On Bayes approach to optimization

Petr Volf [1]

Abstract. In many real optimization problems the objective function is either hardly tractable or its evaluation is expensive. Hence, we have not full information on its form and can afford to evaluate it at just a few points. Then, certain assumptions on the objective function form (shape) must be done. This could be with advantage taken as a prior information in a Bayes scheme. The Bayes approach to optimization, extensively studied in last several decades, then offers the way of effective search for the extremal point. In the present paper we shall recall the ideas behind Bayes optimization procedures, describe the technique using the model of Gauss process and derive a regression-like method dealing with noisy information on objective function.

Keywords: optimization, Bayes method, MCMC, Gauss process, nonparametric regression.

JEL classification: C41, J64
AMS classification: 62N02, 62P25

1 Introduction

The paper deals with the problem of optimization in the case that we have not complete information on the objective function and that its evaluation is costly (computationally or even literally). Hence, we wish to reduce its evaluation just to a few points. Therefore, we have to find an effective way how to ,,reconstruct" objective function, at least in a neighborhood of its extremal point. To make it possible, it is necessary to make some assumptions on the objective function form. These assumptions can as well be taken as a prior information, which is further specified on the basis of other reasonably selected observations. Such a point of view leads quite logically to the use of Bayes approach.

In the present paper two cases are distinguished. In Part 2, it is assumed that the objective function can be evaluated (practically) without error. Then, one of the convenient methods takes the form of objective function as a trajectory of Gauss process, with parameters describing its covariance structure. Such an approach is in detail described in Brochu et al (2010). The authors provide also a rich list of relevant references dated from 70-ties till today and covering also some competing approaches.

However, when the objective function is evaluated with random noise, the interpolating model of Gauss process has problems not only with tracing unknown function shape but also with locating its mode. Therefore in this case we prefer non-parametric regression model estimating the objective function and extrapolating its estimate to the region where the mode is expected. This approach is described in Part 3. The objective function is constructed from regression (i.e. smoothing) splines (De Boor, 1978). Their parameters, namely the location of splines knots and also their numbers, are subjected to bayesian estimation. Practical Bayes search then uses the MCMC (Markov chain Monte Carlo) procedures. More on the MCMC method can be found in a number of papers and monographs, see for instance Gamerman (1997) and also Volf (2006).

In both cases, the search procedure should be adaptive in the sense of a trade-off between *exploration* (revealing the objective function shape) and knowledge *exploitation*, especially in the neighborhood of expected extremal point. In practise it means that after each step the posterior is re-analyzed and the next step selected on the basis of re-computed predictive probabilities.

[1]Institute of Information Theory and Automation of the ASCR, Pod vodárenskou věží 4, Praha 8, Czech Republic, volf@utia.cas.cz

2 Gauss process model

The use of Gauss process as a model for unknown objective function dates back to 70-ties. The advantage is that all finite-dimensional and also conditional distributions are normal. Hence, the process is fully specified by its expectation function $m(z)$ and covariance function $\mathrm{cov}(z_1, z_2)$. In the present case we consider an isotropic (homogeneous) process with constant $m(z) = m$ and covariance depending only on distance $|z_1 - z_2|$. Namely,

$$\mathrm{cov}(z_1, z_2) = d^2 \cdot \exp(-c\,|z_1 - z_2|^2) \tag{1}$$

is a convenient choice (Brochu et al, 2010, Močkus, 1994). Thus, the Gauss process is described by three parameters $m, d > 0, c > 0$.

Let us assume that we already know values of objective function $g_i = g(z_i)$ at several points $z_i, i = 1, ..., n$. Our task is now to select another point z convenient for new evaluation of objective function, having in mind that we wish to approach the $\mathrm{argmax}\, g(z)$. Let $g(z)$ be the value at a new point z, further let us denote $\boldsymbol{m} = (m, ..., m)'$, $\boldsymbol{g} = (g_1, ..., g_n)'$, $\boldsymbol{D} = \mathrm{cov}(\boldsymbol{g})$, and $\boldsymbol{d}_z = \mathrm{cov}(\boldsymbol{g}, g(z))$. Then the joint distribution of ,,old" and new values is

$$\begin{pmatrix} \boldsymbol{g} \\ g(z) \end{pmatrix} \sim N\Big[\begin{pmatrix} \boldsymbol{m} \\ m \end{pmatrix}, \begin{pmatrix} \boldsymbol{D}, \boldsymbol{d}_z \\ \boldsymbol{d}_z', d^2 \end{pmatrix}\Big].$$

Finally, the conditional distribution of the interest is

$$(g(z) - m|\boldsymbol{g}, z) \sim N(\mu(z), \sigma^2(z)), \tag{2}$$

with

$$\mu(z) = \boldsymbol{d}_z'\, \boldsymbol{D}^{-1}\, (\boldsymbol{g} - \boldsymbol{m}), \quad \sigma^2(z) = d^2 - \boldsymbol{d}_z'\, \boldsymbol{D}^{-1}\, \boldsymbol{d}_z.$$

In Bayes setting, the prior space is the space of trajectories of the Gauss process, with its prior distribution given by parameters m, c, d. It means that they are taken as hyper-parameters. The likelihood is then given by the joint Gauss distribution of the data, which actually also stands for a finite-dimensional part of the Gauss process posterior. We are, however, mainly interested in the conditional distribution (2), as it represents the Bayes predictive distribution of the next observation. It is seen that when the hyper-parameters are selected (fixed), the Bayes ‚scent' is presented just implicitly, offering one of several possible interpretations of the method. The choice of covariance function (1) can as well be taken as a selection of a kernel influencing the smoothness of the Gauss process trajectories. Hence, some rules for the choice of optimal smoothing kernel can be applied (again, cf. Brochu et al, 2010).

The main aim of the model construction is to provide a tool for the selection of the next point which, with high probability, is closer to $\mathrm{argmax}\, g(z)$ than points already screened. Let $z^+ = \mathrm{argmax}\, g(z_i), i = 1, ..., n$ be the extremal point from them. Then the ,,probability of improvement", i.e. the probability that at a point z the value is higher, is

$$PI_0(z) := \mathrm{P}(g(z) > g(z^+)) = 1 - \Phi\Big(\frac{g(z^+) - \mu(z)}{\sigma(z)}\Big),$$

where $\Phi(.)$ denotes the distribution function of standard Gauss distribution. Intuitively we should select z maximizing $PI_0(z)$. However, this choice has tendency to dwell in the vicinity of z^+ not supporting jumps to other areas. Therefore, an improved rule is based on the criterion

$$PI(z) := \mathrm{P}(g(z) > g(z^+) + v) = 1 - \Phi\Big(\frac{g(z^+) + v - \mu(z)}{\sigma(z)}\Big),$$

(cf. Kushner and Yin, 1997, Torn and Žilinskas, 1989). Here $v \geq 0$ is a tuning parameter, chosen rather subjectively. It is recommended to decrease it to zero (exponentially, for instance) with growing number of procedure iterations. An analogy can be found in tuning the cooling parameter in the simulated annealing method of randomized optimization. The following simple example illustrates the method on the case of one-dimensional objective function, however, generalization to more dimensions is quite straightforward.

Example 1. Let us consider an objective function

$$g(z) = 10 - z \cdot \sin(\frac{z^2}{10})$$

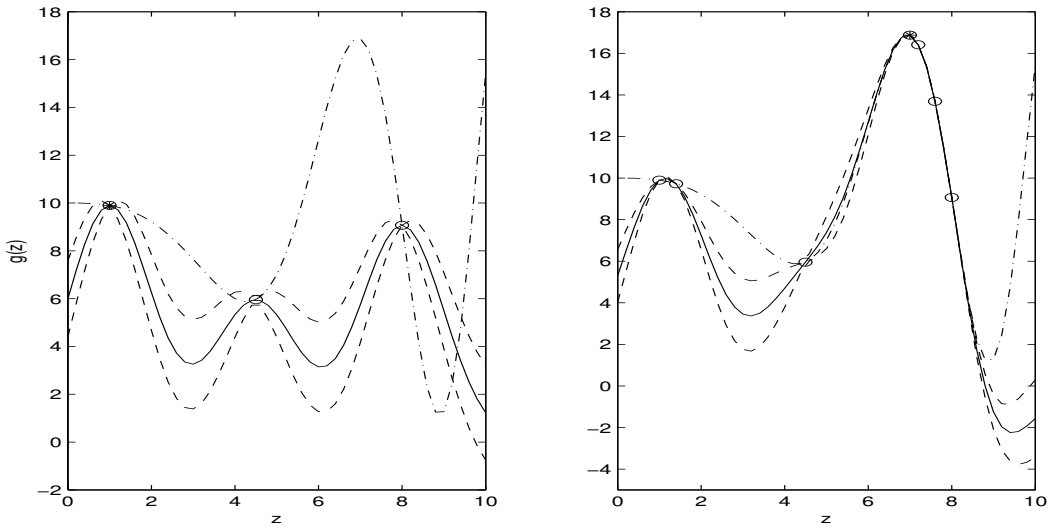

Figure 1 Use of Gauss process model: Initial phase of search with just 3 points (left), state of search after 4 iterations (right)

on interval $\boldsymbol{Z} = (0, 10)$ and let us search for its maximum. Further, let us assume that, initially, the function is evaluated at only 3 points. Figure 1, left plot, shows such a case, maximum of $g(z_i)$, $i = 1, 2, 3$, is at the left point. Solid curve shows the mean $\mu(z)$ of Gauss process constructed from these 3 points, while two dashed curves show $\mu(z) \pm 2\sigma(z)$. Dash-dots curve is then (unknown) objective function $g(z)$. The right plot shows the progress of search after 4 iterations of the search procedure described above. Function has been evaluated sequentially at 4 new points proposed by the PI. They are shown in the plot, the last of them is already quite close to the mode of $g(z)$. In this experiment the parameters of the procedure were fixed to $d = 1, c = 0.5$, m was estimated by the mean of $g(z_i)$. Tuning parameter in j-th iteration was set to $v(j) = 0.5^j$.

3 Nonparametric regression model

Let us now assume that evaluation of the objective function is not precise, that instead $g(z)$ we observe $y(z) = g(z) + \varepsilon(z)$. In the simplest case it is assumed that $\varepsilon(z)$ are independent copies of the same random variable ε possessing the Gauss distribution with zero mean and an unknown variance δ^2. There are essentially two different ways how to estimate unknown (we assume that smooth enough) function $g(z)$. The first consists in the local (e.g. kernel) smoothing. The other approach, utilized here, employs the approximation of $g(z)$ by a combination of functions from some functional basis. For instance, the polynomial splines are the popular choice. Then the model of function $g(z)$ has the form

$$g_M(z) = \boldsymbol{\alpha}' \boldsymbol{B}(z; \boldsymbol{\beta}) = \sum_{j=1}^{M} \alpha_j B_j(z; \boldsymbol{\beta}), \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)'$ is a vector of linear parameters, B_j are basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)'$ is a vector of parameters of the basis functions (e. g. knots of splines). While the estimates of $\boldsymbol{\alpha}$ can be obtained directly from linear regression context, estimation of $\boldsymbol{\beta}$ is more difficult. As a solution to the nonlinear problem for coefficients $\boldsymbol{\beta}$ as well as to optimal choice of number of used units, M, it is possible to use the Bayes methodology in combination with the Markov chain Monte Carlo (MCMC) algorithms. In this framework, the parameter $\boldsymbol{\beta}$ is considered to be a multi-dimensional random vector, with a prior distribution satisfying certain constraint. Simultaneously, M is also regarded as a random variable, with a decreasing prior on $\{0, 1, 2, \ldots, M_{max}\}$. Such a choice lowers the chance to accept a model with high number of units, if the gain of that model is low.

Example 2. Let us consider the same objective function $g(z)$ as in Example 1. However, now it is assumed that it is evaluated with a random error, namely at selected z we observe

$$y(z) = g(z) + \varepsilon(z),$$

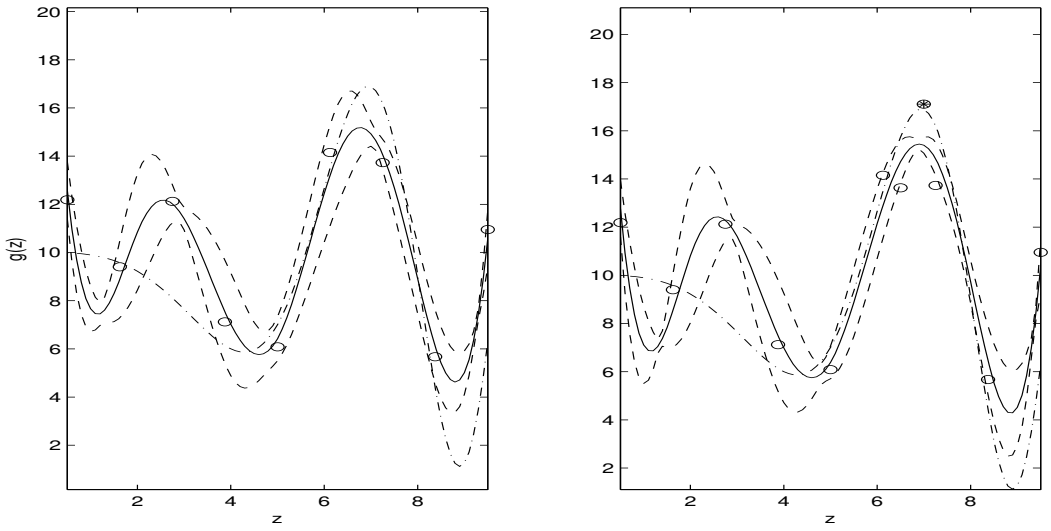

Figure 2 Use of splines model: Initial phase of search for maximum of objective function (left), state of search after 2 iterations (right)

where $\varepsilon(z)$ are the i.i.d. random variables $\varepsilon(z) \sim N(0, \delta^2 = 4)$.

The procedure started from observations at 7 points $z(i)$ located uniformly inside $(0, 10)$, where values $y(z_i) = g(z_i) + \varepsilon(z_i)$ were generated. Notice that here, in the case with considerable noise, the initial number of point is larger than in Example 1. It is necessary for evaluation of the splines.

Values $y(z_i)$ are shown in Figure 2, left plot, again, hidden objective function $g(z)$) is dashed. Estimate of $g(z)$ was then constructed from cubic B-splines. As regards the prior for their knots, we used uniform distribution on the set $\{0 < \beta_1 < \beta_2 < \ \ldots \ < \beta_M < 10\}$. M was bounded by 6 in order to ensure their identifiability. $S = 500$ loops of the Markov chain generation were performed. One loop updated sequentially all components of $\boldsymbol{\beta}$, with possible change of M. It means that it contains up to 6 iterations of model, depending on actual number M.

Only the final result after each loop was registered as a new member of the chain, $g^{(m)}(z)$. The average of this sequence of functions, after skipping first $s = 100$ of them,

$$\mu(z) = \frac{1}{S-s} \sum_{m=s+1}^{S} g^{(m)}(z), \tag{4}$$

serves then as the estimate of $g(z)$. In Figure 2 it is plotted by a full curve. The variability of this set of $S - s$ functions is not constant. The vertical cut at a given z represents Bayes prediction distribution for corresponding $g(z)$. Hence, variance of prediction $\sigma^2(z)$ is computed as a sample variance from values $g^{(m)}(z)$ (compare also discussion in Bishop, 1992, Ch. 10). Dashed curves in the plot again show $\mu(z) \pm 2\,\sigma(z)$. The right plot shows also two new sequentially chosen points tending to the mode of $g(z)$.

Computation of prediction variance

When the nonlinear part of the model (e.g. the knots of splines) is specified, the variance of prediction can also be quantified with the aid of standard linear regression analysis adapted to our case. We then deal with the following linear regression model

$$y_i = \boldsymbol{B}^T(x_i) \cdot \boldsymbol{\alpha} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\alpha}$ are unknown parameters, $\varepsilon_i \sim \mathcal{N}(0, \delta^2)$ are the i.i.d. normal random variables, $\boldsymbol{B}(x_i) = (B_1(x_i), \ldots, B_M(x_i))^T$ are B-splines evaluated at data-points $\boldsymbol{x} = (x_1, \ldots, x_n)^T$. Denote $\boldsymbol{B}$ the $n \times M$ matrix with rows $\boldsymbol{B}^T(x_i)$, $\boldsymbol{A} = (\boldsymbol{B}^T \cdot \boldsymbol{B})^{-1}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Then the least squares method yields the estimate

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{A} \cdot \boldsymbol{B}^T \cdot \boldsymbol{y}, \quad \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \sim \mathcal{N}(O, \delta^2 \cdot \boldsymbol{A}),$$

where O is the null vector. Further, at a new point z the prediction of $g(z)$ is $\hat{g}(z) = \boldsymbol{B}^T(z) \cdot \hat{\boldsymbol{\alpha}}$. Its expectation is 'true' $g(z)$, while its variance equals

$$\sigma^2(z) = \text{var}\left(\hat{g}(z)\right) = \boldsymbol{B}^T(z) \cdot \boldsymbol{A} \cdot \boldsymbol{B}(z) \cdot \delta^2.$$

As expected, it depends both on data ($\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{x})$) from which the model was estimated, and on position of prediction point z.

A case with multidimensional objective function can be, essentially, solved in the same manner. However, multivariate function has to deal with interactions of several predictors. This is as a rule modelled by a tensor product of one-dimensional units. The problem is caused by the fact that their number grows (exponentially) with dimension, so that there also grows the number of 'nonlinear' parameters. Nevertheless, there are some approaches to the regression functions modelling able to reduce the number of parameters, for instance the projection pursuit method.

4 Application to quantile optimization problem

In quantile optimization the main criterion of interest is certain quantile related to the decision maker risk acceptance. Let us consider a stochastic optimization problem with utility function $\varphi(x, z)$, where z are input (decision) variables from a set $\boldsymbol{Z}$ and values x are results of a random variable (or vector) X. Denote $F(y; z)$ the distribution function of random variable $Y = \varphi(X, z)$ with decision z, then, for selected $\alpha \in (0, 1)$ the objective is to maximize, over $z \in \boldsymbol{Z}$, α-quantile $Q(\alpha; z) = \min\{y : F(y; z) \geq \alpha\}$. However, in many cases the evaluation and consequent optimization of quantile criterion is not easy, often it uses iterative procedures or Monte Carlo generation. Then the method described in the present paper can be applied. We shall show such an application of method described in Section 2 on an example from the area of reliability and maintenance optimization.

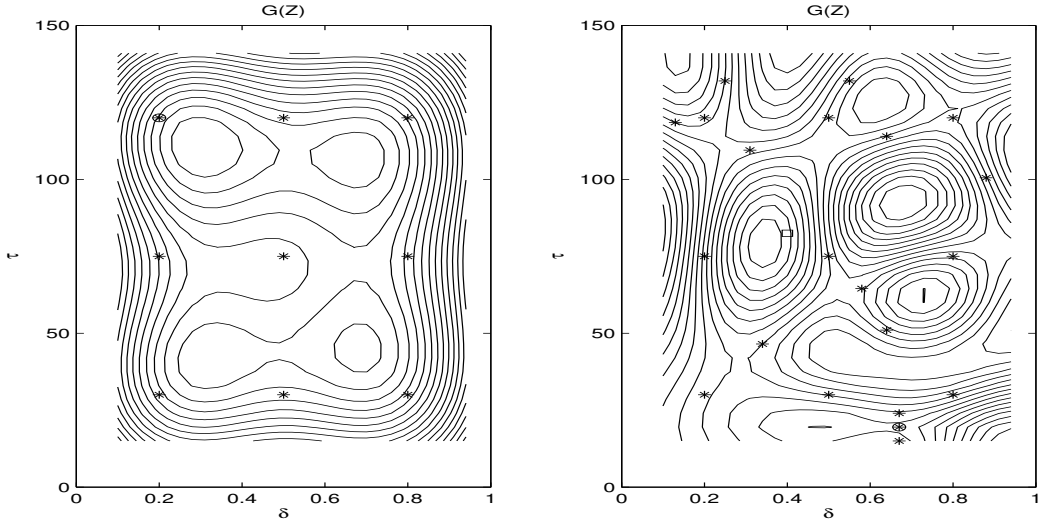

Figure 3 Example 3: Initial phase of search (left), state of search after 12 iterations (right)

Example 3. The Kijima model of non-complete repair (Kijima, 1989) assumes that the device is repaired in its age τ with a degree δ, which means that after repair the virtual age of the device is $(1 - \delta) \cdot \tau$. Thus, $\delta = 1$ means complete repair, renewal, while $\delta = 0$ is the minimal repair.

In the example it is assumed that the Kijima model concerns to preventive repairs, meanwhile after the failure the device has to be renewed completely. We are given the costs of renewal, C_1, and of preventive repair, $C_2(\delta, \tau)$. The objective is to maximize, over τ and δ, an α-quantile of random function $\varphi(X, \delta, \tau)$ equal to proportion of the time to renewal to the costs to renewal. Here X is the random time to failure of the device. This proportion equals

$$\varphi(X, \delta, \tau) = \frac{X}{C_1} \quad \text{with probability} \quad P(X \leq \tau),$$

$$\varphi(X,\delta,\tau)=\frac{\tau+\tau\cdot\delta\cdot(k-1)+X_k}{C_1+k\cdot C_2}\text{ with }P(X>\tau)\cdot P(X_1>\tau)^{k-1}\cdot P(X_k\le\tau),$$

where all $X_k=\{X|X>\tau(1-\delta)\}$ and k is the number of preventive repairs before the failure. It is seen that the direct evaluation of objective function is not easy, moreover, it is strongly non-concave. Therefore, the distribution of variable $Y(\delta,\tau)=\varphi(X,\delta,\tau)$, for different δ,τ, is obtained ‚empirically' by random generation, quantiles $Q(\alpha;\delta,\tau)$ then as sample quantiles.

For numerical illustration we selected $X\sim\text{Weibull}(a=100,b=2)$, with survival function $\overline{F}(x)=\exp\Big(-\big(\frac{x}{a}\big)^b\Big)$, E$X\sim 89$, std$(X)\sim 46$. Further, $\alpha=0.1$ the costs $C_1=40$, $C_2=2+(\delta\cdot\tau)^\gamma$, $\gamma=0.2$. Figure 3 shows the results. Objective function is $Q(\alpha;\delta,\tau)$, procedure started from its Monte Carlo generation in 9 points showed in the left plot. Maximum is denoted by a circle, its value was 0.876. The plot contains also contours of resulting Gauss process surface. The right plot shows the situation after 12 iterations. It is seen how the space was inspected, maximal value was stabilized around 1.124, the corresponding point ($\delta\sim 0.7,\tau\sim 20$) is again marked by a circle.

5 Conclusion

The contribution has studied the problem of optimization in the case when the objective function is not known sufficiently and its evaluation is costly. Two methods of the search for the objective function extremal point were described. Both are based on an appropriate model of the function. However, the aim is not to reconstruct it fully (like in the regression analysis), but, preferably, in the vicinity of its extreme, minimizing simultaneously number of function evaluations. The success of the methods is based also on the design of starting points. Like in other global optimization methods, if they are far from the optimum and the function is not unimodal, the procedure can end in a local extreme.

Acknowledgements

The research has been supported by the grant No. 13-14445S of the Czech Science Foundation.

References

[1] Bishop, C.: *Neural Networks for Pattern Recognition*. Cambridge Univ. Press, Cambridge, 1992.

[2] Brochu, E., Cora, V. M., and de Freitas, N.: A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv:1012.2599v1 [cs.LG]* **12** (2010).

[3] De Boor, C.: *A Practical Guide to Splines*. Springer Verlag, Berlin, 1978.

[4] Gamerman, D.: *Markov Chain Monte Carlo*. Chapman and Hall, New York, 1997.

[5] Kijima, M.: Some results for repairable systems with general repair. *Journal of Applied Probability* **26** (1989), 89–102.

[6] Kushner, H. J. and Yin, G. G.: *Stochastic Approximation Algorithms and Applications*. Springer Verlag, Berlin, 1997.

[7] Močkus, J.: Application of Bayesian approach to numerical methods of global and stochastic optimization. *J. Global Optimization* **4** (1994), 347–365.

[8] Torn, A. and Žilinskas, A.: *Global Optimization*. Springer Verlag, Berlin, 1989.

[9] Volf, P.: Markov Chain Monte Carlo methods in computational statistics and econometrics. In: *Proceedings of the 24th International Conference Mathematical Methods in Economics 2006* (L. Lukáš ed.), ZČU Plzeň, 2006, 525–530.

32nd International Conference

Mathematical Methods in Economics

MME 2014

Conference Proceedings

Olomouc, Czech Republic
September 10 – 12, 2014

Faculty of Science
Palacký University, Olomouc

Editors
Jana Talašová
Jan Stoklasa
Tomáš Talášek

Technical Editor
Tomáš Talášek

Cover Photo
Viktor Čáp

First edition
ISBN 978-80-244-4209-9