On two models for global optimization

Antanas Žilinskas[1]

[1] Institute of Mathematics and Informatics, Vilnius, Lithuania, email: antanasz@ktl.mii.lt

Abstract. Two models for global optimization are considered: statistical model, and radial basis functions. The equivalence of both models in the case of optimization without noise is discussed. Both models are also evaluated with respect to global optimization in the presence of noise by means of experimental testing where approximation errors of passive one dimensional algorithm are estimated.

1 Introduction

Global optimization is a difficult problem. It becomes especially difficult in cases of computationally intensive objective functions, and in cases of optimization in the presence of noise. We consider two approaches aimed to attack the mentioned difficult problems: the first approach is based on statistical models, and the second approach is based on approximation by radial basis functions. The idea to construct a global optimization method using a stochastic function for a model of objective functions was originally proposed in [5, 6], and axiomatically justified in [12, 13]. Various algorithms based on statistical models are considered, e.g. in the books [7, 10, 11]. The idea to develop a global optimization algorithm using the interpolation of an objective function by a radial basis function (RBF) was proposed in [4]; some new results are presented in [8]. As the main application area of algorithms based on both approaches is supposed optimization of expensive objective functions; statistical approach seems also promising for the originally targeted problems of global optimization in the presence of noise [5, 6]. Although basic assumptions of both approaches seem very different there have been mentioned surprising similarities of the algorithms of both types; see [4].

In the present paper we show the identity of the so called P-algorithm (see [12, 13]), and the algorithm based on the interpolation of an objective function by a RBF; our version of the proof is shorter than that in [4]. Further we compare both (statistical and RBF) models with respect to the application for minimization in the presence of noise. The problem of global optimization in the presence of noise using a statistical model is considered, e.g. in [2] where the complexity of the problem is discussed. To the best knowledge of the author of this paper RBF models have not been used yet to construct algorithms for global optimization in the presence of noise. Combining of advantages of both approaches can be helpful to attack this indeed difficult problem.

2 Optimization without Noise

A minimization problem is considered where the objective function is denoted by $f(x)$, $x \in A \subseteq \mathbb{R}^d$. Let n function values be known (computed or observed): $y_i = f(x_i)$, $i = 1, ..., n$.

An optimization algorithm should define the next observation point x_{n+1}, and in the most general case x_{n+1} can depend on all available information on $f(x)$ including x_i, y_i, $i = 1, ..., n$.

The P-algorithm is defined using a stochastic function $\xi(x)$ for a statistical model of objective functions, and the idea to maximize the probability of improvement at current minimization step defined with respect the model [13, 11]:

$$x_{n+1} = \arg\max_{x \in A} \mathbf{P}\{\xi(x) \leq \tilde{y}_{on} | \xi(x_i) = y_i,\ i = 1, ..., n\}, \tag{1}$$

where $\tilde{y}_{on}$ is a level aimed to exceed downwards at $(n+1)$-st minimization step, e.g. $\tilde{y}_{on} = y_{on} - \varepsilon_n$, $y_{on} = \min_{i=1,...,n} y_i$, $\varepsilon_n > 0$. Assuming $\xi(x)$ Gaussian stochastic function the maximization in (1) can be reduced to the maximization of

$$\frac{\tilde{y}_{on} - m_n(x | \xi(x_i) = y_i,\ i = 1, ..., n)}{s_n(x | \xi(x_i) = y_i,\ i = 1, ..., n)}, \tag{2}$$

where $m_n(x | x_i, y_i, i = 1, ..., n)$ and $s_n^2(x | \xi(x_i) = y_i,\ i = 1, ..., n)$ denote the conditional mean and the conditional variance of $\xi(x)$ with respect to $\xi(x_i) = y_i$, $i = 1, ..., n$, correspondingly. Characterization of the P-algorithm by means of maximization of (2) is sufficient for the further analysis in the present paper.

An interesting global optimization method is proposed in [4] using the idea of interpolation by a radial basis function. Let the values of the objective function $y_i = f(x_i)$, $i = 1, ..., n$ be known, and a point for next observation should be chosen. A values of an objective function $f(\cdot)$ at an arbitrary point $x \in \mathbb{R}^d$ can be predicted by the radial basis function

$$\mu_n(x | x_i, y_i, i = 1, ..., n) = \sum_{i=1}^{n} \lambda_i \phi(||x - x_i||), \tag{3}$$

that interpolates the data $(x_i, y_i = f(x_i))$, $i = 1, ..., n$. A naive idea to perform the next observation at the minimum point of the response surface defined by (3) should be rejected because of known disadvantages discussed, e.g. in [11]. The more sophisticated idea of an algorithm proposed by Gutmann is discussed below, after few remarks concerning the interpolating function (3). We use a standard form of RBF [1] without the extra polynomial summands used elsewhere. Different basis functions $\phi(\cdot)$ can be chosen, e.g. the Gaussian function $\phi(r) = \exp(-\gamma r^2)$, $r \geq 0$, $\gamma > 0$. The coefficients λ_i are defined by the system of linear equations $\mu_n(x_i | \cdot) = y_i, i = 1, ..., n$ whose solution is guaranteed by the positive definiteness of the matrix $\Phi = (\phi(||x_i - x_j||))$.

Although statistical models and RBF models root in different theoretical concepts, the heuristic ideas of the algorithms of both types are similar. The P-algorithm for current observation chooses the point where it is most probable to descend below the target level $\tilde{y}_{on}$. By the RBF based algorithm the current observation of $f(\cdot)$ is performed at the point where the value of $f(\cdot)$ equal to the target value $\tilde{y}_{on}$ is most likely. The definition of the radial basis function, e.g. (3), does not directly imply likelihood of various function values. However, some evaluation of the likelihood can be derived from the general concepts of rationality: an interpolating function can be considered most suitable if adding/removing of a point implies minimal changes in characteristics of the interpolating function. A natural criterion to evaluate suitability of an interpolating

function to given data is a norm of the considered interpolating function; for a definition of a (semi)norm of a RBF we refer to [3].

Let the known function values $y_i = f(x_i)$, $i = 1, ..., n$ be interpolated by means of RBF. The point of the next observation is chosen aiming to get the target value of the objective function equal to $\tilde{y}_{on}$, and minimally increasing the norm of the interpolant implied by the augmentation of x_i, y_i, $i = 1, ..., n$ with $x_{n+1}, \tilde{y}_{on}$. Let us find a point x_{n+1} such that the norm of $\mu_{n+1}(x|x_i, y_i, i = 1, ..., n+1)$ was minimal, where $y_{n+1} = \tilde{y}_{on}$. Such a point for the value $\tilde{y}_{on}$ seems most 'likely' assuming that the behavior of the minimal norm interpolating function is most natural of all interpolators. According to the terminology of [4] such choice of x_{n+1} minimizes 'bumpiness' of the response surface. Formally, the algorithm is constructed sequentially tuning the radial function interpolant by means of minimization of the semi-norm with respect to the forecasted global minimum value $\tilde{y}_{on}$.

In the formulas below we use the shorthand $\mu_n(x) = \mu_n(x|\cdot)$, and the notations

$$\Lambda = (\lambda_1, ..., \lambda_n)^T,\ \Phi(x) = (\phi(||x - x_1||), ..., \phi(||x - x_n||))^T. \tag{4}$$

The formula (3) using these notations can be rewritten in the form $\mu_n(x) = \Lambda^T \Phi(x)$. Similarly the RBF interpolator using an extended set of data (x_i, y_i), $i = 1, ..., n+1$ is defined by the formulas

$$\mu_{n+1}(x) = \sum_{i=1}^{n+1} \omega_i \phi(||x - x_i||) = \Omega^T \cdot \begin{pmatrix} \Phi(x) \\ \phi(||x - x_{n+1}||) \end{pmatrix}. \tag{5}$$

The vectors of coefficients Λ^T and Ω^T can be calculated as solutions of systems of linear equations corresponding to the condition of interpolation

$$\Lambda = \Phi^{-1} \cdot Y,\ Y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix},\ \mathcal{M} = \Psi^{-1} \cdot \begin{pmatrix} Y \\ \tilde{y}_{on} \end{pmatrix},$$

$$\Phi = \begin{pmatrix} \phi(0) & ... & \phi(||x_1 - x_n||) \\ ... & ... & ... \\ \phi(||x_n - x_1||) & ... & \phi(0) \end{pmatrix},$$

$$\Psi = \begin{pmatrix} \Phi & \Phi(x_{n+1}) \\ \Phi(x_{n+1})^T & \phi(0) \end{pmatrix}. \tag{6}$$

The squared semi-norm of μ_{n+1} is equal to

$$||\mu_{n+1}(x)||^2 = \mathcal{M}^T \Psi \mathcal{M} = (Y^T, \tilde{y}_{on}) \Psi^{-1} \begin{pmatrix} Y \\ \tilde{y}_{on} \end{pmatrix}, \tag{7}$$

where the expression of $\mathcal{M}$ from (6) is taken into account.

To invert matrix Ψ presented as a block matrix in (6) the formula by Frobenius can be applied

$$\Psi^{-1} = \begin{pmatrix} \Phi^{-1} + \frac{1}{h}\Phi^{-1}\Phi(x_{n+1})\Phi(x_{n+1})^T\Phi^{-1} & -\frac{1}{h}\Phi^{-1}\Phi(x_{n+1}) \\ -\frac{1}{h}\Phi(x_{n+1})^T\Phi^{-1} & \frac{1}{h} \end{pmatrix}, \tag{8}$$

where

$$h = \phi(0) - \Phi(x_{n+1})^T \Phi^{-1} \Phi(x_{n+1}). \tag{9}$$

Calculation of the norm (7) using the latter expression of Ψ^{-1} gives the following result

$$||\mu_{n+1}(x)||^2 = (Y^T, \tilde{y}_{on}) \Psi^{-1} \begin{pmatrix} Y \\ \tilde{y}_{on} \end{pmatrix} = (Y^T, \tilde{y}_{on}) \cdot$$

$$\cdot \begin{pmatrix} \Phi^{-1} Y + \frac{1}{h} \Phi^{-1} \Phi(x_{n+1}) \Phi(x_{n+1})^T \Phi^{-1} Y - \frac{\tilde{y}_{on}}{h} \Phi^{-1} \Phi(x_{n+1}) \\ -\frac{1}{h} \Phi(x_{n+1})^T \Phi^{-1} Y + \frac{\tilde{y}_{on}}{h} \end{pmatrix} =$$

$$= \Lambda \Phi \Lambda + \frac{(\tilde{y}_{on} - \Phi(x_{n+1})^T \Phi^{-1} Y)^2}{\phi(0) - \Phi(x_{n+1})^T \Phi^{-1} \Phi(x_{n+1})}. \tag{10}$$

From (10) the subsequent equality follows

$$||\mu_{n+1}(x)||^2 = ||\mu_n(x)||^2 + \frac{(\tilde{y}_{on} - \Phi(x_{n+1})^T \Phi^{-1} Y)^2}{\phi(0) - \Phi(x_{n+1})^T \Phi^{-1} \Phi(x_{n+1})},$$

where the first summand does not depend on x_{n+1}. Therefore for the next observation a minimum point x_{n+1} of the following function

$$\frac{(\tilde{y}_{on} - \Phi(x_{n+1})^T \Phi^{-1} Y)^2}{\phi(0) - \Phi(x_{n+1})^T \Phi^{-1} \Phi(x_{n+1})}, \tag{11}$$

should be chosen.

Let us consider a homogeneous isotropic Gaussian random field $\xi(x)$, $x \in \mathbb{R}^n$ with zero mean and covariance function $\phi(\cdot)$. The conditional mean and the conditional variance of $\xi(x)$ with respect to $\xi(x_i) = y_i$, $i = 1, ..., n$, is equal to

$$m_n(x | \xi(x_i) = y_i,\ i = 1, ..., n) = \Phi(x)^T \Phi^{-1} Y,$$
$$s_n^2(x | \xi(x_i) = y_i,\ i = 1, ..., n) = \phi(0) - \Phi(x)^T \Phi^{-1} \Phi(x),$$

correspondingly. Therefore the maximization of (2) is reduced to the maximization of

$$\frac{\tilde{y}_{on} - \Phi(x_{n+1})^T \Phi^{-1} Y}{\sqrt{\phi(0) - \Phi(x_{n+1})^T \Phi^{-1} \Phi(x_{n+1})}}. \tag{12}$$

The minimization of (11) is equivalent to the maximization of (12), since the expression of (11) is equal to the squared expression of (12), and the target level $\tilde{y}_{on}$ is naturally chosen less than $\min m_n(x | \xi(x_i) = y_i,\ i = 1, ..., n)$. The latter conclusion means that the statistical model based P-algorithm and the RBF model based algorithm are identical. For the asymptotic analysis of the convergence of the corresponding algorithm the well developed theory of interpolation by RBF (see, [1]) can be useful. On the other hand, a statistical model can be especially useful to justify the statistical estimation of the model parameters, e.g. of the parameter γ in case $\phi(r) = \exp(-\gamma r^2)$). Let us note, that the suitable choice of parameters is crucial for the practical performance of the algorithm

measured after restricted number of iterations; for the use of statistical models with tuned parameters for interpolation we refer to [9].

3 Optimization in the Presence of Noise

The P-algorithm for minimization in the presence of noise is a simple generalization of the corresponding algorithm for minimization without nose; it is defined by an expression similar to (1), where the conditional probability is calculated with respect to the noisy data. The noise is modelled by the independent random variables ξ_i, and the available information about the objective function is (x_i, z_i), $i = 1, ..., n$, $z_i = f(x_i)+\xi_i$. Let $\xi(x)$ (a model of objective functions) be a stationary (homogenous, isotropic) Gaussian stochastic function with zero mean, unit variance, and correlation function $\phi(x)$ depending only on $||x||$. Let ξ_i be Gaussian random variables with zero mean and variance σ_t^2. Similarly as in the case without noise the construction of P-algorithm can be reduced to the maximization of (12) where y_i, $i = 1, ..., n$ are replaced by z_i, $i = 1, ..., n$, and the diagonal elements of the matrix Φ are replaced by $1 + \sigma_t^2$. Although both versions of P-algorithm are similar from the theoretical point of view, the implementation of the version for noisy minimization is more complicated, at least because of considerably larger, than in the case without noise, number of observations needed to achieve an acceptable accuracy.

A generalization of an RBF based algorithm to the noisy case is not so straightforward. The question about the similarity between the P-algorithm and the RBF model based algorithm in the noisy case also is not yet answered.

Prior to the investigation of sophisticated algorithms it seems reasonable to compare both models with respect to the simplest, i.e. passive, algorithm. The randomized version of the passive algorithm is considered where function values (in the presence of noise) are observed at random points uniformly distributed in the feasible region. A statistical model and a RBF model are applied to evaluate function values at arbitrary points as well as global minimum using the noisy observations of the passive algorithm.

Let (x_i, z_i), $i = 1, ..., n$ are known where x_i are random points with uniform distribution over the feasible region, and $z_i = f(x_i) + \xi_i$. An unknown value $f(x)$ can be approximated using the statistical model $\xi(\cdot)$ by the conditional expectation of $\xi(x)$ with respect to $\xi(x_i) + \xi_i = z_i$, $i = 1, ..., n$: $M_n(x|\xi(x) + \xi_i = z_i,\ i = 1, ..., n)$. It is well known that

$$M_n(x|\xi(x) + \xi_i = z_i,\ i = 1, ..., n) = \tag{13}$$

$$= (\phi(||x - x_1||), ..., \phi(||x - x_n||)) \cdot \begin{pmatrix} 1 + \sigma_t^2 & ... & \phi(||x_1 - x_n||) \\ ... & ... & ... \\ \phi(||x_n - x_1||) & ... & 1 + \sigma_t^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} z_1 \\ ... \\ z_n \end{pmatrix}.$$

Originally RBF have been applied for interpolation of multidimensional spatial data. Extension of RBF for approximation of data corrupted by random errors can be based, e.g. on the idea of least squares where the coefficients λ_i in (3) are chosen to satisfy the equality

$$\sum_{i=1}^{n} (z_i - \mu_n(x_i))^2 = n\sigma_t^2, \tag{14}$$

and to minimize the semi norm of $\mu_n(\cdot)$

$$\|\mu_n(\cdot)\| = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \phi(\|x_i - x_j\|) \right)^{1/2}. \tag{15}$$

It can be shown that the vector of optimal coefficients $\Lambda = (\lambda_1, ..., \lambda_n)^T$ is defined by the following system of equations

$$\begin{aligned} \Lambda &= (\Phi + I \cdot \frac{1}{\nu})^{-1} \cdot Z, \\ \nu^2 &= \frac{\|\Lambda\|^2}{n\sigma_t^2}, \end{aligned} \tag{16}$$

where I denotes the unit matrix. The equations (16) imply the following expression of the approximating RBF

$$\Upsilon_n(x|\xi(x) + \xi_i = z_i,\ i = 1, ..., n) = \tag{17}$$
$$= (\phi(||x - x_1||), ..., \phi(||x - x_n||)) \cdot \begin{pmatrix} 1 + 1/\nu & ... & \phi(||x_1 - x_n||) \\ ... & ... & ... \\ \phi(||x_n - x_1||) & ... & 1 + 1/\nu \end{pmatrix}^{-1} \cdot \begin{pmatrix} z_1 \\ ... \\ z_n \end{pmatrix}.$$

The comparison of (13) and (17) enlightens similarity and difference between the statistical model based and the RBF based approximating functions. The structure of expressions defining both functions are identical up to the diagonal elements of the matrices. In (13) the diagonal element is equal to $1+\sigma_t^2$, and the diagonal element in (17) is equal to $1+1/\nu$. The increase of diagonal elements implies the increase of smoothness of the corresponding function. The parameter σ_t^2 has an obvious meaning, and can be estimated by means of a standard method of mathematical statistics. The parameter $1/\nu$ depends not only on σ_t^2 but also on the function values and on the involved RBF. Average values of $1/\nu$ for two test functions are estimated by means of a modelling experiment using the RBF defined by the formula $\phi(r) = \exp(-\gamma r^2)$. The following test functions were used

$$\begin{aligned} f_1(x) &= \sin(x) + \sin(10x/3) + \ln(x) - 0.84x + 3,\ 2.7 \le x \le 7.5, \\ f_2(x) &= \sin(x) + \sin(2x/3),\ 3.1 \le x \le 20.4, \end{aligned}$$

where feasible intervals were rescaled to [0,1], and function values were rescaled to ensure zero mean and variance equal to 1. The sample of size $N = 100$ consisted of $n = 120$ noisy function values defined at random points uniformly distributed over the feasible interval; "noise" was modelled by independent Gaussian random values with zero mean and variance σ_t^2. The parameter $1/\nu$ was estimated for several values of σ_t representing strong and medium noise, and several values of γ similar to the average values of the maximum likelihood estimates equal to 55 for $f_1(\cdot)$ and equal to 52 for $f_2(\cdot)$.

The results presented in Table 1 show that the diagonal elements in (17) are larger than that in (13), meaning that approximating RBF is smoother than the statistical model based approximating function. The variance of estimates is rather large; one of reasons of worsening the estimate is ill conditioning of the correlation matrix.

Table 1. Estimates of mean and standard deviation of $1/\nu$.

Parameters		f_1		f_2	
		mean	std	mean	std
$\sigma_t = 0.5,$	$\gamma = 60$	1.2560	0.9786	1.2109	0.7753
$\sigma_t = 0.5,$	$\gamma = 50$	1.0890	0.9516	0.8947	0.7510
$\sigma_t = 0.5,$	$\gamma = 40$	0.9039	0.7570	0.6555	0.5541
$\sigma_t = 0.3,$	$\gamma = 50$	0.5227	0.3909	0.4325	0.2920
$\sigma_t = 0.1,$	$\gamma = 50$	0.1110	0.0800	0.1180	0.0795

The randomized passive global optimization algorithm evaluates objective function values at n random points uniformly distributed over the minimization interval, and minimum of the approximating function, (13) or (17), is accepted to approximate global minimum. Mean values and standard deviations of approximation errors, measured as differences between the true minimum and approximations, are presented in Table 2 and Table 4 below. The estimates of errors of approximation of $f_1(\cdot)$ are presented in Table 3; mean square errors were evaluated using the same values that were used for minimization, and they were averaged for a sample of size $N = 100$.

Table 2. Precision of estimating global minimum of $f_1(\cdot)$.

n	stat. mod.		RBF	
	mean	std	mean	std
40	-0.0647	0.2078	-0.3639	0.1758
80	-0.0446	0.1715	-0.2153	0.1587
120	-0.0231	0.1621	-0.1536	0.1249
160	-0.0149	0.1377	-0.1086	0.1297
200	-0.0242	0.1219	-0.0971	0.1100

Table 3. Precision of estimating function values of $f_1(\cdot)$.

n	stat. mod.		RBF	
	m.sq.r.	std	m.sq.r.	std
40	0.2920	0.0974	0.3952	0.1104
80	0.1952	0.0538	0.2556	0.0656
120	0.1632	0.0481	0.2107	0.0517
160	0.1381	0.0348	0.1740	0.0455
200	0.1231	0.0354	0.1518	0.0428

Quantitatively results for both test functions are very similar. The estimates of minimum are biased to the direction of overestimation, and the bias is much larger for the RBF based algorithm than for the statistical algorithm. The latter empirical result well corroborates our previous conclusion that the RBF based approximant is smoother (i.e. also more biased towards to average of function values) than that based on the statistical model. The average of square root errors of approximation of values of $f_1(\cdot)$ (see Table 3) is larger than standard deviation of errors in estimating global minimum; this can be explained by more smooth behavior of the objective function in the neighborhood of global minimizer than in average.

Table 4. Precision of estimating global minimum of $f_2(\cdot)$.

n	stat. mod.		RBF	
	mean	std	mean	std
40	-0.1941	0.2436	-0.6187	0.2697
80	-0.0876	0.1868	-0.3616	0.1841
120	-0.0697	0.1593	-0.2464	0.1532
160	-0.0334	0.1311	-0.1693	0.1179
200	-0.0430	0.1259	-0.1521	0.1169

4 Conclusions

The identity between statistical model based P-algorithm and RBF model based algorithm does not extend to noisy case but some similarity between these algorithms is likely. Implementation of adaptive algorithms based on both models is challenging because of difficulties in inverting large ill defined matrices. An extra challenge for RBF model is estimation of parameters of the model.

Bibliography

[1] M. Buhmann. *Radial Basis Functions: Theory and Implementations.* Cambridge University Press, 2003.

[2] J. Calvin. Nonadaptive univariate optimization for observations with noise. In A. Törn and J. Žilinskas, editors, *Models and Algorithms for Global Optimization.* Springer, 2007.

[3] H.-M. Gutman. On the semi norm of radial basis function interpolants. *J. Approx. Theor.*, 111:315–328, 2001.

[4] H.-M. Gutman. A radial basis function method for global optimization. *J. Global Optim.*, 19:201–227, 2001.

[5] H. Kushner. A versatile stochastic model of a function of unknown and time-varying form. *J. Math. Anal. Appl.*, 5:150–167, 1962.

[6] H. Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *J. Basic Engineering*, 86:97–106, 1964.

[7] J. Mockus. *Bayesian Approach to Global Optimization.* KAP, 1988.

[8] R. G. Regis and Ch. A. Shoemaker. Improved strategies for radial basis function methods for global optimization. *J. Global Optim.*, 37:113–135, 2007.

[9] M. Stein. *Interpolation of Spatial Data, Some Theory of Kriging.* Springer, 1999.

[10] R. Strongin and Y. Sergeyev. *Global Optimization with Non-Convex Constraints.* Kluwer, 2000.

[11] A. Törn and A.Žilinskas. *Global Optimization.* Springer, 1989.

[12] A. Zilinskas. Axiomatic approach to statistical models and their use in multimodal optimization theory. *Math. Program.*, 22:104–116, 1982.

[13] A. Zilinskas. Axiomatic characterization of a global optimization algorithm and investigation of its search strategies. *Operat. Res. Letters*, 4:35–39, 1985.