








Deep Learning-Based Phase Retrieval Scheme for Minimum-Phase Signal Recovery

Daniele Orsuti , Cristian Antonelli , *Senior Member, IEEE*, Alessandro Chiuso , *Fellow, IEEE*, Marco Santagiustina , *Member, IEEE*, Antonio Mecozzi , *Fellow, IEEE*, Andrea Galtarossa , *Fellow, IEEE*, and Luca Palmieri , *Senior Member, IEEE*

Abstract—We propose a deep learning-based phase retrieval method to accurately reconstruct the optical field of a single-sideband minimum-phase signal from the directly detected intensity waveform. Our method relies on a fully convolutional Neural Network (NN) model to realize non-iterative and robust phase retrieval. The NN is trained so that it performs full-field reconstruction and jointly compensates for transmission impairments. Compared to the recently proposed Kramers-Kronig (KK) receiver, our method avoids the distortions introduced by the nonlinear operations involved in the KK phase-retrieval algorithm and hence does not require digital upsampling. We validate the proposed phase-retrieval method by means of extensive numerical simulations in relevant system settings, and we compare the performance of the proposed scheme with the conventional KK receiver operated with a 4-fold digital upsampling. The results show that the 7% hard-decision forward error correction (HD-FEC) threshold at BER 3.8×10^{-3} can be achieved with up to 2.8 dB lower carrier-to-signal power ratio (CSPR) value and 1.8 dB better receiver sensitivity compared to the conventional 4-fold upsampled KK receiver. We also present a comparative analysis of the complexity of the proposed scheme with that of the KK receiver, showing that the proposed scheme can achieve the 7% HD-FEC threshold with 1.6 dB lower CSPR, 0.4 dB better receiver sensitivity, and 36% lower complexity.

Index Terms—Deep learning, direct-detection, Kramers Kronig receiver, phase retrieval.

Manuscript received 6 June 2022; revised 6 September 2022 and 27 October 2022; accepted 2 November 2022. Date of publication 7 November 2022; date of current version 15 January 2023. This work was supported in part by the Italian Ministry for Education, University and Research (MIUR), “Departments of Excellence” under Grant law 232/2016, in part by the Project Fiber Infrastructure for Research on Space-division multiplexed Transmission (FIRST) under Grant PRIN 2017, and in part by the University of Padova, Project MACFIBER under Grant BIRD 2019. (Corresponding author: Daniele Orsuti.)

Daniele Orsuti, Alessandro Chiuso, and Andrea Galtarossa are with the Department of Information Engineering, University of Padova, 35131 Padova, Italy (e-mail: daniele.orsuti@phd.unipd.it; alessandro.chiuso@unipd.it; andrea.galtarossa@dei.unipd.it).

Cristian Antonelli and Antonio Mecozzi are with the Department of Physical and Chemical Sciences, University of L’Aquila, 67100 L’Aquila, Italy, and also with the National Laboratory of Advanced Optical Fibers for Photonics, CNIT, 43124 Parma, Italy (e-mail: cristian.antonelli@univaq.it; antonio.mecozzi@univaq.it).

Marco Santagiustina and Luca Palmieri are with the Department of Information Engineering, University of Padova, 35131 Padova, Italy, and also with the National Laboratory of Advanced Optical Fibers for Photonics, CNIT, 43124 Parma, Italy (e-mail: marco.santagiustina@unipd.it; luca.palmieri@dei.unipd.it).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/JLT.2022.3219639>.

Digital Object Identifier 10.1109/JLT.2022.3219639

I. INTRODUCTION

THE ever-increasing traffic growth affecting short-range systems, such as intra- and inter-data center interconnects, requires that the deployed transceivers offer high spectral efficiency while keeping low complexity and low costs [1]. In such systems, Intensity-modulation and direct-detection (IM-DD) is the conventional scheme employed for its simple structure and cost-effectiveness. However, the phase information of light is lost upon DD, removing the ability of digitally compensating for propagation effects. To address these issues, single-side-band (SSB) transmission has gained significant interest. It retains the simplicity of DD while allowing advanced modulation formats and electronic dispersion compensation (EDC). This is enabled by an additional continuous-wave (CW) carrier, typically added at the transmitter to the edge of the optical signal spectrum. At the receiver, by measuring the beating product between the signal and the carrier, phase and amplitude information can be extracted employing a standard single-photodiode receiver. The process of DD is known to introduce an unwanted signal-to-signal beating interference (SSBI) term that represents the main limitation of this scheme. An effective and bandwidth-efficient method to deal with SSBI is the Kramers-Kronig (KK) receiver [2], which has shown superior performance compared to the other SSBI cancellation techniques [3]. It relies on the use of minimum-phase (MP) signals for which the phase information can be retrieved from the detected intensity through digital signal processing (DSP). This holds theoretically, for continuous signals. However, in an analog-to-digital converter (ADC) bandwidth-limited system, exact phase reconstruction may not be achieved even when the MP condition is met. Indeed, spectral broadening generated by the non-linear operations (such as square root and logarithm) entailed by the KK phase-retrieval algorithm needs to be accommodated to avoid aliasing. Consequently, the conventional KK scheme requires upsampling, by a relatively high factor, at the beginning of the DSP chain [4].

Other distortions arise for low CSPRs when the photocurrent samples may often approach zero and the corresponding logarithm produces large negative excursions [5]. These excursions occur over a short time, so have a spectrum that spans the entire sampled bandwidth, worsening the quality of the retrieved phase. Although a higher CSPR leads to effective SSBI cancellation, the carrier component introduces an additional sensitivity penalty as it increases the impact of both carrier-to-amplified-spontaneous-emission (ASE) noise beating [4] and nonlinear fiber

propagation effects [6]. This leads to significant performance degradation in multi-channel transmission settings [6], limiting the number of channels that can be multiplexed. Furthermore, a higher CSPR increases the transmitter requirements. For example, the required resolution of the digital-to-analog converter (DAC) increases with the CSPR in those systems where the carrier is digitally generated (virtual tone) together with the information-bearing signal [7], [8]. Therefore, the CSPR is a key parameter to be optimized for system performance, and the lowest possible CSPR is preferred. Several enhanced KK schemes have been proposed both to reduce the required upsampling factor and to enhance the performance at low CSPR values. Authors in [9], proposed a KK scheme that adopts mathematical approximations to avoid the use of nonlinear operations such as logarithm and exponential functions. However, some nonlinear operations remain in their scheme, meaning that upsampling (albeit of a reduced factor) is still required to avoid aliasing. Authors in [10], instead, proposed two methods for reducing error rates for weak carrier powers. The first is to insert strong clipping to limit the large negative excursions generated by the logarithm. The second is to replace the square-root function with a logarithmic function, to allow analog processing using semiconductor diodes [11]. Alternatively, in [12], the sampling rate and CSPR requirements are relaxed by combining digital upsampling with harmonic filtering. Two low-pass filters are inserted after each nonlinear operation in the KK algorithm to eliminate out-of-band harmonics before downsampling. The above-mentioned works suggest that moving away from a theoretically perfect implementation of the KK receiver may lead to improvements that are not evident from the mathematical analysis of the ideal case.

In this paper, we propose a novel approach, based on deep learning (DL), to reconstruct the phase of a MP optical SSB signal from the detected intensity. The proposed method relies on a fully convolutional neural network (CNN) model, is non-iterative, and does not require digital upsampling (since the KK algorithm nonlinear operations are avoided). We extend the preliminary investigations conducted in [13] by thoroughly analyzing the performance of the proposed method over different transmission scenarios. We first assess the performance in Back-to-Back (B2B) settings, and then we consider linear transmission, single-channel nonlinear transmission, and DWDM transmission over 100 km of standard single-mode fiber. We present a comparative analysis of the conventional KK receiver with two proposed receiver schemes that differ in the NN training procedure. The first receiver scheme is trained in ideal B2B settings so that it emulates the KK processing for full-field reconstruction; the second receiver scheme is trained in a 5-channel DWDM transmission scenario, and the NN retrieves the transmitted IQ components while jointly compensating for linear and nonlinear impairments. In 5-channel DWDM transmission, the proposed schemes comply with the 7% HD-FEC threshold, with a CSPR reduced by 1.6 dB, 0.4 dB better sensitivity, and 36% lower computational complexity than the conventional 4-fold up-sampled KK receiver aided with digital back-propagation (DBP). Alternatively, at the expense of 7.2 times higher complexity, the highest performance improvement is obtained: the 7% HD-FEC threshold is achieved with 2.8 dB

lower CSPR and 1.8 dB better receiver sensitivity. The results show that, after training, the CNN learned to extract and separate the features of the useful information signal from the features of SSBI and other undesired interference terms, providing new avenues to design MP retrieval schemes for DD systems.

II. PROPOSED METHOD

In this section, we first review the KK phase retrieval algorithm. Then, we detail the working principle of the proposed NN-based phase retrieval method.

A. KK Phase Retrieval Algorithm

We denote by $E_s(t)$ the complex envelope of the data-carrying signal whose spectrum is contained within an optical bandwidth B . The optical carrier is assumed to have an amplitude E_0 and to be located at the low-frequency edge of the data-carrying signal spectrum. The complex envelope of the field at the input of the photodiode can thus be written as

$$E(t) = E_0 + E_s(t) \exp(-j\pi Bt). \quad (1)$$

The photocurrent $i(t)$ produced by the photodiode is proportional to the optical intensity $|E(t)|^2$. When E_0 is large enough, so that the MP phase condition is satisfied, the following operations can be performed to retrieve the optical field [2]

$$E_s(t) = \left[\sqrt{i(t)} \cdot \exp\{j\varphi(t)\} - E_0 \right] \exp(j\pi Bt), \quad (2)$$

$$\varphi(t) = \mathcal{H} \left[\ln \sqrt{i(t)} \right]. \quad (3)$$

In (3), $\mathcal{H}[\cdot]$ indicates the Hilbert transform and $\ln(\cdot)$ the natural logarithm function. Since the entire phase retrieval process is performed in the digital domain, $i(t)$ needs to be sampled at least at the Nyquist frequency (i.e., $2B$). Then, to accommodate the spectral broadening caused by the square root and logarithm functions, digital upsampling is required. An upsampling factor of 4 has been shown to be sufficient to accommodate the bandwidth expansion [14].

B. NN-Based Phase Retrieval

In this paper, we propose a DL based model, to which hereinafter we refer simply as G , to recover the phase of SSB and MP signals. The model G is trained to perform a mapping from \mathbf{i} to \mathbf{y} , i.e., $\mathbf{y} = G(\mathbf{i})$, where the input $\mathbf{i} \in \mathbb{R}^n$ is the digitized photocurrent signal at the output of the ADC, and $\mathbf{y} = [\hat{\mathbf{I}}, \hat{\mathbf{Q}}] \in \mathbb{R}^{n \times 2}$ contains the approximate reconstruction of \mathbf{I} and \mathbf{Q} that are the ground truth in-phase and quadrature components generated at the transmitter side. The in-phase and quadrature components are used as targets rather than the amplitude and phase, so to circumvent phase wrapping problems and discontinuities present in the phase signal. We denote by $\{\mathbf{i}_i, \mathbf{y}_i\}_{i=1}^N$ the training set used to learn the parameters θ of the model G . The cost function $\mathcal{L}(\theta)$ we minimize during training reads

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N l_i(\theta), \quad (4)$$

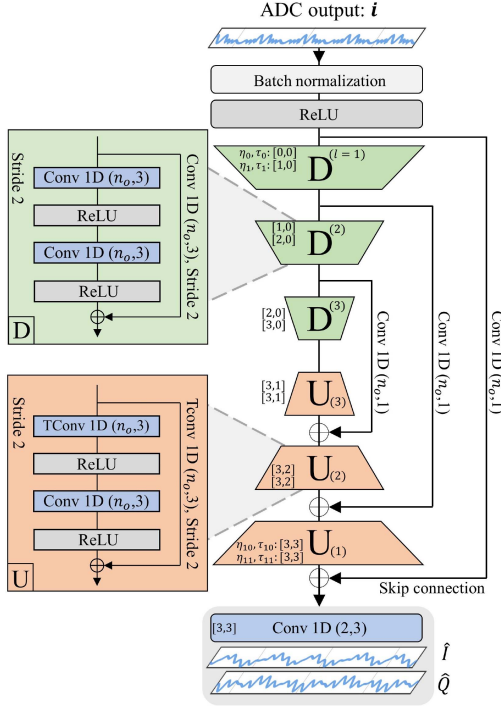


Fig. 1. Diagram of the proposed encoder-decoder temporal fully convolutional architecture. i : input photocurrent signal, $\hat{i}\hat{Q}$ predicted output, l : D/U block index, Conv 1D: 1D convolutional layer, TConv 1D: 1D transposed convolutional layer (also known as fractionally-strided convolutional layer). The figure labels the convolutional layers with stride different from 1. D: downsampling block; U: upsampling block. n_o : number of kernels for the convolutional layers in a D/U block. In the figure $[\cdot, \cdot]$ denotes the pair of values $[\eta_i, \tau_i]$ for the convolutional layers in the main path, whereas (\cdot, \cdot) denotes (number of kernels, kernel size).

where, $l_i(\theta)$ denotes the normalized root mean square error (NRMSE):

$$l_i(\theta) = \sqrt{\frac{\langle |\mathbf{I}_i - \hat{\mathbf{I}}_i|^2 \rangle + \langle |\mathbf{Q}_i - \hat{\mathbf{Q}}_i|^2 \rangle}{\langle |\mathbf{I}_i + j\mathbf{Q}_i|^2 \rangle}}. \quad (5)$$

In (5), $[\hat{\mathbf{I}}_i, \hat{\mathbf{Q}}_i] = G(\mathbf{i}_i)$. Once the training process is completed and the parameters θ are optimized, the NN can be used to perform full-field recovery directly from photocurrent signals (distinct from the training set).

C. Temporal Convolutional Neural Network Model

For the NN architecture we rely on temporal 1D-CNNs, a variant of CNNs that have shown superior performance across several sequence modeling tasks compared to baseline recurrent architectures [15]. Temporal 1D-CNNs are typically implemented using a hierarchy of 1D convolutional layers either in an encoder-decoder structure (using downsampling and upsampling layers) or employing dilated convolutions to effectively capture long-range temporal patterns [15], [16]. In these configurations, the output has the same length as the input, and the model is fully convolutional (without dense layers), hence the number of parameters is reduced, and variable input signal sizes are allowed. The proposed encoder-decoder temporal CNN is shown in Fig. 1. The photocurrent signal at the input of the network undergoes a contraction (upper side of the network)

and an expansion process, where the downsampling blocks (D) and upsampling blocks (U) are detailed on the left side of the figure. Downsampling (upsampling) is performed by repeated application of non-causal convolutional (transposed convolutional) layers with stride 2. Features from both future and past samples are incorporated to predict output samples by relaxing the causality constrain in the convolution operation, which is not sufficient to predict the IQ components from the photocurrent signal. In the architecture, residual learning is enabled by the introduction of skip connections. The latter allows feature maps extracted through the expansion path to be concatenated with features from the contraction path. This information flow between layers has been shown to accelerate the convergence of the training phase and to provide better IQ reconstruction performance [17]. While typical residual blocks (i.e., D and U blocks) contain batch normalization layers [18], we observed performance degradation when such layers were included in our model, and, therefore, batch normalization is applied only after the input layer. Each convolutional layer in the NN model is followed by a ReLU activation function, except for the convolutional layers in the skip connections inside the D and U blocks. In Fig. 1 some ReLU layers are omitted for ease of readability.

We now define some quantities that describe the hyperparameters of the D and U blocks shown on the left side of Fig. 1. The quantities of interest for the D and U blocks are the following.

- k : kernel size of the 1D convolutional layers inside a D/U block.
- s : stride of the 1D convolutional layers inside a D/U block.
- d : model depth, i.e., the number of D and U blocks.
- l : D/U block index ($l = 1, 2, \dots, d$). Fig. 1 shows the values that l assumes across the D and U blocks.
- n_o : number of kernels of the 1D convolutional layers inside a D/U block. n_o can be equivalently defined as the number of channels produced by the 1D convolutional layers inside a D/U block. n_o is set constant throughout the D and U blocks of the NN model, as shown in Fig 1.

In this work, we set $k = 3$; when stride is applied, it is set to $s = 2$ (see Fig. 1); and we investigate the NN model performance over different n_o and d values. For the long skip connections outside the D and U blocks, the convolutional layers have kernel size 1 and n_o kernels. For the output layer, the kernel size is k and the number of kernels is 2 (i.e., $y = [\hat{\mathbf{I}}, \hat{\mathbf{Q}}]$).

A crucial parameter of the considered NN model is the memory size. The memory size controls the number of neighboring samples of the input photocurrent signal that is used to predict a single output sample of the IQ components. Properly selecting the memory size allows, first, to recover the IQ components from the photocurrent signal and, second, to deal with the memory effects of the transmission channel. Conventional NN models adapt the number of neurons of the input layer (for fully-connected NNs [19]) or the filter width of the convolutional layers (for CNNs [20]) to account for a sufficient number of neighboring input samples. Differently, temporal encoder-decoder CNNs control the memory size by adapting both the filter width of the convolutional layers and the number of strided convolutional layers stacked in the NN model [15].

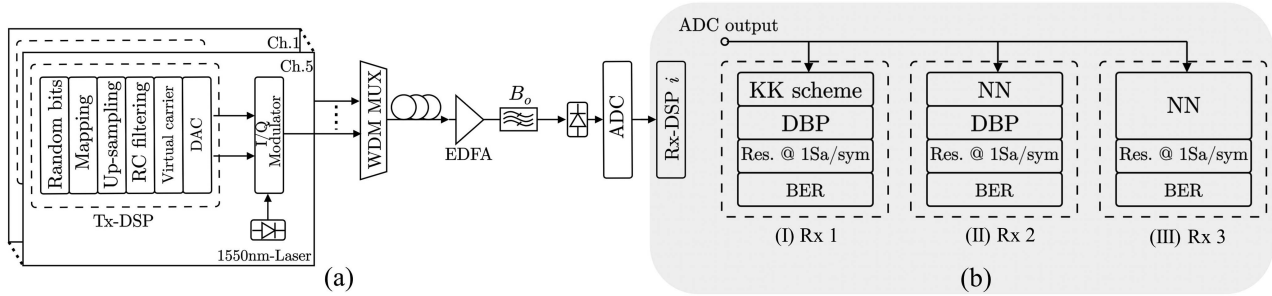


Fig. 2. (a) General simulation setup. The optical filter with bandwidth B_o , selects the central channel. (b) Considered receiver schemes. Insets: Rx 1 (I), Rx 2 (II), Rx 3 (III).

Indeed, both downsampling and upsampling layers modify the way in which the cumulative memory size grows for each new added layer. Specifically, the memory size (in samples) at any convolutional layer i of the model shown in Fig. 1 can be computed with the recursive equation [21]

$$M_i = M_{i-1} + 2^{(\eta_{i-1} - \tau_{i-1})} (k - 1), \quad (6)$$

where, $M_0 = 1$, $\eta_0 = \tau_0 = 0$, and η_i and τ_i are the number of strided convolutions and transposed convolutions used in the NN up to the convolutional layer i^1 , respectively. The factor $2^{(\cdot)}$ in (6) arises because the applied stride is constant and equal to 2 for each strided convolution. A more general expression for the NN model memory is derived in some detail in [22] under the equivalent name of receptive field of the CNN model. Solving (6) up to the output layer of the NN model shown in Fig. 1, which has a model depth $d = 3$, it gives a memory size of 36 symbols. If the model depth is reduced to $d = 2$ or increased to $d = 4$, (6) gives a memory size of 16 and 76 symbols, respectively. Therefore, by varying the model depth, the memory size can be easily tuned to adapt the NN model to the specific transmission system parameters, such as baud rate, transmission distance, and fiber link parameters. Section IV-E, discusses the impact of the model depth and of the number of kernels on performance. Section V, compares the complexity of the proposed scheme with that of the conventional KK receiver.

III. SIMULATION SETUP

In this section, we describe the deployed simulation model. We first provide a general description of the simulated blocks, then we detail the receiver schemes considered in this work.

A. General Simulation Setup

Fig. 2(a) shows the simulation setup and Table I summarizes the simulations parameters. We assume 5 DWDM channels spaced by 40 GHz and evaluate the performance on the central channel. For each transmitter section, random bit sequences are mapped into 16-quadrature-amplitude-modulation (QAM) symbols at a symbol rate of 24 GBaud, which are then upsampled and shaped with a raised-cosine (RC) filter with a roll-off factor of 0.05. Next, a virtual carrier is added, placed exactly at the

¹For transposed convolutional layers τ_i includes the layer i itself (see Fig. 1).

TABLE I
GENERAL SIMULATION PARAMETERS

Parameter	Value
Baud rate	24 GBaud
Modulation format	16-QAM
Number of channels	5
Channel spacing	40 GHz
Center wavelength	1550 nm
Laser linewidth	1 MHz
RIN level	-139 dBc/Hz
Attenuation coefficient	0.2 dB/km
Nonlinear coefficient	1.3 /W/km
CD coefficient	17 ps/nm/km
Span length	100 km
Number of spans	1
Noise figure of EDFA	5 dB
Optical filter bandwidth	36 GHz
Photodiode bandwidth	29 GHz
Photodiode responsivity	1 A/W
ADC sampling rate	$2B$
ADC vertical resolution	8 bits

left edge of the information-bearing signal spectrum, and the resulting signal is sent to an ideal DAC (i.e., without quantization or bandwidth limitation), where electrical to optical (E/O) conversion is performed by an IQ modulator biased at the null-point. The laser source of the central channel is centered at 1550 nm. The employed light sources have 1 MHz linewidth. Relative intensity noise (RIN) of the laser sources is included in the simulations and modelled as white Gaussian noise; its level is set to -139 dBc/Hz, which is a typical value for low-cost distributed feedback laser diodes [23]. After E/O conversion, the output of each transmitter section is multiplexed using a WDM MUX. The channel spectral response of the WDM MUX is assumed to be the same as that of the optical filter used at the receiver. The DWDM signal is launched into a 100 km long G.652 single-mode fiber having an attenuation coefficient of 0.2 dB/km, a chromatic dispersion (CD) coefficient of 17 ps/nm/km, and nonlinear parameter $1.3 \text{ W}^{-1}\text{km}^{-1}$. The waveform evolution inside the fiber is computed using the symmetric split-step Fourier method [24].

At the receiver, the optical signal is amplified by an erbium-doped fiber amplifier (EDFA) with a 5 dB noise figure operating in the transparency condition. A 12 th-order super-Gaussian optical filter with a 3 dB bandwidth of 36 GHz is applied to

select the central channel [2]. Then, a PIN photodiode with responsivity 1 A/W and 29 GHz bandwidth detects the filtered signal; shot noise and thermal noise due to the photodetection are included in the model. The ADC sampling frequency and vertical resolution are fixed at $2B$ and 8 bits, respectively. The ADC output is then fed to the different receiver schemes shown in Fig. 2(b). For each receiver scheme, after full-field recovery and transmission impairments compensation, downsampling to one sample per symbol is applied and the BER evaluated under the assumption of Gray coding.

B. Receiver Schemes

We describe now the DSP of the receiver schemes depicted in Fig. 2(b) (insets I)–(III)). Hereinafter, we refer to these schemes as Rx 1, Rx 2 and Rx 3. For Rx 2 and Rx 3, the training data parameters are sampled from a 2D grid formed by launch power values and CSPR values. For both receiver schemes, the CSPR values are sampled from the range 0–11 dB in steps of 0.2 dB. As detailed below, for Rx 2 the training set span length is 0 km (i.e., B2B settings) and the training set launch power is fixed to 0 dBm. For Rx 3 the training set span length is 100 km and the training set launch power values are 1, 2, 3, and 4 dBm.

1) *Rx 1*: The first receiver configuration is shown in Fig. 2(b) inset (I). It consists in the conventional KK receiver, presented in [2], and it is used as a baseline for our simulations. The digital upsampling factor used at the beginning of the DSP chain is set to $R = 4$. After full-field reconstruction, downsampling is applied to return to the ADC's sampling rate. Both upsampling and downsampling are performed by zero-padding in the frequency domain, and the Hilbert transform in the KK algorithm uses frequency-domain processing. Linear and nonlinear impairments are jointly compensated for using DBP of only the channel of interest. DBP is based on the split-step Fourier method and is carried out with a total of 10 steps (using a logarithmic distribution of step sizes [25]).

2) *Rx 2*: The building blocks of the second receiver scheme are shown in Fig. 2(b) inset (II). The training data for the NN are generated in ideal Back-to-Back (B2B) settings so that the NN processing emulates the KK receiver processing for full-field reconstruction. Once the model has been trained, the DSP chain consists in (1) using the trained model to retrieve the IQ components from the ADC output; (2) adding a CW tone, whose amplitude is estimated from the photocurrent, to the IQ components at the output of the NN; (3) compensating for transmission impairments using the same DBP algorithm as in Rx 1. The training set generation procedure for Rx 2 is now presented. In the simulation set-up of Fig. 2(a) the fiber and the EDFA are removed, all noise sources and the DBP compensation algorithm are switched-off, and only the central channel is considered. Next, N pairs of (digitized photocurrent, IQ components) signals with length 512 symbols are generated from random bits by scanning the CSPR value in the range 0–11 dB in steps of 0.2 dB. The ground truth IQ components are collected after the RC filtering block and are resampled at the same sampling rate as the ADC's output. Both the digitized photocurrent and the IQ components signals are normalized to fit the amplitude range $[0,1]$. The block

length for the training set is fixed to 512 symbols, whereas the test set block length can be arbitrary since the considered NN operates in a sliding window manner. We set $N = 33,600$, hence the CSPR range is scanned 600 times to generate the training set: $N = 600 \cdot |\text{CSPR}_{\text{range}}|$, where $|\text{CSPR}_{\text{range}}|$ denotes the cardinality of the set formed by the CSPR values in training set. The training set launch power for Rx 2 is fixed to 0 dBm since the training results are independent of the launch power (non-linearity is neglected in B2B settings). For higher values of N , no significant improvement in performance was observed. Other CSPR ranges for the training set have been tested, for example 3–11 dB, 3–13 dB, but similar performance were obtained.

3) *Rx 3*: The third considered receiver configuration is shown in Fig. 2(b) inset (III). Differently from Rx 2, the training data are generated in such a way that the NN jointly reconstructs the IQ components and compensates for linear and nonlinear transmission impairments. The training data are generated using the overall simulation setup of Fig. 2(a) with the general simulation parameters shown in Table I. For the central channel, we collect N pairs of (digitized photocurrent, IQ components) signals with a block length of 512 symbols. For each of the collected signals pair of the central channel, the bit sequences of the adjacent WDM channels are randomly selected to vary the inter-channel interference noise. As for Rx 2, we set $N = 33,600$, hence the CSPR range is scanned 150 times for each of the training set launch powers (that are 1, 2, 3, and 4 dBm).

IV. RESULTS AND DISCUSSION

In this section, we discuss and compare the performance of the receiver schemes presented so far. In Section IV-A, a numerical proof-of-concept of the proposed DL-based phase retrieval scheme is given, and its sensitivity to CD investigated. Then, for each transmission scenario, namely linear transmission, single-channel nonlinear transmission, and DWDM transmission, the performance evaluation outline is detailed (Section IV-B). Next, the simulation results for Rx 1, Rx 2, and Rx 3 are presented and discussed (Section IV-C), and a performance comparison between the receiver sensitivities is given (Section IV-D). Finally, the influence of the NN model hyperparameters on Rx 2 and Rx 3 performance is investigated (Section IV-E). Except for Section IV-E, the NN model hyperparameters considered for the investigations carried out in this section are $d = 3$ for the model depth (that corresponds to a memory size of 36 symbols) and $n_o = 32$ for the number of kernels.

For Rx 2 and Rx 3, Adam based optimization with a learning rate of 10^{-3} is used to tune the NNs parameters. Each network is trained for 200 epochs using 256 as batch size. Out of the N 512-symbol-long signals in the training set, 80% were used for training and 20% for validation test. The trainings take ~ 1 hour using Tensorflow on an Nvidia Quadro P2000 GPU. In the performance assessment, several sequences of 2^{15} symbols are transmitted for the extraction of a single BER value (as explained in Section IV-B). We use two independent random number generators for the training and testing phase to verify that the NN has not learned underlying features of the random number generator [26].

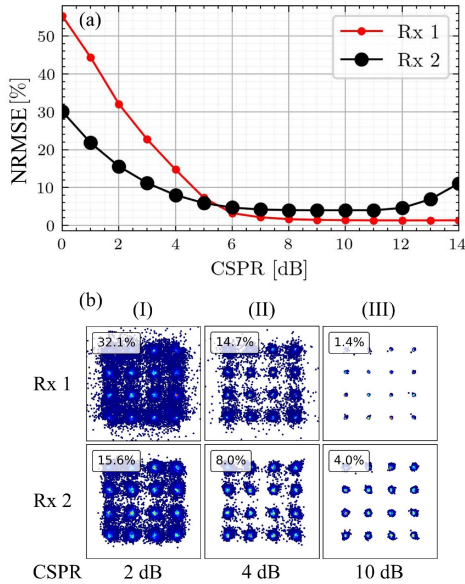


Fig. 3. (a) NRMSE versus CSPPR for Rx 1 and Rx 2 in B2B settings. (b) Reconstructed constellation diagrams at CSPPR 2 dB (I), 4 dB (II), and 10 dB (III). The NN model hyperparameters are $d = 3$ and $n_o = 32$.

A. Proof-of-Concept of NN-Based Phase Retrieval

We start by providing a numerical validation of the proposed NN-based full-field reconstruction scheme. To this end, we evaluate the performance in simplified simulation settings. We first investigate the performance in B2B settings, then we investigate the CD sensitivity of the proposed scheme. In both investigations, we evaluate the full-field reconstruction quality in terms of NRMSE because it has been selected as the loss function for the training phase of the NN model (see (5)). In this section, only Rx 1 and Rx 2 (that emulates KK processing) are considered.

1) *Performance in Ideal Back-to-Back Settings:* Initially, the reconstruction performance of Rx 1 and Rx 2 are evaluated in ideal B2B settings, with all noise sources switched-off. Fig. 3(a) shows the reconstruction capabilities of the two schemes in terms of NRMSE. The results show that, for CSPPR values up to 5 dB, Rx 2 achieves lower NRMSE values than Rx 1. For CSPPR values between 6–11 dB, Rx 2 reaches an NRMSE floor of 4%, whereas Rx 1 is able to drop the NRMSE to $\sim 1\%$. As the CSPPR is further increased above 11 dB, Rx 2 performance starts to degrade and the NRMSE increases from 4% to 11%. For CSPPR values higher than 11 dB the observed performance degradation can be explained by the selected training set CSPPR range for Rx 2 (i.e., 0–11 dB). The insets of Fig. 3(b) provide additional details on the constellations reconstructed by the two receivers at CSPPR values of 2 dB (Inset (I)), 4 dB (Inset (II)), and 10 dB (Inset (III)). Interestingly, for Rx 2, at CSPPRs 2 and 4 dB, the outer constellation corners, that are more likely to violate the MP condition when the carrier is added, experience less amplitude/phase errors than for Rx 1. For a CSPPR value of 10 dB, it can be observed that the constellation points reconstructed by Rx 2 scatter around the true points, whereas for Rx 1 the reconstructed constellation is almost perfect. Although Rx 2 saturates

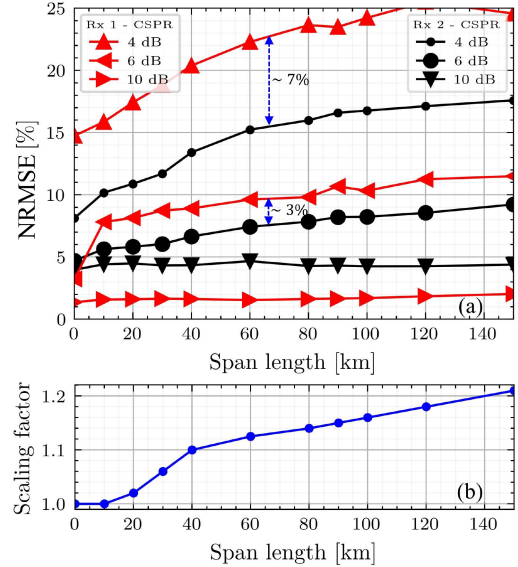


Fig. 4. (a) CD sensitivity as a function of span length for Rx 1 and Rx 2. (b) Scaling factor that minimizes the NRMSE for Rx 2 as a function of span length. The NN model hyperparameters are $d = 3$ and $n_o = 32$.

at a higher NRMSE than Rx 1, the improved performance that Rx 2 offers at low CSPPR will show significant advantages under relevant transmission system settings (see Section IV-C) rather than in the ideal B2B settings considered in this section.

2) *Chromatic Dispersion Sensitivity:* We now investigate the effect of CD on Rx 1 and Rx 2. To this end, we consider a linear transmission scenario with different span lengths (from 0 to 150 km). All noise sources are switched-off, the fiber CD coefficient is set to 17 ps/nm/km, and the nonlinear coefficient is set to 0. In both receiver schemes, the DBP block reduces to a dispersion compensation block based on frequency-domain equalization. The simulation results are presented in Fig. 4(a), where the NRMSE as a function of the span length is plotted for CSPPR values 4, 6 and, 10 dB. The results for Rx 1 (red curves) show that, for CSPPR 4 and 6 dB, the NRMSE increases with the transmission distance prior to saturating, while for CSPPR 10 dB it is constant. This is because CD leads to a higher peak-to-average power ratio (computed without carrier) as the transmission distance increases, and consequently the CSPPR required for the MP condition to be satisfied increases [27]. Considering now the results for Rx 2 (black curves), we recall that the NN has been trained to emulate the KK processing, i.e., on ideal B2B settings. Therefore, in presence of CD, the NN deals with photocurrent signals that were not considered during the training phase. This leads to rapid performance degradation for increasing span length. In order to allow the NN to extrapolate for different span lengths, it is required either to re-train the NN for each span length, or to cleverly include different span lengths in the training set [28], [29]. For the following CD sensitivity analysis of Rx 2, instead of re-training the NN model for each span length, we consider the NN model trained in B2B and analyze the impact of an increasing span length on the reconstructed constellation diagram. Our simulations show that two phenomena occur when testing the NN model in the

presence of CD. First, the NN retrieves the phase waveform up to a constant phase offset, which is also a known feature of the KK receiver [30]. Second, the NN responds to the new class of waveforms by introducing an additional distortion: a constant amplitude scaling on the reconstructed constellation. This amplitude scaling is independent of the CSPR but dependent on the transmission distance (i.e., on the amount of introduced total CD). Since the introduced distortions are systematic, once the span length has been fixed, the phase offset can be compensated for and the scaling factor to correct the constellation amplitude can be obtained by searching for the scaling factor that minimizes the NRMSE. This procedure has been implemented for each of the span lengths considered in our simulation, and the resulting scaling factors are shown in Fig. 4(b). Up to 10 km the NN reconstruction is slightly affected by CD, from 10 km up to 40 km the scaling factor rises rapidly, then the increment slows down from 40 km to 150 km. The NRMSE curves for Rx 2 in Fig. 4(a) have been obtained correcting the constellations with the scaling factors in Fig. 4(b). As for Rx 1, the NRMSE increases with transmission distance for the curves at CSPR 4 and 6 dB, while it is constant and saturates at 4%, as in the B2B case, for CSPR 10 dB. Compared to Rx 1, the NRMSE at CSPR 4 and 6 dB is reduced by 7% and 3%, respectively.

The systematic correction procedure described above for Rx 2 cannot be applied to adapt Rx 3 to different span lengths. Indeed, Rx 3 is trained to include the transmission impairment compensation task for a span length of 100 km. Therefore, the distortions introduced into the constellation diagram when testing Rx 3 for a span length that deviates from that of the training data turn out to be unpredictable.

B. Performance Evaluation Outline and Test Sets Structure

The methods and the test sets (one for each transmission scenario) used to evaluate the performance of the considered receiver schemes are now described. We consider three test sets: Test set 1 (Linear transmission), Test set 2 (Single-channel nonlinear transmission), and Test set 3 (DWDM nonlinear transmission). Hereinafter we refer to each transmission scenario using the corresponding test set. The test set data are generated using the simulation setup shown in Fig. 2, with the proper adjustments depending on the transmission regime (i.e., nonlinearity off, adjacent channels off, etc...). The span length is fixed to 100 km. For each test set, the parameters are sampled from a 2D grid formed by the CSPR values in the range 3-13 dB and the launch power values from -10 to 10 dBm. For each point of the 2D grid, we transmit 40 sequences each of 2^{15} symbols to evaluate the BER performance of the central channel. For numerical convenience, we consider multiple 2^{15} -symbol long sequences rather than a single long sequence. The test sets contain some CSPR values and launch power values that were not included in the training set so to evaluate the extrapolation capabilities of the NN model.

C. Transmission Performance

We present now the performance of the considered receiver schemes over Test sets 1, 2, and 3. Fig. 5(a)–(i) shows the results

of the simulations in terms of BER versus equivalent OSNR for the different CSPR values indicated in the legend. The equivalent OSNR facilitates the comparison with the case of coherent detection since it is defined as the OSNR that would be measured if the CW tone were not transmitted [2]. In the simulation, the equivalent OSNR was varied by varying the signal launch power. Fig. 5 is organized such that columns 1 through 3 of the grid correspond to Rx 1, 2, and 3, whereas rows 1, 2, and 3 correspond to Test set 1, 2, and 3, respectively. For each sub-figure, the dashed curve shows the plot of the analytic expression for the BER of a 16-QAM modulated system impaired by additive white Gaussian noise (AWGN) [2], whereas the horizontal solid black curve shows the 7% HD-FEC threshold limit. The curves are displayed up to a BER value of 10^{-5} for which the number of transmitted sequences guarantees accurate average BER values.

1) *Rx 1 Performance:* We consider first the performance of Rx 1 over Test set 1, 2, and 3.

Fig. 5(a) shows the performance over Test set 1. As expected, the BER decreases for increasing OSNR up to a BER floor. The achieved BER floor decreases at higher CSPR as the full-filed reconstruction quality increases. In the figure, for CSPR values in the range 10 – 12 dB, the BER values achieved by Rx 1 are very close to the ones obtained by an ideal coherent receiver. Further increasing the CSPR above 12 dB would lead to performance degradation due to the increased impact of the carrier-RIN beating term. The above-described trend of the BER curves agrees with that seen in [2], [23], yet for the different simulation parameters.

The results for Test set 2 are shown in Fig. 5(b). We recall that Rx 1 compensates for transmission impairments using DBP (Section III-B1). Compared to linear transmission, we expect the performance to degrade for both higher equivalent OSNR (since it is proportional to the information bearing signal power for a fixed noise power) and CSPR values. Indeed, when considering KK transceivers in presence of fiber nonlinearity, increasing the carrier strength has a twofold effect. On the one hand, increasing the CSPR leads to effective SSBI cancellation, on the other hand it increases the distortion due to fiber nonlinearity. For this reasons, the curves of Fig. 5(b) show an optimum operation point, i.e., at which minimum BER is achieved. The minimum BER improves by increasing the CSPR from 5 dB to 10 dB, and then deteriorates for higher CSPR values.

We finally assess the performance over Test set 3. The simulation results are shown in Fig. 5(c). Since only the channel of interest was back-propagated, we expect additional uncompensated inter-channel nonlinear effects to have an adverse impact on the BER, as also seen in [25], [31]. Indeed, Fig. 5(c) show that a higher minimum BER value is achieved, compared to single-channel operation. As for Test set 2, the minimum BER is achieved at a CSPR of 10 dB.

2) *Rx 2 Performance:* We present now the performance of Rx 2 over Test sets 1, 2, and 3. The results of the simulations have been obtained correcting the output of the NN with the constant amplitude scaling factor corresponding to a span length of 100 km (see Fig. 4).

Fig. 5(d) shows the results for Test set 1. It can be observed that, for CSPR values from 5 to 9 dB, Rx 2 outperforms Rx 1

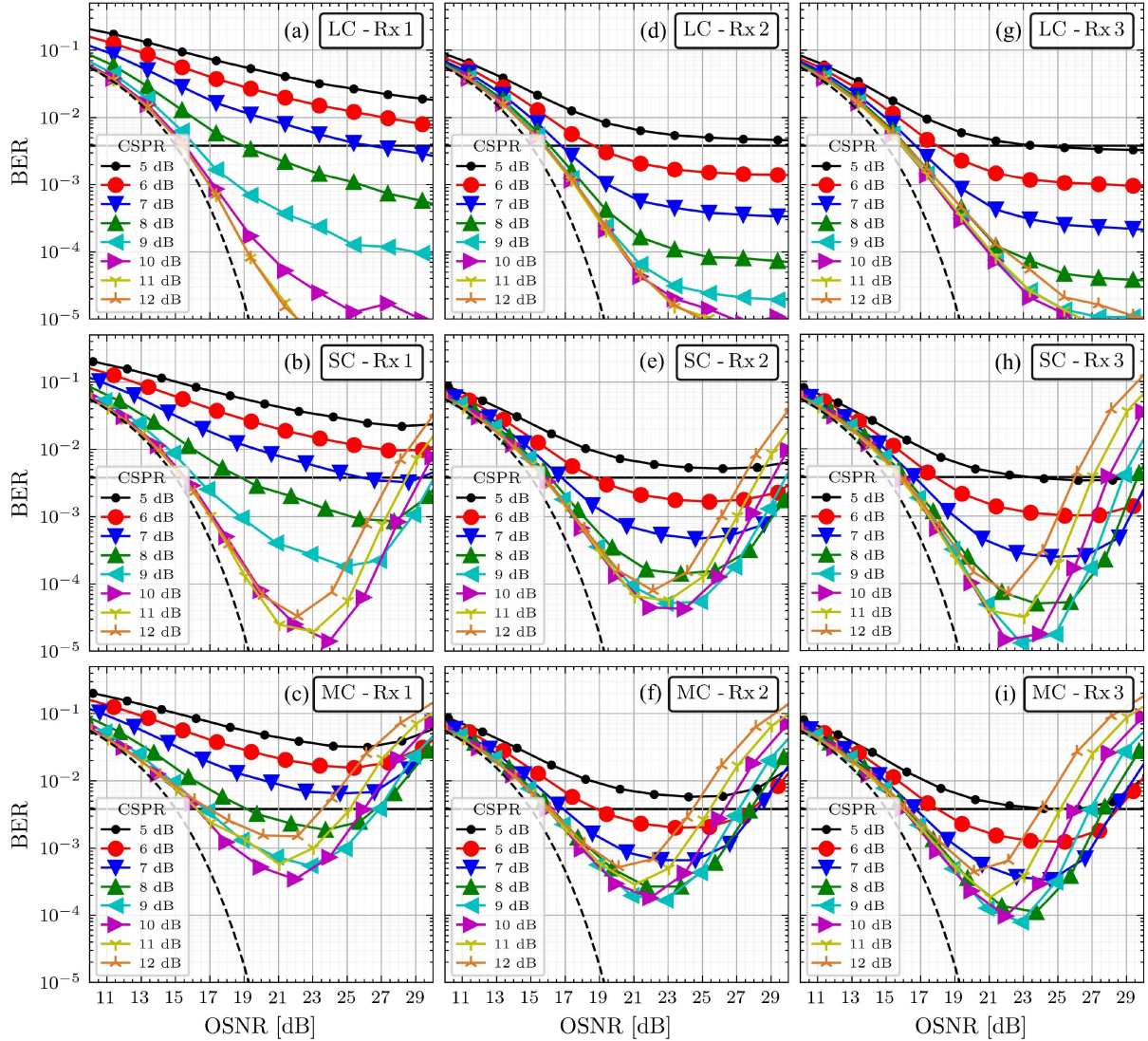


Fig. 5. BER versus equivalent OSNR (i.e., evaluated without carrier) at different CSRR values. The considered receiver scheme varies across columns, whereas the considered transmission regime varies across rows. LC: linear channel transmission; SC: single-channel nonlinear transmission (central channel performance); MC: multi-channel DWDM nonlinear transmission (central channel performance). For each sub-figure, the black dashed curve shows the plot of analytic expression for the BER of a 16-QAM modulated system impaired by AWGN, whereas the horizontal solid black line shows the 7% HD-FEC threshold, i.e., BER at 3.8×10^{-3} . The NN model hyperparameters are $d = 3$ and $n_o = 32$.

over the entire OSNR range considered. For these CSRR values, the BER floor value due to reconstruction error is lower than for Rx 1 and is achieved at lower OSNR values. When the CSRR is increased above 10 dB, the performance of Rx 2 degrades at a faster rate than for Rx 1, and Rx 1 performs better for all OSNRs starting from a CSRR of 11 dB. For CSRRs higher than 10 dB, besides increased carrier-RIN contribution, additional performance degradation stems from the fact that photocurrent signals associated with CSRR values outside the training are provided as input to the NN (as also seen in Section IV-A1). At low CSRR values, the remarkable performance improvement offered by Rx 2 when addressing the full-field reconstruction through the NN can be explained as follows. First, nonlinear operations in the KK algorithm, which are a source of distortions at low CSRRs, are avoided. Second, if the problem is viewed from the standpoint of SSBI cancellation, the NN learned to

extract and separate the features of the useful information signal from the features of SSBI and other undesired interference terms, making it robust to impairments.

The results for Test set 2 are shown in Fig. 5(e). We recall that Rx 2, after full-field reconstruction, compensates for transmission impairments with the same DBP algorithm of Rx 1. Simulation results show a reduced tolerance to nonlinear effects compared to Rx 1. This is related to what has been discussed in Section IV-A2, when the sensitivity to CD of the receiver was investigated. Indeed, this reduced tolerance results from further distortions introduced by the NN full-field reconstruction when dealing with the new class of photocurrent signals (i.e., not seen during training) affected by nonlinear effects. The BER versus OSNR curves of Fig. 5(e) show that Rx 2 (as for Test set 1) outperforms Rx 1 for CSRR values in the range 5 to 9 dB, almost over the entire OSNR range considered (except for the curve

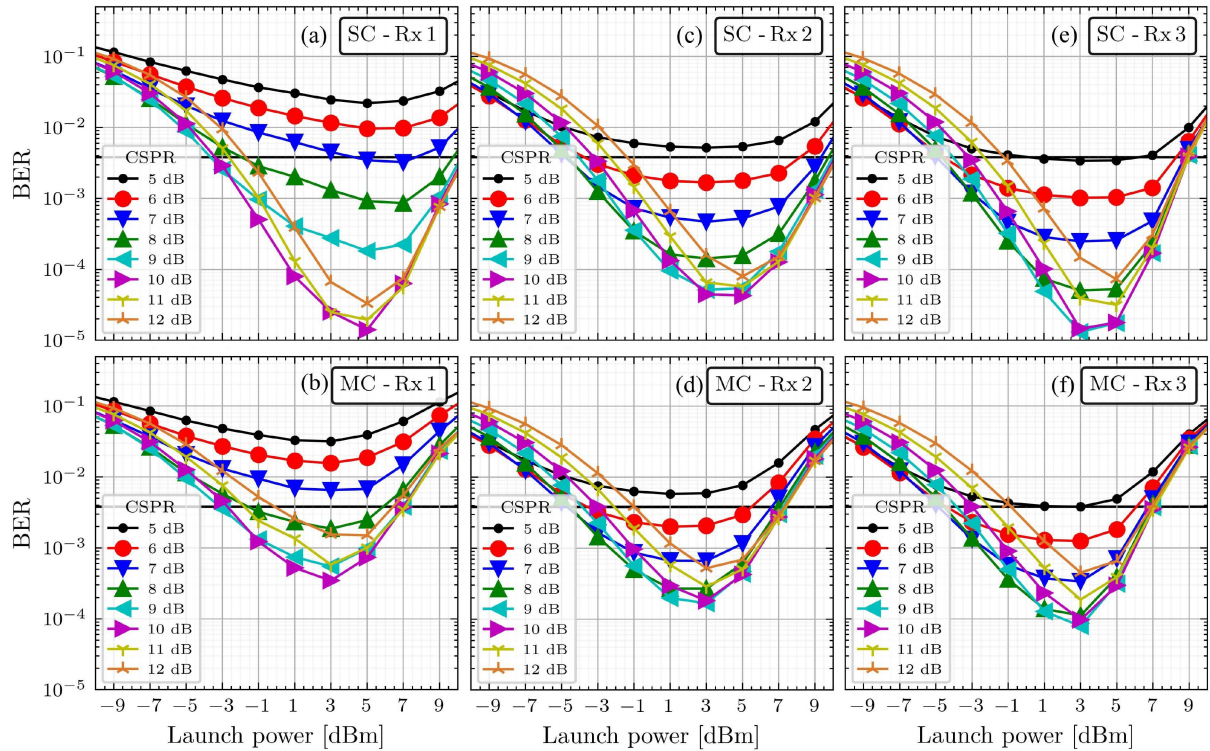


Fig. 6. BER versus launch power (i.e., carrier plus signal powers) per channel at different CSRP values. The considered receiver scheme varies across columns, whereas the considered transmission regime varies across rows. SC: single-channel nonlinear transmission; MC: multi-channel DWDM nonlinear transmission (central channel performance). For each sub-figure, the horizontal solid black line indicates the 7% HD-FEC threshold, i.e., BER at 3.8×10^{-3} . The NN model hyperparameters are $d = 3$ and $n_o = 32$.

at CSRP of 9 dB, where for OSNR higher than 27 dB, Rx 1 performs better). For CSRP above 9 dB, performance degradation due to nonlinear distortions prevails and the minimum BER value achieved is higher than for Rx 1.

Simulation results for Test set 3 are shown in Fig. 5(f). The results show that Rx 2 both improves the performance at low CSRP values, and achieves a lower minimum BER value than Rx 1. Compared to single-channel transmission, the minimum BER value achieved by Rx 2 in DWDM transmission is not limited by distortions introduced by the NN reconstruction, but by the reduced SNR at the receiver due to inter-channel nonlinear effects.

3) *Rx 3 Performance:* The performance of Rx 3 over the considered transmission scenarios are now presented. We recall that the NN in Rx 3 is trained to jointly recover the IQ components and to compensate for transmission impairments.

Fig. 5(g) shows the results over Test set 1. For CSRP values from 5 to 9 dB, Rx 3 further reduces the BER values compared to Rx 2. However, for CSRP higher than 9 dB, both Rx 1 and Rx 2 offer better performance. The rapid performance degradation of Rx 3 at high CSPRs can be explained by the pronounced inter-channel nonlinear distortions affecting the training set data. These nonlinear distortions cause BER saturation when testing Rx 3 in linear transmission settings.

Simulation results for Test set 2 in Fig. 5(h) show that the minimum BER value achieved by Rx 3 is as low as for Rx 1, meaning that Rx 3 effectively compensates for nonlinear transmission impairments. This holds true as long as the highly

nonlinear region, i.e., at high CSRP and OSNR values, is not considered. In this region the BER deteriorates at a faster rate than both for Rx 1 and Rx 2, which use DBP to compensate for transmission impairments. This can be explained by the launch power values selected for the training phase, which control the NN tolerance to nonlinear distortions. The choice for the selected launch power values will be motivated later on when describing Fig. 6, which plots the BER performance versus launch power. As a rule-of-thumb, the higher the launch power included in the dataset, the slower will be the BER degradation at high CSRP and OSNR values.

Fig. 5(i) shows the simulation results for Test set 3. Rx 3 achieves the lowest minimum BER among the considered receiver schemes while offering improved performance at low CSRP values.

We now investigate the transmission performance in terms of BER versus launch power per channel. Fig. 6(a)–(f) show the results of the simulations. For this investigation nonlinear transmission regime is assumed, i.e., only Test set 2 and Test set 3 are considered. The performance improvement achieved by Rx 2 and Rx 3 compared to Rx 1 agree with the results seen in Fig. 5 and described above, however, the following additional comments can be made. The curves of Fig. 6(a)–(f) show that, as expected, the BER decreases with launch power until it reaches a minimum value, and then, it increases again due to fiber nonlinearities. It is worth noticing that, the curves are plotted as a function of the total launch power (i.e., signal and carrier powers). Therefore, increasing the CSRP leads to reduced

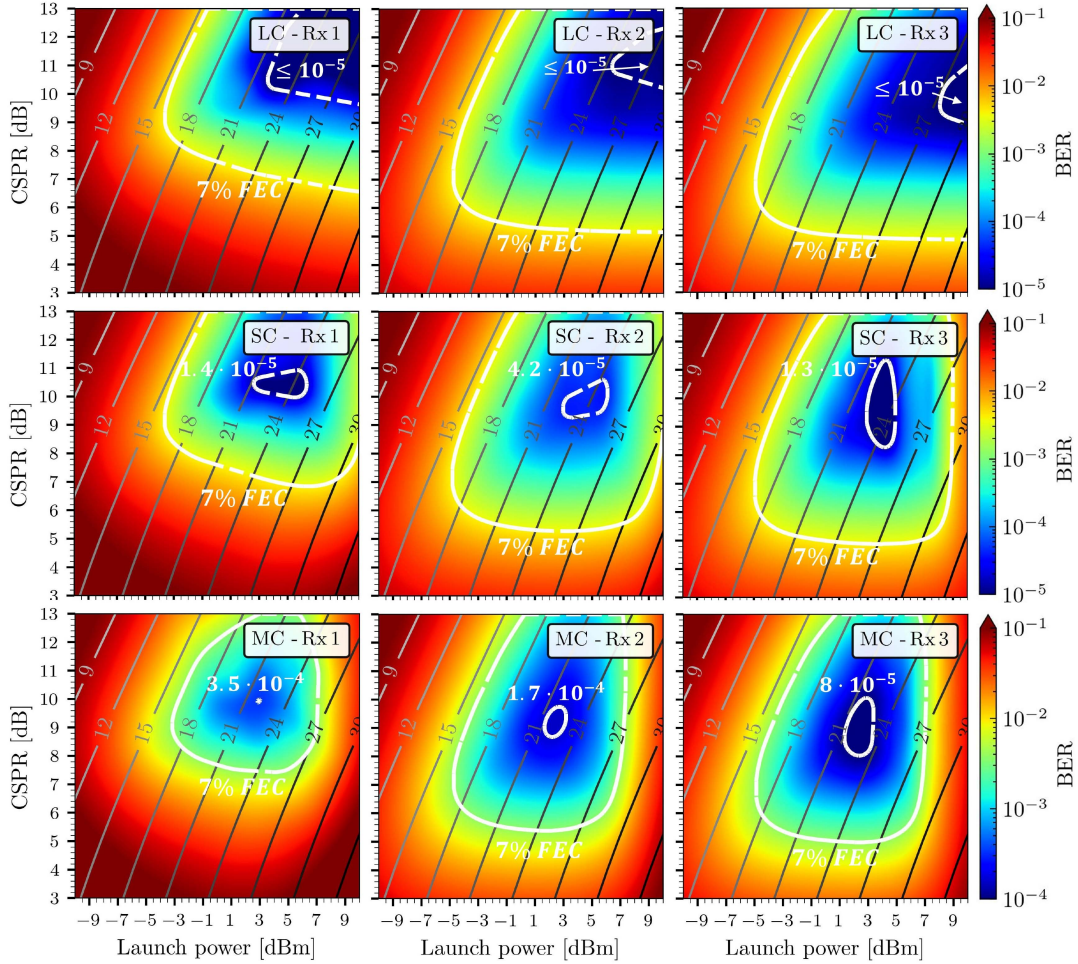


Fig. 7. Maps showing the BER as a function of CSPPR and launch power per channel (i.e., carrier plus signal powers). The considered receiver scheme varies across columns, whereas the considered transmission regime varies across rows. LC: linear channel transmission; SC: single-channel nonlinear transmission; MC: multi-channel DWDM nonlinear transmission (central channel performance). For each sub-figure, the white curve with highest iso-BER value shows the 7% HD-FEC threshold, whereas the curve with the lowest iso-BER value shows the region of parameters where the labeled BER value is achieved. The gray tilted lines show the iso-OSNR curves. The extent of the colorbars is set to different values to improve the color contrast of the maps. The NN model hyperparameters are $d = 3$ and $n_o = 32$.

signal launch power (lower equivalent OSNR) and increases the impact of the carrier-ASE beating term. This explains the non-monotonic trend of the curves for increasing CSPPR [2]. Simulation results shown in Fig. 6(b) motivates the launch power values selected for the training set data concerning Rx 3. Indeed, by selecting 1, 2, 3, and 4 dBm as launch powers for the training set, the NN is trained in a launch power region where Rx 1 (that we use as a baseline) achieves optimum BER performance. The NN nonlinear impairments compensation performance could be further improved by finely sampling the launch power values in the neighborhood of the optimum launch power when generating the training data.

The maps of Fig. 7 summarize the above-described transmission performance results by plotting the BER as a function of CSPPR and launch power per channel. The sub-figures arrangement is the same as in Fig. 5. The maps make clear the less stringent CSPPR and launch power requirements of Rx 2 and 3, compared to Rx 1, to achieve the 7% HD-FEC threshold (shown by the white iso-BER curve in the figure).

The larger areas enclosed by the white curves in the cases of Rx 2 and Rx 3 is a qualitative indication of the superior performance of the proposed schemes, compared to the reference KK scheme.

D. Receiver Sensitivity Comparison

In order to highlight the performance improvement offered by Rx 2 and Rx 3, we measured the receiver sensitivity at BER 3.8×10^{-3} . Fig. 8(a) and (b) show the results of this investigation in terms of receiver sensitivity versus CSPPR (left vertical axis) and in terms of required launch power versus CSPPR (right vertical axis).

Fig. 8(a) shows the performance of Rx 1, 2, and 3 over *Test set 2*. It can be seen that an optimum CSPPR value (i.e., achieving minimum receiver sensitivity/required launch power) exists, which is the result of a changing trade off between the MP condition being met and the increased impact of ASE-carrier beating term for increasing CSPPR values [4]. Remarkably, the

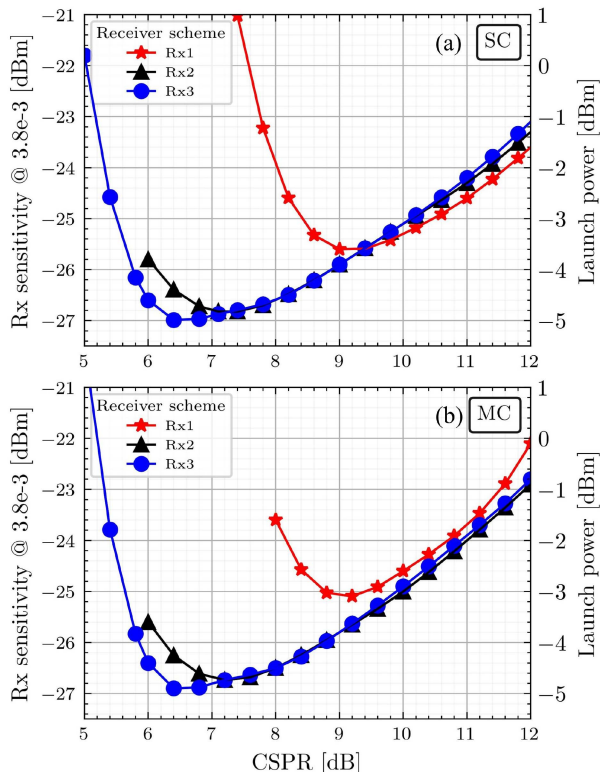


Fig. 8. Receiver sensitivity at BER of 3.8×10^{-3} (left vertical axis) and corresponding total launch power per channel (right vertical axis) as a function of CSPR. The receiver scheme associated to each curve is indicated in the legend. (a) SC: single-channel nonlinear transmission; (b) MC: multi-channel DWDM nonlinear transmission. The NN model hyperparameters are $d = 3$ and $n_o = 32$.

optimum CSPR operation point shifts from 9 dB for Rx 1, to 7.1 dB for Rx 2, and to 6.4 dB for Rx 3. At these optimum operation points, the receiver sensitivities read, -25.6 dBm, -26.8 dBm, and -27 dBm, respectively. Therefore, the relaxed CSPR requirements for Rx 3, allow for the BER threshold at 3.8×10^{-3} to be achieved with 1.4 dB better sensitivity/less total transmit power than for Rx 1.

Fig. 8(b) shows the receivers performance over *Test set 3*. The figure shows that compared to *Test set 2*, at the optimum CSPR operation point, Rx 1 incurs a ~ 0.6 dB sensitivity penalty. On the other hand, for Rx 2 and Rx 3, the incurred sensitivity penalty is less than 0.2 dB. Therefore, when Rx 3 is employed, the target BER threshold can be achieved using 2.8 less CSPR value than for Rx 1, while achieving/requiring ~ 1.8 dB better sensitivity/less total transmit power. These results further confirm the robustness of the proposed NN-based phase retrieval scheme to inter-channel nonlinear effects.

E. Influence of the NN Hyperparameters on Performance

In this section, we investigate the influence of the NN architecture hyperparameters on Rx 2 and Rx 3 performance. To this end, we consider the NN configuration described in Section II-C, and we vary either the model depth d or the number of kernels n_o . For each of the selected NN model hyperparameters we proceed as follows: (1) we train the NN model in Rx 2 and Rx 3 as detailed

in Section III-B; (2) we evaluate the receiver sensitivity at BER of 3.8×10^{-3} versus CSPR. The sensitivity performance are evaluated over *Test set 3*, i.e., in 5-channel DWDM transmission. The results are shown in Fig. 9, where the solid curves show the results of the investigations previously carried out in Section IV-C. For each plotted curve the corresponding n_o value, d value, and the number of NN model parameters are indicated in the legends.

Fig. 9(a) and (b) show the impact of the NN model memory size on the performance of Rx 2 and Rx 3, respectively. The memory of the NN model is set by the model depth d (see Section II-C), and impacts both the full-field recovery performance and the transmission impairments compensation performance. In Figs. 9(a) and Figs. 9(b) the number of kernels n_o is fixed to 32, whereas, the model depth d assumes the values 2, 3, and 4. According to Section II-C, for a model depth of 2 the memory size of the NN model results in 16 symbols, for a model depth of 3 it increases to 36 symbols, and for a model depth of 4 it results in 76 symbols. The results show that a model depth of 3 (i.e., a memory size of 36 symbols) is required for both Rx 2 and Rx 3 to avoid sensitivity penalties. For a model depth of 4 a negligible sensitivity improvement is obtained at the expense of higher NN model complexity (i.e., a higher number of parameters). Notice that a NN model memory size of 36 symbols is sufficient to account for the memory effects introduced by the transmission channel, which can be calculated to affect about 10 symbols. Indeed, for the considered transmission system parameters, namely, 100 km fiber link, 17 ps/nm/km as dispersion coefficient, and ~ 25 GHz (0.2 nm at 1550 nm) as signal bandwidth, the maximum delay experienced by the frequency components of the information bearing signal can be computed to be about 340 ps [32], [33]. Therefore, the total number of symbols that are mixed through CD is 9. The calculation for the number of symbols that are mixed through nonlinear effects is not straightforward and the reader is referred to the analysis in [34] for further insights. The results in [34] show that employing a 10-symbol wide equalizer is sufficient to effectively compensate for nonlinear effects. Therefore, according to the most stringent requirement, the NN model memory needs to be selected higher than or equal to 10 symbols to compensate for transmission impairments.

Fig. 9(c) and (d) show the impact of the number of kernels n_o on the sensitivity performance of Rx 2 and Rx 3, respectively. In the figures, the model depth is fixed to 3, whereas the number of kernels for each convolutional layer assumes the values 8, 12, 16, 32, and 64. It can be seen that increasing the number of kernels has a significant impact on the model's complexity. Indeed, the number of parameters increases by a factor ~ 70 when n_o increases from 8 to 64. Therefore, the number of kernels must be properly tuned to achieve the desired performance improvement while maintaining low complexity. Comparing Figs. 9(c) and (d), it can be seen that Rx 2 and Rx 3 have different sensitivity improvement rates as a function of n_o : up to $n_o = 16$, Rx 2 outperforms Rx 3, vice versa for $n_o = 32$ and $n_o = 64$ Rx 3 outperforms Rx 2. Therefore Rx 3 requires higher n_o values than Rx 2 to achieve similar or higher sensitivity improvements. This is explained by the fact that the NN model in Rx 3 integrates

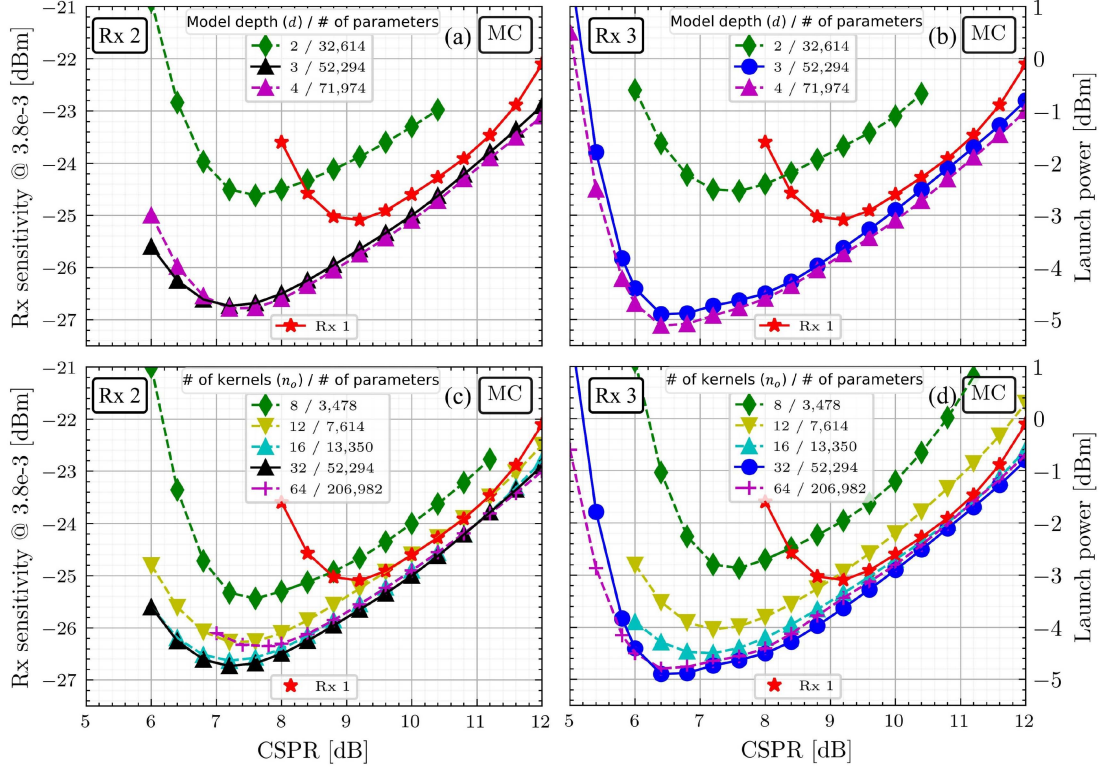


Fig. 9. Receiver sensitivity at BER of 3.8×10^{-3} (left vertical axis) and corresponding total launch power per channel (right vertical axis). The red curve shows the Rx 1 performance. The NN model hyperparameters associated to each curve are indicated in the legends. MC: multi-channel DWDM nonlinear transmission. (a) and (b): The number of kernels n_o is fixed to 32, whereas the model depth d varies. (c) and (d) The model depth is fixed to $d = 3$, whereas n_o varies. The curves with solid lines are the same as in Fig. 8(b).

the transmission impairments compensation task, thus higher n_o values results in better transmission impairments compensation performance. For $n_o = 64$, the sensitivity degradation experienced by Rx 2 is due to NN model overfitting. Section V extends the above trade-off analysis between complexity and performance by evaluating the number of real multiplications required for a NN model prediction.

V. COMPLEXITY ANALYSIS

In this section, we evaluate the computational complexity of the considered receiver schemes in terms of the number of real multiplications per recovered output sample. For the NN-based receivers, an offline training phase is assumed, thus only the computational complexity of the prediction phase is taken into account. In what follows, we first recall the complexity of the conventional KK receiver phase retrieval algorithm and the complexity of the related DBP algorithm (Section V-A). Then, we evaluate the computational complexity required by Rx 2 and Rx 3 to predict the IQ components from the photocurrent signal (Section V-B). Finally, a complexity comparison between the considered schemes is given (Section V-C). The results of the complexity analysis are summarized in Table II and in Fig. 10.

A. Complexity of Rx 1

The computational complexity of Rx 1 can be obtained as $C_{Rx1} = C_{KK} + C_{DBP}$, where C_{KK} and C_{DBP} denotes the KK

TABLE II
COMPUTATIONAL COMPLEXITIES OF THE CONSIDERED SCHEMES

	Real multiplications per sample	Tx impairments compensation
Rx 1	$C_{Rx1} = C_{KK} + C_{DBP}$	iterative (DBP)
Rx 2	$C_{Rx2} = C_{NN} + C_{DBP}$	iterative (DBP)
Rx 3	$C_{Rx3} = C_{NN}$	non iterative

phase retrieval algorithm complexity and DBP complexity, respectively.

1) *Complexity of the KK Phase Retrieval Algorithm:* According to the complexity analysis of the KK receiver performed in [9], [30], in what follows, we consider a low-complexity time-domain implementation of the KK algorithm to determine an expression for C_{KK} . The time-domain implementation of the KK algorithm can achieve similar performance to the FFT-based implementation provided that the number of taps of the employed FIR filters is sufficiently high [30]. Denoting N_S as the number of taps for the up/downsampling FIR filter, N_h as the number of taps for the FIR filter for the Hilbert transform, and $R = 4$ as the digital upsampling factor, the number of multiplications per sample required by the KK phase retrieval algorithm is [9]

$$C_{KK} = (3N_S + 2 + N_h/2) R. \quad (7)$$

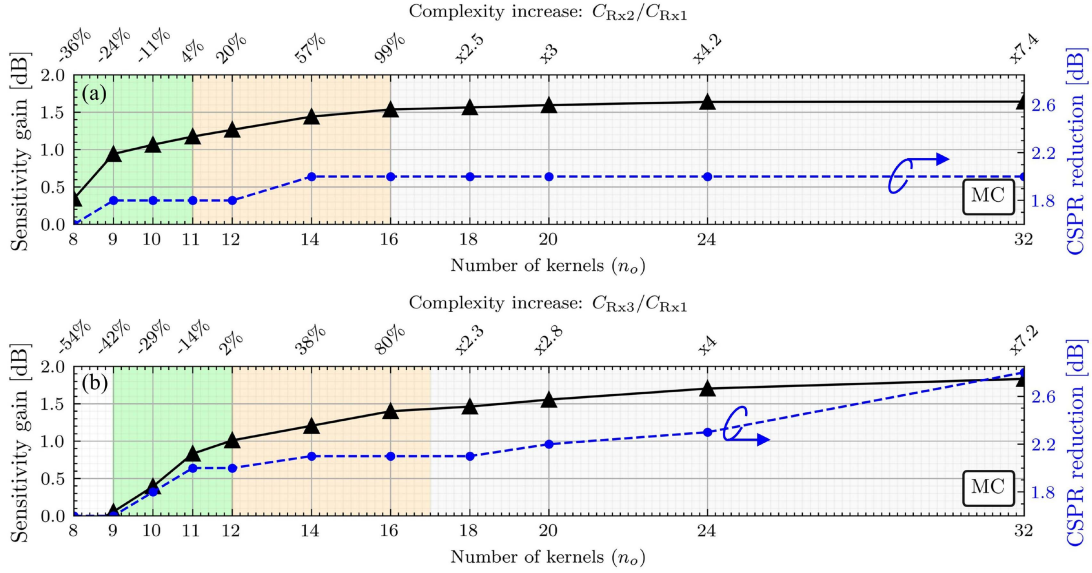


Fig. 10. Receiver sensitivity gain at the optimum CSRR operation point versus n_o (left vertical axis) and versus CSRR reduction (right vertical axis). The computational complexity increase corresponding to each n_o value is shown in the top horizontal axis. The model depth d is fixed to 3. The sensitivity gain (at BER of 3.8×10^{-3}) and the CSRR reduction are relative to Rx 1. The shaded area show the n_o values for which Rx 2 and Rx 3 achieve a complexity lower than Rx 1 (green area), a complexity up to ~ 2 times that of Rx 1 (orange area), and a complexity higher than 2 times that of Rx 1 (light gray area), respectively. Negative percentages indicate a complexity decrease compared to Rx 1. The two sub-plots share the same bottom x-axis but have different top x-axis because Rx 3 avoids the DBP algorithm. MC: multi-channel DWDM nonlinear transmission.

2) *Complexity of the DBP Algorithm:* For the linear step in the DBP algorithm, we assume blockwise frequency domain equalization using the overlap-save method [33]. In what follows, we denote as N_{STEPS} the number of steps of the DBP algorithm, as N_{FFT} the FFT size, and as N_{CD} as the minimum number taps for the CD equalizer [33] (computed in Section IV-E). N_{FFT} must be selected higher than N_{CD} , and should be optimized to minimize the computational complexity (as described later). The total number of real multiplications per equalized output sample required by the DBP algorithm can be estimated to be [20], [33]

$$C_{\text{DBP}} = 4N_{\text{STEPS}} \left(\frac{N_{\text{FFT}} (\log_2 N_{\text{FFT}} + 1)}{N_{\text{FFT}} - N_{\text{CD}} + 1} + 1 \right). \quad (8)$$

In (8), the term $N_{\text{FFT}} - N_{\text{CD}} + 1$ denotes the number of output samples produced by each iteration of the overlap-save algorithm. The optimum value for C_{DBP} is obtained for the N_{FFT} value that minimizes (8).

B. Complexity of Rx 2 and Rx 3

The complexities of Rx 2 and Rx 3 are now evaluated. To this end, we first evaluate the complexity of the NN model (shown in Fig. 1) to predict the IQ components from the photocurrent signal: C_{NN} . Then, we compute the complexity of Rx 2 and Rx 3 as $C_{\text{Rx2}} = C_{\text{NN}} + C_{\text{DBP}}$ and $C_{\text{Rx3}} = C_{\text{NN}}$, respectively. The NN model complexity C_{NN} can be approximated by considering only the computational complexity of the convolutional layers inside the D and U blocks. Namely, by neglecting the contributions of the ReLU activation functions, the long skip connections outside the D and U blocks, and the output convolutional layer in the following calculations. To obtain the expression for C_{NN} , we proceed in two steps: (1) we

evaluate the contributions to the number of real multiplications of each D/U block; (2) we sum the contributions obtained in step (1) to obtain the NN model complexity C_{NN} .

1) *Complexity for the D Blocks:* Referring to the D block structure shown on the left side of Fig. 1, the complexity of the D blocks can be calculated as ($l = 1, 2, \dots, d$) [35]

$$C_{\text{D}}^l = \begin{cases} 2(n_o k/s) + (n_o^2 k/s), & \text{if } l = 1 \\ 3(n_o^2 k/s^l), & \text{otherwise.} \end{cases} \quad (9)$$

The expression for $l = 1$ in C_{D}^l is composed of two terms. The first term accounts for the convolutional layers that have the photocurrent signal as input ($n_i = 1$): the first convolutional layer in the main path and the convolutional layer in the skip connection. The second term accounts for the second convolutional layer in the main path, where the factor n_o^2 arises because the number of input channels is $n_i = n_o$. For the other D blocks (i.e., for $l \neq 1$), $n_i = n_o$ and the factor 3 in the expression C_{D}^l accounts for the three convolutional layers inside a D block.

2) *Complexity for the U Blocks:* For every l , with $l = 1, 2, \dots, d$, the complexity of the U blocks reads

$$\begin{aligned} C_{\text{U}}^l &= 2(n_o^2 k/s^l) + (n_o^2 k/s^{l-1}) \\ &= (2 + s)(n_o^2 k/s^l). \end{aligned} \quad (10)$$

C_{U}^l is composed of two terms: the first term accounts for the two transposed convolutions in the main and residual path, whereas the second term accounts for the convolutional layer applied to the output of the transposed convolution (after upsampling).

Equations (9) and (10) show that the input number of samples (and the required multiplications) reduces/increases by a factor s as the input signal undergoes the downsampling/upsampling

process. This prevents the complexity from increasing linearly with the model depth.

3) *Complexity for a NN Model Prediction:* The complexity required for a NN model prediction can be easily obtained by summing over l - the terms given in (9) and (10). The total number of real multiplications per sample required by the NN model employed by Rx 2 and Rx 3 reads

$$C_{\text{NN}} = C_{\text{D}}^{l=1} + \sum_{l=2}^d C_{\text{D}}^l + \sum_{l=1}^d C_{\text{U}}^l. \quad (11)$$

We do not show the final expression of (11) for the sake of shortness. As expected, since C_{D}^l and C_{U}^l contain the term n_o^2 , the number of kernels n_o has the highest impact on the computational complexity.

C. Complexity Comparison

We now investigate the trade-off between the computational complexities of Rx 2 and Rx 3 and their performance. We use Rx 1 as the reference receiver scheme: we relate the computational complexities and the performance of the proposed schemes to those of Rx 1.

For Rx 1, we set $N_S = N_h = 128$ as the number of taps of the up/downsampling filter and Hilbert transform filter [30], and we consider $R = 4$ as the digital upsampling factor. For the DBP algorithm we set $N_{\text{STEPS}} = 10$, $N_{\text{FFT}} = 128$ as FFT size (obtained as the value that minimizes (8)), and $N_{\text{CD}} = 18$ as the minimum number of taps required for the CD equalizer [33].

For the NN model, we consider the following parameters: $d = 3$ for the model depth and we vary the number of kernels n_o from 8 to 32 (the actual values are 8-12 with step 1, then 12-20 with step 2, 24 and 32). For the following investigations the performance are evaluated over Test set 3, i.e., in 5-channel DWDM transmission.

The results are shown in Fig. 10(a) and (b) in terms of sensitivity gain versus n_o (left vertical axis) and in terms of CSPR reduction versus n_o (right vertical axis). The computational complexity increase corresponding to each n_o value is shown on the top horizontal axis. Hereinafter, the sensitivity gain, the CSPR reduction, and the complexity increase are intended relative to Rx 1. The sensitivity gain and the CSPR reduction at the optimum CSPR operation point are calculated from the results shown in Fig. 9(c) and (d), which show the sensitivity curves for a subset of the n_o values considered in this section. In Fig. 10(a) and (b), the complexity increase shown in the top horizontal axis is computed as the ratio $C_{\text{Rx2}}/C_{\text{Rx1}}$ for Rx 2, and as the ratio $C_{\text{Rx3}}/C_{\text{Rx1}}$ for Rx 3. For a given n_o value, $C_{\text{Rx2}}/C_{\text{Rx1}}$ is higher than $C_{\text{Rx3}}/C_{\text{Rx1}}$ because Rx 3 avoids the DBP algorithm. Each shaded area in Fig. 10 corresponds to a different complexity value as detailed in the figure description. In the figure, complexity increase factors lower than 2 are shown in percentages (negative percentages indicate a decrease in complexity), whereas complexity increase factors higher than 2 are shown using the corresponding increase factor.

The results in Fig. 10(a) and (b) can be summarized as follows. Rx 2 and Rx 3 outperform Rx 1 both in terms of performance and

in terms of complexity for a low number of kernels (i.e., for n_o in the range 8-10 for Rx 2, and $n_o = 10$ and 11 for Rx 3). Rx 2 offers the best trade-off between sensitivity improvement and complexity increase since it requires a lower number of kernels than Rx 3 to achieve a target sensitivity gain (the complexity of the NN model increases quadratically with the number of kernels). For a number of kernels higher than 20, Rx 3 effectively compensates for transmission impairments and achieves higher sensitivity gains than Rx 2. Finally, Rx 3 achieves higher CSPR reduction than Rx 2 with lower computational complexity.

It is worth mentioning that the NN model complexity provided by (11) needs to be intended as the starting expression for the computational complexity of the NN model before further optimizing the NN architecture hyperparameters. As a first approach to further reduce the NN model complexity, the number of kernels n_o , which is now constant throughout all the D and U blocks, can be properly tuned for each D/U block. Alternatively, as a second approach, network compression techniques can be applied to the trained model to reduce the number of required real multiplications with low impact on performance. For example, by pruning filters with the highest redundancy in selected convolutional layers [36].

VI. CONCLUSION

In this paper, we have proposed a DL-based method for MP signal recovery that accurately reconstruct the optical field at low CSPRs from the intensity waveform. Based on numerical simulations, we compared the performance of the conventional 4-fold upsampled KK receiver with two proposed schemes that differed in the NN training data. The first scheme was trained in ideal B2B settings to emulate the KK algorithm. For the second scheme we used an all-embracing approach on which the NN was trained over a 100 km DWDM transmission scenario with 5 channels. As a results, the NN reconstructs the optical signal while compensating for linear and nonlinear transmission impairments. Simulation results in ideal B2B settings show that the NN reconstruction improves the performance at weak carrier powers since it avoids nonlinear operations in the KK algorithm that are the main source of distortions. Simulation results in DWDM transmission validate the feasibility of the all-embracing approach showing sensitivity improvements as high as 1.8 dB achieved with up to 2.8 dB lower CSPR value compared to the conventional 4-fold upsampled KK receiver aided with DBP. To investigate the trade-off between performance and complexity, we performed a comparative analysis of the complexities of the proposed schemes with that of the KK receiver. The results showed that significant performance improvements in terms of receiver sensitivity (up to 0.4 dB better sensitivity) and CSPR reduction (up to 1.6 dB lower CSPR) can be achieved with 36% lower complexity than the conventional 4-fold upsampled KK receiver aided with DBP. The reason for the observed improvements is that the NN benefits from increased robustness to impairments that result from the training being carried out on a specific class of signals, i.e., 16-QAM symbols shaped with a RC waveform. The results also suggest

that more improvements in sensitivity gain, CSPR reduction, and complexity reduction may be possible by properly tuning the training set data and the NN model hyperparameters thus providing new avenues to design MP retrieval schemes for DD systems.

REFERENCES

- [1] C. Kachris, K. Kanonakis, and I. Tomkos, "Optical interconnection networks in data centers: Recent trends and future challenges," *IEEE Commun. Mag.*, vol. 51, no. 9, pp. 39–45, Sep. 2013.
- [2] A. Mecozzi, C. Antonelli, and M. Shtaif, "Kramers–Kronig coherent receiver," *Optica*, vol. 3, no. 11, Nov. 2016, Art. no. 1220.
- [3] Z. Li et al., "SSBI mitigation and the Kramers–Kronig scheme in single-sideband direct-detection transmission with receiver-based electronic dispersion compensation," *J. Lightw. Technol.*, vol. 35, no. 10, pp. 1887–1893, May 2017.
- [4] Z. Li et al., "Joint optimisation of resampling rate and carrier-to-signal power ratio in direct-detection Kramers–Kronig receivers," in *Proc. Eur. Conf. Opt. Commun.*, 2017, pp. 1–3.
- [5] T. Wang and A. J. Lowery, "Minimum phase conditions in Kramers–Kronig optical receivers," *J. Lightw. Technol.*, vol. 38, no. 22, pp. 6214–6220, Nov. 2020.
- [6] H. Zhang, Q. Zhang, Q. Xie, and C. Shu, "Enhanced CSPR for a multichannel Kramers–Kronig receiver by self-seeded stimulated Brillouin scattering," *Opt. Lett.*, vol. 46, no. 3, Feb. 2021, Art. no. 661.
- [7] S. T. Le et al., "1.72-Tb/s virtual-carrier-assisted direct-detection transmission over 200 km," *J. Lightw. Technol.*, vol. 36, no. 6, pp. 1347–1353, Mar. 2018.
- [8] M. van den Hout, S. van der Heide, and C. Okonkwo, "Kramers–Kronig receiver with digitally added carrier combined with digital resolution enhancer," *J. Lightw. Technol.*, vol. 40, no. 5, pp. 1400–1406, Mar. 2022.
- [9] T. Bo and H. Kim, "Toward practical Kramers–Kronig receiver: Resampling, performance, and implementation," *J. Lightw. Technol.*, vol. 37, no. 2, pp. 461–469, Jan. 2019.
- [10] A. J. Lowery, T. Wang, and B. Corcoran, "Enhanced Kramers–Kronig single-sideband receivers," *J. Lightw. Technol.*, vol. 38, no. 12, pp. 3229–3237, Jun. 2020.
- [11] A. J. Lowery and T. Feleppa, "Analog low-latency Kramers–Kronig optical single-sideband receiver," *J. Lightw. Technol.*, vol. 39, no. 10, pp. 3130–3136, May 2021.
- [12] K. Toba, T. Fujita, E. Tsukui, K. I. A. Sampath, and J. Maeda, "A study on sampling penalties reduction of Kramers–Kronig receivers," *J. Lightw. Technol.*, vol. 39, no. 19, pp. 6054–6062, Oct. 2021.
- [13] D. Orsuti et al., "Phase retrieval receiver based on deep learning for minimum-phase signal recovery," in *Proc. Eur. Conf. Opt. Commun.*, 2022, Art. no. We2C.4.
- [14] C. Antonelli, A. Mecozzi, and M. Shtaif, "Kramers–Kronig PAM transceiver and two-sided polarization-multiplexed Kramers–Kronig transceiver," *J. Lightw. Technol.*, vol. 36, no. 2, pp. 468–475, Jan. 2018.
- [15] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1003–1012.
- [16] A. van den Oord et al., "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
- [17] Y. Nishizaki, R. Horisaki, K. Kitaguchi, M. Saito, and J. Tanida, "Analysis of non-iterative phase retrieval based on machine learning," *Opt. Rev.*, vol. 27, no. 1, pp. 136–141, Feb. 2020.
- [18] Y. Rivenson, Y. Zhang, H. Günaydin, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," *Light Sci. Appl.*, vol. 7, no. 2, Feb. 2018, Art. no. 17141.
- [19] T. Kamiyama, H. Kobayashi, and K. Iwashita, "Neural network nonlinear equalizer in long-distance coherent optical transmission systems," *IEEE Photon. Technol. Lett.*, vol. 33, no. 9, pp. 421–424, May 2021.
- [20] O. Sidelnikov, A. Redyuk, S. Sygletos, M. Fedoruk, and S. Turitsyn, "Advanced convolutional neural networks for nonlinearity mitigation in long-haul WDM transmission systems," *J. Lightw. Technol.*, vol. 39, no. 8, pp. 2397–2406, Apr. 2021.
- [21] B. Kayalibay, G. Jensen, and P. van der Smagt, "CNN-based segmentation of medical imaging data," 2017, *arXiv:1701.03056*.
- [22] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB, Supplementary material: "matconvnet-manual.pdf," Chapter 5," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
- [23] W. Yi et al., "Performance of Kramers–Kronig receivers in the presence of local oscillator relative intensity noise," *J. Lightw. Technol.*, vol. 37, no. 13, pp. 3035–3043, Jul. 2019.
- [24] O. V. Sinkin, R. Holzlohner, J. Zweck, and C. R. Menyuk, "Optimization of the split-step fourier method in modeling optical-fiber communications systems," *J. Lightw. Technol.*, vol. 21, no. 1, pp. 61–68, Jan. 2003.
- [25] Z. Li et al., "Performance of digital back-propagation in Kramers–Kronig direct-detection receivers," in *Proc. Opt. Fiber Commun. Conf.*, 2018, Art. no. Tu2D.4.
- [26] 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. Topics Quantum Electron.*, vol. 28, no. 4, Jul./Aug. 2022, Art. no. 7600223.
- [27] C. Sun, D. Che, H. Ji, and W. Shieh, "Study of chromatic dispersion impacts on Kramers–Kronig and SSBI iterative cancellation receiver," *IEEE Photon. Technol. Lett.*, vol. 31, no. 4, pp. 303–306, Feb. 2019.
- [28] B. Karanov et al., "End-to-end deep learning of optical fiber communications," *J. Lightw. Technol.*, vol. 36, no. 20, pp. 4843–4855, Oct. 2018.
- [29] M. Chagnon, B. Karanov, and L. Schmalen, "Experimental demonstration of a dispersion tolerant end-to-end deep learning-based IM-DD transmission system," in *Proc. Eur. Conf. Opt. Commun.*, 2018, pp. 1–3.
- [30] C. Fullner et al., "Complexity analysis of the Kramers–Kronig receiver," *J. Lightw. Technol.*, vol. 37, no. 17, pp. 4295–4307, Sep. 2019.
- [31] L. Galdino et al., "On the limits of digital back-propagation in the presence of transceiver noise," *Opt. Exp.*, vol. 25, no. 4, Feb. 2017, Art. no. 4564.
- [32] S. J. Savory, "Digital filters for coherent optical receivers," *Opt. Exp.*, vol. 16, no. 2, 2008, Art. no. 804.
- [33] B. Spinnler, "Equalizer design and complexity for digital coherent receivers," *IEEE J. Sel. Topics Quantum Electron.*, vol. 16, no. 5, pp. 1180–1192, Sep./Oct. 2010.
- [34] O. Golani et al., "Experimental characterization of nonlinear interference noise as a process of intersymbol interference," *Opt. Lett.*, vol. 43, no. 5, Mar. 2018, Art. no. 1123.
- [35] P. J. Freire et al., "Performance versus complexity study of neural network equalizers in coherent optical systems," *J. Lightw. Technol.*, vol. 39, no. 19, pp. 6085–6096, Oct. 2021.
- [36] Z. Wang, C. Li, and X. Wang, "Convolutional neural network pruning with structural redundancy reduction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14908–14917.