DTU Library

Channel fading-robust pre-coders for low complexity few-mode IM/DD short reach fiber communications

Yankov, Metodi Plamenov; Da Ros, Francesco; Forchhammer, Søren; Gruner-Nielsen, Lars

Published in:
Journal of Lightwave Technology

Link to article, DOI:
10.1109/JLT.2023.3325157

Publication date:
2024

Document Version
Peer reviewed version

Channel fading-robust pre-coders for low complexity few-mode IM/DD short reach fiber communications

Metodi Plamenov Yankov *Senior Member, IEEE, Member, Optica*, Francesco Da Ros *Senior Member, IEEE, Member, Optica*, Søren Forchhammer, *Member, IEEE*, Lars Grüner-Nielsen *Member, IEEE*

***Abstract*—In this paper, an autoencoder (AE) is used to optimize linear and nonlinear digital pre-coders for short reach few-mode fiber transmission. Mach-Zehnder modulation is used to modulate the intensity of the optical carriers, and direct detection is employed at the receiver. The modes are (de)multiplexed using photonic lanterns with air cladding. The AEs are optimized to be robust against cross-talk (XT) and distance variations, which result in time varying power fading across the different modes. The system is exemplified using a simulation of a graded index fiber with the mode groups LP_{01} and LP_{11} used for data transmission. The channel and (de)multiplexer XT parameters and fading distributions are obtained from experimental measurements. In the case of unknown distance, the distance-robust pre-coding allows for on-off keying transmission with an error rate below the chosen forward error correction threshold even at severe XT expectations and variations. In the case of a fixed-length system, pre-coding enables high order modulation such as 4PAM and 8PAM. In this case, nonlinear pre-coding achieves signal-to-noise ratio gains of up to 1.5 dB w.r.t. linear pre-coding and up to 4 dB w.r.t. simple power pre-emphasis.**

***Index Terms*—Few-mode transmission, autoencoder, IMDD, mode mixing, robustness**

I. Introduction

Spatial division multiplexing (SDM) is one of the hottest research topics within the optical fiber communications community. Since it was revealed that standard, single mode fibers (SSMF) have a limited achievable rate due to the fiber Kerr nonlinearity [1], [2], researchers have been exploring novel approaches to dramatically increase the optical network capacity in order to cope with the constantly increasing data rate demands. Exploring the spatial dimension for throughput multiplicity through SDM is one of those approaches. In fibers, SDM can be realized by 1) increasing the number of parallel SSMF connections; 2) adopting multi-core fibers; or 3) exploiting the multiple available orthogonal geometric distributions of the electric field in a waveguide, known as modes. Intricacies and details of each approach have been the subject of extensive studies in the last decade and a summary may be found in [3]. In this paper, the case of multi-mode transmission in optical fibers is exclusively discussed. Such cases find themselves not only as potential solutions to the capacity crunch problem, but also as low-cost alternatives to wavelength division multiplexing (WDM) currently employed in passive optical networks (PONs), e.g. intra-data center (DC) connections and/or last-mile solutions for fiber to the home. For example, current standards for 400 Gbps Ethernet typically employ 4-wavelength WDM, which requires 4 sets of relatively frequency-stable optical frontends with optical filtering in order to avoid cross-talk between the wavelengths. In such networks, cost and energy consumption are key performance indicators. The much cheaper multi-mode fiber (MMF) which supports multi-mode and few-mode transmission can potentially be used as an alternative [4]. In addition to the fiber cost benefits, all channels can potentially be operated near the zero-dispersion wavelength of the fiber (which improves the performance), and without applying filtering at the receiver, which further reduces the system cost. In order to further restrict the cost and power consumption of the transceivers, such links employ intensity-modulation and direct detection (IMDD) with e.g. 4-pulse amplitude modulation (PAM), which requires minimal digital signal processing (DSP) functions for detection.

One challenge in employing few-mode transmission is multiplexing and demultiplexing the transmission modes into and out of the same fiber. For that purpose, several types of mode multiplexers may be employed [5], e.g. air clad photonic lanterns [6], demonstrating 100G transmission over up to 2 km for 2-mode multiplexing. The resulting cross-talk (XT) can be combated with the application of multiple input multiple output (MIMO) processing at the transmitter - e.g. optically [7], and/or receiver - e.g. optically [8] for IMDD link and the more classical digital MIMO used for long-haul transmission [3]. Due to the cost requirement, MIMO and synchronization of the modes at the receiver is typically prohibitive. Instead, the multiplexer (MUX) and demultiplexer (DEMUX) are required to produce negligible XT, which sets a harsh requirement on their fabrication process.

This paper extends [9], where the IMDD optical fiber fading multiple input multiple output (MIMO) channel was studied. In [9], a simplified channel model was considered for the interference between modes which only assumed the interference to present itself as an incoherent power transfer. This model allowed for the computation of channel capacity by building on previous works on the capacity of free-space optical MIMO links. Furthermore, in [9], an autoencoder (AE) was proposed for pre-coding of the spatial streams in order to lift the requirement for power hungry DSP and complicated optical receiver frontends. In this paper, a more fundamental

The authors are with the Department of Electrical and Photonics Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark. e-mail: meya@dtu.dk

channel model is assumed which takes into account the phase beating of the optical carriers in different modes. The effect of this model on the frequency selectivity and coherence time of the XT is discussed together with methods for its mitigation within the AE framework. Furthermore, the toy model for the polarization dependence of the mode XT is replaced by a more mathematically rigorous model [10] and supplemented with XT measurements in order to approach a realistic lab scenario.

The rest of the paper is organized as follows. In Section II, the new MIMO model for the IMDD few-mode transmission is presented, together with the improved models for the expected uncertainty in the channel matrix. In Section III, the AE concept is described in details, especially how robustness in the uncertainty can be imposed on the pre-coder. In Section IV, the performance benefits of the AE-based pre-coder are presented in terms of signal to noise ratio (SNR) by using the end-to-end bit error rate (BER) and achievable information rate (AIR). Section V offers some discussion points, in particular regarding channel capacity and possible extensions of this work. Section VI concludes the paper.

A. Notation

The following notation is used, unless otherwise specified. Lower-case x represents realization of the random variable X. Matrices are represented with capital bold face letters, e.g. $\mathbf{H}$. Vectors are represented using lower-case bold-face $\mathbf{x}$. Other notations are defined on occurrence.

II. Channel model

First, the general multi-mode transmission is described. Then, a 2-mode multiplexing system is used as example. The aim is to transmit spatially multiplexed data with a minimal amount of complexity. Thus, polarization multiplexing is not performed. Similarly, multiplexing data in degenerate modes of the same group is not performed. The transmission system is given in Fig. 1. Independent data on each mode group of the few-mode transmission is considered. Only one polarization of one degenerate mode in each mode group is excited. The data bits modulate the waveform using pulse amplitude modulation (PAM), where the modulation alphabet is $\mathcal{X} = 0, 1, ...M-1$ of size $|\mathcal{X}| = M$. The symbols are independent in time and the signal vector is denoted $\mathbf{x} = [x_1, x_2, ...x_N]^T$, where N is the number of employed modes. The signal may then be pre-coded by a linear or nonlinear function f, $\mathbf{v} = f(\mathbf{x})$. A Mach-Zehnder modulator (MZM) is employed for modulating the intensity of optical carriers of wavelengths $\lambda = \{\lambda_1, \lambda_2, ...\lambda_N\}$ with the corresponding pre-coded symbols. The MZM is perfectly biased in quadrature, infinite extinction ratio is assumed, and its output field E_n^{MZM} (denoted s_n for simplicity) as a function of the modulating voltage v_n can be modeled as

$$E_n^{MZM} = s_n = \sqrt{P_n} \cdot e^{-j\omega_n t + \phi_n(t)} \cdot \cos\left(\frac{\pi}{2}\frac{v_n}{V_\pi}\right), \quad (1)$$

where V_π is the half-wave voltage of the modulator, and P_n is the optical power of the $n-$th carrier. Furthermore in (1), $\omega_n = 2\pi c/\lambda_n$ is the carrier frequency of the $n-$th laser, t represents time and $\phi_n(t)$ is a phase noise term. The phase noise is modeled using a Wiener process.

An example of a MUX is the air clad photonic lantern made of single mode fibers, which are fused together and down tapered. The end of the taper is spliced to a few-mode fiber. After down tapering the surrounding air will act as the cladding. By use of single mode fibers with different core sizes mode selectivity is obtained [6], [4]. The MUX is spliced to an MMF. At the end of transmission, the MMF is spliced to a DEMUX which demultiplexes the modes into independent spatial paths. Each mode is detected by a photodetector (PD), which translates light intensity to electrical current. Optical filtering is not applied in order to simplify the system, which means that optical power from the interfering mode group is present at the mode group of interest regardless of their carrier frequencies. The modes in the same group are added electrically. The electrical noise terms $\mathbf{w} = [w_1, w_2, ...w_N]^T$ are independent, identically distributed samples from a Gaussian distribution with a variance σ_w^2. The noise is assumed dominated by the receiver optical front-end noise, as well as additive noise due to an electrical amplifier. The signals of the degenerate modes in the same group are electrically added together to form a common data received path. The signal at the receiver is denoted $\mathbf{y} = [y_1, y_2, ...y_N]^T$.

A. True channel model

The complete fiber channel model is found in [7] and references therein. The total propagated distance is discretized into $k = 1..K$ sections. Each section k is composed of a unitary transformation $\mathbf{Q}_k(\omega)$ and propagation

$$\begin{aligned}\mathbf{\Lambda}_k(\omega) = diag\{&e^{j(\omega-\omega_0)\cdot 0.5\cdot\Delta\tau z_k},\\ &e^{j(\omega-\omega_0)\cdot 0.5\cdot\Delta\tau z_k},\\ &e^{j(\omega-\omega_0)\cdot(-0.5\cdot\Delta\tau) z_k},\\ &e^{j(\omega-\omega_0)\cdot(-0.5\cdot\Delta\tau) z_k},\\ &e^{j(\omega-\omega_0)\cdot(-0.5\cdot\Delta\tau) z_k},\\ &e^{j(\omega-\omega_0)\cdot(-0.5\cdot\Delta\tau) z_k}\},\end{aligned} \quad (2)$$

where $\Delta\tau$ is the differential group delay (DGD) between the 2 modes, z_k is the length of the section and the following assumptions are made:

- The propagation constants of polarization modes in the same mode group are identical;
- The propagation constants are not frequency dependent within the signal bandwidth (although they do depend on the carrier frequency);
- Chromatic dispersion can be neglected (due to the short considered distance of PONs);
- Kerr nonlinearities can be neglected (due to the short distance and relatively low power);

The [6x6] fiber transfer function is then found as

$$\mathbf{H}_{FIBER}(\omega) = \prod_{k=1:K} \mathbf{Q}_k(\omega)\mathbf{\Lambda}_k(\omega). \quad (3)$$

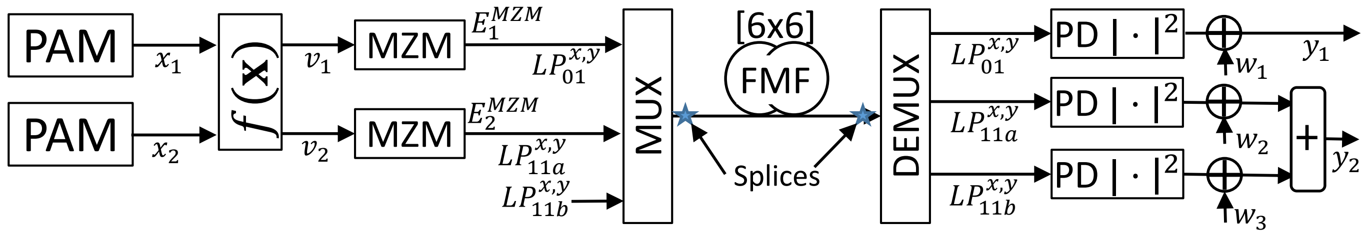

Fig. 1. Two-mode group multiplexing system model.

B. Equivalent model

The following assumptions can be made to simplify the transmission model:

- The distance is so short, that inter-modal crosstalk inside the fiber is negligible[1]. This results in $\mathbf{Q}_k(\omega)$ being block-diagonal, where each block is also unitary;
- For the low-order modes and up to the 10s of GHz baudrates considered here, the coherence bandwidth is much larger (up to 100s of GHz) than the transmission rate [7], which makes $\mathbf{\Lambda}$ and $\mathbf{Q}$ frequency independent within the signal bandwidth;

The block-diagonal structure of the unitary matrix can be described as

$$\mathbf{Q}_k = \begin{bmatrix} \mathbf{Q}_k^{LP_{01}} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_k^{LP_{11}} \end{bmatrix}, \quad (4)$$

where $\mathbf{Q}_k^{LP_{01}}$ and $\mathbf{Q}_k^{LP_{11}}$ are [2x2] and [4x4] unitary matrices, respectively, and $\mathbf{0}$ are all-zero matrices of appropriate sizes. Their product is then

$$\mathbf{Q}_{TOT} = \prod_k \mathbf{Q}_k = \begin{bmatrix} \mathbf{Q}_{TOT}^{LP_{01}} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{TOT}^{LP_{11}} \end{bmatrix}. \quad (5)$$

Together with the assumption on the propagation constants above, the fiber transfer function can be described as

$$\mathbf{H}_{FIBER} = \mathbf{\Lambda}_{TOT} \cdot \mathbf{Q}_{TOT}, \quad (6)$$

where $\mathbf{\Lambda}_{TOT} = \prod_k \mathbf{\Lambda}_k$ is the total propagation matrix for the full fiber propagation distance.

Under these assumptions, the inter-polarization and inter-mode group rotation given by $\mathbf{Q}_{TOT}^{LP_{01}}$ and $\mathbf{Q}_{TOT}^{LP_{11}}$ does not influence the total power delivered to the receiver PDs. The [6x6] fiber model from (3) can then be simplified to the equivalent time-domain [2x2] model

$$\mathbf{H}_{FIBER}^{Equiv} = \begin{bmatrix} e^{-j\beta^{(1)}\cdot z} & 0 \\ 0 & e^{-j\beta^{(2)}\cdot z} \end{bmatrix}, \quad (7)$$

where $\beta^{(n)}$ is the propagation constant for the $n-$th mode and z is the total distance traveled.

In this equivalent [2x2] system model, each of the MUX, DEMUX and splices can be modeled by a MIMO matrix $\mathbf{H}_{MUX}$, $\mathbf{H}_{DEMUX}$ and $\mathbf{H}_{SPL}$, respectively, which have the form (for example for $\mathbf{H}_{MUX}$)

$$\mathbf{H}_{MUX} = \begin{bmatrix} \sqrt{(1-xt_1)/\alpha_1} & \sqrt{xt_2/\alpha_2} \\ \sqrt{xt_1/\alpha_1} & \sqrt{(1-xt_2)/\alpha_2} \end{bmatrix}, \quad (8)$$

[1]For justification, the inter-modal crosstalk is estimated to be $> 80dB$ lower than intra-modal crosstalk for links longer than the ones considered here according to [11, Fig. 5].

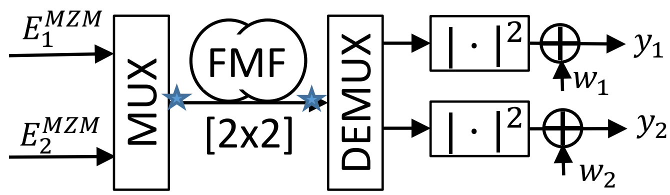

Fig. 2. Equivalent two-mode group system model.

where xt_1 and xt_2 are the XTs from mode group LP_{01} to mode group LP_{11} (LP_{11} to LP_{01}, respectively), and α_1 and α_2 are the losses of the corresponding modes in the component. The latter can potentially be different due to slight difference in the connected single mode fibers and taper inaccuracies for the two modes.

It is beneficial to define a phase noise matrix

$$\mathbf{H}_{PN} = \begin{bmatrix} e^{-j\omega_1 t+\phi_1(t)} & 0 \\ 0 & e^{-j\omega_2 t+\phi_2(t)} \end{bmatrix}, \quad (9)$$

and the phase noise-less field output of the MZM as

$$s_n = \sqrt{P_n} \cdot \cos\left(\frac{\pi}{2}\frac{v_n}{V_\pi}\right). \quad (10)$$

In that case, the time-varying end-to-end channel matrix can be defined as

$$\mathbf{H} = \mathbf{H}_{DEMUX} \cdot \mathbf{H}_{SPL} \cdot \mathbf{H}_{FIBER}^{Equiv} \cdot \mathbf{H}_{SPL} \cdot \mathbf{H}_{MUX} \cdot \mathbf{H}_{PN}, \quad (11)$$

and the end-to-end channel model, as given in Fig. 2 as

$$\mathbf{y} = (\mathbf{Hs}) \odot (\mathbf{Hs})^* + \mathbf{w}, \quad (12)$$

where $\mathbf{s} = [s_1, s_2, ...s_N]^T$, $\odot$ denotes the Hadamard product of two vectors, $()^*$ denotes complex conjugation, and the noise variances are σ_w^2 and $2\sigma_w^2$ for LP_{01} and LP_{11}, respectively

The MZM transfer function imposes an amplitude constraint on the signal $0 \leq v_n \leq V_\pi$, since outside of these boundaries the cosine transfer function becomes non-bijective. Equivalently, the channel in (11) can be considered the standard MIMO IMDD channel [12], [13] with input per mode $\hat{s}_n = s_n^2$ and amplitude constraint $0 \leq \hat{s}_n \leq P_n$. In the case of fiber communications and for IMDD links operating at short reach, there is no practical constraint (within reason) associated with the average power, unlike for example free-space optical wireless links, where safety and interference may be factors. The signal to noise ratio (SNR) is defined using the total transmitted power and total electrical noise power as $SNR = 10\log_{10}\left(\frac{P_1+P_2}{\sqrt{3}\cdot\sigma_w}\right)$ (factor 3 stemming from the 3 employed PDs at the receiver). This is opposed to usual definitions relative to σ_w^2, motivated by the fact that the white noise affects the power of the signal and not its amplitude.

C. Channel fading distribution

Even when the distributed XT along the fiber is neglected, the channel is still subject to variations, also known as fading. The fading is due to two main effects:

1) The cross talk XT in the MUX and DEMUX are polarization dependent, and the polarization drifts randomly;
2) The phase at the output of the fiber depends on the fiber length, the laser phase, $\beta^{(1)}$ and $\beta^{(2)}$. The phase is therefore subject to variation due to different distance between transmitter and receiver at different sites, environmental impact from temperature variations and vibrations, as well as phase noise and carrier wavelengths drift over time.

In order to model the first effect, we adopt the model from [10], according to which the XT in an MMF is distributed according to a χ^2 distribution with $2g_l$ degrees of freedom, mean $\mathbb{E}[XT]$ and a standard deviation given by $\mathbb{E}[XT]/\sqrt{2g_l}$, where g_l is the number of modes in the interfering mode group. In the 2-mode case, the LP_{01} mode has $g_l = 1$ and the LP_{11} mode has $g_l = 2$ (discussion on the validity of this distribution is provided in Section IV-A and Section V). The second effect will require that the transmission is robust to random received phases and thereby random fading.

In this paper, a block-fading channel is assumed, where a new value of the XT and total propagation phase per mode are drawn for each block of 1000 symbols. At a symbol rate on the order of GBd, this corresponds to an assumption of a polarization drift slower than 1 MHz. The uncertainty of the distance and wavelength is modeled by applying a random fiber length, uniformly distributed in the range $z \in [0; 2000]$m, and a random carrier frequency offset of each carrier in the range $f_{off} \in [-1.5; 1.5]$ GHz[14] in each block.

D. Mutual information

In this paper, the performance will be measured mostly by using the mutual information $\mathcal{I}(X;Y)$ between the channel input and output of the *worst performing mode*. That is, the worst case scenario will be reported. Probabilistic shaping is not performed, leading to an entropy of the signal per mode of $\mathcal{H}(X_n) = \log_2(M)$. The formal definition of the worst performing mode is

$$\begin{aligned}
&\min_n \mathbb{E}_H \left[\mathcal{I}(Y_n; X_n|\mathbf{H}, P_{\{1\dots N\}})\right] = \\
&\min_n \mathcal{H}(X_n) - \mathbb{E}_H \left[\mathcal{H}(X_n|Y_n, \mathbf{H}, P_{\{1\dots N\}})\right] = \\
&\log_2(M) - \max_n \mathbb{E}_H \left[\mathcal{H}(X_n|Y_n, \mathbf{H}, P_{\{1\dots N\}})\right], \qquad (13)
\end{aligned}$$

where $P_{\{1\dots N\}}$ is the power allocation per mode and $\mathcal{H}(X_n|Y_n, \mathbf{H}, P_{\{1\dots N\}}) = \mathbb{E}_p\left[-\log_2 p(X_n|Y_n, \mathbf{H}, P_{\{1\dots N\}})\right]$ is the conditional entropy. The motivation for using the worst mode as a performance indicator instead of the joint MI $\mathbb{E}_H\left[\mathcal{I}(\boldsymbol{Y}; \boldsymbol{X}|\mathbf{H}, P_{\{1\dots N\}})\right]$ is two-fold: 1) it better reflects the scenario where low-complexity, independent processing per mode is performed at the receiver: and 2) it better reflects the performance in potential cases where the link is point to multi-point, e.g. a multi-user scenario, in which case fairness between users dictates optimization w.r.t. the worst one.

In cases where the channel law $p(Y_n|X_n, \mathbf{H}, P_{\{1\dots N\}})$ and thus the posterior distribution $p(X_n|Y_n, \mathbf{H}, P_{\{1\dots N\}})$ is not perfectly known at the receiver, an auxiliary distribution $q(X_n|Y_n, \mathbf{H}, P_{\{1\dots N\}})$ will be assumed for calculating the MI in (13). By replacing p by q in the conditional entropy expression, an upper bound on the entropy is obtained

$$\begin{aligned}
&\mathbb{E}_p\left[-\log_2 p(X_n|Y_n, \mathbf{H}, P_{\{1\dots N\}})\right] \leq \\
&\mathbb{E}_p\left[-\log_2 q(X_n|Y_n, \mathbf{H}, P_{\{1\dots N\}})\right], \qquad (14)
\end{aligned}$$

leading to a lower bound on the MI. Any valid distribution q may be used to obtain this lower bound, with equality when $p = q$ [15]. Consequently, applying q for the detection of the signals during transmission results in a limit to the maximum achievable error-free data rate with infinite block lengths, known as the achievable information rate (AIR). Two different choices for an auxiliary channel function are discussed in Section III.

III. AUTOENCODER

As mentioned in the introduction, full MIMO processing may be too complex for the target communication system because of both hardware (in terms of synchronizing the receivers for each mode) and software (in terms of processing for channel estimation and multi-dimensional detection) requirements. However, partial channel knowledge may still be available in terms of the distribution of $\mathbf{H}$ at the transmitter and receiver, which can be obtained e.g. from measurement statistics. For such cases, an auxiliary channel likelihood function must be employed for detection and MI estimation, which is subject to optimization. The likelihood function needs to be robust to the distribution of $\mathbf{H}$. At the same time, the knowledge of $\mathbf{H}$ may allow to design a pre-coder which at least partially mitigates the mode interference. Both of these optimizations can be efficiently approached with an autoencoder (AE) [16]. The AE employs a neural network (NN) for pre-coding, another NN for detection, and wraps them around the channel model at hand. The most closely related to this work AE application is for the Rayleigh fading MIMO channel [17], where the channel is linear. In this paper, both the MZM and the PDs induce nonlinearities. Furthermore, in this paper, the AE is employed to also optimize the power allocation. The alternative to the latter is to perform a grid search, which is still doable for small N, but quickly becomes infeasible when N grows. A basic schematic of the AE is given in Fig. 3. The set of training parameters $\mathcal{W}$ is the set of encoder NN weights $\mathcal{W}_E$, decoder NN weights $\mathcal{W}_D$ and the power allocations $P_{\{1\dots N\}}$.

Three different types of pre-coders are considered:

1) **Power allocation**: The pre-coder is a an identity matrix multiplied by a vector containing the square root of the desired power in each mode;
2) **Linear pre-coding**: The pre-coder is an NxN matrix;
3) **Nonlinear pre-coding**: The pre-coder is a multi-layer NN.

A. Cost function and robust realization

The last layer of each decoder NN is a soft-max layer, and its outputs are posterior probabilities $q_{NN}(X_n|Y_n)$. They

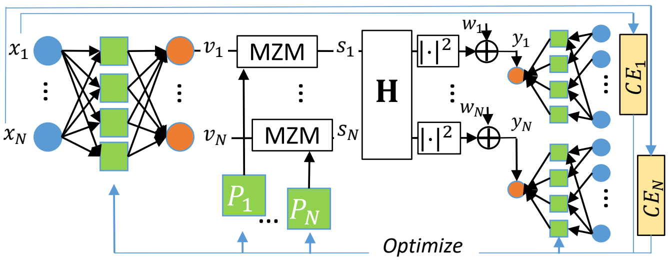

Fig. 3. Autoencoder model for optimization. Showcased with independent detector NN per layer.

TABLE I
OPTIMIZED TOPOLOGIES IN TERMS OF NODES PER LAYER OF THE AE COMPONENTS FOR $N = 2$. ALL NNS ARE BIASED.

	8PAM	4PAM	OOK
Lin. Prec.	[2x2]	[2x2]	[2x2]
Nonlin. Prec.	[2x16x16x2]	[2x8x8x2]	[2x4x4x2]
Decoder per mode	[1x64x64x8]	[1x16x16x4]	[1x4x4x2]

can therefore be used to calculate an upper bound on the conditional entropy in (14). Another term for this bound is the cross entropy (CE) and is a very popular cost function for optimization of the NN, especially when the NN operates as a classifier. This is typically the case for telecommunications. In this paper, the CE is used as a cost function. Furthermore, in order to always optimize the CE of the worst performing mode, the cost function $L(\mathbf{w} \in \mathcal{W})$ is always the maximum of the CEs across the modes

$$L(\mathbf{w}) = \max_n \mathbb{E}_{H,k} \left[-\log_2 q_{NN}(x_n(k)|y_n(k), \mathbf{H}, P_{\{1\dots N\}}) \right], \tag{15}$$

where $x_n(k)$ and $y_n(k)$ are the samples on the $n-$th mode at time k, respectively, and the expectation is over both time and channel distribution.

For all detectors and pre-coders, the NN topologies were optimized by increasing their depth and width by factors of 2 until convergence of the cost function. The resulting topologies for $N = 2$ are given in Table I. The last layer of the detectors is a classifier, and thus has output size of M. All NNs employ the ReLU activation function defined as $ReLU(x) = \max[x, 0]$ and all NNs are biased. The Adam optimizer was employed for stochastic gradient descent optimization of all parameters. The batch size was 1000, and a total of 1.000.000 symbols are used for training and independent 1.000.000 for testing. The testing symbols are also separated into batches of 1000 in order to emulate the block-fading scenario described above. In each batch (training and testing), the channel is drawn from its distribution. The result is an AE robust to changes in the channel. Observe, no explicit constraint is imposed on the encoder part. The AE is naturally forced to operate within the boundaries of the cosine transfer function of the MZM from Eq. (1).

B. Gaussian auxiliary channel for detection

In this paper, the AIR is also estimated using a more conventional Gaussian auxiliary channel

$$q_G(y_n(k)|X_n = \mathcal{X}^j) \propto \exp\left[-\frac{1}{2\hat{\sigma}^2_{w,n}} |y_n(k) - \mu_n^j|^2 \right]. \tag{16}$$

TABLE II
EXPECTED PARAMETERS FOR THE MIMO COMPONENTS.

	$\mathbb{E}[XT_1]$, dB	$\mathbb{E}[XT_2]$, dB	α_1, dB	α_2, dB
MUX	-18	-15	0.7	1.4
DEMUX	-11	-11	1.5	3
SPL	-25	-25	0.04	0.04

The variance of the auxiliary channel is estimated e.g. from training data as $\hat{\sigma}^2_{w,i} = \mathbb{E}_k\left[|y_n(k) - x_n(k)|^2\right]$. Similarly, the means of the auxiliary channel for the $j-$th symbol from the constellation are estimated as $\mu_n^j = \mathbb{E}_{k:x_n(k)=\mathcal{X}^j}\left[y_n(k)\right]$. The Gaussian auxiliary channel provides a better estimate of the rate, achievable by conventional maximum likelihood (or in this case equivalently, minimum distance) receiver.

C. BER estimation

The BER is also reported for completeness. In order to estimate the BER, first, symbol detection is performed using the maximum likelihood decision function e.g. for the Gaussian auxiliary channel

$$\hat{x}_n(k) = \arg\max_j q_G(y_n(k)|X_n = \mathcal{X}^j). \tag{17}$$

Then, Gray labeling is used for mapping of bits to symbols x_n and demapping the symbol decisions $\hat{x}_n$ back to bits, from which the BER can be calculated.

IV. RESULTS

The results are always reported for the *worst performing mode*. The expected cross talk values *of each component* are given in Table II and are obtained from the measurements [4]. The cross talk values are then drawn from their distribution as described in Section II-C.

As shown in [18], a certain offset may be required between the carrier frequencies in the different mode groups. To that end, we consider expected values for $\lambda_1 = 1550$nm and $\lambda_2 = 1550.5$nm. In all cases, $\Delta\tau = 2.1ps/m$, the symbol rate $Rs =$ 10 GBd, and a laser linewidth of 100 kHz is assumed. The AIR results averaged over the entire distribution of $\mathbf{H}$ are given in Fig. 4 for on-off keying (OOK), 4PAM and 8PAM in the first, second and third column, respectively. The top row depicts the MI, and the bottom row depicts the BER. The MI and BER are obtained both from the decoder NN through the CE and corresponding maximum likelihood decisions, and by using an auxiliary Gaussian channel with minimum distance decoding.

In this case, robustness is required to the channel fading. The perceived *end-to-end XT* can be extracted as $XT_1 = |\mathbf{H}_{21}|^2/|\mathbf{H}_{11}|^2$ and $XT_2 = |\mathbf{H}_{12}|^2/|\mathbf{H}_{22}|^2$, where $\mathbf{H}_{ij}$ is the element on the $i-$th row and $j-$th column of the matrix $\mathbf{H}$ obtained as in Eq. (11). The end-to-end XT distributions are given in Fig. 5, with their mean and standard deviation σ in the legend. The distribution has generally a large spread due to two sources of randomness in the XT - the independent phase noise sources which make the phase difference between modes random at the MUX, and the uncertainty in the propagation distance, which makes the phase difference random at the DEMUX. Increased modulation format sizes are therefore probably infeasible for the selected XT values. While

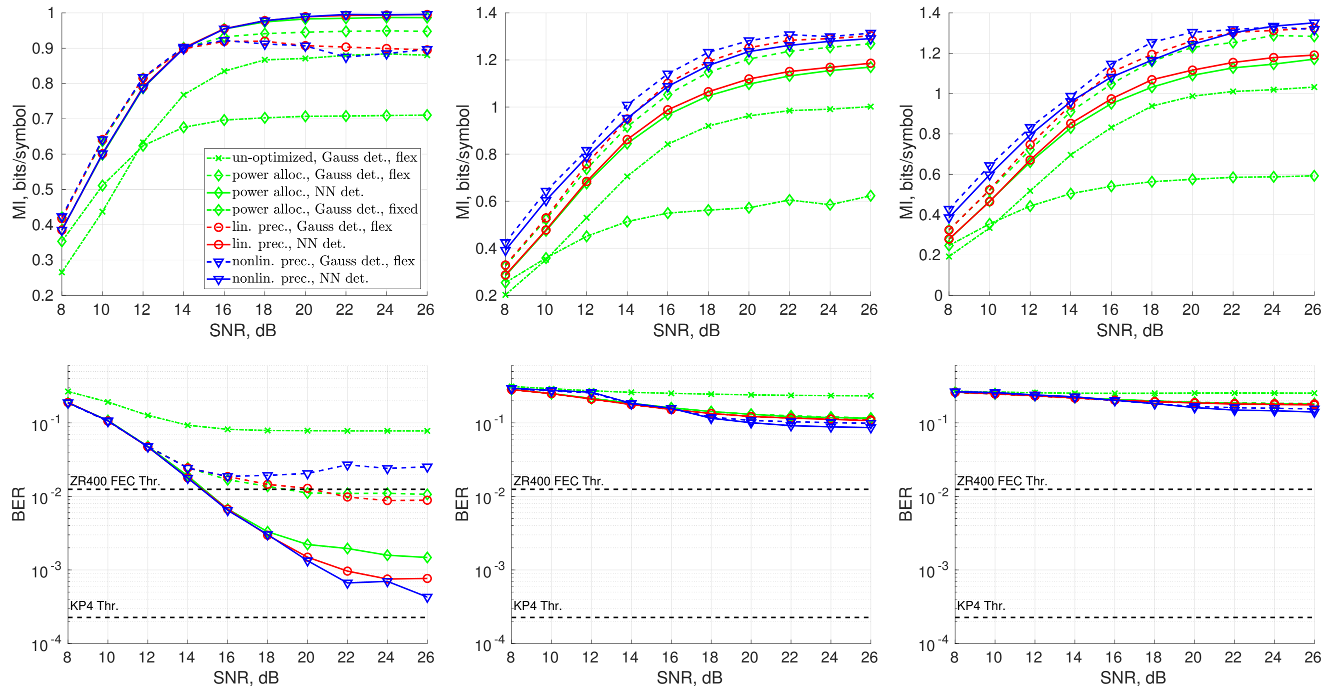

Fig. 4. Performance for fading channel with varying distance and XT distribution as described in Section II-C. Top row: MI. Bottom row: BER. Results for OOK, 4PAM and 8PAM in the first, second and third column, respectively.

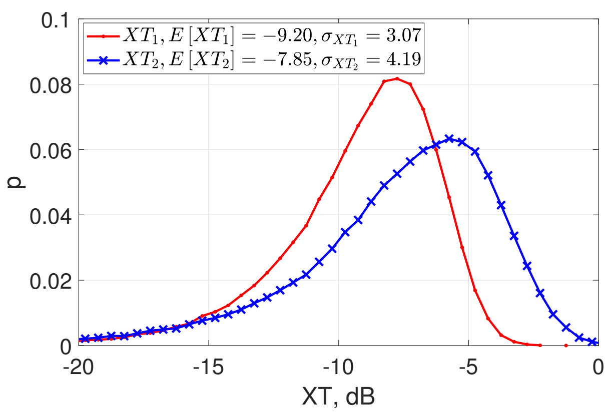

Fig. 5. End-to-end XT from the simulation model given in Section II.

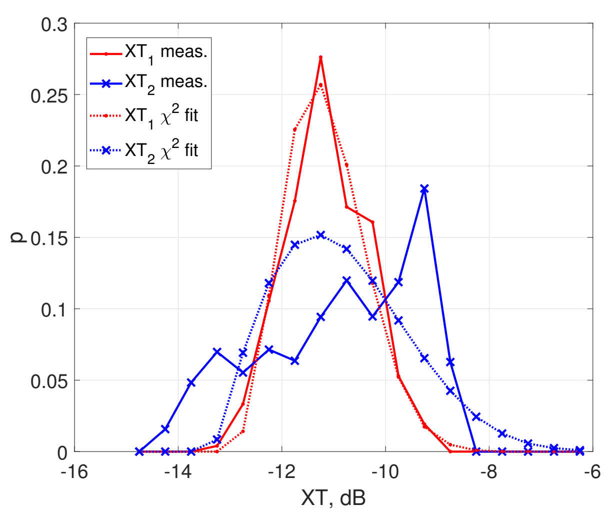

Fig. 6. End-to-end XT measurements and fits.

significant MI and SNR gains can be achieved with high-order modulation at up to ≈ 1.2 bits/symbol, fixed pre-coding is not sufficient to enable them for transmission at high rate. At e.g. the $1.25 \cdot 10^{-2}$ threshold of the FEC from the 400ZR standard [19], which is similar to the hard-decision (HD) low-density parity check (LDPC) FEC threshold of $1.08 \cdot 10^{-2}$ considered for 25G/50G PON [20], linear and nonlinear pre-coding do not achieve a significant gain over simple power allocation for OOK signals. In cases of simpler FEC with lower threshold, e.g. at around 10^{-3} (see examples in Table I in [21]), the pre-coders still enable error-free transmission.

The MI performance at the high SNR end and high order modulation is improved by applying a *flexible* Gaussian receiver w.r.t. using the robust NN decoder. In this case, the parameters of the Gaussian auxiliary channel are estimated in each fading block (as described in Section III-B) and applied for AIR estimation and for decision region calculation. At the high SNR end, the fixed NN decoder, although robust, is slightly penalized by the dominating nature of the XT, which is varying block-to-block. A fixed Gaussian receiver which is not flexible with the channel fading is not competitive, as exemplified with the AIR performance of the power allocation-type optimization. It should be noted that the performance of a fixed, robust Gaussian receiver can potentially be improved by selecting a cost function tailored to this type of processing. This is not straight-forward with the AE described above. The fixed Gaussian receiver is therefore excluded from the BER investigation and AIR investigation of the other pre-coders.

A. Full channel knowledge

In this section, the case when the channel parameters are fixed and known at the transmitter is covered. In such cases, the phase uncertainty reduces only to the phase noise

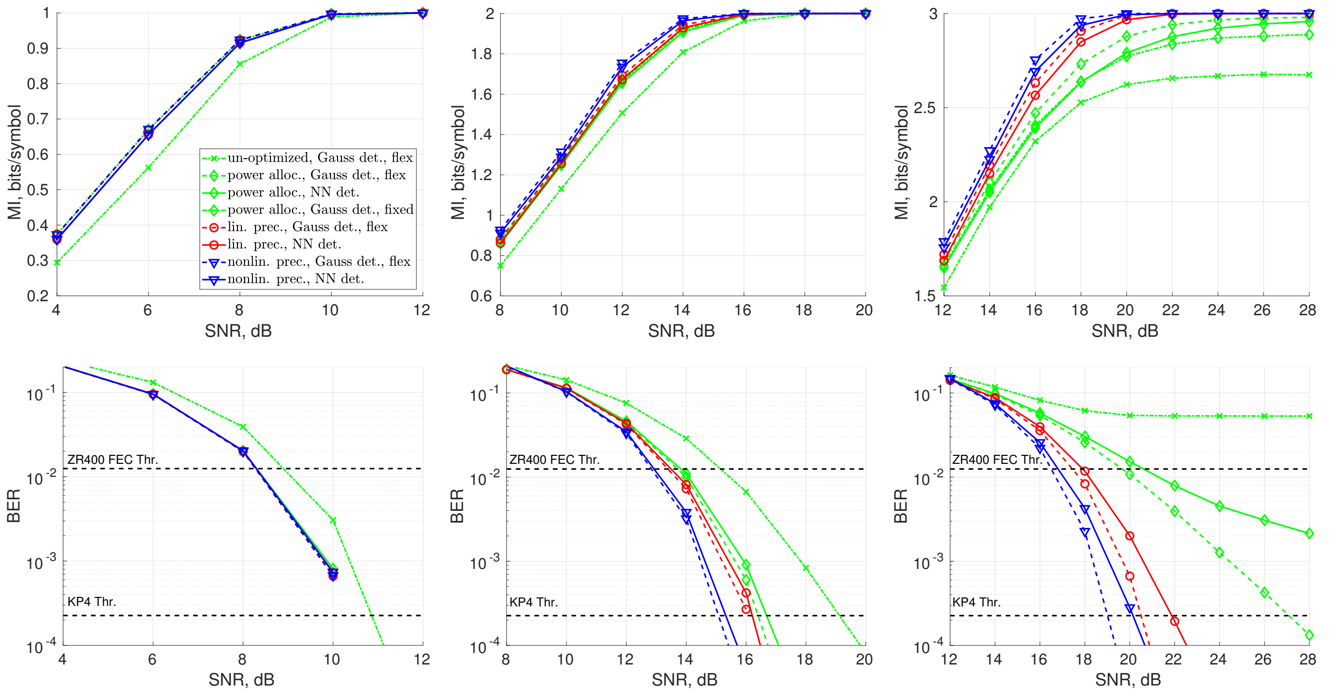

Fig. 7. Performance for fading channel with fixed distance and end-to-end XT distribution measured as described in Section IV-A. Top row: MI. Bottom row: BER. Results for OOK, 4PAM and 8PAM in the first, second and third column, respectively.

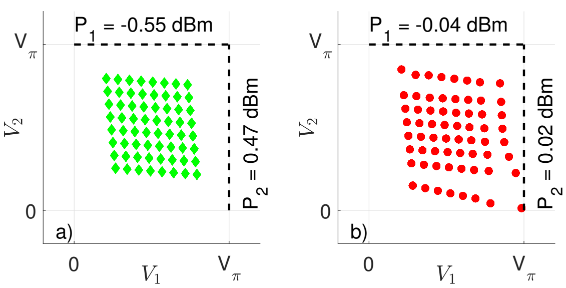

Fig. 8. Pre-coded constellation at an $SNR = 20$ dB, 8PAM, for the fixed-length system.

realization at the transmitter, and the encoder can be tailored better to the channel conditions. This system is relevant when there is for example a low-rate feedback channel from the receiver to the transmitter and/or pre-characterization is allowed for the link prior to deployment. To exemplify this case, the system end-to-end XT measurements from [22] are adopted. The XT is measured with a system similar to that of Fig. 1, where the PDs are replaced by power meters, and the two modes are excited one at a time. The measurements are taken with the power per mode adjusted s.t. the normalized XTs $\mathbb{E}[XT_1] = \mathbb{E}[XT_2] \approx -11dB$. This normalization also implies $\alpha_1 = \alpha_2 = 0$ dB. The distance is fixed to 1630 m. In this case, the compound channel model from (11) collapses to a single transfer matrix $\mathbf{H}$ similar to e.g. $\mathbf{H}_{MUX}$. The XT distribution over time is given in Fig. 6, with mean values calculated to $\mathbb{E}[XT_1] = -11.15$ dB and $\mathbb{E}[XT_2] = -10.89$ dB and standard deviations calculated to $\sigma_{XT_1} = 0.78$ dB and $\sigma_{XT_2} = 1.51$ dB. For completeness, the χ^2 distribution fit described in Section II-C is also shown. The distribution of XT_1 fits rather well with the fitted χ^2 distribution. The distribution of XT_2 seems more flat than the prediction which points to a mismatch in the physical origin of the XT when the latter is dominated by discrete XT sources. This mismatch and its theoretical modeling are subject to further investigation. For the following results, the histogram of the measurements is taken as the distribution of the channel matrix. Applying the AE to the fixed-length system results in the performance curves in Fig. 7.

When the fading is only dominated by the MUX and DEMUX XT variation in Fig. 6, the resulting spread is smaller than in the distance varying case. This means that the pre-coder and decoder can be tailored much better to the channel, enabling high-order modulation for transmission. The gain of nonlinear pre-coder over linear pre-coder is around 0.7 dB for 4PAM and 1.5 dB for 8PAM at the HD LDPC and 400ZR FEC threshold. The gain of the nonlinear pre-coder w.r.t. simple power allocation is around 1 dB and 4 dB for 4PAM and 8PAM, respectively. The gain for OOK w.r.t. un-optimized transmission is a similar ≈1 dB with power allocation, linear, and non-linear pre-coders. The gain increases for the simpler FECs' thresholds, and pre-coding enables error-free transmission with 8PAM at the KP4 threshold of $2.26 \cdot 10^{-4}$[21].

For completeness, the 2-dimensional constellations after the linear and non-linear pre-coding in the case of 8PAM and $SNR = 20$ dB are given in Fig. 8a) and Fig. 8b), respectively. The corresponding power allocation is also given on the sides of the figures. The linear pre-coder only skews the constellation, while the nonlinearities in the NN allow for the constellation to employ larger portion of the dynamic range of the MZMs at no penalty, and thus achieve a better performance.

Histograms of the corresponding received constellations on each mode are given in Fig. 9. The nonlinear precoder significantly improves the interference as evident by the better-

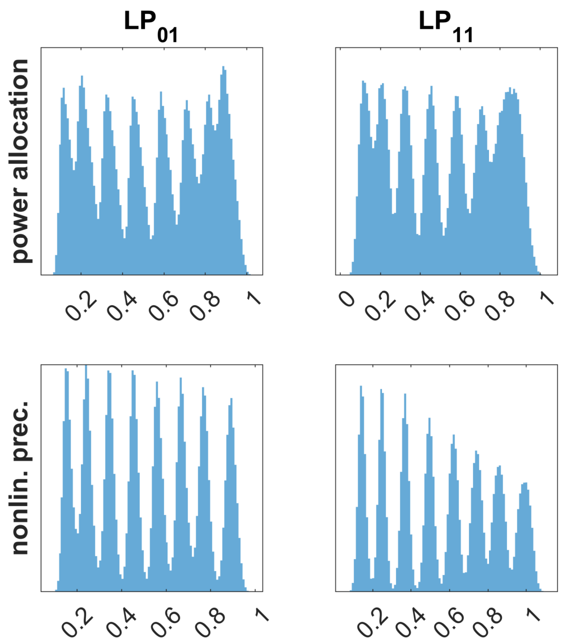

Fig. 9. Histograms of the received signal with power allocation (top row) and nonlinear precoding (bottom row), the LP_{01} mode (left column) and LP_{11} mode (right column) at an $SNR = 20$ dB, 8PAM, for the fixed-length system.

distinguishable PAM points, even though some imbalance still exists between the quality of the signals of the two modes. The latter can be explained by the fact that the cost function aims at maximizing the worse MI, which in this case is already capped at the maximum entropy of 3 bits/QAM symbol per mode for the worse mode (in this case, the LP_{11}), and is not influenced by further visual quality improvements.

V. Discussion and Future Work

A. Channel capacity

The channel capacity of MIMO IMDD systems can be analyzed for a simplified channel model, where the interference is assumed incoherent [9]. In such a case, the output of the PD can be modeled using a standard MIMO channel with a positive input, which represents the output intensity (observe, not the field) of the MZM. In such cases, capacity bounds can be obtained by for example:

1) applying a QR decomposition and assuming a successive interference cancellation receiver [12][13];
2) re-modeling the MIMO channel using a set of parallel channels with a colored Gaussian noise [23], and then using duality and entropy power inequality to arrive at bounds and asymptotic expressions of capacity.

The first option is not applicable to the channel model in (12) due to the absolute square operation of the PD. The second is not very practical as it does not provide a capacity estimate at the moderate SNR region, which is where typical communication systems operate. Tight capacity bounds for the moderate SNR region are an open problem for point to point links due to the implicit amplitude constraint [23]. Furthermore, neither of these results are tight for the broadcast type of channel of interest to this paper. Theoretical capacity estimation is a challenging, yet interesting topic for future work.

B. Channel model

In this paper, a single sample per symbol channel model was treated. The intra- and inter-channel fading frequency selectivity is also neglected due to the assumed large coherence bandwidth. Furthermore, the distributed cross talk along the fiber, as well as chromatic dispersion were neglected. A more accurate representation of the fading may be obtained by finite difference method simulations of the multi-mode optical fiber [24]. Such a task is challenging both computationally and analytically, but is an interesting direction for future extensions of this work.

The model for the distribution of the cross-talk used to generate the results in Fig. 4 is valid when the cross-talk is distributed along an MMF fiber. In contrast, in this paper, the sources of the cross-talk are discrete elements. The validity of the model is thus subject to future investigations. However, we note that the methodology for cross-talk mitigation presented remains the same, and we do not expect significantly different relative performance gains from the proposed method for other distributions under the same cross-talk expected values and spread.

In this paper, it is assumed that the fiber lengths between the MZMs and the MUX for both modes are similar. This is reasonable to achieve with low baudrate signals. Alternatively, a delay can be imposed at the pre-coding stage, which should be able to compensate any differences in the delays of the pre-coded signals, s.t. they are aligned at the MUX. Similarly, it is assumed that the propagation distances between the DEMUX and the detectors of the two degenerate modes in LP_{11} are identical. The effect of any mismatch is a skew, which is problematic if the two degenerate modes are to be added electrically in the analog domain. De-skewing the modes can be done digitally at the cost of an independent analog to digital converters for each degenerate mode, which increases the cost and complexity of the system.

C. Steps towards experimental demonstration

In order to demonstrate the benefits of the AE experimentally, careful characterization of the optical delay lines between the MZMs and the MUX, as well as between the DEMUX and the PDs needs to be performed in order to de-skew the paths. This is required in order to ensure that the pre-coder compensates for the correctly placed interfering symbols. Furthermore, the quantization effects of the necessary analog-to-digital and digital-to-analog conversion need to be taken into account in the model which can be done using the techniques from [25]. Regular IM/DD system equipment (arbitrary waveform generator, 2x lasers, 2x MZMs, 3x PDs, 3x electrical amplifiers) and the necessary FMF system equipment (MUX, FMF, DEMUX) should be sufficient to conduct an experimental demonstration, which we plan for future extension of this work.

We also note that in the case when electrical amplification noise dominates over the optical receiver noise, the system model may be simplified by combining the degenerate modes in the electrical domain prior to electrical amplification, resulting in identical noise variance of both mode groups. This would also require careful engineering of electrical connection lengths and does not allow digital de-skewing.

D. AE for receivers with MIMO

The AE is straight-forward to adapt to the case where joint mode detection can be performed using MIMO processing at the receiver. An example is shown for the simple channel model case in [9]. When MIMO is employed, the [NxN] channel can be estimated at the receiver in each fading block, which potentially allows to dramatically improve the performance, at the expense of significant complexity increase. The non-linear pre-coder in that case is very beneficial and allows to approach the channel capacity, as suggested in [9].

VI. Conclusion

In this paper, a special application of autoencoders (AE) was described for optimization of a few-mode IMDD communication system over a multi-mode fiber. The AE was used to optimize digital pre-coders. In order to maintain the complexity of the system low, the AE was optimized for MIMO-less reception and for equalization of the performance of the difference data streams. The AE can be made robust to channel uncertainties stemming from practical transceiver penalties such as phase noise, as well as to uncertainty in the fiber length, making the AE practical for implementation in various short-reach communication systems, e.g. intra-data centers. Robustness comes at the price of degraded performance and limitation to the maximum AIR with high orders of modulation. On the other hand, when the distance is known, the AE enables high orders of modulation for transmission. In this case, nonlinear pre-coding achieves SNR gains of up to 1.5 dB w.r.t. linear pre-coding and up to 4 dB w.r.t. simple power pre-emphasis.

Acknowledgment

This work was supported by the Danish national research foundation SPOC project, ref. DNRF123, the Villum Young Investigator project OPTIC-AI, ref. 29334, and Innovation Fund Denmark projects INCOM 8057-00059B and GO-SENS 0224-00074B.

References

[1] P. P. Mitra and J. B. Stark, "Nonlinear limits to the information capacity of optical fiber communications," *Nature*, vol. 411, no. 6841, pp. 1027–1030, June 2001.

[2] R.-J. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, and B. Goebel, "Capacity limits of optical fiber networks," *Journal of Lightwave Technology*, vol. 28, no. 4, pp. 662–701, 2010.

[3] B. J. Puttnam, G. Rademacher, and R. S. Luís, "Space-division multiplexing for optical fiber communications," *Optica*, vol. 8, no. 9, pp. 1186–1203, Sep 2021.

[4] L. Grüner-Nielsen, N. M. Mathew, M. H. Nymann, M. Lillieholm, M. Galili, and K. Rottwitt, "Mode division multiplexing on standard 50/125 μm multi mode fiber using photonic lanterns," in *Optical Fiber Communication Conference (OFC) 2021*. Optical Society of America, 2021, p. W6A.41.

[5] N. K. Fontaine, J. Carpenter, S. Gross, S. Leon-Saval, Y. Jung, D. J. Richardson, and R. Amezcua-Correa, "Photonic lanterns, 3-d waveguides, multiplane light conversion, and other components that enable space-division multiplexing," *Proceedings of the IEEE*, vol. 110, no. 11, pp. 1821–1834, 2022.

[6] N. M. Mathew, J. B. Christensen, L. Grüner-Nielsen, M. Galili, and K. Rottwitt, "Air-cladded mode-group selective photonic lanterns for mode-division multiplexing," *Opt. Express*, vol. 27, no. 9, pp. 13 329–13 343, Apr 2019.

[7] A. Vijay and J. M. Kahn, "Effect of higher-order modal dispersion in direct-detection mode-division-multiplexed links," *Journal of Lightwave Technology*, vol. 41, no. 6, pp. 1670–1683, 2023.

[8] K. Choutagunta, I. Roberts, D. A. B. Miller, and J. M. Kahn, "Adapting Mach-Zehnder mesh equalizers in direct-detection mode-division-multiplexed links," *Journal of Lightwave Technology*, vol. 38, no. 4, pp. 723–735, 2020.

[9] M. P. Yankov, F. Da Ros, S. Forchhammer, and L. Grüner-Nielsen, "Capacity and achievable rates of fading few-mode MIMO IM/DD optical fiber channels," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 769–774.

[10] C. Antonelli, A. Mecozzi, M. Shtaif, and P. J. Winzer, "Random coupling between groups of degenerate fiber modes in mode multiplexed transmission," *Opt. Express*, vol. 21, no. 8, pp. 9484–9490, Apr 2013.

[11] K. Choutagunta and J. M. Kahn, "Dynamic channel modeling for mode-division multiplexing," *Journal of Lightwave Technology*, vol. 35, no. 12, pp. 2451–2463, 2017.

[12] A. Chaaban, Z. Rezki, and M.-S. Alouini, "Capacity bounds and high-SNR capacity of MIMO intensity-modulation optical channels," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3003–3017, 2018.

[13] A. Chaaban, O. M. S. Al-Ebraheemy, T. Y. Al-Naffouri, and M.-S. Alouini, "Capacity bounds for the Gaussian IM-DD optical multiple-access channel," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3328–3340, 2017.

[14] T. Mukaihara, T. Kimura, and H. Koshi, "Future narrow linewidth lasers," in *Optical Fiber Communication Conference*. Optica Publishing Group, 2015, p. W1H.6.

[15] D. Arnold, H.-A. Loeliger, P. Vontobel, A. Kavcic, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3498–3508, 2006.

[16] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, 2017.

[17] T. J. O'Shea, T. Erpek, and T. C. Clancy, "Physical layer deep learning of encodings for the MIMO fading channel," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2017, pp. 76–80.

[18] L. Grüner-Nielsen, N. M. Mathew, M. Lillieholm, M. Galili, and K. Rottwitt, "Modeling of MIMO less mode division multiplexed systems," *IEEE Photonics Technology Letters*, vol. 32, no. 18, pp. 1191–1194, 2020.

[19] "Implementation agreement 400ZR, IA 0.10-Draft," OIF Optical Internetworking Forum, Tech. Rep., 2020.

[20] W. Wang, K. Tao, W. Qian, Y. Cai, M. Lei, X. Zhou, H.-C. Chien, J. Liang, S. Zhang, and Z. Liu, "Real-time FPGA verification for 25G-PON and 50G-PON LDPC codes," in *2020 Conference on Lasers and Electro-Optics (CLEO)*, 2020, pp. 1–2.

[21] E. Agrell and M. Secondini, "Information-theoretic tools for optical communications engineers [invited]," in *31st Annual Conference of the IEEE Photonics Society, IPC 2018*, 2018.

[22] L. E. Grüner-Nielsen, N. M. Mathew, M. Galili, L. S. Rishøj, and K. Rottwitt, "Photonic-lantern-based MDM devices," in *Optical Fiber Communication Conference (OFC)*. Optical Society of America, 2023, p. M1B.5.

[23] S. M. Moser, M. Mylonakis, L. Wang, and M. Wigger, "Asymptotic capacity results for MIMO wireless optical communication," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 536–540.

[24] C. Antonelli, M. Shtaif, and A. Mecozzi, "Modeling of nonlinear propagation in space-division multiplexed fiber-optic transmission," *Journal of Lightwave Technology*, vol. 34, no. 1, pp. 36–54, 2016.

[25] O. Jovanovic, F. Da Ros, D. Zibar, and M. P. Yankov, "Geometric constellation shaping for fiber-optic channels via end-to-end learning," *Journal of Lightwave Technology*, vol. 41, no. 12, pp. 3726–3736, 2023.