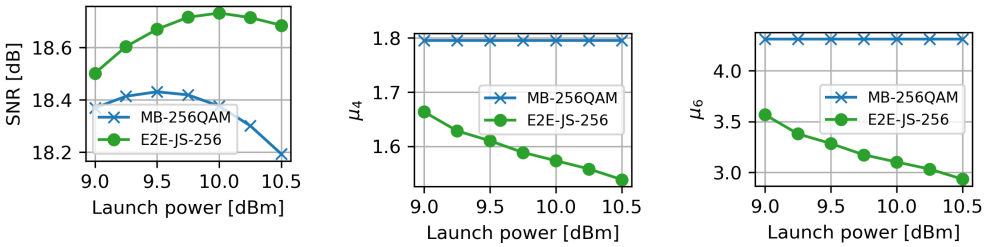


Fig. 7. The comparison between the metrics of reference Maxwell-Boltzmann 256QAM (MB-256QAM) and the E2E learnt JS (E2E-JS-256) constellations.

3.4. Learning the cost-effective pre-distorter via end-to-end learning

Once we have considered the single-symbol conventional constellation shaping, we move to the more advanced case of memory-aware shaping, i.e., the shaping when the transmitted symbol depends not only on the message transmitted in the corresponding time slot, but also on the messages transmitted in the neighbouring slots. To be effective, a constellation shaping should take into account the symbols co-propagating in the neighbouring slots, since they contribute to the deterministic nonlinear distortions introduced into the transmitted symbol of interest. We name this approach as the *multi-symbol constellation shaping* (MSCS).

Thankfully, the RP-based auxiliary channel model, Eq. (4), implemented into the E2E learning algorithm proposed in this work, allows the E2E learning of MSCS, since the RP introduces inter-symbol deterministic nonlinear distortions. Conversely, the EGN-based auxiliary channel model [47], suggested in previous works for end-to-end learning the coherent optical communications [19,45], does not support the MSCS learning: the EGN replaces the actual inter-symbol nonlinear distortion with the symbol-independent Gaussian noise and, therefore, prevents the E2E learning of dependencies between the neighbouring symbols.

We propose the MSCS implementation as a combination of single-symbol JS, described in the previous subsection, and the nonlinear PPD-based pre-distorter, Eq. (2), jointly trained in a single run of E2E learning. Before considering the MSCS training as a whole, we focus on the E2E learning for the cost-effective pre-distorter. We note that it is important to reduce the complexity of pre-distorter, inasmuch as its complexity defines the additional costs of MSCS implementation over the single-symbol JS, and, therefore, denominates the feasibility of the whole MSCS implementation.

As mentioned in Sec. 2.2, the PPD-based pre-distorter performance-to-complexity ratio is defined by the range of nonlinear perturbation terms $T_{m,n}$. An effective way to set it is by keeping in the algorithm only the terms $T_{m,n}$ for which the corresponding coefficient $|C_{m,n}|$ has the absolute value above the fixed cut-off threshold: $|C_{m,n}|>C_{\text{cutoff}}$. We refer to this approach as the pre-distorter pruning, in analogy with the similar technique from neural networks optimization [61].

We seek to find the optimal C_{cutoff} value by a grid search. First, we loaded the unshaped 256QAM constellation to the sampler, and the constellation sampler from E2E learning. Then we trained the PPD with the range of coefficients $|m|, |n| \leq 10$, and pruned it with the range of various cut-off values C_{cutoff}. We measured the performance of the generated family of E2E learnt pre-distorters on the precise SSFM channel model. The dependence of the link performance with the various pruned PPD, on the share of pruned coefficients $C_{m,n}$, is given in Fig. 8. The figure shows that when we keep around 25% of the most significant PPD terms $T_{m,n}$, adding the new ones leads to the negligible improvement in resulting algorithm's performance. This pruning scheme will be used in the following MSCS learning.

Optics EXPRESS

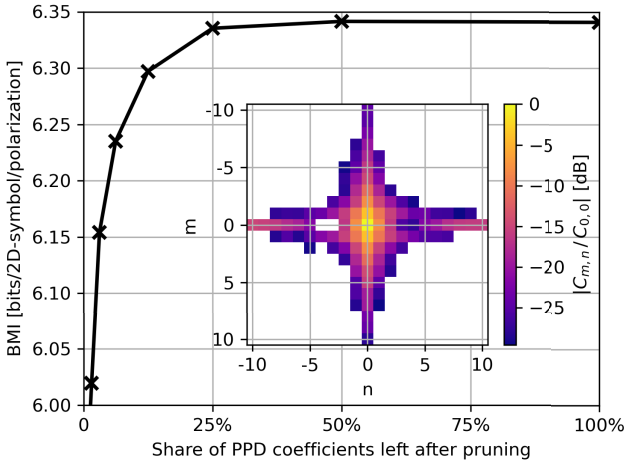


Fig. 8. The dependence of the performance of pruned indirectly learned perturbation-based pre-distorter (PPD) with $|m|, |n| \leq 10$ on the margin of the most significant $C_{m,n}$ coefficients non-zeroed during the pruning procedure. The PPD training was done on the RP model, while the performance was measured in precise SSFM simulation. Inset: The distribution of PPD coefficients $C_{m,n}$ learned and pruned with the cut-off leaving 25% of the coefficients.

Notably, the distribution of the learnt PPD coefficients $|C_{m,n}|$ agrees with the approximation suggested in [62], implying that $|C_{m,n}| \sim 1/|mn|$. The inset of Fig. 8 shows the distribution of the absolute value of coefficients $|C_{m,n}|$ in the aforementioned indirectly trained PPD with 25% coefficients left.

3.5. *Learning the memory-aware constellation shaping*

We have defined the cost-effective perturbation-based pre-distorter (PPD) in the previous subsection, and now we consider the E2E learning of the MSCS enabled by the pre-distorter. The MSCS estimates the performance gains reachable by the proposed E2E learning algorithm, when no complexity-limiting constraints are put on it. The set of non-pruned triplets $T_{m,n}$, defined during the PPD initialization, was kept fixed during the following E2E learning. After the initialization, we simultaneously optimized the PPD with joint probabilistic and geometric single-symbol constellation shaping, thereby arriving at the MSCS shaping.

The pre-distorter initialization followed by the E2E MSCS learning was done separately for a range of launch powers. The performance of the resulting learnt constellation measured on the precise SSFM channel model is given in Fig. 9. For comparison, we also plotted the performance of the single-symbol constellation shapings learnt in Sec. 3.3: MB-256QAM and E2E-JS-256. Compared to the reference MB constellation shaping, the MSCS led to the considerable improvement in the system's BMI, adding 0.47 bits/2D-symbol/pol., and the optimal power level moved up by ≈ 1.25 dBm.

To clarify the reason for the superior MSCS performance, we separately considered its contributing factors: PPD and fine-tuning the JS. To do this, we compared the performance of MSCS, involving the simultaneous training of JS and PPD, with the PPD trained on top of fixed pre-learnt single-symbol E2E-JS-256 (E2E-JS-256 + PPD). In the latter case, the constellation shaping isn't by affected by the presence of PPD in the link, which allows us to isolate the performance gains of PPD. The results of this separate E2E learning are given in Fig. 9. Notably, PPD is found to be the main contributor to the MSCS performance: training the PPD on top of fixed JS already brought 0.39 bits/2D-symbol/pol BMI gain, while E2E learned MSCS brought only 0.012 bits/2D-symbol/pol BMI gain on top of this. Despite this result provides a valuable

Optics EXPRESS

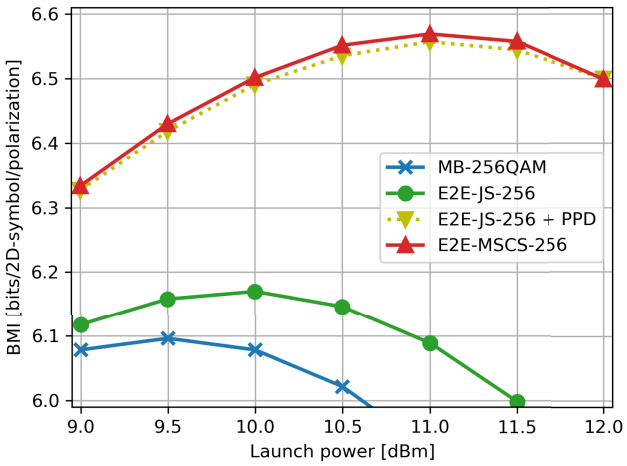


Fig. 9. The performance of the end-to-end learnt multi-symbol constellation shaping (E2E-MSCS-256). For reference, we added here the performance of Maxwell-Boltzmann (MB-256QAM), learnt JS (E2E-JS-256), and the nonlinear perturbation-based pre-distorter trained on top of the fixed E2E-JS-256 constellation (E2E-JS-256 + PPD).

insight into the reasons for the performance of the MSCS, we consider it necessary to conduct a detailed study analysing different link testcases and deep learning hyperparameter values before making any sound conclusions. This study is left beyond the scope of this work.

4. Conclusion

The end-to-end learning of the optical coherent detection communication system offers a possibility to specialize the characteristics of the transmitted signal to the properties of the nonlinear channel. However, we underline that the two main problems are typically associated with the E2E learning implementation.

The first problem relates to the overall complexity of modelling the channel distortions, where the latter constitute an intricate mix of instantaneous nonlinear fiber responses intertwining with a dispersive pulse broadening. The modelling of modern high-baudrate links is especially difficult because of a huge dispersive memory, implying a considerable complexity and time consumption of the split step simulations. In our work, we address this issue by proposing a parallelizable simplified channel model based on the first-order regular perturbation [32].

The second challenge is the difficulty of introducing the concept of trainable discrete probability distribution into the machine learning algorithm. In our paper, this problem is addressed by adopting a novel training procedure proposed first in [33] for the AWGN channel.

The resulting composite solution, proposed and demonstrated in this work, made possible learning the joint probabilistic and geometric shaping of symbol sequences in coherent fiber-optic communication links. Despite the considered approach is still suboptimal, the computed multi-symbol joint probabilistic and geometric shaping has shown a considerable bit-wise mutual information (BMI) improvement of 0.47 bits/2D-symbol/pol over the conventional Maxwell-Boltzmann shaping for a single-channel 64 GBd transmission over the 170 km SMF link.

Moreover, we observed that the proposed end-to-end learning is applicable in the situations when, because of hardware or complexity limitations, we cannot use the multi-symbol shaping. For the same test case, we found that a single-symbol joint probabilistic and geometric shaping gives 0.074 bits/2D-symbol/pol BMI gain over the reference MB shaping. And in the case when

Optics EXPRESS

geometric shaping is not possible, the single-symbol probabilistic shaping can be used: it also outperforms the MB shaping by 0.043 bits/2D-symbol/pol in terms of BMI.

We believe that the end-to-end learning approach, proposed in our work, can lay a path to finding the optimal signal distribution for a variety of nonlinear fiber-optic channels.

Funding. H2020 Marie Skłodowska-Curie Actions (766115); Engineering and Physical Sciences Research Council (EP/R035342/1); Leverhulme Trust (RP-2018-063).

Acknowledgements. This project has received funding from EU Horizon 2020 program under the Marie Skłodowska-Curie grant agreement No. 766115 (FONTE). SKT acknowledges the support of EPSRC project TRANSNET. JEP acknowledges Leverhulme Trust Project RP-2018-063.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. E. Agrell, A. Alvarado, and F. R. Kschischang, "Implications of information theory in optical fibre communications," Phil. Trans. R. Soc. A. **374**(2062), 20140438 (2016).
2. P. J. Winzer, D. T. Neilson, and A. R. Chraplyvy, "Fiber-optic transmission and networking: the previous 20 and the next 20 years," Opt. Express **26**(18), 24190–24239 (2018).
3. C. E. Shannon, "A mathematical theory of communication," The Bell Syst. Tech. J. **27**(3), 379–423 (1948).
4. A. P. T. Lau and F. N. Khan, eds., *Machine Learning for Future Fiber-Optic Communication Systems* (Academic Press, 2022).
5. F. N. Khan, Q. Fan, C. Lu, and A. P. T. Lau, "An optical communication's perspective on machine learning and its applications," J. Lightwave Technol. **37**(2), 493–516 (2019).
6. D. Zibar, M. Piels, R. Jones, and C. G. Schäeffer, "Machine learning techniques in optical communication," J. Lightwave Technol. **34**(6), 1442–1452 (2016).
7. 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 Commun. Surv. Tutorials **21**(2), 1383–1408 (2019).
8. J. W. Nevin, S. Nallaperuma, N. A. Shevchenko, X. Li, M. S. Faruk, and S. J. Savory, "Machine learning for optical fiber communication systems: An introduction and overview," APL Photonics **6**(12), 121101 (2021).
9. P. J. Freire, A. Napoli, B. Spinnler, N. Costa, S. K. Turitsyn, and J. E. Prilepsky, "Neural networks-based equalizers for coherent optical transmission: Caveats and pitfalls," IEEE J. Sel. Top. Quantum Electron. **28**(4), 1–23 (2022).
10. P. J. Freire, D. Abode, J. E. Prilepsky, and S. K. Turitsyn, "Power and modulation format transfer learning for neural network equalizers in coherent optical transmission systems," in *Signal Processing in Photonic Communications*, (Optical Society of America, 2021), pp. SpM5C–6.
11. T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," IEEE Trans. Cogn. Commun. Netw. **3**(4), 563–575 (2017).
12. X. Chen, J. Cheng, Z. Zhang, L. Wu, J. Dang, and J. Wang, "Data-rate driven transmission strategies for deep learning-based communication systems," IEEE Trans. Commun. **68**(4), 2129–2142 (2020).
13. J. Song, C. Häger, J. Schröder, A. G. I. Amat, and H. Wymeersch, "Model-based end-to-end learning for wdm systems with transceiver hardware impairments," IEEE J. Sel. Top. Quantum Electron. **28**(4), 1–14 (2022).
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," J. Lightwave Technol. **36**(20), 4843–4855 (2018).
15. B. Karanov, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-end optimized transmission over dispersive intensity-modulated channels using bidirectional recurrent neural networks," Opt. Express **27**(14), 19650–19663 (2019).
16. B. Karanov, M. Chagnon, V. Aref, D. Lavery, P. Bayvel, and L. Schmalen, "Concept and experimental demonstration of optical im/dd end-to-end system optimization using a generative model," in *Optical Fiber Communication Conference*, (Optica Publishing Group, 2020), pp. Th2A–48.
17. B. Karanov, V. Oliari, M. Chagnon, G. Liga, A. Alvarado, V. Aref, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-end learning in optical fiber communications: Experimental demonstration and future trends," in *2020 European Conference on Optical Communications (ECOC)*, (IEEE, 2020), pp. 1–4.
18. S. Gaiarin, F. Da Ros, R. T. Jones, and D. Zibar, "End-to-end optimization of coherent optical communications over the split-step fourier method guided by the nonlinear fourier transform theory," J. Lightwave Technol. **39**(2), 418–428 (2021).
19. R. T. Jones, T. A. Eriksson, M. P. Yankov, and D. Zibar, "Deep learning of geometric constellation shaping including fiber nonlinearities," in *2018 European Conference on Optical Communication (ECOC)*, (IEEE, 2018), pp. 1–3.
20. R. T. Jones, M. P. Yankov, and D. Zibar, "End-to-end learning for gmi optimized geometric constellation shape," in *2019 European Conference on Optical Communication (ECOC)*, (IET, 2019), pp. 1–3.
21. K. Gümüş, A. Alvarado, B. Chen, C. Häger, and E. Agrell, "End-to-end learning of geometrical shaping maximizing generalized mutual information," in *2020 Optical Fiber Communications Conference and Exhibition (OFC)*, (IEEE, 2020), pp. 1–3.

Optics EXPRESS

22. V. Oliari, B. Karanov, S. Goossens, G. Liga, O. Vassilieva, I. Kim, P. Palacharla, C. Okonkwo, and A. Alvarado, "High-cardinality hybrid shaping for 4d modulation formats in optical communications optimized via end-to-end learning," arXiv arXiv:2112.10471 (2021).
23. Z. Niu, H. Yang, H. Zhao, C. Dai, W. Hu, and L. Yi, "End-to-end deep learning for long-haul fiber transmission using differentiable surrogate channel," J. Lightwave Technol. **40**(9), 2807–2822 (2022).
24. O. Jovanovic, M. P. Yankov, F. Da Ros, and D. Zibar, "End-to-end learning of a constellation shape robust to variations in snr and laser linewidth," in *2021 European Conference on Optical Communication (ECOC)*, (IEEE, 2021), pp. 1–4.
25. J. Song, C. Häger, J. Schröder, A. G. i Amat, and H. Wymeersch, "End-to-end autoencoder for superchannel transceivers with hardware impairment," in *Optical Fiber Communication Conference*, (Optica Publishing Group, 2021), pp. F4D–6.
26. Z. He, J. Song, C. Häger, A. G. i Amat, H. Wymeersch, P. A. Andrekson, M. Karlsson, and J. Schröder, "Experimental demonstration of learned pulse shaping filter for superchannels," in *Optical Fiber Communication Conference*, (Optica Publishing Group, 2022), pp. W2A–33.
27. V. Bajaj, F. Buchali, M. Chagnon, S. Wahls, and V. Aref, "Single-channel 1.61 tb/s optical coherent transmission enabled by neural network-based digital pre-distortion," in *2020 European Conference on Optical Communications (ECOC)*, (IEEE, 2020), pp. 1–4.
28. V. Bajaj, M. Chagnon, S. Wahls, and V. Aref, "Efficient training of volterra series-based pre-distortion filter using neural networks," in *2022 Optical Fiber Communications Conference and Exhibition (OFC)*, (IEEE, 2022), pp. 1–3.
29. V. Bajaj, F. Buchali, M. Chagnon, S. Wahls, and V. Aref, "54.5 tb/s wdm transmission over field deployed fiber enabled by neural network-based digital pre-distortion," in *Optical Fiber Communication Conference*, (Optica Publishing Group, 2021), pp. M5F–2.
30. V. Bajaj, F. Buchali, M. Chagnon, S. Wahls, and V. Aref, "Deep neural network-based digital pre-distortion for high baudrate optical coherent transmission," J. Lightwave Technol. **40**(3), 597–606 (2022).
31. T. Uhlemann, S. Cammerer, A. Span, S. Dörner, and S. ten Brink, "Deep-learning autoencoder for coherent and nonlinear optical communication," in *Photonic Networks; 21th ITG-Symposium*, (VDE, 2020), pp. 1–8.
32. A. Vannucci, P. Serena, and A. Bononi, "The RP method: a new tool for the iterative solution of the nonlinear schrodinger equation," J. Lightwave Technol. **20**(7), 1102–1112 (2002).
33. V. Aref and M. Chagnon, "End-to-end learning of joint geometric and probabilistic constellation shaping," in *2022 Optical Fiber Communications Conference and Exhibition (OFC)*, (IEEE, 2022), pp. 1–3.
34. F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for gaussian channels," IEEE Trans. Inf. Theory **39**(3), 913–929 (1993).
35. T. Fehenberger, A. Alvarado, G. Böcherer, and N. Hanik, "On probabilistic shaping of quadrature amplitude modulation for the nonlinear fiber channel," J. Lightwave Technol. **34**(21), 5063–5073 (2016).
36. V. Neskorniuk, A. Carnio, V. Bajaj, D. Marsella, S. K. Turitsyn, J. E. Prilepsky, and V. Aref, "End-to-end deep learning of long-haul coherent optical fiber communications via regular perturbation model," in *2021 European Conference on Optical Communication (ECOC)*, (IEEE, 2021), pp. 1–4.
37. V. Neskorniuk, A. Carnio, D. Marsella, S. K. Turitsyn, J. E. Prilepsky, and V. Aref, "Model-based deep learning of joint probabilistic and geometric shaping for optical communication," in *2022 Conference on Lasers and Electro-Optics (CLEO)*, (2022), pp. 1–2.
38. G. Böcherer, "Achievable rates for probabilistic shaping," arXiv arXiv:1707.01134 (2017).
39. Z. Tao, L. Dou, W. Yan, L. Li, T. Hoshida, and J. C. Rasmussen, "Multiplier-free intrachannel nonlinearity compensating algorithm operating at symbol rate," J. Lightwave Technol. **29**(17), 2570–2576 (2011).
40. G. P. Agrawal, "Nonlinear fiber optics," in *Nonlinear Science at the Dawn of the 21st Century*, (Springer, 2000), pp. 195–211.
41. C. Häger and H. D. Pfister, "Physics-based deep learning for fiber-optic communication systems," IEEE J. Select. Areas Commun. **39**(1), 280–294 (2021).
42. I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).
43. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), pp. 770–778.
44. S. Li, C. Häger, N. Garcia, and H. Wymeersch, "Achievable information rates for nonlinear fiber communication via end-to-end autoencoder learning," in *2018 European Conference on Optical Communication (ECOC)*, (IEEE, 2018), pp. 1–3.
45. 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," arXiv arXiv:1810.00774 (2018).
46. R. Dar, M. Feder, A. Mecozzi, and M. Shtaif, "Accumulation of nonlinear interference noise in fiber-optic systems," Opt. Express **22**(12), 14199–14211 (2014).
47. A. Carena, G. Bosco, V. Curri, Y. Jiang, P. Poggiolini, and F. Forghieri, "Egn model of non-linear fiber propagation," Opt. Express **22**(13), 16335–16362 (2014).
48. H. Yang, Z. Niu, H. Zhao, S. Xiao, W. Hu, and L. Yi, "Fast and accurate waveform modeling of long-haul multi-channel optical fiber transmission using a hybrid model-data driven scheme," J. Lightwave Technol. **40**(14), 4571–4580 (2022).

Optics EXPRESS

49. Y. Zang, Z. Yu, K. Xu, X. Lan, M. Chen, S. Yang, and H. Chen, "Principle-driven fiber transmission model based on pinn neural network," J. Lightwave Technol. **40**(2), 404–414 (2022).
50. A. Mecozzi and R.-J. Essiambre, "Nonlinear shannon limit in pseudolinear coherent systems," J. Lightwave Technol. **30**(12), 2011–2024 (2012).
51. F. J. García-Gómez and G. Kramer, "Mismatched models to lower bound the capacity of dual-polarization optical fiber channels," J. Lightwave Technol. **39**(11), 3390–3399 (2021).
52. C. Häger and H. D. Pfister, "Nonlinear interference mitigation via deep neural networks," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, (IEEE, 2018), pp. 1–3.
53. M. A. Bolshtyansky, O. V. Sinkin, M. Paskov, Y. Hu, M. Cantono, L. Jovanovski, A. N. Pilipetskii, G. Mohs, V. Kamalov, and V. Vusirikala, "Single-mode fiber sdm submarine systems," J. Lightwave Technol. **38**(6), 1296–1304 (2020).
54. N. Merhav, G. Kaplan, A. Lapidoth, and S. S. Shitz, "On information rates for mismatched decoders," IEEE Trans. Inf. Theory **40**(6), 1953–1967 (1994).
55. 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," J. Lightwave Technol. **34**(22), 5146–5156 (2016).
56. S. Cammerer, F. A. Aoudia, S. Dörner, M. Stark, J. Hoydis, and S. Ten Brink, "Trainable communication systems: Concepts and prototype," IEEE Trans. Commun. **68**(9), 5489–5503 (2020).
57. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv arXiv:1412.6980 (2014).
58. R.-J. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, and B. Goebel, "Capacity limits of optical fiber networks," J. Lightwave Technol. **28**(4), 662–701 (2010).
59. 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," J. Lightwave Technol. **35**(22), 4871–4879 (2017).
60. M. S. Faruk and S. J. Savory, "Digital signal processing for coherent transceivers employing multilevel formats," J. Lightwave Technol. **35**(5), 1125–1141 (2017).
61. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research **3**, 1157–1182 (2003).
62. S. Zhang, F. Yaman, K. Nakamura, T. Inoue, V. Kamalov, L. Jovanovski, V. Vusirikala, E. Mateo, Y. Inada, and T. Wang, "Field and lab experimental demonstration of nonlinear impairment compensation using neural networks," Nat. Commun. **10**(1), 8 (2019).