

NEW BAYESIAN METHOD FOR MULTIEXTREMAL OPTIMIZATION

Natalija Pozniak¹, Leonidas Sakalauskas²

Vilniaus kolegija / University of Applied Sciences, J. Jasinskio. 15, 01111 Vilnius, Lithuania¹

Klaipėda University, H. Manto str. 84, Klaipėda Lithuania²

E-mail: Natalija.Pozniak@gmail.com¹; Leonidas.Sakalauskas@mif.vu.lt²

Abstract. This paper is focused on the Bayes approach to multiextremal optimization problems, based on modelling the objective function by Gaussian random field (GRF) and using the Euclidean distance matrices with fractional degrees for presenting GRF covariances. A recursive optimization algorithm has been developed aimed at maximizing the expected improvement of the objective function at each step, using the results of the optimization steps already performed. Conditional mean and conditional variance expressions, derived by modelling GRF with covariances expressed by fractional Euclidean distance matrices, are used to calculate the expected improvement in the objective function. The efficiency of the developed algorithm was investigated by computer modelling, solving the test tasks, and comparing the developed algorithm with the known heuristic multi-extremal optimization algorithms.

Keywords: Bayesian optimization, Gaussian random fields.

1. Introduction

The Bayes approach provides the "black box" strategy for solving multiextremal problems. Thus, there is no necessity to know derivatives of the function to be optimized. This approach is created by J. Mockus and co-authors (1989). Since the values of the objective functions are unknown and only anticipated, the Bayesian search strategy models these values as realizations of the Gaussian field, using conditional information about the function from the previous iterations for planning the next optimization step. When the new function value is computed, the conditional information is updated. The process is terminated according to the prescribed termination rule. The efficiency of optimization is evaluated using some criteria. The often used criterion is the expected value of the objective function increment during optimization. Of course, the performance of the Bayes method depends on computer speed and memory resources. The Bayes approach is preferred for application, if computation of the response function is related with certain costs.

2. Bayesian optimization method

Let the continuous objective function be minimized on some compact feasible set. The Bayes optimization method is usually built assuming that the values of the objective function are realization of some random field. Note that the objective function $f(x, \omega)$ depends on a random scenario following from some probability space: $\omega \in (\Omega, \Sigma, P)$. Thus, the optimization problem to be solved is minimized:

$$f(x, \omega) \rightarrow \min_{x \in D},$$

where $D \in \mathfrak{R}^N$ is a feasible set. Assume that the objective function is continuous and the feasible set is compact.

Let us consider the iterative optimization way, where each step consists of choice of an optimization point and calculation of the function value at this point.

After performance of k steps, let the sequence of optimization points be known:

$$X^k = (x^1, x^2, x^3, \dots, x^k), \quad (1)$$

where the function values are computed:

$$Y^k = (y^1, y^2, y^3, \dots, y^k), \quad (2.2)$$

which, in turn, are realizations of the random field:

$$y^k = f(x^k) = f(x^k, \omega), \quad x^k \in D \quad (2.3)$$

For the sake of simplicity, here and hereinafter the notation of scenario ω is omitted. It is clear that

$$X^k = (X^{k-1}, x^k), \quad (2.4)$$

$$Y^k = (Y^{k-1}, y^k) \quad (2.5)$$

Hence, each point x^k of the k^{th} optimization step is a function of the results from the previous iterations:

$$x^{k+1} = b_{k+1}(X^k, Y^k). \quad (2.6)$$

The search of optimum is terminated according to the prescribed termination rule. The simplest rule is to terminate the optimization after some given number of steps K .

The sequence $b = (b_1, b_2(X^1, Y^1), \dots, b_K(X^{K-1}, Y^{K-1}))$, determining the dependence of each subsequent step (3.3) on the results of the performed steps is called an optimization method.

Usually, the random field, given by the distribution of values of field realization is defined at any sequence of observation points. Thus, let a sequence of random field realizations after k steps be described by the density function

$$p_X(Y). \quad (2.7)$$

The cost function $\Psi(f, b)$ must be selected for the evaluation of quality of the optimization process. In general, this function depends on the results of optimization steps:

$$\Psi(f, b) = \Psi(Y^K; X^K) = \Psi(Y^{k-1}, y^k; X^{k-1}, x^k). \quad (2.8)$$

Let the efficiency function satisfy the following conditions:

A. The expected value exists: $|E\Psi(f, b)| < \infty$.

B. The function $\Psi(Y^{k-1}, y^k; X^{k-1}, x^k)$ is concave with respect to the variables y^k :

$$\Psi(Y^{K-1}, y_1^k \cdot \alpha + y_2^k \cdot (1-\alpha); X^{k-1}, x^k) \geq \alpha \cdot \Psi(Y^{K-1}, y_1^k; X^{k-1}, x^k) + (1-\alpha) \cdot \Psi(Y^{K-1}, y_2^k; X^{k-1}, x^k), \quad \forall \alpha, 0 \leq \alpha \leq 1.$$

C. The function $\Psi(f, b)$ satisfies the Lipschitz condition:

$$|\Psi(Y^{K-1}, y_1^k; X^{k-1}, x^k) - \Psi(Y^{K-1}, y_2^k; X^{k-1}, x^k)| \leq \|y_1^k - y_2^k\|.$$

D. The function $\Psi(f, b)$ is monotonically increasing:

$$\Psi(Y^{K-1}, y_1^k; X^{k-1}, x^k) \leq \Psi(Y^{K-1}, y_2^k; X^{k-1}, x^k), \quad y_1^k \leq y_2^k.$$

The examples of efficiency functions are as follows:

$$\Psi(f, b) = \min(y^1, y^2, \dots, y^K), \quad (2.9)$$

$$\Psi(f, b) = \min(0, \min(y^1, y^2, \dots, y^{K-1}) - y^K), \quad (2.10)$$

$$\Psi(f, b) = (\min(0, \min(y^1, y^2, \dots, y^{K-1}) - y^K))^p, \quad p > 0.$$

The efficiency functions are related with loss functions (Mockus, 2013), for instance,

$$L(f, b) = \min(y^1, y^2, \dots, y^K) - \min_{x \in D} f(x).$$

The Bayesian optimization method must be constructed so that the expected value of the efficiency function were minimized:

$$E\Psi(f, b) \rightarrow \min_{X^K}. \quad (2.11)$$

Let us prove that the Bayesian method satisfies the recursive Bellman equation system. Denote the conditional probability density functions:

$$p_{X^k}(y^k | Y^{k-1}) = \frac{p_{X^k}(Y^k)}{p_{X^{k-1}}(Y^{k-1})} = \frac{p_{X^k}(Y^{k-1}, y^k)}{p_{X^{k-1}}(Y^{k-1})}, \quad k = 2, 3, \dots, K \quad (2.12)$$

Denote:

$$U_K(Y^K; X^K) = \Psi(Y^{K-1}, y^K; X^{K-1}, x^K), \quad (2.13)$$

$$U_k(Y^k; X^k, x^{k+1}) = E(U_{k+1}(Y^k, y^{k+1}; X^k, x^{k+1}) | Y^k) = \int_{y^{k+1}} U_{k+1}(Y^k, y^{k+1}; X^k, x^{k+1}) \cdot p_{X^{k+1}, x^k}(y^{k+1} | Y^k) dy^{k+1} \quad (2.14)$$

where $U_k(Y^k; X^k, x^{k+1})$ is the expected value of the function $U_{k+1}(Y^k, y^{k+1}; X^k, x^{k+1})$ at the point x^{k+1} with the condition, that the results of the previous k steps are Y^k , $k = 1, 2, \dots, K-1$.

Theorem 1. Let the utility function satisfy conditions A-D. Then the Bayesian optimization method satisfies the system of recursive equations:

$$U_{K-1}(Y^{K-1}) = \min_{x^K \in D} E_{x^K}(\Psi(Y^{K-1}, y^K) | Y^{K-1}), \quad (2.15)$$

$$U_k(Y^k) = \min_{x^{k+1} \in D} E_{x^{k+1}}(U_{k+1}(Y^k, y^{k+1}) | Y^k), k = 1, 2, \dots, K-1 \quad (2.16)$$

Proof. Using the properties of multiple conditional distributions, we can obtain:

$$\begin{aligned} E\Psi(f, b) &= E\Psi(Y^K) = \\ &= \int_{Y^K \in \mathfrak{Y}^K} \Psi(Y^K) \cdot p_{X^K}(Y^K) dY^K = \\ &= \int_{Y^{K-1} \in \mathfrak{Y}^{K-1}} \left(\int_{y^K \in \mathfrak{Y}} \Psi(Y^{K-1}, y^K) \cdot p_{X^K}(y^K | Y^{K-1}) dy^K \right) \cdot p_{X^{K-1}}(Y^{K-1}) dY^{K-1} = \\ &= \int_{Y^{K-1} \in \mathfrak{Y}^{K-1}} E_{x^K}(\Psi(Y^{K-1}, y^K) | Y^{K-1}) \cdot p_{X^{K-1}}(Y^{K-1}) dY^{K-1} = \\ &\quad \dots \\ &= E_{x^1} \left(E_{x^2} \left(\dots E_{x^K}(\Psi(Y^{K-1}, y^K) | Y^{K-1}) \dots | Y^2 \right) y^1 \right). \end{aligned} \quad (2.17)$$

We get that the optimal method in the last step minimizes this integral:

$$x_{opt}^K = \arg \min_{x^K \in D} \int_{y^K \in \mathfrak{Y}} \Psi(Y^{K-1}, y^K) \cdot p_{X^K}(y^K | Y^{K-1}) dy^K. \quad (2.18)$$

Similarly, we obtain that the $k+1$ step point has to satisfy the condition:

$$x_{opt}^{k+1} = \arg \min_{x^{k+1} \in D} \int_{y^{k+1}} U_{k+1}(Y^k, y^{k+1}) \cdot p_{X^{k+1}, x^k}(y^{k+1} | Y^k) dy^{k+1}. \quad (2.19)$$

where $k = 0, 1, \dots, K-2$. The theorem statement follows by induction.

Corollary 1. The function $U_k(Y^k)$, $k = 1, \dots, K$ satisfies conditions A and D.

Actually, boundedness of the function $U_k(Y^k)$ comes from condition A and expression (2.17). The monotone decreasing and concavity of functions follows monotony and concavity conditions of function Ψ .

In general, the solution to this equation is rather a complicated task. Therefore, let us look at the simplified approximate solution to this system. Let us consider utility function (2.9) and use the approximation:

$$\tilde{U}_k(X^k, Y^k) = \min(y^k, c_k). \quad (2.20)$$

Note that the approximation proposed satisfies conditions A-D.

In this case, the design of the Bayesian method splits into K independent optimization tasks:

$$x_{opt}^{k+1} = \arg_{x^{k+1}} \int_{y^{k+1} \in \mathfrak{R}} \min(Y^k, y^{k+1}, c_{k+1}) p(y^{k+1} | Y^k, X^k) dy^{k+1}. \quad (2.21)$$

3. Modelling by using Gaussian fields

Assume the values of the response function are realization of the Gaussian random field. Usually, GRF is defined using a correlation function in accordance with Bochner's theorem (Abrahamsen (1997)). In this paper, a model is considered in which Gaussian field correlations are modelled at Euclidean distances with fractional degrees between the points in which the objective function is calculated.

Let us consider a data set $X = (x_1, x_2, \dots, x_K)$ consisting of different points, $x_i, x_j \in \mathfrak{R}^d$, $x_i \neq x_j$, $i \neq j$, $1 \leq i, j \leq K$ where the values of the response function $Y = (y_1, y_2, \dots, y_K)^T$ are known, obtained by physical measurements, computer simulation, etc. Denote $K \times K$ matrix of fractional degrees δ of Euclidean distance squares among pairs of vectors by $A = \left[\left((x_i - x_j)^T \cdot (x_i - x_j) \right)^\delta \right]_1^K$, where $x_i \in X$, $1 \leq i, j \leq K$, $0 \leq \delta \leq 1$. The properties of fractional Euclidean distance matrices (FEDM) are studied by means of the kernel matrix, defined as follows (see (Schoenberg (1935), Gower (1984), Weinberger (2004), etc.):

$$F = -(I - E \cdot s^T) \cdot A \cdot (I - s \cdot E^T), \quad (3.1)$$

where $s \in \mathfrak{R}^K$, $s^T \cdot E = 1$, E is a vector from units and I is a unit matrix.

Geometrical properties of FEDM (see, Pozniak&Sakalauskas, JMD, 2017, Pozniak&Sakalauskas, 2019) enable us to construct a homogeneous GRF with the mean vector

$$EZ(x, \omega) = \mu \cdot E, \quad (3.2)$$

and a covariance matrix, proportional to the respective kernel matrix

$$\text{cov}(Z(x, \omega)) = \beta^2 \cdot F, \quad (3.3)$$

Assume that only k values of the objective function are calculated and using this information the next point of optimization x^{k+1} should be chosen. Denote the matrix of fractional degrees of Euclidean distances between points of observations as follows:

$$A = \left[\left((x^i - x^j)^T \cdot (x^i - x^j) \right)^\delta \right]_1^k. \quad (3.4)$$

4. One-Step Bayesian Optimization Techniques

Solving of the equation is a very difficult task, so let us consider the one-step Bayesian optimization method. In each step of this method the optimization point is selected in order to solve the problem:

$$x^{k+1} = \arg \min_{x^{k+1} \in D} E_{x^{k+1}} \left(\min(y^{k+1} - \min(Y^k), 0) | Y^k \right), \quad k = 1, 2, \dots, K-1 \quad (4.1)$$

where the conditional mean and conditional variance are calculated under the condition that the function value vector (2.2) is calculated at the points (2.1):

$$y_k(x^{k+1}) = Y_k^T \cdot A_k^{-1} \cdot \left(a^{k+1} + E_k \cdot \frac{(1 - E_k^T \cdot A_k^{-1} \cdot a^{k+1})}{E_k^T \cdot A_k^{-1} \cdot E_k} \right) \quad (4.2)$$

$$\beta_k^2(x^{k+1}) = a_{k+1}^T \cdot A_k^{-1} \cdot a^{k+1} - \frac{(1 - E_k^T \cdot A_k^{-1} \cdot a^{k+1})^2}{E_k^T \cdot A_k^{-1} \cdot E_k}, \quad (4.3)$$

where $a^{k+1} = \left(|x_1 - x^{k+1}|^\delta, |x_2 - x^{k+1}|^\delta, \dots, |x_k - x^{k+1}|^\delta \right)^T$, A_k are vectors X^k FEDM (3.4).

Using (4.2), (4.3), it is possible to write the conditional density of the objective function at the point x :

$$p_x(y|Y) = \frac{e^{-\frac{(y - y_k(x))^2}{2 \cdot \beta_k^2 \cdot s_k^2(x)}}}{\sqrt{2\pi} \cdot \beta_k \cdot s_k(x)}. \quad (4.4)$$

After some manipulations, one can make sure that

$$\begin{aligned} E_{x^{k+1}} \left(\min(y^{k+1} - \min(Y^k), 0) | Y^k \right) &= \int_{-\infty}^{\infty} \min(z - Y_{\min}^k, 0) \cdot \frac{e^{-\frac{(z - y(x^{k+1}))^2}{2 \cdot \beta^2 \cdot s^2(x^{k+1})}}}{\sqrt{2\pi} \cdot \beta \cdot s(x^{k+1})} dz = \\ &= \int_{-\infty}^{Y_{\min}^k} (z - Y_{\min}^k) \cdot \frac{e^{-\frac{(z - y(x^{k+1}))^2}{2 \cdot \beta^2 \cdot s^2(x^{k+1})}}}{\sqrt{2\pi} \cdot \beta \cdot s(x^{k+1})} dz = R(x^{k+1}). \end{aligned} \quad (4.5)$$

The latter expression can be written in the following forms:

$$\begin{aligned} R(x) &= \left(y(x) - Y_{\min}^K \right) \cdot \int_{-\infty}^{\frac{y(x) - Y_{\min}^K}{\beta \cdot s(x)}} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz - \beta \cdot s(x) \cdot e^{-\frac{\frac{Y_{\min}^K - y(x)}{\beta \cdot s(x)}}{2 \cdot \beta^2 \cdot s^2(x)}} = \\ &= \left(y(x) - Y_{\min}^K \right) \cdot \Phi \left(\frac{y(x) - Y_{\min}^K}{\beta \cdot s(x)} \right) - \beta \cdot s(x) \cdot e^{-\frac{\frac{Y_{\min}^K - y(x)}{\beta \cdot s(x)}}{2 \cdot \beta^2 \cdot s^2(x)}} \end{aligned} \quad (4.6)$$

$$R(x) = \int_{-\infty}^{\frac{Y^K - y(x)}{\beta \cdot s(x)}} \int_{-\infty}^t \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz dt = \int_{-\infty}^{\frac{Y^K - y(x)}{\beta \cdot s(x)}} \Phi(z) dz \quad (4.7)$$

Assume

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^x e^{-\frac{z^2}{2}} dz. \quad (4.8)$$

$$(4.9)$$

Corollary 2. Constants c_k satisfy the conditions:

$$c_1 = 0, c_{k+1} = c_k \cdot (1 - \Phi(c_k)) - \frac{e^{-\frac{c_k^2}{2}}}{\sqrt{2\pi}}, \quad k = 0, 1, \dots$$

Note that $\lim_{k \rightarrow \infty} c_k \rightarrow -\infty$, $\lim_{k \rightarrow \infty} c_{k+1} - c_k \rightarrow 0$

Fig. 1 depicts: surfaces of the conditional mean, conditional variance and the expected utility function, using $K = 5$ values of the test function computed at randomly distributed points.

$$G(x, y) = \sqrt{(x-6)^2 + (y-6)^2} \quad (4.10)$$

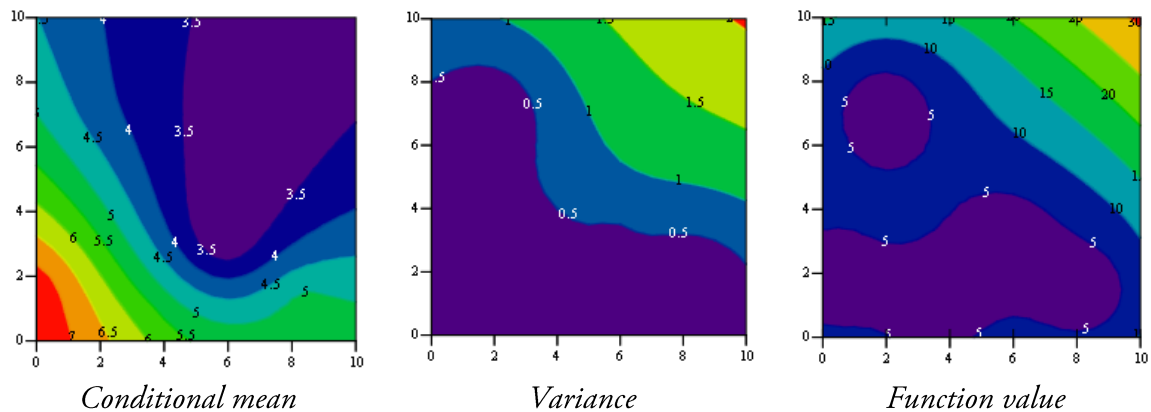


Figure 1. Surfaces of the conditional mean, conditional variance, and the expected utility function

5. Bayesian optimization algorithm

The main calculation time is used to find the FEDM inverse matrix. In general, the calculation of this matrix requires a cubic sequence of operations depending on the size of the matrix.

In calculating the conditional mean of a random field and variance, recursive FEDM inverting formulas can be applied:

$$\begin{bmatrix} A & a \\ a^T & 0 \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} - \frac{A^{-1} \cdot a \cdot a^T \cdot A^{-1}}{a^T \cdot A^{-1} \cdot a} & \frac{A^{-1} \cdot a}{a^T \cdot A^{-1} \cdot a} \\ \frac{a^T \cdot A^{-1}}{a^T \cdot A^{-1} \cdot a} & -\frac{1}{a^T \cdot A^{-1} \cdot a} \end{bmatrix} \quad (5.1)$$

The complexity of this algorithm is quadratic, and thus its application makes it possible to more effectively realize the Bayesian algorithm.

Algorithm:

- Step 1. Random selection of initial points. Attach the initial points to the optimization set.
- Step 2. FEDM inversion of the optimization set.
- Step 3. Calculating the conditional mean and variance of GRF.
- Step 4. Optimization of the expected utility function by the Monte Carlo method.
- Step 5. Optimization of the expected utility function by the local landing method.
- Step 6. FEDM inversion using inversion formulas of block matrices.
- Step 7. If the termination condition is satisfied, the Bayesian method is terminated, in the other cases, add the point of maximum expected utility to the optimization set and repeat step 3.

6. Computer simulation results

The effectiveness of the developed Bayes algorithms was analysed by computer modelling, using the Monte Carlo method. The first test function (4.10), as well as Branin and Rastrigin functions were investigated. Figs. 2-3 depicted the average convergence curve found by the Bayesian method (BM) for optimizing test functions. The convergence curves, presented for comparison, are obtained using the simple Monte Carlo method (MCM) and Simulated Annealing method (SAM).

The computer simulation results show that the Bayesian method is more effective than the known other methods for solving multidimensional problems.

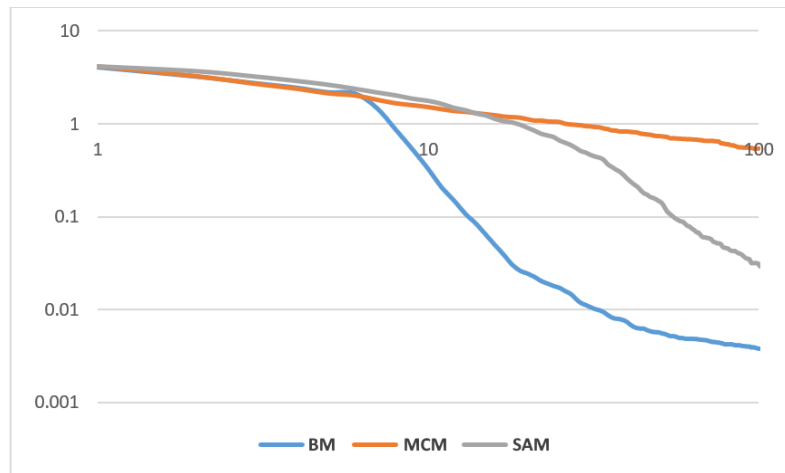


Figure 2. The result of test function (4.10)

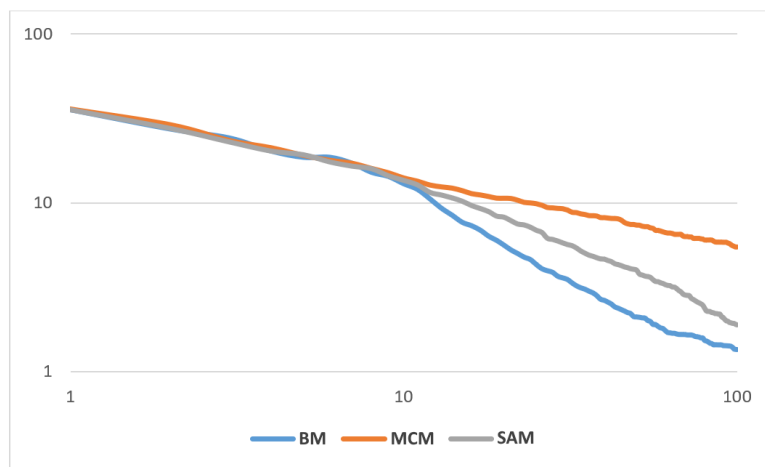


Figure 3. The result of Rastrigin test function

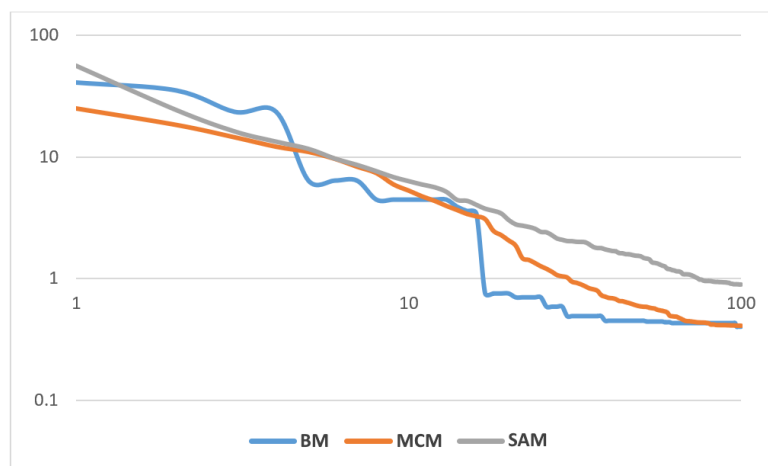


Figure 4. The result of Branin test function

7. Conclusions

The paper is focused on the Bayess approach to global optimization problems, based on modelling the objective function by the Gaussian random field (GRF) and using the Euclidean distance matrices with fractional degrees for presenting GRF covariances. A recursive objective function optimization algorithm has been developed, aimed at maximizing the expected improvement of the objective function at each step, using the results of the optimization steps already performed. The efficiency of the developed algorithm was investigated by computer modelling, solving test problems and comparing the developed algorithm with the known heuristic multi-extremal optimization algorithms. The results of computer modelling have shown that the developed method is more efficient than the other known methods of global optimization. However, the algorithm developed can be refined in future research to calculate the conditional mean and variance, using only a fraction of the results of the experiments performed, thus allowing a reduction in computer time.

References

- Abrahamsen P., 1997, A Review of Gaussian random fields and correlation functions, *Norsk Regnesentral/Norwegian Computing Center*, p. 64.
- Gower J. C., 1984, Multivariate analysis: Ordination, multidimensional scaling and allied topics, *Handbook of applicable mathematics Vol. VI: Statistics. (E. H. Lloyd, Ed.), J. Wiley and Sons, Chichester*, P. 727-781.
- Mockus J., 1989, Bayesian approach to global optimization, *Mathematics and its Applications, Springer Netherlands, Kluwer Academic Publishers*.
- Pozniak N., Sakalauskas L., Saltyte L., 2019, Kriging Model with Fractional Euclidean Distance Matrices, *Informatica*, vol. 30, nr. 2, p. 367-390.
- Pozniak, N., Sakalauskas, L., 2017, Fractional Euclidean distance matrices extrapolator for scattered data. *Journal of Young Scientists*. 47(2), 56–61.
- Schoenberg I. J., 1935, Remarks to Maurice Frechet's article: Sur la definition axiomatique d'une classe d'espaces vectoriels distances applicables vectoriellement sur l'espace de Hilbert," *Ann. Math.*, Vol. 36, P. 724-732.
- Weinberger K.Q, Sha F., Saul L.K., 2004, Learning a Kernel Matrix for Nonlinear Dimensionality Reduction, *Proceedings of the twenty-first international conference on Machine learning (ICML '04)*, P. 106-113.

Natalija Pozniak earned Ph.D. in informatics from Vilnius University in 2019. She has been working as a lecturer at the Vilnius University of Applied Sciences since 2015

Leonidas Sakalauskas, Prof. Habil. Dr., is a principal researcher at Vilnius University Institute of Data Science and Digital Technologies. He is an associate professor at Vilnius Gediminas Technical University.

NAUJAS DAUGIAEKSTREMALINIŲ FUNKCIJŲ BAJESO OPTIMIZAVIMO METODAS

Natalija Pozniak, Leonidas Sakalauskas

Santrauka

Šiame straipsnyje yra nagrinėjamas daugiaekstremalinių funkcijų Bajeso optimizavimo metodas, kuriame kaip optimizuojamos funkcijos modelis yra pasirinktas atsitiktinis Gauso laukas (AGL) su kovariacijų matrica, išreiškiama per trupmeninių Euklido atstumų matricos (TEAM) branduolio matricą. Sukurtas rekursinis optimizavimo algoritmas, kurio tikslas - padidinti tikėtiną tikslo funkcijos pagerėjimą kiekviename žingsnyje, pasinaudojant jau atliktų optimizavimo žingsnių rezultatais. Tikėtino tikslo funkcijos pagerėjimo apskaičiavimui buvo pritaikytos sąlyginio vidurkio ir sąlyginės dispersijos išraiškos, gautos modeliuojant AGL su kovariacijų matricomis, išreikštomis per TEAM. Šio algoritmo efektyvumas buvo tiriamas kompiuterinio modeliavimo būdu. Taip pat, sukurtas algoritmas buvo palygintas su žinomais euristiniais daugiaekstremalio optimizavimo algoritmais.