Error estimation for surrogate models with noisy small-sized training sets

Jeroen Wackers, Hayriye Pehlivan Solak, Riccardo Pellegrini, Andrea Serani, Matteo Diez

International Conference on Adaptive Modeling and Simulation
ADMOS 2023
F.Larsson and P.Díez
CIMNE, Barcelona, 2023

ERROR ESTIMATION FOR SURROGATE MODELS WITH NOISY SMALL-SIZED TRAINING SETS

JEROEN WACKERS[†,*], HAYRIYE PEHLIVAN SOLAK[†], RICCARDO PELLEGRINI[‡], ANDREA SERANI[‡] AND MATTEO DIEZ[‡]

[†] LHEEA Lab, Ecole Centrale de Nantes / CNRS UMR 6598, F–44000 Nantes, France
e-mail: jeroen.wackers@ec-nantes.fr

[‡] CNR-INM, National Research Council-Institute of Marine Engineering
Via di Vallerano 139, 00128, Rome, Italy

Key words: Surrogate Modeling, Uncertainty Estimation, Noise Filtering, Small Data

1 INTRODUCTION

Simulation-driven shape optimization often uses surrogate models, i.e. approximate models fitted through a dataset of simulation results for a limited number of designs. The shape optimization is then performed over this surrogate model. For efficiency, modern approaches often construct the datasets adaptively, adding simulation points one by one where they are most likely to discover the optimum design [3].

The uncertainty estimation of the surrogate model is essential to guide the choice of new sample points: underestimation of the uncertainty leads to sampling in suboptimal regions, missing the true optimum. Gaussian process regression naturally provides uncertainty estimations [4] and Stochastic Radial Basis Functions (SRBF) surrogate models estimate the uncertainty based on the spread of RBF fits with different kernels [5].

In the context of SRBF, this paper discusses two issues with uncertainty estimation. The first is that most existing techniques rely on knowledge about the global behaviour of the data, such as spatial correlations. However, the number of datapoints can be too small to reconstruct this global information from the data. We argue that in this situation, user-provided estimation of the function behaviour is a better choice (section 3).

The second issue is that the dataset may contain noise, i.e. random errors without spatial correlation. Surrogate models can filter out this noise, but it introduces two separate uncertainties: the optimum amount of noise filtering is unknown, and for a small dataset (even with perfect noise filtering) the local mean of the data may not correspond to the true simulation response. In section 4 we introduce estimators for both uncertainties.

2 STOCHASTIC RBF SURROGATE MODEL

The principle of the SRBF model is defined in [5]. Consider $\mathbf{x} \in \mathbb{R}^D$ as a design variable vector of dimension D. Let the true function $f(\mathbf{x})$ be assessed by observations that are (potentially) perturbed by random white noise: $s(\mathbf{x}) = f(\mathbf{x}) + \mathcal{N}(0, \sigma_n)$. Given

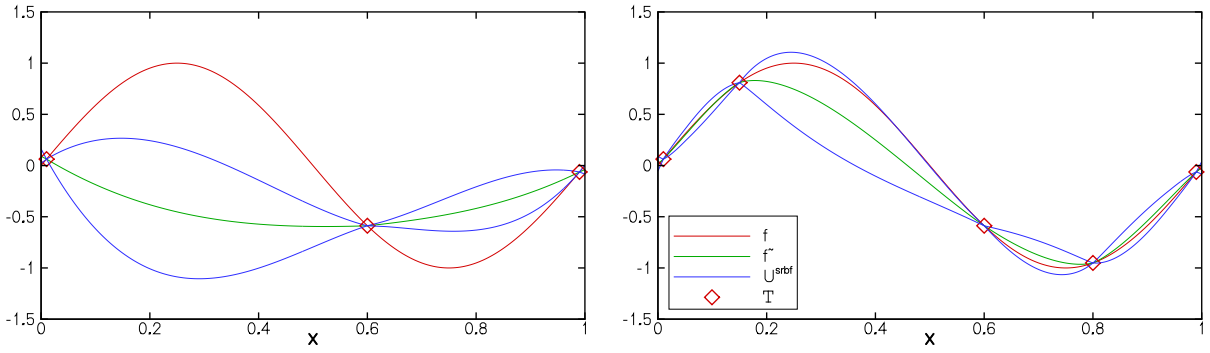

Figure 1: SRBF uncertainty estimation for a sine-wave function. 2 (left) and 3 (right) points per peak.

a training set $\mathcal{T} = \{\mathbf{x}_i, s(\mathbf{x}_i)\}_{i=1}^{J}$ of size J, the SRBF prediction $\widetilde{f}(\mathbf{x})$ (where $\widetilde{\cdot}$ denotes surrogate model prediction) is computed as the expected value (EV) over a stochastic tuning parameter $\tau \sim \text{unif}[1,3]$ of a surrogate model $\widetilde{g}$:

$$\widetilde{f}(\mathbf{x}) = \text{EV}\left[\widetilde{g}(\mathbf{x},\tau)\right]_\tau, \quad \text{with} \quad \widetilde{g}(\mathbf{x},\tau) = \text{EV}\left[\mathbf{s}\right] + \sum_{j=1}^{M} w_j ||\mathbf{x} - \mathbf{c}_j||^\tau, \tag{1}$$

where w_j are unknown coefficients, $||\cdot||$ is the Euclidean norm and $\mathbf{c}_j$ are the RBF centres, with $j = 1, \ldots, M$ and $M \leq J$. The uncertainty $U^{\text{srbf}}(\mathbf{x})$ associated with the SRBF approach is quantified as the 95% confidence interval of the predictions $g(\mathbf{x}, \tau)$.

If the data are not affected by numerical noise ($\sigma_n = 0$), exact interpolation of the training set is imposed and the weights w_j are computed by solving $\mathbf{A}\mathbf{w} = (\mathbf{s} - \text{EV}\left[\mathbf{s}\right])$, with $\mathbf{c}_j = \mathbf{x}_j$ (i.e. $M = J$) and $\mathbf{s} = \{s(\mathbf{x}_i)\}_{i=1}^{J}$. In the presence of noise, [3] choose a number of RBF centres M smaller than the number of training points J, and $\mathbf{c}_j$ coordinates are defined via k-means clustering [1] of the training point coordinates. The weights w_j are determined with least squares regression by solving $\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{w} = \mathbf{A}^\mathsf{T}(\mathbf{s} - \text{EV}\left[\mathbf{s}\right])$.

3 INTERPOLATION UNCERTAINTY IN SMALL-DATA CASES

The SRBF uncertainty estimation U^{srbf} is highly accurate for surrogate models without noise, if enough training points are available to represent f more or less correctly. The 95% confidence interval of $g(\mathbf{x}, \tau)$ is close to $g(\mathbf{x}, 3) - g(\mathbf{x}, 1)$, where $g(\mathbf{x}, 1)$ is $\mathcal{C}^0$ and piecewise linear, while $g(\mathbf{x}, 3)$ is piecewise cubic and $\mathcal{C}^2$. Our tests show that this difference is a good estimator for the missing above-cubic terms, as long as the second derivative of f is approximated correctly by the metamodel (figure 1 right). When insufficient points are available to capture the second derivatives, the uncertainty estimation fails (figure 1 left).

We refer to a "small-data" situation when (a) the true function behaviour cannot be estimated from the data, and (b) the data cannot indicate that the approximation of the true function is incorrect. In this case, the only way to evaluate the uncertainty is with user-provided estimations of the behaviour of f as a supplement to the data. While reliance on user knowledge is a weakness for automatic surrogate model construction, we consider it as inevitable. This section presents a small-data uncertainty estimation.

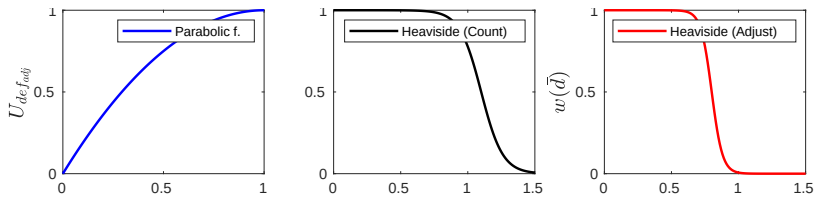

Figure 2: Adjustment functions: default uncertainty U^{def}, counting function $\bar{d}(r)$, and weight $w(\bar{d})$.

Default uncertainty If f is not known well, the uncertainty estimation must be based on assumptions about the function, rather than information from the training points. Since the minimum of f is being sought, the function is assumed to consist of peaks and valleys with a parabolic behaviour, a typical peak width $2r_0$ and a typical peak height U_0. Both these parameters need to be estimated from another source than the training points. In the following, r_0 is estimated as $\frac{1}{4}$ times the domain size, while U_0 is taken as $\max_i s(\mathbf{x}_i) - \min_i s(\mathbf{x}_i)$. These are reasonable default choices for any function, but if more reliable estimates are available for a function, the uncertainty estimation will be better.

The parabolic behaviour between data points is used to define a default uncertainty U^{def}, based only on the distance to the closest training point $r_{i_{min}}(\mathbf{x}) = \min_{i=1\ldots J}\|\mathbf{x}-\mathbf{x}_i\|$:

$$U^{\text{def}}(\mathbf{x}) = \begin{cases} U_0(1-(r_{i_{min}}(\mathbf{x})/r_0 - 1)^2) & r_{i_{min}}(\mathbf{x})/r_0 \leq 1, \\ U_0 & r_{i_{min}}(\mathbf{x})/r_0 > 1. \end{cases} \tag{2}$$

This and other functions used in the modified uncertainty are shown in figure 2.

Blending the uncertainties As noted above, the SRBF uncertainty is reliable when the second derivatives of f are well represented by the data. In each peak or valley, ignoring cross-derivatives, a central point plus 2 points for each dimension are needed to capture the second derivatives. Thus, the original SRBF uncertainty can be considered as valid in points $\mathbf{x}$ where at least $2D+1$ training points are found in a sphere of radius r_0 around $\mathbf{x}$. If fewer training points are close to $\mathbf{x}$, the default uncertainty should be used.

A smooth transition between the two uncertainty estimations is obtained with two smeared Heaviside functions. To prevent a sharp distinction between points on the inside of the region and just outside, the number of training points within the r_0 neighborhood region is counted as:

$$\bar{d}(\mathbf{x}) = \sum_{i=1}^{n} \left(1 - \frac{1}{1+e^{-12.5\left(\frac{r_i(\mathbf{x})}{r_0}-1.1\right)}}\right). \tag{3}$$

A weight w for the default uncertainty is evaluated based on $\bar{d}$:

$$w(\bar{d}) = 1 - \frac{1}{1+e^{-25\left(\frac{\bar{d}}{2D+1}-0.8\right)}}. \tag{4}$$

Using w, the modified interpolation uncertainty estimation is defined as:

$$U^{\text{interp}}(\mathbf{x}) = U^{\text{srbf}}(\mathbf{x})\left(1 - w\left(\bar{d}(\mathbf{x})\right) + U^{\text{def}}(\mathbf{x})\, w(\bar{d}(\mathbf{x})\right). \tag{5}$$

4 NOISE CANCELING AND TRAINING-POINT UNCERTAINTY

LS fitting of the metamodels $\widetilde{g}$ (section 2) is effective for filtering noise, but has two disadvantages. The interpolation error estimator is ill founded for non-interpolating surrogate models and exhibits minima which are not located in the training points. And if the LOOCV procedure chooses a high number of kernels, overfitting can occur [6].

Therefore, we define a new noise canceling approach, where noise-filtered data $\bar{f}_i$ are first reconstructed in the training points. Analogous to SRBF, these data are a weighted average of fits with different noise levels; this reduces the risk of overfitting caused by one extreme fit. Standard SRBF interpolation is then used to construct a surrogate model from $\bar{f}_i$, which eliminates the problem of the interpolation error estimation.

The training data uncertainty has two components. The unknown amount of noise in the data introduces an uncertainty in the training point reconstruction, which can be estimated from the variance of the different fits. And for a small dataset, even with perfect noise filtering, the local mean of the data may not correspond to the true simulation response. This mean-value uncertainty is estimated with the central limit theorem.

Reconstruction and its uncertainty Least-squares fitting is retained here to remove the noise in the training points. For this, fits $\widetilde{g}^M(\mathbf{x}, \tau)$ are computed as in section 2, with numbers of RBF kernels M ranging from $2D$ to $J-1$; the lower M is, the more smoothing is applied. These fits are evaluated in the training points $\mathbf{x}_i$ and the noise-filtered training data $\bar{f}_i$ are based on the average of the fits. Since the uncertainty to be estimated is the noise filtering instead of the interpolation, the fits are computed for only one kernel parameter $\tau = 3$, which gives the most accurate interpolation.

Depending on the noise that is actually present, not all M are equally likely. Leave-one-out cross-validation (LOOCV) can indicate the most likely fits. Let $\widetilde{g}^M_{-i}(\mathbf{x})$ be a surrogate model trained with all the noisy data $\mathcal{T} = \{\mathbf{x}_i, s(\mathbf{x}_i)\}_{i=1}^J$ except the i-th point, using M centres. Then the LOOCV error in the point i is:

$$e_i^M = \left| s(\mathbf{x}_i) - \widetilde{g}^M_{-i}(\mathbf{x}_i) \right|. \tag{6}$$

These LOOCV point errors are translated into an estimated likelihood for each M, which requires an estimation of the noise level σ_n:

$$\sigma_n \approx \sigma_{CV} = \min_M \sqrt{\frac{1}{J} \sum_{i=1}^{J} (e_i^M)^2}, \tag{7}$$

where the minimum of the LOOCV errors is taken as the closest approximation of the noise, since the reconstruction with the lowest error is probably closest to the true function. The likelihood of each M [4] is the probability of the training point data $s(\mathbf{x}_i)$ given the surrogate model $\widetilde{g}^M_{-i}$ and white noise $\sim \mathcal{N}(0, \sigma_{CV})$. This probability is:

$$\mathcal{L}(M) = \prod_{i=1}^{J} p\left(s(\mathbf{x}_i) \,\middle|\, \widetilde{g}^M_{-i}(\mathbf{x}_i) \right) = \prod_{i=1}^{J} \frac{\sqrt{2}}{\sigma_{CV}\sqrt{\pi}} \exp\left(-\tfrac{1}{2} \left(\frac{e_i^M}{\sigma_{CV}} \right)^2 \right). \tag{8}$$

The likelihoods $\mathcal{L}(M)$ for $M = 2D \dots J-1$ are finally scaled to form a partition of unity. $\mathcal{L}$ serves to reconstruct the data $\bar{f}_i$ with a weighted average of the non-LOOCV fits $\widetilde{g}^M$, while the variance provides an uncertainty for the noise filtering:

$$\bar{f}_i = \sum_M \mathcal{L}(M)\widetilde{g}_M(\mathbf{x}_i), \qquad (\sigma_i^{\text{filt}})^2 = \sum_M \mathcal{L}(M)\left(\widetilde{g}_M(\mathbf{x}_i) - \bar{f}_i\right)^2. \tag{9}$$

Since LOOCV is based on interpolation, it does not work if too few training points are available. Therefore, the training points which have fewer than $2D+1$ neighbours within a distance r_0 are ignored for the LOOCV and their training data $s(\mathbf{x}_i)$ are kept unmodified.

Mean value uncertainty The estimation (9) captures the uncertainty in the noise filtering, i.e. the reconstruction of the local mean of the data from the training point results. However, for a small number of data, the local mean does not necessarily correspond to the true function $f(\mathbf{x})$. This introduces a second training point uncertainty.

According to the central limit theorem, the mean of n realizations of a function with stochastic noise is another stochastic variable whose standard deviation is the noise standard deviation divided by $\sqrt{n}$. Therefore, estimating the uncertainty in the mean value of $s(\mathbf{x}_i)$ requires an estimation of the noise level and an indication of how many training points contribute to the local mean value; both vary with M. For safety, it is preferable to overestimate the noise, so a different estimate than (7) is selected, i.e. the highest noise level for which the outcome $s(\mathbf{x}_i)$ is in the 95% confidence interval:

$$\sigma_{n,M} = \sqrt{\frac{1}{F^{-1}(0.025, J)} \sum_{i=1}^{J} \left(s(\mathbf{x}_i) - \widetilde{g}^M(\mathbf{x}_i)\right)^2}, \tag{10}$$

where $F^{-1}(0.025, J)$ is the inverse of the cumulated distribution function for the chi-squared distribution with J samples, evaluated at a probability of 2.5%.

The (probably pessimistic) estimated number of training points which contribute to each local minimum is the number of training points I_i^M which are k-means clustered into the same RBF centre as point i, when M centres are used. The mean-value uncertainty for M centres is then:

$$\sigma_{m,M}(\mathbf{x}_i) = \frac{\sigma_{n,M}}{\sqrt{I_{ij}}}, \tag{11}$$

The final estimated mean-value variance is weighted like the noise-filtering variance:

$$(\sigma_i^{\text{mean}})^2 = \sum_M \mathcal{L}(M)(\sigma_{m,M}(\mathbf{x}_i))^2. \tag{12}$$

Total uncertainty Assuming that the mean-value and noise-canceling uncertainties are independent, the standard deviation of the total datapoint uncertainty in point i is:

$$\sigma_i^{\text{data}} = \sqrt{(\sigma_i^{\text{filt}})^2 + (\sigma_i^{\text{mean}})^2}. \tag{13}$$

The 95% confidence interval $U_i^{\text{data}} = 2\sigma_i^{\text{data}}$ is interpolated like $\widetilde{f}$ and added to U^{interp}.

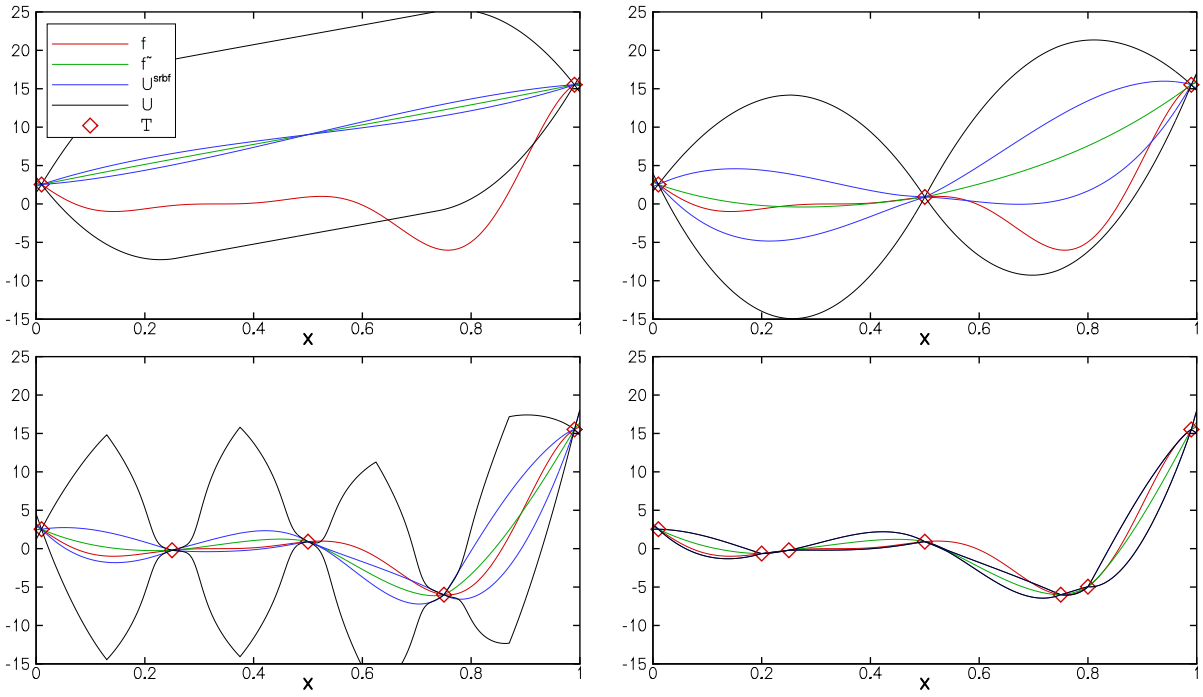

Figure 3: Forrester without noise: interpolation uncertainty with 2, 3, 5 and 7 training points.

5 TEST CASES

Interpolation The interpolation uncertainty estimation is tested on the 1D Forrester function [2]:

$$f_1(x) = (6x - 2)^2 \sin(12x - 4). \tag{14}$$

With 2 and 3 datapoints (figure 3) the default uncertainty is used everywhere, which is the right choice: unlike U^{srbf}, the modified uncertainty interval contains most of the true function. For 5 data points, U^{srbf} becomes reliable and U^{def} starts to be switched off, while for 7 points, U^{srbf} is selected everywhere. Looking in detail, the estimation is pessimistic for 5 points, while for 7 points the true function leaves the uncertainty domain once. This is because the peak width r_0 which was chosen as a compromise to fit many different functions, does not correspond to the actual peak widths for Forrester. Given this limitation, the new estimation predicts the uncertainty with a reasonable accuracy.

Noise filtering Figure 4 shows three surrogate models for the Forrester function with noise. The first one has $\sigma_n = 1.5$ and clustered data. The interpolation uncertainty varies with the distance between sampling points and even U^{def} is used. The datapoint uncertainty is reduced around the cluster, thanks to the lower mean-value uncertainty. The right figure shows that 3 values of M are the most likely. These coincide both with the minimum of the RMSE and with the minimum true error, showing the efficiency of the likelihood estimator. Finally, the noise $\sigma_{n,M}$ is overestimated w.r.t. σ_n as desired, but the order of magnitude is correct. These observations are valid for all three tests.

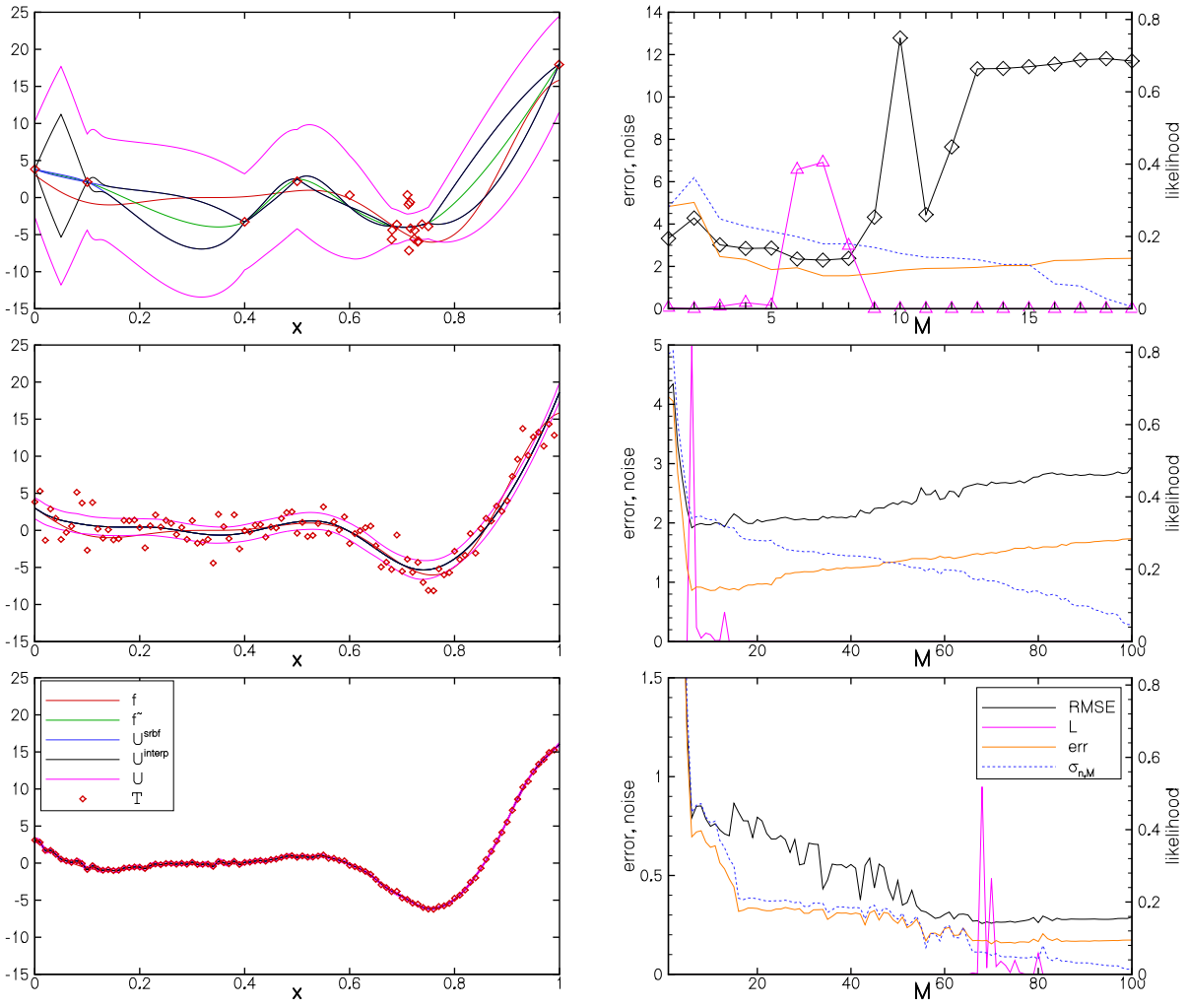

Figure 4: Surrogate models for Forrester with noise: 20 points, $\sigma_n = 1.5$ (top), 100 points, $\sigma_n = 1.5$ (middle), and 100 points, $\sigma_n = 0.15$ (bottom). Right figures: the noise and errors of the fits with different M. $RMSE$ is the function minimised in (7), err the RMS difference between the true f and each fit.

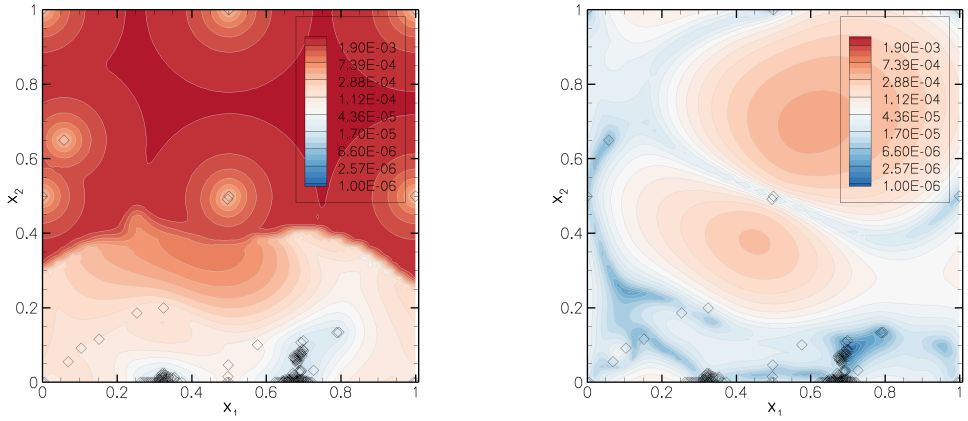

Figure 5: NACA airfoil 2D uncertainty with the new approach (left) and with LS-SRBF from [3] (right).

The middle image retains $\sigma_n = 1.5$ but has 100 equidistributed training points. For this point density, U^{interp} is negligible. The total uncertainty interval is smaller than the spread of the data, indicating effective noise filtering. Also, the uncertainty is smaller than in the cluster for the first case, although the local point density is lower. Thus, the data uncertainty in a given point is non-local; it depends on the data in a region around the point. With a noise level $\sigma_n = 0.15$ (bottom row) the LOOCV automatically detects that less smoothed fits (higher M) are more likely, and changes the chosen fits.

The final test (figure 5) uses the 145-point low-fidelity dataset from the 2-parameter two-fidelity airfoil optimisation of [3], which has a valley-like response shape with a minimum around [0.3, 0] and at least 10% noise. The new approach is compared with the LS-SRBF uncertainty estimation we presented in [3] (see section 2). For the new approach, the neighbour count of equation (3) varies abruptly since the data are highly clustered, which explains the rapid change to the default uncertainty in the top half of the domain. The separation of datapoint and interpolation uncertainty ensures that the uncertainty minima are in the training points. Also, the clustered data reduce the mean-value uncertainty, which leads to minimum zones around the clusters. The LS-SRBF approach however, predicts the minimum uncertainty in positions next to the data, which likely coincide with the RBF centre positions. Altogether, the new uncertainty estimation is credible and appears to be a good basis for adaptive sampling.

ACKNOWLEDGEMENTS

The work at ECN is funded by the Institut Carnot MERS in the ORUP project. CNR-INM is partially supported by the Office of Naval Research through NICOP grant N62909-21-1-2042, administered by Dr. Elena McCarthy and by Dr. Woei-Min Lin of the Office of Naval Research Gobal and the Office of Naval Research, respectively.

REFERENCES

[1] S. Lloyd, "Least squares quantization in PCM". *IEEE Trans. Information Theory* **28**(2), 129–137, (1982).

[2] L. Mainini et al., "Analytical benchmark problems for multifidelity optimization methods". *arXiv preprint arXiv:2204.07867*, (2022).

[3] R. Pellegrini et al., "A multi-fidelity active learning method for global design optimization problems with noisy evaluations". *Engng. with Comput.*, (2022).

[4] C.E. Rasmussen and C.K.I. Williams, *Gaussian processes for machine learning*. The MIT Press, (2006).

[5] S. Volpi et al., "Development and validation of a dynamic metamodel based on stochastic radial basis functions and uncertainty quantification". *Struct. Multidiscipl. Optim.* **51**(2), 347–368, (2015).

[6] J. Wackers et al., "Efficient initialization for multi-fidelity surrogate-based optimization". *J. Ocean Engng. Marine Energy*, (2022).