Blind Polarisation Demultiplexing and Carrier Recovery Using FIR-based Variational AutoEncoder Equaliser for Probabilistic Constellation Shaping in Optical Fibre Communications

Louis Tomczyk, Élie Awwad, Cédric Ware

***Abstract*—We investigate through simulations the potential of Finite Impulse Response (FIR)-based Variational AutoEncoder-inspired (VAE-FIR) equaliser for polarisation demultiplexing, Carrier Phase Recovery (CPR), and Carrier Frequency Offset (CFO) estimation in the context of Probabilistic Constellation Shaped (PCS) transmissions in coherent optical fibre communication systems. Additionally, we compare the performance of this novel estimator with the conventional Constant Modulus Algorithm (CMA) and Pilot-Aided Carrier Phase Recovery (PA-CPR). Our study shows that the VAE-FIR clearly outperforms the conventional approach in terms of polarisation demultiplexing, even with PCS where the CMA fails. We also show the ability of the VAE-FIR to track the phase evolution. Its ability to compensate for the carrier's phase effects is however limited to linewidths of a few dozen kHz and a hundred kHz for CFO, showing that the VAE-FIR may be used to compensate for the small residual phase noise or residual frequency mismatch.**

***Index Terms*—Polarisation demultiplexing, Carrier Phase Recovery, Constant Modulus Algorithm, Variational AutoEncoder, Probabilistic Constellation Shaping**

I. INTRODUCTION

PROBABILISTIC Constellation Shaping (PCS) is increasingly being employed in coherent optical fibre transmission systems due to its ability to approach Shannon's capacity [1], [2]. However, as predicted by [3] and demonstrated by the authors in [4]–[6], the use of PCS may render the traditional and widely industrialised algorithm Constant Modulus Algorithm (CMA) less effective for polarisation demultiplexing.

Minimum Mean Square Error (MMSE) equalisers can address the CMA's limitations, but their complexity—especially due to matrix inversion—can be prohibitive in real-time implementations [7], [8]. Frequency-domain MMSE variants reduce this cost, yet still remain heavier than CMA. Decision-directed schemes like DD-LMS suffer from error propagation in low-SNR regimes and are often combined with CMA, increasing DSP complexity. Multi-Modulus Algorithms (MMA) and their variants further raise the computational load [9], [10].

One possible low-complexity digital equalisation alternative to overcome this limitation is the use of pilot-aided algorithms for both polarisation demultiplexing and Carrier Phase Recovery (CPR) [11]–[14].

Regarding phase recovery, blind techniques such as Blind Phase Search (BPS) exist [15]. However, BPS is computationally expensive [16] and its performance degrades under probabilistic shaping [17]. Pilot-tone-based methods can also be used [18], [19], but they require analogue tone generation, accurate extraction, and spectral planning to avoid interference. Due to these limitations, we adopt, as a benchmark, Pilot-Aided Carrier Phase Recovery (PA-CPR), which provides a robust and near-ideal phase reference using embedded pilot symbols.

While effective, these pilot-aided schemes reduce the net transmission rate. This has motivated the exploration of more efficient Digital Signal Processing (DSP) strategies. In [20], the use of Variational AutoEncoders (VAEs)—originally designed for generative tasks—was proposed for equalisation in data transmission systems. Later, [4] applied the VAE cost function to PCS-QAM formats in coherent optical systems, aiming to preserve a classical DSP structure with adaptive filters. Their results highlighted the VAE's ability, and that of its FIR-based variant (VAE-FIR), to track the State of Polarisation (SoP) in QAM systems. Under uniform-shaped QAM and linear SoP drift conditions, the VAE outperformed CMA in specific scenarios. They also noted that the VAE operates directly on complex-valued signals, potentially enabling phase estimation and recovery. Building on this, we further explored in [6] the VAE-FIR's capacity to track polarisation under PCS-64QAM and various SoP evolution models, without accounting for laser phase noise.

This paper aims to extend our previous work by investigating through simulations the VAE-FIR's capability for tracking of carrier phase evolution and SoP changes in more stringent channel conditions like higher dynamic ranges of variations. Indeed, in this work, we pushed further these scenarios with a wider range of values for the speed of polarisation changes, from $250\ \mathrm{krad/s}$ to $5\ \mathrm{Mrad/s}$. We compare its performance to traditional DSP approaches, including CMA and Pilot-Aided CPR (PA-CPR), referred to collectively as CMA+PA-CPR. The main results are the better polarisation demultiplexing using the VAE-FIR compared to the conventional approach for both equalisation and estimation of the SoP. We also show that the carrier frequency offset (CFO) and CPR can be done using the VAE-FIR up to a few dozens of kilohertz. In section II, we define the used notation. In section III, we provide the optical

---

This work has received funding from BPI France (grant 0168463/00) within the CELTIC-NEXT European project AI-NET Antillas.

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685

channel and transceiver effects taken into account in our system model. In section IV, we present the traditional digital signal processing algorithms for polarisation demultiplexing as well as for the phase recovery. In section V, we describe the use of variational autoencoders both traditionally in artificial intelligence and in the context of telecommunications. In section VI, we tune the parameters of the equalisers to get the best possible performance. In section VII, we present the experimental setups. In section VIII, we present the results. We also provide the source code for our simulator available online [21].

II. NOTATIONS

Here, we define our notation conventions. We denote the estimated value of $\bullet$ as $\widehat{\bullet}$, the complex conjugate of $\bullet$ as $\bullet^*$ and its argument $\arg(\bullet)$ and $\dot{\bullet} = d\bullet/dt$ is the time derivative of $\bullet$. We represent $M_{\bullet_1,\bullet_2}(\mathbb{K})$ as the set of matrices of size $\bullet_1 \times \bullet_2$, defined over the real or complex field $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. The transpose of a matrix $\bullet$ is denoted as $\bullet^T$ and its determinant $\det(\bullet)$.

In dual-polarisation (DP) transmissions, the complex envelope of the optical signal has two complex-valued dimensions, corresponding to the polarisation tributaries $\{H, V\}$, and are sampled at a frequency f_s. When both polarisations are considered, the quantity is written in bold with roman style. In addition, a vector is denoted by an arrow. If we consider n time samples, a generic signal can be represented as a matrix:

$$\mathbf{a} = \begin{pmatrix}\vec{a}_H \\ \vec{a}_V\end{pmatrix} \in M_{2,n}(\mathbb{C}), \underset{\text{“with”}}{\setminus} \vec{a}_\bullet = \left(a_{\bullet,1} \cdots a_{\bullet,n}\right) \in M_{1,n}(\mathbb{C}), \quad (1)$$

where $\bullet \in \{H, V\}$. We also write the square norm: $||\vec{a}||^2 = \sum |a_i|^2$. For operators representing a linear optical fibre channel with memory, we define:

$$\mathbf{h} = \begin{pmatrix}\vec{h}_{HH} & \vec{h}_{HV} \\ \vec{h}_{VH} & \vec{h}_{VV}\end{pmatrix} \in M_{2,2n}(\mathbb{C}). \quad (2)$$

To model the linear transmission of $\mathbf{a}$ through an optical fibre modelled by $\mathbf{h}$, we define the “matrix convolution” $\circledast$ as:

$$\mathbf{h} \circledast \mathbf{a} = \begin{pmatrix}\vec{h}_{HH} * \vec{a}_H + \vec{h}_{HV} * \vec{a}_V \\ \vec{h}_{VH} * \vec{a}_H + \vec{h}_{VV} * \vec{a}_V\end{pmatrix}. \quad (3)$$

where $\bullet_1 * \bullet_2$ is the convolution between $\bullet_1$ and $\bullet_2$. We use the subscripts $\{I, Q\}$ to refer respectively to the {Real, Imaginary} part of the quantity $\bullet$: $\Re/\Im(\bullet) = \bullet_{I/Q}$. The natural logarithm is denoted by $\ln$. The gradient operator is expressed as $\overrightarrow{\text{grad}}\,(\bullet)$ and the interval of integers between $\bullet_1$ and $\bullet_2$ (both included) is written $\llbracket \bullet_1, \bullet_2 \rrbracket$.

For probability-related equations, we make the following distinctions that we believe, are of utmost importance. Thus, we distinguish a random variable (r.v.) X, from the space over which it is defined $X(\Omega)$, and its realisation $(X = x)$. The probability density function (PDF) is written as $x \mapsto p_X(x)$ and $\mathbb{P}(\bullet)$ is the discrete probability of the event $\bullet$ to happen. The law of the r.v. X given Y is written as $X|Y$, the joint PDF of (X, Y) is written as $p_{X,Y}$, and “follows the probability law” is denoted as “$\hookrightarrow$”. To differentiate deterministic functions from stochastic processes, we write the variable as a subscript for the latter. For example, $t \mapsto \theta(t)$ or θ_t: in the first case, θ is predefined, while in the second, it is random. The expectation is either $\mathbb{E}(\bullet)$ or $\langle \bullet \rangle$. Unless otherwise specified, the arithmetic mean is used as the expectation estimator. Since the used code is provided and its potential applications may extend beyond the simulations presented in this work, we may explicitly detail certain notations to facilitate its use and handling.

III. SYSTEM MODEL

A. Optical channel effects

We consider a DP linear transmission operated with a single wavelength with linearly polarised input optical signal. Additive White Circular Gaussian $-\mathcal{CN}-$ noise (AWGCN) (we may shorten in the following simply by “AWGN”) is considered, as well as laser phase noise and channel change. The latter is modelled using the “polarisation linewidth” model of [22]. However, here, the ellipticity of the SoP remains constant and only θ , the parameter of the real-valued Jones rotation matrix, varies according to a normal ($\mathcal{N}$) random walk. More precisely:

$$\Delta\theta \hookrightarrow \mathcal{N}\big(0, \sigma^2_{\text{pol}}[T_{\text{Sb}}]\big) \setminus \sigma^2_{\text{pol}}[T_{\text{Sb}}] = (2\pi)^2 f_{\text{pol}}/(2\pi R_s)\ \text{rad}^2 \quad (4)$$

where k is the time index, $\Delta\theta = \theta_{k+1} - \theta_k$, f_{pol} is the polarisation linewidth, R_s is the symbol rate and $\Delta\theta$ is the evolution of θ over a duration equal to the symbol time. We will see later that precising the symbol time $T_{\text{Sb}} = 1/R_{\text{s}}$ as an argument of the variance can be useful to prevent confusions. We show in Appendix X-A with Eq. (17) that, in the case of a real-valued Jones rotation matrix, the instantaneous speed of channel change v_{ch} is:

$$v_{\text{ch}} = \frac{2}{\Delta t} \arcsin\left(\frac{1}{2}\sqrt{\sum_{i=1}^{3} |\Delta S_i|^2}\right) = \frac{2\Delta\theta}{\Delta t} \quad (5)$$

where S_1, S_2, S_3 are the Stokes parameters that define a given SoP state. As we will discuss in the appendix, *here* there is no difference between the speed of “SoP variations” and the speed of the “channel variations” as in our simulations the input polarisation is linear ($S_3 = 0 \equiv \chi = 0$ in Eq. (17)). Looking back to the definition, we see that here the mean value of v_{ch} is null according to the definition of $\Delta\theta$. Thus, we would like to highlight that in reality we measure the root mean squared value of the instantaneous v_{ch}. It is induced by the inherent time integration when its computation is implemented after a polarimeter or in a coherent receiver. This is why we compute $V_{\text{ch}} = \sqrt{\langle v_{\text{ch}}^2 \rangle} = 2\sigma_{\text{pol}}/\Delta t$ instead.

Eventually, we believe it is worth specifying that Eq.(4) links the channel drift speed with its drift value. However, there may not always be a direct correlation between the twoas explained in Appendix X-A.

B. Transceiver effects

The transmitter and the receiver contribute to the total AWGCN, each with a respective variance $\sigma^2_{N,\text{tx/rx}}$. The noise samples added over each polarisation tributary are drawn from a circularly symmetric Gaussian distribution: $N_{\text{tx/rx},k} \hookrightarrow \mathcal{CN}(0, \sigma^2_{N,\text{tx/rx}})$. In coherent transmission systems, lasers are required at both the transmitter (Tx) and receiver (Rx) ends.

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685

They may exhibit different linewidths, denoted $\Delta\nu_{\mathrm{tx/rx}}$that we take both equal to $\Delta\nu/2$, leading to a polarisation-independent multiplicative phase noise. This phase noise is modelled as a Wiener process, where the k-th phase sample is:

$$\Delta\phi \hookrightarrow \mathcal{N}(0, \sigma^2_{\Delta\phi}[T_{\mathrm{Sb}}]) \setminus \sigma^2_{\Delta\phi}[T_{\mathrm{Sb}}] = (2\pi)^2 \Delta\nu/(2\pi R_s)\ \mathrm{rad}^2 \qquad (6)$$

where $\Delta\phi = \phi_k - \phi_{k-1}$. It is important to note the presence of the square over the 2π, which is uncommon in the literature. We detail in Appendix X-C why this writing ensures dimensional consistency.

IV. TRADITIONAL DIGITAL SIGNAL PROCESSING

A. Generalities

1) Signal notations: Let us denote $\mathbf{s} \in \mathcal{M}_{2,n}(\mathbb{C})$ the symbols at the Tx. These symbols are shaped with a filter $\vec{g} \in \mathcal{M}_{1,N_{\mathrm{T}}}(\mathbb{R})$, where $N_{\mathrm{T}} = 2N_{\mathrm{Sb,T}} - 1$ is the number of taps. The optical pulse-shaped signal sent in the fibre is written $\mathbf{x}$. The signal after propagation, coherent detection, and synchronisation is written $\mathbf{y}$. Finally, the signals after chromatic dispersion compensation (CDC), polarisation demultiplexing, and CPR are respectively written $\mathbf{z_{cdc}}$, $\mathbf{z_{mimo}}$, and $\mathbf{z_{cpr}}$.

2) Zero forcing approach: Both CMA and the VAE-FIR converge towards the Zero-Forcing equaliser at high SNR values as shown by the authors in [23]. This means that the equaliser is a matrix that ideally tends towards the inverse of the channel matrix as the SNR increases: $\mathbf{h_{mimo}} = \widehat{\mathbf{h}}^{-1}_{\mathbf{channel}} \underset{\mathrm{SNR}\uparrow+\infty}{\rightarrow} \mathbf{h}^{-1}_{\mathbf{channel}}$. Applying the obtained operator to the input signal should equalise it, at the cost of modifying the noise characteristics of the incoming signal [24]. In our case, the input signal into the polarisation demultiplexer equaliser, $\mathbf{z_{cdc}}$ is free from chromatic dispersion but still exhibits polarisation mixing (crosstalk). The output signal is $\mathbf{z_{mimo}} = \mathbf{h_{mimo}} \circledast \mathbf{z_{cdc}}$ where $\mathrm{mimo} \in \{\mathrm{CMA}, \mathrm{VAE\text{-}FIR}\}$ and the equaliser matrix $\mathbf{h_{mimo}}$ is:

$$\mathbf{h_{mimo}} = \begin{pmatrix} \vec{h}_{\mathrm{mimo,HH}} & \vec{h}_{\mathrm{mimo,HV}} \\ \vec{h}_{\mathrm{mimo,VH}} & \vec{h}_{\mathrm{mimo,VV}} \end{pmatrix} \setminus$$
$$\vec{h}_{\mathrm{mimo},\bullet_1\bullet_2} = \begin{pmatrix} h_{\mathrm{mimo},\bullet_1\bullet_2,-N_{\mathrm{Sb,T}}} \\ \vdots \\ h_{\mathrm{mimo},\bullet_1\bullet_2,+N_{\mathrm{Sb,T}}} \end{pmatrix}^T \qquad (7)$$

In the following, we may remove the index "mimo" for readability if not necessary.

B. MIMO equaliser

1) Construction of the equaliser: The construction of such a matrix is an optimisation problem for ***both*** CMA and the VAE-FIR that are **both** 2×2 butterfly complex-valued filters. Those filters, expressed as $\vec{h}_{\bullet_1\bullet_2}$, $(\bullet_1, \bullet_2) \in \{H, V\}^2$, are constructed step by step, whose increments are provided by the gradient of a "loss function" J_{mimo} with respect to the symbols at the equaliser input. As the equaliser is blind, after initialising the filters (usually all taps are set to zero apart from the central tap usually called "Dirac initialisation"), a training phase is necessary before using the filters to equalise the signals. In the whole paper, we would like to emphasise that it is *not* a "training" as in the AI community sense that imply the use of labelled data. The "training" phase only refers to the estimation of the matched-filter that is not explicitly implemented at the reception. To take into account the channel memory, the gradient is computed using filters with $N_{\mathrm{Sb,T}}$ taps on each side of the current sample n. Thus, the matrix coefficients are updated as:

$$\vec{h}_{\bullet_1\bullet_2} = \vec{h}_{\bullet_1\bullet_2} - lr \cdot \overrightarrow{\mathrm{grad}}\,(J_{\mathrm{mimo}}) \ \setminus$$
$$\overrightarrow{\mathrm{grad}}\,(J_{\mathrm{mimo}}) = \left(\frac{\partial J_{\mathrm{mimo}}}{\partial h_{\bullet_1\bullet_2,-N_{\mathrm{Sb,T}}}} \cdots \frac{\partial J_{\mathrm{mimo}}}{\partial h_{\bullet_1\bullet_2,+N_{\mathrm{Sb,T}}}} \right) \qquad (8)$$

where lr is the learning rate that controls the convergence speed of the algorithm.

2) State of polarisation extraction: As polarisation demultiplexing is blind, and no explicit matched filter is implemented at reception, the phase-sensitive equaliser matrix coefficients account for this matched filter, the polarisation rotation matrix, the common laser phase, and residual chromatic dispersion. Authors in [25], [26] proposed various techniques for extracting different channel parameters from the equaliser matrix, such as the rate of polarisation change (based on Stokes parameters), as well as polarisation dependent loss using the Jones formalism.

We may draw attention to the Eq. (7) where the central tap, indexed by 0 corresponds to the n-th symbol of the symbol flow. The estimated phase-sensitive channel matrix that will be applied to this n-th symbol is: $h_{\bullet_1\bullet_2,0} = e^{-i\widehat{\phi}_n} \cdot \hat{R}_{\bullet_1\bullet_2}(\theta_n) \cdot \hat{g}(nT_{\mathrm{Sb}})$, where $R_{\bullet_1\bullet_2}(\theta)$ is a real rotation matrix and $\{g(nT_{\mathrm{Sb}})\}_n$ are the samples of the root-raised cosine impulse response sampled at T_{Sb}. The tangent of the angle θ is obtained by [26]: $\tan(\widehat{\theta}) = -h_{\mathrm{HV},0}/h_{\mathrm{HH},0}$.

Additionally, it is reported in [27] that the probability of convergence to a degenerate solution is low in systems without polarisation mode dispersion (PMD). This condition applies to our study, as PMD was not included in the channel model. As a result, singularity unlikely occurred in our simulations, although no explicit tracking was performed. This point could be addressed in future work.

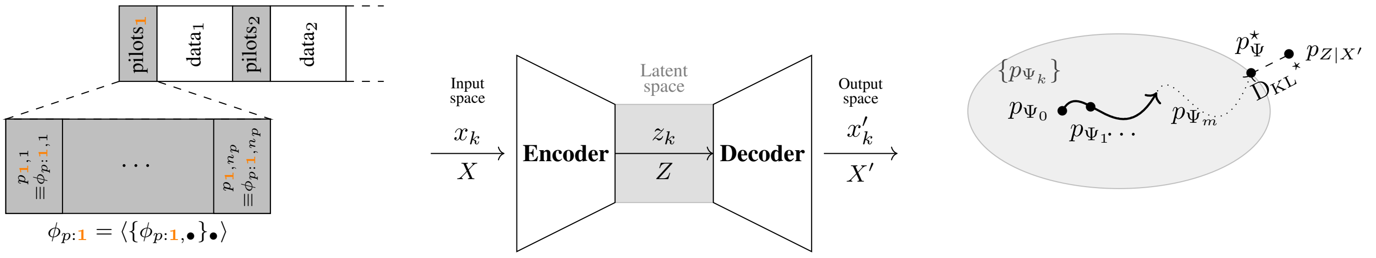
3) Phase extraction: In the time domain, and as explored in [26], the estimated phase $\widehat{\phi}$ is obtained through algebraic operations: $\widehat{\phi} = \arg\left[\det(\mathbf{h_{mimo,0}})/2\right]$.

Regarding the possible cycle-slips, a possible way to mitigate 90° phase ambiguity in quadrature-amplitude modulations is to rotate the received constellation by $\{0°, 90°, 180°, 270°\}$ for each batch of symbols, and to compute the corresponding BER. The rotation that minimises the BER is then selected, as in [4].

C. Pilot-aided Carrier Phase Recovery

The pilot-aided carrier phase recovery (PA-CPR) involves inserting n_p pilot symbols periodically within the data stream after every n_d symbols, as shown in Fig.1a. The pilots are known at both ends of the link and may possess specific statistical properties for channel estimation. For example, orthogonal Golay sequences or Constant Amplitude Zero AutoCorrelation (CAZAC) can be employed for the latter, as in [28], [29]. In this study, we use random QPSK pilots, ensuring that the mean power of the pilot sequences matches that of the payload. We group *p*ilots $p_\bullet$ (resp. *d*ata $d_\bullet$) symbols into sets. The pilot (resp. data) subsets contain n_p (resp. n_d) symbols. Thus p_{mn} (resp. d_{mn}) is the n-th $\in [\![1, n_p]\!]$ pilot (resp. n-th $\in [\![1, n_d]\!]$ data) symbol of the m-th set, as depicted

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685



(a) (b) (c)

Fig. 1: (a). Structure of the payload for pilots-aided CPR. (b). Global structure of an AutoEncoder. (c). The optimisation process. Inspired from [30], Fig.3.5.

in Fig. 1a. Let $\{a_\bullet, \phi_\bullet\}$ represent the amplitude and phase of the symbol $\bullet \in \{p, d\}$:

$$\bullet_{mn} = a_{\bullet,mn} \cdot e^{-i\phi_{\bullet,mn}}, \ \bullet \in \{p, d\} \setminus \\ \langle |\{p_{mn}\}|^2 \rangle = \langle |\{d_{mn}\}|^2 \rangle \tag{9}$$

The phase of each pilot symbol is estimated and denoted $\widehat{\phi}_{p,mn}$. Only the arithmetic average of the estimated phases is retained for each batch of pilots, noted $\widehat{\phi}_{p,m} = \langle \{\widehat{\phi}_{p,mn}\}_n \rangle$ where $n \in [\![1, n_p]\!]$. This averaged phase estimation is assumed constant throughout the payload within the same set, i.e., $\widehat{\phi}_{d,m} = \widehat{\phi}_{p,m}$. Once the pilot phases are estimated, the pilot symbols are removed, and the opposite phase is applied to the data symbols of the corresponding set for phase-noise compensation: $z_{\mathrm{cpr},mn} = d_{mn} \cdot e^{-i(\phi_{d,mn} - \hat{\phi}_{p,m})}$. Clearly, the greater the number of pilot symbols, the more effective the phase recovery will be.

V. VARIATIONAL AUTOENCODER-INSPIRED DIGITAL SIGNAL PROCESSING

A. *From Artificial Intelligence to Communications*

1) What is an AutoEncoder?: AutoEncoders (AEs) are a class of Deep Neural Networks (DNNs) initially designed for generative tasks, such as image or text synthesis. The term “deep” does not refer to the number of layers, but rather to the presence of internal neural layers capable of extracting high-level features from data [31]. AEs consist of two components: an *encoder*, which usually compresses the input into a latent representation, and a *decoder*, which reconstructs the output from this latent space. This structure is illustrated in Fig. 1b. Variational AutoEncoders (VAEs) extend AEs by generating outputs from noisy latent representations [32]. We denote by Θ the set of features targeted by the VAE.

2) Bayesian inference: The central goal is to infer the latent space from the data, i.e., to determine the conditional law $p_{Z|X'}$. A natural approach uses Bayes' rule. If $p_{Z,X'} = p_{Z,X}$, then: $p_{Z|X'} = \frac{p_{X'|Z}}{p_{X'}} \cdot p_Z$. Most works (e.g., [4], [20], [32]) make no distinction between X and X'. Following [33], we deliberately distinguish them to clarify the underlying spaces and to prepare the communication-oriented formulation. In this context, the assumption $p_{Z,X'} = p_{Z,X}$ is non-trivial.

VAEs seek the most likely law for $Z|X'$. This corresponds to an optimisation problem that requires a measure of similarity between probability laws. A natural candidate is the Kullback-Leibler divergence $\mathrm{D_{KL}}$ [30]. Such a divergence is *not* a distance because it does not satisfy the triangle inequality and only has a lower bound equal to zero if the laws are identical.

Given two r.v.s (X, Y) defined on the same universe $X(\Omega) = Y(\Omega)$, the divergence is defined as: $\mathrm{D_{KL}}\left[p_X \parallel p_Y\right] = \mathbb{E}_X\big(\ln[p_X(X)/p_Y(X)]\big)$. where $\mathbb{E}_Y(X) = \int_{Y(\Omega)} \mathbb{E}(X|Y = y) p_Y(y) dy$, when Y is continuous. The standard expression $\mathrm{D_{KL}}\left[p_X \parallel p_Y\right] = \int_U p_X(x) \ln\left[\frac{p_X(x)}{p_Y(x)}\right] dx$ holds only when $X = g(Y)$ if g is an integrable function—a consequence of the transfer theorem. Hence, we prefer the more general formulation using $p_Y(X)$, which makes the usual formula a corollary.

The variational process now seeks the best approximation of the unknown law $p_{Z|X'}$ within a family $\{p_{\Psi_k}\}$: $p^\star_\Psi = \arg\min_{p_\Psi \in \{p_{\Psi_k}\}} \mathrm{D_{KL}}\left[p_\Psi \parallel p_{Z|X'}\right]$. Since $p_{Z|X'}$ is unknown, we cannot evaluate this divergence directly. Instead, the following cost function is optimised:

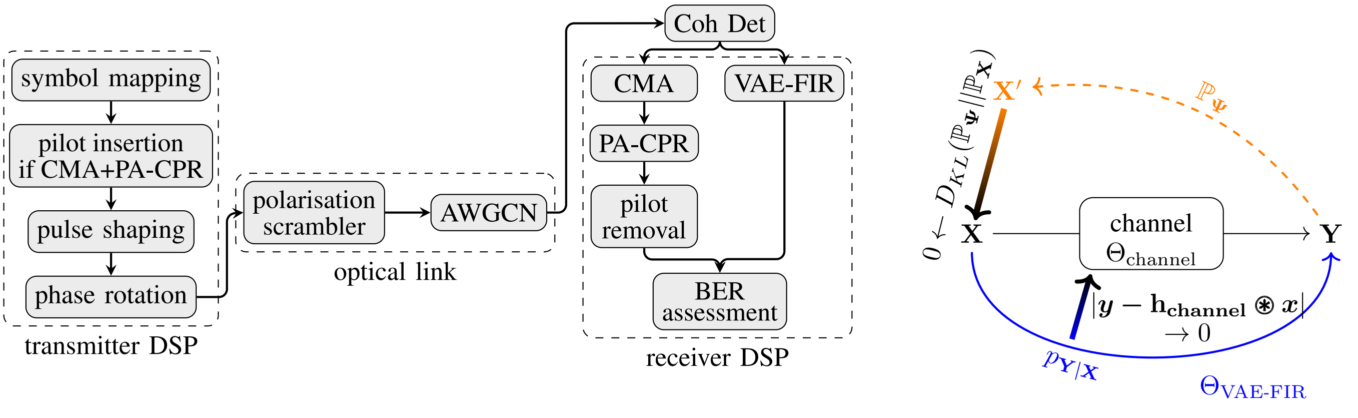
$$J_{\mathrm{VAE}}(p_\Psi) = \mathrm{D_{KL}}\left[p_\Psi \parallel p_Z\right] - \mathbb{E}_\Psi\big(\ln[p_{X'|Z}(X')]\big). \tag{10}$$

The first term quantifies the fidelity to the compressed representation while the second acts as a regulariser to match the prior [30].

3) From a communication perspective: In digital communications, we also model the transmission system with three spaces: the transmitter side, the channel, and the receiver side. Hence, the encoder and decoder can be viewed as operations applied to the transmitted symbols, with the decoder acting as the DSP at the Rx. The objective is to retrieve the sent symbols and the channel parameters ($\Theta_{\text{channel}} = \big\{\{\theta_k\}, \{\phi_k\}, g\big\}$) for communication and monitoring/sensing purposes respectively. The estimated channel characteristics are denoted by $\Theta_{\text{VAE-FIR}} = \widehat{\Theta}_{\text{channel}}$.

Unlike the artificial intelligence approach where we search for the latent space, in the adaptation of VAEs to communications, we already have access to the latent space. Here, the input space is the one sought. The VAE-**FIR** is then a **FIR-based** equaliser playing the role of the “decoder” of the VAE (in AI). Thus, the structure of the VAE-FIR equaliser is similar to the CMA butterfly equaliser that only uses a loss function *inspired* by the VAE in AI. As a consequence, the traditional issues in AI associated to the “training” do not apply here like “overfitting”.

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685



(a)

(b)

Fig. 2: (a) Simplified schematic of the simulation setup. "Coh Det" comprises all the traditional coherent receiver components. (b) Schematic view of the VAE from a communication perspective. The optimisation makes X' tend to X as well as $\Theta_{\text{VAE-FIR}}$ to Θ_{channel}.

The loss function becomes:

$$J_{\text{VAE-FIR}}(\mathbb{P}_{\boldsymbol{\Psi}}) = \mathrm{D}_{\mathrm{KL}}\left[\mathbb{P}_{\boldsymbol{\Psi}} \parallel \mathbb{P}_{\mathbf{X}}\right] - \ln[p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y})] \setminus \\ \mathbf{Y}|\mathbf{X} \hookrightarrow \mathcal{CN}(\mathbf{h}_{\mathbf{channel}} \circledast \mathbf{X}, N_0 \cdot \mathbf{I}), \quad (11)$$

where N_0 is the power spectral density of the additive noise per polarisation and $\mathbf{I}$ is the identity matrix.

One may draw attention on the fact that we switched from a probability density function $p_{\bullet}$ to a probability mass function $\mathbb{P}_{\bullet}$. Indeed, here X (and then ψ) are discrete random variables, contrary to Y as also pointed out in [4]. Secondly, we also used bold font to remind that in DP communications we deal with I/Q quadrature and two polarisation states.

Each term can be interpreted as follows: the divergence is related to the sent constellation or, in probabilistic terms, the prior, and aims to make $\mathbf{X}'$ tend to $\mathbf{X}$, as depicted in Fig. 2b. The second term, the log-likelihood $\ln[p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y})]$, is explicitly linked to the channel model, in this case an AWCGN channel, analogous to the "compression fidelity" mentioned in the AI approach, which aims to make $\Theta_{\text{VAE-FIR}}$ tend to Θ_{channel}.

Unlike CMA, the VAE-FIR performs a batch-wise update of the loss function using variable size of batches of symbols that will impact the computational and time complexity as we discuss in Sec. VI-A.

B. Implementation of SoP and phase variations

1) Division into frames and batches: We process the incoming data stream from the CDC step by cutting it into N_{F} frames, which are further divided into $N_{\text{B/F}}$ batches containing $N_{\text{Sb/B}}$ symbols. We set the same SoP for all batches within the same frame, while the phase differs across each batch.

2) Carrier phase effects: To demonstrate the VAE-FIR's ability to compensate for residual phase noise or carrier frequency offset, we implement a phase evolution where the phase is set by batches. In the case of CFO, the phase evolution across the batches is linear and the offset between batches is:

$$\Delta\Phi_{\text{CFO}} = 2\pi \cdot (\text{CFO}/R_{\text{s}}) \cdot N_{\text{Sb/B}}, \quad (12)$$

where CFO is in [Hz]. For the m-th batch of the k-th frame, the phase is: $\phi_{k,m} = k \cdot m \cdot \Delta\Phi_{\text{CFO}}$ where $(k, m) \in [\![1, N_{\text{F,ch}}]\!] \times [\![1, N_{\text{B/F}}]\!]$ with $N_{\text{F,ch}}$ being the number of frames over which the channel is emulated, after the training phase. When the phase noise is implemented, the phase offset from batch to batch $\Delta\phi_B$ needs to be scaled to the number of symbol per batches such that $\Delta\phi_B = \sqrt{N_{\text{Sb/B}}} \cdot \Delta\phi$. Concerning the phase noise, as it is also implemented by batch, the corresponding variance of the phase noise becomes: $\sigma^2_{\Delta\phi,\text{batch}} = N_{\text{Sb/B}} \cdot \sigma^2_{\Delta\phi}$ where $\sigma^2_{\Delta\phi}$ is given by Eq. (6).

3) State of Polarisation: For the SoP, we set the standard deviation according to the polarisation linewidth f_{pol}, such that on a frame-wise implementation, we have:

$$V_{\text{ch}}(T_{\text{F}}) = R_{\text{s}} \cdot \sigma_{\text{pol}}(T_{\text{F}})/N_{\text{Sb/F}} = \sqrt{2\pi f_{\text{pol}} R_{\text{s}}/N_{\text{Sb/F}}}. \quad (13)$$

where $N_{\text{Sb/F}}$ is the number of symbols per frame.

C. Illustrative Constellation Examples at DSP Stages

To *illustrate* the digital signal processing flow, we provide in Fig. 3 a set of representative constellations at key stages of the transmission and reception chain. The top row (a.i) corresponds to the CMA+PA-CPR scheme, while the middle row (b.i) corresponds to the VAE-FIR. These examples are based on 16-QAM signals for improved pilot visibility. QPSK pilot symbols are visible in (a.1). The input SNR in both cases is 50 dB. Figures (a.2) and (b.2) present the received constellations immediately before equalisation, after propagation through the optical channel. A 20-degree polarisation rotation is applied. The effects of polarisation mixing and noise are clearly visible. In (a.3), only CMA has been applied. The polarisation impairments are partially corrected, revealing the pilot symbols that were indistinguishable in (a.2). Phase recovery is then performed in (a.4) using PA-CPR, yielding a cleaner constellation. Finally, (a.5) shows the decision-ready constellation with pilots removed (labelled "eq norm") and the ideal constellation overlaid in blue as reference.

The VAE-FIR, on the other hand, processes both polarisation and phase impairments jointly, as shown in (b.3). While a

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685

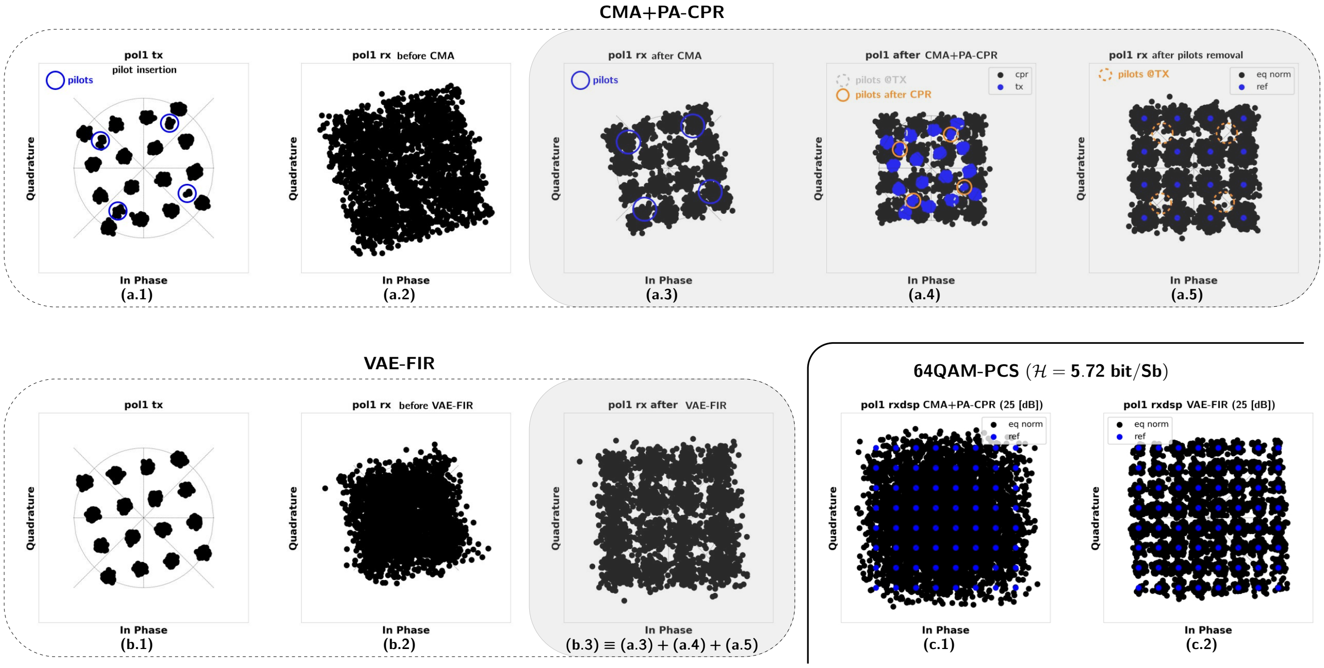


Fig. 3: Evolution of the constellations after key steps in the data flow in the channel and the DSP blocks for the CMA-based (row $\{(a.i)\}$) and for the VAE-FIR-based (row $\{(b.i)\}$) equalisation, for 16QAM transmission. Figs. (a.1,b.1) show the constellation before and Figs. (a.2,b.2) after the channel. Figs. (a.3-5,b.3) after the DSP blocks at the receiver side. The *illustrative* Figs. (c.1,c.2) respectively show the constellation of the equalised signal by the (CMA+PA-CPR,VAE-FIR). for 64QAM-PCS ($\mathcal{H} = 5.72$ bits/Sb)

slight tilt remains, the structure of the constellation demonstrates that the algorithm effectively mitigates both impairments without separate processing stages.

Figures 3(c.1) and 3(c.2) display the final constellations after equalisation for PCS-64QAM with entropy $\mathcal{H} = 5.72$ bits/Sb at SNR = 25 dB. The SNR values were selected to make the constellations visually interpretable. Finally, we highlight that our tests showed that the CMA+PA-CPR also fails for low power of noise ($\text{SNR} = 30$ and 40 dB).

VI. IMPACT OF THE BATCH SIZE ON VAE-FIR PERFORMANCE

While this section may seem out of chronological order, it reflects the actual methodological progression: these preliminary calibrations were necessary to validate the results presented later on. Thus, we first detail the parameter tuning process. Indeed, the optimal settings (e.g., batch sizes, number of training frames, etc.) depend on the channel dynamics and system rate, and must be established *prior* to any performance evaluation.

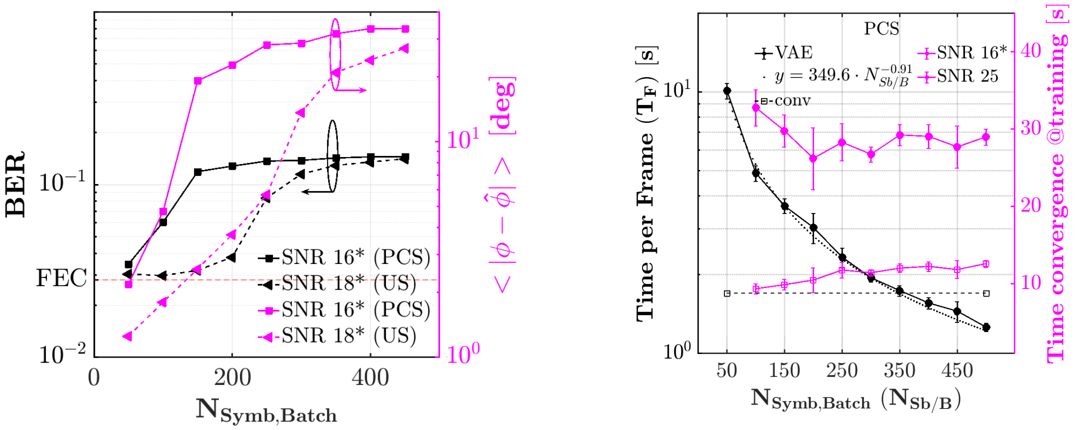
A. Performance with respect to the batch size

1) Setup: In this subsection, we evaluate the impact of the batch size on the VAE-FIR performance under a residual CFO of 100 kHz. The phase evolves linearly within the range $[0, \pi/2]$, allowing us to isolate the effect of additive noise. The baud rate is set to 64 GBd, and SNR to 16 and 18 dB for PCS and US, respectively. For PCS, the entropy is fixed at $\mathcal{H} = 5.72$ bits/Sb, as in [4]. These values correspond to critical working points, as discussed later in subsection VIII-B. Figure 4a shows the impact of three parameters—$N_{\text{Sb/B}}$, SNR, and modulation format—on phase estimation accuracy. The BER (black) and phase estimation error (magenta) are displayed on the left and right axes, respectively.

2) Impact on the BER: Uniform Shaping (US) yields better performance than PCS, demonstrating the VAE-FIR's superior ability to equalise a residual CFO in uniformly distributed QAM formats. At low batch sizes, BER slightly exceeds the FEC threshold: $(3 \cdot 10^{-2}; 3.2 \cdot 10^{-2})$ for PCS and US, respectively. We will later show (see subsection VIII-C) that the corresponding power penalty remains below 1 dB. Interestingly, the BER for US exhibits a sigmoidal dependence on $N_{\text{Sb/B}}$, whereas for PCS it plateaus beyond $N_{\text{Sb/B}} = 150$, indicating no gain from increasing the batch size further.

3) On the estimation error of the phase: The estimation error exhibits similar trends: sigmoidal for US and step-like for PCS, reinforcing the conclusions on the limited benefit of increasing $N_{\text{Sb/B}}$. As expected, smaller batches yield better phase estimation due to more frequent loss updates. For PCS at $\text{SNR} = 16$ dB, the error drops from 31 deg ($N_{\text{Sb/B}} = 450$) to 2.5 deg ($N_{\text{Sb/B}} = 50$). For US, it drops from 29 deg to 1.2 deg over the same range at $\text{SNR} = 18$ dB This improvement due to the frequent updates to the loss function comes at the cost of increased computation time, as discussed in the next section.

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685



(a)

(b)

Fig. 4: (a) Evolution of the average BER (left axis) and the estimation error of the phase (right axis) with respect to the batch size $N_{\mathrm{Sb/B}}$ using the VAE-FIR for CFO = 100 kHz. (b) Time complexity per data frame in black solid line and its fitted curve in dashed line (left axis). Time required for VAE-FIR's convergence in the training phase in order to reach convergence for two SNR values (right axis).

B. Impact of the batch size on the computing requirements

We now take a glance at the computational burden of the VAE-FIR. The execution time at the receiver is measured and compared to the CMA+PA-CPR baseline. All simulations were run on an Intel® Core$^{\mathrm{TM}}$ i5-10310U CPU @1.70GHz × 8. Unlike original CMA, which operates symbol-wise, VAE-FIR processes the signal in batches. As a result, its training time is independent of the number of symbols per batch ($N_{\mathrm{Sb/B}}$). This makes $N_{\mathrm{Sb/B}}$ a crucial hyperparameter for performance and complexity trade-offs. We first examine the influence of $N_{\mathrm{Sb/B}}$ on computational time. Figure 4b summarises the results.

On the left axis (black, circle markers), we show the computation time per frame (T_F), which is invariant to SNR and learning rate. As expected, longer batches reduce T_F. This is because the number of gradient updates per frame is roughly inversely proportional to $N_{\mathrm{Sb/B}}$. The trend follows a power law with exponent -0.91. For comparison, the CMA computation time remains constant with respect to $N_{\mathrm{Sb/B}}$, and under 2 s for $N_{\mathrm{Sb/F}} = 10^4$.

On the right axis (magenta), we plot the convergence time of the VAE-FIR during training, which may be performed offline. We *empirically* considered the model to have converged once a target SNR or SER is reached. These metrics reflect the model's sensitivity to noise. The number of frames required for convergence ($N_{\mathrm{F,c}}$) grows linearly with $N_{\mathrm{Sb/B}}$, as:

$$N_{\mathrm{F,c}}(N_{\mathrm{Sb/B}}, \mathrm{SNR}) = S_{\mathrm{SNR}} \cdot N_{\mathrm{Sb/B}} + \alpha_{\mathrm{SNR}} \setminus \{(S_{\leq 20}, S_{>20}), (\alpha_{\leq 20}, \alpha_{>20})\} \in \{(0.04, 0.02), (2.37, 0)\}. \quad (14)$$

We observed a regime change near SNR = 20 dB. Above this threshold, convergence is empirically declared once $\mathrm{SNR} \geq \mathrm{SNR}_{\mathrm{target}} - 2$ dB. Below it, we use SER $\leq 10^{-2}$ as the convergence criterion. Within each SNR domain ($\leq$ 20 or $>$ 20 dB), the frame count $N_{\mathrm{F,c}}$ varies little, justifying the use of constant slopes and offsets in the model.

Although $N_{\mathrm{F,c}}$ linearly slightly increases with $N_{\mathrm{Sb/B}}$, the frame duration T_F decreases. Hence, for low SNR, the total convergence time T_{c} is then not linear. Its error bars are computed via propagation: $\sigma(T_{\mathrm{c}}) = N_{\mathrm{F,c}} \cdot \sigma(T_F)$.

We report results for two noise levels: $\mathrm{SNR}_{\mathrm{target}} \in \{16, 25\}$ dB. At low SNR, convergence time roughly doubles for $N_{\mathrm{Sb/B}} = 500$ and triples for $N_{\mathrm{Sb/B}} = 100$. Nonetheless, as shown earlier, reducing $N_{\mathrm{Sb/B}}$ yields improved phase estimation accuracy. This reveals a trade-off between accuracy and convergence duration.

VII. TRANSMISSION AND PROCESSING PARAMETERS

In this section, we define the transmission scenarios to evaluate the performance of the VAE-FIR. The simplified schematic of the simulation setup is shown in Fig. 2a. We simulate a linear DP AWGCN channel with a single wavelength where only carrier phase effects and SoP drifts are considered and added separately. The modulating signal is root-raised cosine shaped with a roll-off factor of 0.1. At the Tx side, we add a noise of total constant power -50 dBm as an overall representation of the different noise sources (amplifiers, quantisation, etc.). We operate at two symbol rates $R_{\mathrm{s}} \in \{64, 128\}$ GBd with 64-QAM(-PCS) using the usually inappropriately called "Maxwell-Boltzmann" (MB) distribution. We explain in appendix X-B why this name can lead to confusion with its original use in statistical mechanics, and we will use "discrete Gaussian" as in [34] instead.

We focus on the impact of the source symbol distribution (uniform with entropy $\mathcal{H}_1$ vs. shaped with entropy $\mathcal{H}_2$) on PCS signal processing, regardless of the resulting net information rates. The shaping factor λ of this distribution is set to $\lambda = 0.0254$ corresponding to an entropy of $\mathcal{H} = 5.72$ bits/Sb. This value of entropy has been chosen as it is the value for which [4] has shown the ability of the CMA to converge.

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685

Authors in [5] also showed that the CMA could converge for $\mathcal{H} = 5.8$ bits/Sb.

We set the number of symbols per batch to $N_{\mathrm{Sb/B}} = 250$ as we found this value to be a trade-off between computation time and performance, as shown in Fig. 4a and Fig. 4b. For the CMA+PA-CPR, we introduce 12 pilots every 225 data symbols, corresponding to $\approx 5\%$ overhead. The number of symbols per frame is set to $N_{\mathrm{Sb/F}} \in \{10^4, 2\cdot 10^4\}$ respectively for CMA+PA-CPR and VAE-FIR as we found them to be the best values in terms of achieved BER and computation time. The learning rate is set to $lr \in \{1, 750\} \cdot 10^{-6}$ respectively for the CMA and the VAE-FIR. When phase noise is introduced, we fix the number of frames using its maximum expected absolute value $\left(\equiv \mathbb{E}(|\phi_{N_{\mathrm{F,ch}}}|) = \pi/2\right)$. The phase at the last batch of the simulation run is set to $\pi/2$ when CFO is implemented. Further details of implementation are given in X-D.

VIII. RESULTS

A. Outline

In the previous sections, we tuned the parameters of the algorithms in order to get a trade-off between performance and complexity. Now, we study the performance of both algorithms in terms of error rates and errors of estimation of the channel parameters: SoP and phase for different transmission scenarios. In the following, we mean by "operation point", the SNR value such that the BER obtained after equalising the signal equals the pre-FEC limit. The denomination "pre-FEC limit" may also be shortened by "FEC" when no confusion is possible with the FEC algorithms. We first determine the operation point for each set of transmission parameters $(\mathcal{H}, R_{\mathrm{s}}, \mathrm{CFO}, V_{\mathrm{ch}}, ...)$ by computing the BER versus SNR curve where $\mathrm{SNR} = \mathrm{E_{Sb}/N_0}$ and $N_0 = \sigma^2_{\mathrm{tx}} + \sigma^2_{\mathrm{rx}}$. Then, we compare the ability of each algorithm to track channel effects. We start with the phase drifts modelling carrier phase noise and residual CFO, then we pursue with random polarisation drifts modelled as a Wiener process, known as "polarisation linewidth" model proposed in [22]. In all graphs, when the SNR is given (in dB), a star $*$ is used to mention that it is the one corresponding to the pre-FEC BER (operation point for AWGCN channel). As for the SoP estimation error, we use the mean absolute error as a performance indicator. We will call "conventional approach" the use of CMA+PA-CPR enabling us to shorten the legend entries by "conv" when referring to CMA+PA-CPR, as well as referring to VAE-FIR by "VAE" only. Each point is the average of 10 realisations and each frame contains $N^{\mathrm{conv}}_{\mathrm{Sb/F}} = 10 \cdot 10^3$ or $N^{\mathrm{VAE\text{-}FIR}}_{\mathrm{Sb/F}} = 20 \cdot 10^3$ symbols.

Finally, as the simulations are performed at $R_{\mathrm{s}} = 64$ and $R_{\mathrm{s}} = 128 = 2 \times 64$ GBd, when channel effects are implemented with a unique set of different values for a parameter $x = (x_1, ..., x_p)$, errors of estimation or BERs are plotted with respect to x/R_{s}. This unique set can lead to multiple points for the same x/R_{s} as in Fig. 6c or Fig. 6e that should not be seen as drops and are usually specific of the simulations at $R_{\mathrm{s}} = 128$ GBd. For example if $\Delta\nu = \{1 \cdot 10^4, 2 \cdot 10^4\}$ Hz then $2 \cdot 10^4/128 = 1 \cdot 10^4/64$.

B. Additive White Gaussian Noise Channel

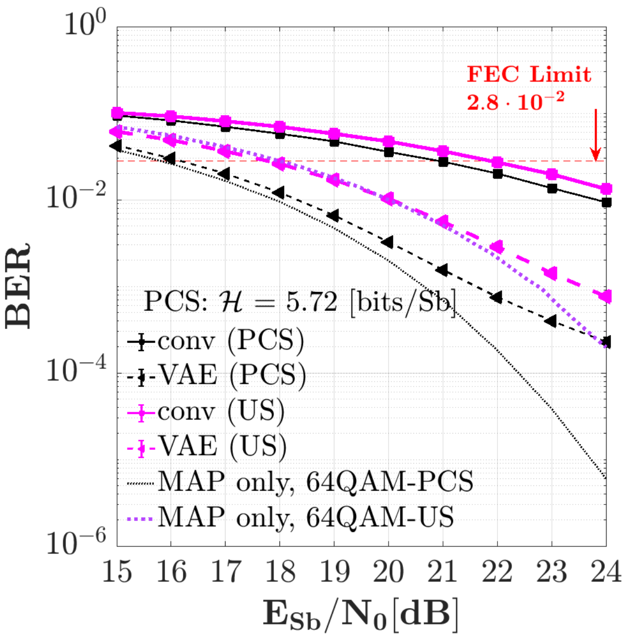


Fig. 5: Evolution of the BER versus SNR for AWGCN channel for both equalisers and both PCS/US. MAP stands for Maximum A Posteriori decoding. The error bars are too small to be visible and are therefore hidden by the markers.

In Fig. 5, we draw the BER versus SNR curves for the uniform 64QAM and 64QAM-PCS at an entropy $\mathcal{H} = 5.72$ bit/Sb using both CMA+PA-CPR and VAE-FIR. We consider the Maximum A Posteriori (MAP) decoding in AWGN as the theoretical limit. The power penalty is the required additional power (or SNR) to get the same BER as the MAP curve. We can firstly notice the lower power penalty to the theoretical limit (Maximum A Posteriori (MAP) only) using the VAE-FIR (≈ 0.3 dB) compared to the CMA+PA-CPR (≈ 5.3 dB) which clearly shows that the VAE- FIR outperforms the CMA+PA-CPR for probabilistic shaping. In the case of uniform 64QAM, the gap diminishes between the algorithms. The power penalty for the VAE-FIR is ≈ 0 dB while it is ≈ 4 dB for the CMA+PA-CPR. From now on, unless otherwise specified, the simulations will be done at the operation points given in Tab. I corresponding to the SNR values at the pre-FEC BER limit for each signalling and each equaliser, and to 1 dB additional margin from these values.

TABLE I: Summary of the SNR (dB) at the operation points for the simulations (* for pre-FEC BER limit).

$\mathcal{H}$ (bits/Sb)	CMA+PA-CPR (aka. "conv")	VAE-FIR (aka. "VAE")
6	22*, 23	18*, 19
5.72	21*, 22	16*, 17

C. Residual Carrier Frequency Offset

In Fig. 6a and Fig. 6d, we show how well both CMA+PA-CPR and VAE-FIR can estimate and mitigate the residual CFO around the pre-FEC operation point with CFO $\in \{10, 50, 100, 200\}$ kHz.

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685

a) Concerning the estimation error: Obviously, CMA+PA-CPR shows very small errors ($< 1\ \mathrm{deg}$) by the essence of the algorithm. The estimation errors using the VAE show that it provides errors below $3\ \mathrm{deg}$ as long as $\mathrm{CFO}/R_\mathrm{s} < 4 \cdot 10^{-7}$ at $16^*\ \mathrm{dB}$. This limit is pushed to $\approx 9 \cdot 10^{-7}$ by setting the operation point $1\ \mathrm{dB}$ beyond the FEC operation point (17 dB). The US cases ($18^*, 19\ \mathrm{dB}$) show similar trends compared to the the PCS case. The error limits are $3\ \mathrm{deg}$ up to $\mathrm{CFO}/R_\mathrm{s} \leq 8 \cdot 10^{-7}$ @17^* dB and $\mathrm{CFO}/R_\mathrm{s} \approx 1.5 \cdot 10^{-6}$ @19^* dB. Hence, it reveals again that the VAE better equalises the uniform shaping.

b) Concerning the Bit Error Rate: We can first notice that, overall, the VAE-FIR is slightly outperformed by CMA+PA-CPR, below $\mathrm{CFO}/R_\mathrm{s} = 8 \cdot 10^{-7}$, and largely outperformed beyond. The conventional algorithm also yields a more stable behaviour, showing its robustness to the fast phase rotation in the range of tested CFO values which is the benefit of the data-aided processing. The efficiency of the VAE-FIR in tracking fast phase rotations rapidly decreases with increasing $\mathrm{CFO}/R_\mathrm{s}$, beyond $\mathrm{CFO}/R_\mathrm{s} > 8 \cdot 10^{-7}$, as the BER soars before stabilising to $\mathrm{BER} = 0.15$. In all cases, we see that a $1\ \mathrm{dB}$ margin yields significant improvement. The additional margin enables to get below the FEC threshold. For uniform shaping and VAE-FIR, we notice a clear improvement in pushing further the maximum tolerable CFO mismatch, from 50 to $\approx 95\ \mathrm{kHz}@R_\mathrm{s} = 64\ \mathrm{GBd}$. Finally, we notice that the slight variations in BER are of the order of 10^{-3} that is around the systemic error ranges that we evaluated earlier.

D. Carrier Phase Recovery

In Fig. 6b and Fig. 6e, we show how well both equalisers can estimate and mitigate the phase noise around the FEC operation points for typical laser linewidth values $\Delta\nu \in \{5, 10, 20, 50\}\ \mathrm{kHz}$.

a) Concerning the estimation error: We can see that two sets of curves come out: one related to the CMA+PA-CPR and another one for the VAE-FIR. The former shows estimation errors below $0.5\ \mathrm{deg}$ as expected given the pilot-aided phase estimation. Its behaviour is stable throughout the whole range of rates and linewidth values. On the contrary, the VAE-FIR yields an increasing error with increasing linewidths, as we expected compared to the conventional approach. Indeed, the pilot-aided version is specifically designed to retrieve the phase as we know the original phase for each pilot symbol. On the other hand, the rapidly increasing behaviour of the estimation error reveals the instability of the VAE-FIR with respect to the phase noise, despite the increase of the SNR. Indeed, we see that adding 1 dB of beyond the operating point, only allow a few kilo hertz of linewidth ($\Delta\nu/R_\mathrm{s}^{(\max)} \approx 5 \cdot 10^{-8} \equiv 6.4\ \mathrm{kHz}@128\ \mathrm{GBd}$).

The VAE-FIR shows estimation errors that are almost always greater than $0.5\ \mathrm{deg}$ revealing its low tolerance to random phase walk. In Fig. 7a, we show typical purely illustrative examples of temporal phase tracking for the CMA+PA-CPR and the VAE-FIR for both signalling schemes at $\Delta\nu = 10\ \mathrm{kHz}$. Significantly, in these situations, the phase sometimes exceeds $90\ \mathrm{deg}$ revealing the effectiveness of the handcrafted unwrapping by avoiding phase jumps. On the right column, we see that the VAE-FIR can catch up the tracking of the phase as long as its evolution is smooth enough.

b) Concerning the Bit Error Rate: We notice, as in the case of the residual frequency offset, that the performance of the CMA+PA-CPR is more or less constant throughout the whole ranges of linewidth values. The conventional approach stays below the FEC limit. Likewise to residual CFO, the performance of the VAE-FIR approach linearly decreases with increasing linewidth. Compared to the use of CMA+PA-CPR, the VAE-FIR hardly handles the phase noise as its performance is always worse after $\Delta\nu/R_\mathrm{s} \approx 5.5 \cdot 10^{-7}$ (corresponding to $\Delta\nu \approx 6.4\ \mathrm{kHz}@128\ \mathrm{GBd}$). Neither the shaping nor increasing the SNR beyond the FEC limit working point improves the equalisation. This leads to the conclusion that the power penalty induced by the random phase walk is greater than $1\ \mathrm{dB}$. Interestingly, we notice that the BER quickly increases after $\Delta\nu/R_\mathrm{s} \approx 8 \cdot 10^{-8}$. It corresponds to an error on the estimation of the phase around $1\ \mathrm{deg}$. This last result reveals that the VAE-FIR is very sensitive to phase estimation mismatches as low as 1 (resp. 3 deg) when phase noise (or CFO mismatch) is implemented.

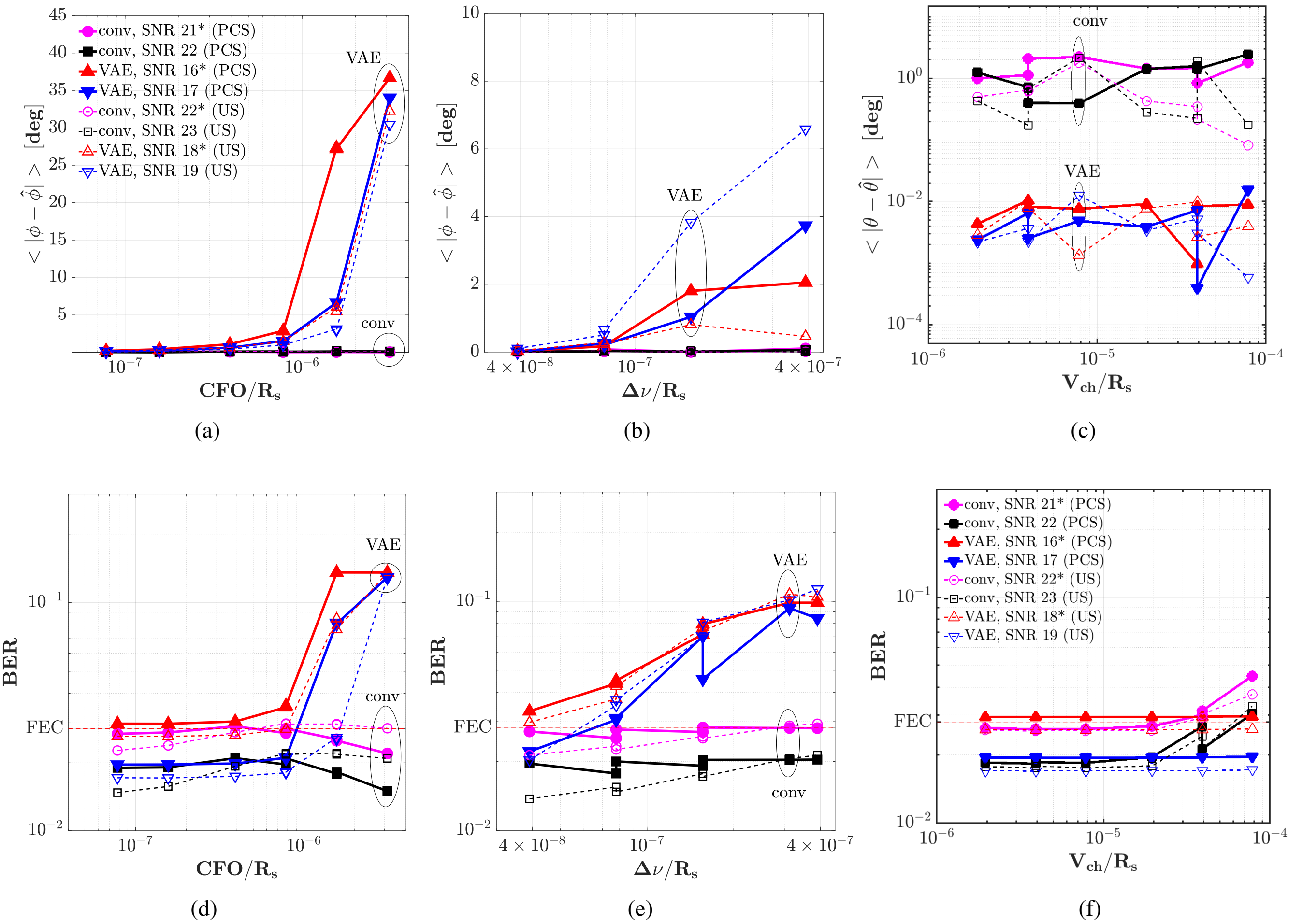
E. Polarisation drifts

Now, we study the polarisation tracking performance. As the channel variations are implemented frame-wise, the required time to reach significant channel drifts using the polarisation linewidth model is much larger than the time required for equivalent phase shifts ($\times 80$ for our setup). Thus, we simulated transmissions at $64\ \mathrm{GBd}$ for $V_\mathrm{ch} \in \{0.25, 0.5, 2.5, 5\}\ \mathrm{Mrad/s}$ and $N_\mathrm{F} = 100$. The phase noise is set to zero.

1) SoP estimation: In Fig. 6c, we display the estimation errors on the polarisation angle with respect to $V_\mathrm{SoP}/R_\mathrm{s}$. We clearly distinguish two sets of curves: one for the CMA+PA-CPR and one for the VAE-FIR. While the former always yields errors greater than $0.08\ \mathrm{deg}$ and even up to a few degrees, the VAE-FIR ensures estimation errors below $0.02\ \mathrm{deg}$. This result shows that VAE-FIR demultiplexes the polarisations better than the CMA for both PCS and uniform signalling at their respective operation points.

2) Bit Error Rate: In Fig. 6f, we show the BER with respect to $V_\mathrm{ch}/R_\mathrm{s}$. Unlike its behaviour with respect to phase noise, after using VAE-FIR equaliser, the BER is flat across the whole range of data rates and SoP rotation speeds. Despite a flat response of the conventional approach up to $\approx 1.5\ \mathrm{Mrad/s}$, the VAE-FIR performs better beyond. For PCS transmission, we see that, at the FEC operation point ($\mathrm{SNR} = 16\ \mathrm{dB}$), the VAE-FIR gets just above the FEC limit ($\mathrm{BER} \approx 2.9 \cdot 10^{-2}$). However, at $17\ \mathrm{dB}$, the BER after the VAE-FIR equalisation, drops to $\approx 2 \cdot 10^{-2}$ with a great stability across the whole values of $V_\mathrm{ch}/R_\mathrm{s}$ revealing a stable and limited power penalty ($< 1\ \mathrm{dB}$). For uniform signalling, the VAE-FIR always outperforms the conventional version and stays below the FEC limit revealing no power penalty at all. In turn, the CMA+PA-CPR exceeds the FEC limit after $\approx 1.5\ \mathrm{Mrad/s}$ and shows more than $1\ \mathrm{dB}$ power penalty.

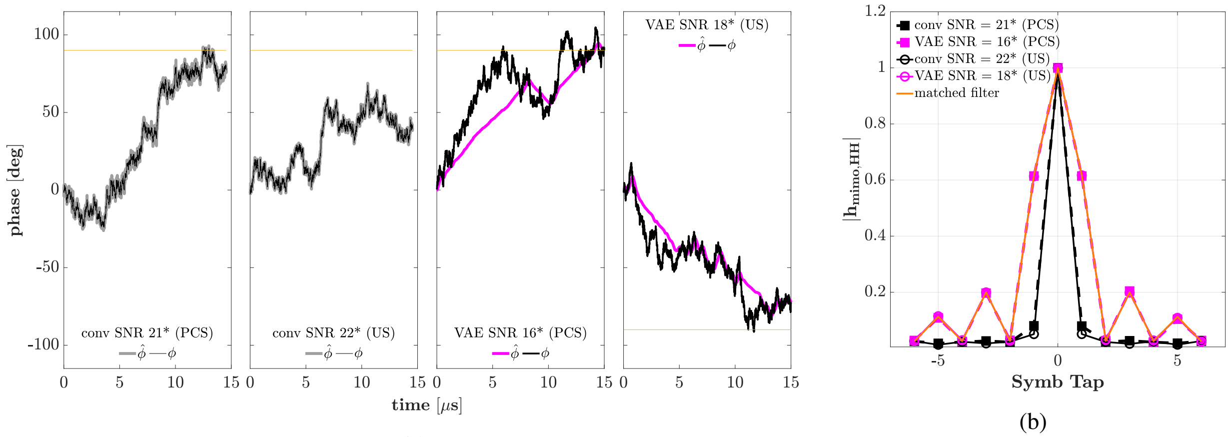
This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685



(a) (b) (c)

(d) (e) (f)

Fig. 6: Error bars are omitted so as not to overload the graph. **(a)** Estimation error of the phase with respect to CFO/R_s. **(b)** Similar to (d) with respect to $\Delta\nu/R_s$. **(c)** Estimation errors of the SoP drift with respect to V_{ch}/R_s. **(d)** Evolution of the BER with respect to the ratio CFO/R_s. **(e)** Similar to (g) with respect to $\Delta\nu/R_s$ch. **(f)** Evolution of the BER with respect to product V_{ch}/R_s.



(a) (b)

Fig. 7: Comparisons with or without constellation shaping between the CMA+PA-CPR and VAE(-FIR) for: (a) phase estimation, examples of time series. (b) time domain equalisers shape for AWGCN.

3) Matched filter estimation: Last, in Fig.7b, we show examples of one of the four FIR filter taps estimated by both algorithms, for the tested entropies, at the FEC operation point. We also superimposed the matched filter to visually check

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685

TABLE II: Power penalties (PP) for the CMA+PA-PR and VAE-FIR.

$\mathcal{H}$ (bits/Sb)	SNR (dB)			PP (dB)	
	MAP	CMA+PA-PR	VAE-FIR	CMA+PA-PR	VAE-FIR
4	11	16	13.7	5	2.7
5	13.7	20.6	14.6	6.9	0.9

how efficiently the algorithms can estimate it. It is crystal clear that the CMA hardly estimates the root-raised cosine while it is almost a perfect match for the VAE-FIR which demonstrates its better channel estimation performance, hence better polarisation mixing equalisation.

F. Tests of various shaping depths and 16-QAM transmission

We perform simulations for two additional entropy values, $\mathcal{H} \in \{4, 5\}$ bits/Sb, over an AWGN channel with the same transmission parameters. Table II presents the SNR and power penalty (PP) values for pre-FEC BER limit measured using CMA+PA-CPR or VAE-FIR, along with the MAP bound. These results show that CMA+PA-CPR continues to converge even at low entropy values, albeit with a large power penalty (typically $\geq 5\,\mathrm{dB}$). In contrast, the power penalty associated with VAE-FIR becomes significant for lower entropies but remains lower than the PP of CMA+PA-CPR. This behaviour is attributed to the instabilities (not shown here) of the VAE-FIR equalisation for lower entropy with lower SNR ($\lessapprox 15$ dB), despite an optimisation of its learning rate (set to $lr = 1{\cdot}10^{-4}$). Regarding the 16QAM-US transmission, it led to a maximum tolerable laser linewidth (resp. CFO) ranging between 15 and 20 kHz (resp. between 50 and 100 kHz) @64 GBd and SNR $=$ 13 dB using the VAE-FIR, without power penalty. Simulations have not been performed for SoP drifts as the use of 64QAM-US modulation already shows a large tolerance to such kinds of channel impairments. Hence, we also expect a large tolerance for 16QAM-US, as Fig. 3 suggests.

IX. CONCLUSIONS

A. Phase management

We explored the ability of the VAE-FIR to tackle the phase dynamics using a linear model for the CFO and a Wiener process for the phase noise. We saw that, although the VAE-FIR shows some tracking capability, its tolerance is globally weak: $(\mathrm{CFO}^{\max}, \Delta\nu^{\max})/R_\mathrm{s} < (8{\cdot}10^{-7}, 5.5{\cdot}10^{-8})$. However, these values show that linear and stable evolution is better equalised as it enables the VAE-FIR to catch up with the tracking errors. For the VAE-FIR, at $(\mathrm{CFO}^{\max}, \Delta\nu^{\max})$, the power penalties for PCS transmission are below 1 dB for CFO and beyond 1 dB for CPR. In uniform signalling, the penalties are beyond 1 dB, while the conventional approach shows no penalties. The error of estimation of the phase is globally rapidly increasing with the CFO or the linewidth of the laser. Moreover, an error of respectively $(1, 0.2)$ deg for the (CFO, CPR) leads to a BER outage, revealing both an unstable and highly sensitive behaviour of the VAE-FIR with respect to the phase, contrary to the CMA+PA-CPR that is, due to the pilot insertion for phase tracking, more robust. Thus, the VAE-FIR should not be used alone to tackle the phase and would need some assistance through pilot insertion.

B. Polarisation demultiplexing and SoP estimation

Our work in [6] confirmed the previous one by showing that the VAE-FIR outperforms the CMA in all scenarios (rates, polarisation linewidth, and constellation shaping), both in terms of decoding and channel estimation, even in scenarios more stringent, in terms of channel variation speed, that the one we simulated previously. Indeed, the power penalty is below 1 dB and we get one order of magnitude lower estimation errors of the channel. These results tend to show that the VAE-FIR is well-suited as a blind alternative to the CMA for polarisation demultiplexing.

C. Perspectives

Future work may include the ability of the VAE-FIR to combat all together chromatic dispersion, polarisation mode dispersion, and polarisation dependent loss when phase evolution is taken into account. Another possible path to explore is to compare the VAE-FIR with pilot-aided polarisation demultiplexing when stronger shaping is applied. Then, an extensive complexity study of the VAE-FIR could also be helpful to better compare it with state-of-the-art algorithms. Finally, one may search for a modification of the cost function such that it takes into account phase noise.

D. Miscellaneous

We proposed some clarifications in Appendix X-A on the definition of the speed of polarisation changes and channel changes where conventional approaches, to the best of our knowledge, may not always question its origins leading to the use of confusing angle representations.

Then, we also shed light on the misunderstood denomination of the Maxwell-Boltzmann distribution that is usually used to mention the discrete Gaussian distribution (as in [34]) used in most works involving probabilistic constellation shaping, even though some have already precised that, to “obtain” a Maxwell-Boltzmann distribution, normalisation of the Gaussian distribution is required [35].

Finally, based on the iconic work of [36], we reminded that the usual denomination “linewidth” for $\Delta\nu$ might be confusing regarding the units for the variance of the phase noise. But also for models inspired by the model of variation of the phase noise, as the “polarisation linewidth”. Though it may look as an anecdote, it is of major importance when it comes to *clearly* quantifying the variations of the studied parameter.

X. APPENDICES

A. Speed of State of Polarisation Change

Here, we demonstrate the relationship between the usual definition of SoP change, our own definition and the approach of T. Ye et al. [37]. We chose an empirical description consisting in first providing the different versions and then linking them when possible.

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685

1) Usual definition of the speed of SoP change: The Poincaré Sphere is a common tool to visualise the evolution of the SoP described by the Stokes vector $\vec{S}_n \in M_{3,1}(\mathbb{R})$ whose component $S_{i,n}$ represents the i-th component of the normalised Stokes vector (divided by S_0) for the n-th time sample, defined by the azimuthal angle and ellipticity $(2\theta, 2\chi)$ respectively, see the left part of Fig. 8:

$$\left\{\begin{array}{lcc} S_{1,n} & = & \cos(2\theta_n)\cos(2\chi_n), \\ S_{2,n} & = & \sin(2\theta_n)\cos(2\chi_n), \\ S_{3,n} & = & \sin(2\chi_n). \end{array}\right. \tag{15}$$

We illustrate two consecutive polarisation states in the right part of Fig. 8 whose displacement is parameterised by the angle $\Delta\alpha_n$.

The usual definition that is stated in the literature [37]–[39] relies on the general case of a complex-valued polarisation rotation matrix in Jones' formalism or equivalently using the Stokes vector, such that the SoP speed relies on this angle $\Delta\alpha_n$ (see the right part of Fig. 8). such that by using $\sin(\Delta\alpha_n/2) = ||\Delta\vec{S}_n||/2$, we have:

$$v_{\mathrm{SoP,n}}^{(\mathrm{usual})} = \frac{\Delta\alpha_n}{\Delta t} = \frac{2}{\Delta t}\arcsin\left(\frac{1}{2}\sqrt{\sum_{i=1}^{3}|\Delta S_{i,n}|^2}\right), \tag{16}$$

One may notice that the commonly used measure of "SoP change" does not correspond either to the standard angular reference on the sphere $(\vec{s}_1, \vec{s}_2, \vec{s}_3)$ or to those defined in the dynamic reference frame $(\vec{u}_{\vec{S}}, \vec{u}_\theta, \vec{u}_\chi)$, shown on the left side of Fig. 8. Our aim is to link the angle $\Delta\alpha_n$ of the variation of the angles of the azimuthal and ellipticity angles of Eq. (15).

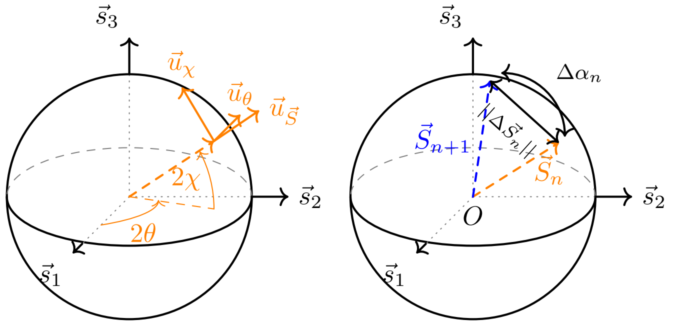


Fig. 8: Left: Representation of coordinates on the Poincaré sphere and illustration of the reference frames associated with motion. Right: Angle considered in the usual definition.

2) Speed of SoP change with real rotation matrix: In the case of a rotation in the plane elevated by a fixed angle $\chi = \chi_n,\ \forall n \in \mathbb{N}$ as in Fig. 9, the polarisation change matrix in Jones' formalism is real-valued and $\Delta S_{3,n} = 0$ which simplifies the sum to $i = \{1, 2\}$. By noting $\Delta\theta_n = \theta_{n+1} - \theta_n$ we have:

$$\Delta\theta_n = \arcsin\left(\frac{\sqrt{|\Delta S_{1,n}|^2 + |\Delta S_{2,n}|^2}}{2\cos(2\chi)}\right) = \frac{\Delta t}{2}\cdot v_{\mathrm{ch}} \tag{17}$$

where $\Delta S_{i,n} = S_{i,n+1} - S_{i,n}$ with $i \in \{1, 2, 3\}$. Indeed, we have $||\Delta S_{1,n}||^2 = 4\sin(\Delta\theta)^2\sin(\theta_n + \theta_{n+1})^2\cos(2\chi)^2$ and $||\Delta S_{2,n}||^2 = 4\sin(\Delta\theta)^2\cos(\theta_n + \theta_{n+1})^2\cos(2\chi)^2$. This can be understood with the aid of the right part of Fig. 8. Hence, it aligns exactly with Eq. (17) in the sole case of $\chi_n = 0$, for which $\Delta\alpha_n$ lies entirely in the plane $(0, \vec{s}_1, \vec{s}_2)$ such that $\Delta\alpha_n = 2\cdot\Delta\theta_n$.

Indeed, reasoning on the two triangles ABD and CBD in Fig. 9, where $AB = AD = \cos(2\chi_n)$ and $CB = CD = 1$, the relevant distance is $||\Delta\vec{S}_n|| = BD = 2\sin(\Delta\alpha_n/2) = 2\sin(\Delta\theta_n)\cos(2\chi_n)$, which leads to our conclusion when $\chi_n = 0$.

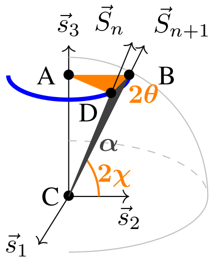


Fig. 9: Linking SoP evolution in the Poincaré sphere using the usual definition of SoP change and ours in the absolute referential in case of a rotation in a parallel.

3) Detailed derivations of [37]: Authors in [37] pointed out the fact that the usual formulation in (16) does not provide an accurate description of the channel changes.

Here, we detail their derivations to better understand the link we want to make with a real rotation matrix. The authors proposed an approach leveraging linear algebra to quantify polarisation state evolution. The variation in the state of polarisation between two time instants is thus $\Delta\vec{S}_n = \vec{S}_n - \vec{S}_{n-1}$. Denoting $\mathbf{M}_n$ as the invertible transfer (or Mueller) matrix of the SoP variation we have: $\vec{S}_n = \mathbf{M}_n\vec{S}_0$. Then, the variation of the SoP is given by:

$$\Delta\vec{S}_n = \mathbf{M}_n\vec{S}_0 - \mathbf{M}_{n-1}\vec{S}_0 = (\mathbf{M}_n\mathbf{M}_{n-1}^{-1} - \mathbf{I})\cdot\vec{S}_{n-1}, \tag{18}$$

where $\mathbf{\Delta R}_n = \mathbf{M}_n\mathbf{M}_{n-1}^{-1}$ is a rotation on a sphere. Such matrix can be diagonalised and, by using the eigenvectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$, the Stokes vector can be re-written: $\vec{S}_{n-1} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \alpha_3\mathbf{u}_3$, with $\sum_i \alpha_i^2 = 1$. In this suitable basis, the rotation matrix can be expressed as:

$$\mathbf{\Delta R}_n = \mathbf{P}_n\cdot\mathbf{D}_n\mathbf{P}_n^T \setminus \mathbf{D}_n = \mathbf{diag}\begin{pmatrix}1 & \lambda_n & \lambda_n^*\end{pmatrix} \tag{19}$$

where $\lambda_n = e^{i\xi_n}$ are the eigenvalues of the rotation depicted on the left side of Fig. 10. The norm becomes:

$$\left|\left|\Delta\vec{S}_n\right|\right| = \left|\left|\mathbf{P}_n\cdot(\mathbf{D}_n - \mathbf{I})\mathbf{P}_n^T\vec{S}_{n-1}\right|\right|. \tag{20}$$

With the change-of-basis matrix expressed as:

$$\mathbf{P}_n^T\vec{S}_{n-1} = \begin{pmatrix}\alpha_1 & \alpha_2 & \alpha_3\end{pmatrix}^T \setminus \mathbf{P}_n = \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3\end{pmatrix}. \tag{21}$$

Thus:

$$\mathbf{P}_n\cdot(\mathbf{D}_n - \mathbf{I})\mathbf{P}_n^T\vec{S}_{n-1} = (\lambda_n - 1)\alpha_2\mathbf{u}_2 + (\lambda_n^* - 1)\alpha_3\mathbf{u}_3. \tag{22}$$

The norm becomes:

$$\left|\left|\Delta\vec{S}_n\right|\right|^2 = |\lambda_n - 1|^2(1 - |\alpha_1|^2). \tag{23}$$

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685

as $\{\mathbf{u}_i\}$ are unitary vectors. Now:

$$|\lambda_n - 1|^2 = 2[1-\cos(\xi_n)] = 4[\sin(\xi_n/2)]^2. \tag{24}$$

Thus:

$$\frac{\xi_n}{2} = \arcsin\left(\frac{\left|\left|\Delta\vec{S}_n\right|\right|}{2\sqrt{1-|\alpha_1|^2}}\right). \tag{25}$$

In Fig. 10, are shown geometrical representations of the previous derivations. On the left, is shown a general case of a rotation of ξ around the rotation axis generated by $\vec{u}_{\text{rot}}$. On the right, is the rotation along a parallel, or equivalently, around $\vec{s}_3$, that helps identifying the cases where ξ and 2θ can match.

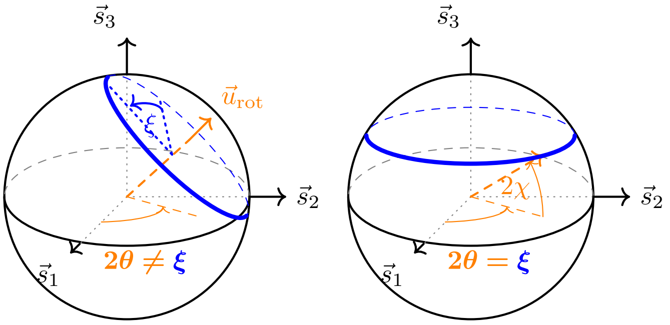


Fig. 10: Displacements on the Poincaré's sphere. On the left: in the general framework of [37]. On the right: in the specific case of $\chi =$ constant (ours).

Thus, Ye's description of SoP change consists in regarding the SoP change as a movement along a small circle of radius (previously R) as depicted in Fig. 10.

Although the authors in [37] highlighted the importance of α_1 for the precision of the polarisation state estimation, they do not comment on its physical or on its geometrical significance. By using Fig. 11, we can show that this factor is actually the projection on the S_3 axis of the rotation axis $\vec{u}_{\text{rot}}$: $|\alpha_1| = |\langle\vec{S}, \vec{u}_{\text{rot}}\rangle|$. Indeed, by noting $R = AB$ the radius of the blue circle (trajectory of the SoP), we have: $\sin(\xi/2) = \frac{BD}{2AB} = \frac{||\Delta\vec{S}||}{2AB}$ & $\sin(\xi/2) = \frac{||\Delta\vec{S}||}{2\sqrt{1-|\alpha_1|^2}} \Rightarrow AB = \sqrt{1-|\alpha_1|^2} \Rightarrow |\alpha_1| = AC = |\langle\vec{S}, \vec{u}_{\text{rot}}\rangle|$.

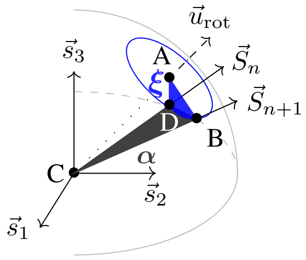


Fig. 11: Linking SoP evolution in the Poincaré sphere using Ye's referential and the usual one in a case of a generic rotation.

4) Inconsistency of the metrics: T. Ye et al. have highlighted a crucial conceptual point: a high rate of channel variation does not necessarily imply a large excursion (radius R previously mentioned) quantified by the "axis angle" in their article of the Stokes vector. They show that the angle θ may rotate rapidly while v_{SoP} remains low, as $||\Delta\vec{S}||$ stays low. Such cases do not affect our results, as our SoP remains near the equator.

These examples underline a key distinction: *SoP changes* (measurable via $\Delta\vec{S}$) do not fully characterise underlying *channel transformations*. A full description would require tracking both angular components (θ, χ) on the Poincaré sphere. Using Eq. (15), a minimum of algebra yields:

$$\dot{\theta} = \frac{\dot{S}_1 S_2 - \dot{S}_2 S_1}{S_1^2 + S_2^2} \;\&\; \dot{\chi} = \frac{\dot{S}_3}{2\sqrt{1-S_3^2}} \tag{26}$$

Moreover, existing metrics—including ours and Ye's—rely on $\|\Delta\vec{S}\|$, the chordal distance in the Poincaré ball, rather than the actual arc length along the sphere. This approximation fails when angular displacements are large (e.g., due to low sampling or fast dynamics). A more accurate description uses the angular speed on the sphere:

$$v_{\text{SoP}} = \|\dot{\vec{u}}_{\vec{S}}\| = \sqrt{(\dot{\chi})^2 + (\cos(2\chi)\dot{\theta})^2}, \tag{27}$$

This, however, remains singular at circular input SoPs ($2\chi = \pm\pi/2$), where the azimuthal coordinate is undefined. To our knowledge, designing a metric that captures all channel dynamics across all regimes remains an open challenge.

B. Probabilistic Constellation Shaping

As this work focuses on the equalisation of PCS (Probabilistic Constellation Shaping) modulation formats, we found relevant to clarify the commonly used terminology regarding the probability law of the symbol amplitude levels. The authors in [40] mentioned that for a given bit rate, the optimal probability law for distributing the symbols to minimise the average energy is the "Maxwell-Boltzmann" (MB) law.

If we denote $R(\Omega)$ as the universe of possible symbols for a given modulation format, and let R be the random variable representing the coordinates of a point in this constellation, with energy $|\vec{r}|^2$, the PDF is written as:

$$p_R(\vec{r}) = \frac{\exp(-\lambda\cdot|\vec{r}|^2)}{\sum_{R(\Omega)}\exp(-\lambda\cdot|\vec{r}|^2)} \setminus \lambda > 0. \tag{28}$$

However, historically and physically speaking, this is not precisely the PDF of the MB law. In brief, one important objective of statistical physics is to count the number of particles $n \in N(\Omega)$ in thermodynamic equilibrium that is function of the energy $\epsilon \in E(\Omega)$. This can be done using the MB law parameterised by λ [41, Chap. 1], with the following PDF:

$$p_N(\epsilon) \equiv \frac{e^{-\lambda\cdot\epsilon}}{Z} \setminus Z \equiv \sum_{\epsilon\in E(\Omega)} e^{-\lambda\cdot\epsilon} \;\&\; \lambda > 0, \tag{29}$$

where Z is called the "partition function". From a mathematical perspective, this PDF is a function of ϵ. Even if the (kinetic) energy of a particle is linked to its speed such that $\epsilon \equiv v^2$, resulting in a numerator resembling $\exp(-\lambda v^2)$, it remains a function of v^2, or equivalently ϵ. This is distinctly different from being a function of v. The most evident discrepancy is the absence of an inflection point in the MB law (as a function of v^2), which is present when considered as a function of v, or equivalently in telecommunications as $|\vec{r}|$. This precision is of importance for physicists familiar with quantum mechanics related topics, such that in the quantum communication domain they avoid using the "Maxwell-Boltzmann" naming, possibly leading to deep misunderstanding, and using the "discrete Gaussian distribution" instead as in [34].

C. Phase noise and laser linewidth

Here, we recall the link between the laser phase noise and the frequency noise. This digression comes from the discrepancy found in terms of the homogeneity of units when we write the variance of the phase noise as: $\sigma^2_{\Delta\phi} = 2\pi \cdot \Delta\nu / R_s$.

We expect the unit of $\sigma^2_{\Delta\phi}$ to be the squared unit of the studied quantity: the phase in radians. However, with the usual writing, we have:

$$[\sigma^2_{\Delta\phi}] = [2\pi] \cdot [\Delta\nu]/[R_s] = \mathrm{rad} \cdot \mathrm{Hz}/\mathrm{Hz} \neq \mathrm{rad}^2. \qquad (30)$$

Recalling the work in [36], we realise that the usual $\Delta\nu$ is actually not the linewidth per se, but rather:

$$\Delta\nu = 2\pi \cdot \sigma^2_{\delta\nu_t}/(2B)\ \mathrm{rad} \cdot \mathrm{Hz}^2/\mathrm{Hz}, \qquad (31)$$

where B is the bandwidth of the spectrum analyser with which the frequency noise is measured. If we note the instantaneous frequency as $\nu_t = \nu_0 + \delta\nu_t$ where $(\nu_0, \delta\nu_t)$ are respectively the central frequency and the frequency variation of the laser emission, the variance of the stochastic frequency process is then $\sigma^2_{\delta\nu_t}$.

D. Practical details for the simulations

1) Additive White Gaussian Circular Noise (AWGCN) : A complex random variable N is circular if $\forall\theta \in [0, 2\pi[$, the probability laws of $N - \mathbb{E}(N)$ and $(N - \mathbb{E}(N))e^{i\theta}$ are identical.

2) Handcrafted phase unwrapping: Here, we explain how the unwrapping of the phase is performed. We developed a customised blind unwrapping function designed specifically for time-domain series with phase jumps. When a phase jumps exceeding $\pi/2$ is identified at a time index k_j, we apply the following correction:

$$\phi_{k \geq k_j} \leftarrow \phi_{k \geq k_j} - \pi \cdot \mathrm{sgn}[\phi_{k_j+1} - \phi_{k_j}] \qquad (32)$$

where sgn denotes the sign function. In Fig. 7a, we show that such phase unwrapping is efficient as there is no phase jump when the phase exceeds ±90 deg.

3) Systemic errors: We assess here the stability of the VAE-FIR across batch sizes through simulations in a controlled scenario (no phase or SoP variations, PCS-64QAM at $R_\mathrm{s} = 64$ GBd, SNR $= 16$ dB, $N_\mathrm{F} = 50$ frames). This SNR corresponds to the pre-FEC BER threshold of $2.8 \cdot 10^{-2}$ reported in [42], below which post-FEC errors become negligible. Symbol-to-bit mapping is assumed Gray-coded, and BER is approximated by BER $\approx$ SER$/\mathcal{H}$. Each frame contains around $20 \cdot 10^3$ symbols, adjusted slightly with $N_{\mathrm{Sb/B}}$ to preserve an integer number of batches. To avoid bias from training transients, BER is averaged over the last $2/3$ of frames. Systemic errors are estimated using standard deviations over this range:

$$\sigma(\{\mathrm{BER}_k\}) = 4 \cdot 10^{-4},\ \big(\sigma(\{\theta_k\}); \sigma(\{\phi_k\})\big) = (30; 5) \cdot 10^{-3}\ \mathrm{deg},$$

with $k \in [\![1/3, 1]\!] \cdot N_{\mathrm{F,ch}}$. These values confirm that the intrinsic noise is well below the FEC threshold and does not affect our conclusions. No dependency on $N_{\mathrm{Sb/B}}$ was observed in this regime.

4) Physics aware training: As mentioned before, we model SoP changes frame-wise even though its associated characteristic time may be of the scale of a batch (or shorter). Indeed, we assume the implemented effects to be commutative as they are linear, which should guarantee the equivalence provided the variance of the processes is changed accordingly. More formally: $\sigma^{\mathrm{frame}}_{\mathrm{pol}} = \sqrt{N_{\mathrm{B/F}}} \cdot \sigma^{\mathrm{batch}}_{\mathrm{pol}}$. In the following, we detail the followed procedure to ensure similar variations of SoP or phase across the different possible setups and their commutativity.

5) Ensuring fair comparison of scenarios: To ensure fair comparisons between CMA and VAE-FIR, we tuned the number of symbols per frame to optimise performance. This tuning, like the symbol rate R_s, directly impacts the variance of the polarisation evolution. Therefore, to evaluate BER variations with respect to channel dynamics, we must ensure equivalent total polarisation (or phase) excursions across configurations. To achieve this, the number of frames per simulation $N_{\mathrm{F,ch}}$ is adjusted such that the expected absolute SoP or phase shift after $N_{\mathrm{F,ch}}$ frames remains constant, regardless of R_s, symbols per frame, or batch size. This expectation is computed from the half-Normal distribution followed by $|\theta_{N_{\mathrm{F,ch}}}|$ or $|\phi_{N_{\mathrm{F,ch}}}|$, yielding:

$$\mathbb{E}(|\{\theta \text{ or } \phi\}_{N_{\mathrm{F,ch}}}|) = 2\sqrt{N_{\mathrm{F,ch}} N_{\mathrm{Sb/F}} \cdot \frac{\{f_{\mathrm{pol}} \text{ or } \Delta\nu\}}{R_\mathrm{s}}} \qquad (33)$$

$$\&\ \mathbb{E}(|\theta_{N_{\mathrm{F,ch}}}|) = \sqrt{\frac{2}{\pi} N_{\mathrm{F,ch}} N_{\mathrm{Sb/F}} \cdot \frac{V_{\mathrm{ch}}(N_{\mathrm{F,ch}} \cdot T_\mathrm{F})}{R_\mathrm{s}}} \qquad (34)$$

Finally, to ensure diversity and mitigate the "seed effect" that could bias the results, the random sequences should be generated with different random seeds for each realisation of a given channel, as done in this work.

REFERENCES

[1] J. Cho and J. P. Winzer, "Probabilistic constellation shaping for optical fiber communications," *J. Lightwave Technol.*, vol. 37, no. 6, pp. 1590–1607, 2019.

[2] H. Sun *et al.*, "800g DSP ASIC design using probabilistic shaping and digital sub-carrier multiplexing," *J. Lightwave Technol.*, vol. 38, no. 17, pp. 4744–4756, 2020.

[3] R. Johnson, P. Schniter, T. Endres, J. Behm, D. Brown, and R. Casas, "Blind equalization using the constant modulus criterion: a review," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 1927–1950, 1998.

[4] V. Lauinger *et al.*, "Blind equalization and channel estimation in coherent optical communications using variational autoencoders," *J. Selected Areas in Communications*, vol. 40, no. 9, pp. 2529–2539, 2022.

[5] M. Kong, J. Shi, B. Sang, J. Ding, K. Wang, W. Li, F. Wang, C. Liu, Y. Wang, Y. Wei, B. Zhu, L. Zhao, W. Zhou, and J. Yu, "800-gb/s/carrier WDM coherent transmission over 2000 km based on truncated ps-64QAM utilizing MIMO volterra equalizer," *J. Lightwave Technol.*, vol. 40, no. 9, pp. 2830–2839, 2022.

This article has been accepted for publication in Journal of Lightwave Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JLT.2025.3603685

[6] l. Tomczyk *et al.*, "Blind state of polarisation monitoring using variational autoencoders-inspired adaptive filter," in *European Conference on Optical Communication*, 2023.

[7] J. Barry, *Equalization*, S. K. Wilson, S. Wilson, and E. Biglieri, Eds. Academic Press, 2016.

[8] M. Kuschnerov *et al.*, "Data-aided versus blind single-carrier coherent receivers," *IEEE Photonics Journal*, vol. 2, no. 3, pp. 387–403, June 2010.

[9] K. Banovic, E. Abdel-Raheem, and M. Khalid, "A novel radius-adjusted approach for blind adaptive equalization," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 37–40, 2006.

[10] G. Giunta and F. Benedetto, "A signal processing algorithm for multi-constant modulus equalization," in *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, 2013, pp. 52–56.

[11] J. Yang *et al.*, "Joint estimation of dynamic polarization and carrier phase with pilot-based adaptive equalizer in PDM-64 QAM transmission system," *Opt. Express*, vol. 29, no. 26, pp. 43 136–43 147, Dec 2021.

[12] M. P. Yankovn *et al.*, "Constellation shaping for WDM systems using 256QAM/1024QAM with probabilistic optimization," *J. Lightwave Technol.*, vol. 34, no. 22, pp. 5146–5156, Nov 2016. [Online]. Available: https://opg.optica.org/jlt/abstract.cfm?URI=jlt-34-22-5146

[13] X. Li *et al.*, "Polarization demultiplexing scheme for probabilistic shaping stokes vector direct detection system using extended kalman filter," *Optics Communications*, vol. 461, p. 125192, 2020.

[14] M. Farsi *et al.*, "Polarization tracking in the presence of pdl and fast temporal drift," *J. Lightwave Technol.*, vol. 40, no. 19, pp. 6408–6416, 2022.

[15] T. Pfau, S. Hoffmann, and R. Noe, "Hardware-efficient coherent digital receiver concept with feedforward carrier recovery for m -QAM constellations," *J. Light. Tech.*, vol. 27, no. 8, pp. 989–999, 2009.

[16] Q. Zhang and C. Shu, "Viterbi and viterbi algorithm based phase recovery for probabilistically shaped signals," *J. Light. Tech.*, vol. 39, no. 5, pp. 1364–1370, 2021.

[17] M. S. Neves *et al.*, "Carrier-phase recovery for coherent optical systems: Algorithms, challenges and solutions," *J. Lightwave Technol.*, vol. 42, no. 3, pp. 1095–1108, 2023.

[18] A. Sano *et al.*, "102.3-tb/s (224 × 548-gb/s) c- and extended l-band all-raman transmission over 240 km using pdm-64qam single carrier fdm with digital pilot tone," in *OFC/NFOEC*, Los Angeles, CA, USA, 2012, pp. 1–3.

[19] S. M. Bilal and G. Bosco, "Pilot tones based polarization rotation, frequency offset and phase estimation for polarization multiplexed offset-qam multi-subcarrier coherent optical systems," in *2016 18th International Conference on Transparent Optical Networks (ICTON)*, Trento, Italy, 2016, pp. 1–4.

[20] A. Caciularu and D. Burshtein, "Blind channel equalization using variational autoencoders," in *Proceedings of International Conference on Communications*, 2018.

[21] l. Tomczyk *et al.*, "Source code." [Online]. Available: https://gitlab.telecom-paris.fr/elie.awwad/vae-fir

[22] C. Czegledi *et al.*, "Polarization drift channel model for coherent fibre-optic systems," *Scientific Reports*, vol. 6, p. 21217, 2016.

[23] C. Laot and N. Le Josse, "A closed-form solution for the finite length constant modulus receiver," in *Proceedings. International Symposium on Information Theory*, 2005, pp. 865–869. [Online]. Available: https://doi.org/10.48550/arXiv.cs/0508126

[24] E. P. da Silva and D. Zibar, "Widely linear equalization for IQ imbalance and skew compensation in optical coherent receivers," *J. Lightwave Technol.*, vol. 34, no. 15, pp. 3577–3586, 2016.

[25] J. C. Geyer *et al.*, "Channel parameter estimation for polarization diverse coherent receivers," *Photon. Technol. Lett.*, vol. 20, no. 10, pp. 776–778, 2008.

[26] F. N. Hauske *et al.*, "Optical performance monitoring in digital coherent receivers," *J. Lightwave Technol.*, vol. 27, no. 16, pp. 3623–3631, 2009.

[27] P. A. Nwakamma, "Efficient adaptive equalization for pmd mitigation in next-generation optical access networks," Ph.D. dissertation, Télécom Paris, 2024, thèse de doctorat dirigée par Ware, Cédric et Jaouën, Yves Réseaux, informations et communications Télécom Paris 2024. [Online]. Available: http://www.theses.fr/2024IPPAT042

[28] E. Awwad *et al.*, "Detection-localization-identification of vibrations over long distance ssmf with coherent $\delta\phi$-OTDR," *J. Lightwave Technol.*, vol. 38, pp. 3089–3095, 2020.

[29] F. Pittalà *et al.*, "Training-aided frequency-domain channel estimation and equalization for single-carrier coherent optical transmission systems," *J. Lightwave Technol.*, vol. 32, no. 24, pp. 4849–4863, 2014.

[30] L. P. Cinelli *et al.*, *Variational methods for machine learning with applications to deep networks*. Springer, 2021.

[31] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *CoRR*, vol. abs/1906.02691, 2019. [Online]. Available: http://arxiv.org/abs/1906.02691

[32] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[33] V. Raj and S. Kalyani, "Design of communication systems using deep learning: A variational inference perspective," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 4, pp. 1320–1334, 2020.

[34] A. Denys, P. Brown, and A. Leverrier, "Explicit asymptotic secret key rate of continuous-variable quantum key distribution with an arbitrary modulation," *Quantum*, vol. 5, p. 540, Sep. 2021.

[35] E. M. Liang and J. M. Kahn, "Probabilistic shaping distributions for optical communications," *J. Lightwave Technol.*, pp. 1–24, 2025.

[36] D. S. Elliott *et al.*, "Extracavity laser band-shape and bandwidth modification," *Phys. Rev. A*, vol. 26, no. 1, pp. 12–18, 1982.

[37] T. Ye, X. Su, K. Zhang, Z. Tao, G. Huang, H. Nakashima, and T. Hoshida, "A polarization change monitor by eigenvalue analysis in coherent receiver," in *Optical Fiber Communications Conference and Exhibition*, 2019, pp. 1–3.

[38] F. Boitier *et al.*, "Proactive fiber damage detection in real-time coherent receiver," in *European Conference on Optical Communication*, 2017.

[39] J. E. Simsarian and P. J. Winzer, "Shake before break: Per-span fiber sensing with in-line polarization monitoring," in *Optical Fiber Communications Conference and Exhibition*, 2017, pp. 1–3.

[40] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for gaussian channels," *Transactions on Information Theory*, vol. 39, no. 3, pp. 913–929, 1993.

[41] G. Chardin *et al.*, *Éléments de physique statistique*, 2nd ed. Les presses de l'Ensta, 2015. [Online]. Available: https://hal.science/hal-01141421

[42] L. Zhang *et al.*, "Real-time FPGA investigation of interplay between probabilistic shaping and forward error correction," *J. Lightwave Technol.*, vol. 40, no. 5, pp. 1339–1345, 2022.