Integral Transforms Induced by Heaviside Perceptrons

Věra Kůrková[1] and Paul C. Kainen[2]

[1] Czech Academy of Sciences, Institute of Computer Science,
Pod Vodárenskou věží 2, 18207 Prague, Czech Republic
vera@cs.cas.cz
[2] Department of Mathematics and Statistics, Georgetown University
Washington, DC, USA 20057
kainen@georgetown.edu

Abstract. We investigate an integral transform with kernel induced by perceptrons with the Heaviside activation function. Representation theorems are given expressing sufficiently smooth functions as "infinite Heaviside perceptron networks." The representation is exploited to obtain estimates of rates of approximation of these functions by networks with increasing numbers of units.

1 Introduction

Integral transforms play an important role in many branches of applied science such as medical imaging, astronomy, seismology, material science, turbulence, multiscale segmentation (see, e.g., [1],[2, pp. 567–569, pp. 591–593]). In addition to these traditional applications, the mathematical theory of neurocomputing utilizes them as a powerful tool to investigate function approximation by networks. An important class of integral operators has the form

$$T_K(w)(x) := \int_A w(a)K(x,a)da, \tag{1}$$

where K is a function of two variables, the *kernel*, and w is a *weight function*.

The term "kernel," derived from the German word "kern," was introduced by Hilbert in 1904 [3, p.291]. Many well-known kernels are named for the mathematicians who introduced them - e.g., Weierstrass, Abel, Laplace, Poisson, Szegő.

Functions computable by units used in neurocomputing also depend on two vector variables, an input vector and a parameter vector, and thus formally they can be considered as kernels. Note that for each appropriate choice of a kernel K, T_K is a linear operator on some normed linear space of functions. Artificial neural networks were introduced as multilayer computational models, but later one-hidden-layer architectures became dominant in applications of feedforward networks (see, e.g., [4, 5] and the references therein). Networks with one hidden layer of computational units, called *shallow*, compute finite linear combinations of functions from parameterized families called *dictionaries of computational units*. *Deep networks* with several hidden layers are mentioned in the last section.

A network with one hidden layer of computational units from the dictionary

$$G_K := \{K(.,a) \,|\, a \in A\}$$

and a single linear output computes input-output functions of the form

$$\sum_{i=1}^{n} w_i K(x, a_i), \tag{2}$$

where w_i are *output weights* and n is the *number of hidden units*.

One can view an integral

$$\int_A f(a)K(x,a)da$$

as an "infinite shallow neural network" with units from the dictionary G_K and output weights $f(a)$. Thus operators T_K map "infinite output-weight vectors" to input-output functions. On the other hand, quadratures of integral with kernels corresponding to computational units generate one-hidden-layer networks.

Originally, computational units, called *perceptrons*, were inspired by a simplified model of a neuron [6]. A perceptron applies an *activation function* (typically sigmoidal) to a weighted sum of its inputs to which is added a bias. So mathematically, it can be described as the composition of an activation function applied to an affine function. Geometrically, functions computable by perceptrons have the form of *plane waves* which are very useful in mathematical physics, as noted by Courant and Hilbert [7, p. 676]:

> ...representations as linear functionals of the data not only lead to many attractive formal relations, but, what is perhaps more important, they allow a study of specific properties. They are based on the decomposition of solutions, and, for that matter, other arbitrary functions, into *plane waves*. 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.

Later, alternative types of computational units were introduced due to their good mathematical properties. Some of these units compute spherical waves and can be highly localized. Nevertheless, perceptrons still remain widely used computational units because of their conceptual and practical advantages.

In this chapter, we explore the analogy between neural networks and integral transforms and show how this provides a conceptual tool for the analysis of shallow networks, which, moreover, can be applied, layer by layer, to deep networks with several layers of computational units. We describe an integral representation of smooth-enough functions in the form of infinite Heaviside perceptron networks that we derived jointly with Vladik Kreinovich [8].

Proof of the theorem was based on Vladik's original idea to employ the derivative of the Heaviside activation function, which is the Dirac delta function,

and to express the d-dimensional delta function with d odd as an integral of one-dimensional delta functions.

In the 20 years since our collaboration with Vladik on the topic of integral formulas, neural networks, and the Heaviside function, we have learned a few additional facts and extended the formula and method to cover even dimensions as well. Further, we substantially weakened some of the constraints. Together with A. Vogt in [9], we proved a version of the integral representation which includes all our previous versions as well as other related work by Ito [10] and Carroll and Dickenson [11]. We review these extensions and sketch their proof techniques.

Further, we review applications of integral representations in the form of infinite networks to estimates of complexity of networks needed for a given accuracy of approximation of functions represented by integral formulas. We describe the concept of variational norm tailored to computational units. Applying the representation in the form of Heaviside plane waves, we derive upper bounds on variation with respect to half-spaces, which plays a role of a critical factor in estimates of network complexity.

The chapter is organized as follows. Section 2 contains an exposition of basics and notation, including distribution theory. Section 3 begins with a brief summary of the proof outline and describes an integral representation for sufficiently smooth functions in the form of Heaviside plane waves. It sketches an argument based on the integral representation of the d-dimensional Dirac delta function. In Section 4, extension to wider classes of functions as well as even dimensions are given. Section 5 is devoted to applications of integral representations to network complexity and Section 6 contains some concluding remarks.

2 Preliminaries

Computational units (such as perceptrons, radial or kernel units) compute functions of two vector variables representing *inputs* and *parameters* (e.g., weights, biases, centroids). So formally computational units can be described as mappings

$$K : X \times A \to \mathbb{R},$$

where $X \subseteq \mathbb{R}^d$ is a set of variables and $A \subseteq \mathbb{R}^s$ is a set of (inner) parameters. Let

$$G_K = G_K(A) = G_K(X, A) := \{K(., a) \,|\, a \in A\}$$

denote the *parameterized set of functions on X induced by K*. We use the shorter notation G_K or $G_K(A)$ when the sets X or A are clear from the context. The set $G_K(X, A)$ is called a *dictionary* of computational units.

If $b \in \mathbb{R}$ and $v \in \mathbb{R}^d$ and $\sigma : \mathbb{R} \to \mathbb{R}$ is any function, then the *perceptron with activation function* σ is the function $K_\sigma : \mathbb{R}^d \times \mathbb{R}^{d+1} \to \mathbb{R}$ defined for $(x, (v, b)) \in \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R}) = \mathbb{R}^d \times \mathbb{R}^{d+1}$ by

$$K_\sigma(x, (v, b)) := \sigma(v \cdot x + b). \tag{3}$$

Typically, activation functions are assumed to be *sigmoidals* - that is, to be monotonic with limits 0 and 1, resp., as the input goes to $-\infty$ or $+\infty$. However, the universal approximation property holds for shallow networks with perceptrons with any sufficiently smooth nonpolynomial activation function [12].

An important type of activation function is the indicator function for the nonnegative reals, called the *Heaviside function* $\vartheta : \mathbb{R} \to \mathbb{R}$ defined as $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. (This function is named for Oliver Heaviside (1850-1925), who used it to construct a quite sophisticated, though heuristic, theory of analysis which has turned out to be accurate. Heaviside's scientific contributions included an explanation for anomalies in radio transmission; he hypothesized an ionized layer in the Earth's atmosphere which is now known to exist.)

A function $f : \mathbb{R}^d \to \mathbb{R}$ is called a *plane wave* if it can be represented as $f(x) = \alpha(v \cdot x)$, where $\alpha : \mathbb{R} \to \mathbb{R}$ is any function of one variable and $v \in \mathbb{R}^d$ is any nonzero vector. Plane waves are constant along hyperplanes

$$H_{v,b} := \{x \in \mathbb{R}^d \,|\, v \cdot x = -b\}.$$

Perceptrons with an activation function σ compute plane waves of the form $\sigma_b(v \cdot x)$, where $\sigma_b(t) = \sigma(t + b)$. If $\sigma = \vartheta$, then $K_\vartheta(\cdot, (v, b))$ is the indicator function of the half-space $\{x \in \mathbb{R}^n \,|\, v \cdot x + b \geq 0\}$. Let S^{d-1} denote the unit sphere in $\mathbb{R}^d$. We denote

$$G_\vartheta = G_\vartheta(S^{d-1} \times \mathbb{R}, X) := \{\vartheta(e \cdot - + b) : X \to \mathbb{R} \,|\, e \in S^{d-1}, b \in \mathbb{R}\},$$

the *dictionary of perceptrons with the Heaviside activation function.*

A *shallow network with a single linear output and with n computational units from a dictionary* $G_K(A)$ computes input-output functions from the set

$$\mathrm{span}_n G_K(A) := \left\{ \sum_{i=1}^{n} w_i K(\cdot, a_i) \,|\, w_i \in \mathbb{R}, a_i \in A \right\}.$$

A network unit computing a function $K : X \times A \to \mathbb{R}$ induces an integral operator. The operator depends on a measure μ on A. For a function $w : A \to \mathbb{R}$ in a suitable space of functions on A such that for all $x \in X$ the integral (4) exists, we denote by $T_{K,\mu}$ the operator defined as

$$T_{K,\mu}(w)(x) := \int_A w(a)\, K(x, a) d\mu(a). \tag{4}$$

When μ is the Lebesgue measure, we drop μ from the notation. Metaphorically, the integral on the right-hand side of the equation (4) can be interpreted as a *one-hidden-layer neural network with infinitely many units* computing functions from a dictionary

$$G_K := \{K(., a) \,|\, a \in A\}.$$

So the operator $T_{K,\mu}$ transforms output-weight functions $w : A \to \mathbb{R}$ of infinite networks with units from the dictionary G_K to input-output functions

$$T_{K,\mu}(w) : X \to \mathbb{R}.$$

Recall (see e.g., [13]) that for a unit vector $e \in S^{d-1}$ and a real-valued function f on $\mathbb{R}^d$, the *directional derivative* of f in the direction e is defined by

$$(D_e f)(y) := \lim_{t \to 0} \frac{f(y+te) - f(y)}{t}$$

and the *k-th directional derivative* is inductively defined by

$$(D_e^{(k)} f)(y) = D_e(D_e^{(k-1)} f)(y).$$

It is well-known (see e.g., [13, p.222] that

$$(D_e f)(y) = e \cdot \nabla f(y),$$

where $\nabla = (\partial_1, \ldots, \partial_d)$ is the vector of partial derivatives w.r.t. the variables. The k-th order directional derivative is a weighted sum of the corresponding k-th order partial derivatives, where the weights are polynomials in the coordinates of e multiplied by multinomials (see e.g., [14, p.130]). Hence existence and continuity of the partials ∂_i implies the same for directional derivatives.

By $\mathcal{C}^d(\mathbb{R}^d)$ we denote the *space of continuous functions on $\mathbb{R}^d$ with continuous derivatives up to order d*, while $\mathcal{C}^\infty(\mathbb{R}^d)$ denotes the space of continuous functions on $\mathbb{R}^d$ with continuous derivatives of *all* orders. The *Schwartz class* $\mathcal{S}(\mathbb{R}^d)$ consists of all functions from $\mathcal{C}^\infty(\mathbb{R}^d)$ which, together with all their derivatives, are rapidly decreasing [15, p.251]).

Let $\mathcal{D} := \mathcal{D}(\mathbb{R}^k)$ denote the linear space of *test functions* which is the intersection of $\mathcal{C}^\infty(\mathbb{R}^k)$ and the linear space of compactly supported functions on $\mathbb{R}^k$. The space $\mathcal{D}$ is nonempty; see, e.g., [16], for the definition of the topology on $\mathcal{D}$.

A *distribution* is a continuous linear functional on the space of test functions. Let $\mathcal{D}' := \mathcal{D}'(\mathbb{R}^d)$ denote the space of all distributions. The *Dirac delta function* δ_k is the distribution on $\mathbb{R}^k$ given by evaluation at zero

$$\delta_k(\phi) := \phi(0).$$

When $k = 1$, we merely write δ.

A function f on $\mathbb{R}^k$ is called *locally integrable* if the integral $\int_C f(x)dx$ exists for any compact $C \subset \mathbb{R}^k$. Every locally integrable function f then defines a distribution T_f whose value on the test function ϕ is

$$\langle T_f, \phi \rangle := \int_{\mathbb{R}^d} f(x)\phi(x)dx.$$

The *convolution* $f * g$ of a compactly supported f and a distribution g on $\mathbb{R}^n$, is defined by

$$(f * g)(x) := \int_{\mathbb{R}^n} f(y)g(x-y)dy.$$

The *distributional derivative* T' of a distribution T is defined by the equation

$$\langle T', \phi \rangle := -\langle T, \phi' \rangle. \tag{5}$$

As $\langle \vartheta', \phi \rangle = -\langle \vartheta, \phi' \rangle = -\int_{-\infty}^{\infty} \vartheta(x)\phi'(x)dx = -\phi(\infty) + \phi(0) = \langle \delta, \phi \rangle$, $\vartheta' = \delta$ (see, e.g., [16, p.47]. Thus,

δ is the distributional derivative of ϑ.

3 Infinite Heaviside Perceptron Networks

In this section, we give a representation of compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$, with d odd, as infinite Heaviside perceptron networks, which we found with V. Kreinovich [8] and published in 1997. Quoting from the abstract:

> We estimate variation with respect to half-spaces in terms of "flows through hyperplanes". Our estimate is derived from an integral representation for smooth compactly supported multivariable functions proved using properties of the Heaviside and delta distributions. Consequently we obtain conditions which guarantee approximation error rate of order $O(n^{1/2})$ by one-hidden-layer networks with n sigmoidal perceptrons.

While our understanding has improved, with 20 years of additional work, we may use the abstract as an outline. Our goal was to find an upper bound on the rate of neural-network approximation.

The Maurey-Jones-Barron Theorem (see Section 5, just before Theorem 3) translates a geometric parameter called "variation with respect to half-spaces" (Section 5), for a suitable target function f, into an upper bound on the least number of Heaviside units used in a one-layer approximation of f (its "rate of approximation"). Variation of f can in turn be estimated using an integral formula expressing f as an integral combination of Heaviside functions. The weighting function for the integral formula (4) corresponds to the "outer" (i.e., linear) output weights in the neural network, while the "inner" variables determine the parameters of the Heaviside units. The weight functions turn out to be the numeric integrals of iterated directional derivatives across the hyperplanes defining the Heavisides.

We derive our representation by exploiting the distributional derivative of the Heaviside function, which is the Dirac delta function, expressing a test function of d variables as its convolution with the d-dimensional delta function, which can be written as an integral of derivatives of 1-dimensional delta functions.

For a positive integer k, δ_k is the identity w.r.t. convolution; that is, every $f \in \mathcal{D}(\mathbb{R}^k)$ satisfies the following equation (e.g., [16])

$$f(x) = (f * \delta_k)(x) := \int_{\mathbb{R}^k} f(z)\delta_k(x - z)dz. \tag{6}$$

For d odd, the delta distribution δ_d can be expressed as an integral over the unit sphere of the $d-1$-st distributional derivatives ${\delta_1}^{(d-1)}$ of δ_1 in the form

$$\delta_d(x) = a_d \int_{S^{d-1}} \delta_1^{(d-1)}(e \cdot x)\, de, \tag{7}$$

where

$$a_d := (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d} \tag{8}$$

see, e.g., [7, p. 680]. For $e \in S^{d-1}$ and $b \in \mathbb{R}$, we denote hyperplanes and half-spaces by

$$H_{e,b} := \{y \in \mathbb{R}^d \,|\, e \cdot y + b = 0\}, \text{ and } H^-_{e,b} := \{y \in \mathbb{R}^d \,|\, e \cdot y + b \leq 0\}, \tag{9}$$

resp. The following theorem from [17] describes an integral representation of a smooth compactly supported function as an uncountably infinite neural network with Heaviside perceptrons.

Theorem 1. *Let d be an odd integer and $f \in \mathcal{C}^d(\mathbb{R}^d)$ be compactly supported. Then for all $x \in \mathbb{R}^d$*

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

where $w_f(e,b) = a_d \int_{H_{e,b}} (D_e^{(d)} f)(y)\, dy$ and a_d is as in (8).

Proof. The proof is based on the relationship between the Heaviside threshold function ϑ and the Dirac delta distribution δ_1. We prove the statement for a test function f. Extension to all compactly supported functions with continuous partial derivatives of order d follows from a basic result of distribution theory: each continuous compactly supported function can be uniformly approximated on $\mathbb{R}^d$ by a sequence of test functions (see e.g., [16, p. 3]).

First, we replace the d-dimensional delta distribution with its integral representation in terms of one-dimensional delta distributions as in (7),

$$\delta_d(x - z) = a_d \int_{S^{d-1}} {\delta_1}^{(d-1)}(e \cdot x - e \cdot z) de.$$

One then obtains from (6) and an application of Fubini's theorem

$$f(x) = a_d \int_{S^{d-1}} \int_{\mathbb{R}^d} f(z) {\delta_1}^{(d-1)}(x \cdot e - z \cdot e) dz de.$$

Rearranging the inner integration, we get for the Lebesgue measure d_H on $H_{e,b}$

$$f(x) = a_d \int_{S^{d-1}} \int_{\mathbb{R}} \int_{H_{e,b}} f(y) {\delta_1}^{(d-1)}(x \cdot e + b) d_H y\, db\, de.$$

Setting $u(e,b) = a_d \int_{H_{e,b}} f(y) d_H y$, we obtain

$$f(x) = \int_{S^{d-1}} \int_{\mathbb{R}} u(e,b) {\delta_1}^{(d-1)}(x \cdot e + b) db\, de. \tag{10}$$

By definition of the distributional derivative, for every $e \in S^{d-1}$ and $x \in \mathbb{R}^d$,

$$\int_{\mathbb{R}} u(e,b) {\delta_1}^{(d-1)}(e \cdot x + b) db = (-1)^{d-1} \int_{\mathbb{R}} \frac{\partial^{d-1} u(e,b)}{\partial b^{d-1}} \delta_1(e \cdot x + b) db.$$

Using integration by parts on the right-hand integral, as d is odd and the distributional derivative of ϑ is δ_1, it follows that for every $e \in S^{d-1}$ and $x \in \mathbb{R}^d$

$$\int_{\mathbb{R}} u(e,b) {\delta_1}^{(d-1)}(e \cdot \mathbf{x} + b) db = - \int_{\mathbb{R}} \frac{\partial^d u(e,b)}{\partial b^d} \vartheta(e \cdot x + b) db.$$

Differentiating w.r.t. b is orthogonal to hyperplane $H_{e,b}$ and so it is in the direction e. Hence,

$$\frac{\partial^d u(e,b)}{\partial b^d} = a_d \frac{\partial^d}{\partial b^d} \int_{H_{e,b}} f(y)dy = a_d \int_{H_{e,b}} D_e^{(d)} f(y)dy.$$

From (10) we obtain the integral representation of f in the form

$$f(x) = a_d \int_{S^{d-1}\times\mathbb{R}} \Big(\int_{H_{e,b}} \Big(D_e^{(d)} f \Big)(y)dy \Big)\, \vartheta(e\cdot x + b)\; db\, de.$$

□

4 Generalizing the Integral Formula

In this section, we explain how one can weaken the conditions for the integral formula to hold and include all dimensions, odd and even.

This entails some additional concepts regarding distributions and analysis. As test functions on $\mathbb{R}^n$ are infinitely differentiable in each of n coordinates, we use operator notation

$$\partial_r^i := \Big(\frac{\partial}{\partial x_r} \Big)^i.$$

For *multi-index* $\alpha \in (\mathbb{N}_0)^n, \alpha = (\alpha_1, \ldots, \alpha_n)$, the differential operator

$$\partial^\alpha := \partial_1^{\alpha_1} \ldots \partial_n^{\alpha_n}$$

indicates differentiating $\alpha_i \geq 0$ times w.r.t. x_i, for $i = 1, \ldots, n$.

The definition of derivative of a distribution is the same adjoint relationship described in (5). So if T is a distribution in $\mathcal{D}'(\mathbb{R}^n)$ and ϕ is a test function, then

$$\langle \partial^\alpha(T), \phi \rangle := (-1)^{|\alpha|} \langle T, \partial^\alpha \phi \rangle.$$

where $|\alpha| := \alpha_1 + \cdots + \alpha_n$, which is the total number of differentiations.

A linear differential operator L is a linear combination of the form

$$a\partial^\alpha + b\partial^\beta + c\partial^\gamma + \cdots .$$

A particularly useful example, the *Laplacian* operator, is given by

$$\Delta := \partial_1^2 + \cdots + \partial_n^2.$$

It turns out that a key step in our generalization involves finding integral formulas for (iterated) Laplacian operators.

We need the notion of a Green's function. A *Green's function associated with a linear operator* L is a function $\mathbf{G}$ such that $L(\mathbf{G}) = \delta$. For example, in dimension 1, differentiation is a linear operator; the Heaviside function is a Green's function for differentiation.

If T is a compactly supported distribution, having a Green's function $\mathbf{G}$ for L, one can find a distribution S satisfying the equation

$$L(S) = T.$$

Indeed, by letting S be the convolution of T and $\mathbf{G}$, $S := T * G$, and using the fact that differentiation can be applied to either factor of a convolution, we have

$$L(S) = \langle L, T * \mathbf{G} \rangle = T * L\mathbf{G} = T * \delta = T.$$

To define the large class of functions for which our most general integral formula holds, we need one more technical notion. A real-valued function f on $\mathbb{R}^d$ *vanishes to order* $r \in \mathbb{R}$ (at ∞), $f(x) = \mathrm{o}\left(\|x\|^{-r}\right)$, if

$$\lim_{\|x\| \to \infty} f(x)\|x\|^r = 0.$$

The *order* of g, ord f, is the supremum of the set of all $r \in \mathbb{R}$ such that f vanishes to order r.

Put $k_d := 2\lceil \frac{d+1}{2} \rceil$, so $k_d = d + 1$ for d odd and $k_d = d + 2$ for d even. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *of controlled decay* if both of the following hold:
(i) f is k_d-times continuously differentiable, and
(ii) $\forall$ multi-index α with $|\alpha| \leq k_d$, ord $\left(\partial^\alpha f\right) > |\alpha|$.

The functions of controlled decay include almost all suitably differentiable, "rapidly vanishing" functions and, in particular, those of compact support. Let

$$\alpha(u) := -u \log(|u|) + u$$

for $u \neq 0$, with $\alpha(0) = 0$. For f of controlled decay and d a positive integer, let

$$w_f(e, b) := a_d \int_{\mathbb{R}^d} \Big(\vartheta(-e \cdot y - b)\Big)^{r_d} \Big(\alpha(e \cdot y + b)\Big)^{1 - r_d} \Big(\Delta^{\frac{(k_d)}{2}} f\Big)(y) dy, \qquad (11)$$

where $e \in S^{d-1}$, $b \in \mathbb{R}^d$, and the various functions of d are defined below. We can now express every function of controlled decay by an integral formula.

Theorem 2. *Let d be a positive integer and let f be a function of controlled decay on $\mathbb{R}^d$. Then for the measure $d(e, b)$ induced by Lebesgue measure on $\mathbb{R}^{d+1}$*

$$f(x) = \int_{S^{d-1} \times \mathbb{R}} w_f(e, b) \vartheta(e \cdot x + b) d(e, b). \qquad (12)$$

To define the (0/1) exponent r_d and the real number a_d, which appear in (11), we introduce several functions which depend on d:

$$r := r_d := d - 2\lfloor (d/2) \rfloor = \begin{cases} 1, & \text{if d is odd,} \\ 0 & \text{if d is even;} \end{cases}$$

$$s := s_d := 2\lceil (d/2) \rceil - 2 = \begin{cases} (d-1)/2, & \text{if d is odd,} \\ (d-2)/2 & \text{if d is even;} \end{cases}$$

$$t := t_d := 2 - k_d = \begin{cases} 1-d, & \text{if d is odd,} \\ -d & \text{if d is even.} \end{cases}$$

Then for all positive integers d

$$a_d := (1/2)^r (-1)^s (2\pi)^t = \begin{cases} (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}, & \text{if d is odd,} \\ (-1)^{(d-2)/2}(2\pi)^{-d} & \text{if d is even;} \end{cases} \quad (13)$$

The ϑ term is present in w_f iff d is odd, while the α term is present iff d is even. Hence, for d odd, in w_f one integrates an iterated Laplacian of f over the negative half-space $H^-_{e,b}$ defined in (9) while for d even, one integrates an iterated Laplacian of f, multiplied by the factor $\alpha(e \cdot y + b)$, over all y in $\mathbb{R}^d$. See [9] where it is shown that Theorem 2 implies previous results some of which hold under slightly different conditions. s For d odd, $f : \mathbb{R}^d \to \mathbb{R}$ is *of weakly controlled decay* [9] if
(i) f is d-times continuously differentiable,
(ii) for all α with $|\alpha| < d$, $\text{ord}\,(\partial^\alpha f) \geq 0$, and
(iii) for all α with $|\alpha| = d$, $\text{ord}\,(\partial^\alpha f) > d + 1$.

Note that the weakly controlled decay is different notion of a "nice" functions than controlled decay. The first two conditions (i) and (ii) are weaker but the third condition is stronger than for controlled decay. However, controlled decay is defined for even d as well.

In the following, we briefly outline the proof, from [9], of the general version of the integral representation in terms of Heaviside perceptron networks.

We first show that both $\|x\|$ and $\log(\|x\|)$ are integrals of plane waves. If de denotes the measure on S^{d-1} induced by Lebesgue measure on $\mathbb{R}^d$ and ω_d is the measure of the sphere S^{d-1}, then one has the following key lemmas:

$$\|x\| = s_d \int_{S^{d-1}} |e \cdot x| de; \quad \text{where } s_d := 2\omega_{d-1}/(d-1), \ d \geq 3, \ x \in \mathbb{R}^n \quad (14)$$

$$\log(\|x\|) = b_d + (1/\omega_d) \int_{S^{d-1}} \log|e \cdot x| de; \qquad d \geq 1, \ x \in \mathbb{R}^n, \ x \neq 0, \quad (15)$$

where b_d is a constant. There is an explicit role for the Laplacian:

$$\log(\|x\|) = b_d + (1/\omega_d) \Delta\Big(\int_{S^{d-1}} \beta(e \cdot x) de \Big); \qquad d \geq 1, \ x \in \mathbb{R}^n, \ x \neq 0, \quad (16)$$

where $\beta(u) := (1/2)u^2 \log|u| - (3/4)u^2$ for $u \neq 0$, $\beta(0) := 0$. Then $\beta'(u) = -\alpha(u)$ for all u and $\beta''(u) = \log|u|$ for $u \neq 0$. The argument for (16) uses calculus.

The theorem is then proved by writing a function of controlled decay as the convolution of its iterated Laplacian with a Green's function, which is in turn represented as an integral combination of plane waves, which are expressed as integral combinations of characteristic functions of half-spaces.

Using Lebesgue dominated convergence, w_f is shown to be both well-defined and continuous. We then find Green's functions for the iterated Laplacians in both the odd and even cases, and the integrability of w_f is also proved for both cases. Finally, we show that the integral formula (12) does hold.

An integral formula involves real-valued functions on a measure space. In [18] this was generalized to functions with values in a Banach space. In this setting, Bochner integrals replace Lebesgue integration. See, e.g., [19] or [20]. We proved in [18] that the Bochner integral $\int w\Phi$ is convergent if w is in $\mathcal{L}_1$ and Φ is essentially bounded. Bochner integrals may allow approximation of nonlinear operators as in [21–23].

5 Network Complexity

In this section, we derive the consequences of Theorem 1 for the number of computational units needed to approximate with a given accuracy smooth functions.

The same integral representation as the one presented in Theorem 1 was derived by Ito [10]. He used a different proof technique based on the inverse Radon transform. Discretizing the integral representation, he proved that smooth functions can be approximated with an arbitrary accuracy by Riemann sums in the form of finite linear combinations of perceptrons. Thus he proved that shallow perceptron networks have the *universal approximation property*. As with all universality type results, this approximation capability of shallow perceptron networks is obtained assuming that the number of units in the approximating network is potentially infinite.

In practical applications, various constraints on numbers and sizes of ntwork parameters limit feasibility of implementations. Thus it is important to describe classes of functions which can be computed or sufficiently well approximated by networks with reasonably bounded numbers of units.

Let

$$f = \sum_{i=1}^{m} w_i g_i \tag{17}$$

be a representation of a function f as an input-output function of a shallow network with units from a dictionary G. The *"l_0-pseudonorm"* of a vector $w \in \mathbb{R}^m$, denoted $\|w\|_0$, is the number of nonzero entries in the vector (see, e.g., [24–26]). So if a neural network with m hidden units calculates f as in (17), then $\|w\|_0$ is the number of computational units with a nonzero output weight. Thus, one can measure the sparsity of a neural network by the "l_0-pseudonorm" of its output weight vector.

However, "l_0-pseudonorm" is neither a norm nor even a pseudonorm. The quantity $\|w\|_0$ is always an integer and thus $\|\cdot\|_0$ does not satisfy the homogeneity

property of a norm ($\|\lambda x\| = |\lambda|\|x\|$ for all λ). Moreover, the "unit ball" $\{w \in \mathbb{R}^n \,|\, \|w\|_0 \leq 1\}$ is nonconvex and unbounded as it is equal to the union of all one-dimensional subspaces of $\mathbb{R}^m$. For any $r > 0$, the ball of radius r is equal to $\text{span}_k \mathbb{R}^m$, where $k = \lfloor r \rfloor$. Minimization of "l_0-pseudonorm" of the vector of output weights is a difficult nonconvex optimization task which, for some dictionaries, is NP-hard [27].

In neurocomputing, instead of "l_0-pseudonorm", l_1 and l_2-norms of output weight vectors $w = (w_1, \ldots, w_m)$ have been minimized in weight-decay regularization techniques [4]. In particular, l_1-norm plays an important role, as solutions with small l_1-norms can be well approximated by networks with small "l_0-pseudonorm"s; see, e.g., [25].

The l_1-norms of output-weight vectors of all networks with units from a dictionary G are minimized by a norm tailored to G. This norm, called *G-variation*, is defined for bounded subsets G of normed linear spaces $(\mathcal{X}, \|.\|)$ as

$$\|f\|_G := \inf \left\{ c \in \mathbb{R}_+ \;\middle|\; \frac{f}{c} \in \text{cl}_{\mathcal{X}} \,\text{conv}\,(G \cup -G) \right\}. \tag{18}$$

In (18) "$\text{cl}_{\mathcal{X}}$" denotes closure with respect to the topology induced by the norm $\|\cdot\|_{\mathcal{X}}$, "conv" is the convex hull, and "$-G$" means $\{-g \mid g \in G\}$. It was shown in [28] that in the definition of G-variation, inf can be replaced with min.

A special case of variational norm is variation with respect to Heaviside perceptrons, also called *variation with respect to half-spaces* as Heaviside perceptrons are the indicator functions for (closed affine) half-spaces. It was introduced by Barron [29] and extended to general dictionaries by Kůrková [30].

A use for G-variation is to estimate the rate of approximation by a shallow network. The next upper bound is a reformulation of a theorem by Maurey [31], Jones [32], Barron [33] in terms of G-variation (see [30, 34]).

Theorem 3. *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Hilbert space, G its bounded nonempty subset, $s_G = \sup_{g \in G} \|g\|_{\mathcal{X}}$, $f \in \mathcal{X}$, and n be a positive integer. Then*

$$\|f - \text{span}_n G\|_{\mathcal{X}}^2 \leq \frac{s_G^2 \|f\|_G^2 - \|f\|_{\mathcal{X}}^2}{n};$$

It was shown in [35] that for every n, the set $\text{span}_n G_\vartheta([0,1]^d)$ of input-output functions of a shallow network with n Heaviside perceptrons is "approximatively compact" (see below for a definition) and hence best approximations (i.e., as close as possible) always exist in $\text{span}_n G_\vartheta([0,1]^d)$ to any suitably nice function f. In particular, by Theorem 3, for every function $f \in \mathcal{L}^2([0,1]^d)$ there exists a function f_n computable by a shallow network with n Heaviside perceptrons with

$$\|f - f_n\|_{\mathcal{L}^2([0,1])} = \|f - \text{span}_n G_\vartheta([0,1]^d)\|_{\mathcal{L}^2([0,1])} \leq \frac{\|f\|_{G_\vartheta([0,1]^d)}}{\sqrt{n}}. \tag{19}$$

So accuracy of approximation of functions from $\mathcal{L}^2([0,1]^d)$ by networks with n Heaviside perceptrons depends on their variations with respect to half-spaces.

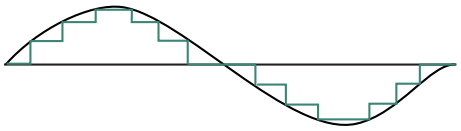

Fig. 1. Variation with respect to half-spaces and total variation

It follows from the definition that, for $d = 1$, variation with respect to half-spaces is, up to a constant, equal to the concept of total variation [14, 36] (see Fig. 1).

To estimate variation with respect to half-spaces, we employ the integral representation of smooth functions as infinite Heaviside perceptron networks. It is easy to see [28, p.164] that for each $f \in \operatorname{span} G$

$$\|f\|_G \leq \min \left\{ \|w\|_1 \,\middle|\, f = \sum_{i=1}^{m} w_i g_i \right\}. \tag{20}$$

So G-variation equals the minimum of the l_1-norms of the output-weight vectors w over all shallow networks (with units from G) which compute f.

A similar upper bound on G_K-variation holds for functions which can be expressed as

$$f(x) = T_{K,\mu}(w) = \int_A w(a)K(x,a)d\mu(a).$$

Under mild conditions on K [23, 28], the following upper bound holds

$$\|f\|_{G_{K,\mu}(A)} \leq \|w\|_{\mathcal{L}^1(A,\mu)} \tag{21}$$

Note that for every continuous sigmoid σ (i.e., a non decreasing $\sigma : \mathbb{R} \to \mathbb{R}$ with $\lim_{t\to-\infty} \sigma(t) = 0$ and $\lim_{t\to\infty} \sigma(t) = 1$)

$$\|.\|_{G_\vartheta(\Omega)} = \|.\|_{G_\sigma(\Omega)},$$

in $\mathcal{L}^p(\Omega)$ with $p \in (1, \infty)$ and Ω compact [8]. Hence, estimates of variation with respect to half-spaces apply also to variation with respect to perceptrons with any continuous sigmoidal function.

Theorem 2 provides an integral representations in terms of infinite Heaviside networks for functions of weakly controlled decay. This class consists of all functions on $\mathbb{R}^d$ which have sufficiently many continuous derivatives and which vanish sufficiently rapidly at infinity and it contains both the compactly supported functions from $\mathcal{C}^d(\mathbb{R}^d)$ and the Schwartz class $\mathcal{S}(\mathbb{R}^d)$. As the Gaussian function belongs to the Schwartz class, it is of weakly controlled decay.

The following corollary estimates rates of approximation of smooth functions by shallow perceptron networks. The value of a_d is as in (8).

Corollary 1. *Let d be an odd positive integer, $\Omega \subset \mathbb{R}^d$ have finite Lebesgue measure $\lambda(\Omega)$, $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous sigmoidal function, and $f \in \mathcal{C}^d(\mathbb{R}^d)$*

be a function of weakly controlled decay. Then for all n,

$$\|f|_\Omega - \mathrm{span}_n G_\sigma(\Omega)\|_{\mathcal{L}^2(\Omega)} \leq \frac{\lambda(\Omega)\|w_f\|_{\mathcal{L}^1(S^{d-1}\times\mathbb{R})}}{\sqrt{n}},$$

where $w_f(e,b) = a(d)\int_{H_{e,b}}(D_e^{(d)}(f))(y)dy$ and $a(d) = (-1)^{(d-1)/2}(1/2)(2\pi)^{1-d}$.

Another consequence is the following upper bound on the half-space variation of the d-dimensional Gaussian $\gamma_d(x) := \exp(-\|x\|^2)$; see [17, Cor. 6.2].

Corollary 2. *Let d and n be positive integers with d odd. If $\Omega \subset \mathbb{R}^d$ has finite measure λ, then*

$$\|\gamma_d - \mathrm{span}_n G_\vartheta(\Omega)\|_{\mathcal{L}^2(\Omega)} \leq (2\pi d)^{3/4}\lambda^{1/2}/\sqrt{n}.$$

Note that versions of the above results hold in sup norm [17, 9].

We now recall some concepts related to best approximation as mentioned above. Let $M \subset X$, where $(\mathcal{X}, \|\cdot\|)$ is a normed linear space. For the following concepts, see, e.g., [37]. Let 2^M denote the set of all subsets of M. The mapping

$$P_M : X \to 2^M$$

is called the *metric projection* of X to M if, for all $g \in P_M(f)$, $\|f-g\| = \|f-M\|$. The subset M is *proximinal* if $P_M(f)$ is nonempty for all $f \in X$. Thus, M is proximinal iff every element in X has at least one *best approximant* in M.

If $f \in X$ and the sequence $(g_i)_{i=1}^\infty \subset M$ satisfies

$$\|f - M\| = \lim_{i\to\infty}\|f - g_i\|,$$

then (g_i) is called a *distance-minimizing sequence for f in M*. The subset M is *approximatively compact* if, for each $f \in X$ and each distance-minimizing sequence (g_i) for f in M, there is a subsequence $(g_{i'})$ which converges to some $g_0 \in M$. For subsets, approximatively compact $\Rightarrow$ proximinal $\Rightarrow$ closed. A closed convex subset of a Banach space is approximatively compact. For Hilbert space, unique best approximation to a closed linear subspace is obtained via orthogonal projection to such a subspace.

A function β from X to M is called a *continuous best approximation* if β is continuous and for every $f \in X$, $\beta(f) \in P_M(f)$. For $\varepsilon > 0$, β is a *continuous ε-near-best approximation* if β is continuous and for all $f \in X$,

$$\|f - \beta(f)\| \leq \|f - M\| + \varepsilon.$$

A Banach space is *strictly convex* if the line segment joining any two distinct points on the unit sphere intersects the sphere only in its endpoints. For instance, $X = \mathcal{L}^p(\Omega)$ is strictly convex iff $1 < p < \infty$. The following theorem is from [38].

Theorem 4. *Let X be strictly convex. If M is either not closed or not convex, then there does not exist a continuous best approximation from X to M.*

As $\mathrm{span}_n G$ is not convex for $n > 1$, it is not possible to continuously choose a best approximation from $\mathcal{L}^2(\Omega)$ to the input-output functions given by a neural network, no matter what type of units are employed for the computation. This result is strengthened in [39], [40] to show that it is not even possible to find an ε-near-best approximation. However, a noncontinuous and nonunique choice of best approximant does exist when $M = \mathrm{span}_n G_\vartheta$ [35] as implied by the following.

Theorem 5. *For n, d positive integers and every $p \in [1, \infty)$, $\mathrm{span}_n G_\vartheta$ is an approximatively compact subset of $(\mathcal{L}^p([0,1]^d), \|\cdot\|)$.*

This theorem can be extended to any compact convex subset of $\mathbb{R}^d$ (not just the unit cube $[0,1]^d$). Another interesting question is how to find, for a given f in $\mathcal{L}^2([0,1]^d)$, some choice of $g_1, \ldots, g_n \in G_\vartheta$ such that the linear subspace they determine contains a best $\mathcal{L}^2$-approximant to f in $\mathrm{span}_n G$, which must then be the orthogonal projection of f onto this subspace.

6 Discussion

One-hidden-layer networks with many common types of computational units are capable of emulating any reasonable function; i.e., they have the so-called "universal approximation" property. Recently, *deep networks* with several convolutional and pooling layers have become state of the art in computer vision and speech recognition tasks largely due to a progress of hardware (computers with graphic processing units strongly accelerate computation, see the survey article [41] and the references therein). But shallow (one-hidden-layer) networks are still widespread and in some cases can perform the same tasks as deep ones with the same numbers of parameters [42]. Theoretical analysis, complementing the experimental evidence, obtained by some comparisons of deep and shallow networks solving the same tasks, is still in its early stages. While there do exist particular problems where multilayer designs outperform single-layer nets with similar numbers of computational units [43], cost per unit might be lower in shallow architectures. In particular, training or learning is more difficult with more layers as responsibilities become blurred. Another advantage of shallow networks is that the computation might be implementable via physics-based operators, for example, in photonic and quantum computers.

Acknowledgments

V.Kůrková was partially supported by the Czech Grant Foundation grants GA15-18108S and institutional support of the Institute of Computer Science RVO 67985807. P. C. Kainen received research support from Georgetown University.

References

1. Wolf, K.B.: Integral transforms in science and engineering,. Plenum Press, New York (1979)
2. Debnath, L., Bhatta, D.: Integral transforms and their applications. Volume 3rd Ed. CRC Press, Boca Raton, FL (2015)
3. Pietch, A.: Eigenvalues and s-numbers. Cambridge University Press, Cambridge (1987)
4. Fine, T.: Feedforward Neural Network Methodology. Springer, New York (1999)
5. Kecman, V.: Learning and Soft Computing. MIT Press, Cambridge (2001)
6. Rosenblatt, F.: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, New York (1962)
7. Courant, R., Hilbert, D.: Methods of Mathematical Physics. Volume 2. Wiley, New York (1962)
8. Kůrková, V., Kainen, P.C., Kreinovich, V.: Estimates of the number of hidden units and variation with respect to half-spaces. Neural Networks **10** (1997) 1061–1068
9. Kainen, P.C., Kůrková, V., Vogt, A.: Integral combinations of Heavisides. Mathematische Nachrichten **283**(6) (2010) 854–878
10. Ito, Y.: Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. Neural Networks **4** (1991) 385–394
11. Carroll, S.M., Dickinson, B.W.: Construction of neural net using the Radon transform. In: Proceedings of IJCN. Volume I. (1989) 607–611
12. Leshno, M., Lin, V.Y., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks **6** (1993) 861–867
13. Rudin, W.: Real and Complex Analysis. MacGraw-Hill, New York (1974)
14. Edwards, C.H.: Advanced Calculus of Several Variables. Dover, New York (1994)
15. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces. Academic Press, Amsterdam (2003)
16. Zemanian, A.H.: Distribution Theory and Transform Analysis. Dover, New York (1987)
17. Kainen, P.C., Kůrková, V., Vogt, A.: A Sobolev-type upper bound for rates of approximation by linear combinations of Heaviside plane waves. Journal of Approximation Theory **147** (2007) 1–10
18. Kainen, P.C., Vogt, A.: Bochner interals and neural networks. In Bianchini, M., Jain, L., Maggini, M., eds.: Handbook on Neural Information Processing. Springer-Verlag, Berlin, Heidelberg (to appear) (2012)
19. Hill, E., Phillips, R.: Functional Analysis and Semi-Groups. AMS, New York (1996)
20. Diestel, J., Jr., J.J.U.: Vector measures. Bulletin of American Mathematical Society **84** (1978) 681–685
21. Girosi, F., Anzellotti, G.: Rates of convergence for Radial Basis Functions and neural networks. In Mammone, R.J., ed.: Artificial Neural Networks for Speech and Vision. Chapman & Hall (1993) 97–113
22. Chen, T., Chen, H.: Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. IEEE Transactins on Neural Networks **6** (1995) 911–917
23. Kainen, P.C., Kůrková, V.: An integral upper bound for neural network approximation. Neural Computation **21**(10) (2009) 2970–2989

24. Mancera, L., Portilla, J.: L0-norm-based sparse representation through alternate projections. In: IEEE Conference on Image Processing. (2006) 2089–2092
25. Candes, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory **52** (2006) 489–509
26. Ramirez, C., Kreinovich, V., Argaez, M.: Why ℓ_1 is a good approximation to ℓ_0: A geometric explanation. Journal of Uncertain Systems **7** (2013) http://www.jus.org.uk
27. Tillmann, A.: On the computational intractability of exact and approximate dictionary learning. IEEE Signal Processing Letters **22** (2015) 45–49
28. Kůrková, V.: Complexity estimates based on integral transforms induced by computational units. Neural Networks **33** (2012) 160–167
29. Barron, A.R.: Neural net approximation. In Narendra, K., ed.: Proc. 7th Yale Workshop on Adaptive and Learning Systems, Yale University Press (1992)
30. Kůrková, V.: Dimension-independent rates of approximation by neural networks. In Warwick, K., Kárný, M., eds.: Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality. Birkhäuser, Boston, MA (1997) 261–270
31. Pisier, G.: Remarques sur un résultat non publié de B. Maurey. In: Séminaire d'Analyse Fonctionnelle 1980-81, vol. I, no. 12, École Polytechnique, Centre de Mathématiques, Palaiseau, France (1981)
32. Jones, L.K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. Annals of Statistics **20** (1992) 608–613
33. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory **39** (1993) 930–945
34. Kůrková, V.: High-dimensional approximation and optimization by neural networks. In Suykens, J., Horváth, G., Basu, S., Micchelli, C., Vandewalle, J., eds.: Advances in Learning Theory: Methods, Models and Applications. IOS Press, Amsterdam (2003) 69–88 (Chapter 4)
35. Kainen, P.C., Kůrková, V., Vogt, A.: Best approximation by linear combinations of characteristic functions of half-spaces. Journal of Approximation Theory **122** (2003) 151–159
36. Chambolle, A., Caselles, V., Novaga, M., Cremers, D., Pock, T.: An introduction to total variation for image analysis. (2009)
37. Singer, I.: Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces. Spriger, Berlin (1970)
38. Kainen, P.C., Kůrková, V., Vogt, A.: Approximation by neural networks is not continuous. Neurocomputing **29** (1999) 47–56
39. Kainen, P.C., Kůrková, V., Vogt, A.: Geometry and topology of continuous best and near best approximations. J. of Approximation Theory **105** (2000) 252–262
40. Kainen, P.C., Kůrková, V., Vogt, A.: Continuity of approximation by neural networks in L_p-spaces. Ann. of Operational Research **101** (2001) 143–147
41. LeCunn, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521** (2015) 436–444
42. Ba, L.J., Caruana, R.: Do deep networks really need to be deep? In Ghahrani, Z., et al., eds.: Advances in Neural Information Processing Systems. Volume 27. (2014) 1–9
43. Kůrková, V.: Constructive lower bounds on model complexity of shallow perceptron networks. Neural Computing and Applications (2017) DOI 10.1007/s00521–017–2965–0