



**HAL**  
open science

# MMSE-Driven Signal Constellation Scatterplot Using Neural Networks-Based Nonlinear Equalizers

Abraham Sotomayor, Vincent Choqueuse, Erwan Pincemin, Michel Morvan

► **To cite this version:**

Abraham Sotomayor, Vincent Choqueuse, Erwan Pincemin, Michel Morvan. MMSE-Driven Signal Constellation Scatterplot Using Neural Networks-Based Nonlinear Equalizers. *Journal of Lightwave Technology*, 2024, pp.1-12. 10.1109/JLT.2024.3421927 . hal-04646576

**HAL Id: hal-04646576**


**<https://hal.science/hal-04646576v1>**

Submitted on 18 Nov 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# MMSE-Driven Signal Constellation Scatterplot Using Neural Networks-Based Nonlinear Equalizers

Abraham Sotomayor , Vincent Choqueuse, Erwan Pincemin, Michel Morvan

**Abstract**—This study investigates a phenomenon observed in signal constellation diagrams when using neural networks (NNs) based nonlinear equalizers optimized with the Minimum Mean Squared Error (MMSE) criterion. This phenomenon is characterized by a concentration of the symbols around the original constellation points, with some scattered along straight lines connecting neighboring points of the original constellation. We refer to this effect as “MMSE-driven signal constellation scatterplot” (MMSE-scatterplot). This phenomenon has harmful implications for subsequent signal processing, particularly in Soft-Decision (SD) Forward Error Correction (FEC) schemes, which require reliable soft information. Indeed, the MMSE-scatterplot behaves like a hard decision or denoiser, resulting in the removal of soft-information. In this paper, we explicitly relate the MMSE-scatterplot with a function named here Soft-Thresholding (STH). Additionally, to avoid the MMSE-scatterplot emergence on the equalized symbols, we propose the inclusion of the STH function as a nonlinear activation function after the NN during the training stage. This approach permits separating the equalization stage, performed by the NN, and the denoising stage, performed by the STH function, the latter giving rise to the MMSE-scatterplot. Consequently, to recover the equalized symbols in the evaluation stage, we remove the STH function, giving as a result a constellation diagram free of MMSE-scatterplot. To assess the effectiveness of this technique, we use a numerical setup DP-64QAM transmission system with  $14 \times 50$  km of SSMF for various input signal powers. We also compare our results, in terms of bit error rate (BER) and Mutual Information (MI), with the obtained ones using an NN optimized with the recently proposed MSE-X loss function. Our results show that both NN+STH (using MSE) and NN (using MSE-X) efficiently permit to recover the equalized signal constellation free of MMSE-scatterplot, with good MI and with a slightly better BER using the NN+STH (MSE) than using the NN (MSE-X).

**Index Terms**—Neural networks, nonlinear equalizers, nonlinear optics, mean square error, Gaussian distribution.

## I. INTRODUCTION

THE application of Machine Learning (ML) in channel equalization has been extensively explored the past two decades. Some ML techniques for channel equalization were already proposed in the mid-90s [1]–[3]. The use of such “data-driven” techniques, which are capable of learning from an “input-output” relationship [4], has demonstrated superior efficiency compared to traditional adaptive methods.

Abraham Sotomayor is with Orange, 22300 Lannion, France, with IMT Atlantique, Technopole Brest-Iroise, 29238 Brest, France, and with Lab’Optic, 22300 Lannion, France (e-mail: abraham.sotomayor@orange.com).

Vincent Choqueuse is with Lab-STICC, UMR CNRS 6285 ENIB, 29238 Brest Cedex 3, France, and with Lab’Optic, 22300 Lannion, France.

Erwan Pincemin is with Orange, 22300 Lannion, France, and with Lab’Optic, 22300 Lannion, France.

Michel Morvan is with with IMT Atlantique, Technopole Brest-Iroise, 29238 Brest, France, and with Lab’Optic, 22300 Lannion, France.

In optical communications, ML has found applications for channel equalization under linear and nonlinear impairments (NLI). Particularly interesting is its application against NLI, which often act as barriers to approaching Shannon’s capacity [5]. Classical deterministic techniques, such as the Digital Backpropagation (DBP), are challenging to apply due to the considerable computation involved [6]–[8].

To address NLI compensation using ML, one promising approach is the use of Neural Networks (NNs). For instance, different types of NNs, such as the Multilayer Perceptron (MLP) [9], [10], Convolutional NNs (CNNs) [11], [12], Recurrent NNs (RNNs) [6], [13], and CNN+RNN [14], have been tested in various numerical and experimental scenarios. A comparison of different NNs architectures in terms of statistical performance and complexity was presented in [15]. In [16], an NN architecture was proposed for optical/electrical nonlinearity compensation, while [7], [17] proposed NNs for digital pre-distortion at the transmitter side. Additionally, several physics-based complex-values NNs, such as the Learned DBP (LDBP), were proposed in [18]–[21]. A comparison between the MLP and the LDBP was also conducted in [22].

When using ML algorithms, there are two main tasks: classification and regression. In classification, the aim is to categorize the output into different target signal classes, usually expressed in terms of probabilities. On the other hand, in regression, the goal is to reconstruct the transmitted signal.

The learning process involves using a loss or cost function to measure the difference between the predicted and the expected result. Based on that difference, the NN parameters are optimized, for instance using a gradient-descent-based approach. While classification tasks are typically associated with the Cross-Entropy Loss (CEL), regression tasks commonly employ the Mean Squared Error (MSE).

NNs-based signal equalizers in classification tasks (hereinafter NN-Class) perform two sub-tasks: signal equalization and soft-demapping concurrently. On the other hand, NNs-based signal equalizers in regression tasks (hereinafter NN-Reg) conduct only signal equalization. The NN-Class has proven to be optimal in the sense that it maximizes the information content [23], [24]. The benefits of using an NN-Class were also highlighted in [25]. Nevertheless, they also revealed that an NN-Class using a CEL faces additional ML-related challenges. These include issues like overfitting, a tendency to converge to local minima losses, and unsuitability for highly accurate systems, as is the case for optical transmissions. Another reason that could blur the NN-Class is the loss of access to the equalized signal, which is crucial for tasks such as carrier phase estimation and synchronization [26]. The

interest in using NN-Reg, therefore, relies on the accessibility of the equalized signal.

Of special interest is the use of NNs as nonlinear equalizers. Focusing on nonlinear NN-Reg equalizers using the MSE loss function in the learning process, it has been observed, particularly in QAM constellations, that the constellation diagram exhibits a specific distortion, referred to as the “jail window” pattern in [27] or the “MSE-grid scatterplot” in [16], [26]. These terms highlight the rectangular nature of QAM constellations. This effect appears as a concentration of the equalized symbols around the original constellation points with some scattered along the straight lines between the neighboring points of the original constellation. This effect is called here “MMSE-driven signal constellation scatterplot” (MMSE-scatterplot) because it appears when using the minimum MSE (MMSE) criterion in the NN-Reg optimization. While the MMSE-scatterplot has been noted in multiple works on NN-Reg using the MSE loss function [9], [16], [28]–[31], the focus on this phenomenon has gained more attention in recent works [26], [27], [32].

The presence of the MMSE-scatterplot can have detrimental consequences on subsequent signal processing blocks, especially on Soft-Decision (SD) Forward Error Correction (FEC) schemes [26], [27], which require reliable soft information<sup>1</sup>. Furthermore, in the classical Digital Signal Processing (DSP) coherent receiver, this soft-information is provided by the demapper based on the equalized signal but also on the optical channel law [33]. However, since this latter is usually unknown, an auxiliary memoryless Additive White Gaussian Noise (AWGN) channel is commonly assumed [33], which even proves to be a good assumption in the nonlinear regime [34]. Therefore, the role of the equalizer is to reconstruct the transmitted signal and provide the necessary information to the demapper in the form of constellation with Gaussian-like noise. The MMSE-scatterplot, however, completely disrupts the Gaussian-like properties expected by the demapper.

Besides this, the MMSE scatterplot also affects the Achievable Information Rate (AIR) estimations, such as the Mutual Information (MI) and the Generalized Mutual Information (GMI), necessary for the Bit Error Rate (BER) after the FEC (postFEC-BER) predictions [33], [35]. Indeed, the MI and GMI, are commonly estimated through closed-form expressions using the auxiliary memoryless AWGN channel [35]–[37], proven to be lower bounds of the MI/GMI of the true channel with memory [38]. The MMSE-scatterplot effect induces a significant alteration in the equalized signal constellation, reducing the precision of these closed-form expressions<sup>2</sup>. Therefore, is required that the noise distribution of the equalized signal approximates a Gaussian.

Some techniques have been proposed to mitigate the MMSE-scatterplot effect. Particularly, in [27] was proposed the monitoring of the MI’s lower bound (MI-LB) in a validation dataset, stopping the training when the MI-LB reaches

its maximum value (early stopping). Another alternative, presented in [26], introduces a novel loss function called MSE-X, which combines the MSE with a regularization term based on the AIR maximization. This term ensures that the noise of the equalized signal follows a Gaussian distribution. However, this regularization term requires a fine-tuning of its noise variance parameter, otherwise, the loss function might fail to converge.

This paper aims to delve deeper into this MMSE-scatterplot effect, building upon previous studies’ findings [25]–[27], [32] and proposing an alternative solution to prevent its occurrence.

The main contributions of this work could be summarized as follows:

- We explicitly associated a mathematical expression derived from the MMSE analysis in previous works with the MMSE-scatterplot effect. This equation is called the Soft-Thresholding (STH) function.
- We proposed to use the STH function as a nonlinear activation function placed after the NN during the training. In the evaluation, the equalized signal, free of the MMSE-scatterplot effect, is obtained before the STH function.

The rest of this paper is organized as follows: In Section II we describe the MMSE-scatterplot effect. Section III explains the fundamental origin of this phenomenon and its equivalent mathematical function. Section IV presents the related works and describes the proposed technique to avoid the MMSE-scatterplot effect. Section V details the numerical setup considered in this study, and Section VI presents the results. The paper concludes by summarizing key findings and suggesting future perspectives.

## II. MMSE-DRIVEN SIGNAL CONSTELLATION SCATTERPLOT

When employing an NN-Reg based nonlinear equalizer using the MSE loss function, it has been observed in QAM modulated signals, a phenomenon termed “jail window pattern” in [27] and “MSE-grid scatterplot” in [26]. These names were given due to the rectangular nature of QAM constellations. To illustrate this phenomenon, let us consider the case of an AWGN channel and a simple NN-Reg. Since we employed the NN-Reg throughout the rest of the paper, let us simplify its notation and denote it as NN.

Let  $X$  be a discrete random variable that represents the sequence of transmitted symbols with an alphabet  $\mathcal{X}$  consisting of  $M$  discrete symbols, i.e.  $\mathcal{X} = \{x_1, \dots, x_M\}$ , and let  $R$  be also a discrete random variable that represents the sequence of received samples. An AWGN channel has the form:

$$R = X + Z, \quad (1)$$

where  $Z$  is a complex Gaussian-distributed random variable with zero mean and total variance  $\sigma^2$ ,  $Z \sim \mathcal{CN}(0, \sigma^2)$ .

Consider an NN with parameters  $\theta$ , with input  $R$  and output  $Y = f(R; \theta)$ , where  $f$  represents the NN and  $Y$  is the estimate of the transmitted sequence  $X$ . An illustration of this setup is shown in Fig. 1, where the equalized signal  $Y$  is the demapper input and  $Q_{X|Y}(\cdot|y)$  is the soft-information in the form of a posterior distribution that feeds the SD-FEC [26], [33].

<sup>1</sup>Soft: a continuous range of probabilities of belonging to a determined class or category. Hard: a real specific value (e.g. 0 or 1 in binary codes)

<sup>2</sup>A similar reasoning could be made for the Q-factor estimate, which is calculated using a Gaussian noise assumption, being not valid when the MMSE-scatterplot appears.

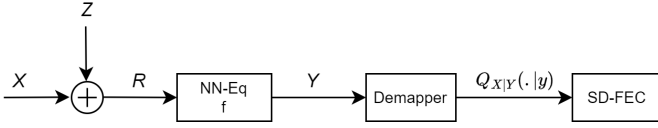


Fig. 1. Classical transceiver model with an NN-based nonlinear equalizer (regressor) in an AWGN channel represented by  $Z$ .

The objective of the NN is to bring  $Y$  closer to  $X$ . The MSE is commonly used to measure the difference between  $Y$  and  $X$ , and can be expressed as [24], [25]:

$$\text{MSE}(X, Y) = \mathbb{E}[|X - Y|^2], \quad (2)$$

where  $\mathbb{E}[\cdot]$  is the expectation operator.

For simplicity, consider NNs using real number parameters. Therefore, the complex inputs were separated into their I and Q components. Consequently, the outputs also corresponded to the separated I and Q parts of the predicted complex symbols.

The chosen NN was the MLP because it is known that it could approximate any nonlinear function [39], but as we will see later, there are not restrictions for the NN-based nonlinear equalizer architecture. The MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer comprises multiple units or neurons. Following the format used in other studies, we represent the MLP as  $N_i|N_1|f_a|\dots|N_L|f_a|N_o$ , where  $N_i$  signifies the number of neurons in the input layer,  $N_h$  denotes the number of neurons in the hidden layer  $h$  (with  $1 \leq h \leq L$ ),  $L$  signifies the total number of hidden layers, and  $N_o$  represents the number of neurons in the output layer. After each layer (linear part), we applied a nonlinear activation function  $f_a$ , except for the output layer. In the case of the regression task, the number of neurons in the output layer is fixed to  $N_o = 2$  (in the case of one polarization), as we aimed to recover the real and imaginary parts of the equalized symbol. The NN architecture utilized in this study for the AWGN channel is outlined in Table I with the specifications for the training.

Fig. 2a and 2c show  $R$  (NN input) for squared 16QAM and rectangular 8QAM, respectively. The noise variance  $\sigma^2$  was set to achieve a received BER of  $10^{-3}$ . Meanwhile, Fig. 2b and 2d show  $Y$  (NN output), that correspond to the observed phenomenon in rectangular constellations. A similar experiment was done in [25], where a pure AWGN channel with only Chromatic Dispersion (CD) was employed to show the MMSE-scatterplot emergence after some epochs using an NN-based nonlinear equalizer.

This phenomenon, to the best of our knowledge, has primarily been attributed to rectangular QAM constellations. However, this effect is not exclusive to QAM signals; a similar one also appears in PSK constellations. For example, if we used an 8PSK signal (Fig. 2e), using a similar NN as in the previous example, we found the outcome presented in 2f. For this constellation, the NN output on the right adopts the geometric shape of an octagon, corresponding to the eight points of the constellation in this case. Comparable results could be achieved with any other PSK constellation.

TABLE I  
NN ARCHITECTURE AND TRAINING SPECIFICATIONS USED IN SECTION II FOR AN AWGN CHANNEL.

NN architecture	Activation function $f_a$	Learning rate	Training epochs	Loss function
2 30  $f_a$  30  $f_a$  2	Tanh	$10^{-3}$	1000	MSE

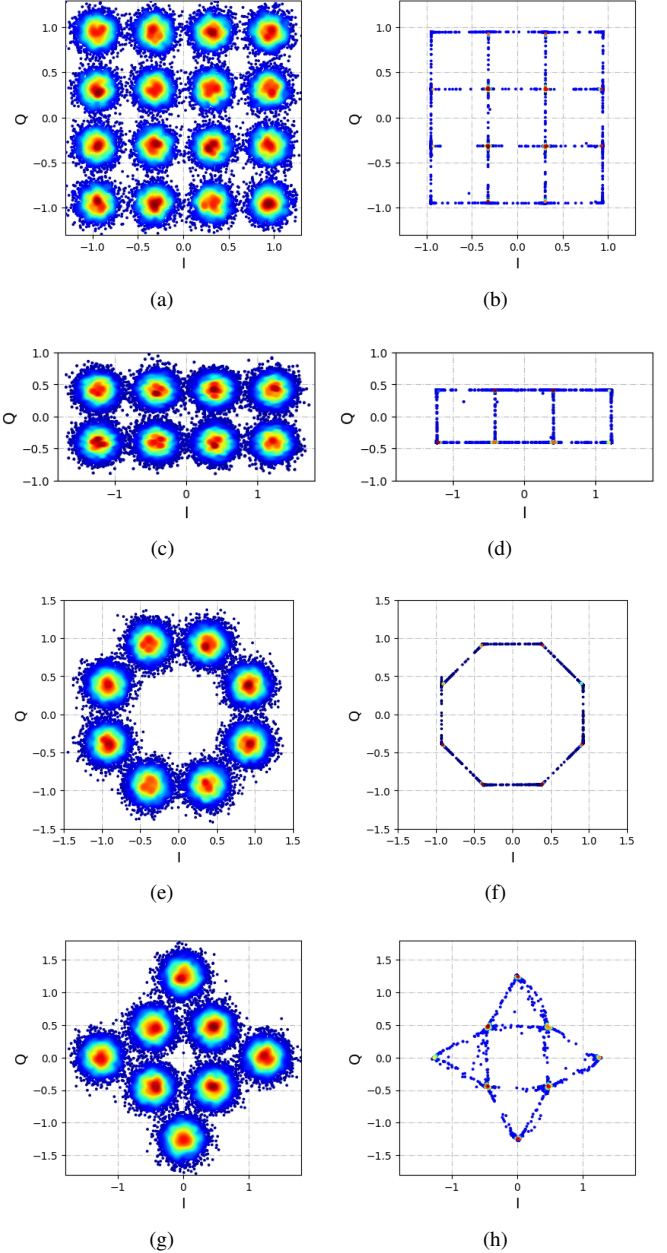


Fig. 2. MMSE-scatterplot effect on signal constellations in an AWGN channel. On the left: NN inputs  $R$ , on the right: NN outputs  $Y$ . (a,b) 16QAM, (c,d) 8QAM, (e,f) 8PSK, (g,h) 8QAM (optimal)

In general, this phenomenon could be observed in any signal constellation with an equal probability of occurrence for all the symbols. For instance, in the optimal 8QAM signal constellation [40] (Fig. 2g), we can also note a distortion in the NN output constellation (2h). Once again, it is observed that the NN induces a specific alteration of the constellation.

All the previously mentioned results share a common characteristic: the NN induces a concentration of the symbols around the original constellation points, with some scattered along straight lines connecting neighboring points of the original constellation. This effect is referred to here as the “MMSE-Driven Signal Constellation Scatterplot” (MMSE-scatterplot). Notably, this phenomenon is consistently observed in experiments involving M-PSK ( $M = 4, 8, 16$ ), M-QAM ( $M = 4, 8, 16, 32, 64$ ), and any signal constellation with a uniform probability of occurrence for all the symbols. It is worth mentioning that this phenomenon is a direct consequence of the choice of the MSE as a loss function.

In the subsequent section, we delve into the fundamental origin of the MMSE-scatterplot effect by means of its equivalent mathematical expression.

### III. ORIGIN OF THE MMSE-SCATTERPLOT

The emergence of the MMSE-scatterplot effect when using the MSE loss function in NN-based nonlinear equalizers necessitated further investigations. This section aims to establish a relationship between the results obtained from the MMSE criterion and the characteristics observed in the MMSE-scatterplot.

Is not worth mentioning that the MMSE-scatterplot only appears when using nonlinear equalizers, i.e. general functions whose outputs are not linear with respect to their inputs, with the MMSE criterion. Therefore, we focus our study on nonlinear equalizers and also we restrain our study to Gaussian-like channels, which is the case of nonlinear channels that could be considered as a Gaussian noise source [34].

#### A. MSE Loss Function

The optimization criterion when training an NN using the MSE loss function is to minimize the MSE, referred to as the MMSE criterion. The rationale behind employing the MSE lies in its functional properties, as discussed in [41]: i) MSE is a differentiable function, making it suitable for gradient backpropagation during the optimization stage. ii) In the case of linear problems, MSE is a convex function, ensuring convergence to the global minima. However, it is still widely used in non-convex optimization problems, where acceptable local minima are achievable [42].

Moreover, a significant property is the relationship between the MSE and the conditional Maximum Likelihood Estimation (MLE). The conditional MLE aims to find the optimal set of parameters  $\theta$ , maximizing the posterior probability to estimate  $X$  given  $R$ ,  $P_{X|R}(x|r)$ , i.e.  $\theta_{ML} = \arg \max_{\theta} P_{X|R}(x|r)$ . Assuming that  $P_{X|R}(x|r)$  is Gaussian distributed, it has been demonstrated that the conditional MLE and the MMSE criterion are equivalent [41], [42].

#### B. MMSE estimate

The optimal estimate of  $X$ , i.e.  $Y = \hat{X}$ , under the MMSE criterion, was studied, for instance, in [43]–[45]<sup>3</sup>. It has been

demonstrated that the MMSE is attained when  $Y$  is the mean of the posterior probability  $P_{X|R}(x|r)$ , i.e.

$$Y = \mathbb{E}[X|R]. \quad (3)$$

In the scenario where  $x \in \mathcal{X}$ , with  $P_X(x) = \frac{1}{M}$  uniform  $\forall x \in \mathcal{X}$  and zero elsewhere, and assuming  $Z \sim \mathcal{CN}(0, \sigma^2)$ , it has been shown in [46, Eq. 3.11], [45, Eq. 5] [43, Eq. 10.9] that the optimal estimate of  $X$  is given by

$$Y = \mathcal{S}_{\mathcal{X}}(R; \sigma^2), \quad (4)$$

where  $\mathcal{S}_{\mathcal{X}}(R; \sigma^2)$  is a Soft-Thresholding (STH) projector onto  $\mathcal{X}$ . This expression is called here STH function and for  $R = r$  is defined as follows:

$$\mathcal{S}_{\mathcal{X}}(r; \sigma^2) = \frac{\sum_{x \in \mathcal{X}} x e^{-\frac{1}{\sigma^2}|r-x|^2}}{\sum_{x \in \mathcal{X}} e^{-\frac{1}{\sigma^2}|r-x|^2}}. \quad (5)$$

In [24, Eq 6.24], the authors analyzed the case of the optimal equalizer using the MMSE criteria. They considered no restrictions on the equalizer, i.e., not necessarily linear, and assumed Gaussian noise and BPSK symbols. They found that the MSE minimization reduces the equalizer function to

$$f(r; \sigma^2) = \tanh\left(\frac{r}{\sigma^2}\right), \quad (6)$$

which is equivalent to the equation (5) for  $x \in \{-1, +1\}$  [Eq. 3.14] [46].

As the noise variance approaches zero ( $\sigma^2 \rightarrow 0$ ), one can note that the STH function reduces to the Hard-Thresholding function  $\mathcal{H}_{\mathcal{X}}(r)$ , defined by

$$\mathcal{H}_{\mathcal{X}}(r) = \lim_{\sigma^2 \rightarrow 0} \mathcal{S}_{\mathcal{X}}(r; \sigma^2) = \arg \min_{x \in \mathcal{X}} |r - x|^2. \quad (7)$$

Note that this function corresponds to the optimal Hard Decision (HD) detector into the signal constellation.

For illustrative purposes, the Soft and Hard-thresholding functions are depicted in Fig. 3 for the particular case of a PAM4 constellation. Observe that, for small values of the noise variance  $\sigma^2$ , most points will concentrate around  $x_1, x_2, x_3$  and  $x_4$ , whereas for large values of  $\sigma^2$ , these points will spread more along the straight lines between neighboring symbols  $x_1 - x_2, x_2 - x_3$ , and  $x_3 - x_4$ .

Consider the inputs  $R$  of the previous examples whose constellations are depicted in Fig. 2a, 2c, 2e and 2g. When using the STH function of (5) with a  $\sigma^2$  equals to the AWGN variance, the outputs show the constellations depicted in Fig. 4 for each case. Upon simple inspection, it is clear that these results exhibit the same distribution as those when using the NN, shown in Fig. 2b, 2d, 2f, and 2h, and the MMSE criterion during the training.

Therefore, we can conclude that in an AWGN channel, a nonlinear equalizer, e.g. an NN, with the objective of minimizing MSE (MMSE criterion) simplifies to the STH function in (5).

It must be mentioned that the use of such a function (5) was also investigated in [46], but in the pursuit of a less complex NN that provides the optimal estimator for the transmitted signal, specifically for PAM constellations.

<sup>3</sup>In the context of MIMO detection, the MMSE estimate has been referred to as the optimal denoiser [45].



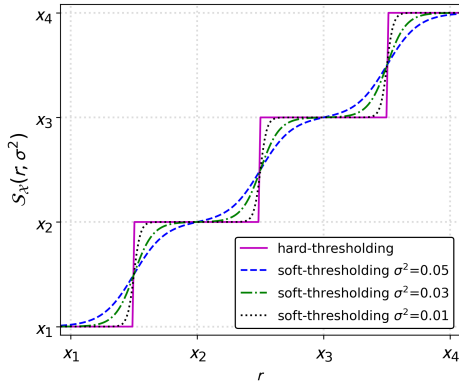


Fig. 3. Soft and Hard-thresholding functions for a PAM4 constellation with alphabet  $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ .

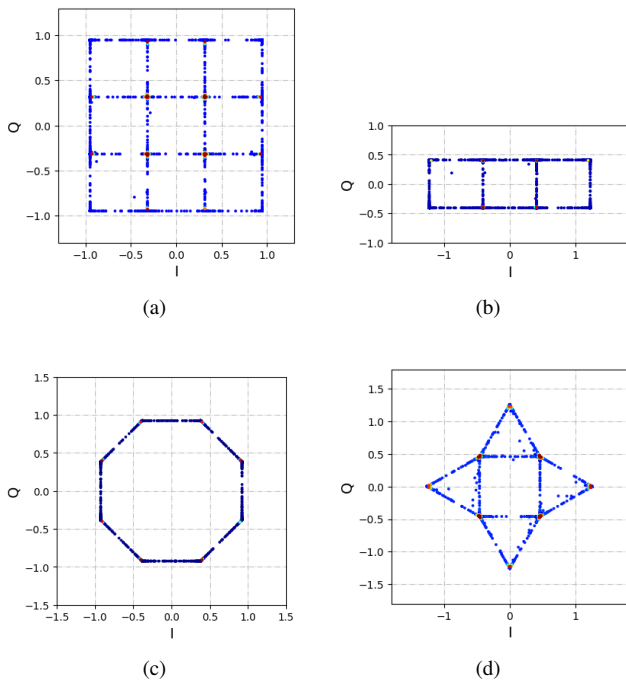


Fig. 4. Estimate of  $X$  using the STH function in (5), where  $R$  are the inputs in Fig. 2a, 2c, 2e and 2g. (a) 16QAM, (b) 8QAM, (c) PSK, (d) 8QAM (optimal).

#### IV. MITIGATION TECHNIQUES OF THE MMSE-SCATTERPLOT EFFECT

This section describes in detail the related works concerning the MMSE-scatterplot effect and presents another alternative to avoid its appearance.

##### A. Related works

In [27], the authors provided some insights about the MMSE-scatterplot effect, which they called the “jail-window” pattern. They highlighted the fact that this phenomenon appears due to the Euclidean distance minimization between the target and predicted symbols performed by the NN when minimizing the MSE loss function. Additionally, they provided possible reasons for its appearance from an ML perspective, e.g. the mismatch between the ultimate goal of the NN-based equalizer (BER improvement) and the NN loss function

(MSE), and the use of not enough large mini-batch sizes. Other important observations were made on that paper. For instance, the necessity of carefully training the NN, being aware to avoid overfitting at most, and making the different datasets employed for training/validation/test highly uncorrelated. In that work, it was also proposed the use of the  $L^2$  regularization technique to mitigate the MMSE-scatterplot effect. However, it was pointed out that it could only partially mitigate the MMSE-scatterplot effect and it still needed large mini-batch sizes.

Another study in [25], provides other solutions to mitigate the MMSE-scatterplot effect. In particular, an early stopping routine based on the maximum AIR estimated in a validation dataset. The AIR estimated in that work was the MI-LB. Indeed, monitoring the MI-LB on the validation dataset is indicated as the best approach to selecting a model that avoids the MMSE-scatterplot effect. However, a signal could still have a good AIR but not be fully equalized. From our point of view, the NN could in effect, avoid the MMSE-scatterplot effect by selecting the model with a good AIR, but not necessarily improving the BER, which is also the goal of the equalizer.

Other authors in [26] also specifically investigated how to mitigate the MMSE-scatterplot effect using a regularization term. They proposed the use of a regularization term based on the *a posteriori* probability distribution  $Q_{X|Y}(x|y)$  of the demapper, where  $X$  is the transmitted data and  $Y$  is the received data after equalization. This function was called MSE-X and defined as

$$\text{MSE-X}(X, f(R)) = \text{MSE}(X, f(R)) - 2\sigma^2 \mathbb{E}[-\log Q_Y(f(R))], \quad (8)$$

where  $Y = f(R)$ ,  $f$  represents the NN channel equalization, and  $\sigma^2$  is the noise power.

The MSE-X loss function requires to set a parameter  $\sigma^2$  related to the variance of the equalized signal. We found that this parameter is difficult to set as the expected quality of the equalized signal is unknown.

In this paper, we propose an alternative approach to mitigate the MMSE-scatterplot effect, by using the STH function as a nonlinear layer at the end of the NN.

##### B. Soft Thresholding-based Output Layer

We propose a new approach that involves adding the STH function (5) after the NN. Instead of directly producing two real numbers as the predicted symbol, we introduce the STH function as an additional nonlinear function immediately after the output layer. A similar “staircase” function was proposed in [28], [47] as a nonlinear activation function to handle nonlinearities in M-QAM systems. They showed that this function effectively minimizes the MSE with BER improvements. However, when observing the MMSE-scatterplot in their results, is clear that the AIR is very poor.

In [46], this function was used as an optimal<sup>4</sup> NN. Indeed, the STH function can be used as a single-layer NN with only a few neurons, with the minimal number of neurons

<sup>4</sup>Notice that the term “optimal” referred to the minimum MSE.

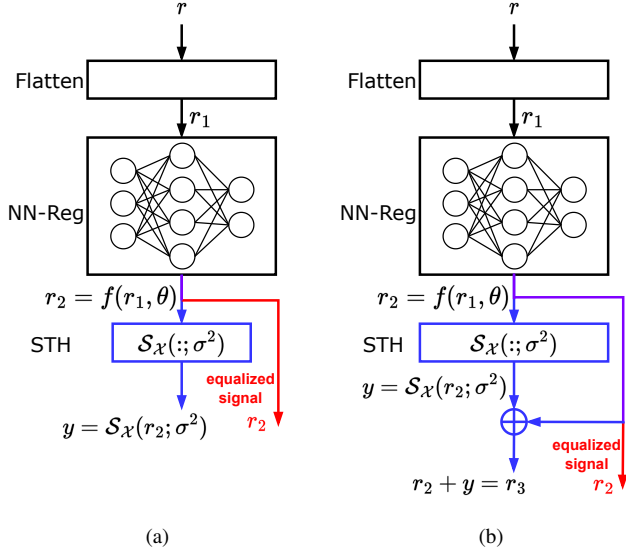


Fig. 5. Proposed NN followed by the STH function. (a) NN + STH, (b) NN + STH with residual connection to solve vanishing gradient problem. Flatten: reshapes input  $r$  into a vector of shape  $(B, :)$ , where  $B$  is the mini-batch size. Color blue indicates only training mode, red indicates only evaluation mode, and purple indicates both training and evaluation modes.

being  $\sqrt{M} - 1$  for a real-valued NN in squared M-QAM constellations. However, this ultra-short NN is not useful as it does not perform any equalization.

We require an NN that can perform equalization, thereby increasing the AIR and decreasing the BER by minimizing the MSE during the training stage. The key to adding the STH function as a nonlinear layer after the NN is the following: on the one hand, the NN addresses the equalization with MI and BER improvement. On the other hand, the STH function takes the role of the MMSE-scatterplot. Indeed, the NN alone plays both roles. Why not separate both roles if the MMSE-scatterplot expression is known?

The NN followed by the STH (NN + STH) architecture is illustrated in Fig. 5a. During the training stage, the output  $Y$  is the outcome of the NN + STH and is used to calculate the  $\text{MSE}(X, Y)$ . However, during the evaluation stage, the equalized signal, denoted as  $R_2$ , is the signal recovered before the STH function, and is free of the MMSE-scatterplot.

The STH-based layer relies on a parameter  $\sigma^2$ , which must be appropriately set to generate the MMSE-scatterplot. Since the gradient loss tends to become null for small  $\sigma^2$  due to this “soft staircase” function, it becomes susceptible to block the gradient backpropagation. To tackle the issue of blocking backpropagation, we can employ a well-known technique, based on a residual connection [48]. Residual connections were initially introduced to alleviate the vanishing gradient issue in deep NNs. In this context, the vanishing gradient problem does not arise from a deep structure but rather from the STH function. A residual connection facilitates gradient propagation through two connections. A modified architecture, featuring a residual connection, is depicted in Fig. 5b. More formally, let  $\mathcal{L}$  be the loss, i.e.,  $\mathcal{L} = E[|R_3 - X|^2]$ , then the variation of NN parameters  $\theta$  will occur through the gradient

descent of  $\mathcal{L}$  with respect to  $\theta$ :

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial r_3} \frac{\partial r_3}{\partial \theta} \quad (9a)$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial r_3} \left( \frac{\partial r_2}{\partial \theta} + \frac{\partial y}{\partial \theta} \right) \quad (9b)$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial r_3} \left( \frac{\partial r_2}{\partial \theta} + \frac{\partial y}{\partial r_2} \frac{\partial r_2}{\partial \theta} \right) \quad (9c)$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial r_3} \frac{\partial r_2}{\partial \theta} \left( 1 + \frac{\partial y}{\partial r_2} \right). \quad (9d)$$

In this manner, even if the gradient  $\frac{\partial y}{\partial r_2}$  becomes zero,  $\frac{\partial \mathcal{L}}{\partial \theta}$  can still propagate due to the second connection. Notice that in the first architecture (Fig. 5a),  $\frac{\partial \mathcal{L}}{\partial \theta}$  becomes zero when  $\frac{\partial y}{\partial r_2}$  approaches zero.

The selection of  $\sigma^2$  using the residual connection was empirically approached in two different ways:

- Calculating the value of  $\sigma^2$  for each batch by using the equalized signal before the STH function along with the transmitted signal. This method requires feeding the NN with the transmitted signal to calculate  $\sigma^2$ .
- Treating  $\sigma^2$  as a learnable parameter of the NN.

In practice, we noticed that the first option did not provide the desired results, as the calculated value of  $\sigma^2$  was too large to generate the required MMSE-scatterplot effect. Therefore, we opted for the second option. However, during the training process, there comes a point where the gradient is unable to backpropagate, due to the very small values of  $\sigma^2$ , even with the residual connection. This eventually stops the training. Nevertheless, this is not an issue as long as the equalization has been performed.

Another alternative that we explored is the use of Kurtosis as a regularization term. Indeed, Kurtosis tends to approach zero for Gaussian-distributed random variables [49], [50]. The idea was to encourage the minimization of the Kurtosis during the training. In doing so, we aimed to force the noise to be Gaussian-distributed, avoiding the MMSE-scatterplot appearance. However, we decided not to use this approach because a Kurtosis value close to zero does not necessarily indicate a Gaussian distribution. While it worked in some cases, it was not easily applicable to other cases.

## V. DESCRIPTION OF THE NUMERICAL SETUP

In this section, we describe the transmission setup under study, the datasets building, the NN architecture, and the training/validation process.

### A. Transmission setup

To investigate the conditions leading to the MMSE-scatterplot effect, we initiated a numerical study based on a dual-polarization transmission setup illustrated in Fig. 6. In this setup, the transmitted symbols  $X$  are oversampled to 8 samples/symbol (SpS) to simulate the digital-to-analog conversion. After pulse shaping using Root-Raised Cosine (RRC) filters, the combined dual-polarization signal propagates through the optical channel. At the receiver, DSP techniques are applied exclusively to address linear impairments.

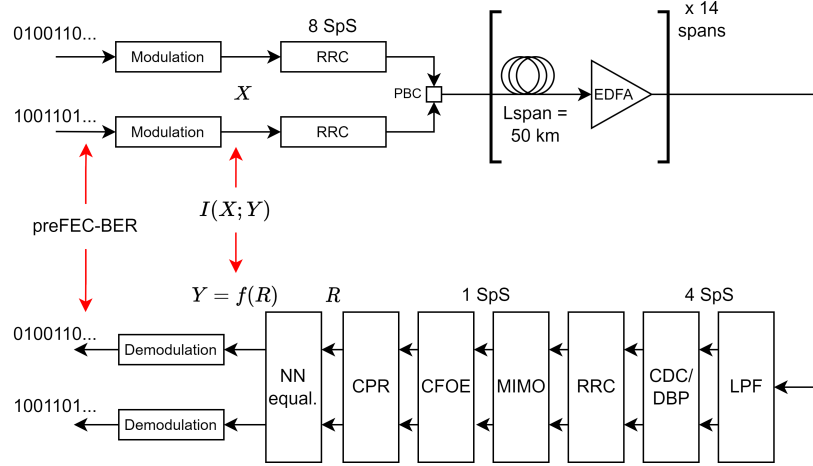


Fig. 6. Dual-Polarization Transmission Setup. SpS: samples per symbol, RRC: root-raised cosine, PBC: polarization beam combiner, EDFA: erbium-doped fiber amplifier, LPF: low pass filter, CDC: chromatic dispersion compensation, DBP: digital backpropagation, MIMO: multiple-input multiple-output, CFOE: carrier frequency offset estimation, CPR: carrier phase recovery.

The resulting signal, denoted as  $R$ , is used as the input for the NN represented by  $f$ . We denote the output of the NN as  $Y = f(R)$ .

The optical channel consisted of 14 spans of standard single-mode fibers (SSMF) with a span length of 50 km. After each span, an erbium-doped-fiber amplifier (EDFA) fully compensates for the fiber loss. To simulate the signal's propagation, we numerically solved the Manakov-PMD equation [51], [52] using the split-step Fourier method (SSFM) [53]. The Rx-DSP low-pass filters the signal to the effective bandwidth. Subsequently, the signal undergoes an undersampling to 4 SpS for full CD compensation (CDC), or using the DBP [52], followed by an identical RRC pulse shaping to mitigate inter-symbol interference. Subsequently, a polarization demultiplexing technique based on a 2x2 Multiple-Input-Multiple-Output (MIMO) equalizer combined with a Fractionally-Spaced Equalizer (FSE) and sequentially using the Constant Modulus Algorithm (CMA) [54] and Radius-Directed Equalizer (RDE) [55], was applied to recover the signal at the symbol rate (1 SpS) [56]. This processing was followed by an estimation of frequency offset [57] and carrier phase [58]. The resulting output  $R$ , is still affected by the interplay between fiber nonlinearity, CD, PMD, and amplified spontaneous-emission (ASE) noise originated by EDFAs. Indeed, this setup allowed us to analyze the impact of significant accumulated fiber nonlinearity, in the widely deployed SMF-28 fiber type. The specific simulation parameters are outlined in Table II.

### B. Datasets

For the training and validation process, we used 14 different datasets. Each dataset was generated with a different random pattern and contained 233,274 symbols, after the classical DSP described in the precedent subsection. For each dataset, 50% of the data was used for training, and the following 50% was used for validation. Then, similar to [27], we reshaped each part in the form  $(B, S, 4)$ , where  $B$  is the mini-batch size equal to 4096,  $S$  accounts for  $N$  neighboring symbols of the input

TABLE II  
PARAMETERS OF NUMERICAL SIMULATION.

PARAMETER	VALUE
System	Dual-Polarization
Modulation	64QAM
Baud Rate	32 Gbaud
Wavelength	1552 nm
Laser linewidth	100 KHz
Frequency offset	200 MHz
Pulse shaper	RRC
RRC roll-off	0.1
SpS (TX)	8
SpS (RX)	4
$N_{\text{spans}}$	14
$L_{\text{span}}$	50 km
Fiber loss	0.2 dB/km
CD coeff.	-21.7 ps <sup>2</sup> /km
PMD coeff.	0.05 ps/ $\sqrt{\text{km}}$
Fiber nonlinear coeff.	1.4 W <sup>-1</sup> .km <sup>-1</sup>
$G_{\text{EDFA}}$	20 dB
$NF_{\text{EDFA}}$	4.5 dB
LPF cutoff freq.	20 GHz
SSFM resolution	1 km/step

TABLE III  
NUMBER OF BATCHES AND SYMBOLS USED IN TRAINING AND VALIDATION.

	Training	Validation
Batches ( $N_B$ )	392	392
Batchsize (B)	4096	4096
Batches $N_B$ /epoch	192	192

symbol ( $S = 2N + 1$ ) and 4 for the real and imaginary parts of each polarization. Under this configuration, the number of neurons of the input layer  $N_i = 4S = 4(2N + 1)$ . After this distribution, the training batches of each dataset were concatenated producing a final training dataset in the form  $(N_B, B, S, 4)$ , where  $N_B$  is the number of batches. An identical procedure was done with the corresponding parts for the validation dataset. Table III details the total number of batches and symbols used during the training.

However, due to limitations in computational resources,



TABLE IV  
ARCHITECTURE OF THE NN-BASED NONLINEAR EQUALIZER.

NN architecture	Activation function $f_a$
86 646  $f_a$  319  $f_a$  365  $f_a$  4	Tanh

we could not use the 392 batches. Therefore, we randomly selected 192 batches from 392 batches at each epoch.

For testing purposes, we used a different unseen dataset consisting of 633,177 effective symbols.

C. NN hyperparameters and Training Process

In this study, we selected an MLP architecture with carefully adjusted hyperparameters to get the lowest MSE values during training and to improve the BER as much as possible in the validation dataset. We utilized the Optuna framework for hyperparameter optimization [59] with 50 trials (candidates), setting the following ranges: the number of taps N should range from 10 to 20 symbols, the number of hidden layers should range from 1 to 3, and the number of hidden units should range from 15 to 1000.

We concur with the viewpoint outlined in [27] regarding the necessity for careful consideration of several crucial factors during NN training. In particular, the following aspects should be taken into account:

- Considering large enough datasets for highly accurate systems, as is the case of optical transport networks.
- Employing distinct data generation patterns for training, validation, and testing datasets, or using cross-fold validation. We opted for cross-fold validation using the 14 datasets.
- Utilizing large mini-batch sizes, with the mini-batch size being as large as possible to ensure it is representative of the entire dataset.

During the training, at each epoch, the training mini-batches were shuffled. This approach was adopted to prevent the NN from learning specific patterns (even if this was highly improbable due to the various datasets with different dataset generation patterns), ensuring a more generalized model. The learning process involved the minimization of the loss function, through an optimization step using Adam optimization, followed by the updating of NN parameters [41]. We found that a learning rate of  $10^{-4}$  and a large minibatch  $B = 4096$  yielded better BER improvements.

We found the NN hyperparameters indicated in Table IV. The numbers in the NN architecture indicate the neurons per layer, with the first and last numbers corresponding to the input and output layers, respectively. The numbers in between are the hidden units and  $f_a$  states for the activation function which is the hyperbolic tangent.

VI. RESULTS

In this section, we have compared the performance of an NN trained with MSE, an NN trained with MSE-X, an NN + STH trained with MSE, and a DBP at 1 step/span. The validation dataset was utilized to monitor the performances

TABLE V  
PARAMETER  $\sigma^2$  UTILIZED IN NNS WITH MSE-X AND NN+STH WITH MSE.

P/ch (dBm)	-4	-2	0	2
$\sigma^2$ (MSE-X)	0.0027	0.0027	0.0027	0.005
$\sigma^2$ initial (STH)	0.025	0.025	0.025	0.03

during the learning process. Finally, the testing dataset was used to calculate accuracy metrics on an unseen dataset.

We selected the model that provides the lowest MSE in the validation dataset for each case. We monitored the MSE in the training and validation datasets and stopped the training process if no MSE improvement was observed in the validation dataset or if we observed overfitting.

The MSE-X loss function requires to set a parameter  $\sigma^2$  related to the variance of the equalized signal. We found this parameter difficult to set as the expected quality of the equalized signal is unknown. Despite that, guided by the details provided in [26], [60], we performed the following steps:

- We fixed  $\sigma^2$  and we trained the NN for a period.
- When we observed a training stabilization or overfitting, we stopped the training and calculated  $\sigma^2$  using a testing dataset.
- We updated  $\sigma^2$  with the recovered value (which corresponded to an equalized signal). The final training of the NN was then performed using this adjusted  $\sigma^2$ .

Table V provides the details of the parameters utilized for  $\sigma^2$  for both the MSE-X and the STH function.

In all configurations, we chose the model that provided the lowest loss in the validation dataset. For instance, for P/ch = 0 dBm, we obtained the curves illustrated in Fig. 7. As we can observe in this example, the NN using MSE shows a gradual loss descent up to eventually occurring overfitting. On the other hand, the NN with MSE-X shows lower training losses compared to the MSE case. This is due to the entropy regularization term added to the MSE, as explained in [26]. The NN + STH using MSE was early stopped owing to the blocking backpropagation. For MSE-X as well as for NN + STH with MSE, the validation losses are calculated using the MSE between the transmitted and equalized symbols.

In Fig. 8 we plot the input constellations denoted previously as  $R$ , followed by the equalized signals  $Y = f(R)$  using the NN with MSE, the NN with MSE-X, and NN + STH with MSE. Notice that both, the NN with MSE-X and NN + STH with MSE, avoid the MMSE-scatterplot. The last column corresponds to the equalized signal using the DBP 1 step/span.

For each launch power and each equalizer, we calculated the BER and the MI-LB using (14). The results are depicted in Fig. 9a and 9b.

In terms of BER, all the NNs performed worse than the DBP 1 step/span. However, the reader should take into account the performance-computational complexity trade-off between the DBP and the NN. This comparison is out of the scope of the present work but the reader could refer, for instance, to [15], [22]. The NN (MSE) is slightly better in the linear regime, at P/ch = -4 and -2 dBm, than the NN (MSE-X) and the NN

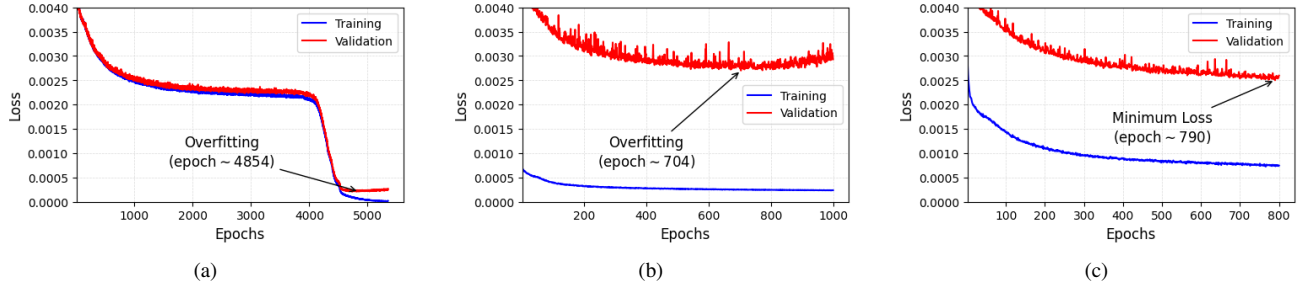


Fig. 7. Curve of losses (average of training batches per epoch) for each equalizer for  $P/\text{ch} = 0 \text{ dBm}$ . (a) NN using MSE, (b) NN using MSE-X, (c) NN + STH using MSE.

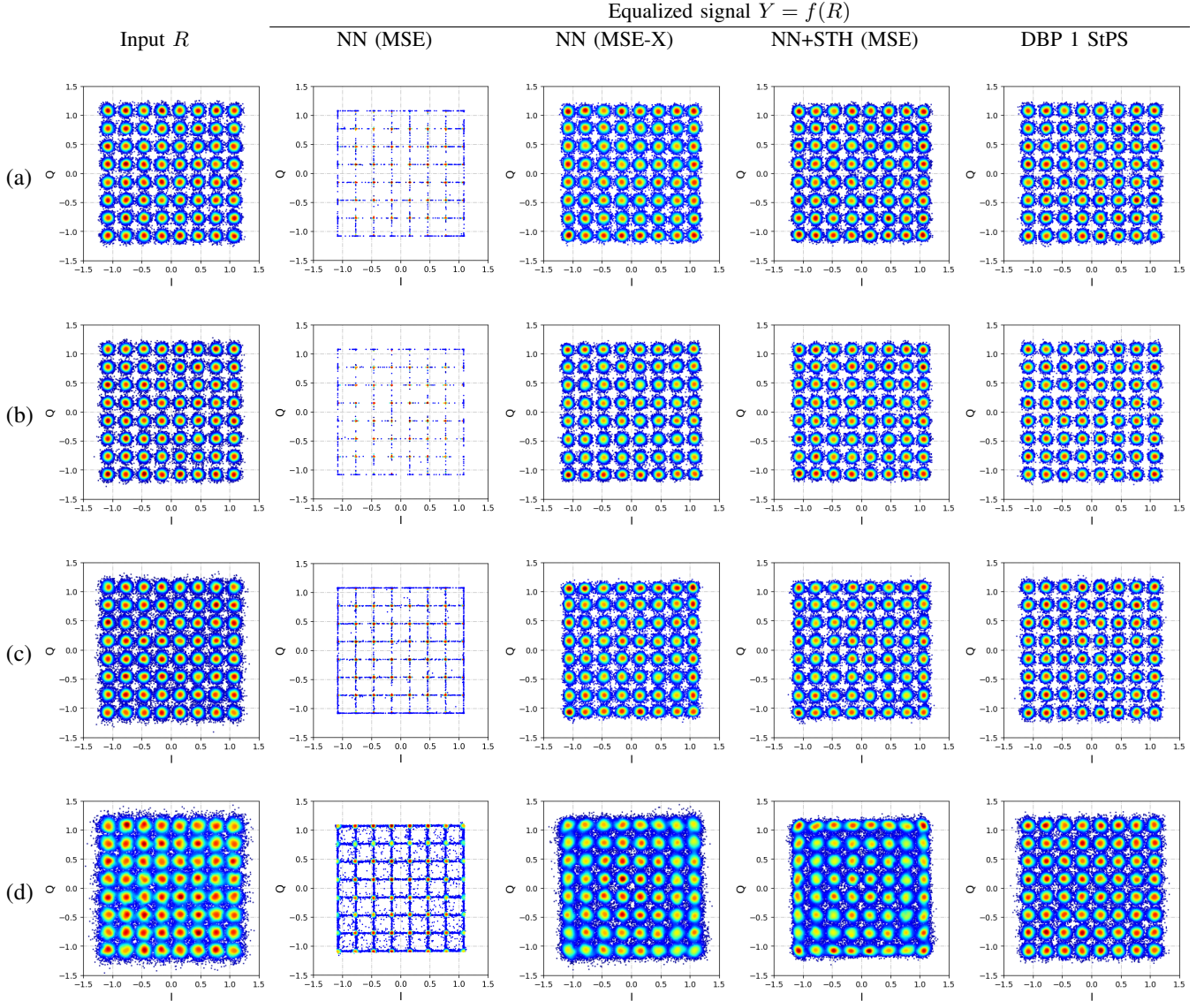
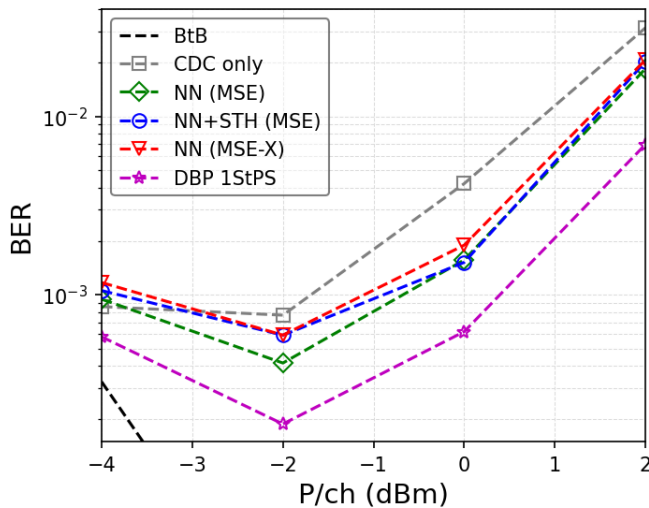


Fig. 8. Symbol constellation diagrams of the received signal (NN input)  $R$  and constellations of equalized signal  $Y = f(R)$  for each NN-based equalizer and using a DBP 1 StPS. (a)  $P/\text{ch} = -4\text{dBm}$ , (b)  $P/\text{ch} = -2\text{dBm}$ , (c)  $P/\text{ch} = 0\text{dBm}$ , (d)  $P/\text{ch} = 2\text{dBm}$ .

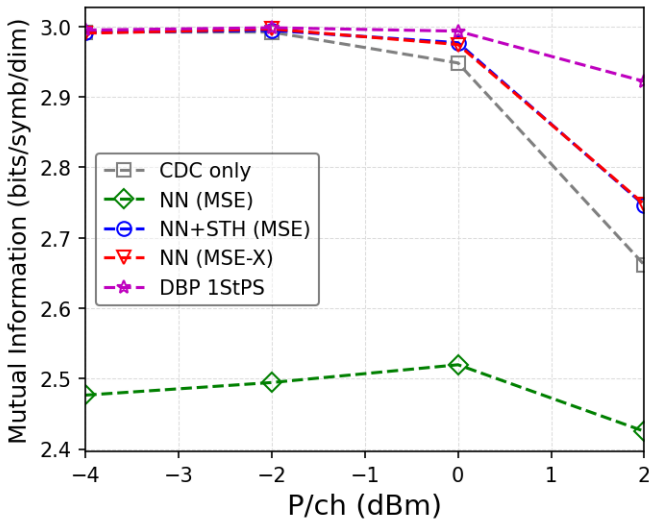
+ STH (MSE), which both show similar performances, with the NN+STH (MSE) slightly outperforming the NN (MSE-X). In the linear regime, both methods hardly improve the CDC or even worsen it. This was not due to the methods themselves but because training an NN in the linear regime

is very challenging, requiring a large amount of data and computational power.

Regarding the MI  $I(X; Y)$ , the NN (MSE) method results in a loss of soft information, showing a very poor MI. On the other hand, all the tested NNs increased the MI, but the DBP



(a) BER vs. P/ch



(b) MI-LB  $I(X; Y)$  vs. P/ch

Fig. 9. Performances obtained for each equalizer.

method was always superior.

### VII. CONCLUSIONS

In this study, our goal was to offer more insights about the MMSE-scatterplot phenomenon that occurs when using nonlinear equalizers based on NNs.

Firstly, we explained the fundamental origin of the MMSE-scatterplot and presented its equivalent mathematical expression, which is the Soft Thresholding (STH) function.

Secondly, we used the STH function as an alternative to avoid the MMSE-scatterplot. The STH function is placed after the NN during the training. In the evaluation, the equalized signal, free of the MMSE-scatterplot, is obtained before the STH function. A comparison between the NN+STH (using MSE) and the NN (using MSE-X), showed slightly better BER using the NN+STH (MSE) than using the NN (MSE-X) with similar MI. The NN+STH approach requires initializing the parameter  $\sigma^2$  with sufficiently small values capable of

generating the MMSE scatterplot. The MSE-X also requires setting a parameter  $\sigma^2$ , however it must be carefully set up. If is too small, the training loss could become negative.

Finally, other strategies could also be explored. For instance a kurtosis-based regularization term, or different hyperparameters optimization strategies with different objectives, for instance, the BER minimization or MI maximization, both requiring more in-depth justifications.

### REFERENCES

- [1] Z. Xiang and G. Bi, "Complex neuron model with its applications to m-qam data communications in the presence of co-channel interferences," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1992, pp. 305–308 vol. 2.
- [2] G. Kechriotis, E. Zervas, and E. Manolakis, "Using recurrent neural networks for adaptive communication channel equalization," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 267–278, 1994.
- [3] W. R. Kirkland and D. Taylor, "Neural network channel equalization," in *Neural networks in telecommunications*. Springer, 1994, pp. 143–171.
- [4] P. Baldi and R. Vershynin, "The capacity of feedforward neural networks," *Neural Networks*, vol. 116, pp. 288–311, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893608019301078>
- [5] M. Secondini, "Chapter 20 - information capacity of optical channels," in *Optical Fiber Telecommunications VII*, A. E. Willner, Ed. Academic Press, 2020, pp. 867–920. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780128165027000233>
- [6] S. Deligiannidis, A. Bogris, C. Mesaritakis, and Y. Kopsinis, "Compensation of fiber nonlinearities in digital coherent systems leveraging long short-term memory neural networks," *Journal of Lightwave Technology*, vol. 38, no. 21, pp. 5991–5999, 2020.
- [7] 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," *Nature communications*, vol. 10, no. 1, p. 3033, 2019.
- [8] V. Lauinger, F. Buchali, and L. Schmalen, "Blind equalization and channel estimation in coherent optical communications using variational autoencoders," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2529–2539, 2022.
- [9] O. Sidelnikov, A. Redyuk, and S. Sygletos, "Equalization performance and complexity analysis of dynamic deep neural networks in long haul transmission systems," *Opt. Express*, vol. 26, no. 25, pp. 32765–32776, Dec 2018. [Online]. Available: <https://opg.optica.org/oe/abstract.cfm?URI=oe-26-25-32765>
- [10] C. Catanese, R. Ayassi, E. Pincemin, and Y. Jaouën, "A fully connected neural network approach to mitigate fiber nonlinear effects in 200g dp-16-qam transmission system," in *2020 22nd International Conference on Transparent Optical Networks (ICTON)*, 2020, pp. 1–4.
- [11] O. Sidelnikov, A. Redyuk, S. Sygletos, M. Fedoruk, and S. Turitsyn, "Advanced convolutional neural networks for nonlinearity mitigation in long-haul wdm transmission systems," *Journal of Lightwave Technology*, vol. 39, no. 8, pp. 2397–2406, 2021.
- [12] C. Li, Y. Wang, J. Wang, H. Yao, X. Liu, R. Gao, L. Yang, H. Xu, Q. Zhang, P. Ma, and X. Xin, "Convolutional neural network-aided dp-64 qam coherent optical communication systems," *Journal of Lightwave Technology*, vol. 40, no. 9, pp. 2880–2889, 2022.
- [13] X. Liu, Y. Wang, X. Wang, H. Xu, C. Li, and X. Xin, "Bi-directional gated recurrent unit neural network based nonlinear equalizer for coherent optical communication system," *Opt. Express*, vol. 29, no. 4, pp. 5923–5933, Feb 2021. [Online]. Available: <https://opg.optica.org/oe/abstract.cfm?URI=oe-29-4-5923>
- [14] A. Shahkarami, M. I. Yousefi, and Y. Jaouen, "Efficient deep learning of nonlinear fiber-optic communications using a convolutional recurrent neural network," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021, pp. 668–673.
- [15] P. J. Freire, Y. Osadchuk, B. Spinnler, A. Napoli, W. Schairer, N. Costa, J. E. Prilepsky, and S. K. Turitsyn, "Performance versus complexity study of neural network equalizers in coherent optical systems," *Journal of Lightwave Technology*, vol. 39, no. 19, pp. 6085–6096, 2021.
- [16] M. Schaedler, C. Bluemm, M. Kuschnerov, F. Pittalà, S. Calabrò, and S. Pachnicke, "Deep neural network equalization for optical short reach communication," *Applied Sciences*, vol. 9, no. 21, 2019. [Online]. Available: <https://www.mdpi.com/2076-3417/9/21/4675>

- [17] 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.*, vol. 40, no. 3, pp. 597–606, Feb 2022. [Online]. Available: <https://opg.optica.org/jlt/abstract.cfm?URI=jlt-40-3-597>
- [18] P. J. Freire, V. Neskornuik, A. Napoli, B. Spinnler, N. Costa, G. Khanna, E. Riccardi, J. E. Prilepsky, and S. K. Turitsyn, "Complex-valued neural network design for mitigation of signal distortions in optical links," *Journal of Lightwave Technology*, vol. 39, no. 6, pp. 1696–1705, 2021.
- [19] C. Häger and H. D. Pfister, "Physics-based deep learning for fiber-optic communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 280–294, 2021.
- [20] Q. Fan, C. Lu, and A. P. T. Lau, "Combined neural network and adaptive dsp training for long-haul optical communications," *Journal of Lightwave Technology*, vol. 39, no. 22, pp. 7083–7091, 2021.
- [21] D. Tang, Z. Wu, Z. Sun, X. Tang, and Y. Qiao, "Joint intra and inter-channel nonlinearity compensation based on interpretable neural network for long-haul coherent systems," *Opt. Express*, vol. 29, no. 22, pp. 36242–36256, Oct 2021. [Online]. Available: <https://opg.optica.org/oe/abstract.cfm?URI=oe-29-22-36242>
- [22] A. Sotomayor, E. Pincemin, V. Choqueuse, and M. Morvan, "A comparison of machine learning techniques for fiber non-linearity compensation: Multilayer perceptron vs. learned digital backpropagation," in *2023 23rd International Conference on Transparent Optical Networks (ICTON)*, 2023, pp. 1–4.
- [23] M. Schädler, G. Böcherer, and S. Pachnicke, "Soft-demapping for short reach optical communication: A comparison of deep neural networks and volterra series," *J. Lightwave Technol.*, vol. 39, no. 10, pp. 3095–3105, May 2021. [Online]. Available: <https://opg.optica.org/jlt/abstract.cfm?URI=jlt-39-10-3095>
- [24] G. Böcherer. Lecture notes on machine learning for communications. [Online]. Available: <http://georg-boecherer.de/mlcomm>
- [25] P. J. Freire, J. E. Prilepsky, Y. Osadchuk, S. K. Turitsyn, and V. Aref, "Deep neural network-aided soft-demapping in coherent optical systems: Regression versus classification," *IEEE Transactions on Communications*, vol. 70, no. 12, pp. 7973–7988, 2022.
- [26] F. Diedolo, G. Böcherer, M. Schädler, and S. Calabró, "Nonlinear equalization for optical communications based on entropy-regularized mean square error," in *European Conference on Optical Communication (ECOC) 2022*. Optica Publishing Group, 2022, p. We2C.2. [Online]. Available: <https://opg.optica.org/abstract.cfm?URI=ECEOC-2022-We2C.2>
- [27] 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 Journal of Selected Topics in Quantum Electronics*, vol. 28, no. 4: Mach. Learn. in Photon. Commun. and Meas. Syst., pp. 1–23, 2022.
- [28] C. You and D. Hong, "Adaptive equalization using the complex back-propagation algorithm," in *Proceedings of International Conference on Neural Networks (ICNN'96)*, vol. 4, 1996, pp. 2136–2141 vol.4.
- [29] J. Patra, R. Pal, R. Baliarsingh, and G. Panda, "Nonlinear channel equalization for qam signal constellation using artificial neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 2, pp. 262–271, 1999.
- [30] M. A. Jarajreh, E. Giacomidis, I. Aldaya, S. T. Le, A. Tsokanos, Z. Ghassemlooy, and N. J. Doran, "Artificial neural network nonlinear equalizer for coherent optical ofdm," *IEEE Photonics Technology Letters*, vol. 27, no. 4, pp. 387–390, 2015.
- [31] S. Liu, P.-C. Peng, C.-W. Hsu, S. Chen, H. Tian, and G.-K. Chang, "An effective artificial neural network equalizer with s-shape activation function for high-speed 16-qam transmissions using low-cost directly modulated laser," in *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, 2018, pp. 269–273.
- [32] A. Sotomayor, E. Pincemin, V. Choqueuse, and M. Morvan, "Optimized cost function of multi-layer perceptron for fibre non-linear impairment mitigation in coherent 200-gbps dp-16qam transmission system," in *Optica Advanced Photonics Congress 2022*. Optica Publishing Group, 2022, p. JTu2A.41. [Online]. Available: <https://opg.optica.org/abstract.cfm?URI=SPPCom-2022-JTu2A.41>
- [33] A. Alvarado, E. Agrell, D. Lavery, R. Maher, and P. Bayvel, "Replacing the soft-decision fec limit paradigm in the design of optical communication systems," *Journal of Lightwave Technology*, vol. 33, no. 20, pp. 4338–4352, 2015.
- [34] P. Poggiolini, "The gn model of non-linear propagation in uncompensated coherent optical systems," *Journal of Lightwave Technology*, vol. 30, no. 24, pp. 3857–3879, 2012.
- [35] A. Alvarado, "Information rates and post-fec ber prediction in optical fiber communications," in *Optical Fiber Communication Conference*. Optica Publishing Group, 2017, p. Th3F.3. [Online]. Available: <https://opg.optica.org/abstract.cfm?URI=OFC-2017-Th3F.3>
- [36] P. P. Mitra and J. B. Stark, "Nonlinear limits to the information capacity of optical fibre communications," *Nature*, vol. 411, no. 6841, pp. 1027–1030, 2001.
- [37] A. Alvarado, T. Fehenberger, B. Chen, and F. M. J. Willems, "Achievable information rates for fiber optics: Applications and computations," *J. Lightwave Technol.*, vol. 36, no. 2, pp. 424–439, Jan 2018. [Online]. Available: <https://opg.optica.org/jlt/abstract.cfm?URI=jlt-36-2-424>
- [38] T. Fehenberger, A. Alvarado, P. Bayvel, and N. Hanik, "On achievable rates for long-haul fiber-optic communications," *Opt. Express*, vol. 23, no. 7, pp. 9183–9191, Apr 2015. [Online]. Available: <https://opg.optica.org/oe/abstract.cfm?URI=oe-23-7-9183>
- [39] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural Networks*, vol. 6, no. 6, pp. 861–867, 1993. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893608005801315>
- [40] J. G. Proakis, *Digital communications*. McGraw-Hill, Higher Education, 2008.
- [41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [42] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.
- [43] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.
- [44] D. Taylor, "The estimate feedback equalizer: A suboptimum nonlinear receiver," *IEEE Transactions on Communications*, vol. 21, no. 9, pp. 979–990, 1973.
- [45] M. Khani, M. Alizadeh, J. Hoydis, and P. Fleming, "Adaptive neural signal detection for massive mimo," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5635–5648, 2020.
- [46] W. R. Kirkland, "On the application of multi-layered perceptrons to nonlinear equalization for frequency selective fading channels and nonlinear prediction for time selective rayleigh fading channels," Ph.D. dissertation, McMaster University, 1994.
- [47] C. You and D. Hong, "Nonlinear blind equalization schemes using complex-valued multilayer feedforward neural networks," *IEEE Transactions on Neural Networks*, vol. 9, no. 6, pp. 1442–1455, 1998.
- [48] 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.
- [49] A. Mansour and C. Jutten, "What should we say about the kurtosis?" *IEEE Signal Processing Letters*, vol. 6, no. 12, pp. 321–322, 1999.
- [50] H. Mathis, "On the kurtosis of digitally modulated signals with timing offsets," in *2001 IEEE Third Workshop on Signal Processing Advances in Wireless Communications (SPAWC'01). Workshop Proceedings (Cat. No.01EX471)*, 2001, pp. 86–89.
- [51] D. Marcuse, C. Manyuk, and P. Wai, "Application of the manakov-pmd equation to studies of signal propagation in optical fibers with randomly varying birefringence," *Journal of Lightwave Technology*, vol. 15, no. 9, pp. 1735–1746, 1997.
- [52] E. Ip, "Nonlinear compensation using backpropagation for polarization-multiplexed transmission," *Journal of Lightwave Technology*, vol. 28, no. 6, pp. 939–951, 2010.
- [53] G. P. Agrawal, "Nonlinear fiber optics," in *Nonlinear Science at the Dawn of the 21st Century*. Springer, 2000, pp. 195–211.
- [54] S. J. Savory, "Digital coherent optical receivers: Algorithms and subsystems," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 16, no. 5, pp. 1164–1179, 2010.
- [55] F. P. Guiomar, S. B. Amado, A. Carena, G. Bosco, A. Nespola, A. L. Teixeira, and A. N. Pinto, "Fully blind linear and nonlinear equalization for 100g pm-64qam optical systems," *Journal of Lightwave Technology*, vol. 33, no. 7, pp. 1265–1274, 2015.
- [56] E. Ip and J. M. Kahn, "Digital equalization of chromatic dispersion and polarization mode dispersion," *Journal of Lightwave Technology*, vol. 25, no. 8, pp. 2033–2043, 2007.
- [57] M. Selmi, Y. Jaouen, and P. Ciblat, "Accurate digital frequency offset estimator for coherent polmux qam transmission systems," in *2009 35th European Conference on Optical Communication*, 2009, pp. 1–2.
- [58] T. Pfau, S. Hoffmann, and R. Noe, "Hardware-efficient coherent digital receiver concept with feedforward carrier recovery for  $m$ -qam constellations," *Journal of Lightwave Technology*, vol. 27, no. 8, pp. 989–999, 2009.

- [59] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
- [60] F. Diedolo. Nonlinear equalization for optical communications based on entropy-regularized mean square error. [Online]. Available: [http://www.georg-boecherer.de/diedolo2022nonlinear\\_slides.pdf](http://www.georg-boecherer.de/diedolo2022nonlinear_slides.pdf)
- [61] T. Fehenberger and N. Hanik, "Mutual Information as a Figure of Merit for Optical Fiber Systems," *arXiv e-prints*, p. arXiv:1405.2029, Apr. 2014.
- [62] L. Hanzo, S. X. Ng, W. Webb, and T. Keller, *Quadrature amplitude modulation: From basics to adaptive trellis-coded, turbo-equalised and space-time coded OFDM, CDMA and MC-CDMA systems*. IEEE Press-John Wiley, 2004.
- [63] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [64] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.
- [65] B. Mukherjee, I. Tomkos, M. Tornatore, P. Winzer, and Y. Zhao, *Springer handbook of optical networks*. Springer, 2020.

## APPENDIX A INFORMATION THEORY

Let  $X$  be the transmitted data from an information source, belonging to a discrete alphabet  $\mathcal{X}$  and with probability mass function  $p_X(x)$ , and let  $Y$  be the received data with probability density function  $f_Y(y)$ <sup>5</sup>. The entropy of  $X$  measures the amount of information generated by the source [63] in bits per symbol or bits per second. It is also interpreted as the amount of information (bits) needed to describe  $X$  [64]. The entropy of  $X$ ,  $H(X)$ , is calculated as follows,

$$\begin{aligned} H(X) &= \mathbb{E}[-\log_2 p_X(x)] \\ &= - \sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x) \\ &\approx - \frac{1}{K} \sum_{i=1}^K \log_2 p_X(x_i) \end{aligned} \quad (10)$$

where  $\mathbb{E}$  is the real expectation and the third equation is the empirical expectation for  $K$  samples from  $p_X(x)$  [24].

The entropy of  $Y$  (differential entropy for the continuous case) is denoted as  $h(Y)$  and is calculated as follows,

$$\begin{aligned} h(Y) &= \mathbb{E}[-\log_2 f_Y(y)] \\ &= - \int_{y \in \mathcal{Y}} f_Y(y) \log_2 f_Y(y) dy \\ &\approx - \frac{1}{K} \sum_{i=1}^K \log_2 f_Y(y_i) \end{aligned} \quad (11)$$

Nevertheless, the meaning of the differential entropy is different from the entropy of the discrete case, as it does not represent an amount of information to describe a random variable. Indeed, the differential entropy could even be negative [63]. The meaning of the differential entropy is related to the log-scale of the smaller set that contains most of the probability [64], meaning a low entropy (more negative) a

more confined set, and high entropies (less negative) a more dispersed set.

The conditional entropy of  $X$  knowing  $Y$  (also called equivocation),  $H(X|Y)$ , measures the uncertainty of  $X$  by the knowledge of  $Y$ .

$$\begin{aligned} H(X|Y) &= \mathbb{E}[-\log_2 p_{X|Y}(x|y)] \\ &\approx - \frac{1}{K} \sum_{i=1}^K \log_2 p_{X|Y}(x_i|y_i) \end{aligned} \quad (12)$$

Similarly, the equivocation of  $Y$  given  $X$  is defined as,

$$\begin{aligned} h(Y|X) &= \mathbb{E}[-\log_2 f_{Y|X}(y|x)] \\ &\approx - \frac{1}{K} \sum_{i=1}^K \log_2 f_{Y|X}(y_i|x_i) \end{aligned} \quad (13)$$

The reduction in uncertainty of  $X$  due to the knowledge of  $Y$  is the MI,  $I(X; Y)$  which is defined as [63], [64],

$$I(X; Y) = H(X) - H(X|Y) = h(Y) - h(Y|X) \quad (14)$$

where the first equality means the amount of information sent less the uncertainty of this information regarding the received data. The second equality is obtained because of the symmetry and means the amount of information received less the uncertainty corresponding to the noise of the channel [63], [64].

The channel capacity is  $\max I(X; Y)$ . In an AWGN channel, this definition leads to the maximization of the Signal-to-Noise ratio (SNR) [63]. Moreover, it is possible to estimate the BER (Q-factor) given the SNR [65], e.g. for a square M-QAM:

$$\begin{aligned} \text{BER} &= \frac{2}{\log_2 M} \left( 1 - \frac{1}{\sqrt{M}} \right) \left( \sqrt{\frac{3}{2(M-1)} \text{SNR}} \right) \\ \text{Q-factor} &= 20 \log \left[ \sqrt{2} \text{erfc}^{-1}(2\text{BER}) \right] \quad (\text{dB}) \end{aligned} \quad (15)$$

<sup>5</sup>Here we considered  $Y$  continuous, though it could be also considered discrete as in [61], [62] by means of an ADC.