Function Approximation by Three-Layered Networks and Its Error Bounds — An Integral Representation Theorem

Noboru Murata

METR 94-19 October 1994

Function Approximation by Three-Layered Networks and Its Error Bounds — An Integral Representation Theorem

Noboru Murata
University of Tokyo

METR 94-19, October 1994

Abstract

Neural Networks are widely noticed to provide a nonlinear function approximation method. In order to make its approximation ability clear, a new theorem on an integral transform of ridge functions is presented. By using this theorem, an approximation bound, which clarifies the quantitative relationship between the approximation accuracy and the number of elements in the hidden layer, can be obtained. This result shows that the approximation accuracy depends on the smoothness of target functions. It also shows that the approximation methods which use ridge functions are free from "curse of dimensionality".

1 Overview

In the middle of the 1980s, computational research on neural networks was activated by the works of the Parallel Distributed Processing (PDP) group, and multi-layered networks having sigmoidal output functions together with back-propagation learning played important roles in this movement. Many kinds of examples provided by the PDP group attracted interest of other researchers, and a large number of subsequent computer simulations showed that the multi-layered networks are applicable to practical problems, such as image processing, speech recognition, system control, and so on.

Advantages of these simple learning machines are summarized in the following two points. One is that the learning algorithm "back-propagation"

can be easily implemented on computers, because the algorithm is essentially a gradient descent method. The gradient of each parameter is easily calculated locally, and the derivative of the sigmoidal function depends only on its output.

The other is that if there are sufficiently many hidden units, a three-layered network can approximate an arbitrary function with desired accuracy. Irie and Miyake[5], Funahashi[3], Cybenko[2], White[7] proved this fact mainly based on the notion of the Fourier transform. Funahashi[3] also gave another proof making use of Kolmogorov and Arnol'd's theorem and Sprecher's theorem. Recently Jones[6], Barron[1] and Girosi and Anzellotti[4] showed that the approximating error with respect to the mean square or sup norm criterion is inversely proportional to the number of the hidden units on a certain class of functions. Their results are interesting and important because it is shown that greedy algorithms such as neural networks and radial basis approximators with back-propagation learning can avoid "curse of dimensionality".

In this paper, we describe the second advantage from another point of view. As the first step to clarify the performance and limitation of three-layered networks as function approximators, we focus on the case where we have complete information about target functions. We don't treat learning from examples in this paper. We show the structure of multi-layered networks has the good property of well-approximating nonlinear functions.

We firstly define an integral representation of functions by using ridge functions. It is a new kind of function transform. This transform gives a rigorous representation of functions and also gives a reasonable interpretation about the structure of three-layered networks. From the correspondence between the transformation coefficients and the parameters of networks, we obtain an intuitive interpretation of the network parameters.

Secondly, applying this result, we give a bound to the approximation accuracy of the three-layered network with a finite number of hidden units. Using the random coding technique, which is commonly known in the field of information theory, it is shown that the mean square error of approximation is inversely proportional to the number of hidden units. It means that as higher accuracy is desired, a linearly large number of hidden units is required, not an exponentially large number. Hence it follows that three-layered networks are free from the restrictions of "curse of dimensionality".

Lastly, we discuss the relationship between the smoothness of functions and the approximation error of the three-layered networks and we give a class of functions which are well-approximated by three-layered networks.

We suggest that the smoothness of target functions is closely connected with the magnitude of the parameters, and that the smoothness plays an important role to prevent the divergence of learning.

2 Integral Transform of Ridge Functions

First of all, we define ridge functions.

Definition 1 When a function $F : R^m \to R$ is written as

$$F(\boldsymbol{x}) = G(\boldsymbol{a} \cdot \boldsymbol{x} - b), \tag{1}$$

with a vector $\boldsymbol{a} \in R^m$, a real number $b \in R$ and an appropriate function $G : R \to R$, it is called *a ridge function.*

In other words, a ridge function takes the same value on certain hyperplanes of R^m whose normal vectors are $\boldsymbol{a}$ (cf. figure 1). Clearly, the output of a neuron with the sigmoidal activation function in an artificial neural network belongs to the class of the ridge functions given in equation (1). We should note that when $m \geq 2$, F is not integrable even if G is integrable on R. In the following, we discuss about the approximation of a certain function $f : R^m \to R$ by a linear combination of ridge functions.

Let us assume that f belongs to $L^1(R^m) \cap L^p(R^m), \quad 1 \leq p < \infty$, or f is bounded and uniformly continuous. Let a pair of two functions $\phi_d, \phi_c \in L^1(R) \cap L^2(R)$ be bounded, satisfying the following condition:

$$\int_R |\omega|^{-m} \left| \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) \right| d\omega < \infty, \tag{2}$$

where $\hat{\cdot}$ denotes the Fourier transform and $\overline{\cdot}$ denotes the complex conjugate. The indices d, c are used to represent the decomposing and composing kernels.

We define

$$C_{\phi_d,\phi_c} = \int_{R^m} |\omega|^{-m} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) d\omega, \quad 0 < |C_{\phi_d,\phi_c}| < \infty. \tag{3}$$

The transform T of f with respect to ϕ_d and ϕ_c is defined by

$$T(\boldsymbol{a}, b) = C_{\phi_d,\phi_c}^{-1} \int_{R^m} \overline{\phi_d(\boldsymbol{a} \cdot \boldsymbol{x} - b)} f(\boldsymbol{x}) d\boldsymbol{x}. \tag{4}$$

Then we can prove the following inversion theorem.

Theorem 1 *Using the transform T given in equation (4), f can be represented as*

$$f(\boldsymbol{x}) = \lim_{\varepsilon \to 0} \int_{R^{m+1}} \phi_c(\boldsymbol{a} \cdot \boldsymbol{x} - b) T(\boldsymbol{a}, b) e^{-\varepsilon |\boldsymbol{a}|^2} d\boldsymbol{a} db. \tag{5}$$

If $f \in L^1(R^m) \cap L^p(R^m) \quad (1 \leq p < \infty)$, $f(\boldsymbol{x})$ in equation (5) converges in the sense of L^p-norm, and if f is bounded and uniformly continuous, it converges in the sense of L^∞-norm.

The proof is given in appendix A.

When $m = 1$, that is, f is a function on R, we can see the close relation between this transform and the wavelet transform. When $m \geq 2$, this kind of relation cannot be found, because the basis of wavelet transforms belongs to $L^2(R^m)$ while $\phi_c(\boldsymbol{a} \cdot \boldsymbol{x} - b)$ is not square-integrable with respect to $\boldsymbol{x}$.

From the fact that f belongs to $L^1(R^m)$ and ϕ_d is bounded, we see that T is bounded, $T(\boldsymbol{a}, b) \in L^\infty(R^{m+1})$. We can also derive

$$\begin{aligned} \int_R |T(\boldsymbol{a}, b)| db &= \int_R \left| C_{\phi_d, \phi_c}^{-1} \int_{R^m} \overline{\phi_d(\boldsymbol{a} \cdot \boldsymbol{x} - b)} f(\boldsymbol{x}) d\boldsymbol{x} \right| db \\ &\leq |C_{\phi_d, \phi_c}^{-1}| \int_R \int_{R^m} |\overline{\phi_d(\boldsymbol{a} \cdot \boldsymbol{x} - b)}| |f(\boldsymbol{x})| db \, d\boldsymbol{x} \\ &< \infty. \end{aligned}$$

Even though $T(\boldsymbol{a}, b)$ is bounded and ϕ_c belongs to $L^1(R)$,

$$f(\boldsymbol{x}) = \int_{R^{m+1}} \phi_c(\boldsymbol{a} \cdot \boldsymbol{x} - b) T(\boldsymbol{a}, b) d\boldsymbol{a} db \tag{5'}$$

is a divergent integral. Then we should note that the convergence of equation (5) is guaranteed by using the convergent factor $e^{-\varepsilon |\boldsymbol{a}|^2}$. In the following discussion, however, we use the expression (5′) instead of equation (5) for notational simplicity.

Equation (4) can be seen to give a map from $L^1(R^m)$ to $L^\infty(R^{m+1})$. Noting that the Fourier transform and inverse Fourier transform give a one-to-one correspondence between $L^1(R^m)$ and $L^\infty(R^m)$, equation (4) has redundancy. This means that f could for example be represented by two different transforms $T^{(1)}(\boldsymbol{a}, b), T^{(2)}(\boldsymbol{a}, b)$ with respect to two decomposing kernel functions $\phi_d^{(1)}, \phi_d^{(2)}$ which satisfy the admissible condition (2) with the same composing kernel ϕ_c. So there is a set $\phi_d^{(1)}, \phi_d^{(2)}, \ldots$ among which we can select a convenient ϕ_d by imposing certain properties such as compactness, smoothness, and so on. We use this advantage to investigate the relationship between the smoothness of target functions and approximation ability in section 3.4.

3 Application to Three-layered Networks

From the previous result, we can evaluate the ability of the three-layered networks which use hidden units with sigmoidal activation functions.

3.1 Representation by Bell-shaped Functions

In the engineering field, a sigmoidal function ψ which has the following characteristics

$$\lim_{z\to-\infty}\psi(z)=0, \quad \lim_{z\to\infty}\psi(z)=1,$$
$$\frac{d}{dz}\psi(z)=\psi'(z)>0, \quad \lim_{z\to\pm\infty}\psi'(z)=0.$$

is usually used as an activation function of hidden units. In the back-propagation learning algorithm, the monotonicity and smoothness of sigmoidal functions are profitable. Since sigmoidal functions do not belong to $L^1(R)$, we cannot apply the previous representation result to this type of networks.

We discard the assumption of monotonicity and define a bell-shaped function ϕ_c which has the following integrability condition,

$$\int_R \phi_c(z)dz<\infty, \quad \lim_{z\to\pm\infty}\phi_c(z)=0,$$
$$\max_z \phi_c(z)=1, \quad \phi_c(z)>0, \quad \phi_c \text{ is unimodal.}$$

Namely, the bell-shaped function is an unimodal $L^1(R)$ function whose maximal value is 1. In the following, we consider the networks composed of ϕ_c's, such that for an input $\boldsymbol{x}\in R^m$ the output $f_n(\boldsymbol{x})\in R$ is calculated as

$$f_n(\boldsymbol{x})=\sum_{i=1}^{n} c_i\phi_c(\boldsymbol{a}_i\cdot\boldsymbol{x}-b_i), \tag{6}$$

where $\boldsymbol{a}_i\in R^m$, $b_i, c_i\in R$ and n denotes the number of hidden units.

A bell-shaped function can be composed of two appropriate sigmoidal functions, for example,

$$\phi_c(z)=\psi(z+h)-\psi(z-h), \quad h>0$$

and we demand that ϕ_c belongs to $L^1(R)$. In this case, networks composed of units with bell-shaped functions can also be written as combinations of

units with sigmoidal functions, the number of which is twice as the number of bell-shaped ones.

We can also represent a bell-shaped function as the derivative of a sigmoidal function,

$$\phi_c(z) = \frac{d}{dz}\psi(z).$$

In this case ϕ_c is clearly integrable.

3.2 Expression of Functions by Three-layered Networks

Using the integral transform defined in the previous section, an arbitrary $f \in L^1(R^m)$ can be written as

$$f(\boldsymbol{x}) = \int_{R^{m+1}} \phi_c(\boldsymbol{a} \cdot \boldsymbol{x} - b) T(\boldsymbol{a}, b) d\boldsymbol{a} db.$$

Then we see that a three-layered network composed of bell-shaped functions can approximate an arbitrary function $f \in L^1(R^m)$ with demanded accuracy if the network has a sufficient number of hidden units. Here $T(\boldsymbol{a}, b)$ is calculated by equation (4) with function ϕ_d which satisfies condition (2).

Since a bell-shaped function can be composed as a linear combination of two sigmoidal functions, a three-layered network composed of the sigmoidal functions can approximate any bounded, uniformly continuous functions.

We can see a straightforward correspondence between the integral representation presented above and the structure of three-layered networks: the parameters $\boldsymbol{a}, b$ correspond to the weights and the thresholds from the input layer to the hidden layer, the spectrum $T(\boldsymbol{a}, b)$ corresponds to the weights from the hidden layer to the output layer (cf. figure 2).

If a composing kernel ϕ_c is sufficiently smooth, we can construct a decomposing kernel ϕ_d as follows.

Choose one $\rho \in C_0^\infty(R)$ which satisfies

$$\begin{aligned} &\rho(z) \geq 0 \\ &\rho(z) = 0 \text{ if } |z| \geq 1 \\ &\int_R \rho(z) dz < \infty, \end{aligned}$$

where $C_0^\infty(R)$ denotes a class of functions on R which belong to $C^\infty(R)$ with compact supports. For example,

$$\rho(z) = \begin{cases} e^{-1/(1-|z|^2)} & |z| < 1 \\ 0 & |z| \geq 1 \end{cases} \tag{7}$$

is available. A decomposing kernel is given by

$$\phi_d(z) = \begin{cases} \dfrac{d^m}{dz^m}\rho(z) & m : \text{even} \\ \dfrac{d^{m+1}}{dz^{m+1}}\rho(z) & m : \text{odd} \end{cases}$$

depending on the dimensionality m of input $\boldsymbol{x}$.

We confirm that these kernels satisfy the admissible condition (2) in appendix B.

3.3 Approximation Bound by Bell-shaped Networks

In the following, we assume that inputs $\boldsymbol{x} \in R^m$ are generated subject to a probability density $\mu(\boldsymbol{x})$, and we evaluate quantitatively the approximation accuracy of the network with n hidden units in terms of the L^2-norm with respect to the probability density $\mu(\boldsymbol{x})$

$$\|f_n(\boldsymbol{x}) - f(\boldsymbol{x})\|^2_{L^2(R^m),\mu} = \int_{R^m} (f_n(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \mu(\boldsymbol{x}) d\boldsymbol{x}, \tag{8}$$

where f is a target function to be approximated.

Firstly, we assume

$$\int |T(\boldsymbol{a}, b)| d\boldsymbol{a} db < \infty \tag{9}$$

holds. Taking into account that f and ϕ_c are real functions, we can derive

$$\begin{aligned} f(\boldsymbol{x}) &= \int T(\boldsymbol{a}, b)\phi_c(\boldsymbol{a}\cdot\boldsymbol{x} - b) d\boldsymbol{a} db \\ &= \mathrm{Re} \int T(\boldsymbol{a}, b)\phi_c(\boldsymbol{a}\cdot\boldsymbol{x} - b) d\boldsymbol{a} db \\ &= \int \mathrm{Re}\, T(\boldsymbol{a}, b)\phi_c(\boldsymbol{a}\cdot\boldsymbol{x} - b) d\boldsymbol{a} db \\ &= \int \frac{|\mathrm{Re}\, T(\boldsymbol{a}, b)|}{C_T} \mathrm{sign}(\mathrm{Re}\, T(\boldsymbol{a}, b)) C_T \phi_c(\boldsymbol{a}\cdot\boldsymbol{x} - b) d\boldsymbol{a} db \\ &= \int c(\boldsymbol{a}, b)\phi_c(\boldsymbol{a}\cdot\boldsymbol{x} - b) p(\boldsymbol{a}, b) d\boldsymbol{a} db, \end{aligned}$$

where

$$\begin{aligned} C_T &= \int |\mathrm{Re}\, T(\boldsymbol{a}, b)| d\boldsymbol{a} db, \\ c(\boldsymbol{a}, b) &= \mathrm{sign}(\mathrm{Re}\, T(\boldsymbol{a}, b)) C_T, \\ p(\boldsymbol{a}, b) &= \frac{|\mathrm{Re}\, T(\boldsymbol{a}, b)|}{C_T}. \end{aligned}$$

As $p(\boldsymbol{a}, b)$ is a positive function whose integral value is 1, $p(\boldsymbol{a}, b)$ is regarded as a probability density function.

Let us think about the following function

$$g_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} c(\boldsymbol{a}_i, b_i) \phi_c(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i), \tag{10}$$

where $(\boldsymbol{a}_i, b_i) \quad (i = 1, \ldots, n)$ are independently chosen, subject to the probability density $p(\boldsymbol{a}, b)$. The expectation and variance of $g_n(\boldsymbol{x})$ satisfy

$$\begin{aligned} E\left[g_n(\boldsymbol{x})\right] &= f(\boldsymbol{x}), \\ V\left[g_n(\boldsymbol{x})\right] &= \frac{1}{n} V\left[c(\boldsymbol{a}, b)\phi_c(\boldsymbol{a} \cdot \boldsymbol{x} - b)\right] \\ &\leq \frac{1}{n} C_T^2 E\left[\phi_c(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i)^2\right] \\ &\leq \frac{1}{n} C_T^2. \end{aligned}$$

The last inequality comes from the fact that the maximum value of ϕ_c is 1. These relations lead to

$$\begin{aligned} & E\left[\int (g_n(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \mu(\boldsymbol{x}) d\boldsymbol{x}\right] \\ &= \int E\left[(g_n(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right] \mu(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int V\left[g_n(\boldsymbol{x})\right] \mu(\boldsymbol{x}) d\boldsymbol{x} \\ &\leq \frac{1}{n} C_T^2. \end{aligned} \tag{11}$$

Equation (11) shows that the expectation of $\|g_n(\boldsymbol{x}) - f(\boldsymbol{x})\|^2_{L^2(R^m),\mu}$ is less or equal to C_T^2/n, and at least there exists one g_n which is a linear combination of ϕ_c's and whose mean square error is as small as C_T^2/n. The previous evaluation is derived under the assumption of equation (9). If the transform T belongs to $L^1(R^{m+1})$, the mean square error of function approximation with a three-layered network is bounded by $O(1/n)$ where n is the number of hidden units.

Now we obtain the following theorem.

Theorem 2 *Let ϕ_c be a bell-shaped function which is an activation function of the hidden units and ϕ_d be a corresponding decomposing kernel function*

which satisfies the admissible condition (2). If the absolute integral C_T of the transform T of the function f is bounded, there exists a bell-shaped three-layered network which satisfies

$$\|g_n(\boldsymbol{x}) - f(\boldsymbol{x})\|^2_{L^2(R^m),\mu} \leq \frac{1}{n} C_T^2.$$

The coefficient given in the theorem is obtained by a rough estimation. In this evaluation, we treated only such networks whose weights from the hidden layer to the output layer are of the same absolute value (cf. figure 3). In other words, the approximation depends only on the density of $(\boldsymbol{a}, b)$ and takes no account of the magnitude of the hidden-output connection weight c_i. Another estimation of an approximation bound is discussed in appendix C, though it shows a bound of the same order of $O(1/n)$.

This theorem shows that three-layered networks have good facility in approximating functions. The approximation bound with respect to the mean square error is inversely proportional to the number of hidden units. Naturally, it can be seen that if there are a sufficiently large number of hidden units, three-layered networks can achieve approximations with any desired accuracy. An important point is that as increase of desired accuracy, an exponentially large number of hidden units are not required, but a linearly large number of them, independent from the input dimensionality, are required. It means that three-layered networks are free from "curse of dimensionality".

In addition, it also shows that a combination of good approximations is not always a good approximation. For example, we divide the target function f into two functions f_1 and f_2

$$f = f_1 + f_2$$

and we have two good approximations g_1 and g_2 which have n-hidden units and satisfy

$$\|g_1(\boldsymbol{x}) - f_1(\boldsymbol{x})\|^2_{L^2(R^m),\mu} \leq \frac{1}{n} C_{T_1}^2,$$
$$\|g_2(\boldsymbol{x}) - f_2(\boldsymbol{x})\|^2_{L^2(R^m),\mu} \leq \frac{1}{n} C_{T_2}^2,$$

where T_1 and T_2 are the transform of f_1 and f_2. In order to approximate the function f, we combine two approximations again. The error bound is evaluated as

$$\|\left(g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})\right) - f(\boldsymbol{x})\|^2_{L^2(R^m),\mu} \leq \frac{1}{n}\left(C_{T_1} + C_{T_2}\right)^2.$$

Note that this combined network has $2n$ hidden units and the equality holds depending on the partition of the target function. From the theorem, we obtain the following bound for an approximator g with $2n$ hidden units

$$\|g(\boldsymbol{x}) - f(\boldsymbol{x})\|^2_{L^2(R^m),\mu} \leq \frac{1}{2n}\left(C_{T_1} + C_{T_2}\right)^2 .$$

It shows that there exist a better approximation than a combination of two approximators if we use a bad partition of the target function. In this sense, function approximation by three-layered networks has nonlinear architecture.

3.4 Relationship between Smoothness of Functions and Integrability of Transforms

Practically, it is difficult to confirm whether the transform of the target function is integrable or not. The integrability of the transform should be translated into other kinds of useful conditions.

There are many kinds of conditions under which the transform T belongs to $L^1(R^{m+1})$. In the following, we only show a sufficient condition. The integrability is restated in terms of the smoothness of target functions. This is an important condition for practical applications.

$C_0^m(R)$ denotes the class of functions with compact supports belonging to $C^m(R)$. The function $f_e(z)$ is defined as

$$f_e(z) = \int_{e\cdot\boldsymbol{x}=z} f(\boldsymbol{x}) d\boldsymbol{x}^{(m-1)}$$

for a unit vector $\boldsymbol{e} \in R^m$, where $d\boldsymbol{x}^{(m-1)}$ denotes the volume element on the hyper-plane $\boldsymbol{e}\cdot\boldsymbol{x} = z$.

Theorem 3 *Any function which satisfies the continuity condition*

$$\begin{aligned} &f_e(z) \in C_0^m(R), \\ &|f_e^{(m)}(z) - f_e^{(m)}(z')| < M|z - z'|^\alpha, \\ &\quad \forall \boldsymbol{e} \in R^m,\ M \in R,\ 0 < \alpha \leq 1 \end{aligned}$$

can be approximated by the three-layered network with n bell-shaped hidden units with $O(1/n)$ accuracy in the sense of mean square errors.

For details of the constructive proof, refer to appendix D.

In practical use, it is almost impossible to confirm the integrability of the transform of a target function. Since the continuity condition is more easily checked than the integrability, theorem 3 can be widely applied. The condition stated in theorem 3 is called Hölder continuity condition. When $f^{(m)}(\boldsymbol{x})$ is Hölder continuous with respect to m-dimensional L^1-norm, it is clear that $f_e^{(m)}(z)$ is Hölder continuous. Moreover, when $f(\boldsymbol{x})$ is a bounded $m+1$ times differentiable function with a compact support, $f(\boldsymbol{x})$ satisfies the condition of theorem 3. And therefore it is expected that most of functions used in practical applications satisfy the condition.

From the result of theorem 3, we conjecture that learning from examples will be unstable when hidden units are augmented carelessly. If the target function is sufficiently smooth, connection weights stay in a bounded area. But if there exists some discontinuity in derivatives of a target function, it is not guaranteed that the weight stays in a bounded area. This indicates the possibility that when a learning machine proceeds to approximate a piecewise linear function constructed by examples, the connection weights diverge because of the unsmoothness. In the case of learning from examples with noise, it is often observed that the huge values of parameters are needed in order to approximate a peculiar data which sticks out like the delta function. This over-learning phenomenon is a serious problem in practical applications. To avoid this problem, kernel methods are commonly used. Our framework validate these methods as follows: taking a convolution with a smooth kernel function makes a discontinuous function smooth and makes its transform integrable.

We expect that our theoretical framework gives a means to analyze various problems of learning from examples.

4 Conclusion

We found an integral representation of functions using ridge functions. Especially when bell-shaped functions are used as the activation functions of the hidden units of three-layered networks, there exists a clear and direct correspondence between the integral representations and the three-layered networks. It gives an intuitive interpretation about the structure of the three-layered networks.

Based on this interpretation and the random coding technique, we gave an approximation bound of three-layered networks in the case of having

complete information about target functions. The bound of the mean square error is inversely proportional to the number of hidden units and does not depend on the dimensionality of the input. From this result, it can be seen that three-layered networks are not subject to "curse of dimensionality".

In addition, we discussed about a sufficient condition for a class of functions to be well approximated. As seen in the previous discussions, we show that the smoothness of functions is closely related with errors of approximating functions by a combination of sigmoidal functions. In practical use, this condition is of wide applications.

Future work will be dedicated to applying our results to problems of learning from examples. Possible applications would be the estimation of the required number of hidden units for given accuracy, the estimation of the error rate, the acceleration of the convergence of learning, warding off "over-learning", and so on.

Acknowledgment The author would like to thank Professor S. Amari, Professor S. Yoshizawa and Dr. K. R. Müller for helpful suggestions and discussions.

References

[1] Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory*, 39(3):930–945, May 1993.

[2] G. Cybenko. Approximation by superpositions of a sigmoid function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

[3] Ken-ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2:183–192, 1989.

[4] F. Girosi and G. Anzellotti. Comvergence rates of approximation by translates. Artificial Intelligence Laboratory Technical Report 1288, Massachusetts Institute of Technology, 1992.

[5] Bunpei Irie and Sei Miyake. Capabilities of three-layered perceptrons. In *International Conference on Neural Nerworks*, pages 641–648. IEEE, 1988.

[6] Lee 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(1):608–613, 1992.

[7] Halbert White. Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3(5):535–549, 1990.

Appendix

A Proof of the Transform Theorem

Let us define

$$\begin{aligned} f_\varepsilon(\boldsymbol{x}) &= \int_{R^{2m+1}} C^{-1}_{\phi_d,\phi_c} \overline{\phi_d(\boldsymbol{a}\cdot\boldsymbol{y}-b)} f(\boldsymbol{y}) \phi_c(\boldsymbol{a}\cdot\boldsymbol{x}-b) e^{-\varepsilon|\boldsymbol{a}|^2} d\boldsymbol{y} db d\boldsymbol{a} \\ &= C^{-1}_{\phi_d,\phi_c} \int_{R^m} \int_R \int_{R^m} \overline{\phi_d(\boldsymbol{a}\cdot\boldsymbol{y}-b)} f(\boldsymbol{y}) d\boldsymbol{y}\, \phi_c(\boldsymbol{a}\cdot\boldsymbol{x}-b) db\, e^{-\varepsilon|\boldsymbol{a}|^2} d\boldsymbol{a} \end{aligned}$$

From the fact that $f(\boldsymbol{y}) \in L^1(R^m)$, ϕ_d is bounded, and $\phi_c \in L^1(R)$, we can see that the right side of equation 12 is absolutely integrable. According to Fubini-Tonelli's theorem, we can exchange the order of the integration of $f_\varepsilon(\boldsymbol{x})$. As $\phi_d(\boldsymbol{a}\cdot\boldsymbol{y}-b)$ and $\phi_c(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ can be seen as functions of $-b$ and shifted by $\boldsymbol{a}\cdot\boldsymbol{y}$ or $\boldsymbol{a}\cdot\boldsymbol{x}$ respectively, we can obtain

$$\begin{aligned} f_\varepsilon(\boldsymbol{x}) &= \int_{R^{2m+1}} C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) e^{i\xi\boldsymbol{a}\cdot(\boldsymbol{x}-\boldsymbol{y})} e^{-\varepsilon|\boldsymbol{a}|^2} f(\boldsymbol{y}) d\xi d\boldsymbol{y} d\boldsymbol{a} \\ &= \int_{R^{2m+1}} C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) e^{-\varepsilon\left\{\boldsymbol{a}-\frac{i\xi(\boldsymbol{x}-\boldsymbol{y})}{2\sigma^2}\right\}^2 - \frac{|\xi(\boldsymbol{x}-\boldsymbol{y})|^2}{4\varepsilon}} f(\boldsymbol{y}) d\boldsymbol{a} d\xi d\boldsymbol{y} \\ &= \int_{R^{m+1}} C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) G_\varepsilon(\xi(\boldsymbol{x}-\boldsymbol{y})) f(\boldsymbol{y}) d\xi d\boldsymbol{y} \\ &= \int_{R^{m+1}} C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} G_{\frac{\varepsilon}{|\xi|^2}}(\boldsymbol{x}-\boldsymbol{y}) f(\boldsymbol{y}) d\boldsymbol{y} d\xi \\ &= \int_R C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) d\xi \end{aligned}$$

by using Parseval's equality, where

$$G_\varepsilon(\boldsymbol{x}) = \frac{1}{(4\pi\varepsilon)^{m/2}} e^{-|\boldsymbol{x}|^2/4\varepsilon}$$

and $*$ denotes the convolution.

Firstly, we prove the theorem in the case of $f(\boldsymbol{x}) \in L^p(R^m)$. In this case, as

$$\begin{aligned} &\|G_\varepsilon * f\|_{L^p(R^m)} \le \|f\|_{L^p(R^m)} \\ &\lim_{\varepsilon\to 0} \|G_\varepsilon * f - f\|_{L^p(R^m)} = 0 \end{aligned}$$

holds, if we take a sufficiently small ε, for any δ's there exists γ such that

$$\|G_{\frac{\varepsilon}{|\xi|^2}} * f - f\|_{L^p(R^m)} < \delta, \quad |\xi| > \gamma$$

and

$$\int_{|\xi|<\gamma} \left| C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \right| d\xi < \delta.$$

Since

$$\begin{aligned}
&\|f_\varepsilon(\boldsymbol{x}) - f(\boldsymbol{x})\|_{L^p(R^m)} \\
&\quad = \left\| \int_R C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right\|_{L^p(R^m)} \\
&\quad \le \left\| \int_{|\xi|>\gamma} C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right\|_{L^p(R^m)} \\
&\qquad + \left\| \int_{|\xi|<\gamma} C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right\|_{L^p(R^m)}
\end{aligned}$$

holds, each right side term can be evaluated as

$$\begin{aligned}
&\left\| \int_{|\xi|>\gamma} C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right\|_{L^p(R^m)} \\
&\quad = \left(\int_{R^m} \left| \int_{|\xi|>\gamma} C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right|^p d\boldsymbol{x} \right)^{\frac{1}{p}} \\
&\quad \le \left(\int_{R^m} \prod_{i=1}^{p} \int_{|\xi_i|>\gamma} \left| C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi_i)} \hat{\phi}_c(\xi_i) \frac{1}{|\xi_i|^m} \right| \left| G_{\frac{\varepsilon}{|\xi_i|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right| d\xi_i d\boldsymbol{x} \right)^{\frac{1}{p}} \\
&\quad = \left(\prod_{i=1}^{p} \int_{|\xi_i|>\gamma} \left| C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi_i)} \hat{\phi}_c(\xi_i) \frac{1}{|\xi_i|^m} \right| \int_{R^m} \left| \prod_{i=1}^{p} G_{\frac{\varepsilon}{|\xi_i|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right| d\boldsymbol{x} d\xi_i \right)^{\frac{1}{p}} \\
&\quad \le \left(\prod_{i=1}^{p} \int_{|\xi_i|>\gamma} \left| C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi_i)} \hat{\phi}_c(\xi_i) \frac{1}{|\xi_i|^m} \right| \left(\prod_{i=1}^{p} \int_{R^m} \left| G_{\frac{\varepsilon}{|\xi_i|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right|^p d\boldsymbol{x} \right)^{\frac{1}{p}} d\xi_i \right)^{\frac{1}{p}} \\
&\quad = \left(\prod_{i=1}^{p} \int_{|\xi_i|>\gamma} \left| C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi_i)} \hat{\phi}_c(\xi_i) \frac{1}{|\xi_i|^m} \right| \prod_{i=1}^{p} \|G_{\frac{\varepsilon}{|\xi_i|^2}} * f - f\|_{L^p(R^m)} d\xi_i \right)^{\frac{1}{p}} \\
&\quad < \delta \int_{|\xi|>\gamma} \left| C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \right| d\xi \\
&\quad \le C' \delta
\end{aligned}$$

$$\left\| \int_{|\xi|<\gamma} C^{-1}_{\phi_d,\phi_c} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right\|_{L^p(R^m)}$$

$$
\begin{aligned}
&\le \left(\prod_{i=1}^{p} \int_{|\xi_i|<\gamma} \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi_i)} \hat{\phi}_c(\xi_i) \frac{1}{|\xi_i|^m} \right| \int_{R^m} \left| \prod_{i=1}^{p} G_{\frac{\varepsilon}{|\xi_i|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right| d\boldsymbol{x} d\xi_i \right)^{\frac{1}{p}} \\
&\le \left(\prod_{i=1}^{p} \int_{|\xi_i|<\gamma} \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi_i)} \hat{\phi}_c(\xi_i) \frac{1}{|\xi_i|^m} \right| \prod_{i=1}^{p} \| G_{\frac{\varepsilon}{|\xi_i|^2}} * f - f \|_{L^p(R^m)} d\xi_i \right)^{\frac{1}{p}} \\
&\le \left(\prod_{i=1}^{p} \int_{|\xi_i|<\gamma} \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi_i)} \hat{\phi}_c(\xi_i) \frac{1}{|\xi_i|^m} \right| d\xi_i \prod_{i=1}^{p} 2\|f\|_{L^p(R^m)} \right)^{\frac{1}{p}} \\
&< 2\delta \|f\|_{L^p(R^m)} \\
&\le C''\delta.
\end{aligned}
$$

The quantities

$$
C' = \int_R \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \right| d\xi, \quad C'' = 2\|f\|_{L^p(R^m)}
$$

are constants which do not depend on δ. Therefore

$$
\lim_{\varepsilon \to 0} f_\varepsilon(\boldsymbol{x}) = f(\boldsymbol{x})
$$

holds in the sense of L^p-norm.

Next, we prove in the case that f on R^m is bounded and uniformly continuous. In this case, as

$$
\begin{aligned}
&\|G_\varepsilon * f\|_{L^\infty(R^m)} \le \|f\|_{L^\infty(R^m)} \\
&\lim_{\varepsilon \to 0} \|G_\varepsilon * f - f\|_{L^\infty(R^m)} = 0
\end{aligned}
$$

hold, if we take a sufficiently small ε, for any δ's there exists γ such that

$$
\|G_{\frac{\varepsilon}{|\xi|^2}} * f - f\|_{L^\infty(R^m)} < \delta, \quad |\xi| > \gamma
$$

and

$$
\int_{|\xi|<\gamma} \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \right| d\xi < \delta.
$$

Similarly in the case of L^∞-norm

$$
\begin{aligned}
&\|f_\varepsilon(\boldsymbol{x}) - f(\boldsymbol{x})\|_{L^\infty(R^m)} \\
&\quad \le \left\| \int_{|\xi|>\gamma} C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right\|_{L^\infty(R^m)} \\
&\qquad + \left\| \int_{|\xi|<\gamma} C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right\|_{L^\infty(R^m)}
\end{aligned}
$$

holds, each right side term can again be evaluated as

$$
\begin{aligned}
&\left\| \int_{|\xi|>\gamma} C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right\|_{L^\infty(R^m)} \\
&\quad = \sup_{\boldsymbol{x}} \left| \int_{|\xi|>\gamma} C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right| \\
&\quad \leq \int_{|\xi|>\gamma} \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \right| \sup_{\boldsymbol{x}} \left| G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right| d\xi \\
&\quad = \int_{|\xi|>\gamma} \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \right| \left\| G_{\frac{\varepsilon}{|\xi|^2}} * f - f \right\|_{L^\infty(R^m)} d\xi \\
&\quad < \delta \int_{|\xi|>\gamma} \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \right| d\xi \\
&\quad \leq C' \delta
\end{aligned}
$$

$$
\begin{aligned}
&\left\| \int_{|\xi|<\gamma} C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \left(G_{\frac{\varepsilon}{|\xi|^2}} * f(\boldsymbol{x}) - f(\boldsymbol{x}) \right) d\xi \right\|_{L^\infty(R^m)} \\
&\quad \leq \int_{|\xi|<\gamma} \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \right| \left\| G_{\frac{\varepsilon}{|\xi|^2}} * f - f \right\|_{L^\infty(R^m)} d\xi \\
&\quad \leq \int_{|\xi|<\gamma} \left| C_{\phi_d,\phi_c}^{-1} \overline{\hat{\phi}_d(\xi)} \hat{\phi}_c(\xi) \frac{1}{|\xi|^m} \right| 2 \left\| f \right\|_{L^\infty(R^m)} d\xi \\
&\quad < \delta 2 \left\| f \right\|_{L^\infty(R^m)} \\
&\quad \leq C'' \delta.
\end{aligned}
$$

Therefore, if f on R^m is bounded and uniformly continuous,

$$\lim_{\varepsilon \to 0} f_\varepsilon(\boldsymbol{x}) = f(\boldsymbol{x})$$

converges uniformly. ∎

B An Example of a Decomposing Kernel

A decomposing kernel

$$\phi_d(z) = \begin{cases} \dfrac{d^m}{dz^m} \rho(z) & m : \text{even} \\ \dfrac{d^{m+1}}{dz^{m+1}} \rho(z) & m : \text{odd} \end{cases}$$

can be confirmed to satisfy the admissible condition (2) as follows, where m is the dimensionality of input $\boldsymbol{x}$ and ρ is given by

$$\rho(z) = \begin{cases} e^{-1/(1-|z|^2)} & |z| < 1 \\ 0 & |z| \geq 1. \end{cases}$$

When m is an even number,

$$\begin{aligned} &\int_R |\omega|^{-m} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) d\omega \\ &\quad = \int_R |\omega|^{-m} \overline{(i\omega)^m \hat{\rho}(\omega)} \hat{\phi}_c(\omega) d\omega \\ &\quad = (-1)^{\frac{m}{2}} \int_R \overline{\hat{\rho}(\omega)} \hat{\phi}_c(\omega) d\omega \\ &\quad = (-1)^{\frac{m}{2}} \int_R \rho(z) \phi_c(z) dz \end{aligned}$$

holds. From the fact that ρ has a bounded support, this integral clearly converges.

When m is an odd number,

$$\begin{aligned} &\int_R |\omega|^{-m} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) d\omega \\ &\quad = \int_R |\omega|^{-m} \overline{(i\omega)^{m+1} \hat{\rho}(\omega)} \hat{\phi}_c(\omega) d\omega \\ &\quad = (-1)^{\frac{m+1}{2}} \int_R \overline{\hat{\rho}(\omega)} |\omega| \hat{\phi}_c(\omega) d\omega \\ &\quad = (-1)^{\frac{m+1}{2}} \int_R \rho(z) \phi^+(z) dz \end{aligned}$$

holds, where $\phi^+(z)$ denotes the inverse Fourier transform of $|\omega|\hat{\phi}_c(\omega)$. The existence of $\phi^+(z)$ depends on the smoothness of ϕ_c. In this case, a sufficient condition for the existence of ϕ^+ is that ϕ_c has a second order derivative and $\hat{\phi}_c$ is bounded. If the inverse Fourier transform $\phi^+(z)$ exists, this integral clearly converges since ρ has a bounded support.

C Another Bound by Bell-shaped Networks

Using any $c(\boldsymbol{a}, b)$ such that $\mathrm{Re}\, T(\boldsymbol{a}, b)/c(\boldsymbol{a}, b)$ is a positive real number and

$$\int \frac{\mathrm{Re}\, T(\boldsymbol{a}, b)}{c(\boldsymbol{a}, b)} d\boldsymbol{a} db = 1,$$

and defining $p(\boldsymbol{a}, b) = \mathrm{Re}\, T(\boldsymbol{a}, b)/c(\boldsymbol{a}, b)$, we can obtain the inequality

$$\begin{aligned} & E\left[\int (g_n(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \mu(\boldsymbol{x}) d\boldsymbol{x}\right] \\ = & \int E\left[(c(\boldsymbol{a}, b) g_n(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right] \mu(\boldsymbol{x}) d\boldsymbol{x} \\ \leq & \frac{1}{n} \int E\left[(c(\boldsymbol{a}, b) \phi_c(\boldsymbol{a} \cdot \boldsymbol{x} - b))^2\right] \mu(\boldsymbol{x}) d\boldsymbol{x} \\ = & \frac{1}{n} E\left[\int (c(\boldsymbol{a}, b) \phi_c(\boldsymbol{a} \cdot \boldsymbol{x} - b))^2 \mu(\boldsymbol{x}) d\boldsymbol{x}\right] \\ \leq & \frac{1}{n} \sup_{\boldsymbol{a}, b} |c(\boldsymbol{a}, b)|^2 \|\phi_c(\boldsymbol{a} \cdot \boldsymbol{x} - b)\|^2_{L^2(R^m), \mu}. \end{aligned}$$

In this case the order of the approximation error is equal to the previous evaluation, that is $O(1/n)$, but the integrability condition of the transform T is reduced from $\int |\mathrm{Re}\, T(\boldsymbol{a}, b)| < \infty$ to

$$\int \frac{\mathrm{Re}\, T(\boldsymbol{a}, b)}{c(\boldsymbol{a}, b)} d\boldsymbol{a} db < \infty. \tag{13}$$

Condition (13) is easily seen to be weaker than $\int |\mathrm{Re}\, T(\boldsymbol{a}, b)| < \infty$. For larger $|\boldsymbol{a}|$, the norm $\|\phi_c(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i)\|^2_{L^2(R^m), \mu}$ becomes smaller and therefore $c(\boldsymbol{a}, b)$ can also take a larger value.

D Proof of Theorem 3

Let us assume that ϕ_c is in $C^2(R)$. Since $\hat{\phi}_c$ is bounded,

$$\int |\omega| \hat{\phi}_c(\omega) d\omega < \infty$$

holds, and we define ϕ_d

$$\phi_d(z) = \begin{cases} \dfrac{d^{m+2}}{dz^{m+2}} \rho(z) & m \text{ is even} \\ \dfrac{d^{m+1}}{dz^{m+1}} \rho(z) & m \text{ is odd} \end{cases}$$

with ρ from equation (7). Noting that ϕ_c is twice differentiable, we can see that ϕ_d and ϕ_c satisfy the admissible condition (2) in a manner similar to section 3.2.

In this case,

$$\int z^k \phi_d(z) dz = \left(-i\frac{d}{d\omega}\right)^k \hat{\phi}_d(\omega)\Big|_{\omega=0} \tag{14}$$

holds, hence the k-th moment as calculated in equation (14) vanishes if m is odd for $k \leq m$, and if m is even for $k \leq m+1$. Namely, the moments vanish up to the m-th order at least.

Let f be in $C_0^m(R^m)$, where $C_0^m(R^m)$ denotes the class of functions with compact supports belonging to $C^m(R^m)$. Let $f_e(z)$ be

$$f_e(z) = \int_{e\cdot x=z} f(\boldsymbol{x}) d\boldsymbol{x}^{(m-1)}$$

for a unit vector $\boldsymbol{e} \in R^m$, where $d\boldsymbol{x}^{(m-1)}$ denotes the volume element on the hyper-plane $\boldsymbol{e}\cdot\boldsymbol{x} = z$. We can rewrite the transform T as

$$\begin{aligned} T_e(a,b) &\equiv T(\boldsymbol{a},b) \\ &= \int \phi_d(\boldsymbol{a}\cdot\boldsymbol{x} - b) f(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int \phi_d(az-b) f_e(z) dz, \end{aligned}$$

where $a = |\boldsymbol{a}|$ and $\boldsymbol{e} = \boldsymbol{a}/a$. We assume that, for any $\boldsymbol{e}$, there exists some $\alpha \quad (0 < \alpha \leq 1)$ such that

$$\begin{aligned} & f_e(z) \in C_0^m(R) \\ & |f_e^{(m)}(z) - f_e^{(m)}(z')| < M|z-z'|^\alpha, \quad M \in R. \end{aligned} \tag{15}$$

Then

$$\begin{aligned} & |T_e(a,b)| \\ &= \left|\frac{1}{a}\int \phi_d(z) f_e\left(\frac{z+b}{a}\right) dz\right| \\ &= \left|\frac{1}{a}\int \phi_d(z) \left\{\sum_{k=0}^{m-1} \frac{1}{k!}\left(\frac{z}{a}\right)^k f_e^{(k)}\left(\frac{b}{a}\right) + \frac{1}{m!}\left(\frac{z}{a}\right)^m f_e^{(m)}\left(\frac{\theta z+b}{a}\right)\right\} dz\right| \\ &= \left|\frac{1}{a}\int \phi_d(z) \frac{1}{m!}\left(\frac{z}{a}\right)^m \left\{f_e^{(m)}\left(\frac{\theta z+b}{a}\right) - f_e^{(m)}\left(\frac{b}{a}\right)\right\} dz\right| \\ &\leq \frac{1}{a}\left|\int \frac{M}{m!}\left(\frac{z}{a}\right)^m \left|\frac{z}{a}\right|^\alpha \phi_d(z) dz\right| \\ &\leq \frac{C_1}{a^{m+1+\alpha}}, \end{aligned}$$

where θ satisfies $|\theta| \leq 1$ and C_1 is a constant. Since $T_e(a,b)$ is bounded, we can obtain an integrable inequality

$$|T_e(a,b)| \leq \frac{C_2}{1+a^{m+1+\alpha}},$$

where C_2 is a constant.

From the fact that f and ϕ_d have compact supports, $T(\boldsymbol{a},b)$ also has a compact support with respect to b when $\boldsymbol{a}$ is fixed. The size of the support is represented as $C_3a + C_4$ for appropriate C_3 and C_4. Let $d\Theta$ be a volume element of a unit hyper-sphere on R^m. C_T is rewritten as

$$C_T = \int |T_e(a,b)| a^{m-1} da d\Theta db,$$

and we see that

$$\int |T_e(a,b)| a^{m-1} da d\Theta db \leq C_5 \int \frac{a^{m-1} C_2 (C_3 a + C_4)}{1+a^{m+1+\alpha}} da < \infty$$

holds, where C_5 is a definite integral with respect to $d\Theta$. Therefore, equation (15) is a sufficient condition of integrability of the transform T.

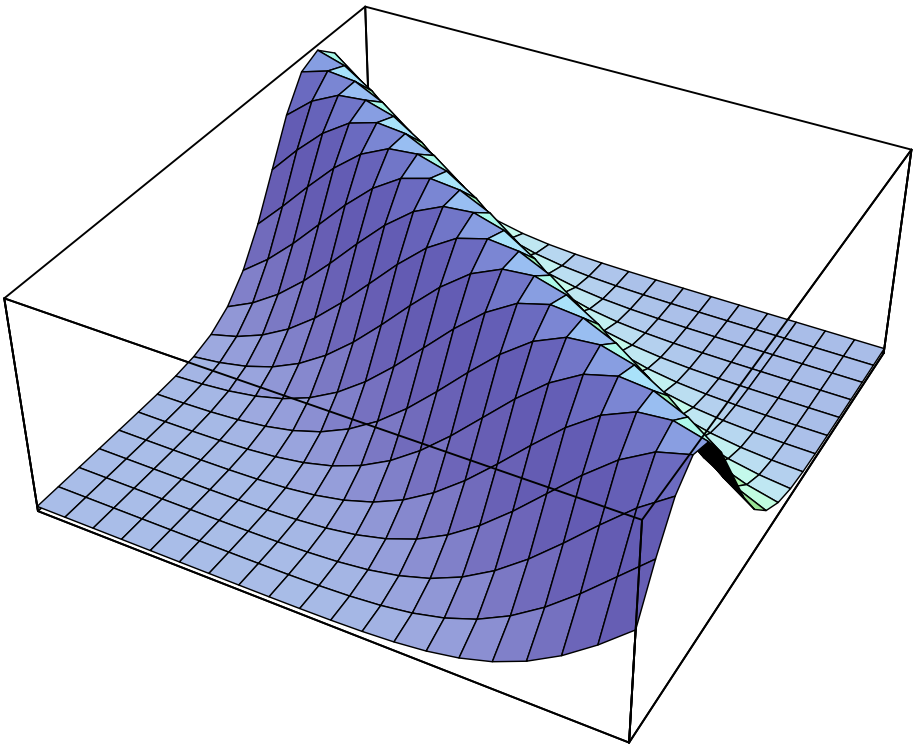

Figure 1: An example of the ridge function on R^2.

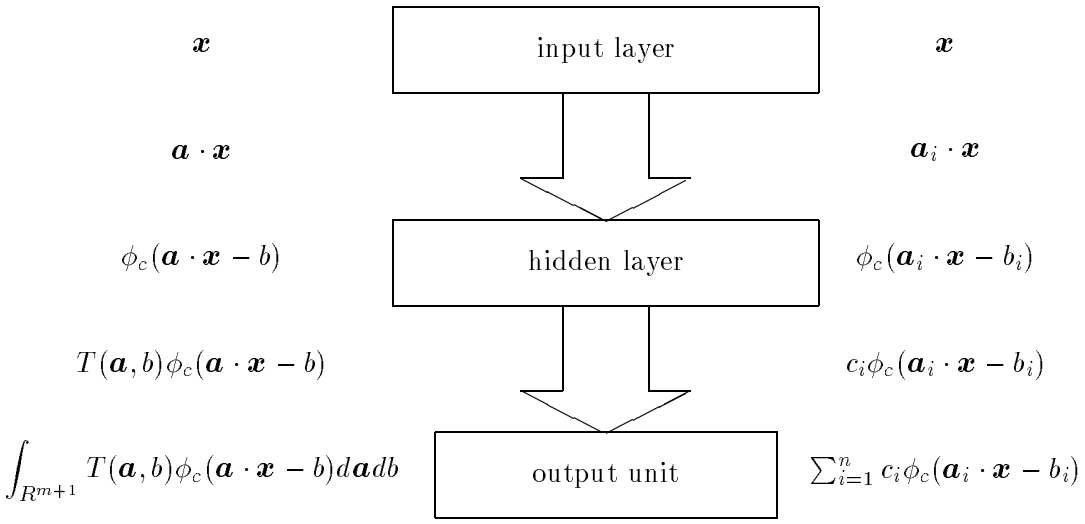

Figure 2: The relationship between the transform $T(\boldsymbol{a}, b)$ and the three-layered network.

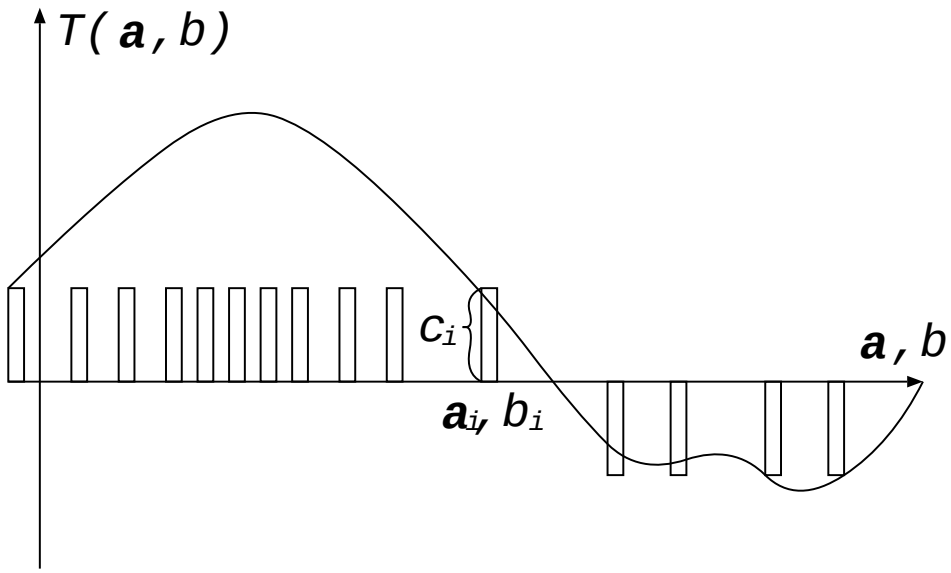

Figure 3: The relationship between the transform $T(\boldsymbol{a}, b)$ and the density of $(\boldsymbol{a}, b)$