Neural Network for Estimating Conditional Distribution

Schiøler, Henrik; Kulczycki, P.

NEURAL NETWORK FOR ESTIMATING CONDITIONAL DISTRIBUTIONS

Henrik Schioler, Piotr Kulczycki

Both authors are with the Department of Control Engineering at Aalborg University, Aalborg, Denmark. Piotr Kulczycki is currently on leave from Cracow University of Technology, Poland. Henrik Schioler can be reached by E-mail at: henrik@control.auc.dk

Abstract

Neural networks for estimating conditional distributions and their associated quantiles are investigated in this paper. A basic network structure is developed on the basis of kernel estimation theory, and consistency is proved from a mild set of assumptions. A number of applications within statistics, decision theory and signal processing are suggested, and a numerical example illustrating the capabilities of the elaborated network is given.

Keywords

Neural Networks, Conditional Distributions, Kernel Estimation, Optimal Control, Data Transmission.

I. Introduction

Relationships between random variables are most often described by characteristic parameters such as mean vectors and covariance matrices or in extraordinary cases moments of higher order. When standard situations are considered, for example, if all variables are jointly Gaussian or when the conditional characteristics are linear or low degree polynomial functions, that approach is to be recommended. On the other hand when the situation is far from the above, more general methods should be considered. Statistical relationships are completely described by the joint distribution of all the variables in consideration, however in some cases it is appropriate to partition the considered random variables into two groups. One group of so called explanatory variables yielding information about or explaining the variables in the second group. From that point of view conditional distributions as well as their associated quantiles are the objects of interest. In some cases such conditional distributions are sufficiently precisely described by standard expressions with only a low number of characteristic parameters to be estimated statistically. In the remaining cases nonparametric methods including neural networks may prove to be usefull, and the purpose of this paper is to develope a neural network applicable for estimating conditional distributions and quantiles in the general nonstandard situation.

Neural networks have in recent years developed into powerful tools for solving optimization problems within e.g. classification, estimation and forecasting. For the majority of cases, the applied neural networks, from a statistical point of view, solve conditional estimation problems. The celebrated Back Propagation Error algorithm used for training Feed Forward Neural Networks is shown to be a special case of gradient optimization in the sense of mean squared error [1]. Feed Forward Neural Networks are analyzed in [2] for consistent estimation of conditional expectation functions, which optimize expected squared error. Optimal classification is concerned with the problem of classifying a set of objects, on the basis of feature measurements, while obtaining a minimal probability of misclassification. This problem is equivalent to conditional estimation, and it is shown in [3] that Feed Forward Neural Networks estimate the optimal discriminating function, when trained with the Back Propagation Error Algorithm. In all of the above cases, some sort of optimization or training algorithm is applied adjusting initially random network parameters optimally w.r.t. average loss functions on a finite set of training data. A more constructive way to follow is indicated by [4], where a Probabilistic Neural Network for classification based on kernel

estimators is investigated, as well as by [5], [6], in which a similar line is followed for proposing neural networks estimating conditional expectation functions. From a certain point of view, this strategy is the basis for suggesting a large class of different neural network architectures, including among others Localized Receptive Fields [7] and Counter Propagation Networks [8]. In this paper such a constructive strategy is pursued in order to design a Feed Forward Neural Network capable of estimating conditional distributions. Initially the nessesary mathematical preliminaries concerning conditional distributions are given along with the basic terminology and notation. In the following section two different applications are suggested, where the latter, which is from the area of digital signal processing, constitutes the basis of a numerical example given in the last section. In the 3 rd section the basic neural network structure is developed based on kernel estimation techniques, and a rather general theoretical result is presented illustrating the wide applicability of the contstructed network.

II. Mathematical preliminaries

Consider a real random variable w with a distribution function F_w, and a number $p \in (0,1)$. Any real number q fulfilling:

$$F_w(q) = p \tag{1}$$

is said to be a quantile of order p [9]. If the distribution function F_w is continuous and strictly monotonous, the quantile of order p is uniquely defined for all values of p by equation (1). In general the p th. order quantile $q(p)$ can be uniquely defined as follows

$$q(p) = \sup\{q|\ F_w(q) \leq p\} \tag{2}$$

Consider two real random variables w and v defined on a common probability space with a joint distribution P_{wv} on $\mathbb{R}^2$. Then the function $F_{w|v} : \mathbb{R}^2 \to [0,1]$, exists ([10], section 33) such that

1. for every $v \in \mathbb{R}$, $F_{w|v}(\cdot, v)$ is a distribution function on $\mathbb{R}$,
2. for every $A = (-\infty, d]$ and every measurable subset B of $\mathbb{R}$

$$P_{wv}(A \times B) = \int_B F_{w|v}(d,v)\ dP_v(v) \tag{3}$$

The function $F_{w|v}$ is called the conditional distribution function of the random variable w with respect to v. In the case where the joint distribution P_{wv} has a density function h_{wv}, a conditional density function $h_{w|v}$ is given as

$$h_{w|v}(w,v) = \frac{h_{wv}(w,v)}{\int_{-\infty}^{\infty} h_{wv}(x,v)dx} \tag{4}$$

for every v where the denominator in the above formula is nonzero. Then the conditional distribution funtion $F_{w|v}$ can be found explicitly by

$$F_{w|v}(d,v) = \int_{-\infty}^{d} h_{w|v}(w,v)dw \tag{5}$$

For any $v \in \mathbb{R}$ the conditional quantile $q_C(p,v)$ of order p can be uniquely defined as in the unconditional case, i.e.

$$q_C(p,v) = \sup\{q|\ F_{w|v}(q,v) \leq p\} \tag{6}$$

The following equations:

$$q_C(F_{w|v}(q,v),v) = q \tag{7}$$

$$F_{w|v}(q_C(p,v),v) = p \tag{8}$$

are readily shown to hold in any point where $F_{w|v}$ is a continuous function of its first argument. Generalization of the definition of $F_{w|v}$ for $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$ is straigthforward, as well as of q_C for $v \in \mathbb{R}^n$ and $w \in \mathbb{R}$. In the most general case where $w \in \mathbb{R}^m$ the quantiles $q(p)$ and $q_C(p,v)$ are to be defined as sets. That is omitted here.

III. Applications

Two applications of conditional distribution functions and quantiles are considered below. One is a time optimal control problem from the area of decision theory, and the other is the task of data compression in the area of digital signal processing.

A. Time optimal control

Let the function $l : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, here after referred to as the loss function be defined by

$$l(W, w) = \begin{cases} -a\,(W - w) & \text{if } W - w \leq 0 \\ b\,(W - w) & \text{if } W - w \geq 0 \end{cases} \tag{9}$$

Obviously $l(W, w) = 0$ for $W = w$, so the loss function describes the losses incurred when the estimate W does not equal the optimal value w. As seen from equation (9) losses may depend strongly on the sign of estimation error, depending on the values of the coefficients a and b. If the optimal value w is a random variable with a probability distribution P_w the expected loss is given by the so called Bayes loss function l_B defined as follows

$$l_B(W) = \int_{\mathbb{R}} l(W, w)\, dP_w(w) \tag{10}$$

The value $l_B(W)$ simply constitutes the expected loss when estimating w by the value W. Any real number W_B such that

$$l_B(W_B) = \inf_{W \in \mathbb{R}} l_B(W) \tag{11}$$

is called a Bayes estimator.

When the loss function l is defined by equation (9) it is readily shown that the Bayes estimator equals the quantile of order

$$p = \frac{a}{a + b} \tag{12}$$

A practical example illustrating the relevance of quantiles as Bayes estimators is described in [11], where they are the solution of a time-optimal control problem. A parameter w, representing motion resistances in a mechanical system, is estimated by the value W, which appear directly in the equations of a time-optimal feedback controller. If $W > w$, overshoots occur which increases the time of reaching the target proportionally to $W - w$ with a coefficient b. In the case where $W < w$, so-called sliding trajectories appear, also prolonging the reaching period proportionally to $w - W$ with a coefficient a. The Bayes optimal estimate of the parameter w therfore exactly constitutes a quantile of order $\frac{a}{a+b}$. That problem has been solved in [12] using a preliminary version of the neural network presented in this paper.

When explanatory variables v are available the Bayes estimator for the above loss function is constituted by the conditional quantile q_C. In the time-optimal control problem, the vector v may contain disturbances possibly influencing the the resistances of motion, as for example temperature or target position. Minimum expected reaching time is then obtained for $W = q_C(p, v)$

B. Signal processing

When a finite capacity channel is used for data transmission, coding is often provided for optimal channel utilization. One example is the ADPCM [13] speech coding used for digital data transmission in mobile telephony. The objective is to transform a sequence of mutually dependent random data $\{.., u_{-1}, u_0, u_1, u_2...\}$ with a certain distribution P_w into a sequence $\{.., uc_{-1}, uc_0, uc_1, uc_2, ...\}$ of mutually independent and uniformly distributed data. The transformed data are then quantized and transmitted across the line to the reciever where they are decoded. ADPCM coding is designed on the assumption that the uncoded data exhibit the properties of linearly filtered white noise, in which case coding can be performed by the inverse linear filter removing the time correlation of the data. When the uncoded data are not correlated but still strongly dependent, linear filtering will of course work.

Assume that the sequence $\{u_i\}$ is Markov, i.e.

$$F_{u_i|u_{i-1},u_{i-2},..} = F_{u_i|u_{i-1}} = F_{w|v} \tag{13}$$

and that the conditional distribution function $F_{w|v}$ is a continuous function of its first variable. Then the sequence $\{uc_i\}$ defined by

$$uc_i = F_{w|v}(u_i, u_{i-1}) \tag{14}$$

will be independent and uc_i is uniformly distributed on $[0, 1]$ for all i. After quantization the sequence $\{ucq_i\}$ is transmitted. At the reciever side the original sequence with quantization error $\{udcq_i\}$ can be restored by

$$udcq_i = q_C(ucq_i, udcq_{i-1}) \tag{15}$$

The coding scheme can directly be generalized to k th. order Markov processes, i.e. where:

$$F_{u_i|u_{i-1},u_{i-2},..} = F_{u_i|u_{i-1},u_{i-2},..,u_{i-k}} = F_{w|v} \tag{16}$$

The above coding scheme is based on the assumptions that the conditional distribution $F_{w|v}$ is known in both ends of the transmission line. Therefore means for transmitting this knowledge without occupying any significant amount of channel capacity is needed.

IV. Neural Networks for estimating conditional distributions.

Feed Forward Neural Networks are most frequently trained by applying some sort of optimization procedure like Back Propagation in order to set weights and offsets optimally w.r.t. some objective function. In most cases the objective function equals the average of some loss function on the available set of data. Thus the objective function constitutes an estimate of the expected loss function, i.e. the Bayes loss function, and the training procedure an attempt to minimize the Bayes loss function. Successful training will force the neural network output to estimate the theoretical optimum, which for mean squared error is the conditional mean, and for the loss function l defined in equation (9) equals the conditional quantile.

In this section, a neural network for solving the more general problem of estimating conditional distribution functions, and their associated quantiles of any order. In [14] a perceptron like structure is trained with Back Propagation to reproduce so called fractional bins representing the conditional density. Here the reasoning follows the constructive line of [4], [5], [6] and is based on the theory of kernel estimation, which will be introduced shortly below.

A. Kernel estimation

Let $\{w_i\}$ in the following be a sequence of identically distributed random variables with a common density h_w. For any $m \in \mathbb{N} \setminus \{0\}$ and $r > 0$ the density estimate $h_w^{m,r} : \mathbb{R} \to \mathbb{R}$ can be defined by

$$h_w^{m,r}(w) = \frac{1}{m\ V(r)} \sum_{i=1}^{m} \phi(\frac{w - w_i}{r}) \tag{17}$$

where the volume function V is expressed as

$$V(r) = \int_{-\infty}^{\infty} \phi(\frac{w}{r})\ dw \tag{18}$$

and the kernel function $\phi : \mathbb{R} \to \mathbb{R}$ obeys

$$\lim_{r \to 0} \frac{1}{V(r)} \int_{-\infty}^{\infty} h(w)\ \phi(\frac{w - d}{r})\ dw\ =\ h(d) \tag{19}$$

for any bounded continuous density function h. The above estimator has been investigated in [15] for the case of the sequence $\{w_i\}$ being i.i.d. (independent identically distributed) random variables with a common continuous density function h_w. For $r \to 0$, and $m \cdot r \to \infty$ as $m \to \infty$, $h_w^{m,r}$ is shown to be a pointwise consistent estimator of h_w and its modes. By interpreting the kernel function ϕ as the nonlinear function of a neuron, and the sequence $\{w_i\}$ as a set of observations serving as training data, it has been demonstrated in [4] how this estimator exhibits properties equivalent to neural networks. From a computational point of view it posesses a massively parallel structure, which allows for high speed implementation on dedicated hardware; functionally, it is capable of learning general probabilistic information from measured data. It should be pointed out, however, that the number of neurons in the network defined from formula (17) equals the number of data in the training set, and that learning more or less takes place by memorizing data. In that respect the network provides no data compression.

In [5], [6] the estimator (17) was transformed to compute conditional expectation functions and a structure equivalent to the kernel smoother described in [16] was obtained. Here this transformation is directed towards estimators of conditional distribution functions.

In the multivariable case the training data is a finite sequence of the form $\{(w_i, v_i)\}$, where v_i denotes an observation of some observable explanatory variable. In that case the multivariable density estimate $h_{wv}^{m,r}$ can be given as

$$h_{wv}^{m,r}(w, v) = \frac{1}{m\ V(r)} \sum_{i=1}^{m} \phi(\frac{w - w_i}{r}) \cdot \phi(\frac{v - v_i}{r}) \tag{20}$$

A conditional distribution estimate $F^{m,r}_{w|v}$ can be obtained by subjecting $h^{m,r}_{wv}$ to a transformation analogous to the one defined by equations (4) and (5), i.e.

$$F^{m,r}_{w|v}(d,v) = \frac{\int_{-\infty}^{d} h^{m,r}_{wv}(w,v)\ dw}{\int_{-\infty}^{\infty} h^{m,r}_{wv}(w,v)\ dw} \tag{21}$$

which leads to the following closed form expression

$$F^{m,r}_{w|v}(d,v) = \frac{\sum_{i=1}^{m} S(\frac{d-w_i}{r}) \cdot \phi(\frac{v-v_i}{r})}{\sum_{i=1}^{m} \phi(\frac{v-v_i}{r})} \tag{22}$$

where S denotes the antiderivative of the function ϕ:

$$S(d) = \int_{-\infty}^{d} \phi(w)\ dw \tag{23}$$

A scaled Gaussian density function may be proposed as a candidate for the function ϕ, that is

$$\phi(d) = exp(-|d|^2) \tag{24}$$

Where $|\cdot|$ denotes the euclidian metric on $\mathbb{R}^D$. This function exhibits all properties required except that its anti derivative is not computable in a closed form expression. Therefore the function S can be chosen not according to equation (23) but as a function exhibiting equivalent properties and which is computable in a closed form expression. The well known sigmoid function then constitutes a natural choice, i.e.

$$S(d) = \frac{1}{1 + exp(-d)} \tag{25}$$

The above elaboration has been based on kernel estimation of a joint density function h_{wv} and leads to an estimate $F^{m,r}_{w|v}$ of the conditional distribution $F_{w|v}$; and serves here merely as motivation to formula (22). In fact, from definitions (24) and (25) it can be shown by fairly standard means, on only a very mild set of assumptions, that a slightly modified version of $F^{m,r}_{w|v}$ consistently estimates $F_{w|v}$, as stated precisely in the following theorem, which is proved in the appendix.

Theorem 1:

Let P_{wv} be a probability measure on $\mathbb{R}^2$ with an associated distribution function F_{wv}, and define the measure P_v on $\mathbb{R}$ by

$$P_v(A) = P_{wv}(\mathbb{R} \times A) \tag{26}$$

Assume a discrete time random process $z = (w,v) : \Omega \times \mathbb{N} \to \mathbb{R}^2$ to be such that empirical distributions converge to F_{wv} at every continuity point of that function, i.e.

$$. \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} U(z - z_i(\omega)) = F_{wv}(z) = \int U(z - (w,y))\ dP_{wv}(w,y)\ \ w.P.1 \tag{27}$$

for every continuity point z of F_{wv}, where the function $U : \mathbb{R}^2 \to [0,1]$ is given as

$$U(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 \geq 0 \text{ and } x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Also let the conditional distribution function $F_{w|v} : \mathbb{R}^2 \to [0,1]$ fulfill the following smoothness condition

$$\lim_{v \to \tilde{v}} F_{w|v}(\tilde{d}, v) = F_{w|v}(\tilde{d}, \tilde{v}) \tag{29}$$

at any point $(\tilde{d}, \tilde{v})$, where $F_{w|v}$ is continuous w.r.t. its first argument.

Then, for the estimator $F^{r,m}_{w|v}$ defined by

$$F^{m,r}_{w|v}(d, v) = \frac{\sum_{i=1}^{m} S(\frac{d - w_i + \sqrt{r}}{r}) \cdot \phi(\frac{v - v_i}{r})}{\sum_{i=1}^{m} \phi(\frac{v - v_i}{r})} \tag{30}$$

the following is true for any v in the support of P_v:

$$\lim_{r \to 0}[\lim_{m \to \infty} F^{r,m}_{w|v}(d, v)] = F_{w|v}(d, v) \quad w.P.1 \tag{31}$$

In Theorem 1 only a very general ergodic property has to be fulfilled, requiring empirical measures of the data sequence to converge to the limit measure P_{wv}, which not even locally is assumed to possess a density function. The conditional distribution function $F_{w|v}$ needs to vary smoothly w.r.t. the explanatory variable v in the sence stated in equation (29), which allows for allmost any degree of discontinuity w.r.t. the explained variable w. The estimator $F^{r,m}_{w|v}$ is redefined in equation (30) from its original definition in equation (22). The term $\sqrt{r}$ now appears in the argument of the function S allowing for convergence to points of discontinuity w.r.t. the explained variable w. It should be noted that the theorem is easily generalized to arbitrairy dimensions only by redefining the function S to an arbitrairy dimension D, i.e.

$$S(d_1, d_2, .., d_D) = \pi_{i=1}^{D} S(d_i) \tag{32}$$

For $D = 1$ the conditional quantile estimate $q_C^{m,r}$ is defined uniquely by

$$F^{m,r}_{w|v}(q_C^{m,r}(p, v), v) \tag{33}$$

B. Data compression

The neural network defined by equation (30) maps its training data directly onto its network parameters implying a number of neurons equal to the number of training data m. At least in the signal processing application discussed earlier some sort of efficient data compression is needed for the network to serve its purpose. Data compression is generally introduced by replacing the estimator $F^{m,r}_{w|v}$ with its compressed modification $F^{n,r}_{com}$

$$F^{n,r}_{com}(d, v) = \frac{\sum_{i=1}^{n} S(\frac{d - x_i^w + \sqrt{r}}{r}) \cdot \phi(\frac{v - x_i^v}{r})}{\sum_{i=1}^{n} \phi(\frac{v - x_i^v}{r})} \tag{34}$$

In this equation n denotes the number of neurons which is considered to be a design parameter restricted by $n << m$ in order to ensure a sufficient level of compression. The parameters $x_i = (x_i^w, x_i^v)$, are viewed as adjustable weights and offsets, subject to some training procedure projecting the statistical information of the training data to the network parameters. Two different procedures for setting the parameters x_i are discussed below.

One approach is based on the following self organising scheme

Algorithm 1:

- The parameters x_i are initially drawn randomly according to the joint density estimate $h_{wv}^{m,r}$ defined in equation (20)., i.e. for every $i \in 1..n$, j is selected from $1..m$ with equal probality for all values. x_i is then initially set by:

$$x_i = w_j + e \tag{35}$$

  where e is independent and Gaussian distributed with a variance $\sqrt{r}/2$.
- All parameters x_i are then submitted to the following correction scheme:
  - repeat
    - for $j = 1..m$
      - if $|x_i - w_j| < |x_k - w_j|$ then
      - $x_i = x_i + \eta \cdot (w_j - x_i)$
  - until all parameters have stabilized

The initial random setting according to the estimated density produces a rather qualified starting guess for the subsequent self organizing, which in turn diminishes the randomness from the initial settings to a level fair to the number of available training data n. That is the statistical uncertainty of $F_{com}^{n,r}(d, v)$ will be comparable to that of $F_{w|v}^{m,r}(d, v)$.

For the signal processing application discussed earlier only the initial settings might be transmitted whereas the subsequent correction is to be done independently at transmitter and reciever ends respectively. Such an approach closely resembles the ADPCM scheme mentioned earlier and even generalizes certain elements of that method.

Another approach to setting the parameters $\{x_i\}$ utilizes the estimator $F_{w|v}^{m,r}$ and an analogous estimator $F_v^{m,r}$, and is based on the following reasoning. Let the function $G : [0,1]^2 \to \mathbb{R}^2$ be defined by

$$G(x^1, x^2) = (q(x^1), q_C(x^2, q(x^1))) \tag{36}$$

where q is the quantile associated to the distribution F_v and q_C is the conditional quantile associated to $F_{w|v}$. If the random variables x_1 and x_2 are independent and uniformly distributed on $[0,1]$ then $G(x_1, x_2)$ is distributed according to F_{wv} on $\mathbb{R}^2$. Also if $\{A_k\}$ is a sequence of finite subsets of $[0,1]^2$ and the empirical distributions of the points in $\{A_k\}$ converge weakly to the uniform distribution on $[0,1]^2$

then the empirical distributions of $\{G(A_k)\}$ converge weakly to F_{wv}. For example $\{A_k\}$ might be crossing points on rectangular grids with gridsizes converging to zero. These properties are used in algorithm (2) to make network parameters imitate the empirical distributions of the data.

Algorithm 2:

- Select $\{x_i^1\}, i = 1..n_1$ equidistantly on $[0,1]$
- Let $y_i = q^{m,r}(x_i^1)$ for $i = 1..n_1$
- Select $\{x_j^2\}, j = 1..n_2$ equidistantly on $[0,1]$
- for $i = 1 : n_1$
  - for $j = 1 : n_2$
    - $k = n_2 \cdot i + j$
    - $x_k^v = y_i$
    - $x_k^w = q_C^{m,r}(x_j^2, y_i)$

The parameter setting scheme above can be generalized to arbitrary dimensions but it is not recommended for high dimensionality as the computational effort tends to grow exponentially with dimension. The advantage of this second approach for parameter setting is that constitutes a deterministic mapping from the available training data onto the network parameters, which makes it feasible for analysis. Convergence properties for the compressed estimator $F_{com}^{n,r}$ could be stated along the same line as for $F_{w|v}^{m,r}$. The possibility for such analysis is only mentioned at this point, whereas the analysis itself is not presented in this paper.

V. Numerical example

In this section a numerical example is presented illustrating how the proposed network can be applied to a signal processing problem, where the data sequence $\{u_i\}$ is to be transmitted over a finite capacity channel. The data sequence is artificially generated according to the following first order auto regression:

$$u_i = s_i \cdot A \cdot u_{i-1} + \sqrt{0.03 \cdot (1 - A^2)} \cdot e_i \quad \text{for } i = 1, 2, ..., 2000 \tag{37}$$

where s_i is drawn independently from $\{-1, 1\}$ for every i and $\{e_i\}$ is a sequence of independent and standard normally distributed random variables. The data sequence is depicted in Fig. 1, and a scatter plot of the data is shown in Fig. 2. It is obvious from equation (37), that u_i and u_j are completely uncorrelated for $i \neq j$, and any coding scheme based on linear filtering like the ADPCM algorithm becomes useless.

The data are encoded in two different ways as mentioned in the former section. It is firstly assumed that the complete data sequence is available before any transmission takes place. An uncompressed estimate of the conditional distribution $F_{w|v}$, where $v_i = u_{i-1}$ and $w_i = u_i$ is defined by $F_{w|v}^{m,r}$ and the network

parameters $\{x_i^w, x_i^v\}, i = 1 : 100$ are set according to algorithm (2), and transmitted to the recieving side in advance. The transmitted parameters are shown in Fig. 3. The performance of the compressed estimate $F_{com}^{n,r}$ is presented by its associated conditional quantiles of orders 0.1, 0.3, 0.7 and 0.9, which are plotted together with the theoretical ones in Fig. 4. The data sequence is encoded according to equation (14) and subsequently quantized in 4 bit precision to produce the sequence $\{ucq_i\}$ of which associated pairs (ucq_i, ucq_{i-1}) are shown in Fig. 5, illustrating how the joint distribution of adjacent values of ucq_i is close to being uniform in $[0,1]^2$. At the reciever side the sequence $\{ucq_i\}$ is decoded according to equation (15)to produce the revieved sequence $\{udcq_i\}$. The average absolute error is given by

$$E_{AV} = \frac{1}{2000} \cdot \sum_{i=1}^{2000} |u(i) - udcq(i)| = 0.004 \tag{38}$$

If the original data sequence $\{u_i\}$ were to be quantized directly in 4 bit precision assuming $u_i \in [-0.5, 0.5]$ for all i, the average absolute error is found to be 0.0156. Consequently 6 bit precision is required to match the above coding scheme by direct quantization. The second coding scheme suggested in the former section requires no network parameters to be transmitted over the line as network parameters are adjusted during transmission by identical algorithms in the transmitter and reciever ends of the line. In the presented example the network parameters are initially set on a rectangular grid in $[-0.5, 0.5]^2$. Only the parameters $\{x_i^w\}$ are adjusted during transmission as this was found to produce the best results. The initial and final parameter settings are shown in Fig. 6. The absolute error is shown slightly smoothened in Fig. 7, and is seen to tend to the average value 0.004 obtained by the previously presented coding scheme.

VI. Conclusion

A neural network for estimating conditional distribtions and their associated quantiles has been constructed in the present paper. Although the network is designed on the basis of kernel estimation of joint probability density functions, theory has been presented showing the network to be valid in more general settings, where only a smoothness condition w.r.t the dependence on the explanatory variable, as well as a very general ergodic property of the training data have to be fulfilled.

The problem of estimating conditional quantiles has been related to Bayes estimation in the case of a special asymmetric loss function feasible for application within a variety of areas in engineering, as well as science and economics. An example from the area of time optimal control is briefly discussed.

An application of the presented neural network within digital signal processing has been suggested. A scheme for encoding and decoding a sequence of data for optimal channel utilization is presented, along with two algorithms for training the network parameters before and during data transmission.

A numerical example where the neural network is applied to the above coding/decoding scheme is given and results are presented for both the two training algorithms. The results are considered satisfying for the presented example.

Appendix

Proof of Theorem 1

Assumption (27) guarantees weak convergence of empirical measures to P_{wv}. This yields, according to equation (30) and Theorem 29.1 of [10]:

$$F^r_{w|v}(d,v) = \lim_{m\to\infty} F^{r,m}_{w|v}(d,v) = \frac{\int S(\frac{d-w+\sqrt{r}}{r})\cdot\phi(\frac{v-y}{r})\,dP_{wv}(w,y)}{\int \phi(\frac{v-y}{r})\,dP_{wv}(w,y)} \tag{39}$$

Now the following expansion can be made for any $\epsilon > 0$, $\delta > 0$ and $\tilde{d} > d$

$$F^r_{w|v}(d,v) - F_{w|v}(d,v) = T_1 + T_2 + T_3 + T_4 + T_5 \tag{40}$$

where

$$T_1 = \frac{\int S(\frac{d-w+\sqrt{r}}{r})\cdot\phi(\frac{v-y}{r})\,dP_{wv}(w,y)}{\int \phi(\frac{v-y}{r})\,dP_{wv}(w,y)} - \frac{\int_{(-\infty,d]\times\mathbb{R}} \phi(\frac{v-y}{r})\,dP_{wv}(w,y)}{\int \phi(\frac{v-y}{r})\,dP_{wv}(w,y)} \tag{41}$$

$$T_2 = \frac{\int_{(-\infty,d]\times B(v,\delta)^C} \phi(\frac{v-y}{r})\,dP_{wv}(w,y)}{\int \phi(\frac{v-y}{r})\,dP_{wv}(w,y)} \tag{42}$$

$$T_3 = \frac{\int_{(-\infty,d]\times B(v,\delta)} \phi(\frac{v-y}{r})\,dP_{wv}(w,y)}{\int \phi(\frac{v-y}{r})\,dP_{wv}(w,y)} - \frac{\int_{B(v,\delta)} F_{w|v}(\tilde{d},y)\cdot\phi(\frac{v-y}{r})\,dP_v(y)}{\int \phi(\frac{v-y}{r})\,dP_v(y)} \tag{43}$$

$$= -\frac{\int_{(d,\tilde{d}]\times B(v,\delta)} \phi(\frac{v-y}{r})\,dP_{wv}(w,y)}{\int \phi(\frac{v-y}{r})\,dP_{wv}(w,y)} \tag{44}$$

$$T_4 = \frac{\int_{B(v,\delta)} F_{w|v}(\tilde{d},y)\cdot\phi(\frac{v-y}{r})\,dP_v(y)}{\int \phi(\frac{v-y}{r})\,dP_v(y)} - F_{w|v}(d,v) \tag{45}$$

$$= \frac{\int_{B(v,\delta)} (F_{w|v}(\tilde{d},y) - F_{w|v}(d,v))\cdot\phi(\frac{v-y}{r})\,dP_v(y)}{\int \phi(\frac{v-y}{r})\,dP_v(y)} \tag{46}$$

$$T_5 = -\frac{\int_{B(v,\delta)^C} F_{w|v}(d,v)\cdot\phi(\frac{v-y}{r})\,dP_v(y)}{\int \phi(\frac{v-y}{r})\,dP_v(y)} \tag{47}$$

in which the superscript "C" denotes complementary set.

The first term T_1 can be further expanded into the following terms

$$T_1 = T_{11} + T_{12} + T_{13} \tag{48}$$

where

$$T_{11} = \frac{\int_{(d+\delta,\infty)\times\mathbb{R}} S(\frac{d-w+\sqrt{r}}{r})\cdot\phi(\frac{v-y}{r})\,dP_{wv}(w,y)}{\int \phi(\frac{v-y}{r})\,dP_{wv}(w,y)} \tag{49}$$

$$T_{12} = \frac{\int_{(d,d+\delta]\times\mathbb{R}} S(\frac{d-w+\sqrt{r}}{r})\cdot\phi(\frac{v-y}{r})\,dP_{wv}(w,y)}{\int \phi(\frac{v-y}{r})\,dP_{wv}(w,y)} \tag{50}$$

$$T_{13} = \frac{\int_{(-\infty,d]\times\mathbb{R}} (S(\frac{d-w+\sqrt{r}}{r}) - 1)\cdot\phi(\frac{v-y}{r})\,dP_{wv}(w,y)}{\int \phi(\frac{v-y}{r})\,dP_{wv}(w,y)} \tag{51}$$

Because any distribution function is upper semi continuous

$$F_w(d+\delta) - F_w(d) = P_{wv}((d, d+\delta] \times \mathbb{R}) < \epsilon \Rightarrow |T_{12}| < \epsilon \tag{52}$$

for a sufficiently small value of $\delta > 0$.

For any $\delta > 0$, $r < \frac{\delta^2}{4}$ can be selected to fulfill

$$\begin{aligned} S(\frac{d-w+\sqrt{r}}{r}) &= \frac{1}{1+\exp(-\frac{d-w+\sqrt{r}}{r})} \\ &< \frac{1}{1+\exp(-\frac{-\delta+\sqrt{r}}{r})} \\ &< \frac{1}{1+\exp(-\frac{-\delta}{2r})} < \epsilon \text{ for } w > d+\delta \end{aligned} \tag{53}$$

and

$$\begin{aligned} S(\frac{d-w+\sqrt{r}}{r}) &\geq \frac{1}{1+\exp(-\frac{\sqrt{r}}{r})} \\ &= \frac{1}{1+\exp(-\frac{1}{\sqrt{r}})} \geq 1-\epsilon \text{ for } w \leq d \end{aligned} \tag{54}$$

which all together implies $|T_{11}| < \epsilon$ and $T_{13} \leq \epsilon$.

The fact that v by assumption belongs to the support of P_v implies

$$\frac{exp(-\frac{3}{4}\frac{\delta^2}{r^2})}{P_v(B(v, \frac{\delta}{2}))} \leq \epsilon \tag{55}$$

for a sufficiently small value of r. Now trivially

$$\left(\frac{v-y}{r}\right)^2 \geq \frac{\delta^2}{r^2} \text{ for } |y-v| \geq \delta \tag{56}$$

and

$$\left(\frac{v-y}{r}\right)^2 \leq \frac{\delta^2}{4 \cdot r^2} \text{ for } |y-v| < \frac{\delta}{2} \tag{57}$$

Inequalities (55),(56) and (57) together imply

$$\begin{aligned} \int_{(-\infty,d]\times B(v,\delta)^C} \phi(\frac{v-y}{r})\, dP_{wv}(w,y) &\leq \\ \int_{\mathbb{R}\times B(v,\delta)^C} \phi(\frac{v-y}{r})\, dP_{wv}(w,y) &\leq exp(-\frac{\delta^2}{r^2}) \end{aligned} \tag{58}$$

$$\begin{aligned} \int \phi(\frac{v-y}{r})\, dP_{wv}(w,y) &\geq \int_{B(v,\frac{\delta}{2})} \phi(\frac{v-y}{r})\, dP_v(y) \\ &\geq exp(-\frac{\delta^2}{4\, r^2}) \cdot P_v(B(v, \frac{\delta}{2})) \end{aligned} \tag{59}$$

so that

$$\bar{T} = \frac{exp(-\frac{\delta^2}{r^2})}{exp(-\frac{\delta^2}{4\, r^2}) \cdot P_v(B(v, \frac{\delta}{2}))} = \frac{exp(-\frac{3}{4}\frac{\delta^2}{r^2})}{P_v(B(v, \frac{\delta}{2}))} \tag{60}$$

by which

$$\begin{aligned} |T_2| &< \bar{T} \leq \epsilon \\ |T_5| &< \bar{T} \leq \epsilon \end{aligned} \tag{61}$$

is obtained for a sufficiently small value of $r > 0$

The distribution functions F_w and $F_{w|v}(\cdot, v)$ can have only a countable number of discontinuities. Also they are upper semi continuous. As a consequence a continuity point $\tilde{d} > d$ of $F_{w|v}(\cdot, v)$ can be found so that

$$F_w(\tilde{d}) < F_w(d) + \epsilon \tag{62}$$

and

$$F_{w|v}(\tilde{d}, v) < F_{w|v}(d, v) + \epsilon \tag{63}$$

which implies $|T_3| < \epsilon$

By assumption (29) $\delta > 0$ can be chosen sufficiently small to fulfill

$$|F_{w|v}(\tilde{d}, y) - F_{w|v}(\tilde{d}, v)| < \epsilon \text{ for } |y - v| < \delta \tag{64}$$

so that

$$\begin{aligned} |F_{w|v}(\tilde{d}, y) - F_{w|v}(d, v)| &= |F_{w|v}(\tilde{d}, y) - F_{w|v}(\tilde{d}, v) + F_{w|v}(\tilde{d}, v) - F_{w|v}(d, v)| \\ &\leq |F_{w|v}(\tilde{d}, y) - F_{w|v}(\tilde{d}, v)| + |F_{w|v}(\tilde{d}, v) - F_{w|v}(d, v)| \\ &\leq 2 \cdot \epsilon \text{ for } |y - v| < \delta \end{aligned} \tag{65}$$

and consequently $|T_4| \leq 2 \cdot \epsilon$

To prove the theorem select firstly $\tilde{d} > d$ to fulfill inequalities (62) and (63). Then choose $\delta > 0$ to fulfill inequalities (52) and (64). Finally pick $0 < r < \frac{\delta^2}{4}$ to fulfill (53), (54) and (61). That all together imply:

$$|F^r_{w|v}(d, v) - F_{w|v}(d, v)| \leq |T_1| + |T_2| + |T_3| + |T_4| + |T_5| \leq 3\epsilon + \epsilon + \epsilon + 2\epsilon + \epsilon \tag{66}$$

by which Theorem (1) is finally proved.

References

[1] D. E. Rumelhart and J. McClelland, *Parallel Distributed Processing*, MIT Press, 1986.

[2] H. White, "Connectionist Nonparametric Regression: Multi Layer Feedforward Network can Learn Arbitrary Mappings", *Neural Networks*, vol. 3, pp. 535–549, 1990.

[3] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function", *IEEE Transactions on Neural Networks*, vol. 1, pp. 296–298, 1990.

[4] D. F. Specht, "Probabilistic Neural Networks for Classification, Mapping, or Associative Memory", *in Proceedings of IEEE International Conference on Neural Networks*, vol. 1, pp. 525–533, 1988.

[5] D. F. Specht, "A General Regression Neural Network", *IEEE Transactions on Neural Networks*, vol. 2, pp. 568–576, 1991.

[6] H. Schioler and U. Hartmann, "Mapping Neural Network Derived from the Parzen Window Estimator", *Neural Networks*, vol. 5, pp. 903–909, 1992.

[7] J. Moody and C. Darken, "Fast Learning in Networks of Locally Tuned Processing Units", *Neural Computation*, vol. 1, pp. 281–294, 1989.

[8] R. H. Nielsen, "Counter propagation networks", *in Proceedings of the First IEEE International Conference on Neural Networks*, vol. 2, pp. 19–33, 1987.

[9] M. Fisz, *Probability Theory and Mathematical Statistics*, New York: Wiley, 1963.

[10] P. Billingsley, *Probability and Measure*, New York: Wiley, 1979.

[11] P. Kulczycki, "Time-Optimal Stochastic Positional Control", *in Proceedings of IFAC 12th World Congress*, vol. 7, pp. 443–448, 1993.

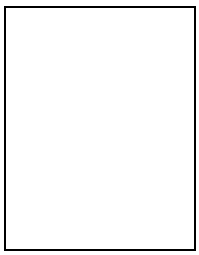
[12] P. Kulczycki and H. Schioler, "Parameter Identification by Bayes Decision and Neural Networks", *in Proceedings of 10th IFAC Symposium on System Identification*, vol. 3, pp. 477–482, 1994.

[13] Simon Haykin, *Communications Systems*, John Wiley & Sons, second edition, 1983.

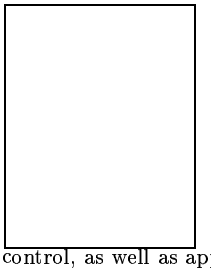
[14] A. S. Weigend and A. N. Srivastava, "Predicting conditional probability distributions: A connectionist approach", *International Journal of Neural Systems*, vol. 6, 1995.

[15] E. Parzen, "On Estimation of a Probability Density Function and Mode", *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.

[16] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, Chapman and Hall, first edition, 1990.



Henrik Schioler was born in 1965 and recieved his M.Sc. and Ph.D degrees from the Department of Control Engineering at Aalborg University. Since 1993 his has been with that department where he is currently an Assistant Professor. During a 2 years leave from 1994 to 1996 he was with CorTech A/S in Pandrup, Denmark, where he was working with methods for improving signal quality in cordless telephones. His research interests include probability theory, control theory, neural networks, fuzzy control, signal processing, e.t.c.



Piotr Kulczycki (M'93) received the M.Sc. and Ph.D. degrees in control engineering from the Academy of Mining and Metallurgy, Cracow, Poland, and the M.Sc. degree (with honors) from the Jagiellonian University, Cracow. Since finishing his studies in 1987 he has been at the Cracow University of Technology, where he is currently an Assistant Professor. In 1993 he held an Visiting Professor position at Aalborg University (Denmark). A member of IEEE, PTM (Polish Mathematical Society), and AMS (American Mathematical Society). Dr. Kulczycki has authored many journal and conference papers in the areas of optimal control, fault detection, neural networks, system identification, and fuzzy control, as well as applications of a probability approach to issues of economy and biology.