CHALMERS

Autoencoders for Physical-Layer Communications: Approaches and Applications

JINXIANG SONG

Department of Electrical Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2023
www.chalmers.se

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Autoencoders for Physical-Layer Communications: Approaches and Applications

Jinxiang Song

CHALMERS
UNIVERSITY OF TECHNOLOGY

Communication Systems Group
Department of Electrical Engineering
Chalmers University of Technology
Göteborg, Sweden, 2023

Autoencoders for Physical-Layer Communications: Approaches and Applications

Jinxiang Song

ISBN 978-91-7905-977-4
Doktorsavhandlingar vid Chalmers tekniska högskola
Ny series No. 5443
ISSN 0346-718X

This thesis has been prepared using LaTeX and Tikz.

Communication Systems Group
Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Göteborg, Sweden
Phone: +46 (0)31 772 1000
www.chalmers.se

Printed by Chalmers Reproservice
Göteborg, Sweden, December 2023

Abstract

The ever-growing demand for higher data rates has driven continuous developments in communication systems over the years. As upcoming high-bandwidth services require even higher data rates, future digital communication infrastructures must undergo continuous upgrades to provide increased capacity. Recently, machine learning has surfaced as a potential tool to augment this capacity further. A particularly promising avenue lies in the application of autoencoders. These can concurrently optimize both the transmitter and receiver tailored to a specific channel model and performance metric, a paradigm commonly referred to as end-to-end autoencoder learning.

In this thesis, we study different aspects of using machine learning for physical-layer communications, spanning wireless and optical communication in terms of applications and unsupervised, supervised, and reinforcement learning in terms of methodologies. The main contributions of this thesis are listed as follows.

Firstly, to overcome the challenge that standard end-to-end autoencoder learning requires a differentiable channel model for gradient-based transmitter optimization, Paper A and Paper B explore reinforcement learning-based transmitter optimization. In Paper A, considering that reinforcement learning-based training necessitates sending a feedback signal from the receiver to the transmitter, we propose a novel method for the feedback signal quantization. Simulation results demonstrate that the proposed quantization scheme facilitates effective transmitter learning with limited feedback. In Paper B, reinforcement learning is applied to mitigate transmitter hardware impairments. A novel digital predistorter based on neural networks is introduced and trained in a back-to-back optical fiber transmission experiment. Experimental results demonstrate that the proposed digital predistorter effectively mitigates transmitter impairments, outperforming commonly used baseline schemes.

Secondly, Paper C and Paper D focus on supervised learning, with an emphasis on improving the interpretability of end-to-end autoencoder learning-based communication systems. In Paper C, a novel model-based autoencoder is proposed for nonlinear systems. By decomposing the autoencoder-based transceivers into concatenations of smaller neural networks, the proposed method allows for the visualization of each learned functional block, improving the interpretability of the learned transmission scheme. Paper D interprets the learned solution from a different perspective by carefully selecting baseline schemes. We demonstrate that, for the linear systems considered in Paper D, machine learning methods do not significantly outperform conventional model-based approaches. Instead, they learn invertible transformations of these model-based solutions.

Lastly, Paper E focuses on unsupervised learning, addressing the problem of blind channel equalization for both linear and non-linear channels. By introducing a constraint to the latent representation of a standard autoencoder, a novel autoencoder-based blind equalizer is formulated. Simulation results demonstrate that, for both linear and nonlinear channels, the proposed equalizer can achieve similar performance as conventional

data-aided equalizers while outperforming state-of-the-art blind methods.

Keywords: machine learning, neural networks, autoencoders, physical-layer communications, digital signal processing, hardware impairments, equalization.

List of Publications

This thesis is based on the following publications:

[A] **J. Song**, B. Peng, C. Häger, H. Wymeersch, and A. Sahai, "Learning physical-layer communication with quantized feedback," *IEEE Transactions on Communications*, vol.67, pp. 645-653, Oct. 2019.

[B] **J. Song**, Z. He, C. Häger, M. Karlsson, A. Graell i Amat, H. Wymeersch, and J. Schröder, "Over-the-fiber digital predistortion using reinforcement learning," in *Proc. European Conference on Optical Communications*, Sept. 2021.

[C] **J. Song**, C. Häger, J. Schröder, A. Graell i Amat, and H. Wymeersch, "Model-based end-to-end learning for WDM systems with transceiver hardware impairments," (invited paper) *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 28, no. 4, pp. 1-14, Aug. 2022.

[D] **J. Song**, C. Häger, J. Schröder, T. J. O'Shea, E. Agrell, and H. Wymeersch, "Benchmarking and interpreting end-to-end learning of MIMO and multi-user communication," *IEEE Transactions on Wireless Communications*, vol. 21, no. 9, pp. 7287-7298, Mar. 2022.

[E] **J. Song**, V. Lauinger, Y. Wu, C. Häger, J. Schröder, A. Greaell i Amat, L. Schmalen, and H. Wymeersch, "Blind channel equalization using latent space constrained autoencoders," submitted to *IEEE Transactions on Communications* (major revision), Oct. 2023.

Publications by the author, not included in the thesis:

[F] **J. Song**, C. Häger, J. Schröder, T. J. O'Shea, and H. Wymeersch, "Benchmarking end-to-end learning of MIMO physical-layer communication," in *Proc. Global Communications Conference*, Dec. 2020.

[G] **J. Song**, C. Häger, J. Schröder, A. Graell i Amat, and H. Wymeersch, "End-to-end autoencoder for superchannel transceivers with hardware impairment," in *Proc. Optical Fiber Communication Conference*, Mar. 2021.

[H] **J. Song**, V. Lauinger, C. Häger, J. Schröder, A. Greaell i Amat, L. Schmalen , and H. Wymeersch, "Blind frequency-domain equalization using vector-quantized variational autoencoders," in *Proc. European Conference on Optical Communications*, Sept. 2023.

[I] Z. He, **J. Song**, C. Häger, K. Vijayan, P. Andrekson, M. Karlsson, A. Graell i Amat, H. Wymeersch, and J. Schröder, "Symbol-based supervised learning predistortion for compensating transmitter nonlinearity," in *Proc. European Conference on Optical Communications*, Sept. 2019.

[J] Y. Wu, **J. Song**, C. Häger, U. Gustavsson, A. Graell i Amat, and H. Wymeersch, "Symbol-based over-the-air digital predistortion using reinforcement learning," in *Proc. European Conference on Optical Communications*, Sept. 2019.

[K] Z. He, **J. Song**, C. Häger, A. Graell i Amat, H. Wymeersch, P. Andrekson, M. Karlsson, and J. Schröder, "Experimental demonstration of learned pulse shaping filter for superchannels," in *Proc. Optical Fiber Communication Conference*, Mar. 2022.

[L] J. M. Mateos-Ramos, **J. Song**, Y. Wu, C. Häger, M. F. Keskin, V. Yajnanaryana, and H. Wymeersch, "End-to-end learning for integrated sensing and communication," in *Proc. International Conference on Communications*, May 2022.

[M] Z. He, **J. Song**, K. Vijayan, C. Häger, A. Graell i Amat, H. Wymeersch, P. Andrekson, M. Karlsson, and J. Schröder, "Periodicity-enabled size reduction of symbol based predistortion for high-order QAM," *IEEE/OSA Journal of Lightwave Technology*, vol. 40, no.18, pp. 6168–6178 Jul. 2022.

[N] M. Srinivasan **J. Song**, C. Häger, J. Schröder, and H. Wymeersch, "Learning optimal PAM Levels for VCSEL-based optical interconnects," in *Proc. European Conference on Optical Communications*, Sept. 2022.

[O] S. Rivetti, J. M. Mateos-Ramos, Y. Wu, **J. Song**, Y. Wu, M. F. Keskin, V. Yajnanaryana, C. Häger, and H. Wymeersch, "Spatial signal design for positioning via end-to-end learning," *IEEE Wireless Communications Letters*, Jan. 2023.

[P] M. Srinivasan, **J. Song**, A. Grabowski, K. Szczerba, H. K. Iversen, M. N. Schmidt, D. Zibar, J. Schröder, A. Larsson, C. Häger, and Henk Wymeersch, (Invited Tutorial) "End-to-end learning for VCSEL-based optical interconnects: state-of-the-art, challenges, and opportunities," *IEEE/OSA Journal of Lightwave Technology* , vol. 41, no.11, pp. 3261–3277, Jun. 2023.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Henk Wymeersch, for recognizing my potential and offering me a PhD position at Chalmers University of Technology. Without your guidance and encouragement, I may not have even considered pursuing a PhD. It has been an incredible five-year journey working together, and I have learned invaluable lessons about what it means to conduct research. Your guidance, support, and patience have been the pillars of this work.

I am also deeply grateful to my co-supervisor, Dr. Jochen Schröder. You have been more than just a mentor; you have been a friend. Thank you for your passion, dedication, and unwavering support. I thoroughly enjoyed the discussions we shared, which greatly enriched my understanding of the subject.

I would also like to extend my gratitude to Prof. Christian Häger. I deeply appreciate the guidance you have provided over the years, especially for all the times you found things to try when I thought I was hopelessly stuck. Additionally, thank you, Prof. Alexandre Graell i Amat, for teaching me to be rigorous in my writing and presentations.

I extend my sincere thanks to Prof. Laurent Schmalen for offering me the opportunity to visit the Communication Engineering Lab at Karlsruhe Institute of Technology, Karlsruhe, Germany. I greatly appreciate the discussions with my collaborators from the Communication Engineering Lab. Special thanks go to my colleagues and friends, Sisi, Haizheng, and Andrej, for making my research visit at KIT so welcoming.

Heartfelt thanks goes to Prof. Erik Ström and Prof. Fredrik Brännström for their dedication and hard work in improving the working environment at Chalmers. I also want to acknowledge all my colleagues in the FORCE group for fostering a diverse and intellectually stimulating working environment. To my friends and colleagues at Chalmers, working with you over these years has been a wonderful experience.

I extend my gratitude to my friends outside of Chalmers, including Chi, Shudi, Yuxin, Tiantian, Cong, Sebastian, and others, for always being there when I needed support.

Last but not least, I am deeply grateful to my family and my girlfriend Suchada, who have always been very supportive.

Jinxiang Song
Göteborg, 2023

Financial Support

This work was funded by the Knut and Alice Wallenberg Foundation, grant No. 2018.0090. I would also like to acknowledge the Ericsson Research Foundation for partially funding my research travels.

Acronyms

ADC:	analog-to-digital converter
AE:	autoencoder
AIR	achievable information rate
AWGN:	additive white Gaussian noise
BCE:	binary cross-entropy
BER:	bit error rate
BPS:	blind phase search
CD:	chromatic dispersion
CE:	cross-entropy
CMA:	constant modulus algorithm
CNN:	convolutional neural network
DAC:	digital-to-analog converter
DL:	deep learning
DPD:	digital predistortion
DSP:	digital signal processing
ENOB:	effective number of bits
FEC:	forward error correction
FIR:	finite impulse response
FOE:	frequency offset estimation
GCS:	geometric constellation shaping
GMI:	generalized mutual information

GMP:	generalized memory polynomials
GN:	Gaussian noise
IQ:	in-phase and quadrature
ISI:	inter-symbol interference
LLR:	log-likelihood ratio
LSC-AE:	latent spaced constrained autoencoder
LO:	local oscillator
MAP:	maximum a posteriori
MFR:	modulation format recognition
MI:	mutual information
MIMO:	multiple-input multiple-output
ML:	maximum likelihood
MLP:	multi-layer perceptron
MSE:	mean-squared-error
MU:	multi-user
MZM:	Mach-Zehnder modulator
NN:	neural network
NLIN:	nonlinear inference noise
NLPN:	nonlinear phase noise
NLSE:	nonlinear Schrödinger equalization
PA:	power amplifier
PAM:	pulse amplitude modulation
PAPR:	peak-to-average power ratio

PCA:	principle component analysis
PCS:	probabilistic constellation shaping
PMD:	polarization mode dispersion
PN:	phase noise
PNC:	phase noise compensation
PSK:	phase shift keying
QAM:	quadrature amplitude modulation
QPSK:	quadrature phase shift keying
RL:	reinforcement learning
SDM:	space-division multiplexing
SE:	spectral efficiency
SER:	symbol error rate
SNDR:	signal-to-noise-and-distortion ratio
SNR:	signal-to-noise ratio
SSFM:	split-step Fourier method
VAE:	variational autoencoder
WDM:	wavelength-division multiplexing

Contents

Part I

Overview

CHAPTER 1

Introduction

In an era of technological breakthroughs and innovations, the worldwide Internet traffic volume has experienced exceptional growth in the last 20 years, and it is expected that this trend will continue in the future [1]. The driving force behind this trend can be attributed to bandwidth-intensive services such as cloud computing, video streaming, and autonomous driving. To meet the ever-growing demand for high-bandwidth services, significant efforts are required to increase data rates and spectral efficiency (SE) in both wireless and wired communications.

One of the key enablers for supporting the ever-growing high-bandwidth services is fiber-optic communication systems. Today, they constitute the backbone of the Internet due to their capabilities to provide extremely high data rates. Traditionally, communication systems, both wireless and wired, have relied on precise and well-understood mathematical models that perform exceptionally well in numerous practical scenarios [2–4]. Yet, the same approach—designing systems based on well-established mathematical models—might prove insufficient for the development of next-generation fiber-optic communication systems for various reasons, including but not limited to the following. Firstly, the model assumptions used in current optical fiber system designs may be inadequate for future generations of systems. For instance, many digital signal processing (DSP) techniques in today's transmission systems, like frequency domain equalizers [5], operate under the assumption of a linear channel. Such an assumption may no longer hold true for future systems, particularly when high data rates necessitate increased transmission power to meet thesignal-to-noise ratio (SNR) requirements associated with advanced modulation schemes [6, 7]. Although researchers today possess a comprehensive understanding of the nonlinear signal propagation within optical fibers, the DSP algorithms

needed to compensate for these nonlinear impairments remain prohibitively expensive for practical implementations [8,9]. Secondly, the mathematical frameworks that researchers use today may prove incomplete for future optic-fiber system design. For example, due to limitations in hardware technology development, practical transceiver hardware are non-ideal and introduce impairments to the transmitted signal [10–12]. Addressing these hardware-induced challenges in the design phase is crucial for preserving signal integrity. However, many existing hardware models frequently fall short in accuracy [13] or are too complex to be included in the transceiver design phase. Thirdly, and perhaps most crucially, developing mathematical models that accurately describes future optical communication systems might be excessively complex [14]. And, even if such models exist, they could be impractical or challenging to solve [3]. Specifically, next-generation systems are expected to support extremely high data rates and low latencies while also enabling dynamicity, flexibility, and efficiency to meet the heterogeneous quality of service requirements of various emerging applications [3, 14]. Consequently, the mathematical complexities underlying these problems can become too intricate to describe and are often too complex to be solved.

Due to these reasons, traditional design methods for fiber-optic communications may fall short in addressing the requirements of upcoming applications, and new design tools are needed to assist with the evolution of future generations of optical transmission systems.

1.1 Background and Motivation

In recent years, the ever-expanding availability of data combined with increased access to computing power has propelled machine learning into a revolutionary force across numerous domains. This advancement has unlocked unprecedented applications and achieved performance benchmarks that significantly surpass traditional methods. In particular, such transformations have been evident in application areas that have limited theoretical foundations and lack robust models, such as image processing [15], natural language processing [16], and autonomous driving [17]. However, the same cannot be said for digital communications, a field deeply rooted in information theory and statistics, and firmly based on well-established mathematical models. At first glance, one might assume that integrating machine learning into the design of fiber-optic communication systems would yield only incremental, if not negligible, gains. Yet, as optical communication systems become more complex, the design of the physical layer, i.e., the development of optimal transmission and detection methods, also becomes more challenging. Machine learning is envisioned as essential for the next generation of systems for the following reasons.

First, machine learning excels when there is a mismatch between models and real-world scenarios [18, 19]. Particularly, communication system design traditionally relies on modeling assumptions, such as linearity, Gaussianity, and stationarity. While these

assumptions make models mathematically tractable and suitable for analysis, they fall short in accurately representing real-world systems affected by both linear (e.g., bandwidth constraints [20]) and nonlinear effects (e.g., transceiver imperfections [21] or fiber nonlinearity [22]). In such cases, machine learning techniques shows great potential to significantly enhance the performance of conventional method, as they are directly learned from data and are not constrained by the modeling assumptions [23, 24].

Second, machine learning is valuable when algorithms involve heuristics for parameter selection [25]. For example, an iterative decoding algorithm (e.g., a belief propagation decoder) requires heuristic selection of iterations to trade-off performance and complexity [26]. Algorithms based on solid theoretical foundations with performance guarantees can become suboptimal if parameters are not appropriately chosen. Optimizing these parameters is often non-trivial and, in some cases, mathematically intractable. In the absence of closed-form solutions, traditional approaches rely on experience, intuition, and extensive manual tuning. Machine learning provides an effective solution to optimize these parameters without explicitly solving complex optimization problems.

Third, machine learning is beneficial when algorithms are computationally prohibitive [9, 27]. While some problems have provably optimal solutions (e.g., maximum-likelihood decoding [28]), these may be too complex for practical implementation, resulting in inefficiencies in terms of execution time and energy consumption. Machine learning offers the opportunity to replace/approximate such algorithms with highly parallelizable structures, for example neural networks (NNs), that execute efficiently and with reduced energy consumption [9, 29, 30].

Finally, machine learning is instrumental when the goal is end-to-end performance optimization [31,32]. Traditional communication systems are typically modeled as a sequence of individual blocks, each designed for a specific task (e.g., channel coding, modulation, pulse shaping, and equalization) [33]. Although this approach has led to the efficient, versatile, and controllable systems we have today, it is not clear that individually optimized processing blocks achieve the best possible end-to-end performance. For example, the separation of channel coding and modulation are known to be sub-optimal [34]. Additionally, this modular approach necessitates optimizing each individual block separately, as joint block optimization becomes exceedingly complex. Machine learning, however, provides a straightforward means of optimizing end-to-end performance by directly modeling the entire system without the constraints of a modular structure.

1.2 Scope of the Thesis

In this thesis, we investigate the potential of applying deep learning (DL) techniques to assist the design of communication systems. Our primary focus is on utilizing end-to-end autoencoder (AE) learning for joint transceiver optimization. Regarding this topic, our main contributions are threefold (in papers A-D), which we summarize as follows:

- In Papers A and B, motivated by the challenge of AE-based transmitter train-

ing imposed by gradient-based optimization, we investigate the use of reinforcement learning (RL) for transmitter NN training. Paper A concerns feedback signal quantization when RL is used for transmitter optimization, while Paper B studies RL-based digital predistortion (DPD) training in an experimental setup.

- In Paper C, inspired by the difficulty of interpreting a learning-based approach that provides a "black-box" solution, we design an AE-based transceiver following the modular structure of a conventional system. By doing this, the proposed method outperforms the conventional solution while also enabling us to interpret the learned solution.
- In Paper D, our focus lies in utilizing AEs to learn multiple-input multiple-output (MIMO) and multi-user (MU) communications. Through carefully selecting benchmarks, we are able to provide additional insights, and sometime full interpretations, of the AE-based solutions.

A second research topic in this thesis concerns blind channel equalization, which we study in Paper E. Motivated by the fact that data-aided equalizers require pilot data transmission (which leads to a loss in information rate), and traditional blind equalizers do not perform well, we propose a novel blind equalization method based on latent space constrained autoencoders (LSC-AEs). We validated the proposed blind equalizer over both linear and nonlinear channels, and simulation results show that the proposed method can achieve similar performance as traditional data-aided equalizers while outperforming state-of-the-art blind equalizers.

1.3 Thesis Organization

The thesis is divided into two parts, where the first part serves as an introduction to the appended papers in the second part. The remainder of the introductory part of this thesis is structured as follows: Chapter 2 first briefly reviews the general setup of a digital communication system, after which the channel models as well as hardware impairments considered in this thesis are introduced. Chapter 3 introduces the basics of DL and its applications to physical-layer communications. Chapter 4 introduces end-to-end AE learning-based communication system design, where recent advances and challenges are discussed. Finally, the introductory part of the thesis is concluded in Chapter 5, where we briefly summarize the contributions in the appended papers.

Notation

The introductory part of this thesis uses the following notation conventions. The sets of integers, real numbers, and complex numbers are denoted by $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$, respectively. A finite set is denoted by $\mathcal{X}$ with $|\mathcal{X}|$ denoting the cardinality. We use boldface letters

to denote vectors and matrices (e.g., $\boldsymbol{x}$ and $\boldsymbol{A}$), and we use $(\,\cdot\,)^\top$ to denote transpose. For a vector $\boldsymbol{x}$, x_i denotes the i-th element of $\boldsymbol{x}$ and $\|\boldsymbol{x}\|^2$ denotes the squared Euclidean norm. $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. $\mathcal{CN}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the distribution of a proper complex Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, evaluated at $\boldsymbol{x}$ ($\boldsymbol{x}$ may be omitted to represent the entire distribution). $\mathbb{E}\{\,\cdot\,\}$ denotes the expected value. The imaginary unit is represented by $j = \sqrt{-1}$. The probability of an event is denoted by $\Pr(\,\cdot\,)$. The probability mass function of a discrete random variable X at x is denoted by $P_X(x)$, and the probability density function of a continuous random variable X at x is denoted by $f_X(x)$. Lastly, for a parametric function characterized by a set of parameters $\boldsymbol{\rho}$, we denote it as $f(\,\cdot\,;\boldsymbol{\rho})$ when it yields a single output and as $\boldsymbol{f}(\,\cdot\,;\boldsymbol{\rho})$ when it generates a vector output.

There are some notational inconsistencies across the introductory part of the thesis and the appended papers. Wherever such inconsistencies appear in the appended papers, we adhere to the notations used in each respective paper.

CHAPTER 2

Coherent Communications

In this chapter, we start with an overview of classical communication systems. We are concerned with coherent communications, where information is modulated onto both the amplitude and phase of the transmitted signal. We then discuss the key functional blocks in modern coherent systems, followed by brief descriptions of the channel models and hardware impairments considered in this thesis.

We highlight that the discussions and nomenclatures in the remainder of this chapter are tailored to coherent fiber-optic communications, despite wireless channels were examined in Paper D and part of Paper E. We justify our choice of presentation as follows. The wireless channel models considered in Paper D and Paper E are rather simplified models. Consequently, the considered wireless systems can potentially be integrated into the subsequent discussions without causing significant confusions.

2.1 Classical Communication System Overview

Fig. 2.1 depicts a high-level point-to-point digital communication link, comprising a transmitter, a channel, and a receiver. The overall goal is to transmit information reliably from one point to another, with the information being in the form of digital messages (i.e., bit sequences). To achieve this goal, the transmitter is designed to produce an analog representation of the digital messages, ensuring that the transmitted signal is resilient to transmission impairments (e.g., through the application of suitable channel coding and modulation schemes). The objective of the receiver is to decode from the received analog signal and reconstruct the transmitted bits with the lowest possible error rate. In the

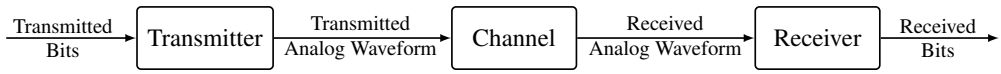

Figure 2.1: A high-level view of a digital communication system consisting of a transmitter that modulates the bit sequence into an analog waveform suitable for transmission and a receiver that demodulates the received waveform back into a bit sequence.

following subsections, we provide a more detailed explanation for each of these elements in this system.

2.2 Classical Transmitter

Fig. 2.2 depicts the various blocks of a transmitter in a coherent fiber-optic communication system. Signal processing in the first five blocks (highlighted in green) operates in the digital domain, while the remaining blocks (highlighted in red) operate in the analog domain. In the subsequent section, we review all the blocks presented in Fig. 2.2, except for channel coding, which falls outside the scope of this thesis and is therefore excluded from our discussions.

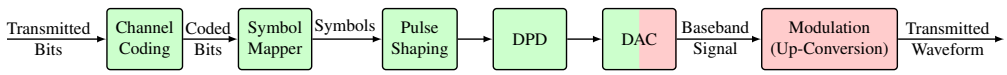

Figure 2.2: Block diagram of a conventional transmitter. Blocks highlighted in green operate in the digital domain, while the rest (i.e., blocks in red) operate in the analog domain.

Modulation Format

The information bits, after being encoded (or protected) by a channel code, are segmented into blocks of m bits. Each of these blocks is then mapped to a complex-valued symbol. The mapping from a bit sequence to a symbol is specified by a modulation format (also known as a constellation), denoted as $\mathcal{X} = \{x_1, \cdots, x_M\}$, containing M distinct complex-valued symbols. Each symbol (or constellation point) is defined as $x = x_{\mathrm{I}} + jx_{\mathrm{Q}}$, where x_{I} and x_{Q} represent the in-phase and quadrature (IQ) components of the symbol x, respectively. These symbols are typically zero mean and have a variance of E_s.

Common choice of modulation formats for coherent fiber-optic communications are quadrature amplitude modulation (QAM) and phase shift keying (PSK). Fig. 2.3 visualizes some of these used formats, including quadrature phase shift keying (QPSK) (also known as 4-QAM, 4-PSK), 8-PSK, 16-, 32-, 64- and 128-QAM. Assuming that all constellation points are selected with equal probability, these formats carry $m = \log_2 M$ bits. If the symbols are transmitted with a symbol duration of T_{sym}, the symbol rate of

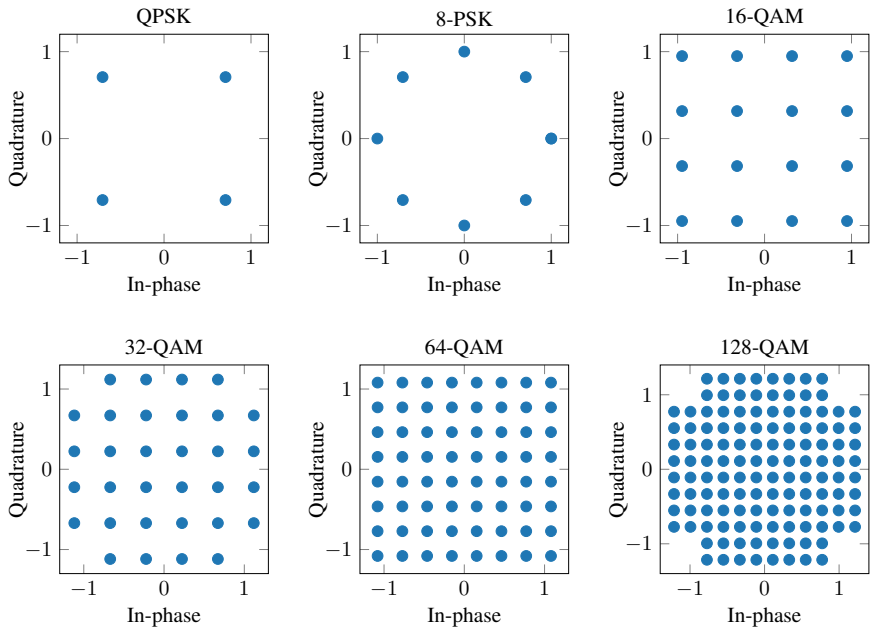

Figure 2.3: Illustration of different PSK and QAM formats with zero mean and unit variance.

this system is

$$R_s = \frac{1}{T_{\text{sym}}}, \tag{2.1}$$

and the corresponding bit rate is $R_b = mR_s$.

The modulation format plays a crucial role in determining the data rate of a communication system. Generally, increasing the number of points (i.e., M) in a constellation $\mathcal{X}$ leads to higher a SE because each constellation point represents a larger number of bits. However, enlarging the constellation size also increases sensitivity to distortions introduced by transmission impairments. Consequently, high-order modulation formats typically exhibit higher detection error probabilities compared to low-order formats. To improve the performance of high SE systems where high-order modulation formats are utilized, constellation shaping has emerged as a prominent research topic in the recent decades [35–42].

Constellation shaping has its origins in information theory established by Shannon [43]. Since all practical channels introduce distortions to transmitted signals, practical communication systems are inherently constrained by the so-called channel capacity, which quantifies how much information a real-world channel can carry. Shannon proved in [43] that channel capacity can be approached by using error correction codes with a large code length, provided that the signal has a capacity-achieving distribution. Neverthe-

less, capacity-achieving distributions are typically unfeasible to implement in real-world systems, and the use of more practical modulation formats, such as constellations with equiprobable constellation points, introduces the so-called shaping gap. The goal of constellation shaping is to reduce this shaping gap caused by the use of sub-optimal modulation formats.

In general, constellation shaping schemes fall into two categories: geometric constellation shaping (GCS) [35–38] and probabilistic constellation shaping (PCS) [39–42]. The former approach involves constellations with non-equally spaced but equiprobable points, meaning that it modifies the geometric locations of the constellation points. The latter approach entails placing constellation points with varying probabilities on a fixed grid, typically using square QAM formats as templates. In this thesis, the use of AEs for GCS is considered directly or indirectly in the appended papers. Detailed discussions regarding GCS with AEs will be presented in Section 4.1.

Pulse Shaping

After mapping the bit sequences to symbols of the chosen modulation format, it is necessary to transform the symbols into a waveform, i.e., through pulse shaping, so that the transmitted signal is suitable for the transmission channel [44–46]. Specifically, since practical channels are bandwidth limited, e.g., due to hardware constraints or only a certain bandwidth is dedicated to a specific user, pulse shaping needs to be performed to limit the bandwidth of the transmitted signal. In practice, pulse shaping can be performed either in the digital domain using digital filters [44, 47] or in analog domain via electrical [46] or optical filtering [48].

Consider performing pulse shaping in the digital domain, the baseband symbols are typically first up-sampled by a factor of N_{os}, e.g., by inserting $N_{\text{os}}-1$ zeros in between every two consecutive baseband symbols. Subsequently, the resulted symbols are convolved with the chosen filter shape $p[n]$, i.e.,

$$x_{\text{ps}}[n] = \sum_{k=1}^{N_s} x_{\text{os}}[n]p[n-k], \tag{2.2}$$

where x_{os} represents the upsampled signal and N_s the is the length of the pulse shaping filter. Common choices of the pulse shape in communication systems are the raised-cosine and the root-raised cosine pulses with a small roll-off factor [45, 49]. While it should be noted that using a small roll-off factor typically leads to a high peak-to-average power ratio (PAPR) [50], which increases the system's sensitivity to both hardware and transmission impairments.

Digital Predistortion

The hardware components, such as the digital-to-analog converters (DACs) and power amplifiers (PAs) at the transmitter, are typically non-ideal, resulting in degraded sys-

tem performance due to a cascade of linear and nonlinear distortions [11].[1] To mitigate the performance degradation caused by transmitter imperfections, state-of-the-art communication systems employ DPDs to compensate for these impairments [21, 51, 52]. Linear filters, also referred to as digital pre-emphases, can be used to compensate for the frequency responses of the DACs [20, 53]. Nonlinear memoryless DPDs based on the *arcsin* function [21, 54] can be utilized to compensate for the intrinsic sinusoidal response (see (2.15)) of the Mach-Zehnder modulators (MZMs), which are commonly employed in fiber-optic communication systems. Since the *arcsin*-based DPD leads to an increase in the signal's PAPR, it is often used in combination with a clipping operation [21]. More sophisticated models based on the generalized memory polynomials (GMPs) [55], Volterra series [51, 52], and NN [12, 56, 57] can also be used to compensate for the cascaded linear and nonlinear responses of the transmitter hardware. These methods often require model parameters optimization, which can be performed using either direct or indirect learning [13].

Digital-to-Analog Conversion

Digital signals cannot be directly transmitted over the channel (e.g., through the air or an optical fiber), and it is necessary to convert digital signals into an analog representation using DACs. Typically, DACs are characterized by features such as bit resolution, sampling frequency, signal-to-noise-and-distortion ratio (SNDR), and effective number of bits (ENOB) [58]. Bit resolution determines the minimum changes in the DAC output, and due to finite bit resolution, DACs inevitably add quantization errors/noise to the transmitted signal. Assuming an ideal DAC with N quantization bits, the quantization noise can be related to the resulted SNR according to

$$\text{SNR(dB)} = 6.02N + 1.76\text{dB}. \tag{2.3}$$

In practice, other distortions, such as sampling and jitter effect [59], will also contribute to this noise and the total amount of distortions are often characterized by a measurable quantity referred to as SNDR. A related parameter that assesses the total amount of noise introduced by DAC is ENOB, a quantity translated from SNDR using the theoretical SNR formula (i.e., (2.3)) of an ideal converter [58],

$$\text{ENOB} = \frac{\text{SNDR(dB)} - 1.76}{6.02}. \tag{2.4}$$

In practice, due to the bandwidth limitations, ENOB is a varying quantity and it changes over frequencies [58].

For optical-fiber communications, the high-speed DACs typically have an 8-bit nominal resolution, which can be translated into ENOB ≤ 6 for operation within the device bandwidth. Quantization noise can be modeled as additive white Gaussian noise (AWGN)

[1] The hardware distortions considered in this thesis are briefly reviewed in Section 2.5.

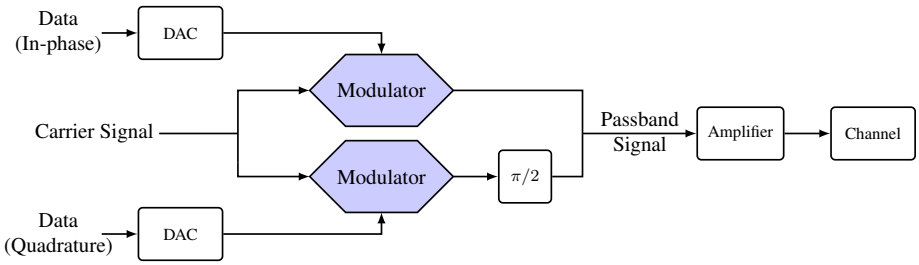

Figure 2.4: Overview of a typical coherent transmitter in a fiber-optic communication system.

with zero mean and a variance determined by the device's ENOB [60], i.e.,

$$\sigma_q^2 = \frac{1}{12}\left(\frac{E_{\text{peak}}}{2^{\text{ENOB}-1}-1}\right)^2, \tag{2.5}$$

where E_{peak} is the peak amplitude of the input signals. One may also model the quantization noise as a uniformly distributed random variable [61].

Modulation (Up-Conversion)

The final stage of the transmitter is to modulate the baseband signal (i.e., the output of the DAC) onto a carrier and then send it over the channel. In complex notation, the modulated signal can be mathematically written by [62, Ch. 4]

$$x_{\text{mod}}(t) = x(t)e^{j2\pi f_c t}, \tag{2.6}$$

where $x(t)$ is the complex-valued baseband signal and f_c is the carrier frequency.[2] Depending on the transmission medium, this procedure is realized using different hardware. Particularly, for fiber-optic communications where information is transmitted in the form of a lightwave, modulation is achieved using the so-called optical modulators. Wireless systems, on the other hand, transmit information in the form of radio waves, and the conversion from an electrical baseband signal to a radio wave involves devices such as mixers, oscillators, and antennas, etc [63]. In the following, we exemplify the modulation procedure in an optical communication system. For detailed modulation procedure for a wireless system, we refer the readers to [63].

Fig. 2.4 visualizes a high-level representation of the modulation procedure in a single-polarized fiber-optic communication system. The optical carrier signal, generated by a laser, is split into two beams, which then enter two modulators responsible for modulating the in-phase and quadrature components of the baseband signal onto the optical carrier. Then, the quadrature component is phase-shifted by $\pi/2$ and combined with the in-phase component. Finally, the resulted lightwave is amplified and transmitted over the channel.

[2] We consider transmitting QAM signal; The modulated/transmitted signals are also often expressed by $x_{\text{tx}}(t) = \text{Re}\{x(t)e^{j2\pi f_c t}\} = \text{Re}\{x(t)\}\cos(2\pi f_c t) - \text{Im}\{x(t)\}\sin(2\pi f_c t)$.

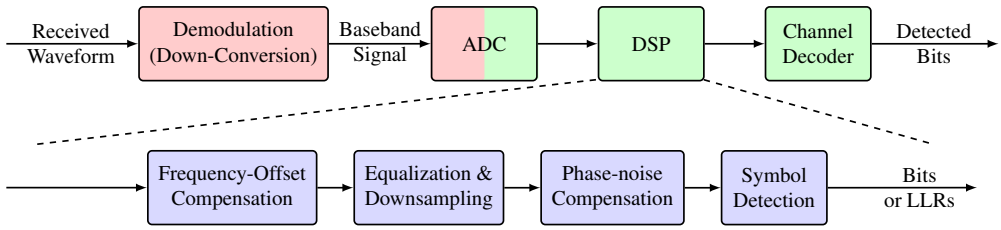

Figure 2.5: Block diagram of a coherent receiver in a fiber-optic communication system. ADC: analog-to-digital converter; LLRs: log-likelihood ratios.

2.3 Classical Receiver

Figure 2.5 shows the different blocks of a receiver in coherent optic-fiber communication systems. The signal transmitted over the channel is detected by the receiver and demodulated to the baseband. Assuming a perfect optical channel that introduces neither distortion nor noise, one may write the demodulated signal (after low-pass filtering and normalization) as [62, Ch. 5]

$$x_{\text{demod}}(t) = x(t)e^{-j(2\pi f_{\text{IF}} t + \Delta\phi(t))}, \tag{2.7}$$

where f_{IF} and $\Delta\phi(t)$ represent the frequency offset and phase difference between the carrier signal and the local oscillator (LO).[3] Since practical channels, as well as the transceiver hardware, introduce distortions to the transmitted signal, the demodulated signal after analog-to-digital converter (ADC) needs to be processed by DSPs before symbol detection. Note that for the sake of complexity considerations, the signals after ADC are often resampled to have two samples/symbol before further processing [5,65,66].

Fig. 2.5 (the bottom branch) depicts a simplified DSP chain in the coherent receiver. Note that the ordering of the DSP steps is not unique [65], and the chain does not include all possible techniques performed in the receiver, such as orthonormalization [67, 68] and timing recovery [69, 70]. In the rest of this section, we review algorithms from the literature that implement the DSP blocks in Fig. 2.5. For more detailed reviews on DSP techniques for coherent fiber-optic communications, we refer the readers to [65, 71].

Frequency Offset Compensation

The coherent receiver in modern communication systems typically performs the so-called heterodyne detection, where a lightwave generated by the so-called LO is mixed with the received signal to extract the in-phase and quadrature components of the transmitted signal [62]. The LO is tuned to approximately match the frequency of the received carrier wave, resulting in a frequency and phase mismatch between the LO and the received

[3] We assume heterodyne detection [64], where the frequency and phase of the LO are not locked to that of the carrier signal.

signal. After analog-to-digital conversion, this frequency mismatch manifests as a linear phase rotation of the received samples (i.e., see (2.7)), and needs to be compensated for ensuring reliable detection.

To compensate for the phase rotation induced by frequency offset, several blind algorithms (i.e., algorithms that do not require pilots transmission) have been proposed for frequency offset estimation (FOE). To name a few, a differential phase-based method was proposed in [72], facilitating a maximum likelihood estimate of the frequency offset. A similar method, presented in [73], conducts FOE recursively, enabling a hardware-efficient implementation. Spectral methods can also be used for FOE. The basic idea is to preprocess the received samples by raising them to the M–th power and then performing a Fourier transform. Subsequently, a frequency offset estimate can be obtained by finding the peak in the resulted spectrum [74]. Based on this concept, an iterative method was proposed [75], improving upon the estimation accuracy and effectiveness for higher-order QAM. While [74] and [75] being feedforward techniques, feedback techniques employing a frequency-controlled loop may also be used, having the advantage of being agnostic to the modulation format [76, 77].

As an alternative to blind methods, data-aided approaches can also be used for FOE. A method based on removing the modulated phase from the received signal using training sequence was proposed in [78], where the frequency offset can be calculated from the averaged phase difference between consecutive symbols. A similar method using asymmetric-shape constellations was proposed in [79] to improve the robustness against timing errors. Various other data-aided methods have also been reported in [80, 81]. A comprehensive review on various blind and pilot-based algorithms for FOE is provided in [74].

Channel Equalization

Channel equalization represents another critical DSP component essential for ensuring reliable and high-quality communication when confronted with channel impairments. While, in principle, equalization could be realized within one DSP block, it is generally beneficial to partition the problem into static and dynamic equalization for fiber-optic communication systems. Static equalization typically requires statics filters with large number of taps, and are often used to compensate for static impairments such as chromatic dispersion (CD) [82]. Dynamic (or adaptive) equalizations, on the other hand, typically use relative short adaptive filters to compensate for time-varying effects, such as polarization rotations and polarization mode dispersion [66]. In the following, we review algorithms from the literature that implements adaptive equalizations. Note that, as mentioned before, adaptive equalizers typically operate with two samples/symbol, whereas the output only has one sample/symbol, i.e., downsampled by using fractional spaced equalizers [5].

Traditionally, linear finite impulse response (FIR) filters have been utilized for adaptive equalization, mainly because of their low implementation complexity. These equalizers

are usually updated recursively by minimizing a cost function through an update algorithm, such as gradient descent, until convergence is reached [66, 83, 84]. Depending on whether the cost function includes known transmitted pilot symbols as input, similar as FOE methods, equalizers are generally classified as either blind or data-aided equalizers (also referred to as pilot-based or non-blind equalizers).

Several blind equalizers have been proposed in the literature, differing mainly in the cost function used to update the filter taps. The most popular algorithm for blind adaptive channel equalization is the constant modulus algorithm (CMA) [83] and its variants, e.g., the modified CMA [85]. The CMA was originally designed for linear channels and phase shift keying constellations, but can also converge for QAM formats. For multi-modulus formats (e.g., QAM signal), the constant-modulus criterion is not fulfilled, indicating that the CMA has suboptimal convergence and steady-state performance. In this case, other variants are more effective, such as the the multi-modulus equalizer [86], the radially-directed equalizer [65], or decision-directed equalizer [87].

Data-aided methods can also be used for adaptive equalization, and have also been extensively researched in the literature. The most popular yet simple non-blind algorithm for adaptive channel equalization is the least-mean-square equalizer, designed to find the filter coefficients that produce the least mean square value of the error signal (i.e., the difference between the desired output and the actual output signal) [66, 88]. Using variable-step-sizes, the convergence speed of conventional least-mean-square equalizers can be improved [84]. The recursive least squares algorithm is also a popular choice for adaptive channel equalization, involving the minimization of an exponentially weighted cost function and treating the minimization problem as deterministic [89–91]. The Stokes-Space method can also be used to update the equalizer coefficients [92]. In general, data-aided equalizers are modulation formats independent, and they offer the merits of fast convergence speed and reliable training. However, due to the fact that pilot symbols do not carry information, the use of data-aided equalizers leads to a reduction in SE.

Finally, it is worth noting that nonlinear equalization has emerged as a popular research topic in recent years, primarily due to the utilization of high-modulation formats in high SE systems. Traditionally, nonlinear equalizers have been commonly implemented using Volterra series [93–95] or their variants, such as GMP [96]. However, nonlinear equalizers based on NNs [97–99] have also been extensively studied in the recent years. In Section 3.5, we review NN-based equalizer in the literature.

Phase-Noise Estimation

The presence of phase noise (PN) in coherent systems necessitate the use of phase noise compensation (PNC) prior to symbol detection. The most commonly used PNC algorithms for fiber-optic communications have traditionally been blind, due to their merit of no reduction in SE. Although blind algorithms lack a priori knowledge of the transmitted symbols, the structure of some modulation formats can be exploited to estimate the PN. As an example, M-PSK comprises M equispaced constellation points on a circle in the

complex plane. When observations corresponding to this modulation formats are raised to the M-th power, the modulated phase is removed and the PN can be estimated in a range of length $2\pi/M$. Subsequently, the phase-noise estimates are processed and then used to remove the phase error. The Viterbi-Viterbi algorithm [100] is based on this concept and work effectively for M-PSK. However, for higher-order QAM, this method works sub-optimally as the constellation points generally do not have equispaced phases. To alleviate this problem, modified Viterbi-Viterbi algorithm based on QPSK partitioning [101,102] has been shown to improve the performance of the standard approach. Another widely-used blind method for PNC is the blind phase search (BPS) algorithm [103], an approach that yields good performance but has a high computational complexity for higher-order modulation formats. Several BPS variants have been proposed to reduce the computational complexity while maintaining or even improving the performance of the original method [104,105]. Furthermore, PNCs based on Kullback-Leibler divergence analysis [106], principal component analysis [107], or hybrid method [108–110] have been proposed in the literature.

Pilot-aided algorithms for PNC have also been extensively researched due to their independence of modulation formats and the ability to provide unambiguous PN estimates. By exploiting the statistical structure of the system model, numerous methods have been reported to find near-optimal PN estimators using probabilistic inference frameworks. To name a few examples, an algorithm employing probabilistic arguments was proposed in [111], where laser PN and nonlinear phase noise (NLPN) were jointly compensated in a wavelength-division multiplexing (WDM) transmission with ideal distributed Raman amplification. A different method based on the Kalman filter was proposed in [112], performing joint laser PN and NLPN compensation. A similar method was reported in [113], where a Kalman filter-based phase estimator was initially trained using pilot symbols and later switched to decision-directed mode once convergence was achieved. PNC can also be implemented through phase interpolation [114], where it is assumed that the phase changes linearly between every two consecutive pilot symbols. Finally, a literature review of various symbol detectors for transmission in the presence of phase noise is provided in [115].

Data Detection

After all impairments have been compensated, data detection is performed. In the case of uncoded transmission, data detection is typically carried out by first performing symbol detection, followed by a demapper that maps the detected symbols back to bit sequences. The optimal symbol detector is the maximum a posteriori (MAP) detector, defined by

$$\hat{x}_{\mathrm{MAP}}(y) = \underset{x\in\mathcal{X}}{\operatorname{argmax}}\, p_X(x) f_{Y|X}(y|x), \tag{2.8}$$

where x and y are the transmitted and received symbols, respectively, and $f_{Y|X}(y|x)$ is the channel transition probability density function. In the case of transmitting symbols

with equal probabilities, i.e., $p_X(x) = 1/M$, the MAP rule is equivalent to the maximum-likelihood (ML) rule

$$\hat{x}_{\text{MAP}}(y) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} f_{Y|X}(y|x). \tag{2.9}$$

For the AWGN channel, the ML detector operates on a symbol-by-symbol basis and detects each symbol by finding the constellation point closest to the received sample in terms of Euclidean distance [116, Ch. 3].

In the case of coded transmission, the data-detection block operates differently depending on the coding scheme being used. For hard-decision binary coded modulation, the data-detection proceeds in the same way as uncoded system (i.e., first symbol decision, and then mapped to bits). However, when soft-decision binary code is used, the demapper maps the detected symbols into posterior probabilities that describes the detected bits being "0" or "1". Then, the soft (channel) decoder takes the posterior probabilities (often in the form of log-likelihood ratios (LLRs)) as input and recover the transmitted bit sequences.

2.4 The Considered Channels

The channel over which the information is transmitted is non-ideal and introduces distortion to the transmitted signal. In this section, we describe the channel models considered in the appended papers, all of which are discussed in discrete time. Additionally, to arrive at the discrete-time model for the optical fiber channel, a brief description of waveform propagation in the optical fiber is also provided.

The Additive White Gaussian Noise Channel

The simplest, yet most commonly used, channel model for analyzing the performance of communication systems is the AWGN channel, defined by

$$y = x + n, \tag{2.10}$$

where x and y represents the complex-valued transmitted and received symbols, respectively, and $n \sim \mathcal{CN}(0, \sigma^2)$ is the complex Gaussian noise. The mapping from x to y is characterized by the conditional probability distribution

$$f_{Y|X}(y|x) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|y-x|^2}{\sigma^2}\right). \tag{2.11}$$

The AWGN channel considers only white Gaussian noise, while other phenomena such as fading, multi-path effects, dispersion, nonlinearity, etc., which occur in wireless or optical fiber channels, are neglected. However, it can still be used for analyzing system performance, providing insights for the system design.

The Optical Fiber Channel

The waveform propagation of a single-polarized signal through the optical fiber and how it evolves with respect to the transmission distance z and time t is described using the nonlinear Schrödinger equation (NLSE) [22]. Considering signal attenuation, CD, and nonlinear effects in an SMF,[4] the NLSE can be written as

$$\frac{\partial A(z,t)}{\partial z} = \underbrace{-\frac{\alpha}{2}A(z,t)}_{\text{Attenuation}} \underbrace{- j\frac{\beta_2}{2}\frac{\partial^2 A(z,t)}{\partial t^2}}_{\text{Dispersion}} + \underbrace{j\gamma|A(z,t)|^2 A(z,t)}_{\text{Nonlinearity}}, \tag{2.12}$$

where $A(z,t)$ is the electrical field in complex baseband propagating along the fiber at distance z and time t, α is the attenuation coefficient, β_2 is the CD coefficient, and γ is the nonlinear Kerr parameter. The optical light has two orthogonal polarizations, and the signal propagation of a dual-polarized signal can be modeled by the Manakov equation [117].

Exact analytical solutions for the NLSE and Manakov equation have not been found in general, rendering these equations challenging for system design and analysis. Nevertheless, it is possible to obtain a numerical evolution of the transmitted waveform using the split-step Fourier method (SSFM). The key idea of this method involves discretizing the signal propagation along a fiber span into K_{step} small spatial steps, allowing for the separation and analytical expression of dispersion and nonlinearity in each step. In general, increasing the number of steps results in higher accuracy, but it comes at the expense of increased computational complexity.

Simpler models, which approximately describe signals that have propagated through the optical-fiber link and potentially undergone some processing at the receiver are also of interest in order to facilitate system design. One of these simplified models is the widely used NLPN model [118–120], which is memoryless and defined by the recursion

$$x^{(k+1)} = x^{(k)} e^{j\gamma L|x^{(k)}|^2/K} + n^{(k+1)}, \quad 0 \le k \le K, \tag{2.13}$$

where $x^{(0)} = x$ and $y = x^{(K)}$ are respectively the channel input and output, $n^{(k+1)} \sim \mathcal{CN}(0, \sigma^2/K)$, L is the fiber length, and σ^2 is the total noise power. Note that this model can be derived from (2.12) by setting $\alpha = \beta_2 = 0$, and it reverts to the AWGN when setting the nonlinear Kerr parameter γ to zero.[5] Alternatives models, including the Gaussian noise (GN) model [122], and the nonlinear interference noise (NLIN) model [10, 123] have also been widely studied in the literature. A detailed review on optical fiber channel model is provided in [124].

[4] Other effects such as Raman scattering and third-order CD are excluded.

[5] (2.13) also considers amplification noise and it is in fact derived from the stochastic NLSE [121].

The MIMO Channel

A third channel consider in this thesis is the MIMO channel. Consider the channel being memoryless and discrete, a MIMO channel is defined as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}, \tag{2.14}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are respectively the complex-valued transmitted and received symbol vectors, $\boldsymbol{H} \in \mathbb{C}^{N_\mathrm{R} \times N_\mathrm{T}}$ denotes the channel matrix of a channel with N_T input ports and N_R output ports, and $\boldsymbol{n} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{N_R})$ is independent and identically distributed Gaussian noise.

The concept of the MIMO communications was initially introduced in the context of wireless communications, where multiple-antenna techniques were employed to transmit parallel data streams. In the last decade, this concept has been adapted to the field of fiber-optic communications, particularly in the context of space-division multiplexing (SDM) systems. In SDM systems, parallel data streams are transmitted simultaneously through different physical dimensions of an optical fiber (e.g., different modes, different cores, or combined) [125]. Depending on the type of fiber being used, the SDM channel exhibits different characteristics, and various channel models have been developed for both multi-core [126–128] and multi-mode [129–131] fibers. Comprehensive reviews on recent advances in SDM techniques can be found in [132–134].

2.5 Hardware Impairments

In addition to the transmission channel, the transceiver hardware used in practical communication systems are commonly non-ideal, adding distortions to the transmitted signal and thereby limiting the performance of practical systems. In this section, we briefly review the hardware-induced distortions considered in this thesis, including bandwidth limitations and transmitter nonlinearities. Other hardware-related distortions, such as IQ-imbalance, are not considered in the appended papers and are consequently not discussed here. Additionally, PN and quantization noise have been described in the previously section and are also excluded from further discussion.

2.5.1 Bandwidth Limitations

The bandwidth of hardware components (e.g., PAs, DACs, or the modulator in an optical system) used in the transmission system determines the maximum transmission rate the system can support. The frequency response of a bandwidth-limited transmitter hardware can be viewed as a linear low-pass filter, as described in (2.2). When a high-speed signal with a bandwidth greater than the supported bandwidth of the transmitter hardware is passed through, the signal pulses broaden and overlap, leading to inter-symbol interference (ISI). According to the Nyquist theorem (also known as the sampling theorem), a signal with a symbol rate of $1/T_\mathrm{sym}$ necessitates a minimum system bandwidth

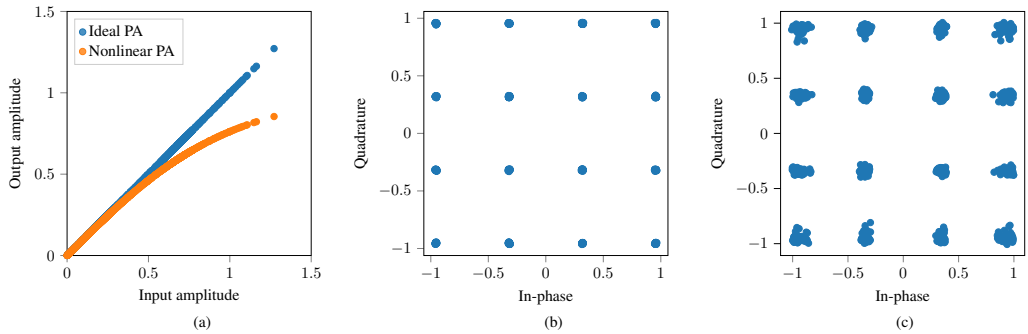

Figure 2.6: An example that illustrates the impact of using a nonlinear PA. (a) PA response; (b) Normalized received signal when PA output power is low; (c) Normalized received signal when PA output power is high.

of $1/T_{\text{sym}}$ for ISI-free transmission [135]. However, in practical applications, the required bandwidth for ensuring ISI-free transmission increases due to limitations imposed by the thermal noise [136].

2.5.2 Nonlinearities

In addition to being bandwidth limited, the PAs as well as the IQ-modulators can add nonlinear distortions to the transmitted signal, further degrading the performance of a communication system. Fig. 2.6 illustrates an example of amplifying 16-QAM signals using a memoryless nonlinear PA, showing the resulting impact on the transmitted and received signals when operating the PA at different output power levels. At low output powers (i.e., when the PA gain is lower), the PA exhibits a linear response, and the output signals remain undistorted. However, when operating in the high-output power region, the PA introduces significant nonlinear distortion to the transmitted signal, resulting in a degradation of received signal quality.

Apart from the PA nonlinearity, the IQ modulator can also introduce nonlinear distortions to the transmitted signal. In particular, the coherent optical transmitter used for high-order modulation schemes such as M–QAM, M–PAM is often based on a dual parallel MZM. Consider an ideal dual parallel MZM biased at the null point, its transfer function can be written as [11]

$$E(t) = E_0 \left[\sin\left(\frac{\pi V_{\text{I}}(t)}{2V_\pi} \right) + j \sin\left(\frac{\pi V_{\text{Q}}(t)}{2V_\pi} \right) \right], \tag{2.15}$$

where E_0 is the amplitude of the magnitude of the optical field, V_π is the required voltage difference to switch ON/OFF the modulator, and $V_{\text{I}}(t)$ and $V_{\text{Q}}(t)$ are the driving voltage of the in-phase and quadrature branches, respectively. The intrinsic sinusoidal response of the MZM leads to strong signal distortions when driving with a high peak voltage

V_p, which must be compensated, e.g., by pre-distortion with an *arcsin* function [21]. Alternatively, one can use a low-driving voltage to operate in the near-linear regime of the modulator. However, this significantly increases the modulator loss, which results in a degraded optical SNR after adding the booster amplifier noise.

2.6 Performance Metrics

In this section, we briefly review the main performance metrics employed in fiber-optic communications. Typically, the metrics of interest revolve around the reliability of information transmission and the capacity to convey information over the channel, while considering reliability criteria. Additionally, other metrics can also be useful for gaining insight during the design of DSP algorithms.

Error Probability

Detection error probability is a common metric in communication systems used to evaluate the reliability of transmission. The most common metrics for approximating error probability in fiber-optic communications are symbol error rates (SERs) and bit error rates (BERs). The SER, defined as

$$P_s = \sum_{x \in \mathcal{X}} p_X(x) \Pr(\hat{x} \neq x | x), \tag{2.16}$$

where $\mathcal{X}$ represents the constellation being used, and x and $\hat{x}$ denote the transmitted and detected symbols, quantifies the average probability of the detector making an error. Analogously, the BER corresponds to the probability that the detector makes a wrong bit decision, i.e.,

$$P_b = \frac{1}{m} \sum_{k=1}^{m} \sum_{x \in \mathcal{X}} p_X(x) \Pr(\hat{b}_k \neq b_k | x), \tag{2.17}$$

where $\hat{b}_k$ and b_k, for $k \in \{1, \ldots, m\}$, are the k-th bit of $\hat{x}$ and x.

Error probabilities are often challenging to compute analytically, but they can be estimated numerically through Monte Carlo simulations. In general, the lower the error probability, the more challenging it becomes to estimate it numerically with reasonable accuracy. In the context of fiber-optic communications, the BER after forward error correction (FEC) (which is often referred to as the post-FEC BER), is often targeted to be as low as 10^{-15}, making direct estimation infeasible. Consequently, it is common to estimate the BER before FEC, i.e., the pre-FEC BER, which is then used to predict the post-FEC BER performance of the system for both hard- and soft-decision decoding schemes [137, 138].

Achievable Information Rates

The use of SERs or BERs is effective for hard-decision decoding. However, in the case of soft-decision decoding, achievable information rates (AIRs) have been found to be more accurate predictors of the post-FEC BER performance. AIRs determine how much information can be conveyed over a channel with an arbitrarily low error rate, assuming the use of a certain modulation format and a capacity-achieving FEC code with ideal decoding [139, 140]. For fiber-optic communications, the common choice of AIRs are the mutual information (MI) and the generalized mutual information (GMI).

The MI between random transmitted and received symbols X and Y is defined as

$$I(X;Y) = \sum_{x\in\mathcal{X}} p_X(x) \int f_{Y|X}(y|x) \log\left(\frac{f_{Y|X}(y|x)}{f_Y(y)}\right) dy. \tag{2.18}$$

Evaluating (2.18) requires an analytical expression to $f_{Y|X}(y|x)$, which is hardly the case for practical optical fiber channels. Consequently, a closed-form expression to (2.18) is generally unknown. A work-around to this challenge is to bound (2.18). For example, a lower bound on (2.18) can be achieved by using an auxiliary channel transition probability $\hat{f}_{Y|X}(y|x)$ instead of the true one $f_{Y|X}(y|x)$ [141]. Alternatively, one may obtain an estimate of the MI using Monte-Carlo approximation

$$\begin{aligned} I(X;Y) &\approx \frac{1}{N_s}\sum_{i=1}^{N_s} \int f_{Y|X}(y|x^{(i)}) \log\left(\frac{f_{Y|X}(y|x^{(i)})}{f_Y(y)}\right) dy \\ &\approx \frac{1}{N_s}\sum_{i=1}^{N_s} \log\left(\frac{f_{Y|X}(y^{(i)}|x^{(i)})}{f_Y(y^{(i)})}\right) \\ &\approx \frac{1}{N_s}\sum_{i=1}^{N_s} \log\left(\frac{f_{Y|X}(y^{(i)}|x^{(i)})}{\sum_{x\in\mathcal{X}} p_X(x) f_{Y|X}(y^{(i)}|x)}\right), \end{aligned} \tag{2.19}$$

where $x^{(i)}$ for $i = 1, \ldots, N_s$ are N_s samples drawn from $p_X(x)$, and $y^{(i)}$ is drawn from the conditional distribution $f_{y|X}(y^{(i)}|x^{(i)})$ when a certain $x^{(i)}$ is given.

The MI is a good performance predictor when a non-binary soft decision channel code is used. However, achieving the MI is only possible for systems using a symbol-wise decoder and is generally unattainable when a bitwise decoder is employed. In the case of binary codes with soft-decision decoding, the GMI, a lower bound on the MI [139], is often a better (more suitable) performance indicator. Assuming the transmission of constellation symbols with equal probability, i.e., $p_X(x) = 1/M$, where each symbol carries m bits, the GMI can be write in a simple form given by

$$\mathrm{GMI} = \frac{1}{M}\sum_{k=1}^{m}\sum_{b\in\{0,1\}}\sum_{x\in\mathcal{X}_k^b} \int f_{Y|X}(y|x) \log\left(\frac{\frac{1}{M}\sum_{x\in\mathcal{X}_k^b} f_{Y|X}(y|x)}{\frac{1}{2} f_Y(y)}\right) dy, \tag{2.20}$$

where $\mathcal{X}_k^b \subset \mathcal{X}$ is the set of constellation points with a bit b at position k in their m-bit binary label. As for the transmission of PCS symbols, the GMI generally has a more complex form, and detailed derivations are provided in [40]. Finally, a closed-form solution is also unavailable to the GMI, and an estimate to the GMI can be obtained from Monte Carlo simulation [139].

Mean-Squared-Error

Another performance metric widely used in communications is the mean-squared-error (MSE). It computes the average squared error, where the error is the difference between the estimate and the ground true. Mathematically, this is written as

$$\text{MSE} = \frac{1}{N_s} \sum_{i=1}^{N_s} |\hat{x}_i - x_i|^2, \tag{2.21}$$

where $\hat{x}_i$ and x_i are the estimates and ground true data, and N_s is the number of samples. This metric is frequently used in the context of channel equalization [89, 90], as well as PNC [104,110]. It is also commonly used for training machine learning algorithms, which we show in the next Chapter.

CHAPTER 3

Deep Learning Basics

This chapter aims to provide a brief introduction to the general theory behind DL and its applications to the physical layer of communications.[1] We begin by outlining the main categories of the DL algorithms in Section 3.1, followed by a brief review of the fundamental concepts of NNs and AEs in Section 3.2. In Section 3.3, we introduce several commonly used loss functions in the machine learning literature and introduce the concept of gradient-based learning in Section 3.4. Finally, the chapter concludes in Section 3.5, where we review the applications of DL techniques in coherent fiber-optic communications.

3.1 Types of Machine Learning Algorithms

Machine learning algorithms can be generally categorized into three sub-fields, namely supervised learning, unsupervised learning, and RL.[2] In the rest of this section, we review the basic concepts of the above-mentioned learning algorithms. The loss functions commonly used in training these different types of algorithms are discussed in Section 3.3.

[1] While this thesis focuses on applications related to DL, the term "machine learning" is sometimes used instead of DL, especially when discussing topics that require a more general perspective.

[2] One may argue that semi-supervised learning is another category of machine learning algorithm that lies at the intersection of supervised and unsupervised learning. It is not studied in this thesis, and its discussions are excluded.

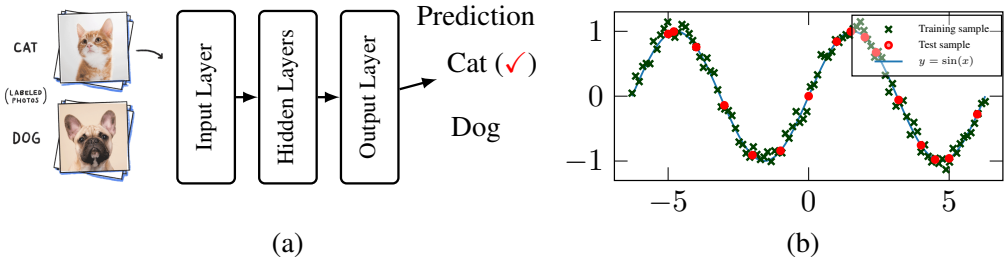

Figure 3.1: Examples of supervised learning. (a) A classification algorithm that learns to distinguish between images of cats and dogs; (b): A regression algorithm aiming to discover the relationship between two continuous variables x and y.

3.1.1 Supervised Learning

Supervised learning involves the task of inferring a function that maps an input to an output based on labeled training data (i.e., using input-output pairs for training). The primary objective is to unveil the relationships between the input features and the output labels, allowing the model to generalize and provide accurate predictions for new or unseen data. Depending on the type of problems being addressed, supervised learning can be further categorized into two groups: classification and regression.[3] A classification algorithm predicts the class associated with the given input data, yielding a *categorical* or *discrete* class label as the output. Regression algorithms, on the other hand, aim to establish a relationship between the given input and a *continuous* output variable.

Fig. 3.1 visualizes two examples, showing a classification algorithm on the left and a regression algorithm on the right. The classification algorithm, whose goal is to train a classifier capable of correctly distinguishing between pictures of cats and dogs [142], outputs a discrete class label (i.e., either a cat or a dog) given an input image. As for the regression algorithm, its goal is to find out that the output y is related to the input x according to $y = \sin(x) + n$, where n represents the random observation noise. If properly trained, given any test data x_{test}, the regression algorithm is expected to output a continuous variable according to $y_{\text{test}} = \sin(x_{\text{test}})$.

3.1.2 Unsupervised Learning

In contrast to supervised learning algorithms, which require labeled datasets for training, unsupervised learning operates with unlabeled data, with the aim of discovering hidden patterns in the dataset without human intervention. Similar to supervised learning, unsupervised learning can also be divided into different categories based on the type of problems being addressed. In the following, we briefly review some common use cases of unsupervised learning.

[3] Classification is also referred to as logistic regression.

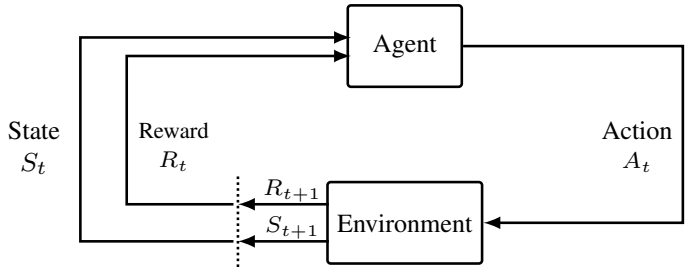

Figure 3.2: A model showing the relations between the key components in RL.

One common application of unsupervised learning is clustering, where the goal is to identify groupings in an unlabeled dataset. Using the earlier example of a dataset containing imagines of cats and dogs (but note that labels are no longer available), a clustering algorithm aims to identify two distinct categories within the dataset. Specifically, the clustering algorithm recognizes the common features of dogs, which contrast with the common features of cats. However, since no labeled data is involved in the training procedure, unlike the supervised learning approach, the clustering algorithm cannot determine which cluster corresponds to cats or dogs.

Another widely used application of unsupervised learning is dimension reduction, where the goal is to decrease the number of input variables or features in a dataset while retaining as much relevant information as possible. The way it typically works is to simplify the dataset by projecting it onto a lower-dimensional space. An example of a dimension reduction algorithm is the widely used principle component analysis (PCA) [143], which identifies orthogonal axes (i.e., principal components) along which the data exhibits the highest variance. The algorithm is linear, efficient for high-dimensional data, and easy to interpret [144]. However, if complex nonlinear relationships exist in the data, PCA may not perform well. As a nonlinear alternative to PCA, dimension reduction techniques based on AEs have garnered significant interest in recent years [145,146]. In Section 3.2, we provide a more detailed description of AEs.

In addition to the applications described above, unsupervised learning has many other uses, including anomaly detection [147], density estimation [148], generative modeling [149] and more. Various unsupervised learning algorithms are reviewed in [150].

3.1.3 Reinforcement Learning

RL is another paradigm of machine learning that focuses on training intelligent agents to make sequential decisions in an environment with the aim of maximizing some cumulative rewards [151]. Unlike supervised learning, where the algorithm learns from labeled examples, or unsupervised learning, where the algorithm discovers patterns in unlabeled data, RL is concerned with training an intelligent agent that learns to make optimal decisions through interactions with an environment.

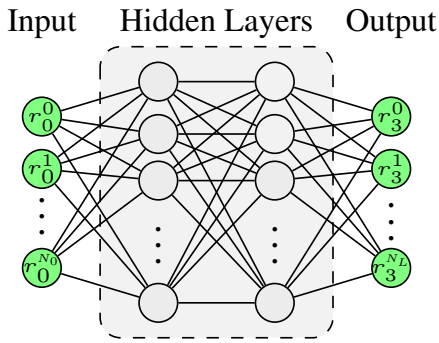

Figure 3.3: An example of a fully-connected NN consists of an input layer, and output layer and two hidden layers.

Fig. 3.2 depicts a standard RL model in which an agent is connected to its environment through perception and action. At each iteration step t, the agent receives the current state S_t of the environment as input and generates an action A_t following some policy (i.e., a set of rules). The action taken by the agent changes the state of the environment, and the value of this state transition is communicated to the agent through a scalar signal known as the reward R_t. The learning goal of an RL algorithm is to find (learn) the optimal policy, according to which the long-term accumulated reward received from the environment is maximized.

3.2 Neural Networks and Autoencoders

Neural Networks

NNs are parametric models that can present highly complex functions through the composition of serval simple operations. The simplest form of NNs is the feedforward NN, which is a parametric function $\boldsymbol{f}(\boldsymbol{r}_0;\boldsymbol{\theta}) : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ that maps an input vector $\boldsymbol{r}_0 \in \mathbb{R}^{N_0}$ to an output vector $\boldsymbol{r}_L \in \mathbb{R}^{N_L}$ through L sequential processing steps according to

$$\boldsymbol{r}_\ell = \boldsymbol{f}_\ell(\boldsymbol{r}_{\ell-1};\boldsymbol{\theta}_\ell), \quad \ell = \{1,\dots,L\}, \tag{3.1}$$

where L is the number of layers, and $\boldsymbol{f}_\ell(\boldsymbol{r}_{\ell-1};\boldsymbol{\theta}_l) : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ is the mapping carried out by the ℓ-th layer. Here, the mapping of the ℓ–th layer is defined by the set of parameters $\boldsymbol{\theta}_\ell$, and the entire NN is defined by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_L\}$.

A commonly used type of feedforward NN is the fully connected NN (also known as multi-layer perceptron (MLP)), in which all layers have the form

$$\boldsymbol{f}_\ell(\boldsymbol{r}_{\ell-1};\boldsymbol{\theta}_\ell) = \sigma(\boldsymbol{W}_\ell \boldsymbol{r}_{\ell-1} + \boldsymbol{b}_\ell), \tag{3.2}$$

where $\boldsymbol{W}_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ is the weight matrix, $\boldsymbol{b}_\ell \in \mathbb{R}^{N_\ell}$ is the bias vector, and $\sigma(\,\cdot\,)$ is the chosen element-wise activation function [152]. Hence, the set of trainable parameters

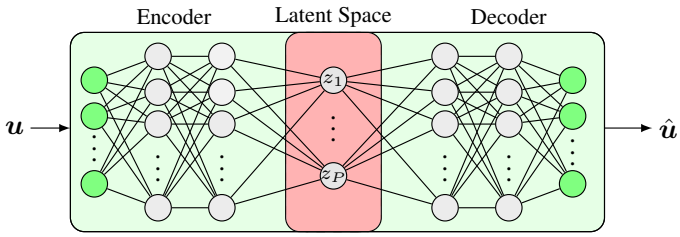

Figure 3.4: An example of an AE composed of two NNs with fully-connected architecture. The encoder and decoder NNs of the AE is connected by a shared layer (i.e., the latent representations).

of the ℓ-th layer is $\boldsymbol{\theta}_\ell = \{\boldsymbol{W}_\ell, \boldsymbol{b}_\ell\}$. In Fig. 3.3, an example of a fully-connected NN consisting of an input layer, an output layer, and two hidden layers, is depicted.

Fully connected NNs excel in capturing complex, nonlinear relationships within data, and it has been proven that a single-layer fully connected NN can serve as a universal function approximator [30]. However, for many practical applications, more specialized network architectures have demonstrated superior performance than the fully-connected NNs in addressing specific challenges within different domains. For example, convolutional neural networks (CNNs) have been shown to be better suited for structured data, and have been widely used for image and video processing [153, 154]. Recurrent neural networks (RNNs), which has strong capability in handling sequential data, have been commonly used for natural language processing [155] and time-series prediction [156,157]. In the machine learning literature, various other specialized (but more sophisticated) network architectures have also been developed for solving diverse tasks. Examples include the long short-term memory [158] and the Transformers [159] for sequential data modeling, as well as the residual NNs [160] and the GoogleNet [161] for imagine processing.

Autoencoders

An AE is an NN designed to learn efficient representations (or encodings) of its input data by training the network to generate a replica of its input data, thereby achieving effective feature extraction and reconstruction of the original data [162, 163]. As depicted in Fig 3.4, an AE is composed of two parts: an encoder $\boldsymbol{f}(\,\cdot\,;\boldsymbol{\tau}) : \mathbb{R}^N \to \mathbb{R}^P$ that transforms the input data $\boldsymbol{u} \in \mathbb{R}^N$ into latent representation $\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{u};\boldsymbol{\tau}) \in \mathbb{R}^P$ (also referred to as encodings) and a corresponding decoder $\boldsymbol{g}(\,\cdot\,;\boldsymbol{\rho}) : \mathbb{R}^P \to \mathbb{R}^N$ that produces a replica $\hat{\boldsymbol{u}} = \boldsymbol{g}(\boldsymbol{z};\boldsymbol{\rho}) \in \mathbb{R}^N$ of original input data from the latent representation. This latent representation, which is typically of lower dimension, i.e., $P < N$, than the input/output, can be considered as a bottleneck in the AE.[4] It enforces the encoder

[4]The overcomplete AEs (e.g., the sparse AEs [164, 165], employ a latent representation of higher dimension than its input/output. They are not used in this thesis and are therefore left out of the discussions.

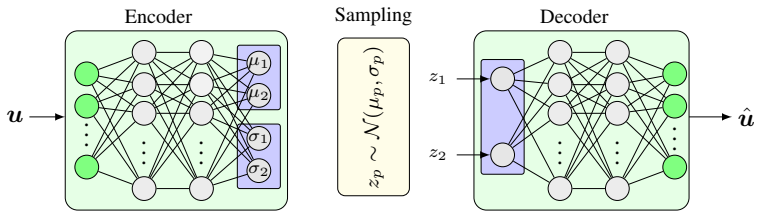

Figure 3.5: An example of a VAE with two latent variables, where these variables are assumed to be independent and each follows a normal distribution. The input to the decoder is mapped to the means and variances of these normal distributions, and samples are generated from the learned distributions when reconstructing the input.

to compress the input data, while ensuring sufficient information (e.g., the important features) are preserved for the decoder to successfully reproduce the input. The goal of an AE training is to minimize the reconstruction error (e.g., the difference between $\boldsymbol{u}$ and $\hat{\boldsymbol{u}}$), and AEs are typically trained in an unsupervised manner. However, in the context of end-to-end AE learning for joint transceiver design, AEs are trained under supervised learning (see Chapter 4).

Variational autoencoders (VAEs) are variants of AEs that introduce probabilistic models into the latent representations [149, 166], and are often used for content generation [167]. The major difference between "conventional" AEs and VAEs is that the the mapping from the input to the latent representation in a AE is deterministic (e.g., given any input, it is mapped to a fixed point in the latent space), while a VAE generates a probabilistic distribution over the latent representation for a given input sample. To elaborate, Fig. 3.5 visualizes an example of a VAE. Given an input data, it is mapped to a probability distribution (typically a multi-variate normal distribution characterized by its mean and variance) that describes the latent representation. To reconstruct the input data, a sample needs to be generated from the learned distribution, which is subsequently used as the decoder input for reconstruction of the original input data.

With the introduction of probabilistic models, the VAE framework has become widely utilized for training generative models. Specifically, after a VAE has been appropriately trained, its decoder can be employed to generate new samples by taking as input the samples (i.e., the latent representations) drawn from the learned probability distributions. For communications, the VAE framework has been applied for blind channel equalization, which we review in Section 4.2.

3.3 Loss Functions

The common practice of training a machine learning algorithm is to formulate a loss function, to which the desired solution is obtained at the minimum of chosen loss function. In the following, we review the loss functions used in this thesis.

Mean-Squared-Error Loss

The MSE that we introduced in Section 2.6 has been commonly used for training both supervised learning and unsupervised learning algorithms. For a supervised learning task with a training dataset $\mathcal{D} \subset \mathcal{R} \times \mathcal{S}$ consisting of $|\mathcal{D}|$ input-output pairs, where each input-output pair is denoted by $(\boldsymbol{r}, \boldsymbol{s})$, the MSE loss is given by

$$J_{\text{MSE}}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{r},\boldsymbol{s})\in\mathcal{D}} ||\boldsymbol{f}(\boldsymbol{r};\boldsymbol{\theta}) - \boldsymbol{s}||^2, \tag{3.3}$$

where $\boldsymbol{f}(\boldsymbol{r};\boldsymbol{\theta})$ is the model prediction, and $\boldsymbol{s}$ is the desired model output (also referred to as labels for a supervised learning task).[5] For unsupervised learning algorithms, e.g., a dimension reduction algorithm, where the training data is not labeled, the desired model prediction for an input data $\boldsymbol{r}$ corresponds to the input data itself (i.e., replacing $\boldsymbol{s}$ by $\boldsymbol{r}$ in (3.3), we get the MSE loss for a dimension reduction algorithm).

Cross-Entropy Loss

Cross-entropy (CE) loss has been ubiquitously employed for training classification algorithms. Consider a classification task containing C distinct classes, with each class denoted by a scalar s, it is customary to design a model that outputs a probability vector of length C for a given input $\boldsymbol{r}$. Here, denoting the model output by $\boldsymbol{q} = \boldsymbol{f}(\boldsymbol{r};\boldsymbol{\theta})$, we have $\boldsymbol{q} \in [0,1]^C$ and all elements in $\boldsymbol{q}$ sum to one.[6] Each element in this probability vector $\boldsymbol{q}$ describes the probability of the input belonging to one class, and classification is typically made by choosing the class with the highest probability.

To train this model, it is popular to minimize the CE loss defined by

$$J_{\text{CE}}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{r},s)\in D} \ell_{\text{CE}}(\boldsymbol{p}, \boldsymbol{f}(\boldsymbol{r};\boldsymbol{\theta})), \tag{3.4}$$

where $\boldsymbol{p}$ is a probability vector associated with the the class label s, and $\ell_{\text{CE}}(\boldsymbol{p},\boldsymbol{q}) = -\sum_{c=1}^{C} p_c \log q_c$ is the CE, a quantitative measure in information theory that assesses the difference between the two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$. In the case where the possible number of classes in the dataset is $C = 2$, each class label is typically represented by either zero or one, i.e., $s \in \{0, 1\}$, and the CE loss simplifies to

$$J_{\text{BCE}}(\boldsymbol{\theta}) = -\frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{r},s)\in D} s \log[(f(\boldsymbol{r};\boldsymbol{\theta})] + (1-s)\log[1-(f(\boldsymbol{r};\boldsymbol{\theta})], \tag{3.5}$$

where it is assumed that model generates the probability of $\boldsymbol{r}$ being class s only (i.e., $f(\boldsymbol{r};\boldsymbol{\theta})$), and the probability of $\boldsymbol{r}$ being the other class is $1 - f(\boldsymbol{r};\boldsymbol{\theta})$. Note that (3.5) is often known as binary cross-entropy (BCE) loss.

[5] The MSE was defined in Section 2.6. Here, it is rewritten in vector form to provide better generality.
[6] We consider the case of single label classification, meaning that one input can only belong to one class.

3.4 Gradient-Based Learning

Training of an NN can be performed in an iterative fashion with data-driven gradient-based methods, for which the goal is to find the set of parameters $\boldsymbol{\theta}^*$ that minimizes the chosen loss function, i.e.,

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\, J(\mathcal{D};\boldsymbol{\theta}), \tag{3.6}$$

where $J(\mathcal{D};\boldsymbol{\theta})$ is the empirical loss associated with the training dataset $\mathcal{D}$, and $\boldsymbol{\Theta}$ is the search space of $\boldsymbol{\theta}$. To achieve (3.6), it is of common practice to minimize $J(\mathcal{D};\boldsymbol{\theta})$ iteratively using gradient descent following

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha\nabla_{\boldsymbol{\theta}} J(\mathcal{D};\boldsymbol{\theta}_t), \tag{3.7}$$

where t is the iteration index, $\alpha > 0$ is the learning rate, and $\nabla_{\boldsymbol{\theta}} J(\mathcal{D};\boldsymbol{\theta}_t)$ is the gradient of $\boldsymbol{\theta}$ averaged over the entire training dataset $\mathcal{D}$ at iteration t. For a small training dataset, $\nabla_{\boldsymbol{\theta}} J(\mathcal{D};\boldsymbol{\theta}_t)$ can be computed (relatively) efficiently via backpropagation. In the case where the training dataset $\mathcal{D}$ is large, evaluating the averaged gradient over the whole training dataset is computationally expensive, and the parameter set $\boldsymbol{\theta}$ is commonly optimized by using a variant of stochastic gradient descent (SGD). In particular, the standard mini-batch SGD approximates the gradient in (3.7) according to $\nabla_{\boldsymbol{\theta}} J(\mathcal{D};\boldsymbol{\theta}_t) \approx \nabla_{\boldsymbol{\theta}} J(\mathcal{B}_t;\boldsymbol{\theta}_t)$, where $\mathcal{B}_t \subset \mathcal{D}$ is a batch of samples randomly sampled from $\mathcal{D}$ at iteration t. In practice, SGD sometimes suffers from slow convergence rate due to problems like small gradients at suboptimal values of $\boldsymbol{\theta}$. To improve the convergence rate of SGD, variants of SGD (such as the Adam optimizer [168]) employing momentum [169] or adaptive learning rate [170] have been proposed.

3.5 Applications of Deep Learning in Communications

In this section, we review the applications of DL techniques in fiber-optic communications. Note that the literature for applications of DL in fiber-optic communication is vast, and we therefore list applications that we deem most representative and most related to this thesis.[7] In particular, we reviews algorithms in the literature that implement modulation format recognition, optical fiber modeling, and nonlinearity mitigation. AE-based techniques are the main focus of the thesis and will be discussed in Chapter 4. Finally, comprehensive reviews on applications of machine learning techniques in fiber-optic communications can be found in [3, 14].

Modulation Format Recognition

Various DL algorithms have been proposed for modulation format recognition (MFR), for which the main motivation (or justification) is that there will be a great need for

[7] We focus on DL applications in the physical layer of fiber-optic communication systems.

flexible transceivers that support multiple data rates and multiple modulation formats in the next generation of fiber-optic networks. To that reason, it is no longer guaranteed that signals arriving at the receiver side will have the same rate and modulation format. Therefore, it is of great importance for coherent receivers to be able to recognize the modulation format of arriving signals to guarantee proper demodulation.

A method that performs MFR was proposed in [171], where the workhorse was a simple MLP. The MLP takes the synchronous amplitude histogram of the received signal as input, and is trained to predict the modulation format associated with the synchronous amplitude histogram using supervised learning. An extended approach was reported in [172], where improved training efficiency was achieved by combining the NN with a genetic method. CNNs have also been used for MFR, wherein the core idea is to treat MFR as image classification. In [173], a CNN was trained to perform MFR using the eye diagram as input. A similar approach was proposed in [174], where the CNN takes the constellation diagrams as input instead of the eye diagram. Various other DL-based algorithms for MFR exist and are reviewed in [175].

Optical Fiber Channel Modeling

As mentioned in the previous chapter, the propagation of signals over an optical fiber is governed by the NLSE, which, in general, cannot be solved in closed form. Although numerical methods, such as the SSFM, allow for accurate simulation of waveform propagation, they typically demand substantial computational resources and become time-consuming for large-scale simulations. To address this challenge, DL techniques based on NNs have been extensively studied for nonlinear fiber modeling.

Several works have focused on simulating waveform propagation using supervised learning [176–178]. In [176], an accurate optical fiber modeling scheme was developed based on training a bidirectional long short-term memory. This method demonstrated similar accuracy to the SSFM-based approach but with significantly reduced computation time. Another hybrid method, reported in [177], achieved reduced running time by incorporating model-based knowledge into the NN-based simulator design. In [178], various NN architectures were compared in terms of both modeling accuracy and computational complexity.

Physics-based methods have also been employed for optical fiber modeling. Drawing inspiration from the observation that feedforward NN can be considered as an iterative solver that alternates between linear and nonlinear steps (which is similar to SSFM), the authors in [9] proposed solving the inverse procedure of waveform propagation (i.e., digital backpropagation) using feedforward NNs. Building on this concept, various fiber nonlinearity compensation schemes have been proposed in the literature [9, 179–181]. While [9] and similar methods explore fiber modeling by leveraging the similarity between a feedforward NN and the SSFM, physics-informed NNs achieve optical fiber modeling by incorporating the NLSE into the NN-based fiber simulator training. This approach has demonstrated that a properly trained fiber simulator can accurately predict the

waveform regardless of the transmission distance [182, 183]. A related line of research is fiber parameter estimation using physics-informed NNs [184, 185].

Generative methods can also be employed for fiber modeling. In [178], an accurate optical fiber simulator was trained using generative adversarial networks. Building on this concept, end-to-end AE-based transceiver learning was demonstrated in [186, 187].

Equalization and Nonlinearity Mitigation

As it has been discussed in the previous chapter, nonlinear effects (e.g., fiber nonlinearity, transceiver imperfection), which degrade the performance of fiber-optic communication systems, need to be compensated in high SE systems. Various nonlinearity mitigation algorithms using DL techniques have been proposed in the literature, differing mainly in the NN architecture used to compensate for the nonlinear effects. A method based on MLP was presented in [188], where it is show a simple MLP can be used to mitigate the performance degradation caused by fiber nonlinearity. A similar approach based on RNN was investigate in [189], where the motivation was based on RNN are good at dealing with sequential data. More sophisticated NN architectures, such as biLSTM [99], have also been studied in the literature, showing either improved performance or reduced computational complexity compared to the simple NN models used in [23, 24].

The methods reviewed above are based on training separate NN equalizers for different operational conditions. More precisely, an NN-based equalizer, often with different hyperparameters, needs to be retrained when the launch power, symbol rate, or fiber length changes. To reduce training overhead (i.e., to avoid training a separate network for each scenario), various methods have been proposed to improve the flexibility of NN-based equalizers. Among these methods, transfer learning-based NN equalizers have been shown to achieve promising performance [190]. The operational principle of a transfer learning-based equalizer is to first train an NN-based equalizer (e.g., offline) using available data for specific scenarios. When deploying the trained equalizer for a new scenario, only certain parameters/layers are retrained. Multi-task training has also been applied to improve the generalization ability of NN equalizers [191–193]. By training a single NN using data collected from different transmission scenarios, the resulting equalizer can generalize well to a wide range of scenarios without the need for retraining [191].

Finally, a review of the principles, performance and complexity for DL based nonlinearity mitigation is provided in [194].

CHAPTER 4

Autoencoders for Designing Communication Systems

AEs can be used to assist the design of communication systems. In this chapter, we first introduce the concept of end-to-end AE learning and its training procedure. This concept is exemplified by using AE learning for GCS. We then introduce AE-based blind channel equalization.

4.1 End-to-End Autoencoder Learning for Physical-Layer Communications

The AE, as described in Section 3.2, resembles a typical communication system. In this analogy, the transmitter, which maps the information bits in a manner that enables reliable transmission through the channel (e.g., by finding a suitable latent representation of the bit sequence), can be viewed as an encoder. Similarly, the corresponding receiver can be seen as a decoder, responsible for mapping the channel observations back to the transmitted data. The channel, which adds distortions to the transmitted signal, can be viewed as the shared latent representation (or the bottleneck) that connects the transmitter and receiver. This concept was initially recognized and investigated in [31], where it was demonstrated that by replacing the traditional transceiver with a pair of NNs, an AE-based communication system can be constructed. In this section, we review the basic idea of end-to-end AE learning-based transceiver design.

Fig. 4.1 visualizes an example of an AE-based communication system. In contrast to conventional communication systems, where the bits-to-waveform (or waveform-to-bits) mapping is carried out by a concatenation of functional blocks, the mapping performed

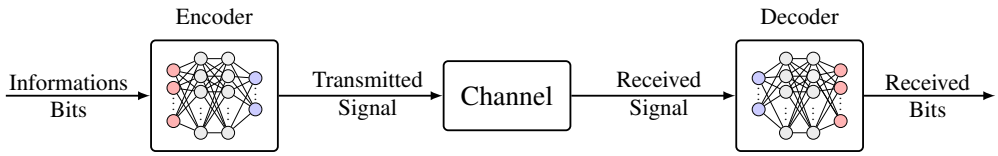

Figure 4.1: An example of an AE-based communication system, where the traditional transceiver is replaced by a pair of NNs.

in an AE-based system is carried out by a single parametric function (e.g., a single NN). To elaborate, given a bit sequence, it is mapped directly to the transmitted waveform by the AE's transmitter NN (i.e., the encoder). Then, the waveform is propagated over the communication channel, after which an impaired version of the transmitted signal is detected by the receiver. The received signal is fed to the corresponding receiver NN (i.e., the decoder), which should then compensate for the transmission impairments and output an estimate of the transmitted bit sequence. The goal of end-to-end learning is to find suitable AE configurations so that the encoder learns a signal representation that is robust to the transmission impairments, while the receiver learns reliable reconstructions of the transmitted information from the channel observations. In the following, we detail end-to-end AE learning using GCS as an example.

4.1.1 Autoencoder-Based Constellation Shaping

We consider performing GCS using a symbol-wise AE,[1] which we illustrate in Fig. 4.2. Without loss of generality, we assume each message is sent over the channel over one complex channel use. Given a message $u \in \mathcal{U} = \{1, \ldots, M\}$, it is mapped to a complex-valued symbol (i.e., a constellation point) according to the following procedures [31]

- "One-hot" encoding: The message u is firstly mapped to an M-dimensional one-hot vector, where the u-th element is set to 1, while all others are set to 0.[2]
- Transmitter NN: The one-hot vector is used as the input to the transmitter NN, which subsequently maps it to a complex-valued symbol $\tilde{x}$. Note that conventionally, the transmitter NN is implemented to have 2 real-valued outputs, representing the real and imaginary components of the transmit symbol. This design choice stems from the fact that early versions of commonly used machine learning frameworks (e.g., TensorFlow or PyTorch) lacked support for complex-valued NNs.
- Normalization: The resulted signal $\tilde{x}$ is then normalized by a normalization layer [31]

[1] Bitwise AEs are introduced in Section 4.1.2.

[2] One-hot encoding is the standard method for representing categorical values in most machine learning algorithms [195]. However, the dimension of the one-hot vector grows exponentially with the number of classes M, which in turn increases the size of the NN. Alternative embeddings [196] and multi-hot sparse categorical CE loss can be used to mitigate this issue.

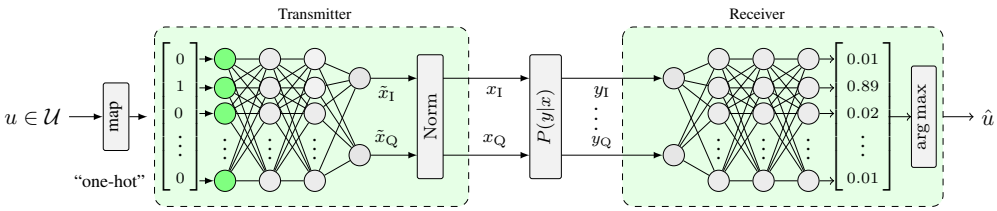

Figure 4.2: An example of a symbol-wise AE for GCS. A complex-valued constellation point x is represented by two real-value numbers, x_I and x_Q.

to fulfill the power constraint (e.g., a peak power constraint or an average power constraint) imposed by the system design. Note that both the one-hot encoding and the normalization layer have no trainable parameters, and the entire transmitter mapping is specified by the set of trainable parameters within the transmitter NN. Consequently, the entire transmitter mapping can be written as $x = f(u; \boldsymbol{\tau})$, where $\boldsymbol{\tau}$ is the trainable parameters of the transmitter NN.

Then, the normalized signal x is sent over the channel, after which a noisy/impaired version of the transmitted signal is observed at the receiver. To determine which message has been transmitted, the receiver NN takes the channel observation y as input and proceeds as follows:

- Receiver NN: The received symbol vector y is firstly processed by the receiver NN to obtain a length M vector, denoted by $\boldsymbol{v} \in \mathbb{R}^M$.
- Softmax activation: The softmax activation function

$$\mathrm{softmax}(\boldsymbol{v})_i = \frac{e^{v_i}}{\sum_{j=1}^{M} e^{v_j}} \tag{4.1}$$

  is applied to $\boldsymbol{v}$ to obtain a M-dimensional probability vector $\boldsymbol{q} \in [0,1]^M$. Here, we have $q_i = \mathrm{softmax}(\boldsymbol{v})_i$, $\sum_{i=1}^{M} q_i = 1$, and each component of $\boldsymbol{q}$ can be interpreted as the estimated posterior probability of the message. The entire receiver mapping is denoted by $\boldsymbol{q} = \boldsymbol{g}(y; \boldsymbol{\rho})$.
- Decision: The transmitted message is estimated according to $\hat{u} = \underset{u \in \mathcal{U}}{\mathrm{argmax}} [\boldsymbol{q}]_u$.

End-to-End Training Procedure

To optimize the transmitter and receiver parameters, it is important to have a suitable optimization criterion. Due to the fact that optimization relies on the empirical computation of gradients, a criterion like SER $\Pr(u \neq \hat{u})$ cannot be used directly (as it is

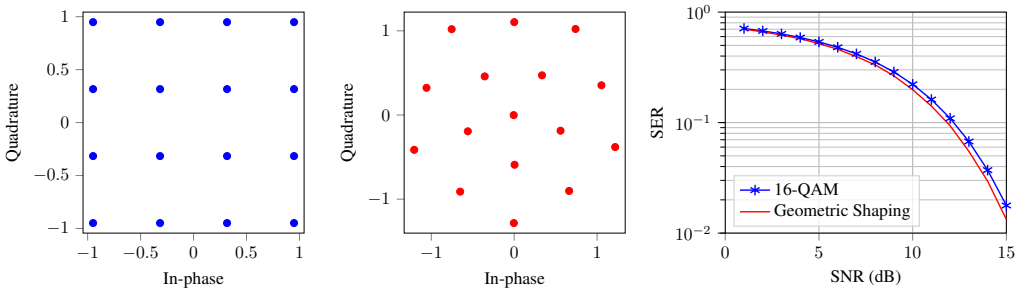

Figure 4.3: Constellation and SER comparison between squared and geometrically shaped 16-QAM. The SER curve assumes transmitting over an AWGN channel.

not differentiable). Instead, a commonly used criterion is the CE loss function given by (3.4). Here, since the target predictions (or labels) are the known transmitted symbols, the CE loss for AE-based GCS can be written as

$$J_{\mathrm{CE}}(\boldsymbol{\tau}, \boldsymbol{\rho}) = -\mathbb{E}_y \left\{ \sum_{u=1}^{M} \log([\boldsymbol{g}(y; \boldsymbol{\rho})]_u) \right\}, \tag{4.2}$$

where it is assumed the symbols are transmitted with equal probability. The dependence of $J_{\mathrm{CE}}(\boldsymbol{\tau}, \boldsymbol{\rho})$ on $\boldsymbol{\tau}$ is implicit through the distribution of y, which is a function of the transmitted symbol $f(u; \boldsymbol{\tau})$. Through iterative minimization of the CE loss across the training dataset, e.g., using SGD, the training algorithm finds the AE configuration that minimizes the error probability during transmission (i.e., by learning the optimal constellation and the corresponding demapper).

Fig. 4.3 illustrates an example of a GCS constellation (for $M = 16$) obtained by training an AE over an AWGN channel at SNR $= 9\,\mathrm{dB}$, where SNR. We consider an averaged power constrain, i.e. $E_s = 1$, and the resulting constellation exhibits a pentagon shape, distinguishing it from the conventional square 16-QAM. Notably, although trained under a fixed SNR, the learned constellation shows the potential to achieve a lower SER compared to the standard QAM constellation across a wide range of SNRs when applied to an AWGN channel. In fact, for an AWGN channel, the "optimal" constellation configuration for a symbol-wise AE has near negligible dependency on the SNRs, which we shall see in Section 4.1.2.

Mutual Information Perspective on Symbol-Wise AEs

The symbol-wise AE trained under CE loss can be used to determine lower bounds on the MI [197]. Indeed, by straightforward manipulations, the MI between random transmitted symbol X (or the message U) and received symbol Y can be rewritten by[3]

[3]For GCS, the mapping from a message to a constellation symbol is fixed, and we have $I(U;Y) = I(X;Y)$.

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{X}} p_X(x) \int f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{p_Y(y)} dy \qquad (4.3)\\
&= \sum_{x\in\mathcal{X}} p_X(x) \int f_{Y|X}(y|x) \log \left(\frac{f_{Y|X}(y|x)\, p_X(x)}{p_Y(y)\, [\boldsymbol{g}(y;\boldsymbol{\rho})]_u} \cdot \frac{[\boldsymbol{g}(y;\boldsymbol{\rho})]_u}{p_X(x)} \right) dy \\
&= \mathrm{KL}(p_{X,Y}(x,y) \| p_Y(y)\, [\boldsymbol{g}(y;\boldsymbol{\rho})]_u) + \sum_{x\in\mathcal{X}} p_X(x) \int f_{Y|X}(y|x) \log \left(\frac{[\boldsymbol{g}(y;\boldsymbol{\rho})]_u}{p_X(x)} \right) dy \\
&\geq \mathbb{E}\{\log([\boldsymbol{g}(y;\boldsymbol{\rho})]_u) - \log[p_X(x)]\} \\
&= H(X) - J_{\mathrm{CE}}(\boldsymbol{\tau},\boldsymbol{\rho}) \,,
\end{aligned}
$$

where the inequality comes from the fact that the Kull-Leibler divergence is non-negative, i.e., $\mathrm{KL}(p_{X,Y}(x,y)\|p(y)\,[\boldsymbol{g}(y;\boldsymbol{\rho})]_u) \geq 0$. The entropy $H(X)$ is a constant assuming the distribution of the transmitted messages is fixed, minimizing the CE loss is therefore equivalent to maximizing a lower bound on the MI.

4.1.2 Bitwise Autoencoder

The symbol-wise AE described in the previous section considers optimizing the transceiver in terms of MI. However, as discussed in Section 2.6, the MI is only achievable for non-binary coded systems (e.g., a system employing multi-level coded modulation [198] or non-binary coded modulation [199]). For the widely deployed bit-interleaved coded modulation systems, training an AE under the CE loss could result in a penalty in the actual AIR after performing bit labeling. Therefore, it is beneficial to train an AE that works directly with the bit sequences. In the rest of this section, we describe the basics of the a bitwise AE and its training.

The bitwise AE was first studied in [200], where it was demonstrated that a bitwise AE can be trained to perform joint GCS and bit labeling. To construct a bitwise AE, the following changes should be made to a symbol-wise AE.

- Instead of having a symbol as input, the input to the bitwise AE is a sequence of m bits.[4] Denoting the set that contains all possible bit sequences by $\mathcal{B}$, it is easy to see that the cardinality of this set is $|\mathcal{B}| = 2^m$.
- Including SNR as an additional parameter for both transmitter and receiver NN. As we will show later, the "optimal" constellation configuration for the bitwise AE varies significantly at different SNRs.[5]

[4] Similar as the symbol-wise AE, the bit sequence can be pre-processed (e.g., one-hot encoded) before fed to the transmitter NN.

[5] One can also include SNR as a hyperparameter when training a symbol-wise AE. However, for an AWGN channel, the constellation obtained using a symbol-wise AE trained at a moderate SNR typically works very well for a wide range of testing SNRs.

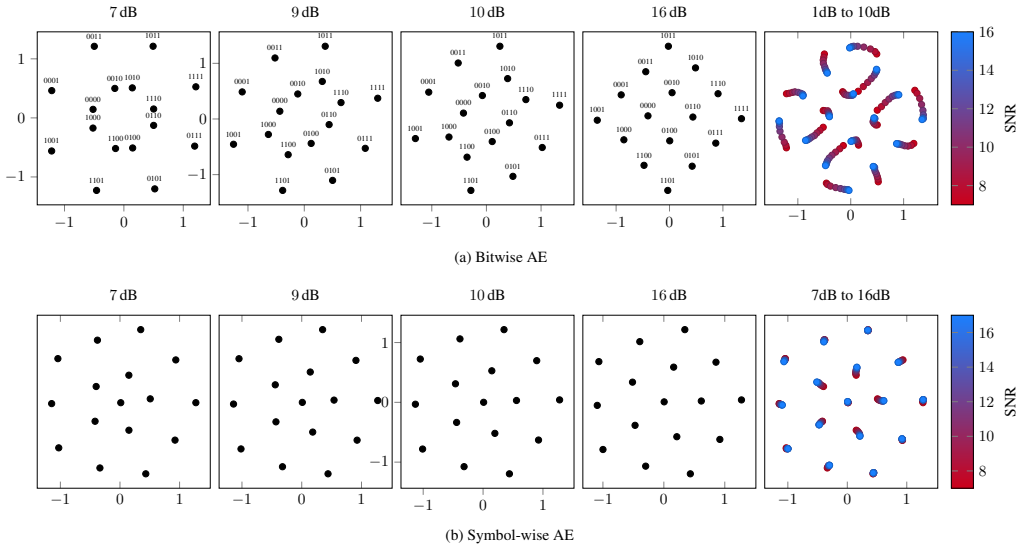

Figure 4.4: Constellations learned over an AWGN channel for $m = 4$ at different SNRs. The figure is based on [200], and the different colors indicate the different training SNR.

- The receiver NN generates m outputs instead of $|\mathcal{B}|$. In particular, given the channel observation y, the receiver NN maps it to a vector of m logits, i.e., $\boldsymbol{l} = [l_1, \ldots, l_m]^\top$. These logits can be related to the LLRs commonly used for soft-decision binary decoder [200].

- The receiver proceeds with applying an element-wise sigmoid function to each of these logits, i.e.,

$$\text{sigmoid}(l_j) = \frac{1}{1 + e^{-l_j}}. \tag{4.4}$$

  The resulted vector represents the posterior probabilities of the bits being "0" or "1".

Optimizing of the bitwise AE is done by minimizing the total BCE loss [200]

$$J_{\text{BCE}} = -\sum_{j=1}^{m} \mathbb{E}_{y,b_j}\{\log[\text{sigmoid}(l_j)]\}, \tag{4.5}$$

which in fact can be used to derive a lower bound for GMI [200], analogous to how CE provides a lower bound for MI.

Fig. 4.4 visualizes the GCS constellations learned over an AWGN channel for $m = 4$, when leveraging the bitwise AE (top) and symbol-wise AE (bottom), respectively. Apart from being bit-labeled, the constellation obtained from training a bitwise AE exhibits

distinct geometric locations compared to that obtained from training a symbol-wise AE across all considered SNRs. In particular, the "optimal" positions of constellation points in the bitwise AE are highly sensitive to the training SNR, whereas such sensitivity is considerably reduced (almost negligible) in the case of a symbol-wise AE.

When looking at the learned constellation together with the corresponding bit labeling, it is also evident that at low SNRs, the bitwise optimized AE tends to cluster constellation points into groups that only differ in one bit position. This clustering phenomenon, absent in a symbol-wise AE, compromises the reliability of this bit position while enhancing the overall constellation's AIR [200]. However, one should note that clustering constellations into groups would lead to a degraded symbol-wise performance, as confusions are more likely to occur within the clustered symbols [200].

4.1.3 End-to-End Autoencoder Training with Non-differentiable Channels

One practical challenge of standard end-to-end AE learning is the requirement for a differentiable channel model to train the transmitter, i.e., transmitter training necessitates backpropagating the gradient through the channel. In practice, such a differentiable channel model is rarely available, and training using an inaccurate channel model leads to a significant performance loss when the system is deployed [201]. To address this challenge, various approaches have been proposed in the literature.

A simple work-around is to first train the transceiver pair on a differentiable channel model and then perform receiver fine-tuning based on measurement data [201]. However, with this approach, the transmitter cannot be fine-tuned to the actual channel, resulting in suboptimal performance. A different approach to circumvent this limitation is to first learn a surrogate channel model, for example, through supervised learning [176] or an adversarial process [31, 186, 202], and then use the surrogate model to train the transceiver. In this case, the performance of the resulting system heavily relies on the accuracy of the learned channel model. Another approach based on a stochastic transmitter was proposed in [203], where RL-based transmitter training was performed by alternating between transmitter and receiver optimization. This approach has been shown to achieve similar performance as standard end-to-end learning both in simulation and experimental environments [203]. A related method based on simultaneous perturbation stochastic optimization was proposed in [204], showing similar performance as standard end-to-end learning (when applied to an AWGN channel) but with slightly degraded convergence rate. Finally, a gradient-free transmitter training approach based on Bayesian filtering was proposed in [205], showing similar performance as standard end-to-end learning.

4.1.4 Applications of End-to-End Autoencoder Learning

Since first reported in [31], the concept of end-to-end AE learning-based communication system design has been extensively studied in both wireless [200, 206–212] and fiber-optic

communications [32, 186, 197, 213–217]. Among these works, we observe that

- Many studies have focused on learning a pair of specific functional blocks. For instance, end-to-end AE learning has gain intensive popularity for designing appropriate modulation formats across various channels and applications. Notable examples include GCS for an AWGN channel [204], GCS for the nonlinear optical fiber channel [197, 216–218], GCS tailored for channels with hardware imperfections [18], and GCS for PN channels utilizing a differentiable BPS [19, 219], among others.

- Some works have explored learning multiple functional blocks jointly. To name a few examples, end-to-end AE learning has been applied for joint source and channel coding in [210, 212]. It has also been employed to perform joint learning of GCS and channel coding [31, 220], joint geometric and probabilistic shaping in [221], and the joint optimization of GCS and transmit waveform [211].

- Finally, several works have focused on optimizing the entire communication link, encompassing the learning of bit-to-waveform mapping and waveform-to-bit mapping. Notable examples include orthogonal frequency-division multiplexing AEs for wireless communication [207], AEs for intensity-modulation direct-detection fiber-optic communication [32,186,214], and AEs for coherent fiber-optic communications in [215].

End-to-end AE learning serves as the primary focus of this thesis. Specifically, this concept has been extensively explored in Paper A, Paper C, and Paper D across various channels and applications. Paper A and Paper C delve into fiber-optic communication. In particular, Paper A studies AE-based GCS for an NLPN channel, while Paper C explores the joint learning of the transceiver chains for a WDM system tailored to nonlinear hardware impairments. Paper D considers wireless communication, where AE-based MIMO and MU communication were investigated.

4.2 Autoencoders for Blind Channel Equalization

AEs can also be used to train blind channel equalizers, although this approach has seldom been explored in the literature. The idea of an AE-based, or more precisely, VAE-based blind equalizer was initially introduced in [222, 223]. The primary objective is to learn a parametric function that approximates the MAP detector. To achieve this, the method involves training two parametric functions, often implemented using FIRs filters [224]. First, an encoding function maps the channel observations to soft estimates of the transmitted symbol sequence. Second, a decoding function aims to reconstruct the channel observations from these estimated symbols. In this framework, the encoding function serves dual purposes, i.e., it acts as both an equalizer and a soft

demapper. By concurrently training these two parametric functions using the evidence lower bound [225], one can derive an approximation to the MAP detector and obtain an ML channel estimate [224].

The VAE-based equalizer has demonstrated comparable performance to a pilot-aided equalizer, surpassing the commonly used blind equalizers [224]. However, it is primarily designed for linear channels, and it may experience performance degradation when directly applied to nonlinear channels.[6] To address this limitation, a novel blind equalizer based on the LSC-AEs was proposed in Paper E. The proposed method is based on introducing a constraint to the latent representation of a standard AE, and is shown to achieve promising performance for both linear and nonlinear channels.

[6] Training the VAE-based equalizer necessitates an analytical solution to the evidence lower bound (see [224]), which is only available for channels of a simple form.

CHAPTER 5

Contributions

This chapter summarizes the contributions of each appended publication, and outlines some potentially interesting ideas for future work.

5.1 Contributions

Paper A

"Learning physical-layer communication with quantized feedback"

In this paper, motivated by the challenge that gradient-based transmitter optimization faces in practice due to the requirement of a known and differentiable channel model, we address the problem of transmitter gradient estimation using RL, assuming that the feedback signal is quantized with a limited number of bits. Our primary contribution is the proposition of a novel scheme for the feedback signal quantization. The effectiveness of the proposed quantization schemes is validated through a comprehensive numerical study, demonstrating that feedback quantization does not significantly affect the learning process. It can lead to performance similar to the case where unquantized feedback is used for training, even with 1-bit quantization. Additionally, we provide a theoretical justification for the effectiveness of the proposed approach. Specifically, we prove that feedback quantization and bit flips simply scale the expected gradient used for parameter training. Furthermore, we derive upper bounds on the variance of the gradient in terms of the Fisher information matrix of the transmitter parameters.

Paper B

"Over-the-fiber digital predistortion using reinforcement learning"

In this paper, we propose a novel NN-based DPD scheme for compensating the transmitter impairments in a practical optical-fiber transmission system. Unlike many existing NN-based DPD schemes, which are trained offline, we demonstrate, for the first time, the online training of DPD over an optical back-to-back channel using RL. Experimental results indicate that the proposed DPD effectively mitigates transmitter impairments, outperforming a widely used baseline scheme.

Paper C

"Model-based end-to-end learning for WDM systems with transceiver hardware impairments"

In this paper, we propose a novel end-to-end AE learning-based transceiver design for WDM systems with hardware impairments. Motivated by the fact that existing AE-based systems are often regarded as "black-box" solutions and are difficult to interpret, we propose to design the AE-based transceiver following the modular structure of traditional fiber-optic communication systems. Specifically, the proposed AE-based transceiver is implemented using a concatenation of small NNs, with the aim that each NN is trained to perform a specific functionality. By doing this, it is shown that the proposed method can achieve superior performance compared to traditional methods. Moreover, the modular structured AE design allows us to visualize and interpret each of the learned block, offering potential insights to guide the improvement of existing communication systems. Finally, we extend the RL-based transmitter training approach to handle systems with memory. Simulation results demonstrate that the RL-based training algorithm achieves similar performance to standard end-to-end learning.

Paper D

"Benchmarking and interpreting end-to-end learning of MIMO and multi-user communication"

In this paper, we study end-to-end AE learning for MIMO and multi-user communications. Our objective is to better understand the potential performance advantages offered by AE-based MIMO and MU systems over traditional methods. Our research shows that, for a wide variety of different scenarios, AE-based communication systems can achieve commendable solutions without a priori knowledge about complex mathematical tools or communication-theoretic principles. However, our work has also highlighted

that these systems do not necessarily outperform state-of-the-art benchmarks, especially when those benchmarks are appropriately selected. A particular emphasis in this study was placed on benchmark selection. This approach allows us to provide deeper insights into AE-based systems and, in certain instances, complete interpretations of the learned communication schemes.

Paper E

"Blind Channel Equalization Using Latent Space Constrained Autoencoders"

In this work, motivated by the fact that data-aided equalizers lead to a loss in SE, we propose a novel blind equalizer based on the LSC-AEs. By introducing a decision block to a standard AEs, the latent representation of the AE is constrained to a fixed codebook (i.e., the modulation format under use), which in return benefits the AE-based equalizer training. The proposed equalizers can be realized using different models, such as NNs or FIR filters, and can be applied to both linear and nonlinear channels. Simulation results shown that the proposed blind equalizer can achieve performance similar to that of a data-aided equalizer while outperforming state-of-the-art blind equalizers. Finally, we demonstrate that the proposed method we demonstrate that the proposed scheme exhibits superior training characteristics compared to the baseline schemes in terms of both convergence speed and robustness to variations in training batch size and learning rate.

5.2 Conclusion

In this thesis, we have studied various aspects of DL for physical-layer communications. Our primary focus has been on the applications of end-to-end AE learning for joint transceiver design, while we have also explored the potential of applying AEs for blind channel equalization. For end-to-end AE learning, we have considered both wireless and fiber-optic communications in terms of applications and supervised and RL in terms of methodology. In general, we found that for a wide variety of different scenarios, end-to-end AE learning has the potential to learn very good solutions without any a priori knowledge about complex mathematical tools or communication-theoretic principles. For linear systems, AE-based solutions do not necessarily outperform state-of-the-art baseline schemes, provided that the baseline schemes are properly chosen. However, for nonlinear systems, end-to-end AE learning holds the potential to discover solutions that surpass traditional model-based methods. The primary reason for this lies in the fact that traditional methods are often designed based on some idealized assumptions (e.g., linearity). When such assumptions do not hold, the performance of these model-based methods degrades. In contrast, learning-based methods do not necessitate such

assumptions and, therefore, have the potential to learn more effective solutions. Regarding the application of blind channel equalization, we introduced a novel AE-based blind equalizer. The proposed method relies on introducing a constraint in the latent space representation of a standard AE. It has been demonstrated that our approach can achieve performance comparable to traditional data-aided equalizers while outperforming state-of-the-art blind equalizers.

5.3 Future Works

While the thesis has explored various aspects of applying AEs in the physical-layer communications, we believe that there are several important aspects related to AE usage that deserve further investigation

- Channel models: Similar to related prior works, we have adopted simplified channel models for end-to-end AE learning. While transmission/transceiver impairments have been considered in some of the appended papers, these channels generally fall short to describe all the different aspects of a practical transmission link. Consequently, it remains unclear how end-to-end AE learning performs in a more practical communication setup where both channel and hardware distortion are present. To more accurately resemble reality, one could consider alternative models to evaluate AE-based transceiver design.

- The results presented in the papers above rely on the assumption that the communication link has a fixed data rate, as does the AE. To address the need for flexible transmission rates, new AE designs are required to offer rate-adaptive transmission. A promising direction could involve using the "many-to-on" mapping [217].

- Complexity: Both training and runtime complexities are crucial considerations for the practical implementation of DL-based algorithms. While the works in this thesis have primarily focused on proof-of-concept demonstrations, little attention has been paid to complexity analysis. For future works, it might be worthwhile, for instance, to optimize NN architectures to reduce training complexity or to prune NN parameters to minimize runtime complexity.

- Scalability: While several works in the appended papers have explored AE-based GCS for two-dimensional constellations with relatively small cardinalities, GCS is expected to offer better performance for multi-dimensional constellations with large cardinalities. Nonetheless, the efficient training of AEs for significantly large constellations remains a topic requiring further exploration.

Bibliography

[1] N. S. Cisco, "Cisco annual internet report (2018-2023)," *White Paper*, 2023.

[2] G. P. Agrawal, *Fiber-optic communication systems*. John Wiley & Sons, 2012.

[3] A. P. T. Lau and F. N. Khan, *Machine learning for future fiber-optic communication systems*. Academic Press, 2022.

[4] D. Wang and M. Zhang, "Artificial intelligence in optical communications: from machine learning to deep learning," *Front. Comms. Net.*, vol. 2, p. 656786, 2021.

[5] M. S. Faruk and K. Kikuchi, "Adaptive frequency-domain equalization in digital coherent optical receivers," *Opt. Exp.*, vol. 19, no. 13, pp. 12 789–12 798, 2011.

[6] P. J. Winzer and R.-J. Essiambre, "Advanced optical modulation formats," in *Opt. Fiber Telecommuni. VB*. Elsevier, 2008, pp. 23–93.

[7] A. Amari *et al.*, "A survey on fiber nonlinearity compensation for 400 Gb/s and beyond optical communication systems," *IEEE Commun. Surv. Tut.*, vol. 19, no. 4, pp. 3097–3113, 2017.

[8] A. Napoli *et al.*, "Reduced complexity digital back-propagation methods for optical communication systems," *IEEE/OSA J. Lightw. Techn.*, vol. 32, no. 7, pp. 1351–1362, 2014.

[9] C. Häger and H. D. Pfister, "Physics-based deep learning for fiber-optic communication systems," *IEEE J. Sel. Areas Commun.*, vol. 39, pp. 280–294, 2020.

[10] C. Antonelli, O. Golani, M. Shtaif, and A. Mecozzi, "Nonlinear interference noise in space-division multiplexed transmission through optical fibers," *Opt. Exp.*, vol. 25, no. 12, pp. 13 055–13 078, 2017.

[11] A. Napoli *et al.*, "Digital pre-compensation techniques enabling high-capacity bandwidth variable transponders," *Opt. Commun.*, vol. 409, pp. 52–65, 2018.

[12] V. Bajaj, F. Buchali, M. Chagnon, S. Wahls, and V. Aref, "Deep neural network-based digital pre-distortion for high baudrate optical coherent transmission," *IEEE/OSA J. Lightw. Techn.*, vol. 40, no. 3, pp. 597–606, 2022.

[13] G. Paryanti, H. Faig, L. Rokach, and D. Sadot, "A direct learning approach for neural network based pre-distortion for coherent nonlinear optical transmitter," *IEEE/OSA J. Lightw. Techn.*, vol. 38, no. 15, pp. 3883–3896, 2020.

[14] F. Musumeci *et al.*, "An overview on application of machine learning techniques in optical networks," *IEEE Commun. Surv. Tut.*, vol. 21, no. 2, pp. 1383–1408, 2019.

[15] V. Aggarwal *et al.*, "A review: Deep learning technique for image classification," *Trans. image Process. Comput. Vis.*, vol. 4, no. 11, p. 21, 2018.

[16] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604–624, 2020.

[17] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, 2020.

[18] O. Jovanovic, F. Da Ros, D. Zibar, and M. P. Yankov, "Geometric constellation shaping for fiber-optic channels via end-to-end learning," *IEEE/OSA J. Lightw. Techn.*, 2023.

[19] A. Rode, B. Geiger, S. Chimmalgi, and L. Schmalen, "End-to-end optimization of constellation shaping for Wiener phase noise channels with a differentiable blind phase search," *J. Light. Techn.*, 2023.

[20] D. Rafique, A. Napoli, S. Calabro, and B. Spinnler, "Digital preemphasis in optical communication systems: On the DAC requirements for terabit transmission applications," *IEEE/OSA J. Lightw. Techn.*, vol. 32, no. 19, pp. 3247–3256, 2014.

[21] A. Napoli, M. M. Mezghanni, S. Calabro, R. Palmer, G. Saathoff, and B. Spinnler, "Digital predistortion techniques for finite extinction ratio IQ mach–zehnder modulators," *IEEE/OSA J. Lightw. Techn.*, vol. 35, no. 19, pp. 4289–4296, 2017.

[22] G. P. Agrawal, "Nonlinear fiber optics," in *Nonlinear Science at the Dawn of the 21st Century*. Springer, 2000, pp. 195–211.

[23] S. Deligiannidis, A. Bogris, C. Mesaritakis, and Y. Kopsinis, "Compensation of fiber nonlinearities in digital coherent systems leveraging long short-term memory neural networks," *IEEE/OSA J. Lightw. Techn.*, vol. 38, no. 21, pp. 5991–5999, 2020.

[24] X. Liu *et al.*, "Nonlinearity mitigation in a fiber-wireless integrated system based on low-complexity autoencoder and BiLSTM-ANN equalizer," *Opt. Exp.*, vol. 31, no. 12, pp. 20 005–20 018, 2023.

[25] S. Zhang *et al.*, "Field and lab experimental demonstration of nonlinear impairment compensation using neural networks," *Nature Commun.*, vol. 10, no. 1, p. 3033, 2019.

[26] J. Hagenauer, E. Offer, and L. Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 429–445, 1996.

[27] J. Kuck *et al.*, "Belief propagation neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 667–678, 2020.

[28] I. Sason, S. Shamai *et al.*, "Performance analysis of linear codes under maximum-likelihood decoding: A tutorial," *Found. Trends Commun. Inf. Theory*, vol. 3, no. 1–2, pp. 1–222, 2006.

[29] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, pp. 127–138, 2016.

[30] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[31] T. J. O'Shea, T. Roy, and N. West, "Approximating the void: Learning stochastic channel models from observation with variational generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Netw. Commun. (ICNC)*, 2019, pp. 681–686.

[32] B. Karanov *et al.*, "End-to-end deep learning of optical fiber communications," *IEEE/OSA J. Lightw. Techn.*, vol. 36, no. 20, pp. 4843–4855, 2018.

[33] B. Sklar, *Digital communications: fundamentals and applications*. Pearson, 2021.

[34] E. Zehavi, "8-PSK trellis codes for a rayleigh channel," *IEEE Trans. Commun.*, vol. 40, no. 5, pp. 873–884, 1992.

[35] E. N. Gilbert, "A comparison of signalling alphabets," *Bell Sys. Techn. J.*, vol. 31, no. 3, pp. 504–522, 1952.

[36] C. Campopiano and B. Glazer, "A coherent digital amplitude and phase modulation scheme," *IRE Trans. Commun. Systems*, vol. 10, pp. 90–95, 1962.

[37] G. Forney and L.-F. Wei, "Multidimensional constellations. I. introduction, figures of merit, and generalized cross constellations," *IEEE J. Sel. Areas Commun.*, vol. 7, no. 6, pp. 877–892, 1989.

[38] B. Chen, C. Okonkwo, H. Hafermann, and A. Alvarado, "Increasing achievable information rates via geometric shaping," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2018.

[39] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4651–4665, 2015.

[40] J. Cho and P. J. Winzer, "Probabilistic constellation shaping for optical fiber communications," *IEEE/OSA J. Lightw. Techn.*, vol. 37, no. 6, pp. 1590–1607, 2019.

[41] L. Schmalen, "Probabilistic constellation shaping: Challenges and opportunities for forward error correction," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2018.

[42] J. Cho, "Probabilistic constellation shaping: An implementation perspective," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2022.

[43] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[44] K. Gentile, "Digital pulse-shaping filter basics," *Appl. Note AN-922. Analog Devices, Inc.*, 2007.

[45] P. J. Winzer, "High-spectral-efficiency optical modulation formats," *IEEE/OSA J. Lightw. Techn.*, vol. 30, no. 24, pp. 3824–3835, 2012.

[46] Z. Dong *et al.*, "Ultra-dense WDM-PON delivering carrier-centralized Nyquist-WDM uplink with digital coherent detection," *Opt. Exp.*, vol. 19, no. 12, pp. 11 100–11 105, 2011.

[47] R. Schmogrow *et al.*, "Pulse shaping with digital, electrical, and optical filters: A comparison," *IEEE/OSA J. Lightw. Techn.*, vol. 31, no. 15, pp. 2570–2577, 2013.

[48] A. M. Weiner, "Ultrafast optical pulse shaping: A tutorial review," *Opt. Commun.*, vol. 284, no. 15, pp. 3669–3692, 2011.

[49] X. Liu, S. Chandrasekhar, and P. J. Winzer, "Digital signal processing techniques enabling multi-Tb/s superchannel transmission: An overview of recent advances in DSP-enabled superchannels," *IEEE Signal Proc. Mag.*, vol. 31, no. 2, pp. 16–24, 2014.

[50] Y. Fujiya and N. Izuka, "PAPR reduction of transmitted signal using modified root roll-off filter," in *Proc. IEEE Asia Pacific Conf. Wirel. Mobile*, 2021, pp. 105–108.

[51] P. W. Berenguer *et al.*, "Nonlinear digital pre-distortion of transmitter components," *IEEE/OSA J. Lightw. Techn.*, vol. 34, no. 8, pp. 1739–1745, 2015.

[52] H. Faig, Y. Yoffe, E. Wohlgemuth, and D. Sadot, "Grading optimization for dimensions-reduced orthogonal Volterra DPD," *IEEE Photon. J.*, vol. 12, pp. 1–10, 2019.

[53] D. Rafique, T. Rahman, A. Napoli, and B. Spinnler, "Digital pre-emphasis in optical communication systems: On the nonlinear performance," *IEEE/OSA J. Lightw. Techn.*, vol. 33, pp. 140–150, 2015.

[54] Z. He, K. Vijayan, M. Mazur, M. Karlsson, and J. Schröder, "Look-up table based pre-distortion for transmitters employing high-spectral-efficiency modulation formats," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2020.

[55] M. U. Hadi and G. Murtaza, "Enhancing distributed feedback-standard single mode fiber-radio over fiber links performance by neural network digital predistortion," *Microwave and Optical Technology Letters*, vol. 63, no. 5, pp. 1558–1565, 2021.

[56] V. Bajaj, F. Buchali, M. Chagnon, S. Wahls, and V. Aref, "54.5 Tb/s WDM transmission over field deployed fiber enabled by neural network-based digital predistortion," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2021, pp. M5F–2.

[57] M. U. Hadi, M. Awais, M. Raza, K. Khurshid, and H. Jung, "Neural network DPD for aggrandizing SM-VCSEL-SSMF-based radio over fiber link performance," in *Proc. Photon.*, vol. 8, no. 1. MDPI, 2021, p. 19.

[58] C. Laperle and M. O'Sullivan, "Advances in high-speed DACs, ADCs, and DSP for optical coherent transceivers," *IEEE/OSA J. Lightw. Techn.*, vol. 32, no. 4, pp. 629–643, 2014.

[59] S. Varughese, D. Lippiatt, S. Tibuleac, and S. E. Ralph, "Frequency dependent ENOB requirements for 400G/600G/800G optical links," *IEEE/OSA J. Lightw. Techn.*, vol. 38, no. 18, pp. 5008–5016, 2020.

[60] A. Napoli *et al.*, "Digital compensation of bandwidth limitations for high-speed DACs and ADCs," *IEEE/OSA J. Lightw. Techn.*, vol. 34, no. 13, pp. 3053–3064, 2016.

[61] R. Gray and D. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[62] S. Kumar and M. J. Deen, *Fiber optic communications: fundamentals and applications.* John Wiley & Sons, 2014.

[63] A. Goldsmith, *Wireless communications.* Cambridge University Press, 2005.

[64] R. Maher and B. Thomsen, "Dynamic linewidth measurement technique using digital intradyne coherent receivers," *Opt. Exp.*, vol. 19, no. 26, pp. B313–B322, 2011.

[65] S. J. Savory, "Digital coherent optical receivers: Algorithms and subsystems," *IEEE J. Sel. Topics Quantum Electron.*, vol. 16, no. 5, pp. 1164–1179, 2010.

[66] ——, "Digital filters for coherent optical receivers," *Opt. Exp.*, vol. 16, no. 2, pp. 804–817, 2008.

[67] I. Fatadin, S. J. Savory, and D. Ives, "Compensation of quadrature imbalance in an optical QPSK coherent receiver," *IEEE Photon. Techn. Lett.*, vol. 20, no. 20, pp. 1733–1735, 2008.

[68] S. H. Chang, H. S. Chung, and K. Kim, "Impact of quadrature imbalance in optical coherent QPSK receiver," *IEEE Photon. Techn. Lett.*, vol. 21, no. 11, pp. 709–711, 2009.

[69] F. Gardner, "A BPSK/QPSK timing-error detector for sampled receivers," *IEEE Trans. Commun.*, vol. 34, no. 5, pp. 423–429, 1986.

[70] K. Mueller and M. Muller, "Timing recovery in digital synchronous data receivers," *IEEE Trans. Commun.*, vol. 24, no. 5, pp. 516–531, 1976.

[71] P. Bayvel, C. Behrens, and D. Millar, "Digital signal processing and its applications in optical communications systems," *Opt. Fiber Telecommun.*, pp. 163–219, 2013.

[72] A. Leven, N. Kaneda, U.-V. Koc, and Y.-K. Chen, "Frequency estimation in intradyne reception," *IEEE Photon. Techn. Lett.*, vol. 19, no. 6, pp. 366–368, 2007.

[73] S. Hoffmann *et al.*, "Frequency and phase estimation for coherent QPSK transmission with unlocked dfb lasers," *IEEE Photon. Techn.y Lett.*, vol. 20, no. 18, pp. 1569–1571, 2008.

[74] M. Morelli and U. Mengali, "Feedforward frequency estimation for PSK: A tutorial review," *Eur. Trans. Telecommun.*, vol. 9, no. 2, pp. 103–116, 1998.

[75] M. Selmi, Y. Jaouen, and P. Ciblat, "Accurate digital frequency offset estimator for coherent PolMux QAM transmission systems," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2009.

[76] K. Piyawanno, M. Kuschnerov, B. Spinnler, and B. Lankl, "Fast and accurate automatic frequency control for coherent receivers," in *Proc. Eur. Conf. Opt. Commun (ECOC)*, vol. 7, 2009.

[77] H. Meyr, M. Moeneclaey, and S. A. Fechtel, *Digital communication receivers: Synchronization, channel estimation and signal processing.* Wiley, 1998.

[78] X. Zhou, X. Chen, and K. Long, "Wide-range frequency offset estimation algorithm for optical coherent systems using training sequence," *IEEE Photon. Techn. Lett.*, vol. 24, pp. 82–84, 2011.

[79] R. Koma *et al.*, "Novel data-aided carrier frequency offset compensation methods using asymmetric-shape constellations for burst-mode coherent reception," *IEEE/OSA J. Lightw. Techn.*, vol. 41, pp. 159–168, 2023.

[80] D. Zhao *et al.*, "Digital pilot aided carrier frequency offset estimation for coherent optical transmission systems," *Opt. Exp.*, vol. 23, no. 19, pp. 24 822–24 832, 2015.

[81] S. Cao, C. Yu, and P.-Y. Kam, "A performance investigation of correlation-based and pilot-tone-assisted frequency offset compensation method for CO-OFDM," *Opt. Exp.*, vol. 21, no. 19, pp. 22 847–22 853, 2013.

[82] G. P. Agrawal, *Lightwave Technology: Telecommunication Systems.* John Wiley & Sons, 2005.

[83] D. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. Commun.*, vol. 28, no. 11, pp. 1867–1875, 1980.

[84] T. Xu, G. Jacobsen, J. Li, M. Leeson, and S. Popov, "Dynamic physical layer equalization in optical communication networks," *Preprint, available online at https://arxiv.org/abs/:1606.04011*, 2016.

[85] K. N. Oh and Y. O. Chin, "Modified constant modulus algorithm: blind equalization and carrier phase recovery algorithm," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 1, 1995, pp. 498–502.

[86] J. Yang, J.-J. Werner, and G. A. Dumont, "The multimodulus blind equalization and its generalized algorithms," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 5, pp. 997–1015, 2002.

[87] G. Picchi and G. Prati, "Blind equalization and carrier recovery using a "stop-and-go" decision-directed algorithm," *IEEE Trans. Commun.*, vol. 35, no. 9, pp. 877–887, 1987.

[88] A. Pandey, L. Malviya, and V. Sharma, "Comparative study of LMS and NLMS algorithms in adaptive equalizer," *Int. J. Eng. Res. Appl.*, vol. 2, no. 3, pp. 1584–1587, 2012.

[89] K.-P. Ho and J. M. Kahn, "Delay-spread distribution for multimode fiber with strong mode coupling," *IEEE Photon. Techn. Lett.*, vol. 24, no. 21, pp. 1906–1909, 2012.

[90] E. Ip and J. M. Kahn, "Feedforward carrier recovery for coherent optical communications," *IEEE/OSA J. Lightw. Techn.*, vol. 25, no. 9, pp. 2675–2692, 2007.

[91] S. Ö. Arık, D. Askarov, and J. M. Kahn, "Adaptive frequency-domain equalization in mode-division multiplexing systems," *IEEE/OSA J. Lightw. Techn.*, vol. 32, no. 10, pp. 1841–1852, 2014.

[92] M. Visintin, G. Bosco, P. Poggiolini, and F. Forghieri, "Adaptive digital equalization in optical coherent receivers with stokes-space update algorithm," *IEEE/OSA J. Lightw. Techn.*, vol. 32, no. 24, pp. 4759–4767, 2014.

[93] F. P. Guiomar, J. D. Reis, A. L. Teixeira, and A. N. Pinto, "Mitigation of intra-channel nonlinearities using a frequency-domain Volterra series equalizer," *Opt. Exp.*, vol. 20, no. 2, pp. 1360–1369, 2012.

[94] F. P. Guiomar and A. N. Pinto, "Simplified Volterra series nonlinear equalizer for polarization-multiplexed coherent optical systems," *IEEE/OSA J. Lightw. Techn.*, vol. 31, no. 23, pp. 3879–3891, 2013.

[95] T. Ogunfunmi and T. Drullinger, "Equalization of non-linear channels using a Volterra-based non-linear adaptive filter," in *Proc. IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS)*, 2011.

[96] J. Gonçalves *et al.*, "Nonlinear compensation with DBP aided by a memory polynomial," *Opt. Exp.*, vol. 24, no. 26, pp. 30 309–30 316, 2016.

[97] K. Burse, R. N. Yadav, and S. Shrivastava, "Channel equalization using neural networks: A review," *IEEE Trans. Syst. Man Cybern. Part C*, vol. 40, no. 3, pp. 352–357, 2010.

[98] T. Kamiyama, H. Kobayashi, and K. Iwashita, "Neural network nonlinear equalizer in long-distance coherent optical transmission systems," *IEEE Photon.s Techn. Lett.*, vol. 33, no. 9, pp. 421–424, 2021.

[99] P. J. Freire *et al.*, "Neural networks-based equalizers for coherent optical transmission: Caveats and pitfalls," *IEEE J. Sel. Topics Quantum Electron.*, vol. 28, no. 4, pp. 1–23, 2022.

[100] A. J. Viterbi and A. M. Viterbi, "Nonlinear estimation of PSK-modulated carrier phase with application to burst digital transmission," *IEEE Trans. on Inf. theory*, vol. 29, no. 4, pp. 543–551, 1983.

[101] I. Fatadin, D. Ives, and S. J. Savory, "Laser linewidth tolerance for 16-QAM coherent optical systems using QPSK partitioning," *IEEE Photon. Techn. Lett.*, vol. 22, no. 9, pp. 631–633, 2010.

[102] H. Louchet, K. Kuzmin, and A. Richter, "Improved DSP algorithms for coherent 16-qam transmission," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2008.

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

[104] X. Zhou, C. Lu, A. P. T. Lau, and K. Long, "Low-complexity carrier phase recovery for square M-QAM based on S-BPS algorithm," *IEEE Photon. Techn. Lett.*, vol. 26, no. 18, pp. 1863–1866, 2014.

[105] S. Chimmalgi, A. Rode, L. Schmid, and L. Schmalen, "Approximate maximum a posteriori carrier phase estimator for Wiener phase noise channels using belief propagation," *Preprint, available online at https://arxiv.org/abs/2307.03517*, 2023.

[106] J. Zhao and L.-K. Chen, "Carrier phase recovery based on KL divergence in probabilistically shaped coherent systems," *IEEE/OSA J. Lightw. Techn.*, vol. 39, no. 9, pp. 2684–2695, 2021.

[107] J. C. M. Diniz *et al.*, "Low-complexity carrier phase recovery based on principal component analysis for square-QAM modulation formats," *Opt. Exp.*, vol. 27, no. 11, pp. 15 617–15 626, 2019.

[108] K. P. Zhong, J. H. Ke, Y. Gao, and J. C. Cartledge, "Linewidth-tolerant and low-complexity two-stage carrier phase estimation based on modified QPSK partitioning for dual-polarization 16-QAM systems," *IEEE/OSA J. Lightw. Techn.*, vol. 31, no. 1, pp. 50–57, 2013.

[109] Y. Gao, A. P. T. Lau, S. Yan, and C. Lu, "Low-complexity and phase noise tolerant carrier phase estimation for dual-polarization 16-QAM systems," *Opt. Exp.*, vol. 19, no. 22, pp. 21 717–21 729, 2011.

[110] K. P. Zhong, J. H. Ke, Y. Gao, J. C. Cartledge, A. P. T. Lau, and C. Lu, "Carrier phase estimation for DP-16QAM using QPSK partitioning and quasi-multiplier-free algorithms," in *Proc. Opt. Fiber Conf. Conf. (OFC)*, 2014.

[111] M. P. Yankov, T. Fehenberger, L. Barletta, and N. Hanik, "Low-complexity tracking of laser and nonlinear phase noise in WDM optical fiber systems," *IEEE/OSA J. Lightw. Techn.*, vol. 33, no. 23, pp. 4975–4984, 2015.

[112] A. Jain and K. P. Kumar, "Tracking linear and nonlinear phase noise in 100G QPSK modulated systems using kalman filter," in *Proc. Opt. Sensors*, 2015, pp. JM3A–8.

[113] A. Jain and P. K. Krishnamurthy, "Phase noise tracking and compensation in coherent optical systems using kalman filter," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1072–1075, 2016.

[114] M. Mazur, J. Schröder, A. Lorences-Riesgo, T. Yoshida, M. Karlsson, and P. A. Andrekson, "Overhead-optimization of pilot-based digital signal processing for flexible high spectral efficiency transmission," *Opt. Exp.*, vol. 27, no. 17, pp. 24 654–24 669, 2019.

[115] R. Krishnan, M. R. Khanzadi, T. Eriksson, and T. Svensson, "Soft metrics and their performance analysis for optimal data detection in the presence of strong oscillator phase noise," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2385–2395, 2013.

[116] U. Madhow, *Fundamentals of digital communication.* Cambridge University Press, 2008.

[117] P. K. A. Wai and C. Menyak, "Polarization mode dispersion, decorrelation, and diffusion in optical fibers with randomly varying birefringence," *IEEE/OSA J. Lightw. Techn.*, vol. 14, no. 2, pp. 148–157, 1996.

[118] L. Beygi, E. Agrell, M. Karlsson, and P. Johannisson, "Signal statistics in fiber-optical channels with polarization multiplexing and self-phase modulation," *IEEE/OSA J. Lightw. Techn.*, vol. 29, no. 16, pp. 2379–2386, 2011.

[119] K.-P. Ho, *Phase-modulated optical communication systems.* Springer Science & Business Media, 2005.

[120] R.-J. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, and B. Goebel, "Capacity limits of optical fiber networks," *IEEE/OSA J. Lightw. Techn.*, vol. 28, no. 4, pp. 662–701, 2010.

[121] A. Mecozzi, "Limits to long-haul coherent transmission set by the kerr nonlinearity and noise of the in-line amplifiers," *IEEE/OSA J. Lightw. Techn.*, vol. 12, no. 11, pp. 1993–2000, 1994.

[122] P. Poggiolini, "The GN model of non-linear propagation in uncompensated coherent optical systems," *IEEE/OSA J. Lightw. Techn.*, vol. 30, no. 24, pp. 3857–3879, 2012.

[123] R. Dar, M. Feder, A. Mecozzi, and M. Shtaif, "Accumulation of nonlinear interference noise in fiber-optic systems," *Opt. Exp.*, vol. 22, no. 12, pp. 14 199–14 211, 2014.

[124] E. Agrell, G. Durisi, and P. Johannisson, "Information-theory-friendly models for fiber-optic channels: A primer," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2015, pp. 1–5.

[125] K. Shibahara *et al.*, "Dense SDM (12-core x 3-mode) transmission over 527 km with 33.2-ns mode-dispersion employing low-complexity parallel MIMO frequency-domain equalization," *IEEE/OSA J. Lightw. Techn.*, vol. 34, pp. 196–204, 2015.

[126] A. Abouseif, G. R.-B. Othman, and Y. Jaouën, "Channel model and optimal core scrambling for multi-core fiber transmission system," *Opt. Commun.*, vol. 454, p. 124396, 2020.

[127] L. Gan *et al.*, "Investigation of channel model for weakly coupled multicore fiber," *Opt. Exp.*, vol. 26, no. 5, pp. 5182–5199, 2018.

[128] L. Gan, J. Zhou, S. Fu, M. Tang, and D. Liu, "Efficient channel model for homogeneous weakly coupled multicore fibers," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 4, pp. 1–11, 2019.

[129] H. R. Stuart, "Dispersive multiplexing in multimode optical fiber," *Science*, vol. 289, no. 5477, pp. 281–283, 2000.

[130] J. M. Kahn, K.-P. Ho, and M. B. Shemirani, "Mode coupling effects in multi-mode fibers," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2012, pp. OW3D–3.

[131] A. Karadimitrakis, A. L. Moustakas, H. Hafermann, and A. Mueller, "Optical fiber MIMO channel model and its analysis," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, 2016, pp. 2164–2168.

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

[133] D. J. Richardson, J. M. Fini, and L. E. Nelson, "Space-division multiplexing in optical fibres," *Nature Photon.*, vol. 7, no. 5, pp. 354–362, 2013.

[134] G. M. Saridis, D. Alexandropoulos, G. Zervas, and D. Simeonidou, "Survey and evaluation of space division multiplexing: From technologies to optical networks," *IEEE Commun. Surv. Tut.*, vol. 17, no. 4, pp. 2136–2156, 2015.

[135] A. V. Oppenheim, A. S. Willsky, S. H. Nawab, and J.-J. Ding, *Signals and systems*. Prentice Hall Upper Saddle River, 1997, vol. 2.

[136] R. Nuyts, L. Tzeng, O. Mizuhara, and P. Gallion, "Effect of transmitter speed and receiver bandwidth on the eye margin performance of a 10-Gb/s optical fiber transmission system," *IEEE Photon. Techn. Lett.*, vol. 9, no. 4, pp. 532–534, 1997.

[137] S. Beppu, K. Kasai, M. Yoshida, and M. Nakazawa, "2048 QAM (66 Gbit/s) single-carrier coherent optical transmission over 150 km with a potential SE of 15.3 bit/s/Hz," *Opt. Exp.*, vol. 23, no. 4, pp. 4960–4969, 2015.

[138] K. Igarashi *et al.*, "114 space-division-multiplexed transmission over 9.8-km weakly-coupled-6-mode uncoupled-19-core fibers," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2015, pp. Th5C–4.

[139] A. Alvarado, T. Fehenberger, B. Chen, and F. M. Willems, "Achievable information rates for fiber optics: Applications and computations," *IEEE/OSA J. Lightw. Techn.*, vol. 36, no. 2, pp. 424–439, 2018.

[140] I. B. Djordjevic, B. Vasic, M. Ivkovic, and I. Gabitov, "Achievable information rates for high-speed long-haul optical transmission," *IEEE/OSA J. Lightw. Techn.*, vol. 23, no. 11, p. 3755, 2005.

[141] N. Merhav, G. Kaplan, A. Lapidoth, and S. S. Shitz, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, 1994.

[142] W. Cukierski, "Dogs vs. cats," 2013. [Online]. Available: https://kaggle.com/competitions/dogs-vs-cats

[143] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Rev. Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[144] M. Greenacre *et al.*, "Principal component analysis," *Nature Rev. Methods Primers*, vol. 2, no. 1, p. 100, 2022.

[145] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[146] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.

[147] S. Varughese, D. Lippiatt, T. Richter, S. Tibuleac, and S. E. Ralph, "Identification of soft failures in optical links using low complexity anomaly detection," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2019, pp. 1–3.

[148] E. Trentin and A. Freno, "Unsupervised nonparametric density estimation: A neural network approach," in *Proc IEEE Int. Joint Con. Neural Netw. (IJCNN)*, 2009, pp. 3140–3147.

[149] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Preprint, available online at https://arxiv.org/abs/:1312.6114*, 2013.

[150] M. E. Celebi and K. Aydin, *Unsupervised learning algorithms*. Springer, 2016, vol. 9.

[151] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[152] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *Towards Data Science*, vol. 6, no. 12, pp. 310–316, 2017.

[153] R. Chauhan, K. K. Ghanshala, and R. Joshi, "Convolutional neural network for image detection and recognition," in *Proc. IEEE Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, 2018, pp. 278–282.

[154] R. Hou, C. Chen, R. Sukthankar, and M. Shah, "An efficient 3D CNN for action/object segmentation in video," *Preprint, available online at https://arxiv.org/abs/:1907.08895*, 2019.

[155] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," *Preprint, available online at https://arxiv.org/abs/:1702.01923*, 2017.

[156] J. Zhang and K.-F. Man, "Time series prediction using rnn in multi-dimension embedding phase space," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, vol. 2, 1998, pp. 1868–1873.

[157] C. Smith and Y. Jin, "Evolutionary multi-objective generation of recurrent neural network ensembles for time series prediction," *Neurocomputing*, vol. 143, pp. 302–311, 2014.

[158] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artif. Intell. Rev.*, vol. 53, pp. 5929–5955, 2020.

[159] A. o. Vaswani, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[160] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Compu. Vis.* Springer, 2016, pp. 630–645.

[161] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. Eur. Conf. Compu. Vis.*, 2015, pp. 1–9.

[162] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[163] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. Int. Conf. Mach. Learn. Workshop (ICML workshop)*, 2012, pp. 37–49.

[164] V. Nair and G. E. Hinton, "3D object recognition with deep belief nets," *Adv. Neural Inf. Process. Syst.*, vol. 22, 2009.

[165] H. Lee, C. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area V2," *Adv. Neural Inf. Process. Syst.*, vol. 20, 2007.

[166] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1278–1286.

[167] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey," *Preprint, available online at https://arxiv.org/abs/:2101.00734*, 2021.

[168] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Preprint, available online at https://arxiv.org/abs/1412.6980*, 2014.

[169] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[170] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *Preprint, available online at https://arxiv.org/abs/1212.5701*, 2012.

[171] F. N. Khan, Y. Zhou, A. P. T. Lau, and C. Lu, "Modulation format identification in heterogeneous fiber-optic networks using artificial neural networks," *Opt. Exp.*, vol. 20, no. 11, pp. 12 422–12 431, 2012.

[172] S. Zhang, Y. Peng, Q. Sui, J. Li, and Z. Li, "Modulation format identification in heterogeneous fiber-optic networks using artificial neural networks and genetic algorithms," *Photon. Netw. Commun.*, vol. 32, pp. 246–252, 2016.

[173] D. Wang *et al.*, "Modulation format recognition and OSNR estimation using CNN-based deep learning," *IEEE Photon. Techn. Lett.*, vol. 29, no. 19, pp. 1667–1670, 2017.

[174] T. Tanimura, T. Hoshida, T. Kato, S. Watanabe, and H. Morikawa, "Intelligent adaptive coherent optical receiver based on convolutional neural network and clustering algorithm," *J. Opt. Commun. Netw.*, vol. 11, pp. A52–A59, 2019.

[175] W. Xiao, Z. Luo, and Q. Hu, "A review of research on signal modulation recognition based on deep learning," *Electron.*, vol. 11, no. 17, p. 2764, 2022.

[176] D. Wang and Sothers, "Data-driven optical fiber channel modeling: A deep learning approach," *IEEE/OSA J. Lightw. Techn.*, vol. 38, no. 17, pp. 4730–4743, 2020.

[177] H. Yang *et al.*, "Fast and accurate waveform modeling of long-haul multi-channel optical fiber transmission using a hybrid model-data driven scheme," *IEEE/OSA J. Lightw. Techn.*, vol. 40, no. 14, pp. 4571–4580, 2022.

[178] N. Gautam, A. Choudhary, and B. Lall, "Comparative study of neural network architectures for modelling nonlinear optical pulse propagation," *Opt. Fiber Techn.*, vol. 64, p. 102540, 2021.

[179] R. M. Bütler, C. Häger, H. D. Pfister, G. Liga, and A. Alvarado, "Model-based machine learning for joint digital backpropagation and pmd compensation," *IEEE/OSA J. Lightw. Techn.*, vol. 39, no. 4, pp. 949–959, 2020.

[180] X. Lin *et al.*, "Perturbation theory-aided learned digital back-propagation scheme for optical fiber nonlinearity compensation," *IEEE/OSA J. Lightw. Techn.*, vol. 40, no. 7, pp. 1981–1988, 2021.

[181] B. I. Bitachon, A. Ghazisaeidi, B. Baeuerle, M. Eppenberger, and J. Leuthold, "Deep learning based digital back propagation with polarization state rotation & phase noise invariance," in *Proc. Opt. Fiber Commun. Conf. (OFC)*.

[182] X. Jiang *et al.*, "Solving the nonlinear schrödinger equation in optical fibers using physics-informed neural network," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2021, pp. M3H–8.

[183] Y. Song, D. Wang, Q. Fan, X. Jiang, X. Luo, and M. Zhang, "Physics-informed neural operator for fast and scalable optical fiber channel modelling in multi-span transmission," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2022.

[184] X. Jiang, D. Wang, X. Chen, and M. Zhang, "Physics-informed neural network for optical fiber parameter estimation from the nonlinear schrödinger equation," *IEEE/OSA J. Lightw. Techn.*, vol. 40, no. 21, pp. 7095–7105, 2022.

[185] D. Wang *et al.*, "Applications of physics-informed neural network for optical fiber communications," *IEEE Commun. Mag.*, vol. 60, no. 9, pp. 32–37, 2022.

[186] B. Karanov *et al.*, "Concept and experimental demonstration of optical im/dd end-to-end system optimization using a generative model," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2020, pp. Th2A–48.

[187] A. Cohen and S. Derevyanko, "Generative adversarial network and end-to-end learning for optical fiber communication systems limited by the nonlinear phase noise," in *IEEE Int. Conf. Microw., Antennas, Communi., Electron. Syst. (COMCAS)*, 2021, pp. 241–246.

[188] H. Zhang *et al.*, "Fiber nonlinearity equalizer using MLP-ANN for coherent optical OFDM," in *Proc. IEEE Int. Conf. Opt. Communi. Netw. (ICOCN)*, 2019.

[189] Y. Zhao and other, "Low-complexity fiber nonlinearity impairments compensation enabled by simple recurrent neural network with time memory," *IEEE Access*, vol. 8, pp. 160 995–161 004, 2020.

[190] P. J. Freire *et al.*, "Transfer learning for neural networks-based equalizers in coherent optical systems," *IEEE/OSA J. Lightw. Techn.*, vol. 39, no. 21, pp. 6733–6745, 2021.

[191] S. Srivallapanondh *et al.*, "Multi-task learning to enhance generazability of neural network equalizers in coherent optical systems," *Preprint, available online at https://arxiv.org/abs/:2307.05374*, 2023.

[192] F. Huang *et al.*, "Multi-task learning aided neural networks for equalization in coherent optical system," in *Proc. IEEE Int. Conf. Commun. China (ICCC)*, 2022, pp. 874–877.

[193] Z. Xu, S. Dong, J. H. Manton, and W. Shieh, "Low-complexity multi-task learning aided neural networks for equalization in short-reach optical interconnects," *IEEE/OSA J. Lightw. Techn.*, vol. 40, no. 1, pp. 45–54, 2022.

[194] Y. Liu, B. Yang, and T. Xu, "Machine learning for fiber nonlinearity mitigation in long-haul coherent optical transmission systems: Invited paper," in *IEEE Proc. Int. Conf. Adv. Infocomm Techn. (ICAIT)*, 2019, pp. 124–127.

[195] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[196] P. Rodríguez, M. A. Bautista, J. Gonzalez, and S. Escalera, "Beyond one-hot encoding: Lower dimensional target embedding," *Image Vis. Comput.*, vol. 75, pp. 21–31, 2018.

[197] S. Li, C. Häger, N. Garcia, and H. Wymeersch, "Achievable information rates for nonlinear fiber communication via end-to-end autoencoder learning," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2018.

[198] Y. Kofman, E. Zehavi, and S. Shamai, "Performance analysis of a multilevel coded modulation system," *IEEE Trans. Commun.*, vol. 42, no. 234, pp. 299–312, 1994.

[199] M. Arabaci, I. B. Djordjevic, L. Xu, and T. Wang, "Nonbinary LDPC-coded modulation for high-speed optical fiber communication without bandwidth expansion," *IEEE Photon. J.*, vol. 4, no. 3, pp. 728–734, 2012.

[200] S. Cammerer *et al.*, "Trainable communication systems: Concepts and prototype," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5489–5503, 2020.

[201] S. Dörner, S. Cammerer, J. Hoydis, and S. t. Brink, "Deep learning based communication over the air," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 132–143, 2018.

[202] H. Ye, G. Y. Li, B.-H. F. Juang, and K. Sivanesan, "Channel agnostic end-to-end learning based communication systems with conditional GAN," in *Proc. IEEE Globecom Workshops*, 2018.

[203] F. A. Aoudia and J. Hoydis, "Model-free training of end-to-end communication systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2503–2516, 2019.

[204] V. Raj and S. Kalyani, "Backpropagating through the air: Deep learning at physical layer without channel models," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2278–2281, 2018.

[205] O. Jovanovic, M. P. Yankov, F. Da Ros, and D. Zibar, "Gradient-free training of autoencoders for non-differentiable communication channels," *J. Light. Techn.*, vol. 39, no. 20, pp. 6381–6391, 2021.

[206] T. J. O'Shea, T. Erpek, and T. C. Clancy, "Deep learning based MIMO communications," *Preprint, available online at https://arxiv.org/abs/:1707.07980*, 2017.

[207] A. Felix, S. Cammerer, S. Dörner, J. Hoydis, and S. Ten Brink, "OFDM-autoencoder for end-to-end learning of communications systems," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wirel.Commun. (SPAWC)*, 2018, pp. 1–5.

[208] F. A. Aoudia and J. Hoydis, "End-to-end learning for OFDM: From neural receivers to pilotless communication," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 2, pp. 1049–1063, 2021.

[209] M. Goutay, F. A. Aoudia, J. Hoydis, and J.-M. Gorce, "End-to-end learning of OFDM waveforms with PAPR and ACLR constraints," in *Proc. IEEE Globecom Workshops*, 2021, pp. 1–6.

[210] K. Choi, K. Tatwawadi, A. Grover, T. Weissman, and S. Ermon, "Neural joint source-channel coding," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2019, pp. 1182–1192.

[211] F. A. Aoudia and J. Hoydis, "Waveform learning for next-generation wireless communication systems," *IEEE Trans. Commun.*, vol. 70, no. 6, pp. 3804–3817, 2022.

[212] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, 2019.

[213] W. Jiang *et al.*, "End-to-end learning based bit-wise autoencoder for optical OFDM communication system," in *Proc. IEEE Asia Commun. Photon. Conf. (ACP)*, 2021, pp. T4A–64.

[214] B. Karanov, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-end optimized transmission over dispersive intensity-modulated channels using bidirectional recurrent neural networks," *Opt. Exp.*, vol. 27, no. 14, pp. 19 650–19 663, 2019.

[215] T. Uhlemann, S. Cammerer, A. Span, S. Doerner, and S. ten Brink, "Deep-learning autoencoder for coherent and nonlinear optical communication," in *Proc. IEEE Photon. Netw. ITG-Symp.* VDE, 2020, pp. 1–8.

[216] R. T. Jones *et al.*, "Geometric constellation shaping for fiber optic communication systems via end-to-end learning," *Preprint, available online at https://arxiv.org/abs/:1810.00774*, 2018.

[217] M. P. Yankov, O. Jovanovic, D. Zibar, and F. Da Ros, "Rate adaptive geometric constellation shaping using autoencoders and Many-To-One mapping," *Preprint, available online at https://arxiv.org/abs/:2307.09897*, 2023.

[218] K. Gümüş, A. Alvarado, B. Chen, C. Häger, and E. Agrell, "End-to-end learning of geometrical shaping maximizing generalized mutual information," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2020.

[219] A. Rode, B. Geiger, and L. Schmalen, "Geometric constellation shaping for phase-noise channels using a differentiable blind phase search," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2022.

[220] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, "Joint channel coding and modulation via deep learning," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wirel. Commun. (SPAWC)*, 2020.

[221] V. Aref and M. Chagnon, "End-to-end learning of joint geometric and probabilistic constellation shaping," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2022.

[222] A. Caciularu and D. Burshtein, "Blind channel equalization using variational autoencoders," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2018.

[223] ——, "Unsupervised linear and nonlinear channel equalization and decoding using variational autoencoders," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 3, pp. 1003–1018, 2020.

[224] V. Lauinger, F. Buchali, and L. Schmalen, "Blind equalization and channel estimation in coherent optical communications using variational autoencoders," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2529–2539, 2022.

[225] X. Yang, "Understanding the variational lower bound," *Inst. Adv. Comp. Stud. Univ. Maryland*, 2017.