3 A functional equation

> Words are a pretty fuzzy substitute
> for mathematical equations.
>
> Isaac Asimov

Series like (2.29) or (2.56) can be directly implemented in a Neural Network architecture to serve as reference solutions, and more general polynomials can be implemented as well by means of the polarization formula. Nevertheless it has the price that a complicated structure like (2.38) is needed. In this Chapter, we review the possibilities offered by a simpler different approach which comes from [24]. This approach generalizes a well known functional equation [41, 42] which has already been encountered in (2.28). The functional equation can be written as

$$f_1^{\mathrm{obj}}(x) = \frac{1}{4}g(x) + \frac{1}{4}f_1^{\mathrm{obj}}(g(x)), \qquad f_1^{\mathrm{obj}}(x) = x - x^2, \tag{3.1}$$

where the objective function f_1^{obj} can be considered as the unknown. Then the identity (2.24) is the expansion of f_1^{obj} in terms of a Neumann series.
Hereafter, a more general functional equation is constructed with three main properties: a) it has general polynomial solutions under the conditions of the main Theorem, b) it is contractive, so is easily solved by any kind of standard fixed point procedure and, c) the converging fixed point iterations can be implemented in a Feedforward Deep Network or in a Residual Neural Network (RNN).

3.1 A contractive functional equation

Consider an equi-distributed subdivision $0 = x_0 < x_1 < \cdots < x_j = jh < \cdots < x_m = 1$ in $m \geq 1$ subintervals $[x_j, x_{j+1}]$, where we note $h = 1/m$. We note $P^n = \{q \text{ real polynomial of degree } \leq n\}$. The set of continuous piecewise linear functions is

$$V_h = \left\{ u \in C^0(I),\ u_{|(x_j,x_{j+1})} \in P^1 \text{ for all } 0 \leq j \leq m-1 \right\}.$$

Similarly with the classical Finite Element setting [16], some basis functions are chosen in a subset of V_h, even if they are not basis functions in the classical Finite Element sense. In the proposed construction, they are taken in subset $E_h \subset V_h$

$$E_h = \{ u \in V_h :\ u(I) \subset I,\ u \text{ is non constant on exactly one subinterval} \}.$$

The assumption $u(I) \subset I$ is critical to get the contraction property under the form of Lemma 3.1.8. Our interest in this set is because functions in E_h and V_h

are easily assembled or implemented in Neural Networks with the ReLU function $R(x) = \max(0, x)$, see [34, 105, 19, 22, 60]. We consider the problem below.

Problem 3.1.1. *Given a real polynomial $f^{\mathrm{obj}} \in P^n$, find*

$$(e_0, e_1, \dots, e_r, \beta_1, \dots, \beta_r) \in V_h \times (E_h)^r \times \mathbb{R}^r$$

such that the identity below holds

$$f^{\mathrm{obj}}(x) = e_0(x) + \sum_{i=1}^{r} \beta_i f^{\mathrm{obj}}(e_i(x)), \quad x \in I, \tag{3.2}$$

with the contraction condition

$$K < 1, \qquad K = \sum_{i=1}^{r} |\beta_i|. \tag{3.3}$$

Because of the external composition by f^{obj} in the last sum, the e_i's are not basis functions in the sense of the Finite Element Method [16]. Once the

$$(e_0, e_1, \dots, e_r, \beta_1, \dots, \beta_r) \in V_h \times (E_h)^r \times \mathbb{R}^r$$

are determined, equation (1.1) can be seen as a functional equation with f^{obj} as a solution. This functional equation is contractive because of the condition (3.3).

3.1.1 Two examples

Equation (3.1) is not exactly (3.2) because the internal function in the composition $f_1^{\mathrm{obj}} \circ g$ is made with g, and $g \notin E_h$. Nevertheless it is not difficult to rewrite (3.15) under the form (3.2). Take $h = 1/2$. Set $e_1(x) = \min(2x, 1)$ and $e_2(x) = \min(2(1-x), 1)$ with $e_1, e_2 \in E_h$. Set $e_0 = \frac{1}{4}g$. One obtains

$$f_1^{\mathrm{obj}}(x) = e_0(x) + \frac{1}{4} f_1^{\mathrm{obj}}(e_1(x)) + \frac{1}{4} f_1^{\mathrm{obj}}(e_2(x)) \tag{3.4}$$

where the contraction property (3.3) is satisfied with a constant $K = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

The second example concerns the function $f_2^{\mathrm{obj}}(x) = x^3$, still for $h = 1/2$. Set $e_3(x) = 1 - e_2(x)$. One can check the formula

$$f_2^{\mathrm{obj}}(x) = e_0(x) + \frac{1}{8} f_2^{\mathrm{obj}}(e_1(x)) + \frac{1}{8} f_2^{\mathrm{obj}}(e_2(x)) + \frac{1}{4} f_2^{\mathrm{obj}}(e_3(x)) \tag{3.5}$$

with $e_0(x) = \frac{3}{4}e_3(x) - \frac{1}{8}$. The contraction number is $K = \frac{1}{8} + \frac{1}{8} + \frac{1}{4} = \frac{1}{2}$.

3.1.2 Number of intervals

Consider the case $m = 1$, that is just one subinterval, and take a polynomial f^{obj} with $\deg(f^{\mathrm{obj}}) \geq 2$. For $f^{\mathrm{obj}}(x) = x^n+$ low order terms, then by equating the coefficients of x^n on both sides, one gets $1 = \sum \beta_i(\mu_i)^n$ with $\mu_i = e_i'(x) \in \mathbb{R}$. Because $e_i(I) \subset I$, then $|\mu_i| \leq 1$. So $1 \leq \sum |\beta_i||\mu_i|^n \leq \sum |\beta_i|$. It means the contraction condition (3.3) is not satisfied and Problem 3.1.1 has only trivial solutions if there is only one subinterval, that is if $m = 1$. Since the contraction property is crucial in the construction, we will consider two or more subintervals in the sequel

$$m \geq 2.$$

3.1.3 Non polynomial solutions

In these notes we concentrate on situations where the functional equation (3.2) has polynomial solutions, but it is not mandatory. This can be infered by comparison of (3.1) which has a polynomial solution and (2.50) which solution is the Takagi function. So if one varies the coefficient β_i in (3.2), all other ingredients kept the same, then non polynomial solutions will be produced by the functional equation. Probably some of these solutions have exotic regularity properties, as the for the Takagi function. Their characterization is an open problem not considered in these notes.

3.1.4 Construction of polynomial solutions

Theorem 3.1.2. *Let $f^{\mathrm{obj}} \in P^n$. There exists a threshold value $m_* \geq 2$ such that the functional equation (3.2) has a solution with the contraction property (3.3) for all $m \geq m_* \Longleftrightarrow h \leq 1/m_*$.*

Taking the basis functions as $e_i(x) = a_i + \frac{b_i - a_i}{h}(x - x_j)$ in the subinterval where they are non constant, the parameters (a_i, b_i) can be taken as in Lemma 3.1.6 in the general case. They can also be taken as in Lemma 3.1.5 if $f^{\mathrm{obj}}(x) = x^n$ is a monomial.

For $n \geq 2$ one can take: a) the number of basis functions which are non constant in a given subinterval equal $n - 1$; and b) the basis functions duplicated by translation from of subinterval to the other. Then the total number of basis functions is

$$r = m(n - 1).$$

The proof is based on decoupled and simpler problems posed in subintervals $[x_j, x_{j+1}]$. Let us note the second derivative of f^{obj} as $q = (f^{\text{obj}})'' \in P^{n-2}$. The collection of reduced problems for all subinterval $[x_j, x_{j+1}]$ writes as follows.

Problem 3.1.3. *For all subintervals $0 \le j \le m-1$, find triples $(a_i, b_i, \gamma_i) \in I \times I \times \mathbb{R}$ $(1 \le i \le s)$ such that $b_i - a_i \neq 0$ for all i and*

$$q(x_j + hy) = \sum_{i=1}^{s} \gamma_i q(a_i + (b_i - a_i)y), \qquad y \in I. \tag{3.6}$$

Lemma 3.1.4. *The equation (1.1) is equivalent to the equations (3.6) for all $0 \le j \le m-1$.*

Proof. The proof is in two parts.
(3.2) ⇒ (3.6): on the interval $[x_j, x_{j+1}]$, one can write $e_i(x) = a_i + \frac{b_i - a_i}{h}(x - x_j)$. By differentiation, a solution to (3.2) gives $q(x) = \sum_{i=1}^{r} \beta_i \left(\frac{b_i - a_i}{h}\right)^2 p(e_i(x))$ for $x_j < x < x_{j+1}$. Retaining in the sum only the indices i such that $b_i - a_i \neq 0$, one gets (3.6) where $x = x_j + hy$ and $e_i(x) = a_i + (b_i - a_i)y$.
(3.6) ⇒ (1.1): rewrite the discrete quantities in (3.6) with another lower index j which refers to the interval in which this equation is considered. It defines $a_{i,j}$, $b_{i,j}$ and $\gamma_{i,j}$. Define

$$e_{i,j}(x) = \begin{cases} a_{i,j} & \text{for } 0 \le x \le x_j, \\ a_{i,j} + \frac{b_{i,j} - a_{i,j}}{h}(x - x_j) & \text{for } x_j \le x \le x_{j+1} = x_j + h, \\ b_{i,j} & \text{for } x_{j+1} \le x \le 1. \end{cases}$$

Define also

$$\beta_{i,j} = \frac{h^2}{(b_{i,j} - a_{i,j})^2} \gamma_{i,j}, \qquad \text{where } b_{i,j} - a_{i,j} \neq 0. \tag{3.7}$$

Consider the function

$$e_0(x) = f^{\text{obj}}(x) - \sum_j \sum_i \beta_{i,j} f^{\text{obj}}(e_{i,j}(x)). \tag{3.8}$$

By construction e_0 is continuous and its second derivative is zero in all open subintervals (x_j, x_{j+1}). Therefore $e_0 \in V_h$ which ends the proof. □

If the polynomials $y \mapsto q(a_{i,j} + (b_{i,j} - a_{i,j})y)$, $1 \le i \le s$, generate a complete system in P^{n-2}, then the equation (3.6) has a solution. That is why we will consider from now on that

$$s = \dim(P^{n-2}) = n - 1. \tag{3.9}$$

Next, by differentiation, the equation (3.6) in $[x_j, x_{j+1}]$ is equivalent to the square linear system

$$M_j X_j = b_j, \qquad 0 \le j \le m-1. \tag{3.10}$$

The square matrix $M_j \in \mathcal{M}_{n-1}(\mathbb{R})$ is

$$M_j = \begin{pmatrix} q(a_{1,j}) & q(a_{2,j}) & \dots & q(a_{n-1,j}) \\ c_{1,j}q'(a_{1,j}) & c_{2,j}q'(a_{2,j}) & \dots & c_{n-1,j}q'(a_{n-1,j}) \\ \dots & \dots & \dots & \dots \\ c_{1,j}^{n-2}q^{(n-2)}(a_{1,j}) & c_{2,j}^{n-2}q^{(n-2)}(a_{2,j}) & \dots & c_{n-1,j}^{n-2}q^{(n-2)}(a_{n-1,j}) \end{pmatrix} \tag{3.11}$$

where we note $c_{i,j} = b_{i,j} - a_{i,j}$. The unknown of the linear system is $X_j = \left(\gamma_{1,j}, \gamma_{2,j}, \dots, \gamma_{n-1,j}\right)^T \in \mathbb{R}^{n-1}$. The right hand side of the linear system is $b_j = \left(q(x_j), hq'(x_j), \dots, h^{n-1}q^{(n-1)}(x_j)\right)^T \in \mathbb{R}^{n-1}$ which is bounded $|b_j|_\infty \le C$ uniformly with respect to the subinterval index j.

One remarks that: a) the matrix M_j is close to a Vandermonde matrix, so natural invertibility conditions arise; b) provided the real numbers $a_{i,j}, b_{i,j} \in [0,1]$ are chosen independently of the subinterval (it will be written $a_{i,j} = a_i$ and $b_{i,j} = b_i$), then $M_j = M$ is independent of the index j. Two cases of invertibility and one case of non invertibility are considered below.

Lemma 3.1.5. *Take $q(x) = x^{n-2}$ with $n-2 \ge 0$. Assume the real numbers $a_{i,j} = a_i \in [0,1]$ and $b_{i,j} = b_i \in [0,1]$ are chosen independently of the subinterval (so $M_j = M$ is independent of the index j) and $b_i - a_i \ne 0$ for all i. Then M is non singular if and only if $a_i b_k - a_k b_i \ne 0$ for all $1 \le i \ne k \le n-2$.*

Proof. The matrix is $M = \left(\frac{n!}{(n-t)!}(b_i - a_i)^t a_i^{n-t}\right)_{1 \le t+1, i \le n-1}$. By assumption $b_i - a_i \ne 0$ for all i, so M is similar to $N = \left(\left(\frac{a_i}{(b_i-a_i)}\right)^{n-t}\right)_{1 \le t+1, i \le n-1}$. It is a Vandermonde matrix, invertible if and only if $\frac{a_i}{b_i - a_i} \ne \frac{a_k}{b_k - a_k}$ for $i \ne k$. The latter condition is equivalent to $a_i b_k - a_k b_i \ne 0$. □

Lemma 3.1.6. *Take $q \in P^{n-2}$ with deg$(q) = n-2 \ge 0$. Assume the real numbers $a_{i,j} = a_i \in [0,1]$ and $b_{i,j} = b_i \in [0,1]$ are chosen independently of the subinterval. Assume $a_i \ne a_k$ and $b_i - a_i = b_k - a_k \ne 0$ for all $1 \le i \ne k \le n-2$. Then the matrix M is non singular.*

Proof. The matrix M is similar to the matrix $N = \left(q^{(t)}(a_i)\right)_{1 \le t+1, i \le n-1}$ which is a reducible to a non singular Vandermonde matrix. □

Lemma 3.1.7. *Take $q(x) = u + x$, $e_1(x) = a_1 + (b_1 - a_1)x$ and $e_2(x) = a_2 + (b_2 - a_2)x$. Then the matrix M is singular if and only if $u(b_2 - a_2 - b_1 + a_1) + a_1 b_2 - a_2 b_1 = 0$.*

Proof. Indeed $q(e_1(x)) = u + a_1 + (b_1 - a_1)x$ and $q(e_2(x)) = u + a_2 + (b_2 - a_2)x$. The condition of linear independence of these two linear polynomials reduces to the claim. □

Proof of Theorem 3.1.2. If $n = 0$ or $n = 1$ the result is trivial, so we consider $n - 2 \geq 0$. If $f^{\text{obj}}(x) = x^n$ is a monomial function, one can takes the first set of basis functions given by Lemma 3.1.5 because the matrices are non singular. Unfortunately, this simple choice is not always possible as shown by Lemma 3.1.7. So to cover the case of general polynomials $f^{\text{obj}} \in P^n$, we continue with basis functions satisfying Lemma 3.1.6.

One notes $\mu = \frac{1}{2(n-1)}$. In a generic subinterval , we construct functions e_i for $1 \leq i \leq n - 1$ by taking $a_i = i\mu$ and $b_i = (i + 1)\mu$. By construction $b_i - a_i = b_k - a_k = \mu > 0$, $0 \leq a_i \leq 1$, $0 \leq b_i \leq \frac{n}{2(n-1)} \leq 1$ (because $n \geq 2$) and $a_i \neq a_k$ for $i \neq k$.

The matrix $M_j = M$ being non singular, then the system (3.10) has a solution $X_j = (\gamma_{1,j}, \ldots, \gamma_{n-1,j})$ such that $\|X_j\|_\infty \leq \|M^{-1}\|_\infty \|b_j\|_\infty \leq C$ uniformly with respect to the index of the subinterval j. So the representation (3.8) of f^{obj} holds for $x \in I$.

The constant is

$$K = \sum_{j=0}^{m-1} \sum_{i=1}^{n-1} |\beta_{i,j}| \leq \sum_{j=0}^{m-1} \sum_{i=1}^{n-1} \frac{h^2 |\gamma_{i,j}|}{(b_i - a_i)^2}$$

$$\leq \sum_{j=0}^{m-1} \sum_{i=1}^{n-1} \frac{h^2 C}{\mu^2} \leq m(n-1)\frac{h^2 C}{\mu^2} = \frac{(n-1)C}{\mu^2 m}.$$

Take $m^* > \frac{(n-1)C}{\mu^2}$. So the contraction property is satisfied for $m \geq m^*(H)$. □

Lemma 3.1.8. *Under the contraction condition (3.3), one has the bounds $\|f^{\text{obj}}\|_{L^\infty(I)} \leq \frac{1}{1-K}\|e_0\|_{L^\infty(I)}$ and $\|\sum_{i\geq 1} \beta_i f^{\text{obj}} \circ e_i\|_{L^\infty(I)} \leq \frac{K}{1-K}\|e_0\|_{L^\infty(I)}$.*

Proof. It is evident but we detail it because the key condition $e_i(I) \subset I$ is used. Consider the linear operator

$$\begin{array}{rccc} \mathcal{H}: & L^\infty(I) & \longrightarrow & L^\infty(I) \\ & g & \longmapsto & \sum_{i\geq 1} \beta_i g \circ e_i. \end{array} \tag{3.12}$$

Then $\|\mathcal{H}\|_{\mathcal{L}(L^\infty(I))} \leq K < 1$, that is $\mathcal{H}$ is a strictly contractive operator. The functional equation rewrites $f^{\text{obj}} = e_0 + \mathcal{H}(f^{\text{obj}})$ from which the inequalities are deduced. □

3.1.5 Fixed point algorithm

We detail a fixed point solution procedure which has a natural interpretation as Neural Networks. The standard fixed point method where the iteration index is $p = 0, 1, \dots$

$$\begin{cases} f_0 = 0, \\ f_{p+1} = e_0 + \sum\limits_{1\leq i\leq r} \beta_i f_p \circ e_i. \end{cases} \tag{3.13}$$

The first terms of the series are $f_1 = e_0$, $f_2 = e_0 + \sum_i \beta_i e_0 \circ e_i$, $f_3 = e_0 + \sum_i \beta_i e_0 \circ e_i + \sum_{i,j} \beta_i \beta_j e_0 \circ e_i \circ e_j$ and more generally

$$f_p = e_0 + \sum_{s=1}^{p-1} \left(\sum_{1\leq i_1,\dots,i_s\leq r} (\beta_{i_1} \dots \beta_{i_s})\, e_0 \circ e_{i_1} \circ \dots \circ e_{i_s} \right), \quad p \geq 1. \tag{3.14}$$

3.2 Application to Neural Networks

We detail three Neural Network implementations of the formula (3.14). The function T (1.24) is clearly well adapted to encode the functions in E_h, that is why we describe in details some implementations features with this function. These implementations can be performed without any difficulty with the ReLU function R.

It is convenient for the rest of the discussion to define a set $\mathcal{E}$ more general than E_h which is included in the new set $\mathcal{E}$

$$\begin{aligned} \mathcal{E} = \big\{ e \in C^0(I) : \quad & \text{there exists } 0 \leq \alpha < \beta \leq 1 \text{ and } 0 \leq a \leq b \leq 1 \text{ such that} \\ & e(x) = a \text{ for } 0 \leq x \leq \alpha, \\ & e(x) = a + (b-a)\tfrac{x-\alpha}{\beta-\alpha} \text{ for } \alpha \leq x \leq \beta, \\ & e(x) = b \text{ for } \beta \leq x \leq 1 \} . \end{aligned} \tag{3.15}$$

We will also use the standard notation $Lx = a + bx$ for affine functions where $a, b \in \mathbb{R}$ (same notations for $L_1 x = a_1 + b_1 x$, and so on, and so forth). A property already noticed in Lemma 1.2.4 is that the composition of two affine functions is also an affine function, that is with natural notations $L_1 \circ L_2 = L_3$.

3.2.1 A first Neural Network implementation

The first Neural Network implementation details a direct use of the representation formula (3.14).

Lemma 3.2.1. *All functions $e \in \mathcal{E}$ are computable with a composition of, first of all an affine function, next the activation function T and finally an affine function (that is $e = L_1 \circ T \circ L_2$).*

Proof. Indeed, with the notation of (3.15), $e(x) = a + (b - a)T\left((x - \alpha)/(\beta - \alpha)\right)$. □

Lemma 3.2.2. *The function f_p (3.14) can be implemented in a neural network with p hidden dense layers with the TReLU as activation function, and with variable widths along the layers $a_0 = 1$, $a_1 = m + 2$, $a_i = mr^{i-1} + 2$ for $1 \leq i \leq p$, and $a_{p+1} = 1$. The accuracy is*

$$\left\| f_p - f^{\text{obj}} \right\|_{L^\infty(I)} \leq K^p \left\| f^{\text{obj}} \right\|_{L^\infty(I)}, \qquad K = \sum_{1 \leq i \leq r} |\beta_i| < 1. \tag{3.16}$$

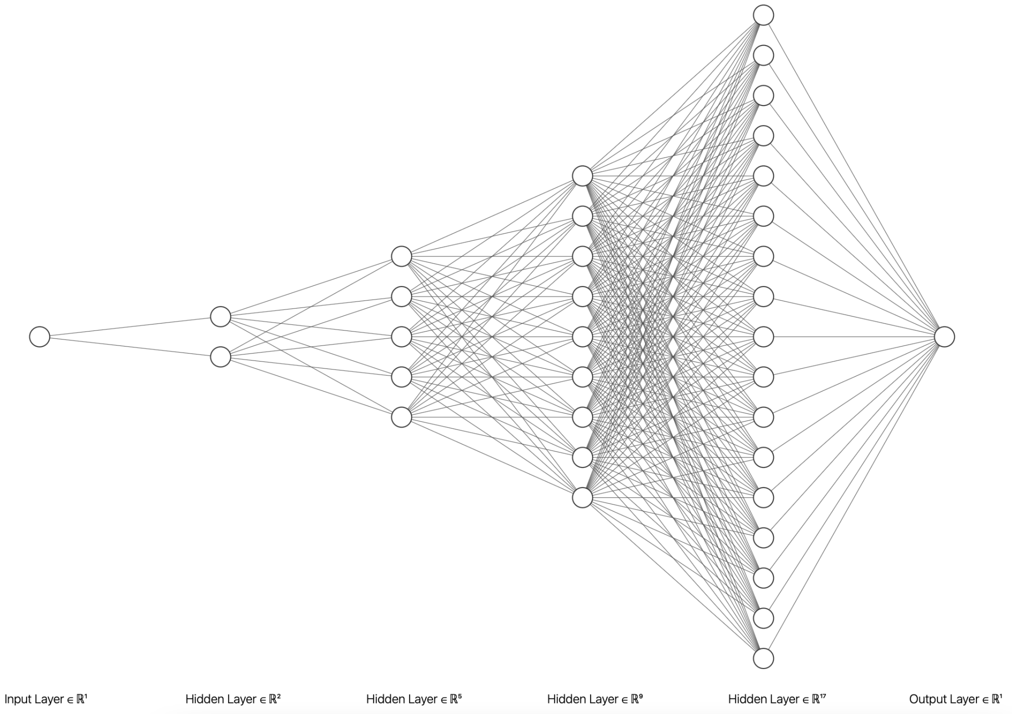

Fig. 3.1: Structure of a Neural Network described in Lemma 3.2.2. The number of neurons or computational units grows exponentially with respect to the layers, until the last one.