



What Can Machine Learning Teach Us about Communications

Downloaded from: <https://research.chalmers.se>, 2021-04-16 22:55 UTC

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

Lian, M., Häger, C., Pfister, H. (2019)

What Can Machine Learning Teach Us about Communications

IEEE International Symposium on Information Theory - Proceedings, 15 January 2019

<http://dx.doi.org/10.1109/ITW.2018.8613331>

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

What Can Machine Learning Teach Us about Communications?

Mengke Lian*, Christian Häger*[†], and Henry D. Pfister*

*Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina

[†]Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden

Abstract—Rapid improvements in machine learning over the past decade are beginning to have far-reaching effects. For communications, engineers with limited domain expertise can now use off-the-shelf learning packages to design high-performance systems based on simulations. Prior to the current revolution in machine learning, the majority of communication engineers were quite aware that system parameters (such as filter coefficients) could be learned using stochastic gradient descent. It was not at all clear, however, that more complicated parts of the system architecture could be learned as well.

In this paper, we discuss the application of machine-learning techniques to two communications problems and focus on what can be learned from the resulting systems. We were pleasantly surprised that the observed gains in one example have a simple explanation that only became clear in hindsight. In essence, deep learning discovered a simple and effective strategy that had not been considered earlier.

I. INTRODUCTION

This paper outlines a few applications of machine-learning techniques to communication systems and focuses on what can be learned from the resulting systems. First, we consider the parameterized belief-propagation (BP) decoding of parity-check codes which was introduced by Nachmani et al. in [1]. Then, we study the low-complexity channel inversion known as digital backpropagation (DBP) for optical fiber communications [2].

II. MACHINE LEARNING

Before discussing the two applications in detail in Secs. III and IV, we start in this section by briefly reviewing the standard supervised learning setup for feed-forward neural networks. Afterwards, we highlight a few important aspects when applying machine learning to communications problems.

A. Supervised Learning of Neural Networks

A deep feed-forward NN with m layers defines a mapping $\mathbf{y} = f(\mathbf{x}; \theta)$ where the input vector $\mathbf{x} = \mathbf{x}^{(0)} \in \mathcal{X}$ is mapped to the output vector $\mathbf{y} = \mathbf{x}^{(m)} \in \mathcal{Y}$ by alternating between affine transformations (defined by $\mathbf{z}^{(i)} = \mathbf{W}^{(i)}\mathbf{x}^{(i-1)} + \mathbf{b}^{(i)}$) and pointwise nonlinearities (defined by $\mathbf{x}^{(i)} = \phi(\mathbf{z}^{(i)})$) [3]. This is illustrated in the bottom part of Fig. 3. The parameter vector θ encapsulates all elements in the weight

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.

matrices $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(m)}$ and all elements in the bias vectors $\mathbf{b}^{(1)}, \dots, \mathbf{b}^{(m)}$. Common choices for the nonlinearities include $\phi(z) = \max\{0, z\}$, $\phi(z) = \tanh(z)$, $\phi(z) = 1/(1 + e^{-z})$.

In a supervised learning setting, one has a training set $S \subset \mathcal{X} \times \mathcal{Y}$ containing a list of desired (\mathbf{x}, \mathbf{y}) input–output pairs. Then, training proceeds by minimizing the empirical training loss $\mathcal{L}_S(\theta)$, where the empirical loss $\mathcal{L}_A(\theta)$ for a finite set $A \subset \mathcal{X} \times \mathcal{Y}$ of input–output pairs is defined by

$$\mathcal{L}_A(\theta) \triangleq \frac{1}{|A|} \sum_{(\mathbf{x}, \mathbf{y}) \in A} L(f(\mathbf{x}; \theta), \mathbf{y})$$

and $L(\hat{\mathbf{y}}, \mathbf{y})$ is the loss associated with returning the output $\hat{\mathbf{y}}$ when \mathbf{y} is correct. When the training set is large, one typically chooses the parameter vector θ using a variant of stochastic gradient descent (SGD). In particular, mini-batch SGD uses the parameter update

$$\theta_{t+1} = \theta_t - \alpha \nabla \mathcal{L}_{B_t}(\theta_t),$$

where α is the step size and $B_t \subseteq S$ is the mini-batch used by the t -th step. Typically, B_t is chosen to be a random subset of S with some fixed size (e.g., $|B_t| = 64$) that matches available computational resources (e.g., GPUs).

B. Machine Learning for Communications

Machine learning for communications differs from traditional machine learning in a number of ways.

1) *Accurate generative modeling and infinite training data supply*: Machine learning is typically applied to fixed-size data sets, which are split into training and test sets. A central problem in this case is the *generalization error* caused by overfitting the model parameters to peculiarities in the training set. On the other hand, communication theory traditionally assumes that one can accurately simulate and/or model the communication channel. In this case, one can generate an infinite supply of training data with which to learn.

2) *Exponential number of classes*: For classification tasks, a different type of generalization error is caused by a lack of class diversity in the training set. For classical machine learning applications, there are typically only few classes and the training set contains a sufficient number of training examples per class. On the other hand, for certain communications problems, e.g., decoding error-correcting codes, the number of classes increases exponentially with the problem size. Training unrestricted NNs (even deep ones) with only a subset of classes leads to poor generalization performance [4].

3) Black-box computation graphs vs. domain knowledge:

Another consequence of having an accurate channel model is that one can actually implement optimal or close-to-optimal solutions in many cases. In that case, learning can be motivated as a means to reduce complexity because there may exist simple approximations with much lower complexity. Moreover, existing domain knowledge can be used to simplify the learning task. Indeed, for both considered applications in this paper, one actually improves existing algorithms by extensively parameterizing their associated computation graphs, rather than optimizing conventional “black-box” NN architectures. Our focus is on examining the trained solutions and trying to understand why they work better and solve the problem more efficiently than the hand-tuned algorithms they are based on.

III. OPTIMIZED BP DECODING OF CODES

Recently, Nachmani, Be’ery, and Burhstein proposed a weighted BP (WBP) decoder with different weights (or scale factors) for each edge in the Tanner graph [1]. These weights are then optimized empirically using tools and software from deep learning. One of the main advantages of this approach is that the decoder automatically respects both code and channel symmetry and requires many fewer training patterns to learn. Their results show that this approach provides moderate gains over standard BP when applied to the parity-check matrices of BCH codes. A more comprehensive treatment of this idea can be found in [5]. In addition, there are other less-restrictive NN decoders that also take advantage of code and channel symmetry [6], [7]

While the performance gains of WBP decoding are worth investigating, the additional complexity of storing and applying one weight per edge is significant. In our experiments, we also consider *simple scaling* models that share weights to reduce the storage and computational burden. In these models, three scalar parameters are used for each iteration: the message scaling, the channel scaling, and the damping factor. They can also be shared for all iterations.

A. Weighted Belief-Propagation Decoding

Consider an $[N, K]$ linear code \mathcal{C} defined by an $M \times N$ parity-check matrix H . Given any parity-check matrix H , one can construct a bipartite Tanner graph $\mathcal{G} = (V, C, E)$, where $V \triangleq [N]$ and $C \triangleq [M]$ are sets of variable nodes (i.e., code symbols) and check nodes (i.e., parity constraints). The edges, $E = \{(v, c) \in V \times C \mid H_{cv} \neq 0\}$, connect all parity checks to the variables involved in them. By convention, the boundary symbol ∂ denotes the neighborhood operator defined by

$$\partial v \triangleq \{c \mid (v, c) \in E\}, \quad \partial c \triangleq \{v \mid (v, c) \in E\}.$$

The log-likelihood ratio (LLR) is the standard message representation for BP decoding of binary variables. The initial channel LLR for variable node $v \in V$ is defined by

$$\ell_v \triangleq \log \left(\frac{\Pr(y_v \mid x_v = 0)}{\Pr(y_v \mid x_v = 1)} \right), \quad (1)$$

where y_v is the v -th symbol in the channel output sequence, and x_v is the corresponding bit in the transmitted codeword.

WBP is an iterative algorithm that passes messages along the edges of the Tanner graph \mathcal{G} . During the t -th iteration, a pair of messages $\hat{\lambda}_{c \rightarrow v}^{(t)}$, and $\lambda_{v \rightarrow c}^{(t)}$ are passed in each direction along the edge (v, c) . This occurs in two steps: the check-to-variable step updates messages $\hat{\lambda}^{(t)} \triangleq \{\hat{\lambda}_{c \rightarrow v}^{(t)}\}_{(v, c) \in E}$ and variable-to-check step updates messages $\lambda^{(t)} \triangleq \{\lambda_{v \rightarrow c}^{(t)}\}_{(v, c) \in E}$. In the variable-to-check step, the pre-update rule is

$$\lambda_{v \rightarrow c}^{(t)} = w_v^{(t)} \ell_v + \sum_{c' \in \partial v \setminus c} w_{v'c}^{(t)} \hat{\lambda}_{c' \rightarrow v}^{(t-1)}, \quad (2)$$

where $w_{vc}^{(t)}$ is the weight assigned to the edge (v, c) and $w_v^{(t)}$ is the weight assigned to the channel LLR ℓ_v . In the check-to-variable step, the pre-update rule is

$$\hat{\lambda}_{c \rightarrow v}^{(t)} = 2 \tanh^{-1} \left(\prod_{v' \in \partial c \setminus v} \tanh \left(\frac{\lambda_{v' \rightarrow c}^{(t)}}{2} \right) \right). \quad (3)$$

To avoid numerical issues, the absolute value of $\hat{\lambda}_{c \rightarrow v}^{(t)}$ is clipped, if it is larger than some fixed value (e.g., 15).

To mitigate oscillation and enhance convergence, we also use a *damping* coefficient $\gamma \in [0, 1]$ to complete the message updates [8]. We note that this method of improving performance was not considered in [1]. In particular, the final BP messages at iteration t are computed using a convex combination of the previous value and the pre-update value:

$$\lambda_{v \rightarrow c}^{(t)} = (1 - \gamma) \lambda_{v \rightarrow c}^{(t-1)} + \gamma \lambda_{v \rightarrow c}^{(t)}, \quad (4)$$

$$\hat{\lambda}_{c \rightarrow v}^{(t)} = (1 - \gamma) \hat{\lambda}_{c \rightarrow v}^{(t-1)} + \gamma \hat{\lambda}_{c \rightarrow v}^{(t)}. \quad (5)$$

For the marginalization step, the sigmoid function $\sigma(x) \equiv (1 + e^{-x})^{-1}$ is used to map the output LLR to an estimate $o_v^{(t)}$ of the probability that $x_v = 0$ defined by

$$o_v^{(t)} = \sigma \left(w_v^{(t)} \ell_v + \sum_{v' \in \partial c} w_{v'c}^{(t)} \hat{\lambda}_{c \rightarrow v'}^{(t)} \right). \quad (6)$$

Setting all weights to 1 and $\gamma = 1$ recovers standard BP.

B. From WBP to Optimized WBP

Any iterative algorithm, such as WBP decoding, can be “unrolled” to give a feed-forward architecture that has the same output for some fixed number of iterations [9]. Moreover, the sections in the feed-forward architecture are not required to be identical. This increases the number “trainable” parameters that can be optimized.

It is well-known that BP performs exact marginalization when the Tanner graph is a tree, but good codes typically have loopy Tanner graph with short cycles. To improve the BP performance on short high-density parity-check (HDPC) codes, one can optimize the weights $w_{vc}^{(t)}$ and $w_v^{(t)}$ in all iterations [1]. The damping coefficient γ can also be optimized.

For supervised classification problems, one typically uses the cross-entropy loss function, and this loss function has also been proposed for the optimized WBP decoding problem [1]. However, our experiments show that minimizing this loss may

not actually minimize the bit error rate. Instead, we use the modified loss function

$$L_s(\mathbf{o}^{(T)}, \mathbf{x}) \triangleq \frac{1}{N} \sum_{v=1}^N \left[1 + \left(\frac{o_v^{(T)}}{1 - o_v^{(T)}} \right)^{1-2x_v} \right]^{-1}, \quad (7)$$

where T is the total number of iterations. More details about the modified loss can be found in [10]. Our experiments also show that the optimization behaves better with the multi-loss approach proposed by [1]. Thus, the results in this paper are based on optimizing the modified multi-loss function

$$L(\{\mathbf{o}^{(t)}\}_{t=1}^T, \mathbf{x}) \triangleq \frac{1}{T} \sum_{t=1}^T L_s(\mathbf{o}^{(t)}, \mathbf{x}). \quad (8)$$

The optimization complexity depends on the number of iterations and how the parameters are shared. For example, one can share the weights temporally (across decoding iterations) and/or spatially (across edges):

- If the weights are shared temporally, i.e.,

$$w_v^{(t)} \equiv w_v, \quad w_{v_c}^{(t)} \equiv w_{v_c}, \quad \forall t,$$

one obtains a recursive NN (RNN) structure.

- If the weights are shared spatially, i.e.,

$$w_v^{(t)} \equiv w_{\text{ch}}^{(t)}, \quad w_{v_c}^{(t)} \equiv w_{\text{msg}}^{(t)}, \quad \forall (v, c) \in E,$$

then there are only two scalar parameters per iteration: one for the channel LLR and one for the BP messages. Compared to the fully weighted (FW) decoder, we call this the simple scaled (SS) decoder.

- Sharing weights both temporally and spatially results in only two weight parameters, w_{ch} and w_{msg} .

C. Random Redundant Decoding (RRD)

A straightforward way to improve BP decoding for HDPC codes is to use redundant parity checks (e.g., by adding dual codewords as rows to the parity-check matrix) [11]. In general, however, the complexity of BP decoding scales linearly with the number of rows in the parity-check matrix.

Another approach is to spread these different parity checks over time, i.e., by using different parity-check matrices in each iteration [12]–[14]. This can be implemented efficiently by exploiting the code’s automorphism group and reordering the code bits after each iteration in a way that effectively uses many different parity-check matrices but stores only one.

In [5], optimized weighted RRD decoders are constructed by cascading several WBP blocks and reordering the code bits after each WBP block. In this work, we also consider optimized RRD decoding based on their approach. But, the input to $(\tau+1)$ -th learned BP block is modified to be a convex combination between the initial channel LLRs and the output of the τ -th learned BP. This procedure is similar to damping and the *mixing* coefficient β is also learned.

For RRD decoding, choosing a good parity-check matrix is crucial because the code automorphisms permute the variable nodes without changing the structure of the Tanner graph. In general, good Tanner graphs have fewer short cycles and can be constructed with heuristic cycle-reduction algorithms [12].

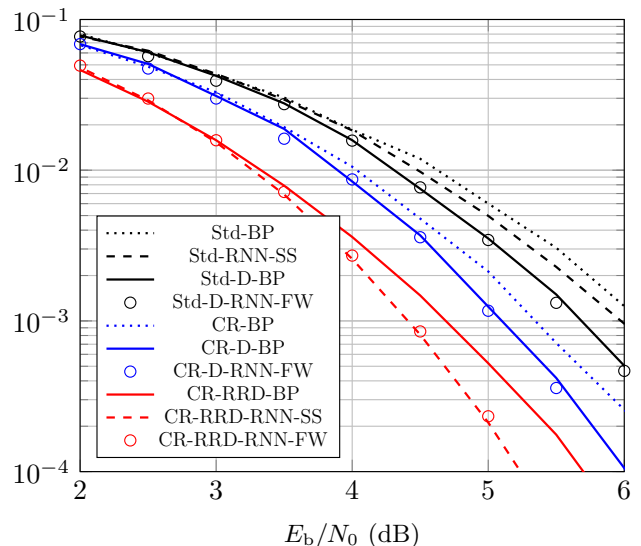


Fig. 1. BER results for the BCH(63, 36) code. Curves are labeled to indicate: whether the parity-check matrix is standard (Std) or cycle reduced (CR), whether damping (D) is used, and also to show the decoder style (BP/RNN-SS/RNN-FW).

D. Experiments and Results

The various feed-forward decoding architectures in this paper are implemented in the PyTorch framework and optimized using the RMSPROP optimizer. The number of decoding iterations is set to $T = 20$. For the RRD algorithm, the code bits are permuted after every second decoding iteration and optimized (iteration-independent) mixing and damping coefficients are used. The decoder architectures are trained using transmit-receive pairs for the binary-input AWGN channel where the E_b/N_0 SNR parameters is chosen uniformly between 1 dB and 6 dB for each training pair. To avoid numerical issues, the gradient clipping threshold is set to 0.1 and the LLR clipping threshold is 15. We define each epoch to be 1000 mini-batches and each mini-batch to be 100 transmit-receive pairs. All decoders are trained for 20 epochs and optimized using the multi-loss function (8).

In Fig. 1, we show the performance curves achieved by the optimized decoders for the BCH(63, 36) code. For the standard parity-check matrix without RRD, the standard BP decoder with damping (Std-D-BP) performs very similarly to the FW optimized decoder (Std-D-RNN-FW). Similarly, for the cycle-reduced parity-check matrix, damping (CR-D-BP) achieves essentially the same gain as the fully-weighted model (CR-S-RNN-FW). Thus, the dominant effects are fully explained by using damping and cycle-reduced parity-check matrices.

For a similar complexity, the RRD algorithm achieves better results. This is true both for standard BP (CR-RRD-BP) with optimized mixing and damping and for optimized weights (CR-RRD-RNN-SS) in the simple-scaling model. However, the fully-weighted model (CR-RRD-RNN-FW) does not provide significant gains over simple scaling. Also, RRD results are shown only for cycle-reduced matrices because they perform much better.

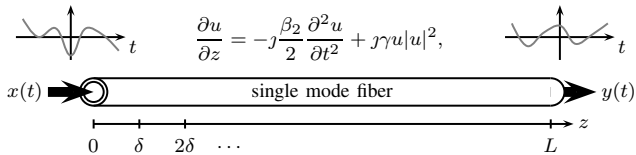


Fig. 2. Conceptual signal evolution in a single-mode fiber. The nonlinear Schrödinger equation implicitly describes the relationship between the input signal $x(t) = u(z = 0, t)$ and the output signal $y(t) = u(z = L, t)$. The parameters β_2 and γ are, respectively, the chromatic dispersion coefficient and the nonlinear Kerr parameter.

IV. MACHINE LEARNING FOR FIBER-OPTIC SYSTEMS

In this section, we discuss the application of machine learning techniques to optical-fiber communications.

A. Signal Propagation and Digital Backpropagation

Fiber-optic communication links carry virtually all inter-continental data traffic and are often referred to as the *Internet backbone*. We consider a simple point-to-point scenario, where a signal with complex baseband representation $x(t)$ is launched into an optical fiber as illustrated in Fig. 2. The signal evolution is implicitly described by the nonlinear Schrödinger equation (NLSE) which captures dispersive and nonlinear propagation impairments [15, p. 40]. After distance $z = L$, the received signal $y(t)$ is low-pass filtered and sampled at $t = kT$ to give the samples $\{y_k\}_{k \in \mathbb{Z}}$.

In the absence of noise, the NLSE is invertible and the transmitted signal can be recovered by solving the NLSE in the reverse propagation direction. This approach is referred to as digital backpropagation (DBP) in the literature. DBP requires a numerical method to solve the NLSE and a widely studied method is the split-step Fourier method (SSFM). The SSFM conceptually divides the fiber into M segments of length $\delta = L/M$ and it is assumed that for sufficiently small δ , the dispersive and nonlinear effects act independently. A block diagram of the SSFM for DBP is shown in the top part of Fig. 3, where $\mathbf{y} = (y_1, \dots, y_n)^\top$. In particular, one alternates between a linear operator \mathbf{A}_δ and the element-wise application of a nonlinear phase-shift function $\sigma_\delta(x) = x e^{j\gamma\delta|x|^2}$. Assuming a sufficiently high sampling rate, the obtained vector \mathbf{z} converges to a sampled version of $x(t)$ as $M \rightarrow \infty$. By comparing the two computation graphs in Fig. 3, one can see that the SSFM has a naturally layered or hierarchical Markov structure, similar to a deep feed-forward NN.

B. Parameter-Efficient Learned Digital Backpropagation

A major issue with DBP is the large computational burden associated with a real-time implementation. Despite significant efforts to reduce complexity (see, e.g., [2], [16], [17]), DBP based on the SSFM is not used in any current optical system that we know. Instead, only linear equalizers are employed. Their implementation already poses a significant challenge; with data rates exceeding 100 Gbit/s, linear equalization of chromatic dispersion is typically one of the most power-hungry receiver blocks [18].

Note that the linear propagation operator \mathbf{A}_δ in the SSFM is a dense matrix. On the other hand, deep NNs are typically

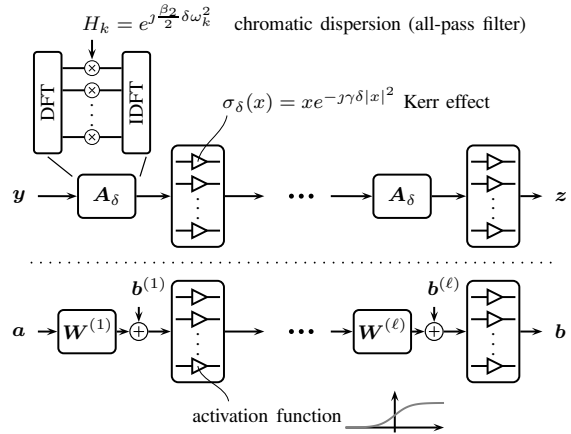


Fig. 3. Block diagram of the split-step Fourier method to numerically solve the nonlinear Schrödinger equation (top) and the canonical model of a deep feed-forward neural network (bottom).

designed to have very sparse weight matrices in most of the layers to achieve computational efficiency. Sparsification of \mathbf{A}_δ can be achieved by switching from a frequency-domain to a time-domain filtering approach using finite-impulse response (FIR) filters. The main challenge in that case is to find short FIR filters in each SSFM step that approximate well the ideal chromatic dispersion all-pass frequency response. In previous work, the general approach is to design either a single filter or filter pair and use is repeatedly in each step [2], [19]–[21]. However, this typically leads to poor parameter efficiency (i.e., it requires relatively long filters) because truncation errors pile up coherently. We have shown in [22], [23] that this truncation error problem can be controlled effectively by performing a *joint* optimization of all filter coefficients in the entire DBP algorithm. In particular, the computation graph of the SSFM is optimized via SGD by simply interpreting all matrices \mathbf{A}_δ as tunable parameters corresponding to the FIR filters, similar to the weight matrices in a deep NN. The nonlinearities are left unchanged, i.e., they correspond to the nonlinear phase-shift functions in the original SSFM and not to a traditional NN activation function. The resulting method is referred to as learned DBP (LDBP).

C. Optimization Results

In Fig. 4, we compare the equalizer accuracy in terms of the effective SNR of LDBP to the conventional approach of designing a single FIR filter (either via least-squares fitting or frequency-domain sampling) and then using it repeatedly in the SSFM. LDBP requires significantly fewer total filter taps (indicated in brackets) to achieve similar or better peak accuracy. The obtained FIR filters are as short as 5 or 3 (symmetric) taps per step, leading to very simple and efficient hardware implementation. This is confirmed by recent ASIC synthesis results which show that the power consumption of LDBP becomes comparable to linear equalization [24]. LDBP can also be extended to subband processing to enable low-complexity DBP for multi-channel or other wideband transmission scenarios [25].

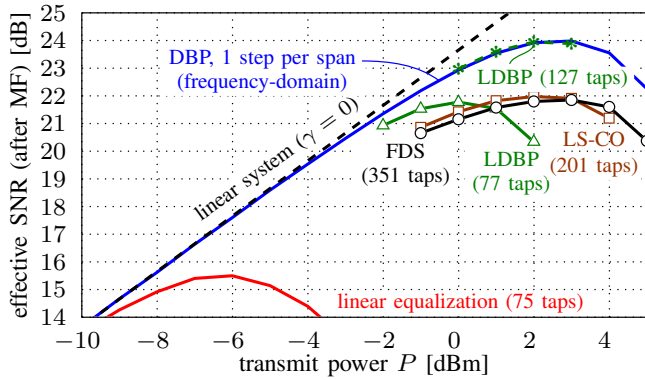


Fig. 4. Results for 25 spans of 80 km single-mode fiber (see parameters in [23]). FDS: frequency-domain sampling, LS-CO: least-squares-optimal constrained out-of-band gain, MF: matched filter.

At first glance, the obtained results in Fig. 4 may be somewhat counterintuitive. Indeed, after examining the optimized individual (per-step) filter responses in LDBP, we found that they are generally worse approximations to the ideal chromatic dispersion frequency response compared to filters obtained by least-squares fitting or other methods. However, the combined response of neighboring filters and also the overall response is better compared to the conventional strategy of using the same filter in each dispersion compensation stage. In fact, using the same filter many times in series magnifies any weakness. By using different filters at each stage, the problem is avoided and shorter filters can achieve the same performance.

V. CONCLUSION

Recent progress in machine learning and off-the-shelf learning packages have made it tractable to add many parameters to existing communication algorithms and optimize. In this paper, we have reviewed this approach with the help of two applications.

For the decoding application, our experiments support the observations in [1], [5] that optimizing parameterized BP decoders can provide meaningful gains. In addition, we observed that many fewer parameters (e.g., damping alone) may be sufficient to achieve very similar gains. Thus, for this general approach, it can be fruitful to also minimize the parameterization necessary to achieve the same gain [10].

For the digital backpropagation application, we were pleasantly surprised that, after analyzing the learned solution, we were able to understand why it worked so well. In essence, deep learning discovered a simple and effective strategy that had not been considered earlier.

REFERENCES

- [1] E. Nachmani, Y. Be'ery, and D. Burshtein, "Learning to decode linear codes using deep learning," in *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Illinois, USA, 2016.
- [2] E. Ip and J. M. Kahn, "Compensation of dispersion and nonlinear impairments using digital backpropagation," *J. Lightw. Technol.*, vol. 26, no. 20, pp. 3416–3425, Oct. 2008.
- [3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

- [4] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink, "On deep learning-based channel decoding," in *Proc. Annual Conf. Information Sciences and Systems (CISS)*, 2017.
- [5] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, and Y. Be'ery, "Deep learning methods for improved decoding of linear codes," *IEEE J. Sel. Topics Signal Proc.*, vol. 12, no. 1, pp. 119–131, Feb. 2018.
- [6] L. G. Tallini and P. Cull, "Neural nets for decoding error-correcting codes," in *Proc. IEEE Technical Applications Conf. and Workshops*, Portland, OR, 1995.
- [7] A. Bennatan, Y. Choukroun, and P. Kisilev, "Deep learning for decoding of linear codes - a syndrome-based approach," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Vail, CO, 2018.
- [8] M. Fossorier, R. Palanki, and J. Yedidia, "Iterative decoding of multi-step majority logic decodable codes," in *Proc. Int. Symp. on Turbo Codes & Iterative Inform. Proc.*, 2003, pp. 125–132.
- [9] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Intl. Conf. on Mach. Learn.*, 2010, pp. 399–406.
- [10] M. Lian, F. Carpi, C. Häger, and H. D. Pfister, "Learned belief-propagation decoding with damping and redundant parity-check matrices," 2018, to be submitted to ISIT 2019.
- [11] J. S. Yedidia, J. Chen, and M. P. Fossorier, "Generating code representations suitable for belief propagation decoding," in *Proc. Annual Allerton Conf. on Commun., Control, and Comp.*, vol. 40, no. 1, 2002, pp. 447–456.
- [12] T. R. Halford and K. M. Chugg, "Random redundant soft-in soft-out decoding of linear block codes," in *Proc. IEEE Int. Symp. Inform. Theory*. IEEE, 2006, pp. 2230–2234.
- [13] I. Dimnik and Y. Be'ery, "Improved random redundant iterative HDPC decoding," *IEEE Trans. Commun.*, vol. 57, no. 7, 2009.
- [14] T. Hehn, J. B. Huber, O. Milenkovic, and S. Laendner, "Multiple-bases belief-propagation decoding of high-density cyclic codes," *IEEE Trans. Commun.*, vol. 58, no. 1, pp. 1–8, 2010.
- [15] G. P. Agrawal, *Nonlinear Fiber Optics*, 4th ed. Academic Press, 2006.
- [16] 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, July 2010.
- [17] 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, April 2011.
- [18] B. S. G. Pillai, B. Sedighi, K. Guan, N. P. Anthapadmanabhan, W. Shieh, K. J. Hinton, and R. S. Tucker, "End-to-end energy modeling and analysis of long-haul coherent transmission systems," *J. Lightw. Technol.*, vol. 32, no. 18, pp. 3093–3111, 2014.
- [19] L. Zhu, X. Li, E. Mateo, and G. Li, "Complementary FIR filter pair for distributed impairment compensation of WDM fiber transmission," *IEEE Photon. Technol. Lett.*, vol. 21, no. 5, pp. 292–294, March 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] 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.
- [22] C. Häger and H. D. Pfister, "Nonlinear interference mitigation via deep neural networks," in *Proc. Optical Fiber Communication Conf. (OFC)*, San Diego, CA, 2018.
- [23] —, "Deep learning of the nonlinear Schrödinger equation in fiber-optic communications," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Vail, CO, 2018.
- [24] C. Fougstedt, C. Häger, L. Svensson, H. D. Pfister, and P. Larsson-Edefors, "ASIC implementation of time-domain digital backpropagation with deep-learned chromatic dispersion filters," in *Proc. European Conf. Optical Communication (ECOC)*, Rome, Italy, 2018.
- [25] C. Häger and H. D. Pfister, "Wideband time-domain digital backpropagation via subband processing and deep learning," in *Proc. European Conf. Optical Communication (ECOC)*, Rome, Italy, 2018.