No. 1964

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

Professor Halbert White
Department of Economics, D008
University of California, San Diego
La Jolla, CA 92093

October 1989

*This research was supported by National Science Foundation Grants SES-8510637, SES-8808015, North Carolina Agricultural Experiment Station Projects NCO-5593, NCO-3879, and the PAMS Foundation. We thank Stephen P. Ellner, Daniel F. Mccaffrey, and Douglas W. Nychka for helpful discussions relating to chaotic dynamics.

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.

1. INTRODUCTION

Recently, Gallant and White (1988) have demonstrated that multiple input, single output, single hidden layer, feedforward networks (e.g., Rumelhart, Hinton, and Williams, 1986) with a particular choice of a monotone squashing function at the hidden layer and no squashing at the output layer can approximate any square integrable function to any desired accuracy by increasing the number of hidden units. Hornik, Stinchcombe and White (1989) relaxed the conditions on the squashing function and expanded the class of functions that can be approximated. These results have the subsidiary implication that any network that embeds these networks as a special case, e.g., additional hidden layers, will inherit their approximating abilities so the results are far more general than they might at first appear. White (1989b) has shown that the approximation potential suggested by these results has practical value in that the appropriate values of the connection strengths and the appropriate number of hidden units can be learned. This result is obtained by verifying that the network together with the learning and expansion rules can be regarded as a weakly consistent estimator in the statistical sense for an element of a function space. The function space is chosen so as to contain the mappings that are to be learned.

In some applications, notably robotics (Jordan, 1989), demand analysis (Elbadawi, Gallant, and Souza, 1983), and chaotic dynamics (Schuster, 1988), approximation of the mapping will not suffice. Close approximation to both the mapping and the derivatives of the mapping are required in these applications. Hornik, Stinchcombe, and White (1989) have demonstrated that multiple input, single output, single hidden layer feedforward networks can not only

approximate the mapping but also its derivatives provided the squashing function is confined to a certain (quite general) class and the inputs are drawn from a suitably restricted domain. In this paper we extend White's (1989b) analysis and provide rules such that these networks can learn both the mapping and its derivatives. This result is obtained by verifying that the network together with the learning and expansion rules can be regarded as a strongly consistent estimator for a particular class of function spaces called Sobolev spaces.

2. PRELIMINARIES

We consider a single hidden layer, feedforward network having network output function

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

where x represents an $r \times 1$ vector of network inputs (a prime $'$ denotes transposition), β_j represents hidden to output layer weights, γ_j represents input to hidden layer weights, $j = 1, 2, \ldots, K$, K is the number of hidden units,

$$\theta' = (\beta_1, \gamma'_1, \beta_2, \gamma'_2, \ldots, \beta_K, \gamma'_K),$$

and G is a given hidden unit activation function. If a bias term is to be incorporated in the specification, read $G[(1,x')\gamma_j]$ for $G(x'\gamma_j)$ above and throughout; with this change, the leading element of γ_j is interpreted as bias term. The set $\mathcal{X} \subset \mathbb{R}^r$ is presumed to contain all admissible inputs. We shall take $\mathcal{X}$ to be the closure of a bounded, open subset of $\mathbb{R}^r$. While assuming a bound on $\mathcal{X}$ may be restrictive in some applications, a key result upon which we rely is not known for unbounded domains. The other restrictions can be relaxed at some inconvenience in verifying the identification condition in Section 4. See Hornik, Stinchcombe, and White (1989) for a detailed discussion of admissible domains and Gallant and Nychka (1987) for an illustration of the difficulties involved in moving to unbounded domains.

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

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

where x_t denotes the observed input and e_t denotes random noise, that the number K_n of hidden units employed depends on the size n of the training set, that training the network is equivalent to finding a network $g_{K_n}(x|\hat{\theta})$ that minimizes some function $s_n(g)$ over all networks $g_{K_n}(x|\theta)$ with K_n hidden units, and that some functional $\sigma(g^*)$ is the feature of the mapping g^* that is supposed to be closely approximated by the network. A common choice of $s_n(g)$ is the least squares criterion

$$s_n(g) = \frac{1}{n}\sum_{t=1}^{n}[y_t - g(x_t)]^2.$$

Often, the choice of objective function is not stated explicitly but rather is implicit in the choice of training procedures. See White (1989a) for the relationship of least squares to the popular backpropagation training rule. A feature of the mapping g^* that might be of interest is an integral such as $\sigma(g) = \int_{\mathcal{X}} g(x)\,dx$. Another feature that might be of interest is a partial derivative at some point x^o such as $\sigma(g) = D^\lambda g(x^o)$.

This is a notation we use heavily. The vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r)$ has nonnegative integers as elements and

$$D^\lambda g(x) = (\partial^{\lambda_1}/\partial x_1^{\lambda_1})\cdot\ldots\cdot(\partial^{\lambda_r}/\partial x_r^{\lambda_r})g(x),$$

where $|\lambda| = \sum_{i=1}^{r}|\lambda_i|$ gives the order of the partial derivative. $D^0 g$ denotes the function itself; that is, $D^0 g(x) \equiv g(x)$.

The purpose of this paper is to obtain general conditions under which a network can be said to learn $\sigma(g)$ with certainty. More precisely, we seek to

determine conditions such that the network $g_{K_n}(x|\hat{\theta})$ learns the mapping g^* in the sense that application of the functional σ to the network $g_{K_n}(x|\hat{\theta})$ gives a an approximation $\sigma[g_{K_n}(x|\hat{\theta})]$ to $\sigma(g^*)$ that can be made as accurate as desired by increasing the size of the training set. This is equivalent to the statistical notion of strong consistency, described as follows.

Following standard conventions, we assume that the errors e_t can be regarded as being determined by functions $E_t(\omega)$, $t = 1, 2, \ldots$ defined over a probability space $(\Omega,\mathcal{A},P)$ where ω is a typical element of Ω and $\mathcal{A}$ is the collection of subsets A of Ω over which the probability measure P(A) is defined, see Tucker (1967) for instance. To each ω in Ω there corresponds a realization of the errors $\{e_t\}_{t=1}^{\infty}$ where $e_t = E_t(\omega)$. Each realization $\{e_t\}$ that can obtain in practice corresponds to some ω and these exhaust the totality of possibilities. Realizations $\{e_t\}$ with specific characteristics can be described by describing the set A of ω to which they correspond. The probability that $\{e_t\}$ with these characteristics occurs is computed as P(A).

Given a specific training procedure and rule K_n for determining the number of hidden units, we are interested in the set A of ω that generate realizations $\{e_t\}$ such that

$$\lim_{n\to\infty} \sigma(\hat{g}_{K_n}) = \sigma(g^*).$$

We shall obtain conditions such that P(A) = 1. This is a strong result because it essentially says that the feature of interest $\sigma(g^*)$ is learned with certainty. One writes

$$\lim_{n\to\infty} \sigma(\hat{g}_{K_n}) = \sigma(g^*) \qquad \text{almost surely}$$

to denote this notion of a limit and says that $\sigma(\hat{g}_{K_n})$ is strongly consistent for $\sigma(g^*)$. This notion of convergence is equivalent to the notion of convergence almost everywhere in measure theory; a probability space is a finite measure space with $P(\Omega) = 1$.

Our strategy is to relate the single hidden layer, feedforward network $g_K(x|\theta)$ described above to the following result of Gallant (1987b).

THEOREM 0. Suppose that $\hat{g}_{K_n}$ is obtained by minimizing a sample objective function $s_n(g)$ over $\mathcal{G}_{K_n}$ where $\mathcal{G}_K$ is a subset of some function space $\mathcal{G}$ on which is defined a norm $\|g\|$. Let $\sigma(g)$ denote the feature of g that is of interest and suppose that $\sigma(g)$ is continuous over $\mathcal{G}$ with respect to $\|g\|$. Consider the following conditions:

(a) **Compactness:** The closure of $\mathcal{G}$ with respect to $\|g\|$ is compact in the relative topology generated by $\|g\|$.

(b) **Denseness:** $\cup_{K=1}^{\infty}\mathcal{G}_K$ is a dense subset of the closure of $\mathcal{G}$ with respect to $\|g\|$ and $\mathcal{G}_K \subset \mathcal{G}_{K+1}$.

(c) **Uniform convergence:** There is a point g^* in $\mathcal{G}$, regarded as the true value of g, and there is a function $\bar{s}(g,g^*)$ that is continuous in g with respect to $\|g\|$ such that

$$\lim_{n\to\infty} \sup_{g\in\mathcal{G}} |s_n(g) - \bar{s}(g,g^*)| = 0 \qquad \text{almost surely.}$$

(d) **Identification:** Any point g^0 in $\bar{\mathcal{G}}$ with

$$\bar{s}(g^0,g^*) \le \bar{s}(g^*,g^*)$$

must have

$$\sigma(g^0) = s(g^*).$$

If conditions (a)-(d) hold, then

$$\lim_{n\to\infty} \sigma(\hat{g}_{K_n}) = \sigma(g^*) \qquad \text{almost surely}$$

provided that $\lim_{n\to\infty} K_n = \infty$ almost surely.∎

With respect to the identification condition, one does not know g^* in advance so one would logically be obligated to verify that the condition holds were g^* an arbitrary point in $\mathcal{G}$. However, usually one does not strive for this level of generality and only verifies the condition for all g^* in some $\mathcal{G}^*$. $\mathcal{G}^*$ is called the parameter space. It is a subset of $\mathcal{G}$ and g^* in $\mathcal{G}^*$ usually possesses some property over and above that possessed by an arbitrary g in $\mathcal{G}$. For instance, g^* may be assumed not to be on the boundary of $\mathcal{G}$ or might be assumed to have more derivatives than membership in $\mathcal{G}$ would imply. $\mathcal{G}$ is called the estimation space and $\|g\|$ the consistency norm (Gallant, 1987).

3. THE SENSE OF THE APPROXIMATION, COMPACTNESS, AND DENSENESS

As seen from Theorem 0 the quality of our results is determined by the consistency norm. The stronger is this norm, the larger the class of functionals σ that are continuous with respect to it, and the more the network can be said to have learned about the mapping g^*. We establish consistency with respect to the Sobolev norm.

The Sobolev norm is defined as

$$\|g\|_{m,p,\mathcal{X}} = \left[\sum_{|\lambda| \le m} \int_{\mathcal{X}} |D^\lambda g(x)|^p \, dx \right]^{1/p} \qquad 1 \le p < \infty$$

$$\|g\|_{m,\infty,\mathcal{X}} = \max_{|\lambda| \le m} \sup_{x \in \mathcal{X}} |D^\lambda g(x)| \qquad p = \infty .$$

We shall apply Theorem 0 with $\|\cdot\|_{m,\infty,\mathcal{X}}$ as the consistency norm where m is the largest derivative to which an approximation is desired in a given application. For instance, if the Jacobian $(\partial/\partial x')g^*(x)$ is to be approximated then $m = 1$.

This is a very strong norm. For instance, consistency with respect to the norm $\|\cdot\|_{m,\infty,\mathcal{X}}$ implies that $\sigma(g^*)$ is learned if

$$\sigma(g) = \int_{\mathcal{X}} f(x) D^\lambda g(x) \, dx \qquad |\lambda| \le m, \ f \text{ bounded on } \mathcal{X},$$

$$\sigma(g) = D^\lambda g(x) \qquad |\lambda| \le m,$$

$$\sigma(g) = \sup_{x \in \mathcal{X}} |D^\lambda g(x)| \qquad |\lambda| \le m,$$

$$\sigma(g) = \inf_{x \in \mathcal{X}} |D^\lambda g(x)| \qquad |\lambda| \le m,$$

or continuous functions of these quantities.

We assume that it is possible to specify an *a priori* bound B on the magnitude of $\|g^*\|_{m+[r/p]+1,p,\mathcal{X}}$ for some p with $1 \leq p < \infty$ where [r/p] denotes the integer part of r/p. Recall that m is the largest derivative of g^* that it is necessary to learn in a given application and that r is the dimension of the domain $\mathcal{X}$; that is, $\mathcal{X} \subset \mathbb{R}^r$. Then we take as the estimation space

$$\mathcal{G} = \{g: \|g^*\|_{m+[r/p]+1,p,\mathcal{X}} \leq B\}.$$

By the Rellich-Kondrachov Theorem (Adams, 1975, Theorem 6.2, Part II), the closure of $\mathcal{G}$ with respect to the norm $\|\cdot\|_{m,\infty,\mathcal{X}}$ is compact in the relative topology generated by $\|\cdot\|_{m,\infty,\mathcal{X}}$. Condition (a) of Theorem 0 is now satisfied.

Hornik, Stinchcombe, and White (1989) set forth mild conditions on the activation function G such that the class of single hidden layer, feedforward networks is dense for the Sobolev space $W_{m,\infty,\mathcal{X}} = \{g: \|g\|_{m,\infty,\mathcal{X}} < \infty\}$. For instance, if G is an m-times continuously differentiable function whose m-th derivative is integrable over $(-\infty,\infty)$ then G is an acceptable choice. The familiar logistic and hyperbolic tangent squashers satisfy this condition. In consequence of the Hornik, Stinchcombe, and White result, we can put

$$\mathcal{G}_K = \{g: g(x) = g_K(x|\theta),\ \theta \in \mathbb{R}^{(r+1)K}\} \cap \mathcal{G}.$$

and Condition (b) of Theorem 0 is satisfied.

We should remark that the intersection with $\mathcal{G}$ in the definition of $\mathcal{G}_K$ above has implications regarding the minimization of $s_n(g)$ over $g \in \mathcal{G}_K$. In principle, the bound $\|g_K(\cdot|\theta)\|_{m+[r/p]+1,p,\mathcal{X}} \leq B$, which is a parametric restriction on θ, must be enforced in the minimization of $s_n(g)$ over $g \in \mathcal{G}_K$, equivalently, in the minimization of $s_n[g_K(\cdot|\theta)]$ over $\theta \in \mathbb{R}^{(r+1)K}$. In

practice, restricting (r+1)K to reasonable values relative to n has the effect of smoothing $\hat{g}_K$ enough that the bound is not binding on the optimum or on any intermediate values of g_K involved in its computation.

This completes the verification of Conditions (a) and (b) of Theorem 0. Conditions (c) and (d) are verified in the next section. Before concluding, we record another consequence of the Rellich-Kondrachov Theorem that is useful in general and used in the next section: Sobolev norms are interleaved in the sense that there is a constant c that does not depend on g such that

$$\|g\|_{m,\infty,\mathcal{X}} \leq c\|g\|_{m+[r/p]+1,p,\mathcal{X}} \leq c\|g\|_{m+[r/p]+1,\infty,\mathcal{X}}$$

4. UNIFORM CONVERGENCE AND IDENTIFICATION

For specificity, we will restrict attention to the case when

$$s_n(g) = \frac{1}{n}\sum_{t=1}^{n}[y_t - g(x_t)]^2$$

as a discussion of identification wanders into vague generalities without the focus of a particular example. This is the most common choice of a sample objective function in applications and our discussion will serve as a template for the determination of the identification status of alternative choices. As remarked previously, minimization of $s_n(g)$ over $\mathcal{G}_K$ is equivalent to minimization of the parametric function

$$s_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}[y_t - \sum_{j=1}^{K}\beta_j G(x'\gamma_j)]^2.$$

While this fact is certainly convenient as regards computations, it plays no role in the theory.

Consider the case when: (i) the observational errors $\{e_t\}$ and network inputs $\{x_t\}$ are independent, (ii) the observational errors $\{e_t\}$ are independently and identically distributed with common distribution function $P(e)$ having mean $\int_{\mathcal{E}} e\,dP(e) = 0$ and variance $\int_{\mathcal{E}} e^2\,dP(e) < \infty$, and (iii) the empirical distribution μ_n of $\{x_t\}_{t=1}^{n}$ converges to a probability distribution $\mu(x)$. That is,

$$\mu_n(x) = \frac{1}{n}\left(\text{number of } x_t \leq x, \text{ coordinate by coordinate, } 1 \leq t \leq n\right)$$

and $\ell\text{im}_{n\to\infty}\mu_n(x) = \mu(x)$ at every point where $\mu(x)$ is continuous. This is a mild restriction on the sequence $\{x_t\}$. An ergodic chaotic process satisfies this

restriction as do realizations from ergodic random processes, deterministic replication schemes, and fill-in rules such as 0, 1, 1/2, 1/4, 3/4, ...

Under these assumptions we have the following Uniform Strong Law: If (i) $f(e,x,g)$ is continuous on $\mathcal{E}\times\mathcal{X}\times\mathcal{G}$ where $\mathcal{E}$ and $\mathcal{X}$ are subsets of a Euclidean space and $\mathcal{G}$ is a compact metric space, and (ii) $f(e,x,g)$ is dominated by an integrable function $f(e,x)$ then

$$\lim_{n\to\infty} \sup_{g\in\mathcal{G}} \left| \frac{1}{n} \sum_{t=1}^{n} f(e_t,x_t,g) - \int_{\mathcal{E}}\int_{\mathcal{X}} f(e,x,g)\, dP(e)d\mu(x) \right| = 0$$

almost surely (Gallant, 1987a, p. 159).

Applying this result to $s_n(g)$ above to get the function $\bar{s}(g,g^*)$ required in Condition (c) of Theorem 0, we have

$$\begin{aligned} s_n(g) &= \frac{1}{n} \sum_{t=1}^{n} [y_t - g(x_t)]^2 \\ &= \frac{1}{n} \sum_{t=1}^{n} [e_t + g^*(x_t) - g(x_t)]^2 \end{aligned}$$

$$\begin{aligned} &\int_{\mathcal{E}}\int_{\mathcal{X}} [e + g^*(x) - g(x)]^2\, dP(e)d\mu(x) \\ &\quad = \int_{\mathcal{E}} e^2 dP(e) + 2\int_{\mathcal{E}} e dP(e) \int_{\mathcal{X}} [g^*(x) - g(x)]d\mu(x) \\ &\qquad + \int_{\mathcal{X}} [g^*(x) - g(x)]^2 d\mu(x) \\ &\quad = \sigma^2 + \int_{\mathcal{X}} [g^*(x) - g(x)]^2 d\mu(x). \end{aligned}$$

The requisite dominating function is

$$f(e,x) = (|e| + 1)^2\, [2\sup_{g\in\mathcal{G}}|g(x)| + 1]^2$$

since e^2 is integrable and $2\sup_{g\in\mathcal{G}}|g(x)| + 1 \leq 2\|g\|_{0,\infty,\mathcal{X}} + 1$ $\leq 2c\|g\|_{m+p/r+1,p,\mathcal{X}} + 1 \leq 2cB + 1$.

Condition (c) of Theorem 0 is now satisfied with

$$\bar{s}(g,g^*) = \sigma^2 + \int_{\mathcal{X}}[g^*(x) - g(x)]^2 d\mu(x).$$

More general Uniform Strong Laws are readily available. For example, Gallant and White (1987) obtain a Uniform Strong Law when $\{(e_t,x_t)\}$ is a heterogeneous, mixing process. The basic requirements of these strong laws are the same as the illustration above. Some sort of stability condition on the process $\{(e_t,x_t)\}$ and a domination condition are required.

As regards Condition (d) of Theorem 0. Let us first consider the case when $\mu(\mathcal{O}) > 0$ for open subsets of $\mathcal{X}$. The implication of $\bar{s}(g^0,g^*) \leq \bar{s}(g^*,g^*)$ is $\int_{\mathcal{X}}[g^*(x) - g^0(x)]^2 d\mu(x) = 0$. Since both g^* and g^0 are continuous on $\mathcal{X}$, as they are both elements of $\mathcal{G} \subset \mathcal{W}_{m,\infty,\mathcal{X}}$, and $\mu(\mathcal{O}) > 0$ for every $\mathcal{O} \subset \mathcal{X}$ the implication of $\int_{\mathcal{X}}[g^*(x) - g^0(x)]^2 d\mu(x) = 0$ is that $g^*(x) \equiv g^0(x)$ for all x in $\mathcal{X}$. Thus, $\bar{s}(g^0,g^*) \leq \bar{s}(g^*,g^*)$ implies $\|g^0 - g^*\|_{m,\infty,\mathcal{X}} = 0$ with the consequence that $\sigma(\hat{g}_{K_n})$ is strongly consistent for every σ that is continuous with respect to the consistency norm $\|\cdot\|_{m,\infty,\mathcal{X}}$, $\sigma(g) = \int_{\mathcal{X}} g(x)dx$ for instance.

Now suppose, for example, that the training sample does not cover the entire input space in the sense that $\mu(\mathcal{O}) > 0$ for $\mathcal{O} \subset \mathcal{Y}$, where $\mathcal{Y}$ is the closure of some open subset of $\mathcal{X}$, and $\mu(\mathcal{O}) = 0$ for $\mathcal{O} \subset \mathcal{X}\cap\tilde{\mathcal{Y}}$. The same argument as above shows that the functionals estimated consistently are those that are continuous with respect to $\|\cdot\|_{m,\infty,\mathcal{Y}}$. The network can no longer estimate the functional $\sigma(g) = \int_{\mathcal{X}} g(x)dx$ consistently but it can estimate $\sigma(g) = \int_{\mathcal{Y}} g(x)dx$ consistently. That is, the network doesn't learn where it isn't trained.

5. SUMMARY AND MAIN RESULT

We summarize by collecting together in one place an internally consistent set of conditions that imply strong consistency. As indicated in the foregoing discussion, these conditions can be modified considerably as required by an application. However, any modification of a condition will usually have side effects that require modification of another.

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

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

where x represents an $r \times 1$ vector of network inputs 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

$$s_n(g) = \frac{1}{n} \sum_{t=1}^{n} [y_t - g(x_t)]^2$$

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 $\mathcal{X}$ 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 $W_{m+[r/p]+1,p,\mathcal{X}}$ and $\|g^*\|_{m+[r/p]+1,p,\mathcal{X}} < B$.

Activation function. The activation function G is in $W_{m,\infty,\mathcal{X}}$ 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 $\mathcal{G} = \{g: \|g\|_{m+[r/p]+1,p,\mathcal{X}} \leq B\}$ 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 $\mu(\mathcal{O}) > 0$ for every open subset $\mathcal{O}$ of $\mathcal{X}$.

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

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

As shown in Sections 3 and 4, the Regularity Conditions are sufficient to verify the conditions of Theorem 0 which implies the following result:

THEOREM 1. Under the Regularity Conditions,

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

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 $\|\cdot\|_{m,\infty,\mathcal{X}}$. I

Note that the condition "provided $\lim_{n\to\infty} K_n = \infty$ almost surely" permits random rules such as cross validation (Stone, 1984).

6. INVERSE DETERMINATION OF THE NONLINEAR MAP OF A CHAOTIC PROCESS.

An exciting new application of neural networks is to the inverse problem of chaotic dynamics: "given a sequence of iterates construct a nonlinear map that gives rise to them" (Casdagli, 1989). There are a number of approximation methods available to estimate the map from a finite stretch of data. Neural nets were found to be competitive with the best of the approximation methods that Casdagli studied and were found by Lapedes and Farber (1987) to perform significantly better than several methods in common use. We shall illustrate the theory of the preceding sections by extending the analysis of these authors with an examination of the accuracy to which neural nets can recover the derivatives of a nonlinear map. We shall use the methods suggested by Casdagli, where for the reader's convenience, we have translated Casdagli's notation to ours.

Casdagli's setup is as follows. $g: \mathcal{X} \to \mathcal{X} \subset \mathbb{R}^r$ is a smooth map with strange attractor $\mathcal{A}$ and ergodic natural invariant measure μ (Schuster, 1988). A time series x_t for $-L \le t < \infty$ has been generated by iterating this map according to

$$\begin{aligned} x_t &= g(x_{t-L-1}, \ldots, x_{t-1}) \\ &= g[g(x_{t-L-2}), \ldots, g(x_{t-2})] \\ &\vdots \\ &= g^t(x_{-L}, \ldots, x_0) \end{aligned}$$

where $x_{-L}, \ldots, x_0$ is a sequence of points from $\mathcal{A}$ that obey the iterative sequence above. Of this series, the stretch of x_t for $-L \le x_t \le N$ is available for analysis and the stretch of x_t for $N < t \le 2N$ is used as a hold-out sample

to assess the quality of estimates. In principle, one can solve the inverse problem by constructing a unique, smooth map g^* that agrees with g on $\mathcal{G}$ from the infinite sequence $\{x_t\}_{t=-L}^{\infty}$. In practice, one should like to find a good approximant $\hat{g}_{K_n}$ to g^* that can be constructed from the finite sequence $\{x_t\}_{t=-L}^{n}$ where $n \leq N$.

The approximant $\hat{g}_{K_n}$ can be put to a variety of uses: detection of chaos, prediction of x_{t+j} given x_t, determination of the invariant measure μ, determination of the attractor $\mathcal{G}$, prediction of bifurcations, and determination of the largest Lyapunov exponent via Jacobian based methods such as discussed in Shimada and Nagashima (1970) and Eckmann *et. al.* (1986). In the last mentioned application, accurate estimation of first derivatives is of critical importance.

Our investigation studies the ability of the single hidden layer network

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

with logistic squasher

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

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

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

This map is of special interest in economic applications because it alone, of many that we tried, can generate a time series that is qualitatively like financial market data (Gallant, Hsieh, and Tauchen, 1989) especially in its ability to generate stretches of extremely volatile data of apparently random duration. Notice that the approximant is handicapped as the dimension of the approximant is higher than is necessary as it has five arguments when a lesser number would have sufficed. We view this as realistically mimicking actual applications as one is likely to overestimate the minimal dimension as a precaution against the worse error of getting it too small. Casdagli's methods for determining dimension suggest that there is a representation of g^* in at most three dimensions $(x_{t-3}, x_{t-2}, x_{t-1})$.

Casdagli suggests that the flexibility of an approximant be increased until improvement in the predictor error $\text{PredErr}(\hat{g}_K)$ becomes negligible. The predictor error can be estimated from the holdout sample using

$$\text{PredErr}^2(\hat{g}_K) \simeq \frac{1}{N} \sum_{t=N+1}^{2N} [x_t - \hat{g}_K(x_{t-5}, \ldots, x_{t-1})]^2/\text{Var}$$

where

$$\text{Var} \simeq \frac{1}{N} \sum_{t=N+1}^{2N} (x_t - \bar{x})^2$$

$$\bar{x} \simeq \frac{1}{N} \sum_{t=N+1}^{2N} x_t.$$

Similarly, the Sobolev norm over $\mathcal{G}$ (not over $\mathcal{X}$) of the approximation error can be estimated from the hold-out sample using

$$\|g^* - \hat{g}_K\|_{m,p,\mathcal{Y}} \simeq \left[\sum_{|\lambda| \le m} \frac{1}{N} \sum_{t=N+1}^{2N} |D^\lambda g(x_{t-5}, x_{t-1}) - D^\lambda \hat{g}_K(x_{t-5}, \ldots, x_{t-1})|^p \right]^{1/p}$$

$$\|g^* - \hat{g}_K\|_{m,\infty,\mathcal{Y}} \simeq \max_{|\lambda| \le m} \max_{N+1 \le t \le 2N} |D^\lambda g(x_{t-5}, x_{t-1}) - D^\lambda \hat{g}_K(x_{t-5}, \ldots, x_{t-1})|$$

We took N as 10,000 in these formulas because we wanted very accurate estimates of PredErr($\hat{g}_K$), $\|g^* - \hat{g}_K\|_{m,p,\mathcal{Y}}$, and $\|g^* - \hat{g}_K\|_{m,\infty,\mathcal{Y}}$. In ordinary applications, one would use a much smaller hold-out sample to estimate PredErr($\hat{g}_K$); $\|g^* - \hat{g}_K\|_{m,p,\mathcal{Y}}$ and $\|g^* - \hat{g}_K\|_{m,\infty,\mathcal{Y}}$ would not ordinarily be estimated since they cannot be determined without knowledge of either g or g^* and if either g or g^* were known the inverse problem has no content. Also, note that

$$\text{PredErr}(\hat{g}_K) = \|g^* - \hat{g}_K\|_{0,2,\mathcal{Y}} / \sqrt{\text{Var}}.$$

For our data, described below, $\sqrt{\text{Var}} = 0.80749892$ so PredErr is about a 20% over-estimate of $\|g^* - \hat{g}_K\|_{0,2,\mathcal{Y}}$.

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 (Gallant, 1989, Ch. 1). We found it helpful to zig-zag by first holding $\hat{\beta}_j$ fixed and iterating on the $\hat{\gamma}_{ij}$, then holding the $\hat{\gamma}_{ij}$ fixed and iterating on the $\hat{\beta}_j$, and so on a few times before going to the full Gauss-Newton iterates. Our rule relating K to n was of the form $K \alpha \log(n)$ because asymptotic theory in a related context (Gallant, 1989) suggests that this is likely to be the relationship that will give stable estimates. The numerical results are in Table 1.

We experimented with other values for n relative to K and found that results were not very sensitive to the choice of n relative to K except in the case n=500 with K=11. The case K=11 has 77 weights to be determined from 500 observations giving a saturation ratio of 6.5 observations per weight, which is rather an extreme case. The results of the sensitivity analysis are in Table 2.

In graphical presentations of $g(x_{-5},0)$ and $\hat{g}(x_{-5},x_{-4},..,x_{-1})$ and their partial derivatives, the effect of x_{-5} totally dominates. Thus, plots of $g(x_{-5},0)$, $(\partial/\partial x_{-5})g(x_{-5},0)$, $\hat{g}(x_{-5},0,0,0,0)$ and $(\partial/\partial x_{-5})\hat{g}(x_{-5},0,0,0,0)$ against x_{-5} give one an accurate visual impression of the adequacy of an approximation. This fact can be confirmed by comparing the error estimates in a row of Table 1 with the scale of the vertical axes of the figures that corresponds to that row. The figures and tables suggest that following Casdagli's (1989) suggestion of increasing the flexibility of an approximation until $\text{PredErr}(\hat{g}_K)$ shows no improvement does lead to estimates of the nonlinear map and its derivatives that appear adequate for the applications mentioned above.

The computations reported in the figures and tables would seem to confirm the findings of Casdagli (1989) and Lapedes and Farber (1987) as to the appropriateness of neural net approximations in addressing the inverse problem of chaotic dynamics. They also suggest that our theoretical results will be of practical relevance in the determination of the derivatives of a map in training samples of reasonable magnitudes.

Table 1. Predictor Error and Error in Sobolev Norm of an Estimate of the Nonlinear Map of a Chaotic Process by a Neural Net.

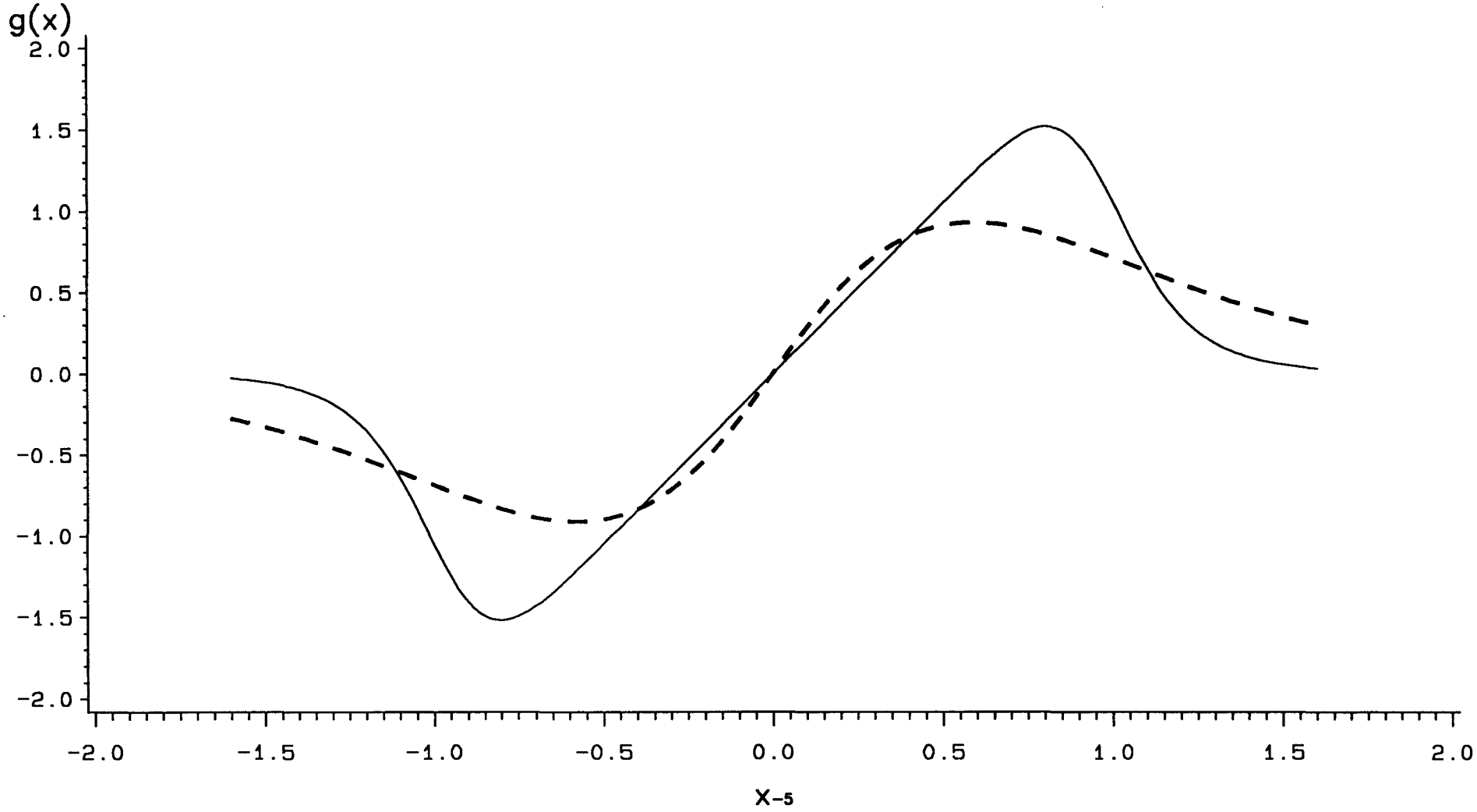
K	n	PredErr($\hat{g}_K$)	$\|g^* - \hat{g}_K\|_{1,\infty,\mathcal{G}}$	$\|g^* - \hat{g}_K\|_{1,2,\mathcal{G}}$	Saturation Ratio
3	500	0.3482777075	3.6001114788	1.3252165780	17.9
5	1000	0.0191675679	0.5522597668	0.1604392912	28.6
7	2000	0.0177867857	0.4145203548	0.1141557050	40.8
9	4000	0.0134447868	0.2586038122	0.0719887443	63.5
11	8000	0.0012308988	0.1263063691	0.0196351730	103.9

Table 2. Sensitivity of Neural Net Estimates.

K	n	PredErr($\hat{g}_K$)	$\|g^* - \hat{g}_K\|_{1,\infty,\mathcal{G}}$	$\|g^* - \hat{g}_K\|_{1,2,\mathcal{G}}$	Saturation Ratio
7	500	0.0184102390	0.3745884157	0.1325439320	10.2
7	2000	0.0177867857	0.4145203548	0.1141557050	40.8
11	500	0.0076063363	0.7141377059	0.1115357981	6.5
11	4000	0.0015057013	0.0858882780	0.0210710677	51.9
11	8000	0.0012308988	0.1263063691	0.0196351730	103.9
15	8000	0.0020546210	0.1125778860	0.0336124596	76.2

Figure 1. Superimposed nonlinear map and neural net estimate

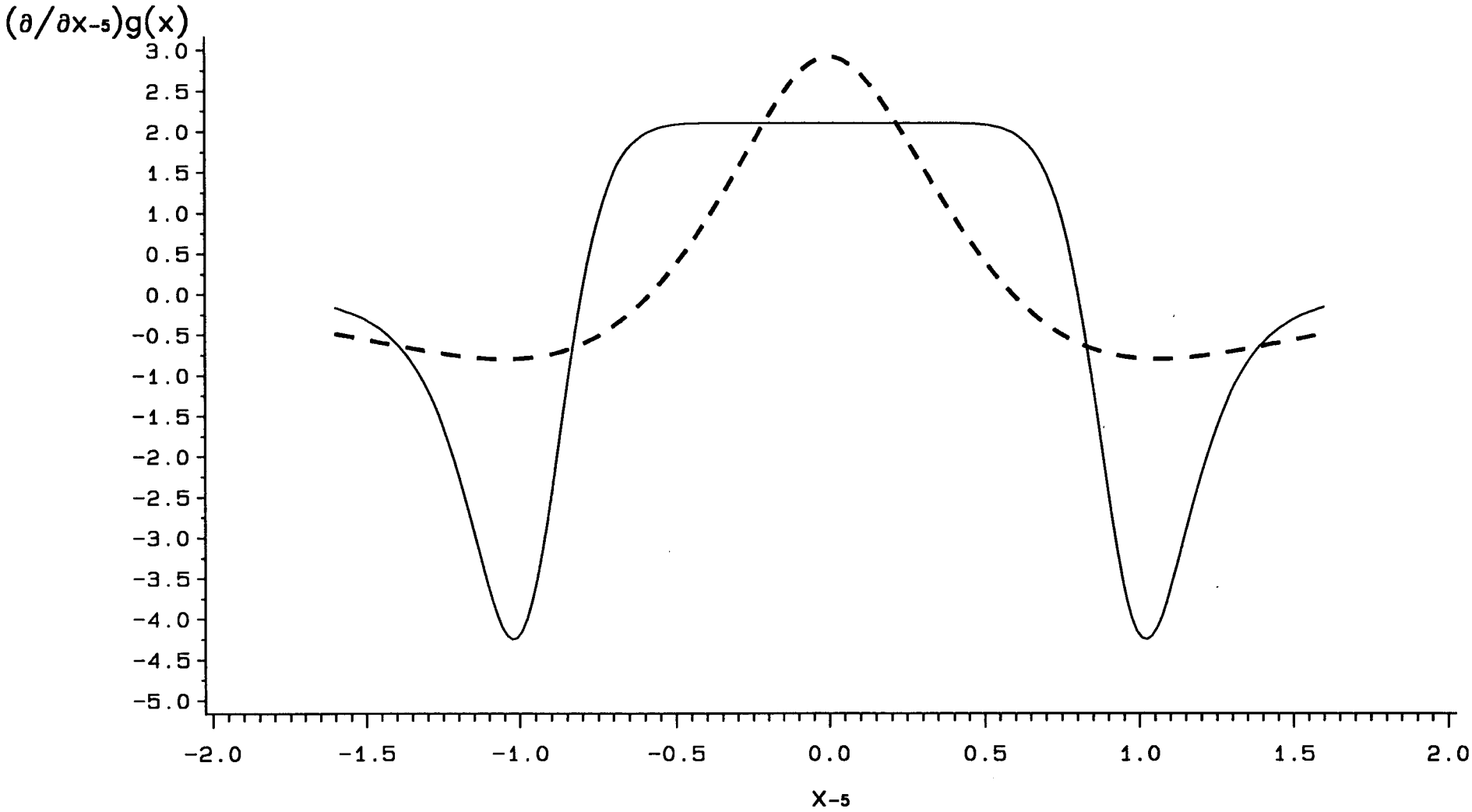
K = 3, n = 500



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

Figure 2. Superimposed derivative and neural net estimate

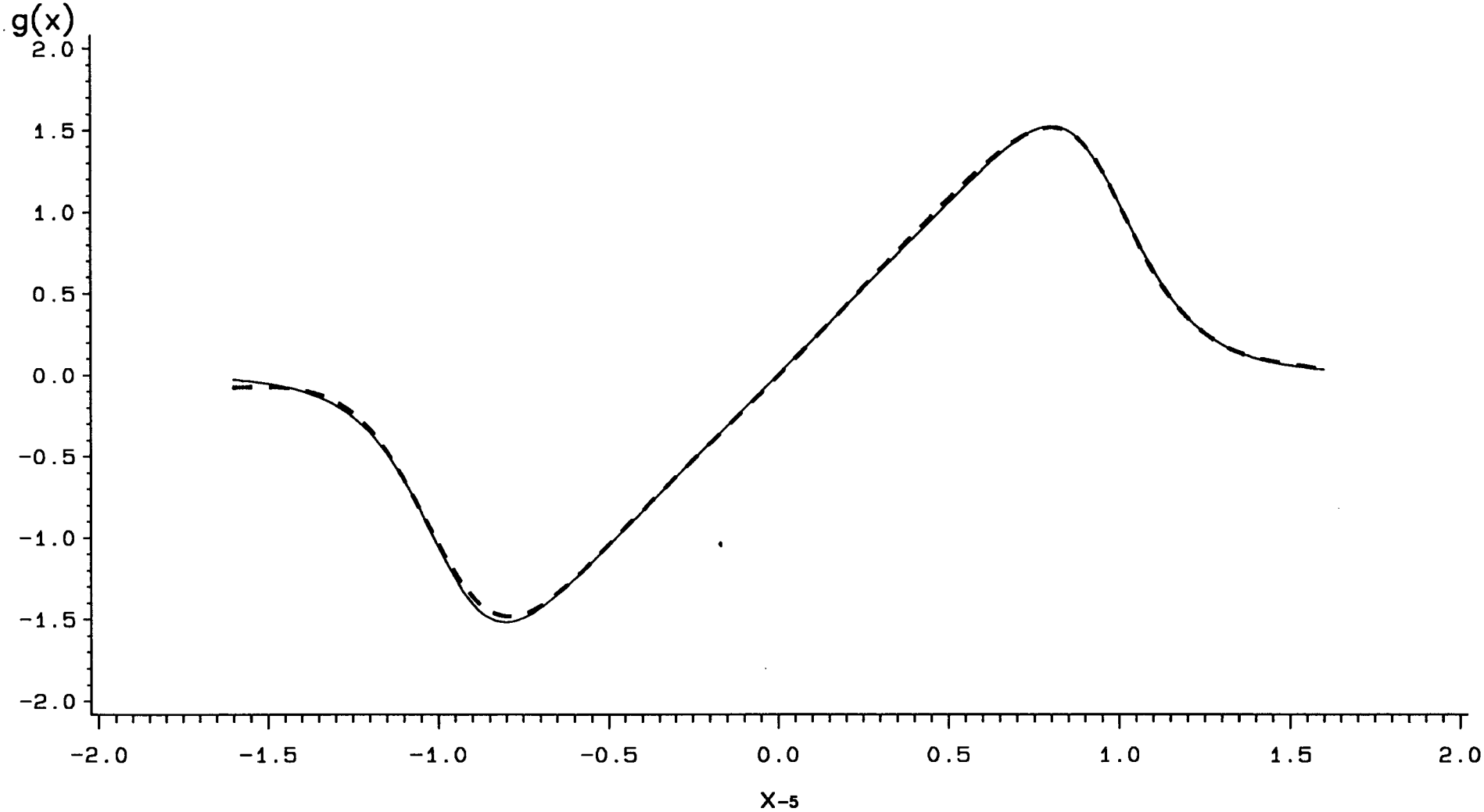
K = 3, n = 500



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

Figure 3. Superimposed nonlinear map and neural net estimate

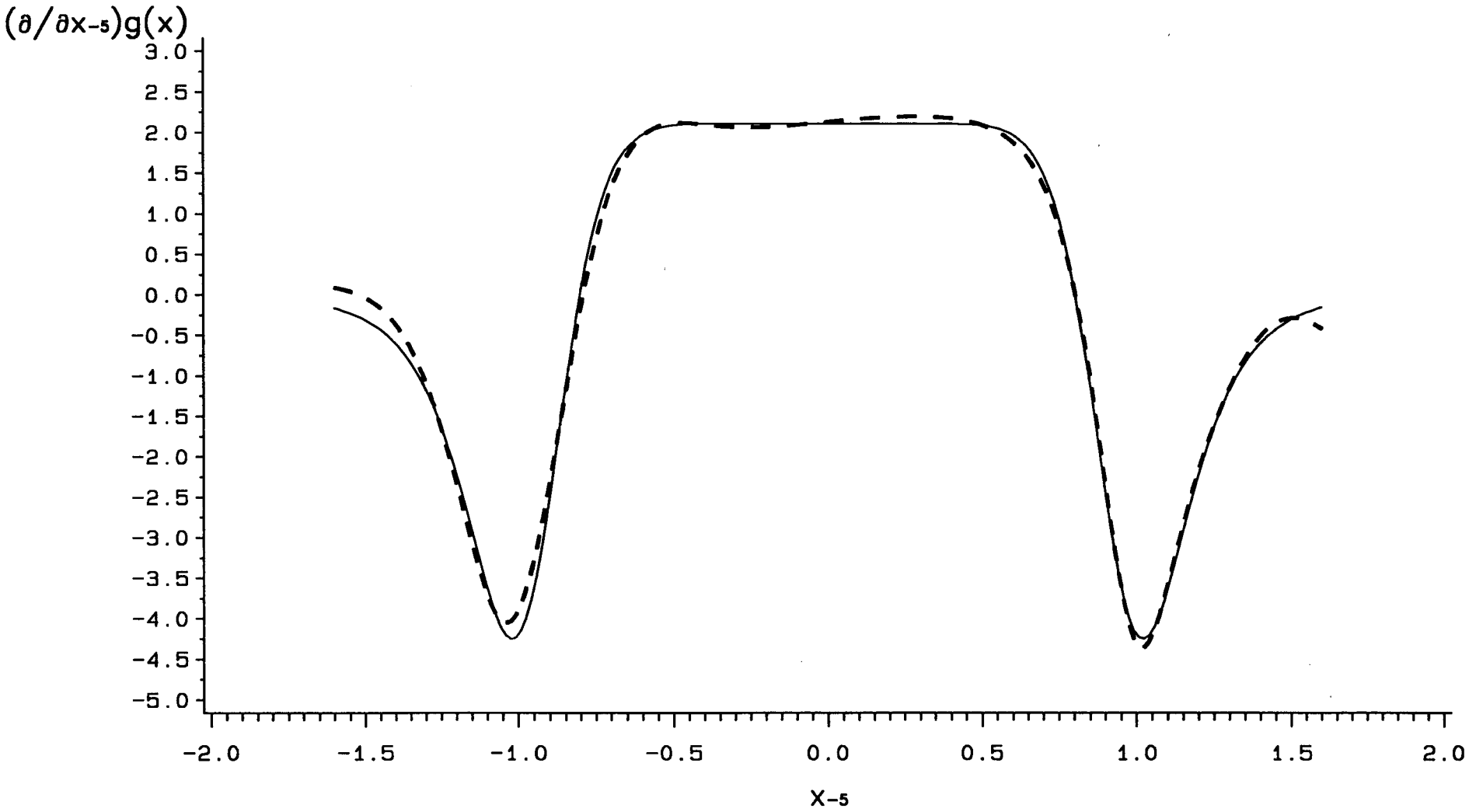
K = 7, n = 2000



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

Figure 4. Superimposed derivative and neural net estimate

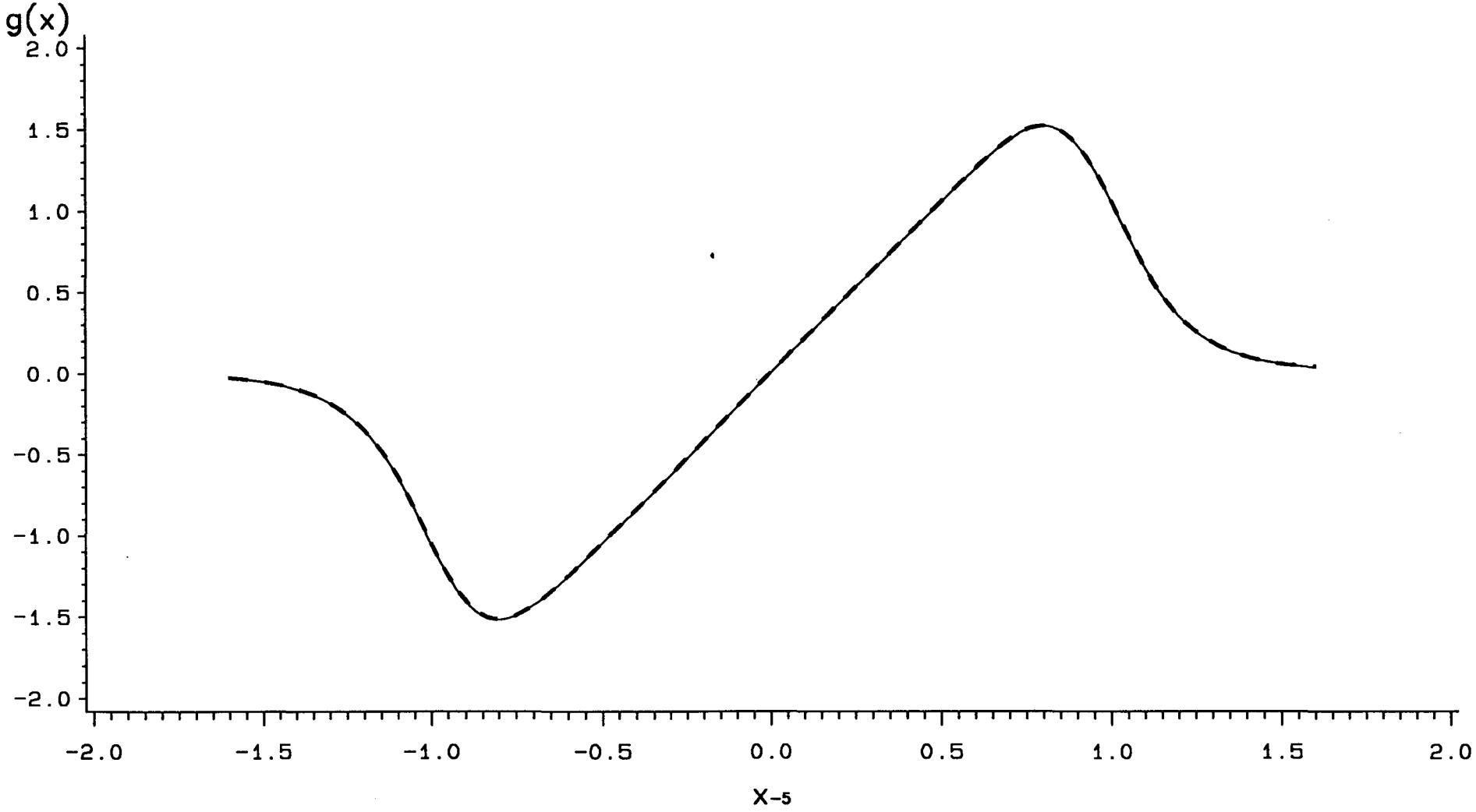
K = 7, n = 2000



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

Figure 5. Superimposed nonlinear map and neural net estimate

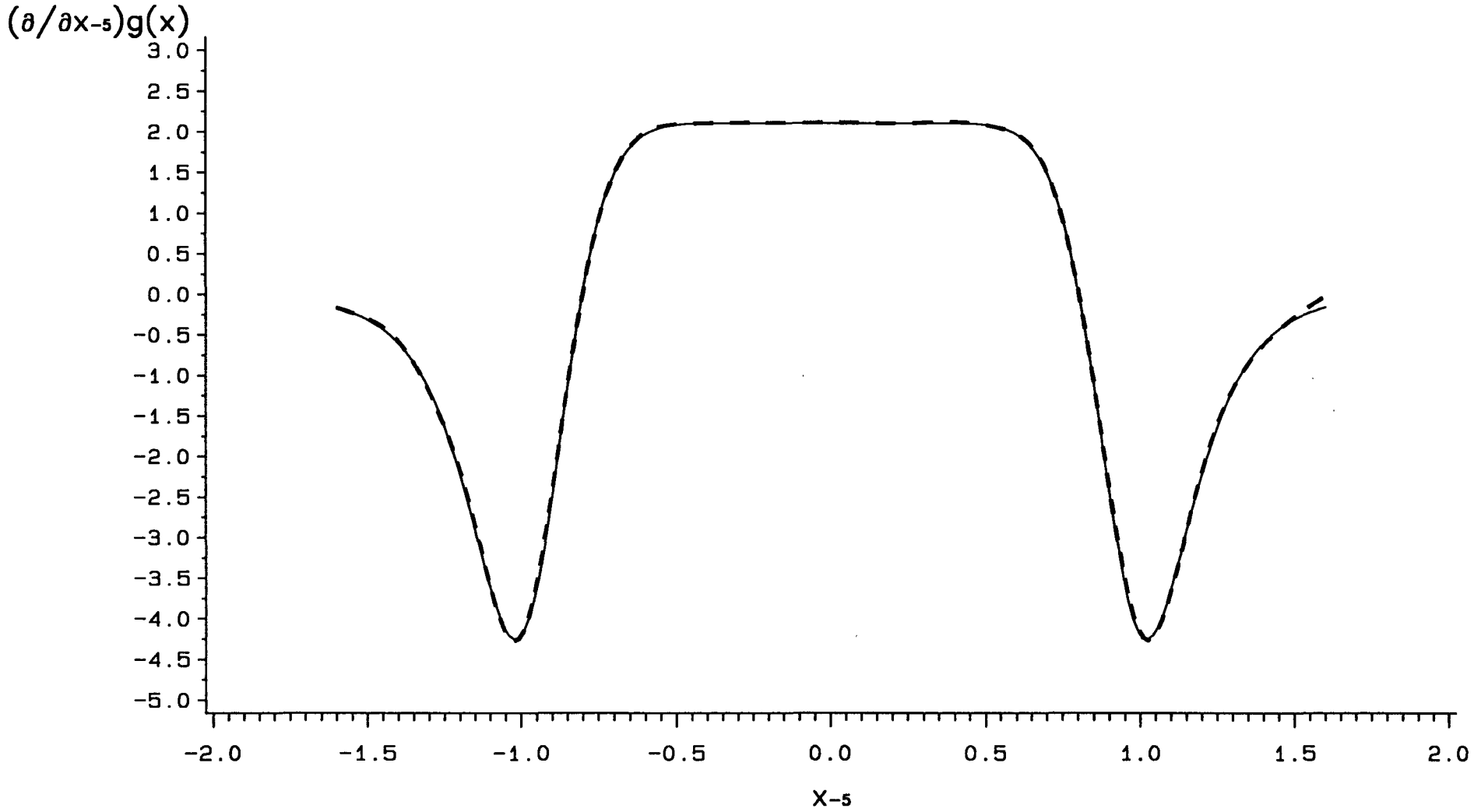
K = 11, n = 8000



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)

REFERENCES

Adams, Robert A. (1975), *Sobolev Spaces*, New York: Academic Press.

Casdagli, Martin (1989), "Nonlinear Prediction of Chaotic Time Series," *Physica D* 35, 335-356.

Eckmann, J.-P., S. Oliffson Kamphorst, D. Ruelle and S. Ciliberto (1986), "Liapunov Exponents from Time Series," *Physical Review A* 34, 4971-4979.

Elbadawi, Ibrahim, A. Ronald Gallant, and Geraldo Souza (1983), "An Elasticity Can be Estimated Consistently Without A Priori Knowledge of Functional Form," *Econometrica* 51, 1731-1752.

Gallant, A. Ronald (1987a), *Nonlinear Statistical Models*. New York: John Wiley and Sons.

Gallant, A. Ronald (1987b), "Identification and Consistency in Seminonparametric Regression", in Truman F. Bewley, ed. *Advances in Econometrics Fifth World Congress, Volume 1*, New York, Cambridge University Press, 145-170. Translated as Gallant, A. Ronald (1985),"Identification et Convergence en Regression Semi-Nonparametrique," *Annals de l'INSEE* 59/60, 239-267.

Gallant, A. Ronald (1989), "On the Asymptotic Normality of When the Number of Regressors Increases and the Minimum Eigenvalue of X'X/n Decreases," Institute of Statistics Mimeograph Series No. 1955, North Carolina State University, Raleigh, NC 27695-8203.

GalTant, A. Ronald, David A. Hsieh, and George E. Tauchen (1989), "On Fitting a Recalcitrant Series: The Pound/Dollar Exchange Rate, 1974-83," in William A. Barnett, James Powell, George E. Tauchen, eds. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge: Cambridge University Press, forthcoming.

Gallant, A. Ronald, and Douglas W. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica* 55, 363-390.

Gallant, A. Ronald, and Halbert L. White, Jr. (1987), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Basil Blackwell Ltd..

Gallant, A. Ronald, and Halbert White (1988), "There Exists a Neural Network that Does Not Make Avoidable Mistakes," *Proceedings of the Second Annual IEEE Conference on Neural Networks, San Diego*. San Diego: SOS Printing, I.657-I.664.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989), "Universal Approximation of an Unknown Mapping and its Derivatives Using Multilayer Feedforward Networks," University of California, San Diego, Department of Economics Discussion Paper.

Jordan, M. (1989), "Generic Constraints on Underspecified Target Trajectories," *Proceedings of the International Joint Conference on Neural Networks, Washington DC.* San Diego: SOS Printing, I.217-I.225.

Lapedes, Alan, and Rober Farber (1987), "Nonlinear Signal Processing Using Neural Networks: Prediction and System Modelling," Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, manuscript.

Rumelhart, D.E., G.E. Hinton and R.J. Williams (1986), "Learning Internal Representations by Error Propagation," in D.E. Rumelhart and J. L. McClelland, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol 1. Cambridge, MA: MIT Press.

Schuster, Heinz Georg (1988), *Deterministic Chaos, An Introduction, Second Revised Edition.* Weinheim, Federal Republic of Germany: VCH Verlagsgesellschaft mbH.

Shimada, Ippei, and Tomomasa Nagashima (1979), "A Numerical Approach to Ergodic Problem of Dissipative Dynamical Systems," *Progress of Theoretical Physics* 61, 1605-1616.

Stone, M. (1974), "Cross-Validitory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Series B 36, 111-133.

Tucker, Howard G. (1967), *A Graduate Course in Probability.* New York: Academic Press.

White, Halbert (1989a), "Some Asymptotic Results for Learning in Single Hidden Layer Feedforward Network Models," *Journal of the American Statistical Association*, forthcoming.

White, Halbert (1989b), "Multilayer Feedforward Networks Can Learn Arbitrary Mappings: Connectionist Nonparametric Regression with Automatic and Semi-automatic Determination of Network Complexity," University of California, San Diego, Department of Economics Discussion Paper.