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

Jaime Nava and Vladik Kreinovich

Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
Emails: jenava@miners.utep.edu,
vladik@utep.edu

1. 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 $s_0(z)$ is a non-linear *activation function*.
- Usually, the "sigmoid" activation function is used:
$$s_0(z) = \frac{1}{1+\exp(-z)}.$$

2. Training a Neural Network: Reminder

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

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

  - $x_1^{(k)}, \ldots, x_n^{(k)}$ $(1 \le k \le N)$ 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 $i \neq i'$ and thus, $y_i = y_{i'}$.

- As a result, we do not fully use the learning capacity of a neural network: we could use fewer hidden neurons.

3. Apolloni's Idea

- *Problem* (reminder):
  - 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 $i \neq i'$ and thus, $y_i = y_{i'}$.
- To avoid this problem, B. Apolloni et al. suggested that we *orthogonalize* the neurons during training.
- In other words, we make sure that the corresponding functions $y_i(x_1, \ldots, x_n)$ remain orthogonal:

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

- Since Apolloni *et al.* idea works well, it is desirable to look for its precise mathematical justification.
- We provide such a justification in terms of symmetries.

4. Why Symmetries?

- At first glance, the use of symmetries in neural networks may sound somewhat strange.
- Indeed, there are no *explicit* symmetries there.
- However, as we will show, *hidden* symmetries have been actively used in neural networks.
- For example, symmetries explain the empirically observed advantages of the sigmoid activation function

$$s_0(z) = \frac{1}{1 + \exp(-z)}.$$

5. Symmetry: a Fundamental Property of the Physical World

- *One of the main objectives of science:* prediction.
- *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.*
- *Example:* we dropped the ball, it fall down.
- *Symmetries:* shift, rotation, etc.
- *In modern physics:* theories are usually formulated in terms of symmetries (not diff. equations).
- *Natural idea:* let us use symmetry to describe uncertainty as well.

6. Basic Symmetries: Scaling and Shift

- *Typical situation:* we deal with the numerical values of a physical quantity.
- Numerical values depend on the *measuring unit.*
- *Scaling:* if we use a new unit which is λ times smaller, numerical values are multiplied by λ: $x \to \lambda \cdot x$.
- *Example:* x meters $= 100 \cdot x$ cm.
- *Another possibility:* change the starting point.
- *Shift:* if we use a new starting point which is s units before, then $x \to x + s$ (example: time).
- Together, scaling and shifts form *linear transformations* $x \to a \cdot x + b$.
- *Invariance:* physical formulas should not depend on the choice of a measuring unit or of a starting point.

7. Basic Nonlinear Symmetries

- Sometimes, a system also has *nonlinear* symmetries.
- 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.
- *Question* (N. Wiener): describe all finite-dimensional groups that contain all linear transformations.
- *Answer* (for real numbers): all elements of this group are fractionally-linear $x \to (a \cdot x + b)/(c \cdot x + d)$.

8. Symmetries Explain the Choice of an Activation Function

- *What needs explaining:* formula for the *activation function* $f(x) = 1/(1 + e^{-x})$.
- A change in the input starting point: $x \to x + s$.
- *Reasonable requirement:* the new output $f(x+s)$ equivalent to the $f(x)$ mod. appropriate transformation.
- *Reminder:* all appropriate transformations are fractionally linear.
- *Conclusion:* $f(x + s) = \dfrac{a(s) \cdot f(x) + b(s)}{c(s) \cdot f(x) + d(s)}$.
- 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.

9. Towards Formulating the Problem in Precise Terms

- We select a basis $e_0(x), e_1(x), \ldots, e_n(x), \ldots$ so that each f-n $f(x)$ is represented as $f(x) = \sum_i c_i \cdot e_i(x)$; e.g.:
  - Taylor series: $e_0(x) = 1$, $e_1(x) = x$, $e_2(x) = x^2$, $\ldots$
  - Fourier transform: $e_i(x) = \sin(\omega_i \cdot x)$.
- We store c_0, c_1, $\ldots$, instead of the original f-n $f(x)$.
- *Criterion:* e.g., smallest # of bits to store $f(x)$ with given accuracy.
- *Observation:* storing c_i and $-c_i$ takes the same space.
- Thus, changing one of $e_i(x)$ to $e_i'(x) = -e_i(x)$ does not change accuracy or storage space, so:
  - if $e_0(x), \ldots, e_{i-1}(x), e_i(x), e_{i+1}(x), \ldots$ is an opt. base,
  - $e_0(x), \ldots, e_{i-1}(x), -e_i(x), e_{i+1}(x), \ldots$ is also optimal.

10. Uniqueness of the Optimal Solution

- *Reminder:* we select the basis $\pm e_0(x)$, $\pm e_1(x)$, ...
- Each function is determined modulo its sign.
- Sometimes, we have several optimal solutions.
- Then, we can use an additional criterion; e.g.:
  - 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, we have a new criterion: A is better than A' if $t^w(A) < t^w(A')$ or $(t^w(A) = t^w(A')$ and $t^a(A) < t^a(A'))$.
- So, non-uniqueness means that the original criterion was not final.
- Relative to a *final* criterion, there is *only one* optimal solution.

11. Uniqueness of the Optimal Basis

- *Reminder:*
  - we select the basis $\pm e_0(x)$, $\pm e_1(x)$, $\pm e_3(x)$, $\ldots$;
  - each function is determined modulo its sign.
- *Optimal solutions* are unique:
  - relative to a *final* criterion,
  - there is *only one* optimal solution.
- *Conclusion:* it is reasonable to require 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

$$\pm e_0(x), \quad \pm e_1(x), \quad \pm e_2(x), \quad \ldots$$

12. How to Describe Average Accuracy

- What is a probability distribution on $f(x)$?
- 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, w.l.o.g., 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 Gaussian distribution can be described by indep. components: $f(x) = \sum_i \eta_i \cdot f_i(x)$, w/$E[\eta_i \cdot \eta_j] = 0$, $i \neq j$.
- We also want to know the mean square values $\int (f(x) - f_{\approx}(x))^2 \, dx$.

13. Kahrunen-Loeve (KL) Basis

- A Gaussian distribution can be described by indep. components: $f(x) = \sum_i \eta_i \cdot f_i(x)$, w/$E[\eta_i \cdot \eta_j] = 0$, $i \neq j$.
- We also want to know $\int (f(x) - f_{\approx}(x))^2\, dx$.
- Idea: use a basis $f_j(x)$ of eigenfunctions of the covariance function $C(x, y) = E[f(x)f(y)]$:

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

- Functions from this *KL basis* are orthogonal; they are usually selected to be orthonormal $\int f_j^2(x)\, dx = 1$.
- If we change some $f_j(x)$ to $-f_j(x)$, we get a KL basis.
- So, criteria depending on $E[f(x) \cdot f(y)]$ and $\int f^2(x)\, dx$ do not change.
- In the general case, when all λ_j are different, each $f_j(x)$ is determined uniquely modulo $f_j(x) \to -f_j(x)$.

14. Proof of the Main Result

- *Let:* $e_i(x)$ be an optimal basis, and let $f_j(x)$ be a KL basis, then $e_i(x) = \sum_j a_{ij} \cdot f_j(x)$.
- *Reminder:* if we change one of the functions $f_{j_0}(x)$ to $-f_{j_0}(x)$, the criterion does not change.
- *Thus:* the following f-ns also form an optimal basis:

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

- *Reminder:* $\forall$ optimal basis has the form $\pm e_i(x)$, thus:

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

- *So:* if $a_{ij_0} \neq 0$, then $a_{ij} = 0$ for all $j \neq j_0$.
- *Thus:* each $e_i(x)$ has the form $e_i(x) = a_{ij_0} \cdot f_{j_0}(x)$ for some j_0.

15. Conclusions

- *We 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.
- *Conclusion:* the elements of the optimal basis are orthogonal.
- *Apolloni's idea:* always make sure that we use an orthogonal basis.
- *Fact:* this idea has been empirically successful.
- *New result:* Apolloni's idea has been theoretically justified.

16. 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.

17. References

- B. Apolloni, 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, November 18–19, 2010, p. 17.
- V. Kreinovich and C. Quintana. "Neural networks: what non- linearity to choose?," *Proc. 4th University of New Brunswick AI Workshop*, Fredericton, N.B., Canada, 1991, pp. 627–637.
- H. T. Nguyen and V. Kreinovich, *Applications of Continuous Mathematics to Computer Science*, Kluwer, Dordrecht, 1997.