This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

Model-aware Deep Learning Method for Raman Amplification in Few-Mode Fibers

Gianluca Marcon, Andrea Galtarossa, *Fellow, IEEE/OSA,* Luca Palmieri, *Member, IEEE/OSA,* and Marco Santagiustina, *Member, IEEE*

***Abstract*—One of the most promising solutions to overcome the capacity limit of current optical fiber links is space-division multiplexing, which allows the transmission on various cores of multi-core fibers or modes of few-mode fibers. In order to realize such systems, suitable optical fiber amplifiers must be designed. In single mode fibers, Raman amplification has shown significant advantages over doped fiber amplifiers due to its low-noise and spectral flexibility. For these reasons, its use in next-generation space-division multiplexing transmission systems is being studied extensively. In this work, we propose a deep learning method that uses automatic differentiation to embed a complete few-mode Raman amplification model in the training process of a neural network to identify the optimal pump wavelengths and power allocation scheme to design both flat and tilted gain profiles. Compared to other machine learning methods, the proposed technique allows to train the neural network on ideal gain profiles, removing the need to compute a dataset that accurately covers the space of Raman gains we are interested in. The ability to directly target a selected region of the space of possible gains allows the method to be easily generalized to any type of Raman gain profiles, while also being more robust when increasing the number of pumps, modes, and the amplification bandwidth. This approach is tested on a 70 km long 4-mode fiber transmitting over the C+L band with various numbers of Raman pumps in the counter-propagating scheme, targeting gain profiles with an average gain in the interval from 5 dB to 15 dB and total tilt in the interval from -1.425 dB to 1.425 dB. We achieve wavelength- and mode-dependent gain fluctuations lower than 0.04 dB and 0.02 dB per dB of gain, respectively.**

***Index Terms*—Space-division multiplexing, Raman amplification, deep learning.**

I. Introduction

NONLINEAR phenomena arising in optical fibers impose an intrinsic limit to their information capacity [1], [2]. During the last three decades, the demand in internet traffic increased exponentially with an annual rate of 40%, while current technologies are rapidly approaching the nonlinear Shannon limit (NSL) of single-mode fibers (SMFs) [3]. In order to avoid bringing the existing optical fiber infrastructure to a "capacity crunch" [4], space-division multiplexing (SDM) has been proposed as the key technology for future lightwave systems operating beyond the NSL [5].

A promising approach to implement SDM is to exploit the spatial diversity of modes in few-mode fibers (FMFs) to transmit independent data streams, so realizing mode-division multiplexing (MDM) [6]. In order to benefit from the added capacity of spatially-multiplexed transmissions, suitable network devices must be designed fully compatible with the already well-established techniques such as wavelength-division multiplexing (WDM). To this end, the role of SDM-compatible amplifiers is of fundalmental importance, with several experimental works demonstrating the effectiveness in MDM scenarios of both erbium-doped fiber amplifiers (EDFAs) [7], [8] and Raman amplifiers (RAs) [9], [10].

The compensation of link losses with minimal signal-to-noise ratio (SNR) reduction has always been a crucial aspect in optical communications, but additional care must be taken with SDM systems to minimize both mode-dependent gain (MDG) and wavelength-dependent gain (WDG), as they can be both detrimental to the multiple-input multiple-output (MIMO) digital signal processing (DSP) algorithms that mitigate the effect of mode-crosstalk to correctly recover the transmitted signals [11].

While the simplicity and power efficiency of EDFAs made them appealing for commercial communication systems, their reduced gain bandwidth has made Raman amplification an attractive solution for wideband WDM schemes [12]. The spectral flexibility of RAs, together with suitable optmization techniques, enables the design of flat gain profiles over large bandwidths by means of multiple wavelength pumps [12]. In the context of SDM, the additional degrees of freedom can lead to higher control of WDG and MDG [13]. Additionally, RAs can offer distributed amplification, resulting in a reduced noise contribution compared to EDFAs [14].

In SMF systems, different approaches have been followed to correctly determine the pump parameters required to obtain pre-determined gain profiles. A recent publication [15] proposed a machine learning (ML) technique to solve this problem. Specifically, a neural network (NN) can be trained to learn the inverse relationship $\mathbf{y} = f^{-1}(\mathbf{G})$ between the vector $\mathbf{y}$ of pump wavelengths and powers and the corresponding gain profile $\mathbf{G}$, using a synthetic dataset $\mathcal{D} = \{(\mathbf{y}_i, \mathbf{G}_i)\}$ of thousands of gain curves generated with random pump parameters. The learned mapping is then used to compute the required pump parameters $\tilde{\mathbf{y}} = \widetilde{f^{-1}}(\mathbf{G}_{target})$ to approximate a given target gain profile. This eliminates the need to solve complex iterative algorithms that require multiple integrations of the propagation equations for every new target profile, making Raman amplification suitable for its application in next-generation self-adaptive and autonomous optical networks, where low-latency automatization is fundamental

Authors are with the Department of Information Engineering, University of Padova, Padova, Italy. A.G., L.P., and M.S. are also with CNIT – National Inter-University Consortium for Telecommunications, Italy. e-mail: gianluca.marcon@dei.unipd.it

Manuscript received October 27, 2020; revised October 27, 2020.

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

[15]. The authors of [15] used two additional techniques to refine the prediction of the NN. The first is *model-averaging*, which consists in training several NNs in parallel, each on a random permutation of the dataset, and finally averaging their output. This approach, while providing some improvements, is significantly heavier in terms of computational time, both for the training and the inference phase. This aspect, together with the memory requirements needed to store hundreds of trained models, could pose a challenge to network controllers, where computational power may be limited. The second technique consists in a fine-tuning phase requiring an additional NN trained to learn the direct mapping $\mathbf{G} = f(\mathbf{y})$. The prediction error on the gain profile obtained with the approximate pump parameters $\tilde{\mathbf{y}}$ is estimated using the learned direct mapping $\tilde{f}$ and minimized using an iterative gradient-descent algorithm without integrating the propagation equations. Publication [15] showed promising results, demonstrating the feasibility of the method with flat and tilted gain profiles using a counter-propagating RA over the C and C+L bands, achieving a maximum prediction error on the considered gain profiles well below 1 dB for different levels of amplifications.

In the context of MDM, a similar approach to design flat gain profiles both for 2-mode and 4-mode fibers has been demostrated in [16]. This work does not use neither model-averaging nor fine-tuning algorithms; therefore, memory and time requirements for both the training and the inference phase are substantially cut. For the 4-mode FMF, [16] showed encouraging results in terms of MDG and gain flatness; nevertheless, the analysis is limited to the C band only.

The main drawback of both methods with respect to iterative optimization algorithms is that while the latter specifically look at minimizing a cost function $\mathcal{C}(\mathbf{G}_{target}, \tilde{\mathbf{G}})$ between a desired and predicted gain profile by taking the propagation model into account, the former is instead optimized to minimize a cost function $\mathcal{C}(\mathbf{y}, \tilde{\mathbf{y}})$ between pump parameters. The NNs are thus unaware of the underlying mathematical and physical relations between pump parameters and gain profile, which has to be learned from the available data. In order to approximate the inverse function $\mathbf{y} = f^{-1}(\mathbf{G})$ using a NN and generate flat gain profiles, the region of space of approximately flat Raman gains, must be properly sampled. This cannot be easily achieved since the training dataset is generated by solving the Raman equations with randomly drawn pump powers and wavelengths, meaning that only the codomain of $f^{-1}(\cdot)$ is sampled with full control. As a result, only a minor part of the generated gains fall inside the region of interest, resulting in the NNs being trained to learn the inverse function on a much bigger domain than required, potentially hindering its performance on flat/tilted gains. This aspect is also more problematic when increasing the amplification bandwidth or the number of modes and pumps, as the dimensionality of the space to explore also increases. The choice of parameters for the generation of the dataset is also critical for the effectiveness of the methods presented in [15], [16]. For example, the powers and wavelengths of the pumps are selected a priori, which requires preliminary supervision and that can finally mean that the trained NNs might not able to predict the optimal pump parameters.

Owing to automatic differentiation (AD) techniques [17], analytical or numerical models describing dynamical systems can be embedded in ML architectures [18]. By recording the series of elementary operations performed on the model input in a computational graph, AD libraries such as PyTorch [19] can compute the exact derivatives of the model output with respect to any parameter to be optimized [20]. In the context of optical communications, this approach has been demonstrated to be able to perform end-to-end (E2E) optimization of a intensity modulation/direct detection system by jointly optimizing the transmitter and receiver using NNs, outperforming classical feed-forward equalization [21]. The effectiveness of this technique has also been demonstrated for coherent transmissions [22] where probabilistic constellation shaping and geometric constellation shaping have shown to be fundamental for achieving record spectral efficiencies in short- and long-haul experiments [23].

In this work we propose an unsupervised ML method which employs AD to embed a differentiable FMF Raman amplification model in the training procedure of a NN to predict the pump parameters able to generate flat and tilted gain profiles over a pre-determined range of amplification levels and gain tilts. The trained NN can then be used to obtain the required pump parameters for a desired gain profile with low time-complexity. The presented method has the advantage to train the NN directly on the searched (e.g. flat and tilted) gain profiles, thereby directly sampling the selected region of space of possible gains, instead of building a dataset by solving the Raman equations using random pump parameters. The supervised dataset design phase, along with the issues related to it, is thus completely avoided, with the relationship between target gain and pump parameters being learned in the training phase of the NN through the differentiable Raman model. The ability to directly target an arbitrary region of the space of Raman gains makes this method easily generalizable to any type of gain profiles, more robust and scalable with respect to the changes in number of modes, Raman pumps, and fiber parameters. For all these reasons this unsupervised method is expected to be more useful in self-adaptive networks. This method is validated on different 4-mode fibers using a counter-propagating scheme with various numbers of Raman pumps, up to 8, predicting the required pump powers and wavelengths to generate gain profiles on the C+L band with average gain and tilt in the interval from 5 dB to 15 dB and from -1.425 dB to 1.425 dB, respectively; results show MDG and gain flatness comparable to those reported in [16], but on a larger bandwidth and quantifying the advantage of higher number of pumps also in terms of the reached root-mean-square error (RMSE).

II. Proposed Method

A. Multi-mode Raman amplifer equations

In a few-mode RA supporting M modes, N_s signal wavelengths and N_p pump wavelengths, the power evolution of the i-th frequency propagating in the m-th mode is described by the following set of nonlinear ordinary differential equations [24], [25]:

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

$$\begin{aligned}\eta_i \frac{dP_i^m}{dz} =& -\alpha_i P_i^m \\ &+ P_i^m \sum_{j=i+1}^{N_s+N_p} \sum_{n=1}^{M} \mathcal{I}_{m,n} g_R(|f_i - f_j|) P_j^n \\ &- P_i^m \sum_{j=1}^{i-1} \sum_{n=1}^{M} \frac{f_i}{f_j} \mathcal{I}_{m,n} g_R(|f_i - f_j|) P_j^n,\end{aligned} \tag{1}$$

where P_i^m is the power in the m-th mode and i-th frequency, where $i \in \{1, \ldots, N_s + N_p\}$, $m \in \{1, \ldots, M\}$, and the frequencies f_i are assumed to be sorted in ascending order; α_i is the attenuation coefficient at the i-th frequency, $g_R(\Delta f)$ is the Raman gain coefficient for the frequency difference Δf, and $\mathcal{I}_{m,n}$ is the overlap integral between mode m and mode n, defined by

$$\mathcal{I}_{m,n} = \frac{\displaystyle\iint_{-\infty}^{+\infty} I_m(x,y) I_n(x,y) dxdy}{\displaystyle\iint_{-\infty}^{+\infty} I_m(x,y) dxdy \iint_{-\infty}^{+\infty} I_n(x,y) dxdy}, \tag{2}$$

where $I_k(x,y)$ is the intensity profile of the k-th mode. The overlap integrals are assumed to be wavelength independent. Finally, η_i determines the relative propagation direction of the i-th frequency, so for the counter-propagating pumps $\eta_i = -1$, $\forall i \in \{N_s + 1, \ldots, N_s + N_p\}$, whereas $\eta_i = 1$ for the first N_s frequencies.

Modes with similar propagation constants, i.e. those within the same mode group, exhibit high coupling efficiency, resulting in the equalization of the amplifier gain for that particular group. For the purpose of RA they can consequently be treated as a unique mode [25], [26]. Conversely, linear mode coupling between different mode groups is weak and will be neglected here, like in [16], [25], [26].

For a fiber of length L, the Raman on-off gain $\mathbf{G} = [G_i^m]$ of the amplifier is defined by

$$G_i^m = \frac{P_i^m(z = L) \text{ with pumps turned on}}{P_i^m(z = L) \text{ with pumps turned off}}, \tag{3}$$

where $i = 1, \ldots, N_s$ and $m = 1, \ldots, M$.

B. Deep Learning Model Architecture

Many of the E2E learning methods in the literature are based on a deep learning (DL) architecture called autoencoder (AE) [27]. An AE is composed of two main blocks: an encoder,

$$\mathcal{E}(\,\cdot\,; \theta_e): \quad \mathbb{R}^p \to \mathbb{R}^q, \tag{4}$$

and a decoder,

$$\mathcal{D}(\,\cdot\,; \theta_d): \quad \mathbb{R}^q \to \mathbb{R}^p, \tag{5}$$

where θ_e, θ_d are learnable parameters and $q < p$. The role of the encoder is to learn a lower dimensionality representation $\hat{\mathbf{x}}$ of its input data $\mathbf{x}$ in a way that enables the decoder to compute an estimate $\tilde{\mathbf{x}}$ of the original data from $\hat{\mathbf{x}}$:

$$\tilde{\mathbf{x}} = \mathcal{D}(\mathcal{E}(\mathbf{x}; \theta_e); \theta_d). \tag{6}$$

Typically, both $\mathcal{E}$ and $\mathcal{D}$ consist in NNs that are jointly trained to minimize the average of the cost function $\mathcal{C}_{\text{AE}}$ between original and reconstructed samples of a dataset $\mathcal{X} = \{\mathbf{x}_i\}$:

$$\theta_e^*, \theta_d^* = \underset{\theta_e, \theta_d}{\operatorname{argmin}} \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \mathcal{C}_{\text{AE}}\big(\mathcal{D}(\mathcal{E}(\mathbf{x}; \theta_e); \theta_d), \mathbf{x}\big). \tag{7}$$

By replacing the decoder $\mathcal{D}$ with a differentiable Raman model $\mathcal{R}$ that maps a vector of pump powers and wavelengths

$$\mathbf{y} = [\boldsymbol{\lambda} \mid \mathbf{P}] \in \mathbb{R}_+^{(M+1)N_p}, \tag{8}$$

to the corresponding on-off gain, we can train the AE using (7) on a dataset $\mathcal{X} = \{\mathbf{G}_i\}$ of gain curves to force the encoder NN to learn a low-dimensional representation that minimizes the reconstruction error through $\mathcal{R}$. That is, the trained encoder approximates the inverse of the Raman model

$$\mathcal{E}(\,\cdot\,; \theta_e^*) \approx \mathcal{R}^{-1}(\,\cdot\,), \tag{9}$$

meaning that the lower dimensionality representation of the input gain $\mathbf{G}$ is the vector $\mathbf{y}$ of pump powers and wavelengths that approximates it.

While the numerical integration of the Raman model $\mathcal{R}$ is still required in the forward-pass of the training process to compute (7), this computational cost is no longer needed to determine the pump parameters that approximate a target gain profile, which are directly obtained by using $\mathcal{E}(\,\cdot\,; \theta_e^*)$.

In this work, the encoder $\mathcal{E}$ is a feed-forward (FF) NN with N_h hidden, fully connected (FC) layers of N_n neurons and rectified linear unit (ReLU) activation functions. Input and ouput layers have size $N_s \times M$ and $N_p \times (M+1)$, respectively.

In order to force a constraint on the predicted pump parameter vector $\mathbf{y}$, a sigmoidal function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{10}$$

is used to limit the output $\mathbf{x}$ of the last FC layer of the NN to the open interval $(0, 1)$. The resulting normalized pump vector $\hat{\mathbf{y}}$ can then be linearly mapped to the desired interval of powers and wavelengths.

The decoder $\mathcal{R}$ consists of a fixed-step, fourth-order Runge-Kutta integrator that solves (1) to compute the on-off gain using the pump parameters generated by the encoder.

C. Training algorithm

The optimal encoder parameters, θ_e^*, are found by solving (7) with an iterative training algorithm and using the RMSE between target and approximated gain as a cost function:

$$\mathcal{C}_{\text{AE}}(\mathbf{G}, \tilde{\mathbf{G}}) = \frac{1}{M} \sum_{m=1}^{M} \underset{i}{\text{RMSE}} \left(G_i^m, \tilde{G}_i^m\right), \tag{11}$$

for $i = 1, \ldots, N_s$. In the k-th iteration of the training algorithm, the AE reconstruction of each curve in the dataset $\mathcal{X}$ is computed as

$$\tilde{\mathbf{G}} = \mathcal{E}(\mathcal{R}(\mathbf{G}); \theta_e(k)) \quad \forall \mathbf{G} \in \mathcal{X}, \tag{12}$$

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

where $\theta_e(k)$ are the encoder parameters at the current iteration. The total cost function for the iteration is then evaluated by averaging (11) over $\mathcal{X}$

$$\mathcal{C}(k) = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{G} \in \mathcal{C}} \mathcal{C}_{\text{AE}}(\mathbf{G}, \tilde{\mathbf{G}}). \tag{13}$$

Finally, the encoder parameters are updated with a gradient descent algorithm

$$\theta_e(k+1) = \theta_e(k) - \epsilon \nabla_{\theta_e(k)} \mathcal{C}(k), \tag{14}$$

where $\epsilon > 0$ is the learning rate (LR) of the algorithm. The exact computation of the gradients is performed by means of AD and backpropagation [27]. Advanced optimization algorithms such as the adaptive moment estimation (Adam) algorithm [28] are typically employed for the update step (14) as they offer robust convergence properties and adaptive LR schemes for each parameter.

During training, the relationship between input target gains and their respective pump powers and wavelengths are learned through the differentiable Raman solver $\mathcal{R}$, meaning the vector of pump parameters $\mathbf{y}_i$ associated to each gain profile $\mathbf{G}_i$ of the dataset is not needed. This fact can be exploited by completely bypassing the dataset generation phase and training the encoder on the targeted family of desired ideal gain profiles.

In this paper we focus on flat and tilted gains, so in each training iteration k we generate a batch $\mathcal{B}_k = \{\mathbf{G}_i\}_{i=1}^{B}$ of B ideal gain profiles with average gain level l_G and tilt t_G (gain variation per unit wavelength) randomly sampled from a uniform distribution

$$l_G \sim \mathcal{U}\left(l_G^{\min}, l_G^{\max}\right), \tag{15}$$

$$t_G \sim \mathcal{U}\left(t_G^{\min}, t_G^{\max}\right), \tag{16}$$

It is important to notice that this approach is completely generalizable and not limited to flat and tilted gains only, but it could be extended to other families of gain profiles by properly including them in the training data.

As detailed in section I, in supervised learning techniques such as those presented in [15] and [16], the underlying physical model is only described by and learned from the provided data, meaning that it is essential to use datasets that are representative of the problem. In the context of RA, this means that the dataset must properly sample the region of possible Raman gains containing approximately flat gain profiles in order for the NN to properly learn the inverse Raman model. This cannot be done efficiently or easily, as there is actually no direct control on which gain profiles are sampled, but rather on the power and wavelength of each pump. Instead, the presented approach avoids this issue by directly sampling the selected space of Raman gains. Consequently, the problem of overfitting is completely avoided, and regularization techniques are not required.

D. Initial conditions

When training the AE using the algorithm described above we face the problem of local minima, which is common when dealing with the optimization of many parameters with complex cost functions. An important aspect to consider when dealing with local minima is the initial conditions of the algorithm, which can significantly affect the outcome of the optimization problem.

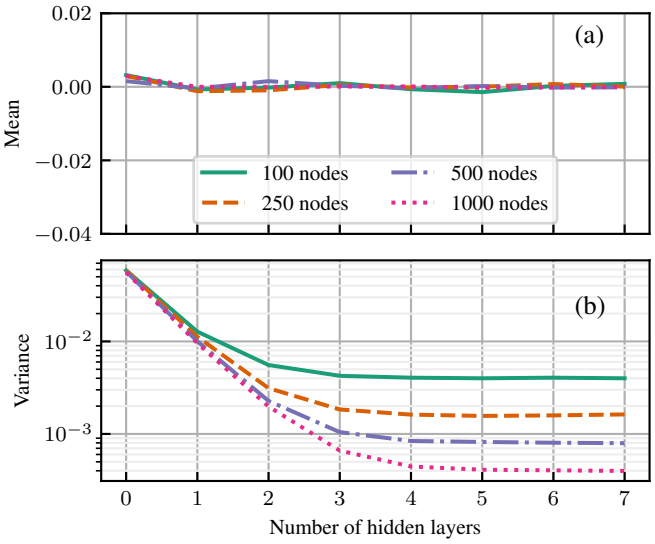

Fig. 1. Mean (a) and variance (b) of the output of the NN during the first training iteration, as a function of the number of hidden layers and for different values of neurons per layer.

The parameters of the encoder's FC layers are initialized by sampling a uniform distribution on the interval $[-\sqrt{n}, \sqrt{n}]$, where n is the inverse of the number of incoming connections to that layer [19]. This approach has been demostrated to be effective to mitigate the problem of *vanishing gradients* when training multi-layer NNs [29]. While this random initialization strategy is beneficial in classic supervised learning models, it affects the initial condition of our AE, as it imposes a random value to the initial normalized pump parameter vector $\hat{\mathbf{y}}_\mathbf{0}$. We analyzed the statistical distribution of the output of the last FC layer, $\mathbf{x_0}$, during the first training iteration for different number of hidden layers and neurons. We found that its elements follow a Gaussian-like distribution with zero mean and a variance that decreases as the number of layers and neurons increases. In Fig. 1 (a) and (b) we show the mean and variance, respectively, of $\mathbf{x_0}$ for the case of a 4-mode fiber with 50 wavelength channels and 8 pumps, resulting into an input layer of 200 neurons and an output layer of 40 neurons. For a sufficiently high number of hidden layers and neurons (which is easily met in practice) we can then use the approximation $\mathbf{x_0} \approx \mathbf{0}$, meaning that by (10) in the first training iteration $\hat{\mathbf{y}}_\mathbf{0} \approx \sigma(\mathbf{0}) = \mathbf{0.5}$, so fixing the initial pump powers and wavelengths to the middle point of the interval of allowed values. By introducing a centering vector $\boldsymbol{\alpha}$ and subtracting it to the input of the sigmoids, we have:

$$\hat{\mathbf{y}} = \sigma(\mathbf{x} - \boldsymbol{\alpha}), \tag{17}$$

which enables us to relate the initial pump parameters to $\boldsymbol{\alpha}$ as follows

$$\hat{\mathbf{y}}_0 = \sigma(\mathbf{x}_0 - \boldsymbol{\alpha}) \approx \sigma(-\boldsymbol{\alpha}) = \frac{1}{1 + e^{\boldsymbol{\alpha}}}. \tag{18}$$

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

We can use this result to force a more desirable initial condition on the pump parameters by computing the appropriate value of $\boldsymbol{\alpha}$ by inverting (18).

E. Counter-propagating pumps

For the case of counter-propagating pumps it is customary to implement a differential equation solver based on a shooting algorithm to determine the correct initial pumps powers to be injected at $z = L$. This however would require significantly more computational resources, as the propagation equation should be solved several times for each training sample and, more importantly, could introduce convergence problems [30].

However, the method proposed here presents a particularly advantageous feature on this regard: in fact, the encoder $\mathcal{E}$ can direcly predict the pumps powers at $z = 0$, $\tilde{P}_i^m(z = 0)$, eliminating the need to employ shooting algorithms. By solving (1) with initial ($z = 0$) conditions for pumps and signals, we obtain the predicted gain $\tilde{\mathbf{G}}$ along with pumps powers at the end of the link, $\tilde{P}_i^m(z = L)$, which are the values of interest. We therefore trade a significant computation cost in the training phase for a single integration of (1) in the inference phase. The resulting AE-based system is represented in the diagram of Fig. 2, highlighting the various components of the architecture and its input-output relations. Green boxes and arrows are related to the training phase of the AE, during which the encoder parameters θ_e are optimized. In order to compensate the significantly higher sensitivity of the predicted gain to the optimization parameters and avoid further problems with local minima, we introduce a modification to the training algorithm by multiplying the output of the last FC layer $\mathbf{x}$ by a mask $\mathbf{H}_k$, where the subscript k indicates the k-th training iteration. The normalized pump parameters for the k-th training iteration are then determined by:

$$\hat{\mathbf{y}}_k = \sigma(\mathbf{x}_k \odot \mathbf{H}_k - \boldsymbol{\alpha}), \tag{19}$$

where $\odot$ indicates the Hadamard (element-by-element) product. $\mathbf{H}_k$ can be suitably designed to "steer" the NN by weighting the computed gradients of the cost function with respect to the pump parameters during backpropagation. In our case, we set:

$$\mathbf{H}_k = [H_k^i] = \begin{cases} 0 & 1 \le i \le N_p \ , \ k < K \\ 1 & N_p + 1 \le i \le (M+1)N_p \ , \ k < K \\ 1 & 1 \le i \le (M+1)N_p \ , \ k \ge K \ , \end{cases} \tag{20}$$

where the superscript i indicates the i-th element of the vector. Using this definition, the pump wavelengths are fixed to their initial conditions for the first K iterations, allowing the encoder to learn just the relationship between predicted pump power and generated gain profile, which is more critical during the first training iterations. The training algorithm is summarized in Algorithm 1 and is completely implemented using the PyTorch DL library [19], which enables us to leverage AD and graphics processing unit (GPU) acceleration.

Algorithm 1 AE training algorithm

Compute centering vector $\boldsymbol{\alpha}$
Initialize encoder parameters: $\theta_e(0)$
for $k = 0$ **to** $N_{\text{iter}} - 1$ **do**
  Compute the mask $\mathbf{H}_k$
  Generate batch $\mathcal{B}_k = \{\mathbf{G}_i\}_{i=1}^B$ of gain profiles
  Propagate batch to obtain the pump parameters from NN: $\hat{\mathbf{Y}}_k = \{\hat{\mathbf{y}}_k^i\} = \{\sigma(\hat{\mathbf{x}}_k^i \odot \mathbf{H}_k - \boldsymbol{\alpha})\}$, $\hat{\mathbf{x}}_k^i = NN(\mathbf{G}_i; \theta_e(k))$
  Map the normalized parameters to the selected range: $\mathbf{Y}_k = \{\mathbf{y}_k^i\} = \text{Scale}(\hat{\mathbf{Y}}_k)$
  Integrate (1) to compute the predicted gain profiles: $\tilde{\mathcal{B}_k} = \{\tilde{\mathbf{G}}_i^k\} = \{\mathcal{R}(\mathbf{y}_k^i)\}$
  Compute the cost function $\mathcal{C}(k)$ using (13)
  Compute gradients with backpropagation: $\nabla_{\theta_e(k)}\mathcal{C}(k)$
  Update the parameters $\theta_e(k+1)$ using (14)
end for

TABLE I
OVERLAP INTEGRALS OF THE FMFS USED FOR SIMULATION, IN UNITS OF 1×10^9 m^{-2}.

	FMF1				FMF2			
	LP_{01}	LP_{11}	LP_{02}	LP_{21}	LP_{01}	LP_{11}	LP_{02}	LP_{21}
LP_{01}	6.24	4.12	4.62	2.85	5.47	3.6	3.87	2.45
LP_{11}	4.12	4.36	2.33	3.81	3.6	5.7	1.95	3.28
LP_{02}	4.62	2.33	6.15	2.12	3.87	1.95	4.94	1.76
LP_{21}	2.85	3.81	2.12	3.88	2.45	3.28	1.76	4.95

III. RESULTS AND VALIDATION

We test the presented method using counter-propagating pumps and a $L = 70\,\text{km}$ long 4-mode step-index fiber (SIF) whose overlap integrals are calculated in [25] and reported in Table I. Hereinafter, we refer to this fiber as FMF1. The Raman gain spectrum is computed using the multi-vibrational-mode model of the Raman response function for silica fibers [31], whereas the peak value for the Raman gain coefficient $g_R = 7 \times 10^{-14}\,\text{W}^{-1}\,\text{m}$ was used [14]. The spectral attenuation coefficient of the fiber is obtained from a parabolic fit of attenuation data of a commercially available SMF, $\alpha(\lambda) = \alpha_0 + \alpha_1\lambda + \alpha_2\lambda^2$, with coefficients $\alpha_0 = 5.788\,\text{dB km}^{-1}$, $\alpha_1 = -7.1246 \times 10^{-3}\,\text{dB km}^{-1}\,\text{nm}^{-1}$, $\alpha_2 = 2.268 \times 10^{-6}\,\text{dB km}^{-1}\,\text{nm}^{-2}$. This approximation is valid for the wavelengths interval from 1385 nm to 1625 nm. As in [24], [25] we assume the absence of mode-dependent losses (MDL).

We consider the transmission on $N_s = 50$ wavelengths on the C+L band, for a total number of spatial and wavelength channels equal to $N_{ch} = M \times N_s = 200$. The input power for each channel is set to $P_{ch} = -10$ dBm.

The encoder NN is composed of $N_h = 5$ hidden layers of $N_n = 1000$ neurons each, and its parameters are optimized using the Adam algorithm with a LR $\epsilon = 1 \times 10^{-4}$. The AE is trained for $N_{\text{iter}} = 1000$ iterations with batches of $B = 1024$ gain curves, which are sufficient to fill the GPU random access memory (RAM) and ensure 100% GPU clock utilization. Each batch is generated according to the strategy described in section II, with average gain level and tilt uniformly sampled

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

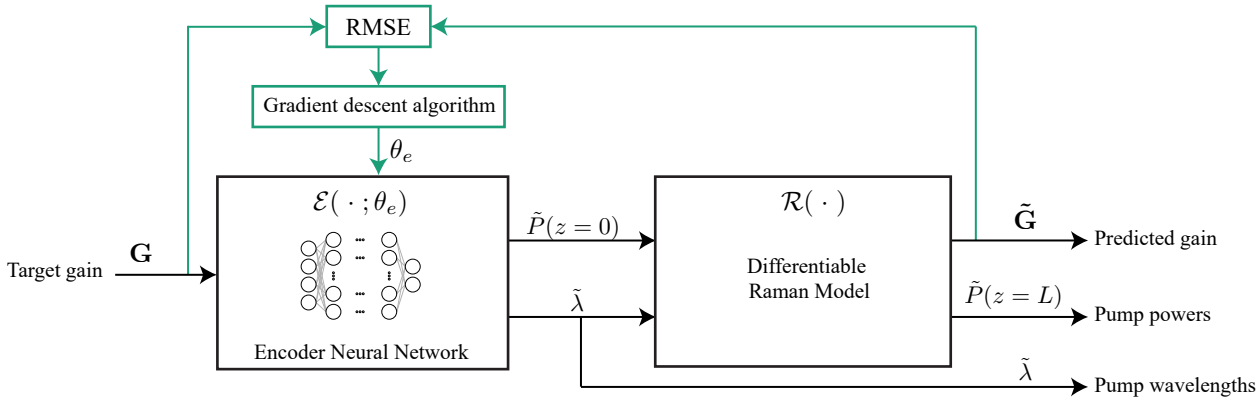

Fig. 2. Diagram of the AE architecture for the design of the Raman gain profile in FMFs. Green arrows and boxes are related to the training phase of the AE.

from the intervals of 5 dB to 15 dB and $-0.015\,\mathrm{dB\,nm^{-1}}$ to $0.015\,\mathrm{dB\,nm^{-1}}$, respectively.

We map the output of the sigmoids to limit the predicted power at $z = 0$ and wavelength of each pump into the intervals $I_{P(z=0)} = [-60, 20]$ dBm and $I_\lambda = [1410, 1520]$ nm, respectively.

Using (18) we set the initial power on each pump to $P_0(z = 0) = 3$ dBm, whereas the wavelengths are uniformly distributed over I_λ. Additionally, we use (20) to fix the pumps wavelengths to their initial value for the first $K = 100$ iterations.

Once the AE is trained, the encoder is used determine pump wavelengths and powers at $z = 0$ to approximate a given target gain profile:

$$\tilde{\mathbf{y}} = [\tilde{\boldsymbol{\lambda}} \mid \tilde{\mathbf{P}}(z=0)] = \mathcal{E}(\mathbf{G}\,;\,\theta_e^*), \tag{21}$$

and the corresponding predicted gain $\tilde{\mathbf{G}}$ and pumps powers at $z = L$ are obtained with a single integration of the Raman equations (1):

$$[\tilde{\mathbf{G}} \mid \tilde{\mathbf{P}}(z=L)] = \mathcal{R}(\tilde{\mathbf{y}}). \tag{22}$$

The total training time for the employed NNs is approximately 45 minutes using an NVIDIA Quadro M4000 GPU. The computational time to perform a prediction for a single target gain profile on an Intel consumer laptop CPU is approximately 11 ms, of which 1 ms is required for computing the output of the encoder NN, and the remaining 10 ms are needed for integrating (1).

A. Flat gain profiles

First, we asses the performance of the presented method using FMF1 for the case of flat target gain profiles in terms of RMSE, gain flatness and MDG, for different levels of amplification and varying the number of Raman pumps. Given that the number of pumps determines the size of the input and output layers of the encoder NN, the training algorithm must be run for every value that this parameter assumes. For each target curve, we obtain the corresponding AE prediction using (21), (22) and compute the RMSE for each mode m as:

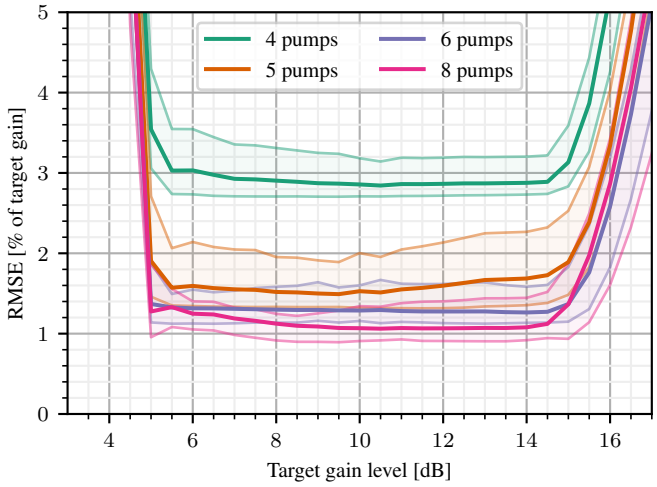

Fig. 3. RMSE as a function of the target gain level, for different number of pumps. Solid lines represent the mean RMSE over the 4 modes, whereas shaded areas indicate the total RMSE variation over the modes.

$$\mathrm{RMSE}_m(\mathbf{G}, \tilde{\mathbf{G}}) = \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s}\left(G_i^m - \tilde{G}_i^m\right)^2}. \tag{23}$$

In Figure 3 we report the RMSE in terms of percentage of the target gain level, as a function of the amplification level and using 4, 5, 6, and 8 Raman pumps. Solid lines and shaded regions represent the average RMSE and maximum to minimum RMSE variation over the modes, respectively. For gain levels inside the target interval of $[5, 15]$ dB, the RMSE curves are almost constant, independently of the number of pumps used. Conversely, the RMSE rapidly grows outside the training interval, as the encoder NN is not able to extrapolate the correct pump parameters. By increasing the number of Raman pumps from 4 to 8 we improve the RMSE, going from 3% to about 1% of the target gain.

A clear picture on the improvements brought by an increased number of pumps is given by the gain flatness or

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

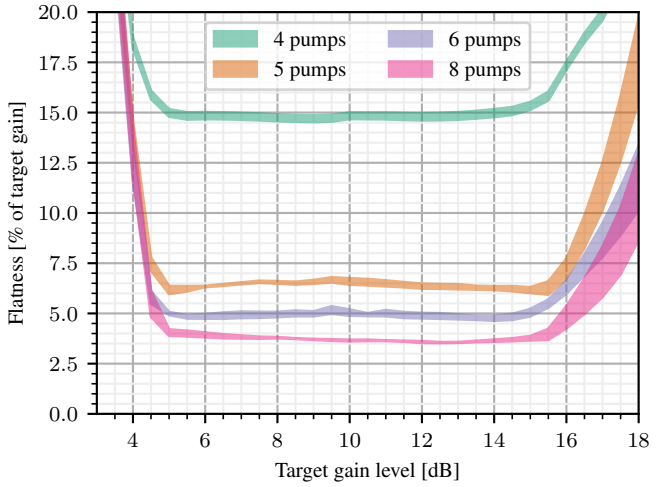

Fig. 4. Gain flatness variation along the modes of FMF1, as a function of the target gain level, using different number of pumps.

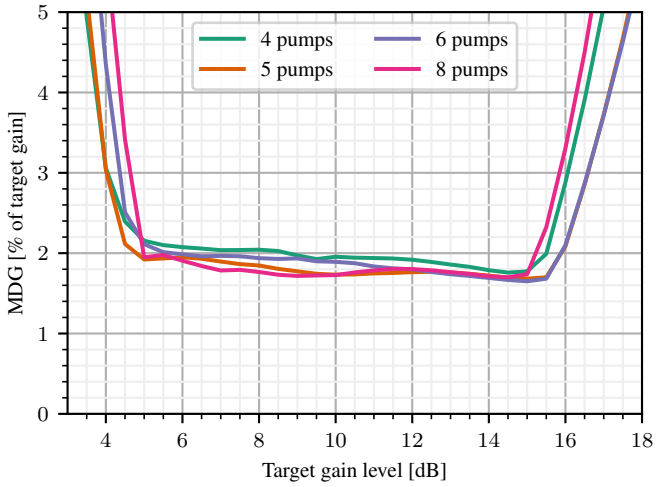

Fig. 6. Relative MDG as a function of the gain level, for different number of pumps.

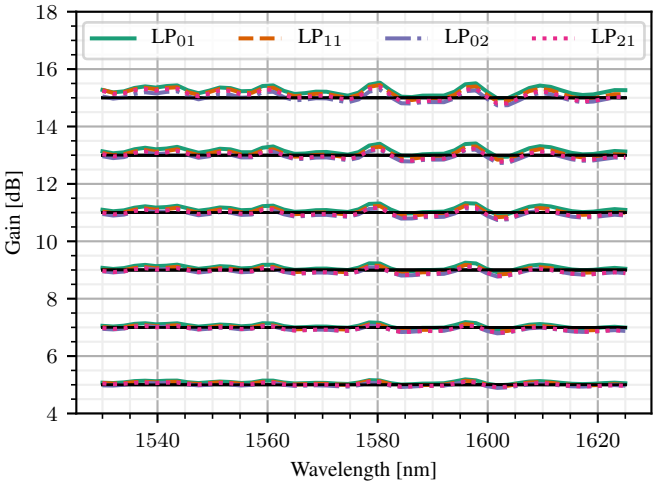

Fig. 5. Target and predicted flat gain profiles for a 4-mode fiber over the C+L–band, using 8 pumps.

WDG, defined for each mode m as

$$\mathrm{F}_m(\mathbf{G}) = \max_i G_i^m - \min_i G_i^m, \tag{24}$$

for $i = 1, \ldots, N_s$ and $m = 1, \ldots, M$. We report gain flatness in terms of percentage of target gain level in Fig. 4, where the shaded areas represent the flatness variation over the modes. The most significant improvement is obtained from 4 to 5 pumps, reducing the flatness from 15% to about 6%. For example, this means that for a 10 dB target gain, the total flatness would be decreased to just 0.6 dB from 1.5 dB; this value is further decreased to 0.35 dB using 8 pumps. Moreover, we can observe that flatness is practically constant among the modes, with fluctuations always lower than 0.5% of the target gain in the interval from 5 dB to 15 dB. We can see an example of the achieved gain profiles for the case of 8 pumps in Fig. 5, where we plot the flat target profiles and the predicted gain curves for different amplification levels inside the training interval. The gain profile for each mode is in fact the same up to a residual MDG, which increases with the gain level.

For a given gain profile $\mathbf{G}$, we quantify its MDG as

$$\mathrm{MDG}(\mathbf{G}) = \max_i \left(\max_m G_i^m - \min_m G_i^m \right), \tag{25}$$

with $i = 1, \ldots, N_s$ and $m = 1, \ldots, M$. In Fig. 6 we report the MDG as percentage of the total gain using 4, 5, 6, and 8 Raman pumps. Differently from the case of gain flatness, the number of pumps does not influence the total MDG, which is practically constant inside the interval of gain levels on which the AE was trained, settling at about 2% of the target gain. This residual MDG is caused mainly by the fact that LP_{01} and LP_{11} modes are systematically over-amplified with respect to the others. By inspecting the values of overlap integrals of FMF1 in Table I, we can observe that the sum of the off-diagonal entries in the columns/rows associated with LP_{01} and LP_{11} are the first and second largest, respectively, meaning that power is more efficiently coupled by the nonlinear Raman interaction in these two modes. In Fig. 7 we plot the total pump power in $z = L$ on each mode of the FMF, as predicted by the AE, as a function of the target gain level; the cases of 4, 5, 6, and 8 pumps are considered, with solid lines representing the average power and shaded areas depicting the power variation by employing different numbers of pumps. Independently of the amplification level, no power is launched in the LP_{01} and LP_{11} modes, with 70% of the total power assigned to LP_{21}, and the remaining 30% to LP_{02}, confirming the results of [24] and [25]. Even though no power is injected in LP_{01} and LP_{11}, these two modes are those that experience the highest amplification, predominantly contributing to the residual MDG of the system. In order to confirm the role of the overlap integrals in determining the MDG we test two additional 4-mode fibers. The first, which we label "FMF2", is a SIF with a core diameter of 18 µm, core refractive index of 1.466, and a relative refractive index difference between core and cladding $\Delta = 0.4\%$, supporting the propagation of the LP_{01}, LP_{11}, LP_{02} and LP_{21} modes over the entire simulation bandwidth. Its overlap integrals are reported in Table I. The second fiber, which we refer to as "FMF3", is instead an ideal 4-mode fiber whose overlap integrals are

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

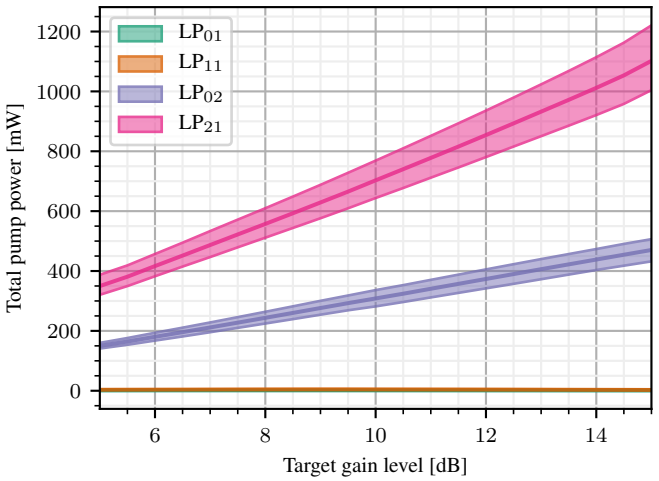

Fig. 7. Total pump power at $z = L$ in each mode of FMF1, as a function of the target gain level. Shaded areas indicate the variation using different number of pumps.

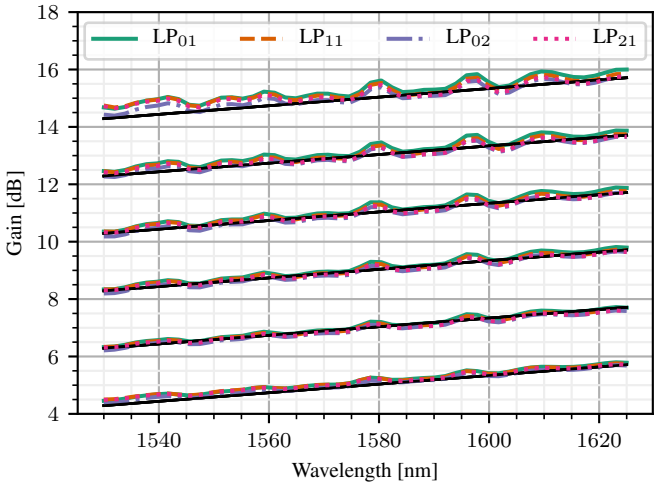

Fig. 9. Target and predicted gain profiles for the tilted case, using 8 pumps and with a total tilt of 1.425 dB, i.e. the maximum considered tilt during training.

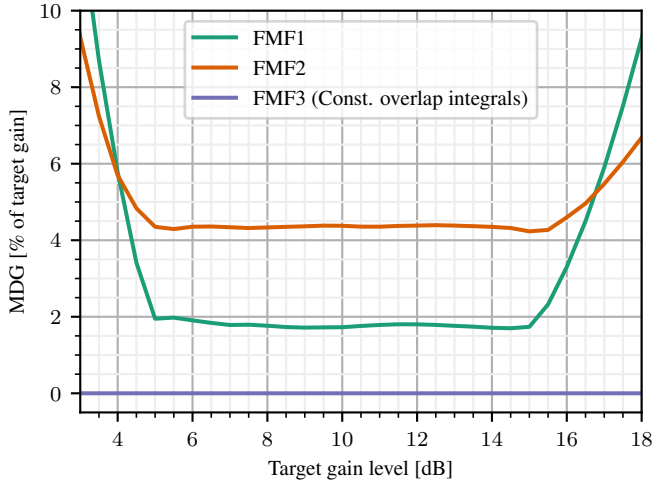

Fig. 8. Relative MDG for flat target gain profiles, as a function of the gain level, for three different fibers. The number of pumps is set to $N_p = 8$.

equal to $5.47 \times 10^{-9}\,\mathrm{m}^{-2}$, i.e the overlap integral for the LP_{01}-LP_{01} mode pair of FMF2. All the other simulation parameters, including the attenuation spectrum and Raman gain coefficient of the fiber, remain unchanged. Training the AE under the same conditions, we can observe the effect of the fiber design on the performance of the system in terms of residual MDG. For the case of 8 Raman pumps, we report the MDG for the three considered fibers as a function of the gain level in Fig. 8: FMF2 exhibits the highest MDG among the fibers, reaching a value of approximately 4% of the target gain inside the training interval of 5 dB to 15 dB, while for FMF3 the AE correctly predicts the power distribution among the modes that results in no MDG, launching power only in the LP_{01} mode.

B. Tilted gain profiles

In order to account for tilted gain profiles, the AE is trained using ideal gain profiles with average gain and tilt uniformly sampled from the two-dimensional training region $\mathcal{T} = [5, 15]\ \mathrm{dB} \times [-0.015, 0.015]\ \mathrm{dB\,nm^{-1}}$, resulting in a maximum total tilt on the C+L band equal to $T_{\max} = 0.015\ \mathrm{dB\,nm^{-1}} \times 95\ \mathrm{nm} = 1.425\ \mathrm{dB}$. In Fig. 9 we report the target gain profiles and corresponding AE predictions using FMF1 and 8 pumps, for a total tilt equal to $T_{\max}$ and for different average gain levels inside the training region. Results show good agreement between targets and predictions, with approximately the same gain profile on each mode, up to the residual MDG.

An analysis similar to that of flat gain profiles is carried out for the case of tilted profiles, evaluating the metrics of interest for FMF1 and varying the number of employed Raman pumps, keeping the other simulation parameters unchanged. We compute RMSE, flatness, and MDG of the predicted gain profiles and visualize them in Fig. 10, representing the metrics as a function of the target gain level and total tilt on the C+L band. Each metric is reported in terms of percentage of the target gain level; for RMSE and flatness we consider the worst-case scenario among the modes, i.e. their maximum value. Fig. 10 is organized such that columns 1 through 4 of the grid correspond to the case of 4, 5, 6, and 8 pumps, whereas row 1, 2, and 3 correspond to RMSE, flatness and MDG, respectively. The color scale for each metric is saturated to different levels in order to improve the contrast of the color maps. In Fig. 10 (a)–(d) we can appreciate the improvements in terms of RMSE by using more pumps: the color map is increasingly darker inside and in the vicinity of the training region $\mathcal{T}$, whose bounds are represented by a dashed rectangle. Additionally, by using 5 or more pumps, the level curves show that a RMSE lower than 3% of the target gain level is achieved for (practically) all the gain level-tilt combinations in $\mathcal{T}$.

Similar observations can be made for the flatness from Fig. 10 (e)–(h), where a value of about 17% is reached for the points inside the training region using 4 pumps; increasing the number of pumps leads to progressively lower flatness values, down to 5% inside $\mathcal{T}$ with 8 pumps.

Similarly, for the MDG, Fig. 10, (i)–(l) show that a higher number of pumps brings no significant changes, as the minimum achievable MDG is determined by overlap integrals of

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

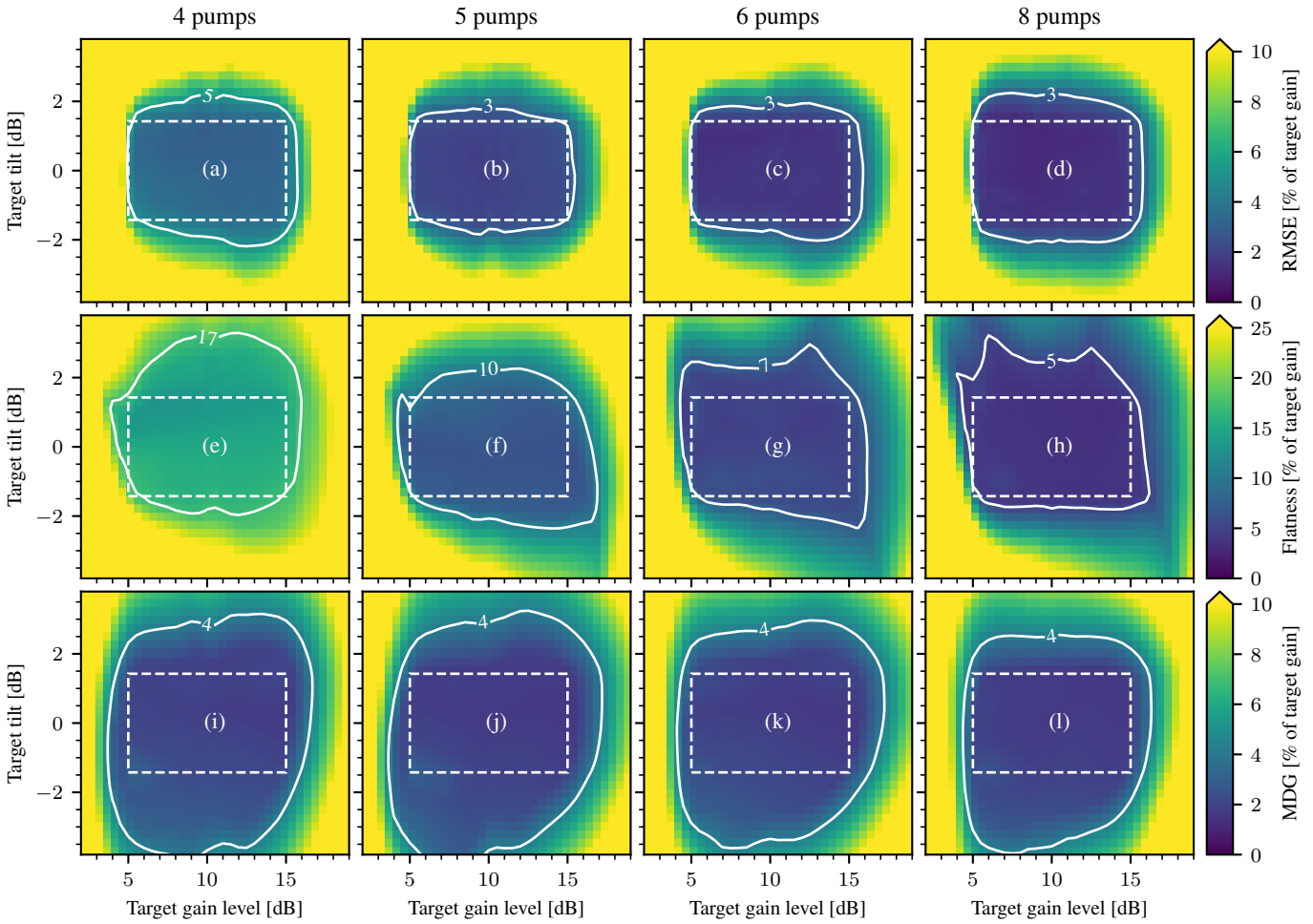

Fig. 10. Calculated metrics for the tilted gain case, varying the number of Raman pumps: RMSE (a)–(d), flatness (e)–(h), and MDG (i)–(l) as a function of the target gain level and target tilt. For RMSE and flatness their maximum value among the modes is reported. Columns 1 through 4 refer to the case of 4, 5, 6, and 8 pumps, respectively.

the fiber. Its value stays infact approximately constant inside the training region regardless of the pump count, with the level curve showing that MDG values lower than 4% are achieved for a region considerably wider than $\mathcal{T}$.

IV. Conclusions

We have demonstrated an unsupervised machine learning method based on autoencoders to predict the required pump parameters to generate flat and tilted gain profiles using Raman amplification in few-mode fibers. Thanks to automatic differentiation, a numerical Raman model is embedded in the autoencoder, allowing to train it directly on ideal gain profiles (e.g. flat or tilted) and obtaining a robust unsupervised learning method that does not rely on a pre-computed dataset to learn the inverse model. In fact, the relationship between input target gain and the pump parameters that best approximate it are learned in the training phase from the embedded numerical model, allowing to accurately sample the targeted region of the space of possible gain profiles. As a result, this method scales well with respect to the number of fiber modes, the number of Raman pumps, and the amplification bandwidth. On this regard, the low root-mean-square error (quantified for various number of pump wavelenghts) demonstrated the achievement of the target profile. Another key advantage of this scheme is that it does not require supervision in selecting simulation parameters (like power and wavelength ranges) that might also affect the quality of the results.

This approach is tested on a 4-mode fiber using the counter-pumping scheme, various numbers of pumps, up to 8, and for the C+L band. The training process is further simplified by the fact that the autoencoder can directly predict the pump powers at $z = 0$, eliminating the need to employ costly shooting algorithms that are typically needed for counter-propagating Raman amplification models. The pump power to be injected in the fiber are in fact computed with a single integration of the propagation equations. We achieved very good results regarding flatness and mode-dependent gain over the entire C+L band and the considered interval of gain levels and tilts, reaching a gain flatness of 3% of the total gain using 8 pumps, and a residual mode-dependent gain of 2% of the total gain, independently of the number of Raman pumps. This method can be extended to the case of co-propagating pumps and even to a mixture of co- and counter-propagating pumps. Finally, if the numerical model is substituted by an experiment (with

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JLT.2020.3034692, Journal of Lightwave Technology

automatic data aquisition), the encoder neural network could, in principle, be trained by the experiments. This will also require the definition of a proper algorithm to update the neural network parameters, to replace automatic differentiation.

ACKNOWLEDGMENTS

This work is partly supported by the Italian Ministry for Education, University and Research (MIUR, "Departments of Excellence"—law 232/2016, and PRIN 2017—project Fiber Infrastructure for Research on Space-division multiplexed Transmission (FIRST)) and by the University of Padova (BIRD 2019—project MACFIBER).

REFERENCES

[1] P. P. Mitra and J. B. Stark, "Nonlinear limits to the information capacity of optical fibre communications," *Nature*, vol. 411, no. 6841, p. 1027, Jun. 2001.

[2] R.-J. Essiambre, G. J. Foschini, G. Kramer, and P. J. Winzer, "Capacity Limits of Information Transport in Fiber-Optic Networks," *Physical Review Letters*, vol. 101, no. 16, p. 163901, Oct. 2008.

[3] A. D. Ellis, N. M. Suibhne, D. Saad, and D. N. Payne, "Communication networks beyond the capacity crunch," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2062, p. 20150191, Mar. 2016.

[4] A. Chralyvy, "The coming capacity crunch," in *European Conference on Optical Communication (ECOC)*, Vienna, Austria, Sep. 2009, Plenary talk.

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

[6] R. Ryf, S. Randel, A. H. Gnauck, C. Bolle, A. Sierra, S. Mumtaz, M. Esmaeelpour, E. C. Burrows, R.-J. Essiambre, P. J. Winzer, D. W. Peckham, A. H. McCurdy, and R. Lingle, "Mode-Division Multiplexing Over 96 km of Few-Mode Fiber Using Coherent 6×6 MIMO Processing," *Journal of Lightwave Technology*, vol. 30, no. 4, pp. 521–531, Feb. 2012.

[7] V. Sleiffer, Y. Jung, V. Veljanovski, R. van Uden, M. Kuschnerov, H. Chen, B. Inan, L. G. Nielsen, Y. Sun, D. Richardson, S. Alam, F. Poletti, J. Sahu, A. Dhar, A. Koonen, B. Corbett, R. Winfield, A. Ellis, and H. de Waardt, "737 Tb/s (96 x 3 x 256-Gb/s) mode-division-multiplexed DP-16QAM transmission with inline MM-EDFA," *Optics Express*, vol. 20, no. 26, p. B428, Dec. 2012.

[8] N. Bai, E. Ip, Y.-K. Huang, E. Mateo, F. Yaman, M.-J. Li, S. Bickham, S. Ten, J. Liñares, C. Montero, V. Moreno, X. Prieto, V. Tse, K. Man Chung, A. P. T. Lau, H.-Y. Tam, C. Lu, Y. Luo, G.-D. Peng, G. Li, and T. Wang, "Mode-division multiplexed transmission with inline few-mode fiber amplifier," *Optics Express*, vol. 20, no. 3, p. 2668, Jan. 2012.

[9] R. Ryf, A. Sierra, R.-J. Essiambre, S. Randel, A. H. Gnauck, C. Bolle, M. Esmaeelpour, P. J. Winzer, R. Delbue, P. Pupalaikise, A. Sureka, D. W. Peckham, A. McCurdy, and R. Lingle, "Mode-Equalized Distributed Raman Amplification in 137-km Few-Mode Fiber," in *Proceedings of European Conference on Optical Communication (ECOC)*, Geneva, Switzerland, Sep. 2011, Paper Th.13.K.5.

[10] M. Esmaeelpour, R. Ryf, N. K. Fontaine, H. Chen, A. H. Gnauck, R.-J. Essiambre, J. Toulouse, Y. Sun, and R. Lingle, "Transmission Over 1050-km Few-Mode Fiber Based on Bidirectional Distributed Raman Amplification," *Journal of Lightwave Technology*, vol. 34, no. 8, pp. 1864–1871, Apr. 2016.

[11] J. Li, L. Wang, J. Du, S. Jiang, L. Ma, C. Cai, L. Zhu, A. Wang, M.-J. Li, H. Chen, J. Wang, and Z. He, "Experimental demonstration of a few-mode Raman amplifier with a flat gain covering 1530–1605 nm," *Optics Letters*, vol. 43, no. 18, p. 4530, Sep. 2018.

[12] J. Bromage, "Raman Amplification for Fiber Communications Systems," *Journal of Lightwave Technology*, vol. 22, no. 1, pp. 79–93, Jan. 2004.

[13] D. Jia, H. Zhang, Z. Ji, N. Bai, and G. Li, "Optical fiber amplifiers for space-division multiplexing," *Frontiers of Optoelectronics*, vol. 5, no. 4, pp. 351–357, Dec. 2012.

[14] C. Headley and G. P. Agrawal, *Raman Amplification in Fiber Optical Communication Systems*. Academic Press, 2005.

[15] D. Zibar, A. M. Rosa Brusin, U. C. de Moura, F. Da Ros, V. Curri, and A. Carena, "Inverse System Design Using Machine Learning: The Raman Amplifier Case," *Journal of Lightwave Technology*, vol. 38, no. 4, pp. 736–753, Feb. 2020.

[16] Y. Chen, J. Du, Y. Huang, K. Xu, and Z. He, "Intelligent gain flattening in wavelength and space domain for FMF Raman amplification by machine learning based inverse design," *Optics Express*, vol. 28, no. 8, pp. 11 911–11 920, Apr. 2020.

[17] M. Bartholomew-Biggs, S. Brown, B. Christianson, and L. Dixon, "Automatic differentiation of algorithms," *Journal of Computational and Applied Mathematics*, vol. 124, no. 1, pp. 171–190, Dec. 2000.

[18] M. Raissi, P. Perdikaris, and G. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, Feb. 2019.

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[20] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic differentiation in machine learning: A survey," *Journal of Machine Learning Research*, vol. 18, no. 153, pp. 1–43, 2018.

[21] B. Karanov, M. Chagnon, F. Thouin, T. A. Eriksson, H. Bulow, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-End Deep Learning of Optical Fiber Communications," *Journal of Lightwave Technology*, vol. 36, no. 20, pp. 4843–4855, Oct. 2018.

[22] R. T. Jones, T. A. Eriksson, M. P. Yankov, and D. Zibar, "Deep Learning of Geometric Constellation Shaping Including Fiber Nonlinearities," in *Proceedings of European Conference on Optical Communication (ECOC)*, Rome, Italy, Sep. 2018, Paper We1F.5.

[23] J. Cho and P. J. Winzer, "Probabilistic Constellation Shaping for Optical Fiber Communications," *Journal of Lightwave Technology*, vol. 37, no. 6, pp. 1590–1607, Mar. 2019.

[24] R. Ryf, R. Essiambre, J. von Hoyningen-Huene, and P. Winzer, "Analysis of Mode-Dependent Gain in Raman Amplified Few-Mode Fiber," in *Proceedings of Optical Fiber Communication Conference (OFC)*, Los Angeles, CA, USA, 2012.

[25] J. Zhou, "An analytical approach for gain optimization in multimode fiber Raman amplifiers," *Optics Express*, vol. 22, no. 18, p. 21393, Sep. 2014.

[26] C. Antonelli, A. Mecozzi, and M. Shtaif, "Raman amplification in multimode fibers with random mode coupling," *Optics Letters*, vol. 38, no. 8, p. 1188, Apr. 2013.

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017.

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, Mar. 2010.

[30] H. M. Jiang and K. Xie, "Efficient and robust shooting algorithm for numerical design of bidirectionally pumped Raman fiber amplifiers," *Journal of the Optical Society of America B*, vol. 29, no. 1, p. 8, Jan. 2012.

[31] D. Hollenbeck and C. D. Cantrell, "Multiple-vibrational-mode model for fiber-optic Raman gain spectrum and response function," *Journal of the Optical Society of America B*, vol. 19, no. 12, p. 2886, Dec. 2002.

0733-8724 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on December 21,2020 at 17:04:48 UTC from IEEE Xplore. Restrictions apply.