Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 5, No. 2, pp. 283–296, May 2009

OPTIMIZATION BY COMBINING EXPERIMENTS WITH SAMPLES OF VARYING PRECISION

Koji Tsuchiya and Hiroshi Morita

Abstract: Efficient sampling and effective use of data are desirable because the number of experiments that might be undertaken is limited by time and cost. The precision of prediction depends on the number and precision of samples, which have a trade-off relation in sampling. Whereas low precision data requires a lot of samples to improve prediction precision, obtaining high precision samples increases the required time and cost. We propose a prediction and optimization method which combines different samples of varying precision. Data is divided into several levels according to its precision and then an auto-regression model with a Gaussian process is assumed between each level. On every level the prediction is updated by adding higher precision samples. Then, we consider the use of Efficient Global Optimization (EGO) as an indicator of the possibility of a solution being optimal. This indicator consists of the predicted value and the accuracy of the predictor at each point. The prediction is sequentially updated by adding higher precision samples which are selected by the indicator. By updating the prediction, the optimal solution can be searched for with an efficient sampling method. Finally, the precision and efficiency of optimization is examined through numerical simulations which utilize the proposed method.

Key words: *efficient global optimization, Kriging, response surface, different precision data*

Mathematics Subject Classification: *65K10*

1 Introduction

Recently, various improved methods for sampling data have been developed which make use of developments in technology, yielding a variety of resultant datasets. The sampling method differs, depending upon whether or not an examination is made in detail, and sampling points are designed in accordance with the region of interest. The time and cost of sampling data often need to be considered when an efficient sampling methodology is desired, because the number of experiments is limited by time and cost. We also have to obtain data suitable for the intended purpose and to select an adequate data analysis method. In data analysis, methods for prediction are widely used. Generally, the accuracy of prediction depends on the number of samples and the precision of sampling. However these two factors are in a trade off relation in terms of the time and cost of sampling. Prediction using high precision data is highly reliable and, by increasing the amount of high precision data, we can ensure increased accuracy of the prediction. However, to obtain high precision data may be expensive; a long measurement time or expensive equipment may be necessary. On the other hand, low precision data can be obtained easily, but a lot of samples are required to improve the accuracy of prediction. Kennedy et al. [7] proposed combining data of different precision obtained via the finite element method. They also attempted to

obtain data efficiently and to improve the precision of prediction. There are several methods of prediction which combine data from different experiments [3, 8].

Efficient Global Optimization (EGO) is one of the methods to search for a global optimal solution with a small number of samples [5]. It is important to consider a balance between global and local regions when searching for such a global optimum. A global optimum may not be obtained if we focus only on certain local regions; however too much time or too many samples might be needed for convergence to the optimum if an exploration of the global region is emphasized. In EGO, this balance between global and local regions is taken into account [6, 12]. At each point the probability that the point is the optimum can be approximately estimated from the prediction value and the accuracy of the predictor at each point; all regions can be evaluated. Sampling is done at the point with the highest probability and then the prediction and probability of the possible optimum values are updated by adding a new sample. The approximate optimal solution is obtained by repeating this sampling and prediction procedure.

The purpose of this study is to solve the trade-off between the precision of prediction and the efficiency of sampling. We consider the prediction of response variables and a global optimization problem. We try to refine the precision of prediction using high and low precision data, improving the prediction by using only high precision data with inefficient sampling. Furthermore, by using the proposed method of prediction with high efficiency and precision, we try to search for the optimal point efficiently by utilizing the EGO algorithm. By combining different data samples of varying precision, the accuracy of optimization and prediction using high precision data is retained while, at the same time, the efficiency of sampling that results from using low precision data is obtained.

The proposed method can be applied for any dimensional problems. However, in this paper, we consider the case of physical experiments to obtain an optimal operational level of parameters. For example, at the statistical optimization by ANOVA (analysis of variance), it usually deals with one or two factors, say one-way and two-way layout. We believe that it has a worthwhile contribution even this size of problems.

In Section 2, we describe the Kriging model, which is a prediction method for a response surface that allows prediction of a value, together with an estimate of its accuracy. Also, the EGO algorithm for prediction and optimization is described in this section. Section 3 deals with different precision data and a corresponding model. The proposed method of prediction is described in Section 4. EGO using the proposed method is then applied to different precision data; the procedure is described in Section 5. The numerical experiments are described in Section 6. Finally, Section 7 contains our conclusion.

2 Efficient Global Optimization with the Kriging Model

2.1 Efficient Global Optimization

EGO is a method for solving a global optimization problem [4, 5, 6]. Using this technique, an optimal solution can be searched for with only a small number of samples. The criterion for the choice of new sampling points with EGO utilizes the global predictor and its accuracy.

Here, we consider a search for a global minimum. Figure 1 shows an example of updating a predicted value with an initial design to approximate a true function. If the neighborhood of a minimum of the predicted values is sampled, it is to be expected that a better solution will be obtained. However, it is not enough to consider only the predicted values to obtain a global minimum. It is necessary to consider the variance of predictors as shown in Figure 2. Here the variance of the predictor at -2 is small since there are several data values

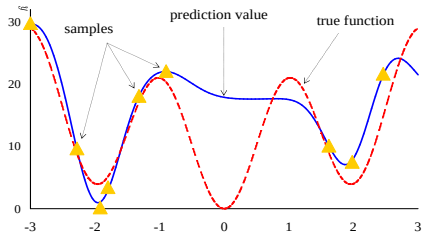

Figure 1: initial prediction

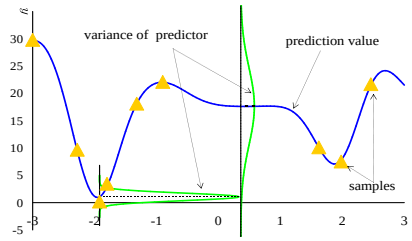

Figure 2: variance of predictors

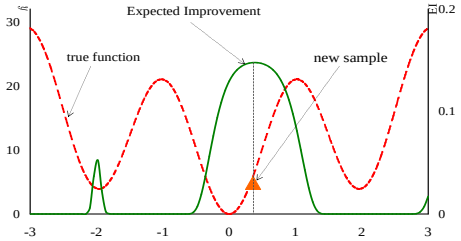

Figure 3: EI and a new sample

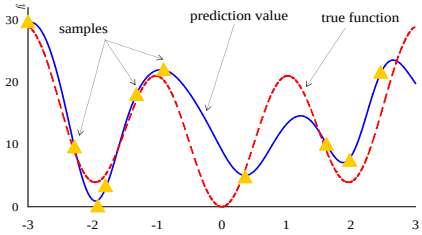

Figure 4: updated prediction

close to -2, but the variance of the predictor around 0 is large since there is no data in a neighborhood of 0. Expected Improvement(EI) is proposed as an indicator for selecting a new sample point, which takes into account a balance between a local and global search. EI stands for the degree of improvement of the minimum in the case of sampling at a point by comparing the current minimum sample value $y_{\min} = \min\{y(\boldsymbol{x}^{(1)}), \ldots, y(\boldsymbol{x}^{(n)})\}$, that is, EI is the expected value that a sampling point becomes a new minimizer. Not only is a local region around a minimum of the predicted value considered, but also the global region including points with a large predicted value, because this expected value includes a measure of the accuracy of the predictor. Figure 3 shows that a new sample will be taken at the point of highest EI value. Then the prediction is updated as shown in Figure 4.

2.2 Kriging Model

We use the Kriging model to predict response variables. The Kriging model was originally used in geostatistics or spatial statistics and, in particular, it has been used as an interpolation method [1]. Furthermore, Kriging has been applied in the design and analysis of computer experiments [10, 11].

In the Kriging model, a response variable $Y(\boldsymbol{x})$ at the point $\boldsymbol{x}$ is assumed to be determined by a regression model and a Gaussian process, as follows:

$$Y(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta} + \eta(\boldsymbol{x}) + \epsilon, \tag{2.1}$$

where ϵ is error or noise, distributed normally with mean zero and variance σ_ϵ. $\boldsymbol{h}(\cdot) = (h_1(\cdot), \ldots, h_p(\cdot))^{\mathrm{T}}$ is known as the regressor vector, $\boldsymbol{\beta}_t = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is an unknown coefficient vector and p is the degree of the regression. In the case of a simple linear function, we set $\boldsymbol{h}(x) = (1, x)^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$, then $\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\beta} = \beta_0 + \beta_1 x$. $\eta(\boldsymbol{x})$ is a Gaussian process with mean zero, and covariance function $\eta(\boldsymbol{x})$ and $\eta(\boldsymbol{x}')$ given by

$$\mathrm{Cov}\{\eta(\boldsymbol{x}), \eta(\boldsymbol{x}')\} = \sigma_\eta^2 c(\boldsymbol{x}, \boldsymbol{x}'), \tag{2.2}$$

where σ_η^2 is the variance, and $c(\boldsymbol{x}, \boldsymbol{x}')$ is a correlation function. This correlation function is dependent on the distance between $\boldsymbol{x}$ and $\boldsymbol{x}'$. If a prediction point is far from any sampling

points, the standard error of the predictor at this point is large. If a prediction point is close to sampling points, the standard error of the predictor at this point is small. Various correlation functions are proposed in [11].

Consider a response variable $Y(\boldsymbol{x}_0)$ at point $\boldsymbol{x}_0$. The following notation is introduced. Let sample points be $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ and a set of samples for these points be $\boldsymbol{Y} = \big(Y(\boldsymbol{x}^{(1)}), \ldots, Y(\boldsymbol{x}^{(n)})\big)^{\mathrm{T}}$. H is the design matrix with elements consisting of $\boldsymbol{h}(\cdot)$ of each sample point, expressed as

$$H = \begin{pmatrix} \boldsymbol{h}(\boldsymbol{x}^{(1)})^{\mathrm{T}} \\ \vdots \\ \boldsymbol{h}(\boldsymbol{x}^{(n)})^{\mathrm{T}} \end{pmatrix}. \tag{2.3}$$

$\boldsymbol{r}(\boldsymbol{x}_0)$ is an $n \times 1$ covariance vector between $Y(\boldsymbol{x}_0)$ and samples $Y(\boldsymbol{x}^{(1)}), \ldots, Y(\boldsymbol{x}^{(n)})$ and R is the $n \times n$ matrix that consists of covariance functions between each sample. The k-th element of $\boldsymbol{r}(\boldsymbol{x}_0)$ and the (i, j)-th element of R are given by

$$\big[\boldsymbol{r}(\boldsymbol{x}_0)\big]_k = \sigma_\eta^2 c(\boldsymbol{x}_0, \boldsymbol{x}^{(k)}) \tag{2.4}$$

$$\big[R\big]_{i,j} = \sigma_\eta^2 c(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) + \sigma_\epsilon \delta_{i,j}, \tag{2.5}$$

respectively, where if $i = j$, $\delta_{i,j} = 1$, otherwise $\delta_{i,j} = 0$. Then, the Best Linear Unbiased Predictor (BLUP) is given by

$$\hat{Y}(\boldsymbol{x}_0) = \boldsymbol{h}(\boldsymbol{x}_0)^{\mathrm{T}} \hat{\boldsymbol{\beta}} + \boldsymbol{r}(\boldsymbol{x}_0)^{\mathrm{T}} R^{-1} (\boldsymbol{Y} - H\hat{\boldsymbol{\beta}}). \tag{2.6}$$

$\hat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$, thus $\hat{\boldsymbol{\beta}} = \big(H^{\mathrm{T}} R^{-1} H\big)^{-1} H^{\mathrm{T}} R^{-1} \boldsymbol{Y}$. Suppose that the response variable for given samples $\boldsymbol{Y}$ is $Y^\dagger(\boldsymbol{x}_0) = [Y(\boldsymbol{x}_0)|\boldsymbol{Y}]$, then the expected value $E\{Y^\dagger(\boldsymbol{x}_0)\}$ coincides with eq.(2.6).

The Mean Squared Error (MSE) of the predictor is

$$\begin{aligned} s^2(\boldsymbol{x}_0) = \sigma_\eta^2 &- \boldsymbol{r}(\boldsymbol{x}_0)^{\mathrm{T}} R^{-1} \boldsymbol{r}(\boldsymbol{x}_0) \\ &+ \big(\boldsymbol{h}(\boldsymbol{x}_0) - H^{\mathrm{T}} R^{-1} \boldsymbol{r}(\boldsymbol{x}_0)\big)^{\mathrm{T}} \big(H^{\mathrm{T}} R^{-1} H\big)^{-1} \big(\boldsymbol{h}(\boldsymbol{x}_0) - H^{\mathrm{T}} R^{-1} \boldsymbol{r}(\boldsymbol{x}_0)\big). \end{aligned} \tag{2.7}$$

The accuracy of the prediction can be expressed using MSE.

2.3 Expected Improvement

Suppose that a predictor $Y^\dagger(\boldsymbol{x})$ is normally distributed with mean $\hat{Y}(\boldsymbol{x})$ and variance $s^2(\boldsymbol{x})$, where $\hat{Y}(\boldsymbol{x})$ is the BLUP and $s^2(\boldsymbol{x})$ is the MSE. The degree to which a sample at point $\boldsymbol{x}$ is smaller than the current minimum is expressed by

$$I(\boldsymbol{x}) = \max\big\{y_{\min} - Y^\dagger(\boldsymbol{x}), 0\big\}. \tag{2.8}$$

The expected value of $I(\boldsymbol{x})$ is the value used to determine whether $\boldsymbol{x}$ may become a new minimizer. Setting $\mathrm{EI}(\boldsymbol{x}) = E\{I(\boldsymbol{x})\}$, we have

$$\begin{aligned} \mathrm{EI}(\boldsymbol{x}) &= \int_{-\infty}^{\frac{y_{\min} - \hat{Y}(\boldsymbol{x})}{s(\boldsymbol{x})}} (y_{\min} - \hat{Y}(\boldsymbol{x}) - s(\boldsymbol{x})\tau)\phi(\tau) d\tau \\ &= \big(y_{\min} - \hat{Y}(\boldsymbol{x})\big)\Phi\left(\frac{y_{\min} - \hat{Y}(\boldsymbol{x})}{s(\boldsymbol{x})}\right) + s(\boldsymbol{x})\phi\left(\frac{y_{\min} - \hat{Y}(\boldsymbol{x})}{s(\boldsymbol{x})}\right), \end{aligned} \tag{2.9}$$

where $\Phi(\cdot)$ is the standard normal probability cumulative density function and $\phi(\cdot)$ is the standard normal probability density function [4, 5, 6, 11]. To obtain a smaller value than the current minimum at a point with large EI, the next sampling point is given by

$$\boldsymbol{x}^{(n+1)} = \arg\max\{\mathrm{EI}(\boldsymbol{x})\} \tag{2.10}$$

Adding the new sample $y(\boldsymbol{x}^{(n+1)})$, the predictor is updated. The process of prediction, estimating EI and adding a sample is repeated to search for a optimal point. This process is indicated schematically in Figures 1, 2, 3 and 4.

3 Combining Data with Different Levels of Precision

3.1 Kennedy's Prediction Model

Kennedy et al. [7] combined data with different levels of precision that were obtained using the finite element method. The finite element method, which is used in many fields, including structural and fluid mechanics, is a numerical technique that subdivides the object with a mesh and an approximate solution is obtained by solving the partial problem in each subregion of the mesh. The solution consists of the sum of those obtained for each partial problem; this solution is then treated as one data item in the subsequent analysis. If the mesh size is small, the approximate solution is close to the true solution, i.e., the data precision is high. However considerable calculation time may be required to obtain the solution. On the other hand, if the mesh size is large, data precision is low, but calculation time is short. In this case, data precision and calculation time have a trade off relation. By combining different precision data from the finite element model with different mesh sizes, both precision of prediction and efficiency are improved. Moreover, error bounds for the data obtained from the finite element model can be calculated. As shown in Figure 5, the same set of inputs always results in the same output.

Kennedy et al. assumed an auto-regression model between the different precision data. Using s finite element models having different grid sizes, outputs from these models are set to $q_1(\cdot), \ldots, q_s(\cdot)$ in the order of largest to smallest, according to the size of the grid. $q_s(\cdot)$ is the output of the highest precision model.

Suppose that the set of inputs is $\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(n_t)}$, where n_t is the number of data elements, and the output is $\boldsymbol{q}_t = \big(q_t(x_t^{(1)}), \ldots, q_t(\boldsymbol{x}_t^{(n_t)})\big)^{\mathrm{T}}$ for the model having the t-th largest size of the grid, i.e., the t-th highest precision model ($t = 1, \ldots, s$). The auto-regression model between these outputs is given by

$$q_1(\boldsymbol{x}) = z_1(\boldsymbol{x}) \tag{3.1}$$

$$q_t(\boldsymbol{x}) = \rho_{t-1} q_{t-1}(\boldsymbol{x}) + z_t(\boldsymbol{x}), \qquad (t = 2, \ldots, s), \tag{3.2}$$

where ρ_{t-1} is an auto-regression coefficient. The Kriging model $z_t(\boldsymbol{x})$ is not independent on $q_{t-1}(\cdot)$, $\ldots$, $q_1(\cdot)$. In this regard, the noise term of eq.(2.1) is ignored for data arising from the computer experiment.

Given all samples $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_s$, output of the highest precision finite element model is predicted. When considering the prediction at point $\boldsymbol{x}_0$, $[q_s(\boldsymbol{x}_0), \boldsymbol{q}_s, \ldots, \boldsymbol{q}_1]$ has a multivariate normal distribution; $z_t(\boldsymbol{x}_0)$ consists of a Gaussian process. The predicted value is given by the expected value $E\big\{[q_s(\boldsymbol{x}_0)|\boldsymbol{q}_1, \ldots, \boldsymbol{q}_s]\big\}$ of the conditional distribution, given all samples.

In this way, using an assumed auto-regression model between outputs from different finite element models, the trade off between precision of prediction and the calculation time in obtaining data is taken into account.

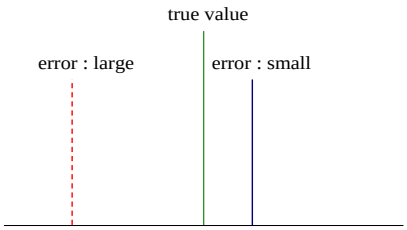

Figure 5: deterministic error

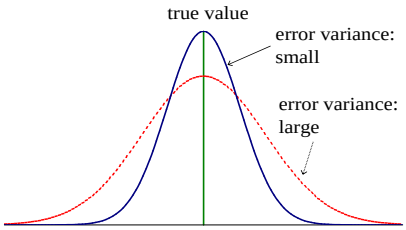

Figure 6: stochastic error

3.2 Expanding the Prediction Model

For actual data modeling, there is uncertainty in observation and measurement, unlike the case of the computer experiment results. When we allow for noise in the observations, the response is stochastic.

This study deals with a degree of uncertainty of precision of data. Suppose that the true function is f and the sample y, then $y = f + \epsilon$, where ϵ is noise, which follows the normal distribution $N(0,\ \sigma_\epsilon^2)$. The error variance σ_ϵ^2 can be regarded as a measure of precision of the data because σ_ϵ^2 represents variation in the sample. Data with small σ_ϵ^2 is regarded as data of high precision.

Data of different precision are divided into several levels according to their precision. Let the t-th lowest precision data belong to the t-th level ($t = 1, \ldots, s$). Higher precision data belong to higher levels; data with errors of greater variance belong to lower levels. Data $y_t(\cdot)$ belonging to the t-th level is expressed as $y_t(\cdot) = f(\cdot) + \epsilon_t$, $\epsilon_t \sim N(0,\ \sigma_{\epsilon t}^2)$, where ϵ_t and $\sigma_{\epsilon t}^2$ are respectively noise and its variance for data at the t-th level. Division into levels is such that

$$\sigma_{\epsilon 1}^2 > \sigma_{\epsilon 2}^2 > \cdots > \sigma_{\epsilon s}^2, \quad (t = 1, \ldots, s). \tag{3.3}$$

The highest precision data elements, which have the smallest error variance, belong to the highest (s-th) level. Data at the first level have the lowest precision.

Our study is using different precision data; we are using the error variance as a measure of the precision of the data. Data elements having the highest precision, i.e., having the smallest error variance, belong to the highest level s. When the error variance of data elements is larger, they belong to a lower level. We assume that the level for each data element is known, although the actual values of the variances $\sigma_{\epsilon 1}^2, \sigma_{\epsilon 2}^2, \cdots, \sigma_{\epsilon s}^2$ are supposed to be unknown. So, our study treats data divided into each levels, according to eq.(3.3).

An auto-regression model is adopted which uses the relation between levels, similar to that of Kennedy's method. At the t-th level, the response variable $Y_t(\boldsymbol{x})$ is assumed to satisfy

$$Y_1(\boldsymbol{x}) = Z_1(\boldsymbol{x}) \tag{3.4}$$

$$Y_t(\boldsymbol{x}) = \rho_{t-1} Y_{t-1}(\boldsymbol{x}) + Z_t(\boldsymbol{x}), \quad (t = 2, \ldots, s), \tag{3.5}$$

where ρ_{t-1} is an auto-regression coefficient. Data in the t-th level is not always closer to true value than that in $(t-1)$-th level because of the stochastic error. However, data in higher level is more reliable than data in lower level in the sense of the data precision. So, let ρ_{t-1} that shows a weight factor between levels be in the range from 0 to 1. $Z_t(\cdot)$ is assumed to satisfy a Kriging model:

$$Z_t(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\beta}_t + \eta_t(\boldsymbol{x}) + \epsilon_t. \tag{3.6}$$

$\boldsymbol{\beta}_t$ is a $p \times 1$ vector of regression coefficients, $\eta_t(\cdot)$ is a Gaussian process and $\epsilon_t \sim N(0,\ \sigma_{\epsilon t}^2)$ is noise. $\eta_t(\cdot)$ has mean zero and covariance function

$$\mathrm{Cov}\{\eta_t(\boldsymbol{x}), \eta_t(\boldsymbol{x}')\} = \sigma_{Y_t}^2 c_t(\boldsymbol{x}, \boldsymbol{x}') \tag{3.7}$$

$$c_t(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\gamma_t(\boldsymbol{x} - \boldsymbol{x}')^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}')\}, \tag{3.8}$$

where $\sigma_{\eta t}^2$ and $c_t(\boldsymbol{x}, \boldsymbol{x}')$ are t-th level variance and correlation functions, respectively. The exponential form (3.8) is widely used as a correlation function [7, 8, 10].

Because the auto-regression model includes the Kriging model, at each level not only a predicted value is obtained but also the accuracy of the predictor can be estimated. Therefore, EI can be used at each level. A sample point is selected by applying EI as an indicator; higher precision predicted values are added.

The accuracy of the predictor depends on the distance between the sample and the predicted point; the variance is small (and consequently the accuracy is high) where this distance is small. If a predicted point is far from any sampling points, the variance of the predictor at this point is large, so the predictor is of low accuracy. When using EGO with different precision data, adding to higher precision prediction according to EI, high precision data elements are used only in a limited region.

4 Prediction and Optimization with Data of Varying Precision

4.1 Best Linear Unbiased Pedictor with Data of Varying Precision

When the predictor of the extended model is considered, a methodology to update the predictor using data divided into several levels is necessary, each time adding higher precision samples. Let $\boldsymbol{Y}_t^{n_t} = \big(Y_t(\boldsymbol{x}_t^{(1)}), \ldots, Y_t(\boldsymbol{x}_t^{(n_t)})\big)^{\mathrm{T}}$ be at the t-th level. At the first level, a predictor is equal to that of the Kriging model, and it is the BLUP of $Y_1(\boldsymbol{x})$ as shown in eq.(2.6). The predicted value is the expected value of the predictor $Y_1^{\dagger}(\boldsymbol{x})$ given $\boldsymbol{Y}_1$.

The covariance function between $Y_1^{\dagger}(\boldsymbol{x})$ and $Y_1^{\dagger}(\boldsymbol{x}')$ is dependent on the distance between $\boldsymbol{x}$ and $\boldsymbol{x}'$ as shown in eq.(3.7) and (3.8). Let $C_1(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}\{Y_1^{\dagger}(\boldsymbol{x}), Y_1^{\dagger}(\boldsymbol{x}')\}$, then

$$\begin{aligned} C_1(\boldsymbol{x}, \boldsymbol{x}') = {} & \sigma_{Y1}^2 c_1(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{r}_1(\boldsymbol{x})^{\mathrm{T}} R_1^{-1} \boldsymbol{r}_1(\boldsymbol{x}') \\ & + (\boldsymbol{h}(\boldsymbol{x}) - H_1^{\mathrm{T}} R_1^{-1} \boldsymbol{r}_1(\boldsymbol{x}))^{\mathrm{T}} \big(H_1^{\mathrm{T}} R_1^{-1} H_1\big)^{-1} (\boldsymbol{h}(\boldsymbol{x}') - H_1^{\mathrm{T}} R_1^{-1} \boldsymbol{r}_1(\boldsymbol{x}')), \end{aligned} \tag{4.1}$$

where H_1, r_1 and R_1 are given by eq.(2.3), (2.4) and (2.5), respectively. When $\boldsymbol{x} = \boldsymbol{x}'$, eq.(4.1) expresses MSE at the first level, for which ${s_1}^2(\boldsymbol{x}) = C_1(\boldsymbol{x}, \boldsymbol{x})$.

Next, we consider a predictor at the $t(\geq 2)$-th level. Given sample points $\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(n_t)}$, a set of predictors of the $(t-1)$-th level is expressed by $\boldsymbol{Y}_{t-1}^{\dagger} = \big(Y_{t-1}^{\dagger}(\boldsymbol{x}^{(1)}), \ldots, Y_{t-1}^{\dagger}(\boldsymbol{x}_t^{(n_t)})\big)^{\mathrm{T}}$. At the new sample points, a set of lower level predictors $\boldsymbol{Y}_{t-1}^{\dagger}$ is compared with that from the new samples $\boldsymbol{Y}_t$, and then the predictor is updated. From eq.(3.5), the predictor is expressed by $Y_t(\boldsymbol{x}) = \rho_{t-1} Y_{t-1}^{\dagger}(\boldsymbol{x}) + \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\beta}_t + \epsilon_t$.

Consider the BLUP of $Y_t(\boldsymbol{x}_0)$. Let

$$\hat{Y}_t(\boldsymbol{x}_0) = \boldsymbol{a}(\boldsymbol{x}_0)^{\mathrm{T}} \boldsymbol{Y}_t^{n_t}. \tag{4.2}$$

Then the BLUP is obtained by selecting $\boldsymbol{a}(\boldsymbol{x}_0)$ that minimizes $\mathrm{MSE}\{\hat{Y}_t(\boldsymbol{x}_0)\}$ subject to the unbiased constraint, which can be expressed as the following problem:

$$\begin{array}{ll} \min & E\big\{\big(\boldsymbol{a}(\boldsymbol{x}_0)^{\mathrm{T}} \boldsymbol{Y}_t^{n_t} - Y_t(\boldsymbol{x}_0)\big)^2\big\} \\ \text{subject to} & E\big\{\boldsymbol{a}(\boldsymbol{x}_o)^{\mathrm{T}} \boldsymbol{Y}_t^{n_t} - Y_t(\boldsymbol{x}_o)\big\} = 0 \end{array}. \tag{4.3}$$

Here, let $\boldsymbol{r}_t(\boldsymbol{x}_0)$ be the covariance vector between $Y_t(\boldsymbol{x}_0)$ and each sample $Y_t(\boldsymbol{x}_t^{(1)})$, …, $Y_t(\boldsymbol{x}_t^{(n_t)})$, and similarly R_t be the covariance matrix consisting of covariance functions between each $Y_t(\boldsymbol{x}_t^{(1)})$, …, $Y_t(\boldsymbol{x}_t^{(n_t)})$. The k-th element of $\boldsymbol{r}_t(\boldsymbol{x}_0)$ and the (i,j)-th element of R_t are given by

$$\big[\boldsymbol{r}_t(\boldsymbol{x}_0)\big]_k = \rho_{t-1}^2 C_{t-1}(\boldsymbol{x}_0, \boldsymbol{x}_t^{(k)}) + \sigma_{Y_t}^2 c_t(\boldsymbol{x}_0, \boldsymbol{x}_t^{(k)}) \tag{4.4}$$

$$\big[R_t\big]_{i,j} = \rho_{t-1}^2 C_{t-1}(\boldsymbol{x}_t^{(i)}, \boldsymbol{x}_t^{(j)}) + \sigma_{Y_t}^2 c_t(\boldsymbol{x}_t^{(i)}, \boldsymbol{x}_t^{(j)}) + \sigma_{\epsilon_t}^2 \delta_{i,j}, \tag{4.5}$$

respectively, where if $i = j$, $\delta_{i,j} = 1$, otherwise $\delta_{i,j} = 0$. At sample points $(\boldsymbol{x}^{(1)}), \ldots, (\boldsymbol{x}_t^{(n_t)})$ at the t-th level, let a set of predicted values at this $(t-1)$-th level be $\hat{\boldsymbol{Y}}_{t-1} = \big(\hat{Y}_{t-1}(\boldsymbol{x}^{(1)}), \ldots, \hat{Y}_{t-1}(\boldsymbol{x}_t^{(n_t)})\big)^{\mathrm{T}}$, and additionally, let $\boldsymbol{g}_t(\boldsymbol{x}_0) = \big(\hat{Y}_{t-1}(\boldsymbol{x}_0), \boldsymbol{h}(\boldsymbol{x}_o)^{\mathrm{T}}\big)^{\mathrm{T}}$ and $F_t = \big(\hat{\boldsymbol{Y}}_{t-1}^{n_t}, H_t\big)$. Then eq.(4.3) can be expressed as the following problem :

$$\begin{array}{ll} \min & \boldsymbol{a}(\boldsymbol{x}_0)^{\mathrm{T}} R_t \boldsymbol{a}(\boldsymbol{x}_0) - 2\boldsymbol{a}(\boldsymbol{x}_0)^{\mathrm{T}} \boldsymbol{r}(\boldsymbol{x}_0) \\ \text{subject to} & F_t^{\mathrm{T}} \boldsymbol{a}(\boldsymbol{x}_0) - \boldsymbol{g}_t(\boldsymbol{x}_0) = 0 \end{array} . \tag{4.6}$$

Introducing Lagrange multipliers $\boldsymbol{\lambda}(\boldsymbol{x}_0)$, eq.(4.6) yields

$$\begin{pmatrix} 0 & F_t^{\mathrm{T}} \\ F_t & R_t \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}(\boldsymbol{x}_0) \\ \boldsymbol{a}(\boldsymbol{x}_0) \end{pmatrix} = \begin{pmatrix} \boldsymbol{g}(\boldsymbol{x}_0) \\ \boldsymbol{r}_t(\boldsymbol{x}_0) \end{pmatrix}. \tag{4.7}$$

Solutions for $\boldsymbol{\lambda}$ and $\boldsymbol{a}$ can then be calculated. Therefore we have the BLUP at the t-th level as

$$\begin{aligned} \hat{Y}_t(\boldsymbol{x}_0) = \boldsymbol{g}_t(\boldsymbol{x}_0)^{\mathrm{T}} (H_t^{\mathrm{T}} R_t^{-1} H_t)^{-1} H_t^{\mathrm{T}} R_t^{-1} \boldsymbol{Y}_t^{n_t} \\ + \boldsymbol{r}_t(\boldsymbol{x}_0)^{\mathrm{T}} R_t^{-1} \big(\boldsymbol{Y}_t^{n_t} - F_t (H_t^{\mathrm{T}} R_t^{-1} H_t)^{-1} H_t^{\mathrm{T}} R_t^{-1} \boldsymbol{Y}_t^{n_t}\big). \end{aligned} \tag{4.8}$$

Additionally setting the covariance function between $Y_t^{\dagger}(\boldsymbol{x})$ and $Y_t^{\dagger}(\boldsymbol{x}')$ as $C_t(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}\big\{Y_t^{\dagger}(\boldsymbol{x}), Y_t^{\dagger}(\boldsymbol{x}')\big\}$, we have

$$\begin{aligned} C_t(\boldsymbol{x}_0, \boldsymbol{x}_0') = \rho_{t-1}^2 C_{t-1}(\boldsymbol{x}, \boldsymbol{x}') + \sigma_{Y_t}^2 c_t(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{r}_t(\boldsymbol{x})^{\mathrm{T}} R_t^{-1} \boldsymbol{r}_t(\boldsymbol{x}') \\ + (\boldsymbol{g}_t(\boldsymbol{x}) - F_t^{\mathrm{T}} R_t^{-1} \boldsymbol{r}_t(\boldsymbol{x}))^{\mathrm{T}} (F_t^{\mathrm{T}} R_t^{-1} F)^{-1} (\boldsymbol{g}_t(\boldsymbol{x}') - F_t^{\mathrm{T}} R_t^{-1} \boldsymbol{r}_t(\boldsymbol{x}')) \end{aligned} \tag{4.9}$$

When $\boldsymbol{x} = \boldsymbol{x}'$, eq.(4.9) expresses theMSE at the t-th level and ${s_t}^2(\boldsymbol{x}) = C_t(\boldsymbol{x}, \boldsymbol{x})$.

4.2 Estimating the Parameters

To estimate the parameters, we make the same two assumptions as Kennedy et al. [7]. First, all parameters from the first to the $(t-1)$-th level are treated as fixed at the t-th level . Firstly, we assume that parameters at the t-th level are little influenced by those at lower levels. Secondly, only current data and lower predictors are utilized. This means that data elements belonging to the $(t+1)$-th or higher levels are not considered during estimation at the t-th level. If we make these two assumptions, parameters at each level can be estimated to the order of lower levels.

The parameters belonging to the first level are $\boldsymbol{\beta}_1$, $\sigma_{Y_1}^2$, $\sigma_{\epsilon_1}^2$ and γ_1. Let $\sigma_{\mathrm{sum1}}^2 = \sigma_{Y_1}^2 + \sigma_{\epsilon_1}^2$, ${\sigma_1^*}^2 = \frac{\sigma_{Y_1}^2}{\sigma_{\mathrm{sum1}}^2}$, and $R_1^* = \frac{1}{\sigma_{\mathrm{sum1}}^2} R_1$. Then the diagonal elements of R_1^* are 1, and the others

are $\sigma_1^{*2} c_1(\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_1^{(j)})$. The logarithm likelihood function is expressed by

$$l_1(\boldsymbol{\beta}_1, \sigma^2_{\text{sum1}}, \sigma_1^{*2}, \gamma_1) \propto -n_1 \log \sigma^2_{\text{sum1}} - \log |R_1^*| + \frac{1}{\sigma^2_{\text{sum1}}} (\boldsymbol{Y}_1^{n_1} - H_1 \boldsymbol{\beta}_1)^{\mathrm{T}} R_1^{*-1} (\boldsymbol{Y}_1^{n_1} - H_1 \boldsymbol{\beta}_1). \quad (4.10)$$

The maximum likelihood estimators of $\boldsymbol{\beta}_1$ and σ^2_{sum1} are given by $\hat{\boldsymbol{\beta}}_1 = (H_1^{\mathrm{T}} R_1^{-1} H_1)^{-1} H_1^{\mathrm{T}} R_1^{-1} \boldsymbol{Y}_1^{n_1}$ and $\hat{\sigma}^2_{\text{sum1}} = \frac{1}{n_1} (\boldsymbol{Y}_1^{n_1} - H_1 \boldsymbol{\beta}_1)^{\mathrm{T}} R_1^{*-1} (\boldsymbol{Y}_1^{n_1} - H_1 \boldsymbol{\beta}_1)$. By finding σ_1^{*2} and γ_1 that maximize the logarithm likelihood function in eq.(4.10), these estimators can be obtained as solutions to the following problem :

$$\{\hat{\sigma_1^*}^2, \hat{\gamma}_1\} = \arg \min_{0<\sigma_1^{*2}<1, \gamma_1>0} \{n_1 \log \hat{\sigma}^2_{\text{sum1}} + \log |R_1^*|\} \quad (4.11)$$

The parameters at the t-th level are $\boldsymbol{\beta}_t$, σ^2_{Yt}, $\sigma^2_{\epsilon t}$, γ_t and ρ_{t-1}, and here, setting $T_t = \boldsymbol{Y}_t^{n_t} - \rho_{t-1} \hat{\boldsymbol{Y}}_{t-1}$, the logarithm likelihood function is expressed by

$$l_t(\boldsymbol{\beta}_t, \sigma^2_{Y_t}, \sigma^2_{\epsilon_t}, \gamma_t, \rho_{t-1}) \propto -\log |R_t| + (T_t - H_t \boldsymbol{\beta}_t)^{\mathrm{T}} {R_t}^{-1} (T_t - H_t \boldsymbol{\beta}_t). \quad (4.12)$$

The maximum likelihood estimate of $\boldsymbol{\beta}_t$ is given by $\hat{\boldsymbol{\beta}}_t = (H_t^{\mathrm{T}} R_t^{-1} H_t)^{-1} H_t^{\mathrm{T}} R_t^{-1} T_t$. By finding $\sigma^2_{Y_t}$, $\sigma^2_{\epsilon_t}$, γ_t and ρ_{t-1} that maximize the logarithm likelihood function in eq.(4.12), these estimates can be determined, by solving the following problem :

$$\{\hat{\sigma}^2_{Y_t}, \hat{\sigma}^2_{\epsilon_t}, \hat{\gamma}_t, \hat{\rho}_{t-1}\} = \arg \min_{\sigma^2_{Y_t}, \sigma^2_{\epsilon_t}, \gamma_t > 0,\ 0 \le \rho_{t-1} \le 1} \{\log |R_t| - (T_t - H_t \hat{\boldsymbol{\beta}}_t)^{\mathrm{T}} {R_t}^{-1} (T_t - H_t \hat{\boldsymbol{\beta}}_t)\}. \quad (4.13)$$

5 Applying Efficient Global Optimization to Data of Varying Precision

5.1 Integrated Mean Square Error for Data Divided into Levels

The methodology for selecting only one sample is given by eq.(2.10). However, the number of additional samples must be greater than the number of parameters at each level, so a sampling criterion to obtain several samples is needed. There are methods to structure a uniform design such as a Latin Hypercube design, minimax and maxmin distance designs, or criteria for the optimal design matrix H in eq.(2.3) may be utilized, such as D-optimality, A-optimality [2]. A uniform design is used to infill and spread sample point evenly throughout the area, and the optimal design has a criterion to minimize the particular variance when the relation between sample points and the response surface is examined. Now, in addition to obtain new samples according to EI, it is also necessary the weight the points in the sample.

The Integrated Mean Squares Error (IMSE) is one approach to optimal design. This method has the criterion of minimizing the MSE with a weighted sampling function at each point [2, 10]. Let the design $\boldsymbol{D}^n$ be the structure of sample points, and set $\boldsymbol{D}^n = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$. The criterion of IMSE is to choose the design $\boldsymbol{D}^n$ which minimizes the integration of MSE, given a weight function of the sample points. For the Kriging model, because $s(\boldsymbol{x})^2$ represents the MSE, the IMSE design is given by

$$\min_{\boldsymbol{D}^n \subset \chi} \text{IMSE} = \min_{\boldsymbol{D}^n \subset \chi} \int_{\chi} s(\boldsymbol{x})^2 w(\boldsymbol{x}) d\boldsymbol{x}, \quad (5.1)$$

where χ is the sampling region, $w(\cdot)$ is a weight function that satisfies $\int_\chi w(\boldsymbol{x})d\boldsymbol{x} = 1$, and also, from eq.(2.7), IMSE is expressed by

$$\text{IMSE} = \sigma_\eta^2 - \text{trace}\left[\begin{pmatrix} O & H(\boldsymbol{D}^n) \\ H(\boldsymbol{D}^n) & R(\boldsymbol{D}^n) \end{pmatrix}^{-1} \int_\chi \begin{pmatrix} \boldsymbol{h}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}} & \boldsymbol{h}(\boldsymbol{x})\boldsymbol{r}(\boldsymbol{x})^{\mathrm{T}} \\ \boldsymbol{r}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}} & \boldsymbol{r}(\boldsymbol{x})\boldsymbol{r}(\boldsymbol{x})^{\mathrm{T}} \end{pmatrix} w(\boldsymbol{x})d\boldsymbol{x}\right]. \quad (5.2)$$

Suppose that, when the higher level sample is added, these points are selected by the IMSE design criterion. The design $\boldsymbol{D}_t^{n_t} = \{\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(n_t)}\}$ at the t-th level is therefore selected to minimize IMSE, when the parameters of IMSE use alternative for estimating at the $(t-1)$-th level and $w(\cdot)$ is given by normalized EI$(\cdot)$ obtained at the $(t-1)$-th level . Let the parameters at the $(t-1)$-th level be $\boldsymbol{\theta}_{t-1}$, and normalized EI at the $(t-1)$-th level be $w_{t-1}(\cdot)$, given by $\text{EI}(\cdot)/\int_\chi \text{EI}(\boldsymbol{x})d\boldsymbol{x}$. Then $\boldsymbol{D}_t^{n_t}$ is determined by solving the following problem :

$$\max_{\boldsymbol{D}_t} \text{trace}\left[\begin{pmatrix} O & H_t \\ H_t & R_t(\boldsymbol{\theta}_{t-1}) \end{pmatrix}^{-1} \int_\chi \begin{pmatrix} \boldsymbol{h}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}} & \boldsymbol{h}(\boldsymbol{x})\boldsymbol{r}_t(\boldsymbol{x};\boldsymbol{\theta}_t)^{\mathrm{T}} \\ \boldsymbol{r}_t(\boldsymbol{x};\boldsymbol{\theta}_{t-1})\boldsymbol{h}(\boldsymbol{x})^{\mathrm{T}} & \boldsymbol{r}_t(\boldsymbol{x};\boldsymbol{\theta}_{t-1})\boldsymbol{r}_t(\boldsymbol{x};\boldsymbol{\theta}_{t-1})^{\mathrm{T}} \end{pmatrix} w_{t-1}(\boldsymbol{x})d\boldsymbol{x}\right] \quad (5.3)$$

We consider the grid points in χ. The maximization problem for the choice of the IMSE design in eq.(5.3) is approximately solved by Simulated Annealing, where the selection of points on χ is a combinatorial optimization problem.

5.2 Proposed Efficient Global Optimization

Here, we consider how to search for a global minimizer. For each level $t = 1, \ldots, s$, suppose that the number of initial and additional data elements is set to $n_t^{(ini)}$ and $n_t^{(add)}$ respectively in advance, such that the number of initial data elements exceeds the number of parameters.

The following shows the algorithm of the proposed EGO.

[First level]

Step 1: Set $j = 0$. Choose an initial design $\boldsymbol{D}_1^{(0)}$ and obtain a set of initial samples $\boldsymbol{Y}_1(\boldsymbol{D}_1^{(0)})$

Step 2: Given $\hat{Y}(\boldsymbol{x})$ and $s(\boldsymbol{x})$ at all points $\boldsymbol{x}$ by eq.(2.6), (2.7) from $\{\boldsymbol{D}_1^{(j)}, \boldsymbol{Y}_1(\boldsymbol{D}_1^{(j)})\}$, respectively. Then, obtain EI$(\boldsymbol{x})$ in eq.(2.9) at all points $\boldsymbol{x}$ from $\hat{Y}(\boldsymbol{x})$ and $s(\boldsymbol{x})$.

Step 3: If $j > n_1^{(add)}$, $t = t+1$ and go to Step 5. Otherwise go to Step 4.

Step 4: Sample at the point where $\boldsymbol{x}_1^{new} = \arg\max\{\text{EI}(\boldsymbol{x})\}$, and obtain a new sample $Y(\boldsymbol{x}_1^{new})$. Set $\{\boldsymbol{D}_1^{(j)}, \boldsymbol{Y}_1(\boldsymbol{D}_1^j)\} = \{\boldsymbol{D}_1^{(j)}, \boldsymbol{Y}_1(\boldsymbol{D}_1^j)\} \cup \{\boldsymbol{x}_1^{new}, Y(\boldsymbol{x}_1^{new})\}$, then go to Step 2.

Step 5: If $t < s$, go to the next level, otherwise stop.

[t-th level ($t \geq 2$)]

Step 1: Set $j = 0$. Choose the initial design $D_t^{(0)}$ that minimizes IMSE in eq.(5.3) from EI at the $(t-1)$-th level, and obtain a set of initial samples $\boldsymbol{Y}_t(\boldsymbol{D}_t^{(0)})$.

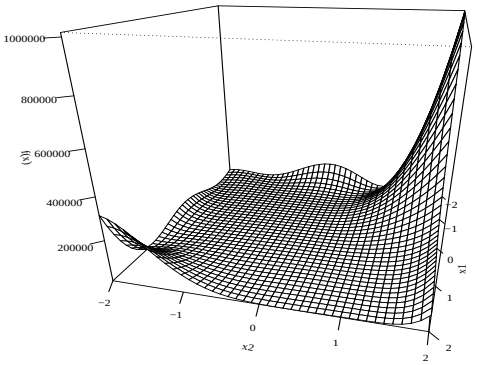

Figure 7: Goldstein-Price's function

Step 2: Obtain $\hat{Y}_t(\boldsymbol{x})$ and $s_t(\boldsymbol{x})$ in eq.(4.8), (4.9) at all points $\boldsymbol{x}$ from $\{\boldsymbol{D}_1^{(j)}, \boldsymbol{Y}_t(\boldsymbol{D}_t^{(j)})\}$, respectively. Then, obtain $\mathrm{EI}(\boldsymbol{x})$ in eq.(2.9) at all points $\boldsymbol{x}$ from $\hat{Y}_t(\boldsymbol{x})$ and $s_t(\boldsymbol{x})$.

Step 3: If $j > n_t^{(add)}$, set next level or stop, otherwise go to Step 4.

Step 4: Sample at the point where $\boldsymbol{x}_t^{new} = \arg\max\{\mathrm{EI}(\boldsymbol{x})\}$, and obtain a new sample $Y(\boldsymbol{x}_t^{new})$. Set $\{\boldsymbol{D}_t^{(j)}, \boldsymbol{Y}_t(\boldsymbol{D}_t^j)\} = \{\boldsymbol{D}_t^{(j)}, \boldsymbol{Y}_t(\boldsymbol{D}_t^j)\} \cup \{\boldsymbol{x}_t^{new}, Y(\boldsymbol{x}_t^{new})\}$, then go to Step 2.

Step 5: If $t < s$, go to the next level, otherwise stop.

6 Numerical Experiments

In this section, we consider some numerical experiments, in which we used the proposed method with different precision data to search for an optimal solution, and the precision of the solution and the efficiency of sampling was evaluated from the results.

Goldstein-Price's function [9] was used as a test function, expressed by

$$\begin{aligned} f(x_1, x_2) = \{1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2)\} \\ \times\{30 + (2x_1 - 3x_2)^2(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)\}, \\ (-2 \le x_1 \le 2, -2 \le x_2 \le 2), \quad (6.1) \end{aligned}$$

where a global minimizer is $\boldsymbol{x}^* = (x_1, x_2) = (0, -1)$ and a global minimum is $f(\boldsymbol{x}^*) = 3$. Goldstein-Price's function is shown in Figure 7.

This experiment treated 41×41 points on the grid $\chi = [-2, -1.9, \ldots, 2] \times [-2, -1.9, \ldots, 2]$, and data of two different levels of precision, where lower precision data elements $\boldsymbol{Y}_1$ belonged to the first level and higher precision data elements $\boldsymbol{Y}_2$ belonged to the second level. Suppose that these have error variances $\sigma_{\epsilon 1}^2 = 0.4^2$, and $\sigma_{\epsilon 2}^2 = 0.2^2$ respectively; however both variances were unknown for the experiments. The regression function was taken as constant, which is expressed by $\boldsymbol{h}(x_1, x_2) = 1$. The initial and additional number of samples at the first and second level, $n_1^{(ini)}$, $n_1^{(add)}$, $n_2^{(ini)}$ and $n_2^{(add)}$ were decided in advance. The minimization problems for the estimation of parameters in eq.(4.11), (4.13) were numerically solved using the Nelder-Mead method.

The initial design was structured by points chosen by random sampling at the first level. For each sampling strategy, numerical experiments were carried out using 50 initial designs

Table 1: Evaluation of $\boldsymbol{x}_{\min}$ and $y(\boldsymbol{x}_{\min})$ with data of varying precision

$n_1^{(ini)}$	$n_1^{(add)}$	$n_2^{(ini)}$	$n_2^{(add)}$	$\mathrm{E}\{d(\boldsymbol{x}_{\min})\}$	$\mathrm{Var}\{d(\boldsymbol{x}_{\min})\}$	$\mathrm{E}\{d(y(\boldsymbol{x}_{\min}))\}$	$\mathrm{Var}\{d(y(\boldsymbol{x}_{\min}))\}$
30	10	8	2	0.360	0.085	25.98	378.2
20	20	8	2	0.191	0.155	29.85	796.2
20	10	8	7	0.413	0.189	35.58	1113
20	10	10	5	0.443	0.191	41.30	2789
20	0	10	10	0.677	0.460	56.93	2826
15	5	10	10	0.346	0.052	34.54	1080
10	10	10	10	0.454	0.109	47.10	2412

Table 2: Evaluation of $\boldsymbol{x}_{\min}$ and $y(\boldsymbol{x}_{\min})$ with data of a single precision

$n_1^{(ini)}$	$n_1^{(add)}$	$n_2^{(ini)}$	$n_2^{(add)}$	$\mathrm{E}\{d(\boldsymbol{x}_{\min})\}$	$\mathrm{Var}\{d(\boldsymbol{x}_{\min})\}$	$\mathrm{E}\{d(y(\boldsymbol{x}_{\min}))\}$	$\mathrm{Var}\{d(y(\boldsymbol{x}_{\min}))\}$
20	40	-	-	0.217	0.039	14.64	188.3
30	30	-	-	0.261	0.064	14.76	143.8
40	20	-	-	0.365	0.185	27.58	571.1
-	-	10	20	0.566	0.319	59.03	4801
-	-	15	15	0.546	0.285	52.65	2737
-	-	20	10	0.565	0.237	53.45	1483

chosen by random sampling. The minimizer $\boldsymbol{x}_{\min}$ of samples obtained during searching is given by

$$\boldsymbol{x}_{\min} = \arg\min\{y(\boldsymbol{x}_1^{(1)}), \ldots, y(\boldsymbol{x}_1^{(n_1)}), y(\boldsymbol{x}_2^{(1)}), \ldots, y_1(\boldsymbol{x}_2^{(n_2)})\}, \tag{6.2}$$

and then the evaluation used the Euclidean distance between the true global minimizer $\boldsymbol{x}^*$ and $\boldsymbol{x}_{\min}$, which is expressed by

$$d(\boldsymbol{x}_{\min}) = \sqrt{\textstyle\sum_{j=1}^{2}({x^*}_j - x_{\min j})^2}, \tag{6.3}$$

where ${x^*}_j$ and $x_{\min j}$ were the coordinates of $\boldsymbol{x}^*$ and $\boldsymbol{x}_{\min}$ in dimension j. The expected value $\mathrm{E}\{d(\boldsymbol{x}_{\min})\}$ and variance $\mathrm{Var}\{d(\boldsymbol{x}_{\min})\}$ of each sampling strategy using the different precision data are shown in Table 1. Also, the minimum sample value $y(x_{\min j})$ is evaluated. The Euclidean distance between the global minimum $f(\boldsymbol{x}^*)$ and $y(\boldsymbol{x}_{\min})$ is expressed by

$$d(y(\boldsymbol{x}_{\min})) = |f(\boldsymbol{x}^*) - y(\boldsymbol{x}_{\min})|, \tag{6.4}$$

The expected value $\mathrm{E}\{d(y(\boldsymbol{x}_{\min}))\}$ and variance $\mathrm{Var}\{d(y(\boldsymbol{x}_{\min}))\}$ are shown in Table 1.

For comparison, searches for a global optimum using only single precision data were made. The number of initial and additional data elements were set in advance, as was the case for the search using data of varying precision. The results for each sampling strategy using low and high precision data are shown in Table 2.

Table 2 shows that the precision of the search using only high precision data was low; on the other hand, the precision of the search using only low precision data was high, but many samples were required. In particular, the precision of the search was high when the number of additional data elements was large. The searches using data of varying precision were of greater precision than those using only high precision data. If it is difficult to obtain a lot of high precision data, the precision of the search can be made greater by combining

low precision data. The precision of the search with $n_1^{(ini)} = 20$, $n_1^{(add)} = 20$, $n_2^{(ini)} = 8$ and $n_2^{(add)} = 2$ was greater than the precision of those using only low precision data. As was the case for data of a single precision, the search was of high precision when the number of additional data elements was large.

This paper proposes one approach to combine the varying precision samples, and shows the possibility of the improvement of searching efficiency. We have applied for some test functions. However, we did not mention in the manuscript, because the results are almost duplicates of Goldstein-Price's function. Our approach is based on the sampling rather than optimization, where the procedure of finding the optimum point may be a sequence of sampling point which maximizes $EI(x)$. The optimum point is given by taking the minimum from the searching points as eq.(6.2).

7 Conclusion

This paper deals with the search for a global optimum considering efficiency of sampling, and for this purpose, a method utilizing EGO and combining data of varying precision is proposed. An auto-regression model with a Kriging model is applied between respective levels where data elements are classified into levels according to their precision. Since not only a predicted value can be obtained but also the accuracy of the predictor can be estimated, the search can be executed with EGO applied to data of varying precision. Numerical experiments showed that the proposed method is effective, resulting in higher efficiency and precision of an obtained solution than was possible using only data of a single precision.

There are some further problems to consider. In this paper, selection of initial data at each level was performed by the IMSE; however data elements of varying precision were not considered. Sampling points would have to be changed to take account of data of varying precision.

When we consider efficiency of sampling, it is necessary to consider practical consequences. In the case that there are constraints of time or cost, or constraints on sampling points, we must consider how to treat the problem including the constraints.

References

[1] N. Cressie, Kriging nonstationary data, *J. Amer. Statist. Assoc.* 81 (1986) 625–634.

[2] K.T. Fang, R. Li and A. Sudjiants, *Design and Modeling for Computer Experiments*, Chapman & Hall/CRC, New York, 2006.

[3] D. Hirst, G. Storvik and A.R. Syversveen, A hierarchical modelling approach to combining environmental data at different scales, *J. Roy. Statist. Soc. Ser. C* 52 (2003) 377–390.

[4] D. Huang, T.T. Allen, W.I. Notz and N. Zeng, Global optimization of stochastic black-box systems via sequential kriging meta-models, *J. Global Optim.* 34 (2006) 441–466.

[5] D.R. Jones, M. Schonlau and W.J. Welch, Efficient global optimization of expensive black-box functions, *J. Global Optim.* 13 (2000) 455–492.

[6] D.R. Jones, A taxonomy of global optimization methods based on response surfaces, *J. Global Optim.* 21 (2001) 345–383.

[7] M.C. Kennedy and A. O'Hagan, Predicting the output from a complex computer code when fast approximations are available, *Biometrika* 87 (2000) 1–13.

[8] M.C. Kennedy and A. O'Hagan, Bayesian calibration of computer models, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 63 (2001) 425–464.

[9] A.A. Goldstein and I.F. Price, On descent from local minima, *Math. Comput.* 25 (1971) 569–574.

[10] J. Sacks, W.J. Welch, T.J. Mitchell and H.P. Wynn, Design and analysis of computer experiments, *Statist. Sci.* 4 (2001) 409–435.

[11] T.J. Santner, B.J. Williams and W.I. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, 2003.

[12] A.Sóbester, S. Leary and A. Keane, On the design of optimization strategies based on global response surface approximation models, *J. Global Optim.* 33 (2005) 31–59.

Manuscript received 31 March 2008
revised 22 September 2008, 21 January 2009
accepted for publication 24 February 2009

Koji Tsuchiya
Graduate School of Information Science and Technology, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871 Japan
E-mail address: koji.tsuchiya@ist.osaka-u.ac.jp

Hiroshi Morita
Graduate School of Information Science and Technology, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871 Japan
E-mail address: morita@ist.osaka-u.ac.jp