Neural Network World 10: 313–320, 2000

An integral formula for Heaviside neural networks

Paul C. Kainen
Dept. of Mathematics
Georgetown University
Washington, D.C. 20057

Věra Kůrková *
Institute of Computer Science
Acad. of Sci. of the Czech Republic
P.O. Box 5, 182 07 Prague 8, Czech Republic

Andrew Vogt
Dept. of Mathematics
Georgetown University Washington, D.C. 20057

Abstract

A connection is investigated between integral formulas and neural networks based on the Heaviside function. The integral formula developed by Kůrková, Kainen and Kreinovich is derived in a new way for odd dimensions and extended to even dimensions. In particular, it is shown that well-behaved functions of d variables can be represented by integral combinations of Heavisides with weights depending on higher derivatives.

Keywords: Feedforward neural network, one-hidden-layer network, perceptron, Heaviside function, plane wave, integral formula, numerical quadrature, Laplacian, Green's function.

1 Introduction

An integral formula of the form

$$\int_A w(\mathbf{a})\phi(\mathbf{a},\mathbf{x})d\mathbf{a}.$$

can be metaphorically seen as a one-hidden-layer neural network with a single linear output unit and a continuum of hidden units. Each hidden unit computes a value of the function ϕ depending on an input vector $\mathbf{x}$ and a parameter vector $\mathbf{a}$.

Integral formulas have been used to derive the universal approximation property of one-hidden-layer networks (see, e.g., Funahashi [6], Carroll and Dickinson [3], Ito [8]). Integral representations have also been used to estimate how accuracy of approximation varies with the number of hidden units (see, e.g., Barron [1], Girosi and Anzellotti [7] and Kůrková, Kainen and Kreinovich [12]).

[*] V. K. was supported by GA ČR grant number 201/99/0092

Arguments applying integral formulas make use of the fact that integrals with respect to the parameter vector can be approximated by Riemann sums. A neural network can even be thought of as a kind of numerical quadrature, a generalization of the midpoint, trapezoid and Simpson rules for approximating integrals. Implications of integral representations for the development of practical neural network algorithms motivated our investigation here.

We derive integral formulas corresponding to one-hidden-layer Heaviside perceptron networks, extending results of Kůrková, Kainen and Kreinovich [12].

2 Preliminaries

2.1 Feedforward neural networks

Feedforward neural networks compute parametrized sets of functions depending on the type of units as well as their interconnections. The *computational units* depend on two vector variables (*input* and *parameter*) and compute functions of the form $\phi : \mathcal{R}^p \times \mathcal{R}^d \to \mathcal{R}$, where ϕ corresponds to the type of the unit, p and d are the dimensions of the parameter and input space, resp., and $\mathcal{R}$ denotes the set of real numbers.

A one-hidden-layer network with hidden units using a function ϕ and a single linear output unit compute functions $g : \mathcal{R}^d \to \mathcal{R}$ of the form

$$g(\mathbf{x}) = \sum_{i=1}^{n} w_i \phi(\mathbf{a}_i, \mathbf{x}),$$

where all $w_i \in \mathcal{R}$, all $\mathbf{a}_i \in \mathcal{R}^p$, and n is the number of hidden units.

A *perceptron* with activation function $\psi : \mathcal{R} \to \mathcal{R}$ computes a function of the form $\phi((\mathbf{v}, b), \mathbf{x}) = \psi(\mathbf{v} \cdot \mathbf{x} + b) : \mathcal{R}^{d+1} \times \mathcal{R}^d \to \mathcal{R}$, where $\mathbf{v} \in \mathcal{R}^d$ is an *input weight* vector and $b \in \mathcal{R}$ is a *bias*; thus, the parameter vector is the pair $(\mathbf{v}, b) \in \mathcal{R}^{d+1}$.

We denote by ϑ the threshold *Heaviside function* $\vartheta(t) = 0$ for $t < 0$, $\vartheta(t) = 1$ for $t \geq 0$.

Let S^{d-1} denote the $(d-1)$-dimensional sphere in $\mathcal{R}^d$. Since for every $a > 0$, $\vartheta(at) = \vartheta(t)$, we can represent any function $\vartheta(\mathbf{v} \cdot \mathbf{x} + b)$ with $\mathbf{v} \neq 0$ as $\vartheta(\mathbf{e} \cdot \mathbf{x} + b')$, where $\mathbf{e} = \frac{\mathbf{v}}{\|\mathbf{v}\|} \in S^{d-1}$ is a unit vector and $b' = \frac{b}{\|\mathbf{v}\|}$.

For $\mathbf{e} \in S^{d-1}$ and $b \in \mathcal{R}$ we denote by $H_{\mathbf{e},b}$ the cozero hyperplane of the function $\mathbf{x} \mapsto \mathbf{e} \cdot \mathbf{x} + b$

$$H_{\mathbf{e},b} = \{\mathbf{x} \in \mathcal{R}^d : \mathbf{e} \cdot \mathbf{x} + b = 0\}.$$

The half-spaces bounded by this hyperplane, where ϑ equals 1 and 0, resp., are denoted

$$H^+_{\mathbf{e},b} = \{\mathbf{x} \in \mathcal{R}^d : \mathbf{e} \cdot \mathbf{x} + b \geq 0\}$$

and

$$H^-_{\mathbf{e},b} = \{\mathbf{x} \in \mathcal{R}^d : \mathbf{e} \cdot \mathbf{x} + b < 0\}.$$

A function $f : \mathcal{R}^d \to \mathcal{R}$ is called a *plane wave* if it can be represented as $f(\mathbf{x}) = \alpha(\mathbf{v} \cdot \mathbf{x})$, where $\alpha : \mathcal{R} \to \mathcal{R}$ is any function of one variable and $\mathbf{v} \in \mathcal{R}^d$ is any nonzero vector. Notice that plane waves are constant along hyperplanes parallel to $H_{\mathbf{v},0} = \{\mathbf{x} \in \mathcal{R}^d : \mathbf{v} \cdot \mathbf{x} = 0\}$. Perceptrons with activation function ψ compute plane waves of the form $\psi_b(\mathbf{v} \cdot \mathbf{x})$, where $\psi_b(t) = \psi(t + b)$.

2.2 Distributions, operators, and Green's functions

The theory of distributions frees calculus from the chore of checking differentiability by extending the set of functions to a larger set of *distributions* (or *generalized functions*), where all elements are differentiable and the formal rules of calculus hold.

For convenience we review some definitions here, see, e.g., Zemanian [16]. Let $\mathcal{D}(\mathcal{R}^d)$ denotes the set of *test functions*, i.e., compactly supported infinitely differentiable functions on $\mathcal{R}^d$. As usual, $\mathcal{D}'(\mathcal{R}^d)$ denotes the set of continous linear functionals on $\mathcal{D}(\mathcal{R}^d)$, where continuity is with respect to certain seminorms (see, e.g., Zemanian [16, p.7]). Members of $\mathcal{D}'(\mathcal{R}^d)$ are called *distributions*.

A locally integrable function g on $\mathcal{R}^d$ induces a distribution by the formula $< g, f >= \int_{\mathcal{R}^d} g(\mathbf{x})f(\mathbf{x})d\mathbf{x}$ where f is a test function. Distributions obtained in this way are called *regular distributions* and the same expression is sometimes used to denote both the regular distribution and the function that induces it.

An important distribution is the *(Dirac) delta function* δ defined for all $f \in \mathcal{D}(\mathcal{R}^d)$ by $< \delta, f >= f(0)$. Notice that δ plays the role of evaluation at 0. Recall that for all $g \in \mathcal{D}'(\mathcal{R}^d)$, $g * \delta = g$, where $*$ denotes convolution. Convolution of distributions is an extension of convolution of functions; the latter is defined by $(g*h)(\mathbf{x}) = \int_{\mathcal{R}^d} g(\mathbf{y})h(\mathbf{x}-\mathbf{y})d\mathbf{y}$, for the former see [16, p. 123].

The *Laplacian operator* $\triangle : \mathcal{D}'(\mathcal{R}^d) \to \mathcal{D}'(\mathcal{R}^d)$ is defined by:

$$\triangle(g) = \sum_{i=1}^{d} \frac{\delta^2 g}{\delta x_i^2}.$$

For a positive integer m, $\triangle^m$ denotes the Laplacian iterated m times, and $\triangle^0$ is defined as the identity operator.

A *Green's function* for an operator $T : \mathcal{D}'(\mathcal{R}^d) \to \mathcal{D}'(\mathcal{R}^d)$ is a distribution $g \in \mathcal{D}'(\mathcal{R}^d)$ such that $T(g) = \delta$. Green's functions typically are regular distributions and satisfy boundary conditions that do not concern us here. Standard Green's functions for the Laplacian operator $\triangle$ in $\mathcal{R}^d$ are: $G(\mathbf{x}) = \frac{1}{2\pi}\log\|\mathbf{x}\|$ when $d = 2$, and $G(\mathbf{x}) = \frac{1}{(2-d)\omega_d}\|\mathbf{x}\|^{2-d}$ when $d \neq 2$, where ω_d is the area of the sphere $S^{d-1} \subset \mathcal{R}^d$(see [4, p.64]).

3 Integral representation as a Heaviside perceptron network with a continuum of hidden units

3.1 The Representation theorem

Our basic assertion is that a smooth real-valued function on $\mathcal{R}^d$ with compact support can be written as an integral combination of characteristic functions of closed half-spaces.

Theorem 1 *Let d be a positive integer and let $f : \mathcal{R}^d \to \mathcal{R}$ he compactly supported and $d+2$-times continuously differentiable. Then*

$$f(\mathbf{x}) = \int_{S^{d-1}\times\mathcal{R}} w_f(\mathbf{e}, b)\vartheta(\mathbf{e}\cdot\mathbf{x} + b)d\mathbf{e}db,$$

where for d odd

$$w_f(\mathbf{e}, b) = a_d \int_{H^-_{\mathbf{e},b}} \triangle^{k_d} f(\mathbf{y}) d\mathbf{y},$$

$k_d = \frac{d+1}{2}$, *and* a_d *is a constant independent of* f*, while for* d *even,*

$$w_f(\mathbf{e}, b) = a_d \int_{H^-_{\mathbf{e},b}} \triangle^{k_d} f(\mathbf{y}) \alpha(\mathbf{e} \cdot \mathbf{y} + b) d\mathbf{y},$$

where $\alpha(t) = -t \log |t| + t$ *for* $t \neq 0$ *and* $\alpha(0) = 0$, $k_d = \frac{d+2}{2}$, *and* a_d *is a constant independent of* f*.*

A special case of Theorem 1 is the following consequence of the fundamental theorem of calculus.

Proposition 1 *Let* $g : \mathcal{R} \to \mathcal{R}$ *be continuously differentiable with* $\lim_{t \to \pm\infty} g(t) = 0$*. Then for every* $x \in \mathcal{R}$*,*

$$g(x) = \int_{-\infty}^{\infty} \frac{1}{2} g'(t) \vartheta(x - t) dt - \int_{-\infty}^{\infty} \frac{1}{2} g'(t) \vartheta(t - x) dt.$$

Sketch of proof of Theorem 1

Let $f : \mathcal{R}^d \to \mathcal{R}$ be given sufficiently smooth with compact support. Then

$$f = f * \delta = f * \triangle G = f * \triangle^{m_d} F = \triangle^{m_d} f * F.$$

Here G is the Green's function in $\mathcal{R}^d$ for the Laplacian $\triangle$. The Green's function with argument $\mathbf{x}$ in $\mathcal{R}^d$ depends only on $\|\mathbf{x}\|$ and has a singularity at the origin. However, it can be represented as a Laplacian, iterated m_d times, applied to a function F that is a scalar multiple of $\|\mathbf{x}\|$ (d odd) or $\log \|\mathbf{x}\|$ (d even). The quantity m_d equals $\frac{d+1}{2}$ for d odd, and $\frac{d}{2}$ for d even. Details may be found in Courant and Hilbert [4, pp. 677-681]. Because f is compactly supported (it suffices for f to die out rapidly at infinity), integration by parts allows one to shift $\triangle^{m_d}$ through the convolution so that it acts on f rather than F.

In the even and odd cases, F is an integral combinations of plane waves of the form $|\mathbf{x} \cdot \mathbf{e}|$ (d odd) or $\log |\mathbf{x} \cdot \mathbf{e}|$ (d even) where the variable of integration is $\mathbf{e}$ and integration is over the unit sphere S^{d-1}.

Indeed, for every positive integer d,

$$\|\mathbf{x}\| = s_d \int_{\mathbf{e} \in S^{d-1}} |\mathbf{x} \cdot \mathbf{e}| d\mathbf{e},$$

where s_d is a constant. Likewise, for every positive integer d,

$$\log \|\mathbf{x}\| = b_d + s_d \int_{S^{d-1}} \log |\mathbf{x} \cdot \mathbf{e}| d\mathbf{e} = b_d + s_d \triangle \left(\int_{S^{d-1}} \beta(\mathbf{e} \cdot \mathbf{x}) d\mathbf{e} \right),$$

where b_d, s_d are constants and $\beta(t) = \frac{t^2}{2} \log |t| - \frac{3t^2}{4}$ for $t \neq 0$ and $\beta(0) = 0$.
These identities, to be found in [4, pp. 678-9], can be proved by means of rotational invariance and homogenity arguments.

In the odd case, by a variant of Proposition 1, $|\mathbf{x} \cdot \mathbf{e}|$ can be represented as an integral combination of Heavisides $\vartheta(\mathbf{x} \cdot \mathbf{e} + b)$ where integration is with respect to b in $\mathcal{R}$. In the even case $\log(|\mathbf{x} \cdot \mathbf{e}|)$, although it has a singularity, is itself the Laplacian of a continuous function β of $\mathbf{x} \cdot \mathbf{e}$. This Laplacian commutes with the integration with respect to $\mathbf{e}$, and another integration by parts moves it to the "f"-side of the convolution. The continuous function β is again a simple integral combination of Heavisides $\vartheta(\mathbf{x} \cdot \mathbf{e} + b)$ with weight function $\alpha = -\beta'$ by a variant of Proposition 1. This yields the equations in Theorem 1, with $k_d = m_d$ in the odd case and $k_d = m_d + 1$ in the even case. □

The assumption that f is compactly supported can be replaced by the weaker assumption that f vanishes sufficiently rapidly at infinity. The integral representation also applies to certain nonsmooth functions that generate tempered distributions.

By an approach reminiscent of Radon transform but based directly on distributional techniques from Courant and Hilbert [4], it was shown in [12] that if f is compactly supported function on $\mathcal{R}^d$ with continuous d-th order partials, where d is *odd*, then f can be represented as

$$f(\mathbf{x}) = \int_{S^{d-1} \times \mathcal{R}} v_f(\mathbf{e}, b) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) d\mathbf{e} db,$$

where $v_f = a_d \int_{H_{\mathbf{e},b}} (D_{\mathbf{e}}^{(d)} f)(\mathbf{y}) d\mathbf{y}$, $a_d = (-1)^{k-1}(1/2)(2\pi)^{1-d}$ for $d = 2k+1$, $D_{\mathbf{e}}^{(d)} f$ is the directional derivative of f in the direction $\mathbf{e}$ iterated d times, $d\mathbf{e}$ is the $(d-1)$-dimensional volume element on S^{d-1}, and $d\mathbf{y}$ is likewise on a hyperplane.

Although the coefficients v_f are obtained by integration over hyperplanes, while the w_f arise from integration over half-spaces, these coefficients can be shown to coincide by application of the Divergence Theorem [2, p. 423] to the half-spaces $H^-_{\mathbf{e},b}$.

Theorem 1 extends the representation of [12] to *even* values for d and target functions f which are not compactly supported but which decrease sufficiently rapidly at infinity.

3.2 Integral operators with Heaviside kernel function

Theorem 1 can be formulated in terms of integral operators.
For $w \in \mathcal{L}_1(S^{d-1} \times \mathcal{R})$ define

$$T_H(w)(\mathbf{x}) = \int_{S^{d-1} \times \mathcal{R}^d} w(\mathbf{e}, b) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) d\mathbf{e} db.$$

For $f \in \mathcal{D}(\mathcal{R}^d)$ define

$$S_H(f)(\mathbf{e}, b) = w_f(\mathbf{e}, b).$$

Theorem 1 shows that for each $f \in \mathcal{D}(\mathcal{R}^d)$,

$$T_H(S_H(f)) = f.$$

3.3 Approximation by Heaviside perceptron networks

"A *quadrature formula* is a numerical rule whereby the value of a definite integral is approximated by the use of information about the integrand only at discrete points (where the integrand is defined)" (Engels, [5, p. 1]).

Thus any quadrature of the integral formula from Theorem 1 determines parameters of a Heaviside perceptron network and should be useful information for designing a learning algorithm.

In [10] and [11] we studied functions defined on $[0,1]^d$. If such functions are smooth, they can be extended to compactly supported smooth functions on $\mathcal{R}^d$, and the integral formulas of this paper can be applied to them.

Let H_d denote the set of Heaviside functions in $\mathcal{R}^d$ restricted to $[0,1]^d$, and let $span_n H_d$ denote the set of all linear combinations of at most n elements of H_d. We showed in [10] and [11] that for all positive integers d, n and any $p \in [1, \infty)$ there exists a best approximation mapping from $\mathcal{L}_p([0,1]^d)$ to $span_n H_d$, but for $p \in (1, \infty)$ such a mapping cannot be continuous. Thus there exists a quadrature with n terms that approximates on $[0,1]^d$ the integral formula derived in Theorem 1 within the smallest error achievable using n variable points. But such an optimal quadrature rule cannot vary continuously with the target function in $\mathcal{L}_p([0,1]^d)$.

4 Discussion

For smooth functions of d variables vanishing sufficiently rapidly at infinity, Theorem 1 gives an integral representation as a Heaviside perceptron network with a continuum of hidden units. The plane-wave integral representation demonstrates the remark of Courant and Hilbert [4, p. 676]: "But always the use of plane waves fails to exhibit clearly the domains of dependence and the role of characteristics. This shortcoming, however, is compensated by the elegance of explicit results."

Work on the theory of neural nets has begun to give specific computational guarantees for neural net approximation. Thus, finite neural networks may be measured against their progenitor, the integral formula of this paper. It remains for these issues to be further addressed in terms appropriate to engineering.

Representation of any suitable target function as an integral combination of Heaviside threshold functions, where the weights come from integration of directional derivatives over hyperplanes, or Laplacians over half-spaces, requires data that might be obtained from field measurements. An averaging process might estimate the weights and give efficient means for reconstructing a sufficiently smooth function from sparse but frequently resampled data. Modern computational power could make practical the discretization methods suggested by Sobolev [15].

Integral operators can be implemented physically - e.g., optically or in a semiconducting medium. Holographic [9], [14] and analogue VLSI [13] embodiments of neural network architectures might be ideal for this purpose.

The theoretical properties of integral formulas that correspond to continuum neural networks may thus be useful in guiding the evolution of new hardware and software approaches.

References

[1] Barron, A. R. (1993). Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, **39**, 930–945.

[2] Buck, R. C. (1965). *Advanced Calculus*, New York: McGraw-Hill.

[3] Carroll, S. M. & Dickinson, B. W. (1989). Construction of neural nets using the Radon transform. In *Proceedings of IJCNN* (pp. I. 607–611). New York: IEEE Press.

[4] Courant, R. & Hilbert, D. (1962). *Methods of Mathematical Physics*, vol. 2. New York: Wiley.

[5] Engels, H. (1980). *Numerical Quadrature and Cubiture.* London: Academic Press.

[6] Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, **2**, 183–192.

[7] Girosi, F., & Anzellotti, G. (1993). Rates of convergence for radial basis function and neural networks. In *Artificial Neural Networks for Speech and Vision* (pp. 97–113). London: Chapman & Hall.

[8] Ito, Y. (1991). Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. *Neural Networks*, **4**, 385–394.

[9] Jenkins, B. K. & Tanguay, A. R., Jr. (1992). Photonic implementations of neural networks, Chap. 9 in *Neural Networks for Signal Processing*, (Ed. B. Kosko) (pp. 287–382). Englewood Cliffs, NJ: Prentice-Hall.

[10] Kainen, P. C., Kůrková, V. & Vogt, A. (2000). Geometry and topology of continuous best and near best approximations. *Journal of Approximation Theory* (to appear).

[11] Kainen, P. C., Kůrková, V. & Vogt, A. (2000). Best approximation by Heaviside perceptron networks (submitted).

[12] Kůrková, V., Kainen, P. C. & Kreinovich, V. (1997). Estimates of the number of hidden units and variation with respect to half-spaces, *Neural Networks*, **10**, 1061–1068.

[13] Mead, C. A. (1989). *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, MA.

[14] Psaltis, D., Brady, D. & Gu, X. G. (1990). Holography in artificial neural networks, *Nature*, **343**, 325–330.

[15] Sobolev, S. L. (1964). *Lectures on the Theory of Cubature Formulae* I, II (Russian). Novosibirsk.

[16] Zemanian, A. H. (1987). *Distribution Theory and Transform Analysis.* New York: Dover.