N91-21780

NEURAL NETWORK REPRESENTATION AND LEARNING OF MAPPINGS AND THEIR DERIVATIVES

April 1990

Halbert White, Ph.D.
University of California, San Diego

Collaborative research with:

K. Hornik, M. Stinchcombe and A.R. Gallant

ABSTRACT

We discuss recent theorems proving that artificial neural networks are capable of approximating an arbitrary mapping and its derivatives as accurately as desired. This fact forms the basis for further results establishing the learnability of the desired approximations, using results from non-parametric statistics. These results have potential applications in robotics, chaotic dynamics, control, and sensitivity analysis (physics, chemistry, and engineering). We discuss an example involving learning the transfer function and its derivatives for a chaotic map.

Jordan (1989), "Generic Constraints on Underspecified Target Trajectories," *Proceedings IJCNN*, Washington D.C.:

The Jacobian matrix $\partial z/\partial x$ . . . is the matrix that relates small changes in the controller output to small changes in the task space results and cannot be assumed to be available a priori, or provided by the environment. However, all of the derivatives in the matrix are *forward* derivates. They are easily obtained by differentiation if a forward model is available. The forward model itself must be learned, but this can be achieved directly by system identification. Once the model is accurate over a particular domain, its derivatives provide a learning operator that allows the system to convert errors in task space into errors in articulartory space and thereby change the controller.

UNIVERSAL APPROXIMATION OF AN UNKNOWN MAPPING AND ITS DERIVATIVES USING MULTILAYER FEEDFORWARD NETWORKS *

by

Kurt Hornik, Maxwell Stinchcombe

and

Halbert White

January 1990

* We are indebted to Angelo Melino for pressing us on the issue addressed here and to the referees for numerous helpful suggestions. White's participation was supported by NSF Grant SES-8806990.

ABSTRACT

We give conditions ensuring that multilayer feedforward networks with as few as a single hidden layer and an appropriately smooth hidden layer activation function are capable of arbitrarily accurate approximation to an arbitrary function and its derivatives. In fact, these networks can approximate functions that are not differentiable in the classical sense, but possess only a generalized derivative, as is the case for certain piecewise differentiable functions. The conditions imposed on the hidden layer activation function are relatively mild; the conditions imposed on the domain of the function to be approximated have practical implications. Our approximation results provide a previously missing theoretical justification for the use of multilayer feedforward networks in applications requiring simultaneous approximation of a function and its derivatives.

Relevant Application Areas:

1. Robotics

2. Chaotic Dynamics

3. Control

4. Sensitivity Analysis (Physics, Chemistry, Engineering)

Intuition suggests that networks having smooth hidden layer activation functions ought to have output function derivatives that will approximate the derivatives of an unknown mapping. However, the justification for this intuition is not obvious. Consider the class of single hidden layer feedforward networks having network output functions belonging to the set

$$\Sigma(G) \equiv \{g : \mathbb{R}^r \to \mathbb{R} \mid g(x) = \sum_{j=1}^{q} \beta_j G(\tilde{x}^T \gamma_j);$$

$$x \in \mathbb{R}^r, \beta_j \in \mathbb{R}, \gamma_j \in \mathbb{R}^{r+1}, j = 1,..., q, q \in \mathbb{N}\},$$

where x represents an r vector of network inputs ($r \in \mathbb{N} \equiv \{1, 2, ...\}$), $\tilde{x} \equiv (1, x^T)^T$ (the superscript T denotes transposition), β_j represents hidden to output layer weights and γ_j represents input to hidden layer weights, $j = 1,..., q$, where q is the number of hidden units, and G is a given hidden unit activation function. The first partial derivatives of the network output function are given by

$$\partial g(x) / \partial x_i = \sum_{j=1}^{q} \beta_j \gamma_{ji} DG(\tilde{x}^T \gamma_j), \quad i = 1,..., r,$$

where x_i is the ith component of x, γ_{ji} is the ith component of γ_j, $i = 1,..., r$ (γ_{j0} is the input layer bias to hidden unit j), and DG denotes the first derivative of G.

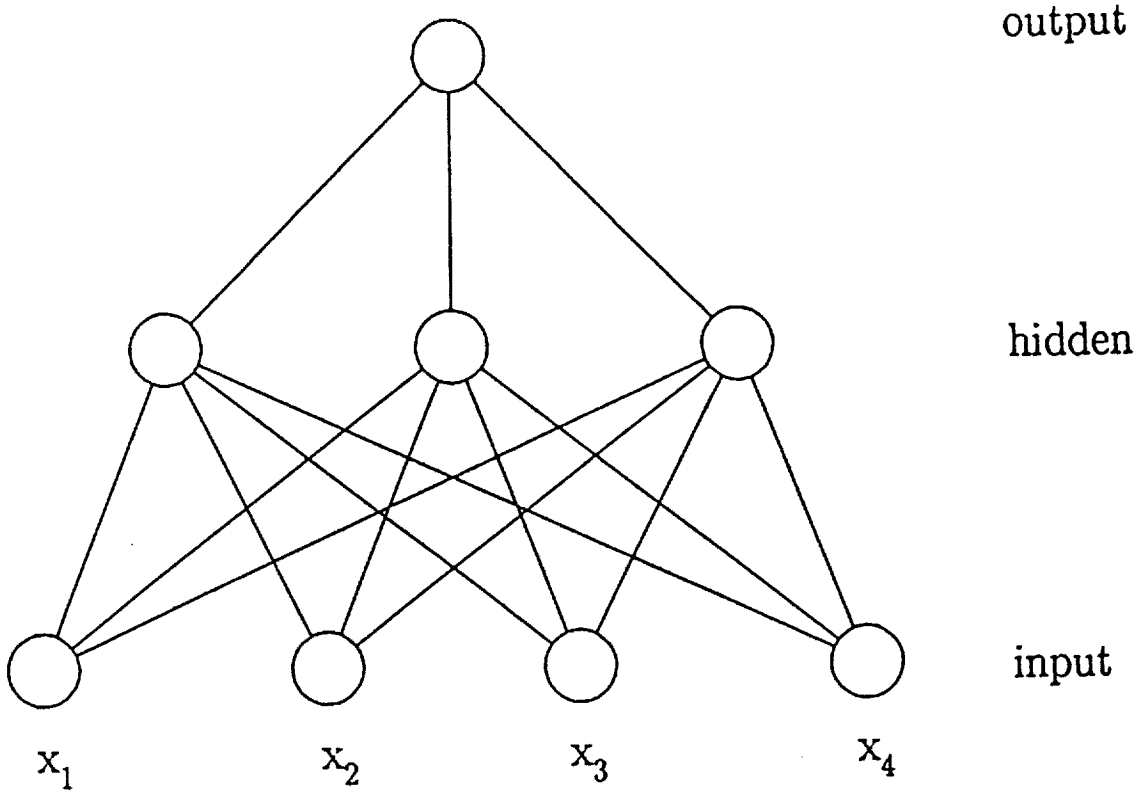


Figure 2

Single Hidden Layer Feedforward Network

Outline:

1. Mathematical Background

2. Approximation Results

3. Learning Results

4. Example: Learning Chaotic Map

1. MATHEMATICAL BACKGROUND

Let U be an open subset of $\mathbb{R}^r$, and let $C(U)$ be the set of all functions continuous on U. Let α be an r-tuple $\alpha = (\alpha_1, \ldots, \alpha_r)^T$ of non-negative integers (a "multi-index"). If x belongs to $\mathbb{R}^r$, let $x^\alpha \equiv x_1^{\alpha_1} \cdot \ldots \cdot x_r^{\alpha_r}$. Denote by D^α the partial derivative

$$\partial^{|\alpha|}/\partial x^\alpha \equiv \partial^{|\alpha|}/(\partial x_1^{\alpha_1}\ \partial x_2^{\alpha_2} \ldots \partial x_r^{\alpha_r})$$

of order $|\alpha| \equiv \alpha_1 + \alpha_2 + \ldots + \alpha_r$. For non-negative integers m, we define $C^m(U) \equiv \{f \in C(U)$: $D^\alpha f \in C(U)$ for all α, $|\alpha| \leq m\}$ and $C^\infty(U) = \cap_{m \geq 1} C^m(U)$. We let D^0 be the identity, so that $C^0(U) = C(U)$. Thus, the functions in $C^m(U)$ have continuous derivatives up to order m on U, while the functions in $C^\infty(U)$ have continuous derivatives on U of every order. We shall be interested in approximating elements of $C^m(U)$ using feedforward networks. When $U \neq \mathbb{R}^r$, the fact that network output functions (elements of $\Sigma(G)$) will belong to $C^m(\mathbb{R}^r)$ necessitates considering their restriction to U, written $g|_U$ for g in $\Sigma(G)$. Recall that $g|_U(x) = g(x)$ for x in U and is not defined for x not in U, thus $g|_U \in C^m(U)$, as desired.)

DEFINITION 2.1: Let U be a subset of $\mathbb{R}^r$, let S be a collection of functions f: $U \to \mathbb{R}$ and let ρ be a metric on S. For any g in $\Sigma(G)$ (recall $g: \mathbb{R}^r \to \mathbb{R}$) define the restriction of g to U, $g_{|U}$ as $g_{|U}(x) = g(x)$ for x in U, $g_{|U}(x)$ unspecified for x not in U.

Suppose that for any f in S and $\varepsilon > 0$ there exists g in $\Sigma(G)$ such that $\rho(f, g_{|U}) < \varepsilon$. Then we say that $\Sigma(G)$ *contains a subset* ρ*-dense in* S. If in addition $g_{|U}$ belongs to S for every g in $\Sigma(G)$, we say that $\Sigma(G)$ is ρ*-dense in* S. □

DEFINITION 2.2: Let $m, l \in \{0\} \cup \mathbb{N}$, $0 \le m \le l$, and $U \subset \mathbb{R}^r$ be given, and let $S \subset C^l(U)$. Suppose that for any f in S, compact $K \subset U$ and $\varepsilon > 0$ there exists g in $\Sigma(G)$ such that $\max_{|\alpha| \le m} \sup_{x \in K} | D^\alpha f(x) - D^\alpha g(x) | < \varepsilon$. Then we say that $\Sigma(G)$ is *m-uniformly dense on compacta* in S. □

When $\Sigma(G)$ is m-uniformly dense on compacta in S, then no matter how we choose an f in S, a compact subset K of U, or the accuracy of approximation $\varepsilon > 0$, we can always find a single hidden layer feedforward network having output function g (in $\Sigma(G)$) with all derivatives of $g_{|U}$ on K up to order m lying within ε of those of f on K. This is a strong and very desirable approximation property.

The space $L_p(U,\mu)$ is the collection of all measurable functions f such that $\|f\|_{p,U,\mu} \equiv [\int_U |f|^p d\mu]^{1/p} < \infty, 1 \le p < \infty$, where the integral is defined in the sense of Lebesgue. When $\mu = \lambda$ we may write either $\int_U f\,d\lambda$ or $\int_U f(x)dx$ to denote the same integral. We measure the distance between two functions f and g belonging to $L_p(U,\mu)$ in terms of the metric $\rho_{p,U,\mu}(f,g) \equiv \|f-g\|_{p,U,\mu}$. Two functions that differ only on sets of μ-measure zero have $\rho_{p,U,\mu}(f,g) = 0$. We shall not distinguish between such functions.

The first Sobolev space we consider is denoted $S_p^m(U,\mu)$, defined as the collection of all functions f in $C^m(U)$ such that $\|D^\alpha f\|_{p,U,\mu} < \infty$ for all $|\alpha| \le m$. We define the Sobolev norm $\|f\|_{m,p,U,\mu} \equiv (\sum_{|\alpha| \le m} \|D^\alpha f\|_{p,U,\mu}^p)^{1/p}$. The Sobolev metric is

$$\rho_{p,\mu}^m(f,g) \equiv \|f-g\|_{m,p,U,\mu} \quad f,g \in S_p^m(U,\mu).$$

Note that $\rho_{p,\mu}^m$ depends implicitly on U, but we suppress this dependence for notational convenience. The Sobolev metric explicitly takes into account distances between derivatives. Two functions in $S_p^m(U,\mu)$ are close in the Sobolev metric $\rho_{p,\mu}^m$ when all derivatives of order $0 \le |\alpha| \le m$ are close in L_p metric.

We also consider the Sobolev spaces

$$W_p^m(U) \equiv \{ f \in L_{1,\mathrm{loc}}(U) \mid \partial^\alpha f \in L_p(U,\lambda), 0 \le \mid \alpha \mid \le m \}.$$

This is the collection of all functions having generalized derivatives belonging to $L_p(U,\lambda)$ of order up to m. Consequently, $W_p^m(U)$ includes $S_p^m(U,\lambda)$, as well as functions that do not have derivatives in the classical sense, such as piecewise differentiable functions.

The norm on $W_p^m(U)$ generalizes that on $S_p^m(U,\lambda)$; we write it as

$$\| f \|_{m,p,U} \equiv \Big(\sum_{\mid \alpha \mid \le m} \| \partial^\alpha f \|_{p,U,\lambda}^p \Big)^{1/p} \quad f \in W_p^m(U).$$

For the metric on $W_p^m(U)$ we suppress the dependence on U and write

$$\rho_p^m(f,g) \equiv \| f - g \|_{m,p,U} \quad f, g \in W_p^m(U).$$

Two functions are close in the Sobolev space $W_p^m(U)$ if all generalized derivatives are close in $L_p(U,\lambda)$ distance.

Our results make fundamental use of one last function space, the space $C^{\infty}_{\downarrow}(\ I\!R^r)$ of rapidly decreasing functions in $C^{\infty}(\ I\!R^r)$. $C^{\infty}_{\downarrow}(\ I\!R^r)$ is defined as the set of all functions in $C^{\infty}(\ I\!R^r)$ such that for all multi-indices α and β, $x^{\beta}D^{\alpha}f(x) \to 0$ as $|\ x\ | \to \infty$, where $x^{\beta} \equiv x_1^{\beta_1}x_2^{\beta_2}...x_r^{\beta_r}$ and $|\ x\ | \equiv \max_{1 \le i \le r} |x_i\ |$. Note that $C_0^{\infty}(\ I\!R^r) \subset C^{\infty}_{\downarrow}(\ I\!R^r)$.

Desired results:

1.) $\Sigma(G)$ is m-uniformly dense on compacta in $C^{\infty}_{\downarrow}(\ I\!R^r)$, $S_p^m(U, \lambda)$

2.) $\Sigma(G)$ is $\rho^m_{p,\mu}$-dense in $S_p^m(\ I\!R^r\ ,\ \mu)$

3.) $\Sigma(G)$ is ρ_p^m-dense in $W_p^m(U)$

2. APPROXIMATION RESULTS

THEOREM 3.1: Let $G \neq 0$ belong to $S_1^m(\ \mathbb{R},\lambda)$ for some integer $m \geq 0$. Then $\Sigma(G)$ is m-uniformly dense on compacta in $C_{\downarrow}^{\infty}(\ \mathbb{R}^r)$. □

DEFINITION 3.2: Let $l \in \{0\} \cup \mathbb{N}$ be given. G is *l-finite* if $G \in C^l(\ \mathbb{R})$ and $0 < \int | D^l G | \, d\lambda < \infty$. □

LEMMA 3.3: If G is l-finite then for all $0 \leq m \leq l$ there exists $H \in S_1^m(\ \mathbb{R},\lambda), H \neq 0$, such that $\Sigma(H) \subset \Sigma(G)$. □

l-finite activation functions G with $\int D^l G \, d\lambda \neq 0$ have $\int |D^m G| \, d\lambda = \infty$ for all $m < l$, and for $m > l$ all l-finite activation functions G have $\int D^m G \, d\lambda = 0$ (provided $D^m G$ exists).

It is informative to examine cases not satisfying the conditions of the theorems. For example, if $G = \sin$ then $G \in C^{\infty}(\ \mathbb{R})$, but for all l, $\int | D^l G | \, d\lambda = \infty$. If G is a polynomial of degree m then again $G \in C^{\infty}(\ \mathbb{R})$, but for $l \leq m$ we have $\int | D^l G | \, d\lambda = \infty$, although $\int | D^l G | \, d\lambda = 0$ for $l > m$. Consequently, neither trigonometric functions nor polynomials are l-finite.

COROLLARY 3.4: If G is l-finite, then for all $0 \leq m \leq l$, $\Sigma(G)$ is m-uniformly dense on compacta in $C^{\infty}(\mathbb{R}^r)$. □

COROLLARY 3.5: If G is l-finite, $0 \leq m \leq l$, and U is an open subset of $\mathbb{R}^r$ then $\Sigma(G)$ is m-uniformly dense on compacta in $S_p^m(U,\lambda)$ for $1 \leq p < \infty$. □

COROLLARY 3.6: If G is l-finite and μ is compactly supported, then for all $0 \leq m \leq l$ $\Sigma(G) \subset S_p^m(\mathbb{R}^r, \mu)$ and $\Sigma(G)$ is $\rho_{p,\mu}^m$-dense in $S_p^m(\mathbb{R}^r, \mu)$.

COROLLARY 3.8: If G is l-finite, $0 \leq m \leq l$, U is an open bounded subset of $\mathbb{R}^r$ and $C_0^{\infty}(\mathbb{R}^r)$ is ρ_p^m-dense in $W_p^m(U)$ then $\Sigma(G)$ is also ρ_p^m-dense in $W_p^m(U)$.

These results rigorously establish that sufficiently complex multilayer feedforward networks with as few as a single hidden layer are capable of arbitrarily accurate approximation to an unknown mapping and its (generalized) derivatives in a variety of precise senses. The conditions imposed on G are relatively mild; the conditions required of U have practical implications.

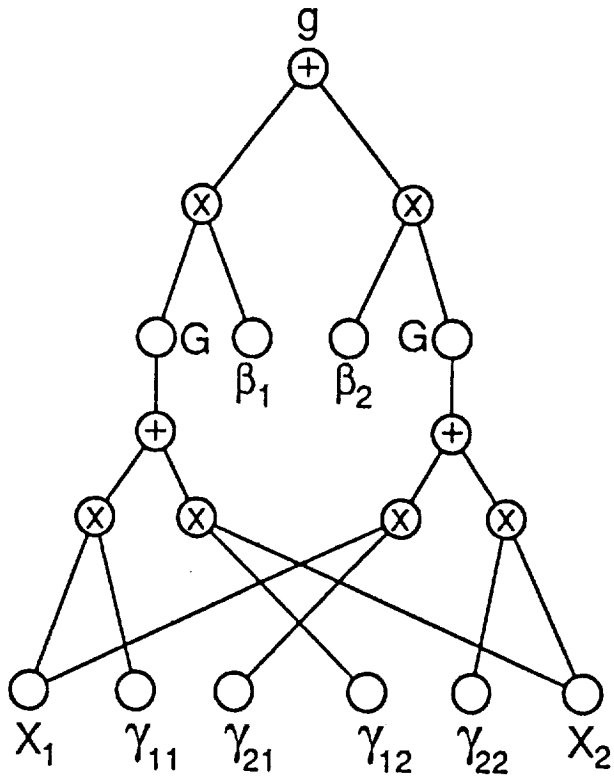


Figure 1. Feedforward Network

○ *input unit* ⊗ *multiplication unit*

G○ *activation unit* ⊕ *addition unit*

Note: biases not shown

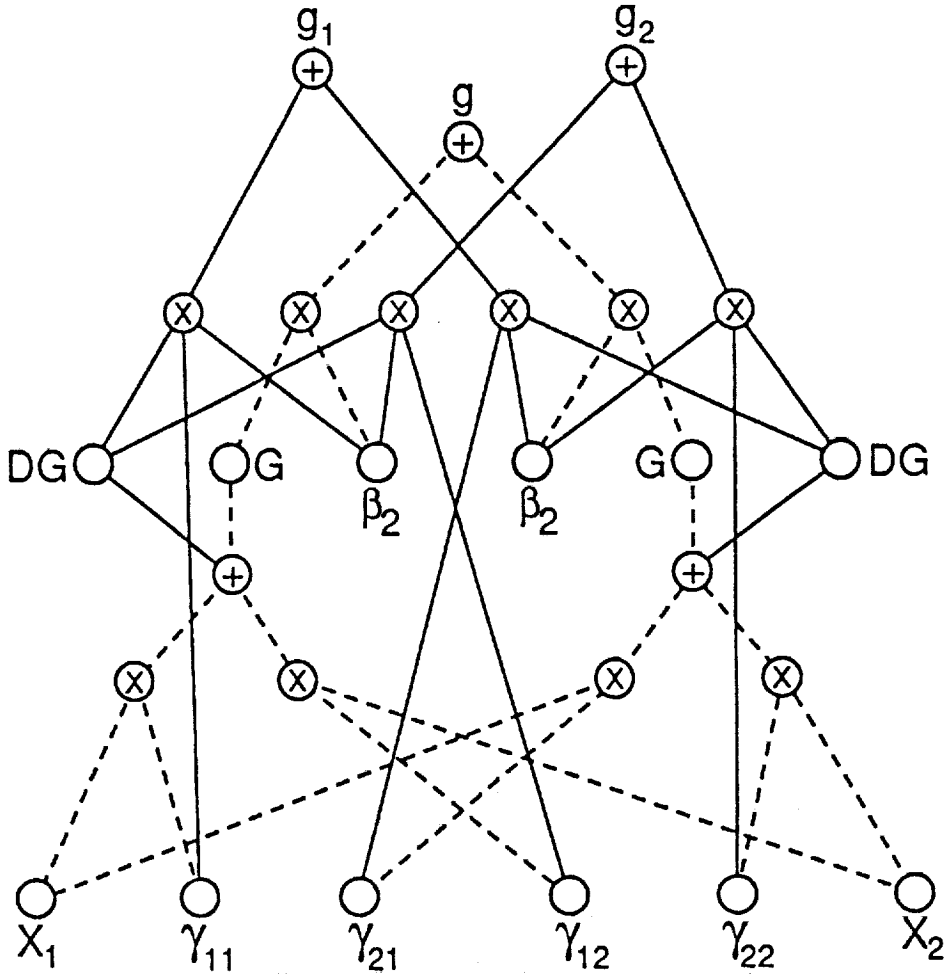


Figure 2. Derivative Network

○ *input unit* ⊗ *multiplication unit*

G○ *activation unit* ⊕ *addition unit*

DG○ *activation derivative unit*

Note: biases not shown

On Learning the Derivatives of an Unknown Mapping with Multilayer Feedforward Networks

by

A. Ronald Gallant
Department of Statistics
North Carolina State University
Raleigh, NC 27696-8203 USA

Halbert White
Department of Economics, D-008
University of California, San Diego
La Jolla, CA 92093

October 1989

ABSTRACT

Recently, multiple input, single output, single hidden layer, feedforward neural networks have been shown to be capable of approximating a nonlinear map and its partial derivatives. Specifically, neural nets have been shown to be dense in various Sobolev spaces (Hornik, Stinchcombe and White, 1989). Building upon this result, we show that a net can be trained so that the map and its derivatives are learned. Specifically, we use a result of Gallant (1987b) to show that least squares and similar estimates are strongly consistent in Sobolev norm provided the number of hidden units and the size of the training set increase together. We illustrate these results by an application to the inverse problem of chaotic dynamics: recovery of a nonlinear map from a time series of iterates. These results extend automatically to nets that embed the single hidden layer, feedforward network as a special case.

3. LEARNING RESULTS

SETUP. We consider a single hidden layer feedforward network having network output function

$$g_K(x, \theta) = \sum_{j=1}^{K} \beta_j G(x^T \gamma_j)$$

where x represents an $r \times 1$ vector of network inputs (including a "bias unit"), β_j represents hidden to output layer weights, γ_j represents input to hidden layer weights, K is the number of hidden units,

$$\theta' = (\beta_1, \gamma_1', \beta_2, \gamma_2', \ldots, \beta_K, \gamma_K'),$$

and G is the hidden unit activation function.

We assume that the network is trained using data $\{y_t, x_t\}$ generated according to

$$y_t = g^*(x_t) + e_t \qquad t = 1, 2, \ldots, n .$$

x_t denotes the observed input and e_t denotes random noise. The number K_n of hidden units employed depends on the size n of the training set. The network is trained by finding $g_{K_n}(x, \hat{\theta})$ that minimizes

$$s_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} [y_t - \sum_{j=1}^{K_n} \beta_j G(x_t^T \gamma_j)]^2 ,$$

subject to the restriction that $g_{K_n}(x, \hat{\theta})$ is a member of the estimation space $\mathcal{G}$.

REGULARITY CONDITIONS:

Input space. The input space X is the closure of a bounded, open subset of $\mathbb{R}^r$.

Parameter space. For some integer m, $0 \leq m < \infty$, some integer p, $1 \leq p < \infty$, and some bound B, $0 < B < \infty$, g^* is a point in the Sobolev space $\mathcal{W}_{m+[r/p]+1, p, X}$ and $\|g^*\|_{m+[r/p]+1, p, X} < B$.

Activation function. The activation function G belongs to $C^m(\mathbb{R})$ and $\int_{-\infty}^{\infty} (d^m/du^m) G(u)\, du < \infty$. See Section 3 of Hornik, Stinchcombe and White (1989).

Estimation space. $g_{K_n}(x, \hat{\theta})$ is restricted to $\mathcal{G} = \{g: \|g\|_{m+[r/p]+1, p, X} \leq B\}$ in the optimization of $s_n(g)$.

Training set. The empirical distribution of $\{x_t\}_{t=1}^n$ converges to a distribution $\mu(x)$ and $\mu(O) > 0$ for every open subset O of X.

Error process. The errors $\{e_t\}$ are independently and identically distributed with common probability law P having $\int_{\mathcal{E}} eP(de) = 0$ and $0 \leq \int_{\mathcal{E}} e^2 P(de) < \infty$. ($\int_{\mathcal{E}} e^2 P(de) = 0$ implies $e_t = 0$ for all t.)

Independence. The probability law P of the errors does not depend on $\{x_t\}_{t=1}^{\infty}$; that is, $P(A)$ can be evaluated without knowledge of $\{x_t\}_{t=1}^{n}$, $\lim_{n\to\infty}(1/n)\sum_{t=1}^{n} x_t$, etc.

THEOREM 1. Under the Regularity Conditions

$$\lim_{n\to\infty} \| g^* - g_{K_n}(\cdot, \hat{\theta}) \|_{m,\infty,x} = 0 \qquad \text{almost surely}$$

provided $\lim_{n\to\infty} K_n = \infty$ almost surely. In particular,

$$\lim_{n\to\infty} \sigma[g_{K_n}(x, \hat{\theta})] = \sigma(g^*) \qquad \text{almost surely}$$

provided σ is continuous with respect to $\| \cdot \|_{m,\infty,x}$. □

4. EXAMPLE: LEARNING CHAOTIC MAP

Our investigation studies the ability of the single hidden layer network

$$g_K(x_{t-5}, \ldots, x_{t-1}) = \sum_{j=1}^{K} \beta_j G(\gamma_{5j} x_{t-5} + \cdots + \gamma_{1j} x_{t-1} + \gamma_{0j})$$

with logistic squasher

$$G(u) = 1/[1 + \exp(-u)]$$

to approximate the derivatives of a discretized variant of the Mackey-Glass equation (Schuster, 1988, p. 120)

$$g(x_{t-5}, x_{t-1}) = x_{t-1} + (10.5)\left[\frac{(0.2)x_{t-5}}{1 + (x_{t-5})^{10}} - (0.1)x_{t-1}\right].$$

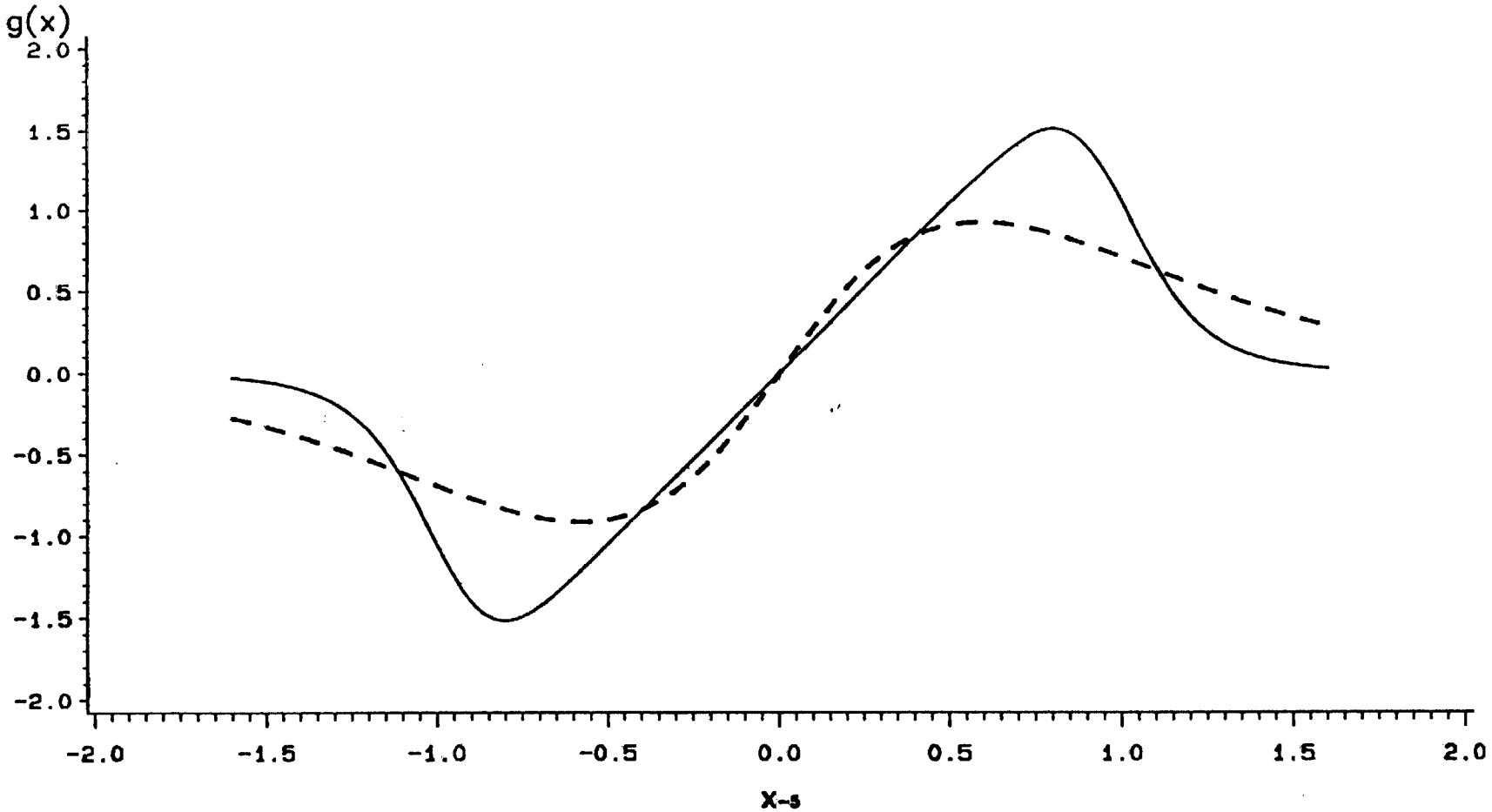
The values of the weights $\hat{\beta}_j$ and $\hat{\gamma}_{ij}$ that minimize

$$s_n(g_K) = \frac{1}{n} \sum_{t=1}^{n} [x_t - g_K(x_{t-5}, \ldots, x_{t-1})]^2$$

were determined using the Gauss-Newton nonlinear least squares algorithm. Our rule relating K to n was of the form $K \propto \log(n)$ because asymptotic theory in a related context (Gallant, 1989) suggests that this is likely to be the relationship that will give stable estimates.

Figure 1. Superimposed nonlinear map and neural net estimate

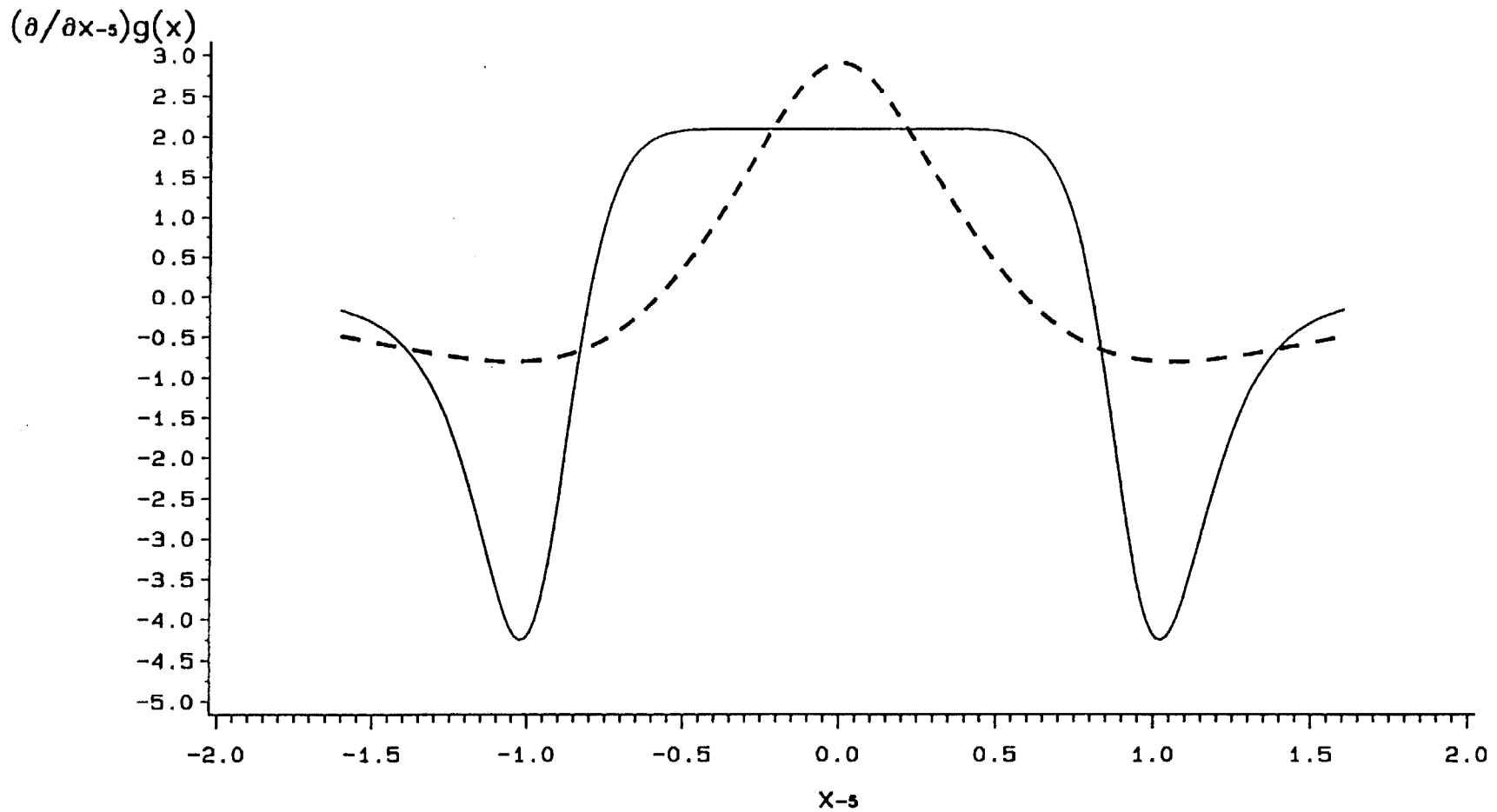
K = 3, n = 500



Note: Estimate is dashed line, x = (x_{-s}, 0, 0, 0, 0)

Figure 2. Superimposed derivative and neural net estimate

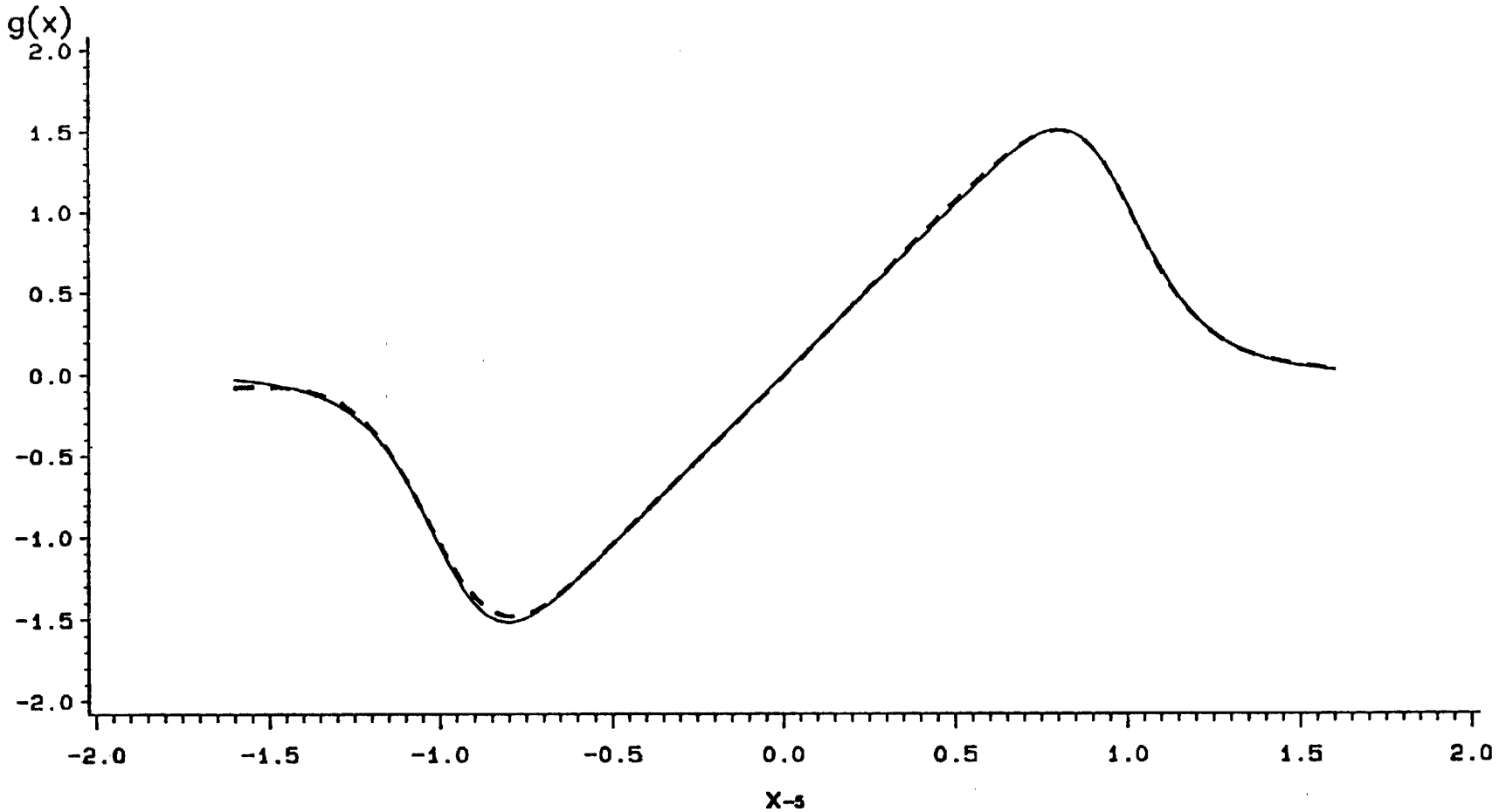
K = 3, n = 500



Note: Estimate is dashed line, x = (x-s, 0, 0, 0, 0)

Figure 3. Superimposed nonlinear map and neural net estimate

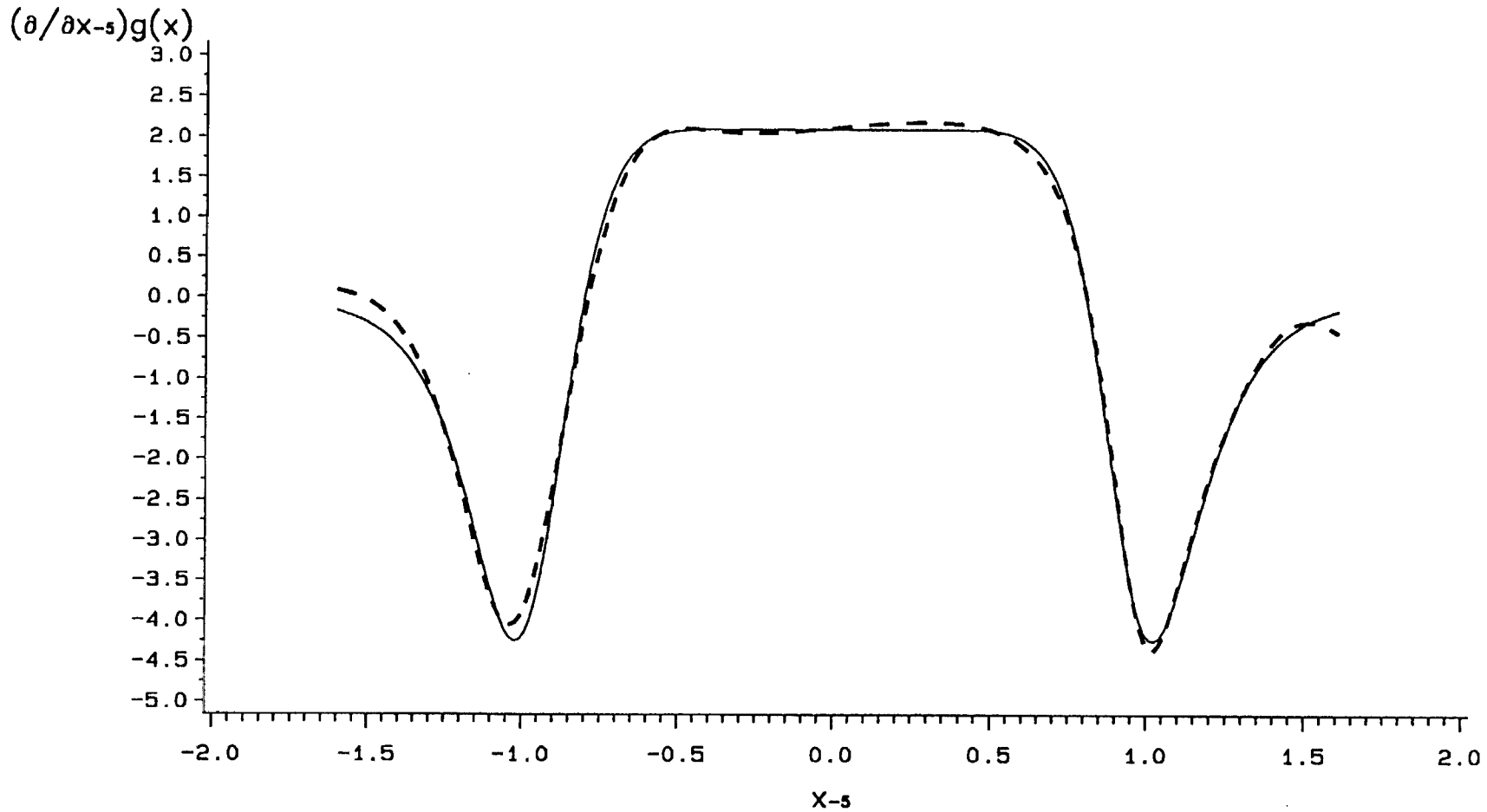
K = 7, n = 2000



Note: Estimate is dashed line, x = (x_{-5}, 0, 0, 0, 0)

Figure 4. Superimposed derivative and neural net estimate

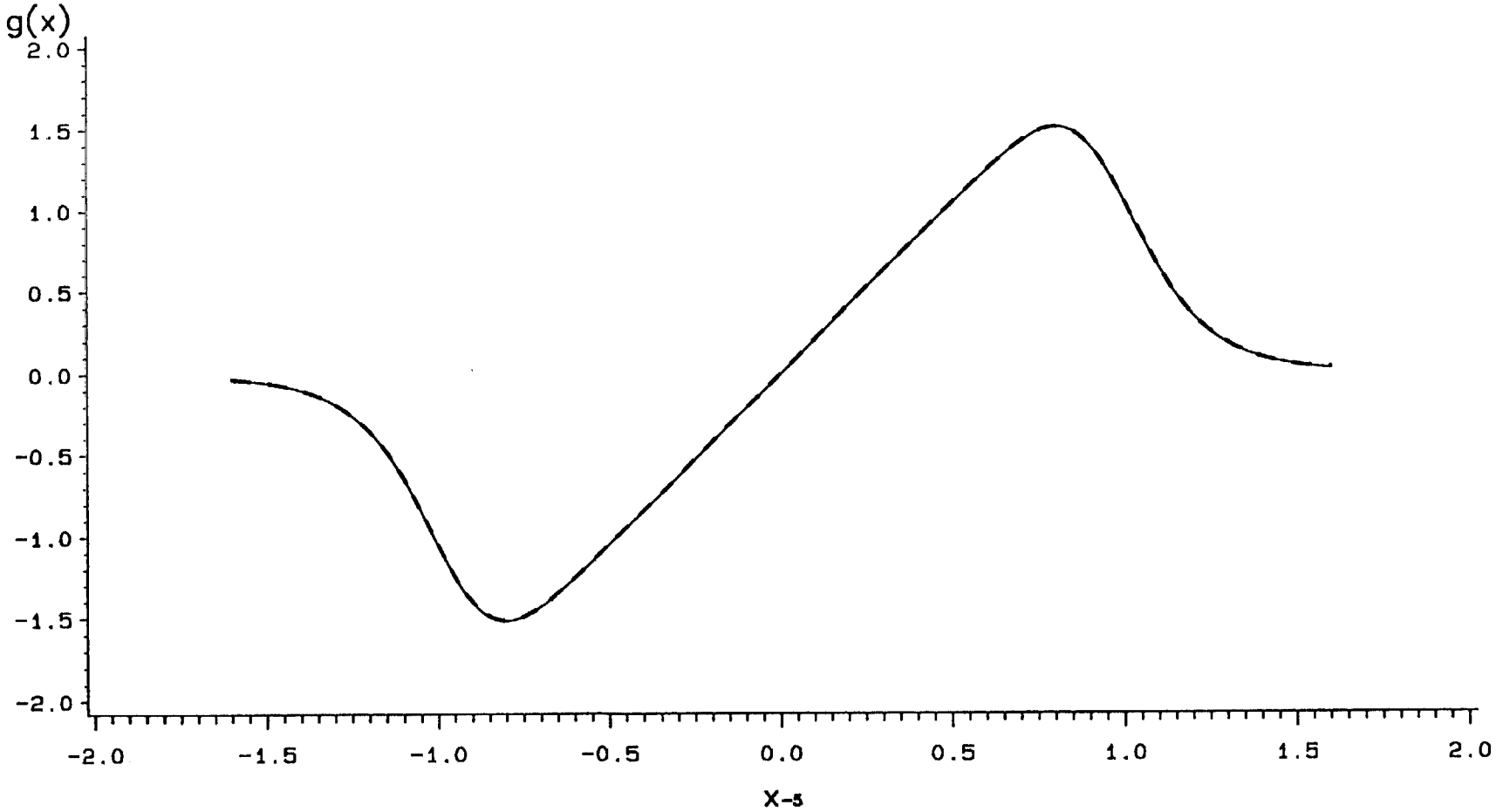
K = 7, n = 2000



Note: Estimate is dashed line, x = (x_{-5}, 0, 0, 0, 0)

Figure 5. Superimposed nonlinear map and neural net estimate

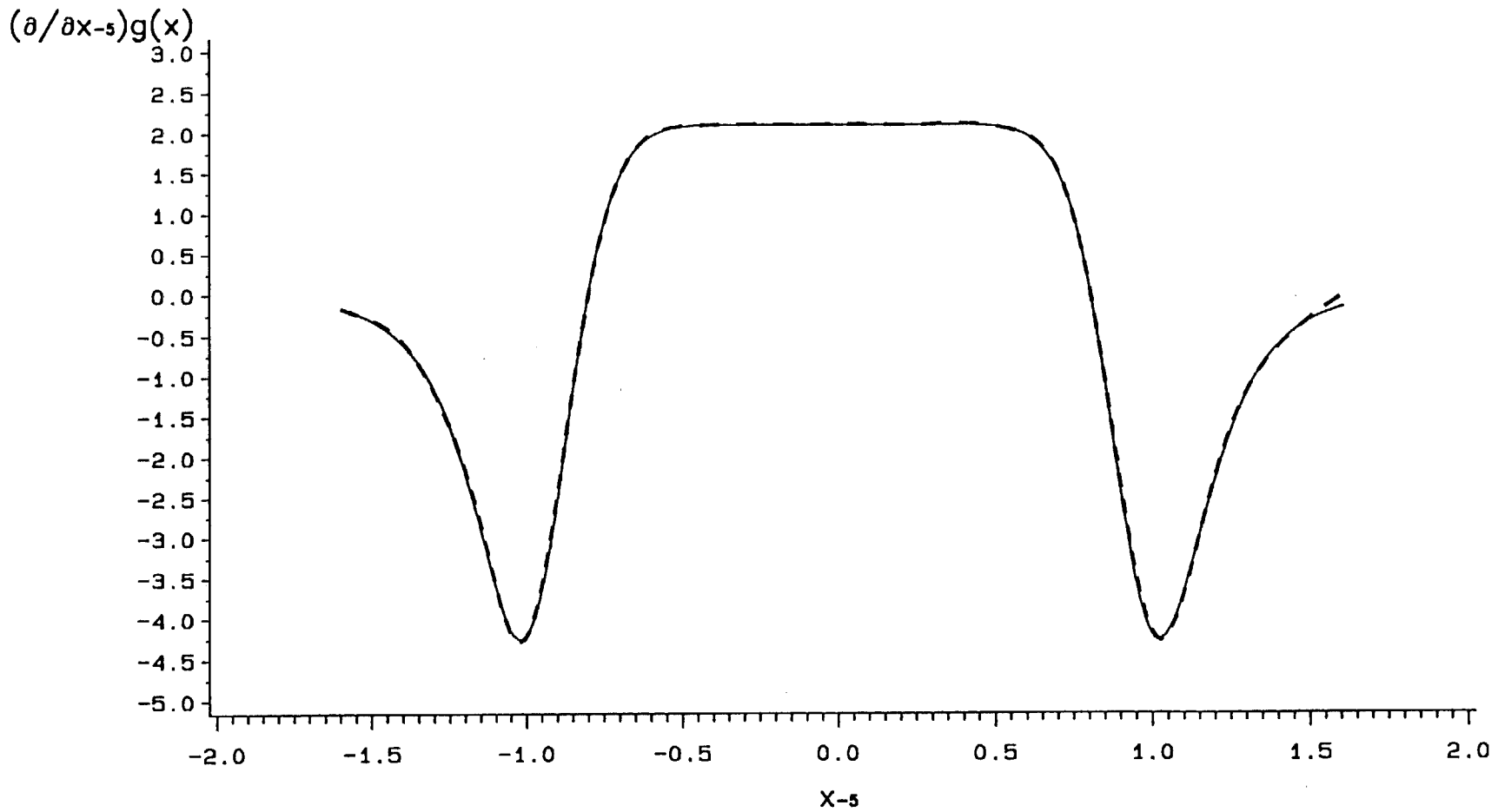
K = 11, n = 8000



Note: Estimate is dashed line, x = (x-5, 0, 0, 0, 0)

Figure 6. Superimposed derivative and neural net estimate

K = 11, n = 8000



Note: Estimate is dashed line, x = (x-5, 0, 0, 0, 0)

Impact of Application of Fuzzy Theory to Industry

(Paper not provided by publication date.)

Time-sweeping Mode Fuzzy Computer -- Forward and Backward Fuzzy Inference Engine

(Paper not provided by publication date.)