On the Training of a Kolmogorov Network

Mario Köppen

Fraunhofer IPK Berlin, Department Pattern Recognition
Pascalstr. 8-9, 10587 Berlin, Germany
e-mail: mario.koeppen@ipk.fhg.de

Abstract. The Kolmogorov theorem gives that the representation of continuous and bounded real-valued functions of n variables by the superposition of functions of one variable and addition is always possible. Based on the fact that each proof of the Kolmogorov theorem or its variants was a constructive one so far, there is the principal possibility to attain such a representation. This paper reviews a procedure for obtaining the Kolmogorov representation of a function, based on an approach given by David Sprecher. The construction, adapted to our purposes, is considered in more detail for an image function. It comes out that such a representation is featureless (with regard to analytical properties of the represented function), and basically resembles a look-up procedure, employing fuzzy singletons around functions values that were looked up for generalization.

1 Introduction

This paper gives a study of the superposition of a continuous, bounded real-valued function of n variables by the superposition of functions of one variable and addition. The universal possibility of such a representation is granted by the Kolmogorov theorem [1]. It was Hecht-Nielsen [2], who rediscovered the importance of the Kolmogorov theorem for the theoretical understanding of the abilities of neural networks. The Kolmogorov theorem was also pointed out to be of importance for other designs of soft computing, as e.g. normal forms in fuzzy logic [8] [9]. Here, we focus on the constructive aspects of this representation, thus following a line of studies presented so far e.g. in [3], [4], [5], [6] and [7].

The starting point is the Kolmogorov theorem in a notation used by Sprecher:

Theorem 1 (Sprecher, 1996). *Every continuous function $f : I^n \to R$ can be represented as a sum of continuous real-valued functions:*

$$f(x_1, \dots, x_n) = \sum_{q=0}^{2n} \chi_q(y_q). \tag{1}$$

In this representation, the $x_1, \dots, x_n$ are the parameters of an embedding of I_n into R_{2n+1}:

$$y_q = \eta_q(x) = \sum_{p=1}^{n} \lambda_p \psi(x_p + qa) \tag{2}$$

with a continuous real-valued function ψ and suitable constants λ_p and a. This embedding is independent of f.

In [6] and [7], Sprecher gives a numerical procedure for computing the "inner" function ψ and setting the "outer" function" χ. However, the inner function given easily shows to be non-continuous, so the definition presented here has been corrected accordingly in order to give a continuous ψ function.

This paper is organized as follows: section 2 describes the ingredients needed for performing the construction. Then, section 3 shows how this procedure might be used to gain understanding of a given function f (an image function). The paper concludes with a discussion.

Due to space limitations, it will not be possible to provide the complete proof of the Kolmogorov theorem from the construction. We hope to be able to present this in a future communication.

2 The revised Sprecher algorithm

2.1 The inner function

Be $n \geq 2$ the dimensionality of the function f, and $m \geq 2n$ the number of terms in eq. (1). Also, be $\gamma \geq m + 2$ a natural number, which is used as radix in the following (a good choice for $n = 2$ is $m = 5$ and $\gamma = 10$). A constant a is given by $a = (\gamma(\gamma - 1))^{-1}$, and x will asign a vector $(x_1, \ldots, x_n)$. For the decimal base, a has the representation $0.0111\ldots$.

One further definition: $\beta(r) = (n^r - 1)/(n - 1) = 1 + n + n^2 + \ldots + n^{(r-1)}$ (e.g. for $n = 2$ the sequence 1, 3, 7, 15, 31, $\ldots$).

The constants λ_p in theorem 2 will be computed by the expression

$$\lambda_p = \begin{cases} 1 & p = 1 \\ \sum_{r=1}^{\infty} \gamma^{-(p-1)\beta(r)} & p > 1 \end{cases} \quad (3)$$

Using those definitions, the inner function$\psi(x)$ can be defined. The construction is given for all terminating rational numbers $d_k \in I$, which have in the decimal notation to the base γ not more than k digits (as 0.031, 0.176, 0.200, if $\gamma = 10$ and $k = 3$).
The notation $d_k = [i_1, \ldots, i_k]_\gamma$ means that d_k to the base γ has the notation $0.i_1 i_2 \ldots i_k$.
Then, the inner function is defined as follows:

$$\psi_k(d_k) = \begin{cases} d_k & \text{for } k = 1; \\ \psi_{k-1}(d_k - \frac{i_k}{\gamma^k}) + \frac{i_k}{\gamma^{\beta(k)}} & \text{for } k > 1 \text{ and } i_k < \gamma - 1; \\ \frac{1}{2}\left(\psi_k(d_k - \frac{1}{\gamma^k}) + \psi_{k-1}(d_k + \frac{1}{\gamma^k}\right) & \text{for } k > 1 \text{ and } i_k = \gamma - 1. \end{cases} \quad (4)$$

It can be easily seen that this recursive definition always terminates.

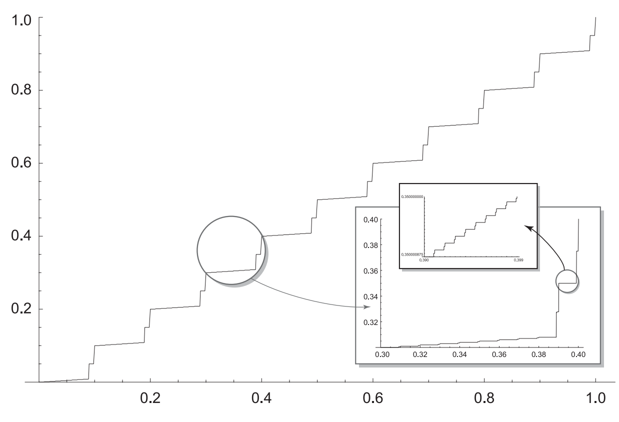

Fig. 1. Graph of the inner function ψ_x for $n = 2$ and $\gamma = 10$.

Figure 1 gives the graph of $\psi(x)$ for $n = 2$ und $\gamma = 10$. It features some kind of structural self-similarity. The assembling part of this self-similarity is shown in fig. 2. There are two self scalings. The section a with increase α is replaced by γ segments, with the first

$(\gamma - 2)$ segments having identical increase $\alpha_1 < \alpha$ and going to 0, and the last 2 segments having increase $\alpha_2 > \alpha$ going to ∞ for growing k. The "switching point", where the two replacement schemes change, approximates the point $d_k + \delta_k$ with

$$\delta_k = \frac{\gamma - 2}{\gamma - 1}\gamma^{-k} = (\gamma - 2) \sum_{r=k+1}^{\infty} \gamma^{-k}. \tag{5}$$

The function ψ_x has some mathematical features, as being continuous, strict monotone increasing, being flat nearly everywhere and concave.

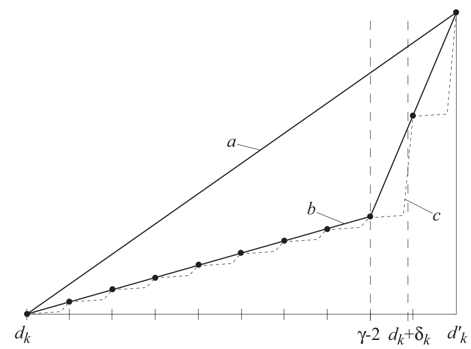

Fig. 2. Self-similar part of the inner function ψ_x for $n = 2$ and $\gamma = 10$, and definition of δ_k.

2.2 Inner superposition

Now, we have to consider the superposition of the inner functions, as given by

$$\eta_q(\mathbf{d}_k) = \sum_{p=1}^{n} \lambda_p \psi(d_{kp} + qa) \tag{6}$$

for a fixed $0 \leq q \leq m = 2n$. Here, $\mathbf{d}_k$ shall be the vector composed of n rational numbers from the d_{kp}.

This superpositions has an important property: For two arbitrary vectors $\mathbf{d}_k, \mathbf{d}'_k \in I^n$, the distance of their images under $\eta()$ is never smaller than $\gamma^{-n\beta(k)}$.

2.3 The outer function

The construction of the outer function is basically a look-up. Each value from I_n is assigned a different "height" under the mapping η. Now, the values of f at point $x \in I_n$ gives the value of χ at the height assigned to x by η.

However, this would not suffice to gain an exact representation of f for growing k. This is, where the $m \geq 2n + 1$ shifted versions of η come into play, but the final proof can not be given here.

If such a look-up is made for all $\mathbf{d}_k$, the function χ has to extended for being defined over its full domain. This is done by the help of fuzzy singletons. Be

$$\Gamma_k = \frac{1}{2}\left[\frac{1}{\gamma^{\beta(k+1)}} - \frac{1}{\gamma^{n\beta(k+1)}}\right]. \tag{7}$$

Then, a fuzzy singleton is designed around $\eta(\mathbf{d}_k)$ by

$$\begin{aligned} \theta(\mathbf{d}_k; x) = \sigma[\mathit{\Gamma}_k^{-1}(x - \eta(\mathbf{d}_k)) + 1] \\ - \quad \sigma[\mathit{\Gamma}_k^{-1}(x - \eta(\mathbf{d}_k) - (\gamma - 2)b_k)] \end{aligned} \tag{8}$$

Here, $\sigma(x)$ is a continuous function with $\sigma(x) \equiv 0$ for $x \leq 0$ and $\sigma(x) \equiv 1$ for $x \geq 1$.

From the minimum separation of images of η, it can be seen that the fuzzy singletons around $\eta(\mathbf{d}_k)$ are not overlapping. The principal design of the outer function is given in fig. 3.

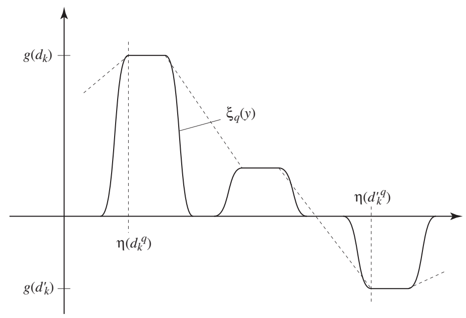

Fig. 3. Construction of the outer function $\chi_q(x)$ by putting fuzzy singletons around looked up values $\eta(d_k)$.

2.4 The Sprecher algorithm

Based on the constructions given in the foregoing sections, the Sprecher algorithm gives an iterative procedure for approximating the representation of a given function f by the Kolmogorov theorem.

The values $\epsilon > 0$ and $\delta > 0$ being choosen such that they fulfill

$$0 < \frac{m - n - 1}{m + 1}\epsilon + \frac{2n}{m + 1} \leq \delta < 1 \tag{9}$$

Therefrom $\epsilon < 1 - (n/(m - n + 1))$ (e.g. $\epsilon < 1/3$ for $n = 2$).

The algorithm starts with $f_0 \equiv f$ and the number $k_0 = 0$. Then, for each $r \geq 1$, from f_{r-1} a function f_r is computed in four layers. The sum of all functions f_r converges to f, thus approximating a representation of f.

I. Layer 1 computes all $\psi(x)$.
Function $f_{r-1}(x)$ with $x \in I^n$ is known. Now, a natural number $k_r > k_{r-1}$ is determined fulfilling that for all $p = 1, \ldots, n$ from $|x_p - y_p| \leq \gamma^{-k_r}$ it follows that $|f_{r-1}(x) - f_{r-1}(y)| \leq \epsilon||f_{r-1}||$. Via

$$\mathbf{d}^q_{k_r} = \mathbf{d}_{k_r} + q\sum_{r=2}^{k_r} \frac{1}{\gamma^r} = \mathbf{d}_{k_r} + qa_{k_r} \tag{10}$$

we obtain the values $\mathbf{d}^q_{k_r}$ for $q = 0, \ldots, m$. Now, we can compute the values $\psi(\mathbf{d}^q_{k_r})$ from eq. (4).

II. Layer 2 computes the linear combinations $\eta_q(x)$ of $\psi(x)$.
This is achieved by the equation

$$\eta_q(\mathbf{d}^q_{k_r}) = \sum_{p=1}^{n} \lambda_p \psi(d^q_{k_r p} + qa). \tag{11}$$

III. In Layer 3 the values $\chi_q[\eta_q(x)]$ are computed.
The $m+1$ functions of one variable $\chi_q^r(y)$ are given by

$$\chi_q^r(y) = \frac{1}{m+1} \sum_{\mathbf{d}_{k_r}^q} f_{r-1}(\mathbf{d}_{k_r}) \theta(\mathbf{d}_{k_r}^q; y) \tag{12}$$

The y-values are substituted by the η_q-values of layer 2:

$$\chi_q^r[\eta_q(x)] = \frac{1}{m+1} \sum_{\mathbf{d}_{k_r}^q} f_{r-1}(\mathbf{d}_{k_r}) \theta(\mathbf{d}_{k_r}^q; \eta_q(x)). \tag{13}$$

IV. In Layer 4 the $f_r(x)$ are computed as linear combination of the $\chi_q[\eta_q(x)]$.
This is done by computing

$$f_r(x) = f_{r-1}(x) - \sum_{q=0}^{m} \chi_q^r[\eta_q(x)]. \tag{14}$$

This ends the rth iteration step.

3 Image function

The procedure may be applied for the representation of image functions. From the Sprecher algorithm, the following algorithm can be derived for getting the representation of an image subsampled for e.g. $k = 2$ and resampling it for a higher k, as e.g. $k = 3$:

I. (**Offline-Phase**) For all $d_{31}, d_{32} = 0.000, \ldots, 0.999$ and $q = 0, \ldots, 4$:
1. Compute

$$\eta_3 = \eta(\mathbf{d}_3 + qa) = \eta(d_{31} + qa, d_{32} + qa).$$

2. Find a pair $\mathbf{d}_2$ and $\mathbf{d}_2'$ of points from D_2, which are neighbors in the ranking induced by η in D_2 and fulfilling:

$$\eta(\mathbf{d}_2) \leq \eta_3 \leq \eta(\mathbf{d}_2').$$

3. Now, be $d^q = (\mathbf{d}_x^q, \mathbf{d}_y^q)$ the value from $\mathbf{d}_2$ and $\mathbf{d}_2'$, for which in case $\theta(\mathbf{d}_2; \eta_3)\theta(\mathbf{d}_2'; \eta_3) \neq 0$ holds that $\theta(d^q; \eta_3) \neq 0$, otherwise one of both values is arbitraily choosen .
4. Enter into a table T_{q23} at position (d_{31}, d_{32}) the triple $(d_x^q - 0.01, d_y^q - 0.01, \theta(d^q; \eta_3))$.

II. (**Online-Phase**) A result array I_r of 1000×1000 positions is initialised with values 0. For all $d_{31}, d_{32} = 0.000, \ldots, 0.999$ and $q = 0, \ldots, 4$ then:
1. Get the value (d_x^q, d_y^q, θ) in T_{q23} at position (d_{31}, d_{32}).
2. Add the value

$$\frac{f(d_x^q, d_y^q)}{5} \cdot \theta$$

to the value at position (d_{31}, d_{32}) in I_r.

Figure 4 shows the result of resampling the Lena image from the representation of its image function for $k = 2$ at the next level $k = 3$. The right subfigure illustrates the manner how the Sprecher algorithm for one value of k attempts to generalize the unknown parts of the image function by placing fuzzy singletons around each already looked up position of the unit square. This gives the granular structure of the resampled image.

Fig. 4. Resampling of the Lena image according to the Kolmogorov theorem for $k = 3$ at all positions from D_2.

4 Summary

The construction of a representation of functions of n variables by superposition of functions of one variable and addition was considered in this paper in more detail. Following the approach given by David Sprecher, an algorithm was given for approximating such a representation. Based on the fact that the inner function of the Kolmogorov theorem induces a ranking of the points in the unit square, a qualitative analysis of such a representation can be derived. The generalization of one approximation step to the next was considered by providing the technical procedure of such a computation for an image (the Lena image).

References

1. Kolmogorov, A.N.: On the representation of continuous functions of several variables by superpositions of continuous functions of one variable and addition. Doklady Akademii Nauk SSSR **114** (1957) 679–681 (in Russisch).
2. Hecht-Nielsen, R.: Kolmogorov's mapping neural network existence theorem. In: Proceedings of the First International Conference on Neural Networks. Volume III., IEEE Press, New York (1987) 11–13
3. Sprecher, D.: On the structure of continuous functions of several variables. Transcations of the American Mathematical Society **115** (1965) 340–355
4. Cotter, N.E., Guillerm, T.J.: The CMAC and a theorem of Kolmogorov. Neural Networks **5** (1992) 221–228
5. Nakamura, M., Mines, R., Kreinovich, V.: Guaranteed intervals for Kolmogorov's theorem (and their possible relation to neural networks). Interval Computations **3** (1993) 183–199
6. Sprecher, D.: A numerical implementation of Kolmogorov's superpositions I. Neural Networks **9** (1996) 765–772
7. Sprecher, D.: A numerical implementation of Kolmogorov's superpositions II. Neural Networks **10** (1997) 447–457
8. Kreinovich, V., Nguyen, H., Sprecher, D.: Normal forms for fuzzy logic — an application of kolmogorov's theorem. Technical Report UTEP-CS-96-8, University of Texas at El Paso (1996)
9. Nguyen, H., Kreinovich, V.: Kolmogorov's theorem and its impact on soft computing. In Yager, R.R., Kacprzyk, J., eds.: The Ordered Weighted Averaging operators. Theory and Applications. Kluwer Academic Publishers (1997) 3–17