


RESEARCH ARTICLE  

# Optical network physical layer parameter optimization for digital backpropagation using Gaussian processes

Josh W. Nevin<sup>1</sup>, Eric Sillekens<sup>2</sup>, Ronit Sohanpal<sup>2</sup>, Lidia Galdino<sup>3</sup>, Sam Nallaperuma<sup>4</sup> ,  
Polina Bayvel<sup>3</sup> and Seb J. Savory<sup>4</sup>

<sup>1</sup>CloudNC, London, United Kingdom

<sup>2</sup>Optical Networks Group, Department of Electronic and Electrical Engineering, University College London, London, United Kingdom

<sup>3</sup>Corning Optical Fiber Communications, Ewloe, United Kingdom

<sup>4</sup>Fibre Optic Communication Systems Laboratory (FOCSLab), Electrical Engineering Division, Department of Engineering, University of Cambridge, Cambridge, United Kingdom

**Corresponding author:** Sam Nallaperuma; Email: [snn26@cam.ac.uk](mailto:snn26@cam.ac.uk)

**Received:** 29 March 2023; **Revised:** 14 July 2023; **Accepted:** 19 July 2023

**Keywords:** digital backpropagation; Gaussian processes; machine learning; optical networks

## Abstract



We present a novel methodology for optimizing fiber optic network performance by determining the ideal values for attenuation, nonlinearity, and dispersion parameters in terms of achieved signal-to-noise ratio (SNR) gain from digital backpropagation (DBP). Our approach uses Gaussian process regression, a probabilistic machine learning technique, to create a computationally efficient model for mapping these parameters to the resulting SNR after applying DBP. We then use simplicial homology global optimization to find the parameter values that yield maximum SNR for the Gaussian process model within a set of a priori bounds. This approach optimizes the parameters in terms of the DBP gain at the receiver. We demonstrate the effectiveness of our method through simulation and experimental testing, achieving optimal estimates of the dispersion, nonlinearity, and attenuation parameters. Our approach also highlights the limitations of traditional one-at-a-time grid search methods and emphasizes the interpretability of the technique. This methodology has broad applications in engineering and can be used to optimize performance in various systems beyond optical networks.

## Impact Statement

This article presents a powerful methodology for optimizing fiber-optic communication systems that can greatly impact various engineering applications. First, our methodology offers a highly efficient tool for optimizing fiber-optic communication systems, which can lead to faster and more reliable communication in various industries. Second, this approach has the potential to provide significant impact within other engineering disciplines, enabling efficient, and interpretable system parameter optimization.

## 1. Introduction

Optical fiber communication networks are a vital component of the global internet, but they suffer from nonlinear interference which reduces system capacity (Agrawal, 2019). Digital backpropagation (DBP) is

  This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.



a powerful technique for compensating for these nonlinear impairments, allowing lost capacity to be recovered (Ip and Kahn, 2008). However, to achieve high gain from DBP, the physical layer parameters of the system, such as fiber dispersion  $D$ , fiber nonlinearity coefficient  $\gamma$ , and fiber loss  $\alpha$ , must be optimized with respect to DBP gain. These parameters are uncertain, even at the beginning of the system's life (Pointurier, 2021), and can change with time due to the aging of system components (Pointurier, 2017). Currently, these parameters are manually optimized for DBP using a time-consuming one-at-a-time grid search process that does not guarantee optimality. This is not a suitable approach, as these parameters have complex interdependencies.

Here, we present a method that uses Gaussian process (GP) regression (Rasmussen and Williams, 2006), a class of probabilistic machine learning (ML) models, and a powerful global optimizer, simplicial homology global optimization (SHGO), to optimize these physical layer parameters with respect to the DBP gain. This technique provides both systematic optimization of DBP gain and estimation of the system physical layer parameters, which is beneficial for emerging applications such as digital twins (Nevin et al., 2021) and reducing network margins. However, it should be noted that the parameters that optimize the DBP gain may differ from those that of the physical system (Fan et al., 2020; Yi et al., 2021). Nevertheless, these estimates are useful for reducing network margins and for building virtual system models, in the absence of more accurate, time-consuming measurements using dedicated equipment. Specifically, we optimize the fiber nonlinearity coefficient  $\gamma$ , fiber dispersion  $D$ , and fiber loss  $\alpha$ . Lumped parameters are estimated, rather than distributed ones, as these are sufficient for the optimization of the DBP gain, which is the focus of this work. Extending this methodology to consider distributed physical layer parameters forms part of the planned future work. We demonstrate the effectiveness of our method through simulation, where we know the ground truth values of  $D$ ,  $\gamma$ , and  $\alpha$ , before applying it to experimentally measured traces. We also provide parameter sweeps of the learned GP models to aid in interpreting the surrogate model mapping from  $D$ ,  $\gamma$ , and  $\alpha$  to the signal-to-noise ratio (SNR) after DBP, and we use these models to highlight the limitations of using one-at-a-time grid search to optimize the DBP parameters.

Our methodology has applications beyond optical networks and can be used to optimize performance in various systems. By leveraging GP regression and SHGO, we can optimize complex parameters efficiently, providing valuable insights and improvements to engineering design and optimization.

The remainder of this article is organized as follows. We present the theoretical background in Section 2 followed by a review of existing work in Section 3. We present the proposed parameter optimization methodology in Section 4. The experimental optical fiber communications system is detailed in Section 5, followed by the parameter estimation results in simulation in Section 6 and experiment in Section 7. Finally, we summarize the key conclusions and outline future work in Section 8.

## 2. Background

This section describes the techniques referred in the study: the Split-step Fourier method (SSFM) for modeling the physical layer of the optical network, the DBP technique for joint compensation of both linear and nonlinear impairments in optical fibers, the Gaussian process regression (GPR) technique for probabilistic supervised ML, and the Simplicial Homology Global Optimisation (SHGO) algorithm for finding the global optimum of noncontinuous, nonconvex and nonsmooth functions. The SSFM uses the Manakov equation, and GPR uses kernel methods and the squared exponential kernel function.

### 2.1. Split-step Fourier method

In this work, we model the optical network physical layer using the Manakov equation, given by (Millar et al., 2010)

$$\begin{aligned}
 & \frac{\partial E_X}{\partial z} + \frac{\alpha}{2} E_X + \frac{j\beta_2}{2} \frac{\partial^2 E_X}{\partial t^2} \\
 & -j\frac{8}{9}\gamma(|E_X|^2 + |E_Y|^2)E_X = 0, \\
 & \frac{\partial E_Y}{\partial z} + \frac{\alpha}{2} E_Y + \frac{j\beta_2}{2} \frac{\partial^2 E_Y}{\partial t^2} \\
 & -j\frac{8}{9}\gamma(|E_X|^2 + |E_Y|^2)E_Y = 0,
 \end{aligned} \tag{1}$$

where  $\beta_2$  is the group velocity dispersion coefficient of the fiber,  $\alpha$  is the fiber attenuation coefficient,  $t$  and  $z$  denote the retarded time and the spatial coordinate along the direction of propagation respectively,  $E_X$  and  $E_Y$  are the electric field components along the  $X$  and  $Y$  axes, respectively, and  $\gamma$  is the fiber nonlinearity coefficient, given by

$$\gamma = \frac{2\pi n_2}{\lambda A_{\text{eff}}}, \tag{2}$$

where  $A_{\text{eff}}$  is the fiber effective mode area,  $n_2$  is the nonlinear index coefficient, and  $\lambda$  is the center operating wavelength.  $\beta_2$  relates to the dispersion coefficient  $D$  via  $D = -\frac{2\pi c}{\lambda^2}\beta_2$ . Writing (1) in terms of its constituent linear and nonlinear parts can be shown to yield an exact solution given by (Millar et al., 2010)

$$\mathbf{E}(z+h, T) = \exp(h(\hat{D} + \hat{N}))\mathbf{E}(z, T), \tag{3}$$

after propagation over distance  $h$  for time  $T$ , where

$$\begin{aligned}
 \hat{N} &= -j\frac{8}{9}\gamma \mathbf{E}^H \mathbf{E} - \frac{\alpha}{2}, \\
 \hat{D} &= -\frac{j\beta_2}{2} \frac{\partial^2}{\partial t^2},
 \end{aligned} \tag{4}$$

where  $\mathbf{E}^H$  denotes the Hermitian transpose of  $\mathbf{E}$ . We can then assume (Millar et al., 2010)

$$\exp(h(\hat{D} + \hat{N}))\mathbf{E}(z, T) \approx \exp(h\hat{D}) \exp(h\hat{N})\mathbf{E}(z, T). \tag{5}$$

In this work, we utilize the symmetric SSFM, which involves taking a dispersive step of size  $\frac{h}{2}$ , followed by a nonlinear step of size  $h$  and an additional dispersive step of size  $\frac{h}{2}$ . This can be written as (Millar et al., 2010; Agrawal, 2019)

$$\mathbf{E}(z+h, T) \approx \exp\left(\frac{h}{2}\hat{D}\right) \exp(L_{\text{eff}}\hat{N}) \exp\left(\frac{h}{2}\hat{D}\right)\mathbf{E}(z, T), \tag{6}$$

where  $L_{\text{eff}} = \frac{1 - \exp(-\alpha h)}{\alpha}$  is the fiber nonlinear effective length. It should be noted that here  $\alpha$  is measured in units of  $\text{km}^{-1}$ , rather than  $\text{dB km}^{-1}$ . Equation (6) can be solved iteratively using a fast Fourier transform to obtain an approximate solution for the transmission in the fiber (Ip and Kahn, 2008).

We also simulate polarization mode dispersion (PMD) using the waveplate model for birefringent fiber, such that each step of the SSFM includes a random polarization rotation and a frequency-dependent birefringence. We obtain the differential group delay between fast and slow axes  $\Delta\tau_p$  for a given fiber PMD  $\langle\Delta\tau\rangle$  as (Poole and Favin, 1994)

$$\Delta\tau_p = \sqrt{\frac{3\pi}{8n_{\text{sps}}}\langle\Delta\tau\rangle}, \tag{7}$$

where  $n_{\text{sps}}$  is the number of steps per fiber span. To include PMD, a random polarization rotation is applied at the start of each symmetric SSFM step and the differential group delay is then included as part of each dispersive step of size  $\frac{h}{2}$ .

## 2.2. Digital backpropagation

DBP is a technique for the joint compensation of both linear and nonlinear impairments that accumulate in optical fibers. This is achieved by dividing the fiber link of interest up into small sections  $\Delta z$  and inversely solving the nonlinear Schrödinger equation (NLSE) which governs the propagation of optical signals through the optical fiber (Ip and Kahn, 2008). Specifically, we solve the Manakov equation using the SSFM, with the sign of the physical layer parameters reversed, that is,  $\alpha \rightarrow -\alpha$ ,  $\gamma \rightarrow -\gamma$ , and  $\beta_2 \rightarrow -\beta_2$ . Thus, we simulate the propagation of the signal in reverse, in order to estimate the linear and nonlinear impairments accumulated in the span. We utilize a symmetric SSFM implementation, as outlined above. In this work, we utilized DBP with a fixed step size and a constant number of steps per span, although this can be adapted whilst performing DBP to improve the computational efficiency (Shao et al., 2014). All SSFM simulations, for both forward and reverse propagation, were performed with 100 steps per fiber span.

## 2.3. Gaussian process regression

GP regression is a probabilistic supervised ML method. GPs leverage Bayesian statistics to find the most likely function describing the relationship between a set of inputs and a set of outputs, given the data and a set of priors (Rasmussen and Williams, 2006). GP regression is a nonparametric ML method, meaning that rather than assuming a given parametric form and finding a set of parameters that describe the mapping, the space of functions is searched directly in a probabilistic way. The main benefit of using GPs over other supervised regression methods, such as neural networks (NNs), is that GPs provide a rigorous prediction error, the predictive variance. This can be used to judge the level of confidence associated with a given prediction, providing an extra layer of interpretability for the model user.

GPs are kernel methods, meaning that a kernel function is used to model the covariance between the data. Specifically, the kernel function specifies the covariance between the evaluation of the function being learned by the GP  $f(\mathbf{x})$  at two input points (Rasmussen and Williams, 2006):

$$k(\mathbf{x}_i, \mathbf{x}_j) = \text{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)). \quad (8)$$

In this work, we use the squared exponential kernel function, defined as (Rasmussen and Williams, 2006)

$$k(\mathbf{x}_i, \mathbf{x}_j) = h_1^2 \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\mathbf{h}_2^2}\right) + h_3^2 \delta_{ij}, \quad (9)$$

where  $x_i$  and  $x_j$  are the input values of the points being compared,  $h_1$  and  $\mathbf{h}_2$  are the hyperparameters of the squared exponential kernel,  $\delta_{ij}$  is the Kronecker delta, that is,  $\delta_{ij} = 1$  if  $i = j$  and 0 otherwise, and  $\|\cdot\|$  represents the Euclidean distance. It should be noted that  $\mathbf{h}_2$  is a vector with the same number of elements as the dimensionality of  $\mathbf{x}$ , whereas  $h_1$  and  $h_3$  are both scalars, as the targets have a single dimension. Commonly, the set of kernel hyperparameters are denoted by  $\theta$ , meaning for equation (9),  $\theta \triangleq \{h_1, \mathbf{h}_2, h_3\}$ .  $h_1$  acts as a global covariance scale for the GP, as it affects the magnitude of the covariances between all pairs of sample points  $\mathbf{x}_i, \mathbf{x}_j$ . The role of  $\mathbf{h}_2$  is to scale the distance between inputs  $\mathbf{x}_i, \mathbf{x}_j$  that will yield a large covariance  $\text{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j))$ , where each component of  $\mathbf{h}_2$  scales a given input dimension. Therefore, larger values of the correlation length indicate that the function  $f(\mathbf{x})$  varies less with respect to changes in a given component of  $\mathbf{x}$ , and vice versa.  $h_3$  represents the noise of the observations—larger values of  $h_3$  indicate that the GP is interpreting more of the variation in the data as noise, rather than the underlying signal. Using a kernel function to compute inner products in the original space is equivalent to using a large set of basis functions, facilitating efficient learning of complex functions (Rasmussen and Williams, 2006). When we fit the GP to data, we do so by maximizing the likelihood function with respect to the hyperparameters  $\theta = h_1, \mathbf{h}_2$ , and  $h_3$  (Rasmussen and Williams, 2006). This corresponds to maximization of the probability of the targets given the data and hyperparameters—thus we find the underlying mapping between the inputs and

targets in a probabilistic way. Specifically, this was achieved by computing the gradient of the log marginal likelihood with respect to the hyperparameters and performing a gradient-based maximization using the Limited memory Broyden–Fletcher–Goldfarb–Shanno Box constrained (L-BFGS) method (Liu and Nocedal, 1989). This gradient is given by the expression (2.30) (Rasmussen and Williams, 2006, chap. 2):

$$\frac{\partial}{\partial \theta_j} \log p(\mathbf{y}|X, \theta) = \frac{1}{2} \mathbf{y}^T K^{-1} \frac{\partial K}{\partial \theta_j} K^{-1} \mathbf{y} - \frac{1}{2} \text{tr} \left( K^{-1} \frac{\partial K}{\partial \theta_j} \right), \quad (10)$$

where  $K$  is the matrix of the kernel evaluated at all pairs of input points and  $\mathbf{y}$  is the vector of targets. An in-depth discussion of this process is provided in Rasmussen and Williams (2006, chap. 5, 5.3–5.7). Choosing a given kernel corresponds to making some assumptions about the data, as different kernels result in a different prior distribution over functions. We choose equation (9) because we expect the underlying functions to be well-described by a single-length scale with Gaussian noise, and we do not expect the data to exhibit properties such as periodicity and decay that may require more complex kernels (Rasmussen and Williams, 2006). We also tried using the more general Matérn kernel (Rasmussen and Williams, 2006, chap. 4), but saw no improvement relative to equation (9), and therefore we used the more simplistic kernel which was faster to fit. In this work, we utilize an implementation provided by the Multi-Output Gaussian Process Emulator library (Daub and Barlow, 2021).

The GP used in this work is constructed using a zero mean function and the covariance function given by equation (9). Therefore, we have a GP prior over our unknown function  $f \sim \text{GP}$ , and we can write the prior over the function value at a specific input  $\mathbf{x}$  as:

$$P(f(\mathbf{x})) = \mathcal{N}(f(\mathbf{x}); 0, K), \quad (11)$$

where  $P$  denotes a probability,  $\mathcal{N}$  is a multivariate Gaussian distribution, and  $K$  is the kernel matrix, such that  $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ . It should be noted that the majority of applications of GPs utilize a zero mean function, as these models are able to learn a vast number of functions accurately within the range of the measured data (Rasmussen and Williams, 2006). As we have strong a priori bounds on the range over which we wish to make predictions, the incorporation of a specific mean function is not required. This GP model was observed to be of sufficient complexity to act as a high-accuracy surrogate model for this problem. For a detailed description of the theory of GPs, we refer the reader to Bishop (2006, chap. 6) and Rasmussen and Williams (2006). Moreover, we note that relevance vector machines (Candela and Hansen, 2004; Martino and Read, 2021) (RVMs) are very similar in formulation to GPs. There is an inherent trade-off between these two approaches—RVMs offer greater flexibility in the choice of basis functions than GPs, however, GPs have more interpretable predictive variance than RVMs. Exploring the use of RVMs in this technique forms part of the planned future work.

#### 2.4. Simplicial homology global optimization

SHGO is a general-purpose global optimization algorithm based on algebraic topology and simplicial integral homology (Endres et al., 2018). SHGO has been chosen to optimize the function learned by the GP in this work as it offers a number of theoretical and practical advantages. First, SHGO is theoretically proven to find the global optimum, even for the highly general case of noncontinuous, nonconvex, and nonsmooth functions (Endres et al., 2018). In addition, SHGO has been shown to be highly performant compared to other state-of-the-art global optimization algorithms (Endres et al., 2018). Also, SHGO accepts bounds on the variables being optimized, allowing us to input some a priori knowledge in the form of broad parameter ranges derived from fiber specification sheets and the literature. Finally, SHGO does not require the derivative of the objective function, eliminating an additional source of error given that the objective function is in this case learned by the GP. We utilized the SciPy implementation of SHGO in this work (Virtanen et al., 2020).

### 3. Related Work

A number of previous works have considered ML approaches for the optimization of DBP. In Häger and Pfister (2018), the similarities between the structure of split-step Fourier-method (SSFM)-based DBP and NNs are leveraged to reduce the computational complexity of DBP. Here, NNs are used to perform DBP and an interpretable approach is taken in which the chosen structure of the NN is informed by the structure of traditional DBP. Such approaches are known as learned DBP. An extension of the learned DBP approach to joint DBP and PMD has also been presented (Bütler et al., 2020). Further extensions include learned DBP with polarization state rotation and state noise invariance (Bitachon et al., 2020). Moreover, learned time-domain DBP (Sillekens et al., 2020) has also been proposed as a lower-complexity alternative to standard learned DBP, in which nonlinearity compensation is done in the time domain and chromatic dispersion is compensated in the frequency domain. Thus, repeated transforms between the time and frequency domains via Fourier transform result in high computational complexity. Using the time-domain learned DBP approach, it is possible to avoid this conversion by performing all compensation in the time domain.

However, all DBP approaches require tuning of the physical layer parameters, specifically the nonlinearity, attenuation, and dispersion coefficients,  $\gamma$ ,  $\alpha$ , and  $D$ . Errors in these parameters lead to reduced DBP gain. These parameters are poorly known for deployed systems (Pointurier, 2021) and may also change with time due to the aging of components (Pointurier, 2017). Thus, a number of techniques have been proposed in the literature to perform data-driven estimation of these physical layer parameters. A simple least-squares minimization technique leveraging the enhanced Gaussian noise model was proposed (Ives et al., 2017). Moreover, other works have proposed to update the parameters of a physics-based model based on monitoring data using a range of techniques, including maximum likelihood estimation (Bouda et al., 2018), Markov chain Monte Carlo (Meng et al., 2017), and gradient ascent (Seve et al., 2018). In each case, monitoring data is used to update the parameters in the physics-based model based on measurements of the system QoT. An additional approach leveraging GP-driven history matching for physical layer parameter estimation was presented in Nevin et al. (2022). Here, GPs were utilized to learn a computationally cheap surrogate for a simulation of the forward channel based on the SSFM, which was then used to perform history matching to find the most likely values for a set of uncertain physical layer parameters.

In addition, a number of recent works have investigated how DBP can be utilized to estimate these physical layer parameters. Sasai et al. (2020) used learned DBP based on NNs to estimate  $\gamma$  and the group velocity dispersion  $\beta_2$  on a span-by-span basis, and demonstrated the detection of spans with excessively high loss along a long-haul transmission link. The proposed method utilized gradient descent to minimize the mean squared error (MSE) between the backpropagated signal and a reference signal, to find optimal values for these parameters simultaneously. Optical attenuators were used to model points of excessive loss along the link, which were detected using this scheme by comparing with the expected loss for a normal state. An extension of this method was also applied to estimation of the loss profile and the detection of frequency detuning in optical filters (Sasai et al., 2020). Self-phase modulation was exploited to separate out the filter responses, which is not possible in the linear regime. Again, gradient descent was utilized to estimate the parameters, and the detection of an anomalous optical filter was demonstrated. Another work focused on using DBP to estimate the output power of an experimental cascade of EDFAs (Tang et al., 2021), in order to help with power budget adjustment and detection of anomalous EDFA behavior. Similarly, DBP has also been utilized for estimation of the power profile and gain spectra of Raman amplifiers from the received signal traces (Sasai et al., 2021). Thus, DBP has proven useful in the estimation of physical layer parameters at the receiver.

In this work, we propose a novel methodology that optimizes the physical layer parameters  $\gamma$ ,  $\alpha$ , and  $D$  in terms of DBP gain. This is achieved by using a probabilistic ML method, GP regression, to learn the functional mapping between the physical layer parameters and the DBP gain for a given received trace. This functional mapping is then used to find the optimal parameters via SHGO (Endres et al., 2018). The parameters that optimize DBP gain are found for this trace in a systematic way, which also provides a

data-driven estimate of the parameters themselves, although it should be noted that the parameters that optimize the DBP gain may differ from the physical characteristics of the equipment (Fan et al., 2020; Yi et al., 2021). Moreover, the optimizer used returns local optima, which allows the user to avoid falling into a single local optimum. We demonstrate how we can leverage a priori domain knowledge to select the most likely estimates of the parameters. This methodology is of use for both learned and traditional DBP techniques, as all require optimization of the physical layer parameters. It also provides a method for estimation of the physical layer parameters from a single measured trace, thus it avoids incurring a significant SNR penalty. Reducing the uncertainties for these parameters is crucial for reducing the errors of physics-based models of the system, which are used to make network planning and operation decisions. These uncertainties are typically accounted for with excessive margins, and thus more accurate parameter estimates allow us to unlock extra capacity by reducing these margins. Also, the proposed method provides an interpretable way to understand the parameter sensitivity of the DBP chain, via the trained GP surrogate models.

We summarize our contributions relative to those in the literature here. We use a powerful global optimizer that is theoretically proven to converge to the optimum, even for nonconvex, noncontinuous, and nonsmooth functions (Endres et al., 2018). The problem is therefore reduced to finding a highly accurate probabilistic surrogate model for the DBP chain, with a well-quantified predictive uncertainty. Thus, we do not rely on gradient descent, as in some of the previous approaches, which can be prone to finding a local optimum. Rather, we demonstrate how multiple sets of parameters can yield equivalent DBP gain to within negligible tolerance, and demonstrate how to select the most likely parameter estimates via application of tighter a priori bounds. Moreover, some methods require measurements over a range of launch powers, which may incur significant SNR penalties, whereas we perform our estimation at the receiver, from a single trace with maximum SNR penalty of 0.4 dB and a minimum SNR penalty of 0 dB (i.e., using the optimum launch power). In addition, the proposed method is very flexible with respect to the parameters being optimized—it is trivial to include any other parameter that is set in the DBP chain, although it is likely that more samples will be required to learn an accurate surrogate model in a higher-dimensional space. On the other hand, some of the previously proposed methods are relatively inflexible, requiring significant modifications to estimate new parameters. Finally, many of the previous parameter estimation approaches have been demonstrated on experimental data only without a deterministic simulated ground truth comparison. In this work, we first demonstrate the approach in simulation before moving on to experimental data, that is, we show that we are able to accurately estimate a set of known, noiseless, ground truths, before applying the method to experimental data.

#### 4. Proposed Method

Here, we describe the proposed method for optimization of the parameters  $\alpha$ ,  $\gamma$ , and  $D$  in terms of the DBP gain, given a set of received traces. First, we define a set of a priori ranges for the physical layer parameters considered, obtained from component datasheets. Then, we sample from these ranges to generate a set of input points for our training data, using a Latin hypercube design for efficient sampling of the input space (Daub and Barlow, 2021). Given a single trace, we decode the trace using DBP for each set of sampled inputs  $\alpha$ ,  $\gamma$ , and  $D$  and record the resulting SNR. These SNR values form the set of targets for our training data, whilst the sampled inputs are the training data inputs. Also, we take additional samples in order to perform validation of the GP model, again from a Latin hypercube design. A GP is then trained to learn the mapping from the input space to the SNR yielded by DBP. It should be noted that the GP is trained on SNR in linear units. This learned function is a computationally cheap surrogate for the entire DBP chain. Having validated the accuracy of the GP by comparing its predictions with the validation samples, we then use the SHGO algorithm to find the global and local optima of the GP surrogate model, which correspond to the physical layer inputs that maximize the SNR after DBP. This optimizer calls the GP model thousands of times. This method is described below in Algorithm 1.

**Algorithm 1.** DBP parameter optimization process

SET UP: Let  $X = \{X_i = \{x_1, x_2, \dots, x_j, \dots, x_m\} : j_L \leq x_j \leq j_U, 1 \leq i < \infty, 1 \leq j \leq m\}$  be the continuous sample space containing the samples  $X_i$  consisting of a set of  $m$  physical layer parameters  $x_j$  with specified ranges bounded by upper and lower limits  $j_U$  and  $j_L$ , respectively.

Let  $P_{\text{DBP}}$  be the launch power of the trace for which we perform the estimation, and  $P_p$  be the launch power of the  $p$ th trace used in the validation of the estimated parameters.

Step 1: Draw training validation samples:

**for**  $k := [1, \dots, n_{\text{sam}}]$  **do**

    Draw sample  $X_k := \text{LHD}(X)$ —draw training input samples using Latin hypercube design

$\text{SNR}_k := \text{DBP}(X_k, P_{\text{DBP}})$ —compute SNR target for each sample using DBP chain

**end for**

**for**  $l := [1, \dots, n_{\text{val}}]$  **do**

    Draw sample  $X_l := \text{LHD}(X)$ —draw validation input samples using Latin hypercube design

$\text{SNR}_l := \text{DBP}(X_l, P_{\text{DBP}})$ —compute SNR target for each sample using DBP chain

**end for**

Step 2: Fit GP hyperparameters via maximum likelihood using training data

Step 3: Validate GP—compare validation targets to GP predictions with error metrics:

$$\text{MAE} = \sum_{l=1}^{n_{\text{val}}} |\text{SNR}_l - \text{GP}(X_l)|$$

$$\text{RMSE} = \sum_{l=1}^{n_{\text{val}}} \sqrt{(\text{SNR}_l - \text{GP}(X_l))^2}$$

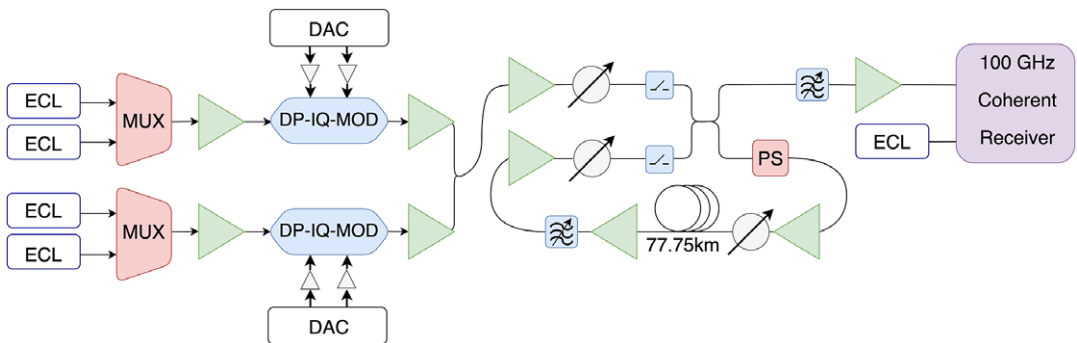
Step 4: Perform global optimization of GP inputs via SHGO to find optima  $X_1^*, \dots, X_n^*$ , such that  $\text{GP}(X^*) > \text{GP}(X) \forall X$

**5. Experimental System**

**5.1. Experimental setup**

The experimental setup consists of transmission of a four-subcarrier 49.5 Gbd polarization division multiplexed DP-256QAM superchannel across 1,010.75 km using a recirculating fiber loop. The experimental setup used is outlined in Figure 1.

Four external cavity lasers (ECLs) with <100 kHz linewidth and 50 GHz spacing were used as sources for two odd and two even subcarriers, which were modulated by two separate dual-polarization IQ

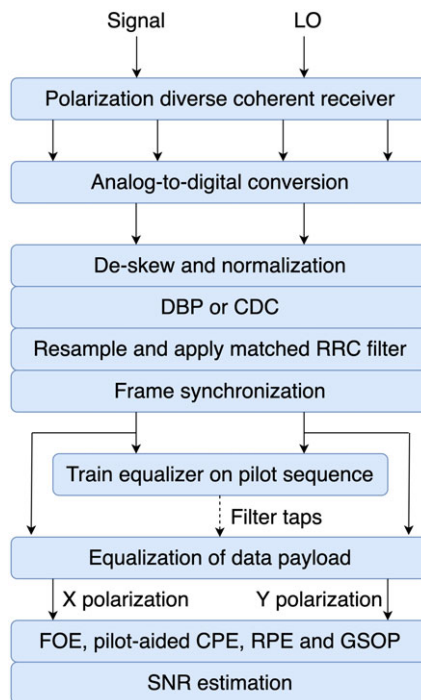


**Figure 1.** Diagram of the experimental setup, consisting of a recirculating fiber loop that was used to transmit a  $4 \times 49.5$  Gbd superchannel over 13 spans of length 77.75 km. DAC, digital-to-analog converter; DP-IQ-MOD, dual-polarization IQ modulator; ECL, external cavity laser; MUX, multiplexer; PS, polarization scrambler.



modulators driven by two 33-GHz 92-GSa/s digital-to-analog converters (DACs). The 256-QAM driving signals were each spectrally shaped by matched root-raised cosine (RRC) filters with 1% roll-off. The frequency response of the transmitter components was compensated by applying digital pre-emphasis to the generated signals. The odd and even signals were modulated with independent data and combined with a 3-dB coupler to form a  $4 \times 49.5$  GBd DP-256QAM superchannel with 200 GHz optical bandwidth. The superchannel was then amplified and launched into a recirculating fiber loop, which comprised of a polarization scrambler (PS), a span of 77.75 km of Corning<sup>®</sup> SMF-28<sup>®</sup> ULL fiber, two variable optical attenuators (VOAs), three erbium-doped fiber amplifiers (EDFAs) with 4.5 dB noise figure and a 2 nm optical bandpass filter to remove noise outside the signal bandwidth. The transmitted superchannel circulated through the loop 13 times, corresponding to a total transmission distance of 1,010.75 km. It was then mixed with a <100 kHz linewidth local oscillator (LO) laser using a 90° hybrid and detected by four balanced photo-detectors with 100 GHz electrical bandwidth. Finally, the received signals were digitized using a 256 GSa/s real-time digital sampling oscilloscope with 100 GHz analog electrical bandwidth.

The offline digital signal processing (DSP) chain was conducted as outlined in Wakayama et al. (2019) using QPSK pilot symbols. Specifically, the pilot symbols are used for equalization and frame synchronization, as well as for carrier phase estimation (CPE). A block diagram of the DSP is given in Figure 2. After the signal is received, it is converted to the electrical domain using an analog-to-digital converter (ADC). It is then deskewed and normalized. Following this, if desired, DBP is performed on the entire four-subcarrier superchannel. If we do not perform DBP, that is, to quantify the DBP gain, we perform chromatic dispersion compensation (CDC) instead. Following either of these blocks, we then shift the signal by 25 GHz to move the third subchannel to baseband. This is then followed by resampling to two samples per symbol, RRC-matched filtering to extract the third subchannel, and then frame synchronization. The filtered subchannel is equalized by an adaptive FIR filter, which utilizes two algorithms for



**Figure 2.** Schematic of the DSP chain, adapted from Wakayama et al. (2019). CDC, chromatic dispersion compensation; CPE, carrier-phase estimation; DBP, digital backpropagation; FOE, frequency offset estimation; GSOP, Gram–Schmidt orthogonalization procedure; LO, local oscillator; RPE, residual phase estimation; RRC, root-raised cosine.

training the equalizer filter weights; recursive least squares for coarse-tuning and least mean squares for fine-tuning, which are both trained on the received pilot symbols (Wakayama et al., 2019). Having trained the equalizer, the optimized filter weights are then used to equalize the payload symbols. The signal is then demultiplexed into its two constituent orthogonal polarizations before we perform frequency offset estimation and CPE. The CPE algorithm is implemented as a two-step process, consisting of initial linear interpolation followed by more detailed residual phase estimation (RPE), as outlined in detail in Wakayama et al. (2019). Following this, we apply the Gram–Schmidt organization procedure (GSOP) and then estimate the SNR. This estimation consists of taking the noisy symbols outputted by the DSP chain and calculating the SNR as

$$\text{SNR} = \frac{\mathbb{E}[|X|^2]}{\mathbb{E}[|X - Y|^2]}, \quad (12)$$

where  $\mathbb{E}[\cdot]$  denotes the expectation,  $X$  are the transmitted symbols, and  $Y$  is the received symbols after DSP, and a  $k$ -means-based algorithm was used to extract the constellation centroid.

## 5.2. Experimental a priori knowledge

The data sheet of the Corning® SMF-28® ULL fiber used in this experiment provides us with some a priori knowledge. First,  $\alpha$  is stated to have a maximum value of  $0.17 \text{ dB km}^{-1}$ . Thus, we consider a range of  $\alpha$  values between  $0.14$  and  $0.17 \text{ dB km}^{-1}$ . As GPs are a kernel-based method, they estimate the similarity between neighboring data to infer the underlying function. As a result, the model accuracy is limited near the edge of the range of training data, as there are fewer neighboring data. Thus, we want to avoid the ground truth value lying too close to the edge of the range, where the model accuracy will suffer. If any of the estimated parameters were observed to fall close to any of the corresponding a priori bounds, the bounds should be increased. The lower bound in this case is chosen to be lower than the world record of  $0.142 \text{ dB km}^{-1}$  (Hasegawa et al., 2018), to avoid the estimated value falling too close to the lower bound. It was observed that the upper bound was sufficiently conservative, such that the estimated values were not close to this value. Additionally, this fiber type has a specified expected nonlinear refractive index component value  $n_2 = 2.1 \times 10^{-20}$ , however, no range is given. Similarly, the expected fiber effective area is given as  $83 \text{ } \mu\text{m}^2$ , with no range or tolerance. However, the value of  $\gamma$  has significant uncertainty for deployed fibers, which may suffer from aging effects, or in some cases, the fiber type may even be unknown (Seve et al., 2019). Moreover, variations during the manufacturing process can modify the value of  $\gamma$ . Thus, we define a broad range for  $\gamma$ , from  $0.9$  to  $1.2 \text{ W}^{-1}\text{km}^{-1}$ . As for the dispersion, we have stronger a priori knowledge as it is possible to obtain a measured estimate of the dispersion from the adaptive equalizer in the DSP chain (Corsini et al., 2013, eq. 4). Thus, we use a tight range around this estimated value, from  $15.9$  to  $16.1 \text{ ps nm}^{-1}\text{km}^{-1}$ , to allow the SHGO optimizer to tweak the dispersion to the value that yields optimal gain. It should be noted that, for deployed systems, the uncertainty surrounding these parameters is significantly higher than in laboratory experiments (Pointurier, 2017, 2021).

Thus, as well as being a useful research tool that avoids time-consuming and nonoptimal grid search, the proposed method provides a systematic way to obtain a high accuracy characterization of deployed systems from broad a priori ranges. To demonstrate the feasibility of using this methodology in deployed networks, we assume relatively broad a priori bounds, defined from our approximate a priori knowledge.

## 6. Parameter Estimation from Simulated Data

### 6.1. Parameter estimates

We first test the efficacy of this technique using simulated traces, generated for a simulated system using the SSFM. This system was designed to emulate the experimental system outlined in Figure 1, with a

**Table 1.** *A priori physical layer parameter ranges.*

| Parameter | Simulated ground truth | Range          | Unit   |
|-----------|------------------------|----------------|--|
| $D$       | 16.00                  | $U[15.9,16.1]$ | $\text{ps}\cdot\text{nm}^{-1}\text{km}^{-1}$ |
| $\gamma$  | 1.03                   | $U[0.9,1.2]$   | $\text{W}^{-1}\text{km}^{-1}$                |
| $\alpha$  | 0.156                  | $U[0.14,0.17]$ | $\text{dB}\cdot\text{km}^{-1}$               |

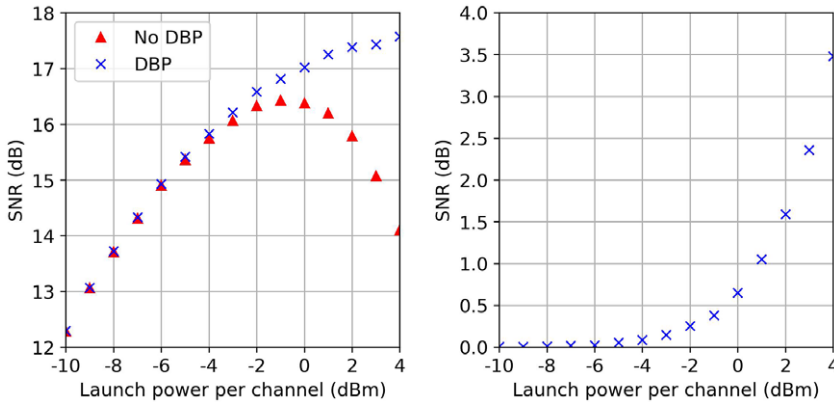
superchannel consisting of four subcarriers which are used to transmit a 49.5 GBd DP-256QAM signal over 13 spans of length 77.75 km, for a total transmitted distance of 1,010.75 km. As we know exactly the parameters that we used to generate the traces, this provides us with a ground truth that we can use to verify the accuracy of the physical layer parameter estimation. The ground truth values of  $D$ ,  $\gamma$ , and  $\alpha$  are shown in Table 1, along with the a priori search ranges used. The same ranges were used for the experimental traces. It should be noted that we model the effects of polarization mode dispersion (PMD) in our simulation, as outlined above, using a value of  $0.04 \text{ ps km}^{-1/2}$ . This value is the link design value taken from the manufacturer specification sheet of the Corning<sup>®</sup> SMF-28<sup>®</sup> ULL fiber used in this experiment, a statistical upper limit on the total link PMD.

During training of the GP, we take 100 additional samples from the parameter space and record the SNR after application of DBP and DSP, for verification of the trained models. We compare the predictions of the GP models at launch powers of 1, 2, 3, and 4 dBm with the corresponding DBP chain SNRs. The mean absolute error (MAE) values were 0.009, 0.016, 0.010, and 0.012 and the MSE values were 0.013, 0.022, 0.013, and 0.017, respectively for the traces at 1, 2, 3, and 4 dBm. As the GP is trained on SNR in linear units, these errors are given in terms of linear SNR. This indicates that 500 samples are sufficient for training a high-accuracy GP model, which learns the functional mapping from the physical layer parameters to the SNR gain with a small prediction error. This is crucial, as the GP is used as a computationally cheap surrogate model for the DBP chain, which is used by the SHGO optimizer to find the parameters that maximize the DBP gain. Exploring the precise minimum number of samples required for a good estimation forms part of the planned future work. It is important to note the motivation for using the GP to learn a surrogate model for the SNR after DBP in conjunction with SHGO, rather than simply using SHGO to optimize the SNR after DBP directly. Specifically, the bottleneck in terms of performance is running the DBP chain. It was observed that training the GP surrogate model required running the DBP chain 500 times, whereas SHGO required approximately 1,500 function evaluations to converge. Thus, using SHGO directly on the DBP chain would be more computationally expensive, as the DBP chain would need to be run a larger number of times. In addition, a trained GP surrogate model is itself a useful tool, providing an interpretable model of the SNR as a function of  $D$ ,  $\gamma$ , and  $\alpha$  that can be used to perform a sensitivity analysis. This is explored further in Section 7.2.

We use the simulated traces at launch powers per channel of 1, 2, 3, and 4 dBm to perform the estimation, and the optimal parameter values obtained are shown in Table 2. For the traces at 1, 2, and 3 dBm, one global optimum was found, whereas for the trace at 4 dBm two optima with equivalent

**Table 2.** *Parameter estimates for simulated traces.*

| Power (dBm) | $D$ ( $\text{ps nm}^{-1}\text{km}^{-1}$ ) | $\gamma$ ( $\text{W}^{-1}\text{km}^{-1}$ ) | $\alpha$ ( $\text{dB km}^{-1}$ ) |
|-------------|---|--|----------------------------------|
| 1           | 16.00                                     | 1.01                                       | 0.155                            |
| 2           | 16.00                                     | 1.02                                       | 0.157                            |
| 3           | 16.00                                     | 1.02                                       | 0.156                            |
| 4           | 16.00                                     | 1.02                                       | 0.156                            |
| 4           | 16.00                                     | 1.03                                       | 0.157                            |

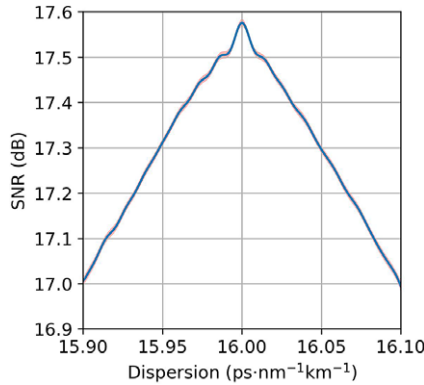


**Figure 3.** Simulated traces comparison with and without DBP using parameters  $D = 16.00 \text{ ps nm}^{-1} \text{ km}^{-1}$ ,  $\gamma = 1.02 \text{ W}^{-1} \text{ km}^{-1}$ , and  $\alpha = 0.156 \text{ dB km}^{-1}$ , estimated using Algorithm 1 using the simulated trace at a launch power of 4 dBm. The SNR versus launch power curves with and without DBP are shown, as well as the DBP gain at each launch power.

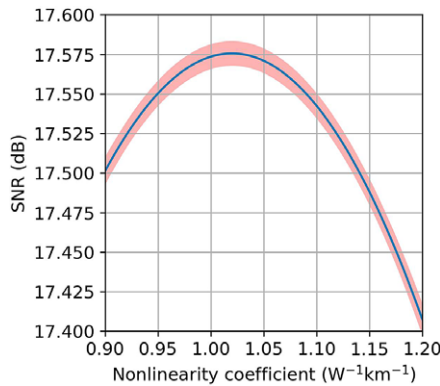
performance were found. These traces were chosen as they correspond to the four values around the peak in the experimental system after DBP is applied. We can see that we are able to correctly estimate the ground truth value of  $D$  to a precision of 4 significant figures. Moreover, we are able to estimate  $\gamma$  with a maximum error of  $0.02 \text{ W}^{-1} \text{ km}^{-1}$ . As for  $\alpha$ , the estimation is accurate to within an error of  $0.001 \text{ dB km}^{-1}$ . Thus, this technique is able to obtain highly accurate estimates of  $D$ ,  $\alpha$ , and  $\gamma$  from broad a priori parameter ranges for a simulated system. Sources of error in the simulated system include PMD, numerical errors introduced by the DSP and interpolation errors introduced by the GP surrogate model. As a representative example, Figure 3 shows the comparison of the SNR after decoding the simulated traces both with and without DBP using the parameters  $D = 16.00 \text{ ps nm}^{-1} \text{ km}^{-1}$ ,  $\gamma = 1.02 \text{ W km}^{-1}$ , and  $\alpha = 0.156 \text{ dB km}^{-1}$ , estimated using the trace at 4 dBm. We show the achieved SNR in Figure 3a and the SNR gain at each launch power in Figure 3b. The gain in maximum SNR was 1.14 dB with all five sets of parameters found. Moreover, the sum of the gain for each trace, that is, the difference between the performance with and without DBP for each trace, was 10.10 dB for the parameters found using all traces. Thus, as expected, these parameter sets give practically equivalent performance.

## 6.2. GP model parameter sweeps

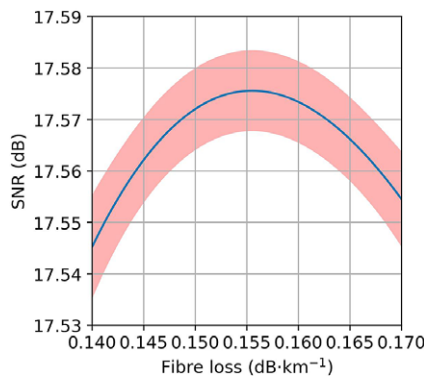
In order to further interpret the trained GP model, we perform parameter sweeps and plot the predictive mean and predictive variance. Specifically, this is done by sweeping each parameter in turn with the others held fixed at the optimum found, for each of the GP models trained. The results of these sweeps are shown for  $D$ ,  $\gamma$ , and  $\alpha$  in Figures 4–6, respectively. The ranges for each sweep are the same as those defined in Table 1, meaning that we are sweeping across the full a priori search range learned by each GP. A confidence region corresponding to one predictive standard deviation is also shown, where the predictive standard deviation is simply given by the square root of the predictive variance of the GP model. From the confidence bands, we can see that the GP model has the lowest uncertainty with respect to  $D$ , followed by  $\gamma$  and  $\alpha$ . Additionally, from the variation of the predictive mean of the GP across the range, we can see that  $D$  has the largest effect on the SNR after DBP, followed by  $\gamma$  and then  $\alpha$ . Specifically, the ranges of variation are 0.58, 0.17, and 0.03 dB for  $D$ ,  $\gamma$ , and  $\alpha$ , respectively. We also consider these values as a fraction of the DBP gain achieved for the trace used in the estimation. For the trace at 4 dBm shown, the ranges for  $D$ ,  $\gamma$ , and  $\alpha$  correspond to 16.7, 4.8, and 0.8% of the achieved gain respectively. However, it is important to contextualize these results—the objectives of this methodology are to estimate the parameters  $D$ ,  $\alpha$ , and  $\gamma$  at the receiver, as well as to provide a methodology for systematically finding the optimal parameters in



**Figure 4.** *D* GP model sweep, in which we sweep *D* across the a priori search range with  $\gamma=1.03 \text{ W}^{-1}\text{km}^{-1}$  and  $\alpha=0.156 \text{ dB km}^{-1}$ . The hyperparameters for this model are  $h_1=21.5$ ,  $\mathbf{h}_2=(0.01 \text{ psnm}^{-1}\text{ km}^{-1}, 0.84 \text{ W}^{-1}\text{ km}^{-1}, 0.11 \text{ dBkm}^{-1})$ , and  $h_3=0.01$ .



**Figure 5.**  $\gamma$  GP model sweep, in which we sweep  $\gamma$  across the a priori search range with  $D=16.00 \text{ ps nm}^{-1}\text{ km}^{-1}$  and  $\alpha=0.156 \text{ dB km}^{-1}$ . The hyperparameters for this model are  $h_1=21.5$ ,  $\mathbf{h}_2=(0.01 \text{ psnm}^{-1}\text{ km}^{-1}, 0.84 \text{ W}^{-1}\text{ km}^{-1}, 0.11 \text{ dBkm}^{-1})$ , and  $h_3=0.01$ .



**Figure 6.**  $\alpha$  GP model sweep, in which we sweep  $\alpha$  across the a priori search range with  $D=16.00 \text{ ps nm}^{-1}\text{ km}^{-1}$  and  $\gamma=1.03 \text{ W}^{-1}\text{ km}^{-1}$ . The hyperparameters for this model are  $h_1=21.5$ ,  $\mathbf{h}_2=(0.01 \text{ psnm}^{-1}\text{ km}^{-1}, 0.84 \text{ W}^{-1}\text{ km}^{-1}, 0.11 \text{ dBkm}^{-1})$ , and  $h_3=0.01$ .

terms of DBP gain. Even if the effect on the DBP is small for  $\alpha$  and  $\gamma$ , estimating these parameters with high accuracy is crucial for building high-accuracy network models, which in turn can be used to unlock extra capacity by reducing network margins.

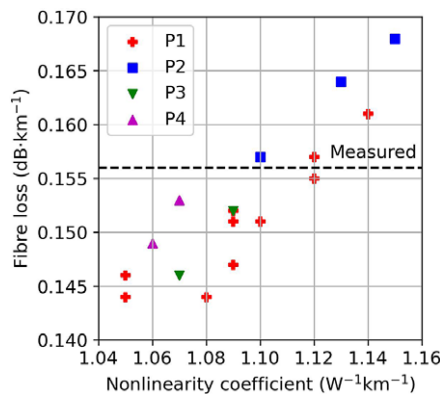
## 7. Parameter Estimation from Experimental Data

### 7.1. Parameter estimates

Having confirmed the efficacy of Algorithm 1 for simulated traces, we demonstrate the DBP parameter estimation process for experimentally measured traces, using the experimental system outlined in Figure 1. We perform the parameter estimation process outlined in Algorithm 1 for the experimental traces measured at launch powers 1.22, 2.27, 3.24, and 4.25 dBm, denoted  $P_1$ ,  $P_2$ ,  $P_3$ , and  $P_4$ , respectively. These particular traces were chosen due to their proximity to the optimum launch power, once DBP is applied.

In order to validate the GP model learned, we again draw an additional 100 validation samples from the DBP chain and compare these samples to the predictions of the trained GP model at these sample inputs. This yielded a linear SNR mean absolute error (MAE) of 0.015, 0.032, 0.012, and 0.010 for the GPs trained on traces at  $P_1$ ,  $P_2$ ,  $P_3$ , and  $P_4$ , respectively. The corresponding linear SNR RMSE values were 0.020, 0.049, 0.019, and 0.013, respectively. Thus, the GP trained on the trace at  $P_2$  has the highest validation error, whereas the model trained at  $P_4$  has the lowest validation error. The magnitude of the validation errors are very small in practical terms, indicating that 500 samples is sufficient for the GP to learn an accurate probabilistic mapping from  $D$ ,  $\gamma$ , and  $\alpha$  to the SNR after DBP.

In Figure 7, we show the estimation results for each trace. SHGO returned 10, 3, 2, and 2 local optima respectively for the traces at  $P_1$ ,  $P_2$ ,  $P_3$ , and  $P_4$ . These local optima were observed to give equivalent performance in terms of DBP gain, to within a maximum tolerance of 0.006 dB across all the experimentally measured traces. Specifically, this tolerance was calculated by performing DBP for all the experimental traces with each set of parameters and calculating the difference between the mean SNR after DBP is applied for the best and worst-performing parameter sets. This highlights the limitations of considering approaches such as gradient descent, which are prone to getting stuck in a given local optima. It should be noted that  $D$  was estimated to be  $16.05 \text{ ps nm}^{-1} \text{ km}^{-1}$  for all optima using traces  $P_2$ ,  $P_3$ , and  $P_4$ , whereas  $D = 16.06 \text{ ps nm}^{-1} \text{ km}^{-1}$  for all optima estimated using the trace measured at  $P_1$ . It is clear that there is a correlation between  $\alpha$  and  $\gamma$ : we can achieve equivalent performance with higher  $\gamma$  if  $\alpha$  is increased, and vice versa. The reason for using a search range that extends down to  $0.140 \text{ dB km}^{-1}$  is that



**Figure 7.** Plot of local optima obtained from the experimental traces at launch powers  $P_1 = 1.22$ ,  $P_2 = 2.27$ ,  $P_3 = 3.24$ , and  $P_4 = 4.25$  dBm. All optima have equivalent performance in terms of DBP gain, to a maximum tolerance of 0.006 dB. For traces at  $P_2$ ,  $P_3$ , and  $P_4$ ,  $D = 16.05 \text{ ps nm}^{-1} \text{ km}^{-1}$ , whereas for  $P_1$ ,  $D = 16.06 \text{ ps nm}^{-1} \text{ km}^{-1}$ . The independently measured fiber loss of  $\alpha > 0.156 \text{ dB km}^{-1}$  is indicated on the plot.

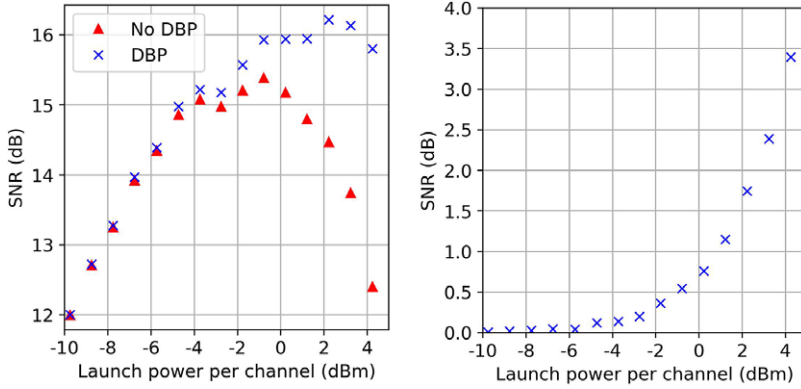
**Table 3.** Most likely parameter estimates for the experimental system.

| Power (dBm) | $D$ (ps nm <sup>-1</sup> km <sup>-1</sup> ) | $\gamma$ (W <sup>-1</sup> km <sup>-1</sup> ) | $\alpha$ (dB km <sup>-1</sup> ) |
|-------------|---|--|---------------------------------|
| 1.22        | 16.06                                       | 1.09   | 0.152                           |
| 1.22        | 16.06                                       | 1.09   | 0.151                           |
| 1.22        | 16.06                                       | 1.10   | 0.151                           |
| 2.27        | 16.05                                       | 1.10   | 0.157                           |
| 3.24        | 16.05                                       | 1.09   | 0.152                           |
| 4.25        | 16.05                                       | 1.07   | 0.153                           |

GPs are a kernel-based method. Thus, the accuracy of the model will be much lower at the boundaries of the training data range, and so we want to avoid the optimum falling too close to the edge of the range, where the model accuracy will be reduced. If the optimum fell near the edge of the range, it would indicate to the user that the a priori ranges should be increased, such that the optimum is not too close to the edge of the range. In addition, we indicate that lower values of  $\gamma$  are more likely to reflect the true system values. This is because the fiber specifications list an effective area of 83  $\mu\text{m}^2$  and  $n_2 = 2.1 \times 10^{-20}$ , which corresponds to  $\gamma = 1.03 \text{ W}^{-1}\text{km}^{-1}$ . Thus, we would expect parameter sets with a smaller deviation from this value to be more likely. For clarity, we show the effective area values for each  $\gamma$  value in Figure 7. Thus, we consider parameter sets with the lowest value of  $\gamma$  in the feasible  $\alpha$  range to be the most likely parameter estimates. For the traces measured at launch powers  $P_2$ ,  $P_3$ , and  $P_4$ , there is one clear choice for the most likely set of parameters. For the trace measured at  $P_1$ , there are three traces which we consider to be equally likely. These parameter best estimates are shown in Table 3 for each of the traces, including the three estimates from the trace at  $P_1$ .

Additionally, estimated parameter values that are more consistent across multiple traces are more likely to be correct—we can see that a grouping of similar values emerges for all the traces in the range for  $\gamma$  values from 1.07 to 1.10  $\text{W}^{-1}\text{km}^{-1}$  and  $\alpha$  values from 0.151 to 0.157  $\text{dB km}^{-1}$ . In deployment, parameter estimates from measurements of multiple lightpaths occupying the same fiber segment could be used to cross-reference and identify more likely parameter estimates. We also notice that the number of local optima is observed to increase as we decrease the launch power. This is due to the fact that, at higher launch power, we have more nonlinear interference noise. Thus, the sensitivity of the SSFM to  $\gamma$  increases, making it easier to estimate, as fewer values of  $\gamma$  have equivalent performance. As a demonstrative example, we performed the parameter estimation process using the trace measured at  $-9.75$  dBm, which returned 129 local optima. In this linear regime,  $\gamma$  has negligible effect, so SHGO finds a large number of local optima with equivalent performance.

To consider how these parameters perform across all the experimentally measured traces, we record the DBP gain, both in terms of the maximum SNR achieved and the SNR gain for each trace. All sets of estimated parameters yield a gain in maximum SNR of 0.83 dB. As for the total gain, taken as the sum of individual gain at each trace, gain values of 10.92, 10.95, 10.94, and 10.95 dB were achieved using the parameters estimated using the traces at  $P_1$ ,  $P_2$ ,  $P_3$ , and  $P_4$ , respectively. Here, 10.92 dB was achieved for each of the local optima estimated using the trace at  $P_1$ . Thus, there is a small difference in the achieved gain, with the parameters estimated using the  $P_4$  trace giving the highest gain, and the parameters estimated using the trace at  $P_1$  giving the lowest gain. As an representative example, Figure 8 shows the DBP gain given by the parameters estimated for the trace at  $P_4$ , for each of the experimental traces. The SNR achieved is compared with and without DBP in Figure 8a and the SNR gain is presented in Figure 8b. As in simulation, sources of error include PMD, DSP errors, and inaccuracies in the GP surrogate model. For the experimental case, there are additional sources of error arising from the nonideal nature of the experimental components. These sources of error lead to increased variation in the estimation of the physical layer parameters  $D$ ,  $\gamma$ , and  $\alpha$ , as compared to the simulated case. Moreover, due to these error

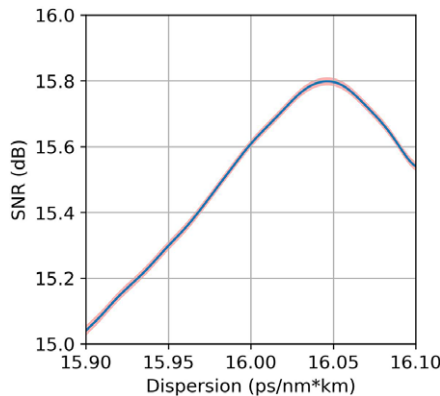


**Figure 8.** Experimental traces comparison with and without DBP using parameters  $D=16.05 \text{ ps nm}^{-1}\text{km}^{-1}$ ,  $\gamma=1.07 \text{ W}^{-1}\text{km}^{-1}$ , and  $\alpha=0.153 \text{ dB km}^{-1}$ , estimated using Algorithm 1 using the experimental trace at a launch power of 4.25 dBm. The SNR versus launch power curves with and without DBP are shown, as well as the DBP gain at each launch power.

sources, in the experimental case we obtain multiple local optima with equivalent gain for all traces, whereas in simulation, only one trace returned more than one optimum.

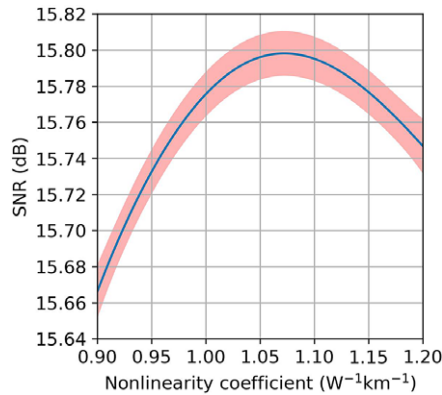
**7.2. GP model parameter sweeps**

We again perform parameter sweeps of the trained GP model for the experimental case, sweeping each parameter in turn, with the other two parameters fixed at the optimal values. These sweeps are shown in Figures 9–11 for  $D$ ,  $\gamma$ , and  $\alpha$ , respectively. The hyperparameters found to maximize the marginal likelihood during the GP fitting process are also reported. As with the simulated traces,  $D$  has the largest variation in SNR across the a priori search range, followed by  $\gamma$  and  $\alpha$ . Thus, as in simulation,  $D$  is the most important parameter in terms of maximizing the DBP gain, followed by  $\gamma$  and then  $\alpha$ . Specifically, the ranges found are 0.76, 0.13, and 0.02 dB for  $D$ ,  $\gamma$ , and  $\alpha$ , respectively. We also consider these values as a fraction of the DBP gain achieved for the trace used in the estimation. For the trace at 4.25 dBm shown, the ranges for  $D$ ,  $\gamma$ , and  $\alpha$  correspond to 22.4, 3.9, and 0.7% of the achieved gain respectively. Also, we observe that the predictive variance is smallest for  $D$ , followed by  $\gamma$ , and then  $\alpha$ . An intuitive explanation

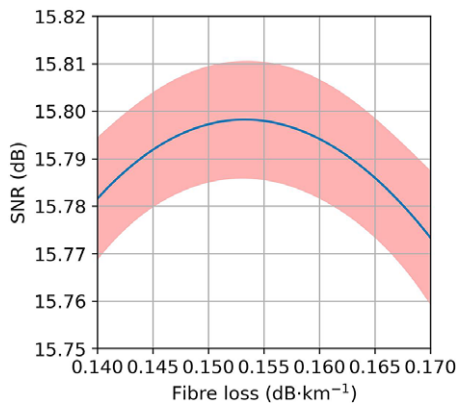


**Figure 9.**  $D$  GP model sweep, in which we sweep  $D$  across the a priori search range with  $\gamma=1.07 \text{ W}^{-1}\text{km}^{-1}$  and  $\alpha=0.153 \text{ dB km}^{-1}$ . The hyperparameters for this model are  $h_1=13.6$ ,  $\mathbf{h}_2=(0.03 \text{ ps nm}^{-1}\text{km}^{-1}, 0.26 \text{ W}^{-1}\text{km}^{-1}, 0.08 \text{ dB km}^{-1})$ , and  $h_3=0.01$ .





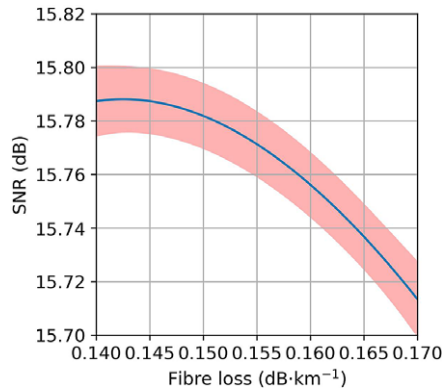
**Figure 10.**  $\gamma$  GP model sweep, in which we sweep  $\gamma$  across the a priori search range with  $D = 16.05 \text{ ps nm}^{-1} \text{ km}^{-1}$  and  $\alpha = 0.153 \text{ dB km}^{-1}$ . The hyperparameters for this model are  $h_1 = 13.6$ ,  $\mathbf{h}_2 = (0.03 \text{ ps nm}^{-1} \text{ km}^{-1}, 0.26 \text{ W}^{-1} \text{ km}^{-1}, 0.08 \text{ dB km}^{-1})$ , and  $h_3 = 0.01$ .



**Figure 11.**  $\alpha$  GP model sweep, in which we sweep  $\alpha$  across the a priori search range with  $D = 16.05 \text{ ps nm}^{-1} \text{ km}^{-1}$  and  $\gamma = 1.07 \text{ W}^{-1} \text{ km}^{-1}$ . The hyperparameters for this model are  $h_1 = 13.6$ ,  $\mathbf{h}_2 = (0.03 \text{ ps nm}^{-1} \text{ km}^{-1}, 0.26 \text{ W}^{-1} \text{ km}^{-1}, 0.08 \text{ dB km}^{-1})$ , and  $h_3 = 0.01$ .

for this is that the larger the sensitivity of the SNR to a given parameter, the more confident the GP model is in the learned underlying function.

GP sweeps can also be used to illuminate the issues associated with manually optimizing  $D$ ,  $\alpha$ , and  $\gamma$  in terms of DBP. Crucially, if one or more of the other parameters is set to a nonoptimal value, the sweeps can change significantly, resulting in an incorrect optimum value. This is because  $\alpha$  and  $\gamma$  jointly contribute to the nonlinear noise, meaning that there will not be uniquely optimal values of  $\alpha$  and  $\gamma$  with respect to the SNR after DBP. This can be seen in Chen and Shieh (2010, eqs. 31–33). Thus, performing a one-at-a-time grid search is likely to lead to an incorrect optimum. To demonstrate the benefits of the multi-parameter optimization framework proposed as compared to single-parameter-at-a-time grid search, Figure 12 shows parameter sweeps over  $\alpha$  for the 4.25 dBm experimental trace with  $\gamma$  set to the nonoptimal value of  $\gamma = 1.00$ . We can see that for  $\gamma = 1.00 \text{ W}^{-1} \text{ km}^{-1}$ , the optimum value of  $\alpha$  occurs at  $0.143 \text{ dB km}^{-1}$ , with a lower maximum SNR than the global optimum reported in Table 3. This highlights the limitations of doing a one-at-a-time grid search—the optimum found in one dimension is not in general independent of the values of the other parameters.



**Figure 12.** Demonstrative example a GP model sweep, in which we sweep  $a$  across the a priori search range with  $D = 16.05 \text{ ps nm}^{-1} \text{ km}^{-1}$  and  $\gamma = 1.00 \text{ W}^{-1} \text{ km}^{-1}$ —a deliberately nonoptimal value. This highlights the limitations of one-at-a-time grid search optimization: the optimal value for a single parameter sweep depends on the values of the other parameters.

It is clear that having an incorrect value of  $\gamma$  results in an incorrect optimum at  $0.143 \text{ dB km}^{-1}$ , at a lower SNR than that obtained using SHGO.

## 8. Conclusions and Future Work

We propose a methodology for optimizing physical layer parameters in optical fiber communication networks with respect to DBP gain, which has significant implications for improving system performance and reducing network margins. By systematically and efficiently optimizing the physical layer parameters with respect to the DBP gain, the proposed approach provides valuable insights into complex interdependencies of the system whilst maximizing the SNR. The accuracy of this method was demonstrated through simulation and experimental testing, achieving optimal values of dispersion, nonlinearity, and attenuation parameters with maximum errors of  $0.02 \text{ W}^{-1} \text{ km}^{-1}$  and  $0.01 \text{ dB km}^{-1}$ , respectively. Moreover, the proposed methodology addresses the limitations of traditional one-at-a-time grid search and gradient-based methods and emphasizes the interpretability of the technique. The potential impact of our approach extends beyond optical networks and has applications in various engineering systems where optimization of system parameters is required. Future work will focus on reducing offline computational complexity and exploring how the method can be extended to consider finer-grain mapping of the parameters. Specifically, the extension of this methodology to optimizing the distributed physical layer parameters on a span-by-span basis will be considered. This will deliver insights about the variation of the parameters along the link, at the cost of increasing the dimensionality of the GP surrogate model and the SHGO optimization. In addition, the use of RVMS instead of GPs to learn the mapping from parameters to SNR will be explored. Specifically, the increased flexibility in basis functions will be weighed against the associated cost of reduction of interpretability of the predictive variance.

### Abbreviations

|           |   |
|-----------|---|
| ADC       | analog-to-digital converter                       |
| CDC       | chromatic dispersion compensation                 |
| CPE       | carrier phase estimation                          |
| DAC       | digital-to-analog converter                       |
| DBP       | digital backpropagation                           |
| DP-IQ-MOD | dual-polarization IQ modulator                    |
| DP-QAM    | duel polarization quadrature amplitude modulation |
| DSP       | digital signal processing                         |

|      |   |
|------|---|
| ECL  | external cavity lasers                  |
| EDFA | Erbium-doped fiber amplifier            |
| GP   | Gaussian process                        |
| GPR  | Gaussian process regression             |
| GSOP | Gram–Schmidt organization procedure     |
| LO   | local oscillator                        |
| ML   | machine learning                        |
| MSE  | mean squared error                      |
| MUX  | multiplexer                             |
| NLSE | nonlinear Schrodinger equation          |
| NN   | neural network                          |
| PMD  | polarization mode dispersion            |
| PS   | polarization scrambler                  |
| QoT  | quality of transmission                 |
| RPE  | residual phase estimation               |
| RRC  | root-raised cosine                      |
| SHGO | simplicial homology global optimization |
| SMF  | single mode fiber                       |
| SNR  | signal-to-noise ratio                   |
| SSFM | split-step Fourier method               |
| VOA  | variable optical attenuators            |

**Acknowledgments.** The authors thank Corning for the loan of the Corning<sup>®</sup> SMF-28<sup>®</sup> ULL Fiber. The research is carried out as a part of the PhD thesis work of J.W.N. with the thesis title “Machine Learning for Optical Fibre Communication Systems,” University of Cambridge, 2023.

**Author contribution.** Conceptualization: All authors; Data curation: J.W.N., E.S., R.S.; Data visualization: J.W.N., E.S., R.S.; Investigation: All authors; Methodology: All authors; Supervision: L.G., P.B., S.J.S.; Writing—original draft: J.W.N.; Writing—review and editing: All authors. All authors approved the final submitted draft.

**Competing interest.** The authors declare no competing interests exist.

**Data availability statement.** Data underlying this research is available at <https://doi.org/10.17863/CAM.94561>.

**Funding statement.** This work was supported by the Engineering and Physical Sciences Research Council Programme Grant TRANSNET (grant number EP/R035342/1).

## References

- Agrawal GP** (2019) *Nonlinear Fiber Optics*. Amsterdam: Elsevier.
- Bishop CM** (2006) *Pattern Recognition and Machine Learning*. Singapore: Springer.
- Bitachon BI, Ghazisaeidi A, Baeuerle B, Eppenberger M and Leuthold J** (2020) Deep learning based digital back propagation with polarization state rotation & phase noise invariance. In *2020 Optical Fiber Communications Conference and Exhibition (OFC)*. San Diego, CA: IEEE, pp. 1–3.
- Bouda M, Oda S, Vassilieva O, Miyabe M, Yoshida S, Katagiri T, Aoki Y, Hoshida T and Ikeuchi T** (2018) Accurate prediction of quality of transmission based on a dynamically configurable optical impairment model. *Journal of Optical Communications and Networking* 10(1), A102–A109.
- Bütler RM, Häger C, Pfister HD, Liga G and Alvarado A** (2020) Model-based machine learning for joint digital backpropagation and pmd compensation. *Journal of Lightwave Technology* 39(4), 949–959.
- Candela JQ and Hansen LK** (2004) *Learning with Uncertainty—Gaussian Processes and Relevance Vector Machines*. Copenhagen, Denmark: Technical University of Denmark, pp. 1–152.
- Chen X and Shieh W** (2010) Closed-form expressions for nonlinear transmission performance of densely spaced coherent optical OFDM systems. *Optics Express* 18(18), 19039–19054.
- Corsini R, Peracchi A, Matarazzo E, Foggi T, Nijhof J, Meloni G, Poti L, Magri R and Ciaramella E** (2013) Blind adaptive chromatic dispersion compensation and estimation for DSP-based coherent optical systems. *Journal of Lightwave Technology* 31(13), 2131–2139.

- Daub E and Barlow N** (2021) Multi-Output Gaussian Process Emulator. Available at <https://github.com/alan-turing-institute/mogp-emulator> (accessed 7 January 2021).
- Endres SC, Sandrock C and Focke WW** (2018) A simplicial homology algorithm for lipschitz optimisation. *Journal of Global Optimization* 72(2), 181–217.
- Fan Q, Zhou G, Gui T, Lu C and Lau APT** (2020) Advancing theoretical understanding and practical performance of signal processing for nonlinear optical communications through machine learning. *Nature Communications* 11(1), 3694.
- Häger C and Pfister HD** (2018) Nonlinear interference mitigation via deep neural networks. In *Optical Fiber Communications Conference (OFC)*. Washington, DC: Optical Society of America, p. W3A.
- Hasegawa T, Tamura Y, Sakuma H, Kawaguchi Y, Yamamoto Y and Koyan Y** (2018) The first 0.14-dB/km ultra-low loss optical fiber. *SEI Technical Review* 86, 18–22.
- Ip E and Kahn JM** (2008) Compensation of dispersion and nonlinear impairments using digital backpropagation. *Journal of Lightwave Technology* 26, 3416–3425.
- Ives DJ, Chin H-M, Caballero FJV and Savory SJ** (2017) Single channel probe utilizing the EGN model to estimate link parameters for network abstraction. In *2017 European Conference on Optical Communication*. Gothenburg, Sweden: IEEE, pp. 1–3.
- Liu DC and Nocedal J** (1989) On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45 (1–3), 503–528.
- Martino L and Read J** (2021) A joint introduction to Gaussian processes and relevance vector machines with connections to Kalman filtering and other kernel smoothers. *Information Fusion* 74, 17–38.
- Meng F, Yan S, Wang R, Ou Y, Bi Y, Nejabati R and Simeonidou D** (2017) Robust self-learning physical layer abstraction utilizing optical performance monitoring and markov chain Monte Carlo. In *2017 European Conference on Optical Communication*. Gothenburg, Sweden: IEEE.
- Millar DS, Makovejs S, Behrens C, Hellerbrand S, Killey RI, Bayvel P and Savory SJ** (2010) Mitigation of fiber nonlinearity using a digital coherent receiver. *IEEE Journal of Selected Topics in Quantum Electronics* 16(5), 1217–1226.
- Nevin JW, Nallaperuma S and Savory SJ** (2022) Gaussian process-driven history matching for physical layer parameter estimation in optical fiber communication networks. In *Workshop on AI for Design and Manufacturing (ADAM), AAAI Conference on Artificial Intelligence*. Vancouver, BC: AAAI.
- Nevin JW, Nallaperuma S, Shevchenko NA, Li X, Faruk MS and Savory SJ** (2021) Machine learning for optical fiber communication systems: An introduction and overview. *APL Photonics* 6(12), 121101.
- Pointurier Y** (2017) Design of low-margin optical networks. *Journal of Optical Communications and Networking* 9(1), A9–A17.
- Pointurier Y** (2021) Machine learning techniques for quality of transmission estimation in optical networks. *Journal of Optical Communications and Networking* 13(4), B60–B71.
- Poole CD and Favin DL** (1994) Polarization-mode dispersion measurements based on transmission spectra through a polarizer. *Journal of Lightwave Technology* 12(6), 917–929.
- Rasmussen CE and Williams CKI** (2006) *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.
- Sasai T, Nakamura M, Kobayashi T, Kawakami H, Yamazaki E and Kisaka Y** (2021) Revealing Raman-amplified power profile and Raman gain spectra with digital backpropagation. In *Optical Fiber Communications Conference*. San Francisco, CA: IEEE, pp. 1–3.
- Sasai T, Nakamura M, Okamoto S, Hamaoka F, Yamamoto S, Yamazaki E, Nishizawa AMH and Kisaka Y** (2020) Simultaneous detection of anomaly points and fiber types in multi-span transmission links only by receiver-side digital signal processing. In *Optical Fiber Communications Conference (OFC)*. San Diego, CA: Optica Publishing Group, pp. 1–3.
- Sasai T, Nakamura M, Yamazaki E, Yamamoto S, Nishizawa H and Kisaka Y** (2020) Digital backpropagation for optical path monitoring: Loss profile and passband narrowing estimation. In *European Conference on Optical Communications (ECOC)*. Brussels, Belgium: IEEE, pp. 1–4.
- Seve E, Pesic J, Delezoide C, Bigo S and Pointurier Y** (2018) Learning process for reducing uncertainties on network parameters and design margins. *Journal of Optical Communications and Networking* 10(2), A298–A306.
- Seve E, Pesic J, Delezoide C, Giorgetti A, Sgambelluri A, Sambo N, Bigo S and Pointurier Y** (2019) Automated fiber type identification in sdn-enabled optical networks. *Journal of Lightwave Technology* 37(7), 1724–1731.
- Shao J, Liang X and Kumar S** (2014) Comparison of split-step Fourier schemes for simulating fiber optic communication systems. *IEEE Photonics Journal* 6(4), 1–15.
- Sillekens E, Yi W, Semrau D, Ottino A, Karanov B, Lavery D, Galdino L, Bayvel P, Killey RI, Zhou S, Law K and Chen J** (2020) Time-domain learned digital back-propagation. In *2020 IEEE Workshop on Signal Processing Systems (SiPS)*. Coimbra, Portugal: IEEE, pp. 1–4.
- Tang D, Wu Z, Sun Z and Qiao Y** (2021) Non-uniform EDFA power map monitoring via power-guided learned digital backpropagation. In *Asia Communications and Photonics Conference*. Shanghai, China: IEEE, pp. 1–3.
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, Van der Plas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F and van Mulbregt P** (2020) Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods* 17(3), 261–272.

- Wakayama Y, Sillekens E, Galdino L, Lavery D, Killey RI and Bayvel P** (2019) Increasing achievable information rates with pilot-based DSP in standard intradyne detection. In *45th European Conference on Optical Communication*. Dublin, Ireland: IEEE.
- Yi X, Huang X, Zhang J, Xu B, Li F and Li Z** (2021) Imbalanced digital back-propagation for nonlinear optical fiber transmissions. *Journal of Lightwave Technology* 39(14), 4622–4628.

---

**Cite this article: Nevin JW, Sillekens E, Sohanpal R, Galdino L, Nallaperuma S, Bayvel P and Savory SJ** (2023). Optical network physical layer parameter optimization for digital backpropagation using Gaussian processes. *Data-Centric Engineering*, 4, e19. doi:10.1017/dce.2023.15