Active Learning with Gaussian Process Regression and Physical Models for Robust SNR Estimation

Ye, Xiaoyan; Mansour, Mariane ; Faruk, Md Saifuddin; Laperle, Charles; Reimer, Michael; O'Sullivan, Maurice; Savory, Seb J.

Published: 01/01/2025

Peer reviewed version

Active Learning with Gaussian Process Regression and Physical Models for Robust SNR Estimation

Xiaoyan Ye[(1,*)], Mariane Mansour[(1)], Md Saifuddin Faruk[(2)], Charles Laperle[(3)], Michael Reimer[(3)], Maurice O'Sullivan[(3)], and Seb J. Savory[(1)]

(1) Electrical Engineering Division, Department of Engineering, University of Cambridge, Cambridge, CB3 0FA, UK.
(2) School of Computer Science and Engineering, Bangor University, LL57 1UT Bangor, U.K.
(3) Ciena Corporation, Ottawa, Ontario K2K 0L1, Canada.
**xy356@cam.ac.uk*

Abstract: We demonstrate improved performance using active learning for both GPR and hybrid models to predict SNR using experimental data from a 15-channel WDM system over 1000km. Physical model interpreted GPR agrees with interpreting measured data.

1. Introduction

To meet the required high demands for optical networks capacity, extensive efforts have been made to further reduce the system margin. To achieve this goal, an accurate quality of transmission (QoT) tool is needed to estimate key metrics such as bit error rate (BER), signal-to-noise ratio (SNR), noise-to-signal ratio (NSR), etc. Analytical and semi-analytical models have been developed [1-2]; however, they suffer from parameter uncertainty. Recently, machine learning (ML) algorithms have gained popularity [3-4] for their ability to solve this problem, but they usually lack interpretability and require an extensive training set. Gaussian process regression (GPR) seems to be a better candidate since it requires fewer data and provides more information than the conventional "black-box" ML solutions: GPR is a probabilistic ML approach, and thus it also computes the confidence interval for each prediction [5].

This work is an extension of [6], where different ways of implementing the physical models and hybrid models have been investigated. The physical-based models are interpretable "white-box" solutions based on the Gaussian noise model [1]. The hybrid model is composed of the physics-based model combined with a GPR to predict the residual error (this residual error arises from phenomena not accounted for, due to the simplified assumptions made when formulating the physical model).

In this paper, we present an active learning-based solution providing accurate and interpretable models using experimental data. We show that a pool-based active learning framework [7] can speed up the convergence and increase the accuracy of both GPR and hybrid models. Two different approaches are presented. For the first one, we apply the physical model to predictions generated by the active learning-enhanced GPR to interpret this trained model by finding the transmission system's physical parameters. Alternatively, the second approach uses active learning to train the GPR of the interpretable hybrid model to enhance its accuracy. The effectiveness of both methods has been verified with 15-channel WDM transmission over 1000 km.

2. Proposed methods for SNR estimation

We consider a 15-channel WDM system where the SNR of each channel is estimated from the 15-input powers launch into each span. Thus, one datapoint is composed of 15 input channel powers and 15 corresponding output SNRs.

First, a 15-input 15-output GPR is trained with the squared exponential kernel function. The GPR was trained following the standard training process (training sets are sampled randomly) or using a pool-based active learning represented in Fig. 1(a). The active learning-based algorithm starts training the GPR with an initial dataset of 10 randomly selected samples. Then, at each step, the trained GPR model picks the most significant samples to be added to the training set for the next step. Here, at each iteration, the trained GPR adds two new datapoints to the training set by picking the most certain point and the least certain point from the pool of 20 datapoints based on the 20 predicted variances. A 15-input 15-output artificial neural network (ANN) with two hidden layers with 64 neurons was also trained as a reference. The Adam optimizer, mean squared error loss function, and the exponential linear unit activation function: $\mathrm{ELU}(x) = x$ if $x > 0; \exp(x) - 1$ if $x \geq 0$ were used with a batch size of 32 datapoints. Models' hyperparameters were optimized using the validation set.

Second, a hybrid approach with active learning is presented. The hybrid model consists of one trained physics-based model followed by an active learning-based GPR to predict the NSR residual error defined as $\Delta NSR_{GPR} = NSR_{meas} - NSR_{phy}$. The physical model is based on the SNR expression presented in Fig. 2. Using the measured launch powers and SNRs, the equation can be solved to estimate the unknown system parameters: back-to-back SNR (SNR_{TRX}), ASE noise power (P_{ASE}), and nonlinear interference

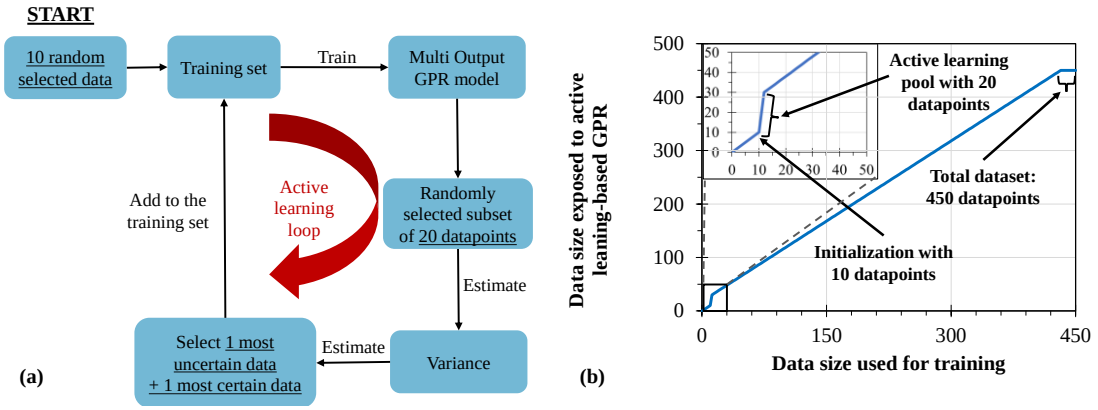

Fig. 1. (a) Schematic representation of the chosen pool-based active learning framework. (b) Data size exposed to the GPR with the active learning framework *vs.* Data size used for training the GPR

coefficients (η). Consequently, the trained physical model can predict any SNR of any given launch power configuration. Here, we assume in the physical model that only channel separation is the determining factor for the nonlinear interference coefficients η, due to the superior performance of this assumption [6]. We have trained the physical model with 100 datapoints before combining it with a GPR to build the hybrid model. Next, the GPR component of the hybrid model is trained using either active learning (same process described in Fig. 1(a).) or a standard training process with random selection.

To ensure a fair comparison, we benchmarked the models with respect to the "data size exposed to the model", rather than the actual training data size. Indeed, in a pool-based active learning method, more data is required than just the training data, as the algorithm selects from a larger pool at each step. Fig. 1(b) illustrates the relationship between the training data size and the exposed data size.

3. Experimental setup

To generate the experimental data, we use the setup shown in Fig. 2. The transmitted signal is synthesized using the 16 integrated tunable laser assemblies (iTLAs) bulk modulated by a modified Ciena WaveLogic 3 line card. The modulated output is then passed through a wavelength selective switch (WSS). In this paper, the transmitter WSS (Tx-WSS) was used to select 15 channels centered around 1550 nm with central frequency at 193.4 THz and 50 GHz channel spacing, and each channel is modulated with 34.5 GBd 16-QAM signals. The WDM channels are amplified by a booster amplifier before being transmitted through the link which consists of 10 spans of 100 km standard fibers. The span loss is compensated at the end of each span by a fixed gain inline-EDFA. All of the connections are made using a Polatis 32 × 32 fiber switch, which also allows the control of the total power launch into each span. At the receiver end, a Rx-WSS allows channel selection for demodulation with a Ciena WaveLogic 3 line card receiver. We measure the BER at the receiver and we can calculate the SNR using the relation $BER = 3/8 \times \mathrm{erfc}\left(\sqrt{SNR/10}\right)$ with erfc(.) the complementary error function.

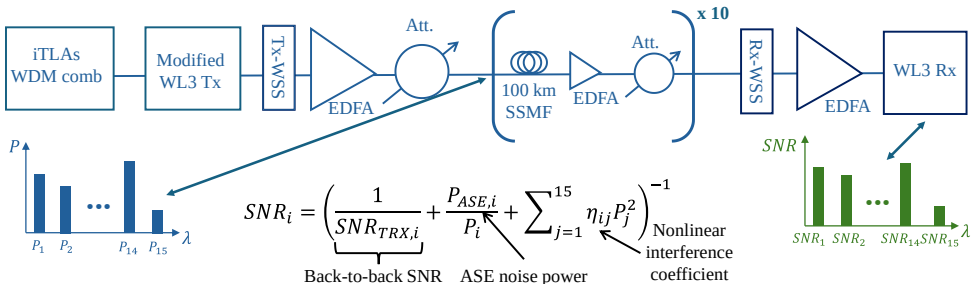

Fig. 2. Experimental setup:15 WDM channels are transmitted in a 10-spans standard single mode fiber (SSMF) link.

We captured a total of 700 datapoints, each having 15 channel launch powers and the corresponding measured SNRs. The channel power was randomly chosen between [-3, 3] dBm and this dataset cover the linear, quasi-linear, and nonlinear regimes. The primary source of uncertainty in this experiment is the launch power measurements, with an uncertainty of ±0.05 dB. The measured dataset is split into training, validation, and test sets, with 450, 50, and 200 datapoints allocated to each, respectively.

4. Results and discussion

In Fig. 3(a), the evolution of the overall RMSE in dB for the test set is plotted for the physics-based model, the ANN and the GPR models with a standard training process or with our proposed active

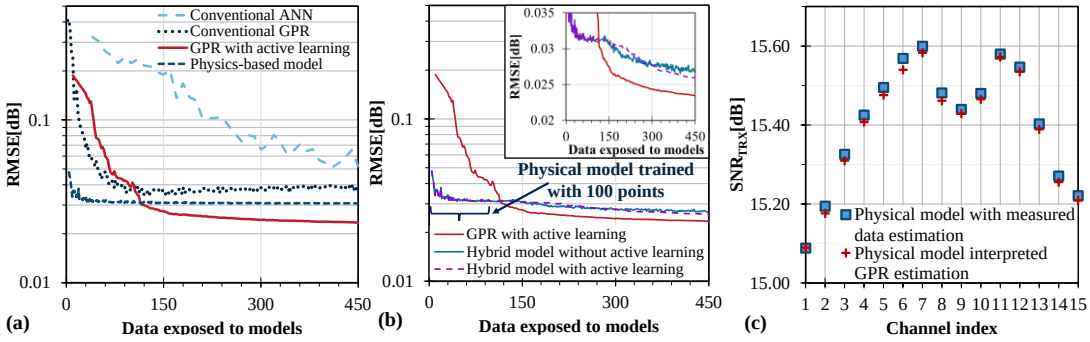

Fig. 3. (a) Comparison between physics-based model and GP model (with or without active learning), (b) RMSE *vs*. exposed data size for the active learning-trained GP model and the hybrid models (trained with or without active learning), (c) Back-to-back SNR estimation with physical model applied to measured data or to GPR generated new data.

learning training framework. As shown, the active learning-based GPR outperforms standard random sampling training process in terms of RMSE, with smoother and faster convergence, achieving an RMSE of approximately 0.023 dB compared to 0.038 dB for conventional GPR. The physical model converges the fastest, but the RMSE which plateaus at a value of 0.031 dB, which is higher than that of our proposed method. The ANN has the largest RMSE due to the complexity of the problem (15 channel powers as input and 15 SNR values as output) and the relatively small size of the training set (450 datapoints).

The performance of the hybrid models is shown in Fig. 3(b). The inset of the figure indicates that the hybrid model can also benefit from our proposed active learning framework as it converges faster (after 300 datapoints) and reaches a slightly lower level, resulting in a more accurate model with an RMSE of 0.026 dB. Nevertheless, the active learning-based GPR still outperforms.

We finally interpreted the trained GPR, by applying the physical model to the GPR-predicted dataset. This approach allows us to evaluate the reliability of the GPR model by estimating the physical parameters of the transmission systems (SNR_{TRX}, P_{ASE} and η). Fig. 3(c) illustrates the resulting estimation of SNR_{TRX}, as an example. The estimation based on a dataset of 1000 GPR-generated datapoints, compared to the estimation based on the 450 measured datapoints, demonstrates an RMSE of 0.016 dB and a maximum error of 0.029 dB. These results indicate that the GPR model is performing reliably.

4. Conclusion

We proposed and experimentally validated an active learning-based method to train a GPR and a hybrid model for SNR estimation for a 15-channels WDM systems over 1000 km transmission. The GPR model trained with pool-based active learning process outperforms the other solutions, achieving an RMSE at 0.023 dB and demonstrating faster convergence. Thus, it requires less measured data to provide accurate estimates.

Both hybrid and pure GPR models are robust and “interpretable”. Indeed, the hybrid model is based on the explainable physical model followed by a GPR to predict the residual error. As for the GPR model, we have shown that the physics-based model can be used to interpret it, by extracting physical parameters of the transmission system with the GPR generated data. The proposed approaches offer more reliable and explainable data-driven solutions compared to other popular ML-based QoT estimation tools.

Acknowledgements XY, MM, MSF and SJS thank Ciena for the donation of equipment and support. This work was supported by the EPSRC Programme Grant TRANSNET [EP/R035342/1] and the EPSRC Grant TITAN [EP/Y037243/1]. MM thanks Ciena for funding her Ph.D. studentship. Data underlying the results presented in this paper are available at https://doi.org/10.17863/CAM.113009

References

[1] P. Poggiolini, G. Bosco, A. Carena, V. Curri, Y. Jiang, and F. Forghieri, "The GN-model of fiber non-linear propagation and its applications," *J. Lightwave Technol.* 32, 694-721(2014).
[2] A. Mecozzi and R. -J. Essiambre, "Nonlinear Shannon limit in pseudolinear coherent systems," in *Journal of Lightwave Technology*, vol. 30, no. 12, pp. 2011-2024, June15, 2012.
[3] I. Pointurier. “Machine learning techniques for quality of transmission estimation in optical networks,” in *Journal of Optical Communications and Networking*, 13(4) :B60–B71, 2021.
[4] R. Ayassi, A. Triki, N. Crespi, R. Minerva and M. Laye, "Survey on the use of machine learning for quality of transmission estimation in optical transport networks," in *Journal of Lightwave Technology*, vol. 40, no. 17, pp. 5803-5815, 1 Sept.1, 2022.
[5] C. E. Rasmussen, “Gaussian processes for machine learning”. *MIT Press*, 2006.
[6] M. Mansour, M. S. Faruk, C. Laperle, M. Reimer, M. O'Sullivan and S. J. Savory, "Physics-based modeling for hybrid data-driven models to estimate SNR in WDM systems," in *Journal of Lightwave Technology*, 2024.
[7] B. Settles, “Active learning, volume 6 of synthesis lectures on artificial intelligence and machine learning,” *Morgan & Claypool*, 2012.