physical sciences forum

MDPI

Proceeding Paper

Global Variance as a Utility Function in Bayesian Optimization †

Roland Preuss * and Udo von Toussaint

Max–Planck–Institut für Plasmaphysik, 85748 Garching, Germany; udt@ipp.mpg.de

* Correspondence: preuss@ipp.mpg.de

† Presented at the 40th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, online, 4–9 July 2021.

Abstract: A Gaussian-process surrogate model based on already acquired data is employed to approximate an unknown target surface. In order to optimally locate the next function evaluations in parameter space a whole variety of utility functions are at one's disposal. However, good choice of a specific utility or a certain combination of them prepares the fastest way to determine a best surrogate surface or its extremum for lowest amount of additional data possible. In this paper, we propose to consider the global (integrated) variance as an utility function, i.e., to integrate the variance of the surrogate over a finite volume in parameter space. It turns out that this utility not only complements the tool set for fine tuning investigations in a region of interest but expedites the optimization procedure in toto.

Keywords: global optimization; Bayesian optimization; utility function; global variance

PACS: 02.50.-r; 52.65.-y

Citation: Preuss, R.; von Toussaint, U. Global Variance as a Utility Function in Bayesian Optimization. *Phys. Sci. Forum* **2021**, *3*, 3. https://doi.org/10.3390/psf2021003003

Academic Editor: Sascha Ranftl

Published: 5 November 2021

1. Introduction

In many experimental or theoretical approaches the effort of acquiring data may be costly, time consuming or both. The goal is to get insights in either the overall or extremal behaviour of a target quantity with respect to a set of parameters. If insights to functional dependencies between target and parameters are only to be obtained from a computationally expensive function, which may be considered as a black box function, it is instructive to employ surrogate modelling: already acquired data serve as a starting basis for establishing a surrogate surface in parameter space which then gets explored by Bayesian optimization [1]. An overall survey about Bayesian optimization may be found in [2], though it concentrates on an expected improvement (EI) utility and considers noise in the data only in the last paragraph, again by concentrating on EI. In contrast to this nice study we propose to alternate the different utilities at hand. Moreover, a fast information theory related to Bayesian optimization is shown in [3], though this approach approximates any black-box function by a parabolic form which differs from our ansatz letting the black-box function untouched. Interesting insights to multi-objective Bayesian optimization are provided by [4], which considers "multi-objective" in the sense of seeking the extrema–each is free of choice maximum or minimum–of a bunch of single-objective functions. However, the present paper concentrates on finding a common extremum depending on multiple parameters.

For the surrogate modelling we use the Gaussian process method (GP) [5] whose early stages date back to the middle of last century with very first efforts in geosciences [6] tackling the problem of surrogate modelling by so-called kriging [7]. Afterwards, GP has been appreciated much in the fields of neural networks and machine learning [8–12] and further work showed the applicability of active data selection via variance based criterions [13,14]. Our implementation of the GP method in this paper was already introduced at [15], and follows in notation–and apart from small amendments—the very instructive book of Rasmussen & Williams [5].

While in a previous work [16] we investigated the performance of utility functions for the expected improvement of an additional data point or for a data point with the maximal variance, in this paper we would like to introduce the global variance, i.e., the integral over the variance for a target surrogate within a region of interest with respect to a newly added data point. It is the substantial advantage of the Gaussian process method that such a task may be tackled simply on the basis of already acquired data, i.e., before new data have to be determined.

2. Global Variance for Gaussian Process-Based Model

In the following we concisely report the formulas leading to the results in this paper. For a thorough discussion of Gaussian processes please refer to the above mentioned papers, especially to the book of Rasmussen & Williams [5].

The problem of predicting function values in a multi-dimensional space supported by given data is a regression problem for a non-trivial function of unknown shape. Given are n target data $\boldsymbol{y}$ for input data vectors $\boldsymbol{x}_i$ of dimension N_{dim} with matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$ written as

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iN_{\mathrm{dim}}} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N_{\mathrm{dim}}} & x_{2N_{\mathrm{dim}}} & \dots & x_{nN_{\mathrm{dim}}} \end{pmatrix}. \tag{1}$$

We assume that target data y_i are blurred by Gaussian noise with $\sigma_{d_i}^2$. Further, we assume that the black box function interconnecting input $\boldsymbol{X}$ and target $\boldsymbol{y}$ is at least uniformly continuous and thereby justifies a description of a target surface with a surrogate from the Gaussian process method. Despite the experimental data and the physics background all quantities throughout this paper are without units.

The decisive quantity of a Gaussian process is the covariance function k describing the distance between two vectors $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$ defined by

$$k(\boldsymbol{x}_p, \boldsymbol{x}_q) = \sigma_f^2 \exp\left\{ -\frac{1}{2} \left| \frac{\boldsymbol{x}_p - \boldsymbol{x}_q}{\lambda} \right|^2 \right\}. \tag{2}$$

with the signal variance σ_f^2 and length scale λ. A covariance matrix $(\boldsymbol{K})_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ considers the covariances of all input data $\boldsymbol{X}$. The GP method describes a target value f_* at test input vector $\boldsymbol{x}_*$ by a normal distribution $p(f_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) \propto \mathcal{N}(\bar{f}_*, \mathrm{var}(x_*))$ with mean $\bar{f}_* = \boldsymbol{k}_*^T \left(\boldsymbol{K} + \sigma_n^2 \Delta\right)^{-1} \boldsymbol{y}$, and variance $\mathrm{var}(\boldsymbol{x}_*) = \boldsymbol{k}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T \left(\boldsymbol{K} + \sigma_n^2 \Delta\right)^{-1} \boldsymbol{k}_*$, where the term $\sigma_n^2 \Delta$ represents the degree of information in the data. While Δ is the diagonal matrix of the given variances $\sigma_{d_i}^2$, the variance σ_n^2 accounts for an overall noise in the data. Then the full covariance matrix $\boldsymbol{M}$ of the Gaussian Process is

$$\boldsymbol{M} = \boldsymbol{K} + \sigma_n^2 \Delta \tag{3}$$

In Bayesian probability theory the three parameters $\boldsymbol{\theta} = (\lambda, \sigma_f, \sigma_n)^T$ are considered to be hyper-parameters which show up in the marginal likelihood as

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}) = \mathrm{const} - \frac{1}{2} \boldsymbol{y}^T \left[\boldsymbol{K}(\lambda, \sigma_f) + \sigma_n^2 \Delta \right]^{-1} \boldsymbol{y} - \frac{1}{2} \log \left| \boldsymbol{K}(\lambda, \sigma_f) + \sigma_n^2 \Delta \right|. \tag{4}$$

In [16], we showed, that for a sufficiently large data base the target surrogate is well described by using the expectation values of the hyper-parameters in the formulas for f_* and $\mathrm{var}(f_*)$, at least well enough to determine a global optimum in a region of interest (RoI). The global optimum is found by employing utility functions, as there are the expected improvement $U_{\mathrm{EI}}(\boldsymbol{x}_*) = \langle I \rangle = \int_{f_{\max}}^{\infty} f_* p(f_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) \mathrm{d}f_*$ and the maximum variance

$U_{\mathrm{MV}}(\boldsymbol{x}_*) = \mathrm{var}(f_*)$. For both the respective maximum at $\boldsymbol{x}_{*\max} = \arg\max_{\{\boldsymbol{x}_*\}} U_{\mathrm{EI/MV}}$ is sought. While the first one (U_{EI}) evaluates the possible information gain from a new data point, the second utility (U_{MV}) simply estimates the vector in input space with largest variance in the target surrogate.

In order to have a look on the implications of an additional data point in the surrogate, we propose a further utility function, i.e., the global variance defined on the multi-dimensional RoI$\in [-\mathbf{1} : \mathbf{1}]$ by

$$U_{\mathrm{GV}} = \int_{-\mathbf{1}}^{\mathbf{1}} \mathrm{var}(\boldsymbol{x}) \mathbf{d}\boldsymbol{x} \,. \tag{5}$$

The exact integration shown in the Appendix A leads to

$$\begin{aligned} U_{\mathrm{GV}}^{\mathrm{exact}} &= 2^{N_{\mathrm{dim}}}\sigma_f^2 - \sigma_f^4 \left(\frac{\sqrt{\pi}\lambda}{2}\right)^{N_{\mathrm{dim}}} \sum_{ij}^{n} \left(\boldsymbol{M}^{-1}\right)_{ij} \\ &\quad \cdot \prod_{k}^{N_{\mathrm{dim}}} \left\{ \mathrm{erf}\left[\frac{1}{\lambda} - \frac{x_{ik}+x_{jk}}{2\lambda}\right] - \mathrm{erf}\left[-\frac{1}{\lambda} - \frac{x_{ik}+x_{jk}}{2\lambda}\right] \right\} \end{aligned} \tag{6}$$

Though Equation (6) represents the correct result, it may turn out in computation runs that the determination of the error-function is substantially time consuming compared to the total expenditure. Therefore, we would like to introduce two alternatives to the exact integration in Equation (5).

The first one is kind of an approximation. Since outside of the RoI the integrand in Equation (5) shows only trivial contributions we shift the upper and lower integral bounds to $\pm$ infinity and get from the simple Gaussian integrals

$$U_{\mathrm{GV}}^{\mathrm{inf}} \approx -\sigma_f^4 \left(\sqrt{\pi}\lambda\right)^{N_{\mathrm{dim}}} \sum_{ij}^{n} \left(\boldsymbol{M}^{-1}\right)_{ij} \exp\left\{ -\frac{1}{4\lambda^2} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^T \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) \right\} . \tag{7}$$

We dropped the first term in the integral over $[\sigma_f^2 \boldsymbol{x}]_{-\infty}^{\infty}$ for being infinity, since at least it is a constant contribution regardless of changes in the input $\boldsymbol{X}$. Although the utility U_{GVinf} in Equation (7) is an approximation only, it has the advantage of being much easier accessible by numerical means and its computation performs much faster compared to Equation (6).

A second much more sophisticated approach is to insert an enveloping Gaussian function with adjustable location $\boldsymbol{x}_G$ (guiding center) and variance σ_G in the integral of Equation (5). Again the integration limits are shifted to $\pm$ infinity, however this time the enveloping Gaussian function takes care of the integrability and we get

$$\begin{aligned} U_{\mathrm{GV}}^{\mathrm{env}} &= \int_{-\infty}^{\infty} \mathrm{var}(\boldsymbol{x}) \left(\frac{1}{\sqrt{2\pi\sigma_G^2}}\right)^{N_{\mathrm{dim}}} \exp\left[\frac{1}{2\sigma_G^2} (\boldsymbol{x} - \boldsymbol{x}_G)^T (\boldsymbol{x} - \boldsymbol{x}_G)\right] \mathbf{d}\boldsymbol{x} \\ &= \sigma_f^2 - \sigma_f^4 \left(\frac{\lambda}{\sqrt{2\sigma_G^2 + \lambda^2}}\right)^{N_{\mathrm{dim}}} \sum_{ij}^{n} \left(\boldsymbol{M}^{-1}\right)_{ij} \\ &\quad \cdot \exp\left\{ -\frac{1}{2} \left[\frac{\boldsymbol{x}_i^T \boldsymbol{x}_i + \boldsymbol{x}_j^T \boldsymbol{x}_j}{\lambda^2} + \frac{\boldsymbol{x}_G^T \boldsymbol{x}_G}{\sigma_G^2} - \frac{\left(\frac{\boldsymbol{x}_i + \boldsymbol{x}_j}{\lambda^2} + \frac{\boldsymbol{x}_G}{\sigma_G^2}\right)^T \left(\frac{\boldsymbol{x}_i + \boldsymbol{x}_j}{\lambda^2} + \frac{\boldsymbol{x}_G}{\sigma_G^2}\right)}{\frac{2}{\lambda^2} + \frac{1}{\sigma_G^2}} \right] \right\} . \end{aligned} \tag{8}$$

The two parameters $\boldsymbol{x}_G$ and σ_G provide a toolset to guide the search for the next target data evaluations: a smaller standard deviation σ_G shifts the attention to the center of

the enveloping Gaussian, while $\boldsymbol{x}_G$ gives us the possibility to focus on certain regions in the RoI.

All three utilities employing the global variance, $U_{\mathrm{GV}}^{\mathrm{exact}}$, $U_{\mathrm{GV}}^{\mathrm{inf}}$, $U_{\mathrm{GV}}^{\mathrm{env}}$ require the inversion of the full covariance matrix $\boldsymbol{M}$ of Equation (3). Since the inversion has to be performed for every newly proposed test input $\boldsymbol{x}_*$ this is the main time consuming part in the whole procedure. Let us remind the reader, that the method we are proposing fully resides on input space (together with already acquired data) and the bottle neck is the generation of new target data. Therefore, the starting condition of very expensive (aka time consuming) data acquisition still holds. However, we can beneficially use blockwise matrix inversion [17] since a new input vector $\boldsymbol{x}_{n+1}$ expands the covariance matrix for one additional row and line only. Consequently, we reduced the computational effort to n^2-behaviour instead of n^3 for standard inversion.

3. Proof of Principle

We follow the global optimization scheme from Section 4 of [18]. Again we give proof of principle with a "black box" model featuring a broad parabolic maximum $2-\sum_k^{N_{dim}} x_k^2 + (-1)^k 0.3$ together with a smaller cosine structure $0.1\cos[2\pi(x_k-0.3)/\Delta_{cos}]$ on top of it, while we focus on a decent ripple on Δ_{cos}=0.6 in one and two dimensions (N_{dim}=1, 2).

Figures 1–3 show in left and right panels the results for one and two dimensions, respectively. The x axis to the right counts the number of newly acquired data for the utility comparisons in Figure 3 and in the bottom rows (**c**), (**d**) of Figures 1 and 2.

For every newly added data point proposed by the various utilities, the distance between the true location of the global optimum and the maximal value of the surrogate residing on the data at hand is calculated in Figure 1. In a similar fashion, the search for the best surrogate description of the hidden model is shown in Figure 2.

Eventually Figure 3 demonstrates the use of an enveloping Gaussian function in the integral of the global variance by varying its center x_G, e.g., if an educated guess about the location of the extremal structure is at hand, i.e., the guiding center x_G is preset to the positive axis (1d) or the quadrant (2d) with the true model maximum. Consequently, Table 1 displays for 1d the specific number of data and for 2d the saturation level for which the target surrogate enters the stage of resembling the true model, i.e. the summed up (absolute) differences between all grid points of the target surrogate and the model starts to diminish with the number of target data only.

Table 1. Comparison of enveloping Gaussian utilities with different integral weights and guiding centers in finding the best surrogate. **1d**: Changing step to solution. **2d**: Saturation level of the solution.

Integral Weight of env. Gaussian within RoI	**0.6**	**0.8**	**0.95**
1d: Corresponding width of env. Gaussian σ_G	1.38	0.95	0.71
U_{GV} with $x_G = 0$	15	14	22
U_{GV} with $x_G = 0.5$	13	12	13
2d: Corresponding width of env. Gaussian σ_G	0.99	0.79	0.67
U_{GV} with $\boldsymbol{x}_G$=(0;0)	0.23	0.61	0.21
U_{GV} with $\boldsymbol{x}_G$=(0.5;-0.5)	0.06	0.14	0.05

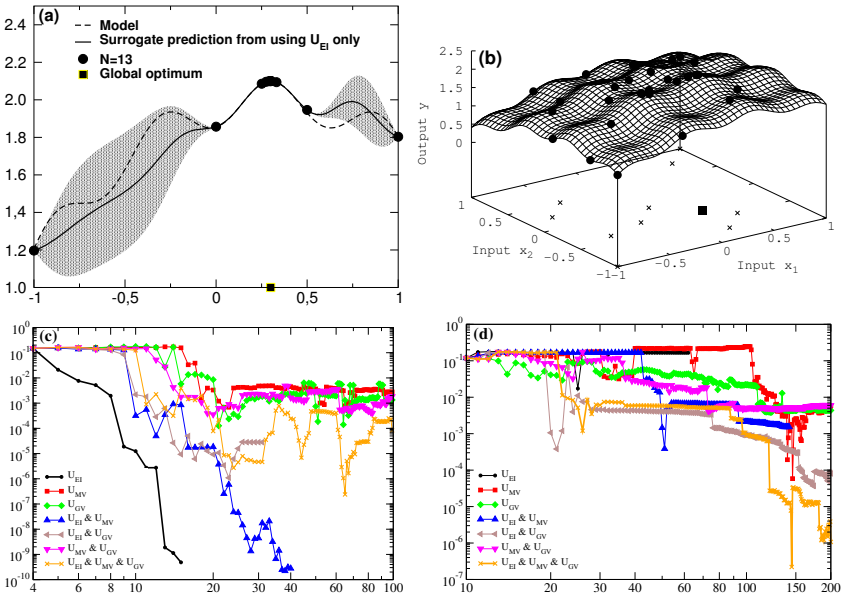

Figure 1. (**a**,**b**): One- and two-dimensional model with target data (full circles). The square in the bottom line/surface represents the true maximum. On the left the gray shaded area represents the uncertainty region of the surrogate (full line) from using the expected improvement utility only. On the right the points in the bottom surface are input data. Full circles represent additional data proposed by combination of all three utilities. (**c**,**d**): Distance surrogate/true maximum for different utilities employed in the global optimization procedure.

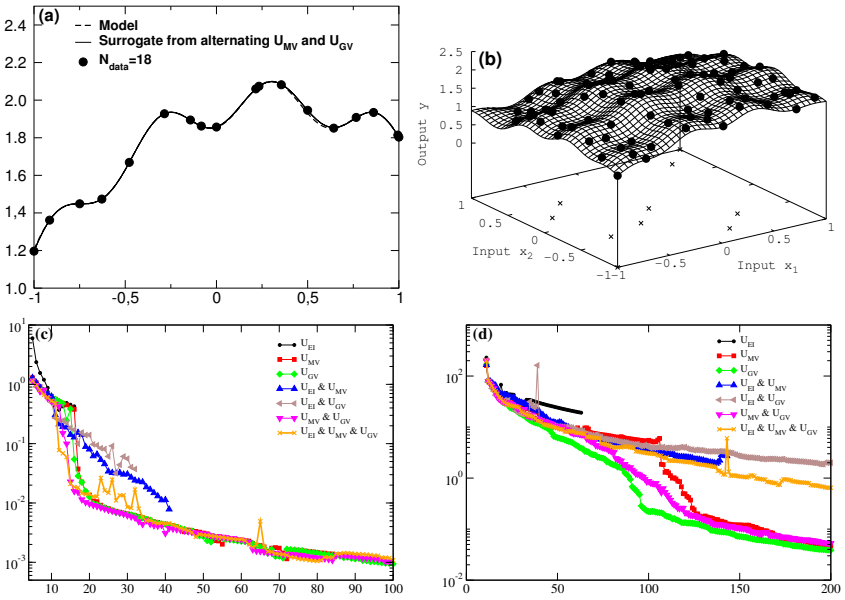

Figure 2. (**a**,**b**): One- and two-dimensional surrogate with newly acquired data. Surrogate solution (full line) on the left from using combination of U_{MV} and U_{GV}. Surrogate surface on the right from employing U_{GV} only. (**c**,**d**): Comparison of the differences between surrogate and true model integrated over RoI for various utilities.

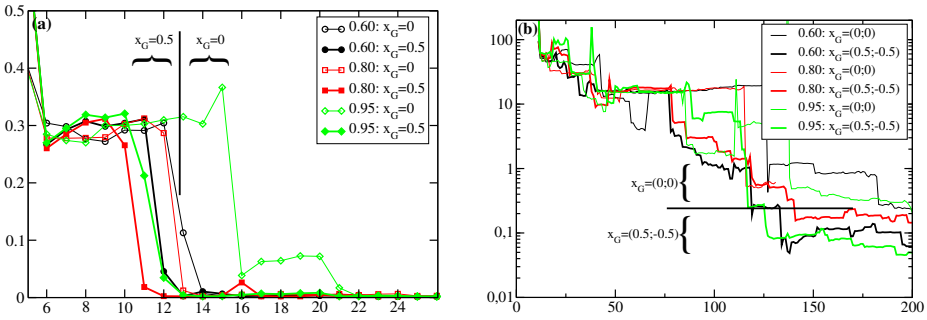

Figure 3. Summation over difference between grid points of target surrogate and true model as function of additionally acquired data. (**a**,**b**): One and two-dimensional case for enveloping Gaussian utility with various weights and guiding center at origin and at (0.5) or (0.5; −0.5).

4. Discussion

The results above show the usage of various utilities as a toolbox for surrogate modeling. Depending on the task—either to find an extremum or to get a best surrogate description of an unknown "black box" model—and depending on the prior knowledge at hand—presumption of location of the sought extremum or concentration on the region of interest—it is advisable to choose the most eligible utility function. However, even more promising is the combination of utilities of different character to profit from their benefits in toto and to compensate for pitfalls and drawbacks of one or the other utility.

As can be seen in the very first example in Figure 1 for the one-dimensional case the global optimum is found very fast with help of the expected improvement utility U_{EI} (starting to enter the bump with the correct extremum already below $N = 10$). However, a known drawback of this utility is that it gets stuck in local extrema and that it takes an unreasonably high number of additional data to get distracted from this pitfall.

This is taken into account for the two-dimensional case where the best result with lowest difference to the exact result is obtained by acting in combination of all three utilities U_{EI}, U_{MV} and U_{GV}. Focusing the maximum search on the utility regarding expected improvement alone (black line in Figure 1d would have got stuck in a local extremum with $y = 2.03$ in the "wrong" quadrant at $(-0.26; -0.29)$ for not recovering from this at all at about N = 63 (internal stop of algorithm for no improvement after entering computing accuracy level) and totally missing the true optimum with $y = 2.2$ at $(0.3; -0.3)$.

The situation changes for the task of getting a best overall description within the region of interest. To accomplish this the newly introduced global variance utility U_{GV} is of tremendous help both in one and two dimensions—either alone or in combination with at least the maximum variance utility U_{MV}. As shown in Figure 2d the best surrogate can already be established around ninety data points by employing U_{GV} only (full circles in the target surface of Figure 2b, with very few deviations from the true model left.

A guess about the approximate occurrence of an extremal structure–without excluding another region—can be emphasized by a further refinement to the global variance utility. In letting act an enveloping Gaussian within the global variance integral Equation (5) the result is not only much easier to be tackled from a computational point of view, but also the focus of the numerical search for the global optimum can be guided by predetermining the center of Gaussian x_G and its integral weight (aka width σ_{x_G}).

Figure 3 shows the results for three different integral weights (0.6; 0.8; 0.9) of the enveloping Gaussian function at two guiding centers: the first one at the origin corresponds to an ignorant scenario where one is not sure about a certain position of some global optimum at all. In the second approach we suppose that the extremal structure may be found in one dimension for positive values and thereby set $x_G = 0.5$, while in two dimensions it may be located within the quadrant with positive values for x_1 and negative ones for x_2 resulting in $x_G = [0.5; -0.5]$. As can be seen already in Figure 3, but all the more learned from the numbers of Table 1, displacing the center of the enveloping Gaussian

function to the real center of the optimum of the hidden model facilitates the development of a best—regarding similarity to the true model—surrogate surface.

Author Contributions: The authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Global Variance: Derivation of the Exact Integration

The variance at some (test) point $\boldsymbol{x}^T = (x_1, x_2, \ldots, x_{N_{dim}})$ in a region confined to $[-\mathbf{1}, \mathbf{1}]$ of dimension N_{dim} is

$$\mathrm{var}(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}^T (\boldsymbol{K} + \sigma_n^2 \Delta)^{-1} \boldsymbol{k} \quad . \tag{A1}$$

The covariance $k(\boldsymbol{x}_p, \boldsymbol{x}_q)$ between pairs of input variables $(\boldsymbol{x}_p, \boldsymbol{x}_q)$ is defined by

$$k(\boldsymbol{x}_p, \boldsymbol{x}_q) = \sigma_f^2 \exp\left[-\frac{(\boldsymbol{x}_p - \boldsymbol{x}_q)^T (\boldsymbol{x}_p - \boldsymbol{x}_q)}{2\lambda^2} \right] \quad . \tag{A2}$$

While the first term in Equation (A1) is simply $k(\boldsymbol{x}, \boldsymbol{x}) = \sigma_f^2$, we need for the second term

$$\boldsymbol{k} = \sigma_f^2 \begin{pmatrix} \exp\left[-\frac{1}{2\lambda^2} (\boldsymbol{x} - \boldsymbol{x}_1)^T (\boldsymbol{x} - \boldsymbol{x}_1) \right] \\ \exp\left[-\frac{1}{2\lambda^2} (\boldsymbol{x} - \boldsymbol{x}_1)^T (\boldsymbol{x} - \boldsymbol{x}_2) \right] \\ \vdots \\ \exp\left[-\frac{1}{2\lambda^2} (\boldsymbol{x} - \boldsymbol{x}_1)^T (\boldsymbol{x} - \boldsymbol{x}_N) \right] \end{pmatrix} \tag{A3}$$

and the inversion of the matrix $\boldsymbol{M} = \boldsymbol{K} + \sigma_n^2 \Delta$, where the matrix elements are $\Delta_{ii} = \sigma_{d_i}$ ($\Delta_{ij} = 0$ for $i \neq j$) and $K_{ij} = \sigma_f^2 \exp\left[-\frac{1}{2\lambda^2} (\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \right]$. For a given set of hyperparameters the matrix $\boldsymbol{M}$ does not depend on the test vector $\boldsymbol{x}$ and may be treated as a constant in integration over d$\boldsymbol{x}$. Therefore, after the inversion has been performed, $\boldsymbol{M}^{-1}$ can easily be regarded as a pure number. So the second term in Equation (A1) is just a sum over all components with indices $\{i, j\} \in (1, \ldots, N)$,

$$\boldsymbol{k}_*^T (\boldsymbol{K} + \sigma_n^2 \Delta)^{-1} \boldsymbol{k}_* = \sum_{i,j=1}^{N} k_{*i} \left(\boldsymbol{M}^{-1} \right)_{ij} k_{*j} \tag{A4}$$

$$= \sigma_f^4 \sum_{i,j=1}^{N} \left(\boldsymbol{M}^{-1} \right)_{ij} e^{-\frac{(\boldsymbol{x} - \boldsymbol{x}_i)^T (\boldsymbol{x} - \boldsymbol{x}_i)}{2\lambda^2}} e^{-\frac{(\boldsymbol{x} - \boldsymbol{x}_j)^T (\boldsymbol{x} - \boldsymbol{x}_j)}{2\lambda^2}} \quad . \tag{A5}$$

Further let us concentrate on the terms in the nominator of the exponential:

$$(\boldsymbol{x} - \boldsymbol{x}_i)^T (\boldsymbol{x} - \boldsymbol{x}_i) + (\boldsymbol{x} - \boldsymbol{x}_j)^T (\boldsymbol{x} - \boldsymbol{x}_j) = 2 \left[-\boldsymbol{x}^T \frac{\boldsymbol{x}_i + \boldsymbol{x}_j}{2} - \frac{\boldsymbol{x}_i^T + \boldsymbol{x}_j^T}{2} \boldsymbol{x} \right] + \boldsymbol{x}_i \boldsymbol{x}_i^T + \boldsymbol{x}_j \boldsymbol{x}_j^T \quad . \tag{A6}$$

Completing the square gives

$$2 \left[\left(\boldsymbol{x} - \frac{\boldsymbol{x}_i + \boldsymbol{x}_j}{2} \right)^T \left(\boldsymbol{x} - \frac{\boldsymbol{x}_i + \boldsymbol{x}_j}{2} \right) \right] + \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \quad . \tag{A7}$$

We insert Equation (A7) in Equation (A5) and finally get for the variance

$$\mathrm{var}(\boldsymbol{x}) = \sigma_f^2 - \sigma_f^4 \sum_{i,j=1}^{N} \left(\boldsymbol{M}^{-1}\right)_{ij} e^{-\frac{1}{4\lambda^2}(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{x}_i-\boldsymbol{x}_j)} e^{-\frac{1}{\lambda^2}\left[\left(\boldsymbol{x}-\frac{\boldsymbol{x}_i+\boldsymbol{x}_j}{2}\right)^T\left(\boldsymbol{x}-\frac{\boldsymbol{x}_i+\boldsymbol{x}_j}{2}\right)\right]} . \quad \text{(A8)}$$

Only the second exponential in Equation (A8) depends on $\boldsymbol{x}$ and therefore needs to be considered in the integral of the global variance:

$$\int_{-\mathbf{1}}^{\mathbf{1}} \mathbf{d}\boldsymbol{x}\, \mathrm{var}(\boldsymbol{x}) \,. \quad \text{(A9)}$$

We insert Equation (A8) into Equation (A9) and let the integral stay only for the term with $\boldsymbol{x}$ dependency:

$$\begin{aligned}\int_{-\mathbf{1}}^{\mathbf{1}} \mathbf{d}\boldsymbol{x}\, \mathrm{var}(\boldsymbol{x}) \quad &= \quad 2^{N_D}\sigma_f^2 - \sigma_f^4 \sum_{i,j=1}^{N} \left(\boldsymbol{M}^{-1}\right)_{ij} e^{-\frac{1}{4\lambda^2}(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{x}_i-\boldsymbol{x}_j)} \\ &\quad \cdot \int_{-\mathbf{1}}^{\mathbf{1}} \mathbf{d}\boldsymbol{x}\, e^{-\frac{1}{\lambda^2}\left[\left(\boldsymbol{x}-\frac{\boldsymbol{x}_i+\boldsymbol{x}_j}{2}\right)^T\left(\boldsymbol{x}-\frac{\boldsymbol{x}_i+\boldsymbol{x}_j}{2}\right)\right]} . \quad \text{(A10)}\end{aligned}$$

Since the term in the exponential is quadratic it separates into a sum, which itself facilitates the separation of the integral into each dimension. Being simplified to a number of N_d one-dimensional integrals they can easily be solved by employing the error function. To prove this, let us have a closer look at the integral only:

$$\int_{-\mathbf{1}}^{\mathbf{1}} \mathbf{d}\boldsymbol{x}\, e^{-\frac{1}{\lambda^2}\left[\left(\boldsymbol{x}-\frac{\boldsymbol{x}_i+\boldsymbol{x}_j}{2}\right)^T\left(\boldsymbol{x}-\frac{\boldsymbol{x}_i+\boldsymbol{x}_j}{2}\right)\right]} = \prod_{k}^{N_{dim}} \int_{-1}^{1} \mathrm{d}x_k e^{-\frac{1}{\lambda^2}\left[\left(x_k-\frac{x_{ik}+x_{jk}}{2}\right)^2\right]} . \quad \text{(A11)}$$

Focusing on a the kth integral and substituting $\tau_k = (x_k - \frac{x_{ik}+x_{jk}}{2})/\lambda$ some error functions evolve to end up finally in:

$$\begin{aligned}\int_{-1}^{1} \mathrm{d}x_k e^{-\frac{1}{\lambda^2}\left[\left(x_k-\frac{x_{ik}+x_{jk}}{2}\right)^2\right]} &= \lambda \int_{\frac{1}{\lambda}-\frac{x_{ik}+x_{jk}}{2\lambda}}^{-\frac{1}{\lambda}-\frac{x_{ik}+x_{jk}}{2\lambda}} \mathrm{d}\tau_k e^{-\tau_k^2} \quad \text{(A12)} \\ &= \lambda \left[\int_{\frac{1}{\lambda}-\frac{x_{ik}+x_{jk}}{2\lambda}}^{0} \mathrm{d}\tau_k e^{-\tau_k^2} + \int_{0}^{-\frac{1}{\lambda}-\frac{x_{ik}+x_{jk}}{2\lambda}} \mathrm{d}\tau_k e^{-\tau_k^2}\right] \quad \text{(A13)} \\ = \frac{\sqrt{\pi}}{2}\lambda &\left\{\mathrm{erf}\left[\frac{1}{\lambda}-\frac{x_{ik}+x_{jk}}{2\lambda}\right] - \mathrm{erf}\left[-\frac{1}{\lambda}-\frac{x_{ik}+x_{jk}}{2\lambda}\right]\right\} . \quad \text{(A14)}\end{aligned}$$

This concludes the study. Simply inserting Equation (A14) into Equation (A10) succeeds in the result reported in the paper:

$$\begin{aligned}\int_{-\mathbf{1}}^{\mathbf{1}} \mathbf{d}\boldsymbol{x}\, \mathrm{var}(\boldsymbol{x}) \quad &= \quad 2^{N_D}\sigma_f^2 - \sigma_f^4 \left(\frac{\sqrt{\pi}}{2}\lambda\right)^{N_{dim}} \sum_{i,j=1}^{N} \left(\boldsymbol{M}^{-1}\right)_{ij} e^{-\frac{1}{4\lambda^2}(\boldsymbol{x}_i-\boldsymbol{x}_j)^T(\boldsymbol{x}_i-\boldsymbol{x}_j)} \\ &\quad \cdot \prod_{k}^{N_{dim}} \left\{\mathrm{erf}\left[\frac{1}{\lambda}-\frac{x_{ik}+x_{jk}}{2\lambda}\right] - \mathrm{erf}\left[-\frac{1}{\lambda}-\frac{x_{ik}+x_{jk}}{2\lambda}\right]\right\} . \quad \text{(A15)}\end{aligned}$$

References

1. Mockus, J. *Bayesian Approach to Global Optimization*; Springer: Berlin/Heidelberg, Germany, 1989.
2. Frazier, P.I. A Tutorial on Bayesian Optimization. 2018. Available online: http://xxx.lanl.gov/abs/1807.02811 (accessed on 11 September 2021).

3. Ru, B.; McLeod, M.; Granziel, D.; Osborne, M.A. Fast Information-theoretic Bayesian Optimisation. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80.
4. Hernández-Lobato, D.; Hernández-Lobato, J.; Shah, A.; Adams, R.P. Predictive Entropy Search for Multi-objective Bayesian Optimization. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; Volume 48.
5. Rasmussen, C.; Williams, C. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, UK, 2006.
6. Krige, D.G. A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. *J. Chem. Metal. Mining Soc. S. Afr.* **1951**, *52*, 119–139.
7. Matheron, G. Principles of geostatistics. *Econ. Geol.* **1963**, *58*, 1246–1266. [CrossRef]
8. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.
9. Bishop, C. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1996.
10. Cohn, D. Neural Network Exploration Using Optimal Experiment Design. *Neural Netw.* **1996**, *9*, 1071–1083. [CrossRef]
11. MacKay, D.J.C. *Bayesian Approach to Global Optimization: Theory and Applications*; Kluwer Academic: Dordrecht, The Netherlands, 2013.
12. Neal, R.M. *Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification*; Technical Report 9702; Department of Statistics, University of Toronto: Toronto, ON, Canada, 1997.
13. Seo, S.; Wallat, M.; Graepel, T.; Obermayer, K. Gaussian process regression: active data selection and test point rejection. In Proceedings of the International Joint Conference on Neural Networks, Como, Italy, 24–27 July 2000; pp. 241–246.
14. Gramacy, R.B.; Lee, H.K.H. Adaptive Design and Analysis of Supercomputer Experiments. *Technometrics* **2009**, *51*, 130–145. [CrossRef]
15. Preuss, R.; von Toussaint, U. Prediction of Plasma Simulation Data with the Gaussian Process Method. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*; Niven, R., Ed.; AIP Publishing: Melville, NY, USA, 2014; Volume 1636, p. 118.
16. Preuss, R.; von Toussaint, U. Global Optimization Employing Gaussian Process-Based Bayesian Surrogates. *Entropy* **2018**, *20*, 201. [CrossRef] [PubMed]
17. Invertible Matrix, Section 3.7: Blockwise Inversion. Available online: https://en.wikipedia.org/wiki/Invertible_matrix#Blockwise_inversion (accessed on 25 May 2021).
18. Preuss, R.; von Toussaint, U. Optimization Employing Gaussian Process-Based Surrogates. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; Volume 239.