Available online at www.eccomasproceedia.org
Eccomas Proceedia UNCECOMP (2025) 1-15

ECCOMAS
Proceedia

UNCECOMP 2025
6th ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Science and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Rhodes Island, Greece, 15-18 June 2025

HETEROSCEDASTIC GAUSSIAN PROCESS THROUGH KERNEL SMOOTHING

Ghifari A. Faza[1,3], **Nasrulloh R.B.S. Loka**[2], **and David Moens**[1,3]

[1] LMSD, Department of Mechanical Engineering, KU Leuven, Belgium e-mail: {adam.faza,david.moens}@kuleuv en.be

[2] Department of Computer Science, Helsinki University
e-mail: nasrulloh.satrio@helsinki.fi

[3] Flanders Make @KU Leuven, Belgium

Abstract. *Gaussian process (GP) regression is widely used in practice due to its ability to handle noisy problems. In many scientific experiments or stochastic simulators, observation noise can vary across the input space (i.e. heteroscedastic). This paper proposes a new approach to handle heteroscedastic noise that is inherently present in the data. Here, we follow a post-modelling learning strategy similar to the most likely heteroscedastic Gaussian process (MLHGP) algorithm. Unlike the MLHGP and its variations, which require multiple GP models, our proposed methodology only requires one GP model to fit the main function and uses the trained kernel parameters to estimate the noise level via kernel smoothing regression. We test our proposed model on three benchmark experiments: a one-dimensional analytical case, a four-dimensional computational simulation case, and a one-dimensional simple Bayesian optimisation problem to analyse how heteroscedastic modelling performs in a risk-averse setting. Our proposed method is able to reduce the computational complexity from $\mathcal{O}(2\mathcal{N}^3)$ to $\mathcal{O}(\mathcal{N}^3 + \mathcal{N}^2)$. As we don't need to train the kernel smoothing, our empirical results show that the proposed method is able to achieve speedup up to almost 2 times during training while not compromising the predictive capability. Additionally, in our Bayesian optimisation test case, the result shows that the proposed method outperforms MLHGP in case of a low number of initial observations and remains competitive in medium and high initial observation settings, all while being faster in every case.*

Keywords: Gaussian process, Heteroscedastic noise, Kernel smoothing.

ISSN:2623-3339
Peer-review under responsibility of the organizing committee of UNCECOMP 2025.
doi: 10.7712/120225.12336.21075

1 INTRODUCTION

Data-driven methods have gained traction as a viable alternative to numerical simulation for solving problems. Especially in problems that demand extensive computational resources such as design exploration, optimisation and reliability analysis, a data-driven surrogate model is often utilized to replace the original numerical simulation. These machine learning surrogate models, such as support vector regression [1], Gaussian process [2], and neural networks [3, 4], leverage data to approximate complex relationships, circumventing the need for computationally intensive simulations for every configuration

Compared to the other methods, the Gaussian process (GP) model is very attractive since it naturally provides a predictive distribution as an uncertainty estimate besides the predictive mean. This capability of GPs to characterize uncertainty has been leveraged in various applications, such as Bayesian optimisation [5–8], sensitivity analysis [9] and reliability analysis with active learning [10, 11]. The capability underscores the importance of developing more appropriate GP models that can effectively capture uncertainty, as it directly impacts the performance of applications leveraging this property.

In the standard GP regression, the noise level that is inherently present in the data is typically presumed to be constant throughout the input space (homoscedastic). However, in many real-world problems [12–14], the observation variability can heavily depend on the input (heteroscedastic). Hence, assuming that the data is homoscedastic can lead to underestimating the variance of high-noise regions and overestimating it in low-noise regions. As a result, a homoscedastic GP model may misrepresent uncertainty and provide suboptimal predictions in heteroscedastic settings. The illustration given in Figure 1 depicts the main difference between homoscedastic and heteroscedastic problems.

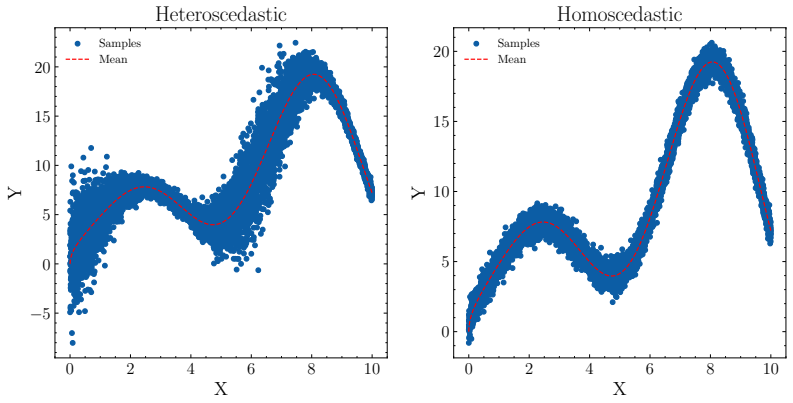


Figure 1: Varying noise in heteroscedastic function (left) and constant noise in homoscedastic function (right).

Various heteroscedastic Gaussian process (HGP) models [15–24] have been proposed to properly handle the noise level that varies across the input space instead of assuming a constant noise. The typical HGP configuratio uses two GPs, with the firs GP modelling the main function and the second GP learning the input-dependent noise level. Despite the apparent simplicity of the idea, the combination of the two GPs produces a joint posterior distribution over the main function and the heteroscedastic noise which is analytically intractable.

Markov chain Monte Carlo (MCMC) sampling is often viewed as the gold standard to obtain the joint posterior distribution [15]. However, due to its expensive cost, especially in large datasets, various alternatives have been developed to achieve a trade-off between accuracy and

computational efficien y such as the expectation propagation (EP) [16, 17], Laplace approximation [18], variational inferences [19–21], and most likely noise approaches [22–24]. Among these alternatives, the most likely noise approaches, including the most likely heteroscedastic GP (MLHGP) [22], improved MLHGP (IMLHGP) [23], and the nearest neighbour point estimate HGP (NNPEHGP) [24] are the simplest and most computationally attractive, in which the noise posterior is replaced by a point estimate at its most likely level. Thus, the predictive posterior distribution can be obtained analytically.

In this research, we propose a modificatio to the most likely noise approaches by using a kernel smoothing (KS) function [25–27] in place of the second GP. Our proposed method, named KS-MLHGP and KS-IMLHGP, is designed to enhance the adaptability of both MLHGP and IMLHGP. The primary goal of our proposed method is to address the issue of noise level overfitting particularly prevalent in IMLHGP when the training data is limited, while also reducing the computational cost of MLHGP. To validate the performances of the proposed methods, we compare their predictive capabilities in terms of the predicted distribution on several heteroscedastic problems consisting of one 1-dimensional analytical case, one 4-dimensional computational simulation, and one simple Bayesian optimisation case.

This paper is organized as follows: In section 2 we review the related work on heteroscedastic GP. Then we discuss our proposed algorithm using Kernel smoothing regression to approximate the noise level in section 3. Section 4 discusses the performance of our proposed algorithm compared to the current ones and finall , the conclusion and future works are given in Section 5.

2 PRELIMINARIES AND RELATED WORKS

The goal of using a Gaussian process as a surrogate model is to recover an unknown black-box function $f(\boldsymbol{x})$ from a dataset $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where $\boldsymbol{x} \in \mathbb{R}^d$ is the input vector of dimension d. The observed output of the unknown function is a scalar denoted by $y_i \in \mathbb{R}$ such that

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \tag{1}$$

where the observation error ε_i is typically assumed to be:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_o^2). \tag{2}$$

The observation noise σ_o^2 can be constant (homoscedastic) or varying in the input space governed by another unknown function $\sigma_o = g(\boldsymbol{x}_i)$ (heteroscedastic).

We can place a GP prior on the unknown black-box function $f(\boldsymbol{x})$ in equation 1, so it follows a multivariate normal distribution fully specifie by the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{K}$, written as:

$$p(f|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}), \tag{3}$$

where $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$ is a collection of data point $\boldsymbol{x_i}$. A zero-mean GP prior is often assumed over the function value for simplification hence $\boldsymbol{\mu} = \boldsymbol{0}$. The entries of the covariance matrix $\boldsymbol{K}$ are calculated from the covariance function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ at input points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. In the present study, we mainly focus on using the squared exponential (SE) kernel [2] as the covariance function, since we assume that our target function is smooth. The SE kernel is written as:

$$k_{SE}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_f^2 \exp\left(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||^2}{2l^2}\right). \tag{4}$$

The l variable usually refers to the kernel length-scale parameter, which roughly translates to how far information of a data point could affect others, σ_f^2 is the signal amplitude or the scaling

parameter, and $||\cdot||$ denotes the Euclidean distance between input locations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. This means the SE kernel is driven by two parameters $\theta_f = \{\sigma_f^2, l\}$ that measures the similarity between two observations.

The assumption of homoscedasticity and a zero-mean GP prior leads to the formation of the homoscedastic GP hyperparameters $\boldsymbol{\theta}_y = \{\boldsymbol{\theta}_f, \sigma_o^2\}$, where the model hyperparameters can be estimated from the data by maximizing the log marginal likelihood, written as:

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}_y) = -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T(\boldsymbol{K}+\sigma_o^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - \frac{1}{2}\log|\boldsymbol{K}+\sigma_o^2\boldsymbol{I}| - \frac{n}{2}\log(2\pi). \quad (5)$$

This can be obtained by using any optimisation algorithm with the goal of reaching an acceptable local maximum or ideally, global optimum.

In the heteroscedastic Gaussian process (HGP), the main idea is to employ another GP model to approximate the observed noise level $g(\boldsymbol{x}_i) = \sigma_o$, with a separate covariance function $k_z(\boldsymbol{x}_i, \boldsymbol{x}_j)$ parametrised by $\boldsymbol{\theta}_z$. Therefore, two GPs are involved in HGP modelling, where the firs GP learns the unknown main function (the y-process) and the second GP learns the unknown function of varying noise (the z-process). Now, the predictive posterior distribution over the test output y_* is given by:

$$p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}_f, \boldsymbol{\theta}_z, D) = \iint p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}_f, \boldsymbol{z}, z_*, D)p(\boldsymbol{z}, z_*|\boldsymbol{x}_*, \boldsymbol{\theta}_z, D)d\boldsymbol{z}dz_*, \quad (6)$$

where $\boldsymbol{z}$ is the noise level at input training and z_* is the noise level at $\boldsymbol{x}_*$. However the full posterior of the noise level $p(\boldsymbol{z}, z_*|\boldsymbol{x}_*, \boldsymbol{\theta}_z, D)$ is not solvable analytically and therefore one has to provide an approximation to solve it. Since the integral is intractable, various approximation methods have been proposed such as MCMC [15], variational inference [19–21], expectation propagation [16, 17], and Laplace approximation [18].

Other approaches that are also popular are the Most likely heteroscedastic Gaussian process (MLHGP) [22] and its variations, including the improved MLHGP (IMLHGP) [23] and the nearest neighbour point estimate HGP (NNPEHGP) [24]. These methods propose a simple and computationally cheap approach to approximate HGP. MLHGP replaces the full posterior of the varying noise with a point estimate at the most likely value such that the predictive posterior can be treated analytically. This reads as

$$p(\boldsymbol{z}, z_*|\boldsymbol{x}_*, \boldsymbol{\theta}_z, D) \approx \delta(\tilde{\boldsymbol{z}}, \tilde{z}_*), \quad (7)$$

where $(\tilde{\boldsymbol{z}}, \tilde{z}_*)$ is the most likely (logarithmic) noise level, and δ is the Dirac delta function with $\delta(\tilde{\boldsymbol{z}}, \tilde{z}_*) = 1$ when $\tilde{\boldsymbol{z}} = \tilde{z}_*$ and zero otherwise. Thus, the predictive posterior distribution over the output y_* is approximated as:

$$\begin{aligned} p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}_f, \boldsymbol{\theta}_z, D) &\approx \iint p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}_f, \boldsymbol{z}, z_*, D)\delta(\tilde{\boldsymbol{z}}, \tilde{z}_*)d\boldsymbol{z}dz_* \\ &\approx p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}_f, \tilde{\boldsymbol{z}}, \tilde{z}_*, D), \end{aligned} \quad (8)$$

where the most likely noise level is given by:

$$(\tilde{\boldsymbol{z}}, \tilde{z}_*) = \arg\max_{(\tilde{\boldsymbol{z}}, \tilde{z}_*)} p(\boldsymbol{z}, z_*|\boldsymbol{x}_*, \boldsymbol{\theta}_z, D). \quad (9)$$

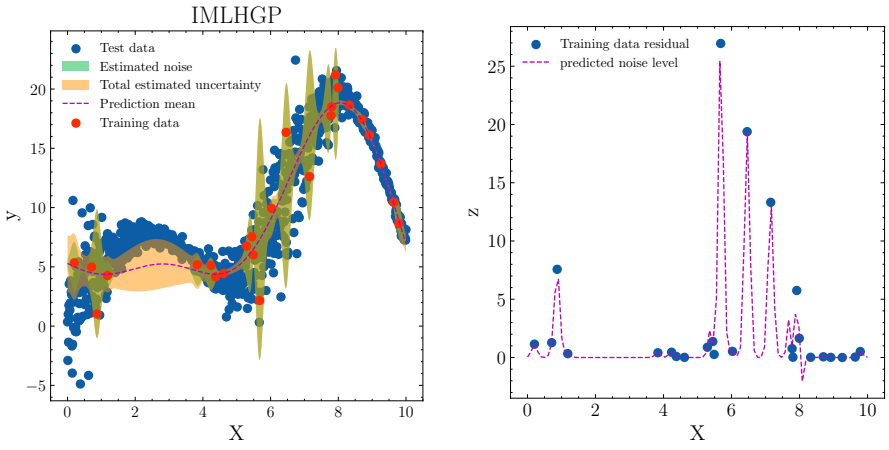
As the input-dependentnoise is also modelled by a GP regression, its most likely noise level is expressed as $(\tilde{\boldsymbol{z}}, \tilde{z}_*) = (\boldsymbol{\mu}_{\boldsymbol{z}}, \mu_{z_*})$. Thus, the posterior distribution over the output can be further expressed as:

$$p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}_f, \boldsymbol{\theta}_z, D) \approx p(y_*|\boldsymbol{x}_*, \boldsymbol{\theta}_f, \boldsymbol{\mu}_{\boldsymbol{z}}, \mu_{z_*}, D) \quad (10)$$

The main difference between MLHGP, IMLHGP, and NNPEHGP is that in MLHGP [22], an iterative approach similar to the expectation-maximization (EM) algorithm is used to obtain the noise estimation. IMLHGP [23] uses a correction factor to approximate the fina result, making the algorithm faster. Meanwhile the NNPEHGP [24] was developed to address the overfittin issue in IMLHGP by using k-nearest neighbour regression to transform the scattered Bayesian residual into a smoother function.

3 HETEROSCEDASTIC GAUSSIAN PROCESS WITH KERNEL SMOOTHING

The training procedure of the MLHGP involves an iterative process similar to the EM algorithm, making the computational complexity of MLHGP becomes prohibitively high. To address this problem, the IMLHGP [23] model was then developed. In some cases, IMLHGP even provides a better prediction of the noise and the full distribution compared to MLHGP under the condition of a large number of training data sizes. However, when only a few training data are available, we observe that IMLHGP tends to overfi the noise level as shown in Figure 2a. The overfittin phenomenonis reasonable when only a few observations are available. When the number of training samples is low, the noise residuals are likely scattered and it becomes difficul for a GP to model the noise level properly, as shown in Figure 2b. To avoid the overfittin problem, NNPEHGP [24] then proposes to utilize k-nearest neighbour regression to transform the scattered residual into a smoother function. However, this approach requires the correct definitio of the number of the neighbours k to be able to accurately model the heteroscedastic noise.



(a) (b)

Figure 2: (a) Noise level overfittin behaviour in IMLHGP, (b) Second GP overfittin in estimating the residual.

In contrast to NNPEHGP, our proposed approach aims to avoid overfittin in the noise level approximation and reduce the computational complexity of training multiple GPs to approximate the noise level by adopting the parameters of the trained main function GP. Since we use the squared exponential kernel, define in equation 4, this means we use the lengthscales l parameter from the trained GP directly without retraining for the noise approximation.

In practice, we relax the procedure of the typical heteroscedastic GP that uses two GPs by replacing the noise-GP with a kernel smoothing function [25–27] with a length scale identical to the firs Gaussian process (GP). The main motivation of this study comes from our observation

that the GP model assumes neighbouring data points are similar and correlated following the kernel function 4. As the second GP training is prone to overfitt ng in case of a low number of observations, we hypothesise that the length scale parameter from the firs GP is already a good approximation to describe how neighbouring data noise correlates to each other.

Given $\boldsymbol{l_1}$ as the length scale parameter from the trained firs GP, the Nadaraya-Watson estimator (kernel smoothing) function [27] is define as:

$$\hat{y}(\boldsymbol{x}_*; \boldsymbol{l_1}) = \sum_{i=1}^{n} W_i(\boldsymbol{x}_*) Y_i, \tag{11}$$

where $W_i(x)$ is the weighting function derived from the kernel function:

$$W_i(\boldsymbol{x}_*) := \frac{k_{l_1}(\boldsymbol{x}_*, \boldsymbol{x}_i)}{\sum_{j=1}^{n} k_{l_1}(\boldsymbol{x}_*, \boldsymbol{x}_j)}. \tag{12}$$

The subscript in k_{l_1} indicates that the squared exponential kernel in equation 4 is using $\boldsymbol{l_1}$ as its length scale parameter. The σ_f parameters in the kernel will be eventually cancelling each other in equation 12. Thus, we don't need to provide a specifi value. Since our proposed method is very fl xible, it can be used in the fashion of either MLHGP or IMLHGP. Thus, as a comparison we introduce the kernel smoothing variant of both cases named KS MLHGP and KSIMLHGP respectively.

4 EXPERIMENTAL RESULTS

To quantitatively assess the model predictive performance and uncertainty quantificatio (UQ) quality, we follow [22–24] to use several metrics. First, to assess the noise level prediction, we use the standardized mean square error (SMSE) between the true noise standard deviation and the noise level prediction by the GP model, written as

$$SMSE_{(g)} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \frac{(\tilde{g}_i - g_{*,i})^2}{\text{var}(g_*)}, \tag{13}$$

where n_{test} is the number of test points, $\tilde{g}_i$ is the predicted noise standard deviation, $g_{*,i}$ is the true noise standard deviation, and $\text{var}(g_*)$ is the variance of the true noise standard deviation at all test points.

The goal of heteroscedastic modelling is to predict the full distribution of a specifi point. Hence, to assess the GP model in predicting the full distributions, we follow [24, 28] using the expectation of the normalized 2-Wasserstein distance at all test points $x_{*,i}$, define as

$$\varepsilon = \mathbb{E}[d(Y(x_*), \hat{Y}(x_*)], \tag{14}$$

where d is the normalized 2-Wasserstein distance between the probability distribution over the true response $Y(x_*)$ and the prediction from GP model $\hat{Y}(x_*)$

$$d(Y_1, Y_2) = \frac{d_{WS}(Y_1, Y_2)}{\sigma(Y_1)}, \tag{15}$$

and the Wasserstein distance d_{WS} is written as

$$d_{WS}(Y_1, Y_2) := ||Q_1 - Q_2||_2 = \sqrt{\int_0^1 (Q_1(u) - Q_2(u))^2 du}. \tag{16}$$

Q_1 and Q_2 are the quantile function of Y_1 and Y_2 respectively.

Additionally, we also employ the average of the negative log probability density (NLPD) to quantify the predictive quality of the test outputs, since in practice, the true standard deviation in noise is sometimes unavailable. Such metric is evaluated by:

$$NLPD_{(y)} = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_{*,i}|x_{*,i}, D). \tag{17}$$

4.1 One-dimensional analytical problem

We consider a one-dimensional mathematical function with heteroscedastic noise, define as:

$$\begin{aligned} g(x) &= x\sin(x) + 4\sqrt{x}, \\ h(x) &= \exp(\cos(x)), \\ f(x) &\sim \mathcal{N}(g(x), (h(x))^2), \end{aligned} \tag{18}$$

with input variable x within $[0, 10]$. This one-dimensional function has a nonlinear mean function $g(x)$ and nonlinear noise level $h(x)$ that varies in the input domain x. We use the training dataset $n = 25$ for this problem, and the number of test points is set to 1000. One realization of different models is evaluated as shown in Figure 3.

Here, we observe that IMLHGP is overfittin on the noise level prediction indicated by the non-smoothnoise level prediction. Even NNPEHGP which is supposed to solve the overfittin problem from IMLHGP still looks overfitte at $X = [4, 10]$. We can also observe that using kernel smoothing approximation alleviates the overfittin problem in KSIMLHGP. Meanwhile, comparing MLHGP and KSMLHGP we observe that the prediction distribution looks very similar.

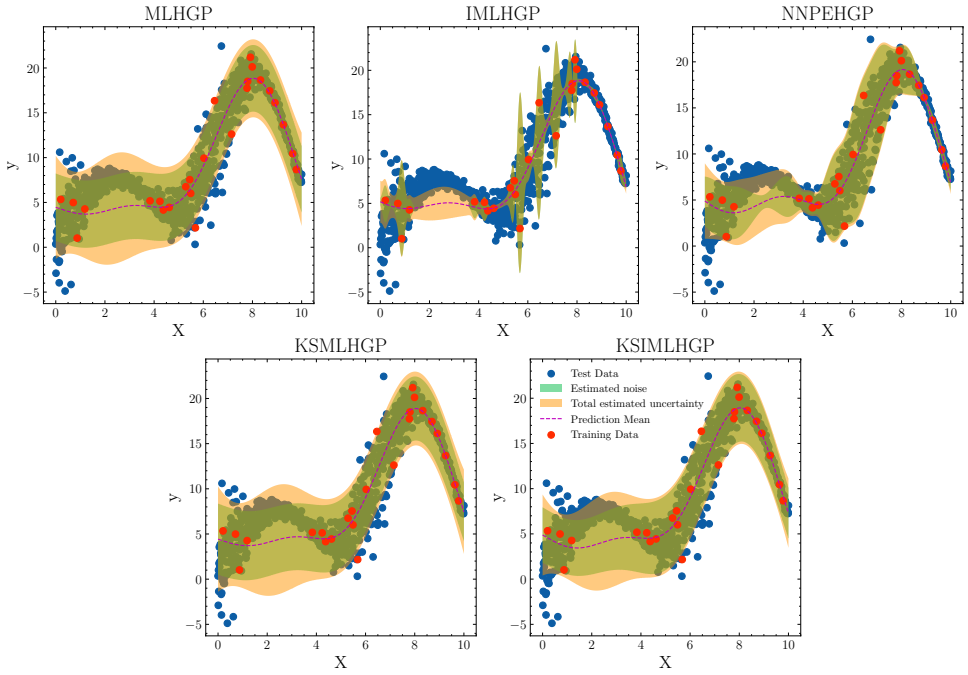
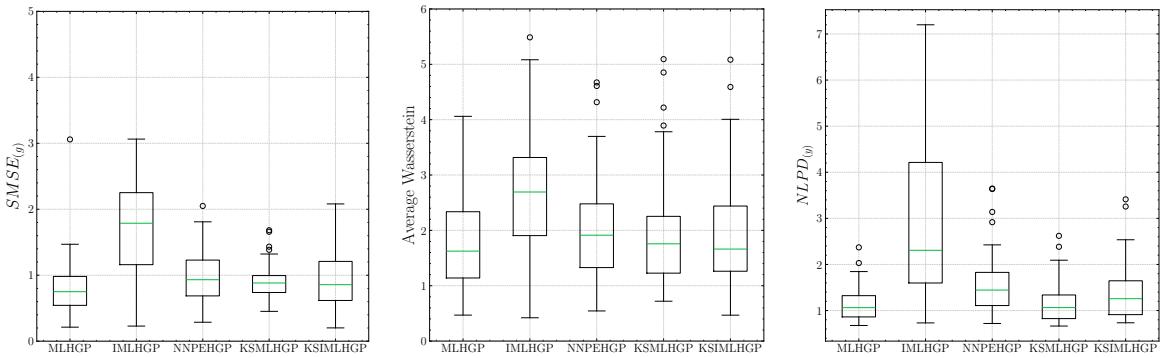


Figure 3: One realization of multiple model runs for the one-dimensional analytical case.

The performances of the models are evaluated on standardized mean square error (SMSE), averaged 2-Wasserstein distance (ε), and negative log probability density (NLPD) which are

shown in Figure 4a, 4b, and 4c respectively. In these evaluations, we only consider a small number of training datasets ($n = 25$) and repeat each model's training and evaluations with 50 different training datasets with different random seeds. The results show that in this problem IMLHGP has the worst performance compared to the other and employing the kernel smoothing for noise level prediction provides improvement in its predictive capability. Meanwhile, KSMLHGP holds a similar performance to the MLHGP while reducing the computational complexity.



(a) (b) (c)

Figure 4: (a) Standardized mean squared error, (b) Averaged 2-Wasserstein distance, and (c) Negative log probability density for one-dimensional analytical problem.

4.2 Four-dimensional computational simulation

To represent a more realistic engineering problem, we follow [24] to evaluate the 4-dimensional computational simulation of a solid sphere immersed in flui [13]. This heat transfer problem models the temperature of the sphere in a flui with a higher temperature as the initial condition, where the heat is transferred between the sphere and the flui via convection and via conduction inside the sphere. The temperature inside the sphere T is described as the solution of a partial differential equation. The solution is explicitly given by

$$T = u_2 + u_3 \sum_{i=1}^{\infty} \frac{4(\sin \eta_i - \eta_i \cos \eta_i)}{2\eta_i - \sin(2\eta_i)} \exp(-\eta_i^2 \pi_2) \frac{\sin(\eta_i u_1)}{\eta_i u_1}, \tag{19}$$

where η_i is the solution of:

$$\begin{aligned} 1 - \eta_i \cot \eta_i = \pi_1, \quad \eta_i \in ((i-1)\pi, i\pi) \\ \pi_1 = \frac{u_4 u_7}{u_6}, \quad \pi_2 = \frac{u_6 u_5}{u_8 u_9 u_7^2}. \end{aligned} \tag{20}$$

We defin the input variable as the distance from the sphere's centre u_1, the flui initial temperature u_2, the initial temperature of the sphere u_3, and the convective heat transfer coefficien u_4. Each variable is specifie in Table 1. The time variable u_5 is set to $800\ s$, and the other variables $u_6, \ldots, u_9$ are treated as random variables that are distributed normally, given in Table 2.

Although the PDE solution in equation 19 is explicit, the exact noise function is unknown. Typically in this situation, we evaluate the negative log probability density (NLPD). However, since the problem is computationally cheap, we estimate the output distribution using multiple realizations per test input. This allows us to use SMSE and 2-Wasserstein distance as metrics.

Table 1: Input variable ranges for the heat transfer problem.

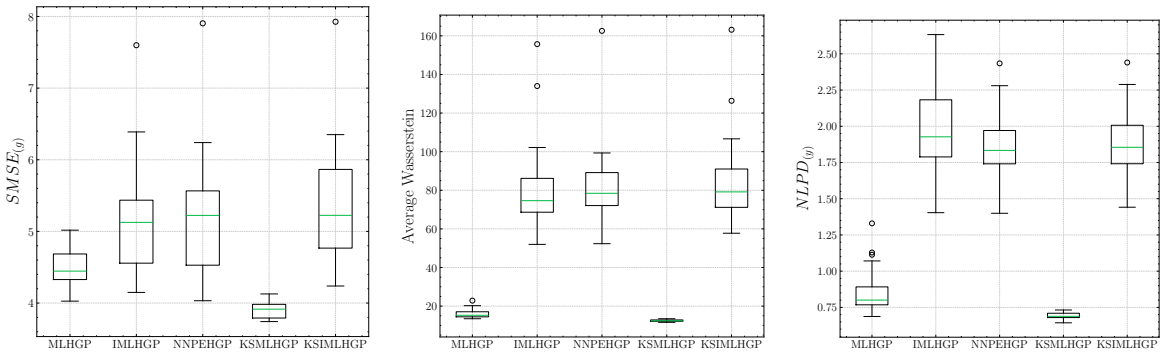
Variable	Description	Units	Minimum	Maximum
u_1	Distance from the center of sphere	-	0	1
u_2	Temperature of flui	K	250	270
u_3	Initial sphere temperature	K	-100	-30
u_4	Convective heat transfer coefficien	$kgs^{-3}K^{-1}$	180	210

Table 2: Fixed and random variable parameters for the heat transfer problem.

Variable	Description	Units	Value/Distribution
u_5	Time	s	800
u_6	Thermal conductivity	$kgs^{-3}K^{-1}$	$\mathcal{N}(65, 10.3^2)$
u_7	Sphere's radius	m	$\mathcal{N}(0.1, 0.02^2)$
u_8	Specifi heat	$m^2s^{-2}K^{-1}$	$\mathcal{N}(400, 33.3^2)$
u_9	Density	kgm^{-3}	$\mathcal{N}(8000, 333.3^2)$

Here, we use 10^3 test data points in the input spaces with 10^4 realizations at each point. The number of training samples is set to 160. In this problem, we repeat each experiment with a different random seed 25 times.

Figure 5 clearly shows that MLHGP-based methods perform better than IMLHGP-based methods on a fairly complex problem with noticeable differences. Here, we especially highlight the KSMLHGP method that outperforms the IMLHGP-based methods and is still relatively comparable to the original MLHGP while reducing the computational complexity from $\mathcal{O}(2\mathcal{N}^3)$ to $\mathcal{O}(\mathcal{N}^3 + \mathcal{N}^2)$.



(a) (b) (c)

Figure 5: (a) Standardized mean squared error, (b) Averaged 2-Wasserstein distance, and (c) Negative log probability density for the four-dimensional heat conduction simulation.

4.3 Risk-averse Bayesian optimisation on one-dimensional function

Bayesian optimisation (BO) [7] is a common application of GPs that aims to identify the global optimum (in this work, we consider minimization) of a black-box function, represented as $x^* \in \min_{x \in \mathcal{X}} f(x)$, where $\mathcal{X} \in \mathbb{R}^d$. Such functions are often expensive to evaluate, noisy, or both. The common approach in BO is an iterative process where a surrogate model, typically a GP, is firs trained to approximate the target function. Subsequently, an acquisition function (AF)

is optimized to determine the next query point for efficientl locating the function's minimum.

This section focuses on a scenario where the function has heteroscedastic noise. Rather than seeking a simple global minimumlike in a non-noisy BO setting, the goal is to identify a risk-averse optimum. To tackle the risk aversion problem, we follow the framework proposed by [29] and adopt the widely used mean-variance (MV) objective [30]. This objective define a trade-off between the mean return $f(x)$ and the risk represented by a variance proxy $\rho^2(x)$:

$$\mathrm{MV}(x) = f(x) + \alpha\rho^2(x), \tag{21}$$

where $\alpha > 0$ is the absolute risk tolerance parameter, assumed to be known to the learner. In this work, we focus on a minimization task, aiming to achieve a solution where $\mathrm{MV}(x)$ is as low as possible (i.e., a solution with a low function value and minimal uncertainty).

To evaluate our model in BO , we apply the Risk-Averse Heterogeneous Bayesian optimisation (RAHBO) AF [29, 31]. RAHBO extends the GP Upper Confidenc Bound (GP-UCB) [32] AF by incorporating input dependent noise prediction. It is define as:

$$\mathrm{UCB}(x) = \mu(x) + \beta\sigma(x), \qquad \mathrm{RAHBO}(x) = \mathrm{UCB}(x) + \alpha g^2(x). \tag{22}$$

We evaluate our approach using four metrics: the *risk-averse regret*, define as $\mathrm{Rt} = \mathrm{MV}(x^{**}) - \mathrm{MV}(x_t)$, where $x^{**} \in \arg\max_{x\in\mathcal{X}} \mathrm{MV}(x)$; the *cumulative risk-averse regret*, given by $\sum_{t=1}^{T} (\mathrm{MV}(x^{**}) - \mathrm{MV}(x_t))$; the *simple regret*, expressed as $f(x^*) - f(x_t)$; and the *cumulative simple regret*, computed as $\sum_{t=1}^{T} (f(x^*) - f(x_t))$. The cumulative metrics are useful for online settings, while the simple regret is used to compare the method against non-risk-averse baseline.

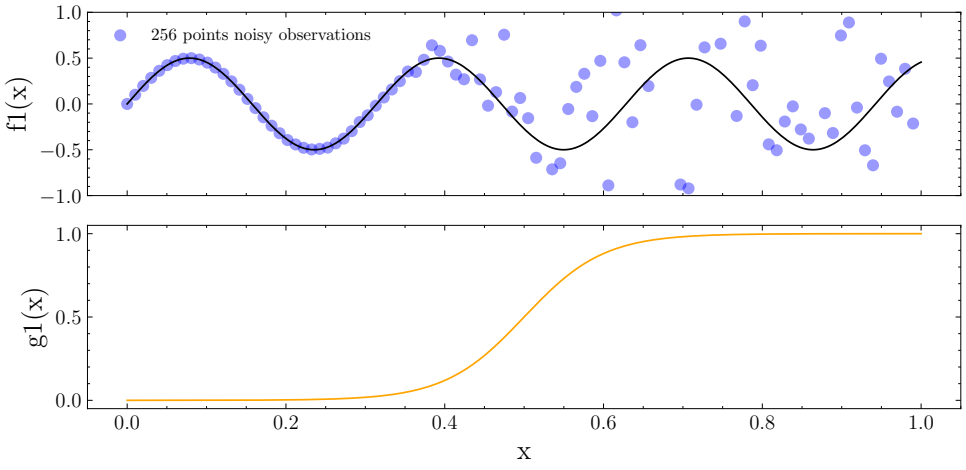


Figure 6: Synthetic heteroscedastic Bayesian optimisation function with 256 noisy evaluation points.

We consider a one-dimensional noisy optimisation problem, illustrated in Figure 6. The problem features three global (mean) minima, each subject to varying levels of noise. The function is define as follows:

$$f(x) = \mathrm{f1}(x) + \epsilon_1, \qquad \epsilon_1 \sim \mathcal{N}(0, \mathrm{g1}(x)), \tag{23}$$

where $\mathrm{f1}$ and $\mathrm{g1}$ are define as:

$$\mathrm{f1}(x) = 0.5\sin(20x), \qquad \mathrm{g1}(x) = \frac{1}{1 + e^{-(20x-10)}}. \tag{24}$$

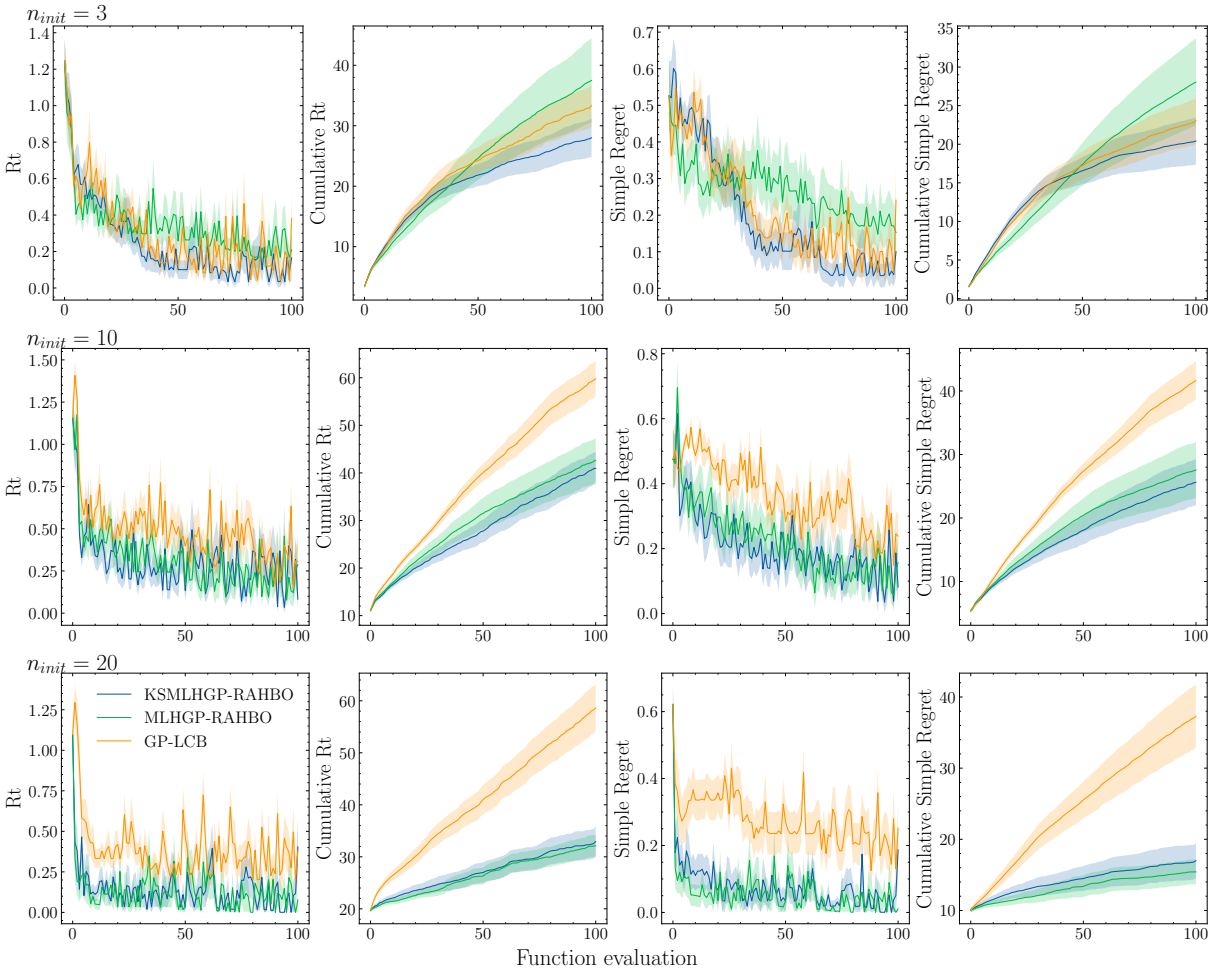


Figure 7: Risk-averse heteroscedastic Bayesian optimisation result, from top to bottom, initial sample $n_{init} = 3, 10, 20$.

We conduct experiments to test the performance of the KSMLHGP with the RAHBO AF on the above 1D function problem. We compare it with two baselines: first the MLHGP with the RAHBO AF, and second, a non-risk-averse baseline—a GP model with the lower confidenc bound (LCB) AF. Notably, we use LCB instead of the UCB [32] in this work, as our problem involves minimisation instead of maximisation. We assume a fi ed, known $\alpha = 1$ for both $\mathrm{MV}(x)$ in the evaluation metrics calculation and the RAHBO AF. The β parameters for RAHBO and LCB are set to 0.2, following the default setting of the popular BO package BoTorch [33][1]. We optimize the AF using random search with 10K candidates[2]. We ran the experiments with 100 BO iterations starting from initial observations and repeated 10 times with different initial observation configuration for consistency. We use different numbers of initial observations—3, 10, and 20—to examine how the models behave under varying numbers of initial observations.

The results of the experiments are presented in Figure 7. We observe that KSMLHGP-RAHBO outperforms MLHGP-RAHBO in the low initial observation setting ($n_{init} = 3$) while remaining competitive in the mid and high initial observation settings ($n_{init} = 10, 20$, respectively). This demonstrates that KSMLHGP is more robust when trained on a small number of

[1] https://botorch.org/

[2] For higher-dimensional problems, a more sophisticated optimizer, such as a multi-start gradient-based or evolutionary algorithm, is preferable due to the curse of dimensionality.

initial observations compared to MLHGP. Moreover, in the cumulative regret metrics, risk-averse methods exhibit significan performance gains compared to non-risk-averse ones. While this result is expected, it highlights that adopting a risk-averse approach might be highly beneficia in online settings, particularly in safety-critical applications.

Additionally, we tested the average runtime for each model (independent of n_{init}) with $n_{\text{iter}} = 100$. The results are as follows: GP-LCB: 75 seconds, KSMLHGP-RAHBO: 372 seconds, MLHGP-RAHBO: 665 seconds. All experiments were conducted on a MacBook with an M1 Pro chip. While our method is still slower compared to the non-risk-averse method, it shows a significan speed improvement over the state-of-the-art MLHGP while maintaining good performance and even outperforming it in low initial observation settings. The code that we used for the experiments is available on GitHub [3].

5 CONCLUSION

In this paper, we introduce kernel smoothing as a viable alternative to the second Gaussian process (GP) for modelling the noise level in heteroscedastic data. In this approach, we set the kernel length scale of the kernel smoothing method to be identical to the firs GP model that approximates the mean function of the data by assuming that neighbouring data points should have similar noise levels. This approach presents an enhancement over the IMLHGP by enhancing its resilience to overfitti g in scenarios where the observed data is limited. Furthermore, it also improves upon the MLHGP by decreasing the computational complexity, as it only necessitates training the firs GP, while still maintaining comparable predictive performance to the standard MLHGP method.

We conduct a comparative analysis of our proposed kernel smoothing methods, KSMLHGP and KSIMLHGP, against established post-modeling HGP methods, including MLHGP, IMLHGP, and NNPEHGP, to validate the feasibility and effectiveness of our approaches. In our experiments, we utilize three metrics - the noise level SMSE, averaged 2-Wasserstein distance, and NLPD - to quantitatively evaluate the predictive distribution of the heteroscedastic data. The results of the experiments demonstrate that our proposed methods outperform IMLHGP and NNPEHGP in scenarios with limited data and maintain strong predictive performance compared to MLHGP. Moreover, in the simple heteroscedastic Bayesian optimisation test case we show that our proposed model outperforms MLHGP in small initial sample settings and remains competitive in larger initial samples while reducing the computational time.

While the heteroscedastic GP methods in general are quite powerful for modelling regression problems with input-dependent noise, the predictive interval from GP relies on the Gaussian assumption and the well-specificatio of the priors that are not always appropriate. Furthermore, in the absence of prior information, it might be difficul to fully specify the exact parameter of the GP model. Therefore it would be desirable in the future to conduct research related to conformalized-GP [34, 35] as well as imprecision in GP [36].

ACKNOWLEDGEMENTS

This work was supported by granted H2020 FETOPEN-2018-2019-2020-01 European project, *Epistemic AI* under grant agreement No. 964505 (E-pi).

[3] https://github.com/fazaghifari/hetgp-eval

REFERENCES

[1] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

[2] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning series. MIT Press, London, England, November 2005.

[3] Rohit K Tripathy and Ilias Bilionis. Deep uq: Learning deep neural network surrogate models for high dimensional uncertainty quantificatio . *Journal of computational physics*, 375:565–588, 2018.

[4] Gang Sun and Shuyue Wang. A review of the artificia neural network surrogate modeling in aerodynamic design. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 233(16):5863–5872, 2019.

[5] Donald R Jones, Matthias Schonlau, and William J Welch. Efficien global optimization of expensive black-box functions. *Journal of Global optimization*, 13:455–492, 1998.

[6] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

[7] Roman Garnett. *Bayesian optimization*. Cambridge University Press, 2023.

[8] Timothy M.S. Jim, Ghifari A. Faza, Pramudita S. Palar, and Koji Shimoyama. A multiobjective surrogate-assisted optimisation and exploration of low-boom supersonic transport planforms. *Aerospace Science and Technology*, 128:107747, September 2022.

[9] Lavi Rizki Zuhal, Ghifari Adam Faza, Pramudita Satria Palar, and Rhea Patricia Liem. Performance assessment of kriging with partial least squares for high-dimensional uncertainty and sensitivity analysis. *Structural and Multidisciplinary Optimization*, 66(5), April 2023.

[10] Benjamin Echard, Nicolas Gayton, and Maurice Lemaire. Ak-mcs: an active learning reliability method combining kriging and monte carlo simulation. *Structural safety*, 33(2):145–154, 2011.

[11] Lavi Rizki Zuhal, Ghifari A. Faza, Pramudita S. Palar, and Rhea P. Liem. Fast and adaptive reliability analysis via kriging and partial least squares. In *AIAA Scitech 2021 Forum*. American Institute of Aeronautics and Astronautics, January 2021.

[12] B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 47(1):1–21, September 1985.

[13] Matthias H. Y. Tan. Monotonic quantile regression with bernstein polynomials for stochastic simulation. *Technometrics*, 58(2):180–190, April 2016.

[14] Haichen Shi, Keith Worden, and Elizabeth J. Cross. A cointegration approach for heteroscedastic data based on a time series decomposition: An application to structural health monitoring. *Mechanical Systems and Signal Processing*, 120:16–31, April 2019.

[15] Paul Goldberg, Christopher Williams, and Christopher Bishop. Regression with input-dependent noise: A gaussian process treatment. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997.

[16] Luis Muñoz-González, Miguel Lázaro-Gredilla, and Aníbal R. Figueiras-Vidal. Heteroscedastic gaussian process regression using expectation propagation. In *2011 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2011.

[17] Ville Tolvanen, Pasi Jylänki, and Aki Vehtari. Expectation propagation for nonstationary heteroscedastic gaussian process regression. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2014.

[18] Marcelo Hartmannand Jarno Vanhatalo. Laplace approximation and natural gradient for gaussian process regression with heteroscedastic student-t model. *Statistics and Computing*, 29(4):753–773, October 2018.

[19] Miguel Lazaro-Gredilla and Michalis K. Titsias. Variational heteroscedastic gaussian process regression. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 841–848, Madison, WI, USA, 2011. Omnipress.

[20] Marianne Menictas and Matt P. Wand. Variational inference for heteroscedastic semiparametric regression. *Australian and New Zealand Journal of Statistics*, 57(1):119–138, March 2015.

[21] Haitao Liu, Yew-Soon Ong, and Jianfei Cai. Large-scale heteroscedastic regression via gaussian process, 2020.

[22] Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 393–400, New York, NY, USA, 2007. Association for Computing Machinery.

[23] Qiu-Hu Zhang and Yi-Qing Ni. Improved most likely heteroscedastic gaussian process regression via bayesian residual moment estimator. *IEEE Transactions on Signal Processing*, 68:3450–3460, 2020.

[24] Muhammad D. Robani, Pramudita S. Palar, and Lavi Rizki Zuhal. Heteroscedastic gaussian process regression using nearest neighbor point estimates. In *AIAA Scitech 2021 Forum*. American Institute of Aeronautics and Astronautics, January 2021.

[25] E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142, January 1964.

[26] Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, 1964.

[27] Herman J Bierens. The nadaraya–watson kernel regression function estimator. In *Topics in advanced econometrics*, chapter 10, page 212–247. Cambridge University Press, Cambridge, UK, 1994.

[28] Xujia Zhu and Bruno Sudret. Emulation of stochastic simulators using generalized lambda models. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1345–1380, January 2021.

[29] Anastasia Makarova, Ilnura Usmanova, Ilija Bogunovic, and Andreas Krause. Risk-averse heteroscedastic bayesian optimization. *Advances in Neural Information Processing Systems*, 34:17235–17245, 2021.

[30] Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. *Advances in neural information processing systems*, 25, 2012.

[31] Marshal Arijona Sinaga, Julien Martinelli, Vikas Garg, and Samuel Kaski. Heteroscedastic preferential bayesian optimization with informative noise distributions, 2024.

[32] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1015–1022. Omnipress, 2010.

[33] Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficien Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020.

[34] Harris Papadopoulos. Guaranteed coverage prediction intervals with gaussian process regression, 2023.

[35] Edgar Jaber, Vincent Blot, Nicolas Brunel, Vincent Chabridon, Emmanuel Remy, Bertrand Iooss, Didier Lucor, Mathilde Mougeot, and Alessandro Leite. Conformal approach to gaussian process surrogate evaluation with coverage guarantees, 2024.

[36] Francesca Mangili. A prior near-ignorance gaussian process model for nonparametric regression. *International Journal of Approximate Reasoning*, 78:153–171, November 2016.