Bayesian Regularization of Function Approximation Using Orthogonalized Bases

Nuzhny A.S., Kuchugov P.A.

Nuclear Safety Institute of Russian Academy of Sciences
Bol'shaja Tul'skaja 12, Moscow 115191 Russian Federation
nuzhny@inbox.ru
National Research Nuclear University "MEPhI"
Kashirskoe shosse 31, Moscow 115409 Russian Federation
pkuchugov@gmail.com

Abstract

A point-by-point approximation method for multidimensional scalar function is discussed. In accordance with Tikhonov theory a stabilizing functional was introduced in the solution set for the purpose of approximation regularization. Bayesian method was used to determine the regularization parameter. The developed algorithm has unique analytical solution for the regularization parameter unlike other approximation models which use Bayesian regularization.

1 Introduction

The heart of the problem of scalar function approximation is to build a functional dependence $h(\vec{x})$ able to produce the given dataset $D = \{y_i, \vec{x_i}\}_{i=1}^{L}$. The data are presented as points in $m+1$-dimensional space consisting of m-dimensional subspace of inputs X and one-dimensional subspace of outputs Y. If there is noise in the data the function $h(\vec{x})$ is admitted not to go exactly through the points.

Function approximation, as most of the inverse problems, is ill-posed i.e. has many solutions. Mathematically it is usually formulated as mean-square error minimization problem (the least squares method):

$$E = \sum_{i=1}^{L} (y_i - h(x_i))^2 .$$

To guarantee uniqueness of the solution, in accordance with Tikhonov theory of regularization, a stabilizing functional $\Omega(h)$ was introduced in the field of solutions. It reflects the degree of preference of possible functions h. As a result the problem comes to another optimization problem: minimization of mean-square error with stabilization term:

$$F = \sum_{i=1}^{L} (y_i - h(x_i))^2 + \lambda\Omega(h). \qquad (1)$$

Minimization of the functional (1) actually is a trade-off between accuracy of data description (the first term minimization) and prior preferences of solutions (the second term minimization). The regularization parameter λ in this formula defines which of these two terms is more significant.

One of the possible ways to define this parameter is validation approach (*Zhu, 1996*). The data $D = \{y_i, \vec{x_i}\}_{i=1}^{L}$ are divided into two parts: training set D_L, and validation set D_V: $D = D_L \oplus D_V$. As the first step the parameter λ is chosen (perhaps randomly) and minimum of (1) is found for the training data set D_L. Then λ is corrected to minimize the validation error $\sum_{i \in D_V} (y_i - h(x_i))^2$. After that new approximation function is determined using new and so on. That is so-called Expectation Maximization algorithm (EM-algorithm) (*Dempster, 1977*).

The weakness of this method is a strong dependence on the way of division into training D_L and validation D_V parts. To avoid this problem a cross-validation method was suggested (*Zhu, 1996*). In this case the data were divided into training and validation parts many times and each time the procedure of approximation function building

and λ correction is fulfilled. The cross-validation technique provides increase of solution reliability but the expensiveness of the procedure rises too. Indeed, the number of ways of data fragmentation into training and validation parts increases as a factorial of data points number L.

A new method of regularization parameter determination based on the Bayesian formula has become popular since the end of eighties of last century. It is assumed that there is some functional dependence of probability of approximation error on the magnitude of this error $P(\beta E)$. This functional dependence includes β as a parameter. The same assumptions are introduced for prior probabilities of the solutions $P(h) = P(\alpha\Omega(h))$. The solution of the problem is the function which maximizes the product $P(\beta E)P(\alpha\Omega(h))$. It depends on the parameters α and β. In probabilistic paradigm these parameters perform the same role as λ does in (1). Their values are determined using maximum likelihood method. We will discuss Bayesian method in details in the next section.

Then we will describe a new model of approximation using Bayesian regularization. This model has a unique analytical solution for the regularization parameters in difference of the others.

2 Bayesian approach to the regularization problem

In Bayesian paradigm we choose the solution taking into account some terms set by the model H (*MacKay, 1992*). The most probabilistic solution is defined by Bayesian formula:

$$P(h|D,H) = \frac{P(D|h,H)P(h|H)}{P(D|H)}. \qquad (2)$$

In this expression $P(h|D,H)$ is a conditional probability that the solution h is chosen for data D description under the conditions H, $P(D|h,H)$ is a probability of generation of data D by the function h under the conditions H, and $P(h|H)$ is a prior probability of solution h choosing the model H.

Likelihood of the model called *Evidence* is defined by the denominator of Bayesian formula.

$$P(D|H) = \sum_h P(D|h,H)P(h|H) = \sum_h P(D,h|H).$$

Bayesian formula also can be written for models. In this case it compares different prior terms:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}.$$

We have to define priorities for models $P(H)$ to maximize *Evidence* i.e. have to define next-level model. In practice two-levels scheme is usually used. We assume all models are equiprobable:$P(H) = const$. Under this assumption we choose the model with maximal *Evidence*:

$$H_{ML} = arg\max_h P(D|H).$$

After the prior terms are defined we choose the function (2).

3 Prior function selection

Let us suppose the solution depends on some set of fitting parameters $h(\vec{x}) = h(\vec{x},\vec{a})$, where $\vec{a} = \{a_n\}_{n=1}^N$. Some assumptions must be done for the probabilities in the formula (2). In the review (*Zhu, 1996*) Bayesian regularization of interpolation is discussed for various models. The probabilities considered there are:

$$P(D|h,H) = \frac{1}{Z_X} exp(-\alpha E_X),$$

$$P(h|H) = \frac{1}{Z_A} exp(-\alpha\Omega(h)).$$

Here E_X is an approximation error, $\Omega(h)$ is a function of parameters $\vec{a} = (a_1, a_2, \ldots, a_N)$. As a rule the function is symmetrical i.e. it depends on absolute values of $(a_1, a_2, \ldots, a_N)$. Z_X and Z_A are normalization constants. There are two the most popular and simplest stabilization functionals which suit this requirement:

- Gaussian prior $\Omega(h) = \sum_{n=1}^N a_n^2$ (for example, it was used in (*Sollich, 1999*))
- Laplace prior $\Omega(h) = \sum_{n=1}^N |a_n|$ (for example, it was used in (*Poggio, 1998*))

They correspond to probabilities

$$P(h|H) = \frac{1}{Z_A} exp(-\alpha \sum_{n=1}^N a_n^2), \qquad (3)$$

$$P(h|H) = \frac{1}{Z_A} exp(-\alpha \sum_{n=1}^N |a_n|).$$

Below we use Gaussian prior probability function (3). It allows building the scheme of function approximation with some calculation advantages.

4 Approximation by functional basis

In this section algorithm of approximation suggested by authors in (*Nuzhny, 2003*) is described briefly. Let's search the solution $h(\vec{x})$ as series of functions $\{\psi_n(\vec{x})\}_{n=1}^N$:

$$h(\vec{x}) = \sum_{n=1}^{N} a_n \psi_n(\vec{x}).$$

It is also assumed that noise in the data is Gaussian. Then the probability of data point $\{y_i, \vec{x}_i\}$ generation by the solution is

$$P(x_i|h) = \sqrt{\frac{\pi}{\beta}} exp\left(-\beta(y_i - h(x_i))^2\right)$$

and full data set $D = \{y_i, \vec{x}_i\}_{i=1}^L$ generation probability is

$$P(D|h) = \frac{1}{Z_X} exp\left(-\beta \sum_{n=1}^{L} (y_i - h(x_i))^2\right).$$

Let the stabilization term to be of Gaussian form. In this case the prior probability is

$$P(h|H) = \frac{1}{Z_A} exp\left(-\alpha \sum_{n=1}^{N} a_n^2\right).$$

Maximization of the total probability

$$P(h|D, H) = \frac{1}{Z_M} e^{-M}$$

where

$$M = \beta \sum_{i=1}^{L} (y_i - h(x_i))^2 + \alpha \sum_{n=1}^{N} a_n^2$$

equals to minimization of (1) and comes to solving of the system of linear algebraic equations problem:

$$\sum a_m A_{mn} = B_n \tag{4}$$

where

$$A_{mn} = \beta \sum_{i=1}^{L} \psi_{mi}\psi_{in},\ n \neq m,$$

$$A_{nn} = \beta \sum_{i=1}^{L} \psi_{ni}^2 + \alpha,$$

$$B_n = \beta \sum_{i=1}^{L} y_i \psi_{ni}.$$

Now we have to maximize *Evidence* to define the parameters α and β. It is easy to show that

$$P(D|H) = \frac{Z_M}{Z_A Z_X}$$

or

$$ln\ P(D|H) = ln\ Z_M - ln\ Z_A - ln\ Z_X$$

where

$$Z_A = \int_{-\infty}^{\infty} exp\left(-\alpha \sum_{n=1}^{N} a_n^2\right) d^N a_n$$

$$Z_X = \left(\frac{\pi}{\beta}\right)^{\frac{L}{2}}$$

$$Z_M = \int_{-\infty}^{\infty} d^N a_n exp(-\beta \sum_{i=1}^{L} (y_i - \\ - \sum_{n=1}^{N} a_n \psi_{ni})^2 - \alpha \sum_{n=1}^{N} a_n^2) \tag{5}$$

The last integral can be calculated approximately.

$$ln\ P(D|H) = \sum_{n=1}^{N} \Big(\frac{\left(\beta \vec{y} \vec{\psi}_n\right)^2}{A_{nn}} - \frac{1}{2} ln\ A_{nn} - \\ -\beta \sum_{m \neq n}^{N} \vec{\psi}_n \vec{\psi}_m a_n a_m \Big) + \frac{N}{2} ln\ \alpha - \beta \vec{y}^2 + \frac{L}{2} ln\ \beta - \frac{L}{2} ln\ \pi$$

This functional contains the coefficients $\{a_n\}_{n=1}^N$ defined by the parameters α and β. In this case EM-algorithm (*Dempster, 1977*) can be used:

- On the first step we initialize the parameters α and β and calculate the coefficients $\{a_n\}_{n=1}^N$ resolving the system (4).
- On the second step we fix these coefficients and calculate the regularization parameters and so on.

These steps are repeated until $\{a_n\}_{n=1}^N$ are stable.

5 The Case of orthogonal basic vectors

In this section the case when basic functions meet the condition (6) is discussed.

$$\sum_{i=1}^{L} \psi_{mi}\psi_{ni} = \delta_{mn} \tag{6}$$

For example, many wavelet bases defined on regular grid satisfy this term. Normalization of these sums on the unit is an unnecessary condition but it makes calculations easier.If the term (6) is true the algorithm of approximation is simplified essentially:

1. The expensive procedure of system (4) calculation (the number of operations increases as the number of basic functions to the third power N^3) is replaced by $a_n = \frac{B_n}{A_{nn}}$ or taking into account (6)

$$a_n = \frac{\beta}{\beta + \alpha} \sum_{i=1}^{L} y_i \psi_{ni} \tag{7}$$

2. Logarithm of *Evidence* does not include the coefficients a_n:

$$ln\, P(D|H) = \sum_{n=1}^{N} \left(\frac{\left(\beta \vec{y} \vec{\psi}_n\right)^2}{A_{nn}} - \frac{1}{2} ln\, A_{nn} \right) + \\ + \frac{N}{2} ln\, \alpha - \beta \vec{y}^2 + \frac{L}{2} ln\, \beta - \frac{L}{2} ln\, \pi \tag{8}$$

It allows not to use EM-algorithm but to perform a single iteration. Optimal α and β are found by maximizing of (8) and then coefficients (7) are calculated.

3. The expression for logarithm of *Evidence* is computed exactly. The details of computation are in appendix A.

4. Functional (8) has a unique extreme point. It corresponds to the functional maximum and the parameters can be expressed analytically:

$$\alpha = \frac{1}{2} \frac{(L-N)}{\left(\frac{L}{N} S - \vec{y}^2\right)} \tag{9}$$

$$\beta = \frac{1}{2} \frac{(L-N)}{\vec{y}^2 - S} \tag{10}$$

$$\lambda = \frac{\alpha}{\beta} = \frac{\vec{y}^2 - S}{\left(\frac{L}{N} S - \vec{y}^2\right)}$$

where S is given by (14).

For details see appendix B. As a result the algorithm of approximation is built:

1. Calculate basis functions $\{\vec{\psi}_n(x)\}_{n=1}^N$ in points $\{x_i\}_{i=1}^L$

2. Find parameters α and β by formulas (9) and (10).

3. Using these parameters find the coefficients of solution decomposition $\{a_n\}_{n=1}^N$.

The parameters α and β must be positive. Otherwise the expressions for probabilities lose meaning. This requirement is satisfied when $L > N$. The other cases are outside of our model. Indeed if $N \geq L$ matrix (4) is irregular and term (6) is not valid.

6 Examples of approximation

The described algorithm was used for approximation of a model data set on a regular mesh. The points were generated by the function $f(x) = exp\,(x/x_0)^2\,(1 - cos\,(x/10))$, where x_0 - parameter which determines damping. Here this parameter equals the length of variation interval for X variable. The number of the points was $L = 128$. Haar functions (*Daubechies, 1992*) were used as basic functions. They are given by scaling and shifting of the "mother function":

$$\psi(x) = \begin{cases} 1 & 0 \leq x < 1/2 \\ -1 & 1/2 \leq x < 1 \\ 0 & otherwise \end{cases}$$

With the "scaling function"

$$\phi(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

they form an orthonormal basis on a regular mesh.

Accordingly to the developed algorithm the regularization parameters and series coefficients for initial function were calculated. The results of approximation for 16 and 64 basis functions are shown in Fig. 1-2 respectively.
This method can be extended to the case of nonorthogonal basis functions. If term (6) is not true the set of vectors $\{\psi_n(x)\}_{n=1}^N$ can be orthogonalized. Consider matrix Ψ which columns are the basis vectors. Transformation

$$W = \Psi \left(\Psi^T \Psi\right)^{-\frac{1}{2}} \tag{11}$$

gives a new matrix which columns are new orthogonal basis vectors $\{\vec{\omega}_n\}_{n=1}^N$ (*Hyvarinen, 2000*). We apply the above described algorithm for this new basis. The final decision can be written as

$$h(x) = \sum_{n=1}^{N} \tilde{a}_n \psi_n(x)$$

where

$$\vec{\tilde{a}} = \vec{a} \left(\Psi^T \Psi\right)^{-\frac{1}{2}}.$$

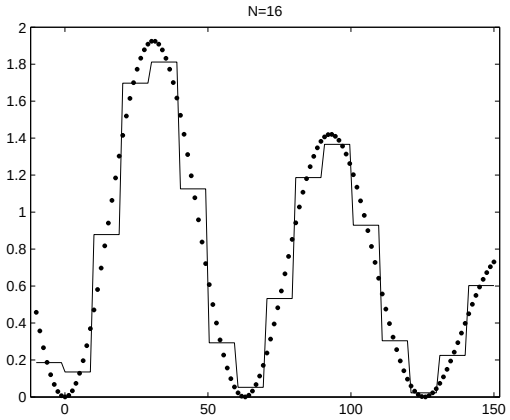

Figure 1: The results of approximation with Haar wavelets for $N = 16$

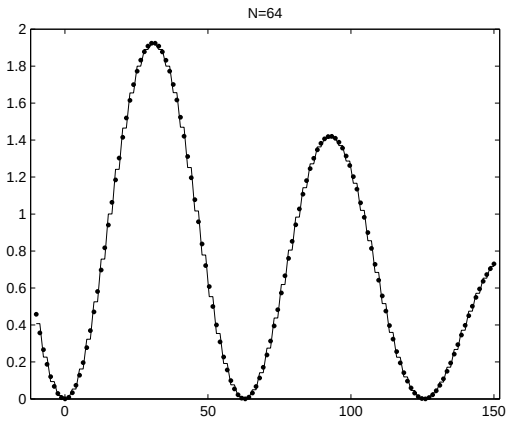

Figure 2: The results of approximation with Haar wavelets for $N = 64$

The inverse matrix in equation (11) was defined by Gauss-Jordan elimination (*Strang, 2003*). In the case of singular matrix $\Psi^T\Psi$ the number of basic functions was decreased to avoid singularity.

The results of application of the modified algorithm are shown in Fig. 3. The data set defined on irregular mesh was approximated by Fourier basis $\left\{1, \sin\frac{2\pi}{x_0}kx, \cos\frac{2\pi}{x_0}kx\right\}_{k=1}^{(N-1)/2}$.

The algorithm described above was applied for two-variable data set approximation. This data was given on irregular mesh. Radial Basis Functions (RBF) $\psi_n = \exp\left(-\frac{(\vec{x}-\vec{x}_n)^2}{2\sigma_n^2}\right)$ (*Buhmann, 2003*) were used as a function basis. The data were

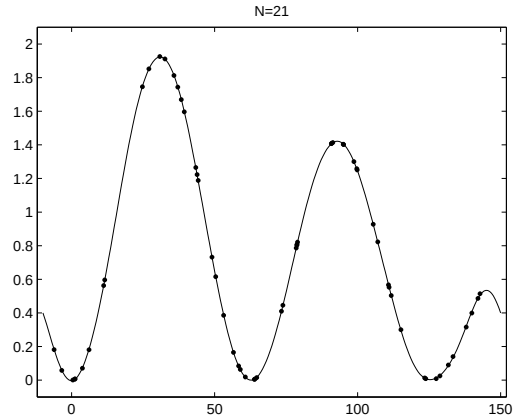

Figure 3: The results of approximation of the data set defined on irregular mesh using Fourier basis and $N = 21$

clustered using Kohonen maps algorithm (*Kohonen, 1982*). The centers of clusters were selected as centers $\vec{x}_n$ of RBFs. Parameters σ_n were determined as standard deviations for corresponding clusters

$$\sigma_n = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\vec{x}_j - \vec{x}_n)^2}$$

where $\vec{x}_j,\ j = \overline{1, J}$ are representatives of the n-th cluster.

The result is shown in Fig. 4.

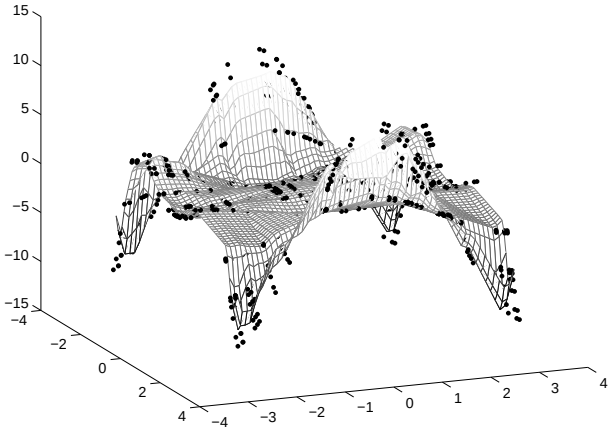

Figure 4: The result of RBF network approximation

7 Conclusions

The algorithm of point-by-point function approximation was discussed. The solution was being searched as a series of basis functions. For regularization parameter determination Bayesian method was used. The proposed algorithm gives unique analytical solution for regularization parameters and approximation function. Besides it does not use any iteration procedures like the gradient methods or EM-algorithm. That is why it can have computing advantages compared with the others.

Appendix A

Using (5) the following formula can be obtained:

$$Z_M = \int d^N a_n exp - \beta\vec{y}^2 + 2\beta \sum_{n=1}^{N} \vec{y}\vec{\psi}_n a_n - \\ - \sum_{n=1}^{N} A_{nn} a_n^2$$

$$Z_M = \int d^N a_n exp - \beta\vec{y}^2 - \\ \sum_{n=1}^{N} A_{nn} \left(a_n^2 - 2\frac{\beta\vec{y}\vec{\psi}_n}{A_{nn}} a_n \right)$$

$$Z_M = \int d^N a_n exp - \beta\vec{y}^2 + \sum_{n=1}^{N} \frac{\left(\beta\vec{y}\vec{\psi}_n\right)^2}{A_{nn}} - \\ \sum_{n=1}^{N} A_{nn} \left(a_n - \frac{\beta\vec{y}\vec{\psi}_n}{A_{nn}} \right)^2$$

$$Z_M = exp \left(-\beta\vec{y}^2 + \sum_{n=1}^{N} \frac{\left(\beta\vec{y}\vec{\psi}_n\right)^2}{A_{nn}} \right) \\ \int d^N a'_n exp \left(- \sum_{n=1}^{N} A_{nn} a'^2_n \right)$$

where

$$a'_n = a_n - \frac{\beta\vec{y}\vec{\psi}_n}{A_{nn}}$$

After integration, substitution all normalization terms in (4) and finding the logarithm expression (8) is obtained.

Appendix B

Differentiate functional (8) over parameters α and β and get the set of equations to determine the values of these parameters:

$$\frac{\partial \ln\ P(D|H)}{\partial\alpha} = -\sum_{n=1}^{N} \frac{(\beta\vec{y}\vec{\psi}_n)^2}{A_{nn}^2} + \frac{N\beta}{2\alpha A_{nn}} = 0,$$

$$\frac{\partial \ln P(D|H)}{\partial\beta} = \sum_{n=1}^{N} \left(\frac{2\beta(\vec{y}\vec{\psi}_n)^2}{A_{nn}} - \frac{\left(\beta\vec{y}\vec{\psi}_n\right)^2}{A_{nn}^2} \right) - \\ - \frac{N}{2A_{nn}} - \vec{y}^2 + \frac{L}{2\beta}. \quad (12)$$

From the first one it can be found

$$\frac{\beta S}{A_{nn}} = \frac{N}{2\alpha} \quad (13)$$

where

$$S = \sum_{n=1}^{N} \left(\vec{y}\vec{\psi}_n\right)^2. \quad (14)$$

Then substituting expression (13) in (12) the following term can be calculated:

$$\frac{N}{2\alpha} - \vec{y}^2 + \frac{L}{2\beta} = 0.$$

Using this expression and equation (13) one can obtain the parameter (10).

Reference

Tikhonov A. N. (1963). "Solution of incorrectly formulated problems and the regularization method" Soviet Math Dokl. vol.4 pp. 1035-1038

Tikhonov A. N., V. Y. Arsenin (1977). *Solutions of Ill-posed Problems* Winston: Washington, DC

Zhu H. and R. Rohwer (1996) "No free lunch for cross-validation" Neural Computation. vol. 8. no. 7. pp. 1421-1426.

Dempster A. P., N. M. Laird and D. B. Rubin (1977) "Maximum likelyhood from incomplete data via the EM algorithm" Journal of the Royal Statistical Society (B). vol. 39. no. 1. pp. 1-38.

MacKay D. (1992) "Bayesian interpolation" Neural Computation. vol. 4. no. 3. pp. 415-447.

Sollich P. (1999) "Probabilistic interpretations and Bayesian methods for Support Vector Machines" Artificial Neural Networks Conference Publication No. 4700 IEE, 7 - 10 September 1999.

Poggio T., F. Girosi (1998) "A sparse representation for Function approximation" Neural Computation. vol. 10. no. 6. pp. 1445-1454.

Nuzhny A., S. Shumsky (2003) "Bayesian regularization in the problem of multivariable function approximation" Mathematical modeling. vol. 15. no. 9. pp. 55-63.

Hyvarinen A., J. Karhunen, E. Oja (2000) *Independent Component analysis* A Wiley-Interscience Publication: New York.

Strang G. (2003) *Introduction to Linear Algebra* Wellesley-Cambridge Press: Wellesley.

Buhmann M. D. (2003). *Radial Basis Functions: Theory and Implementations* Cambridge University Press: Cambridge.

Daubechies I. (1992). *Ten Lectures on Wavelets* Society for Industrial and Applied Mathematics: Philadelphia

Kohonen T. (1982). "Self-organized formation of topologically correct feature maps" Biol. Cybernetics. vol.43. pp. 56-69.