A Stochastic Active Learning Strategy for Gaussian Process Models with Application to the Uncertainty Quantification of Signal Integrity

A Stochastic Active Learning Strategy for Gaussian Process Models with Application to the Uncertainty Quantification of Signal Integrity

Paolo Manfredi
EMC Group, Department of Electronics and Telecommunications
Politecnico di Torino
10129 Torino, Italy
paolo.manfredi@polito.it

***Abstract*—This paper introduces an effective active learning strategy to iteratively refine the training of Gaussian process models with application to uncertainty quantification. Compared to traditional deterministic approaches or naive random sampling, the proposed approach uses a stochastic measure, i.e., the predictive variance of the output mean, to drive the acquisition of additional training points. The advocated algorithm is shown to outperform alternative strategies in the uncertainty quantification of the insertion loss of a microstrip transmission line with a discontinuity in the ground plane.**

***Index Terms*—Active learning, Bayesian estimation, crosstalk, Gaussian processes, Kriging, machine learning, surrogate modeling, training, uncertainty quantification.**

I. Introduction

Machine learning is rapidly burgeoning in the field of design, optimization, and uncertainty quantification (UQ) of modern electronics. Among the plethora of available techniques, kernel-based methods benefit from a relatively simple and flexible structure, good scalability to higher dimensions, and reasonable training cost. In particular, Gaussian process (GP) models were successfully applied to the design exploration and optimization of electronic devices [1]–[3]. Besides their relative simplicity, an attractive feature of GPs is that they leverage a Bayesian setting that allows associating confidence levels to model predictions. This feature was extended from standard deterministic predictions to UQ measures in [4].

One of the open issues with machine learning methods is the appropriate selection of training data, which ideally should be as limited as possible. In the context of Bayesian optimization [1], [3] and surrogate modeling [5], the associated prediction uncertainty is used to drive the iterative query of additional training data. However, classical strategies usually assume deterministic input parameters and may be therefore non-optimal in stochastic settings.

In this paper, a more effective active learning strategy for GPs applied to UQ tasks is proposed and investigated. The algorithm leverages the prediction uncertainty of a stochastic measure, i.e., the predictive variance of the output mean.

II. GP Regression

This section introduces the basis notions about GP modeling as needed for the development of the proposed active learning method.

Let us consider a generic system that we shall denote with

$$y = \mathcal{M}(\boldsymbol{x}), \tag{1}$$

where $\mathcal{M} : \mathcal{X} \to \mathbb{R}$, with $\mathcal{X} \subseteq \mathrm{R}^d$, is a (usually, implicit) map between a set of input parameters $\boldsymbol{x} = (x_1, \ldots, x_d)$ and the corresponding output y. For the sake of simplicity, in this paper paper we focus on a single (scalar) output quantity y.

A. Classical Deterministic Setting

GP regression starts by the assumption that the map (1) is one particular realization of a *prior* GP with a given mean or *trend* function $\mu(\boldsymbol{x})$ and kernel function $k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 r(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta})$, where σ^2 is the kernel variance and $r(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta})$ is the correlation function, which depends on a set of *hyperparameters* $\boldsymbol{\theta}$ [6].

Without loss of generality, we assume a constant mean function, i.e., $\mu(\boldsymbol{x}) = \beta_0$. Many correlation functions are available for the kernel. The most popular are the squared-exponential and the Matérn 5/2 functions, the latter reading:

$$r(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta}) = \left(1 + \sqrt{5}u + \frac{5}{3}u^2\right) \exp\left(-\sqrt{5}u\right), \tag{2}$$

with $u = \sqrt{\sum_{j=1}^{d} (x_j - x'_j)^2/\theta_j^2}$. Again without loss of generality, we shall use the above kernel in this paper. The kernel is characterized by hyperparameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, which are lengthscales dictating the GP smoothness along each input dimension. These hyperparameters are optimized based on the available training data.

B. Training

The GP model is trained using observations collected from the actual system (1), which we shall denote with $\{(\boldsymbol{x}_l^{\mathrm{tr}}, y_l^{\mathrm{tr}})\}_{l=1}^{L}$, where $y_l^{\mathrm{tr}} = \mathcal{M}(\boldsymbol{x}_l^{\mathrm{tr}})$. The training involves the estimation of the trend coefficient β_0 as well as of the kernel variance σ^2 and hyperparameters $\boldsymbol{\theta}$. This is a standard routine that is implemented in dedicated toolboxes (e.g., [6], [7]), therefore we omit the details in this paper.

Oftentimes, the training is performed once with a fixed dataset. However, active learning strategies exist for the iterative acquisition of additional training samples [2], [5], thereby refining the model and improving its accuracy.

979-8-3503-8376-8/23/$31.00

C. Prediction

Once the prior parameters are known, the model predictions are readily computed as [6]

$$m(\boldsymbol{x}|\boldsymbol{x}^{\mathrm{tr}}) = \beta_0 + \boldsymbol{r}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{R}^{-1}(\boldsymbol{y}^{\mathrm{tr}} - \beta_0), \tag{3}$$

where $\boldsymbol{y}^{\mathrm{tr}} = (y_1^{\mathrm{tr}}, \ldots, y_L^{\mathrm{tr}})^{\mathsf{T}}$ is the vector of training observations, $\boldsymbol{R}$, with entries $R_{lm} = r(\boldsymbol{x}_l^{\mathrm{tr}}, \boldsymbol{x}_m^{\mathrm{tr}}|\boldsymbol{\theta})$, is the correlation matrix of the training samples, and $\boldsymbol{r}(\boldsymbol{x})$, with entries $r_l = r(\boldsymbol{x}_l^{\mathrm{tr}}, \boldsymbol{x}|\boldsymbol{\theta})$, is the vector of correlations between the training samples and the prediction point. The notation $m(\boldsymbol{x}|\boldsymbol{x}^{\mathrm{tr}})$ is used to emphasize that the prediction is related to the specific selection of training points.

An important feature of GP models is that the covariance

$$c(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{x}^{\mathrm{tr}}) = \sigma^2\left(r(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta}) - \boldsymbol{r}^{\mathsf{T}}(\boldsymbol{x})\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x}')\right) \tag{4}$$

can be associated to model predictions obtained via (3). Specifically, the prediction at a given point is a Gaussian random variable with mean (3) and variance

$$\begin{aligned} v(\boldsymbol{x}|\boldsymbol{x}^{\mathrm{tr}}) &= c(\boldsymbol{x}, \boldsymbol{x}|\boldsymbol{x}^{\mathrm{tr}}) \\ &= \sigma^2\left(r(\boldsymbol{x}, \boldsymbol{x}|\boldsymbol{\theta}) - \boldsymbol{r}^{\mathsf{T}}(\boldsymbol{x})\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x})\right), \end{aligned} \tag{5}$$

for which confidence levels are readily obtained.

D. Application to UQ

In UQ settings, the GP model is used to surrogate the actual system (1) and speed up a Monte Carlo analysis. Hence, a (large) number of samples $\{\boldsymbol{x}_i^{\mathrm{mc}}\}_{i=1}^N$ is drawn and the corresponding predictions are obtained using (3) in place of (1). It is important to mention that expected values, variance, and confidence levels are obtained for pertinent statistical information, such as moments and distributions of the output y, either in closed form or numerically [4].

For the subsequent developments, it is useful to report the predictive variance of the Monte Carlo mean, which reads

$$v_\mu(\boldsymbol{x}^{\mathrm{tr}}) = \frac{1}{N^2}\sum_{i,j=1}^{N} c(\boldsymbol{x}_i^{\mathrm{mc}}, \boldsymbol{x}_j^{\mathrm{mc}}|\boldsymbol{x}^{\mathrm{tr}}). \tag{6}$$

For a fixed choice of Monte Carlo samples $\boldsymbol{x}^{\mathrm{mc}}$, v_μ is a scalar defining with how much uncertainty the mean of the Monte Carlo samples is predicted via the GP model trained with $\boldsymbol{x}^{\mathrm{tr}}$.

III. Active Learning Strategies

This section introduces three alternative sampling strategies for the iterative query of additional training samples.

1) Plain Random Sampling: In this naive scenario, an additional training point $\boldsymbol{x}_\star^{\mathrm{tr}}$ is generated randomly, e.g., according to the distribution of the input parameters.

2) Standard Active Learning (Deterministic): In the classical active learning strategy [5], the predictive variance (5) is used as a measure to identify locations of the input space where the prediction is less accurate. Hence, the new training sample is selected as the point at which the prediction uncertainty is maximum, i.e.,

$$\boldsymbol{x}_\star^{\mathrm{tr}} = \arg\max_{\boldsymbol{x}\in\mathcal{X}} v(\boldsymbol{x}|\boldsymbol{x}^{\mathrm{tr}}). \tag{7}$$

The selection is therefore based exclusively on the currently available training points. This approach is deterministic since it disregards the distribution of the input parameters $\boldsymbol{x}$, which may be non-optimal. For example, if the selected point lies in a corner of the input space $\mathcal{X}$ and the input distribution is Gaussian, it may add little information in an UQ scenario.

3) Proposed Stochastic Active Learning: The proposed approach starts from the observation that the predictive variance of the output mean (6) does not (directly) depend on the observations $\boldsymbol{y}^{\mathrm{tr}}$, but only on the training points $\boldsymbol{x}^{\mathrm{tr}}$ at which they are evaluated. Therefore, it is possible to "virtually" add a new training point and check how much (6) is reduced, without the need to actually calculate the corresponding observation. In more formal terms, the new training point is chosen as

$$\boldsymbol{x}_\star^{\mathrm{tr}} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} v_\mu(\boldsymbol{x}^{\mathrm{tr}} \cup \boldsymbol{x}), \tag{8}$$

where $\boldsymbol{x}^{\mathrm{tr}} \cup \boldsymbol{x}$ is a "loose" notation to indicate that (6) is computed with the training set obtained by the union of the currently available points $\boldsymbol{x}^{\mathrm{tr}}$ and the new candidate point $\boldsymbol{x}$. At this stage, the hyperparameters are not updated and the last available estimate is used.

In practice, the predictive variance does depend on the training observations through the kernel parameters σ^2 and $\boldsymbol{\theta}$, which appear in the correlation function r and are estimated based on the available training data. Hence, the proposed approach relies on the assumption that the hyperparameters are not significantly affected by the additional point. It is important to point out that the distribution of the input parameters is now inherently accounted for in the calculation of the predictive variance of the output mean.

Regardless of the adopted sampling method, once a new training point is selected, the corresponding observation is queried and included in the training dataset. At this point, the model is re-trained and the hyperparameters are also updated.

IV. Application Example and Numerical Results

The three sampling techniques introduced in Section III are applied to the UQ of the insertion loss of a microstrip line with a discontinuity in the ground plane [8]. The stochastic input parameters are the location and the length of the slot in the ground plane, which are considered as two independent Gaussian random variables, both with a nominal value of 15 mm and a relative standard deviation of 10%. In particular, we focus the attention on the magnitude of S_{21} at 7.5 GHz, a frequency at which the insertion loss exhibits a large variability. The scattering parameters are computed by means of CST Studio Suite® [9]. The Reader is referred to [8] for additional details on this test case.

Without loss of generality, the outlined sampling techniques are implemented based on a search over a finite set of 1000 candidate points, randomly generated using Latin hypercube sampling. The corresponding outputs are pre-simulated to serve also as a reference for assessing the accuracy (see below). For the deterministic and stochastic active learning strategies, the pertinent optimization, i.e., (7) or (8), is performed over these candidate points. The best selected point

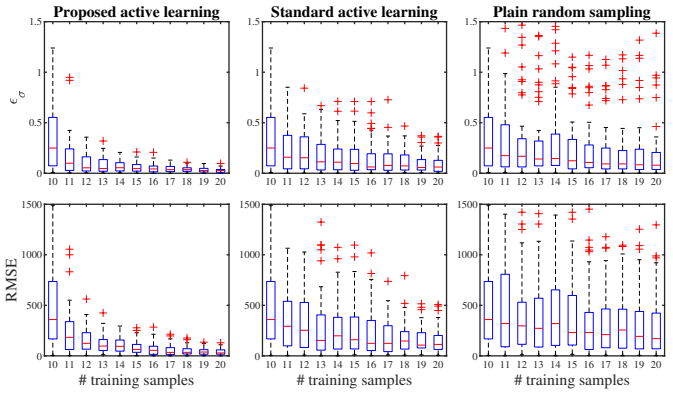

Fig. 1. Box plots showing the behavior of ϵ_σ (top panels) and RMSE (bottom panels) for the various sampling strategies. Left panels: proposed stochastic active learning; central panels: standard deterministic active learning; right panels: plain random sampling.

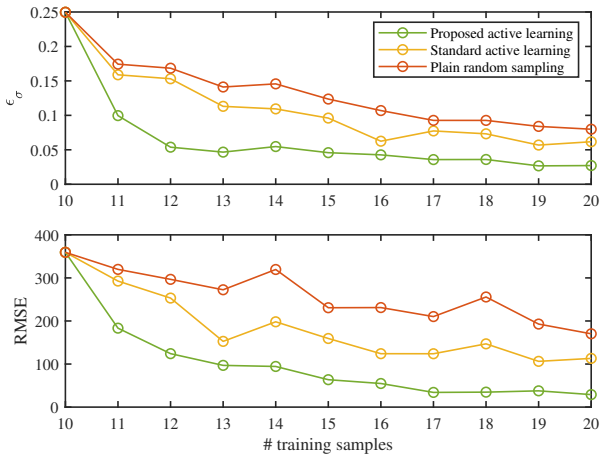

Fig. 2. Behavior of the median over 50 runs of ϵ_σ (top panel) and RMSE (bottom panel) as the iterations increase. Red curves: plain random sampling; yellow curves: standard active learning; green curves: proposed method.

and the corresponding simulation output are added to the training dataset and removed from the candidate set. For the plain random sampling instead, the additional training point is simply selected randomly from the search set.

The GP models are trained using the MATLAB® Statistics and Machine Learning Toolbox™ toolbox [7]. The accuracy of the GP models is assessed based on the following two figures of merit: 1) the relative error on the variance, defined as

$$\epsilon_\sigma = \frac{|\hat{\sigma}_y^2 - \sigma_y^2|}{\sigma_y^2}, \tag{9}$$

where σ_y^2 is the variance of the reference observations $\{y_i\}_{i=1}^N$ and $\hat{\sigma}_y^2$ is the variance of the corresponding GP predictions $\{\hat{y}_i\}_{i=1}^N$, with $N = 1000$; 2) the root-mean-square error (RMSE) between observations and predictions, defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}. \tag{10}$$

The box plots in Fig. 1 show the behavior of the above error metrics for the three sampling strategies and for 50 independent runs. For each run, the same initial set of 10 training samples, randomly picked from the search set, is considered for all strategies. Then, 10 additional training samples are iteratively selected from the search set based on the corresponding algorithm. From the box plots, it is observed that that the proposed active learning method outperforms both the standard deterministic strategy and the plain random sampling in both metrics, as it provides lower error and lower dispersion at each iteration.

For a clearer comparison, Fig. 2 shows the medians of the achieved metrics over the 50 runs. It is observed that the proposed algorithm (green curve) achieves a faster decreasing trend of the median error compared to both the standard deterministic method (yellow curve) and the plain random sampling (red curve).

V. Conclusions

This paper introduced an active learning algorithm for the iterative training of GP models with application to UQ. The proposed strategy starts from a limited set of initial training samples and uses the predictive uncertainty of the output mean as the objective function to minimize in the selection of the additional training points. Compared to the standard deterministic active learning strategy and naive random sampling, the advocated method achieves faster convergence and lower error for the same number of training samples. The proposed method was successfully applied to UQ of the insertion loss of a microstrip line with ground plane discontinuity.

References

[1] H. M. Torun and M. Swaminathan, "High-dimensional global optimization method for high-frequency electronic design," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 6, pp. 2128–2142, Jun. 2019.

[2] M. Swaminathan, H. M. Torun, H. Yu, J. A. Hejase, and W. D. Becker, "Demystifying machine learning for signal and power integrity problems in packaging," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 10, no. 8, pp. 1276–1295, Aug. 2020.

[3] D. De Witte, J. Qing, I. Couckuyt, T. Dhaene, D. Vande Ginste, and D. Spina, "A robust bayesian optimization framework for microwave circuit design under uncertainty," *Electronics*, vol. 11, no. 14, p. 2267, Jul. 2022.

[4] P. Manfredi, "Probabilistic uncertainty quantification of microwave circuits using Gaussian processes," *IEEE Trans. Microw. Theory Techn.*, vol. 71, no. 6, pp. 2360–2372, Jun. 2023.

[5] R. Trinchero and F. Canavero, "Use of an active learning strategy based on Gaussian process regression for the uncertainty quantification of electronic devices," *Eng. Proc.*, vol. 3, no. 1, p. 3, 2020.

[6] C. Lataniotis, D. Wicaksono, S. Marelli, and B. Sudret, "UQLab user manual – Kriging (Gaussian process modeling)," Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland, Tech. Rep., 2022, report UQLab-V2.0-105.

[7] *Statistics and Machine Learning Toolbox, Version 12.5*. Natick, MA, USA: The MathWorks Inc., 2023. [Online]. Available: https://www.mathworks.com

[8] P. Manfredi and S. Grivet-Talocia, "Fast stochastic surrogate modeling via rational polynomial chaos expansions and principal component analysis," *IEEE Access*, vol. 9, pp. 102 732–102 745, 2021.

[9] *CST Studio Suite, Version 2022.00*. Vélizy-Villacoublay, France: Dassault Systemès, 2022. [Online]. Available: https://www.3ds.com/products-services/simulia/products/cst-studio-suite/