Conservative Surrogate Modeling of Crosstalk with Application to Uncertainty Quantification

Conservative Surrogate Modeling of Crosstalk with Application to Uncertainty Quantification

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

***Abstract*—Machine learning methods are attracting a great interest as surrogate modeling tools for signal and power integrity problems. However, an open issue is that it is often difficult to assess the model trustworthiness in generalizing beyond the training data. In this regard, Gaussian process (GP) models notably provide an indication of the prediction confidence due to the limited amount of training samples. They are wildly used as surrogates in design exploration, optimization, and uncertainty quantification tasks. Nevertheless, their prediction confidence does not account for the uncertainty introduced by the estimation of the GP parameters, which is also part of the training process. In this paper, we discuss two improved GP formulations that take into account the additional uncertainty related to the estimation of (some) GP parameters, thereby leading to more reliable and conservative confidence levels. The proposed framework is applied to the uncertainty quantification of the maximum transient crosstalk in a microstrip interconnect.**

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

I. Introduction

Surrogate modeling and machine learning are increasingly used in signal and power integrity applications to perform efficient design exploration, optimization, and uncertainty quantification (UQ) tasks. A wide range of techniques were employed for this purpose, including techniques based on polynomial chaos expansion (PCE), neural networks, and kernel-based machine learning methods [1], [2], [3], [4].

The PCE-based methods are particularly suitable for UQ, since statistical information like moments and sensitivity indices are derived analytically from the model coefficients [5]. The surrogate model consists in this case in an expansion of suitable orthogonal polynomials depending on the distribution of the uncertain design parameters. One of the main limitations of this technique is the fact that the model complexity grows exponentially with the number of uncertain parameters, thereby making it inefficient for high-dimensional problems.

More recently, data-driven methods belonging to the broad class of machine learning methods attracted a growing interests thanks to their model-free structure. This yields enhanced flexibility and adaptability, as well as better scaling to high-dimensional problems. These techniques include a wide class of neural network architectures as well as kernel-based methods like support-vector regression, least-square support-vector machines, and Gaussian process (GP) regression, also known as Kriging [6]. In particular, GP regression is a popular and flexible tool that is used either as a plain surrogate or as a target function approximation in Bayesian optimization [7], [8] and UQ [9].

When it comes to surrogate models, an open issue that so far received rather limited attention is the trustworthiness in generalizing beyond training samples. Indeed, when reference data is not available, the accuracy of the surrogate is usually assessed based on the same data that is used to build it, possibly leading to overfitting and making it difficult to predict how well the model generalizes to unseen data.

In this regard, an attractive feature of GP models is that they inherently carry an estimate of the prediction uncertainty. The method assumes that the target function can be assimilated to a realization of a certain GP. If the GP is fully known, the information on the prediction confidence is rigorous and accurate, and it reflects the model uncertainty due to the fact that a limited amount of data is used to "train" the model. In practice, some of the GP parameters are unavoidably left as degrees of freedom that are tuned as part of the training process, so that a more flexible and general model is put forward and later adapted to the specific problem at hand. The fact that the GP parameters are also estimated based on the available data adds uncertainty to the model prediction, but this aspect is often overlooked in the literature.

In this paper, we discuss a more rigorous GP formulation, which accounts for the additional uncertainty related to the estimation of – at least some of – the GP parameters. This leads to a more conservative and reliable prediction uncertainty, which helps prevent overconfidence in the model. The proposed method is applied to the UQ of the maximum transient crosstalk occurring in a microstrip interconnect and its performance in comparison with the standard GP formulation is illustrated.

II. GP Regression

Let us consider a quantity of interest y that depends on d design parameters $\boldsymbol{x} = (x_1, \ldots, x_d)$ through a – typically implicit – functional dependence

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

where $\mathcal{M} : \mathbb{R}^d \to \mathbb{R}$ is a computational model.

GP regression seeks to model the function (1) as a realization of a GP called "prior", which we denote as

$$y \sim \mathcal{GP}\left(\mu(\boldsymbol{x}), \sigma^2 r(\boldsymbol{x}, \boldsymbol{x}')\right) \tag{2}$$

where $\mu(\boldsymbol{x})$ is the mean function, or *trend*, and $\sigma^2 r(\boldsymbol{x}, \boldsymbol{x}')$ is the covariance function, or *kernel* [6], [10].

Bayesian inference is used to "identify" the specific GP realization that best fits to the target function based on a limited number of observations. Indeed, if L observations $\{y_l\}_{l=1}^L$ are collected for as many configurations $\{\boldsymbol{x}_l\}_{l=1}^L$ of the input parameters, with samples $y_l = \mathcal{M}(\boldsymbol{x}_l)$ computed using the computational model (1), the target function is approximated using the mean function of the resulting "posterior" GP, which reads

$$y \approx \mathcal{M}_{\mathrm{GPR}}(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \boldsymbol{r}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \tag{3}$$

where:

- $\boldsymbol{y} = (y_1, \ldots, y_L)^{\mathsf{T}}$ is the vector of observations;
- $\boldsymbol{\mu} = (\mu(\boldsymbol{x}_1), \ldots, \mu(\boldsymbol{x}_L))^{\mathsf{T}}$ is the vector of the GP trend evaluated at the training samples;
- $\boldsymbol{R}$ is the correlation matrix of the training samples, with entries $R_{lm} = r(\boldsymbol{x}_l, \boldsymbol{x}_m)$, $l, m = 1, \ldots, L$;
- $\boldsymbol{r}(\boldsymbol{x}) = (r(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, r(\boldsymbol{x}, \boldsymbol{x}_L)^{\mathsf{T}}$ is the vector of cross-correlations between the prediction point and the training samples.

Equation (3) provides the GP prediction at an arbitrary point $\boldsymbol{x}^*$. If no noise is assumed on the data, the prediction interpolates the training samples.

Furthermore, a posterior covariance function, computed as

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

is associated to the predictions. The posterior covariance describes the correlation between the predictions at different points. Given a single design point $\boldsymbol{x}^*$, the model prediction is a Gaussian random variable with expectation (i.e., most likely prediction) $\mu(\boldsymbol{x}^*)$, computed from (3), and variance $c(\boldsymbol{x}^*, \boldsymbol{x}^*)$, computed with (4). Therefore, the GP model is notably a stochastic model!

The model variability is an expression of the uncertainty due to the fact that a limited amount of data is used to train it. For an arbitrary point that is not part of the training data, the prediction variance provides an estimate of the prediction uncertainty, which allows assessing its confidence. It is important to note that, if the target function (1) does come from the GP described by $\mu(\boldsymbol{x})$ and $\sigma^2 r(\boldsymbol{x}, \boldsymbol{x}')$, the covariance information is rigorous and the variance quantitatively accounts for the prediction uncertainty, which in that case is solely due to the lack of data.

A. Parameterization of the Prior

In signal and power integrity problems, the GP assumption may hardly hold. Nevertheless, we can still assume that there exist a *certain* GP of which the target function may be a possible realization! Even in that case, however, it is difficult to guess a priori a good prior GP. What is typically done in engineering applications is to assume a generic, parameterized form of the prior, and to leave some prior parameters as degrees of freedom to be optimized during training.

Typically, the trend is assumed to be a linear combination of predefined basis functions (e.g., polynomials up to a given order), i.e.,

$$\mu(\boldsymbol{x}) = \sum_{j=1}^{P} \beta_j h_j(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{h}(\boldsymbol{x}) \tag{5}$$

with $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_P(\boldsymbol{x}))^{\mathsf{T}}$ and unknown coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^{\mathsf{T}}$. The trend function can already capture the main behavior of the target function w.r.t. the input parameters. For stochastic problems, a PCE can be used for the trend, where the basis functions in (5) are the classical orthogonal polynomials of the Wiener-Askey scheme, leading to the so-called polynomial-chaos-based Kriging [11]. Since in this work we aim at UQ, we shall choose this trend without loss of generality.

The kernel function is also parameterized. For relatively smooth problems, popular choices are the squared-exponential or the Matérn 5/2 kernels, the latter reading

$$\sigma^2 \, r(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \left(1 + \sqrt{5}\rho + \frac{5}{3}\rho^2 \right) \exp\left(-\sqrt{5}\rho \right), \tag{6}$$

where

$$\rho = \sqrt{\sum_{j=1}^{d} \frac{(x_j - x_j')^2}{\theta_j^2}} \tag{7}$$

The kernel is parameterized by the variance σ^2 and the smoothness parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, also called "length-scales". Again without loss of generality, we assume in this work the lengthscale to be the same for all inputs, i.e., $\theta_j = \theta$, $\forall j = 1, \ldots, d$, which makes the kernel "isotropic".

The triplet of parameters $(\boldsymbol{\beta}, \sigma^2, \theta)$ is estimated and optimized during the training phase. It is important to point out that the posterior covariance (4) does *not* account for the additional uncertainty related to the estimation of these parameters, thereby leading to an overoptimistic estimate of the prediction uncertainty, especially when a low number of training samples is used.

B. Application to UQ

Assuming that the prior parameters are known, the outlined GP framework can be used for UQ. In particular, the GP surrogate is used as an emulator of the true computational model (1) for the fast predictions of samples in a Monte Carlo (MC)-like analysis.

To this end, a randomly drawn set of design parameters $\{\boldsymbol{x}_i^*\}_{i=1}^N$ is generated according to their distribution. Corresponding predictions are generated by evaluating (3), leading to

$$\bar{\boldsymbol{y}} = \boldsymbol{\mu}_* + \boldsymbol{R}_*^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \tag{8}$$

where $\boldsymbol{\mu}_*$ is the prior trend evaluated at the MC samples and $\boldsymbol{R}_*$ is the cross-correlation matrix between the MC samples and the training samples (i.e., $R_{*,li} = r(\boldsymbol{x}_l, \boldsymbol{x}_i^*)$, with $l = 1, \ldots, L$ and $i = 1, \ldots, N$). Moreover, the covariance matrix of the MC predictions is obtained from (4) as

$$\boldsymbol{C}_0 = \sigma^2 \left(\boldsymbol{R}_{**} - \boldsymbol{R}_*^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{R}_* \right) \tag{9}$$

where $R_{**,ij} = r(\boldsymbol{x}_i^*, \boldsymbol{x}_j^*)$, with $i, j = 1, \ldots, N$.

While (8) provides the best prediction of the MC samples, (9) can be used to assess its confidence. Specifically, the vector of MC predictions is a multivariate Gaussian random variable with mean and covariance given by (8) and (9), respectively, which can be expressed as

$$\boldsymbol{y}^* = \bar{\boldsymbol{y}} + \boldsymbol{C}_0^{\frac{1}{2}} \boldsymbol{\xi} \tag{10}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^\mathsf{T}$ is a vector of N independent standard normal random variables, i.e., $\xi_i \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, N$. A different realization of $\boldsymbol{\xi}$, plugged into (10), provides a different prediction of the MC samples taking into account the uncertainty of the surrogate model due to the limited training data, thus allowing the estimation, e.g., of confidence bounds on the statistical information. Analytical estimates are available for the mean and the variance of the output y [9].

C. Estimation of the Prior Parameters

Usually, the GP parameters are estimated empirically from the training data. The trend coefficients $\boldsymbol{\beta}$ are estimated using a generalized least-square estimate [6], [10], leading to

$$\boldsymbol{\beta} = \left(\boldsymbol{H}^\mathsf{T} \boldsymbol{R}^{-1} \boldsymbol{H} \right)^{-1} \boldsymbol{H}^\mathsf{T} \boldsymbol{R}^{-1} \boldsymbol{y} \tag{11}$$

where $\boldsymbol{H}$ is a matrix with the trend basis functions evaluated at the training samples, i.e., $H_{lj} = h_j(\boldsymbol{x}_l)$, with $l = 1, \ldots, L$ and $j = 1, \ldots, P$. The kernel parameters σ^2 and θ are instead usually optimized by either maximum likelihood estimation or cross-validation error minimization using global optimization algorithms [6], [10].

Once the prior parameters are known, they are plugged in in the expressions of the trend (5) and kernel (6) to compute the surrogate model prediction of the MC samples via (8)–(10). The estimate of the prior parameters is avoidably inexact but, as already mentioned, the posterior covariance (4) does not account for this contribution to the overall model uncertainty.

D. Impact of Parameter Estimation on Prediction Uncertainty

This section illustrates how the additional uncertainty related to the estimation of the prior parameters can be included in the predictions. Only the main results are discussed here, whereas a detailed discussion is deferred to an expanded paper.

Using Bayesian settings, it has been shown that the uncertainty in the estimation of the trend coefficients $\boldsymbol{\beta}$ can be accounted for by including an additional term in the posterior covariance [6], [10], [12], leading to a modified covariance matrix that we can express as

$$\boldsymbol{C}_1 = \boldsymbol{C}_0 + \Delta \boldsymbol{C} \tag{12}$$

The posterior distribution remains Gaussian, which means that we can still generate the MC predictions by means of (10), using $\boldsymbol{C}_1$ in place of $\boldsymbol{C}_0$.

It is possible to use Bayesian inference also to estimate the kernel variance σ^2. This leads to a different expression of the covariance matrix, which we denote as $\boldsymbol{C}_2$. However, in this case, the posterior distribution becomes no longer Gaussian, but rather a Student's t-distribution [10], [12]. Therefore, the MC predictions can be computed as

$$\boldsymbol{y}^* = \bar{\boldsymbol{y}} + \boldsymbol{C}_2^{\frac{1}{2}} \boldsymbol{\tau} \tag{13}$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)^\mathsf{T}$ is a vector of N independent t-distributed random variables, i.e., $\tau_i \sim t_\nu$, with $\nu = L - P$ degrees of freedom. It should be noted that a t-distribution is similar to a standard normal distribution but with larger variance ($= \nu/(\nu - 2) > 1$), which in general leads to a lower confidence of the surrogate model predictions. The t-distribution approaches the Gaussian distribution (i.e., its variance approaches 1) when the degrees of freedom $\nu \to \infty$.

Interestingly, ν increases (i.e., the uncertainty reduces) when the number of training samples L is increased and/or the number of trend basis functions is reduced. This is reasonable, because the former yields a more accurate training, whereas the latter requires the estimation of a higher number of coefficients.

Unfortunately, it is much more difficult to account also for the uncertainty due to the estimation of the lengthscale θ, because in this case the posterior distribution is no longer available in closed form [10], [12]. Therefore, in this work we use the following compromise:

- We estimate the kernel lengthscale θ using a standard empirical method, in this case maximum likelihood (but leave-one-out cross validation could be alternatively used);
- Once the lengthscale scale is available, we use Bayesian estimators to account for the uncertainty in the estimation of the trend coefficients $\boldsymbol{\beta}$ and kernel variance σ^2.

The above approach unavoidably neglects the uncertainty in the estimation of θ. Nonetheless, it still provides more conservative prediction confidence compared to the standard GP formulation.

III. Numerical Results

In this section, we apply the outlined GP framework to the UQ of the maximum transient crosstalk in the coupled embedded microstrip line originally introduced in [13] and already investigated in [14]. In particular, two test cases are considered, with $d = 2$ and $d = 6$ uncertainty parameters. In the first test case (TC-1), the uncertainty is in the substrate thickness and the line gap. In the second test case (TC-2), the uncertainty is in the line widths and in their distance from the ground plane. All parameters are assumed to be independent and Gaussian distributed with a 10% standard deviation from the mean. We refer to [14] for additional details.

The transient crosstalk in the microstrip interconnect, produced by a 1-ns pulse with an amplitude of 5 V and a risetime of 100 ps, is simulated in SPICE. For each random configuration of the uncertain parameters, the maximum over time is considered as the target quantity y. We train a GP surrogate of the maximum crosstalk with a second-order PCE as trend function and an isotropic Matérn 5/2 kernel. The

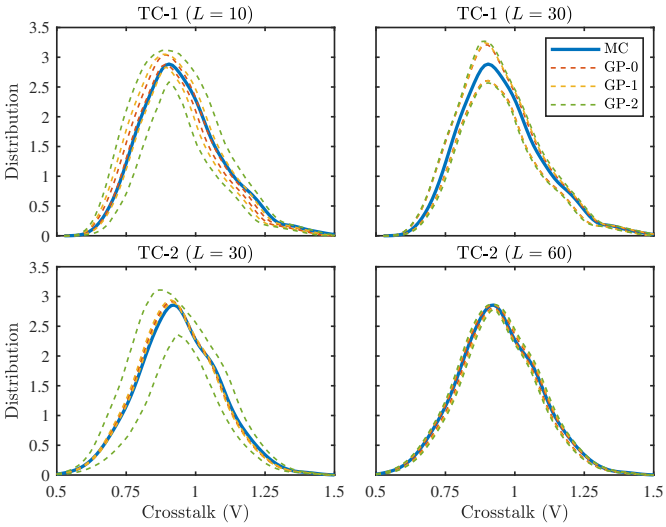

Fig. 1. Distribution of the maximum crosstalk for the two test cases and different number of training samples. Blue line: reference distribution of the MC samples; dashed lines: confidence levels of the various GP models.

number of trend basis functions is $P = 6$ for TC-1 and $P = 28$ for TC-2. The training data is generated using Latin hypercube sampling over the domain $[-3, 3]^d$ and we use the MATLAB® Statistics and Machine Learning Toolbox™ [15] toolbox to estimate the kernel lengthscale. Finally, a MC simulation with 1000 samples is run to generate reference results.

Figure 1 shows the results for the probability distribution of the maximum crosstalk. The top and bottom panels refer to TC-1 and TC-2, respectively, whereas the results of the left and right panels refer to a different number of training samples. In particular, either $L = 10$ (left) or $L = 30$ (right) is used for TC-1, whilst for TC-2, we use either $L = 30$ or $L = 60$.

The distribution of the MC samples is shown by the solid blue lines. The dashed lines are instead the 95% confidence bounds of the distribution predicted with the GP surrogates. In particular:

- in red are the confidence bounds obtained with the standard GP implementation, labeled as "GP-0", with the covariance matrix (9), which does not account for the uncertainty introduced by the estimation of the prior parameters;
- in yellow are the confidence bounds obtained with the corrected covariance matrix (12), labeled as "GP-1", which only accounts for the additional uncertainty introduced by estimation of the trends coefficients $\boldsymbol{\beta}$;
- the green lines are the confidence bounds obtained from (13), labeled as "GP-2", which also account for the uncertainty in the estimation of the kernel variance σ^2.

We notice that the GP confidence bounds are increasingly wider as we include more contributions to the model uncertainty. GP-0 and GP-1 confidence bounds are overoptimistic, especially for the lower number of training samples. This is clearly observed in particular for TC-2 and $L = 30$: very narrow bounds are predicted, which however do not enclose the actual MC distribution. GP-2 provides much more conservative bounds, better reflecting the actual model uncertainty. The confidence difference between the three GP models bounds reduces by increasing L, since the estimation of the prior parameters becomes more accurate. In this case, despite still being narrower, the GP-0 and GP-1 confidence bounds include the reference MC distribution, indicating that the GP model is overall more accurate.

IV. Conclusions

In this paper, we discussed more conservative GP regression formulations that account also for the prediction uncertainty arising from the estimation of some prior parameters from the training data. The results were illustrated based on the simulation of the maximum transient crosstalk in a microstrip interconnect, for which more conservative and reliable confidence bounds were obtained.

References

[1] R. Trinchero, P. Manfredi, I. S. Stievano, and F. G. Canavero, "Machine learning for the performance assessment of high-speed links," *IEEE Trans. Electromagn. Compat.*, vol. 60, no. 6, pp. 1627–1634, Dec. 2018.

[2] R. Trinchero, M. Larbi, H. M. Torun, F. G. Canavero, and M. Swaminathan, "Machine learning and uncertainty quantification for surrogate models of integrated devices with a large number of parameters," *IEEE Access*, vol. 7, pp. 4056–4066, 2018.

[3] 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.

[4] T. Nguyen, B. Shi, H. Ma, E.-P. Li, X. Chen, A. C. Cangellaris, and J. Schutt-Aine, "Comparative study of surrogate modeling methods for signal integrity and microwave circuit applications," *IEEE Trans. Compon., Packag., Manuf. Technol*, vol. 11, no. 9, pp. 1369–1379, Sep. 2021.

[5] P. Manfredi and D. Vande Ginste, "Polynomial chaos based uncertainty quantification in electrical engineering: theory," in *Uncertainty Quantification of Electromagnetic Devices, Circuits, and Systems*, S. Roy, Ed. Stevenage, U.K.: IET, 2021.

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

[7] 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.

[8] F. Garbuglia, D. Spina, D. Deschrijver, I. Couckuyt, and T. Dhaene, "Bayesian optimization for microwave devices using deep GP spectral surrogate models," *IEEE Trans. Microw. Theory Techn.*, 2022.

[9] P. Manfredi, "Probabilistic uncertainty quantification of microwave circuits using Gaussian processes," *IEEE Trans. Microw. Theory Techn.*, 2022.

[10] V. Dubourg, "Adaptive surrogate models for reliability analysis and reliability-based design optimization," Ph.D. dissertation, Université Blaise Pascal-Clermont-Ferrand II, 2011.

[11] R. Schobi, B. Sudret, and J. Wiart, "Polynomial-chaos-based Kriging," *Int. J. Uncertainty Quantification*, vol. 5, no. 2, 2015.

[12] T. J. Santner, B. J. Williams, W. I. Notz, and B. J. Williams, *The Design and Analysis of Computer Experiments*, 2nd ed. New York, NY, USA: Springer, 2003.

[13] M. Ahadi and S. Roy, "Sparse linear regression (spliner) approach for efficient multidimensional uncertainty quantification of high-speed circuits," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst*, vol. 35, no. 10, pp. 1640–1652, 2016.

[14] P. Manfredi and R. Trinchero, "Statistical crosstalk analysis via probabilistic machine learning surrogates," in *Proc. 30th IEEE Conf. Elect. Perform. Electron. Packag. Syst.*, Oct. 17-20, 2021, pp. 1–3.

[15] *Statistics and Machine Learning Toolbox, Version 12.1*. Natick, MA, USA: The MathWorks, Inc., 2021.