Dimension-Independent Rates of Approximation by Neural Networks and Variation with Respect to Half-spaces

Kůrková, Věra
1995

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Dimension-independent rates of approximation by neural networks and variation with respect to half-spaces

Věra Kůrková, Paul C. Kainen, Vladik Kreinovich

Technical report No. 626

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+422) 66414244 fax: (+422) 8585789
e-mail: vera@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Dimension-independent rates of approximation by neural networks and variation with respect to half-spaces

Věra Kůrková[1], Paul C. Kainen[2], Vladik Kreinovich[3]

Technical report No. 626

Abstract

For f a function from $\mathcal{R}^d$ to $\mathcal{R}$, we prove that variation with respect to half-spaces for f on a box J is at most $\lambda(J) \cdot \sup \left\{ \left| \int_{J \cap H_{\mathbf{e}b}} (\bigtriangledown f(\mathbf{y}) \cdot \mathbf{e})\, d\mathbf{y} \right| ; \mathbf{e} \in \mathcal{S}^{d-1}, b \in \mathcal{R} \right\}$, where $H_{\mathbf{e}b}$ is the hyperplane orthogonal to $\mathbf{e}$ with offset b and $\lambda(J)$ denotes the Lebesgue measure of J. As a result we obtain conditions on when functions can be approximated within $\mathcal{O}(\frac{1}{\sqrt{n}})$ by one-hidden-layer feedforward neural networks with n hidden units. Our bound on variation is derived from an integral representation theorem based on the fact that the first distributional derivative of the Heaviside function is the delta distribution.

Keywords

One-hidden-layer feedforward neural network, Heaviside activation function, rate of approximation, estimate of the number of hidden units, variation with respect to half-spaces

[1]V. K. was partially supported by GACR Grant 201/93/0427.

[2]Industrial Math, 3044 N St., N.W., Washington, D.C. 20007

[3]Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968

V.K. was partially supported by NSF Grant No. CDA-9015006 and NASA Research Grant No. NAG 9-757.

1 Introduction

Approximating functions from $\mathcal{R}^d$ to $\mathcal{R}^m$ by feedforward neural networks has been widely studied in recent years, and the existence of an arbitrarily close approximation, for any continuous or $\mathcal{L}_p$ function defined on a d-dimensional box, has been proven for one-hidden-layer networks with perceptron or radial-basis-function units with quite general activation functions (see, e.g. Mhaskar and Micchelli [8], Park and Sandberg [9]).

However, estimates of the number of hidden units that guarantee a given accuracy of an approximation are less understood. Most upper estimates grow exponentially with the number of input units, i.e. with the number d of input variables of the function f to be approximated (e.g., Mhaskar and Micchelli [8], Kůrková [6]). A general result by deVore et al. [4] confirms that there is no hope for a better estimate when the class of multivariable functions being approximated is defined in terms of the bounds of partial derivatives. But in applications, functions of hundreds of variables are approximated sufficiently well by neural networks with only moderately many hidden units (e.g., Sejnowski and Yuhas [12]).

Jones [5] introduced a recursive construction of approximants with "dimension-independent" rates of convergence to elements in convex closures of bounded subsets of a Hilbert space and together with Barron proposed to apply it to the space of functions achievable by a one-hidden-layer neural network. Applying Jones' estimate Barron [1] showed that it is possible to approximate any function satisfying a certain condition on its Fourier transform within $\mathcal{L}_2$ error of $\mathcal{O}(\frac{1}{\sqrt{n}})$ by a network whose hidden layer contains n perceptrons with a sigmoidal activation function.

Using a probabilistic argument Barron [2] extended Jones' estimate also to supremum norm. His estimate holds for functions in the convex uniform closure of the set of characteristic functions of half-spaces multiplied by a real number less than or equal to B. He called the infimum of such B the *variation with respect to half-spaces* and noted that it could be defined for any class of characteristic functions.

In this paper, we show that variation with respect to half-spaces is bounded above by the sup of integrals of absolute values of directional derivatives multiplied by the d-dimensional volume of the box where the function is defined (Theorem 2.2.). Consequently, we obtain conditions which guarantee approximations for both $\mathcal{L}_2$ and supremum norm with error rate at most $\mathcal{O}(\frac{1}{\sqrt{n}})$ by one-hidden-layer neural networks. We utilize an integral representation theorem (Theorem 2.1) proved using properties of the Heaviside and delta distributions. Precise definitions and statements of results follow, with proofs deferred to the last section.

2 Variation with respect to half-spaces

Let J be any box in $\mathcal{R}^d$. For a function $\psi : \mathcal{R} \to \mathcal{R}$ let us denote $\mathcal{E}_d(\psi, B, J) = \{g : J \to \mathcal{R};\ g(\mathbf{x}) = w\psi(\mathbf{v} \cdot \mathbf{x} + b), \mathbf{v} \in \mathcal{R}^d, w, b \in \mathcal{R}, |w| \leq B\}$. So $\mathcal{E}_d(\psi, B, J)$ denotes the set of functions computable by a network with d inputs, one hidden perceptron with an activation function ψ and one linear output unit.

Let ϑ denote the Heaviside function ($\vartheta(x) = 0$ for $x < 0$ and $\vartheta(x) = 1$ for $x \geq 0$). It is easy to see that $\mathcal{E}_d(\vartheta, B, J) = \{g : J \to \mathcal{R};\ g(\mathbf{x}) = w\vartheta(\mathbf{e} \cdot \mathbf{x} + b), \mathbf{e} \in S^{d-1}, w, b \in \mathcal{R}, |w| \leq B\}$, where S^{d-1} denotes the unit sphere in $\mathcal{R}^d$. Let $\|.\|_2$ denote the $\mathcal{L}_2$-norm and $\|.\|_{sup}$ the supremum norm with $\mathcal{L}_2$ and $\mathcal{C}$, respectively, the induced topologies. For a subset X of a set of functions $\mathcal{S}$ from J to $\mathcal{R}$ and topology τ on $\mathcal{S}$ we write $cl_\tau(X)$ for the closure of X with respect to the topology τ.

Let $\mathcal{S}$ be a set of functions from J to $\mathcal{R}$ containing $\mathcal{E}_d(\vartheta, B, J)$ and τ be a topology on $\mathcal{S}$. If $f \in \mathcal{S}$, put

$$V(f, \tau, J) = \inf\{B \in \mathcal{R}; f \in cl_\tau(conv(\mathcal{E}_d(\vartheta, B, J)))\}$$

and call $V(f, \tau, J)$ the *variation of f on J with respect to half-spaces and topology τ*.

Since for every $X \subseteq \mathcal{L}_2(J)$ we have $cl_{\mathcal{C}}(X) \subseteq cl_{\mathcal{L}_2}(X)$ surely $V(f, \mathcal{L}_2, J) \leq V(f, \mathcal{C}, J)$. Also, it is easy to verify that for every $f, g \in \mathcal{L}_2(J)$, $V(f + g, \mathcal{C}, J) \leq V(f, \mathcal{C}, J) + V(g, \mathcal{C}, J)$ and for every $a \in \mathcal{R}$, $V(af, \mathcal{C}, J) = |a|V(f, \mathcal{C}, J)$; and similarly for $V(f, \mathcal{L}_2, J)$.

Recall that for a function $f : \mathcal{R} \to \mathcal{R}$ and an interval $[x, y] \subset \mathcal{R}$ *total variation* of f on $[x, y]$ denoted by $T(f, [x, y])$ is defined by $T(f, [x, y]) = \sup\{\sum_{i=1}^{k} |f(x_{i+1}) - f(x_i)|; x = x_1 < \ldots < x_k = y, k \in \mathcal{N}\}$ (see e.g. [7]). For functions of one variable, the concept of variation with respect to half-spaces and the topology of uniform convergence coincides with the concept of total variation since $T(f, J) = V(f, \mathcal{C}, J)$ (see [2], also Darken et al. ([3, Theorem 6]).

When generalizing to functions of several variables, there is no unique way to extend the notion of total variation since we lose the linear ordering property. One well-known method divides d-dimensional cubes into boxes with faces parallel to the coordinate hyperplanes. One defines $T(f, J) = \sup\{\sum_{i=1}^{k} |f(J_i)|$, where $\{J_i; i = 1, \ldots, k\}$ is a subdivision of J into boxes $\}$, $f(J_i) = \sum_{j=1}^{2^d} (-1)^{\nu(j)} f(\mathbf{x}_{ij})$, $\{\mathbf{x}_{ij}; j = 1, \ldots, 2^d\}$ are the corner points of J_i and $\nu(j) = \pm 1$ is a parity (see [7]). For $d \geq 2$ this concept is different from Barron's variation with respect to half-spaces. For example, the characteristic function χ of the set $\{(x_1, x_2) \in [0, 1]^2;\ x_1 \geq x_2\}$ has the variation w.r.t. half-spaces and any topology equal to 1, while it is easy to verify that its total variation $T(\chi, [0, 1]^2)$ is infinite.

We will need the following integral representation theorem. Its proof is based on properies of delta and Heaviside distributions. Recall ([11, p.33]) that a function $f : \mathcal{R}^d \to \mathcal{R}$ is called a *test function* if $f \in \mathcal{C}^\infty(\mathcal{R}^d)$ and is compactly supported.

Theorem 2.1 *Let d be a positive integer and $f : \mathcal{R}^d \to \mathcal{R}$ be a test function. Then for every $\mathbf{x} \in \mathcal{R}^d$*

$$f(\mathbf{x}) = \int_{S^{d-1}} \int_{\mathcal{R}} \left(- \int_{H_{\mathbf{e}b}} (\bigtriangledown f(\mathbf{y}) \cdot \mathbf{e})\, d\mathbf{y} \right) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) d\mathbf{e} db,$$

where $H_{\mathbf{e}b} = \{\mathbf{y} \in \mathcal{R}; \mathbf{y} \cdot \mathbf{e} = -b\}$.

For a differentiable function, total variation can be characterized as an integral of the absolute value of its derivative. Formally, if $J \subset \mathcal{R}$ is an interval and $f' \in \mathcal{L}_1(J)$ then $T(f, J) = \int_J |f'(x)| dx$ (see e.g. [7, p.242]). Our characterization of variation with respect to half-spaces is of a similar type.

Since each continuous compactly supported function can be uniformly approximated by a sequence of test functions ([13], p.3) we can derive an estimate of variation with respect to half-spaces for differentiable functions. Let $\lambda(J)$ denote the Lebesgue measure of J and $\bigtriangledown f$ denote the gradient of f.

Theorem 2.2 *Let $J \subset \mathcal{R}^d$ be a box and $f : J \to \mathcal{R}$ be a differentiable function. Then*

$$V(f, \mathcal{C}, J) \leq \lambda(J) \sup \left\{ \left| \int_{J \cap H_{\mathbf{e}b}} (\bigtriangledown f(\mathbf{y}) \cdot \mathbf{e})\, d\mathbf{y} \right| ; \mathbf{e} \in \mathcal{S}^{d-1}, b \in \mathcal{R} \right\}.$$

Jones [5] estimated rates of approximation of functions from convex closures of bounded subsets of a Hilbert space. It follows from his results that for any activation function ψ with a norm less than or equal to 1, every function f with $V(f, \mathcal{L}_2, J) \leq B$ can be approximated by a function f_n computable by a network with n hidden ψ-perceptrons within an error $\|f - f_n\|_2 \leq \sqrt{\frac{c}{n}}$, where c is any real number satisfying $c > B^2 - \|f\|_2{}^2$.

Theorem 2.2 together with Jones [5] implies the following estimate:

Corollary 2.3 *Let $J \subset \mathcal{R}^d$ be a box and $f : J \to \mathcal{R}$ be a differentiable function such that $\lambda(J) \sup\{\left|\int_{J \cap H_{\mathbf{e}b}} (\bigtriangledown f(\mathbf{y}) \cdot \mathbf{e})\, d\mathbf{y}\right| ; \mathbf{e} \in \mathcal{S}^{d-1}, b \in \mathcal{R}\} \leq B$. Then for every $n \in \mathcal{N}$ there exists a function f_n computable by a neural network with a linear output unit and n Heaviside perceptrons in the hidden layer such that $\|f - f_n\|_2 \leq \sqrt{\frac{c}{n}}$ where c is any real number satisfying $c > B^2 - \|f\|_2{}^2$.*

Barron [2, Theorem 2] extended Jones' result to supremum norm so we have the following.

Corollary 2.4 *Let $J \subset \mathcal{R}^d$ be a box and $f : J \to \mathcal{R}$ be a differentiable function such that $\lambda(J) \sup\{\left|\int_{J \cap H_{\mathbf{e}b}} (\bigtriangledown f(\mathbf{y}) \cdot \mathbf{e})\, d\mathbf{y}\right| ; \mathbf{e} \in \mathcal{S}^{d-1}, b \in \mathcal{R}\} \leq B$. Then for every sigmoidal function ψ there exists a real number c such that for every $n \in \mathcal{N}$ there exists a function f_n computable by a neural network with a linear output unit and n ψ-perceptrons in the hidden layer such that $\|f - f_n\|_{sup} \leq \frac{cB}{\sqrt{n}}$.*

3 Discussion

Suppose that J has no side of length less than 1. Then for any hyperplane H, $\lambda(J \cap H) \leq \lambda(J) \cdot C(J)$, where $C(J)$ is the geometric constant that describes the ratio of the largest possible $\lambda(J \cap H)$ divided by the smallest $\lambda(J')$, where J' is a face of J. Using the Cauchy-Schwartz inequality, the right-hand side of Theorem 2.2 is at most $\sup\{\| \bigtriangledown f(\mathbf{y})\| ; \mathbf{y} \in J\} \lambda(J) \sup\{\lambda(J \cap H_{\mathbf{e}b}); \mathbf{e} \in \mathcal{S}^{d-1}, b \in \mathcal{R}\}$ which is less than or equal to $\sup_{\mathbf{y} \in J} \| \bigtriangledown f(\mathbf{y})\| C(J) \lambda(J)^2$.

A result of DeVore et al. [4] shows that an upper bound on gradient is not sufficient to guarantee dimension-independent rates of approximation by one-hidden-layer neural networks. Our results show that it *is* sufficient to bound the gradient multiplied by the square of d-dimensional volume of the box where the function is defined. Since

volume of a d-dimensional cube grows exponentially with increasing dimension, to keep $\| \bigtriangledown f(\mathbf{y})\| C(J)\lambda(J)^2$ bounded by the same bound B, the gradient must be decreasing with increasing d.

So the dimension-independent rates of approximation must be interpreted and used carefully. The size of spaces of functions that can be approximated with rates of approximation $\mathcal{O}(\frac{1}{\sqrt{n}})$ is decreasing with increasing input dimension d. Also the constant factor (depending on a bound B and the $\mathcal{L}_2$-norm of the function to be approximated which is a d-dimensional integral) can, at realistic scales, dominate the $\frac{1}{\sqrt{n}}$ factor.

4 Sketches of proofs

To prove Theorem 2.1 we need two technical lemmas. The first one can be verified using approximation of the delta distribution by a sequence of sharp pulse functions (e.g., functions defined in [13, p.12]) converging uniformly in distributional sense to delta.

Lemma 4.1 *For every positive integer d* $\delta(\mathbf{x}) = \int_{S^{d-1}} \delta(\mathbf{e}\cdot\mathbf{x})d\mathbf{e}$.

Recall [10] that the *directional derivative* $D_{\mathbf{e}}f(\mathbf{y})$ of f in the direction $\mathbf{e}$ is defined by $D_{\mathbf{e}}f(\mathbf{y}) = \lim_{t\to 0} t^{-1}(f(\mathbf{y}+t\mathbf{e}) - f(\mathbf{y}))$

Lemma 4.2 *For every positive integer d, for every differentiable function $f : \mathcal{R}^d \to \mathcal{R}$ and for every unit vector $\mathbf{e} \in \mathcal{R}^d$ and for every $b \in \mathcal{R}$* $\frac{\partial}{\partial b}\int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} = \int_{H_{\mathbf{e}b}} (\bigtriangledown f(\mathbf{y})\cdot\mathbf{e})\, d\mathbf{y}$.

Proof.

$$\frac{\partial}{\partial b}\int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} = \lim_{t\to 0} t^{-1}\left(\int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} - \int_{H_{\mathbf{e}b+t}} f(\mathbf{y})d\mathbf{y}\right) = \lim_{t\to 0} t^{-1}\int_{H_{\mathbf{e}b}} (f(\mathbf{y}+t\mathbf{e}) - f(\mathbf{y}))$$

$d\mathbf{y} = \int_{H_{\mathbf{e}b}} \lim_{t\to 0} t^{-1}(f(\mathbf{y}+t\mathbf{e}) - f(\mathbf{y})) = \int_{H_{\mathbf{e}b}} D_{\mathbf{e}}f(\mathbf{y})d\mathbf{y}$. By the definition of gradient [10, p.222], $\int_{H_{\mathbf{e}b}} D_{\mathbf{e}}f(\mathbf{y})d\mathbf{y} = \int_{H_{\mathbf{e}b}} (\bigtriangledown f(\mathbf{y})\cdot\mathbf{e})\, d\mathbf{y}$. □

Proof of Theorem 2.1. Since f is a test function $f(\mathbf{x}) = (f * \delta)(\mathbf{x}) = \int_{\mathcal{R}^d} f(\mathbf{z})\delta(\mathbf{x}-\mathbf{z})d\mathbf{z}$ (see [13]). By Lemma 4.1 $\delta(\mathbf{x}-\mathbf{z}) = \int_{S^{d-1}} \delta(\mathbf{e}\cdot\mathbf{x} - \mathbf{e}\cdot\mathbf{z})d\mathbf{e}$. Thus, $f(\mathbf{x}) = \int_{S^{d-1}}\int_{\mathcal{R}^d} f(\mathbf{z})\delta(\mathbf{x}\cdot\mathbf{e} - \mathbf{z}\cdot\mathbf{e})d\mathbf{z}d\mathbf{e}$. So rearranging the inner integration, we have $f(\mathbf{x}) = \int_{S^{d-1}}\int_{\mathcal{R}}\int_{H_{\mathbf{e}b}} f(\mathbf{y})\delta(\mathbf{x}\cdot\mathbf{e}+b)d\mathbf{y}dbd\mathbf{e}$, where $H_{\mathbf{e}b} = \{\mathbf{y}\in\mathcal{R}; \mathbf{y}\cdot\mathbf{e} = -b\}$. Let $u(\mathbf{e},b) = \int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y}$, so $f(\mathbf{x}) = \int_{S^{d-1}}\int_{\mathcal{R}} u(\mathbf{e},b)\delta(\mathbf{x}\cdot\mathbf{e}+b)dbd\mathbf{e}$.

Since the first distributional derivative of the Heaviside function is the delta distribution [13, p.47], it follows that for every $\mathbf{e}\in S^{d-1}$ and $\mathbf{x}\in\mathcal{R}^d$ $\int_{\mathcal{R}} u(\mathbf{e},b)\delta(\mathbf{e}\cdot\mathbf{x}+b)db = -\int_{\mathcal{R}} \frac{\partial u(\mathbf{e},b)}{\partial b}\vartheta(\mathbf{e}\cdot\mathbf{x}+b)db$. By Lemma 4.2 $\frac{\partial u(\mathbf{e},b)}{\partial b} = \frac{\partial}{\partial b}\int_{H_{\mathbf{e}b}} f(\mathbf{y})d\mathbf{y} = \int_{H_{\mathbf{e}b}} (\bigtriangledown f(\mathbf{y})\cdot\mathbf{e})d\mathbf{y}$. Hence, $f(\mathbf{x}) = \int_{S^{d-1}}\int_{\mathcal{R}}\left(-\int_{H_{\mathbf{e}b}} (\bigtriangledown f(\mathbf{y})\cdot\mathbf{e})\, d\mathbf{y}\right)\vartheta(\mathbf{e}\cdot\mathbf{x}+b)dbd\mathbf{e}$. □

Proof of Theorem 2.2. For every $\delta > 0$ consider an extension of f to a function $f_\delta : \mathcal{R}^d \to \mathcal{R}$ differentiable on $\mathcal{R}^d$ such that f_δ is zero outside of a δ-neighbourhood J_δ of J. Let $\{f_{\delta n}; n\in\mathcal{N}\}$ be a sequence of test functions converging uniformly on $\mathcal{R}^d$ to

f_δ such that all $f_{\delta n}$ are zero outside the $\frac{1}{n}$-neighbourhood $J_{\delta n}$ of J_δ (such a sequence can be constructed using convolutions with a sequence of test functions converging uniformly to zero and having integrals equal to 1, see [13]). Note that as n goes to ∞, $\lambda(J_{\delta n})$ approaches $\lambda(J_\delta)$ and as δ goes to 0, $\lambda(J_\delta)$ goes to $\lambda(J)$.

By Theorem 2.1 $f_{\delta n}(\mathbf{x}) = \int_{S^{d-1}} \int_{\mathcal{R}} \left(-\int_{H_{\mathbf{e}b}} (\bigtriangledown f_{\delta n}(\mathbf{y}) \cdot \mathbf{e})\, d\mathbf{y} \right) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) db d\mathbf{e}$. Put $B_{\delta n} = \lambda(J_{\delta n}) \sup\{ \left| \int_{H_{\mathbf{e}b}} (\bigtriangledown f_{\delta n}(\mathbf{y}) \cdot \mathbf{e})\, d\mathbf{y} \right| ; \mathbf{e} \in \mathcal{S}^{d-1}, b \in \mathcal{R}\}$. Approximating integrals by sums we get $f_{\delta n} \in cl_{\mathcal{C}}(conv(\mathcal{E}_d(\vartheta, B_{\delta n}, J_{\delta n})))$. Since $f_\delta = \lim_{n\to\infty} f_{\delta n}$ uniformly on $\mathcal{R}^d$, we have $\bigtriangledown f_\delta = \lim_{n\to\infty} \bigtriangledown f_{\delta n}$. So, $B_\delta = \lim_{n\to\infty} B_{\delta n}$. Put $B = \lambda(J) \sup\{ \left| \int_{H_{\mathbf{e}b} \cap J} (\bigtriangledown f(\mathbf{y}) \cdot \mathbf{e})\, d\mathbf{y} \right| ; \mathbf{e} \in \mathcal{S}^{d-1}, b \in \mathcal{R}\}$. Since for every $\delta > 0$ the restriction $f_\delta|_J = f$, we have $B = \lim_{\delta\to 0} B_\delta$. $\square$

Bibliography

[1] A.R. Barron: Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**, 930-945, 1993.

[2] A.R. Barron: Neural net approximation. In *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems* (pp. 69-72), 1992.

[3] C. Darken, M. Donahue, L. Gurvits, E. Sontag: Rate of approximation results motivated by robust neural network learning. In *Proceedings of the 6th ACM Conference on Computational Learning Theory* (pp. 303-309), New York: ACM.

[4] R. DeVore, R. Howard, C.A. Micchelli: Optimal nonlinear approximation. *Manuscripta Mathematica* **63**, 469-478, 1989.

[5] L.K. Jones: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics* **20**, 601-613, 1992.

[6] V. Kůrková: Kolmogorov's theorem and multilayer neural networks. *Neural Networks* **5**, 501-506, 1992.

[7] E. J. McShane: *Integration*, Princeton: Princeton University Press, 1944.

[8] H. N. Mhaskar, C. A. Micchelli: Approximation by superposition of a sigmoidal function. *Advances of Applied Mathematics* **13**, 350-373, 1992.

[9] J. Park, I. W. Sandberg: Approximation and radial-basis-function networks. *Neural Computation* **5**, 305-316, 1993.

[10] W. Rudin: *Principles of Mathematical Analysis*, New York: McGraw-Hill, 1964.

[11] W. Rudin: *Functional Analysis*. New York: McGraw-Hill, 1973.

[12] T.J. Sejnowski, B.P. Yuhas: Mapping between high-dimensional representations of acoustic and speech signal. In *Computation and Cognition* (pp. 52-68). Philadelphia: Siam.

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