Nonlinear Channel Equalization Using Gaussian Processes Regression in IMDD Fiber Link

Xiang Li, Yixin Zhang, Desheng Li, Perry Ping Shum, *Senior Member, IEEE*, and Tianye Huang

***Abstract*—Gaussian processes regression (GPR)-aided nonlinear channel equalizer (CE) is experimentally demonstrated in a multi-level intensity modulation and direct detection fiber link. In this scheme, the GPR model is used to estimate the transmitted symbols or the corresponding nonlinear distortions after pre-processing. The experimental results show that GPR-aided nonlinear CE has better nonlinear tolerance than conventional linear and nonlinear filter-based CE. It is also shown that the GPR model in the nonlinear channel equalization process can be understood as an optimized single-layer neural network model with infinite width. Finally, we reveal the relationship between the key coefficients in GPR model and parameters in fiber link through both experiment and simulation.**

***Index Terms*—Direct detection, Gaussian processes regression, intensity modulation, nonlinear channel equalizer.**

I. Introduction

NONLINEAR channel equalization is a major issue in fiber transmission systems because the nonlinear effects fundamentally limit the achievable information rates and transmission distance [1]. Traditionally, the most popular nonlinear channel equalizers (CEs) for intensity modulation and direct detection (IMDD) fiber link are maximum-likelihood sequence equalizer (MLSE) [2] and Volterra series transfer function (VSTF) based nonlinear filter [3]. However, the performances improvement of those nonlinear CEs is limited due to inaccurate nonlinear modelling.

Recently, machine learning (ML)-aided nonlinear CEs have shown great potential in improving the nonlinear tolerance in IMDD fiber links, inculding neural networks (NNs) [4], [5], radial basis function networks (RBFNs) [6], support vector machines (SVMs) [7], and long short-term memory recurrent neural networks (LSTM-RNN) [8]. One common issue associated with these nonlinear CEs is that the physical meaning of the parameters in these ML models are not clear. Therefore, it is unknown whether the parameters in the model have been adjusted to be optimal or how the parameters may affect the system performance since the ML technique is used as a "black box".

In this paper, we focus on nonlinear channel equalization with interpretable Gaussian processes regression (GPR). The GPR model for nonlinear distortion mitigation is presented by assuming the linear channel impairments are compensated. In the experimental demonstration, we build an IMDD link by transmitting 28-GBaud 4-level pulse amplitude modulation (PAM-4) signal over 100-km standard single mode fiber (SSMF). The experimental results show that the GPR-aided nonlinear CE has better nonlinear tolerance than conventional linear and nonlinear filter-based CE. We also compare the performances of GPR model and NN model in the nonlinear channel equalization scheme. It is shown that the output of the GPR model can be viewed as the mean of the output of the NN model with optimized parameters and infinite width. Finally, we give an explanation on how the parameters in GPR model are related with the parameters in the fiber link through both experiment and simulation.

II. Gaussian Processes Regression

The nonlinear CE based on GPR model in a multi-path communication system can be expressed as [9, Chapter 2]:

$$y = f(\boldsymbol{x}) + \nu \tag{1}$$

where $y \in \mathbb{R}$ is the CE output scaler and $\boldsymbol{x} \in \mathbb{R}^{\mathrm{n}}$ is the CE input vector, which also corresponds to the received samples. The noise term ν is assumed to be zero mean with variance σ_{ν}^2. It is noted that (1) doesn't assume $p(y)$ is Gaussian distributed. However, it believes that $p(y|x)$ is Gaussian distributed, which means ν is zero-mean Gaussian [10]. Similar to other CEs, GPR can recover the symbol as $\mathrm{f}_* = \mathrm{f}(\mathbf{x}_*)$ with input $\mathbf{x}_*$ and training set $\mathcal{D}_{\mathrm{m}}$, where $\mathcal{D}_{\mathrm{m}} \in \{\mathbf{X}_{\mathrm{m}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathrm{m}}],\ \mathbf{y}_{\mathrm{m}} = [\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_{\mathrm{m}}]^{\mathrm{T}}\}$. $\mathbf{x}_{\mathrm{i}}$ in $\mathbf{X}_{\mathrm{m}}$ represent the i-th received vector and y_{i} in $\mathbf{y}_{\mathrm{m}}$ represents the corresponding transmitted symbol in the training stage. Ideally, the function $f(\cdot)$ in the test stage may satisfy the minimum mean square error criterion as $f_* = \arg\min_{f(\cdot)} \mathrm{E}[(\mathrm{y}_* - f(\mathbf{x}_*))^2]$, where y_*represents the transmitted symbol. In GPR model, the estimation f_* should follow Gaussian distribution as

Manuscript received 17 August 2022; revised 21 September 2022; accepted 1 October 2022. Date of publication 4 October 2022; date of current version 13 October 2022. This work was supported by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Tianye Huang)*

Xiang Li, Desheng Li, and Tianye Huang are with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan 430074, China (e-mail: 1240912861@qq.com; 920887105@qq.com; tianye_huang@163.com).

Yixin Zhang is with the College of Engineering and Applied Science, Nanjing University, Nanjing 210046, China, and also with the Key Laboratory of Intelligent Optical Sensing and Manipulation, Ministry of Education, Nanjing University, Nanjing 210093, China (e-mail: zyixin@nju.edu.cn).

Perry Ping Shum is with the EEE Department, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: shenp@sustech.edu.cn).

Digital Object Identifier 10.1109/JPHOT.2022.3211906

[9, Chapter 2]:

$$p(f_*|\boldsymbol{x}_*, \mathcal{D}_m) \sim \mathcal{N}\left(\bar{f}_*, cov\left(f_*\right)\right) \tag{2}$$

where $\bar{f}_* = \mathbf{k}^{\mathrm{T}}\, \mathbf{C}_m^{-1}\mathbf{y}_m$ and $cov\ (f_*) = \ k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^{\mathrm{T}}\mathbf{C}_m^{-1}\mathbf{k}$ with

$$\mathbf{k} \ = \left[k\left(\mathbf{x}_1, \mathbf{x}_*\right), k\left(\mathbf{x}_2, \mathbf{x}_*\right), \ldots, k\left(\mathbf{x}_m, \mathbf{x}_*\right)\right]^{\mathrm{T}} \tag{3}$$

$$\mathbf{C}_m = \mathbf{K}_m \ + \sigma_\nu^2 \mathbf{I}_m \tag{4}$$

In (3) and (4), $k(\mathbf{x}_i, \mathbf{x}_j)$ represents covariance function and $\mathbf{C}_m$ denotes the covariance matrix, where $(\mathbf{K}_m)_{ij} = \ k(\mathbf{x}_i, \mathbf{x}_j)$, and $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_m$. From (2), we can see that the estimation f_* is actually a variable following the Gaussian distribution with mean $\bar{f}_*$ and covariance $cov(f_*)$. For simplicity the mean $\bar{f}_*$ is regarded as the estimation result f_*. It is noted that both $\mathbf{k}^T$ and $\mathbf{C}_m^{-1}$ are affected by the choice of covariance function. Therefore, the design of covariance function is significant to an accurate estimation.

In wireless communication system, the following covariance matrix has been applied [10]:

$$(\mathbf{C}_m)_{ij} = \alpha_1\ exp\left(-\sum_{l\,=\,1}^{n} \frac{(x_{il} - x_{jl})^2}{\gamma_l}\right) + \alpha_2 \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + \alpha_3\delta_{ij} \tag{5}$$

where $\boldsymbol{\theta} \ = [\alpha_1, \alpha_2,\ \alpha_3, \gamma_1,\ \gamma_2,\ \ldots,\ \gamma_n]^{\mathrm{T}}$ is called the hyper-parameters, which will be determined in the training stage. x_{il} and x_{jl} represent the l-th scaler in the received symbol vector $\mathbf{x}_i$ and $\mathbf{x}_j$. In (5), the first term is the squared exponential (SE) covariance function, which is infinitely differentiable. Since SE covariance function is a function of $\mathbf{x}_i - \mathbf{x}_j$, it holds the crucial assumption in supervised learning that close points between input vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ may have similar target y. The parameter α_1 is related to the variance of target y. In SE covariance function, the value of length-scale γ_l determines the dependence degree of the elements between $\mathbf{x}_i$ and $\mathbf{x}_j$. The second term is designed for linear regression problem, which is also called the dot product covariance function. The third term α_3 is related to the variance of noise, where δ_{ij} represents the Kronecker's delta function.

In practical scenario, the values of hyperparameters $\boldsymbol{\theta}$ are usually unknown. In order to solve this issue, the *log* marginal likelihood function is first introduced as [9, Chapter 5]:

$$log\ p\left(\mathbf{y}_m|\mathbf{X}_m, \boldsymbol{\theta}\right) = -\frac{1}{2}\mathbf{y}_m^T\mathbf{C}_m^{-1}\mathbf{y}_m - \frac{1}{2}log\left|\mathbf{C}_m\right| - Const. \tag{6}$$

where $Const. = m/2 \cdot log(2\pi)$. According to (6), the optimal setting of hyperparameters $\boldsymbol{\theta}$ corresponds to the maximization of *log* marginal likelihood function. A computational efficient method can be used by calculating the partial derivatives of *log* marginal likelihood function with regarded to hyperparameters $\boldsymbol{\theta}$ as:

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} log\ p\left(\mathbf{y}_m|\mathbf{X}_m, \boldsymbol{\theta}\right) = \frac{1}{2}\ \mathrm{tr}\left(\left(\boldsymbol{\beta}\boldsymbol{\beta}^T - \mathbf{C}_m^{-1}\right)\frac{\partial \mathbf{C}_m}{\partial \boldsymbol{\theta}_i}\right) \tag{7}$$

where $\boldsymbol{\beta} \ = \mathbf{C}_m^{-1}\ \mathbf{y}_m$. Therefore, the hyperparameters $\boldsymbol{\theta}$ can be optimized iteratively by applying the gradient decent method. The notation $\mathrm{tr}(\mathbf{A})$ represents the trace operation to the matrix

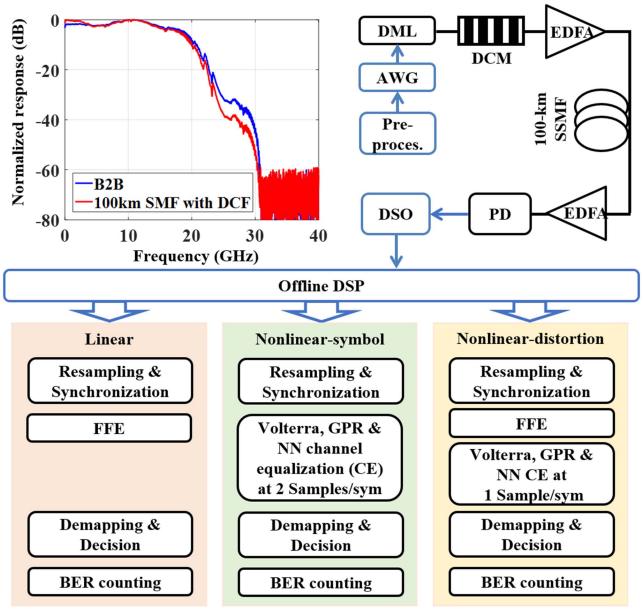

Fig. 1. Experimental setup of the IMDD link and DSP scheme of the proposed GPR-aided nonlinear CE.

A. In this paper, we use the tools from scikit-learn [11] to find the optimal values of the hyperparameters in GPR model.

In order to further reduce the computational complexity, we assume the elements in the input vector $\mathbf{x}_l$ are equally important by setting the same value for all the length-scales, which can also be used to describe the effect of variance of the target y. Therefore, the parameter α_1 can be omitted. After simplification, covariance function in our GPR-aided nonlinear CE is then simplified as:

$$(\mathbf{C}_m)_{ij} = \ exp\left(-\sum_{l\,=\,1}^{n}\left(x_{il} - x_{jl}\right)^2/\gamma\right) + \alpha_2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + \alpha_3\delta_{ij} \tag{8}$$

III. Experimental Setup and Results

To verify the effectiveness of GPR-aided nonlinear CE in the nonlinear distortion estimation mode and symbol estimation mode, we conduct an IMDD experiment to transmit 28-GBaud PAM-4 signal over 100 km SSMF. The experimental setup is shown in Fig. 1. The signal frame consists of 256 OOK symbols for time synchronization followed by payload PAM-4 symbols. The digital signal is generated offline with digitally oversampling factor of 2 and roll-off factor of 0.01. The digital signal is then loaded into an arbitrary waveform generator running at 56 GSa/s to achieve digital to analog conversion. The peak-to-peak voltage of the analog signal is maximized to be 1 V to avoid nonlinear impairment. The analog electrical signal is finally converted to optical signal by a DML operating at 1550.8 nm with bandwidth of 18 GHz and extinction ratio of 5 dB. The bias of the EML is optimized to operate in the linear region. In order to overcome the power fading effect due to fiber chromatic dispersion (CD), a dispersion compensation module (DCM) is used to pre-compensate the CD in the 100-km

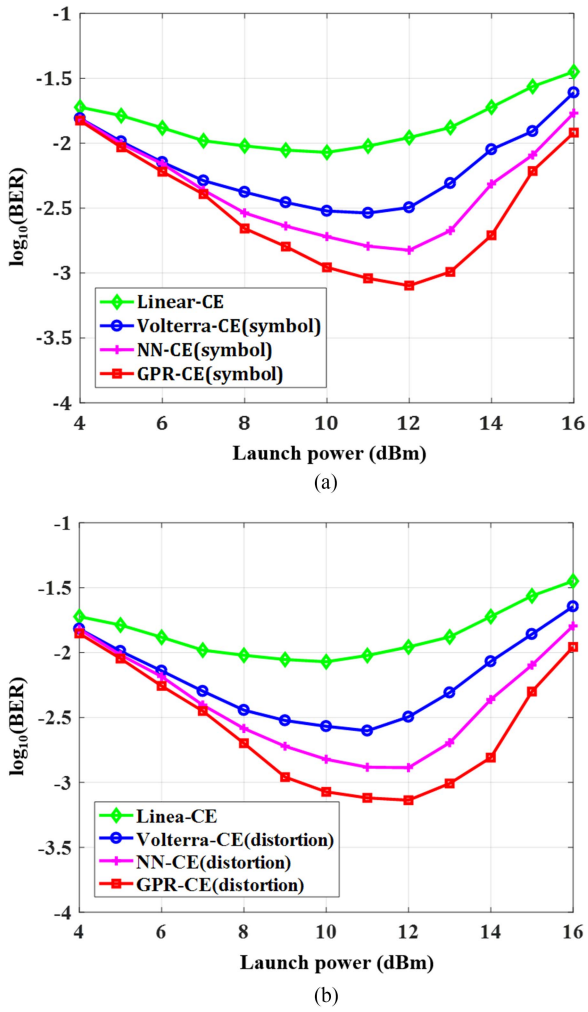

Fig. 2. BER performances of linear-CE, Volterra-CE, NN-CE and GPR-CE versus launch power with (a) symbol and (b) distortion estimation modes.

fiber link. The frequency responses at back-to-back (B2B) and 100-km transmission cases are shown in the inset of Fig. 2. It can be seen that the power fading effects are effectively mitigated by DCM and the residual dispersion can be ignored. Then an erbium doped fiber amplifier (EDFA) is applied to compensate the power loss due to DCM and adjust the input power to a piece of SSMF with length of 100 km. After fiber transmission, the optical signal is amplified by another EDFA before converted to electrical signal by a photodetector (PD) with bandwidth of 40 GHz. The second EDFA is required because the PD is not integrated with trans-impedance amplifier, which can only work properly with input power smaller than 0 dBm. Finally, analog-to-digital conversion is achieved by a digital sampling oscilloscope (DSO) at sampling rate of 80 GSa/s with bandwidth of 33 GHz.

As shown in Fig. 1, we characterize the offline DSP into three categories. The first one is the linear CE including re-sampling to oversampling factor of 2, time synchronization based on self-correlation, feedforward equalization (FFE), de-mapping, decision and bit error rate (BER) calculation. The second one directly applies the nonlinear equalizer to the received samples at 2 samples per symbol to recover the symbols, which is called symbol estimation mode. Here, we consider three nonlinear equalization schemes based on VSTF, NNs, and GPR, which are referred as Volterra-CE, NN-CE and GPR-CE, respectively. For GPR-CE in the symbol estimation mode, the i-th transmitted symbol y_i is designed as the target value in the GPR model at the training stage, and $f^{sym}(\mathbf{x}_*)$ is then regarded as the estimated symbol $\hat{y}_*$ at the transmitter side. The third one combines the first two channel equalization schemes. In this scheme, the nonlinear distortions are estimated after FFE, which is called the nonlinear distortion estimation mode. In this mode, we assume the target value in the nonlinear model as ξ_i, which represents the i-th symbol noise at the training stage. It is noted that the value of noise contains both the nonlinear distortion and linear noise. The value of ξ_i can be realized by subtracting the true symbol y_i from the corresponding symbol $\tilde{y}_i$ after FFE. The estimated symbol $\hat{y}_*$ is then calculated as:

$$\hat{y}_* = \tilde{y}_* - \xi_* \tag{9}$$

where $\tilde{y}_*$ is the recovered symbol after FFE and ξ_* is the estimated nonlinear distortion as $f^{dist}(\mathbf{x}_*)$. One advantage of nonlinear distortion estimation mode is the value of ξ_* is proportional to the optical launch power [12]. Therefore, (9) can be modified as:

$$\hat{y}_* = \tilde{y}_* - 10^{0.1(P_{ch}-P_{ref})}\xi_* \tag{10}$$

where P_{ch} denotes the channel power of the test data set and P_{ref} represents the channel power of the training set. Therefore, it is not necessary to re-train the channel model when the value of input power is changed.

In the experimental demonstration, the bandwidth limitation issue is first pre-compensated at the transmitter side based on back-to-back performance as described in Fig. 1. For the post channel equalization, the tap number of linear FFE filter is 21. The memory lengths of the Volterra-CE are 21, 13 and 9 for the first, second and third orders. The corresponding number of taps is 277. The number of training symbols for linear FFE and Volterra-CE is 4096, which is sufficient to convergence based on recursive least square (RLS) algorithm. The NN model is constructed from an input layer with a vector of dimension (set to 11×1 initially), hidden layer with 200 nodes, and one output node corresponding to the estimated nonlinear distortion. The rectified linear unit function is chosen as the activation function. The weights are updated by adaptive moment estimation algorithm with a learning rate of 0.001. To avoid overfitting during training stage, early stopping and dropout regularization with a dropout rate of 0.2 are applied. For GPR-CE, the dimension n of the input vector $\mathbf{x}$ in (8) is set to 11 initially, which is the same as NN model.

To avoid the issue of overfitting and ensure the randomness of training/testing sets, each PAM-4 transmitted sequence is generated by applying SIGN$(\cdot)$ function to two independent binary random sequences, each drawn from an AWGN sequences. In the experimental demonstration, one PAM-4 sequence with length of 4096 is used as training set. Another three different PAM-4 sequences with length of 65536 are then generated as

testing sets. The BER is calculated by averaging the BER of all the three PAM-4 sequences after testing process.

The performances of linear and nonlinear CEs with two operational modes are shown in Fig. 2(a) and (b), respectively. The channel model is trained individually for each launch power. As shown in Fig. 2, GPR-CE can provide better performance than linear CE, Volterra-CE and NN-CE. The optimal launch power of GPR-CE is 12 dBm, which is 1 dB and 2 dB higher than Volterra-CE and linear-CE, respectively. The performances of the two operational modes are similar in almost all the launch power range.

Next, we investigate the effect of reference power for GPR-CE in the noise estimation mode. It is shown in Fig. 3(a) that the performances are degraded at higher launch power when the reference power is 12 dBm. It means that the nonlinear distortion cannot be calculated with high accuracy at optimal launch power in (10). However, we can also observe small degradation at low launch power region when the reference power is 16 dBm. It means the nonlinear model in such high launch power may overestimate the effect of nonlinear distortion in the low launch power region. From the results in Fig. 3(a), the optimal reference power is 14 dBm, which can provide similar result as that by individual training.

We also investigate how the length of training symbols and number of dimension n affect the performance. As shown in Fig. 3(b), the performance is better with larger length of training symbols. This is mainly because the approximation of multi-dimensional Gaussian distribution can be more accurate if more training symbols are applied in the training stage. The effect of dimension n is shown in Fig. 3(c). It can be seen that the optimal number of dimension n is 11. In our view, the dimension n is related to the degree of correlation among symbols, which can be optimized according to the fibre length. From (2)–(4), we can see that the computational complexity of symbol recovery is related to the size of vector $\mathbf{k}$ in (3), which is determined by the number of training symbols and dimension. It is noted that the vector value $\mathbf{C}_m^{-1}\mathbf{y}_m$ is only related to the training symbols, which can be pre-computed in the training process. Therefore, it can be assumed that the computational complexity of GPR-CE is proportional to the product of number of training symbols (128) and dimension (11), which is 1408 in our experimental demonstrations.

We then evaluate the effect of number of nodes in the hidden layer in Fig. 4. Specifically, the BER is averaged by running the NN model 100 times with different stochastic initializations for each number of nodes. As shown in Fig. 3(d), the performances of NN model are continuously approaching those of GPR model with the increase of the number of nodes. Therefore, the output of the GPR-CE can be viewed as the mean of the output of the NN-CE with infinite width and optimized parameters. An explanation of the results in Fig. 3(d) is that the function computed by the NN model is a function drawn from a Gaussian process in the sense of multidimensional Central Limit Theorem if the width of the NN model is infinite [13]. Therefore, it is believed that GPR can perform better than NN in the nonlinear modelling when the NN model is trained with stochastic optimization.

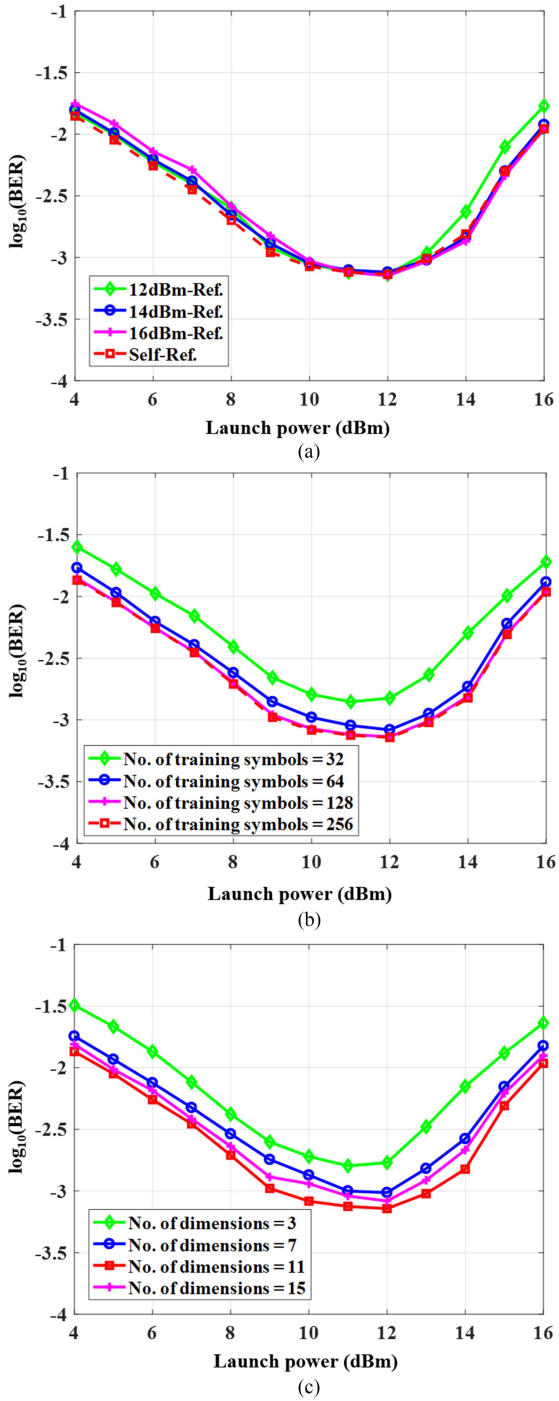

Fig. 3. BER performances of GPR-CE with (a) different values of reference power, (b) different lengths of training symbols, and (c) different lengths of dimensions in distortion estimation modes.

IV. Simulation Analysis

To further explain the physical meaning of parameters in (8), we conduct a simulation by transmitting 28-GBaud PAM-4 signal over IMDD fiber link. The simulation model is shown in Fig. 5. The PAM-4 signal is oversampled by a factor of 32 and added with a bias value to emulate the modulation process after normalization. In this simulation, we mainly focus on the fiber nonlinear effects in the transmission process and the nonlinear

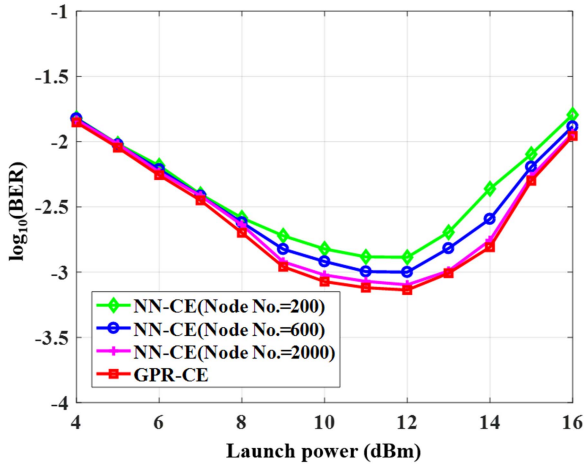

Fig. 4. BER performances of NN model versus number of nodes in the hidden layer.

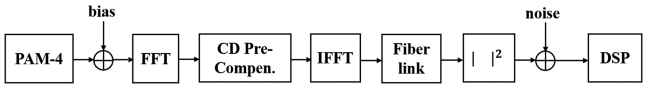

Fig. 5. Simulation model of 28-GBaud PAM-4 signal over IMDD fiber link.

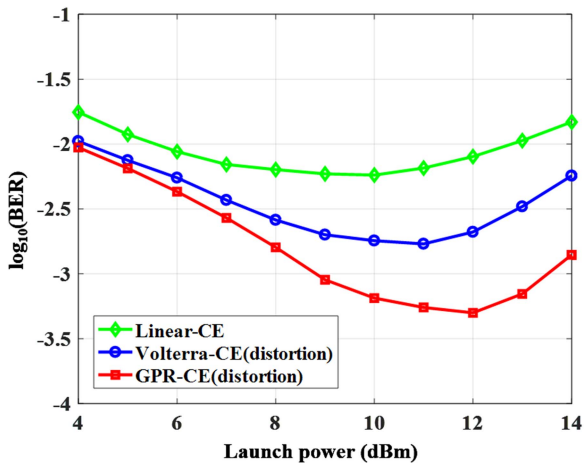

Fig. 6. BER performances of linear-CE, Volterra-CE, and GPR-CE versus launch power with distortion estimation mode.

distortion in the modulation process is not considered. The effectiveness of GPR-CE to specific other nonlinear impairments will be left for future work. The CD pre-compensation is achieved in frequency domain and the fiber link is modeled by split-step Fourier method with distance resolution of 0.01 km, dispersion coefficient of 16.8 ps/nm/km, nonlinear coefficient of 1.3/W/km and fiber loss of 0.2 dB/km [14]. After fiber transmission, the signal is squared to emulate the optical-to-electrical process and then combined with linear noise. Therefore, the nonlinear noise and linear noise are separated, which facilitate the investigation and explanation of parameters in the GPR-CE. The conditions of post channel equalization in the simulation are the same as those in the experimental demonstration.

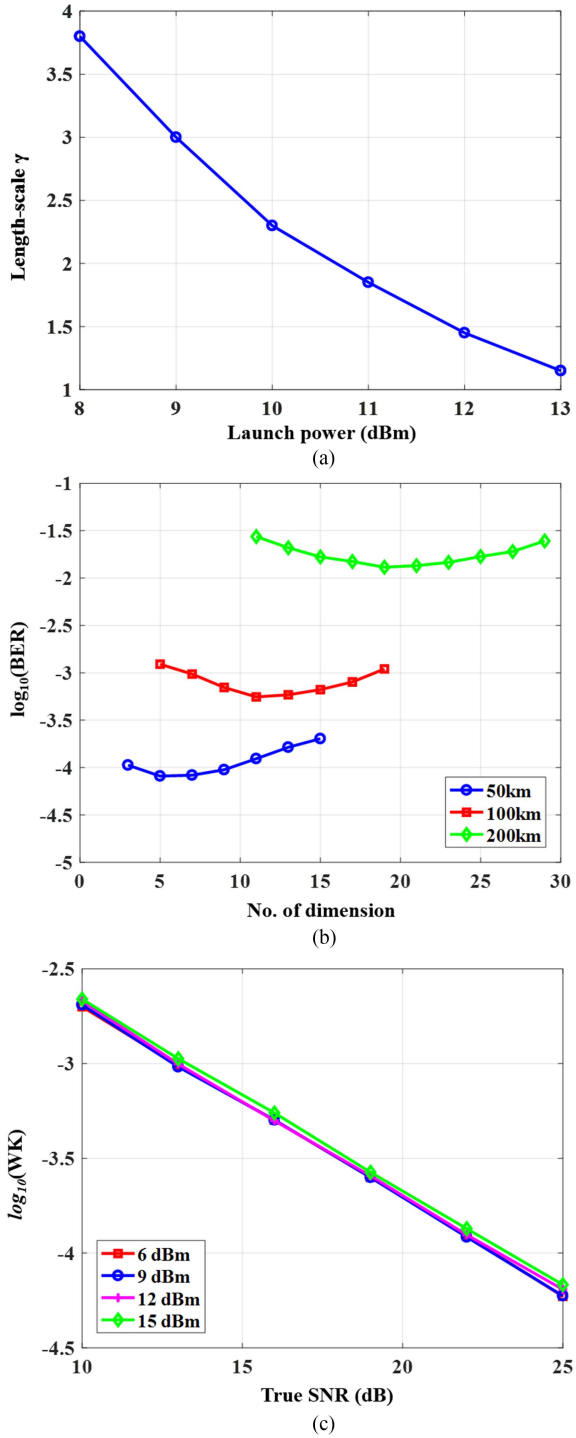

Fig. 7. (a) Trends of estimated length-scale γ with different launch power. (b) BER versus number of dimensions n under different fibre length. (c) Estimated value of white kernel versus true SNR value under different launch power.

Similar to the results in the experimental demonstration, the GPR-CE shows better performance than linear CE and Volterra-CE after 100-km fiber transmission, as shown in Fig. 6. The optimal launch power is also increased by 2 dB and 1 dB over linear CE and Volterra-CE, respectively.

Considering the physical meaning of parameters, it can be seen in Fig. 7(a) that the estimated length-scale γ is reduced with the increase of launch power, which confirms the fact that the input vector $\mathbf{x}_i$ and $\mathbf{x}_j$ are more dependent on each other when the nonlinear distortion is large. As shown in Fig. 7(b), the optimal number of dimensions n is increased when the fiber length is changed from 50 km to 200 km. Therefore, the optimal number of dimensions n is related to the CD in the fiber link, which agrees well with the experimental demonstration. The parameter α_3 is also called white kernel (WK), which is related to the linear noise. By changing the signal-to-noise ratio (SNR) in the simulation, we can see in Fig. 7(c) that the estimated value of $log_{10}(\mathrm{WK})$ is linearly proportional to the true SNR value at different launch power. The deviation at higher launch power is caused by the strong nonlinear distortion in the training stage, which affect the estimation accuracy of hyperparameters.

V. Conclusion

To conclude, we propose a GPR-CE to mitigate the nonlinear distortions in IMDD fiber link. It is shown that the GPR-CE can provide better nonlinear tolerance than conventional linear and nonlinear CEs. The performance comparison between GPR and NN models in channel equalization is also conducted to prove the superiority of GPR model over NN model. Finally, the physical meaning of parameters in GPR model is also investigated through both experiment and simulation to show their connections with CD, linear/nonlinear noise and launch power in the fiber link. In future work, we will conduct further simulations to investigate how much performance improvement can be realized for the specific nonlinear distortions and check which nonlinear mitigations can be more accurately described as Gaussian type.

References

[1] R. Dar and P. J. Winzer, "Nonlinear interference mitigation: Methods and potential gain," *J. Lightw. Technol.*, vol. 35, no. 4, pp. 903–930, Feb. 2017.

[2] Q. Guo and A. V. Tran, "Improving performance of MLSE in RSOA- based WDM-PON by partial response signaling," *Opt. Exp.*, vol. 19, no. 26, pp. B181–B190, Dec. 2011.

[3] S. Zhou, X. Li, L. Yi, Q. Yang, and S. Fu, "Transmission of 2× 56 Gb/s PAM-4 signal over 100 km SSMF using 18 GHz DMLs," *Opt. Lett.*, vol. 41, no. 8, pp. 1805–1808, Apr. 2016.

[4] L. Yi, T. Liao, L. Huang, L. Xue, P. Li, and W. Hu, "Machine learning for 100 Gb/s/λ passive optical network," *J. Lightw. Technol.*, vol. 37, no. 6, pp. 1621–1630, Mar. 2019.

[5] T. Liao, L. Xue, W. Hu, and L. Yi, "Unsupervised learning for neural network-based blind equalization," *IEEE Photon. Technol. Lett.*, vol. 32, no. 10, pp. 569–572, May 2020.

[6] Z. Yang et al., "Radial basis function neural network enabled C-band 4× 50 Gb/s PAM-4 transmission over 80 km SSMF," *Opt. Lett.*, vol. 43, no. 15, pp. 3542–3545, Aug. 2018.

[7] G. Chen et al., "Nonlinear distortion mitigation by machine learning of SVM classification for PAM-4 and PAM-8 modulated optical interconnection," *J. Lightw. Technol.*, vol. 36, no. 3, pp. 650–657, Feb. 2018.

[8] X. Dai, X. Li, M. Luo, Q. You, and S. Yu, "LSTM networks enabled nonlinear equalization in 50-Gb/s PAM-4 transmission links," *Appl. Opt.*, vol. 58, no. 22, pp. 6079–6084, Aug. 2019.

[9] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[10] F. Cruz, J. M.-Fuentes, and S. Caro, "Nonlinear channel equalization with Gaussian processes for regression," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5283–5286, Oct. 2008.

[11] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Feb. 2011.

[12] S. Zhang et al., "Field and lab experimental demonstration of nonlinear impairment compensation using neural networks," *Nature Commun.*, vol. 10, no. 1, Jul. 2019, Art. no. 3033.

[13] A. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani, "Gaussian process behavior in wide deep neural networks," in *Proc. Int. Conf. Learn. Representations*, Vancouver, Canada, 2018, pp. 1–36.

[14] G. P. Agrawal, *Nonlinear Fiber Optics*, 5th ed. London, U.K.: Academic, 2013.