# A uniform central limit theorem for neural network based autoregressive processes with applications to change-point analysis*

Claudia Kirch† Joseph Tadjuidje Kamgaing‡

March 16, 2011

**Abstract**

We consider an autoregressive process with a nonlinear regression function that is modeled by a feedforward neural network. We derive a uniform central limit theorem which is useful in the context of change-point analysis. We propose a test for a change in the autoregression function which – by the uniform central limit theorem – has asymptotic power one for a large class of alternatives including local alternatives.

**Keywords:** Uniform central limit theorem, nonparametric regression, neural network, autoregressive process

**AMS Subject Classification 2000:** 60F05, 62J02, 62M45

## 1 Introduction

Limit theory – in general – can be described as the heart of probability and mathematical statistics. Uniform central limit theorems for dependent random variables – in particular

*The work was supported by the DFG graduate college 'Mathematics and Practice' as well as by the DFG grant 565216. The position of the first author was financed by the Stifterverband für die Deutsche Wissenschaft by funds of the Claussen-Simon-trust.

†Karlsruhe Institute of Technology (KIT), Institute for Stochastics, Kaiserstr. 89 D–76133 Karlsruhe, Germany; claudia.kirch@kit.edu

‡University Kaiserslautern, Department of Mathematics, Erwin-Schrödinger-Straße, D–67 653 Kaiserslautern, Germany; tadjuidj@mathematik.uni-kl.de

– have given a new dimension to the asymptotic theory for sums of processes indexed by families of sets by giving rise to many applications out of reach of the classical central limit theorem.

In this paper we consider a nonlinear autoregressive time series, where we model the autoregression function by a feedforward neural network. Due to its universal approximation property, a large class of functions can be approximated by a neural network to any degree of accuracy (confer e.g. White [10] or Franke et al. [4]). Therefore, this setup is very general and able to model many real-life time series while – at the same time – being mathematical feasible and computationally easier to handle due to its parametric nature.

For $\theta = (\nu_0, \ldots, \nu_H, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_H, \beta_1, \ldots, \beta_H)$, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jp})$,

$$f(\mathbf{x}, \theta) = \nu_0 + \sum_{h=1}^{H} \nu_h \psi(< \boldsymbol{\alpha}_h, \mathbf{x} > + \beta_h), \tag{1.1}$$

denotes a one layer feedforward neural network with $H$ hidden neurons, $<,>$ is the classical scalar product on $\mathbb{R}^p$. In this paper we assume that $\psi$ belongs to the class of sigmoid activation functions that satisfy

$$\lim_{x \to -\infty} \psi(x) = 0, \qquad \lim_{x \to \infty} \psi(x) = 1, \qquad \psi(x) + \psi(-x) = 1. \tag{1.2}$$

A popular example is the logistic function $\psi(x) = (1 + e^{-x})^{-1}$ which also fulfills Assumption C.1.

The time series model, we have in mind in this paper, is given by

$$X_t = f\left(\mathbb{X}_{t-1}, \theta_0\right) + e_t, \tag{1.3}$$

where $\mathbb{X}_{t-1} = (X_{t-1}, \ldots, X_{t-p})$, $\theta_0$ is fixed but unknown, $e_t$ independent of $\mathcal{F}_{t-1} = \sigma\{X_u, u \leq t-1\}$ the $\sigma$-algebra generated by the observations up to time $t-1$. Furthermore, $\{e_t : 1 \leqslant t \leqslant n\}$ are independent identically distributed random errors with a positive variance.

Stockis et al. [8] use these time series as building blocks in a regime-switching model, so called CHARME-models, in the context of financial time series. In their model the duration time in each regime is random and driven by a hidden Markov chain, while in classical change-point analysis the duration time is usually fixed and deterministic. Motivated by the CHARME time series Kirch and Tadjuidje-Kamgaing [6] developed change-point tests in such a setup. Change-point analysis deals with the question whether the stochastic structure of the observations has changed at some unknown point in the sample, which is an important question in diverse areas such as economy, finance, geology, physics or quality control. For a detailed discussion we refer to the book by Csörgő and Horváth [2].

In Section 2 we develop a uniform central limit theorem involving the above autoregressive time series. Based on these uniform central limit theorems we can enlarge in Section 3 the class of alternatives for which the change-point tests developed in Kirch and Tadjuidje-Kamgaing [6] have asymptotic power one allowing in particular for local changes. Finally, in Section 4 the proofs can be found.

# 2 Uniform Central Limit Theorem

In the following $\{X_t\}$ is an arbitrary stationary time series and not necessarily of the form (1.3).

The uniform central limit theorem for empirical processes traces back to Dudley [3] and is founded on notions like VC-subgraphs or covering numbers. We will make use of this methodology.

Denote by $\mathcal{F} = \{f(\theta, \cdot); \theta \in \Theta\}$ the set of all feedforward networks as defined previously with parameter $\theta \in \Theta$.

To obtain the uniform central limit theorem we need the following assumptions.

**C. 1.** The set $\Theta$ is compact. Furthermore $\psi$ is a sigmoid activation function as in (1.2) which is continuously differentiable with bounded derivative.

The latter assumption is e.g. fulfilled for the logistic function.

**C. 2.** $X_t$ is stationary and $\beta$-mixing with mixing coefficient $\beta(\cdot)$ fulfilling for some $\tau > 2$ as $k \to \infty$

$$k^{\tau/(\tau-2)}(\log k)^{2(\tau-1)/(\tau-2)}\beta(k) \to 0.$$

Furthermore $\mathbb{E}|X_1|^\nu < \infty$ for some $\nu > 2$.

This is a classical condition in nonlinear time series analysis and can be derived with little effort using the stability theory for Markov processes, see e.g. Meyn and Tweedie [7]. Indeed, this property is a consequence of the existence of a stationary solution that is geometric ergodic. In this situation the assumption on the rate is therefore fulfilled since the time series is $\beta$-mixing with an exponential rate.
In a more general setup and for the related CHARME-models this $\beta$-mixing property has been proven by Stockis et al. [8].

For the next assumption we first need to recall the definition of covering numbers.

**Definition 2.1.** *(Covering Numbers) The covering number $\mathcal{N}(\epsilon, \mathcal{F}, \|.\|)$ is the minimal number of balls of radius $\epsilon$, i.e. $\{g : \|g - f\| < \epsilon\}$, needed to cover $\mathcal{F}$, where it is not necessary that $f \in \mathcal{F}$.*

**C. 3.** Assume that for some $0 < q < \infty$

$$\int_0^\infty (\log \mathcal{N}(u, \mathcal{F}, \|.\|_q))^{1/2} du < \infty.$$

This uniform entropy condition is classical in the context of weak convergence of empirical processes (cf. e.g. Van der Vaart and Wellner [9]).

We are now ready to prove the uniform central limit theorem.

**Theorem 2.1.** *Assume C.1– C.3 hold. Then,*

$$\left\{ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (f(\mathbb{X}_t, \theta) - \mathbb{E} f(\mathbb{X}_t, \theta)); \theta \in \Theta \right\} \xrightarrow{w} \{G(\theta); \theta \in \Theta\} \quad \text{in } l_\infty(\Theta),$$

*where* $\{G(\theta)\}$ *has a version with uniformly bounded and uniformly continuous paths with respect to the* $\|.\|_2$*-norm. In particular*

$$\sup_{\theta \in \Theta} \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (f(\mathbb{X}_t, \theta) - \mathbb{E} f(\mathbb{X}_t, \theta)) \right| = O_P(1). \tag{2.1}$$

# 3 Change-point tests under local alternatives

We observe a time series $Z_t = X_t 1_{\{t \leqslant k^*\}} + Y_{t,n} 1_{\{t > k^*\}}$ with a possible switch from $X_t$ to $Y_{t,n}$ at some unknown time point $1 \leqslant k^* = k^*(n) \leqslant n$. The time series $X_t$ is not necessarily of the form (1.3) but we assume that it is well approximated by $\widetilde{X}_t = f(\widetilde{\mathbb{X}}_{t-1}, \widetilde{\theta}_0) + e_t$ for some $\widetilde{\theta}_0$ in the interior of $\Theta$, $Y_{t,n}$ is an arbitrary time series. If $X_t$ follows (1.3), then $\widetilde{\theta}_0 = \theta_0$. For more details on the derivation of $\widetilde{\theta}_0$ we refer to Kirch and Tadjuidje-Kamgaing [6]. The unknown parameter $k^*$ is called the change-point and we are interested in the testing problem

$$H_0 : k^* = n \qquad \text{vs.} \qquad H_1 : k^* < n.$$

Our testing procedures are based on various functionals of the partial sums of estimated residuals

$$\hat{S}_n(k) = \sum_{t=p+1}^{k} \widehat{e}_t = \sum_{t=p+1}^{k} \left( Z_t - f(\mathbb{Z}_{t-1}, \widehat{\theta}_n) \right), \tag{3.1}$$

where $\widehat{\theta}_n$ is the least-squares estimator of $\widetilde{\theta}_0$ under $H_0$, i.e. the minimizer of $\sum_{t=p+1}^{n} (Z_t - f(\mathbb{Z}_{t-1}, \theta))^2$. Kirch and Tadjuidje-Kamgaing [6] prove consistency and asymptotic normality of $\widehat{\theta}_n$ under the null hypothesis under some mild regularity conditions - in addition to some asymptotic results under alternatives as well as misspecification. If the estimator is not in the interior of $\Theta$ we reject the null hypothesis as asymptotically this can only happen under the alternative – or if the model (1.3) is not suitable for the data set at hand. Consequently, we can assume w.l.o.g. $\widehat{\theta}_n \in \Theta^\circ$, the interior of $\Theta$, for asymptotic power considerations.

Typical test statistics in this context are given by

$$\begin{aligned} T_{1,n} &= \max_{p<k<n} \left( \sqrt{\frac{n-p}{k(n-p-k)}} \, |\hat{S}_n(k)| \right), \\ T_{2,n}(q) &= \max_{p<k<n} \left( \frac{1}{\sqrt{n-p}\, q(\frac{k}{n-p})} \, |\hat{S}_n(k)| \right), \\ T_{3,n}(G) &= \max_{p+G<k\leqslant n} \frac{1}{\sqrt{G}} \left| \hat{S}_n(k) - \hat{S}_n(k-G) \right| \end{aligned} \tag{3.2}$$

where $q(\cdot)$ is a positive weight functions defined on $(0,1)$ fulfilling certain conditions, e.g. $q(t) = (t(1-t))^{\gamma}$, $0 \leqslant \gamma < 1/2$, and $G \to \infty$ with a certain rate. In Kirch and Tadjuidje-Kamgaing [6] the asymptotics of the above statistics under the null hypothesis (and certain regularity conditions) are obtained.
$T_{2,n}(q)$ converges in distribution to $\sup_{0<t<1} \frac{|B(t)|}{q(t)}$, where $B(\cdot)$ is a Brownian bridge. The other two statistics converge a.s. to infinity but such that $\alpha_n T_n - \beta_n$ for some $\alpha_n$, $\beta_n$ converges to a Gumbel distribution, which suffices to obtain asymptotic critical values. To obtain tests with asymptotic power one in these cases, it suffices to show that $\frac{\alpha_n}{\beta_n} T_n \xrightarrow{P} \infty$. For $T_{1,n}$ it holds $\frac{\alpha_{1,n}}{\beta_{1,n}} \sim (\log\log n)^{-1/2}$ and for $T_{3,n}(G)$ it holds $\frac{\alpha_{3,n}}{\beta_{3,n}} \sim (\log(n/G))^{-1/2}$, where $c_n \sim d_n \iff \frac{c_n}{d_n} \to c$ for some constant $c > 0$.

In Kirch and Tadjuidje-Kamgaing [6] it was shown that the above tests have asymptotic power one for certain fixed alternatives, where $Y_{t,n} = Y_t$, using a uniform law of large numbers. Using the uniform central limit theorem of the previous section instead allows us to generalize this result to a larger class of alternatives, including local alternatives such as $Y_{t,n} = X_t + d_n Z_t$ for some $d_n \to 0$ or $X_t$ as in (1.3) and $Y_{t,n} = f(\mathbb{Y}_{t-1,n}, \theta_n) + e_t$ with $\theta_n \to \theta_0$. Local alternatives are often considered in statistics to get an idea about the sensitivity of the test for small differences.

**A. 1.** The change-point fulfills $k^* = \lfloor \lambda n \rfloor$ for some $0 < \lambda < 1$.

**A. 2.** $\{X_t\}$ fulfills (2.1) as well as $\sum_{t=1}^{n} (X_t - \mathbb{E}X_1) = O_P(\sqrt{n})$.

**A. 3.** It holds $b_n \left( \mathbb{E} f(\mathbb{X}_p, \theta)|_{\theta=\widehat{\theta}_n} - \mathbb{E}X_1 \right)^2 \xrightarrow{P} \infty$ for $b_n \to \infty$ specified below.

Under local alternatives the estimator $\widehat{\theta}_n$ will typically converge to $\widetilde{\theta}_0$ with a rate depending on the rate of convergence of the local alternative, i.e. the rate with which $d_n \to 0$ resp. $\theta_n \to \theta_0$ in the examples above. For a general neural network it is difficult to quantify these rates due to the highly nonlinear structure of $f$. In the classical special case of a mean change (i.e. a trivial neural network with $H = 0$ and $X_i = \mu + e_i$, $Y_{i,n} = \mu + d_n + e_i$), Assumption A. 3 reduces to the well known condition $d_n \to 0$ but $b_n |d_n| \to \infty$.

The type of mean change condition as in A.3 is typical if the test statistics are based on estimated residuals and already arise in linear regression models (cf. e.g. Hušková and Koubkova [5]). For some simulations concerning detectability of different types of changes we refer to Kirch and Tadjuidje-Kamgaing [6].

The next theorem shows that the tests corresponding to the statistics in (3.2) have asymptotic power one under the above conditions.

**Theorem 3.1.** *Assume that A.1 – A.3 hold with*

$$b_n = \begin{cases} \frac{n}{\log\log n}, & \text{for a)}, \\ n, & \text{for b)}, \\ \frac{G}{\log(n/G)}, & \text{for c)}. \end{cases}$$

*Then it holds*

$$a) \quad (\log\log n)^{-1/2} T_{1,n} \xrightarrow{P} \infty, \qquad b) \quad T_{2,n}(q) \xrightarrow{P} \infty,$$
$$c) \quad (\log(n/G))^{-1/2} T_{3,n} \xrightarrow{P} \infty,$$

*if* $\min_{\eta \leqslant t \leqslant 1-\eta} q(t) > 0$ *for any* $\eta > 0$ *and* $G \to \infty$, $(\log n)/G \to 0$, $G/n \to 0$.

# 4 Proofs

We start with some auxiliary lemmas.

**Lemma 4.1.** *Let* $f(\mathbf{x}, \theta)$, *be a neural network fulfilling C.1,* $\mathbf{x} = (x_1, \ldots, x_p)^T$. *Then, for any* $\theta_1, \theta_2 \in \Theta$ *there exists a constant* $D > 0$ *not depending on* $\theta_1$, $\theta_2$ *or* $\mathbf{x}$ *such that*

$$|f(\mathbf{x}, \theta_1) - f(\mathbf{x}, \theta_2)| \leqslant D \|\theta_1 - \theta_2\|_2 \max_{i=1,\ldots,p} |x_i|,$$

*where* $\|\cdot\|_2$ *is the Euclidian norm.*

**Proof.** An application of the mean value theorem with respect to $\theta$ yields for any $\theta_1, \theta_2 \in \Theta$

$$f(\mathbf{x}, \theta_1) - f(\mathbf{x}, \theta_2) = \nabla f(\mathbf{x}, \xi)^T (\theta_1 - \theta_2)$$

for some $\xi \in \mathrm{con}(\Theta)$, the convex hull of $\Theta$. Since $\Theta$ is compact, $\mathrm{con}(\Theta)$ is compact. By this and the boundedness of the derivative of $\psi$ we find $D > 0$ such that

$$|\nabla f(\mathbf{x}, \xi)^T (\theta_1 - \theta_2)| \leqslant \|\theta_1 - \theta_2\|_2 \, \|\nabla f(\mathbf{x}, \xi)\|_2 \leqslant D \, \|\theta_1 - \theta_2\|_2 \max_{i=1,\ldots,p} |x_i|.$$

■

**Lemma 4.2.** *Let* $f(\mathbf{x}, \theta)$, *be a neural network fulfilling C.1. Then, there exists* $R > 0$, *such that for* $0 < \varepsilon \leqslant R$ *it holds for any probability measure* $Q$

$$\mathcal{N}(\varepsilon, \mathcal{F}, \|.\|_{L^2(Q)}) \leqslant \left( \frac{R \|F\|_{L^2(Q)}}{\varepsilon} \right)^d$$

*where* $F(x) = D \max_{i=1,\ldots,p} |x_i|$ *for some constant* $D > 0$, $d = H(p+2) + 1$ *and* $\|f(x)\|_{L^2(Q)} = \left( \int f^2(x) \, dQ(x) \right)^{1/2}$.

**Proof.** Denote by $\mathcal{N}_{[\,]}(\epsilon, \mathcal{F}, \|.\|)$ the bracketing number and by $\mathcal{M}(\epsilon, \mathcal{F}, \|.\|)$ the packing number, then it holds (cf. Van der Vaart and Wellner [9])

$$\mathcal{N}(\varepsilon, \mathcal{F}, \|.\|) \leq \mathcal{N}_{[\,]}(2\varepsilon, \mathcal{F}, \|.\|), \tag{4.1}$$
$$\mathcal{N}(\varepsilon, \mathcal{F}, \|.\|) \leq \mathcal{M}(\varepsilon, \mathcal{F}, \|.\|). \tag{4.2}$$

Lemma 4.1 and Theorem 2.7.11 in Van der Vaart and Wellner [9] yield for any $\varepsilon > 0$ $\mathcal{N}_{[\,]}(2\varepsilon\|F\|_{L^2(Q)}, \mathcal{F}, \|.\|_{L^2(Q)}) \leqslant \mathcal{N}(\varepsilon, \mathcal{F}, \|.\|_2)$. Hence by (4.1) (w.l.o.g. $\|F\|_{L^2(Q)} < \infty$)

$$\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{L^2(Q)}) \leqslant \mathcal{N}\left(\frac{\varepsilon}{\|F\|_{L^2(Q)}}, \mathcal{F}, \|\cdot\|_2\right)$$

for which (4.2) implies

$$\mathcal{N}\left(\frac{\varepsilon}{\|F\|_{L^2(Q)}}, \mathcal{F}, \|.\|_2\right) \leqslant \mathcal{N}\left(\frac{\varepsilon}{\|F\|_{L^2(Q)}}, \widetilde{\mathcal{F}}, \|.\|_2\right) \leqslant \mathcal{M}\left(\frac{\varepsilon}{\|F\|_{L^2(Q)}}, \widetilde{\mathcal{F}}, \|.\|_2\right)$$
$$\leqslant \left(\frac{3R\|F\|_{L^2(Q)}}{\varepsilon}\right)^d,$$

for any $0 < \varepsilon \leqslant R$, where $\widetilde{\mathcal{F}} = \{f(\theta, \cdot); \theta \in B(0, R)\}$ for a suitable $R > 0$, $B(0, R)$ is the ball around 0 of radius $R$ in $\mathbb{R}^d$. The last line follows from Exercise 6, page 94, of Van der Vaart and Wellner [9]. ■

**Proof of Theorem 2.1.** The assertion follows from Lemma 2.1 of Arcones and Yu [1]. The statement there is given for the minimal envelope function but the proof shows that it remains true for any envelope function $F$. Furthermore it is sufficient if their condition (2.10) holds for all $0 < \varepsilon \leqslant \varepsilon_0$ for some $\varepsilon_0 > 0$. In our case we consider $F$ as in Lemma 4.1. It remains to check the conditions of Lemma 2.1 of Arcones and Yu [1]. Let $p = \min(\tau, \nu)$, then their Condition (2.3) holds by $\mathbb{E}|X_t|^p < \infty$ according to C.2. Condition (2.4) holds by C.2, (2.10) by C.3 and (2.11) for small $\varepsilon$ by Lemma 4.2. ■

**Proof of Theorem 3.1.** By (3.1), A.1 and A.2 it holds

$$\hat{S}_n(k^*) = \sum_{j=p+1}^{k^*} (X_j - f(\mathbb{X}_{j-1}, \widehat{\theta}_n))$$
$$= \sum_{i=p+1}^{k^*} X_i + k^* \mathbb{E} f(\mathbb{X}_p, \theta)|_{\theta=\widehat{\theta}_n} + O\left(\sup_{\theta\in\Theta} \left|\sum_{i=p+1}^{k^*} (f(\mathbb{X}_{i-1}, \theta) - \mathbb{E} f(\mathbb{X}_p, \theta))\right|\right)$$
$$= \lambda n \left(\mathbb{E}X_1 - \mathbb{E} f(\mathbb{X}_p, \theta)|_{\theta=\widehat{\theta}_n}\right) + O_P\left(\sqrt{n}\right)$$

By A.3 this implies

$$(\log\log n)^{-1/2}\, T_{1,n} \geqslant \lambda \sqrt{\frac{n}{\log\log n}} \left|\mathbb{E}X_1 - \mathbb{E} f(\mathbb{X}_p, \theta)|_{\theta=\widehat{\theta}_n}\right| + O_P((\log\log n)^{-1/2}) \xrightarrow{P} \infty.$$

Since by A.1 and $\min_{\eta\leqslant t\leqslant 1-\eta} q(t) > c$ for any $\eta > 0$ it holds $q(k^*) \geqslant \tilde{c}$ for some $\tilde{c} > 0$, which together with A.3 implies

$$T_{2,n}(q) \geqslant \lambda\sqrt{n}\ \left|\mathbb{E}X_1 - \mathbb{E} f(\mathbb{X}_p, \theta)|_{\theta=\widehat{\theta}_n}\right| + O_P(1) \xrightarrow{P} \infty.$$

Similarly

$$\left|\hat{S}_n(k^*) - \hat{S}_n(k^* - G)\right| = G \left|\mathbb{E}X_1 - \mathbb{E} f(\mathbb{X}_p, \theta)|_{\theta=\widehat{\theta}_n}\right| + O_P(\sqrt{G})$$

hence by A.3

$$(\log(n/G))^{-1/2}\, T_{3,n}(G) \geqslant \sqrt{\frac{G}{\log(n/G)}} \left|\mathbb{E}X_1 - \mathbb{E} f(\mathbb{X}_p, \theta)|_{\theta=\widehat{\theta}_n}\right| + O_P\left(\log(n/G)^{-1/2}\right) \xrightarrow{P} \infty.$$

■

# References

[1] Arcones, M. A. and Yu, B. Central limit theorems for empirical and u-processes of stationary mixing sequences. *Econometrica*, 7:48–71, 1994.

[2] Csörgő, M., and Horváth, L. *Limit Theorems in Change-Point Analysis*. Wiley, Chichester, 1997.

[3] Dudley, R.M. Central limit theorem for empirical processes. *Annals of probabilities*, 6:899–929, 1978.

[4] Franke, J. and Mabouba, D. Estimating market risk with neural networks. *Statistic Decision*, 30:63–82, 2006.

[5] Hušková, M. and Koubková, A. Monitoring jump changes in linear models. *J. Statist. Res.*, 39:59–78, 2005.

[6] Kirch, C. and Tadjuidje K., J. . Testing for parameter stability in nonlinear autoregressive models. *preprint*, 2010.

[7] Meyn, S.P. and Tweedie, R.L. . *Markov Chains and Stochastic Stability*. Springer, London, 1993.

[8] Stockis, J.-P., Franke, J., and Tadjuidje K., J. On geometric ergodicity of charme models. *J. Time Series Analysis*, 31:141–152, 2010.

[9] Van der Vaart, A. W. and Wellner, J. A. *Weak convergence and empirical processes*. Springer, New York, 1996.

[10] White, H. Connectionist nonparametric regression: Multilayer feedforward netwoks can learn arbitrary mappings. *Neural Networks*, 3:535–549, 1990.