Approximation with Diffusion-Neural-Network

Zoltán Makó

Department of Mathematics and Informatics, Sapientia University
530104 – Csíkszereda, Romania
e-mail: makozoltan@sapientia.siculorum.ro

Abstract: Neural information processing models largely assume that the samples for training a neural network are sufficient. Otherwise there exist a non-negligible error between the real function and estimated function from a trained network. To reduce the error in this paper we suggest a diffusion-neural-network (DNN) to learn from a small sample. First, we show the principle of information diffusion using properties of quasi-triangular fuzzy numbers. After that, we apply this principle to construct the DNN. Finally, we give an example to show that the approximation with DNN is better than the conventional back propagation network.

Keywords: artificial neural network, information diffusion, fuzzy number

1 Introducion

Generally, the data are facts characterizing the phenomena of the real world and information is such a structured sample of these datas that helps the exploration of phenomena. In many cases the data are only a part of the facts so the information deducted from them is uncertain. For example: if there are only few observations to the examination of a phenomenon then the information concluded will be uncertain.

If there are only few data available in the examination of a phenomenon we can assign these to some already existing statistical distribution (the Bayes method, Martiz and Lewin, 1989) The structured sample will have an informational value. The question arises: what to do in the case when we don't know a priori statistical distribution?

Many successful solutions of the practical problems show that in such a case the theory of *fuzzy sets* can be applied with a very good efficiency. This theory enables the processing of uncertain information, to be more precise, it writes down the fuzzy logical assertions in an exact mathematical form (L. A. Zadeh, 1975).

In a given space X let A be an sample of data that come from the observation of some phenomenon. Our aim is to search for a real relation among the state parameters of the phenomenon with the help of the data. Let us denote the real relation with R. The method that defines the R from sample A is called an *operator*. Examples for operators: data series analysis, correlation examination, hypothesis examination, the method of artificial neural networks, etc.

Since the membership functions of the fuzzy sets diffuse the information among the fuzzy information sets, the methods (operators) that search for relations with the help of such membership functions are called information diffusion methods.

The present paper deals with the diffusion principle known from the theory of the fuzzy sets and the application of the principle. In the first part we explain the basic concepts of fuzzy sets, fuzzy numbers and quasi-triangular fuzzy numbers with the help of the triangular norm. In the second part we prove the principle of information diffusion with the help of the quasi-triangular fuzzy numbers. In the third part we apply the principle of information diffusion to artificial neural networks.

2 Preliminaries

This section reviews the definitions and basic propositions applied in this paper.

Let X be a set. The mapping $\mu : X \to [0, 1]$ is called *membership function*, and the set $A = \{ (x, \mu(x)) \ / \ x \in X\}$ is called *fuzzy set* on X. The membership function of A is denoted by μ_A .

A mapping $T : [0,1] \times [0,1] \to [0,1]$ is a *triangular norm* if T is symmetric, associative, non-decreasing in each argument and $T(x, 1) = x$, for all $x \in [0,1]$.

A triangular norm T is said to be *Archimedean* if T is continuous and $T(x, x) < x$, for all $x \in [0,1]$.

Every Archimedean triangular norm is representable by a continuous and decreasing function $g : [0,1] \to [0,+\infty]$ with $g(1) = 0$ and $T(x, y) = g^{[-1]}(g(x)+ g(y))$, where

$$g^{[-1]}(x) = \begin{cases} g^{-1}(x) & \text{if } 0 \le x \le g(0), \\ 0 & \text{if } x > g(0). \end{cases}$$

Let $p \in [1, +\infty]$ and $g : [0, 1] \to [0, +\infty]$ be a continuous, strictly decreasing function with boundary properties $g(1) = 0$ and $\lim_{t\to 0} g(t) = g_0 \le +\infty$. The *set of quasi-triangular fuzzy numbers* is

$$N_g = \{A \in F(\mathrm{R}) \;\; / \text{ there is } a \in \mathrm{R} \text{ and } d > 0 \text{ such that}$$

$$\mu_{\mathrm{A}}(t) = g^{[-1]}\left(\frac{|t-a|}{d}\right), \text{ for all } t \in \mathrm{R}\}$$

$$\bigcup \{A \in F(\mathrm{R}) \;\; / \text{ there is } a \in \mathrm{R} \text{ such that } \mu_{\mathrm{A}}(t) = \chi_{\{a\}}(t), \text{ for all } t \in \mathrm{R}\},$$

where χ_A is characteristic function of the set A. The elements of N_g will be called *quasi- triangular fuzzy numbers* generated by g with *centre a* and *spread d* and we will denote them by $< a, d >$ (M. Kovács, 1992).

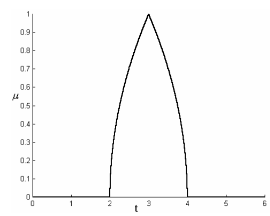

Figure 2.1
Quasi-triangular fuzzy number $< 3, 1 >$ if $g(t) = 1 - t^2$

As follows from the definition of T_{gp}-*Cartesian product,* the membership function of quasi-triangular fuzzy numbers pair $(< a, d >, < b, e >)$ is

$$\mu_{(<a,d>,<b,e>)}(x,y) = T_{gp}\left(\mu_{<a,d>}(x), \mu_{<b,e>}(y)\right) =$$
$$= g^{[-1]}\left(\left[g^{p}\left(g^{[-1]}\left(\frac{|x-a|}{d}\right)\right) + g^{p}\left(g^{[-1]}\left(\frac{|y-b|}{e}\right)\right)\right]^{1/p}\right) \qquad (2.1)$$

for all $(x, y) \in \mathrm{R} \times \mathrm{R}$.

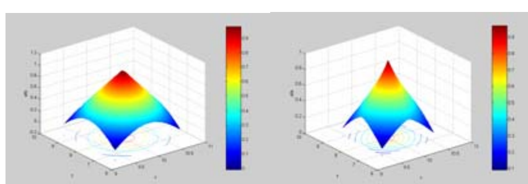

Figure 2.2
The quasi-triangular fuzzy numbers pair $(< 10, 1 >, < 8, 2 >)$ if $g(t) = 1 - t$, $p = 2$ and $p = 1.5$ respectively

Proposition 2.1 (Z. Makó, 2002**)** If $p \in [1, +\infty)$, $d > 0$ and $e > 0$, then the α level of quasi-triangular fuzzy numbers pair $(< a, d >, < b, e >)$ is

$$[(< a,d >,< b,e >)]^{\alpha} = \left\{ (x,y) \in R \times R / \frac{|x-a|^p}{d^p} + \frac{|y-b|^p}{e^p} \leq g^p(\alpha) \right\}$$

and if $p = +\infty$, then

$$[(< a,d >,< b,e >)]^{\alpha} = \left\{ (x,y) \in R \times R / \max\left\{ \frac{|x-a|}{d}, \frac{|y-b|}{e} \right\} \leq g(\alpha) \right\}$$

3 The Diffusion of Information

Let A be an sample of data in a given normed space X that, come from the observation of some phenomenon. Let us denote a real relation with R. The method that defines the R from sample A is called an *operator*. The set of all operators we denote by Γ.

Definition 3.1 Let R be a real relation in X. The sample A is a *correct-data set* to R on universe $U \subset X$ if there exists an operator γ such that we can obtain a relation $R(\gamma, A)$ equal to the restriction of R at U.

Definition 3.2 Let R be a real relation in X. The sample A is an *incomplete-data set* to R on universe $U \subset X$ if there doesn't operator that we can obtain the restriction of R at U from A.

The concept of incompleteness has captured very many bright minds. According to the monumental work of B. Russell and A. N. Whitehead, published between 1910 and 1915, the Principia Matematica logic meant the certain, unambiguous foundation of mathematics. They were wrong, since K. Gödel's theory of incompleteness from 1931 (Gödel, 1931) says: "all the axiomatic formulations of number theory contain uncertain statements."

Is there – even if only in theory – such a method (algorithm) that can solve all mathematical problems? – A. Turing asked the question rhyming with Gödel's theorem. Analising logic-based methodological processes performed by man and the functio of a theoretical computer he reached the conclusion that such algorithm did not exist. A. Church, American logician reached the same standpoint. The Church-Turing theorem says that every problem that can be solved with an algorithm with a finite procedure can also be calculated with the Turing machine.

G. J. Chaitin (1990) examined the concept of incompleteness from the aspect of coincidence and showed that, the breakpoint probability of the non deterministic Turing machine program is algorithmically randomized.

According to A. Turing's and Chaitin's ideas incompleteness and uncertainty are correlated. Let's see the two bit series below as an example:

01010101010101010101   01101100110111100101

One of the questions arising: are these series randomized? The answer is ambiguous, they may be, but it is possible to construct such an algorithm (and even more algorithms) for both series that will return its first 20 bits identical with the given series. Moreover, may we continue the series in any way, theoretically there is an algorithm that return the continuation. So the incompleteness of the series makes the answer uncertain. Algorithms in their own do not decide mathematical truth. The validity of the algorithms can always be stated with the help of external devices.

3.1 Characteristic Function of the Sample

Definition 3.3 Let $A = \{\mathbf{x}_k \,/\, k = 1,2,...n\}$ be a deterministic sample in universe $U \subset X$. The *characteristic function* of A is $\chi_A : A \times U \to \{0,1\}$, where $\chi_A(\mathbf{x}_k, \mathbf{u}) = 1$ if $\mathbf{u} = \mathbf{x}_k$ and $\chi_A(\mathbf{x}_k, \mathbf{u}) = 0$ if $\mathbf{u} \neq \mathbf{x}_k$. The relation derived from sample A and operation γ we denote by $R(\gamma,A)$.

For example, if we inted to determine the elastic constant of a spring then we do measurements. We measure the relative elongation $\Delta l = l - l_0$ and the spring force induced by elongation Δl. In this case the sample is $A = \{(\Delta l_i, F_i) \in X / i = 1,2,...,n\}$ and $U \subset X = \mathrm{R}^2$. Let γ be the least squares method, and the searched real relation $R(\gamma,A)$ is the Hook's formula: $F = k \cdot \Delta l$. The function $f(k) = \sum_{i=1}^{n} (k\Delta l_i - F_i)^2$ achieves the minimal value at $k = \sum_{i=1}^{n} F_i \Delta l_i \,/\, \sum_{i=1}^{n} \Delta l_i^2$. Theoretically, the sample A is correct data-set even if $i = 1$, because from Hook's formula it possible to determine the value of k with the first measurement.

3.2 Scattering Function of the Information

Definition 3.4 We consider a division U_j, $j = 1,...,$m of universe U, i.e.

$$U = \bigcup_{j=1}^{m} U_j, \; U_j \cap U_k = \varnothing \text{ if } j \neq k.$$

The *characteristic function of the division* U_j is $\chi_m : A \times U \to \{0,1\}$, where $\chi_A(\mathbf{x}_k, \mathbf{u}) = 1$ if $\mathbf{x}_k \in U_j$ and $\chi_A(\mathbf{x}_k, \mathbf{u}) = 0$ if $\mathbf{x}_k \notin U_j$, for all $\mathbf{u} \in U_j$.

The characteristic function is replaceable with membership function $\mu : A \times U \to [0,1]$. In this case, the value $\mu(\mathbf{x}_k, \mathbf{u})$ shows that the sample's element $\mathbf{x}_k$ how much is in set U_j. For example, if X = R then the membership function of quasi-triangular fuzzy number

$$\mu(x_k, u) = g^{[-1]}\left(\frac{|x_k - u|}{d}\right)$$

is a membership function of division to interval with centre x_k and length $2d$.

Therefore, $\mu(:,\mathbf{u}_j)$ is the membership function of U_j, for all $\mathbf{u}_j \in U_j$.

Definition 3.5 The family of membership functions μ (:, $\mathbf{u}_j$) : $U \rightarrow [0,1]$, j = 1,...,m is a *fuzzy division* of the U.

Since, the membership functions μ (:, $\mathbf{u}_j$) diffuse the information $\mathbf{x}$ among the fuzzy sets U_j, hence the relations searching methods (operators) that use these membership functions will be called *information-diffusion methods* (operators).

Definition 3.6 Let A be a sample of universe U. The function $\mu : A \times U \rightarrow [0,1]$ is a *scattering function of the information*, if

i) $\mu(\mathbf{x}_k, \mathbf{x}_k) = 1$, for all $\mathbf{x}_k \in A \cap U$;

ii) for all $\mathbf{x}_k \in A$ and for all $\mathbf{u},\mathbf{v} \in U$, if $\|\mathbf{x}_k - \mathbf{u}\| \leq \|\mathbf{x}_k - \mathbf{v}\|$ then

$\mu(\mathbf{x}_k, \mathbf{u}) \geq \mu(\mathbf{x}_k, \mathbf{v})$.

For all elements $\mathbf{x}_k$ of the sample A the scattering function define a fuzzy number with centre in $\mathbf{x}_k$ and membership function $\mu(\mathbf{x}_k,:):U \rightarrow [0,1]$. The simplest scattering function is $\mu_t = \chi$. This function will be called *trivial scattering function* of the information.

The scattering function of the information shows that the data $\mathbf{u}$ how much can be the correct data of a phenomena. For example, if $\mathbf{u}$ is in sample A then $\mathbf{u}$ is absolutely correct data of the phenomena. Using the scattering function μ, the sample A can be expand with new elements and so we get a sample notated by $A(\mu,U)$ with elements $(\mathbf{x}_k, \mathbf{u}_j, \mu(\mathbf{x}_k, \mathbf{u}_j,)) \in A \times U \times [0,1]$, where $\mathbf{u}_j \in U$, j = 1,...,p.

If $X = R^n$ then it possible to define scattering function with help of quasi-triangular fuzzy numbers. Let $p \in [1, +\infty]$ and $g : [0, 1] \rightarrow [0, +\infty]$ be a continuous, strictly decreasing function with boundary properties $g(1)$=0 and $\lim_{t \to 0} g(t) = g_0 \leq +\infty$. The triangular norm generated by g^p is $T_{gp}(x,y)$ given by formula (2.1). If we fuzzyfied all elements of sample A, i.e. for all components x_{ki} of vector $\mathbf{x}_k \in A$ we assign a quasi-triangular fuzzy number $< x_{ki}, \lambda(x_{ki})>$ with spread $\lambda(x_{ki}) > 0$, i = 1,...,n.

As follows from the definition of T_{gp}-Cartesian product (2.1) the scattering function of information is given by

$$\mu\left((x_{k1},x_{k2},\ldots,x_{kn}),(u_1,u_2,\ldots,u_n)\right)=$$
$$\begin{cases} g^{-1}\left(\left[\left(\dfrac{|x_{k1}-u_1|}{\lambda(x_{k1})}\right)^p+\ldots+\left(\dfrac{|x_{kn}-u_n|}{\lambda(x_{kn})}\right)^p\right]^{1/p}\right) & if \quad |u_i-x_{ki}|\le\lambda(x_{ki})g_0,\ i=\overline{1,n}, \\ & otherwise. \end{cases} \tag{3.1}$$

3.3 The Principle of Information Diffusion

Let R be a relation on universe $U\subset X$ and γ be an operator. If we are using the sample $A=\{\mathbf{x}_k\in X/\ k=1,2,\ldots,n\}$ to estimate the relation R then our *method is a nondiffusion estimator*, and if we are using the sample $A(\mu,U)$, where μ is a nontrivial information scattering function, then our method is a *diffusion estimator*. The trivial information scattering function yield a nondiffusion estimator.

Theorem 3.1 (Principle of information diffusion) Let R be a relation on universe $U\subset X=R^n$, where U is a convex set. Let $A=\{\mathbf{x}_k\in X\ /\ k=1,2,\ldots,n\}$ be a deterministic sample for estimation of R on universe $U\subset X$. Assuming that γ is the best operator of relation R for some measurement of the error. The sample A is incomplete-data set of the relation R on U if and only if there exist an nontrivial information scattering function μ such that, if we apply the operator γ to fuzzyfied sample $A(\mu,U)$ then we get a better estimation of R.

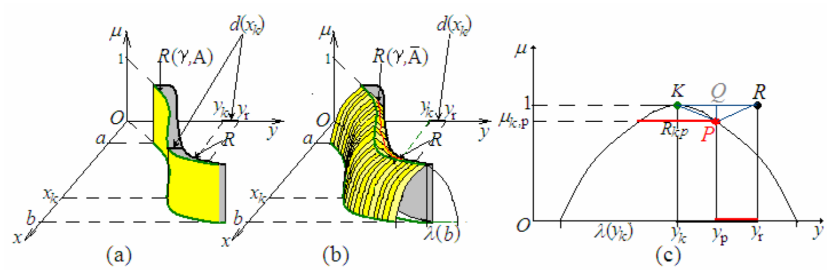

Figure 3.1
Principle of information diffusion

Proof. For the intelligibility we proof this theorem in $X=R^2$. The proof of the general case is similar. In this case $U=[a,b]\times R$ and the relation R is a subset of U. The first component of the R relation's elements is input (independent) and the second component is output (dependent) variable of R. In figure (3.1) the relation R is showed by a curve. Let $A=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ be a sample in U. If A is incomplete-data set of the relation R then for all operator γ the difference between R and $R(\gamma,A)$ is greater than zero. The error is

$$\sup\left\{ \left[(x'-x)^2 + (y'-y)^2\right]^{1/2} / (x,y)\in R \text{ and } (x',y')\in R(\gamma,A) \right\} = \varepsilon > 0. \quad (3.2)$$

We diffuse the information derived from sample A using the scattering function (3.1). Therefore, we get the fuzzyfied sample

$$\overline{A} = A(\mu,U) = \left\{ \left(< x_k, \lambda(x_k) >, < y_k, \lambda(y_k) >\right) / k = 1,\ldots,n \right\}, \quad (3.3)$$

where $\left(< x_k, \lambda(x_k) >, < y_k, \lambda(y_k) >\right)$ is a quasi-triangular fuzzy numbers pair with $\lambda(x_k)$, $\lambda(y_k) > 0$. For p =2 the scattering function is

$$\mu((x_k,y_k),(u,v)) = \begin{cases} g^{-1}\left(\left[\left(\frac{|x_k - u|}{\lambda(x_k)} \right)^2 + \left(\frac{|y_k - v|}{\lambda(y_k)} \right)^2 \right]^{1/2} \right) & \text{if } \quad |u - x_i| \le \lambda(x_k) g_0, \text{ and } |v - y_i| \le \lambda(y_k) g_0, \\ & \text{otherwise.} \end{cases}$$

(3.4)

It follows from proposition 2.1 that the level sets of μ are ellipses. Our problem is that: to determine the spreads $\lambda(x_k)$ and $\lambda(y_k)$ such that the approximation with operator γ and $\overline{A}$ let be less than ε, i.e. the difference between level set μ_{kp} and curve R is less than ε (see figure 3.1. c). The graph (a) of figure 3.1 shows, that, if we apply the operator γ to sample A then all elements of relation $R(\gamma,A)$ absolutelly is in $R(\gamma,A)$ (the membership value to $R(\gamma,A)$ of all elements are one). Consequently, all level sets are equal. The graph (b) of figure 3.1 shows that, if we apply the operator γ to sample $\overline{A}$ then all elements of relation $R(\gamma, \overline{A})$ are quasi-triangular fuzzy numbers pair. We will denote by R_{kp} the μ_{kp} level set of $R(\gamma, \overline{A})$. We show the projection of $R(\gamma, \overline{A})$ on graph (c) of figure 3.1. As shown in the illustration, the projection of error to axis of reference Oy is $| y_r - y_k | = f$, similarly the projection of error to axis of reference Ox is $| x_r - x_k | = e$. We select an element $P(x_p,y_{\text{p}})$ in R_{kp} such that the distance between R and P is less than the distance between R and K. We denote by $l = | y_p - y_r |$ and by $h = | x_p - x_r |$. Then $\mu_{kp}= \mu((x_k,y_k),(x_p,y_p))$ and

$$g^2(\mu_{kp}) = \left(\frac{e-h}{\lambda(x_k)} \right)^2 + \left(\frac{f-l}{\lambda(y_k)} \right)^2$$

If we consider that $\lambda(x_k) = \lambda(y_k)$, then

$$\lambda(x_k) = \lambda(y_k) < \frac{\varepsilon}{g(\mu_{kp})}. \quad (3.5)$$

It possible to select

$$\lambda(x_k) = \lambda(y_k) = \frac{\varepsilon}{2g(\mu_{kp})}.$$

If A is correct-data set of relation R then ε = 0. Therefore, from (3.5) follows that the spread of all quasi-triangular fuzzy numbers are zero. In this case the information scattering function is trivial. □

4 The Approximation Property of BP Artificial Neural Network

The neural network can be understood as a mapping $f : R^n \to R^m$, defined by $\mathbf{y} = f(\mathbf{x}) = g(W \cdot \mathbf{x})$, where **x** is the input vector, **y** is the output vector, W is the weight matrix and g is the activation function. The mapping f can be decomposed into a chaining of mappings; the result is a multi-layer network $R^n \to R^p \to R^q \to \ldots \to R^m$. The algorithm for computing W is often called the training algorithm. The most popular neural network are the multi-layer back-propagation networks whose training algorithm is the well-known gradient descendent method. Such networks are called back-propagation (BP) networks.

An artificial neural network is a learning machine whose function depends on the training examples. So, the machine does not recognize the real relation but it determine a numerical relation among the state parameters. According to the principle of information diffusion we can increase the certainty of the determined relation if we multiply the number of the training examples with the help of an appropriate information scattering function or if we apply a banded approach. Neural networks trained in this manner are called *diffusion-neural-networks* (C.F. Huang and Y. Shi, 2002; C. F. Huang and C. Moraga, 2004).

A number of authors have discussed the universal approximation property of BP networks. For example, in 1989 G. Cybenko showed that any f: $[a,b] \to R$ continuous function can be approximated by a neural network with one internal hidden layer using sigmoidal activation function (G. Cybenko, 1989). Also in 1989 K. Hornik et al. proved that the multi-layer networks can approximate the continous function to any degree of accuracy, i.e. multi-layer networks have the universal approximation property (K. Hornik et al., 1989). After that, in 1995 J. Wray and G. G. R. Green showed that, the universal approximation property does not hold in practice for networks implemented on computers (J. Wray and G. G. R. Green, 1995). For illustration this property, we consider the function f : $[0,1] \to R$, $f(x)=x-x^2$. Our task is to learn the function with the sample A = {(0,0),(0.25,0.1875),(0.5,0.25),(0.75,0.1875),(1,0)}. It is easy to show by example that, the BP networks, for any topology, can not approximate the function f to any degree of accuracy. In our experiments the minimal value of difference between function f and function defined by BP network is greater than 0.001, for any topology of network (fig. 4.1). The approximation is not efficient because the sample A is incomplete-data set to relation f. The approximation with BP network is efficient if A is correct-data set to f.

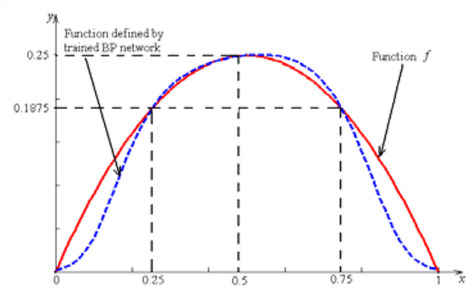

Figure 4.1
Approximation with BP network

From information diffusion principle follows that, the accuracy of approximation is increasable if we use a scattering function μ to diffuse the information derived from sample A.

4.1 Banded Approximation

Let $f : [a,b]\rightarrow R$ be a given continuous function and $A=\{(x_k,f(x_k)) \in [a,b]\times R \;/\; k=1,2,...,n\}$ be a given sample. We diffuse the information derived from this sample with the generator function $g:[0,1]\rightarrow[0,\infty]$, where $g(1)=0$ and $\lim_{t\to\infty} g(t) = g_0 \leq \infty$. Thus, we obtain the fuzzyfied

$$\overline{A} = \{(<x_k,\alpha_k>,<f(x_k),\beta_k>)/\,k=1,...,n\},$$

sample, where $\alpha_k, \beta_k \geq 0$ are the spread of quasi-triangular fuzzy numbers $<x_k, \alpha_k>$ and $<f(x_k), \beta_k>$. Above derivative sample can be used to train a conventional BP-network with two input value x_k and α_k, and two output value $o_1(x_k, \alpha_k)$ and $o_2(x_k, \alpha_k)$. After the training we get a weight system where

$$H = \sum_{k=1}^{n}\left[(f(x_k) - o_1(x_k,\alpha_k))^2 + (\beta_k - o_2(x_k,\alpha_k))^2\right] \qquad (4.1)$$

the sum of square errors is less than a given number $\delta > 0$. The trained network for any input values x and 0 return the output values y and β. Using the generator function g, we can construct a band $[y\text{-}\, g(\gamma)\beta, y+g(\gamma)\beta]$ around to function f for any level value $\gamma\in[0,1]$. The approximation has precision ε on the level γ, if the distance between f and γ level set $[y\text{-}g(\gamma)\beta\,,\, y+g(\gamma)\beta]$ is less than ε.

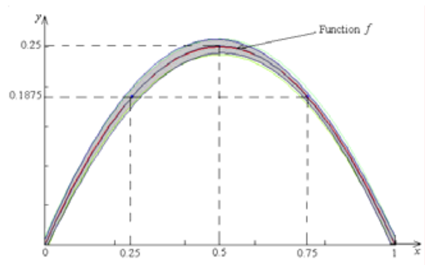

Figure 4.2
The banded approximation

4.2 Approximation with Derivative Sample

Let f : $[a,b]\to R$ be a given continuous function and $A=\{(x_k, f(x_k)) \in [a,b]\times R$ / $k=1,2,...,n\}$ be a given sample. We diffuse the information derived from this sample with the generator function $g:[0,1]\to[0,\infty]$, where $g(1)=0$ and $\lim_{t\to\infty} g(t) = g_0 \leq \infty$. If we consider two points $x_k - \delta$ and $x_k + \delta$ around on points x_k then the membership value of these points to fuzzy set B =proj$_{Ox}(A)$ are

$$\mu_k = \mu_B(x_k - \delta) = \mu_B(x_k + \delta) = \begin{cases} g^{-1}\left(\dfrac{\delta}{\lambda(x_k)}\right) & if \quad \delta \leq g_0\lambda(x_k), \\ 0 & otherwise. \end{cases} \tag{4.2}$$

Since, in practice δ is relative small therefore $f(x_k \pm \delta) \approx f(x_k) \pm \delta$ and

$$\mu_k = \mu_{proj_{Oy}(A)} f(x_k \pm \delta).$$

Consequently, the derived sample is

$$\overline{A} = \{(x_k, 1, y_k, 1), (x_k \pm \delta, \mu_k, f(x_k) \pm \delta, \mu_k) / k = 1,...,n\}.$$

Above derivative sample can be used to train a conventional BP-network, where the first two components are input values and second two components are output value of BP network. We show the absolute accuracy of the approximation with samples A and $\overline{A}$ on the figure 4.3. We can see that, the approximation with $\overline{A}$ is better than with A. The average of square errors (ASE) is

$$E = \frac{1}{p}\sum_{k=1}^{p}\left(f(x_i) - o(x_i)\right)^2,$$

where $x_i \in [0,1]$, $I = 1,\ldots,p$ are the check-test points and $o(x_i)$ are the values calculated by network if the input value is x_i. In our case the ASE for sample A is $E(A) = 0.000002051706$ and for sample $\overline{A}$ is $E(\overline{A}) = 0.0000005069$.

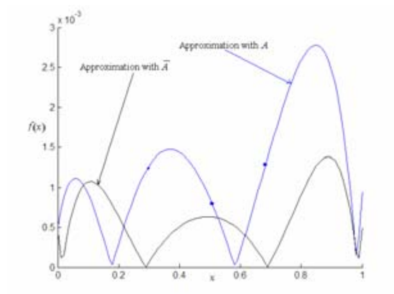

Figure 4.3
Approximation with derivative sample

The above example shows that, the approximation is better with derived sample than with original sample. The principle of information diffusion pronounces this fact, but not gives the method of expansion.

References

[1] G. J. Chaitin: *Information, Randomness & Incompleteness*, Word Scientific, Singapore, 1990

[2] G. Cybenko: Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals, and Systems*, 2(4) (1989) 303-314

[3] K. Gödel: Über formal unentsceidbare Sätze der Principia Mathematica und Verwandter Systeme I., *Monatsshefte für Mathematik und Physik*, Vol. 38 (1931), 173-198

[4] K. Hornik, M. Stinchcombe, H. White: Multilayer feedforward networks are universal approximators, *Neural Networks*, 2(5) (1989), 359-366

[5] C. F. Huang and Y. Shi: *Towards Efficient Fuzzy Information Processing – Using the Principle of Information Diffusion*, Physica-Verlag, Heidelberg, 2002

[6] C. F. Huang and C. Moraga: A diffusion-neural-network for learning from small samples, Internat. *J. of Approx. Reason*. 35 (2004), 137-161

[7] M. Kovács: A stable embedding of ill-posed linear systems into fuzzy systems, *Fuzzy Sets and Systems* 45 (1992) 305-312

[8] Z. Makó: The opposite of quasi-triangular fuzzy number, *Proceedings of the 3rd Computational Intelligence Conference, Budapest* (2002), 229-238

[9] J. Wray and G. G. R. Green: Neural networks, approximation theory, and finite precision computation, *Neural Networks*, 8(1) (1995) 31-37

[10] L. A. Zadeh: The concept of a linguistic variable and its application to approximate reasoning, Parts 1 and 2, *Inform. Sci.* 8,9 (1975), 199-249, 301-357