Journal of Uncertain Systems
Vol.6, No.2, pp.122-127, 2012
Online at: www.jus.org.uk

Orthogonal Bases are the Best: A Theorem Justifying Bruno Apolloni's Heuristic Neural Network Idea

Jaime Nava, Vladik Kreinovich*
Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA

Received 19 August 2011; Revised 14 April 2012

Abstract

One of the main problems with neural networks is that they are often very slow in learning the desired dependence. To speed up neural networks, Bruno Apolloni proposed to othogonalize neurons during training, i.e., to select neurons whose output functions are orthogonal to each other. In this paper, we use symmetries to provide a theoretical explanation for this heuristic idea.

Keywords: neural networks, symmetries, optimality of orthogonal bases, Apolloni's orthogonalization of neural networks

1 Formulation of the Problem

Neural networks: brief reminder. In the traditional (3-layer) neural networks, the input values $x_1, \ldots, x_n$:

- first go through the non-linear layer of "hidden" neurons, resulting in the values

$$y_i = s_0\left(\sum_{j=1}^{n} w_{ij} \cdot x_j - w_{i0}\right), \quad 1 \le i \le m,$$

- after which a linear neuron combines the results y_i into the output $y = \sum_{i=1}^{m} W_i \cdot y_i - W_0$.

Here, W_i and w_{ij} are *weights* selected based on the data, and $s_0(x)$ is a non-linear *activation function*. Usually, the "sigmoid" activation function $s_0(x) = 1/(1+\exp(-x))$ is used.

The weights W_i and w_{ij} are selected so as to fit the data, i.e., that

$$y^{(k)} \approx f\left(x_1^{(k)}, \ldots, x_n^{(k)}\right) \text{ for all } k,$$

where:

- $x_1^{(k)}, \ldots, x_n^{(k)}$ $(1 \le k \le N)$ are given values of the inputs, and
- $y^{(k)}$ are given values of the output.

One of the problems with the traditional neural networks is that in the process of learning – i.e., in the process of adjusting the values of the weights to fit the data – some of the neurons are duplicated, i.e., we get $w_{ij} = w_{i'j}$ for some $i \ne i'$ and thus, $y_i = y_{i'}$.

As a result, we do not fully use the learning capacity of a neural network, since when $y_i = y_{i'}$, we can get the same approximation with fewer hidden neurons.

Apolloni's idea. To avoid the above redundancy problem, B. Apolloni and others suggested [1] that we *orthogonalize* the neurons during training, i.e., that we make sure that the corresponding functions $y_i(x_1, \ldots, x_n)$ remain orthogonal in the sense that

$$\langle y_i, y_j \rangle = \int y_i(x) \cdot y_j(x)\, dx = 0.$$

*Corresponding author. Email: vladik@utep.edu (V. Kreinovich).

Challenge. Since Apolloni *et al.* heuristic idea works well, it is desirable to look for its precise mathematical justification.

What we do in this paper. We provide such a justification in terms of symmetries.

Comment. This result was first presented at a conference [3].

2 Why Symmetries?

Why symmetries. At first glance, the use of symmetries in neural networks may sound somewhat strange, because there are no *explicit* symmetries there, but *hidden* symmetries have been actively used in neural networks. For example, they are the only way to explain the empirically observed advantages of the sigmoid activation function; see, e.g., [2, 4].

Symmetry: a fundamental property of the physical world. One of the main objectives of science is prediction. What is the usual basis for prediction? We observed similar situations in the past, and we expect similar outcomes. In mathematical terms, similarity corresponds to *symmetry*, and similarity of outcomes – to *invariance.*

For example, we dropped the ball, it fell down. We conclude that if we drop it at a different location and/or at a different orientation, it will also fall down. Why – because we believe that the process is invariant with respect to shifts, rotations, etc.

This fundamental role of symmetries is well recognized in modern physics, to the extent that, starting with the quark theory, theories are usually formulated in terms of the corresponding symmetries – and not in terms of differential equations as it was in Netwon's time and later. Of course, once the symmetries are known, we can determine the equations, but they are no longer the original formulation.

It is therefore natural to apply symmetries to neural networks as well.

Basic symmetries: scaling and shift. What are the basic symmetries? Typically, we deal with the numerical values of a physical quantity. Numerical values depend on the *measuring unit.* If we use a new unit which is λ times smaller, numerical values are multiplied by λ: $x \to \lambda \cdot x$. For example, x meters $= 100 \cdot x$ cm. The transformation $x \to \lambda \cdot x$ is usually called *scaling.*

Another possibility is to change the starting point. For example, instead of measuring time from year 0, we can start measuring it from some more distant year in the past. If we use a new starting point which is s units smaller, then the quantity which was originally represented by the number x is now represented by the new value $x + s$. The transformation $x \to x + s$ is usually called a *shift.*

So, we arrive at the following natural requirement: that the physical formulas should not depend on the choice of a measuring unit or of a starting point. Together, scaling and shifts form *linear transformations* $x \to a \cdot x + b$. Thus, in mathematical terms, this means that the physical formulas be invariant under linear transformations.

Basic nonlinear symmetries. Sometimes, a system also has *nonlinear* symmetries. To find such non-linear symmetries, we can take into account that if a system is invariant under f and g, then

- it is invariant under their composition $f \circ g$, and
- it is invariant under the inverse transformation f^{-1}.

In mathematical terms, this means that symmetries form a *group.*

In practice, at any given moment of time, we can only store and describe finitely many parameters. Thus, it is reasonable to restrict ourselves to *finite-dimensional* groups.

One of the first researcher to explore this idea was Norbert Wiener, the father of cybernetics. He formulated a question: describe all finite-dimensional groups that contain all linear transformations. For transformations from real numbers to real numbers, the answer to this question are known: all elements of this group are fractionally-linear functions $x \to (a \cdot x + b)/(c \cdot x + d)$.

Symmetries explain the choice of an activation function. Let us show that such non-linear symmetries explain the formula for the *activation function* $f(x) = 1/(1 + \exp(-x))$.

Indeed, a change in the input starting point has the form $x \to x + s$. It is reasonable to require that the new output $f(x+s)$ is equivalent to the $f(x)$ modulo an appropriate transformation. We have just mentioned that appropriate transformations are fractionally linear. Thus, we conclude that for every s, there exist values $a(s)$, $b(s)$, $c(s)$, and $d(s)$ for which

$$f(x+s) = \frac{a(s) \cdot f(x) + b(s)}{c(s) \cdot f(x) + d(s)}.$$

Differentiating both sides by s and equating s to 0, we get a differential equation for $f(x)$. Its known solution is the sigmoid activation function – which can thus be explained by symmetries.

3 Explanation of Apolloni's Heuristic Idea

Towards formulating the problem in precise terms. We must select a basis $e_0(x)$, $e_1(x)$, ..., $e_n(x)$, ...so that each function $f(x)$ is represented as $f(x) = \sum_i c_i \cdot e_i(x)$. For example,

- an expansion in Taylor series corresponds to choosing the basis $e_0(x) = 1$, $e_1(x) = x$, $e_2(x) = x^2$, ...
- an expansion in Fourier series corresponds to selecting the basis $e_i(x) = \sin(\omega_i \cdot x)$.

Once the basis is selected, to store the information about the function $f(x)$, we store the coefficients c_0, c_1, ..., corresponding to this basis.

From this viewpoint, one of the possible criteria for selecting the basis can be that the selected basis should require, on average, the smallest number of bits to store $f(x)$ with given accuracy. We can come up with several similar criteria.

For all these criteria, we can take into account that storing a number c_i and storing the opposite number $-c_i$ take the same space. Thus, changing one of the basis function $e_i(x)$ to $e_i'(x) = -e_i(x)$ (which we lead to exactly this change $c_i \to -c_i$) does not change accuracy or storage space. So, we conclude that

- if $e_0(x)$, ..., $e_{i-1}(x)$, $e_i(x)$, $e_{i+1}(x)$, ...is an optimal basis,
- then the basis $e_0(x)$, ..., $e_{i-1}(x)$, $-e_i(x)$, $e_{i+1}(x)$, ...is also optimal.

Uniqueness of the optimal solution. Due to the previous argument, we do not select a single basis, we select a family $\pm e_0(x)$, $\pm e_1(x)$, ..., in which each function is determined modulo its sign. Out of all such families, we should select the optimal one.

In general, an optimization problem may have have several optimal solutions. In this case, we can use this non-uniqueness to optimize something else. For example,

- if two sorting algorithms are equally fast in the worst case $t^w(A) = t^w(A')$,
- we can select the one with the smallest average time $t^a(A) \to \min$.

In effect, by introducing the additional criterion, we now have a new criterion: A is better than A' if either $t^w(A) < t^w(A')$ or $(t^w(A) = t^w(A')$ and $t^a(A) < t^a(A'))$.

If this new criterion also has several optimal solutions, we can optimize something else, etc., until we end up with a unique optimal solution. So, non-uniqueness means that the original criterion was not final. Relative to a *final* criterion, there is *only one* optimal solution.

For our problem, this uniqueness means that

- once we have one optimal basis
$$e_0(x), \quad e_1(x), \quad e_2(x), \quad \ldots,$$
- all other optimal bases have the form
$$\pm e_0(x), \quad \pm e_1(x), \quad \pm e_2(x), \quad \ldots.$$

How to describe average accuracy. Our objective is to describe average accuracy, or average number of bits, etc. For example, we may want to know the average value of the We also want to know the mean square distance $\int (f(x) - f_{\approx}(x))^2\,dx$ between the original function $f(x)$ and its approximation $f_{\approx}(x)$.

To describe these averages, we need to know the corresponding probability distribution on the set of all possible functions $f(x)$.

Dependencies $f(x)$ come from many different factors. Due to Central Limit Theorem, it is thus reasonable to assume that the distribution on $f(x)$ is Gaussian. If $m(x) \stackrel{\text{def}}{=} E[f(x)] \neq 0$, we can store differences $\Delta f(x) \stackrel{\text{def}}{=} f(x) - m(x)$, for which $E[\Delta f(x)] = 0$. Thus, without losing generality, we can assume that $E[f(x)] = 0$.

Such Gaussian distributions are uniquely determined by their covariances $C(x,y) \stackrel{\text{def}}{=} E[f(x) \cdot f(y)]$. A general Gaussian distribution can be described by independent components: $f(x) = \sum\limits_i \eta_i \cdot f_i(x)$, where $E[\eta_i \cdot \eta_j] = 0$, $i \neq j$. The corresponding functions $f_i(x)$ are eigenfunctions of the covariance function $C(x,y) = E[f(x)f(y)]$:

$$\int C(x,y) \cdot f_j(y)\,dy = \lambda_j \cdot f_j(x).$$

The basis formed by these functions is known as the *Kahrunen-Loeve* (KL) basis. The functions from the KL basis together with the corresponding eigenvalues λ_i uniquely determine the corresponding probability distribution – and thus, the value of the optimality criterion.

Functions from this *KL basis* are orthogonal; they are usually selected to be orthonormal, i.e., satisfy the condition $\int f_j^2(x)\,dx = 1$.

In the general case, when all eigenvalues λ_j are different, each eigenfunction $f_j(x)$ is determined uniquely modulo $f_j(x) \to -f_j(x)$.

One can easily see that if we change one of functions $f_j(x)$ from the KL basis to to $-f_j(x)$, we get a KL basis. Under these change, the values $E[f(x) \cdot f(y)]$ and $\int f^2(x)\,dx$ do not change – and thus, optimality criteria based on these values do not change. Thus, we arrive at the following formulation of the problem.

Formulation of the problem in precise terms. We have an optimality criterion described in terms of a sequence of orthonormal functions $f_j(x)$ and a sequence of corresponding numbers λ_i. We know that functions $\pm f_j(x)$ determine the exact same criterion as the original functions $f_j(x)$.

We consider the generic case, if which all the eigenvalues λ_j are different.

Based on this criterion, we must select an optimal basis $e_0(x)$, $e_1(x)$, ..., $e_i(x)$, ... Each function from the desired basis can be represented as a linear combination of functions from the KL basis

$$e_i(x) = \sum_j a_{ij} \cdot f_j(x).$$

Thus, selecting an optimal basis is equivalent to selecting the matrix of values a_{ij}, and the optimality criterion is equivalent to selecting a class of all matrices corresponding to optimal functions.

Of course, since the vectors $e_i(x)$ must form a basis, we cannot have $e_i(x) \equiv 0$, i.e., for every i, at least one value a_{ij} must e different from 0. We will call such matrices *non-trivial*.

We have mentioned that if we change one of the functions $f_{j_0}(x)$ to $-f_{j_0}(x)$, the criterion does not change. Thus, the following functions also form an optimal basis

$$e_i'(x) = \sum_{j \neq j_0} a_{ij} \cdot f_j(x) - a_{ij_0} \cdot f_{j_0}(x).$$

These functions correspond to the new matrix a'_{ij} for which $a'_{ij_0} = -a_{ij_0}$ and $a'_{ij} = a_{ij}$ for all $j \neq j_0$.

We also require that every optimal basis has the form $e_i'(x) = \pm e_i(x)$. Thus, we arrive at the following definition.

Definition. *Let $f_i(x)$ be a sequence of linearly independent functions.*

- *We say that a matrix a_{ij} is* non-trivial *if for every i, there exists a j for which $a_{ij} \neq 0$.*
- *By an* optimality criterion*, we mean a class A of non-trivial matrices a_{ij}.*

- *For each matrix $a_{ij} \in A$, the functions $e_i(x) = \sum_j a_{ij} \cdot f_j(x)$ are called* optimal *functions corresponding to this matrix.*
- *We say that the optimality criterion is* invariant *if for every matrix $a_{ij} \in A$ and for every j_0, the matrix a'_{ij}, for which $a'_{ij_0} = -a_{ij_0}$ and $a'_{ij} = a_{ij}$ for all $j \neq j_0$, also belongs to the class A.*
- *We say that the optimality criterion is* final *if for every two matrices $a_{ij}, a'_{ij} \in A$ and for every i, the corresponding optimal functions $e_i(x) = \sum_j a_{ij} \cdot f_j(x)$ and $e'_i(x) = \sum_j a'_{ij} \cdot f_j(x)$ differ only by sign, i.e., either $e'_i(x) = e_i(x)$ or $e'_i(x) = -e_i(x)$.*

Theorem. *If an optimality criterion is invariant and final, then each optimal function $e_i(x)$ has the form $e_i = a_{ij_0} \cdot f_{j_0}(x)$ for some j_0.*

Proof. Indeed, let a_{ij} be a matrix from the optimal criterion. Since the matrix is non-trivial, for every i, there exist a j_0 for which $a_{ij_0} \neq 0$. Since the optimality criterion is invariant, the class A also contains the matrix a'_{ij} for which $a'_{ij_0} = -a_{ij_0}$ and $a'_{ij} = a_{ij}$ for all $j \neq j_0$. For this new matrix, the corresponding optimal functions have the form

$$e'_i(x) = \sum_{j \neq j_0} a_{ij} \cdot f_j(x) - a_{ij_0} \cdot f_{j_0}(x).$$

Since the optimality criterion is final, this expression must be equal either to $e_i(x)$ or to $-e_i(x)$.

If $e'_i(x) = e_i(x)$, we would have

$$\sum_{j \neq j_0} a_{ij} \cdot f_j(x) - a_{ij_0} \cdot f_{j_0}(x) = \sum_{j \neq j_0} a_{ij} \cdot f_j(x) + a_{ij_0} \cdot f_{j_0}(x).$$

The difference between the two sides is equal to 0, hence $a_{ij_0} \cdot f_{j_0}(x) = 0$ and $a_{ij_0} = 0$, but we have selected j_0 for which $a_{ij_0} \neq 0$. Thus, $e'_i(x) = e_i(x)$ is impossible, so we must have $e'_i(x) = -e_i(x)$, i.e.,

$$\sum_{j \neq j_0} a_{ij} \cdot f_j(x) - a_{ij_0} \cdot f_{j_0}(x) = -\sum_{j \neq j_0} a_{ij} \cdot f_j(x) - a_{ij_0} \cdot f_{j_0}(x).$$

Since the functions $f_j(x)$ are linearly independent, this equality implies that the coefficients at all $f_j(x)$ in both sides must coincide. In particular, by comparing the coefficients at $f_j(x)$ for every $j \neq j_0$, we conclude that $a_{ij} = -a_{ij}$ hence $a_{ij} = 0$. So, $a_{ij} = 0$ for all $j \neq j_0$. The theorem is proven.

Discussion. We have proved that for the optimal basis $e_i(x)$ and for the KL basis $f_j(x)$, each $e_i(x)$ has the form

$$e_i(x) = a_{ij_0} \cdot f_{j_0}(x) \text{ for some } a_{ij_0}.$$

We know that the elements $f_j(x)$ of the KL basis are orthogonal. So, we conclude that the elements $e_i(x)$ of the optimal basis are orthogonal as well.

Apolloni's idea was to always make sure that we use an orthogonal basis. This idea has been empirically successful. Our new result provides a theoretical justification for Apolloni's idea.

Acknowledgments

This work was supported in part:

- by the National Science Foundation grants HRD-0734825 and DUE-0926721, and
- by Grant 1 T36 GM078000-01 from the National Institutes of Health.

References

[1] Apolloni, B., S. Bassis, and L. Valerio, A moving agent metaphor to model some motions of the brain actors, *Abstracts of the Conference "Evolution in Communication and Neural Processing from First Organisms and Plants to Man ... and Beyond"*, Modena, Italy, p.17, 2010.

[2] Kreinovich, V., and C. Quintana, Neural networks: what non-linearity to choose?, *Proceedings of the 4th University of New Brunswick AI Workshop*, Fredericton, N.B., Canada, pp.627–637, 1991.

[3] Nava, J., and V. Kreinovich, Orthogonal bases are the best: a theorem justifying Bruno Apolloni's heuristic neural network idea, *Abstracts of the 9th Joint NMSU/UTEP Workshop on Mathematics, Computer Science, and Computational Sciences*, Las Cruces, New Mexico, 2011.

[4] Nguyen, H.T., and V. Kreinovich, *Applications of Continuous Mathematics to Computer Science*, Kluwer, Dordrecht, 1997.