Including the Derivative Information into Statistical Models Used in Global Optimization

Antanas Žilinskas[1]

[1] *Vilnius University, Institute of Data Science and Digital Technologies, Akademijos 4, LT-08663 Vilnius, Lithuania*

antanas.zilinskas@mii.vu.lt

Abstract. Forecasting of unknown values of the objective function in question is needed to search rationally for the global minimum. Statistical models are used to define the probabilistic characteristics of an unknown value with respect to the values computed at the previous iterations. A new idea, discussed in this paper, is to include the known derivatives into information aimed at forecasting.

INTRODUCTION

The global optimization approach based on statistical models is aimed at the so called ”black box” problems with ”expensive” objective functions. The term ”black box” means the scarcity of information about the properties of the objective functions in question. To represent the uncertainty in a problem, frequently statistical/probabilistic models are most adequate. Correspondingly, the Gaussian random fields are normally applied as the statistical models of ”black box” functions to aid the development of global optimization algorithms. The assumptions of ”expensiveness” substantiates the aspiration to develop the algorithms which satisfy the rationality/optimality postulates [1, 2]. Such algorithms require considerable computing resources. To extend the field of applicability over ”expensive” problems, the computational burden should be reduced. This can be done by implementing the algorithm according to the mode of partition of the feasible region where a simplified statistical model is used to substantiate the selection of an appropriate subregion for the subdivision. The simplicial and rectangular partition is used, e.g. in the algorithm [3] and [4, 5] correspondingly. The partition mode is similarly used in Lipschitz optimization [6, 7, 8, 9, 10, 11].

To our best knowledge, the statistical model based algorithms, including the partition aided versions, are derivative free. It can be explained either by unavailability or ”expensiveness” of derivatives. However, for the prospective Infinity Computer [12, 13], the computation of derivatives as supplement to the function value requires almost no extra time. Therefore, the development of the versions with derivatives seems an apropos challenge. The Infinity Computer-oriented algorithms for different numeric problems are presented e.g. in [14, 15, 16].

Gaussian random fields normally are used as the statistical models of objective functions. The stationarity of fields is also normally assumed. Theoretically, in such models gradients of an objective function can be easily taken into account [17]. However, the involvement of gradients would essentially increase the complexity of the original Bayesian and P-algorithm which is itself restrictively high [18]. In case the algorithm would be implemented in the partition mode, the increase of complexity implied by the involvement of gradients could be avoided. At the conference we will present a statistical model based algorithm which is implemented in the rectangular partition mode. A sub-rectangle for the partition is selected using information about the function value and gradient at the center of the sub-rectangle. The selection criterion is defined in terms of the chosen statistical model. Because of length restrictions of the paper, we present here only the main ideas, and results in full will be presented at the conference and in the subsequent publication.

THE BASIC ALGORITHM

The global optimization problem $\min_{x \in A} f(x)$, where $A = \prod_{i=1}^{d}[a_i, b_i] \subset \mathbb{R}^d$ and $f(x)$ is considered, which is supposed "black box" type. The algorithm is implemented in the rectangular partition mode. The feasible region is subsequently partitioned into the sub-rectangles A_i, $\bigcup_{i=1}^{n} A_i = A$ where A_i are closed sets, and $mess(A_i \bigcap A_j) = 0$, $mess(\cdot)$ denotes the Lebesque measure. At a current iteration a rectangle is selected and partitioned into three equal parts by the hyperplanes rectangular to the longest edge, and subdividing this edge into three equal sub-intervals. Such a partition is used by the DIRECT algorithm [6] and its numerous subsequent modifications where the selection criterion is based on the Lipschitzian model of the objective functions. The selection criterion in our algorithm is based on a statistical model, and it is discussed below.

THE CHOICE OF A STATISTICAL MODEL

The Gaussian random field $\xi(x)$, $x \in A$, is accepted as a statistical model of the objective functions. It is assumed also that the field is stationary, i.e. it is homogeneous and isotropic. The latter (isotropic) assumption can be relaxed if it is not compatible with the available information about the objective function in question. Properties of the considered Gaussian random field are greatly defined by the covariance function $r(t)$. Therefore, the choice of $r(t)$ is very important for the construction of an algorithm, and should be motivated by the available information on the aimed optimization problems. The problem of choice of an appropriate covariance function is broader than the problem of choice of statistical model for global optimization. It is important also in kriging, and the design and analysis of computer experiments [19, 20]. Despite being urgent from different points of view, yet it has no satisfactory solution.

The Gaussian random field with the covariance function $r(t) = \sigma^2 \exp(-(t/s)^2)$ is attractive because of smoothness of the sampling functions. However, the estimation of parameters of this model is problematic [21, 22]. The severely ill definiteness of the correlation matrices of even medium size is the other disadvantage hindering the use of this model for development of Bayesian and P-algorithm in their original form [18]. However, in the algorithm considered below the inversion of correlation matrices is not needed. The estimation problem should be solved anyway. Therefore, for the construction of the algorithm we have chosen the Gaussian random field with the covariance function $r(t) = \sigma^2 \exp(-(t/c)^2)$ as a model of the objective functions of interest. An example sample function is shown in Figure 1.

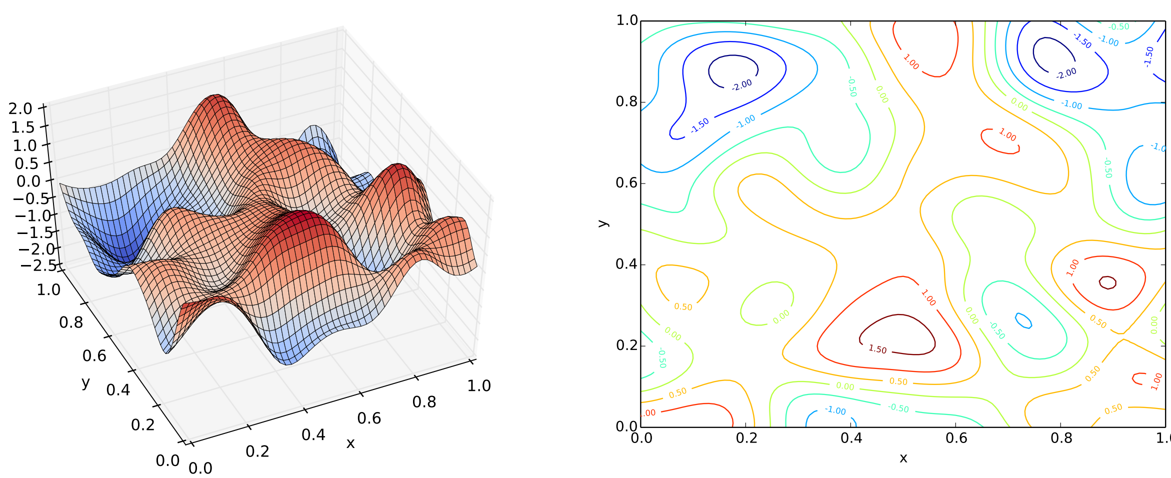

FIGURE 1. An example of a sample function of the 2-dimensional Gaussian stationary random field with mean equal to zero, variance equal to 1, and $r(t) = \exp(-(t/c)^2)$, $c = 0.1$.

CONDITIONAL CHARACTERISTICS WITH RESPECT TO FUNCTION VALUE AND DERIVATIVE

The statistical model is needed to evaluate the sites of potential computation of the objective function by forecasting there the objective function values. Let us consider the single variable case $d = 1$. Assume that the function value and its derivative are computed at the point $x = 0$, and they are equal to y and y' correspondingly. Let the Gaussian random process $\xi(x)$ be chosen for the statistical model of the considered function where $\mathbb{E}\{\xi(x)\} = 0$, $\mathbb{D}\{\xi(x)\} = 1$, and $r(t) = \exp(-(t/c)^2$. The conditional mean

$$m(x|y, y') = \mathbb{E}\{\xi(x) \mid \xi(0) = y, \xi'(0) = y'\},$$

is the best forecast of an unknown function value with respect to the chosen statistical model, and the conditional variance

$$s^2(x|y, y') = \mathbb{D}\{\xi(x) \mid \xi(0) = y, \xi'(0) = y'\},$$

evaluates the uncertainty of the forecast. For the development of an algorithm it is important that both characteristics can be expressed by simple formulas. Since the covariance between $\xi(x)$ and $\xi'(t)$ is defined via partial derivatives of $r(x-t)$ [17], it is easy to obtain the needed formulas. The variance of $\xi'(0)$ is equal to $2/c^2$. The covariance between $\xi(0)$ and $\xi(x)$ is equal to $\exp(-x^2)$, and covariance between $\xi'(0)$ and $\xi(x)$ is equal to $\frac{2x}{c^2}\exp(-(x/c)^2)$.

$$m(x|y, y') = (y, y')\begin{pmatrix} 1 & 0 \\ 0 & 2/c^2 \end{pmatrix}^{-1}\begin{pmatrix} \exp(-(x/c)^2) \\ \frac{2x}{c^2}\exp(-(x/c)^2) \end{pmatrix} = (y + x \cdot y')\exp(-(x/c)^2), \tag{1}$$

$$\begin{aligned} s^2(x|y, y') &= 1 - (\exp(-(x/c)^2), \frac{2x}{c^2}\exp(-(x/c)^2)\begin{pmatrix} 1 & 0 \\ 0 & 2/c^2 \end{pmatrix}^{-1}\begin{pmatrix} \exp(-(x/c)^2) \\ \frac{2x}{c^2}\exp(-(x/c)^2) \end{pmatrix} = \\ &= 1 - \left(1 + \frac{2x^2}{c^2}\right)\exp(-2(x/c)^2). \end{aligned} \tag{2}$$

The influence of the inclusion of derivatives is illustrated in Figure 2 where the graphs of conditional mean values and variances are presented.

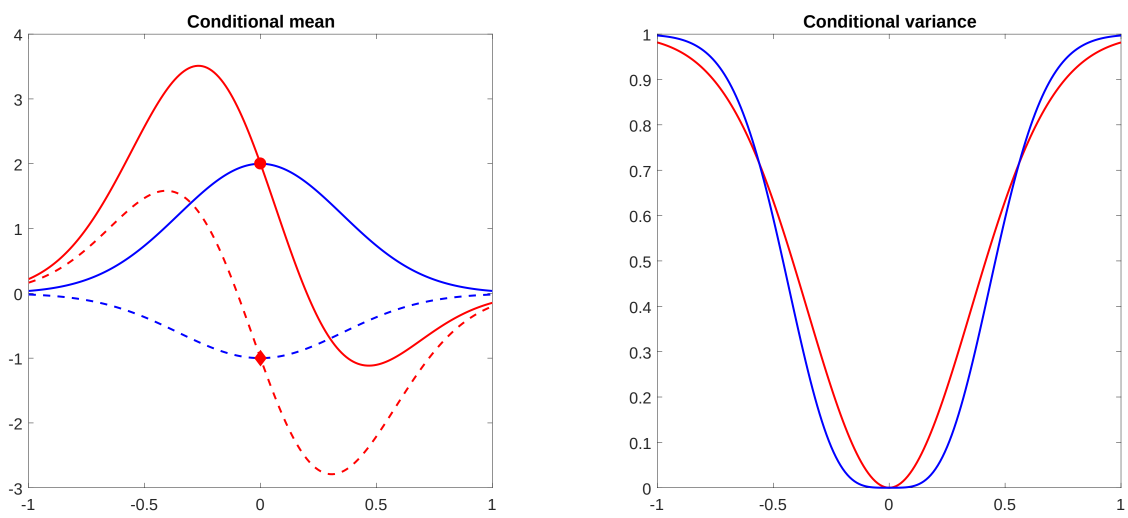

FIGURE 2. Conditional means and variances presented in red and blue for the cases with and without derivative correspondingly; $y = 2$ (circle), $y = -1$ (diamond), and $c = 0.5$, $y' = -10$

SELECTION OF A HYPER-RECTANGLE

A partition algorithm is defined by two main rules: selection and partition of the selected subset. We use the partition analogous to one used in the known DIRECT algorithm [6]. Two types of selection criteria are considered. First: the probability to improve the best found value of the objective function which is similar to that used in [5, 23]. Second: bi-objective selection recently proposed in [24]. In the conference talk the implementation details will be presented, the advantage of the inclusion of gradients will be discussed, and performance of the developed versions of the algorithm will be illustrated.

CONCLUSION

In the case of availability of the derivatives of the objective function, they can be quite easily included into statistical models used for the construction of global optimization algorithms.

ACKNOWLEDGMENTS

This work was supported by the Research Council of Lithuania under Grant No. P-MIP-17-61.

REFERENCES

[1] E. Nikolaidis, V. Pandey, and Z. P. Mourelatos, *Design Decisions under Uncertainty with Limited Information* (CRC Press, 2011).
[2] A. Žilinskas and A. Zhigljavsy, Informatica **27**, 229–256 (2016).
[3] A. Žilinskas and J. Žilinskas, Computers and Mathematics with Applications **52**, 211 – 224 (2006).
[4] A. Žilinskas and J. Žilinskas, Communications in Nonlinear Science and Numerical Simulation **21**, 89 – 98 (2015).
[5] J. M. Calvin, G. Gimbutienė, W. O. Phillips, and A. Žilinskas, Journal of Global Optimization **71**, 165 – 191 (2018).
[6] D. R. Jones, C. D. Perttunen, and C. D. Stuckman, Journal of Optimization Theory and Applications **79**, 157–181 (1993).
[7] D. E. Kvasov and Y. D. Sergeyev, Journal of Computational and Applied Mathematics **236**, 4042–4054 (2012).
[8] R. Paulavičius and J. Žilinskas, *Simplicial Global Optimization. Springer Briefs in Optimization* (Springer, 2014).
[9] J. Pinter, *Global Optimization in Action* (Kluwer Academic Publisher, 1996).
[10] Y. D. Sergeyev and D. E. Kvasov, SIAM Journal on Optimization **16**, 910–937 (2006).
[11] Y. D. Sergeyev and D. E. Kvasov, Communications in Nonlinear Science and Numerical Simulation **21**, 99–111 (2015).
[12] Y. D. Sergeyev, Computer system for storing infinite, infinitesimal, and finite quantities and executing arithmetical operations with them, USA patent 7,860,914,, 2010.
[13] Y. D. Sergeyev, EMS Surveys In Mathematical Sciences **4**, 219–320 (2017).
[14] R. De Leone, Applied Mathematics and Computation **318**, 290–297 (2018).
[15] Y. D. Sergeyev, M. S. Mukhametzhanov, F. Mazzia, F. Iavernaro, and P. Amodio, J. Unconventional Comput **12**, 3–23 (2016).
[16] A. Žilinskas, Applied Mathematics and Computation **218**, 8131–8136 (2012).
[17] A. M. Yaglom, *Correlation Theory of Stationary and Related Random Functions, vol. 1* (Springer, 1987).
[18] A. Zhigljavsky and A. Žilinskas, *Stochastic global optimization* (Springer, 2007).
[19] T. Santler, B. Williams, and W. Notz, *Design and analysis of computer experiments* (Springer, 2003).
[20] M. L. Stein, *Interpolation of spatial data: some theory of kriging* (Springer, 1999).
[21] W. Xu and M. Stein, SIAM/ASA J. Uncertainty Quantification **5**, 138–175 (2017).
[22] A. Pepelyshev, Statistical Methodology **8**, 356 – 362 (2011).
[23] J. M. Calvin and A. Žilinskas, Journal of Optimization Theory and Applications **102**, 479 – 495 (1999).
[24] A. Žilinskas and J. M. Calvin, Journal of Global Optimization, published in Internet https://doi.org/10.1007/s10898–018–0622–5 (2018).
[25] Y. D. Sergeyev, D. E. Kvasov, and M. S. Mukhametzhanov, Communications in Nonlinear Science and Numerical Simulation **59**, 319–330 (2018).