Working Paper Series

Stationary and Integrated Autoregressive Neural Network Processes

Adrian Trapletti
Friedrich Leisch
Kurt Hornik

Working Paper No. 24
November 1998

Working Paper Series

November 1998

SFB
'Adaptive Information Systems and Modelling in Economics and Management Science'

Vienna University of Economics
and Business Administration
Augasse 2–6, 1090 Wien, Austria

in cooperation with
University of Vienna
Vienna University of Technology

http://www.wu-wien.ac.at/am

This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modelling in Economics and Management Science').

STATIONARY AND INTEGRATED AUTOREGRESSIVE NEURAL NETWORK PROCESSES

ADRIAN TRAPLETTI, FRIEDRICH LEISCH, AND KURT HORNIK

ABSTRACT. We consider autoregressive neural network (ARNN) processes driven by additive noise. Sufficient conditions on the network weights (parameters) are derived for the ergodicity and stationarity of the process. It is shown that essentially the linear part of the ARNN process determines whether the overall process is stationary. A generalization to the case of integrated ARNN processes is given. Least squares training (estimation) of the stationary models and testing for non-stationarity are discussed. The estimators are shown to be consistent and expressions on the limiting distributions are given.

1. INTRODUCTION

Over the past decade neural network (NN) models have been extensively used for a wide range of applications in the domain of time series analysis and forecasting. Although the NN time series application literature has grown in a spectacular fashion, relatively few theoretical advancements have been made.

The aim of this paper is to provide a rigorous analysis of the stochastic properties for the most popular class of NNs for time series analysis and forecasting, the class of *autoregressive neural network* (ARNN) processes driven by additive noise. The scalar *ARNN*(p) process can be seen as a natural generalization of the linear autoregressive AR(p) process and is defined by stochastic difference equations of the form

$$y_t = h(x_{t-1}, \theta) + \varepsilon_t. \tag{1}$$

$h(x_{t-1}, \theta)$ denotes a multi layer perceptron (MLP) with input vector $x_{t-1} = (y_{t-1}, \ldots, y_{t-p})'$ and weight (parameter) vector θ. ε_t is a sequence of independent and identically distributed (i.i.d.) random variables and represents the noise process. If $\mathsf{E}\varepsilon_t = 0$, then $h(x_{t-1}, \theta)$ equals the conditional expectation $\mathsf{E}[y_t|x_{t-1}]$.

The output of the MLP with q hidden units, shortcut connections ψ, and the activation function $G(\cdot)$ is given by

$$h(x_{t-1}, \theta) = \nu + \psi' x_{t-1} + \sum_{i=1}^{q} \beta_i G(\phi_i' x_{t-1} + \mu_i). \tag{2}$$

The weight vectors are ψ $(p \times 1)$, $\beta \equiv (\beta_1, \ldots, \beta_q)'$ $(q \times 1)$, $\phi \equiv (\phi_1', \ldots, \phi_q')'$ $(pq \times 1)$, $\mu \equiv (\mu_1, \ldots, \mu_q)'$ $(q \times 1)$, and the intercept ν (scalar) collected together in the $r \times 1$ network weight vector $\theta = (\psi', \beta', \phi', \mu', \nu')'$ with $r = 1 + p + q(2 + p)$. The activation function $G(\cdot)$ is assumed to be bounded, e.g., the logistic function $G(x) = (1 + \exp(-x))^{-1}$. However, for most of the results in this article the exact structure of the MLP is not important, provided that the non-linear part of the network is bounded.

It is well known that MLPs can provide arbitrarily accurate approximations to arbitrary functions in a variety of normed function spaces if the dimension of the weight space is sufficiently large (e.g., Hornik et al., 1989). However, we focus

Date: November 30, 1998.

here on bounded MLP complexity, i.e., the network weight space Θ is defined as a subset of the finite-dimensional space $\mathbb{R}^r$. Nevertheless, this treatment of MLP as parametric models still allows to approximate an arbitrary function to some degree. For example, White (1989) approximated the Hénon map with five hidden units.

Of particular importance for statistics in time series analysis is the question of stationarity and ergodicity of a process. Since for stationary ergodic processes a single trajectory displays the whole probability law of the process, averaging over time gives the same result as averaging over population. Thus, the sample moments converge to their population mean. Whereas for linear AR processes conditions for stationarity and ergodicity are well known (e.g., Brockwell and Davis, 1991), this question has not gained much interest in the neural networks literature. There are — up to our knowledge — no results giving conditions for the stationarity of ARNN processes. There are results for Hopfield nets (Wang and Sheng, 1996), but these nets cannot be used to model the conditional expectation for time series prediction.

In Section 2 we obtain sufficient conditions on the network weights for the ARNN process to be stationary, (geometrically) ergodic, and thus strong mixing. It is shown that essentially the shortcut connections of the ARNN process determine whether the overall process is stationary. Moreover, conditions for the existence of moments and expressions for the shape of the stationary distribution are provided. These conditions are established using results for Markov chains and non-linear time series analysis.

Non-stationary time series play an important role in a variety of application fields such as economics and physical sciences. In Section 3 we focus on non-stationary ARNN processes which exhibit either explosive or random walk behaviour. We introduce integrated ARNN (ARINN) processes which are a particularly interesting class of non-stationary ARNN processes, since they can be made stationary by simple differencing. It is shown that ARINN processes "converge" to a Wiener process when appropriately standardized.

In Section 4 the results are applied to the problem of training (estimation) and testing in the context of ARNN processes. In particular, least squares estimators of the weights of the stationary ARNN model are shown (under mild regularity conditions) to be consistent and asymptotically normal. We introduce the hypothesis test for a unit root of Perron (1988); Phillips (1987); Phillips and Perron (1988) (PP) as a tool to differentiate between stationary ARNN and ARINN models. This test is shown to be applicable for ARNN processes. The limiting distribution follows from functional central limit theory.

All proofs are deferred to the appendix.

2. Ergodicity and Stationarity

2.1. **Some Markov Chain Theory.** To start, we write the ARNN(p) process as a Markov chain on $\mathbb{R}^p$ and outline some relevant Markov chain theory; see Chan and Tong (1985); Feigin and Tweedie (1985); Meyn and Tweedie (1993); Tong (1996); White and Domowitz (1984) for details.

Defining $\eta_t = (\varepsilon_t, 0, \dots, 0)'$, equation (1) becomes

$$x_t = H(x_{t-1}) + \eta_t, \tag{3}$$

where $H(x_{t-1}) = (h(x_{t-1}, \theta), y_{t-1}, \dots, y_{t-p+1})'$. Let

$$P(x, \mathcal{A}) = \mathsf{P}(x_t \in \mathcal{A} \mid x_{t-1} = x)$$

denote the probability that $\{x_t\}$ moves from x to the set $\mathcal{A} \in \mathcal{B}$ in one step. Then $\{x_t\}$ with $P(x, \mathcal{A})$ forms a Markov chain with state space $(\mathbb{R}^p, \mathcal{B}, \lambda)$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}^p$ and λ denotes the Lebesgue measure on $(\mathbb{R}^p, \mathcal{B})$ (cf. Chan and Tong, 1985, pp. 666, 667).

We are basically interested in conditions on the weights for which $\{x_t\}$ is a strictly stationary process. In Markov chain theory this problem is closely related to the asymptotic behaviour of $\{x_t\}$ as t becomes large. The Markov chain $\{x_t\}$ is *geometrically ergodic* if there exists a probability measure π on $(\mathbb{R}^p, \mathcal{B})$ and a constant $\varrho > 1$ such that

$$\lim_{n\to\infty} \varrho^n \|P^n(x,\cdot) - \pi(\cdot)\| = 0 \tag{4}$$

for each $x \in \mathbb{R}^p$ and $\|\cdot\|$ denotes the total variation norm. If (4) holds for $\varrho = 1$, then $\{x_t\}$ is called *ergodic*. It is easy to see that π in (4) satisfies the *invariance equation*

$$\pi(\mathcal{A}) = \int P(x,\mathcal{A})\pi(dx), \quad \forall \mathcal{A} \in \mathcal{B}. \tag{5}$$

Thus, π is called the *stationary measure*. Suppose that $\{x_t\}$ is ergodic. Then (4) implies that the distribution of $\{x_t\}$ converges to π. Hence, $\{x_t\}$ is asymptotically stationary. If $\{x_t\}$ is started either with initial distribution π, i.e., $x_0 \sim \pi$, or in the infinite past, then $\{x_t\}$ is strictly stationary. This solution will be called the *stationary* or *steady state solution* of (3).

To establish the ergodicity of ARNN processes, we further need the concepts of irreducibility and aperiodicity. The Markov chain $\{x_t\}$ is called *irreducible* if

$$\sum_{n=1}^{\infty} P^n(x,\mathcal{A}) > 0, \quad \forall x \in \mathbb{R}^p,$$

whenever $\lambda(\mathcal{A}) > 0$. This basically means that all parts of the state space can be reached by the Markov chain irrespective of the starting point.

A irreducible Markov chain is *aperiodic* if there exists an $\mathcal{A} \in \mathcal{B}$ with $\lambda(\mathcal{A}) > 0$ and for all $\mathcal{C} \in \mathcal{B}$, $\mathcal{C} \subseteq \mathcal{A}$ with $\lambda(\mathcal{C}) > 0$, there exists a positive integer n such that

$$P^n(x,\mathcal{C}) > 0 \ \text{ and } \ P^{n+1}(x,\mathcal{C}) > 0, \quad x \in \mathbb{R}^p,$$

Tong (1996), Proposition A1.2. Hence, for an aperiodic chain it is impossible that the chain returns to given sets only at specific time points.

2.2. **Geometrically Ergodic and Stationary ARNN Processes.** For most general time series models irreducibility and aperiodicity cannot be assumed automatically. However, for an ARNN process these conditions can be checked easily. Essentially, it is enough if the distribution of the noise process exhibits an absolutely continuous component with respect to Lebesgue measure and if the support of the probability density function (p.d.f.) is sufficiently large. In this case every non-null p-dimensional hypercube is reached in p (and also in $p+1$) steps with positive probability (and hence every non-null Borel set $\mathcal{A}$).

Lemma 1. *Let $\{x_t\}$ be the Markov chain of the ARNN process with continuous activation function $G(\cdot)$. Let ε_t be i.i.d. with absolutely continuous distribution with respect to Lebesgue measure λ. If the p.d.f. $f(\cdot)$ of ε_t is positive everywhere in $\mathbb{R}$ and lower semi-continuous, then $\{x_t\}$ is a irreducible and aperiodic Markov chain.*

Obviously, e.g., Gaussian white noise fulfils the conditions of Lemma 1. We can now state conditions for which the stochastic difference equations (1) and (2), or equivalently (3), have stationary solutions.

Theorem 1. *Suppose the Markov chain $\{x_t\}$ of the ARNN process satisfies the conditions of Lemma 1, the activation function $G(\cdot)$ is bounded, and $\mathsf{E}|\varepsilon_t| < \infty$.*

Let

$$\psi(z) := 1 - \sum_{i=1}^{p} \psi_i z^i, \quad z \in \mathbb{C} \tag{6}$$

denote the characteristic polynomial associated with the shortcut connections. Then a sufficient condition for the geometric ergodicity of the Markov chain $\{x_t\}$ *is that*

$$\psi(z) \neq 0 \quad \forall z,\ |z| \leq 1. \tag{7}$$

Furthermore, if (7) *holds, then the associated ARNN process* $\{y_t\}$ *is asymptotically stationary.*

This result is not surprising since boundedness is a strong form of stability. It shows that essentially the linear part of the ARNN process determines whether the overall process is stationary. Thus, the usual conditions for AR processes can be used. Moreover, an ARNN without shortcut connections leads always to a stationary solution.

The geometric rate of convergence in Theorem 1 implies that the memory of the ARNN process vanishes exponentially fast. A convenient way of describing the memory of a process is through the concept of mixing processes. Let $\{y_t\}$ be a stochastic process on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$, and let $\mathcal{F}_a^b = \sigma(y_t;\ a \leq t \leq b)$ be the σ-algebra generated by the random variables $y_a, y_{a+1}, \ldots, y_b$. Define

$$\alpha(m) := \sup|\mathsf{P}(\mathcal{E} \cap \mathcal{F}) - \mathsf{P}(\mathcal{E})\mathsf{P}(\mathcal{F})|, \tag{8}$$

where the supremum is taken over all $\mathcal{E} \in \mathcal{F}_{-\infty}^n$, $\mathcal{F} \in \mathcal{F}_{n+m}^{\infty}$, and n. This quantity measures the dependence between events separated by at least m time periods. We call the process $\{y_t\}$ *strong* or α*-mixing* if $\alpha(m) \to 0$ as $m \to \infty$.

The following result establishes strong mixing for geometrically ergodic ARNN processes.

Corollary 1. *Let* $\{y_t\}$ *be a stationary ARNN process satisfying the conditions of Theorem 1. Then* $\{y_t\}$ *is strong mixing with mixing coefficients*

$$\alpha(m) \leq \kappa \varrho^m,$$

for some $\kappa < \infty$ *and* $\varrho \in (0, 1)$.

This result is, as we will see later, particularly important for the asymptotic theory of inference and testing, since mixing processes are sufficiently well behaved to allow laws of large numbers and central limit theorems to be established.

If the memory of the stationary ARNN process vanishes exponentially fast, then the autocovariance function approximates zero with an exponential rate.

Corollary 2. *Let* $\{y_t\}$ *as in Corollary 1. If* $\mathsf{E}|y_t|^4 < \infty$, *then*

$$\gamma(\tau) := \mathsf{Cov}(y_t, y_{t+\tau}) \leq \tilde{\kappa} \tilde{\varrho}^{\tau},$$

for some $\tilde{\kappa} < \infty$ *and* $\tilde{\varrho} \in (0, 1)$.

Sometimes the "frequency domain" analysis of time series provides an illuminating alternative way of viewing a stationary process. An important tool is the spectral density as the Fourier transform of the covariance function of the process. Although not every stationary process has a spectral density, an absolutely summable autocovariance function as in Corollary 2 implies the existence of a spectral density.

Corollary 3. *Every ARNN process* $\{y_t\}$ *which satisfies the conditions of Corollary 2 has a spectral density*

$$f(\lambda) := \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} e^{-i\tau\lambda} \gamma(\tau) \geq 0, \quad \forall \lambda \in [-\pi, \pi],$$

and the autocovariance function has the spectral representation

$$\gamma(\tau) = \int_{-\pi}^{\pi} e^{-i\lambda\tau} f(\lambda) d\lambda.$$

2.3. **Ergodic and Stationary ARNN Processes.** Condition (7) is sufficient but not necessary for ergodicity. This will be shown in the following.

Suppose a root of the characteristic polynomial (6) (characteristic root) of an ARNN process lies on the unit circle and all other roots are outside the unit circle. Based on linear time series theory random walk behaviour with or without drift would be expected, depending upon the non-linear part and the intercept of the ARNN process. However, in contrast to a purely linear process, for an ARNN process it is possible that the drift generated by the non-linear part is always towards the "centre" of the state space. We consider the following special case of an ARNN(1) process as an example

$$y_t = y_{t-1} + G(y_{t-1}) + \varepsilon_t, \tag{9}$$

where $G(\cdot)$ is a bounded continuous activation function such that

$$\lim_{y\to\infty} G(y) = -a \quad \text{and} \quad \lim_{y\to-\infty} G(y) = b \quad \text{for some } a, b > 0.$$

Proposition 1. *Let ε_t satisfy the conditions of Lemma 1 and $\mathsf{E}\varepsilon_t = 0$. If $\mathsf{E}|\varepsilon_t|^2 < \infty$, then the solution $\{y_t\}$ of (9) is ergodic, asymptotically stationary, and strong mixing.*

Since the drift toward the stationary solution is at most linear, we can state ergodicity but not geometric ergodicity. Nevertheless, the ARNN(1) process (9) remains strong mixing. However, for simplicity we use in the remainder of this article the term stationary ARNN process for strictly stationary or state state solutions of the equations (1) and (2) for which condition (7) holds.

2.4. **The Stationary Distribution π.** We now turn to the question of the existence of moments of $\{y_t\}$ in its stationary regime. Obviously, the bounded part of the ARNN process has no influence on the existence of moments. We get the following simple result.

Theorem 2. *Let $\{y_t\}$ be a stationary ARNN process as in Theorem 1. Then*

$$\mathsf{E}|y_t|^k = \int |x|^k \pi(dx) < \infty \iff \mathsf{E}|\varepsilon_t|^k < \infty.$$

Hence, e.g., for Gaussian white noise the ARNN process is k-integrable for all $k < \infty$.

While we have used the concept of stationarity in the strict sense, it is often of interest to consider second order stationarity.

Corollary 4. *Let $\{y_t\}$ as in Theorem 1. Then $\{y_t\}$ is weakly stationary if and only if $\mathsf{E}|\varepsilon_t|^2 < \infty$.*

Unfortunately it is not possible to describe the stationary distribution π in general. Although an implicit solution of π is always defined by (5), for most cases this equation cannot be solved in closed form. There exist different numerical techniques for solving this problem; see Tong (1996) for an overview.

However, a partial result about the shape of π is obtained in the following. Since this result depends on the particular structure of the ARNN, we state it only for the logistic activation function $G(x) = (1 + \exp(-x))^{-1}$. First we impose a minimality condition on the non-linear part of the ARNN process as in Hwang and Ding (1997); Sussmann (1992).

Assumption 1. Each hidden unit makes a non-trivial contribution to the overall ARNN process, i.e., $\beta_i \neq 0$ and $\phi_i \neq 0$ for all $i = 1, \ldots, q$. No two of the functions $g_i(x) = \phi_i' x + \mu_i$ are sign equivalent, i.e., $(\phi_i', \mu_i') \neq \pm(\phi_j', \mu_j')$ for all $i \neq j$ (two functions $g_1(x)$ and $g_2(x)$ are called *sign equivalent* iff $|g_1(x)| = |g_2(x)|$ for all $x \in \mathbb{R}$).

For the logistic activation function, assumption 1 ensures that it is not possible to state equation (2) with fewer hidden units, i.e., with a new q, say $\tilde{q} < q$. See Hwang and Ding (1997) concerning the logistic function and Sussmann (1992) concerning the essentially equivalent $\tanh(\cdot)$ squasher.

Proposition 2. *Suppose Assumption 1 holds and $\{y_t\}$ is a stationary ARNN process with activation function $G(x) = (1 + \exp(-x))^{-1}$. Let ε_t have a symmetric p.d.f. $f(\cdot)$ about the origin. Then all finite stationary joint p.d.f.s of $y_{t_1}, y_{t_2}, \ldots, y_{t_k}$ are symmetric about the origin for all $t_1, t_2, \ldots, t_k$ and for $k \geq 1$ if and only if $\mu_i = 0$ for all $i = 1, \ldots, q$ and $2\nu + \sum_{i=1}^{q} \beta_i = 0$.*

Since a symmetric joint p.d.f. implies a symmetric marginal p.d.f. for y_t, the expectation of y_t is zero, i.e., $\mathsf{E} y_t = 0$.

Corollary 5. *Let the assumptions of Proposition 2 hold and k be an odd integer. Then $\mathsf{E} y_t^k = 0$.*

3. Non-Stationary and Integrated Processes

3.1. **Transience.** Using linear time series theory it seems obvious that the ARNN process exhibits explosive behaviour if at least one characteristic root lies inside the unit circle. In Markov chain theory the concept of transience is used to formalize explosive behaviour of a chain.

The state ω of a Markov chain on a countable space is called *transient* if the expected number of visits to ω is finite. The Markov chain is called *transient* if every state is transient (for a definition on a general state space see, e.g., Meyn and Tweedie, 1993, Chapter 8). Transience of a Markov chain $\{x_t\}$ on a general state space implies that $\mathsf{P}_x\{x_t \to \infty\} = 1$, for each $x \in \mathbb{R}^p$. This means that the trajectory of $\{x_t\}$ visits each compact set only finitely often for each initial state x.

We now consider transience for the ARNN process.

Theorem 3. *Let $\{x_t\}$ be the irreducible and aperiodic Markov chain of an ARNN process with $\psi(\cdot)$ as in Theorem 1. Then $\{x_t\}$ is transient if $\mathsf{E}[\exp(|\varepsilon_t|)] < \infty$, $\psi(\cdot)$ has distinct roots, and $\psi(z) = 0$ for at least one z with $|z| < 1$.*

We believe that the distinctness assumption on the roots is not necessary and that the moment condition could be relaxed.

3.2. **ARINN Processes.** In linear time series modeling it is a rather common practice to detrend a time series by taking differences if the series exhibits non-stationary features. An example is the class of autoregressive integrated moving average (ARIMA) processes (e.g., Brockwell and Davis, 1991), where the difference operator must be applied several times before a stationary representation is appropriate. This approach has proven to be useful in many applications.

Let d be a non-negative integer. As in linear time series analysis we define the dth order difference operator

$$\Delta^d z_t := \Delta(\Delta^{d-1} z_t), \quad d \geq 2,$$
$$\Delta^1 z_t \equiv \Delta z_t := z_t - z_{t-1}.$$

Then z_t is said to be an *ARINN*(p, d) process if $y_t = \Delta^d z_t$ is a stationary ARNN(p) process.

Hence, an ARINN process reduces to a stationary ARNN process after differencing finitely many times. Moreover, an ARINN(p,d) process is stationary if and only if $d=0$, in which case it reduces to a stationary ARNN(p) process. A stochastic process $\{z_t\}$ is called *integrated of order* d, briefly $z_t \sim I(d)$ if $\Delta^d z_t$ is stationary and $\Delta^{d-1} z_t$ is not stationary. Thus, an ARINN(p,d) process is $I(d)$.

We now analyze the long term behaviour of an ARINN$(p,1)$ process $\{z_t\}_{t=0}^{\infty}$. Following Phillips (1987) we write the ARINN$(p,1)$ process as partial sum

$$z_t = \sum_{i=1}^{t} y_t, \quad t = 1, 2, \ldots$$

where $\{y_t\}$ is a stationary ARNN(p) process with $\mathsf{E}y_t = 0$. Without loss of generality we assume $z_0 = 0$. From the sequence $\{z_t\}$ we construct the random variables

$$Z_t(r) = \frac{1}{\sqrt{t}\sigma}\, z_{\lfloor tr \rfloor}, \quad (j-1)/t \le r \le j/t \quad (j = 1, \ldots, t),$$

where $\lfloor \cdot \rfloor$ denotes the integer part of its argument. We further assume

$$\sigma^2 = \lim_{t\to\infty} \mathsf{E}[t^{-1} z_t^2] = \mathsf{E}y_1^2 + 2\sum_{k=2}^{\infty} \mathsf{E}[y_1 y_k]$$

being positive.

Theorem 4. *If* $\mathsf{E}|y_t|^{\delta} < \infty$ *for some* $\delta > 2$, *then*

$$Z_t(r) \Rightarrow W(r), \quad t \to \infty,$$

where $W(r)$ *is the Wiener process on the unit interval and* $\Rightarrow$ *denotes weak convergence.*

This result is often referred to as a functional central limit theorem. Here, it basically states that a standardized zero-mean ARINN$(p,1)$ process "converges" to a Wiener process.

4. Training and Testing

4.1. **Consistency.** The general results from the previous sections are applied to training of ARNN models in the context of ARNN processes for fixed model orders p and q; see equations (1) and (2). We follow White and Domowitz (1984), although other results from the literature could be used. The following assumption specifies the data generation process (DGP).

Assumption 2. The DGP for the sequence of scalar real valued observations $\{y_t\}_{t=1}^{n}$ is a stationary ARNN(p) process with continuous activation function $G(\cdot)$ and $r \times 1$ network weight vector $\theta_0 \in \Theta$. The network weight space Θ is a compact subset of $\mathbb{R}^r$ for some $r \in \mathbb{N}$.

Clearly, the stationarity condition is a condition on the weight space Θ such that for each $\theta_0 \in \Theta$ the roots of the characteristic polynomial $\psi(\cdot)$ lie outside the unit circle.

The goodness of fit of an ARNN model as a function of θ for a given time series $\{y_t\}_{t=1}^{n}$ can be measured by $Q_n(\theta) = n^{-1}\sum_{t=1}^{n}(y_t - h(x_{t-1},\theta))^2$. However, it is of interest to consider the overall model performance, i.e., the performance for future observations $\bar{Q}_n(\theta) = \mathsf{E}Q_n(\theta)$, where the expectation is with respect to the stationary distribution π. Unfortunately, choosing θ to solve $\inf_{\theta\in\Theta} \bar{Q}_n(\theta)$ is not possible since the stationary distribution π is unknown. Nevertheless, an approximate solution $\hat{\theta}_n$ can be found solving the problem

$$Q_n(\hat{\theta}_n) = \inf_{\theta\in\Theta} Q_n(\theta). \tag{10}$$

This so-called *non-linear least squares estimator* $\hat{\theta}_n$ provides a good approximation since for large n, $Q_n(\theta)$ approximates well $\bar{Q}_n(\theta)$.

A fundamental problem for statistical inference with MLPs is the unidentifiability of the network weights (e.g., Hwang and Ding, 1997; White, 1996). The next assumptions deal with this problem. First, we rule out equivalent minima of equation (10) achieved by permuting the hidden units and changing the sign of the network weights on each side of a hidden unit (cf. Kůrková and Kainen, 1994, Proposition 3.9).

Assumption 3. The network weights from the hidden layer to the output unit are strictly positive, i.e., $\beta_i > 0$ for all $i = 1, \ldots, q$, and these weights are ordered in the sense that $\beta_i < \beta_{i+1}$ for all $i = 1, \ldots, q-1$.

Although this Assumption only specifies the interior of a minimal search set, this is not a significant restriction since every weight vector belonging to the minimal search set can be approximated arbitrarily well. For certain activation functions $G(\cdot)$ Assumption 1 and 3 ensure that equation (10) has a unique global minimum. These activation functions are specified in the following Assumption (e.g., Hwang and Ding, 1997, Condition A).

Assumption 4. If no two of the functions $g_i(x)$ are sign equivalent (see Assumption 1), then the functions $G(g_1(x)), \ldots, G(g_k(x))$, and the constant function 1 are linear independent for any positive integer k and any $x \in \mathbb{R}^p$.

For a discussion of identifiability concepts see, e.g., Hwang and Ding (1997); Kůrková and Kainen (1994); Sussmann (1992).

We can now state the strong consistency of the non-linear least squares estimator $\hat{\theta}_n$.

Theorem 5. *Suppose Assumptions 1-4 hold. If* $\mathsf{E}|\varepsilon_t^2|^{\delta_1} < \infty$ *for some* $\delta_1 > 1$, *then*

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_0, \quad n \to \infty,$$

where $\xrightarrow{a.s.}$ *denotes convergence almost sure (a.s.).*

4.2. **Asymptotic Normality.** We have to add the following two conditions to state the asymptotic normality of $\hat{\theta}_n$.

Assumption 5. The network weight vector θ_0 is interior to Θ and $G(\cdot)$ is continuously differentiable of order 2.

Assumption 6. If no two of the functions $g_i(x)$ are sign equivalent, then the functions $G(g_1(x)), \ldots, G(g_k(x)), G'(g_1(x)), \ldots$
$G'(g_k(x)), xG'(g_1(x)), \ldots, xG'(g_k(x))$, and the constant function 1 are linear independent for any positive integer k and any $x \in \mathbb{R}^p$.

This Assumption on the activation function $G(\cdot)$ is clearly stronger than Assumption 4 and essentially ensures together with Assumption 1 that the information matrix is regular at θ_0. Fortunately, two popular choices of the activation function, the logistic function and the $\tanh(\cdot)$ squasher, satisfy Assumption 6 (see Hwang and Ding, 1997, Condition B and Lemma 2.7).

It follows the asymptotic normality.

Theorem 6. *Suppose Assumptions 1-3, 5, and 6 hold. If* $\mathsf{E}|\varepsilon_t^3|^{\delta_2} < \infty$ *for some* $\delta_2 > 1$, *then*

$$(\frac{1}{2\sigma^2}\nabla^2\bar{Q}_n(\theta_0))^{1/2}\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathbb{1}_p), \quad n \to \infty,$$

where $\nabla^2\bar{Q}_n(\theta_0)$ *is the Hessian matrix of* $\bar{Q}_n(\theta)$ *evaluated at* θ_0, σ^2 *is the variance of* ε_t, *and* $\xrightarrow{d}$ *denotes convergence in distribution.*

This result forms the basis for hypotheses tests, e.g., testing whether certain network weights are zero. A practical difficulty is that the Hessian $\nabla^2 \bar{Q}_n(\theta_0)$ and σ^2 are unknown. They can, however, be consistently estimated by $\widehat{\nabla^2 \bar{Q}_n(\theta_0)} = \nabla^2 Q_n(\hat{\theta}_n)$ and $\widehat{\sigma^2} = n^{-1} \sum_{t=1}^{n} \hat{\varepsilon}_t^2$, where $\{\hat{\varepsilon}_t\}$ is the sequence of the observed residuals; see e.g., White and Domowitz (1984); Gallant and White (1988).

4.3. **Unit Root Test.** The methods presented in subsections 4.1 and 4.2 are for stationary time series. However, many time series encountered in practice are non-stationary. A popular way to solve the problem of training in a non-stationary environment is to look for a transformation of the data which generates a new stationary series. This can frequently be achieved by simple differencing. Once the data has been suitably transformed, the results from subsections 4.1 and 4.2 can be used.

We introduce the PP unit root test as a formal test concerning the need for differencing in ARNN modeling. To establish the corresponding asymptotic results we follow Perron (1988); Phillips (1987); Phillips and Perron (1988), where the reader can find a detailed discussion and several test statistics, e.g., for a non-zero drift, not presented here.

Let us write a zero-mean ARINN$(p,1)$ process $\{z_t\}_{t=1}^{\infty}$ as

$$z_t = \alpha z_{t-1} + y_t, \quad t = 1, 2, \ldots \tag{11}$$

$$\alpha = 1, \tag{12}$$

where by definition $\{y_t\}$ is a stationary zero-mean ARNN(p) process. Further suppose z_0 is equal to zero.

Now, a formal test of the null hypothesis that $\{z_t\}$ is an ARINN$(p,1)$ process against the alternatives of $\{z_t\}$ being a stationary ARNN$(p+1)$ or a transient ARNN$(p+1)$ process can be expressed as hypothesis on the weight α

$$\begin{cases} H_0: & \alpha = 1, \\ H_1: & \alpha \neq 1. \end{cases}$$

If the null hypothesis is accepted, the difference operator has to be applied before a stationary ARNN(p) model is appropriate.

Given a series of $n+1$ observations $\{z_t\}_{t=0}^{n}$, the basic idea for the PP test is to estimate α by the least squares estimator

$$\hat{\alpha} = \frac{\sum_{t=1}^{n} z_t z_{t-1}}{\sum_{t=1}^{n} z_{t-1}^2}$$

from regressing z_t on z_{t-1} and to consider the conventional t-statistic of $\hat{\alpha}$. Unfortunately, the limiting distribution of this statistic is non-normal and depends upon nuisance parameters. However, the nuisance parameters may be consistently estimated and a transformation of the test statistic exists, which eliminates the nuisance dependence asymptotically. The transformed test statistic is defined as

$$Z_\alpha = n(\hat{\alpha} - 1) - (1/2)(s_{nl}^2 - s_y^2) \left(n^{-2} \sum_{t=1}^{n} z_{t-1}^2 \right)^{-1}, \tag{13}$$

where $s_{nl}^2 = n^{-1} \sum_{t=1}^{n} (z_t - z_{t-1})^2 + 2n^{-1} \sum_{\tau=1}^{l} \sum_{t=\tau+1}^{n} (z_t - z_{t-1})(z_{t-\tau} - z_{t-\tau-1})$ and $s_y^2 = n^{-1} \sum_{t=1}^{n} (z_t - z_{t-1})^2$. The limiting distribution of (13) is given in the following Theorem.

Theorem 7. *Suppose $\{z_t\}$ is generated by (11) where $\{y_t\}$ follows a stationary ARNN process. Let $\mathsf{E}|\varepsilon_t|^{2\delta} < \infty$ for some $\delta > 2$. If $l \to \infty$ as $n \to \infty$ such that*

$l = o(n^{1/4})$, then

$$Z_\alpha \Rightarrow \frac{(W(1)^2 - 1)/2}{\int_0^1 W(r)^2 dr}, \tag{14}$$

under H_0, i.e., if $\{z_t\}$ comes from an ARINN$(p, 1)$ process.

The distribution in (14) may be found tabulated in econometric textbooks such as Banerjee et al. (1993).

Furthermore, $\hat{\alpha}$ is a consistent estimator for the unit root.

Proposition 3. *Let $\{z_t\}$ be generated by* (11) *and* (12). *If $\mathsf{E}|\varepsilon_t|^\delta < \infty$ for some $\delta > 2$, then $\hat{\alpha}$ is a (weakly) consistent estimator*

$$\hat{\alpha} \xrightarrow{p} 1, \quad n \to \infty,$$

where $\xrightarrow{p}$ denotes convergence in probability.

Hence, for practical applications Theorem 7 might provide the basis to decide whether a given time series has to be differenced: if the empirical autocovariance (or equivalently autocorrelation) function of the considered time series vanishes not sufficiently fast as suggested in Corollary 2, we can use the formal test of Theorem 7. In Leisch et al. (1999) the implications of modeling a random walk without taking differences are discussed.

5. Conclusions

In this paper we have studied several classical concepts of linear time series analysis in the context of ARNN processes driven by additive noise.

We have derived conditions on the network weights for the ergodicity and stationarity of ARNN processes. Our results show that an ARNN process having all its characteristic roots outside the unit circle is geometrically ergodic and asymptotically stationary. Furthermore, if the process is started either in its stationary regime or in the infinite past, then the process is strictly stationary. If at least one characteristic root lies inside the unit circle and a technical assumption is satisfied, then the process is proved to be explosive. Concerning ARNN processes with characteristic roots lying on and outside the unit circle the long term behaviour is determined by the "state-dependent intercept", i.e., the non-linear part and the intercept of the process: driftless processes exhibit random walk behaviour; a drift towards $+\infty$ or $-\infty$ results in a transient process; a state dependent drift towards the "centre" of the state space gives an ergodic and asymptotically stationary solution.

We have obtained results on the memory of the geometrically ergodic ARNN process. This article shows that the geometric ergodic ARNN process is strong mixing with a geometric mixing rate and its autocovariance (autocorrelation) function approaches zero exponentially fast. Hence, all geometrically ergodic ARNN process have a spectral density.

Concerning the stationary regime of an ARNN process, we have stated conditions for the existence of moments. In particular, the existence of the second moment of the noise process is a necessary and sufficient condition for the weak stationarity of the ARNN process. Moreover, we have presented conditions for which the shape of the stationary distribution is symmetric about the origin.

Since many time series in practice are non-stationary we have introduced the class of ARINN processes. These processes "converge" to a Wiener process when appropriately standardized.

Finally, we have discussed training and testing of ARNN models in the context of ARNN processes. We have shown the consistency and asymptotic normality

(under mild regularity conditions) of the non-linear least squares estimator for the stationary ARNN models. These results form the basis for a number of statistical tests.

A popular way to find an appropriate model for non-stationary time series is to transform the data to a stationary series by simple differencing. Once the data has been transformed, the results for stationary models can be applied. We suggest to use the unit root test of Perron (1988); Phillips (1987); Phillips and Perron (1988) concerning the need for differencing in ARNN modeling. Our results show that this test can be used to differentiate between stationary ARNN and ARINN processes.

The research reported in this article is currently being extended in various ways. Since all the statistical results are asymptotic, we have to be careful when applying them to actual data. Thus, it is of interest to study the small sample performance of the suggested estimators and tests by simulation.

Another interesting topic currently under investigation is to extend the results of this paper to the multivariate case. Most of the concepts could be transformed straightforward and various new concepts such as non-linear cointegration in a system of non-stationary variables could be formulated (Granger, 1995; Granger and Hallmann, 1991; Granger and Swanson, 1996).

6. Appendix: Mathematical Proofs

Proof of Lemma 1. This follows immediately from Chan and Tong (1985), pp. 668, 669, and the definition of an ARNN(p) process or Markov chain, respectively. In the following "ARNN" is used for ARNN processes, models, and Markov chains. □

Proof of Theorem 1. We note that this can also be obtained by the decomposition techniques of Chan and Tong (1985), p. 673, or Theorem 4.2 of Tong (1996). However, we verify geometric ergodicity by applying a so-called drift criterion for Markov chains. This has the advantage that it can be extended to proof other results as well.

We first note that every λ-non-null compact set is small (Chan and Tong, 1985, pp. 668, 669) and petite (Meyn and Tweedie, 1993, pp. 121).

We now verify the drift criterion (15.3) of Theorem 15.0.1 in Meyn and Tweedie (1993) to obtain the desired result. The proof is based on a similar proof by Tjøstheim (1990) for recurrence of vector threshold models. Define the matrix

$$\Psi := \begin{bmatrix} \psi_1 & \psi_2 & \dots & \psi_{p-1} & \psi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

Then we can write (3) as

$$x_t = \Psi x_{t-1} + F(x_{t-1}) + \eta_t, \tag{15}$$

where $F(\cdot)$ is the non-linear part and the intercept. Now, there exists a transformation Σ such that $\Upsilon = \Sigma\Psi\Sigma^{-1}$ has the eigenvalues of Ψ along its diagonal and arbitrarily small off-diagonal elements (cf. Bellmann, 1970, p. 205). Let the test function be $V(x) = \|\Sigma x\|$ and the test set $\mathcal{C} = \{x \in \mathbb{R}^p : V(x) \leq c\}$ for some $c < \infty$. $\|\cdot\|$ denotes the Euclidean norm for a vector and the spectral norm for a matrix. Then we have

$$\begin{aligned} \mathsf{E}[V(x_t)|x_{t-1} = x] &\leq \|\Sigma\Psi x\| + \|\Sigma F(x)\| + \mathsf{E}\|\Sigma\eta_t\| \\ &\leq (\|\Lambda\| + \|\Delta\|)V(x) + \|\Sigma F(x)\| + \mathsf{E}\|\Sigma\eta_t\|, \end{aligned}$$

where $\Lambda = \mathrm{diag}(\Upsilon)$ and $\Delta = \Upsilon - \Lambda$. By assumption $\|\Lambda\| < 1$, and Σ can be chosen such that $(\|\Lambda\| + \|\Delta\|) < 1 - \epsilon$ for some $\epsilon > 0$. Since the second and third term are bounded we can choose c such that $\mathsf{E}[V(x_t)|x_{t-1} = x] \leq (1-\epsilon)V(x) + \delta\mathbb{1}_{\mathcal{C}}(x)$ for some $\delta < \infty$ and for all x. This is also valid for the test function $V(x) + 1$ and we have the desired result. □

Proof of Corollary 1. Since the drift criterion of Theorem 15.0.1 (Meyn and Tweedie, 1993) is satisfied, the ARNN is also V-uniformly ergodic (Meyn and Tweedie, 1993, Theorem 16.0.1). This implies V-geometric mixing (Theorem 16.1.5 and discussion on p. 388 of Meyn and Tweedie, 1993), i.e.,

$$|\mathsf{E}[g(x_n)h(x_{n+m})] - \mathsf{E}[g(x_n)]\mathsf{E}[h(x_{n+m})]| \leq \kappa\varrho^m$$

for all $g^2(\cdot), h^2(\cdot) \leq V(\cdot)$ and $n, m \in \mathbb{Z}$, which is equivalent to the desired result (cf. proof of Theorem A, pp. 883, 884, Athreya and Pantula, 1986). □

Proof of Corollary 2. Omitted (e.g., Billingsley, 1995, Lemma 3, p. 365). □

Proof of Corollary 3. Omitted (cf. Brockwell and Davis, 1991, Corollary 4.3.2.). □

Proof of Proposition 1. To establish ergodicity the drift criterion (9.1), (9.2) of Theorem 9.1 in Tweedie (1976) is verified for the test function $V(y) = |y|$ and the test set $\mathcal{C} = \{y : |y| \leq c\}$.

$$\begin{aligned}\mathsf{E}[V(y_t)|y_{t-1} = y] &= \mathsf{E}[(y + G(y) + \varepsilon_t)\mathbb{1}_1 - (y + G(y) + \varepsilon_t)\mathbb{1}_2] \\ &= (y + G(y))(P_1 - P_2) + \mathsf{E}[\varepsilon_t\mathbb{1}_1] - \mathsf{E}[\varepsilon_t\mathbb{1}_2] \\ &= (y + G(y)) - 2(y + G(y))P_2 - 2\mathsf{E}[\varepsilon_t\mathbb{1}_2],\end{aligned}$$

where $\mathbb{1}_1 = \mathbb{1}_{\{\varepsilon_t \geq -y-G(y)\}}$, $\mathbb{1}_2 = \mathbb{1}_{\{\varepsilon_t < -y-G(y)\}}$, $P_1 = \mathsf{P}(\varepsilon_t \geq -y - G(y))$, and $P_2 = \mathsf{P}(\varepsilon_t < -y - G(y))$. The last equality follows from $P_1 + P_2 = 1$ and $\mathsf{E}[\varepsilon_t\mathbb{1}_1] + \mathsf{E}[\varepsilon_t\mathbb{1}_2] = 0$. Now, for y positive and large enough

$$\mathsf{E}[V(y_t)|y_{t-1} = y] \leq y - \epsilon = V(y) - \epsilon$$

for some $\epsilon > 0$, since $\lim_{y\to\infty}(y + G(y))P_2 = 0$, $\lim_{y\to\infty}\mathsf{E}[\varepsilon_t\mathbb{1}_2] = 0$, and $\lim_{y\to\infty} G(y) = -a$. The case y negative is proved analogously. Since for $y \in \mathcal{C}$, $\mathsf{E}[V(y_t)|y_{t-1} = y]$ is bounded, ergodicity and asymptotic stationarity follow.

Strong mixing is a direct consequence of ergodicity; see Athreya and Pantula (1986). □

Proof of Theorem 2. Sufficiency: Rewriting (15) and taking expectations of the norm gives

$$\mathsf{E}\|\eta_t\| \leq \mathsf{E}\|x_t\| + \|\Psi\|\mathsf{E}\|x_{t-1}\| + \mathsf{E}\|F(x_{t-1})\| < \infty.$$

Necessity: The use of the test function $V(x) = \|\Sigma x\|^k + 1$ in the proof of Theorem 1 establishes the desired result (Meyn and Tweedie, 1993, Theorem 14.3.7, Theorem 15.0.1). □

Proof of Corollary 4. Omitted. □

Proof of Proposition 2. From Tong (1996), Theorem 4.6, a symmetric joint p.d.f. of $\{y_t\}$ is equivalent to $h(-x, \theta) = -h(x, \theta)$ (π-a.s.). Therefore,

$$h(x, \theta) + h(-x, \theta) = 2\nu + \sum_i \beta_i[G(\phi_i'x + \mu_i) + G(-\phi_i'x + \mu_i)] = 0.$$

Now assume $\mu_j \neq 0$ for some j and $\mu_i = 0$ for all $i \neq j$. Then

$$\begin{aligned}2\nu + \sum_{i\neq j} \beta_i[G(\phi_i'x) + G(-\phi_i'x)] + \beta_j[G(\phi_j'x + \mu_j) + G(-\phi_j'x + \mu_j)] \\ = 2\nu + \sum_i \beta_i + \beta_j G(\phi_j'x + \mu_j) - \beta_j G(\phi_j'x - \mu_j) = 0,\end{aligned}$$

which implies that $\beta_j = 0$, contradicting Assumption 1. This follows because $G(\phi_j'x + \mu_j)$, $G(\phi_j'x - \mu_j)$, and the constant 1 are linear independent functions for the logistic $G(\cdot)$ (Lemma 2.7 Hwang and Ding, 1997) and the support of π is $\mathbb{R}^p$. Therefore, $\mu_i = 0$ for all i and $2\nu + \sum_i \beta_i = 0$. □

Proof of Corollary 5. Omitted. □

Proof of Theorem 3. We verify the drift criterion (11.4) of Theorem 11.3 in Tweedie (1976). The proof is based on a similar proof by Tjøstheim (1990) for transience of vector threshold models.

Representation (15) is used. Since $\psi(\cdot)$ has distinct roots, there exists an orthonormal basis of eigenvectors $\{\chi_i\}_{i=1}^p$ of Ψ, such that an arbitrary $x \in \mathbb{R}^p$ has the representation $x = \sum_i \vartheta_i(x)\chi_i$. Furthermore, there exists a transformation Σ such that $\Upsilon = \Sigma\Psi\Sigma^{-1}$ is diagonal and has the eigenvalues of Ψ along its diagonal. Let j be the coordinate number associated with an eigenvalue λ_j, $|\lambda_j| > 1$. We define the test function and the test set as $V(x) = \exp(-|\vartheta_j(\Sigma x)|)$, $\mathcal{C} = \{x : |\vartheta_j(\Sigma x)| \leq c\}$. Obviously, condition (11.4b) of Tweedie (1976) holds. Since $|\vartheta_i(x)| \leq \|x\|$ and the coordinate functions $\vartheta_i(x)$ are linear,

$$\begin{aligned}
\mathsf{E}[V(x_t)|x_{t-1} = x] &\leq \exp(-|\vartheta_j(\Sigma\Psi x)|)\exp(|\vartheta_j(\Sigma F(x))|)\exp(|\vartheta_j(\Sigma\eta_t)|) \\
&\leq \exp(-|\lambda_j||\vartheta_j(\Sigma x)|)\exp(\|\Sigma F(x)\|)\exp(\|\Sigma\eta_t\|) \\
&\leq \exp(-|\vartheta_j(\Sigma x)|) = V(x),
\end{aligned}$$

for x large enough. In addition, we have used $\vartheta_j(\Sigma\Psi x) = \vartheta_j(\Upsilon\Sigma x) = \vartheta_j(\Upsilon\sum_i \vartheta_i(\Sigma x)\chi_i)$ $= \lambda_j\vartheta_i(\Sigma x)$, $\mathsf{E}[\exp(|\varepsilon_t|)] < \infty$, and $F(\cdot)$ is bounded. The desired result follows directly since both, $\mathcal{C}$ and its complement, are Lebesgue non-null Borel sets. □

Proof of Theorem 4. This result follows directly from Lemma 2.2 of Phillips (1987) and the properties of ARNNs. □

Proof of Theorem 5. We use Theorem 3.1 from White and Domowitz (1984) and check their Assumptions 1-4.

Assumption 1 and 3 are trivial.

Assumption 2: Let $d_t = \sup_{\theta\in\Theta}(h(x_{t-1},\theta_0) - h(x_{t-1},\theta))^2$. Then d_t dominates $(h(x_{t-1},$ $\theta_0) - h(x_{t-1},\theta))^2$. Successive application of the Triangle and Cauchy-Schwarz inequality to d_t and using the moment condition in Theorem 5 provides the desired result, because Θ is compact and $h(x_{t-1},\theta)$ is continuous and bounded on Θ.

Assumption 4: Our Assumptions 1, 3, and 4 ensure that $(h(x_{t-1},\theta_0) - h(x_{t-1},\theta))^2 > 0$ if and only if $\theta \neq \theta_0$ (e.g., Hwang and Ding (1997), Theorem 2.3a and Kůrková and Kainen (1994), Proposition 3.9), which is (in our context) enough to ensure identifiable uniqueness of θ_0.

Applying Theorem 3.1 of White and Domowitz (1984) and using the geometric mixing for ARNNs provides the final result. □

Proof of Theorem 6. Theorem 3.2 of White and Domowitz (1984) establishes the desired result. We check their Assumptions 1′-9.

Assumptions 1′ and 5 are trivial.

Assumption 6: Let $h_i(x_{t-1},\theta)$ denote $\partial h_i(x_{t-1},\theta)/\partial\theta_i$ $(i = 1,\ldots,r)$. Then straightforward calculations yield that each term $\mathsf{E}[h_i^2(x_{t-1},\theta_0)\varepsilon_t^2]$ is bounded by quadratic terms, that is, $\mathsf{E}y_t^2$, $\mathsf{E}\varepsilon_t^2$. Hence, the moment condition in Theorem 6 suffices to establish Assumption 6.

Assumption 8: Denote the elements of the Hessian as $h_{ij}(x_{t-1},\theta)$ $(i,j = 1,\ldots,r)$. A dominating function is given by taking the supremum over Θ, that is, $d_t = \sup_{\theta\in\Theta}(h_i(x_{t-1},$ $\theta)h_j(x_{t-1},\theta) - h_{ij}(x_{t-1},\theta)(y_t - h(x_{t-1},\theta))$. Using the assumptions and successive application of the Triangle and Cauchy-Schwarz inequality to d_t yield the desired result.

Assumption 7 and 9: We first show that $\det(\mathsf{E}[\nabla h_t'\nabla h_t]) > 0$ at θ_0, where ∇h_t is the $1 \times r$ gradient of $h(x_{t-1},\theta)$ with respect to θ. However, this follows directly from our Assumptions 3 and 6 and Hwang and Ding (1997), Theorem 2.3b. This also establishes Assumption 7 (see remark below Assumption 7 in White and Domowitz, 1984). Since the Hessian $\nabla^2\bar{Q}_n(\theta_0)$ is just $2\mathsf{E}[\nabla h_t'\nabla h_t]$ and $\det(\cdot)$ is a continuous function in θ the Hessian has constant rank in some neighborhood of θ_0.

The geometric mixing property of ARNNs concludes the proof. □

Proof of Theorem 7. It follows from Theorem 5.1 of Phillips (1987) and the properties of ARNNs. □

Proof of Proposition 3. Omitted (Phillips, 1987, Theorem 3.1d). □

References

Athreya, K. B. and Pantula, S. G. (1986). Mixing properties of harris chains and autoregressive processes. *Journal of Applied Probability*, **23**, 880–892.

Banerjee, A., Dolado, J. J., Galbraith, J. W., and Hendry, D. F. (1993). *Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

Bellmann, R. (1970). *Introduction to Matrix Analysis*. New York: McGraw-Hill.

Billingsley, P. (1995). *Probability and Measure*. New York: Wiley, 3rd edn.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. New York: Springer Verlag, 2nd edn.

Chan, K. S. and Tong, H. (1985). On the use of the deterministic lyapunov function for the ergodicity of stochastic difference equations. *Advances in Applied Probability*, **17**, 666–678.

Feigin, P. D. and Tweedie, R. L. (1985). Random coefficient autoregressive processes: A markov chain analysis of stationarity and finiteness of moments. *Journal of Time Series Analysis*, **6**, 1–14.

Gallant, A. R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Basil Blackwell.

Granger, C. W. J. (1995). Modelling non-linear relationships between extended-memory variables. *Econometrica*, **63**, 265–279.

Granger, C. W. J. and Hallmann, J. (1991). Long memory series with attractors. *Oxford Bulletin of Economics and Statistics*, **53**, 11–26.

Granger, C. W. J. and Swanson, N. (1996). Future developments in the study of cointegrated variables. *Oxford Bulletin of Economics and Statistics*, **58**, 537–553.

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

Hwang, J. T. G. and Ding, A. A. (1997). Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, **92**, 748–757.

Kůrková, V. and Kainen, P. C. (1994). Functionally equivalent feedforward neural networks. *Neural Computation*, **6**, 543–558.

Leisch, F., Trapletti, A., and Hornik, K. (1999). Stationarity and stability of autoregressive neural network processes. In *Advances in Neural Information Processing Systems*, vol. 11. To appear.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. London: Springer Verlag.

Perron, P. (1988). Trends and random walks in macroeconomic time series. *Journal of Economic Dynamics and Control*, **12**, 297–332.

Phillips, P. C. B. (1987). Time series regression with a unit root. *Econometrica*, **55**, 277–301.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biomètrika*, **75**, 335–346.

Sussmann, H. J. (1992). Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, **5**, 589–593.

Tjøstheim, D. (1990). Non-linear time series and markov chains. *Advances in Applied Probability*, **22**, 587–611.

Tong, H. (1996). *Non-Linear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press.

Tweedie, R. L. (1976). Criteria for classifying general markov chains. *Advances in Applied Probability*, **8**, 737–771.

Wang, T. and Sheng, Z. (1996). Asymptotic stationarity of disrete-time stochastic neural networks. *Neural Networks*, **9**, 957–963.

White, H. (1989). Some asymptotic results for learning in single hidden-layer feed-forward network models. *Journal of the American Statistical Association*, **84**, 1003–1013.

White, H. (1996). Parametric statistical estimation with artificial neural networks. In Smolensky, P., Mozer, M. C., and Rumelhart, D. E. (eds.), *Mathematical Perspectives on Neural Networks*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

White, H. and Domowitz, I. (1984). Non-linear regression with dependent observations. *Econometrica*, **52**, 143–161.

Department of Operations Research, Vienna University of Economics and Business Administration, Augasse 2-6, A-1090 Vienna, Austria

Department of Statistics and Probability Theory, Technical University of Vienna, Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria