



## Deep Learning of the Nonlinear Schrödinger Equation in Fiber-Optic Communications

Downloaded from: <https://research.chalmers.se>, 2025-02-19 03:38 UTC

Citation for the original published paper (version of record):

Häger, C., Pfister, H. (2018). Deep Learning of the Nonlinear Schrödinger Equation in Fiber-Optic Communications. IEEE International Symposium on Information Theory - Proceedings, 2018-June: 1590-1594. <http://dx.doi.org/10.1109/ISIT.2018.8437734>

N.B. When citing this work, cite the original published paper.

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, or reuse of any copyrighted component of this work in other works.

# Deep Learning of the Nonlinear Schrödinger Equation in Fiber-Optic Communications

Christian Häger<sup>\*†</sup> and Henry D. Pfister<sup>†</sup>

<sup>\*</sup>Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden

<sup>†</sup>Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina

**Abstract**—An important problem in fiber-optic communications is to invert the nonlinear Schrödinger equation in real time at the receiver to reverse the deterministic effects of the channel. Interestingly, the popular split-step Fourier method (SSFM) leads to a computation graph that is reminiscent of a deep feed-forward neural network. This observation allows one to leverage tools from machine learning to reduce complexity. In particular, the main disadvantage of the SSFM is that its complexity using  $M$  steps is at least  $M$  times larger than a linear equalizer. This is because the linear SSFM operator is a dense matrix. In previous work, truncation methods such as frequency sampling, wavelets, or least-squares have been used to obtain “cheaper” operators that can be implemented using filters. However, a large number of filter taps are typically required to limit truncation errors. For example, Ip and Kahn showed that for a 2000 km optical link, a truncated SSFM with 25 steps would require 70-tap filters in each step and 100 times more operations than linear equalization. We find that, by jointly optimizing all filters with deep learning, the complexity can be reduced significantly for similar accuracy. Using optimized 5-tap and 3-tap filters in an alternating fashion, one requires only around 2–6 times the complexity of linear equalization, depending on the implementation.

## I. INTRODUCTION

In a single-mode optical fiber, narrowband signals propagate according to the nonlinear Schrödinger equation (NLSE) [1, p. 40]. This is schematically illustrated in Fig. 1. In the absence of noise, the transmitted signal can thus be recovered by solving an initial value problem (IVP) using the received signal as a boundary condition. In practice, the received signal first passes through an analog-to-digital converter and the IVP can then be solved via receiver digital signal processing (DSP). This approach is referred to as digital backpropagation (DBP) and was inspired by a similar idea where optical components were used for the processing [2]. DBP was first studied as a transmitter pre-distortion technique [3], [4].

A major issue with DBP is the large computational burden associated with a real-time DSP implementation. Thus, various techniques have been proposed to reduce its complexity [5]–[13]. In essence, the task is to approximate the solution of a partial differential equation using as few computational resources as possible. We approach this problem from a

machine-learning perspective. In contrast to [7], [9], [10], we focus on deep learning and deep neural networks (NNs), which have attracted tremendous interest in recent years [14]. Our approach is to obtain a multi-layer computation graph similar to a deep NN by applying the split-step Fourier method (SSFM) [1]. This can be seen as an example of a more general methodology where domain knowledge is used to generate computation graphs with many layers, see, e.g., [15].

Deep NNs have achieved record-breaking performance for various tasks such as speech or object recognition [14]. In order to explain this success, the authors in [16] argue that most data of practical interest is generated by some form of hierarchical or Markov process, often obeying physical principles such as locality and symmetry. This makes it plausible that there exist efficient multi-layer computation graphs that can approximate these processes with few parameters. Our design choices are directly motivated by such considerations. In particular, our computation graph exploits the hierarchical problem structure that is introduced by the transmission process. Moreover, we choose the linear operators in the graph to be short and symmetric finite impulse response (FIR) filters.

This paper is a continuation of our work outlined in a recent summary paper [17]. It contains several novel contributions. Most importantly, we provide a theoretical justification for the deep learning approach. In particular, while FIR filter design for chromatic dispersion has been studied extensively in the past [5], [18]–[22], the designed filters have shown relatively poor efficiency when used in a split-step method for DBP due to truncation errors. Indeed, we argue that, for computational efficiency, the filters used in each step should be different and that they should be optimized *jointly*. We also compare our approach with multiple truncation methods (for the filter coefficients) and with “few-step” perturbation approaches.

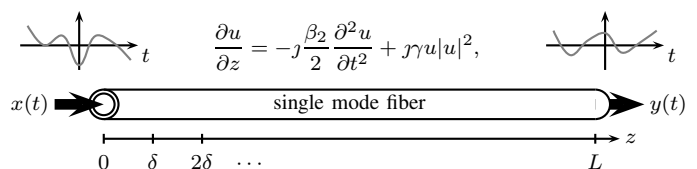


Fig. 1. Conceptual signal evolution in a single-mode fiber. The nonlinear Schrödinger equation implicitly describes the relationship between the input signal  $x(t) = u(z = 0, t)$  and the output signal  $y(t) = u(z = L, t)$ . The parameters  $\beta_2$  and  $\gamma$  are, respectively, the chromatic dispersion coefficient and the nonlinear Kerr parameter of the fiber. The loss term  $\alpha u/2$ , where  $\alpha$  is the attenuation parameter, is ignored for simplicity.

Author e-mails: christian.haeger@chalmers.se and henry.pfister@duke.edu. This work is part of a project that has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant No. 749798. The work was also supported by the National Science Foundation (NSF) under grant No. 1609327. Any opinions, findings, recommendations, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of these sponsors.

## II. DIGITAL BACKPROPAGATION

We assume that the signal  $x(t)$  is launched into an optical fiber where it propagates according to the NLSE as shown in Fig. 1. After distance  $z = L$ , the received signal  $y(t)$  is low-pass (LP) filtered and sampled at  $t = kT$  to give a sequence of samples  $\{y_k\}_{k \in \mathbb{Z}}$ . Our goal is to *efficiently* recover the signal  $x(t)$  (or a sampled version thereof) from  $\{y_k\}_{k \in \mathbb{Z}}$ .

### A. Split-Step Fourier Method

The popular SSFM is based on a block-wise receiver processing. To that end, assume that we collect  $n$  received samples into a vector  $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{C}^n$ . Consider now the time-discretized NLSE

$$\frac{d\mathbf{u}(z)}{dz} = \mathbf{A}\mathbf{u}(z) + \mathcal{J}\gamma\rho(\mathbf{u}(z)), \quad (1)$$

where  $\mathbf{A} = \mathbf{W}^{-1} \text{diag}(H_1, \dots, H_n)\mathbf{W}$ ,  $\mathbf{W}$  is the  $n \times n$  discrete Fourier transform (DFT) matrix,  $H_k = -j\frac{\beta_2}{2}\omega_k^2$ ,  $\omega_k = 2\pi f_k$  is the  $k$ -th DFT angular frequency, and  $\rho$  is defined as the element-wise application of  $\rho(x) = x|x|^2$ . To derive the SSFM, the fiber is conceptually divided into  $M$  segments of length  $\delta = L/M$ . Then, it is assumed that for sufficiently small  $\delta$ , the effects stemming from the two terms on the right hand side of (1) can be separated. More precisely, for  $\gamma = 0$ , (1) is linear with solution  $\mathbf{u}(z) = \mathbf{A}_z\mathbf{u}_0$ , where  $\mathbf{A}_z \triangleq e^{z\mathbf{A}}$ . For  $\beta_2 = 0$ , the solution is  $\mathbf{u}(z) = \sigma_z(\mathbf{u}_0)$ , where  $\sigma_z$  is the element-wise application of  $\sigma_z(x) = xe^{\mathcal{J}\gamma z|x|^2}$ . Alternating between these two operators for  $z = -\delta$  leads to the block diagram shown in the top part of Fig. 2.

The degree to which the obtained vector  $\mathbf{z}$  constitutes a good approximation of  $x(t)$  is now a question of choosing  $M$ ,  $T$ , and  $n$ . In practice,  $1/T$  is typically an integer multiple of the baud rate and  $n$  is chosen to minimize the overhead in overlap-and-save techniques for continuous data transmission. Increasing  $M$  leads to a more accurate approximation, but also increases complexity, as discussed in the next section.

### B. Implementation Complexity and Few-Step Approaches

Ignoring the complexity of the nonlinear steps, the SSFM can be implemented using  $M$  DFT/IDFT pairs, utilizing the fast Fourier transform. On the other hand, a linear equalizer can be implemented using a single DFT/IDFT pair. Based on this reasoning, the SSFM is at least  $M$  times more complex than linear equalization. This motivates a number of approaches that focus on reducing the number of steps, see, e.g., [6], [8], [11] and references therein.

## III. DEEP FEED-FORWARD NEURAL NETWORKS

Deep feed-forward NNs map an input vector  $\mathbf{a}$  to an output vector  $\mathbf{b}$  by alternating between affine transformations and pointwise nonlinearities [14], [16, Eq. (6)]. This is illustrated in the bottom part of Fig. 2. The matrices  $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(\ell)}$  and vectors  $\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(\ell)}$  are the network weights and biases, respectively, and  $\ell$  is the number of layers. The nonlinearities typically correspond to some activation function, e.g., the logistic or sigmoid function.

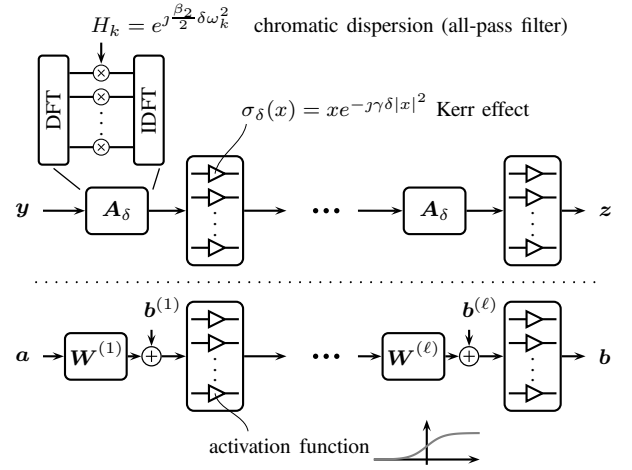


Fig. 2. Block diagram of the split-step Fourier method to numerically solve the nonlinear Schrödinger equation (top) and the canonical model of a deep feed-forward neural network (bottom).

While the similarity between the two computation graphs in Fig. 2 is apparent, there are, however, important differences. The one that is most relevant for this paper is the sparsity level of the linear operators. In order to be computationally efficient, deep NNs are typically designed to have very sparse weight matrices in most of the layers, whereas the linear propagation operator  $\mathbf{A}_\delta$  in the SSFM is a dense matrix.

*Remark 1.* In that regard, one may argue that (1) is a “computationally inefficient” time-discretization of the NLSE, in the sense that it relates local propagation changes to all time instances. A different time-discretization approach is via partial discretization or finite-difference methods. Indeed, finite-difference methods can be more computationally efficient than the SSFM in some applications [1, Sec. 2.4.2]. However, to the best of our knowledge, finite-difference methods have not been studied for real-time DBP. One reason for this might be that many methods that show good performance are implicit, i.e., they require solving a system of equations at each step. This makes it challenging to satisfy a real-time constraint.

## IV. FILTER DESIGN FOR CHROMATIC DISPERSION

According to the NLSE, chromatic dispersion acts as an all-pass filter with frequency response  $H(\omega) = e^{\mathcal{J}\kappa\omega^2}$ , where  $\kappa \triangleq -\beta_2\delta/(2T^2)$ ,  $\omega \triangleq 2\pi fT$ , and  $\delta$  is the transmission distance. Various approaches have been proposed to approximate this response (over a fixed bandwidth) with an FIR filter. For example, since the inverse Fourier transform of  $H(\omega)$  can be computed analytically, filter coefficients may be obtained through direct sampling and truncation [18]. Other approaches include frequency-domain sampling (FDS) [5], wavelets [20], and least-squares (LS) [21], [22].

### A. Parameter Efficiency in Split-Step Methods

Time-domain FIR filtering has been suggested for DBP in, e.g., [5], [12], [13], [19], [20]. In the SSFM, approximating  $H(\omega)$  with a short FIR filter can be interpreted as a truncation of  $\mathbf{A}_\delta$  to obtain a sparse banded matrix.

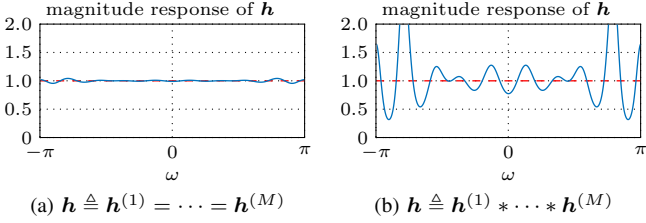


Fig. 3. Schematic illustration of the truncation error when using the same (or very similar) FIR filters in a split-step method, where  $*$  denotes convolution.

To estimate the required filter length, one may use the fact that chromatic dispersion leads to a group delay difference of  $2\pi\beta_2\Delta f\delta$  over a bandwidth  $\Delta f$  and distance  $\delta$ . Normalizing by the sampling interval  $T$ , this confines the memory to

$$K_{\text{cd}} = 2\pi\beta_2\Delta f\delta/T \quad (2)$$

samples. For example, we have  $\beta_2 = -21.668 \text{ ps}^2/\text{km}$ ,  $\delta = 80 \text{ km}$ , and  $1/T = 32.1 \text{ GHz}$  for the system studied in [5]. The receiver bandwidth is 32.1 GHz, but it is limited by an LP filter with 3-dB cutoff at 12.4 GHz. Thus, FIR filters with 4–12 taps should be sufficient. However, 70-tap filters are required to obtain acceptable accuracy using FDS [5]. Similar observation apply to the results in [12], [13], [19], [20], i.e., the required filter length is significantly longer than predicted by (2).

### B. Joint Filter Optimization

In previous work, a single filter or filter pair is designed and then used repeatedly in the SSFM. In this case, the truncation error accumulates coherently, leading to an undesired overall magnitude response as illustrated in Fig. 3. The effect is well known and a simple way to control it is by increasing the filter length. We propose instead to optimize *all*  $M$  filters jointly.

*Remark 2.* In [12], [13], filter coefficient quantization is studied for time-domain DBP. They highlight the effect of correlated quantization errors and propose random dithering [12] and co-optimization of quantization levels of filter pairs [13]. While this does not address the truncation error problem directly, it does alleviate it somewhat.

In this section, we illustrate how a joint filter optimization can be done in a way such that the problem admits a (possibly suboptimal) solution strategy via iteratively solving a set of weighted LS problems. This approach is simple and provides valuable insight into the problem. The optimized coefficients are then used as the initial starting point for the gradient-based deep learning approach discussed in the next section.

For simplicity, it is assumed that each of the  $M$  filters  $\mathbf{h}^{(i)} = (h_{-K}^{(i)}, \dots, h_0^{(i)}, \dots, h_K^{(i)})^\top$  for  $i = 1, \dots, M$  has  $2K + 1$  taps. The generalization to unequal filter lengths is straightforward. Let  $\mathcal{F}(\mathbf{h}^{(i)}) = \sum_{k=-K}^K h_k^{(i)} e^{-j\kappa\omega}$  be the discrete-time Fourier transform of  $\mathbf{h}^{(i)}$ . We use  $\mathcal{F}(\mathbf{h}^{(i)}) \stackrel{\circ}{=} e^{j\kappa\omega^2}$  to denote an objective, i.e., the symbol  $\stackrel{\circ}{=}$  may be interpreted as “should be close to”. The standard filter design uses the same objective for each of the  $M$  filters, i.e., each filter should approximate, as closely as possible, the chromatic dispersion transfer function  $e^{j\kappa\omega^2}$  over some frequency range.

In this case, one finds that all  $M$  filters should be the same. In particular, after discretizing the problem with  $\omega_i = 2\pi i/N$  for  $i = -N/2, \dots, N/2$ , one may use standard techniques to solve the linear LS problem  $\min_{\mathbf{h}^{(i)}} \|\mathbf{B}\mathbf{h}^{(i)} - \mathbf{d}\|^2$ , where  $\mathbf{d} = (d_{-N/2}, \dots, d_0, \dots, d_{N/2})^\top$  with  $d_i \triangleq e^{j\kappa\omega_i^2}$  and  $\mathbf{B}$  is an  $(N + 1) \times (2K + 1)$  DFT matrix.

On the other hand, by sacrificing some accuracy for the individual frequency responses, it may be possible to achieve a better combined response of neighboring filters and also a better overall response. This leads to the set of objectives

$$\begin{aligned} \mathcal{F}(\mathbf{h}^{(i)}) &\stackrel{\circ}{=} e^{j\kappa\omega^2}, & i = 1, 2, \dots, M \\ \mathcal{F}(\mathbf{h}^{(i)} * \mathbf{h}^{(i+1)}) &\stackrel{\circ}{=} e^{j2\kappa\omega^2}, & i = 1, 2, \dots, M - 1 \\ &\vdots \\ \mathcal{F}(\mathbf{h}^{(1)} * \dots * \mathbf{h}^{(M)}) &\stackrel{\circ}{=} e^{jM\kappa\omega^2}. \end{aligned} \quad (3)$$

Keeping the coefficients for all but one filter constant, (3) can be written as a standard weighted LS problem. Since, e.g.,  $\mathcal{F}(\mathbf{h}^{(i)} * \mathbf{h}^{(i+1)}) = \mathcal{F}(\mathbf{h}^{(i)})\mathcal{F}(\mathbf{h}^{(i+1)})$ , we have  $(\mathbf{B}\mathbf{h}^{(i)}) \circ (\mathbf{B}\mathbf{h}^{(i+1)})$  in the discretized problem, where  $\circ$  denotes element-wise multiplication. Hence, one obtains

$$\min_{\mathbf{h}^{(i)}} \sum_{j=1}^{O_i} \lambda_j \|(\mathbf{B}\mathbf{h}^{(i)}) \circ \mathbf{e}_j - \mathbf{d}_j\|^2, \quad (4)$$

where  $O_i$  is the number of objectives,  $\lambda_j > 0$  are weights,  $\mathbf{e}_j$  are constant vectors representing the influence of other filters and  $\mathbf{d}_j$  are the discretized objective vectors. A simple strategy for the joint optimization is then to solve (4) for each of the  $M$  filters in an iterative fashion. The weights  $\lambda_1, \dots, \lambda_{O_i}$  can be chosen based on a suitable system criterion.

We assume  $x(t) = \sum_{k=-\infty}^{\infty} x_k p(t - k/R_s)$ , where  $x_k \in \mathbb{C}$  are the data symbols,  $p(t)$  is the pulse shape, and  $R_s$  is the baud rate. For the block-wise processing, the estimated symbol vector  $\hat{\mathbf{x}}$  is obtained by passing  $\mathbf{z}$  (see Fig. 2) through a digital matched filter (MF) followed by a phase-offset rotation. The mean squared error  $\|\mathbf{x} - \hat{\mathbf{x}}\|^2$  is then used as a criterion to be minimized. Assuming that  $\|\mathbf{x}\|^2$  is constant for all  $\mathbf{x}$ , this is equivalent to maximizing the effective signal-to-noise ratio (SNR)  $\|\mathbf{x}\|^2 / \|\mathbf{x} - \hat{\mathbf{x}}\|^2$ .

### V. LEARNED DIGITAL BACKPROPAGATION

In [17], we have proposed to use deep learning for the joint filter optimization. The resulting method is referred to as learned DBP (LDBP). For LDBP, the computation graph of the SSFM is modified by interpreting all matrices  $\mathbf{A}_\delta$  as tunable parameters corresponding to the filters  $\mathbf{h}^{(1)}, \dots, \mathbf{h}^{(M)}$ , similar to the weight matrices in a deep NN. The nonlinearities are changed to  $\sigma^{(i)}: \mathbb{C}^n \rightarrow \mathbb{C}^n$  which act element-wise using  $\sigma^{(i)}(x) = x e^{-\gamma_i |x|^2}$ , where  $\gamma_i \in \mathbb{R}$  is a tunable parameter.

The computation graph including the MF and phase-offset rotation is implemented in TensorFlow. All parameters  $\theta = \{\mathbf{h}^{(1)}, \dots, \mathbf{h}^{(M)}, \gamma_1, \dots, \gamma_M\}$  are optimized by using many pairs  $(\mathbf{y}, \mathbf{x})$  of input and desired-output examples and adjusting the parameters such that the loss  $\|\mathbf{x} - \hat{\mathbf{x}}\|^2$  decreases. For this, we use the built-in *Adam* optimizer with a mini-batch size

of 30 and a fixed learning rate. To find a good starting point for the filter coefficients, we employ the LS method described in Sec. IV-B. While it is possible to use random starting points, we observe that a better final solution can be obtained with pre-optimized coefficients.

## VI. RESULTS AND DISCUSSION

We revisit the parameters in [5], using a different LP filter and transmit signals. Extensions to wavelength division multiplexing (WDM) systems and higher baud rates are discussed below. The optical link consists of 25 spans of 80 km fiber and an amplifier is inserted after each span to compensate for the signal attenuation. All parameters are summarized in Fig. 4.<sup>1</sup> Forward propagation is simulated with 6 samples/symbol using the SSFM with 50 steps per span (StPS).

LDBP is based on 1 StPS, alternating between 5-tap and 3-tap filters in the linear steps. The effective SNR after training is shown in Fig. 4 by the green line (triangles). As a reference, we show the performance of linear equalization (red) and DBP with 1 StPS using frequency-domain filtering (blue). The linear equalizer uses LS-optimal coefficients with constrained out-of-band gain (LS-CO) [22]. LDBP achieves a peak SNR of 21.9 dB using  $13 \cdot 4 + 12 \cdot 2 + 1 = 77$  total taps. After increasing the filter lengths to 7 and 5 (127 total taps), one obtains essentially the same peak SNR as frequency-domain filtering.

### A. Comparison to Other Truncation Methods

The performance of FDS (circles) and LS-CO (squares) is shown in Fig. 4 as a comparison. The same filter is used in each step and the length is chosen such that the peak SNR is around 22 dB. For this, 15-tap filters are required for FDS (351 total taps) and 9-tap filters for LS-CO (201 total taps). This is roughly 5 and 3 times more than required for LDBP.

In [5], 70-tap filters based on FDS are required for similar accuracy. This is likely due to the higher oversampling factor used (3 samples/symbol). While a higher oversampling factor may increase the maximum SNR achievable via DBP, it can also adversely affect the performance if truncation errors are taken into account. In general, it is difficult to predict how truncation errors affect the SNR in a nonlinear system.

### B. Complexity Compared to Linear Equalization

We use multiplications as a surrogate for complexity and assume that the exponential function is implemented with a look-up table, similar to [5]. For the nonlinear steps, one needs to square each sample (2 real multiplications), multiply by  $\gamma_i$ , and compute the phase rotation (4 real multiplications). This gives  $25 \cdot 7 = 175$  real multiplications per sample. For the linear steps, one has to account for 13 filters with 5 taps and 12 filters with 3 taps. All filters have symmetric coefficients and can be implemented using a folded structure with  $h_0$ -normalization as shown in Fig. 5. This gives  $39 \cdot 4 = 156$  real multiplications per sample. In comparison, the fractionally-spaced linear equalizer in [5] requires 188 real multiplications per data symbol operating at  $3/2$  samples/symbol. Thus,

<sup>1</sup>A Butterworth LP filter and QPSK modulation are assumed in [5].

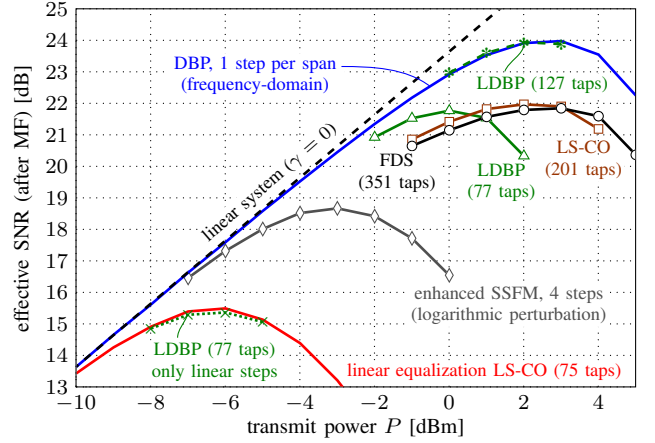


Fig. 4. Results for  $25 \times 80$  km single-mode fiber ( $\beta_2 = -21.668$  ps<sup>2</sup>/km,  $\gamma = 1.3$  1/W/km,  $\alpha = 0.2$  dB/km), 5 dB amplifier noise figure, Gaussian root-raised cosine pulses (0.1 roll-off),  $R_s = 10.7$  Gbaud, 2 samples/symbol ( $1/T = 21.4$  GHz), 15 GHz brick-wall low-pass filter,  $n = 2048$ . FDS: frequency-domain sampling, LS-CO: least-squares-optimal constrained out-of-band gain, LDBP: learned digital backpropagation, MF: matched filter.

LDBP requires 3.5 times more multiplications per symbol. For the same oversampling factor as LDBP, the linear equalizer has 75 taps (cf. (2) with  $\Delta f = 1.2 \cdot 10.7$  GHz). This leads to  $38 \cdot 4 = 152$  real multiplications with a folded implementation. Thus, LDBP requires around 2 times more operations. If the linear equalizer is implemented in the frequency domain, the number of real multiplication is reduced to  $n(4 \log_2 n + 4)/(n - 75) \approx 50$  per sample (see, e.g., [11, Sec. 4]), which increases the estimated complexity overhead factor to 6.

### C. Comparison to Few-Step Approaches

The enhanced SSFM (ESSFM) modifies the nonlinear step based on a logarithmic perturbation [11]. As a result, the sampled intensity waveform  $\{|y_k|^2\}_{k \in \mathbb{Z}}$  is filtered before applying the nonlinear phase shift. This gives the same functional form as previous approaches (e.g., [6], [8]), albeit with potentially different performance due to different choices or heuristics for the filter coefficients used in the modified nonlinear steps.

Excluding the overhead due to overlap-and-save techniques, one ESSFM step requires  $4 \log_2 n + 11 + N_c$  real multiplications per sample, where  $2N_c + 1$  is the filter length in the modified nonlinear steps [11, Sec. 4, single pol.]. We perform 4 ESSFM steps with  $N_c = 20$ , which gives roughly the same number of multiplications as LDBP. The filter coefficients are optimized from data as suggested in [11]. The performance is shown by the grey line (diamonds) in Fig. 4. The ESSFM achieves a smaller peak SNR by around 3 dB than LDBP.

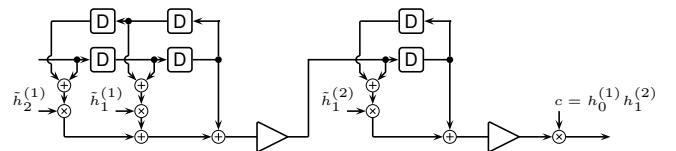


Fig. 5. Folded FIR filter implementation for two LDBP steps using  $h_0^{(i)}$ -normalization, i.e., we have  $\tilde{h}_j^{(i)} \triangleq h_j^{(i)}/h_0^{(i)}$ . The Kerr parameters in the nonlinearities have to be scaled accordingly.

The performance of only the linear steps in LDBP after training reverts approximately to that of the linear equalizer, as shown by the dotted green line (crosses) in Fig. 4. This leads to an intuitive interpretation of the task that is accomplished by deep learning. In particular, the optimized filter coefficients represent an approximate factorization of the overall linear inverse fiber response. At first, this may seem trivial because the linear matrix operator  $e^{LA}$  can be factored as  $e^{\delta A} \cdot \dots \cdot e^{\delta A}$  with  $L = \delta M$  for arbitrary  $M$  to represent shorter propagation distances. However, the factorization task becomes nontrivial if we also require the individual operators  $e^{\delta A}$  to be “cheap”, i.e., implementable using short filters.

*Remark 3.* We also experimented with factoring the  $z$ -transform polynomial of the 75-tap linear equalizer into a cascade of 3-tap filters. However, this gives no control over the individual filter responses, other than the choice of how to distribute the overall gain factor. Moreover, it is not obvious how to achieve a good ordering of sub-filters in the SSFM.

#### E. Wavelength Division Multiplexing and Higher Baud Rates

In a WDM system, the performance improvements of ideal single-channel (or few-channel) DBP are limited due to nonlinear interference from neighboring channels. This implies that it may be desirable to sacrifice some accuracy (i.e., target a lower effective SNR), and further simplify the design of LDBP, e.g., by pruning additional filter taps. A relaxed accuracy requirement also leaves some margin for practical impairments such as noise caused by filter coefficient quantization [12].

The memory introduced by chromatic dispersion increases quadratically with the considered bandwidth and linearly with the transmission distance, see (2). For longer links and/or higher baud rates, this seems to favor frequency-domain equalization (e.g., a DFT-based linear equalizer) over time-domain equalization in terms of complexity. On the other hand, the Kerr effect and its compensation are naturally described in the time domain. One possible approach to achieve a good performance–complexity trade-off is through digital sub-band processing. This entails a potential performance loss (due to possibly uncompensated sub-band interference), but it also reduces the effective system memory per sub-band. A closer investigation of this trade-off for LDBP is the subject of ongoing research.

## VII. CONCLUSION

We have considered the problem of reducing the complexity of DBP to facilitate a real-time DSP implementation. Our approach, called learned DBP (LDBP), is based on a multi-layer computation graph generated by the SSFM with many steps. Computational efficiency is achieved by using, in each step, very short and symmetric FIR filters that are jointly optimized with deep learning. Numerical results show that for a single-channel transmission scenario, LDBP can achieve a favorable performance–complexity trade-off compared to other filter design methods and perturbation-based “few-step” DBP.

- [1] G. P. Agrawal, *Nonlinear Fiber Optics*, 4th ed. Academic Press, 2006.
- [2] C. Paré, A. Villeneuve, P.-A. A. Bélanger, and N. J. Doran, “Compensating for dispersion and the nonlinear Kerr effect without phase conjugation,” *Optics Letters*, vol. 21, no. 7, pp. 459–461, 1996.
- [3] R.-J. Essiambre and P. J. Winzer, “Fibre nonlinearities in electronically pre-distorted transmission,” in *Proc. European Conf. Optical Communication (ECOC)*, Glasgow, UK, 2005.
- [4] K. Roberts, C. Li, L. Strawczynski, M. O’Sullivan, and I. Hardcastle, “Electronic precompensation of optical nonlinearity,” *IEEE Photon. Technol. Lett.*, vol. 18, no. 2, pp. 403–405, Jan. 2006.
- [5] E. Ip and J. M. Kahn, “Compensation of dispersion and nonlinear impairments using digital backpropagation,” *J. Lightw. Technol.*, vol. 26, pp. 3416–3425, Oct. 2008.
- [6] L. B. Du and A. J. Lowery, “Improved single channel backpropagation for intra-channel fiber nonlinearity compensation in long-haul optical communication systems,” *Opt. Express*, vol. 18, no. 16, pp. 17075–17088, Jul. 2010.
- [7] T. S. R. Shen and A. P. T. Lau, “Fiber nonlinearity compensation using extreme learning machine for DSP-based coherent communication systems,” in *Proc. Optoelectronics and Communications Conf. (OECC)*, Kaohsiung, Taiwan, 2011.
- [8] D. Rafique, M. Mussolin, M. Forzati, J. Mårtensson, M. N. Chughtai, and A. D. Ellis, “Compensation of intra-channel nonlinear fibre impairments using simplified digital back-propagation algorithm,” *Opt. Express*, vol. 19, no. 10, pp. 9453–9460, Apr. 2011.
- [9] A. M. Jarajreh, E. Giacoumidis, I. Aldaya, S. T. Le, A. Tsokanos, Z. Ghassemlooy, and N. J. Doran, “Artificial neural network nonlinear equalizer for coherent optical OFDM,” *IEEE Photon. Technol. Lett.*, vol. 27, no. 4, pp. 387–390, Feb. 2015.
- [10] E. Giacoumidis, S. T. Le, M. Ghanbarisabagh, M. McCarthy, I. Aldaya, S. Mhatli, M. A. Jarajreh, P. A. Haigh, N. J. Doran, A. D. Ellis, and B. J. Eggleton, “Fiber nonlinearity-induced penalty reduction in CO-OFDM by ANN-based nonlinear equalization,” *Opt. Lett.*, vol. 40, no. 21, pp. 5113–5116, Nov. 2015.
- [11] M. Secondini, S. Rommel, G. Meloni, F. Fresi, E. Forestieri, and L. Poti, “Single-step digital backpropagation for nonlinearity mitigation,” *Photon. Netw. Commun.*, vol. 31, no. 3, pp. 493–502, 2016.
- [12] C. Fougstedt, M. Mazur, L. Svensson, H. Eliasson, M. Karlsson, and P. Larsson-Edefors, “Time-domain digital back propagation: Algorithm and finite-precision implementation aspects,” in *Proc. Optical Fiber Communication Conf. (OFC)*, Los Angeles, CA, 2017.
- [13] C. Fougstedt, L. Svensson, M. Mazur, M. Karlsson, and P. Larsson-Edefors, “Finite-precision optimization of time-domain digital back propagation by inter-symbol interference minimization,” in *Proc. European Conf. Optical Communication (ECOC)*, Gothenburg, Sweden, 2017.
- [14] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [15] K. Gregor and Y. Lecun, “Learning fast approximations of sparse coding,” in *Proc. Int. Conf. Mach. Learning*, 2010.
- [16] H. W. Lin, M. Tegmark, and D. Rolnick, “Why does deep and cheap learning work so well?” *J. Stat. Phys.*, vol. 168, no. 6, pp. 1223–1247, 2017.
- [17] C. Häger and H. D. Pfister, “Nonlinear interference mitigation via deep neural networks,” in *Proc. Optical Fiber Communication Conf. (OFC), to be presented*, San Diego, CA, 2018.
- [18] S. J. Savory, “Digital filters for coherent optical receivers,” *Opt. Express*, vol. 16, no. 2, pp. 804–817, 2008.
- [19] L. Zhu, X. Li, E. Mateo, and G. Li, “Complementary fir filter pair for distributed impairment compensation of wdm fiber transmission,” *IEEE Photonics Technology Letters*, vol. 21, no. 5, pp. 292–294, 2009.
- [20] G. Goldfarb and G. Li, “Efficient backward-propagation using wavelet-based filtering for fiber backward-propagation,” *Opt. Express*, vol. 17, no. 11, pp. 814–816, May 2009.
- [21] A. Eghbali, H. Johansson, O. Gustafsson, and S. J. Savory, “Optimal least-squares FIR digital filters for compensation of chromatic dispersion in digital coherent optical receivers,” *J. Lightw. Technol.*, vol. 32, no. 8, pp. 1449–1456, Apr. 2014.
- [22] A. Sheikh, C. Fougstedt, A. Graell i Amat, P. Johansson, P. Larsson-Edefors, and M. Karlsson, “Dispersion compensation FIR filter with improved robustness to coefficient quantization errors,” *J. Lightw. Technol.*, vol. 34, no. 22, pp. 5110–5117, Nov. 2016.