George Cybenko

Neural Networks, Vol. 5, pp. 129–138, 1992
Printed in the USA.

0893-6080/92 $5.00 + .00

ORIGINAL CONTRIBUTION

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

A. RONALD GALLANT[1] AND HALBERT WHITE[2]

[1]North Carolina State University and [2]University of California, San Diego

(*Received* 29 *November* 1989; *revised and accepted* 20 *June* 1991)

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. 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's 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.*

Keywords—Estimating derivatives, Chastic dynamics, Sobolev spaces, Denseness.

1. INTRODUCTION

Recently, a number of authors have shown that single hidden-layer activation functions are capable of approximating arbitrary functions arbitrarily well, provided sufficiently many hidden units are available (see, for example, Carroll & Dickinson, 1989; Cybenko, 1989; Funahashi, 1989; Gallant & White, 1988; Hecht-Nielsen, 1989; Hornik, Stinchcombe & White, 1989; Stinchcombe & White, 1989). White (1990) has shown that the approximation potential suggested by these results has practical value by proving that arbitrarily accurate approximations to arbitrary functions can be learned; White's proof relies on methods of nonparametric statistics, specifically Grenander's (1981) method of sieves. In this approach, the number of hidden units grows with the size of the training set at just the right rate to ensure good approximation without overfitting.

In some applications, notably robotics (Jordan, 1989), demand analysis (Elbadawi, Gallant, & Souza, 1983), and chaotic dynamics (Schuster, 1988), approximation of the mapping alone will not suffice. Close approximation to both the mapping and the derivatives of the mapping are required in these applications. Hornik, Stinchcombe, and White (1990) (referred to hereafter as HSW) have demonstrated that multiple input, single output, single hidden-layer feedforward networks can approximate not only the mapping, but also its derivatives, provided the hidden layer activation 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 (1990) analysis and provide learning rules ensuring that these networks can learn both the mapping and its derivatives.

Acknowledgements: This research was supported by National Science Foundation Grants SES-8806990, SES-8808015, SES-8921382, North Carolina Agricultural Experiment Station Projects NCO-5593, NCO-3879, and the PAMS Foundation. We thank Stephen P. Ellner, Daniel McCaffrey, and Douglas W. Nychka for helpful discussions relating to chaotic dynamics.

Requests for reprints should be sent to A. Ronald Gallant, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203.

2. HEURISTICS

We consider situations in which training data are generated according to

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

where $\{y_t\}$ is an observable sequence of targets (scalar for simplicity), g^* is an unknown mapping whose derivatives are of interest, $\{x_t\}$ is a sequence of observable inputs taking values in $X \subset R^r, r \in N$, where X is the closure of an open bounded subset of R^r, and $\{e_t\}$ is a sequence of unobserved independently identically distributed (i.i.d.) errors (a noise process) independent of $\{x_t\}$. Some further conditions will be imposed in stating our formal results, but these suffice to set the stage and motivate our approach.

In discussing the derivatives of a function g we use the following standard notation. Let $\lambda = (\lambda_1, \ldots, \lambda_r)'$ be a "multi-index," i.e., a vector of nonnegative integers, and if g has a derivative at x of order $|\lambda| = \sum_{i=1}^{r} |\lambda_i|$, write

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

where $x = (x_1, \ldots, x_r)'$. We also write $D^0 g(x)$ to denote $g(x)$.

We assume that g^* is an element of a Sobolev space $\mathfrak{W}_{m,p,X}$, $m \in N \cup \{0\}$, $p \in N \cup \{\infty\} \equiv \bar{N}$. Elements of this space are functions having continuous derivatives of order m on the domain X that satisfy

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

$$\|g\|_{m,\infty,X} \equiv \max_{|\lambda| \le m} \sup_{x \in X} |D^\lambda g(x)| < \infty.$$

We refer to $\|\cdot\|_{m,p,X}$ or $\|\cdot\|_{m,\infty,X}$ as a "Sobolev norm." See HSW for additional background on Sobolev spaces relevant in the present context.

In applications, interest may attach not just to certain specific derivatives $D^\lambda g$, but also to particular functions of these derivatives, such as

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

the λ-th derivative evaluated at a point x;

$$\sigma(g) = \sup_{x \in X} |D^\lambda g(x)|$$

or

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

the supermum or infinum of the λ-th derivative over X; or

$$\sigma(g) = \int_X f(x) D^\lambda g(x) \, dx, \quad \lambda \le m,$$

the cross-moment of the bounded function f with $D^\lambda g$ over X. Each of these functions σ is continuous over $\mathfrak{W}_{m,\infty,X}$ with respect to $\|\cdot\|_{m,\infty,X}$. Accordingly, we shall seek conditions ensuring that such functions can be learned by our networks. We refer to $\sigma(g^*)$ as the feature of g^* that is of interest.

To approximate g^*, we use single hidden-layer feedforward networks with output given by

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

where $\tilde{x} \equiv (1, x')'$ (a prime $'$ denotes transposition), $x \in X$ is the $r \times 1$ vector of network inputs, G is a given hidden unit activation function, β_j represents hidden to output unit weights, γ_j represents input to hidden unit weights (including a bias), $j = 1, 2, \ldots, K$, and K is the number of hidden units. We collect all the weights together as

$$\delta' = (\beta_1, \ldots, \beta_K, \gamma_1', \ldots, \gamma_K') \in R^{(r+2)K}.$$

In demonstrating that such networks are capable of learning arbitrary mappings, White (1990) considered least squares learning rules of the form

$$\min_{\delta \in D_{K_n}} n^{-1} \sum_{t=1}^{n} [y_t - g_{K_n}(x_t|\delta)]^2,$$

where D_{K_n} is an appropriately restricted subset of $R^{(r+2)K_n}$. We consider the same learning rule here. Note that the number of hidden units, K_n, is taken explicitly to depend on n, the number of available training examples. By permitting $K_n \to \infty$ as $n \to \infty$ we create the opportunity for learning the features of interest of an arbitrary function.

Denote the solution to the least squares problem above as $\hat{\delta}_n$ and define $\hat{g}_{K_n} = g_{K_n}(\cdot|\hat{\delta}_n)$. We can view $\hat{g}_{K_n}$ as the solution to the problem (equivalent to that above)

$$\min_{g \in \mathcal{G}_{K_n}} s_n(g) = n^{-1} \sum_{t=1}^{n} [y_t - g(x_t)]^2,$$

where $\mathcal{G}_{K_n} = \{g_{K_n}(\cdot|\delta), \delta \in D_{K_n}\}$.

With this structure, our goal is to find conditions ensuring that $\sigma(\hat{g}_{K_n}) \to \sigma(g^*)$ as $n \to \infty$ almost surely, as this says that the network learns the features of interest as $n \to \infty$ with probability one.

3. MAIN RESULT

To achieve our goal, we can make use of the following general result of Gallant (1987b). It delivers exactly the conclusion that we are after in a setting that applies directly to our context.

THEOREM 3.1. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, and let $\mathcal{G}$ be a function space on which is defined a norm $\|\cdot\|$. Suppose that $\hat{g}_{K_n} : \Omega \to \mathcal{G}$ is obtained by minimizing a sample objective function $s_n(\cdot)$ over $\mathcal{G}_{K_n}$ where $\mathcal{G}_{K_n}$ is a subset of $\mathcal{G}$. Let $\sigma(\cdot)$ be continuous over $\mathcal{G}$ with respect to $\|\cdot\|$. Suppose the following Conditions hold:

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

(b) Denseness: $\cup_{K=1}^{\infty} \mathcal{G}_K$ is a dense subset of $\bar{\mathcal{G}}$, 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") and there is a function $\bar{s}(g, g^*)$ that is continuous in g with respect to $\|\cdot\|$ such that

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

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

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

must have

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

Conclusion: If $\lim_{n \to \infty} K_n = \infty$ *a.s.*, then

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

For convenience, the proof is given in the Mathematical Appendix.

This is a general purpose theorem. The objects it refers to are abstract, and do not need to coincide with the objects defined in the preceding section. Of course, it is by identifying the objects previously defined with those referred to here that we can accomplish our goal.

Thus, it will suffice to provide additional structure that will permit us to apply this result. We must satisfy Conditions (a)–(d) for appropriate choices of $\mathcal{G}$, $\mathcal{G}_K$, $\|\cdot\|$ and $\bar{s}$.

To ensure the compactness required in (a), we begin by assuming that g^* belongs to $\mathfrak{W}_{m+\lfloor r/p \rfloor + 1, p, X}$ for some p with $1 \le p < \infty$, where $[r/p]$ denotes the integer part of r/p, and m is the largest derivative of g^* that we are interested in. Further, suppose there is available an *a priori* fiite bound B on $\|g^*\|_{m+\lfloor r/p \rfloor + 1, p, X}$. We therefore can restrict our attention to

$$\mathcal{G} = \{g \in \mathfrak{W}_{m+\lfloor r/p \rfloor + 1, p, X} | \|g\|_{m+\lfloor r/p \rfloor + 1, p, X} \le B\}.$$

Then we take $\|\cdot\|$ in Theorem 3.1 to be $\|\cdot\|_{m,\infty,X}$. By the Rellich–Kondrachov Theorem (Adams, 1975, Theorem 6.2 Part II), the closure of $\mathcal{G}$ with respect to the norm $\|\cdot\|_{m,\infty,X}$ is compact in the relative topology generated by $\|\cdot\|_{m,\infty,X}$. Condition (a) of Theorem 3.1 is now satisfied.

Note that the stronger is the norm $\|\cdot\|$ in Theorem 3.1, the larger the class of functions continuous with respect to it, and the more the network can be said to have learned about g^*. The norm $\|\cdot\|_{m,\infty,X}$ is very strong, so our result will imply that the network learns a great deal about g^*.

Now consider Condition (b) of Theorem 3.1. HSW (Corollary 3.4) give mild conditions on the activation function G ensuring that $\cup_K \bar{\mathcal{G}}_K$ is dense in $\mathfrak{W}_{m,\infty,X}$, where

$$\bar{\mathcal{G}}_K = \{g : R^r \longrightarrow R | g(x) = g_K(x|\delta), \delta \in R^{(r+2)K}\}.$$

The sufficient condition on the activation function G is that it be "m-finite," i.e., continuously differentiable of order m, with $\int |D^l G(a)| \, da < \infty$ for some $0 \le l \le m$. The familiar logistic and hyperbolic tangent squashers satisfy this condition. By taking

$$\mathcal{G}_K = \bar{\mathcal{G}}_K \cap \mathcal{G},$$

we therefore ensure that Condition (b) of Theorem 3.1 is satisfied.

We 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 $\|g_K(\cdot|\delta)\|_{m+\lfloor r/p \rfloor + 1, p, X} \le B$, which is a parametric restriction on δ, must be enforced in the minimization of $s_n(g)$ over $g \in \mathcal{G}_K$, equivalently, in the minimization of $s_n[g_K(\cdot|\delta)]$ over $\delta \in R^{(r+2)K}$. In practice, restricting $(r + 2)K$ 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.

Next, we verify Condition (c) with

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

The function $\bar{s}$ required by Condition (c) is delivered by an appropriate uniform law of large numbers (ULLN). A convenient strong ULLN is given by Gallant (1987a). To state this result, we say that the empirical distribution μ_n of $\{x_t\}_{t=1}^{n}$ converges weakly to a probability distribution μ almost surely if $\mu_n(x) \to \mu(x)$ at every point x where μ is continuous, almost surely, where

$$\mu_n(x) = n^{-1}(\#\text{ of } x_t \le x \text{ coordinate by coordinate}, \quad 1 \le t \le n);$$

and we write $\mu_n \Rightarrow \mu$ *a.s.* This is a mild condition on $\{x_t\}$. It holds for ergodic chaotic processes as well as ergodic random processes, deterministic replication schemes, and fill-in rules such as 0, 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{3}{4}$, ... Gallant's (1987a, p. 159) ULLN can be stated as follows:

THEOREM 3.2. Let $\{e_t\}$ and $\{x_t\}$ be independent sequences of random vectors taking values in $\mathcal{E}$ and X, respectively, subsets of finite dimensional Euclidean spaces, and suppose that: (i) $\{e_t\}$ is an i.i.d. sequence with common distribution P on $\mathcal{E}$; and (ii) $\{x_t\}$ is such that $\mu_n \Rightarrow \mu$ *a.s.*

Let $\mathcal{G}$ be a compact metric space, and suppose that $f : \mathcal{E} \times X \times \mathcal{G} \to R$ is a continuous function dominated by an integrable function $d : \mathcal{E} \times X \to R^+$ (i.e., $|f(e, x, g)| \le d(e, x)$ for all g in $\mathcal{G}$ and $\int_{\mathcal{E}} \int_X d(e, x) P(de) \mu(dx) < \infty$).

Then $g \to \int_{\mathcal{E}} \int_X f(e, x, g) P(de) \mu(dx)$ is continuous on $\mathcal{G}$, and as $n \to \infty$

$$\sup_{g \in \mathcal{G}} \left| n^{-1} \sum_{t=1}^{n} f(e_t, x_t, g) - \int_{\mathcal{E}} \int_X f(e, x, g) P(de) \mu(dx) \right| \longrightarrow 0 \ a.s. \qquad \blacksquare$$

Applying this result on the compact metric space $(\mathcal{G}, \rho)$, with $\mathcal{G}$ as above, ρ the metric induced by $\|\cdot\|_{m,\infty,X}$ with P such that $E(e_t) = \int_{\mathcal{E}} eP(de) = 0$, $E(e_t^2) = \int_{\mathcal{E}} e^2 P(de) \equiv \sigma_e^2 < \infty$; and with $f(e, x, g) = [y - g(x)]^2 =$

$[e + g^*(x) - g(x)]^2$ gives

$$s_n(g) = n^{-1}\sum_{t=1}^{n} f(e_t, x_t, g)$$

$$= n^{-1}\sum_{t=1}^{n} [e_t + g^*(x_t) - g(x_t)]^2$$

converging to

$$\bar{s}(g, g^*) = \int_{\mathcal{E}}\int_{X} [e + g^*(x) - g(x)]^2 P(de)\mu(dx)$$

$$= \int_{\mathcal{E}} e^2 P(de) + 2\int_{\mathcal{E}} eP(de)\int_{X} [g^*(x) - g(x)]\mu(dx)$$

$$+ \int_{X} [g^*(x) - g(x)]^2\mu(dx)$$

$$= \sigma_e^2 + \int_{X} [g^*(x) - g(x)]^2\mu(dx),$$

provided that f is appropriately dominated. Because $|a + b|^2 \le 2|a|^2 + 2|b|^2$, we have $f(e, x, g) \le 2|e|^2 + 2|g^*(x) - g(x)|^2$, and we can take

$$d(e, x) = 2|e|^2 + 4\sup_{g\in\mathcal{G}} |g(x)|^2.$$

Now $|g(x)| \le \|g\|_{0,\infty,X} \le \|g\|_{m,\infty,X}$. By the Rellich–Kondrachov Theorem, the Sobolev norms are interleaved in the sense that there exists a constant c not depending on g such that

$$\|g\|_{m,\infty,X} \le c\|g\|_{m+\lfloor r/p\rfloor+1,p,X} \le c\|g\|_{m+\lfloor r/p\rfloor+1,\infty,X}.$$

Therefore $|g(x)| \le c\|g\|_{m+\lfloor r/p\rfloor+1,p,X}$, so that for all x in X and all g in $\mathcal{G}$ $|g(x)| \le cB$. Consequently, $d(e, x) \le 2|e|^2 + 4c^2B^2$, and

$$\int_{\mathcal{E}}\int_{X} d(e, x) \le \int_{\mathcal{E}}\int_{X} (2|e|^2 + 4c^2B^2)P(de)\mu(dx)$$

$$= 2\sigma_e^2 + 4c^2B^2 < \infty,$$

as required. Condition (c) therefore holds.

Now consider Condition (d) of Theorem 3.1. Let us first treat the case when $\mu(O) > 0$ for open subsets O of X. The implication of $\bar{s}(g^0, g^*) \le \bar{s}(g^*, g^*)$ is $\int_X [g^*(x) - g^0(x)]^2\mu(dx) = 0$. Since both g^* and g^0 are continuous on X, as they are both elements of $\mathcal{G} \subseteq \mathcal{W}_{m,\infty,X}$, and $\mu(O) > 0$ for every $O \subseteq X$ the implication of $\int_X [g^*(x) - g^0(x)]^2\mu(dx) = 0$ is that $g^*(x) \equiv g^0(x)$ for all x in X. Thus, $\bar{s}(g^0, g^*) \le \bar{s}(g^*, g^*)$ implies $\|g^0 - g^*\|_{m,\infty,X} = 0$ with the consequence that $\sigma(g^0) = \sigma(g^*)$ whenever σ is continuous with respect to $\|\cdot\|_{m,\infty,X}$, as assumed.

Next suppose that the training sample does not cover the entire input space in the sense that $\mu(O) > 0$ for $O \subset \mathcal{Z}$, where $\mathcal{Z}$ is the closure of some open subset of X and $\mu(O) = 0$ for $O \subseteq X \cap \mathcal{Z}^c$. The same argument as above ensures that $\sigma(g^0) = \sigma(g^*)$ whenever σ is continuous with respect to $\|\cdot\|_{m,\infty,\mathcal{Z}}$. For example, $\sigma(g) = \int_{\mathcal{Z}} g(x)\,dx$ is continuous with respect to $\|\cdot\|_{m,\infty,\mathcal{Z}}$, but $\sigma(g) = \int_X g(x)\,dx$ is not. Consequently, and as should be expected, the network will not be able to learn where it isn't trained.

Collecting together the structure set out above, we can state a set of formal conditions that permit us to verify the conditions of Theorem 3.1, and thus assert its conclusion. Our first assumption describes how the training data are generated.

ASSUMPTION A.1. The training observations are generated as

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

where $\{e_t\}$ and $\{x_t\}$ are independent sequences of random vectors taking values in $\mathcal{E} \subseteq R$ and $X \subset R^r$, $r \in N$, respectively, with X the closure of an open bounded set; and $g^* \in \mathcal{G}$, where

$$\mathcal{G} = \{g \in \mathcal{W}_{m+\lfloor r/p\rfloor+1,p,X} : \|g\|_{m+\lfloor r/p\rfloor+1,p,X} \le B\},$$

for some $m \in N \cup \{0\}$, $p \in \overline{N}$, and $B < \infty$.

Further, the errors $\{e_t\}$ are i.i.d. sequence with common distribution P on $\mathcal{E}$, $\int_{\mathcal{E}} eP(de) = 0$, $\sigma_e^2 \equiv \int_{\mathcal{E}} |e|^2 P(de) < \infty$. The inputs $\{x_t\}$ are a sequence such that $\mu_n \Rightarrow \mu$ *a.s.*, where μ_n is the empirical distribution of $\{x_t\}_{t=1}^n$ and μ is a probability distribution on $(X, B(X))$ such that $\mu(O) > 0$ for every open subset of X. ■

Next we formally specify the networks to be trained.

ASSUMPTION A.2. For $K = 1, 2, \ldots$, let $\mathcal{G}_K = \tilde{\mathcal{G}}_K \cap \mathcal{G}$, where

$$\tilde{\mathcal{G}}_K = \left\{g: R^r \longrightarrow R : g(x) = \sum_{j=1}^{K} \beta_j G(x'\gamma_j),\right.$$

$$\left.\beta_j \in R,\ \gamma_j \in R^{r+1},\ j = 1, \ldots, K\right\},$$

where G is an m-finite activation function. ■

The discussion above establishes the following result for least-squares learning of a function and its derivatives.

THEOREM 3.3. Suppose Assumptions A.1 and A.2 hold, and let $\hat{g}_{K_n}$ be a solution to the problem

$$\min_{g\in\mathcal{G}_{K_n}} s_n(g) = n^{-1}\sum_{t=1}^{n} [y_t - g(x_t)]^2.$$

Let $\sigma(\cdot)$ be continuous with respect to $\|\cdot\|_{m,\infty,X}$. If $K_n \to \infty$ as $n \to \infty$ *a.s.*, then $\sigma(\hat{g}_{K_n}) \to \sigma(g^*)$ as $n \to \infty$ *a.s.* In particular, $\|\hat{g}_{K_n} - g^*\|_{m,\infty,X} \to 0$ as $n \to \infty$ *a.s.* ■

The last conclusion follows by taking $\sigma(g) = \|g - g^*\|_{m,\infty,X}$. This is easily seen to be continuous as required.

Thus, the method of least squares can be used to train a single hidden-layer feedforward network to learn an unknown mapping and its derivatives

in this satisfyingly precise sense. The derivatives are learned without having to make a special effort to provide them as targets; this is quite convenient in applications.

Note that the condition $K_n \to \infty$ *a.s.* permits random rules for determining network complexity, such as cross-validation (Stone, 1974). In White (1990), specific rates for the growth of K_n with n are given. These are unnecessary here because unlike White (1990) the space $\mathcal{G}$(Θ in White's (1990) notation) is compact.

Because we assumed that the training set covers all of X, we obtain the strong result $\|\hat{g}_{K_n} - g^*\|_{m,\infty,X} \to 0$ *a.s.* As illustrated in the discussion above, had we assumed that the training set covered only a subset $\mathcal{Z}$ of X, then we would have the weaker result $\|\hat{g}_{K_n} - g^*\|_{m,\infty,\mathcal{Z}} \to 0$ *a.s.* In this case, the only functions σ that can be estimated consistently are those invariant to a redefinition of g on $X - \mathcal{Z}$. The example in the next section in which the training set is the realization of a chaotic process is a practical illustration of this situation. Because transients die out, the training set eventually must lie entirely within a strange attractor $\mathcal{Z}$, which is a subset of the phase space X over which the nonlinear map g that describes the dynamics is defined.

The specific conditions on the stochastic processes imposed in Assumption A.1 can be considerably modified and relaxed. Their primary function is to ensure the validity of the ULLN. ULLNs are available for quite general stochastic processes with compact space $\mathcal{G}$ (e.g., Andrews, 1990).

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

An exciting recent 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 other methods in common use. We 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 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: $X \to X \subseteq R^r$ is a smooth map with strange attractor $\mathcal{Z}$ and ergodic natural invariant measure μ (Schuster, 1988). A time series x_t for $-L \le t \le \infty$ is generated by iterating this map according to

$$x_t = g(x_{t-L-1}, \ldots, x_{t-1})$$
$$\vdots$$
$$= g^t(x_{-L}, \ldots, x_0),$$

where $x_{-L}, \ldots, x_0$ is a sequence of points from $\mathcal{Z}$ 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 $\mathcal{Z}$ from the infinite sequence $\{x_t\}_{t=-L}^{\infty}$. In practice, one would 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 $n \le N$.

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 $\mathcal{Z}$, prediction of bifurcations, and determination of the largest Lyapunov exponent via Jacobian-based methods such as discussed in Shimada and Nagashima (1979) 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 eqn (Schuster, 1988, p. 120)

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

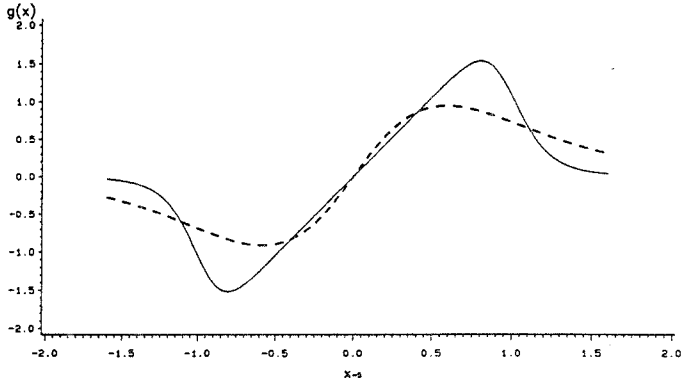
This map is of special interest in economics applications because it alone, of many that we tried, can generate a time series that is qualitatively like financial market data (Gallant, Hsieh, & Tauchen, 1991), especially in its ability to generate stretches of extremely volatile data of apparently random duration. Notice that the approximant is handicapped, as its dimension is higher than is necessary; it has five arguments when a lesser number would have sufficed. We view this as realistically mimicing 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

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($\hat{g}_K$)	$\|g^* - \hat{g}_K\|_{1,\pi,3}$	$\|g^* - \hat{g}_K\|_{1,2,3}$	Saturation Ratio
3	500	0.3482777075	3.6001114788	1.3252165780	17.9
5	1,000	0.0191675679	0.5522597668	0.1604392912	28.6
7	2,000	0.0177867857	0.4145203548	0.1141557050	40.8
9	4,000	0.0134447868	0.2586038122	0.0719887443	63.5
11	8,000	0.0012308988	0.1263063691	0.0196351730	103.9

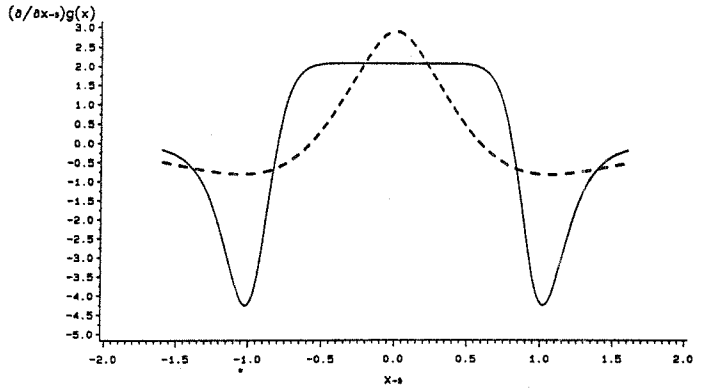
TABLE 2
Sensitivity of Neural Net Estimates

K	n	PredErr($\hat{g}_K$)	$\|g^* - \hat{g}_K\|_{1,\pi,3}$	$\|g^* - \hat{g}_K\|_{1,2,3}$	Saturation Ratio
7	500	0.0184102390	0.3745884175	0.1325439320	10.2
7	2,000	0.0177867857	0.4145203548	0.1141557050	40.8
11	500	0.0076063363	0.7141377059	0.1115357981	6.5
11	4,000	0.0015057013	0.0858882780	0.0210710677	51.9
11	8,000	0.0012308988	0.1263063691	0.0196351730	103.9
15	8,000	0.0020546210	0.1125778860	0.0336124596	76.2



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

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; $K = 3$, $n = 500$.

determining dimension suggest that there is a representation g^* of g in at most three dimensions $(x_{t-3}, x_{t-2}, x_{t-1})$.

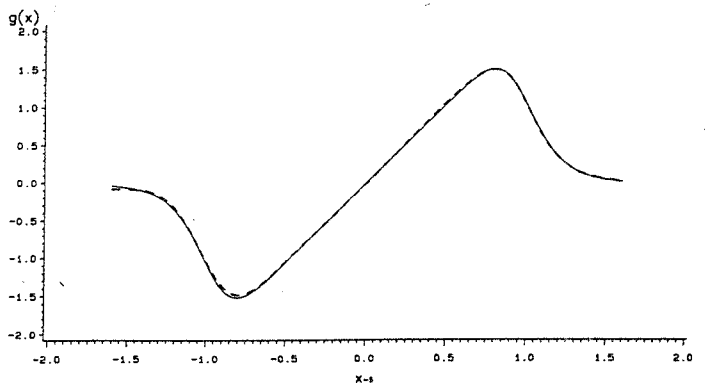
Note that in terms of the theory of the preceding section we have $e_t = 0$.

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

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

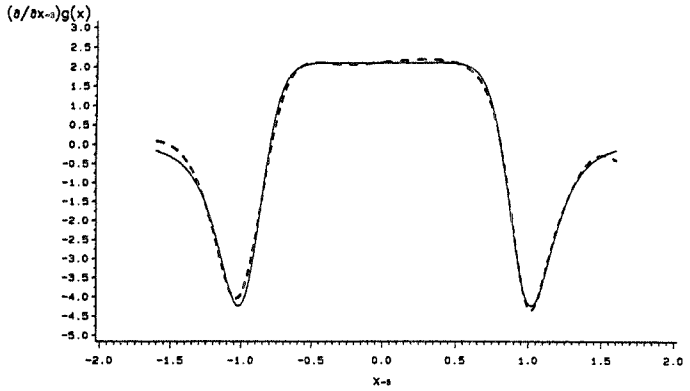
were determined using the Gauss–Newton nonlinear least squares algorithm (Gallant, 1987a, chap. 1). We found it helpful to zigzag by first holding $\hat{\beta}_j$ fixed and iterating on the $\hat{\gamma}_{ij}$, then holding the $\hat{\gamma}_{ij}$ fixed and iterating on the $\hat{\beta}_j$, and so on a few times, before going to the full Gauss–Newton iterates. Our rule relating K to n was of the form $K \propto \log(n)$ because asymptotic theory in a related context (Gallant, 1989) suggests that this is likely to 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.



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

FIGURE 3. Superimposed nonlinear map and neural net estimate; $K = 7$, $n = 2{,}000$.



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

FIGURE 4. Superimposed derivative and neural net estimate; $K = 7$, $n = 2{,}000$.

: estimate the prediction error from the holdout e using

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

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

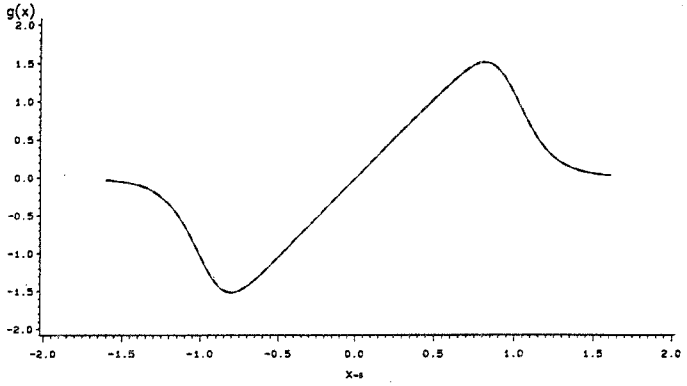
$$\bar{x} \cong \frac{1}{N}\sum_{t=N+1}^{2N} x_t.$$

rly, the Sobolev norm over Ξ (not over X) of the approximation error can be estimated from the hold-out sample using

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

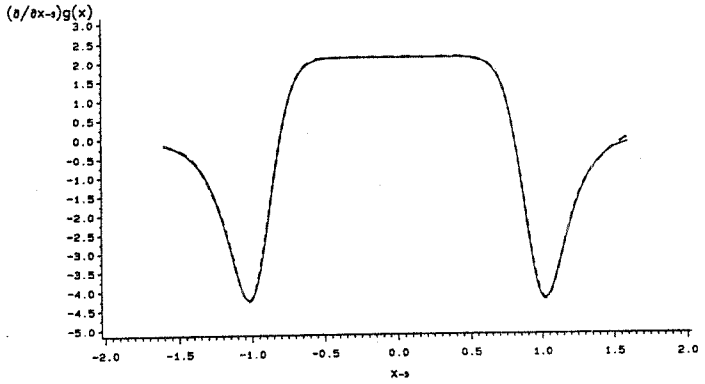
$$\|g^* - \hat{g}_K\|_{m,\infty,\Xi} \cong \max_{|\lambda|\le m}\max_{N+1\le s\le 2N} \times |D^\lambda g^*(x_{t-5}, x_{t-1}) - D^\lambda \hat{g}_K(x_{t-5}, \ldots, x_{t-1})|.$$

We took N as 10,000 in applying these formulas because we wanted very accurate estimates of



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

FIGURE 5. Superimposed nonlinear map and neural net estimate; $K = 11$, $n = 8{,}000$



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

FIGURE 6. Superimposed derivative and neural net estimate; $K = 11$, $n = 8{,}000$.

$\text{PredErr}(\hat{g}_K)$, $\|g^* - \hat{g}_K\|_{m,p,\Xi}$, and $\|g^* - \hat{g}_K\|_{m,\infty,\Xi}$. In ordinary applications, one would use a much smaller hold-out sample to estimate $\text{PredErr}(\hat{g}_K)$; $\|g^* - \hat{g}_K\|_{m,p,\Xi}$ and $\|g^* - \hat{g}_K\|_{m,\infty,\Xi}$ would not ordinarily be estimated since they cannot be determined without knowledge of g^*, and if g^* were known the inverse problem has no content. Also, note that

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

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

In graphical representations, Figures 1 through 6, of $g(x_{-5}, x_{-1})$ and $\hat{g}(x_{-5}, x_{-4}, \ldots, x_{-1})$ 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 correspond 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 appear 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.

REFERENCES

Adams, R. A. (1975). *Sobolev spaces*. New York: Academic Press.

Andrews, D. W. K. (1990). *Generic uniform convergence*. New Haven, CT: Yale University. (Cowles Foundation Discussion Paper No. 940.)

Carroll, S. M., & Dickinson, B. W. (1989). Construction of Neural Nets Using the Radon Transform. In *Proceedings of the International Joint Conference on Neural Networks, Washington, D.C.* (pp. I:607–I:611). New York: IEEE Press.

Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D*, **35**, 335–356.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2**, 303–314.

Eckmann, J.-P., Oliffson Kamphorst, S. Ruelle, D., & Ciliberto, S. (1986). Lyapunov exponents from time series. *Physical Review A*, **34**, 4971–4979.

Elbadawi, I., Gallant, A. R., & Souza, G. (1983). An elasticity can be estimated consistently without a priori knowledge of functional form. *Econometrica*, **51**, 1731–1752.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, **2**, 183–192.

Gallant, A. R. (1987a). *Nonlinear statistical models*. New York: John Wiley and Sons.

Gallant, A. R. (1987b). Identification and consistency in semi-nonparametric regression. In Truman F. Bewley (Ed.), *Advances in econometrics fifth world congress, Vol. I* (pp. 145–170). New York: Cambridge University Press. Translated as Gallant, A. R. (1985). Identification et Convergence en Regression Semi-Nonparametrique. *Annals de l'INSEE*, **59/60**, 239–267.

llant, A. R. (1989). *On the asymptotic normality of series estimators when the number of regressors increases and the minimum eigenvalue of* X'X/n *decreases* (Institute of Statistics Mimeograph Series No. 1955). Raleigh, NC: North Carolina State University.

llant, A. R., Hsieh, D. A., & Tauchen, G. E. (1991). On fitting a recalcitrant series: The pound/dollar exchange rate, 1974–83. In W. A. Barnett, J. Powell, & G. E. Tauchen (Eds.), *Nonparametric and semiparametric methods in econometrics and statistics* (pp. 199–240). Cambridge: Cambridge University Press.

llant, A. R., & White, H. (1988). There exists a neural network that does not make avoidable mistakes. *Proceedings of the second annual IEEE conference of neural networks, San Diego* (pp. I:657–I:664). New York: IEEE Press.

enander, U. (1981). *Abstract inference*. New York: John Wiley and Sons.

cht-Neilsen, R. (1989). Theory of the back-propagation neural network. In *Proceedings of the International Joint Conference on Neural Networks, Washington D.C.* (pp. I:593–I:606). New York: IEEE Press.

rnik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366.

rnik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, **3**, 551–560.

dan, M. (1989). Generic constraints on underspecified target trajectories. *Proceedings of the International Joint Conference on Neural Networks, Washington D.C.* (pp. I:217–I:225). New York: IEEE Press.

edes, A. & Farber, R. (1987). *Nonlinear signal processing using neural networks: Prediction and system modelling*. Los Alamos, NM: Theoretical Division, Los Alamos National Laboratory.

uster, H. G. (1988). *Deterministic chaos, an introduction, Second Revised Edition*. Weinheim, Federal Republic of Germany: VCH Verlagsgesellschaft mbH.

nada, I., & Nagashima, T. (1979). A numerical approach to ergodic problem of dissipative dynamical systems. *Progress of Theoretical Physics*, **61**, 1605–1616.

hcombe, M., & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In *Proceedings of the International Joint Conference on Neural Networks, Washington D.C.* (I:613–I:617). New York: IEEE Press.

e, M. (1974). Cross-validitory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**, 111–133.

e, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, **3**, 535–550.

MATHEMATICAL APPENDIX

Proof of Theorem 3.1. For the sake of clarity, we make explicit the dependence of $\hat{g}_K$, K_n and $s_n(g)$ on $\omega \in \Omega$ by writing $\hat{g}_K(\omega)$, $K_n(\omega)$ and $s_n(\omega, g)$. Let $F = \{\omega\colon K_n(\omega) \to \infty\} \cap \{\omega\colon \sup_{g\in\bar{\mathcal{G}}}|s_n(\omega, g) - \bar{s}(g, g^*)| \to 0\}$. As each of the intersected events has probability one, $P(F) = 1$.

Pick $\omega \in F$. Because σ is continuous with respect to $\|\cdot\|$ on the compact set $\mathcal{G}$, $\sigma(\mathcal{G})$ is compact. Therefore, $\sigma(\hat{g}_{K_{n'(\omega)}}(\omega)) \to \sigma(g^*)$ provided every subsequence $\{\sigma(\hat{g}_{K_{n''(\omega)}}(\omega))\}$ of $\{\sigma(\hat{g}_{K_{n'(\omega)}}(\omega))\}$ has accumulation point $\sigma(g^*)$. If this holds for every ω in F, the theorem is proven, because $P(F) = 1$.

To prove that every subsequence has accumulation point $\sigma(g^*)$ with given $\omega \in F$, let $\{n'\}$ index any subsequence of $\{n\}$ and pick a further subsequence $\{n''\}$ of $\{n'\}$ such that $\|\hat{g}_{K_{n''(\omega)}}(\omega) - g^0\| \to 0$ as $n'' \to \infty$ for $g^0 \in \bar{\mathcal{G}}$. Such a subsequence exists because $\{\hat{g}_{K_{n'(\omega)}}(\omega)\}$ is a sequence on the compact set $\bar{\mathcal{G}}$, and thus has an accumulation point g^0. By the triangle equality,

$$|s_{n''}(\omega, \hat{g}_{K_{n''(\omega)}}(\omega)) - \bar{s}(g^0, g^*)| \le |s_{n''}(\omega, \hat{g}_{K_{n''(\omega)}}(\omega)) - \bar{s}(\hat{g}_{K_{n''(\omega)}}(\omega), g^*)| + |\bar{s}(\hat{g}_{K_{n''(\omega)}}(\omega), g^*) - \bar{s}(g^0, g^*)|.$$

Given $\varepsilon > 0$, there exists $N_1(\omega, \varepsilon) < \infty$ such that for all $n'' > N_1(\omega, \varepsilon)$, $|s_{n''}(\omega, \hat{g}_{K_{n''(\omega)}}(\omega)) - \bar{s}(\hat{g}_{K_{n''(\omega)}}(\omega), g^*)| < \varepsilon/2$ by the uniform convergence condition (c). Also, there exists $N_2(\omega, \varepsilon) < \infty$ such that for all $n'' > N_2(\omega, \varepsilon)$, $|\bar{s}(\hat{g}_{K_{n''(\omega)}}(\omega), g^*) - \bar{s}(g^0, g^*)| < \varepsilon/2$ by continuity of $\bar{s}$. Consequently for all $n'' > \max(N_1(\omega, \varepsilon), N_2(\omega, \varepsilon))$

$$|s_{n''}(\omega, \hat{g}_{K_{n''(\omega)}}(\omega)) - \bar{s}(g^0, g^*)| < \varepsilon$$

and in particular

$$\bar{s}(g^0, g^*) < s_{n''}(\omega, \hat{g}_{K_{n''(\omega)}}(\omega)) + \varepsilon.$$

Next, because $\cup_K \mathcal{G}_K$ is dense in $\bar{\mathcal{G}}$ by condition (b), there exists a sequence $\{g^*_{K_{n''(\omega)}} \in \mathcal{G}_{K_{n''(\omega)}}\}$ such that $\|g^*_{K_{n''(\omega)}} - g^*\| \to 0$. Because $\hat{g}_{K_{n''(\omega)}}(\omega)$ minimizes $s_{n''}(\omega, \cdot)$ on $\mathcal{G}_{K_{n''(\omega)}}$

$$s_{n''}(\omega, \hat{g}_{K_{n''(\omega)}}(\omega)) \le s_{n''}(\omega, g^*_{K_{n''(\omega)}}).$$

Argument identical to that above with $g^*_{K_{n''(\omega)}}$ replacing $\hat{g}_{K_{n''(\omega)}}(\omega)$ and g^* replacing g^0 gives

$$|s_{n''}(\omega, g^*_{K_{n''(\omega)}}) - \bar{s}(g^*, g^*)| < \varepsilon,$$

and in particular

$$s_{n''}(\omega, g^*_{K_{n''(\omega)}}(\omega)) < \bar{s}(g^*, g^*) + \varepsilon,$$

for all n sufficiently large.

Collecting together all the above inequalities, we have

$$\bar{s}(g^0, g^*) < \bar{s}(g^*, g^*) + 2\varepsilon,$$

and because $\varepsilon > 0$ is arbitrary, it follows that

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

By the identification condition (d), we have that $\sigma(g^0) = \sigma(g^*)$.

Continuity of σ ensures that $\sigma(\hat{g}_{K_{n''(\omega)}}(\omega)) \to \sigma(g^0) = \sigma(g^*)$ as $n \to \infty$. Hence, $\{\sigma(\hat{g}_{K_{n'(\omega)}}(\omega))\}$ has accumulation point $\sigma(g^*)$, and the result follows because $\{n'\}$ and $\omega \in F$ are arbitrary. ■