CONSTRUCTION OF NEURAL NETS USING THE RADON TRANSFORM[1]

S. M. Carroll and B. W. Dickinson
Department of Electrical Engineering
Princeton University
Princeton, New Jersey 08544

ABSTRACT

This paper presents a method for constructing a feedforward neural net implementing an arbitrarily good approximation to any $\mathcal{L}_2$ function over $[-1,1]^n$. The net uses n input nodes, a single hidden layer whose width is determined by the function to be implemented and the allowable mean square error, and a linear output neuron. Error bounds and an example are given for the method.

This paper is concerned with the problem of analyzing and constructing neural nets with a single hidden layer of sigmoidal nodes. These nets are of particular interest because, as shown recently by Cybenko [1], they are capable of realizing essentially arbitrary net input-output mappings (*i.e.* functions from the n-dimensional cube $[-1,1]^n$ to the interval $(-1,1)$). The limitations of networks with no hidden layers, perceptrons, have long been known, so Cybenko's result is the best that can be achieved. In fact, Cybenko's result is stated for modified networks with a linear output neuron rather than the usual sum-and-sigmoid node, and this allows implementation of real-valued functions. Such linear-output nets are more suitable for applications requiring continuous-valued mappings [2].

Cybenko's proof of this remarkable approximation result is an existential, not constructive one. This paper provides an explicit constructive proof of a similar approximation theorem by developing a method for specifying the weights of a net that estimates a given $\mathcal{L}_2$ function to within a given mean square error. The construction can also be applied to the situation in which it is desired to initialize some learning algorithm in a way that reflects certain *a priori* partial information about the input-output function.

The heart of the derivation comes from recognizing that the functional form of a linear output net with a single hidden layer is a finitely parameterized, approximate form of the back-projection operator, a component of the inverse Radon transform. The Radon transform is commonly used in medical and geophysical imaging, and its inverse is the basis of CAT scan image reconstruction. For an investigation of the properties and applications of the Radon transform, see [3] and [4]. The transform represents a function exactly by the set of all possible integrals over hyperplanes in $\mathbb{R}^n$, which are indexed by the unit normal vector to the hyperplane, $\mathbf{u}$, and by the least distance of the hyperplane from the origin, α. Thus R takes a function f over $\mathbb{R}^n$ to a function $\hat{f}$ over $\mathbb{R} \times \mathbf{S}^{n-1}$ (where $\mathbf{S}^{n-1}$ is the unit sphere in n dimensional Euclidean space). The inverse Radon transform is usually expressed as the composition of two operators, $R^{-1} = B \circ F$. The first to be applied, F, takes $(n-1)$ derivatives (and for n even, a Hilbert Transform) with respect to α, producing the *filtered back-projection data*, denoted here as $F(\hat{f}) = h(\alpha, \mathbf{u})$. B, the back-projection operator, takes the filtered Radon Transform back to the original space:

$$f(\mathbf{x}) = Bh(\alpha, \mathbf{u}) = \int_{\|\mathbf{u}\|_2=1} h(\mathbf{u}\cdot\mathbf{x}, \mathbf{u}) d\mu(\mathbf{u}),$$

where $d\mu(\mathbf{u})$ is a unit of surface area on the unit sphere, $\mathbf{S}^{n-1}$. If B is discretized in the angle variable $\mathbf{u}$ we have

$$f(\mathbf{x}) = Bh \approx \sum_{i=1}^{K} \mu_i h(\mathbf{u}_i \cdot \mathbf{x}, \mathbf{u}_i).$$

If it is possible to represent each of the K terms above by a linear combination of sigmoids,

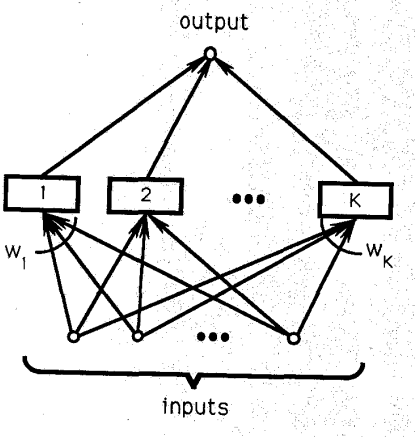
$$h_i = h(\mathbf{u}_i \cdot \mathbf{x}, \mathbf{u}_i) = \sum_{j=1}^{m_i} \alpha_{ij} \sigma(\mathbf{u}_i \cdot \mathbf{x} + \beta_{ij})$$

then this is precisely the functional form of the linear output neural net with one hidden layer:

$$O(\mathbf{x}) = \sum_{i=1}^{N} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$$

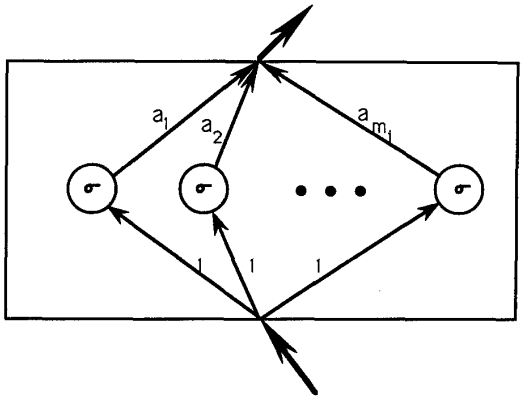
where $\mathbf{w}_i$ is the vector of weights connecting the inputs to the ith hidden unit, and a_i is the weight connecting that unit to the output.

A neural net representing $f = \sum_{i=1}^{K} h_i$ is



[1]This work was supported in part by the HP Faculty Development Program and by grant AFOSR-88-0227 from the Air Force Office of Scientific Research

where each numbered block has structure



So the N nodes of the hidden layer are arranged in K blocks, one for each of the angles, or normal vectors, $\mathbf{u}_i$ of the discretized transform. Each of the blocks has an output that represents the corresponding back projection function h_i by a sum of m_i sigmoids. Setting $K = N$ and $m_i = 1 \quad (i = 1, \ldots, N)$ gives the fully-connected single hidden layer structure. Thus, any one-hidden-layer neural net can be interpreted as approximating the discretized back-projection of some function, Bh.

To claim that a net may be accurately synthesized using this approach requires investigation of the errors produced by the two approximations made in the general case: that the discretized transform approaches the value of the continuous version, and that the resulting transform is well-behaved for small deviations of $h(\mathbf{u}_i \cdot \mathbf{x}, \mathbf{u}_i)$ so that approximation of h by a finite sum of sigmoids causes small deviations in the output function. Helgason has shown that the Radon Transform carried out at a finite set of angles has a nontrivial null space. This implies that some features of a function can be arbitrarily large, yet will be neither detected nor reconstructed by the discretized transform and inverse method given here. However, Helgason also shows that any infinite set of angles is sufficient to determine a compactly supported function with a Radon Transform, and that there exists a sequence of transforms with finitely many angles converging to any such function. This underlies the proof of correctness for this method: convergence implies that a net of sufficiently large size can achieve any nonzero error bound. The error is bounded explicitly by Theorem 1 of the appendix. The bound is quite loose, and requires knowledge of the complete Radon transform, but its form is amenable to asymptotic analysis and reveals the effect of input dimension on performance.

Theorem 1 also provides a basis for tracing the effect of small errors in the discretized back-projection function $h(\alpha, \mathbf{u}_i)$. This is the subject of Theorem 2, and the result is encouraging; the errors produced by discretization and approximation of $h(\alpha, \mathbf{u}_i)$ add in RMS, with a constant dependent only on the dimension of the input space. The method proposed is therefore capable of representing any function with an arbitrarily small error by some finite set of parameters, provided that function is sufficiently smooth. It remains only to show that sigmoids can be used for such a parameterization, and to extend the result to less-smooth functions.

That continuous functions are constructed from step functions, which are approximated well by sigmoids, leads us to expect that we should be able to construct the necessary approximations to h_i, and the idea that "all functions are nearly continuous" suggests that smoothness is not a unreasonable constraint for our approximating set. Both of these are the case, as is proven in Theorem 3, which states essentially that "All $\mathcal{L}_2$ functions with compact support can be approximated arbitrarily well in the $\mathcal{L}_2$ norm by a single hidden layer neural net with finitely many nodes."

So we have established an approximation result like Cybenko's using an alternative method, one which constructs a function whose form can be directly implemented on a single hidden layer net. By smoothing the original function f, taking its Radon transform $\hat{f}$ at finitely many angles, and filtering each to find h at those angles, we perform the work of a learning algorithm analytically. We then approximate h with a linear combination of sigmoids. The angles $\mathbf{u}_i$ chosen are the input weights for one block, and the weights of the linear combination approximating h_i are the output weights for the same block. The usual structure of a net simply uses many angles and a very crude approximation to each filtered back-projection. This amounts to taking advantage of the smoothness of h to approximate the average behavior of the filtered projections rather than approximating a single sample of the filtered back-projection data more accurately.

For functions with multiple outputs, these results guarantee existence of an implementation on nets with a single hidden layer: it is sufficient to produce one net dedicated to each output, and run all concurrently. The usual fully-connected structure allows sharing of intermediate results in the calculation, sometimes greatly reducing the size of the net required to approximate the outputs collectively.

Application of the method of this proof is straightforward. We will give an example problem, that of constructing a net to produce a continuous form of the three variable parity function. The necessary smoothness is provided by replacing the value ± 1 at each corner of the 3-cube by a Gaussian peak centered at that corner and scaled to have the correct maximum. The Radon transform of such a function can be taken analytically fairly easily by use of the scaling and translation properties of the transform, and the transform of the Gaussian centered at the origin [see 3]. The form of f is

$$f(\mathbf{x}) = \sum_{n=1}^{8} (-1)^n e^{-\|\mathbf{x}-\mathbf{v}_n\|_2^2}$$

where

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

$$\mathbf{v}_4 = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} \mathbf{v}_5 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \mathbf{v}_6 = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}$$

$$\mathbf{v}_7 = \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix} \mathbf{v}_8 = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}.$$

The Radon transform of f is

$$\hat{f}(\alpha, \mathbf{u}) = \pi \sum_{n=1}^{8} (-1)^n e^{-(\alpha - \mathbf{v}_n \cdot \mathbf{u})^2}.$$

and the back-projection data $h(\alpha, \mathbf{u}) = \frac{-1}{8\pi^2} \frac{\partial^2}{\partial \alpha^2} \hat{f}$ is

$$h(\alpha, \mathbf{u}) = \frac{-1}{4\pi} \sum_{n=1}^{8} (-1)^n (2(\alpha - \mathbf{u} \cdot \mathbf{v}_n)^2 - 1) e^{-(\alpha - \mathbf{v}_n \cdot \mathbf{u})^2}.$$

Using the projection angles in (Appendix (5)) with $k = 3$ we need to evaluate h at

$$\mathbf{u} \in \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/2 \\ \pm\sqrt{3}/2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/2 \\ 0 \\ \pm\sqrt{3}/2 \end{pmatrix}, \begin{pmatrix} 1/4 \\ \pm\sqrt{3}/4 \\ \pm\sqrt{3}/2 \end{pmatrix} \right\}.$$

Due to the symmetry of our example $h(\alpha, \mathbf{u})$ is zero for all α if $\mathbf{u}$ has any element 0, which makes evaluation at five of the nine angles trivial. Evaluating h at each of the remaining four angles gives the same function of α except for sign. Let us label these four vectors as $\mathbf{u}_1, \ldots, \mathbf{u}_4$ so that $h(\alpha, \mathbf{u}_1) = -h(\alpha, \mathbf{u}_2) = h(\alpha, \mathbf{u}_3) = -h(\alpha, \mathbf{u}_4)$. To approximate h by sigmoids, we choose the usual back-propagation sigmoid scaled by 2, and centered at the origin

$$\sigma(\alpha) = \frac{2}{1 + e^{-\alpha}} - 1$$

A sum of three such sigmoids gives a good approximation to h, so our net will be based on four three-node blocks. The block for the projection from direction $\mathbf{u}_2 = [1/4\ \sqrt{3}/4\ \sqrt{3}/2]^T$ has function

$$\eta(\alpha) = .22\sigma(3\alpha - 3.2) + .22\sigma(3\alpha + 3.2) - .32\sigma(3\alpha)$$

so the function of the 12-node network is

$$\tilde{f}(\mathbf{x}) = \frac{4\pi}{9} \sum_{n=1}^{4} (-1)^n \eta(\mathbf{x} \cdot \mathbf{u}_n)$$

and has a mean square error of about 0.013, relative to a function with a mean square value of 0.13. Using h in place of η gives an error of 0.011, implying that the number of projections is the critical parameter rather than the sigmoidal approximation of h. The qualitative character of the original function is reproduced: each octant retains a distinctive sign, and points on the boundary between octants carry an output value of 0, to within machine precision limits. The only adaptive step used in this process was fine-tuning η, and this illustrates the power of the construction by reducing the approximation of a continuous function in 3-space to the approximation of a scalar function by three sigmoids. In fact, for this example, it is possible to produce a reasonable estimate η by inspection, and as already noted the small change in error due to approximation of h suggests that in this example the critical parameter is the number of projections rather than the quality of η.

The method proposed here serves not only as a proof of the constructiblity of single-hidden-layer nets that perform to any given specification of error and function, but provides an algorithm for reducing the design of such nets to the design of segments of real-valued functions of a single variable. There are efficient solutions to this simpler problem for some classes of sigmoids, so we regard this problem as well-solved.

Bibliography

[1] CYBENKO, G., "Approximation by Superpositions of a Sigmoidal Function," *to appear, Math. Control Sytems Signals.*

[2] LAPEDES, A., FARBER, R. "Nonlinear Signal Processing Using Neural Networks: Prediction and System Modeling," *to appear, Proc. IEEE.*

[3] DEANS, S. R., *The Radon Transform and Some of Its Applications* , Wiley, 1983.

[4] HELGASON, S., *The Radon Transform*, Birkhauser, 1980.

[5] MUKHERJEA, A., POTHOVEN, K., *Real and Functional Analysis, Part A: Real Analysis*, Plenum, 1984.

[6] TITCHMARSH, E. *Introduction to the Theory of Fourier Integrals*, Clarendon, 1937.

Appendix

THE RADON TRANSFORM IN $\mathbb{R}^n$

For $f \in \mathcal{D} = \mathcal{C}^\infty(\mathcal{S})$, the space of functions continuously differentiable to all orders taking nonzero values only on a compact subset $\mathcal{S}$ of $\mathbb{R}_n$, we define the *Radon Transform* of $f(\mathbf{x})$ as

$$\hat{f}(\alpha, \mathbf{u}) = \int_{\mathbf{u}\cdot\mathbf{x}=\alpha} f(\mathbf{x}) d\mu(\mathbf{x}) \tag{1}$$

where $\mu(\cdot)$ is the measure of area on a $(n-1)$ dimensional differential manifold, here a hyperplane in $\mathbb{R}^n$. The inversion formula uses μ on the unit sphere (measured by the Euclidean norm throughout this paper)

$$f(\mathbf{x}) = \int_{\|\mathbf{u}\|=1} h(\mathbf{x}\cdot\mathbf{u}, \mathbf{u}) d\mu(\mathbf{u}) \tag{2}$$

and we define the back-projection data h using $\mathcal{H}_\alpha$ to represent the Hilbert transform with respect to α (*i. e.* convolution with $1/\alpha$ holding $\mathbf{u}$ constant):

$$h(\alpha, \mathbf{u}) = \begin{cases} \frac{1}{2(2\pi i)^{n-1}} \left(\frac{\partial}{\partial\alpha}\right)^{n-1} \hat{f}(\alpha, \mathbf{u}) & n \text{ odd}, \\ \frac{1}{(2\pi i)^n} \mathcal{H}_\alpha \left[\left(\frac{\partial}{\partial\alpha}\right)^{n-1} \hat{f}(\alpha, \mathbf{u}) \right] & n \text{ even}. \end{cases} \tag{3}$$

We are interested in a discretized form of the inversion formula valid over $[-1, 1]^n$, so we define

$$\omega(\mathbf{u}) = (r, \theta_1, \ldots, \theta_{n-1}) \tag{4}$$

to be the hyperspheroidal coordinates of $\mathbf{u}$, and use ω to partition the half-sphere

$$\mathcal{U} = \{\|\mathbf{u}\| = 1\} \cap \{\mathbf{u}\cdot\mathbf{e}_1 > 0\}$$

into k^{n-1} sets $\mathcal{U}(i)$, $i = 1, 2, \ldots, k^{n-1}$:

$$\begin{aligned} \mathcal{U}_i &= \left\{ \mathbf{u} : \omega(\mathbf{u}) = \left(\omega(\mathbf{u}_i) + \begin{bmatrix} 0 \\ \theta \end{bmatrix} \right), \theta \in \left[\frac{-\pi}{k}, \frac{\pi}{k} \right]^{n-1} \right\} \\ \omega(\mathbf{u}_i) &= \left\{ \begin{bmatrix} 1 \\ \theta \end{bmatrix} : \theta \in \left\{ -\tfrac{k-1}{k}\pi, -\tfrac{k-3}{k}\pi, \ldots, \tfrac{k-1}{k}\pi \right\}^{n-1} \right\} \end{aligned} \tag{5}$$

$$\mathbf{u}_i \neq \mathbf{u}_j \quad (i \neq j)$$

It is not necessary to partition the entire sphere because $\hat{f}(\alpha, \mathbf{u}) = \hat{f}(-\alpha, -\mathbf{u})$.

We use (5) to rewrite (2) as

$$f = 2 \sum_{i=1}^{k^{n-1}} \int_{\mathcal{U}_i} h(\mathbf{x}\cdot\mathbf{u}, \mathbf{u}) d\mu(\mathbf{u}). \tag{6}$$

We now approximate the ith term ($i = 1, \ldots, k^{n-1}$) in the sum by a function requiring evaluation only at $\mathbf{u}_i$. This defines a function $\bar{f}(\mathbf{x})$, the *Discrete-in-Angle Inverse Radon Transform* of $\hat{f}(\alpha, \mathbf{u})$, or the *Discrete-in-Angle Backprojection* of $h(\alpha, \mathbf{u})$. For $\hat{f}$ sufficiently smooth, we can analytically bound the difference between f and $\bar{f}$ at any point. The Paley-Weiner theorem for Radon transforms assures us that if $f \in \mathcal{D}$ then $\hat{f} \in \mathcal{D}$, so we may proceed to bound the error due to discretization:

THEOREM 1

Let $f : \mathbb{R}^n \to \mathbb{R}$ be in $\mathcal{D}$. Assume the closed ball $\mathcal{B}$ of radius $r_{\mathcal{B}}$ centered at the origin contains the set $\mathcal{S} = \{\mathbf{x} : f(\mathbf{x}) \neq 0\}$. Let $\hat{f}$ be the Radon transform of f and let $h(\alpha, \mathbf{u})$ be the corresponding back-projection data. Define

$$\bar{f}(\mathbf{x}) = 2 \sum_{i=1}^{k^{n-1}} \mu(\mathcal{U}_i) h(\mathbf{u}_i \cdot \mathbf{x}, \mathbf{u}_i)$$

where $\mathbf{u}_i$ is defined as in (5). Then

$$|f(\mathbf{x}) - \bar{f}(\mathbf{x})| \leq E_1 = \frac{2}{k} \left(\frac{\pi^{\left(\frac{n+3}{2}\right)} \sqrt{(r_{\mathcal{B}}+1)n}}{\Gamma\left(\frac{n+1}{2}\right)} \right) \max_{-1 \leq \alpha \leq 1} \max_{\mathbf{u} \in \mathcal{U}} \|\nabla h(\alpha, \mathbf{u})\|$$

PROOF

$$\begin{aligned} |f(\mathbf{x}) - \bar{f}(\mathbf{x})| &\leq 2 \sum_{i=1}^{k^{n-1}} \int_{\mathcal{U}_i} |h(\mathbf{u} \cdot \mathbf{x}, \mathbf{u}) - h(\mathbf{u}_i \cdot \mathbf{x}, \mathbf{u}_i)| \, d\mu(\mathbf{u}) \\ &\leq 2\mu(\mathcal{U}) \max_{\alpha} \max_{\mathcal{U}} \|\nabla h\| \max_{i} \max_{\mathbf{u} \in \mathcal{U}_i} \left\| \begin{bmatrix} (\mathbf{u} - \mathbf{u}_i) \cdot \mathbf{x} \\ (\mathbf{u} - \mathbf{u}_i) \end{bmatrix} \right\| \\ &\leq 2\mu(\mathcal{U}) \max_{\alpha} \max_{\mathcal{U}} \|\nabla h\| \sqrt{r_{\mathcal{B}}+1} \max_{i} \max_{\mathbf{u} \in \mathcal{U}_i} \|\mathbf{u} - \mathbf{u}_i\| \end{aligned}$$

■

The surface area of the unit sphere and hence the value of $2\mu(\mathcal{U})$, is

$$2\mu(\mathcal{U}) = \Omega_n = \frac{2\pi^{\left(\frac{n+1}{2}\right)}}{\Gamma\left(\frac{n+1}{2}\right)} \tag{7}$$

and the greatest distance between $\mathbf{u} \in \mathcal{U}_i$ and $\mathbf{u}_i$ is

$$\max_{i} \max_{\mathbf{u} \in \mathcal{U}_i} \|\mathbf{u} - \mathbf{u}_i\| \leq \left(\frac{\pi\sqrt{n}}{k} \right) \tag{8}$$

so the pointwise error bound is as given in the Theorem.

We have shown that we can represent approximately an n-dimensional function on a compact set with finitely many scalar functions on intervals, which can of course be represented approximately by a finite number of parameters. Can we then represent the n-dimensional function with finitely many parameters? Let us assume that $h(\alpha, \mathbf{u})$ is approximated by some easily represented function $\eta(\alpha, \mathbf{u})$. Let

$$\tilde{f}(\mathbf{x}) = 2 \sum_{i=1}^{k^{n-1}} \mu(\mathcal{U}_i) \eta(\mathbf{x} \cdot \mathbf{u}_i, \mathbf{u}_i) \tag{9}$$

The following theorem investigates the mean square error of the approximation $\tilde{f}(\mathbf{x})$.

THEOREM 2

Assume the hypotheses of Theorem 1. Let $\tilde{f}$ be defined by (9). Assume that for $i = 1, \ldots, k^{n-1}$

$$\int_{-r_{\mathcal{B}}}^{r_{\mathcal{B}}} [h(\alpha, \mathbf{u}) - \eta(\alpha, \mathbf{u})]^2 d\alpha \leq \Sigma^2 \tag{10}$$

Then the mean square error

$$MSE = \frac{1}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} [f(\mathbf{x}) - \tilde{f}(\mathbf{x})]^2 d\mathbf{x} \tag{11}$$

(where we have used $V_n(r_{\mathcal{B}}) = \Omega_{n-1}/n = 2\pi^{\frac{n}{2}}/n\Gamma(\frac{n}{2})$ for the volume of an n-sphere of radius $r_{\mathcal{B}}$) is bounded above by

$$MSE \leq \left(E_1 + \mu(\mathcal{U}) \Sigma \sqrt{2 \frac{V_{n-1}(r_{\mathcal{B}})}{V_n(r_{\mathcal{B}})}} \right)^2 \tag{12}$$

PROOF

$$\begin{aligned} MSE &= \frac{1}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} [(f(\mathbf{x}) - \bar{f}(\mathbf{x})) + (\bar{f}(\mathbf{x}) - \tilde{f}(\mathbf{x}))]^2 d\mathbf{x} \\ &\leq \frac{1}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} [f(\mathbf{x}) - \bar{f}(\mathbf{x})]^2 d\mathbf{x} \\ &\quad + \frac{1}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} [\bar{f}(\mathbf{x}) - \tilde{f}(\mathbf{x})]^2 d\mathbf{x} \qquad (13) \\ &\quad + \frac{2}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} [f(\mathbf{x}) - \bar{f}(\mathbf{x})][\bar{f}(\mathbf{x}) - \tilde{f}(\mathbf{x})] d\mathbf{x} \end{aligned}$$

Now the first term of (13) is easily evaluated in terms of Theorem 1:

$$\begin{aligned} \frac{1}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} [f(\mathbf{x}) - \bar{f}(\mathbf{x})]^2 d\mathbf{x} &= \frac{1}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} |f(\mathbf{x}) - \bar{f}(\mathbf{x})|^2 d\mathbf{x} \\ &\leq \frac{1}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} E_1^2 d\mathbf{x}. \qquad (14) \end{aligned}$$

The second term can be bounded by Jensen's inequality applied to the definitions of $\bar{f}$ and $\tilde{f}$:

$$\begin{aligned} &\frac{1}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} [\bar{f}(\mathbf{x}) - \tilde{f}(\mathbf{x})]^2 d\mathbf{x} \\ &= \frac{2}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} [\sum_{i=1}^{k^{n-1}} \mu(\mathcal{U}_i) h(\mathbf{x} \cdot \mathbf{u}_i, \mathbf{u}_i) \\ &\qquad - \sum_{i=1}^{k^{n-1}} \mu(\mathcal{U}_i) \eta(\mathbf{x} \cdot \mathbf{u}_i, \mathbf{u}_i)]^2 d\mathbf{x} \\ &\leq \frac{2\mu(\mathcal{U})}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} \sum_{i=1}^{k^{n-1}} \mu(\mathcal{U}_i) [h(\mathbf{x} \cdot \mathbf{u}_i, \mathbf{u}_i) - \eta(\mathbf{x} \cdot \mathbf{u}_i, \mathbf{u}_i)]^2 d\mathbf{x} \\ &\leq 2\mu(\mathcal{U}) \sum_{i=1}^{k^{n-1}} \mu(\mathcal{U}_i) \int_{-r_{\mathcal{B}}}^{r_{\mathcal{B}}} [h(\alpha, \mathbf{u}_i) - \eta(\alpha, \mathbf{u}_i)]^2 \\ &\qquad \times \left(\frac{1}{V_n(r_{\mathcal{B}})} \int_{\substack{\mathbf{x} \in \mathcal{B} \\ \mathbf{x} \cdot \mathbf{u}_i = \alpha}} 1 d\mathbf{x} \right) d\alpha \\ &\leq 2\mu^2(\mathcal{U}) \Sigma^2 \frac{V_{n-1}(r_{\mathcal{B}})}{V_n(r_{\mathcal{B}})} \qquad (15) \end{aligned}$$

where the last step uses the maximal cross-section of the n-sphere to dominate the area of any cross-section $\mathcal{B} \cap \{\mathbf{x} : \mathbf{x} \cdot \mathbf{u}_i = \alpha\}$. The third term of (13) is bounded using (15), Theorem 1, and the Schwarz inequality:

$$\begin{aligned} &\frac{2}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} [f(\mathbf{x}) - \bar{f}(\mathbf{x})][\bar{f}(\mathbf{x}) - \tilde{f}(\mathbf{x})] d\mathbf{x} \\ &\leq \frac{2E_1}{V_n(r_{\mathcal{B}})} \int_{\mathcal{B}} (\bar{f}(\mathbf{x}) - \tilde{f}(\mathbf{x})) sgn(\bar{f}(\mathbf{x}) - \tilde{f}(\mathbf{x})) d\mathbf{x} \\ &\leq \frac{2E_1}{V_n(r_{\mathcal{B}})} \| \bar{f}(\mathbf{x}) - \tilde{f}(\mathbf{x}) \|_2 \cdot \| sgn(\bar{f}(\mathbf{x}) - \tilde{f}(\mathbf{x})) \|_2 \\ &\leq \frac{2E_1}{V_n(r_{\mathcal{B}})} \sqrt{2\mu^2(\mathcal{U}) \Sigma^2 V_{n-1}(r_{\mathcal{B}})} \sqrt{V_n(r_{\mathcal{B}})} \qquad (16) \end{aligned}$$

Adding (14), (15) and (16), and a bit of algebra give the result.

■

The class of functions $\mathcal{D}$ used here can be extended readily to the set of all continuous functions in $\mathcal{L}_2(\mathcal{B})$. To do so, we note that the Stone-Weierstrass Theorem implies that there exists $f \in \mathcal{D}$ which agrees with any continuous function to within any $\varepsilon > 0$ at every point, so any continuous function can be approximated simply by approximating the corresponding $\mathcal{C}^\infty$ function.

Still further, for any given $g \in \mathcal{L}_2$, there exists a continuous function f such that $|f - g| < \varepsilon$. This implies that it is possible to approximate any function in $\mathcal{L}_2$ with a $\mathcal{C}^\infty$ function with arbitrarily small error.

Since the function $h(\alpha, \mathbf{u})$ is of bounded variation in α, we can approximate each such function with a sample-and-hold version of itself, that is, with a piecewise constant function taking on the value of $h(\alpha, \cdot)$ at the midpoint of each "piece". If these midpoints are uniformly separated by $\Delta\alpha_i$, the resulting function differs from h with mean square error no greater than

$$(1/12)(\Delta\alpha_i)^2 \left(\max_{\alpha} \left| \frac{\partial}{\partial \alpha} h(\alpha, \mathbf{u}_i) \right| \right)^2 . \qquad (17)$$

Together with the fact that any sigmoid can be scaled in domain so that it approximates a step function to within any desired mean square error, this implies that any continuous function can be represented by a finite linear combination of scaled sigmoids with arbitrarily small error. We can therefore construct $\eta(\alpha, \mathbf{u})$ of Theorem 2 using a finite sum of scaled sigmoids and set Σ as we like. We can now state our main result,

THEOREM 3

Any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f \in \mathcal{L}_2$ with compact support S can be approximated arbitrarily well in the norm of $\mathcal{L}_2$ by a function of form

$$\tilde{f} = \sum_{i=1}^{K} \sum_{j=1}^{m_i} a_{ij} \sigma_i(b_i(\mathbf{u}_i \cdot \mathbf{x})$$

where $a_{ij} \in \mathbb{R}$ is chosen by sampling h, σ_i is the sigmoid implemented at node i, $b_i \in \mathbb{R}$ is chosen sufficiently large to fit σ sufficiently close to a step, and $\mathbf{u}_i$ is defined in (5).

PROOF

A simple $\varepsilon/2$ type argument will suffice, given Theorem 2 and the approximation arguments above. Specifically, choose any $\varepsilon > 0$. There is always a $g \in C^\infty$ such that $\|f - g\|_2 < \varepsilon/3$ as stated above. Then we require that MSE of Theorem 2 satisfy $\sqrt{V_n(r_B)} \times \overline{MSE} < \frac{2\varepsilon}{3}$. This can be done by choosing k large enough so that E_1 of Theorem 1 is less than $\frac{\varepsilon}{3\sqrt{V_n(r_B)}}$ and choosing b_i large enough and $\Delta\alpha$ of (17) small enough so that Σ of Theorem 2 is less than $\frac{\varepsilon}{3\sqrt{V_{n-1}(r_B) \times \Omega_n}}$. Then by the triangle inequality applied to the 2-norm, we know that $\tilde{g}$, the approximate discrete-in-angle inverse radon transform of the Radon transform of g, agrees closely with f in the sense $\|f - \tilde{g}\| < \varepsilon$.

■