Appeared in *Proceedings of the 2nd IASTED Int. Conf. Artificial Intelligence and Applications, Malaga, Spain, Spetember 2002,* pp. 539-543, © ACTA Press

# MODELING OF DISTRIBUTIONS WITH NEURAL APPROXIMATION OF CONDITIONAL QUANTILES*

Pawel Wawrzynski and Andrzej Pacut
Institute of Control and Computation Engineering
Warsaw University of Technology
00–665 Warsaw, Poland
P.Wawrzynski@elka.pw.edu.pl, http://home.elka.pw.edu.pl/~pwawrzyn
A.Pacut@ia.pw.edu.pl, http://www.ia.pw.edu.pl/~pacut

## Abstract

We propose a method of recurrent estimation of conditional quantiles stemming from stochastic approximation. The method employs a sigmoidal neural network and specialized training algorithm to approximate the conditional quantiles. The approach may by used in a wide range of fields, in partricular in econometrics, medicine, data mining, and modeling.
**Keywords**: quantile regression, neural networks, data mining, modeling, discretization.

## 1 Introduction

Stochastic dependence between variables is commonly employed in various applications. Typically, the dependence is presented in a form of the conditional expected value as a function of the condition. Quite often yet the expected value is not a sufficient representation of this dependence, and the conditional distribution as a function of the condition is required. Estimation of conditional quantiles as a powerful tool for modeling the conditional distributions has been widely recognized in the field of econometrics (see, e.g. [1, 2, 3, 7]) and medicine (see, e.g. [6]). Parametric estimation of conditional quantiles from a finite sample (quantile regression) leads to a problem of minimization of a functional that is not differentiable in parameters, which require non-standard minimization techniques, like the one proposed by Koenker and Park [4].

In this paper we propose a method of conditional quantiles as a function of condition based on stochastic approximation (see e.g. [5]). The problem of non-differentiability is then avoided. The proposed method is recurrent and can be applied "on line". We investigate the use of sigmoidal neural networks, which allows to estimate an entire family of conditional quantiles. Such conditional multi-quantile models can be used in problems occurring in econometrics, and also in data mining and modeling. Some of potential applications are discussed in the paper.

*This paper was supported by the grant 134/E-365/SPUB-M5.PR/DZ320/2000-2002 of the State Committee for Scientific Research

## 2 Quantiles

In this section we introduce the notation and recall some definitions related to quantiles and conditional quantiles. Random variables will be denoted by capital letters, and their values by lower case letters. For multi-dimensional random variables and their values we use bold letters. We typically represent distributions by their distribution functions, and denote the distribution function of $Y$ by $F_Y$. Values of the conditional distribution of $Y$ conditioned on $X = x$ will be denoted by $F_{Y|X=x}(y)$ or $F_{Y|X}(y|x)$.

**Definition 1 (Quantiles)** *Consider one dimensional random variable $Y$ of distribution function $F_Y$. For a fixed $\alpha \in (0,1)$, the $\alpha$-quantile $q_\alpha$ is defined as any number $y$ that fulfills the relations*

$$\begin{aligned} P(Y \leq y) &\geq \alpha \\ P(Y \geq y) &\geq 1-\alpha \end{aligned} \tag{1}$$

It can be easily seen that while such $y$ always exists, it is not necessarily unique, since (1) may be fulfilled by all points of some interval. To avoid non-uniqueness, it is customary in such cases to choose a representative of the set of points that fulfill (1), typically by taking the lower end of the interval. With this addition, $q_\alpha$ as a function of $\alpha$ is an inverse function to the distribution function and is defined everywhere on $(0,1)$. This enable to formulate the following property.

**Proposition 1 (Representation Property)** *Any one-dimensional distribution is defined by the family of all its quantiles.*

The equation $q_\alpha = F^{-1}(\alpha)$ takes place for all $\alpha \in (0,1)$ such that $F^{-1}(\alpha)$ is determined.

The representation property enables to approximate distribution functions through their quantiles. More specifically, we may formulate the following proposition.

**Proposition 2 (Approximation)** *Any scalar distribution can by arbitrarily well approximated by a finite number of quantiles.*

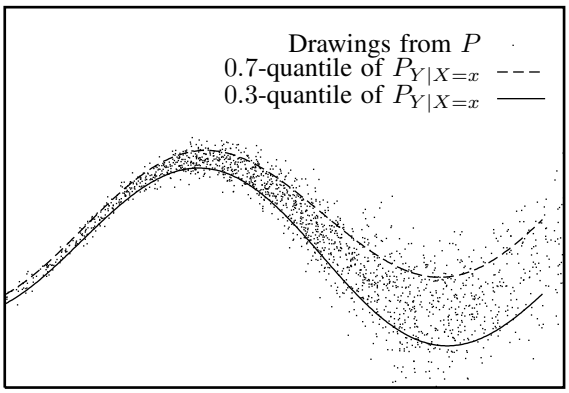


Figure 1. A sample from a two-dimensional distribution $F_{X,Y}$ and two conditional quantiles: $q_{0.3}(x)$ and $q_{0.7}(x)$. As little as two quantiles already give some representation of the distribution function.

Consider a discrete distribution $S_n$ concentrated at $n$ equi-probable points, where $i$-th point is located at $(i-0.5)/n$-quantile of a distribution $F$. It is straightforward to show that $S_n$ weekly converges with $n$ to $F$, hence $S_n$ may serve as an approximate discrete distribution of $F$.

Note that the Approximation Property may serve as the base for sampling from the approximating distribution described above.

**Definition 2 (Conditional Quantiles)** *Let $Y$ be a scalar random variable $Y$, and $\mathbf{X}$ be a random vector, and assume that the conditional distribution $F_{Y|\mathbf{X}=x}$ is well defined. The* conditional quantile *$q_\alpha(x)$ is defined as the $\alpha$-quantile of the conditional distribution $F_{Y|\mathbf{X}=x}$*

The conditional quantiles are functions of the condition and may approximate the conditional distributions. Figure 1 shows how only two conditional quantiles $q_\alpha(x)$ can approximately represent the conditional distribution $F_{Y|X=x}$ for any given value of $x$. If the quantiles of $X$ are also known, the joint distribution of $(X, Y)$ can also be approximated.

## 3 Approximation of conditional quantiles

To approximate the conditional quantiles we may use any parameterized approximator, for example the multilayer perceptron. Denote by $\mathbf{N}_\mathbf{w}(\mathbf{x}) = [N_{1,\mathbf{w}}(\mathbf{x}), \ldots, N_{n,\mathbf{w}}(\mathbf{x})]^T$ the output vector of a neural network $\mathbf{N}_\mathbf{w}$ parameterized by the weight vector $\mathbf{w}$, and whose input is equal to $\mathbf{x}$. We assume that $N_{i,\mathbf{w}}(\mathbf{x})$ are continuously differentiable functions of $\mathbf{w}$ for all $\mathbf{x}$ and $i = 1 \ldots n$. This assumption is satisfied, for instance, by sigmoidal multilayer perceptrons.

Let $(\mathbf{X}, Y)$ be a random variable and consider the problem of approximation the conditional distribution $F_{Y|\mathbf{X}}$ by the use of quantiles. In this order we need to approximate a set of quantiles of the conditional distribution, of orders $\alpha_1, \ldots, \alpha_n \in (0, 1)$ evenly covering

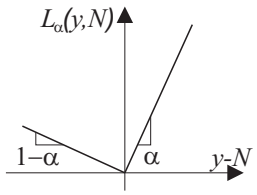


Figure 2. The loss function used in quantile estimation

the interval $(0, 1)$. Here we design a $n$-output neural network $\mathbf{N}_\mathbf{w}(.)$ that approximates such quantiles. First, we introduce a certain convex minimization problem for a neural network and prove that its solution appropriates the desired set of conditional quantiles. Then, we propose a training regime that make weights $\mathbf{w}$ of the network $\mathbf{N}_\mathbf{w}(.)$ to approximate this solution.

For simplicity, we first consider a scalar $\alpha \in (0, 1)$ and design a network $N_\mathbf{w}$ with a single output that approximates the $\alpha$-quantile of the conditional distribution $F_{Y|\mathbf{X}}$. Define a loss function $L$. (Fig. 2)

$$L_\alpha(y, N) = \begin{cases} (y - N)\,\alpha, & y \geq N \\ (N - y)\,(1 - \alpha), & y < N \end{cases} \tag{2}$$

This function can be understood as a momentary loss for a network whose output is equal to $N$ while the desired value is equal to $y$. Now we assume that the desired output $Y$ is random, and consider the expected loss for a given value $x$ of the random variable $X$, namely

$$Q_\alpha(\mathbf{x}, N) = E\big(L_\alpha(Y, N)|\mathbf{X} = \mathbf{x}\big) \tag{3}$$

We consider the following minimization problem

$$\min_{h(\cdot)} E\big(L_\alpha(Y, h(\mathbf{X}))\big) \tag{4}$$

Since

$$\begin{aligned} E\big(L_\alpha(Y, h(\mathbf{X}))\big) &= E\big(E\big(L_\alpha(Y, h(\mathbf{X}))|\mathbf{X}\big)\big) \\ &= E\big(Q_\alpha(\mathbf{X}, h(\mathbf{X}))\big) \end{aligned}$$

problem (4) is solved if $h(\mathbf{x})$ minimizes $Q_\alpha(\mathbf{x}, h(\mathbf{x}))$ for almost all $\mathbf{x}$.

**Theorem 1** *Suppose the conditional expected value $E\big(Y\big|\mathbf{X}\big)$ is defined almost everywhere. The solution to (4) is equal to $\alpha$-quantile of $F_{Y|\mathbf{X}}$.*

**Proof:** By definition

$$\begin{aligned} Q_\alpha(\mathbf{x}, N) = (1 - \alpha) \int_{-\infty}^{N} (N - y) P(\mathrm{d}y|\mathbf{X} = \mathbf{x}) \\ + \alpha \int_{N}^{\infty} (y - N) P(\mathrm{d}y|\mathbf{X} = \mathbf{x}) \end{aligned} \tag{5}$$

We transform the formual above in two steps. First we split the integration interval into two parts for an arbitrary constant $c$. Second we differentiate by parts. We

obtain:

$$Q_\alpha(\mathbf{x}, N) = \overline{C}(c) - \alpha N + \int_c^N P(Y < y|\mathbf{X} = \mathbf{x})\mathrm{d}y \quad (6)$$

where $\overline{C}$ is another constants. Let fix $\mathbf{x}$ and consider the derivative

$$\begin{aligned} \frac{\mathrm{d}Q_\alpha(\mathbf{x}, N)}{\mathrm{d}N} &= P(Y < N|\mathbf{X} = \mathbf{x}) - \alpha \quad (7) \\ &= 1 - \alpha + P(Y \geq N|\mathbf{X} = \mathbf{x}) \quad (8) \end{aligned}$$

Note that the derivative does not depend on $c$. If $N$ is smaller then the $\alpha$-quantile, then by (7) the derivative is smaller then 0. Reversely, by (8), if $N$ is greater than the $\alpha$-quantile, the derivative is greater then zero. In the non unique case, the derivative is equal to zero inside the quantile interval. Consequently, $Q_\alpha(\mathbf{x}, .)$ is minimized for the $\alpha$-quantile of $P_{Y|\mathbf{X}}$. ■

**Corollary 1** *Solution to the minimization problem*

$$\min_{h_1(.),\ldots,h_n(.)} E\Big(\sum_{i=1}^n L_{\alpha_i}(Y, h_i(\mathbf{X}))\Big) \quad (9)$$

*is the set of $\alpha_1, \ldots, \alpha_n$ quantiles of $F_{Y|\mathbf{X}}$.*

**Corollary 2** *Minimization of the risk functional*

$$R(\mathbf{w}) = E\Big(\sum_{i=1}^n L_{\alpha_i}(Y, N_{i,\mathbf{w}}(\mathbf{x}))\Big) \quad (10)$$

*makes $N_{i,\mathbf{w}}(.)$ to approximate $\alpha_i$-quantile of $P_{Y|\mathbf{X}}$ in a sense of minimization problem (4).*

The corollary is a straightforward consequence of previous theorem and the fact that minimization problem is convex.

**Theorem 2** *Suppose pairs $\langle y, \mathbf{x}\rangle$ are drawn independently from a continuous distribution whose conditional expected values $E\big(Y|\mathbf{X}\big)$ exist. The weight update mechanism*

$$\mathbf{w} := \mathbf{w} - \beta\delta(\mathbf{w}, \mathbf{x}, y) \quad (11)$$

*where*

$$\delta(\mathbf{w}, \mathbf{x}, y) = \sum_{i=1}^n \frac{\mathrm{d}N_{i,\mathbf{w}}(\mathbf{x})}{\mathrm{d}\mathbf{w}} \begin{cases} -\alpha_i & y \geq N_{i,\mathbf{w}}(\mathbf{x}) \\ (1-\alpha_i) & y < N_{i,\mathbf{w}}(\mathbf{x}) \end{cases}$$

*minimizes the cost functional (10) if $\beta$ is a sequence that satisfies the standard stochastic approximation conditions.*

**Proof:** Proof consists of showing that the expected value of $\delta$ is equal to the gradient of the risk functional (10). Modification of the weights according to (11) becomes identical to minimization of the risk functional with Robbins-Monroe procedure of stochastic approximation.

Consider a function:

$$R_i(\mathbf{w}) = E\big(Q_{\alpha_i}(\mathbf{X}, N_{i,\mathbf{w}}(\mathbf{X}))\big) \quad (12)$$

We can calculate the gradient of $R_i$, namely

$$\begin{aligned} \frac{\mathrm{d}R_i(\mathbf{w})}{\mathrm{d}\mathbf{w}} &= \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} E\Big(Q_{\alpha_i}(\mathbf{X}, N_{i,\mathbf{w}}(\mathbf{X}))\Big) \\ &= E\Big(\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} Q_{\alpha_i}(\mathbf{X}, N_{i,\mathbf{w}}(\mathbf{X}))\Big) \\ &= E\Big(\frac{\mathrm{d}N_{i,\mathbf{w}}(\mathbf{X})}{\mathrm{d}\mathbf{w}} \times \\ &\qquad \times\Big(-\alpha_i + P(Y < N_{i,\mathbf{w}}(\mathbf{X})|\mathbf{X})\Big)\Big) \quad (13) \end{aligned}$$

On the other hand, consider a function:

$$\delta_i(\mathbf{w}, \mathbf{x}, y) = \frac{\mathrm{d}N_{i,\mathbf{w}}(\mathbf{x})}{\mathrm{d}\mathbf{w}} \begin{cases} -\alpha_i & y \geq N_{i,\mathbf{w}}(\mathbf{x}) \\ (1-\alpha_i) & y < N_{i,\mathbf{w}}(\mathbf{x}) \end{cases}$$

and its expected value:

$$\begin{aligned} E\delta_i(\mathbf{w}, \mathbf{X}, Y) = E\Big(&\frac{\mathrm{d}N_{i,\mathbf{w}}(\mathbf{X})}{\mathrm{d}\mathbf{w}} \times \quad (14) \\ &\times\Big(-\alpha_i P(Y \geq N_{i,\mathbf{w}}(\mathbf{X})|\mathbf{X}) \\ &+ (1-\alpha_i)P(Y < N_{i,\mathbf{w}}(\mathbf{X})|\mathbf{X})\Big)\Big) \end{aligned}$$

By (13) and (14), we obtain

$$E\delta_i(\mathbf{w}, \mathbf{X}, Y) = \frac{\mathrm{d}R_i(\mathbf{w})}{\mathrm{d}\mathbf{w}}$$

Straightforward calculation

$$E\,\delta(\mathbf{w}, \mathbf{X}, Y) = \sum_{i=1}^n E\,\delta_i(\mathbf{w}, \mathbf{X}, Y) = \frac{\mathrm{d}R(\mathbf{w})}{\mathrm{d}\mathbf{w}}$$

completes the proof. ■

According to the above theorem, (11) leads to the quantile approximation when the underlaying distribution is continuous. Some technical difficulties emerge in general case that will not be discussed here.

To give some intuition on how the approximator works, suppose that $n = 1$, and rewrite (11) in the form

$$\mathbf{w} := \mathbf{w} + \beta\frac{\mathrm{d}N_{1,\mathbf{w}}(\mathbf{x})}{\mathrm{d}\mathbf{w}} \begin{cases} \alpha_1 & N_{1,\mathbf{w}}(\mathbf{x}) \leq y \\ -(1-\alpha_1) & N_{1,\mathbf{w}}(\mathbf{x}) > y \end{cases}$$

Whenever $N_{1,\mathbf{w}}(\mathbf{x})$ happens be smaller than $y$, its value is increased and, reversely, whenever $N_{1,\mathbf{w}}(\mathbf{x})$ is greater then $y$ it is decreased. The average change of $N_{1,\mathbf{w}}(\mathbf{x})$ for a given $\mathbf{x}$ is related to the weighted frequency of those two actions, weighted by $\alpha_1$ or $(1-\alpha_1)$ and, on the average, moves $N_{1,\mathbf{w}}(\mathbf{x})$ toward $q_{\alpha_1}(\mathbf{x})$.

## 4 From conditional quantiles to multidimensional distributions

Modeling a multidimensional distribution with the use of conditional quantiles is possible through the application of the Bayes rule, namely

$$\begin{aligned} &F_{X_1,\ldots,X_m}(x_1, \ldots, x_m) = \\ &\quad = F_{X_1}(x_1)\; F_{X_2|X_1}(x_2|x_1)\; \cdots \times \\ &\quad \times F_{X_m|X_{m-1},\ldots,X_1}(x_m|x_{m-1}, \ldots, x_1) \end{aligned}$$

Note that the problem of approximation of an $m$-dimensional distribution is converted here to an equivalent problem of modeling 1 one-dimensional distribution and $m-1$ one-dimensional conditional distributions. A multidimensional conditional distribution $F_{\mathbf{Y}|\mathbf{X}}$ can be modeled in a similar way, namely

$$\begin{aligned} F_{Y_1,\ldots,Y_m|\mathbf{X}}(y_1,\ldots,y_m|\mathbf{x}) = \\ = F_{Y_1|\mathbf{X}}(y_1|\mathbf{x})\ F_{Y_2|Y_1,\mathbf{X}}(y_2|y_1,\mathbf{x})\ \cdots\ \times \\ \times\ F_{Y_m|Y_{m-1},\ldots,Y_1,\mathbf{X}}(y_m|y_{m-1},\ldots,y_1,\mathbf{x}) \end{aligned}$$

Again, approximation of an $m$-dimensional conditional distribution is converted here to an equivalent problem of modeling $m$ one-dimensional conditional distributions. As it results from Proposition 2, if the ranges of the quantiles are distributed evenly over the interval $(0,1)$, drawing them is identical to drawing from the source distribution.

## 5 Illustration

For illustration, we present some results of a numerical experiments, Fig. 3. The data $(x_i, y_i)$ were calculated as $x_i = r_i \sin\varphi_i$, $y_i = r_i \cos\varphi_i$, where $\{r_i\}_{i\geq 1}$ and $\{\varphi_i\}_{i\geq 1}$ were generated independently from the normal distribution $\mathbf{N}_{1,\,0.15^2}$ and from the uniform distribution $\mathbf{U}_{(0,2\pi)}$, resp. Two models of this distribution were investigated. In the first model we approximated the distribution of $(X,Y)$ conditioned on $\Phi=\varphi$ for various values of $\varphi$. The approximating system consists of two neural networks that model the distributions $P_{X|\Phi=\varphi}$ and $P_{Y|X=x,\Phi=\varphi}$. The first network, approximating quantiles of the distribution $P_{X|\Phi=\varphi}$, was a two-layer perceptron with 1 input, 6 neurons in the hidden layer and 6 neurons in the output layer which approximated $(i-0.5)/6$-quantiles, $i=1,\ldots,6$. The second network, modeling quantiles of the distribution $P_{Y|X=x,\Phi=\varphi}$, was implemented as a two-layer perceptron of exactly the same structure as the first one, except that it has two inputs.

In the second model we approximated the distribution of $(X,Y)$. We first approximated the quantiles of $X$. Then we approximated quantiles of $P_{Y|X=x}$ by a two-layer perceptron with a single input, 6 neurons in the hidden layer, and 8 neurons in the output layer which approximated the quantiles.

## 6 Applications

**Data Mining** An important issue in data mining is to discover relations between variables, hence the method of neural approximation of conditional quantiles can be conveniently applied. As an example, suppose the objects are described by $n+1$ variables: $x_1,\ldots,x_n,y$. Then, for the validation data that describe the objects only by $x_1,\ldots,x_n$, one is to derive, for every object in the validation set, an interval as narrow as possible such that the probability that $y$ will hit the interval is not less than $1-2\alpha$ for some small $\alpha$. The problem leads to the estimation of conditional $\alpha$- and $(1-\alpha)$-quantiles of $y$ as functions of $x_1,\ldots,x_n$. Figure 1 gives some

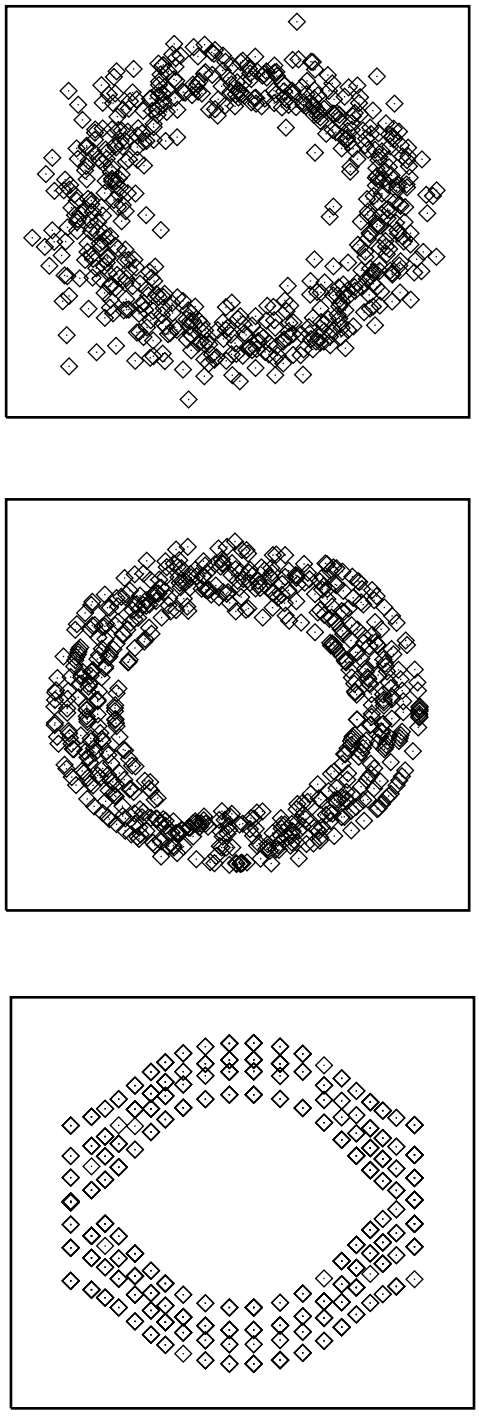

Figure 3. Results of sampling from a distribution and its models. (*top:*) A 600-element sample from a two-dimensional distribution $(X,Y)$ where $X = R\sin\Phi$, $Y = R\cos\Phi$, with independent $R$ of normal distribution and $\Phi$ of uniform distribution. (*middle:*) Results of sampling from the model of $(Y,X)$ conditioned on $\Phi=\varphi$ for varying $\varphi$. (*bottom:*) Results of sampling from the model of $(Y,X)$.

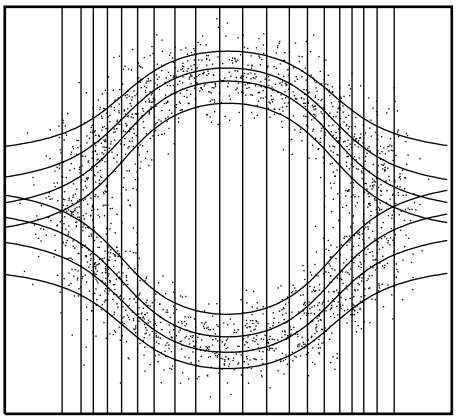

Figure 4. Illustration of discretisation with quantiles

intuition related to the problem and its solution. The method gives, for any given $x$, an interval that is likely to contain $y$. For small $x$, the value of $y$ is determined almost certainly — the interval is narrow. Conversely, for large $x$, the stochastic dependence between $x$ and $y$ is more „fuzzy" and the interval is wider.

**Modeling** Problem of modeling multidimensional distributions arise in many fields. As already discussed, the $m$-dimensional distribution can be modeled by putting together a quantile model of one-dimensional distribution and $m$–1 quantile models of conditional distribution.

**Discretization** Conditional quantiles make an important tool in discretization of multi-dimensional distributions. First, a model of the underlying distribution is builded. The model domain is than partitioned to coherent areas. The space is first partitioned regarding to the quantiles of $x_1$ distribution. The resulting areas are again partitioned, with conditional quantiles $x_2|x_1$. The partitioning is continued until all the dimensions are taken into account. The last division is done by conditional quantiles $x_n|x_{n-1}, \ldots, x_1$. To continue Illustration, the two-dimensional space obtained above is first partitioned with the use of $x$-quantiles (Fig. 4, vertical lines) and then with the use of conditional quantiles $y|x$ (Fig. 4, "horizontal" curves).

## 7 Conclusions

This paper proposes method of estimating conditional quantiles, which can be conviniently used to train a sigmoidal neural network. The presented weight update mechanism allows neural network to approximate an antire family of conditional quantiles. A few applications of conditional quantiles in artificial intelligence are also presented.

## References


[1] M. Buchinsky, Changes in the U.S. wage structure 1963-1987: Application of quantile regression, *Econometrica* 62, 1994, 405-458.

[2] M. Buchinsky, Recent Advances in Quantile Regression Models: A practical guide for empirical research, *Journal of Human Resources*, 33, 1998, 88-126.

[3] W. Hendricks, R. Koenker, Hierarchical spline model for conditional quantiles and demand for electricity, *Journal of American Statistical Associacion* 87, 1992, 58-68.

[4] R. Koenker, B. J. Park, An Interior Point Algorithm for Nonlinear Quantile Regression, *Journal of Econometrics*, 71, 1997, 265-283.

[5] H.J. Kushner and G.G. Yin, *Stochatic Approximation Algorithms and Applications*. (New York, Springer-Verlag, 1997).

[6] S.R. Lipsitz, G.M. Fitzmaurice, G. Molenberghs and L.P. Zhao, Quantile regression methods for longitudinal data with drop-outs, *Applied Statistics*, 46, 1997, 463-476.

[7] J.A.F. Machado, J. Mata, Box-Cox Quantile Regression and the Distribution of Firm Sizes, *Journal of Applied Econometrics*, 15, 1997, 253-274.