Simplifying the Supervised Learning of Kerr Nonlinearity Compensation Algorithms by Data Augmentation

Vladislav Neskorniuk[(1,*)], Pedro J. Freire[(1)], Antonio Napoli[(2)], Bernhard Spinnler[(2)],
Wolfgang Schairer[(2)], Jaroslaw E. Prilepsky[(1)], Nelson Costa[(3)], Sergei K. Turitsyn[(1)]

[(1)] Aston Institute of Photonic Technologies, Aston University, Birmingham, UK
[(2)] Infinera, Munich, Germany   [(3)] Infinera, Carnaxide, Portugal   [(*)] E-mail: v.neskorniuk@aston.ac.uk

Abstract *We propose a data augmentation technique to improve performance and decrease complexity of the supervised learning of nonlinearity compensation algorithms. We demonstrate both numerically and experimentally that the augmentation allows reducing the training dataset size up to 6 times while keeping the same post-compensation bit-error rate.*

Introduction

The Kerr nonlinearity is one of the main factors limiting the capacity of modern fiber-optic communication systems[1]. Among others, the nonlinearity compensation (NLC) algorithms trained via supervised learning (SL-NLC) like perturbative post-distortion[2] or machine-learning-based methods[3]–[5] have been intensively studied recently. The SL-NLC algorithms obtain the system parameters required for their operation by analyzing the pre-collected dataset of the algorithm inputs and desired outputs[6]. The main advantages of the SL-NLC algorithms are: i) not requiring the prior knowledge of the link parameters; ii) the ability to automatically adapt to changes in the link by re-training on a newly collected dataset.

Nevertheless, SL approaches may require huge training datasets which can lead to technical challenges. The first emerges from the increase of the training complexity with the dataset size[7], imposing an upper bound on the size. The second is the idleness of the communication line during SL-NLC training and dataset collection. These factors may limit the capacity of the link if frequent SL-NLC re-training on big data is needed due to the rapid change of model parameters or usage of flexible routing. Therefore, the efficient use of a limited dataset is highly desirable for the practical implementation of SL-NLC algorithms. We propose here to synthetically expand the dataset by generating new artificial input objects from the original ones. This technique is referred to as data augmentation (DA) in the general ML context[8]. DA generally improves the diversity of dataset objects and, thereby, makes the dataset a more representative sample of the set of all possible transmitted signals. Such a technique is frequently used in computer vision applications[8]. In optical communications, DA was already explored for predicting failures[9],[10] and traffic peculiarities[11].

In this work, we propose for the first time, to the best of our knowledge, the DA for SL-NLC training. We demonstrate numerically and experimentally that the SL-NLC algorithms, trained with datasets augmented using the proposed technique, achieve similar bit-error rate (BER) as SL-NLC algorithms trained with 4 to 6 times larger pure datasets. The suggested DA can be used to improve the performance of a general SL-NLC trained with insufficient data or, for the case of abundant data, to decrease its training complexity by further limiting the dataset used.

Data augmentation technique

We demonstrate the proposed technique for DA using the Manakov equations[1],[12] describing the practical case of the averaged evolution of a dual-polarized (DP) optical signal along a fiber-optic link:

$$\frac{\partial u_{h/v}}{\partial z} = \frac{G(z)}{2} u_{h/v} - i \frac{\beta_2(z)}{2} \frac{\partial^2 u_{h/v}}{\partial t^2} +$$
$$+ i \frac{8\gamma(z)}{9} (|u_h|^2 + |u_v|^2) u_{h/v} + \xi(z,t), \quad (1)$$

where $u_h(z,t)$ and $u_v(z,t)$ are, respectively, the horizontal (h) and vertical (v) polarizations of the optical signal waveform $u(z,t)$; $\beta_2(z)$ is the group velocity dispersion (GVD) coefficient; $\gamma(z)$ is the effective nonlinear coefficient; $G(z) = -\alpha(z) + \sum_{m=1}^{N_{OA}} \Gamma_m \delta(z - L_m)$ describes the optical losses $\alpha(z)$ fully compensated by lumped optical amplifiers (OAs) with gain Γ_m located at the end of every fiber span L_m; and $\xi(z,t)$ is the amplified spontaneous emission (ASE) noise injected by OAs.

Let us define the solution of Eq. (1) as a pair of signal waveforms at the channel input
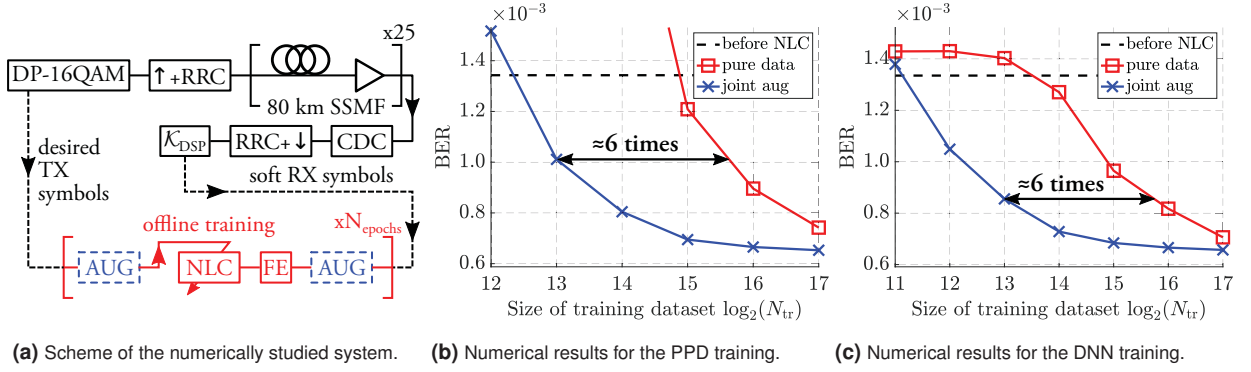
(a) Scheme of the numerically studied system. **(b)** Numerical results for the PPD training. **(c)** Numerical results for the DNN training.

Fig. 1: Numerical study of the dependence of BER (after SL-NLC) on the size of dataset N_{tr} for a 25×80 km SSMF system.

and output: $\{u_{h/v}(0,t),\, u_{h/v}(z,t)\}$. Several transformations can be used to generate the new solutions of Eq. (1) out of the known ones, for instance: i) the discrete phase shift $(\Delta\varphi_{\mathrm{disc}})$ — $\{\bar{u}_{h/v}(0,t)\cdot\exp^{i\varphi},\, \bar{u}_{h/v}(z,t)\cdot\exp^{i\varphi}\}$ $\forall\varphi\in\{0,\pi/2,\pi,3\pi/2\}$; ii) the time-inversion (t_{inv}) — $\{\bar{u}_{h/v}(0,-t),\, u_{h/v}(z,-t)\}$; and iii) the polarization swapping (H/V_{swap}) — $\{u_{v\leftrightarrow h}(0,t),\, u_{v\leftrightarrow h}(z,t)\}$.

Essentially, the NLC aims to predict the transmitted signal $u(0,t)$ from the received one $u(z,t)$. Therefore, the SL-NLC training dataset is formed by Eq. (1) solutions. We propose augmenting this dataset before every training epoch, i.e., a single run over the dataset of the algorithm optimizing the SL-NLC model, by replacing a randomly chosen part of the original dataset with the objects generated via the transformations itemized above. In more detail, we apply a randomly chosen discrete phase shift $\Delta\varphi_{\mathrm{disc}}$ to every dataset object, while the time-inversion t_{inv} and polarization swapping H/V_{swap} are each applied to separately randomly chosen halves of the dataset objects. To augment the received signal features used by NLC algorithms, like nonlinear perturbation triplets (NPTs) $T_{j,k}$ from Eq. (3), one can, first, apply the transformations (AUG) to the received signal $u(z,t)$ and then extract the features out of it (FE) (see Fig. 1a). Thus, during every epoch, SL-NLC is effectively trained on a new dataset and, hence, obtains more information about the channel than via the conventional training on a fixed dataset.

Notably, the proposed DA is applicable to the SL-NLC algorithms operating after conventional digital signal processing (DSP)[13]. The collateral distortion introduced by DSP into the restored received signal can be viewed as its normalization by a constant $\mathcal{K}_{\mathrm{DSP}}\in\mathbb{C}$ chosen to satisfy:

$$\mathcal{K}_{\mathrm{DSP}} = \min_{\mathcal{K}}\left\|\mathcal{K}\cdot u_{h/v}(z,t) - u_{h/v}(0,t)\right\|_2, \qquad (2)$$

where $\|\cdot\|_2$ is the Euclidean norm. Since the augmentation transformations are still valid if the received signal $u_{h/v}(z,t)$ is multiplied by an arbitrary complex-valued coefficient $\mathcal{K}\in\mathbb{C}$, they are robust to the side-effect DSP normalization (Eq. (2)).

The condsidered SL-NLC algorithms

The potential of DA was evaluated considering two SL-NLC algorithms. Firstly, we consider the adaptive perturbative post-distortion (PPD)[2]. PPD estimates the nonlinear distortion $\Delta u_h = u_h(z,t_0) - u_h(0,t_0)$ in the transmitted symbol of interest (SOI) $H_{\mathrm{TX}} = u_h(0,t_0)$ via a weighted sum of NPTs $T_{j,k}$:

$$\Delta u_h = \sum_{j,k} C_{j,k}\ \underbrace{H_j(H_k H_{j+k}^* + V_k V_{j+k}^*)}_{T_{j,k}}. \qquad (3)$$

Here $H_j = u_h(z,t_0+j\Delta t)$, $V_j = u_v(z,t_0+j\Delta t)$ are the elements of the received symbol sequences $\bar{H}, \bar{V}$ in h- and v- polarizations, respectively; Δt is the symbol period; $C_{j,k}\in\mathbb{C}$ are the learnt coefficients; and j and k are the symbol indices with respect to the received SOI $u_h(z,t_0) = H_0$. Secondly, we examine the deep neural network (DNN) used in[14]. The DNN input is formed by separated real and imaginary parts of NPTs $T_{j,k}$ along with the received symbol sequences $\bar{H}, \bar{V}$ centered around the SOI. The DNN has two hidden layers with, correspondingly, 2 and 10 neurons employing SeLU[15] activation function. The second hidden layer is followed by a 20% dropout[16]. The output layer has 2 neurons without activation, each predicting the real and imaginary part of nonlinear distortion Δu_h.

For the NLC input we used the symbol sequences $\bar{H}, \bar{V}$ of length $2*75+1$, i.e. $|j|\leq 75$ and NPTs $T_{j,k}$ fulfilling the condition $|k|\leq \min\{\mathrm{ceil}(75/|j|), 75\}$, as in[17]. This procedure led us to 301 complex-valued input symbols from both polarizations and 1929 NPTs. The algorithms were trained by Adam optimizer[18] min-
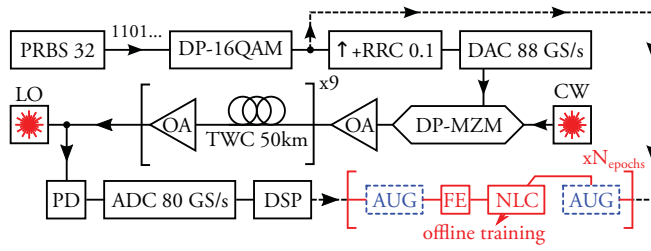
(a) Scheme of the experimentally studied system.
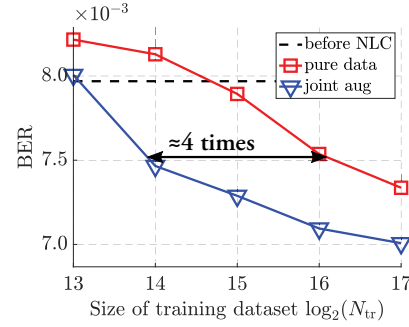
(b) Experimental results for DNN training.

Fig. 2: Experimental study of the dependence of BER after SL-NLC on the size of dataset N_{tr} for 9×50 km TWC system.

imising the mean-squared error (MSE) $\|H_{TX} - (H_0 - \Delta u_h)\|_2$. The training was carried out for 200 epochs with a batch size 100. The testing dataset, used to evalute BER, had 2^{17} objects.

Numerical setup

We numerically simulated the transmission of a single-channel DP-16QAM 64 Gbaud signal with root-raised-cosine (RRC) 0.06 roll-off pulse over a system consisting of 25×80 km standard single-mode fiber (SSMF) spans (see Fig. 1a). The SSMF parameters are: $\alpha = 0.21$ dB/km, $D = 16.8$ ps/(nm*km), and $\gamma = 1.14$ (W*km)$^{-1}$. Every span was followed by an ideal OA with noise figure NF $= 4.5$ dB. The waveform evolution was simulated by solving Eq. (1) via split-step Fourier method[19]. At the receiver, after full electronic chromatic dispersion compensation (CDC) by the frequency-domain equalizer (FDE) and downsampling to the symbol rate, the received symbols were normalized by Eq. (2). The SL-NLC algorithms were applied off-line, like in[17],[20], on the pre-collected datasets. The symbol patterns forming the training and testing datasets were randomly generated by the Mersenne Twister[21].

Experimental setup

In practice, not only Kerr nonlinearities but also transceiver impairments are a significant source of distortion[22]. To demonstrate the full potential of the proposed method, we consider an experimental setup where both impairments contribute to signal distortion (see Fig. 2a). We transmitted DP-16QAM 34.4 Gbaud RRC 0.1 roll-off symbol sequences in a single 37.5 GHz channel centered at $\lambda = 1.55 \ \mu m$. The used digital-to-analog converter (DAC) operated at 88 Gsamples/s. The optical path consisted of 9×50 km spans of TrueWave Classic (TWC) optical fiber with each span followed by an erbium-doped-fiber amplifier (EDFA) fully compensating for the losses. The parameters of the TWC fiber are: $\alpha = 0.23$ dB/km, $D = 2.8$ ps/(nm*km), and $\gamma = 2.1$ (W*km)$^{-1}$. The

EDFA noise figures were in the 4.5 to 5 dB range. After the link, the optical signal was received by an integrated coherent receiver (ICR), followed by balanced photodiodes (PD) and an 80 Gsamples/s analog-to-digital converter (ADC) and then processed using the DSP. After the DSP, the soft symbols were collected for off-line processing by NLC. The testing and training data were randomly generated using PRBS $2^{32} - 1$ bit sequences.

Results and discussion

We compared the BERs achieved for different training dataset sizes N_{tr}: i) before the NLC; after the SL-NLC trained ii) with non-augmented datasets (pure data) or iii) with the dataset augmented by joint application of several transformations (joint aug). In the numerical study, the joint augmentation was formed by $\Delta\varphi_{disc} + t_{inv} + H/V_{swap}$ transformations, while in the experiment the time-inversion t_{inv} was excluded, since it led to a higher BER compared to training with pure dataset, supposedly, because of the asymmetrical memory introduced by the transceiver impairments. For every considered size N_{tr} and augmentation type, we separately optimized the power level and the learning rate to reach the best BER. The BERs were averaged over 20 epochs to remove local performance fluctuations.

In the numerical study, the joint augmentation led to the same BER achieved on a 6 times smaller pure dataset for both PPD (Fig. 1b) and DNN (Fig. 1c). In the experiment, the joint augmentation led to the reduction of the dataset size by 4 times while keeping the same BER after the DNN (see Fig. 2b). As expected, PPD did not show BER improvement in the experiment, since it is not designed for the considered metro link with significant transceiver distortions[22],[23].

References

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

[2] M. Malekiha, I. Tselniker, and D. V. Plant, "Efficient nonlinear equalizer for intra-channel nonlinearity compensation for next generation agile and dynamically reconfigurable optical networks", *Optics Express*, vol. 24, no. 4, pp. 4097–4108, 2016.

[3] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, and M. Tornatore, "An overview on application of machine learning techniques in optical networks", *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1383–1408, 2018.

[4] F. N. Khan, Q. Fan, C. Lu, and A. P. T. Lau, "An optical communication's perspective on machine learning and its applications", *Journal of Lightwave Technology*, vol. 37, no. 2, pp. 493–516, 2019.

[5] D. Zibar, M. Piels, R. Jones, and C. G. Schäeffer, "Machine learning techniques in optical communication", *Journal of Lightwave Technology*, vol. 34, no. 6, pp. 1442–1452, 2015.

[6] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[8] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning", *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[9] L. Cui, Y. Zhao, B. Yan, D. Liu, and J. Zhang, "Deep-learning-based failure prediction with data augmentation in optical transport networks", in *17th International Conference on Optical Communications and Networks (ICOCN2018)*, International Society for Optics and Photonics, vol. 11048, 2019, p. 110482I.

[10] H. Zhuang, Y. Zhao, X. Yu, Y. Li, Y. Wang, and J. Zhang, "Machine-learning-based alarm prediction with gans-based self-optimizing data augmentation in large-scale optical transport networks", in *2020 International Conference on Computing, Networking and Communications (ICNC)*, IEEE, 2020, pp. 294–298.

[11] J. Li, D. Wang, S. Li, M. Zhang, C. Song, and X. Chen, "Deep learning based adaptive sequential data augmentation technique for the optical network traffic synthesis", *Optics express*, vol. 27, no. 13, pp. 18 831–18 847, 2019.

[12] S. V. Manakov, "On the theory of two-dimensional stationary self-focusing of electromagnetic waves", *Soviet Physics-JETP*, vol. 38, no. 2, pp. 248–253, 1974.

[13] M. S. Faruk and S. J. Savory, "Digital signal processing for coherent transceivers employing multilevel formats", *Journal of Lightwave Technology*, vol. 35, no. 5, pp. 1125–1141, 2017.

[14] V. Kamalov, L. Jovanovski, V. Vusirikala, S. Zhang, F. Yaman, K. Nakamura, T. Inoue, E. Mateo, and Y. Inada, "Evolution from 8qam live traffic to ps 64-qam with neural-network based nonlinearity compensation on 11000 km open subsea cable", in *Optical Fiber Communication Conference*, Optical Society of America, 2018, Th4D–5.

[15] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks", in *Advances in neural information processing systems*, 2017, pp. 971–980.

[16] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors", *arXiv preprint arXiv:1207.0580*, 2012.

[17] S. Zhang, F. Yaman, K. Nakamura, T. Inoue, V. Kamalov, L. Jovanovski, V. Vusirikala, E. Mateo, Y. Inada, and T. Wang, "Field and lab experimental demonstration of nonlinear impairment compensation using neural networks", *Nature Communications*, vol. 10, no. 1, pp. 1–8, 2019.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[19] G. P. Agrawal, *Nonlinear Fiber Optics*, Fifth. Boston: Academic Press, 2013. DOI: https://doi.org/10.1016/B978-0-12-397023-7.00002-4. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123970237000024.

[20] V. Oliari, S. Goossens, C. Hager, G. Liga, R. M. Butler, M. van den Hout, S. van der Heide, H. D. Pfister, C. M. Okonkwo, and A. Alvarado, "Revisiting efficient multi-step nonlinearity compensation with machine learning: An experimental demonstration", *Journal of Lightwave Technology*, vol. 38, no. 12, pp. 3114–3124, 2020.

[21] M. Matsumoto and T. Nishimura, "Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator", *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 8, no. 1, pp. 3–30, 1998.

[22] L. Galdino, D. Semrau, D. Lavery, G. Saavedra, C. B. Czegledi, E. Agrell, R. I. Killey, and P. Bayvel, "On the limits of digital back-propagation in the presence of transceiver noise", *Optics Express*, vol. 25, no. 4, pp. 4564–4578, 2017.

[23] V. Oliari, E. Agrell, and A. Alvarado, "Regular perturbation on the group-velocity dispersion parameter for nonlinear fibre-optical communications", *Nature Communications*, vol. 11, no. 1, pp. 1–11, 2020.