Recent advances in constellation optimization for fiber-optic channels

Metodi P. Yankov(1), Ognjen Jovanovic(1), Darko Zibar(1), Francesco Da Ros(1)

(1) Department of Electrical and Photonics Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark meya, ognjo, dazi, fdro@dtu.dk

Abstract *The autoencoder concept for geometric constellation shaping is discussed. Applications in coherent optical fiber communications are presented. Several popular training algorithms are compared. The quantization problem of finite precision DAC and ADC is addressed. ©2022 The Author(s)*

Introduction

Optical fiber wavelength division multiplexing (WDM) communication systems are becoming pressed to increase their spectral efficiency in order to accommodate a higher throughput in standard, single mode fibers. Spectral efficiency close to the Shannon limit can only be achieved through optimization of the modulation format, either probabilistically or geometrically. Probabilistic amplitude shaping (PAS) is a proven strategy for increasing the spectral efficiency and provide rate adaptivity[1]. However, as the line rates of coherent transceivers increase beyond the 100-Gbd range and the single-carrier speed exceed 1 Tbps, due to the serial nature of the distribution matcher and dematcher, PAS may become problematic for high-speed hardware implementation[2]. Geometric constellation shaping (GCS) presents a low-complexity alternative, since conventional bit-interleaved coded modulation (BICM) can in principle be supported by simply optimizing the locations of the constellation points on the I/Q plane. Due to its versatility w.r.t. the channel model employed, the autoencoder (AE) concept is gathering traction as a method to approach GCS.

The AE principle for communication was first popularized for wireless[3] and Rayleigh fading channels[4], but its popularity is increasing in the optical communications community. It has been applied for geometric shaping in several notable scenarios, namely (a non-complete list):

- long-haul coherent communications for symbol-wise[5] and then extended to bit-wise performance maximization[6]–[8];
- robustness to phase noise[9]–[13];
- intensity modulation direct detection channels[14]–[16], including robustness to varying dispersion[17];
- joint GCS, pulse shaping, predistortion and robustness to other hardware impairments[18],[19], as well as joint GCS and PS[20];
- nonlinear frequency division multiplexing[21],[22].

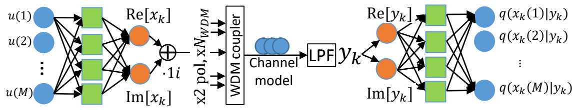

Fig. 1: Illustration of an autoencoder. The channel model may be any function. When CE is used for training, the decoder output are estimates of the posterior probabilities.

In this paper, the AE principle is reviewed for coherent communication and different training algorithms are compared w.r.t. their performance. The problem of quantization noise and how to address it using an AE is also discussed.

Autoencoder training

An AE is composed of an encoder neural network (NN), a decoder NN, and a channel model in between. The task of the encoder is to map the one-hot encoded vectors $\mathbf{u} = [u\,(1)\,, u\,(2)\,, ...u\,(M)]$ of size M onto symbols in the I/Q plane using some linear or nonlinear transformation. The task of the decoder is to provide estimates of the posterior probabilities of the one-hot encoded vectors, which coincide with the posterior probabilities of the constellation symbols $\{p(x(i)|y), i = 1..M\}$ from the continuous channel observations y. An illustration of an AE is given in Fig. 1. For communications applications, the AE is typically trained using the cross-entropy (CE) cost function, which in this case coincides with the conditional entropy $CE = \mathcal{H}(X|Y) = \mathbb{E}_k\left[-\log_2 p(x_k|y_k)\right]$, where x_k and y_k are the output of the encoder and input to the decoder at the discrete time index k, respectively. The achievable information rate (AIR) of the system is estimated from the testing cost as the mutual information (MI) between the channel input and output $\mathcal{I}(X;Y) = \mathcal{H}(X) - \mathcal{H}(X|Y)$. The CE cost function is well-justified because for a fixed input distribution $p(X)$, e.g. uniform, its minimization directly results in maximization of the AIR. It is worth noting that the decoder NN essentially represents an auxiliary distribution $q(x|y)$ to the true posterior distribution $p(x|y)$. The AIR with the decoder NN is thus a lower-bound on the channel capacity. Another lower bound can be estimated by replacing the decoder NN's distribution $q(x|y)$ by any other function. A practically relevant one and what is considered in this paper is given by $q_G(x|y) \propto p(x) \cdot \exp\left(-\frac{1}{\sigma^2}|y-x|^2\right)$, known as the Gaussian auxiliary channel, where σ^2 is the empirically estimated variance of the white Gaussian noise.

The standard approach to train an AE is to use back-propagation (BP) of gradients from the cost function to the optimization variables, which are the weights of the encoder and decoder NNs. This approach is mathematically rigorous and especially in recent years with the explosion of the development of automatic differentiation tools like PyTorch and Tensorflow has become the benchmark. Two main drawbacks of this approach are 1) the channel model needs to be differentiable; 2) the channel model needs to be computationally tractable. The first drawback gives rise to issues when non-differentiable portions of the effective channel need to be included. This is especially the case for channel models that need to represent the true practical channel as well as possible. One example is the classical blind phase search (BPS) algorithm, wherein the argmax function precludes computation of gradients. Another example is performing optimization on experimental setups. The second drawback gives rise to practical issues related to computing resources, especially GPU memory and running time for forward and BP through the channel model. An example here is a channel model described by the split step Fourier method (SSFM) for solving the nonlinear Schrödinger equation for fiber propagation. Each step in the SSFM is essentially a nonlinear layer in a NN, and 100s of 1000s steps are necessary for accurate description of a long-haul optical fiber transmission. Each step needs to be stored as part of the computational graph, which quickly becomes infeasible. A solution to this problem is to apply *checkpoints* at given instances of the forward pass through the graph[22]. These can be e.g. every 100 steps in the SSFM. The basic idea is that the output of each checkpoint is stored, and when performing the backpropagation, the forward pass is re-computed in between checkpoints instead of storing the output of each step. Essentially, memory is traded-off for computation speed.

Tab. 1: Comparison between AE training algorithms.

	BP	RL	CKF
# of SSFM props. per batch	1	20	2*#of weights ≈ 4500
# of epochs for conversion	30	45	40
Processing	serial	serial	parallel
TX and RX optimization	joint	iterative	joint
Usable in experiment	no	yes	yes
Require memory for gradient	yes	no	no

Both of these drawbacks can potentially be addressed with semi-heuristic optimization strategies. Here, we focus on two such options. The first relies on introducing a state space model for the AE, and learning the AE weights using conventional Bayesian inference. An example here is the cubature Kalman filter (CKF), which was applied to learn an AE for a channel containing the non-differentiable BPS algorithm[23]. In a nutshell, the AE is represented with a state-space model, and the parameters of the state-space representation, i.e. the NN weights, are estimated using the CKF rules.

The second relies on the general class of action/reward algorithms, in which many candidate weight sets are transmitted through the forward channel, and specific update rules are designed to select the suitable candidates for the next iteration. Examples here are genetic algorithms, the reinforcement learning (RL) strategy[24] and the model-free training concepts[25],[26].

We also mention the method of training a generative-adversarial network (GAN)[27] for the channel which is time consuming and complex, but shows some promise for going around the differentiability requirement of BP.

Results

We study a link of 1000 km, 5 WDM channels, 0.01 square-root cosine roll-off, 64QAM. A summary of the algorithms described above is given in Table 1, together with a comparative estimate of their computational complexity when applied to learning an AE for communication through a specific fiber link modeled using the SSFM. In all cases, the encoder NN is a linear mapping to the output and the decoder NN has a hidden layer of 32 nodes with the ReLU activation function. All algorithms are pre-trained on a simple, shorter link of 500 km and a large SSFM step of 10 km for computation speed, and then a few epochs are performed on the desired link of 10 spans with a 100 m SSFM step for convergence. For BP after the pre-convergence, checkpointing is employed at every 10 steps, corresponding to every 1 km. The AE is trained to optimize the performance of

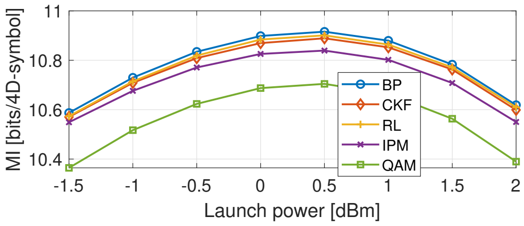

Fig. 2: Comparison of training algorithms. The AWGN channel-optimized iterative polar modulation[28] also plotted for comparison.

one of the polarizations of the central channel, indicated by the low pass filter (LPF) in Fig. 1. The receiver also includes chromatic dispersion compensation and static phase de-rotation. Both polarizations of all channels are generated using identical encoder NNs. In principle, joint shaping and detection across polarizations and frequency channels is straight-forward with AEs. In such cases, special care must be taken with overfitting the NNs due to the exponentially growing number of trainable parameters. The performance of the three algorithms is given in Fig. 2 in terms of the symbol-wise AIR with a Gaussian auxiliary channel: i.e., a decoder NN is used to train the AE with the CE cost, while in order to resemble a more practical coded modulation scheme, the Gaussian receiver is employed for testing. The best performing algorithm is BP with a gain of 0.21 bits/4D w.r.t. square QAM. The RL and CKF perform similar to each other with a penalty w.r.t. BP of 0.015 and 0.027 bits/4D, respectively.

One of the main potential drawbacks of GCS w.r.t. PAS is the irregular projections of the points on the I and Q axes of the plane, which may require a finer quantization. The quantization issue is especially important at high effective signal to noise ratio (SNR) relative to the constellation format size. That is, when the target AIR is close to the entropy $H(X)$. However, we mention that shaping in general is more effective at lower rates corresponding to the operating points of powerful soft-decision FEC of overheads between 40% and 20%. The quantization noise may be modeled using a uniformly distributed random variable $w_q \propto \mathcal{U}_{[-D/2^{nQbits-1},D/2^{nQbits-1}]}$, where $nQbits$ are the number of quantization bits, and D is an optimized dynamic range of the ADC and DAC. Here, we assume $D = 1.2 \cdot \max_i Re\,[x(i)]$. For simplicity, an ideal, frequency flat quantization noise is considered. The quantization noise is added for each training and testing batch to $Re\,[y]$, $Im\,[y]$, $Re\,[x]$ and $Im\,[x]$, assuming again for simplicity identically distributed quantization noise from DAC and ADC. The performance of

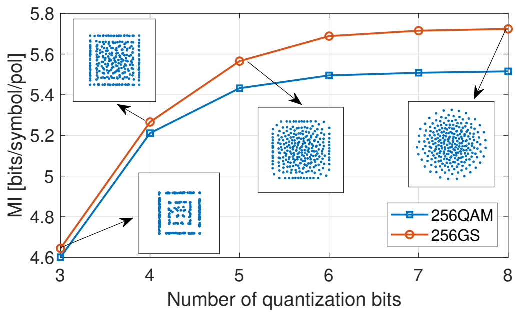

Fig. 3: Effect of quantization noise on the shaping gain and the optimized constellations using an AE.

the BP algorithm on a quantized channel modeled using the NLIN model[6] is given in Fig. 3 for 256QAM, 5 channel, 10 spans x 100km in [bits/QAM symbol]. For 5 bit-DAC and ADC, the GCS gain deteriorates from ≈ 0.2 to ≈ 0.15 bits/symbol. At 4 bits, the gain disappears, and at 3 bits the optimized constellation tends to a rectangular one, with a few distinct discrete points on each I and Q, which we speculate converge on the optimum quantization level selection for the system under test. This is left for future work to confirm.

Finally, we mention that in this paper, for the sake of algorithm comparison the symbol-wise MI and CE were discussed. Typically, coherent optical communication systems employ a bit-interleaved coded modulation (BICM), where the bit-wise MI is of interest[29]. Optimization of an AE targeting bit-wise performance can also be approached[6] and we do not expect that the conclusions will be altered in that case.

Conclusions

Several training algorithms were compared for AE-based GCS. When the channel model is differentiable and reasonably simple, BP is the superior choice. For complicated channels which are still differentiable, BP can be applied using checkpointing. For non-differentiable and experimental channels, RL and CKF present decent alternatives to BP with slight degradation in performance. Finally, it was demonstrated that for notable shaping gains to be achieved with GCS w.r.t 256QAM, at least 5 bits ADC and DAC resolution is required. Experimental demonstrations of the presented methods are of obvious interest for confirming both the quantization analysis, and the non-differentiable model support of RL and CKF.

Acknowledgements

This work was supported by the Danish national research foundation SPOC project, ref. DNRF123, the European Research Council through the ERC-CoG FRECOM project (grant agreement no. 771878), and the Villum Young Investigator OPTIC-AI project (grant no. 29334).

References

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

[2] D. S. Millar, T. Fehenberger, T. Koike-Akino, K. Kojima, and K. Parsons, “Distribution matching for high spectral efficiency optical communication with multiset partitions”, *J. Light. Tecnh.*, vol. 37, no. 2, pp. 517–523, 2019.

[3] T. O’Shea and J. Hoydis, “An Introduction to Deep Learning for the Physical Layer”, *IEEE Trans. on Cogn. Commun. and Netw.*, vol. 3, no. 4, pp. 563–575, 2017, ISSN: 23327731. arXiv: `1702.00832`.

[4] T. J. O’Shea, T. Erpek, and T. C. Clancy, “Physical layer deep learning of encodings for the MIMO fading channel”, in *2017 55th Annual Allerton*, 2017, pp. 76–80.

[5] R. T. Jones, T. A. Eriksson, M. P. Yankov, and D. Zibar, “Deep Learning of Geometric Constellation Shaping Including Fiber Nonlinearities”, *ECOC*, 2018. arXiv: `1805.03785`.

[6] R. T. Jones, M. P. Yankov, and D. Zibar, “End-to-end learning for GMI optimized geometric constellation shape”, in *ECOC*, 2019, pp. 1–3.

[7] M. Schaedler, S. Calabrò, F. Pittalà, *et al.*, “Neural network assisted geometric shaping for 800Gbit/s and 1Tbit/s optical transmission”, *OFC*, pp. 3–5, 2020.

[8] K. Gümüs, A. Alvarado, B. Chen, C. Häger, and E. Agrell, “End-to-end learning of geometrical shaping maximizing generalized mutual information”, *OFC*, pp. 10–12, 2020. arXiv: `1912.05638`.

[9] S. Li, C. Häger, N. Garcia, and H. Wymeersch, “Achievable Information Rates for Nonlinear Fiber Communication via End-to-end Autoencoder Learning”, *ECOC*, 2018.

[10] V. Talreja, T. Koike-Akino, Y. Wang, D. S. Millar, K. Kojima, and K. Parsons, “End-to-End Deep Learning for Phase Noise-Robust Multi-Dimensional Geometric Shaping”, in *ECOC*, 2020, pp. 1–3.

[11] A. Rode, B. Geiger, and L. Schmalen, “Geometric constellation shaping for phase-noise channels using a differentiable blind phase search”, in *OFC*, 2022, Th2A.32.

[12] O. Jovanovic, M. P. Yankov, F. Da Ros, and D. Zibar, “End-to-end Learning of a Constellation Shape Robust to Variations in SNR and Laser Linewidth”, in *ECOC*, 2021, pp. 1–3.

[13] ——, “End-to-end Learning of a Constellation Shape Robust to Channel Condition Uncertainties”, *Journal of Lightwave Technology*, 2022.

[14] M. P. Yankov, F. Da Ros, S. Forchhammer, and L. Grüner-Nielsen, “Capacity and Achievable Rates of Fading Few-mode MIMO IM/DD Optical Fiber Channels”, in *accepted, International Conference on Communications, ICC*, 2022. [Online]. Available: `https://arxiv.org/abs/2201.11538`.

[15] B. Karanov, M. Chagnon, F. Thouin, *et al.*, “End-to-End Deep Learning of Optical Fiber Communications”, *J. Light. Techn.*, vol. 36, no. 20, pp. 4843–4855, 2018.

[16] B. Karanov, D. Lavery, P. Bayvel, and L. Schmalen, “End-to-end optimized transmission over dispersive intensity-modulated channels using bidirectional recurrent neural networks”, *Optics Express*, vol. 27, no. 14, p. 19 650, 2019.

[17] B. Karanov, L. Schmalen, and A. Alvarado, “Distance-Agnostic Auto-Encoders for Short Reach Fiber Communications”, in *OFC*, 2021, pp. 1–3.

[18] J. Song, C. Häger, J. Schröder, A. G. i Amat, and H. Wymeersch, “End-to-end Autoencoder for Superchannel Transceivers with Hardware Impairment”, in *OFC*, Optical Society of America, 2021, F4D.6.

[19] J. Song, C. Hager, J. Schroeder, A. Graell i Amat, and H. Wymeersch, “Model-Based End-to-End Learning for WDM Systems With Transceiver Hardware Impairments”, *IEEE J. of Sel. Topics in Quant. Elec.*, pp. 1–1, 2022.

[20] V. Aref and M. Chagnon, “End-to-end learning of joint geometric and probabilistic constellation shaping”, in *OFC*, 2022, W4I.3.

[21] S. Gaiarin, R. T. Jones, F. Da Ros, and D. Zibar, “End-to-end optimized nonlinear Fourier transform-based coherent communications”, *CLEO*, SF2L.4, 2020.

[22] S. Gaiarin, F. Da Ros, R. T. Jones, and D. Zibar, “End-to-End Optimization of Coherent Optical Communications Over the Split-Step Fourier Method Guided by the Nonlinear Fourier Transform Theory”, *J. Light. Tecnh.*, vol. 39, no. 2, pp. 418–428, 2021.

[23] O. Jovanovic, M. P. Yankov, F. Da Ros, and D. Zibar, “Gradient-Free Training of Autoencoders for Non-Differentiable Communication Channels”, *J. Light. Tecnh.*, vol. 39, no. 20, pp. 6381–6391, 2021.

[24] J. Song, Z. He, C. HÃ¤ger, *et al.*, “Over-the-fiber digital predistortion using reinforcement learning”, in *ECOC*, 2021, pp. 1–4.

[25] F. A. Aoudia and J. Hoydis, “End-to-End Learning of Communications Systems Without a Channel Model”, in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, 2018, pp. 298–303.

[26] ——, “Model-Free Training of End-to-End Communication Systems”, *IEEE J. on Sel. Areas in Commun.*, vol. 37, no. 11, pp. 2503–2516, 2019.

[27] B. Karanov, M. Chagnon, V. Aref, D. Lavery, P. Bayvel, and L. Schmalen, “Concept and experimental demonstration of optical IM/DD end-to-end system optimization using a generative model”, *OFC*, pp. 1–3, 2020.

[28] I. B. Djordjevic, H. G. Batshon, L. Xu, and T. Wang, “Coded polarization-multiplexed iterative polar modulation (PM-IPM) for beyond 400 Gb/s serial optical transmission”, in *OFC/NFOEC 2010*, 2010, pp. 1–3.

[29] A. Alvarado, E. Agrell, D. Lavery, R. Maher, and P. Bayvel, “Replacing the soft-decision FEC limit paradigm in the design of optical communication systems”, *J. Light. Tecnh.*, vol. 34, no. 2, pp. 707–721, 2016.