G-Variation in Lp Spaces and Integral Representation

Šámalová, Terezie
2008

Institute of Computer Science
Academy of Sciences of the Czech Republic

$\mathcal{G}$-variation in $\mathcal{L}^p$-spaces and integral representation

Terezie Šámalová

Technical report No. 1021

June 2008

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 051 111, fax: +420 286 585 789,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

$\mathcal{G}$-variation in $\mathcal{L}^p$-spaces and integral representation

Terezie Šámalová

Technical report No. 1021

June 2008

Abstract:

This paper brings some improvements on known estimates on rates of approximation by neural networks. We proceed along the line proposed by V. Krkov applying integral representations. We prove that existence of integral representation of a function is equivalent to limit of classical neural networks. We give two proofs for finiteness of $\mathcal{G}$-variation for $\mathcal{L}^p$ activation functions. This enables Maurey-Jones-Barron-type estimates to be applied in this more general setting. We show that the known estimates cannot distinguish between sigmoidal activation functions and provide limitations of the approach as well. Applying presented results we finally give estimates for some concrete approximation schemas.

Keywords:
Neural networks, rates of approximation, $\mathcal{G}$-variation, integral representation

In this paper we address a crucial question of interest when building a neural network: how precisely can we approximate a given function using a limited number of units. We proceed along the lines initiated by Barron in the respect that we study approximation by convex combinations of "basic" functions and try to derive classes of functions that can be approximated in such a way.

In Section 1 we first review the pioneering work of Maurey [Ps81], Jones [Jo92], and Barron [Ba93], and the extension by Darken, Donahue, Gurvits, and Sontag [DDGS93]. Then we show how these results were utilized by Krkov, Kainen, and Kreinovich [KKK97] who (implicitly) used so-called $\mathcal{G}$-variation (explicitly defined in [Ku97]) of the function f to be approximated. We will see that bounded $\mathcal{G}$-variation is a sufficient (though not necessary) condition for good rates of approximation.

In Section 2 we first present results of [KKK97] where bounds on $\mathcal{G}$-variation are obtained for functions in the form of integral representation using continuous or Heaviside functions. We then extend their results to more general function spaces. We do not require continuity of the functions involved in the integral representation; we also present simpler proof of the estimate from [KKK97]. The obtained improvements enable more direct and more general application of results of Maurey, Jones, Barron and Darken et al. giving approximation error rate of order $O(n^{1/q})$ for one-hidden-layer networks with n hidden units. Here q is a constant depending on the "type" of the involved function space, but not on the "dimension". E.g., if we are dealing with functions in $\mathcal{L}^p(\mathbb{R}^d)$ $(1 < p < 2)$ then $q = p/(p-1)$ is the conjugate exponent; in particular, q does not depend on d (note that for high d we may obtain large constant in the $O(\cdot)$, though, [KHS98]). Using $\mathcal{L}^p$ spaces for $p \neq 2$ is of practical interest, as by using $\mathcal{L}^p$-norm for $1 < p < 2$ one can cope better with functions with peaks, which are probably errors in measurement, so-called outliers [Re83, HaBu88]. We also present an interesting property of $\mathcal{G}$-variation for neural network approximation schema with sigmoidal activation functions – we show that the presented estimates on approximation rates cannot distinguish between sigmoidal activation functions (Theorem 2.9).

For our estimates on $\mathcal{G}$-variation we need to have function f represented in form of an integral representation. In Section 3 we listed examples of functions where such integral representation exists. We generalise integral representation of function by using measure instead of weights. This enables us to provide in Section 3.2 explanation and justification of the metaphor "neural network with continuum many neurons", which is used in [KKK97] to motivate special type of integral representation of functions. By an application of Helly's theorem on w^* sequential compactness we get that, in a proper setting, such representation is equivalent to a limit of "classical" finite neural networks.

In Section 4 we combine the three previous sections and thus provide a few concrete estimates on rates of approximation of the type: If a function is "smooth enough" then it can be approximated by one-hidden-layer neural network with n units with rate of approximation of $O(n^{1/q})$. We also discuss possibilities to weaken the smoothness assumptions.

Some of the results presented in this paper have been published in [S03a, S03b].

1 Rates of Approximation in Banach Spaces

A general topic (not only) in mathematics is, how to approximate some complicated object using limited resources. To be more specific, we have a Banach space X of functions, and a set $\mathcal{G} \subseteq X$ of functions we are allowed to use for approximation of a given function $f \in X$, while we want to use as few functions from $\mathcal{G}$ as possible.

In Section 1.1 we show results from approximation theory that provide good rates of approximation for function f in the closure of convex hull of the set of approximating functions. In Section 1.2 we show how the above results have been reformulated in a more explicit form taking into account relationship between the set of approximating functions and the function to be approximated. We will see that the condition of f being in convex hull is sufficient (though not necessary) for the existence of efficient approximations of f. On the other hand, if f is merely in the closed linear span, the rates of approximation of f may be arbitrary bad (Corollary 1.8).

1.1 Approximations in the Closed Convex Hull

A frequent approach in approximation theory is to iteratively construct a sequence of approximants f_n to a function f, where at each step we add an appropriate element of $\mathcal{G}$:

$$f_{n+1} = f_n + g, \qquad g \in \mathcal{G}\,. \tag{1.1}$$

Here, g is chosen to minimize the norm $\|f_{n+1} - f\|$ (or to make it close to $\inf\{\|(f_n + g) - f\| : g \in \mathcal{G}\}$). A natural setting for this is when X is a Hilbert space. Huber [Hu85] conjectured that for *projection pursuit regression* (which corresponds to $\mathcal{G}$ consisting of all ridge functions) this method always produces a sequence f_n converging to f. This was affirmatively resolved by Jones [Jo87]. However, the convergence can in general be very slow.

In a subsequent work [Jo92] Jones studies approximations when slightly more general iterative step is allowed – instead of adding some $g \in \mathcal{G}$ to the previous function, we take a convex combination:

$$f_{n+1} = \alpha f_n + (1-\alpha)g, \qquad g \in \mathcal{G}, \alpha \in [0,1] \tag{1.2}$$

where g and α are chosen (approximately) optimal. Somewhat surprisingly, this modification significantly increases the speed of convergence:

Theorem 1.1 (Maurey-Jones-Baron – Iterative rates in Hilbert sp. [Ba93, Jo92, Ps81]) *Let $\mathcal{G}$ be a set of functions, subset of a Hilbert space $\mathcal{H}$ of functions on $\mathbb{R}^d$. Suppose f is in $\operatorname{cl}\operatorname{conv}\mathcal{G}$, and that for every $g \in \mathcal{G}$ we have $\sqrt{\|g\|_{\mathcal{H}}^2 - \|f\|_{\mathcal{H}}^2} \le \rho$ for some constant $\rho \in \mathbb{R}$. Then it is possible to find a sequence $\{f_n\}$ satisfying*

$$\|f - f_n\|_{\mathcal{H}} \le \frac{\rho}{\sqrt{n}}\,,$$

by using the recurrence (1.2), when the functions g and numbers α are chosen sufficiently close to the optimum. Observe that we have $f_n \in \operatorname{conv}_n \mathcal{G}$.

Note that ρ does not depend on n, however it depends on $\mathcal{H}$, $\mathcal{G}$, and, in particular, on f. We will present estimates of this constant later (Theorem 1.5 and 1.6), the dependence on f is actually the topic of the rest of the paper. The above result in slightly weaker version is attributed to Maurey by Pisier [Ps81]. Barron [Ba93] was the first who noticed that it is applicable to neural networks. He also provides the improved bound: instead of $\frac{1}{\sqrt{n}}\sup_{g\in\mathcal{G}}\|f-g\|_{\mathcal{H}}$ (Maurey) he obtains $\frac{1}{\sqrt{n}}\sup_{g\in\mathcal{G}}\sqrt{\|g\|_{\mathcal{H}}^2 - \|f\|_{\mathcal{H}}^2}$, which in the natural applications is lower.

We feel obliged to comment here that the title of the theorem (Iterative rates) is slightly misleading but for a good reason. Maurey's proof of the theorem is in fact probabilistic but we retain the title iterative to stress that an iterative proof is possible as this is interesting from algorithmic point of view. We follow this approach also in titles of further theorems.

The above result was extended in various ways. The strongest result obtained in this direction is due to Makovoz [Mk96]. He replaces the bound $\rho/\sqrt{n}$ by $\varepsilon_n(\mathcal{G})/\sqrt{n}$, where $\varepsilon_n(\mathcal{G}) \in (0,\rho]$ depends on $\mathcal{G}$ and on n:

$$\varepsilon_n(\mathcal{G}) = \inf\{\varepsilon > 0 : \mathcal{G} \text{ can be covered by at most } n \text{ sets of diameter} \le \varepsilon\}\,.$$

When $\mathcal{G}$ is finite-dimensional, $\varepsilon_n(\mathcal{G}) = o(1)$ (as $n \to \infty$), so this is a stronger result. Consider the particular case where $\mathcal{G}$ corresponds to neural networks with Heaviside activation functions, with inputs in $\mathbb{R}^d$. In this case $\mathcal{G} = \mathcal{G}_\vartheta = \{\vartheta(a\cdot x+b), a \in \mathbb{R}^d, b \in \mathbb{R}\}$. This yields [Mk96] an improved bound on the error of the n-term approximation, namely $O(1/n^{\frac{1}{2}+\frac{1}{2d}})$. We will not pursue this direction, as our particular interest is on the case of large d, where the improvement is only slight. The drawback of Makovoz' approach is that it does not yield existence of approximants that can be computed in an iterative manner, as in (1.2).

Let us pause here to explain the dependence on the "dimension of the data". In early proofs of universal approximation property of neural networks, the "amount of work" needed for efficient approximation of a function on $\mathbb{R}^d$ seemed to depend exponentially on the dimension d. This so-called "curse of dimensionality" is obviously a major obstacle in applications of neural networks, as many

interesting applications are intrinsically multi-dimensional. Theorem 1.1 tells us, that for a fixed function f, space $\mathcal{H}$ and approximating functions $\mathcal{G}$, the error of approximation decreases fast with n, the number of approximants. This is certainly useful (and in particular proves superiority of the approximation schema (1.2) over (1.1)), the dimension, however, is "cursed" in more ways than this. One other problem is, that with increasing dimension of inputs, we are likely to see larger constants ρ in Theorem 1.1. In [KHS98] a sequence of functions is presented, where ρ grows exponentially with the dimension. Yet another problem is met when we consider the algorithmic point of view. The amount of work to do "elementary operations" (estimating the norm, scalar product, etc.) with functions on $\mathbb{R}^d$ grows exponentially with d. This can be remedied by using more sophisticated numerical methods (as Monte Carlo), however. We address some of these issues in [SS08].

It is natural to ask, whether the above-mentioned result can be generalized to arbitrary Banach spaces. Not only this is an interesting question in itself, it was motivated by the fact, that spaces $\mathcal{L}^p$ (for $p < 2$) possess better approximation properties than $\mathcal{L}^2$: namely, they can cope better with an "error in measurement" of the function to be approximated [Re83, HaBu88].

In Darken et al. [DDGS93] this question was addressed in a great detail. It is shown, that Theorem 1.1 can be extended to any Banach space with unit ball that is not too "pointed" – namely to any *uniformly smooth* space. We say that a Banach space X has *modulus of smoothness* ϱ if $\varrho : [0, \infty) \to [0, \infty)$ is a function given by

$$\varrho(r) := \sup_{\|f\|_X = \|g\|_X = 1} \Big(\frac{\|f + rg\|_X + \|f - rg\|_X}{2} - 1 \Big)$$

(the supremum is taken over all $f, g \in X$ of unit norm). It is easy to observe that $\varrho(r) \leq r$ for any Banach space, and that in a Hilbert space $\varrho(r) = \sqrt{1 + r^2} - 1 = O(r^2)$ (as $r \to 0$). A Banach space is termed *uniformly smooth* if $\varrho(r) = o(r)$ (as $r \to 0$). This is in particular satisfied for $\mathcal{L}^p$ spaces with $1 < p < \infty$, the modulus of smoothness is (see [DDGS93])

$$\varrho(r) \leq \begin{cases} r^p/p & \text{if } 1 < p \leq 2 \\ \frac{p-1}{2} r^2 & \text{if } 2 \leq p < \infty. \end{cases}$$

Darken et al. [DDGS93] prove a result about approximating functions in Banach spaces based on modulus of smoothness of these spaces. This theorem applied to $\mathcal{L}^p$ spaces yields Theorem 1.2. It also turns out, that the convex combination in (1.2) can be chosen so that $\alpha = n/(n + 1)$.

Theorem 1.2 (Rates in $\mathcal{L}^p$ spaces – iterative [DDGS93]) *Let $\mathcal{G}$ be a bounded subset of an $\mathcal{L}^p$-space X ($1 < p < \infty$), with $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}$ given. Put $q = p/(p-1)$ and let $\rho > 0$ be such, that $\|f - g\| \leq \rho$ for all $g \in \mathcal{G}$. Then for every $\varepsilon > 0$ there is a sequence $\{g_n\} \subset \mathcal{G}$ such that the sequence $\{f_n\} \subset \operatorname{conv}\mathcal{G}$ defined by*

$$f_1 = g_1, \qquad f_{n+1} = \frac{n}{n+1} f_n + \frac{1}{n+1} g_n$$

satisfies

$$\|f - \operatorname{conv}_n \mathcal{G}\| \leq \|f - f_n\| \leq \frac{2^{1/p}(\rho + \varepsilon)}{n^{1-1/p}} \big(1 + \frac{(p-1)\log_2 n}{n}\big)^{1/p} \qquad \text{if } 1 < p \leq 2$$

and

$$\|f - \operatorname{conv}_n \mathcal{G}\| \leq \|f - f_n\| \leq \frac{(2p-2)^{1/2}(\rho + \varepsilon)}{n^{1/2}} \big(1 + \frac{\log_2 n}{n}\big)^{1/2} \qquad \text{if } 2 \leq p < \infty.$$

When we lift the condition to construct the approximants iteratively, it is possible to get somewhat better bounds. The improvement is only in the constant factor – in this case, however, the result is tight for $p \in (1, 2]$; for $p > 2$ it is still "only" asymptotically tight. Theorem 1.3 is obtained by a different approach than Theorem 1.2 – by using the probabilistic method in Banach spaces. We discuss algorithmic consequences of this approach in [SS08].

Theorem 1.3 (Rates in $\mathcal{L}^p$ spaces – probabilistic [DDGS93]) *Let $\mathcal{G}$ be a bounded subset of an $\mathcal{L}^p$-space X ($1 < p < \infty$), with $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}$ given. Put $q = \max\{p/(p-1), 2\}$ and let $\rho > 0$ be such, that $\|f - g\| \le \rho$ for all $g \in \mathcal{G}$. Then for all n*

$$\|f - \operatorname{conv}_n \mathcal{G}\| \le \frac{\rho C_p}{n^{1/q}}\,.$$

Here $C_p = 1$ if $p \le 2$ and $C_p = \sqrt{2}\big(\Gamma(\frac{p+1}{2})/\sqrt{\pi}\big)^{1/p}$ for $p > 2$. For large p, $C_p \sim \sqrt{p/e}$.

Further we go into more detail regarding the constants that appear in the presented estimates.

1.2 Approximation Rates using $\mathcal{G}$-variation

The results of Jones and of Darken et al. were used by Krkov [Ku97, Ku03], Krkov, Kainen, and Kreinovich [KKK97] and Krkov, Kainen and Vogt [KKV07]. Krkov [Ku97] exhibited a natural way to obtain functions f and system of functions $\mathcal{G}$, such that $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}$ and the constant ρ from the previous section can be estimated. As we will build on and extend their results, we explain them now in some detail.

Consider a set $\mathcal{G}$ of functions, a bounded subset of a Banach space X. For convenience, we will assume that $g \in \mathcal{G}$ implies $-g \in \mathcal{G}$.[1] A function $f \in X$ can be approximated arbitrarily well by a linear combination of elements of $\mathcal{G}$ if and only if $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$. To apply the results of [Jo92, DDGS93] we need a set $\mathcal{G}'$ such that $f \in \operatorname{cl}\operatorname{conv}\mathcal{G}'$. As

$$\operatorname{cl}\operatorname{span}\mathcal{G} = \operatorname{cl}\bigcup_{c>0}\operatorname{conv} c\mathcal{G}$$

we may try to put $\mathcal{G}' = c\mathcal{G}$ for some c. To this end, we follow Krkov [Ku97][2] and define *$\mathcal{G}$-variation* as the Minkowski functional of the set $\operatorname{cl}\operatorname{conv}\mathcal{G}$. Note that $\mathcal{G} = -\mathcal{G}$ implies $\operatorname{conv}\mathcal{G} = \{\sum_i c_i g_i : g_i \in \mathcal{G}, c_i \ge 0, \sum_i c_i \le 1\}$. Consequently, the set $\operatorname{cl}\operatorname{conv}\mathcal{G}$ is convex, bounded, *balanced* (that is, $h \in \operatorname{conv}\mathcal{G}$ and $|a| \le 1$ implies $ah \in \operatorname{conv}\mathcal{G}$) and closed. Thus, we may put

$$\|f\|_{\mathcal{G}} = \inf\{c > 0 \mid f \in \operatorname{cl}\operatorname{conv} c\mathcal{G}\} \tag{1.3}$$

and we get a norm on the subspace $\{f : \|f\|_{\mathcal{G}} < \infty\}$. Note that (1.3) defines $\|f\|_{\mathcal{G}} = \infty$ if we do not have $f \in \operatorname{cl}\operatorname{conv} c\mathcal{G}$ for any c. This will certainly happen if $f \notin \operatorname{cl}\operatorname{span}\mathcal{G}$, in which case we can not get arbitrary close approximations. It will also happen when $f \in \operatorname{cl}\operatorname{span}\mathcal{G} \setminus \operatorname{span}\mathcal{G}$. As the next example shows, this may occur even for "reasonable" f and $\mathcal{G}$.

Example 1.4 (Infinite $\mathcal{G}$-variation) *Consider $X = \ell^2$ and let $\{e_k\}_{k=1}^{\infty}$ be the orthonormal basis ($e_k = (0, \dots, 0, 1, 0, \dots)$ is the sequence with 1 exactly at the k-th place). Then we put $f = (1/k)_{k\ge1}$, $f_n = (1, 1/2, \dots, 1/n, 0, \dots)$, and $\mathcal{G} = \{\pm e_k \mid k \ge 1\}$. It is easy to see that f_n is the best approximant to f in $\operatorname{span}_n\mathcal{G}$ and that $\|f - f_n\|_2 \to 0$. However, $\|f_n\|_{\mathcal{G}} = \sum_{i=1}^{n}\frac{1}{i} \sim \log n$, and so $\|f\|_{\mathcal{G}} = \infty$.*

In the example above, the error of approximation of f by combination of n terms is $\sqrt{\sum_{k>n}\frac{1}{k^2}} = O(1/\sqrt{n})$, so one may think, that the assumption on finite $\mathcal{G}$-variation is not that crucial. However, this is not the case, as we will show later, in Theorem 1.7.

We need one more definition to describe the approach of [Ku97, Ku03]. Let $s_{\mathcal{G}} = \sup\{|g| : g \in \mathcal{G}\}$. Recall that we are assuming $\mathcal{G}$ to be bounded, so $s_{\mathcal{G}} < \infty$. A consequence of the definition of $\|f\|_{\mathcal{G}}$ is that $f \in \operatorname{cl}\operatorname{conv} c\,\mathcal{G}$ for $c \ge \|f\|_{\mathcal{G}}$. If $c = \|f\|_{\mathcal{G}}$, and $g \in c\,\mathcal{G}$, then the following bound holds:

$$\|f - g\| \le \|f\| + \|g\| \le 2\sup\{\|h\| : h \in c\mathcal{G}\} = 2s_{\mathcal{G}}\|f\|_{\mathcal{G}}\,.$$

Consequently, one gets the following corollaries of results of Jones/Barron and of Darken et al., as stated by Krkov [Ku97, Ku03], see also [KKK97, KKV07].

[1] This will simplify some of the expressions and is satisfied for the practically interesting applications.

[2] who did extend the concept of *variation with respect to half-spaces* introduced in [Ba92].

Theorem 1.5 (Rates with $\mathcal{G}$-variation – iterative [Ku97]) *Let $\mathcal{H}$ be a Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$ and let $\mathcal{G}$ be a bounded subset of $\mathcal{H}$. Let us denote $s_{\mathcal{G}} = \sup_{g\in\mathcal{G}} \|g\|_{\mathcal{H}}$. Then, for every $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$ with finite $\|f\|_{\mathcal{G}}$ and for every natural number n the following holds:*

$$\|f - \operatorname{span}_n \mathcal{G}\|_{\mathcal{H}} \leq \frac{\sqrt{(s_{\mathcal{G}}\|f\|_{\mathcal{G}})^2 - \|f\|_{\mathcal{H}}^2}}{\sqrt{n}}.$$

Theorem 1.6 (Rates with $\mathcal{G}$-variation – probabilistic [DDGS93, Ku03]) *Let $\mathcal{G}$ be a bounded subset of an $\mathcal{L}^p$-space X ($1 < p < \infty$) and $s_{\mathcal{G}} = \sup_{g\in\mathcal{G}} \|g\|_p$. Let $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$ have finite $\|f\|_{\mathcal{G}}$. Then for every n*

$$\|f - \operatorname{span}_n \mathcal{G}\|_p \leq \frac{2C_p s_{\mathcal{G}}\|f\|_{\mathcal{G}}}{n^{1-1/t}},$$

where $t = \min\{p, 2\}$, $C_p = 1$ if $p \leq 2$ and $C_p = \sqrt{2}\big(\Gamma(\frac{p+1}{2})/\sqrt{\pi}\big)^{1/p}$ for $p > 2$. For large p, $C_p \sim \sqrt{p/e}$.

Note that in both the theorems above one could instead of $\|f - \operatorname{span}_n \mathcal{G}\|$ write the more accurate expression $\|f - \operatorname{conv}_n c\mathcal{G}\|$, where $c = \|f\|_{\mathcal{G}}$. This actually gives a stronger result: we do not need to use the entire span of $\mathcal{G}$ to attain good approximation. This is interesting also from the numerical point of view: it will not happen that we need to work with big numbers to approximate small ones as convex combinations work with $c_i \in (0, 1), \sum c_i = 1$.

Further one would like to estimate $\|f\|_{\mathcal{G}}$ in concrete instances of approximation schemas; we do this for neural networks in the next section. Before that, we discuss assumptions in the above estimates on rate of convergence.

Analogues of Theorem 1.6 are false in many spaces of interest, including $C[0, 1]$ and $\mathcal{L}^1[0, 1]$. By Theorem 2.3 and 2.4 in [DDGS93], in such spaces we may see arbitrary slow convergence even for elements of $\mathcal{G}$-variation equal to 1. We complement this by showing that the same happens in ℓ^p spaces for elements of infinite $\mathcal{G}$-variation. (Note that by the obvious embedding this yields the same result for $\mathcal{L}^p$ spaces as well.)

Theorem 1.7 (Slow rate of approximation) *Suppose $1 < p < \infty$ and let $(a_n)_{n=0}^{\infty}$ be a sequence of real numbers decreasing to 0 so that the sequence (a_n^p) is convex (that is, $a_{n-1}^p + a_{n+1}^p \geq 2a_n^p$). Then there is a set $\mathcal{G} \subseteq \ell^p$ and an element $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$ so that*

$$\|f - \operatorname{span}_n \mathcal{G}\|_p = a_n.$$

So we have $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$ and the rate of convergence is a_n.

This is in particular possible for $a_n = 1/n^{\alpha}$ (for any $\alpha > 0$), and $a_n = 1/\log^k n$ (where $\log^k$ denotes k-times iterated logarithm). More generally, we may have $a_n = 1/g(n)$ whenever g is a concave increasing function with limit ∞.

Proof: We let $\mathcal{G} = \{\pm e_i : i \geq 0\}$. Put $b_n = (a_n^p - a_{n+1}^p)^{1/p}$ for $n \geq 0$ and $f = (b_0, b_1, b_2, \dots)$. As $a_n \geq a_{n+1}$, the numbers b_n are well-defined and an easy computation shows $\|f\|_p = a_0$, in particular f is in ℓ^p. Convexity of the sequence (a_n^p) implies that $b_n^p = a_n^p - a_{n+1}^p$ is decreasing, so the element of $\operatorname{span}_n \mathcal{G}$ closest to f is $f_n = (b_0, \dots, b_{n-1}, 0, 0, \dots)$. Now

$$\|f - f_n\|_p = \sum_{i\geq n}(a_n^p - a_{n+1}^p) = a_n,$$

as claimed.

For the specific examples of sequences (a_n): $(x^{-p\alpha})'' = p\alpha(p\alpha+1)x^{-p\alpha-2}$ is positive for $x > 0, \alpha > 0$, so $x^{-p\alpha}$ is a convex function, thus $n^{-p\alpha}$ is a convex sequence. If $a_n = 1/g(n)$ then $a_n^p = 1/g(n)^p$. The first derivative (replacing again n by a continuous variable) is

$$\Big(\frac{1}{g(x)^p}\Big)' = \frac{-pg'(x)}{g(x)^{p+1}},$$

which is an increasing function (as g' is decreasing and g increasing), thus a_n^p is convex as required. Computing the first derivative of the iterated logarithm reveals that it is a concave function, which finishes the proof. □

The convexity assumption in the above theorem may be a bit misleading. It is in fact possible to do away with this assumption, if we only want a lower bound on the error of approximation.

Corollary 1.8 (Slow rate of approximation – lower bound) *Suppose $1 < p < \infty$ and let $(a_n)_{n=0}^{\infty}$ be a strictly decreasing sequence of real numbers converging to 0. Then there is a set $\mathcal{G} \subseteq \ell^p$ and an element $f \in \operatorname{cl}\operatorname{span}\mathcal{G}$ so that*

$$\|f - \operatorname{span}_n \mathcal{G}\|_p \geq a_n.$$

Proof: We find a sequence $a'_n \geq a_n$ so that a'_n decreases to 0 and $(a'_n)^p$ is convex. This is a standard exercise in analysis, we may for example take $a'_n = \max\{a_n, (2(a'_{n-1})^p - (a'_{n-2})^p)^{1/p}]\}$. Then we apply Theorem 1.7 for (a'_n). □

2 Properties of $\mathcal{G}$-variation

So far we showed several results describing how efficiently can we approximate a function f using functions in some set $\mathcal{G}$, provided $f \in \operatorname{cl}\operatorname{conv} c\mathcal{G}$ holds for some constant c. (Recall that cl and "approximation" are to be understood with respect to some Banach space that contains f and $\mathcal{G}$.) Several questions come up. Given $\mathcal{G}$, for which functions f such finite constant c exists? How can we estimate it?

In Theorem 1.7 we have shown that for elements of $\operatorname{cl}\operatorname{span}\mathcal{G}$, the rate of approximation can be arbitrarily bad; this happens if the $\mathcal{G}$-variation is infinite. Perhaps surprisingly, the situation turns out to be different for systems $\mathcal{G}$ corresponding to neural networks – such systems $\mathcal{G}$ are sufficiently rich, so that $\|f\|_{\mathcal{G}}$ is finite for large class of functions f.

In Section 2.1 we start with the general set-up and with bounds for the $\mathcal{G}$-variation due to Krkov et al. [KKK97]. In Sections 2.2 and 2.3 we follow up with developing bounds for $\mathcal{G}$-variation that are applicable in a more general setting of Banach spaces. Finally, in Section 2.4 we clarify the dependence on the activation function.

2.1 $\mathcal{G}$-variation: Continuous / Heaviside Activation Functions

To answer the questions regarding existence and finiteness of c in the expression $c\mathcal{G}$ we have to be more specific as to the task investigated: We consider one-hidden-layer neural networks, which consist of interconnected computational units with activation functions depending on parameters and input variables: Consider a function $\varphi(x, a) : H \times A \to \mathbb{R}$, where $x \in H$ are inputs and $a \in A$ parameters, $H \subseteq \mathbb{R}^d$, $A \subseteq \mathbb{R}^k$. For $a \in A$ we let $\varphi_a = \varphi(\cdot, a)$ be the function parametrized by a. One-hidden-layer network with n units of type φ computes a function of d variables of the form:

$$f(x) = \sum_{i=1}^{n} w_i \varphi_{a_i}(x)\,,$$

where $w_i \in \mathbb{R}$, $a_i \in A$, and $x \in H$. More specifically, in the case of neural networks we would typically let

$$\varphi(x, a) = \sigma(w \cdot x + \theta), \quad \text{where } a = (w, \theta) \in \mathbb{R}^{d+1}\,. \tag{2.1}$$

for perceptron–type networks or

$$\varphi(x, a) = \sigma\big(\frac{x - w}{\theta}\big), \quad \text{where } a = (w, \theta) \in \mathbb{R}^{d+1} \tag{2.2}$$

for RBF networks.

Following Krková [Ku03] and Krková et al. [KKK97], we extend this notion to a "continuum of hidden units". That is, we consider functions with integral representation

$$f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a\,, \tag{2.3}$$

with a weight function $w : A \to \mathbb{R}$. (We will discuss the relation between such generalized neural networks and ordinary neural networks later, in Section 3.2.)

We wish to apply the results of the previous section to a more specific case of the set $\mathcal{G}$, namely we put

$$\mathcal{G} = \{\pm\varphi_a : a \in A\}\,.$$

In the case that φ is given by (2.1), we use $\mathcal{G}_\sigma$ to denote this particular set $\mathcal{G}$. We will consider the set $\mathcal{G}$ as a subset of various spaces of functions from H to $\mathbb{R}$. The results to follow show in various circumstances how function f can be approximated by convex combinations of elements of $\mathcal{G}$. In view of Theorem 1.5 and 1.6, this amounts to estimating $\mathcal{G}$-variation of f, that is $\|f\|_{\mathcal{G}}$. (A surprising phenomenon is that $\|f\|_{\mathcal{G}_\sigma}$ does in fact not depend on σ, for a large class of functions σ. This first appeared implicitly in [KKK97], we will prove this more generally as Theorem 2.9.) The following result appears as Corollary 2.3 in [KKK97] (without explicitly using the term $\mathcal{G}$-variation).

Theorem 2.1 ($\mathcal{G}$-variation for continuous activation functions [KKK97]) *Let d, k be positive integers, $H \subseteq \mathbb{R}^d$ and $A \subseteq \mathbb{R}^k$ compact sets. Finally, let $w \in C(A)$, $\varphi \in C(H \times A)$ and $\mathcal{G} = \{\pm\varphi_a\}$.*

Let $f \in C(H)$ be represented as

$$f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a\,. \tag{2.4}$$

Then $f \in \mathrm{cl}_C \operatorname{conv} c\,\mathcal{G}$, where $c = \int_A |w(a)|\,\mathrm{d}a$. Using our previous terminology, $\|f\|_{\mathcal{G}} \le \|w\|_1$.

We have to give some remarks here: The theorem speaks about supremum norm. Careful reader might have noticed, that theorems on rates of approximation (Section 1.1) do not hold for this norm. However, when the measure of H is finite (as it is when H is compact), then closure in $C(H)$ is contained in the closure in $\mathcal{L}^p(H)$. Consequently, the above theorem can be combined with the results in Section 1.1.

Note that in [KKK97] a slightly sharper version is presented, with $c = \int_{A_\varphi} |w(a)|\,\mathrm{d}a$, where A_φ is the set of $a \in A$, such that $\varphi(x,a) \neq 0$ for some $x \in H$. We present here the shorter statement, because for natural choices of activation function φ we have $A_\varphi = A$.

Without going into details, the main idea of the proof of Theorem 2.1 in [KKK97] is using the definition of Riemann integral to approximate an integral by a sum. In Section 2.2 we generalize this result to bounded functions in $\mathcal{L}^p$. To prove that, we will use Luzin's theorem to approximate a measurable function by a continuous function (and Fubini's theorem to deal with the error of the approximation). In Section 2.3 we will use more abstract functional-analytic approach to generalize this result even further, with easier (though non-constructive) proofs.

Theorem 2.1 was further extended in [KKK97] to $\varphi(x,a)$ given by the Heaviside function ($\vartheta(x) = 1$ for $x \ge 0$, and $\vartheta(x) = 0$ otherwise):

Theorem 2.2 ($\mathcal{G}$-variation for Heaviside activation functions [KKK97]) *Let d be a positive integer, $A \subseteq S^{d-1} \times \mathbb{R}$, where S^{d-1} denotes the unit sphere in $\mathbb{R}^d$. Let H be a compact subset of $\mathbb{R}^d$ and let $f \in C(H)$ be any function that can be represented as*

$$f(x) = \int_A w(\mathbf{e},b)\vartheta(\mathbf{e}\cdot\mathbf{x} + b)\,\mathrm{d}(\mathbf{e},b)$$

where $w \in C(S^{d-1} \times \mathbb{R})$ is compactly supported and $\mathrm{supp}(w) \subseteq A$. Then $f \in \mathrm{cl}_C \operatorname{conv} c\mathcal{G}_\vartheta$, where $c = \int_A |w(\mathbf{e},b)|\,\mathrm{d}(\mathbf{e},b)$. Using our previous notation, $\|f\|_{\mathcal{G}_\vartheta} \le \|w\|_1$.

2.2 $\mathcal{G}$-variation in $\mathcal{L}^p$ Spaces

In this section we shall use the Luzin's theorem to extend Theorem 2.1 of Kůrková, Kainen and Kreinovich [KKK97] to a more general setting, where the activation functions need no longer be continuous. This also in a sense generalizes Theorem 2.2: we do not prove that some function is in cl_C (closure in the supremum norm), but only that it is in $\mathrm{cl}_{\mathcal{L}^p}$ (closure in the $\mathcal{L}^p$-norm), which is, however, what we will use later to obtain rates of approximation using results in Section 4.

Theorem 2.3 ($\mathcal{G}$-variation in $\mathcal{L}^p$ spaces) *Let k, d be positive integers, let $p \in [1, \infty)$. Consider sets $A \subseteq \mathbb{R}^k$ and $H \subseteq \mathbb{R}^d$ of finite measure, that is $\lambda_k(A) < \infty$ and $\lambda_d(H) < \infty$.*

Consider functions $w \in \mathcal{L}^1(A, \lambda_k)$ and $\varphi \in \mathcal{L}^p(H \times A, \lambda_{d+k})$ such that there exists $b \in \mathbb{R}$ so that

- *$|w| \leq b$ holds λ_k-almost everywhere on A and*
- *$|\varphi| \leq b$ holds λ_{d+k}-almost everywhere on $H \times A$.*

Put $f(x) = \int_A w(a)\varphi(x, a)\,\mathrm{d}a$ and $\mathcal{G} = \{\pm\varphi(\cdot, a) \mid a \in A\} \subseteq \mathcal{L}^p(H)$. Then

$$\|f\|_{\mathcal{G}} \leq \|w\|_1.$$

Note: an almost everywhere bounded function on a set of finite measure is clearly in $\mathcal{L}^p$ for any p. We still included the condition $\varphi \in \mathcal{L}^p$ and $w \in \mathcal{L}^1$ to indicate the "right" spaces to consider these functions in. In particular, let us emphasize that in the definition of $\|f\|_{\mathcal{G}}$ the $\mathcal{L}^p$ norm is used. See also the end of Section 2.3, where the achieved results are stated in terms of bounds on certain operators.

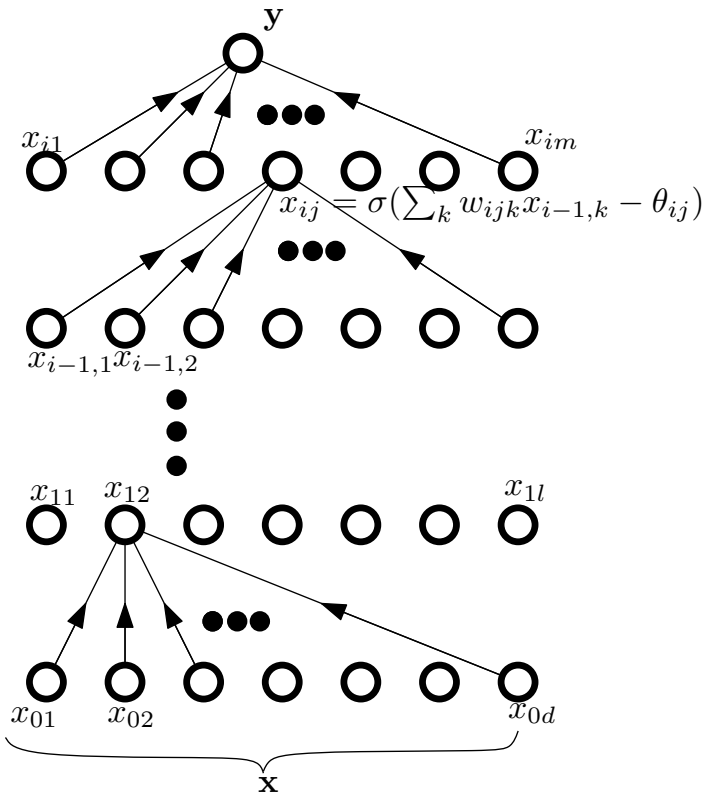

Figure 2.1: Illustration of proof of Theorem 2.3.

Proof: In the definition of $\|f\|_{\mathcal{G}}$, the underlying space (in our case $\mathcal{L}^p$) is used. Thus, the statement we are proving can be equivalently rewritten as follows

$$f(x) \in \mathrm{cl}_p \,\mathrm{conv}\{c\varphi(x, a) : a \in A, |c| \leq \|w\|_1\}.$$

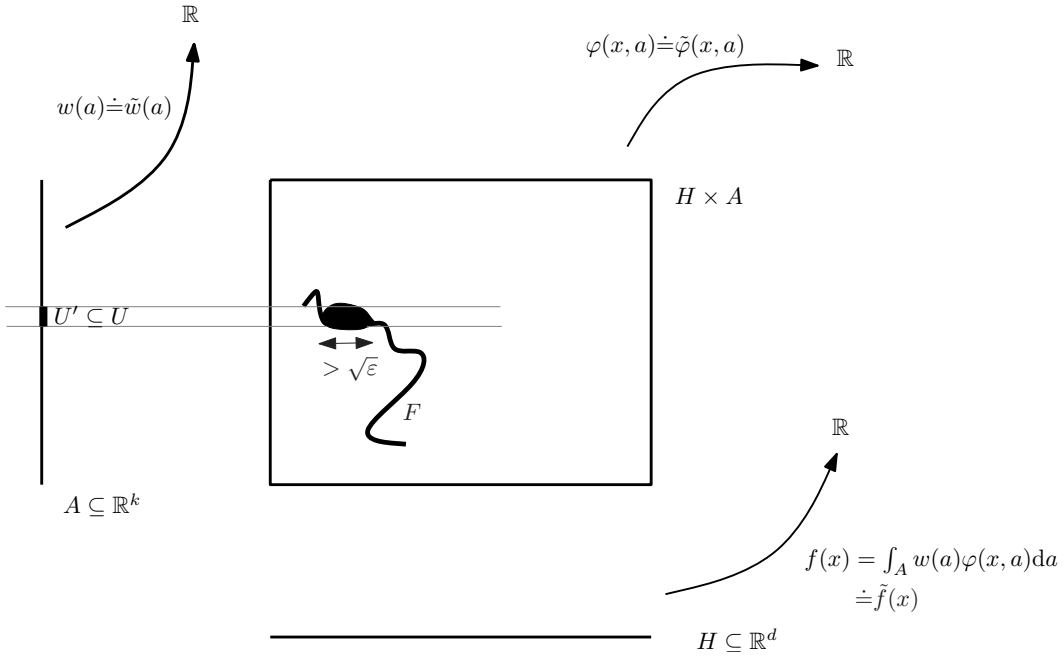

Figure 2.2: ?

We will prove that f can be approximated arbitrarily well (in $\mathcal{L}^p$-norm) by functions of the type

$$\sum_i c_i\varphi(x,a_i) \qquad \text{with } a_i \in A,\ c_i \geq 0, \text{ and } \textstyle\sum_i c_i \leq \|w\|_1.$$

To achieve this we first approximate functions w and φ by continuous functions (using a version of Luzin's theorem) and then apply Theorem 2.1.

We start, however, by showing that we can restrict to the case when A and H are compact. A finite measure subset of $\mathbb{R}^n$ can be approximated arbitrarily closely by its compact subsets (Lemma 15.3 and Theorem 26.1 of [LuMa95]). So let us choose $\varepsilon > 0$ and find compact sets A^{cp}, H^{cp}, so that $A^{cp} \subseteq A$, $H^{cp} \subseteq H$, $\lambda_k(A \setminus A^{cp}) < \varepsilon$, and $\lambda_d(H \setminus H^{cp}) < \varepsilon$. We apply the theorem for sets A^{cp}, H^{cp} instead of A, H. That is, we put

$$f^{cp}(x) = \int_{A^{cp}} w(a)\varphi(x,a)\,\mathrm{d}a$$

for $x \in H^{cp}$ and find an approximation of f^{cp} in $\mathcal{L}^p(H^{cp})$ by a function $\widetilde{f} = \sum_i c_i\varphi(x,a_i)$, where $a_i \in A^{cp}$, $c_i \geq 0$, and $\sum_i c_i \leq \|w\|_{\mathcal{L}^1(A^{cp})} \leq \|w\|_{\mathcal{L}^1(A)} (= \|w\|_1)$. We can demand

$$\|f^{cp} - \widetilde{f}\|_{\mathcal{L}^p(H^{cp})} < \varepsilon$$

and we only need to observe, that $\widetilde{f}$ (extended to H) is close to f. Clearly, $|f(x) - f^{cp}(x)| < \varepsilon \cdot b^2$ whenever $x \in H^{cp}$. For any $x \in H$, we have $\max\{|f(x)|, |\widetilde{f}(x)|\} < b\|w\|_1$. Together we have

$$\begin{aligned}
\|f - \widetilde{f}\|^p_{\mathcal{L}^p(H)} &= \int_H |f(x) - \widetilde{f}(x)|^p \,\mathrm{d}x \\
&= \int_{H^{cp}} |f(x) - \widetilde{f}(x)|^p + \int_{H \setminus H^{cp}} |f(x) - \widetilde{f}(x)|^p \\
&\leq \|(f - f^{cp}) + (f^{cp} - \widetilde{f})\|^p_{\mathcal{L}^p(H^{cp})} + \int_{H \setminus H^{cp}} (|f(x)| + |\widetilde{f}(x)|)^p \\
&\leq \big(\|f - f^{cp}\|_{\mathcal{L}^p(H^{cp})} + \|(f^{cp} - \widetilde{f})\|_{\mathcal{L}^p(H^{cp})}\big)^p + \int_{H \setminus H^{cp}} (|f(x)| + |\widetilde{f}(x)|)^p \\
&\leq \big(\varepsilon \cdot b^2 \cdot \lambda_d(H^{cp}) + \varepsilon\big)^p + (2b \cdot \|w\|_1)^p \cdot \varepsilon \\
&= O(\varepsilon), \text{as } \varepsilon \text{ tends to } 0.
\end{aligned}$$

Thus we will assume further on, that A, H are compact sets.

Let us fix an $\varepsilon > 0$, we may assume that $\varepsilon < 1$. Using Luzin's Theorem we find a continuous function $\widetilde{w}$ on A and a set $E \subseteq A$ such that

$$\widetilde{w} = w \text{ on } A \setminus E,\ |\widetilde{w}| \leq b \text{ on } A, \text{ and } \lambda_k(E) < \varepsilon\,.$$

Similarly, we find function $\widetilde{\varphi}$ on $H \times A$ and a set $F \subseteq H \times A$ such that

$$\widetilde{\varphi} = \varphi \text{ on } (H \times A) \setminus F,\ |\widetilde{\varphi}| \leq b \text{ on } H \times A, \text{ and } \lambda_{d+k}(F) < \varepsilon\,.$$

In order to apply Theorem 2.1 we need to define another small "exceptional set" to describe where our approximation fails, namely the set of such a's that for many x's the functions φ and $\widetilde{\varphi}$ differ on (a, x). To be precise, put

$$U' = \big\{a \in A \mid \lambda_d\{x; (a, x) \in F\} > \sqrt{\varepsilon}\big\} \cup E\,.$$

By an application of Fubini's theorem we get that $\lambda_k(U') < \sqrt{\varepsilon} + \varepsilon$. Continuity of measure implies that we can choose an open set $U \supseteq U'$ such that $\lambda_k(U) < 2\sqrt{\varepsilon}$ (recall that $\varepsilon < 1$). Finally we define

$$\begin{aligned}
\widetilde{f}(x) &= \int_{A \setminus U} \widetilde{\varphi}(x, a)\widetilde{w}(a) \,\mathrm{d}a \\
\widetilde{\mathcal{G}} &= \{\pm\widetilde{\varphi}(\cdot, a) \mid a \in A \setminus U\}\,.
\end{aligned}$$

Next we use Theorem 2.1 for the functions $\widetilde{w}$, $\widetilde{\varphi}$, and $\widetilde{f}$, set $\widetilde{\mathcal{G}}$ and with the set $A \setminus U$ in place of A. We conclude that

$$\|\widetilde{f}\|_{\widetilde{\mathcal{G}}} \leq \|\widetilde{w}\|_1\,.$$

This means that there is $n \in \mathbb{N}$, and $c_i \in \mathbb{R}$, $a_i \in A \setminus U$ $(i = 1, \ldots, n)$ such that $\sum_{i=1}^n |c_i| \leq \|\widetilde{w}\|_1$, and for the function $\widetilde{f}_1$ defined by

$$\widetilde{f}_1(x) = \sum_{i=1}^n c_i \widetilde{\varphi}(x, a_i)$$

we have $|\widetilde{f}(x) - \widetilde{f}_1(x)| < \varepsilon$ for all $x \in H$. We use these parameters to define our desired approximant, f_1:

$$f_1(x) = \sum_i c_i \varphi(x, a_i)\,.$$

We have $\|\widetilde{w}\|_1 = \int_{A \setminus U} |\widetilde{w}| = \int_{A \setminus U} |w| \leq \int_A |w| = \|w\|_1$. To finish the proof, we need to establish an upper bound on $\|f - f_1\|_p$. To this end, we first use the triangle inequality

$$\|f - f_1\|_p \leq \|f - \widetilde{f}\|_p + \|\widetilde{f} - \widetilde{f}_1\|_p + \|\widetilde{f}_1 - f_1\|_p\,.$$

Now we deal with these three terms one by one.

(A) $\|\widetilde{f} - \widetilde{f}_1\|_p$
We know that $|\widetilde{f}(x) - \widetilde{f}_1(x)| < \varepsilon$ on H, thus $\|\widetilde{f} - \widetilde{f}_1\|_p < \varepsilon\lambda_d(H)$.

(B) $\|\widetilde{f}_1 - f_1\|_p$
Observe first, that

$$\begin{aligned} |\widetilde{f}_1(x) - f_1(x)| &= \Big|\sum_i c_i\big(\widetilde{\varphi}(x, a_i) - \varphi(x, a_i)\big)\Big| \\ &\leq \sum_i |c_i|\big|\widetilde{\varphi}(x, a_i) - \varphi(x, a_i)\big| \,. \end{aligned}$$

Due to the bounds on φ and $\widetilde{\varphi}$, each of the absolute values in the last sum is at most $2b$ for every $x \in H$. Moreover, each of these absolute values is equal to 0 for most of $x \in H$, namely up to a set of measure $\sqrt{\varepsilon}$ (recall that $a_i \notin U$). Now, we have

$$\begin{aligned} \|\widetilde{f}_1 - f_1\|_p^p &= \int_H |\widetilde{f}_1 - f_1|^p \\ &= \int_H |\widetilde{f}_1 - f_1|^{p-1}|\widetilde{f}_1 - f_1| \\ &\leq \int_H (2b\|w\|_1)^{p-1}|\widetilde{f}_1 - f_1| \\ &\leq (2b\|w\|_1)^{p-1} \int_H \sum_{i=1}^n |c_i|\big|\widetilde{\varphi}(x, a_i) - \varphi(x, a_i)\big| \,\mathrm{d}x \\ &= (2b\|w\|_1)^{p-1} \sum_{i=1}^n |c_i| \int_H \big|\widetilde{\varphi}(x, a_i) - \varphi(x, a_i)\big| \,\mathrm{d}x \end{aligned}$$

According to the previous paragraph, we can bound each of the integrals in the last sum by $2b\sqrt{\varepsilon}$, yielding

$$\begin{aligned} \|\widetilde{f}_1 - f_1\|_p^p &\leq (2b\|w\|_1)^{p-1} \sum_i |c_i| \cdot 2b\sqrt{\varepsilon} \\ &\leq (2b\|w\|_1)^{p-1}\|\widetilde{w}\|_1 \cdot 2b\sqrt{\varepsilon} \\ &\leq (2b\|w\|_1)^{p-1}\|w\|_1 \cdot 2b\sqrt{\varepsilon} \\ &= (2b\|w\|_1)^p\sqrt{\varepsilon} \,. \end{aligned}$$

(C) $\|f - \widetilde{f}\|_p$
Here we proceed similarly as in part (B):

$$\begin{aligned} |f(x) - \widetilde{f}(x)| &= \left|\int_{A\setminus U} (w(a)\varphi(x, a) - \widetilde{w}(a)\widetilde{\varphi}(x, a)) \,\mathrm{d}a + \int_U w(a)\varphi(x, a) \,\mathrm{d}a\right| \\ &\leq \int_{A\setminus U} \big|w(a)\varphi(x, a) - \widetilde{w}(a)\widetilde{\varphi}(x, a)\big| \,\mathrm{d}a + \int_U \big|w(a)\varphi(x, a)\big| \,\mathrm{d}a \end{aligned}$$

We will use that both $|w(a)\varphi(x, a)|$ and $|\widetilde{w}(a)\widetilde{\varphi}(x, a)|$ are at most b^2, the set U is small, and $|w(a)\varphi(x, a) - \widetilde{w}(a)\widetilde{\varphi}(x, a)|$ is "usually" zero. In particular, $|f(x) - \widetilde{f}(x)| \leq 2b^2\lambda_k(A)$.

$$
\begin{aligned}
\|f-\widetilde{f}\|_p^p &= \int_H |f-\widetilde{f}|^p \\
&= \int_H |f-\widetilde{f}|^{p-1}|f-\widetilde{f}| \\
&\le \int_H (2b^2\lambda_k(A))^{p-1}|f-\widetilde{f}| \\
&\le (2b^2\lambda_k(A))^{p-1}\int_H \left(\int_{A\setminus U} |w(a)\varphi(x,a)-\widetilde{w}(a)\widetilde{\varphi}(x,a)|\,\mathrm{d}a \right. \\
&\qquad \left. + \int_U |w(a)\varphi(x,a)|\,\mathrm{d}a\right)
\end{aligned}
$$

Next we use Fubini's theorem – note that we integrate a nonnegative measurable function:

$$
\begin{aligned}
&\le (2b^2\lambda_k(A))^{p-1}\left(\int_{H\times(A\setminus U)} |w(a)\varphi(x,a)-\widetilde{w}(a)\widetilde{\varphi}(x,a)|\,\mathrm{d}(x,a)\right. \\
&\qquad \left. + \int_{H\times U} |w(a)\varphi(x,a)|\,\mathrm{d}(x,a)\right) \\
&\le (2b^2\lambda_k(A))^{p-1}\left(\int_F |w(a)\varphi(x,a)-\widetilde{w}(a)\widetilde{\varphi}(x,a)|\,\mathrm{d}(x,a) + \int_{H\times U} b^2\,\mathrm{d}(x,a)\right) \\
&\le (2b^2\lambda_k(A))^{p-1}\left(\varepsilon 2b^2 + \lambda_d(H)2\sqrt{\varepsilon}\;b^2\right)
\end{aligned}
$$

By combining (A), (B), and (C) we see that we can choose ε small enough to get as good approximation as desired. □

We have proven $\|f\|_{\mathcal{G}} \le \|w\|_1$ for f computed by one-hidden-layer neural network with $\mathcal{L}^\infty$ (almost everywhere bounded) activation function. Together with Theorem 1.6 we derive rates of approximation for this approximation schema:

Corollary 2.4 (Rates in $\mathcal{L}^p$) *Let k, d be positive integers, A a compact subset of $\mathbb{R}^k$ and H a compact subset of $\mathbb{R}^d$. Let $w \in \mathcal{L}^1(A,\lambda_k)$ and $\varphi \in \mathcal{L}^p(H\times A, \lambda_{d+k})$ for some $1<p<\infty$. Additionally let w and φ be bounded almost everywhere on A and $H\times A$ respectively. Let $\mathcal{G}=\{\varphi(\cdot,a) : a\in A\}$ be bounded and $s_{\mathcal{G}}=\sup_{\varphi\in\mathcal{G}}\|\varphi\|_p$. Let f be any function that can be represented as $f(x)=\int_A w(a)\varphi(x,a)\,\mathrm{d}a$. Then*

$$\|f-\operatorname{span}_n \mathcal{G}\|_p \le \frac{2C_p s_{\mathcal{G}}\|w\|_1}{n^{1-1/t}},$$

where $t=\min\{p,2\}$ and $C_p=1$ if $p\le 2$ and $C_p=\sqrt{2}\big(\Gamma(\frac{p+1}{2})/\sqrt{\pi}\big)^{1/p}$ for $p>2$.

As in Theorems 1.5 and 1.6 one could write instead of $\|f-\operatorname{span}_n\mathcal{G}\|_p$ the more accurate expression $\|f-\operatorname{conv}_n c\mathcal{G}\|_p$, with $c=\|w\|_1$.

2.3 Estimates of $\mathcal{G}$-variation via Hahn-Banach Theorem

In this section we provide a generalization (and also an alternative proof) of the result of the previous section. We extend Theorems 2.1 and 2.3 (the estimate of $\mathcal{G}$-variation) to more general Banach spaces in place of $C(K)$, resp. $\mathcal{L}^p\cap\mathcal{L}^\infty$. We also generalize the integral formula to employment of signed measures.

The proof of the main result in this section is in fact shorter than previously presented proofs of weaker results. This is achieved by using more advanced tools from functional analysis. A slight drawback is that this approach (relying on Hahn-Banach theorem) is no longer constructive: given formula $f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a$, the previous proofs suggested a technique to really obtain a sequence of convex combinations that converge to f. The functional-analytic approach, on the other hand, only proves that such a sequence exists. This, however, has no implications for our present considerations; we revisit this issue in [SS08].

The main tool we will use in this section is the following version of geometric Hahn-Banach theorem [Lax02, LuMa95].

Theorem 2.5 (Geometric Hahn-Banach [Lax02]) *Let X be a Banach space, consider $x \in X$ and $T \subseteq X$. Then $x \in \operatorname{cl}\operatorname{conv} T$, unless there is a functional $\ell \in X^*$ and $z \in \mathbb{R}$ such that*

$$\ell(x) > z \quad \text{and} \quad \ell(t) < z \text{ for every } t \in T\,. \tag{2.5}$$

First we present an alternative proof (a generalization) of Theorem 2.1.

Theorem 2.6 ($\mathcal{G}$-variation for continuous activation functions using measure) *Let d, k be positive integers, $H \subseteq \mathbb{R}^d$ and $A \subseteq \mathbb{R}^k$ compact sets. Suppose ν is a signed Radon measure on A. Finally, let $\varphi \in C(H \times A)$ and $\mathcal{G} = \{\pm\varphi_a\}$.*

Let the function $f \in C(H)$ be represented as $f(x) = \int_A \varphi(x,a)\,\mathrm{d}\nu(a)$.

Then $f \in \operatorname{cl}_C \operatorname{conv} c\mathcal{G}$, where $c = \|\nu\|$ is the norm of ν. Using our previous terminology, $\|f\|_{\mathcal{G}} \le \|\nu\|$.

Proof: Let (P, N) be a Hahn decomposition for the measure ν. That is, A is the disjoint union of P and N, and $\nu(E) \ge 0$ (resp. ≤ 0) whenever $E \subseteq P$ (resp. $E \subseteq N$). Define the function s by

$$s(a) = \begin{cases} +1 & \text{for } a \in P \\ -1 & \text{for } a \in N \end{cases}$$

that is, $s(a)$ is the "sign of ν at a". In particular, we have $c = \|\nu\| = \int_A s(a)\,\mathrm{d}\nu(a)$.

If $c = 0$ then $\nu(E) = 0$ for any set E, thus $f(x) \equiv 0$ and the assertion is true. So we may assume $c > 0$; note that $c = \|\nu\| < \infty$, as ν is a signed measure.

We need to prove that $f \in \operatorname{cl}\operatorname{conv} c\mathcal{G}$. Suppose the contrary; according to the geometric Hahn-Banach theorem (Theorem 2.5), there is a constant z, and a functional $\ell \in C(H)^*$ such that (2.5) is true with $x = f$ and $T = c\mathcal{G}$. Let μ be the signed measure defining ℓ.

We have $\ell(f) > z$ and for every a

$$\ell(\pm c\varphi_a) = \pm c \int_H \varphi_a\,\mathrm{d}\mu < z\,. \tag{2.6}$$

By definition,

$$\begin{aligned} \ell(f) &= \int_H f\,\mathrm{d}\mu \\ &= \int_H \int_A \varphi(x,a)\,\mathrm{d}\nu(a)\,\mathrm{d}\mu(x)\,. \end{aligned}$$

Note, that $\varphi(x,a)$ is a continuous function and it is only integrated over a compact set. So, the integral of the absolute value is finite and, obviously, both ν and μ are σ-finite (they are even finite). Thus we can use Fubini's theorem to get

$$\begin{aligned} \ell(f) &= \int_A \int_H \varphi(x,a)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(a) \\ &= \frac{1}{c} \int_A s(a) \int_H s(a) c\varphi(x,a)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(a)\,. \end{aligned}$$

Next, we use (2.6) and the definition of s

$$\begin{aligned}&\leq \frac{1}{c}\int_A s(a)z\,\mathrm{d}\nu(a)\\&= \frac{z}{c}\|\nu\|\\&= z\,.\end{aligned}$$

This contradiction finishes the proof. □

As a corollary, we obtain an alternative proof for Theorem 2.1 (that appears as Corollary 2.3 in [KKK97]). Actually, we obtain a stronger version, as we do not require w to be continuous.

Corollary 2.7 ($\mathcal{G}$-variation for continuous activation functions (weaker assumptions)) *Let d, k be positive integers, $H \subseteq \mathbb{R}^d$ and $A \subseteq \mathbb{R}^k$ compact sets. Let $w \in \mathcal{L}^1(A)$, $\varphi \in C(H \times A)$ and $\mathcal{G} = \{\pm\varphi_a\}$.*

Finally let the $f \in C(H)$ be represented as $f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a$.

Then $f \in \mathrm{cl}_C \operatorname{conv} c\mathcal{G}$, where $c = \int_A |w(a)|\,\mathrm{d}a$. Using our previous terminology, $\|f\|_{\mathcal{G}} \leq \|w\|_1$.

Proof: We define a signed measure ν by letting for any Lebesgue-measurable $E \subseteq A$

$$\nu(E) = \int_E w(a)\,\mathrm{d}a\,.$$

We easily get

$$\|\nu\| = \int_A |w(a)|\,\mathrm{d}a = \|w\|_1$$

and

$$f(x) = \int_A w(a)\varphi(x,a) = \int_A \varphi(x,a)\,\mathrm{d}\nu(a)\,.$$

It only suffices to apply Theorem 2.6 and we conclude. □

Now we present a theorem bounding $\mathcal{G}$-variation for $\mathcal{L}^p$ activation function.

Theorem 2.8 ($\mathcal{G}$-variation for $\mathcal{L}^p$ activation functions using measure) *Let d, k be positive integers, let $p \in (1,\infty)$. Consider sets $H \subseteq \mathbb{R}^d$ and $A \subseteq \mathbb{R}^k$, and a signed Radon measure ν on A. Let φ be a measurable function such that there is $b \in \mathbb{R}$ so that for any $a \in A$ the function $\varphi_a = \varphi(\cdot, a)$ is in $\mathcal{L}^p(H, \lambda_d)$, and $\|\varphi_a\|_p \leq b$. Put $\mathcal{G} = \{\pm\varphi_a, a \in A\}$.*

Let the function f be represented as $f(x) = \int_A \varphi(x,a)\,\mathrm{d}\nu(a)$ (that is, the integral exists for almost every x).

Then $f \in \mathrm{cl}_{\mathcal{L}^p} \operatorname{conv} c\mathcal{G}$, where $c = \|\nu\|$ is the norm of ν. Using our previous terminology, $\|f\|_{\mathcal{G}} \leq \|\nu\|$.

Proof: We proceed similarly as in the proof of Theorem 2.6. Again, we use Hahn decomposition of ν to define $s(a)$ as the "sign of ν at a" and put $c = \|\nu\| = \int s(a)\,\mathrm{d}\nu(a)$. We first remark that f is in $\mathcal{L}^p(H, \lambda_d)$: For any linear functional l in $(\mathcal{L}^p)^*$ we will derive in (2.7) that $l(f)$ is finite. This shows, that f is an element in $(\mathcal{L}^p)^{**}$, and as $\mathcal{L}^p$ is reflexive, we see that indeed $f \in \mathcal{L}^p$.

If $f \notin \mathrm{cl}\operatorname{conv} c\mathcal{G}$ then, using Theorem 2.5 again, there is an $\ell \in (\mathcal{L}^p)^*$ such that

$$\ell(f) > z > \ell(\pm c\varphi_a)$$

for some z and all $a \in A$. Let $\psi \in \mathcal{L}^q$ (with $1/p + 1/q = 1$) be the representant of ℓ. Similarly as before, we get

$$\begin{aligned}\ell(f) &= \int_H f\psi\\&= \int_H \int_A \varphi(x,a)\psi(x)\,\mathrm{d}\nu(a)\,\mathrm{d}\lambda(x)\end{aligned}$$

To apply Fubini's theorem, we observe that $\varphi(x,a)\psi(x)$ is a measurable function, and that $\int_H |\varphi(x,a)\psi(x)| \leq \left(\int_H |\varphi(x,a)|^p\right)^{1/p}\left(\int_H |\psi(x)|^q\right)^{1/q} = \|\varphi_a\|_p\|\psi\|_q$ (Hölder inequality for $|\varphi_a|$ and $|\psi|$). Consequently,

$$\int_H \int_A |\varphi\psi| \leq \|\nu\| \cdot b \cdot \|\psi\|_q \tag{2.7}$$

and we may use Fubini's theorem to obtain

$$\begin{aligned}
\ell(f) &= \int_A \int_H \varphi(x,a)\psi(x)\,\mathrm{d}\lambda(x)\,\mathrm{d}\nu(a) \\
&= \frac{1}{c}\int_A s(a) \int_H s(a)c\varphi(x,a)\psi(x)\,\mathrm{d}\lambda(x)\,\mathrm{d}\nu(a) \\
&\leq \frac{1}{c}\int_A s(a)z\,\mathrm{d}\nu(a) \\
&= z
\end{aligned}$$

Again, by Hahn-Banach Theorem we found a contradiction. □

Note, that we get Theorem 2.3 as a corollary. Also we recovered a version of a result of [KKK97] (Theorem 2.2). We do not obtain that f is in the closure in the supremum norm. This, however, is not needed to apply the Maurey-Jones-Barron theorem or any of the other theorems on rates of approximation.

The technique of using Hahn-Banach theorem can be applied to other spaces as well – all we need is to have elements of the dual to that space "behave nicely with respect to integration", that is, some version of Fubini's theorem holds. Natural candidates to consider in this setting would be Sobolev spaces, leading to a simultaneous approximation of a function and its derivatives. We will not elaborate on this topic, as it was already researched by Hornik et al. [HSW89].

Before we end this section, let us remark that we can express the obtained results in terms of functional analysis. Define an operator T_φ by

$$T_\varphi(\nu) = \int_A \varphi(\cdot,a)\,\mathrm{d}\nu(a)\,.$$

We consider T_φ as an operator from $M(A)$ (the space of all signed measures on A) to a subspace of $C(H)$ (or $\mathcal{L}^p(H)$, etc.) with the norm $\|\cdot\|_{\mathcal{G}}$ (the subspace consists of functions of finite $\|\cdot\|_{\mathcal{G}}$-norm). Then the above results say that the operator norm of T_φ is at most (in fact exactly) equal to 1.

2.4 Surprising Property of $\mathcal{G}$-variation

In this short section we are going to prove a surprising property of the $\mathcal{G}_\sigma$-variation, namely its independence of σ for a large class of activation functions.

We will assume σ is a sigmoidal function (that is $\lim_{x\to-\infty}\sigma(x) = 0$, $\lim_{x\to\infty}\sigma(x) = 1$ and σ is nondecreasing). Note that we do not require continuity: after all, from practical perspective, the easiest functions to evaluate are step functions, that is linear combinations of characteristic functions of intervals.

If we consider $\mathcal{G}_\sigma$ as a subset of $\mathcal{L}^p(H)$ (for a compact set H) then the $\mathcal{G}_\sigma$-variation $\|f\|_{\mathcal{G}_\sigma}$ does not depend on σ. A version of this result appears implicitly in [KKK97]. However, there σ was assumed to be either continuous, or the Heaviside function.

Theorem 2.9 ($\mathcal{G}_\sigma$-variation independent of σ) *Suppose $1 < p < \infty$, let $H \subseteq \mathbb{R}^d$ be a compact set and $f \in \mathcal{L}^p(H)$. Then there is $c_f \in [0,\infty]$ so that for any sigmoidal function σ we have*

$$\|f\|_{\mathcal{G}_\sigma} = c_f\,.$$

(Recall that sigmoidal function means a function $\mathbb{R} \to \mathbb{R}$ that is nondecreasing with limits in $\pm\infty$ being 0 and 1; we do not demand continuity.)

Proof: We put $c = \|f\|_{\mathcal{G}_\vartheta}$ and show that for any sigmoidal function σ we have $\|f\|_{\mathcal{G}_\sigma} = c$. To this end, we prove an auxiliary claim first, reducing to a question about functions of one real variable. Then we utilize this claim by letting either σ_1 or σ_2 be the Heaviside function ϑ.

Claim Let σ_1, σ_2 be two sigmoidal functions so that for each finite interval $J \subseteq \mathbb{R}$

$$\sigma_2(t) \in \operatorname{cl}\operatorname{conv}\{\sigma_1(rt+s) : r, s \in \mathbb{R}\}\,,$$

where the closure is taken in $\mathcal{L}^p(J)$. Then for any function $f \in \mathcal{L}^p(H)$ we have $\|f\|_{\mathcal{G}_{\sigma_1}} \le \|f\|_{\mathcal{G}_{\sigma_2}}$.

Indeed, by definition of the $\mathcal{G}$-variation, there are functions $f_{apx}(x)$ that are arbitrarily close to $f(x)$ (in $\mathcal{L}^p(H)$-norm), and that are of form

$$f_{apx} = \sum_i c_i \sigma_2(a_i \cdot x + b_i), \qquad \sum_i |c_i| \le \|f\|_{\mathcal{G}_{\sigma_2}}\,.$$

If the assumptions of the Claim are satisfied, then we can approximate each of $\sigma_2(t)$ by a finite convex combination $g_i(t) = \sum_j k_{i,j}\sigma_1(r_{i,j}t + s_{i,j})$ in $\mathcal{L}^p(J)$ for a finite interval J containing $\cup_i\{a_i \cdot x + b_i : x \in H\}$. If we put $E(t) = |g_i(t) - \sigma_2(t)|^p$, we get

$$\int_H E(a_i \cdot x + b_i)\,\mathrm{d}x \le \frac{B}{|a_i|}\int_J E(t)\,\mathrm{d}t\,.$$

Here B is the upper bound on $\lambda_{d-1}(H_{a_i,c})$, where $H_{a_i,c} = \{x \in H : a_i \cdot x = c\}$ are the sections of H. In particular, B can be chosen as a constant depending only on H. So we can for any given $\varepsilon > 0$ find functions g_i so that $\|g_i(a_i \cdot x + b_i) - \sigma_2(a_i \cdot x + b_i)\|_{\mathcal{L}^p(H)} < \varepsilon$. By triangle inequality it follows that $\|f_{apx}(x) - \sum_i c_i g_i(a_i \cdot x + b_i)\|_{\mathcal{L}^p(H)} < \varepsilon$, too.

Also $\sum_i \sum_j |c_i k_{i,j}| = \sum_i |c_i| \sum_j k_{i,j} = \sum_i |c_i| \le \|f\|_{\mathcal{G}_{\sigma_2}}$, which finishes the proof of the claim.

(A) $\|f\|_{\mathcal{G}_\sigma} \le \|f\|_{\mathcal{G}_\vartheta}$ for any f (This part appears in [KKK97], we repeat the simple argument for reader's convenience.) According to the Claim, we only need to observe, that for any M, $\|\sigma(Nt) - \vartheta(t)\|_{\mathcal{L}^p([-M,M])}$ tends to zero as $N \to \infty$. To observe this, we only need for any $\varepsilon > 0$ choose N so large that $\sigma(\varepsilon \cdot N) > 1 - \varepsilon$ and $\sigma(-\varepsilon \cdot N) < \varepsilon$. Then

$$\begin{aligned}\|\sigma(Nt) - \vartheta(t)\|^p_{\mathcal{L}^p([-M,M])} &\le \int_{[-\varepsilon,\varepsilon]} |\sigma(Nt) - \vartheta(t)|^p + \int_{[-M,-\varepsilon]\cup[\varepsilon,M]} |\sigma(Nt) - \vartheta(t)|^p \\ &\le 2\varepsilon \cdot 1 + 2M \cdot \varepsilon^p\end{aligned}$$

A choice of arbitrarily small ε finishes the proof.

(B) $\|f\|_{\mathcal{G}_\vartheta} \le \|f\|_{\mathcal{G}_\sigma}$ for any f Now we pursue with the more surprising part of the proof. We will actually prove something stronger than required by the Claim. Namely, for any $\varepsilon > 0$ there is a function g of form

$$g(t) = \sum_{i=1}^{k} c_i \vartheta(t - b_i) \tag{2.8}$$

so that $|g(t) - \sigma(t)| \le \varepsilon$ for every $t \in \mathbb{R} \setminus \{b_1, \ldots, b_k\}$. This clearly implies that g and σ are close in $\mathcal{L}^p$ norm on any set of finite measure.

We will construct g by inductively finding points b_i. We will rely heavily on the result from first-year analysis: all points of discontinuity of a nondecreasing function σ are jumps, that is, for any x there exists $\sigma(x_-)$ (the limit from the left) and $\sigma(x_+)$ (the limit from the right).

To start the process, we put $b_0 = -\infty$ and $h_0 = 0$. Now, whenever b_i was defined (and $b_i < \infty$), we put

$$b_{i+1} = \sup\{x \in \mathbb{R} : \sigma(x) \le h_i + \varepsilon\}$$

and

$$h_{i+1} = \sigma((b_{i+1})_+)\,.$$

Now for any $y > b_{i+1}$ we have $\sigma(y) > h_i + \varepsilon$, so in particular $h_{i+1} = \sigma((b_{i+1})_+) \geq h_i + \varepsilon$. This implies our process ends after at most $\lfloor 1/\varepsilon \rfloor + 1$ steps (recall that σ is bounded by 1). When we reach $b_{k+1} = \infty$, we define g by (2.8) with $c_i = h_i - h_{i-1}$.

Next, observe that for any $i = 0, 1, \ldots, k$ we have $\sigma((b_{i+1})_-) \leq h_i + \varepsilon$ (by definition of b_{i+1}. As for $t \in [b_i, b_{i+1})$ the constructed function $g(t)$ is equal to $\sum_{1 \leq j \leq i}(h_j - h_{j-1}) = h_i$, we see that $|g(t) - \sigma(t)| \leq \varepsilon$ unless t is one of the points b_i. Thus we conclude that (B) holds as well and this finishes the proof. □

Let us now comment about implications of the above result. The result applies whenever we want to estimate the rate of convergence in $\mathcal{L}^p$ norm, using results of Maurey, Jones, Barron, and Darken et al. As far as these estimates are concerned, all sigmoidal functions are of equal utility. Let us mention some limitations for practical applications, though:

- In part (A) of the above proof ("any σ is at least as good as the Heaviside function") we need to use large multiplicative coefficients, which is not numerically feasible.
- It says nothing about convergence in the supremum norm. (For supremum norm the analog of Maurey-Jones-Barron theorem is false [DDGS93]. See the discussion preceding Theorem 1.7 for more details.)
- To elaborate further on the previous point, we can extend part (B) of the above proof to get the following simple bound for estimates in the supremum norm: If $\|f - \operatorname{span}_n \mathcal{G}_\sigma\| \leq \varepsilon$ then $\|f - \operatorname{span}_N \mathcal{G}_\vartheta\| \leq 2\varepsilon$ with $N = n/\varepsilon$.
- Theorem 2.9 implies equality of *bounds* on the rate of convergence. It is quite possible, that for problems of practical interest, convergence will be faster (but perhaps not for all activation functions). This question deserves further study.

3 Integral Representations

As we have seen in the previous section we needed integral representation of the function f to estimate its $\mathcal{G}$-variation. Thus a natural question is: when does such a representation exist?

In Section 3.1 we present several specific examples of functions where integral representation is known to exist. In Section 3.2 we discuss relationship between integral representation and neural network with number of units going to infinity.

3.1 Integral Representations for Specific Classes of Functions

In this section we present known integral representations for specific types of function f.

A. Absolutely continuous functions Let us consider one-dimensional functions first. Let f be an absolutely continuous function on $[a, b]$. It is known (see, e.g., Corollary 23.5 of [LuMa95]) that f' exists almost everywhere as a function in $\mathcal{L}^1[a, b]$. Moreover,

$$f(x) = f(a) + \int_a^x f'(t)\,\mathrm{d}t\,.$$

Assume now that $f(a) = 0$ and recall that $\vartheta(x)$ is the Heaviside function ($\vartheta(x) = 1$ if $x \geq 0$, $\vartheta(x) = 0$ otherwise). Then the above formula can be expressed as

$$f(x) = \int_a^b f'(t)\vartheta(x - t)\,\mathrm{d}t\,. \tag{3.1}$$

B. Integral representation of $f(x)$ based on Poisson's theorem / inverse Radon transform To apply the above mentioned types of bounds we need function f expressed in the form of an integral as in (2.4). To this end, the following consequence of Poisson's theorem of potential theory was proved in [KKK97]. (The same result, but only for functions in the Schwartz space, is obtained in [Ito91] using inverse Radon transform [He99]. In [KKV06] a variant of Theorem 3.1 (for functions of *weakly controlled decay*) is proved and in [KKV07] this is utilized to find bounds on $\mathcal{G}$-variation in terms of the Sobolev norm.)

Let D_e be the operator of directional derivative in the direction given by e, that is $D_e f(y) = \lim_{h\to 0} \frac{f(y+h\cdot e)-f(y)}{h}$. For a positive integer d, $D_e^{(d)}$ is d-fold iteration of D_e. Note, that if f is C^d, that is the partial derivatives of order at most d exist and are continuous, then one can use the partial derivatives to express all directional derivatives. Finally, $H_{eb} = \{y \in \mathbb{R}^d : y \cdot e + b = 0\}$.

Theorem 3.1 (Integral representation in $C^d(\mathbb{R}^d)$ [KKK97]) *For every odd positive integer d every compactly supported function $f \in C^d(\mathbb{R}^d)$ can be represented as*

$$f(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \Big(\int_{H_{eb}} D_e^{(d)} f(y)\, \mathrm{d}y \Big) \vartheta(e \cdot x + b)\, \mathrm{d}b\, \mathrm{d}e$$

where $a_d = \frac{(-1)^{(d-1)/2}}{2(2\pi)^{d-1}}$.

Thus from Theorems 1.6, 1.5, and 2.2 it follows that if $f \in C^d(\mathbb{R}^d)$, then it can be approximated efficiently by neural networks with Heaviside activation functions, that is with rate $O(\frac{1}{n^{1-1/p}})$ in the space $\mathcal{L}^p$, $1 < p \leq 2$ and with rate $O(\frac{1}{\sqrt{n}})$ for $p > 2$. We can get the same conclusions with somewhat weaker assumptions. Namely, instead of requiring d continuous derivatives, we only ask for weak derivatives (as members of an $\mathcal{L}_p$ space):

Theorem 3.2 (Integral representation in $W^{d,p}(\Omega)$) *Let d be an odd integer, $p > 1$ and let $\Omega \subseteq \mathbb{R}^d$ be a bounded open set with a C^1 boundary. Then every f in the Sobolev space $\mathcal{W}^{d,p}(\Omega)$ can be represented as*

$$f(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \Big(\int_{H_{eb}} D_e^{(d)} f(y)\, \mathrm{d}y \Big) \vartheta(e \cdot x + b)\, \mathrm{d}b\, \mathrm{d}e$$

where $a_d = \frac{(-1)^{(d-1)/2}}{2(2\pi)^{d-1}}$.

Proof: Let f be a function in $\mathcal{W}^{d,p}$. It is known [Lan93] that we can find functions $f_n \in C^\infty(\Omega)$ such that $\|f_n - f\|_{d,p} < 1/n$. For f_n we know the formula

$$f_n(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \Big(\int_{H_{eb}} D_e^{(d)} f_n(y)\, \mathrm{d}y \Big) \vartheta(e \cdot x + b)\, \mathrm{d}b\, \mathrm{d}e \tag{3.2}$$

from several sources ([He99], [Ito91], Theorem 2.1, [KKV07]). It remains to show, how we can derive the same formula for f itself.

By definition of $\mathcal{W}^{d,p}$-norm we easily conclude that $D_e^{(d)} f_n \to D_e^{(d)} f$ in $\mathcal{L}_p$-sense. Consequently, for any given $\varepsilon > 0$ and every sufficiently large n we have $\|D_e^{(d)} f_n - D_e^{(d)} f\|_p < \varepsilon$. Then we have for each $e \in S^{d-1}$

$$\begin{aligned}
&\Big| \int_{\mathbb{R}} \Big(\int_{H_{eb}} (D_e^{(d)} f_n(y) - D_e^{(d)} f(y))\, \mathrm{d}y \Big) \vartheta(e \cdot x + b)\, \mathrm{d}b \Big| \\
&\leq \int_{\mathbb{R}} \Big(\int_{H_{eb}} |D_e^{(d)} f_n(y) - D_e^{(d)} f(y))|\, \mathrm{d}y \Big) \vartheta(e \cdot x + b)\, \mathrm{d}b \\
&\leq \int_{\Omega} |D_e^{(d)} f_n(y) - D_e^{(d)} f(y))|\, \mathrm{d}y
\end{aligned}$$

and by the power mean inequality we get that for some C depending only on the measure of Ω

$$\begin{aligned}&\leq C\|D_e^{(d)}f_n - D_e^{(d)}f\|_p\\&\leq C\varepsilon\,.\end{aligned}$$

Consequently, the right-hand side of (3.2) for f and for f_n differ by at most $a_d\lambda_{d-1}(S^{d-1})C\varepsilon$. The difference of the left-hand sides of (3.2) can be estimated using the Sobolev inequality:

$$\|f_n - f\|_{C(\Omega)} \leq C'\|f_n - f\|_{d,p} \leq c_1\varepsilon\,.$$

Here c_1 depends only on d, p, and Ω. It follows, that there is a constant $c_2 > 0$ such that for each $\varepsilon > 0$ the representation (3.2) holds for f with the error at most $c_2\varepsilon$. Letting $\varepsilon > 0$ go to 0 finishes the proof. □

C. Wavelets For set $\mathcal{G}$ obtained from functions of RBF type (2.2), the theory of wavelets is of use. The basic result there is the following. Let σ be an $\mathcal{L}^2$ function with $\|\sigma\|_2 = 1$, such that $\int \frac{|\hat{\sigma}(a)|^2}{|a|}\,\mathrm{d}a$ is finite (such σ is called a wavelet). Under suitable conditions (which we will not describe here in detail) one has

$$f = \int w_{a,b}\sigma\left(\frac{x-b}{a}\right)\mathrm{d}(a,b)\,,$$

where $w_{a,b}$ are suitable "weights". For more details, any book about wavelets, e.g. [Bl98] can be of use.

D. Integral representation of $f(x)$ based on Fourier transform Another approach to bounds on $\|f\|_{\mathcal{G}}$ (although without this notation) is due to Barron [Ba93]. Let $B \subseteq \mathbb{R}^d$ be bounded. Let $\Omega_{B,\rho}$ be the set of all functions $f : B \to \mathbb{R}$ such that

1. For some complex-valued measure $\hat{F}(\,\mathrm{d}\omega)$ and for any $x \in B$

$$f(x) = f(0) + \int (e^{i\omega\cdot x} - 1)\hat{F}(\,\mathrm{d}\omega)\,.$$

2. We have $\int |\omega|_B F(\,\mathrm{d}\omega) \leq \rho$. Here $F(\,\mathrm{d}\omega)$ denotes the magnitude distribution of $\hat{F}(\,\mathrm{d}\omega)$ from part 1, and $|\omega|_B = \sup_{x\in B}|x\cdot\omega|$.

Examples of such functions include functions f for which the Fourier transform $\hat{f}$ exists, the inverse Fourier transform produces f, and $\omega\hat{f}(\omega)$ is integrable. Many more examples (positive definite functions, functions in C^s where $s = \lfloor d/2\rfloor + 2$, etc.) are listed in [Ba93].

Theorem 3.3 (Integral representation based on Fourier transform [Ba93]) *Let σ be any sigmoidal function, let $f \in \Omega_{B,\rho}$. Then $\|f(x) - f(0)\|_{\mathcal{G}_\sigma} \leq \rho$. Consequently, $f(x) - f(0)$ can be approximated well by elements in $\mathcal{G}_\sigma$:*

$$\|(f(x) - f(0)) - \mathrm{conv}_n\,\rho\mathcal{G}_\sigma\|_2 \leq \frac{\rho}{\sqrt{n}}\,.$$

To compare with Theorem 2.2, this method is more widely applicable; it does not yield an explicit formula for $f(x)$, though.

3.2 Networks with Continuum Many Units

The term *neural network with continuum many units* was metaphorically used in [KKK97] to describe a function in integral representation

$$f(x) = \int_A w(a)\varphi(x,a)\,\mathrm{d}a \tag{3.3}$$

where it is to be understood that for every a we have a function $\varphi(\cdot,a)$ as the activation function of one neuron; we take this neuron with weight $w(a)$. This concept enabled an interesting application of results of Section 1.1. It is not clear, however, what is the relation between the class of functions representable as (3.3), functions realizable by finite neural networks, that is expressible as

$$f(x) = \sum_{i=1}^{n} c_i\varphi(x,a_i) \tag{3.4}$$

and functions that can be approximated by finite networks.

In this section we will try to clarify these relationships. To this end, we extend the notion of neural network with continuum many neurons even further. For any signed measure ν on A, we consider the function

$$f(x) = \int_A \varphi(x,a)\,\mathrm{d}\nu(a)\,. \tag{3.5}$$

Any function representable by (3.3) can be represented as (3.5), when ν has density $w(a)$.

We recall (and introduce) some notation. We have a continuous function φ on $H \times A$. As in previous sections, we put $\mathcal{G} = \{\pm\varphi(\cdot,a), a \in A\}$. In this notation, finite neural networks compute functions in $\operatorname{span}\mathcal{G}$. As we want to restrict the size of the weights towards the output neuron, it makes more sense to consider functions in $\operatorname{conv} c\mathcal{G}$ for some real c. Functions that can be approximated by such bounded finite networks are those in $\operatorname{cl}\operatorname{conv} c\mathcal{G}$. Finally, we let $I(c,\mathcal{G})$ denote the set of functions f that can be represented as (3.5) for some measure ν on A with $\|\nu\| \le c$.

It is obvious that $\operatorname{conv}\mathcal{G} \subseteq \operatorname{cl}\operatorname{conv}\mathcal{G}$ and $\operatorname{conv} c\mathcal{G} \subseteq I(c,\mathcal{G})$. Less obvious is the relationship between $\operatorname{cl}\operatorname{conv}\mathcal{G}$ and $I(c,\mathcal{G})$:

Theorem 3.4 (Sum $\Rightarrow$ Integral) *Let $\varphi \in C(H \times A)$, H and A compact subsets of $\mathbb{R}^d$ and $\mathbb{R}^k$, respectively. Then for every real c*

$$\operatorname{cl}\operatorname{conv} c\mathcal{G} \subseteq I(c,\mathcal{G})\,.$$

Explicitly, every function that can be approximated by functions of form $\sum_{i=1}^m c_i\varphi(x,a_i)$ for $\sum_{i=1}^m |c_i| \le c$ can be expressed as $f(x) = \int \varphi(x,a)\,\mathrm{d}\nu(a)$ for some signed measure ν of norm at most c.

Proof: Let f be a function in $\operatorname{cl}\operatorname{conv} c\mathcal{G}$, and choose a sequence f_n converging to f. We can write $f_n = \sum_{i=1}^{m_n} c_{n,i}\varphi(a_{n,i},x)$, where $\sum_i |c_{n,i}| \le c$. We let ν_n be the weighted counting measure, that is for any set $E \subseteq A$ we put

$$\nu_n(E) = \sum_{i:a_{n,i}\in E} c_{n,i}\,.$$

Recall that the space $M(A)$ of signed measures on A is the dual to $C(A)$. As $C(A)$ is separable, Helly's theorem implies that the ball of radius c in $M(A)$ is w^*-sequentially-compact. This in particular implies that there is a measure ν and a subsequence ν_{n_k} converging to ν in the w^* topology (as ν is w^*-limit of measures with norms at most c its norm is at most c as well). This in turn means that for every function $g \in C(A)$ we have $\int g\,\mathrm{d}\nu_{n_k} \to \int g\,\mathrm{d}\nu$. We apply this for $g = \varphi(\cdot,a)$ for every $x \in H$. We obtain

$$f_{n_k}(x) = \int_A \varphi(x,a)\,\mathrm{d}\nu_{n_k}(a) \to \int_A \varphi(x,a)\,\mathrm{d}\nu(a)\,.$$

As $\lim_n f_n(x) = f(x)$ by our choice of f_n, this finishes the proof. □

In Theorem 2.6 we showed the converse to the above theorem: if a function $f(x)$ is in form (3.5) then it is a limit of functions of form (3.4). This tells us that certain functions can be approximated well. Moreover, combination of Theorem 2.6 and 3.4 concludes our intention to compare these two ways to extend the notion of neural networks to infinity.

4 Applications

In this section we combine results regarding rates of approximation (Section 1), $\mathcal{G}$-variation (Section 2) and integral representation (Section 3) to derive practically applicable results.

We start with a result that appears already in [KKK97].

Corollary 4.1 (Approximation for $C^d(\mathbb{R}^d)$ functions [KKK97]) *Let d be an odd positive integer and $f \in C^d(\mathbb{R}^d)$ a compactly supported function. Let σ be a continuous sigmoidal function. Then there is a constant C so that*

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_2 \leq \frac{C}{\sqrt{n}}\,.$$

Corollary 4.2 (Approximation for $C^d(\mathbb{R}^d)$ functions in $\mathcal{L}_p$, gen. sigm. function) *Let $1 < p < \infty$, let d be an odd positive integer and $f \in C^d(\mathbb{R}^d)$ a compactly supported function. Let σ be a nondecreasing sigmoidal function (not necessarily continuous). Then there is a constant C so that*

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_p \leq \frac{C}{n^{1-1/t}}\,,$$

where $t = \min\{p, 2\}$.

Proof: By Theorem 3.1 we have integral representation of f using Heaviside functions:

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

Thus by Theorem 2.2 we obtain bounded $\mathcal{G}_\vartheta$-variation, bound given by integral of directional derivatives:

$$\mathcal{G}_\vartheta \leq \int_{S^{d-1}} \int_{\mathbb{R}} a_d \Big| \int_{H_{eb}} D_e^{(d)} f(y)\,\mathrm{d}y \Big|\,\mathrm{d}b\,\mathrm{d}e).$$

Using Theorem 2.9 we observe that $\mathcal{G}_\vartheta(f) = \mathcal{G}_\sigma(f)$ for any sigmoidal activation function σ. Now it remains to use Theorem 1.6, we observe that $\mathcal{G}_\sigma$ is $\mathcal{L}^p$-bounded on support of f and having shown that $\mathcal{G}_\sigma(f)$ is finite we conclude the proof. □

Corollary 4.3 (Approximation for $\mathcal{W}^{d,p}(\mathbb{R}^d)$ functions in $\mathcal{L}_p$, gen. sigm. function) *Let $1 < p < \infty$, let d be an odd positive integer, let $\Omega \subseteq \mathbb{R}^d$ be a bounded open set with a C^1 boundary and consider an $f \in \mathcal{W}^{d,p}(\Omega)$. Let σ be a nondecreasing sigmoidal function (not necessarily continuous). Then there is a constant C so that*

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_p \leq \frac{C}{n^{1-1/t}}\,,$$

where $t = \min\{p, 2\}$.

Proof: By Theorem 3.2 we have integral representation of f using Heaviside functions:

$$f(x) = -a_d \int_{S^{d-1}} \int_{\mathbb{R}} \Big(\int_{H_{eb}} D_e^{(d)} f(y)\,\mathrm{d}y \Big) \vartheta(e \cdot x + b)\,\mathrm{d}b\,\mathrm{d}e$$

(the derivatives are taken in the weak sense). Thus by Theorem 2.8 we find that the $\mathcal{G}_\vartheta$-variation is bounded. The bound is given by integral of directional derivatives:

$$\mathcal{G}_\vartheta \leq \int_{S^{d-1}} \int_{\mathbb{R}} a_d \Big| \int_{H_{eb}} D_e^{(d)} f(y)\,\mathrm{d}y \Big|\,\mathrm{d}b\,\mathrm{d}e\,.$$

In [KKV06] this computation is carried on to provide an upper bound on $\mathcal{G}_\vartheta$ variance in terms of Sobolev norm (even Sobolev seminorm) $\|\cdot\|_{d,1}$. As Ω is of finite measure, this implies a bound

$\|f\|_{\mathcal{G}_\vartheta} = O(\|\cdot\|_{d,p})$. Using Theorem 2.9 we observe that $\mathcal{G}_\vartheta(f) = \mathcal{G}_\sigma(f)$ for any sigmoidal activation function σ. Now it remains to use Theorem 1.6, we observe that $\mathcal{G}_\sigma$ is $\mathcal{L}^p$-bounded on support of f and having shown that $\mathcal{G}_\sigma(f)$ is finite we conclude the proof. □

By using Theorem 2.9 instead of results in [KKK97], we can weaken the assumption on σ – we do not need σ continuous, it is enough, if σ is nondecreasing and bounded. More disagreeable, though, are the assumptions required on f, which are perhaps too strong for applications. We mostly care about rather large d, so we need f to be very smooth. Next, we will discuss the possibilities to weaken this requirement. First, we will see that for $d = 1$ such weakening is possible. (Similar result for σ being the Heaviside function is suggested in [KKV06].)

Theorem 4.4 (Rates for absolutely continuous functions) *Let f be an absolutely continuous function on $[a, b]$. Let σ be any sigmoidal function (not necessarily continuous). Then there is a constant C so that*

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_p \leq \frac{C}{n^{1-1/p}}\,.$$

Proof: We represent $f(x)$ in form (3.1) (Section 3.1, part A). Theorem 2.8 implies that $\|f\|_{\mathcal{G}_\vartheta} \leq \|f'\|_1$ (which we know is finite). From Theorem 2.9 we know that $\|f\|_{\mathcal{G}_\sigma} = \|f\|_{\mathcal{G}_\vartheta}$, so it remains to use Theorem 1.6. □

We see that we lowered the smoothness assumption – we require f to be absolutely continuous, instead of being C^1. However, it is possible to weaken the assumptions on f even further and at the same time improve the approximation, at least in the one-dimensional case. (This result may be known in the analysis community, we have been unable to find it in the literature, though.)

Theorem 4.5 (Rates for bounded variation functions) *Let f be a bounded variation function on $[a, b]$. Then*

$$\|f - \operatorname{span}_n \mathcal{G}_\vartheta\|_\infty \leq \frac{\|f\|_{BV[a,b]}}{n-1}\,.$$

If σ is any sigmoidal function (not necessarily continuous) then we have for any $p \in (1, \infty)$ and a constant $c = c(a, b, p)$

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_p \leq \frac{c\|f\|_{BV[a,b]}}{n-1}\,.$$

Proof: It is known from calculus (see, e.g., Theorem 1.2 in Section X.1 of [Lan93]) that a bounded variation function can be expressed as a difference of two nondecreasing functions, $f = f_1 - f_2$ in such a way, that $\|f\|_{BV[a,b]} = d_1 + d_2$, where $d_i = f_i(b) - f_i(a)$. Using the technique in the proof of Theorem 2.9 (part (B)) we approximate $f_i(x) - f_i(a)$ by a function $g_i(x)$, which is a linear combination of n_i shifts of the Heaviside function ϑ, so that $n_i \leq \left\lfloor \frac{d_i}{\varepsilon} \right\rfloor$ and for all but finitely many values of x we have $0 \leq \big(f_i(x) - f_i(a)\big) - g_i(x) \leq \varepsilon$. Consequently,

$$\big|\big(g_1(x) - g_2(x) + f(a)\big) - f(x)\big| \leq \varepsilon$$

for all but finitely many values of x. We may realize addition of $f(a)$ as one extra Heaviside function, so we found an approximation using $n_1 + n_2 + 1 \leq \frac{d_1+d_2}{\varepsilon} + 1$ Heaviside functions and achieved an $\mathcal{L}^\infty$ error ε.

The second assertion follows immediately by approximating $\vartheta(t)$ by $\sigma(Nt)$ for N large enough. □

We remark that a weaker version of the above theorem (with the usual rate of convergence $O(1/n^{1-1/p})$ in $\mathcal{L}^p$-norm) could be proved also using Theorem 2.8: if f is a bounded variation function on $[a, b]$ and μ_f the corresponding Riemann-Stieltjes measure, then we have the following formula (Proposition 1.8 in Section X.1 of [Lan93])

$$f(x) - f(a) = \int_a^x 1 \,\mathrm{d}\mu_f\,,$$

whenever f is continuous at both a and x. As bounded variation function is continuous at all but finitely many points, we can indeed apply Theorem 2.8.

Next, we will consider the case of larger d.

In Theorem 3.2 we decreased the differentiability requirement – instead of existence of continuous d-fold derivatives as in 4.1 and 4.2 we only require that d weak derivatives exist (and are bounded in the $\mathcal{L}_p$ norm). This may not seem as a tremendous improvement. On the other hand, in this setting we have the following result that presents a limit on how much can we weaken the assumptions on the function to be approximated.

Theorem 4.6 (Good rates $\implies$ many weak derivatives) *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function. Suppose that for each function $f \in \mathcal{W}^{m,2}(B^d)$ (where B^d is the unit ball in $\mathbb{R}^d$) there is a constant C so that*

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_2 \leq \frac{C}{\sqrt{n}}\,.$$

Then $m \geq (d-1)/2$.

Proof: By Theorem 4.7 of Maiorov [Ma99] there is a function $f \in \mathcal{W}^{m,2}(B^d)$ such that

$$\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_2 \geq C' n^{-m/(d-1)}\,.$$

So we have $\frac{C}{\sqrt{n}} \geq C' n^{-m/(d-1)}$. Considering the limit as $n \to \infty$ finishes the proof. □

For convenience of the reader we give full version of Mairov's theorem we used here:

Theorem 4.7 (Lower bound on rates of approximation for perceptron [Ma99]) *Let $m \geq 1$ and $d \geq 2$. Then for each n there exists an $f \in \mathcal{W}^{m,2}(B^d)$; $\|f\|_{m,2} \leq 1$ for which*

$$\inf_{g \in \mathcal{R}_n} \|f - g\|_2 \geq C n^{-m/(d-1)}.$$

Here the positive constant C is independent of f and n, B^d denotes the unit ball in $\mathbb{R}^d$.

We finish this discussion by mentioning the connection with Theorem 4.8:

Theorem 4.8 (Upper bound on rates of approximation for perceptron [Mh96]) *Let I be an open interval. Assume $\sigma : \mathbb{R} \to \mathbb{R}$ is such that $\sigma \in C^\infty(I)$ and σ is not a polynomial on I. Then for each $p \in [1, \infty]$, $m \geq 1$ and $d \geq 2$*

$$\sup_{f \in \mathcal{W}^{m,p}(B^n); \|f\|_{m,p} \leq 1} \quad \inf_{g \in \mathcal{M}_n(\sigma)} \|f - g\|_p \leq C n^{-m/d},$$

for some constant C independent of n.

Indeed, this theorem actually gives better bounds on $\|f - \operatorname{span}_n \mathcal{G}_\sigma\|_p$ than the results of this chapter. The drawback, however, is that we need to use linear combinations with unbounded coefficients. (The inspection of the proof, as presented in [Pi99] shows that, indeed, unbounded coefficients are crucial for the proof.) This renders the result useless for practical applications: we can find good approximation of f in form

$$\sum_{i=1}^{n} c_i \sigma(a_i \cdot x + b_i) \tag{4.1}$$

for small n, but at the expense of using large coefficients c_i. Consequently, we need to do the computations with a high precision – which only shows that n is not an appropriate measure of complexity of the expression (4.1). This problem is partially avoided by using convex combinations (or, rather, combinations with bounded sum of the coefficients). However, a detailed study of the numerical issues involved remains to be done.

Similar results as in Theorem 4.4 can be easily derived for wavelets and Baron's representation – paragraphs C and D in Section 3.1.

5 Conclusion

In this chapter we studied properties of approximations of functions using convex combinations. Results of Maurey, Jones and Barron and of Darken et al. show that, when applicable, such convex combinations yield good rates of approximation (independent of input dimension).

Further study of constants that appear in these rates bring the notion of $\mathcal{G}$-variation (as defined in [Ku97]). To maintain the mentioned rates when approximating a function f by functions from $\mathcal{G}$ we have seen that fininte $\mathcal{G}$-variation of f is needed. Pursuing this idea Krkov [Ku97] shows that for continuous approximating functions in $\mathcal{G}$ for f representable as integral of these functions weighted by a continuous function $\mathcal{G}$-variation is finite. She proves this result also for Heaviside activation functions. We extend these results to $\mathcal{L}^p$ almost everywhere bounded activation functions and weights by a constructive proof (Theorem 2.3) and nonconstructively using Hahn-Banach Theorem to continuous or $\mathcal{L}^p$ activation functions and weights represented by any signed measure (Theorems 2.6 and 2.8). We further investigate the notion of $\mathcal{G}$-variation and show that for f with infinite $\mathcal{G}$-variation we can have arbitrarily slow convergence of approximation (Theorem 1.7). A surprising result comes from Theorem 2.9 - we show that the presented rates of approximation cannot distinguish between sigmoidal functions.

As mentioned above, all the presented results require f to representable as an integral. In Section 3 we overview known results towards this direction and also show that integral representation is a necessary condition for f to be approximable with good rates of approximation by convex sums of continuous activation functions (Theorem 3.4).

In Section 4 we combine the above mentioned and present a few instances of theoretical bounds on rates of approximation for specific functions f showing how to easily derive corollaries of the type using results of previous sections. One more interesting and less obvious result of this section is the less optimistic information presented in Theorem 4.6 – if we wish to have good rates of approximation for function f we have to demand it to have many weak derivatives.

Bibliography

[Ba93] A. R. Barron: *Universal approximation bounds for superposition of a sigmoidal function*, IEEE Transactions on Information Theory, **39** (1993), 930–945.

[Ba92] A. R. Barron: *Neural net approximation*, Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems, (K. S. Narendra, Ed.), (1992), 69–72.

[Bl98] Ch. Blatter: *Wavelets: a primer*, A K Peters, (1998).

[DDGS93] C. Darken, M. Donahue, L. Gurvits, E. Sontag: *Rate of Approximation Results Motivated by Robust Neural Network Learning*, Proceedings of the 6th Annual ACM Conference on Computational Learning Theory, Santa Cruz, CA, (1993), 303–309.

[HaBu88] S.J. Hanson, D.J. Burr: *Mikowski-r back-propagation: learning in connectionist models with non-Euclidean error signals*, Neural Information Processing Systems, New York: American Institute of Physics, (1988), p. 348.

[He99] S. Helgason: *The Radon Transform*, Progrees in Mathematics, Birkhauser, (1999).

[HSW89] K. Hornik, M. Stinchcombe, H. White: *Multilayer feedforward networks are universal approximators*, Neural Networks **2** (1989), 359–366.

[Hu85] P.J. Huber: *Projection Pursuit*, Annals of Statistics, **13** (1985), 435–475.

[Ito91] Y. Ito: *Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory*, Neural Networks **4** (1991), no. 3, 385–394.

[Jo92] L. K. Jones: *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), no. 1., 608–613.

[Jo87] L. K. Jones: *On a conjecture of Huber concerning the convergence of projection pursuit regression*, Annals of Statistics **15** (1987), 880-882.

[Ku03] V. Kůrková: *High-dimensional approximation and optimization by neural networks*, Chapter 4 in Advances in Learning Theory: Methods, Models and Applications, (Eds. J. Suykens et al.), IOS Press, Amsterdam, (2003), 69–88.

[Ku97] V. Krkov: *Dimension-independent rates of approximation by neural networks*, in Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, (K. Warwick and M. Krn, eds.), Birkhäuser, Boston, (1997), pp. 261–270.

[KHS98] V. Kůrková, K. Hlaváčková, P. Savický: *Representations and rates of approximation of real-valued Boolean functions by neural networks*, Neural Networks **11** (1998), no. 4, 651–659.

[KKK97] V. Kůrková, P.C. Kainen, V. Kreinovich: *Estimates of the Number of Hidden Units and Variation with Respect to Half-Spaces*, Neural Networks, **10** (1997), 1061–1068.

[KKV07] V. Kůrková, P.C. Kainen, A. Vogt: *A Sobolev-Type Upper Bound for Rates of Approximation by Linear Combinations of Heaviside Plane Waves*, Journal of Approximation Theory, **147** (2007), no. 1, 1–10.

[KKV06] V. Kůrková, P.C. Kainen, A. Vogt: *Integral combinations of Heavisides*, ICS AS CR, Technical Report, 968, (2006).

[Lan93] S. Lang: *Real and Functional Analysis*, Springer, 1993.

[Lax02] P. D. Lax: *Functional Analysis*, J. Wiley, (2002).

[LuMa95] J. Lukeš,J. Malý: *Measure and Integral*, Matfyzpress, Praha, (1995).

[Ma99] V. E. Maiorov: *On best approximation by ridge functions*, Journal of Approximation Theory, **99** (1999), no. 1, 68–94.

[Mk96] Y. Makovoz: *Random approximants and neural networks*, Journal of Approximation Theory **85** (1996), 98–109.

[Mh96] H. N. Mhaskar, *Neural networks for optimal approximation of smooth and analytic functions*, Neural Computation, **8** (1996), 164–177.

[Pi99] A. Pinkus: *Approximation theory of the MLP model in neural networks*, Acta Numerica (1999), 143–195.

[Ps81] G. Pisier: *Remarques sur un resultat non publi'e de B. Maurey*, in Seminaire D'Analyse Fonctionnelle, 1980-1981, 'Ecole Polytechnique, Centre de Math'ematiques, Palaiseau, France (1981).

[Re83] W.J. Rey: *Introduction to Robust and Quasi-Robust Statistical Methods*, Springer-Verlag, Berlin, (1983).

[S03a] T. Šidlofová: *Bounds on Rates af Approximation by Neural Networks in Lp-spaces*, Artificial Neural Nets and Genetic Algorithms, SpringerVerlag, (2003), 23–27.

[S03b] T. Šidlofová: *Estimates of Rates of Approximation by Neural Networks in Lp-Spaces*, Kognicia, umel ivot a potaov inteligencia, ELFA, (2003), 365–368.

[SS08] T. Šámalová, R. Šámal: *Pruning algorithms for one-hidden-layer feedforward neural networks* ICS AS CR, Technical Report, V-1022, (2008).

[Zi89] W.P. Ziemer: *Weakly Differentiable Functions*, Springer, (1989).