Learning CHARME models with neural networks

José G. Gómez García[1,3*], Jalal Fadili[2†] and Christophe Chesneau[3†]

[1*]Department MMIP, MIA-Paris, Université Paris-Saclay, AgroParisTech, 16 Rue Claude Bernard, 75005, Paris, France.
[2]CNRS, GREYC, Normandie Université, ENSICAEN, UNICAEN, 6 Bd du Maréchal Juin, 14050, Caen, France.
[3]Department of Mathematics, LMNO, Normandie Université, UNICAEN, Bd du Maréchal Juin, 14032, Caen, France.

*Corresponding author(s). E-mail(s): jose.gomez-garcia@agroparistech.fr;
Contributing authors: Jalal.Fadili@ensicaen.fr; christophe.chesneau@unicaen.fr;
†These authors contributed equally to this work.

Abstract

In this paper, we consider a model called CHARME (Conditional Heteroscedastic Autoregressive Mixture of Experts), a class of generalized mixture of nonlinear (non)parametric AR-ARCH time series. The main objective of this paper is to learn the autoregressive and volatility functions of this model with neural networks (NN). This approach is justified thanks to the universal approximation capacity of neural networks. On the other hand, in order to build the learning theory, it is necessary first to prove the ergodicity of the CHARME model. We therefore show in a general nonparametric framework that under certain Lipschitz-type conditions on the autoregressive and volatility functions, this model is stationary, ergodic and $\boldsymbol{\tau}$-weakly dependent. These conditions are much weaker than those in the existing literature. Moreover, this result forms the theoretical basis for deriving an asymptotic theory of the underlying parametric estimation, which we present for this model in a general parametric framework. Altogether, this allows to develop a learning theory for the NN-based autoregressive and volatility functions of the CHARME model, where strong consistency and asymptotic normality of the considered estimator of the NN weights and biases are guaranteed under weak conditions. Numerical experiments are reported to support our theoretical findings.

Keywords: Nonparametric AR-ARCH, deep neural network, mixture models, ergodicity, stationarity, consistency

1 Introduction

Statistical models such as AR, ARMA, ARCH, GARCH, ARMA-GARCH, etc. are still popular today in time series analysis (see [47, Part III]). These time series are part of the general class of models called conditional heteroscedastic autoregressive nonparametric (CHARN) process, which takes the form

$$X_t = f(X_{t-1}, \dots, X_{t-p}) + g(X_{t-1}, \dots, X_{t-p})\epsilon_t, \tag{1}$$

with unknown functions f, g and independent identically distributed zero-mean innovations ϵ_t. It provides a flexible class of models for many applications such as in econometrics or finance, see [25] and [22]. However, in practice, note that it is not always realistic to assume that the observed process has the same trend function f and the same volatility function g at each time point (this is for instance the case of EEG signals, see [39], and financial datasets, see [31]). In particular, if those functions change slowly over time, local stationarity can be assumed (see [9]), in which there is already a good list of appropriate models. Anyway, estimation procedures for those models are mainly based on applying estimators for stationary processes locally in time which do not work well if the structure of the time series generating mechanism changes more or less abruptly. In this paper, we consider a more general class of (non)parametric models (called CHARME), which adapt to situations where explosive phases may be included. The basics of this new class are presented below.

1.1 The CHARME model

Let $(E, \|\cdot\|)$ be a Banach space, and E endowed with its Borel $\sigma-$algebra $\mathcal{E}$. The product Banach space E^p is naturally endowed with its product $\sigma-$algebra $\mathcal{E}^{\otimes p}$. The conditional heteroscedastic $p-$autoregressive mixture of experts CHARME(p) model, with values in E, is the random process defined by

$$X_t = \sum_{k=1}^{K} \xi_t^{(k)} \left(f_k(X_{t-1}, \dots, X_{t-p}) + g_k(X_{t-1}, \dots, X_{t-p})\epsilon_t \right), \ t \in \mathbb{Z}, \tag{2}$$

where

- for each $k \in [K] := \{1, 2, \ldots, K\}$, $f_k : E^p \longrightarrow E$ and $g_k : E^p \longrightarrow \mathbb{R}$ are the so-called autoregressive and volatility functions, which are $\mathcal{E}^{\otimes p}$ measurable functions.
- $(\epsilon_t)_t$ are $E-$valued independent identically distributed (iid) zero-mean innovations;
- $\xi_t^{(k)} = \mathbb{I}_{\{R_t=k\}}$, with $\mathbb{I}_{\mathcal{C}}$ the characteristic function of $\mathcal{C}$ (takes 1 on $\mathcal{C}$ and 0 otherwise), where $(R_t)_{t\in\mathbb{Z}}$ is an iid sequence with values in a finite set of states $[K]$, which is independent of the innovations $(\epsilon_t)_{t\in\mathbb{Z}}$. In the sequel, we will denote $\pi_k = \mathbb{P}(R_0 = k)$.

Model (2) can be extended to the case where $p = \infty$, called CHARME with infinite memory, denoted by CHARME(∞) for short. For the related setting, we will define the subset of $E^{\mathbb{N}}$ as

$$E^{\infty} := \left\{(x_k)_{k>0} \in E^{\mathbb{N}} : \ x_k = 0 \text{ for } k > N, \text{ for some } N \in \mathbb{N}^*\right\},$$

which will be considered with its product $\sigma-$algebra $\mathcal{E}^{\otimes\mathbb{N}}$.

It is obvious that the model (2) contains the model (1) (corresponding to the case $K = 1$ in (2)). On the other hand, applications of the CHARME model (2) have been directly and indirectly seen in various areas, such as financial analysis [55] (for asset management and risk analysis) and [59] (for predictions of daily probability distributions of S&P returns), hydrology [32] (for the detection of structural changes in hydrological data), electroencephalogram (EEG) signals [38] (for the analysis of EEG recordings from human subjects during sleep), among others.

1.2 Contributions

The objective of this article is to build an estimation theory for the feedforward neural network (NN) based CHARME models. In this regard, we first approach the CHARME model in a general nonparametric context, showing its τ-weak dependence, ergodicity and stationarity under weak conditions. Considering model (2) in a general parametric form: $f_k(\cdot) := f_k(\cdot, \theta_k^0)$, $g_k(\cdot) := g_k(\cdot, \lambda_k^0)$, $k \in [K]$, together with ergodicity and simple conditions, allow us to establish strong consistency for the estimator of the parameters $(\theta_k^0, \lambda_k^0)_{k\in[K]}$ of the model (2), which are the minimizers of a general loss function, not necessarily differentiable. Addressing non-differentiable losses and non iid samples is rather challenging and necessitate to invoke intricate arguments from the calculus of variations (in particular on normal integrands and epi-convergence; see Section 4). Such arguments are not that common in the statistical literature and allow us to investigate new cases that have not been considered before. Additionally, under the same weak assumptions to obtain ergodicity and stationarity together with usual regularity conditions on the autoregressive functions, we prove the asymptotic normality of the conditional least-squares estimator of a simpler CHARME model (i.e., (2) with $g_k \equiv \sigma_k^2$ constant, in order to simplify the presentation of the section).

For the NN-based CHARME(p) model (i.e., the CHARME(p) model with NN-based autoregressive and volatility functions), we specialize the above results that will ensure establish learning consistency guarantees.

Our results are not limited to the case where p is finite. Indeed, we will show that the stationary solution of the CHARME(∞) model can be approximated by the stationary solution of its associated CHARME(p) model (see Remark 3.1 and (7) in Section 3), when p is large enough. Moreover, in Section 6.3, we will argue that CHARME(p) models can be universally approximated by NN-based CHARME(p) models. Altogether, this will provide us with a provably controlled way to learn infinity memory CHARME models with neural networks.

1.3 Relation to prior works

Stockis *et al.* [52] show geometric ergodicity of CHARME(p) models, with $p < \infty$, under certain conditions, including regularity. Specifically, they demand that the iid random variables ϵ_t have a continuous density function, positive everywhere. In contrast, in this paper, the innovations are not supposed to be absolutely continuous and our approach can also be applied, for example, to discrete state space processes. Note also that [52] uses this regularity condition in order to obtain some mixing conditions of $\eta_t = (X_t, \xi_t)_{t\in\mathbb{Z}}$ for deriving asymptotic stability of the model through the results of [42]. However, observe that taking a simple model as the AR(1)-input, solution of the recursion

$$X_t = \frac{1}{2}(X_{t-1} + \epsilon_t), \qquad t \in \mathbb{Z}, \tag{3}$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ iid such that $\mathbb{P}(\epsilon_0 = 0) = \mathbb{P}(\epsilon_0 = 1) = 1/2$, we can see that the assumptions in [52] are not satisfied. In fact, this model is not mixing, see [1]. On the other hand, this model is τ-weakly dependent and satisfies all our assumptions, see [12].

Karmakar *et al.* [31] study in detail the volatility function for the tv(G)ARCH model, which could be extended to more complicated cases. However, their results are obtained under stronger assumptions than ours. For instance, they use locally stationary models (in the sense of Dahlhaus, R. [9]), and as we indicated at the beginning of this section, unlike CHARME models, these locally stationary models do not correctly describe time series with abrupt regime changes. On the other hand, they use fairly standard regularity conditions, such as twice continuously differentiability of the loss function as well as the compactness of the parameter space. Whereas in our work, these regularity conditions (except for the CLT) are not necessary due to normal integrands and epi-convergence arguments.

Though Karmakar *et al.* [31]'s AR-ARCH results are interesting and powerful, their approach is different from ours. They are rather focused on statistical inference (confidence bands and tests) for the tvAR-tvARCH and tv(G)ARCH series. We focus on Regime-Switching processes of AR-ARCH models with constant coefficients (not modelizable with a locally stationary process) and which can be approximated by neural networks, for which consistency of the estimator of parameters (weights and biases) does not need regularity conditions of the loss function.

1.4 Paper organization

The paper is organized as follows: In Section 2 we start with the preliminaries such as the definition and most important properties of τ-weak dependence which characterize our model, and a summary of neural networks. In Section 3 we study the properties of ergodicity and stationarity of the CHARME model in a general nonparametric framework, which will be essential for developing a theory of estimation of the model. To create a smoother transition between neural networks and estimation of the autoregressive (f_k) and volatility (g_k) functions, we have considered f_k and g_k in a general parametric form: $f_k(\cdot) = f(\cdot, \theta_k)$ and $g_k(\cdot) = g_k(\cdot, \lambda_k)$ in Sections 4 and 5. In this way, in Section 4 we provide parameter estimators of the parametric form of (2) and we prove its strong consistency under very weak conditions. Asymptotic normality of the conditional least-squares estimator is established in Section 5, but for a simpler CHARME model (the parametric form of (2) with g_k constant, for each $k \in [K]$) in order to write more simplified hypotheses and make this section more readable. These parametric autoregressive and volatility functions are exactly feedforward neural networks in Section 6, where the parameters θ_k and λ_k are the weights and biases of the networks. Here, we discuss the previous results in the context of NN-based CHARME models and examine the difference between approximation and exact modeling by NNs. Numerical experiments are included in Section 7 and the proofs in Section 8.

2 Preliminaries

Let $(E, \|\cdot\|)$ be a Banach space and $h : E \longrightarrow \mathbb{R}$. We define $\|h\|_\infty = \sup_{x\in E} |h(x)|$ and the Lipschitz constant/modulus of h as

$$\mathrm{Lip}(h) = \sup_{x\neq y\in E} \frac{|h(x) - h(y)|}{\|x - y\|}.$$

For an $E-$valued random variable X defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and $m \geq 1$, we denote by $\|\cdot\|_m$ the $\mathbb{L}^m$-norm, i.e., $\|X\|_m = (\mathbb{E}\|X\|^m)^{1/m}$, where $\mathbb{E}$ denotes the expectation.

2.1 Weak dependence

The appropriate notion of weak dependence for the model (2) was introduced in [12]. It is based on the concept of the coefficient τ defined below.

Definition 2.1 (τ-dependence) Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{M}$ a σ-subalgebra of $\mathcal{A}$ and X a random variable with values in E such that $\|X\|_1 < \infty$. The coefficient τ is defined as

$$\tau(\mathcal{M}, X) = \mathbb{E}\left|\sup\left\{\left|\int_E h(x)\mathbb{P}_{X|\mathcal{M}}(dx) - \int_E h(x)\mathbb{P}_X(dx)\right| : \ h \ \text{ s.t. } \ \mathrm{Lip}(h) \leq 1\right\}\right|.$$

Note that if Y is any random variable with the same distribution as X and independent of $\mathcal{M}$, then

$$\tau(\mathcal{M}, X) \leq \|X - Y\|_1.$$

This is a coupling argument that allows us to easily bound the τ coefficient. See the examples in [12]. On the other hand, if the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is rich enough (which we always assume in the sequel), there exists X^* with the same distribution as X and independent of $\mathcal{M}$ such that $\tau(\mathcal{M}, X) = \|X - X^*\|_1$.

Using the definition of this τ coefficient with the σ-algebra $\mathcal{M}_p = \sigma(X_t, t \leq p)$ and the norm $\|x - y\| = \|x_1 - y_1\| + \cdots + \|x_k - y_k\|$ on E^k, we can assess the dependence between the past of the sequence $(X_t)_{t\in\mathbb{Z}}$ and its future k-tuples through the coefficients

$$\tau_k(r) = \max_{1\leq l\leq k} \frac{1}{l} \sup\{\tau(\mathcal{M}_p, (X_{j_1}, \dots, X_{j_l})) \quad \text{with} \qquad p + r \leq j_1 < \cdots < j_l\}.$$

Finally, denoting $\tau(r) := \tau_\infty(r) = \sup_{k>0} \tau_k(r)$, the time series $(X_t)_{t\in\mathbb{Z}}$ is called τ-*weakly dependent* if its coefficients $\tau(r)$ tend to 0 as r tends to infinity.

2.2 Neural networks

Neural networks produce structured parametric families of functions that have been studied and used for almost 70 years, going back to the late 1940's and the 1950's [27, 48]. An often cited theoretical feature of neural networks, known since the 1980's, is their universal approximation capacity [30], i.e., given any continuous target function f and a target accuracy $\epsilon > 0$, neural networks with enough judiciously chosen parameters give an approximation to the function within an error of size ϵ.

It appears then natural to use this property when it comes to model the functions f_k and g_k, $k \in [K]$, of the process (2).

Definition 2.2 Let $d, L \in \mathbb{N}$. A fully connected feedforward neural network with input dimension d, L layers and activation map $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$, is a collection of weight matrices $\left(W^{(l)}\right)_{l\in[L]}$ and bias vectors $\left(b^{(l)}\right)_{l\in[L]}$, where $W^{(l)} \in \mathbb{R}^{N_l\times N_{l-1}}$ and $b^{(l)} \in \mathbb{R}^{N_l}$, with $N_0 = d$, and $N_l \in \mathbb{N}$ is the number of neurons for layer $l \in [L]$. Let's gather these parameters in the vector

$$\theta = \left((W^{(1)}, b^{(1)}), (W^{(2)}, b^{(2)}), \dots, (W^{(L)}, b^{(L)})\right) \in \bigtimes_{l=1}^{L} \mathbb{R}^{N_l\times N_{l-1}} \times \mathbb{R}^{N_l}.$$

Then, a neural network parametrized by[1] θ produces a function

$$f:(x,\theta)\in\mathbb{R}^d\times\left(\bigtimes_{l=1}^{L}\mathbb{R}^{N_l\times N_{l-1}}\times\mathbb{R}^{N_l}\right)\mapsto f(x,\theta)=x^{(L)}\in\mathbb{R}^{N_L},$$

where x_L results from the following recursion:

$$\begin{cases} x^{(0)} & := x, \\ x^{(l)} & := \varphi(W^{(l)}x^{(l-1)}+b^{(l)}), \qquad \text{for } l=1,\dots,L-1, \\ x^{(L)} & := W^{(L)}x^{(L-1)}+b^{(L)}, \end{cases}$$

where φ acts componentwise, that is, for $y=(y_1,\dots,y_N)\in\mathbb{R}^N$, $\varphi(y)=(\varphi(y_1),\dots,\varphi(y_N))$.

The rectified linear unit (ReLU) is the activation map of preference in many applications, but other examples of activation maps in the literature include the sigmoid, softplus, ramp or other activations [51, Chapter 20.4].

Remark 2.1 Modern machine learning emphasizes the use of deep architectures (as opposed to shallow networks popular in the 1980's-1990's). A few recent works have focused on the advantages of deep versus shallow architectures in neural networks by showing that deep networks can approximate many interesting functions more efficiently, per parameter, than shallow networks (see [11, 26, 57, 61, 62] for a selection of rigorous results). In particular, the work of [11] has shown that neural networks with sufficient depth and appropriate width, possess greater expressivity and approximation power than traditional methods of nonlinear approximation. They also exhibited large classes of functions which can be exactly or efficiently captured by neural networks whereas classical nonlinear methods fall short of the task.

3 Ergodicity and Stationarity of CHARME models

In this section we study the properties of ergodicity and stationarity of the model (2) for the general case, i.e., for the case $p=\infty$, because the case $p<\infty$ is a straightforward corollary. In turn, these properties will be instrumental in establishing statistical inference guarantees.

Theorem 3.1 *Consider the CHARME(∞) model, i.e., (2) with $p=\infty$. Assume that there exist non-negative real sequences $(a_i^{(k)})_{i\geq1,k\in[K]}$ and $(b_i^{(k)})_{i\geq1,k\in[K]}$, such that, for any $x,y\in E^\infty$, and $\forall k\in[K]$,*

$$\begin{aligned} \|f_k(x)-f_k(y)\| &\leq \sum_{i=1}^{\infty} a_i^{(k)}\|x_i-y_i\| \\ \textit{and}\quad |g_k(x)-g_k(y)| &\leq \sum_{i=1}^{\infty} b_i^{(k)}\|x_i-y_i\|. \end{aligned} \tag{4}$$

[1] We intentionally omit the explicit dependence on φ since the latter is chosen once for all.

Denote $A_k = \sum_{i=1}^{\infty} a_i^{(k)}$, $B_k = \sum_{i=1}^{\infty} b_i^{(k)}$ *and*

$$C(m) = 2^{m-1} \sum_{k=1}^{K} \pi_k \left(A_k^m + B_k^m \|\epsilon_0\|_m^m \right).$$

Then, the following statements hold:

(i) if $c := C(1) < 1$, *then there exists a* $\tau-$*weakly dependent strictly stationary solution* $(X_t)_{t\in\mathbb{Z}}$ *of CHARME(*∞*) which belongs to* $\mathbb{L}^1$, *and such that*

$$\tau(r) \leq 2\frac{\mu_1}{1-c} \inf_{1\leq s\leq r} \left(c^{r/s} + \frac{1}{1-c} \sum_{i=s+1}^{\infty} c_i \right) \xrightarrow[r\to\infty]{} 0, \tag{5}$$

where $\mu_1 = \sum_{k=1}^{K} \pi_k \left(\|f_k(0)\| + |g_k(0)|\,\|\epsilon_0\|_1 \right)$ *and* $c_i = \sum_{k=1}^{K} \pi_k \left(a_i^{(k)} + b_i^{(k)} \|\epsilon_0\|_1 \right)$.

(ii) if moreover $C(m) < 1$ *for some* $m > 1$, *then the stationary solution belongs to* $\mathbb{L}^m$.

Corollary 3.1 *Consider the CHARME(p) model (2) and suppose that the inequalities (4) hold (in this case* $a_i^{(k)} = b_i^{(k)} = 0$ *for all* $i > p$ *and all* $k \in [K]$*). Under the notations of Theorem 3.1, if* $c < 1$, *then there exists a* $\tau-$*weakly dependent stationary solution* $(X_t)_{t\in\mathbb{Z}}$ *of CHARME(p) which belongs to* $\mathbb{L}^1$ *and such that* $\tau(r) \leq 2\mu_1(1-c)^{-1}c^{r/p}$ *for* $r \geq p$. *Moreover, if* $C(m) < 1$ *for some* $m > 1$, *then this solution belongs to* $\mathbb{L}^m$.

Remark 3.1 1. Consider the assumptions of Theorem 3.1. The Lipschitz-type assumption (4) entails continuity of $f_k(\cdot)$ and $g_k(\cdot)$, whence we deduce continuity of F as defined in (24). It then follows from [16, Lemma 5.5] and the completeness of $\mathbb{L}^m$, that there exits a measurable function H such that the CHARME(∞) process can be written as $X_t = H(\tilde{\xi}_t, \tilde{\xi}_{t-1}, \ldots)$, where $\tilde{\xi}_t := (\epsilon_t, \xi_t^{(1)}, \ldots, \xi_t^{(K)}) = (\epsilon_t, \xi_t) \in E \times \{e_1, \ldots, e_K\}$, where $e_1, \ldots, e_K$ are the canonical basis vectors for $\mathbb{R}^K$. In other words, the CHARME(∞) process can be represented as a causal Bernoulli shift. Moreover, under these assumptions, $(X_t)_{t\in\mathbb{Z}}$ is the unique causal Bernoulli shift solution to (2) with $p = \infty$. Therefore, the solution $(X_t)_{t\in\mathbb{Z}}$ is automatically an ergodic process. Finally, the ergodic theorem implies the SLLN for this process. This consequence of Theorem 3.1 will be a key to establish strong consistency when it comes to estimating the autoregressive and volatility functions of the CHARME(p) model.

2. Using the arguments in [16], it can be shown that the stationary solution of CHARME (∞) can be approximated by a stationary solution of the CHARME(p) model (2) for some large value of p. In fact, the bounds of the weak dependence coefficients of Doukhan and Wintenberger [16, Theorem 3.1] come from an approximation with Markov chains of order p along with its weak dependence and stationarity properties (see [16, Corollary 3.1]). Indeed, let X_t be the stationary solution of the CHARME(∞) model and let $X_{p,t}$ be the stationary solution of its associated CHARME(p) model, i.e.,

$$X_{p,t} = F(X_{p,t-1}, \ldots, X_{p,t-p}, 0, 0, \ldots;\ \tilde{\xi}_t), \tag{6}$$

where F is defined in (24). Then, [16, Lemma 5.5] gives

$$\mathbb{E}\|X_t - X_{p,t}\| \leq \frac{\mu_1}{(1-c)^2} \sum_{i=p+1}^{\infty} c_i. \tag{7}$$

3. In [52], the authors show that CHARME(p) is geometrically ergodic for $p < \infty$ considering the process $(R_t)_{t\in\mathbb{Z}}$ as a first-order irreducible and aperiodic strictly stationary Markov chain, together with a list of conditions. In particular, they demand that the iid random variables ϵ_t have a continuous density function, positive everywhere. In contrast, in this paper the innovations are not supposed to be absolutely continuous and our approach can also be applied to discrete state space processes. We refer the reader to [13–15, 18–21].

   Additionally, in [52], the geometric ergodicity of $\eta_t = (X_t, \xi_t)$, $t \in \mathbb{Z}$, has been shown in order to obtain some mixing conditions of $(\eta_t)_{t\in\mathbb{Z}}$ for deriving asymptotic stability of the model and, therefore, for formalizing consistency of parameter estimates of shallow-NN-based CHARME models. However, note that by taking the simple AR(1) model defined in (3), we can see that this does not satisfy some of the assumptions in [52]. In fact, the AR(1) process (3) is not mixing, see [1]. It turns out that the main restrictions of the mixing processes are the regularity conditions required for the noise process $(\epsilon_t)_{t\in\mathbb{Z}}$. These regularity conditions, however, are not needed within the framework of $\tau-$dependence. For example, the process (3) is $\tau-$weakly dependent with $\tau(r) \leq 2^{-r}\sqrt{1/6}$; see [12, Application 1].

4 Estimation of CHARME parameters: Consistency

In the sequel, we suppose that the true autoregression and volatility functions have a parametric form: $f_k(\cdot) := f(\cdot, \theta_k^0)$ and $g_k(\cdot) := g_k(\cdot, \lambda_k^0)$, with $k \in [K]$. Here, $f_k : E^p \times \Theta_k \longrightarrow E$ and $g_k : E^p \times \Lambda_k \longrightarrow \mathbb{R}$ are, respectively, $\mathcal{E}^{\otimes p} \times \mathcal{B}(\Theta_k)$– and $\mathcal{E}^{\otimes p} \times \mathcal{B}(\Lambda_k)$–measurable functions, where Θ_k and Λ_k are the respective parameter spaces and $\mathcal{B}(\Theta_k)$ is the Borel field on Θ_k and similarly for Λ_k. We will also denote the space of parameters as the product spaces $\Theta := \bigtimes_{k=1}^{K} \Theta_k$ and $\Lambda := \bigtimes_{k=1}^{K} \Lambda_k$.

Let $(X_t)_{-p+1\leq t\leq n}$[2] be $n + p$ observations of a strictly stationary solution $(X_t)_{t\in\mathbb{Z}}$ of the model (2) (which exists by Theorem 3.1). We assume that the number of states K is known, and that we have access to observations of the hidden iid variables $\big(R_t\big)_{-p+1\leq t\leq n}$, or equivalently, the variables $\big(\xi_t^{(k)}\big)_{-p+1\leq t\leq n, k\in[K]}$.

Remark 4.1 One may wonder how strong these two assumptions are. In general, a careful analysis of the model usually provides interpretation for the number of states K in terms of physical significance or economical meaning. As far as the assumption that $\big(R_t\big)_{-p+1\leq t\leq n}$ are observed is concerned, it is rather common in the literature, see, e.g., [52, 55] for special cases of CHARME. If both K and p

[2] With a slight abuse of notations, we use the same symbol for the observations.

still happen to be unknown, one may appeal to BIC-type model selection criteria to estimate them. Nevertheless, given the additional challenges that this would be bring to the estimators, we leave it to a future work (including other extensions of the model such as removing the iid assumption on $\big(R_t\big)_{-p+1\leq t\leq n}$ or considering K increasing with the number of data).

Our goal now is to design consistent estimators of the parameters

$$(\theta^0,\lambda^0) := (\theta^0_1,\dots,\theta^0_K,\lambda^0_1,\dots,\lambda^0_K)$$

of the CHARME(p) model (2) from observations $\big(X_t\big)_{-p+1\leq t\leq n}$ and $\big(\xi^{(k)}_t\big)_{-p+1\leq t\leq n, k\in[K]}$. This will be achieved through solving the minimization problem

$$(\widehat{\theta}_n,\widehat{\lambda}_n) \in \operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda} Q_n(\theta,\lambda), \text{ where}$$
$$Q_n(\theta,\lambda) := \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{K}\xi^{(k)}_t \ell\big(X_t, f_k(X_{t-1},\dots,X_{t-p},\theta_k), g_k(X_{t-1},\dots,X_{t-p},\lambda_k)\big). \tag{8}$$

Here, $\ell : E\times E\times\mathbb{R}\to\mathbb{R}\cup\{+\infty\}$ is some loss function. Typically, ℓ would satisfy $\ell(u,u,\tau)=0$, $\forall\tau$. Observe that we allow ℓ to be extended-real-valued (i.e., possibly taking value $+\infty$). This will allow to deal equally well with non-classical (and challenging) situations as would be the case if we wanted to include some information/constraints one might have about certain parameters and the relationships between them in the estimation process. Handling extended-real-valued functions when establishing consistency theorems is very challenging which will necessitate more sophisticated arguments.

It will be convenient to define the processes

$$Y_t = (X_{t-p}, X_{t-p+1},\dots,X_t) \quad \text{and} \quad \xi_t = (\xi^{(1)}_t,\dots,\xi^{(K)}_t), \quad t\in\mathbb{Z}.$$

Observations $(X_t)_{-p+1\leq t\leq n}$ yield observations $(Y_t)_{1\leq t\leq n}$. Denote $\{e_1,\dots,e_K\}$ be the set of canonical basis vectors for $\mathbb{R}^K$. Let $(E^{p+1}\times\{e_1,\dots,e_K\}, \mathcal{E}^{\otimes(p+1)}\otimes\Xi, P)$ the (common) probability space on which the random vectors Y_t and ξ_t are defined. We use the shorthand notation

$$h(Y_t,\xi_t,\theta,\lambda) := \sum_{k=1}^{K}\xi^{(k)}_t \ell\big(X_t, f_k(X_{t-1},\dots,X_{t-p},\theta_k), g_k(X_{t-1},\dots,X_{t-p},\lambda_k)\big). \tag{9}$$

Consistency will be established under the following assumptions.
We will denote $\operatorname{ran}_g := \bigcup_{k\in[K]} g_k(E^p,\Lambda_k)\subset\mathbb{R}$; i.e., the union of the ranges of the functions g_k.

(A.1) $\mathcal{E}^{\otimes(p+1)}\otimes\Xi$ is P-complete, namely, a subset of a null set in $\mathcal{E}^{\otimes(p+1)}\otimes\Xi$ also belongs to $\mathcal{E}^{\otimes(p+1)}\otimes\Xi$.

(**A.2**) For each $k \in [K]$, Θ_k and Λ_k are Polish spaces, i.e., a complete, separable, metric spaces.

(**A.3**) For any $k \in [K]$, f_k and g_k are Carathéodory mappings, i.e., $f_k(X_1, \ldots, X_p, \theta_k)$ (resp. $g_k(X_1, \ldots, X_p, \lambda_k)$) is $\mathcal{E}^{\otimes p}$-measurable in $(X_1, \ldots, X_p)$ for each fixed θ_k (resp. λ_k) and continuous in θ_k (resp. λ_k) for each fixed $(X_1, \ldots, X_p)$.

(**A.4**) ℓ is $\mathcal{E} \otimes \mathcal{B}(E) \otimes \mathcal{B}(\mathrm{ran}_g)$-measurable, and for every $u \in E$, $(v, \tau) \in E \times \mathrm{ran}_g \mapsto \ell(u, v, \tau)$ is lower semicontinuous (lsc).

(**A.5**) $\inf(\ell) \geq 0$.

(**A.6**) For each $k \in [K]$ and $t \in [n]$, there exists $\bar{\theta}_k \in \Theta_k$ such that

$$f_k(X_{t-1}, \ldots, X_{t-p}, \bar{\theta}_k) = 0.$$

(**A.7**) There exist non-negative constants C and c, and $\gamma > 0$, such that for all $k \in [K]$ and $t \in [n]$,

$$\inf_{\lambda_k \in \Lambda_k} \ell\big(X_t, 0, g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k)\big) \leq C\|X_t\|^\gamma + c.$$

Before proceeding, some remarks on these assumptions are in order.

Remark 4.2

1. The completeness assumption (**A.1**) is harmless and for technical convenience. Standard techniques can be used to eliminate it.

2. Functions verifying Assumption (**A.4**) are known as *random lsc* or *normal integrands*. The concept of a random lsc function is due to [45], who introduced it in the context of the calculus of variations under the name of normal integrand. Properties of random lsc functions are studied in [46, Chapter 14]. The proof of our consistency theorem will rely on stability properties of the family of random lsc functions under various operations, and on their powerful ergodic properties set forth in the series of papers [34–36]. Unlike other works on the Law of Large Numbers for random lsc functions [2, 4, 29], which postulate iid sampling, only stationarity is needed in our context.

3. Lower-semicontinuity wrt the parameters is a much weaker assumption than those found in the literature. In addition to allowing to handle constraints on the parameters easily (see the discussion after Theorem 4.1), it will also allow for non-smooth activations maps in NN-based learning such as the very popular ReLU. In fact, even continuity is not needed in our context whereas differentiability is an important assumption in existing works; see, e.g., [52, 55].

4. Assumption (**A.5**) can be weakened to lower-boundedness by a negative combination of powers (with appropriate exponents) of the norm. We leave the details to the interested reader.

5. Assumption (**A.6**) is quite natural and is verified in most applications we have in mind (e.g., neural networks).

6. Our proof technique does not really need p to be finite. Thus our result can be extended equally well to the CHARME(∞) model by considering the process Y_t as valued in E^{∞} and assume $\mathcal{E}^{\otimes \mathbb{N}}$-measurability in our assumptions.

Example 1 A prominent example in applications is where the loss function ℓ takes the form

$$\ell(u, v, \tau) = \frac{\|u - v\|^{\gamma}}{|\tau|^{\gamma}}, \quad \gamma > 0.$$

In view of the role played by τ, it is natural to impose the following assumption on g_k:

($\mathbf{A}_g$) $\exists \delta > 0$ such that $\forall k \in [K]$, $\inf_{x_1,\ldots,x_p,\lambda_k} |g_k(x_1, \ldots, x_p, \lambda_k)| \geq \delta$.

Let us show that ℓ complies which assumptions (**A.4**), (**A.5**) and (**A.7**). First, (**A.5**) is obviously verified. As for (**A.7**), we have from ($\mathbf{A}_g$) that

$$\ell\big(X_t, 0, g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k)\big) \leq \delta^{-\gamma} \|X_t\|^{\gamma},$$

whence assumption (**A.7**) holds with $C = \delta^{-\gamma}$ and $c = 0$. It remains to check (**A.4**). Since ($\mathbf{A}_g$) implies that $0 \notin \mathrm{ran}_g = \mathbb{R} \backslash] - \delta, \delta[$, continuity of the norm and ($\mathbf{A}_g$) entails that ℓ, which is the ratio of continuous functions on Borel spaces, is continuous, hence a Borel function.

We are now in position to state our consistency theorem.

Theorem 4.1 *Let $(X_t)_{t\in\mathbb{Z}}$ be a strictly stationary ergodic solution of* (2)*, which exists under the assumptions of Theorem 3.1 with $C(m) < 1$ for some $m \geq 1$. Let $(\widehat{\theta}_n, \widehat{\lambda}_n)$ the estimator defined by* (8)*, and assume that* (**A.1**)*-*(**A.7**) *are verified with $\gamma = m$. Then, the following statements hold:*

(i) each cluster point of $(\widehat{\theta}_n, \widehat{\lambda}_n)_{n\in\mathbb{N}}$ belongs to $\mathrm{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda} \mathbb{E}h(Y, \xi, \theta, \lambda)$ a.s.

(ii) if moreover the sequence $(Q_n)_{n\in\mathbb{N}}$ is equi-coercive, and

$$\mathrm{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda} \mathbb{E}h(Y, \xi, \theta, \lambda) = \{\theta^0, \lambda^0\},$$

then

$$(\widehat{\theta}_n, \widehat{\lambda}_n) \to (\theta^0, \lambda^0) \quad \textit{and} \quad Q_n(\widehat{\theta}_n, \widehat{\lambda}_n) \to \mathbb{E}h(Y, \xi, \theta^0, \lambda^0) \quad \textit{a.s.}$$

Recall that a sequence of functions $(\phi_n)_{n\in\mathbb{N}}$ is equi-coercive if there exists a lsc coercive function ψ such that $\phi_n \geq \psi$, $\forall n \in \mathbb{N}$, see [10, Definition 7.6 and Proposition 7.7]. This entails in particular that the sublevel sets of the functions ϕ_n are compact[3] uniformly in n.

For instance, a sufficient condition to ensure equi-coerciveness in our context is that, for each $k \in [K]$, there exists a $\mathcal{B}(\Theta_k)\otimes\mathcal{B}(\Lambda_k)$-measurable compact subset $\mathcal{C}_k \subset \Theta_k \times \Lambda_k$ such that[4]

$$\mathrm{dom}\,\Big(\ell\big(X_t, f_k(X_{t-1}, \ldots, X_{t-p}, \cdot), g_k(X_{t-1}, \ldots, X_{t-p}, \cdot)\big)\Big) \subset \mathcal{C}_k, \quad \forall t \in [n].$$

[3] We here specialized [10, Definition 7.6] to metric spaces (see(**A.2**)) where compactness implies closeness and countable compactness.

[4] Observe that accounting for this constraint does not compromise assumption (**A.4**) thanks to compactness of $\mathcal{C}_k$.

Indeed, it is immediate to see that such a condition implies that

$$\operatorname{dom}(Q_n) \subset \bigtimes_{k=1}^{K} \mathcal{C}_k,$$

which is then a compact set.

The sets $\mathcal{C}_k$ can be used to impose some prior constraints on the parameters (θ_k, λ_k) which might follow from certain physical, economic or mathematical considerations. For instance, these parameters can be constrained to comply with the strict stationarity assumption in Theorem 3.1. Other constraints can be also used to promote some desirable properties such robustness and generalization for the case of neural networks (see Section 6 for further discussion). In general, to account for constraints, one sets

$$\ell\big(X_t, f_k(X_{t-1}, \dots, X_{t-p}, \theta_k), g_k(X_{t-1}, \dots, X_{t-p}, \lambda_k)\big) = \\ \widetilde{\ell}\big(X_t, f_k(X_{t-1}, \dots, X_{t-p}, \theta_k), g_k(X_{t-1}, \dots, X_{t-p}, \lambda_k)\big) + \iota_{\mathcal{C}_k}(\theta_k, \lambda_k), \quad \forall k \in [K],$$

where $\widetilde{\ell}$ is a full-domain loss verifying **(A.4)**, **(A.5)** and **(A.7)**, and $\iota_{\mathcal{C}_k}$ is the indicator function of $\mathcal{C}_k$, taking 0 on $\mathcal{C}_k$ and $+\infty$ otherwise. By assumptions on $\mathcal{C}_k$, $\iota_{\mathcal{C}_k}$ is $\mathcal{B}(\Theta_k) \otimes \mathcal{B}(\Lambda_k)$-measurable and lsc, and thus ℓ inherits **(A.4)** from $\widetilde{\ell}$. **(A.5)** is trivially verified, and for **(A.6)** to hold, it is necessary and sufficient that for each $k \in [K]$ and $t \in [n]$, $f_k(X_{t-1}, \dots, X_{t-p}, \cdot)^{-1}(0) \times \Lambda_k \cap \mathcal{C}_k \neq \emptyset$.

We finally stress that the constraints above do not need to be separable, as soon as one takes $h(Y_t, \xi_t, \theta, \lambda)$ as

$$h(Y_t, \xi_t, \theta, \lambda) = \sum_{k=1}^{K} \xi_t^{(k)} \widetilde{\ell}\big(X_t, f_k(X_{t-1}, \dots, X_{t-p}, \theta_k), g_k(X_{t-1}, \dots, X_{t-p}, \lambda_k)\big) \\ + \iota_{\mathcal{C}}(\theta, \lambda),$$

where $\mathcal{C} \subset \Theta \times \Lambda$ is a $\mathcal{B}(\Theta) \otimes \mathcal{B}(\Lambda)$-measurable compact set. Thus, depending on the application at hand, our reasoning above can be extended to more complicated situations.

5 Estimation of CHARME parameters: Asymptotic normality

To establish asymptotic normality, we need to restrict ourselves to a finite-dimensional framework where $E = \mathbb{R}^d$ and $\Theta_k = \mathbb{R}^{d_k}$. Throughout this section, $\|\cdot\|$ denotes the standard Euclidean norm and the corresponding (Euclidean) space is to be understood from the context.

In this section, we consider the following constant-volatility special case of the model in (2):

$$X_t = \sum_{k=1}^{K} \xi_t^{(k)} \left(f_k(X_{t-1}, \ldots, X_{t-p}, \theta_k^0) + \epsilon_{k,t} \right), \quad t \in \mathbb{Z}, \tag{10}$$

where $\epsilon_{k,t} = \sigma_k^2 \epsilon_t$, with $\sigma_k^2 > 0$ constant for each $k \in [K]$. This is to write more simplified assumptions and make the section more readable.
We then specialize the estimator in (8) to (10) and the quadratic loss, which now reads

$$\widehat{\theta}_n \in \mathrm{Argmin}_{\theta\in\Theta} \left\{ Q_n(\theta) := \frac{1}{n} \sum_{t=1}^{n} \sum_{k=1}^{K} \xi_t^{(k)} \left\| X_t - f_k(X_{t-1}, \ldots, X_{t-p}, \theta_k) \right\|^2 \right\}. \tag{11}$$

This corresponds to the conditional least-squares method. We focus on this simple loss although our results hereafter can be extended easily, through tedious calculations, to any loss ℓ which is three-times continuously differentiable wrt its second argument.

For a three-times continuously differentiable mapping $h : \nu \in \mathbb{R}^{d_k} \mapsto h(\nu) \in \mathbb{R}^d$, we will denote $\partial h / \partial \nu_i(\mu) \in \mathbb{R}^d$ the derivative of h wrt to the i-th entry of ν evaluated at $\mu \in \mathbb{R}^{d_k}$, and $J[h](\mu) = (\partial h/\partial \nu_1(\mu) \ldots \partial h / \partial \nu_{d_k}(\mu))$ the Jacobian of h. Similarly the second and third order (mixed) derivatives are denoted as $\partial^2 h/(\partial \nu_i \partial \nu_j)(\mu)$ and $\partial^3 h/(\partial \nu_i \partial \nu_j \partial \nu_l)(\mu)$, respectively. For a differentiable scalar-valued function on an Euclidean space, ∇ will denote its gradient operator (the vector of its partial derivatives).

From Example 1, Theorem 4.1 applies, hence showing consistency of the estimator (11). On the other hand, to establish asymptotic normality of this estimator, we will invoke [56, Theorem 3.2.23 or 3.2.24] (which are in turn due to [33]). This requires to impose the following more stringent regularity assumptions:

(B.1) For each $k \in [K]$, the function $\theta_k \in \Theta_k \mapsto f_k(X_p, \ldots, X_1, \theta_k)$ is three-times continuously differentiable almost everywhere in an open neighborhood $\mathcal{V}$ of $\theta^0 = (\theta_1^0, \ldots, \theta_K^0)$.

(B.2) For all $k \in [K]$ and all $i, j \in [d_k]$,

$$\mathbb{E} \left\| \frac{\partial f_k(X_p, \ldots, X_1, \theta_k^0)}{\partial \theta_{k,i}} \right\|^2 < \infty \quad \text{and} \quad \mathbb{E} \left\| \frac{\partial^2 f_k(X_p, \ldots, X_1, \theta_k^0)}{\partial \theta_{k,j} \partial \theta_{k,i}} \right\|^2 < \infty.$$

(B.3) The vectors $\{\partial f_k(X_p, \ldots, X_1, \theta_k^0)/\partial \theta_{k,i}\}_{i\in[d_k],k\in[K]}$, are linearly independent in the sense that if $(a_{k,i})_{i\in d_k, k\in[K]}$ are arbitrary real numbers such that

$$\mathbb{E} \left\| \sum_{k=1}^{K} \sum_{i=1}^{d_k} a_{k,i} \frac{\partial f_k(X_p, \ldots, X_1, \theta_k^0)}{\partial \theta_{k,i}} \right\|^2 = 0,$$

then $a_{k,i} = 0$ for all $i \in [d_k]$ and all $k \in [K]$.

(B.4) For $k \in [K]$ and $i, j, r \in [d_k]$

$$G_k^{ijr} := \mathbb{E}\left|\left(\frac{\partial f_k(X_p, \ldots, X_1, \theta_k^0)}{\partial \theta_{k,i}}\right)^\top \frac{\partial^2 f_k(X_p, \ldots, X_1, \theta_k^0)}{\partial \theta_{k,j} \partial \theta_{k,r}}\right| < \infty$$

and

$$H_k^{ijr} := \mathbb{E}\left|\left(X_{p+1} - f_k(X_p, \ldots, X_1, \theta_k^0)\right)^\top \frac{\partial^3 f_k(X_p, \ldots, X_1, \theta_k^0)}{\partial \theta_{k,i} \partial \theta_{k,j} \partial \theta_{k,r}}\right| < \infty.$$

(B.5) For all $k \in [K]$ and all $i, j \in [d_k]$, $|W_{k,ij}| < \infty$, where

$$W_{k,ij} = \mathbb{E}\Bigg[\xi_t^{(k)} \left(X_{p+1} - f_k(X_p, \ldots, X_1, \theta_k^0)\right)^\top \frac{\partial f_k(X_p, \ldots, X_1, \theta_k^0)}{\partial \theta_{k,i}} \cdot \left(X_{p+1} - f_k(X_p, \ldots, X_1, \theta_k^0)\right)^\top \frac{\partial f_k(X_p, \ldots, X_1, \theta_k^0)}{\partial \theta_{k,j}}\Bigg].$$

Let us denote by $W = (W_{kl})_{1 \le k, l \le K}$ the block–diagonal matrix defined by the sub-matrices

$$W_{kl} = \begin{cases} 0_{d_k \times d_l} & \text{if } k \neq l \\ (W_{k,ij})_{1 \le i, j, \le d_k} & \text{if } k = l. \end{cases} \tag{12}$$

We are now in shape to formalize our asymptotic normality result.

Theorem 5.1 *Let $(X_t)_{t \in \mathbb{Z}}$ be a strictly stationary ergodic solution of (10) with $\mathbb{E}\|X_t\|^2 < \infty$, which exists under the assumptions of Theorem 3.1 with $C(m) < 1$ for $m = 2$. Suppose that **(B.1)**-**(B.5)** hold. Then there exists a sequence of estimators $\widehat{\theta}_n$ such that*

$$\widehat{\theta}_n \to \theta^0 \quad a.s.$$

and for any $\varepsilon > 0$, there exists N large enough and an event with probability at least $1 - \varepsilon$ on which, for all $n > N$, $\nabla Q_n(\widehat{\theta}_n) = 0$, and Q_n attains a relative minimum at $\widehat{\theta}_n$. Furthermore,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^0\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, V^{-1} W V^{-1}\right),$$

as $n \to \infty$, where $V = (V_{kl})_{1 \le k, l \le K}$ is the block-diagonal matrix defined by the sub-matrices

$$V_{kl} = \begin{cases} 0_{d_k \times d_l} & \text{if } k \neq l \\ \pi_k \; \mathbb{E}\left[\left(J[f_k(X_p, \ldots, X_1, \cdot)](\theta_k^0)\right)^\top J[f_k(X_p, \ldots, X_1, \cdot)](\theta_k^0)\right] & \text{if } k = l. \end{cases} \tag{13}$$

Observe that the covariance matrix $V^{-1} W V^{-1}$ is also block-diagonal with diagonal blocks $V_{kk}^{-1} W_{kk} V_{kk}^{-1}$.

6 Learning CHARME models with Neural Networks

In this section, we apply our results to the case where $E = \mathbb{R}^d$ and each of the functions f_k and g_k in the CHARME(p) model (2) is exactly modeled by a feedforward neural network (see Section 2.2), we call this NN-based CHARME(p) model. More precisely, given an activation map φ, and for each $k \in [K]$, f_k and g_k are feedforward neural networks according to Definition 2.2, parameterized by weights and biases given respectively by

$$\begin{aligned} \theta_k &= \left((W_k^{(1)}, b_k^{(1)}), \ldots, (W_k^{(L_k)}, b_k^{(L_k)})\right) \\ \lambda_k &= \left((\bar{W}_k^{(1)}, \bar{b}_k^{(1)}), \ldots, (\bar{W}_k^{(\bar{L}_k)}, \bar{b}_k^{(\bar{L}_k)})\right). \end{aligned} \tag{14}$$

For each $k \in [K]$, we have:

- for each layer $l \in [L_k]$ of the k-th NN modeling f_k, $W_k^{(l)} = (w_{k,ij}^{(l)})_{(i,j)\in[N_{k,l}]\times[N_{k,l-1}]}$ and $b_k^{(l)} = (\beta_{k,1}^{(l)}, \ldots, \beta_{k,N_{k,l}}^{(l)})^\top$ are respectively the matrix of weights and vector of biases;
- for each layer $l \in [\bar{L}_k]$ of the k-th NN modeling g_k, $\bar{W}_k^{(l)} = (\bar{w}_{k,ij}^{(l)})_{(i,j)\in[\bar{N}_{k,l}]\times[\bar{N}_{k,l-1}]}$ and $\bar{b}_k^{(l)} = (\bar{\beta}_{k,1}^{(l)}, \ldots, \bar{\beta}_{k,\bar{N}_{k,l}}^{(l)})^\top$ are respectively the matrix of weights and vector of biases;
- $N_{k,0} = \bar{N}_{k,0} = d \cdot p$, $N_{k,L} = d$ and $\bar{N}_{k,L} = 1$.

We throughout make the standard assumption that the activation map φ is Lipschitz continuous[5].

6.1 Ergodicity and stationarity

Considering the notations of Theorem 3.1, let $x^\top = (x_1, \ldots, x_{dp}) \in \mathbb{R}^{dp}$ and $y^\top = (y_1, \ldots, y_{dp}) \in \mathbb{R}^{dp}$. Split the matrix $W_k^{(1)}$ into p column blocks $W_{k,i}^{(1)} \in \mathbb{R}^{N_{k,1}\times d}$ such that $W_k^{(1)} = \left(W_{k,1}^{(1)}\ W_{k,2}^{(1)} \ldots W_{k,p}^{(1)}\right)$. It is easy to see that

$$\begin{aligned} &\|f_k(x,\theta_k) - f_k(y,\theta_k)\| \\ &\le \mathrm{Lip}(\varphi)\mathrm{Lip}\left((W_k^{(L_k)} \cdot - b_k^{(L_k)}) \circ \cdots \circ \varphi \circ (W_k^{(2)} \cdot - b_k^{(2)})\right) \left\|W_k^{(1)}(x-y)\right\| \\ &= \mathrm{Lip}(\varphi)\mathrm{Lip}\left((W_k^{(L_k)} \cdot - b_k^{(L_k)}) \circ \cdots \circ \varphi \circ (W_k^{(2)} \cdot - b_k^{(2)})\right) \left\|\sum_{i=1}^{p} W_{k,i}^{(1)}(x_i - y_i)\right\| \\ &\le \mathrm{Lip}(\varphi)\mathrm{Lip}\left((W_k^{(L_k)} \cdot - b_k^{(L_k)}) \circ \cdots \circ \varphi \circ (W_k^{(2)} \cdot - b_k^{(2)})\right) \sum_{i=1}^{p} \left|\left|\left|W_{k,i}^{(1)}\right|\right|\right| \|x_i - y_i\|, \end{aligned}$$

where $|||\cdot|||$ stands for the spectral norm. Similarly, we have

[5] Actually the Lipschitz constant is even 1 in general, *e.g.*, ReLU, Leaky ReLU, SoftPlus, Tanh, Sigmoid, ArcTan or Softsign.

$$
\begin{aligned}
&|g_k(x,\lambda_k) - g_k(y,\lambda_k)| \\
\leq\ & \mathrm{Lip}(\varphi)\mathrm{Lip}\left((\bar{W}_k^{(\bar{L}_k)}\cdot - \bar{b}_k^{(\bar{L}_k)})\circ\cdots\circ\varphi\circ(\bar{W}_k^{(2)}\cdot - \bar{b}_k^{(2)})\right)\sum_{i=1}^{p}\left|\left|\left|\bar{W}_{k,i}^{(1)}\right|\right|\right| \|x_i - y_i\|.
\end{aligned}
$$

Identifying with (4), we may take the above bounds as estimates for A_k and B_k, i.e.,

$$
A_k = \mathrm{Lip}(\varphi)\mathrm{Lip}\left((W_k^{(L_k)}\cdot - b_k^{(L_k)})\circ\cdots\circ\varphi\circ(W_k^{(2)}\cdot - b_k^{(2)})\right)\sum_{i=1}^{p}\left|\left|\left|W_{k,i}^{(1)}\right|\right|\right|
$$

and

$$
B_k = \mathrm{Lip}(\varphi)\mathrm{Lip}\left((\bar{W}_k^{(\bar{L}_k)}\cdot - \bar{b}_k^{(\bar{L}_k)})\circ\cdots\circ\varphi\circ(\bar{W}_k^{(2)}\cdot - \bar{b}_k^{(2)})\right)\sum_{i=1}^{p}\left|\left|\left|\bar{W}_{k,i}^{(1)}\right|\right|\right|. \tag{15}
$$

Therefore, if $C(m) = 2^{m-1}\sum_{k=1}^{K}\pi_k\left(A_k^m + B_k^m\|\epsilon_0\|_m^m\right) < 1$ for some $m \geq 1$, there exists a stationary solution of the NN-based CHARME(p) model such that the coefficient $\tau(r) \leq M\left(C(1)\right)^{r/p}$ for $r > p$ and some $M > 0$.

Remark 6.1 The expression of $C(m)$ and the corresponding condition $C(m) < 1$ is the crux of the stability of our model. Thus, checking this condition in practice, as for the case of neural networks with A_k and B_k given by (15), is key. This in turn relies on having a good estimate of the Lipschitz constant of the neural network[6] which is captured in the first part of these expressions. It is known however that computing exactly this Lipschitz constant, even for two layer neural networks, is a NP-hard problem, [50, Theorem 2].

A simple upper-bound is given in [54], i.e.,

$$
\mathrm{Lip}\left((W_k^{(L_k)}\cdot - b_k^{(L_k)})\circ\cdots\circ\varphi\circ(W_k^{(2)}\cdot - b_k^{(2)})\right) \leq \mathrm{Lip}(\varphi)^{L_k - 1}\prod_{l=2}^{L_k}\left|\left|\left|W_k^{(l)}\right|\right|\right|, \tag{16}
$$

and this bound can be computed efficiently with a forward pass on the computational graph. However, the bound (16) depends exponentially on the numbers of layers, L_k, and can provide very pessimistic estimates with a gap in the upper-bound that is in general off by factors or orders of magnitude especially as L_k increases; see the discussion in [50]. In turn, such a crude bound may harm the condition $C(m) < 1$ when L_k becomes large. This gap can be explained by the fact that for differentiable activations, with the chain rule[7], the equality in (16) can only be attained if the activation Jacobian at each layer maps the left singular vectors of $W_k^{(l)}$ to the right singular vectors of $W_k^{(l+1)}$. But these Jacobians being diagonal, this is unlikely to happen causing misaligned singular vectors. Starting from this observation, and using Rademacher's theorem together with the chain rule for differentiable activation maps, a much better bound is proposed in [50, Theorem 3]. This computational burden to get this bound lies in computing the SVD of the weight matrices and solving a

[6] Excluding the first layer.

[7] The reasoning is only valid for differentiable activation maps unlike what is done in [50], and thus excludes the ReLU; see [6] for a thorough justification on the chain rule for neural networks.

maximization problem in each layer. The latter is itelf given an explicit estimate for large number of neurones in [50, Lemma 2].

6.2 Learning guarantees

6.2.1 Consistency

We will now explain the consistency of the weights and biases estimator of the NN-based CHARME model.

Proposition 6.1 *Let $(X_t)_{t\in\mathbb{Z}}$ be a strictly stationary ergodic solution of the NN-based CHARME(p) model, which exists if $C(m) = 2^{m-1}\sum_{k=1}^{K}\pi_k\left(A_k^m + B_k^m\|\epsilon_0\|_m^m\right) < 1$ for some $m \geq 1$, with A_k and B_k defined in (15). Assume that the loss function ℓ is defined as in Exemple 1 with $\gamma = m$ and the NN-based volatility functions g_k satisfy* **(A***$_g$***)**. *Then, the estimator $(\widehat{\theta}_n, \widehat{\lambda}_n)$ defined by (8) verifies Theorem 4.1(i). Moreover, if there exists a lsc coercive function $\tilde{Q}$ such that $Q_n \geq \tilde{Q}$, $\forall n \in \mathbb{N}$, and*

$$\operatorname{Argmin}_{(\theta,\lambda)\in\Theta\times\Lambda}\mathbb{E}h(Y,\xi,\theta,\lambda) = \{\theta^0, \lambda^0\},$$

with h defined in (9), then

$$(\widehat{\theta}_n, \widehat{\lambda}_n) \to (\theta^0, \lambda^0) \quad \textit{and} \quad Q_n(\widehat{\theta}_n, \widehat{\lambda}_n) \to \mathbb{E}h(Y,\xi,\theta^0,\lambda^0) \quad \textit{a.s.}$$

Proof: It suffices to invoke the consistency result of Theorem 4.1, that is to say we need to check that f_k and g_k verify the corresponding assumptions. As $E = \mathbb{R}^d$, **(A.1)** is obviously verified. Similarly it is obvious that the Euclidean spaces of parameters Θ_k and Λ_k obey **(A.2)**. As for **(A.3)**, it is also fulfilled thanks to obvious continuity properties of NN functions, defined as composition of affine and Lipschitz continuous mappings. **(A.6)** is obviously verified, for instance, by zeroing both the weight matrix and bias vector at any same layer (a fortiori, this is true for $\theta_k = 0$). As we have assumed that the NN-bases volatilty functions g_k satisfy **(A**$_g$**)**, from Exemple 1, ℓ complies with assumptions **(A.4)**, **(A.5)** and **(A.7)**.

Thus, since there exists a stationary solution of the NN-based CHARME(p) model under the condition of the previous section, the statement of Theorem 4.1(i) applies to the estimator (8) of the NN parameters. Since we have assumed the assumptions of Theorem 4.1(ii), from this it follows that the estimator (8) of the NN parameters is (strongly) consistent.

Remark 6.2

1. When the volatility functions g_k are not (non-zero) constant, **(A**$_g$**)** is verified if for example φ is positive-valued (as for the ReLU) and for any $k \in [K]$, the weights $\bar{W}_k^{(\bar{L}_k)}$ of the last layer are non-negative and the bias $\bar{b}_k^{(\bar{L}_k)} \geq \delta$ for some $\delta > 0$.
2. To be able to apply Theorem 4.1(ii), we need some equi-coerciveness and uniqueness of the true parameters (θ^0, λ^0). First, it is important to note that neural

networks are often non-identifiable models, which means that different parameters can represent the same function, or equivalently, $f_k(\cdot,\theta_k) = f_k(\cdot,\theta'_k) \not\Rightarrow \theta_k = \theta'_k$. In fact there are invariances in the NN parametrization which induce ambiguities in the solutions of the estimation problem (8). Clearly, this is a non-convex problem which may not have a global minimizer, not to mention uniqueness of the latter, even with the population risk $\mathbb{E}h(Y,\xi,\cdot,\cdot)$ if the weights and biases are allowed to vary freely over the parameters space [8]. Clearly, there is a need to appropriately constraining the weights and biases to get the neessary compactness in our case.

While there is empirical evidence that suggests that when the size of the network is large enough and ReLU non-linearities are used all local minima could be global, there is currently no complete rigorous theory that provides a precise mathematical explanation for these observed phenomena. This is the subject of intense research activity which goes beyond the scope of this paper; see the review paper [58]. A few sufficient deterministic conditions for the existence of global minimizers of (8)[9] can be found in [24, 64]. In [24], it is shown that for certain network architectures with positively homogeneous activations and regularizations, any sparse local minimizer is a global one. The work in [64] deals with general architectures but with smooth activations but no regularization, and delivers conditions under which any critical point is a global minimizer.

Regularizing a neural network by constraining its Lipschitz constant has been proven an effective and successful way to ensure good stability and generalization properties, see, e.g., [5, 8, 40, 43, 44, 60, 63]. In our context, from Section 6.1, this amounts to imposing a constraint of the form

$$\mathrm{Lip}\left((W_k^{(L_k)}\cdot -b_k^{(L_k)}) \circ \cdots \circ \varphi \circ (W_k^{(2)}\cdot -b_k^{(2)}) \circ \varphi \circ (W_k^{(1)}\cdot -b_k^{(1)})\right) \le L,$$

where $L > 0$. As discussed in Remark 6.1, even computing this bound is hard not to mention a constraint based on it. Many authors, e.g., [43, 63] and others, use the simplest strategy that consists in constraining each layer of the network to be Lipschitz, i.e., $\left|\left|\left|W_k^{(l)}\right|\right|\right| \le L^{1/L_k}$, where we used the bound (16) and that the activation maps are also 1-Lipschitz. In [44], the authors imposed an even cruder bound by constraining group norms of the weights. All these bounds define a compact constraint, whose radius L can be chosen such that it satisfies $C(m) < 1$ for $m \ge 1$ known.

To summarize, if (8) is solved with ℓ and compact constraint sets $\mathcal{C}_k$ (with appropriate diameter), or more generally any lsc coercive regularizers, see the discussion after Theorem 4.1, then equi-coerciveness holds true.

6.2.2 Asymptotic normality

We now turn to asymptotic normality of the estimator (11) for the CHARME(p) model (10), where f_k is neutral network.

[8] This is the case for rescaling when the activation is positively homogeneous, in which case multiplying one layer of a global minimizer by a positive constant and dividing another layer by the same constant produces a pair of different global minimizers

[9] More precisely, in all the works cited here, their framework amounts to considering g_k as a constant and ℓ as quadratic in our setting.

Proposition 6.2 *Let $(X_t)_{t\in\mathbb{Z}}$ be a strictly stationnary ergodic solution of the NN-based CHARME(p) model (10) such that $C(m) = 2^{m-1}\sum_{k=1}^K \pi_k A_k^m < 1$ for $m \in \{2,4,6\}$ and with A_k defined as in (15). Assume that the activation function of the NN is three-times continuously differentiable with bounded derivatives*[10] *and the NN-based autoregressive functions f_k satify Assumption (**B.3**). Then, for this particular case, the conclusions of Theorem 5.1 hold true, i.e., strong consistency and asymptotic normality of the estimator (11) of the NN-parameters of the CHARME(p) model (10) are obtained.*

Proof: We need to check the assumptions of Theorem 5.1. Indeed, as the activation map of the NN is three-times continuously differentiable with bounded derivatives, this entail that for all $k \in [K]$, $\theta_k \mapsto f_k(X_p,\ldots,X_1,\theta_k)$ is almost surely three-times continuously differentiable at any $\theta_k \in \Theta_k$, i.e., (**B.1**) holds. On the other hand, in view of the derivatives of f_k in (A1) (see Section A), boundedness of the derivatives of φ and stationarity, it is not difficult to check that

$$\mathbb{E}\left\|\frac{\partial f_k(X_p,\ldots,X_1,\theta_k^0)}{\partial w_{k,ij}^{(l)}}\right\|^2 = O(\mathbb{E}\left\|X_t\right\|^2),$$

$$\mathbb{E}\left\|\frac{\partial^2 f_k(X_p,\ldots,X_1,\theta_k^0)}{\partial w_{k,ij}^{(l)}\partial w_{k,i'j'}^{(l)}}\right\|^2 = O(\max_{s\in\{2,4\}}(\mathbb{E}\left\|X_t\right\|^s)),$$

$$\mathbb{E}\left\|\frac{\partial^3 f_k(X_p,\ldots,X_1,\theta_k^0)}{\partial w_{k,ij}^{(l)}\partial w_{k,i'j'}^{(l)}\partial w_{k,i''j''}^{(l)}}\right\|^2 = O(\max_{s\in\{2,4,6\}}(\mathbb{E}\left\|X_t\right\|^s)).$$

The derivatives wrt biases $\beta_{k,i}^{(l)}$ as given in (A2) are bounded in view of boundedness of the derivative of φ. $C(m) < 1$, for $m \in \{2.4.6\}$, implies that $\max_{s\in\{2,4,6\}}(\mathbb{E}\left\|X_t\right\|^s) < \infty$, whence conditions (**B.2**), (**B.4**), and (**B.5**) hold. Assumption (**B.3**) is assumed to be verified, which concludes the proof.

Remark 6.3 Assumption (**B.3**) captures the fact that θ^0 is a strict local minimizer of (11), which is in turn closely related to our discussion on uniqueness in the previous section. Given the complexity of this problem, we have assumed that it holds in the statement of Proposition 6.2 because it is not the objective of this paper.

6.3 Approximation vs exact modeling by neural networks

Until now, in this section we have assumed that the autoregressive and volatility functions f_k are g_k are *exactly* modeled by feedforward NNs with

[10] This is the case for softplus, smoothed ReLU, sigmoid, etc.

finitely many neurons. A natural question we ask is: what are the consequences if the NN architecture (depth and width) is such that it provides only ε-approximations to f_k and g_k ?

To settle this question, let X_t be the CHARME process given in (2), and $\widetilde{X}_t$ be the CHARME process defined by the same innovations and hidden process $(R_t)_{t\in\mathbb{Z}}$ but with parametric functions $\widetilde{f}_k$ and $\widetilde{g}_k$, i.e.,

$$\widetilde{X}_t = \sum_{k=1}^{K} \xi_t^{(k)} \left(\widetilde{f}_k(\widetilde{X}_{t-1}, \dots, \widetilde{X}_{t-p}, \widetilde{\theta}_k) + \widetilde{g}_k(\widetilde{X}_{t-1}, \dots, \widetilde{X}_{t-p}, \widetilde{\lambda}_k)\epsilon_t \right), \quad t \in \mathbb{Z}. \tag{17}$$

The functions $\widetilde{f}_k$ and $\widetilde{g}_k$ are supposed to be two neural networks providing approximations to f_k and g_k. Denote the approximation accuracy as

$$\begin{aligned}\varepsilon_k := \sup_{(x_1,\dots,x_p)\in E^p} \Big(&\|\widetilde{f}_k(x_1, \dots, x_p, \widetilde{\theta}_k) - f_k(x_1, \dots, x_p)\|, \\ &\Big|\widetilde{g}_k(x_1, \dots, x_p, \widetilde{\lambda}_k) - g_k(x_1, \dots, x_p)\Big|\Big). \end{aligned} \tag{18}$$

To compare the two processes, it is natural to assume that the functions $(\widetilde{f}_k, \widetilde{g}_k)_{k\in\mathbb{N}}$ verify the assumptions of Theorem 3.1 so that $\big(\widetilde{X}_t\big)_{t\in\mathbb{Z}}$ is a strictly stationary solution of (17). Thus, $\forall t \in \mathbb{Z}$, we have

$$\begin{aligned}\|\widetilde{X}_t - X_t\| = \Bigg\| \sum_{k=1}^{K} \xi_t^{(k)} \Bigg(&\Big(\widetilde{f}_k(\widetilde{X}_{t-1}, \dots, \widetilde{X}_{t-p}, \widetilde{\theta}_k) - f_k(\widetilde{X}_{t-1}, \dots, \widetilde{X}_{t-p})\Big) \\ &+ \Big(\widetilde{g}_k(\widetilde{X}_{t-1}, \dots, \widetilde{X}_{t-p}, \widetilde{\lambda}_k) - g_k(\widetilde{X}_{t-1}, \dots, \widetilde{X}_{t-p})\Big)\epsilon_t \\ &+ \Big(f_k(\widetilde{X}_{t-1}, \dots, \widetilde{X}_{t-p}) - f_k(X_{t-1}, \dots, X_{t-p})\Big) \\ &+ \Big(g_k(\widetilde{X}_{t-1}, \dots, \widetilde{X}_{t-p}) - g_k(X_{t-1}, \dots, X_{t-p})\Big)\epsilon_t \Bigg)\Bigg\| \\ \le \sum_{k=1}^{K} \xi_t^{(k)} &\left(\varepsilon_k(1 + \|\epsilon_t\|) + \sum_{i=1}^{p} \Big(a_i^{(k)} + b_i^{(k)}\|\epsilon_t\|\Big) \|\widetilde{X}_{t-i} - X_{t-i}\| \right).\end{aligned}$$

Taking expectations in both sides and thanks to stationarity of both processes, and by assumptions on ϵ_t and $\xi_t^{(k)}$, we get

$$\begin{aligned}\mathbb{E}\|\widetilde{X}_t - X_t\| &\le (1 + \mathbb{E}\|\epsilon_0\|) \sum_{k=1}^{K} \pi_k\varepsilon_k + \mathbb{E}\|\widetilde{X}_t - X_t\| \left(\sum_{k=1}^{K} \pi_k \left(A_k + B_k\mathbb{E}\|\epsilon_0\|\right) \right) \\ &\le (1 + \mathbb{E}\|\epsilon_0\|) \sum_{k=1}^{K} \pi_k\varepsilon_k + \mathbb{E}\|\widetilde{X}_t - X_t\| \left(\sum_{k=1}^{K} \pi_k \left(A_k + B_k\mathbb{E}\|\epsilon_0\|\right)^m \right)^{1/m}\end{aligned}$$

$$\leq (1+\mathbb{E}\|\epsilon_0\|)\sum_{k=1}^{K}\pi_k\varepsilon_k + \mathbb{E}\|\widetilde{X}_t - X_t\|\left(\sum_{k=1}^{K}\pi_k 2^{m-1}\left(A_k^m + B_k^m\|\epsilon_0\|_m^m\right)\right)^{1/m}$$

$$\leq (1+\mathbb{E}\|\epsilon_0\|)\sum_{k=1}^{K}\pi_k\varepsilon_k + \mathbb{E}\|\widetilde{X}_t - X_t\|C(m)^{1/m},$$

where we have used that $m \geq 1$ in the second line and Jensen's inequality in the third. Since by assumption $C(m) < 1$ for some $m \geq 1$, see Theorem 3.1, we get that

$$\mathbb{E}\|\widetilde{X}_t - X_t\| \leq \frac{(1+\mathbb{E}\|\epsilon_0\|)\sum_{k=1}^{K}\pi_k\varepsilon_k}{1-C(m)^{1/m}}. \tag{19}$$

In a nutshell, this inequality highlights the fact that, as expected, the mean error between $\widetilde{X}_t$ and the true process X_t is within a factor of the average approximation accuracy of f_k and g_k. This bound also casts a new light on the role of $C(m)$, and the smaller, the better.

Notice also that if $\bar{X}_t$ is the stationary solution of the CHARME(∞) model ((2), for $p = \infty$) and X_t is the stationary solution of its associated CHARME(p) model (defined in (6)), we can then approximate this solution by $\widetilde{X}_t$, for some large integer value of p and ε_k small enough for all $k \in [K]$. Precisely, we would get that

$$\mathbb{E}\|\bar{X}_t - \widetilde{X}_t\| \leq E\|\bar{X}_t - X_t\| + \mathbb{E}\|X_t - \widetilde{X}_t\| \leq (7) + (19) \longrightarrow 0,$$

as $\varepsilon_k \to 0$ for all $k \in [K]$ and $p \to \infty$. This justifies that one could learn infinity memory CHARME models with neural networks, by approximating them by a NN-based CHARME(p) model for p finite but sufficiently large. Of course, strictly speaking, learning a CHARME(∞) would necessitate infinitely many observations.

7 Numerical experiments

In order to assess numerically the performance (consistency and asymptotic normality) of our estimator and support our theoretical predictions, we here report some numerical experiments. The CHARME(p) models in (10) were generated in two scenarios: (i) when the autoregressive functions f_k are generated by feedforward NNs, in which case the functions f_k are exacly modeled by neural networks; and (ii) when they are not, that is a neural network may provide only an ε_k-approximations to each function f_k. We also show the NN-based CHARME model-fitting performance when the data is generated by an integer-valued time series. We include a study of the variability of the estimated residuals as a function of the number of NN parameters (weights and biases) as well as a Monte Carlo simulation to assess the asymptotic normality of the estimated NN parameters. In all cases, we parametrize the functions

f_k with feedforward NNs, and we train the NNs by minimizing (11) to estimate the corresponding weights and biases θ_k. The estimation/training step is accomplished using stochastic (sub)gradient descent (SGD). For smooth activation maps, the gradient is computed via the chain rule through reverse mode automatic differentiation (i.e., backpropagation algorithm); see [23]. For non-smooth activations such as the ReLU, we invoke the theory of conservative fields and definability proposed recently in [7] to justify our use of the non-smooth chain rule and automatic differentiation.

All experiments were conducted under R with an interface to Keras 2.6.0 [17]. R files that allow to reproduce our experiments are publicly available for download at https://github.com/jose3g/Learning_CHARME_models_with_DNN.git.

Experiment 1 (Learning NNs from NN-based CHARME data) We simulate a NN-based CHARME(p) model as in (10) with $K = 3$ and $p = 30$, where $f_k(\cdot) = f_k(\cdot, \theta_k^0)$, $k = 1, 2, 3$, are neural networks with $\#\text{neu}(\theta_1^0) = (N_{1,0}, \dots, N_{1,5}) = (30, 50, 60, 40, 20, 1)$, $\#\text{neu}(\theta_2^0) = (N_{2,0}, \dots, N_{2,3}) = (30, 20, 5, 1)$ and $\#\text{neu}(\theta_3^0) = (N_{3,0}, \dots, N_{3,3}) = (30, 25, 30, 1)$, all with a ReLU activation function. We have taken the weights $w_{k,ij}^{(l)}$ arbitrarily (randomly uniform over a small interval $[-\delta, \delta]$) and $(\pi_1, \pi_2, \pi_3) = (0.1, 0.4, 0.5)$ such that $C(1) < 1$ (the explicit expression is provided in (15)) in order to guarantee the stationarity of the model. Precisely, $C(1) = 0.8353307$ for this model. The biases $b_k^{(l)} = (\beta_{k,1}^{(l)}, \dots, \beta_{k,N_{k,l}}^{(l)})^\top$ are also taken arbitrarily but in $\mathbb{R}$ and we have set particularly $(b_1^{(5)}, b_2^{(3)}, b_3^{(3)}) = (1, 0, -1)$. Then, from this model and with innovations $\epsilon_t \sim \mathcal{N}(0, 1)$, we have generated a dataset of $n = 10^5$ observations.

Let us turn to the estimation/training step. For this, we consider the quadratic loss function defined in (11) with the same configurations of the model that generates the data, that is, with $K = 3$, $(\pi_1, \pi_2, \pi_3) = (0.1, 0.4, 0.5)$ and f_k such that $\#\text{neu}(\theta_1) = (30, 50, 60, 40, 20, 1)$, $\#\text{neu}(\theta_2) = (30, 20, 5, 1)$ and $\#\text{neu}(\theta_3) = (30, 25, 30, 1)$, and the ReLU activation function. We run 20 iterations of the SGD algorithm with learning rate/step-size 0.01. Let $\widehat{\theta}_n^* = (\widehat{\theta}_{n,1}^*, \widehat{\theta}_{n,2}^*, \widehat{\theta}_{n,3}^*)$ be the parameters obtained in the last iteration.

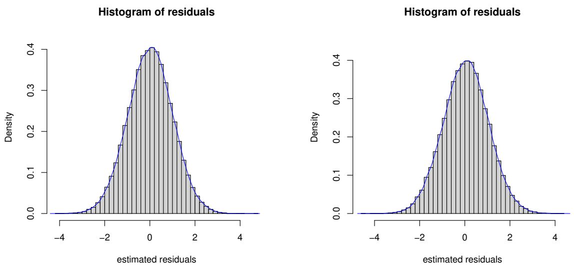
In Figure 1a, we show the histogram of estimated errors/residuals $\widehat{\epsilon}_t = X_t - \widehat{X}_t$, where

$$\widehat{X}_t = \sum_{k=1}^{K} \xi_t^{(k)} f_k(X_{t-1}, \dots, X_{t-p}, \widehat{\theta}_{n,k}^*). \tag{20}$$

We can observe how most of the residuals are concentrated around zero. In particular, due to the construction of the dataset with Gaussian residuals, these estimated residuals are standard Gaussians. We left the diagnostic graphs on the Rmarkdown.

Experiment 2 (Learning NNs from non NN-based CHARME data) In this experiment we simulate a CHARME(5) model as follows:

$$\begin{aligned} X_t = \epsilon_t &+ (X_{t-1} + 3)\mathbb{I}_{\{R_t=1\}} \\ &+ (\sqrt{0.2X_{t-1}^2 + 0.1X_{t-2}^2 + 0.25X_{t-3}^2 + 0.2X_{t-4}^2 + 0.05X_{t-5}^2} - 3)\mathbb{I}_{\{R_t=2\}} \end{aligned}$$



(a) NN-based data (b) Non NN-based data (c) Non-mixing data

Fig. 1: Histograms of estimated residuals $\widehat{\epsilon}_t$ after fitting a NN-based CHARME model. In Figure 1a, the data has been generated with a NN-based CHARME(30) process with $K = 3$ (Experiment 1). In Figures 1b and 1c the data has been generated with processes (21) and (22), respectively.

$$+(0.05X_{t-1} + 0.2X_{t-2} + 0.15X_{t-3} + 0.03X_{t-4} + 0.01X_{t-5} + 0.1)\mathbb{I}_{\{R_t=3\}} \tag{21}$$

with $(\pi_1, \pi_2, \pi_3) = (0.15, 0.35, 0.5)$. Note that the first autoregressive process $X_t^{(1)} = X_{t-1} + 3 + \epsilon_t$ is not stationary, although the entire process is stationary (because $C(1) < 1$). By taking $\epsilon_t \sim \mathcal{N}(0, 1)$, we generate again a dataset of $n = 10^5$.

For the estimation/training procedure, we consider also the quadratic loss function (11) with three NNs $f_k(\cdot, \theta_k)$, $k = 1, 2, 3$, such that $\#\text{neu}(\theta_1) = (5, 30, 40, 20, 1)$, $\#\text{neu}(\theta_2) = (5, 50, 60, 40, 1)$ and $\#\text{neu}(\theta_3) = (5, 30, 40, 20, 1)$, all with a ReLU activation map. We run 20 iterations of the SGD algorithm with learning rate/step-size 0.001. Let $\widehat{\theta}_n^* = (\widehat{\theta}_{n,1}^*, \widehat{\theta}_{n,2}^*, \widehat{\theta}_{n,3}^*)$ be the parameters obtained in the last iteration.

Similarly to Experiment 1, we show on Figure 1b the histogram of the errors/residuals $\widehat{\epsilon}_t = X_t - \hat{X}_t$, where $\widehat{X}_t$ is as given in (20).
We find the same behavior as in the first experiment: residuals concentrated around zero and with a Gaussian behavior due to the construction of the data with Gaussian residuals.

Experiment 3 (Learning NNs from a INAR (non-mixing) data) In this experiment we consider the Integer-valued first order autoregressive process (INAR(1)) defined as follows:

$$X_t = \pi \circ X_{t-1} + Z_t, \qquad (\in E = \mathbb{N}) \tag{22}$$

where $\pi \in (0, 1)$ is a fixed value and $(Z_t)_{t\in\mathbb{N}}$ is a sequence of integer-valued i.i.d. random variables. Here, the operator $\circ$ is defined as follows:

$$\circ : (\pi, x) \in (0, 1) \times E \mapsto \pi \circ x = \sum_{i=1}^{x} B_i \in E,$$

with $B_1, \ldots, B_n, \ldots$ i.i.d. Bernoulli(π).

Note that the model (22) can be seen as a CHARME(1) model with $K = 1$, $g \equiv 1$ and $f(x, \pi) = \pi \circ x$. Moreover, for all $x, y \in E$, $\mathbb{E}\,|f(x, \pi) - f(y, \pi)| = \pi\,|x - y|$, which implies the existence of a solution stationary, ergodic and τ-weakly dependent (follow the steps of the proof of Theorem 3.1 with $F(x;\ (z, b_1, b_2, \ldots)) = \sum_{i=1}^{x} b_i + z$).

For the simulation, we generate a dataset of $n = 10^5$, with $\pi = 0.7$ and $Z_i \sim \mathcal{P}(1)$ (Poisson of mean 1). For the estimation/training procedure, we consider the classical quadratic loss function ((11) with $K = 1$) with a NN $\tilde{f}(\cdot, \theta)$ such that $\#neu(\theta) = (1, 150, 150, 150, 150, 1)$ and the sigmoid activation map. Here, we run 20 iterations of the SGD algorithm with learning rate/step-size 0.001. In the same way, let $\hat{\theta}^*_n$ be the parameters obtained in the last iteration. Now, we show on Figure 1c the histogram of errors/residuals $\hat{\epsilon}_t = X_t - \hat{X}_t$, where $\hat{X}_t = \tilde{f}(X_t, \hat{\epsilon}^*_n)$.

Due to the irregularity of the data, we obtain a slightly irregular density. However, these estimated residuals are also concentrated around zero.

Experiment 4 (Asymptotic normality of trained NNs parameters) We set a CHARME(p) model as in (10) with $K = 3$ and $p = 16$, where $f_k = f_k(\cdot, \theta^0_k)$, $k = 1, 2, 3$, are NNs with $\#\mathrm{neu}(\theta^0_1) = (16, 32, 64, 32, 1)$, $\#\mathrm{neu}(\theta^0_2) = (16, 64, 32, 1)$, $\#\mathrm{neu}(\theta^0_3) = (16, 32, 64, 1)$, all with sigmoid activation function (this is because for the CLT result of Theorem 5.1 to apply, the activation function must be three-times continuously differentiable). Of course, the weights generated satisfy the condition $C(1) < 1$. In particular, $C(1) = 0.9743731$ for the weights generated in this model.

We now perform the following steps $N = 125$ times:

(i) By taking normal standard innovations with the aforementioned model, we generate a dataset of $n = 2 \cdot 10^4$,

(ii) By considering the quadratic loss function (11) with the same configurations of the model that generates the data and the sigmoid activation function, we run 2000 iterations of the SGD algorithm with learning rate/step-size 0.01 and decay rate 10^{-6}, in order to obtain an approximation $\widehat{\theta}^*_n$ of $\widehat{\theta}_n$.

Let $\widehat{\theta}^*_n(t)$, $t = 1, \ldots, 125$, be the estimates[11] obtained in each step of the Monte Carlo simulation and let $\eta_n(t) := \sqrt{n}\left(\widehat{\theta}^*_n(t) - \theta^0\right)$, $t = 1, \ldots, 125$. On can easily check that the number of parameters to learn is 10691, i.e., $\theta^0 \in \mathbb{R}^{10691}$, and in turn each $\eta_n(t)$ is a vector in $\mathbb{R}^{10691}$.

Figure 2 shows the box-plots of the coordinates of η_n. For the sake of readability, we only show 100 arbitrarily selected coordinates.

To test normality of η_n, as predicted by Theorem 5.1, we apply three multivariate normality tests: Mardia, Henze-Zirkler and Royston test (for the details of these tests, see [28, 37, 41, 49]). Given that the dimension of $\eta_n(t)$ is quite large (anyway larger than $N = 125$), to avoid numerical instabilities due to matrix inversion, these tests were not applied to the entire set of coordinates of $\eta_n(t)$, but to an arbitrary subset of 15 parameters (i.e., 15 arbitrary coordinates of η_n that we will call $\eta_n|_B$, where $B \subset [10691]$), which yield the results shown in Table 1.

We also report the Chi-Square Q-Q plot for Squared Mahalanobis Distance from $\eta_n|_B$ to 0 on Figure 3. We can see that the Q-Q plot is, in fact, almost along the straight line. Therefore, observing this behavior and the p-values obtained in the three tests of normality on Table 1, we can conclude that the vector $\eta_n|_B$ has indeed the predicted Gaussian behavior.

[11] These are really the SGD-approximations of the conditional least-squares estimates.

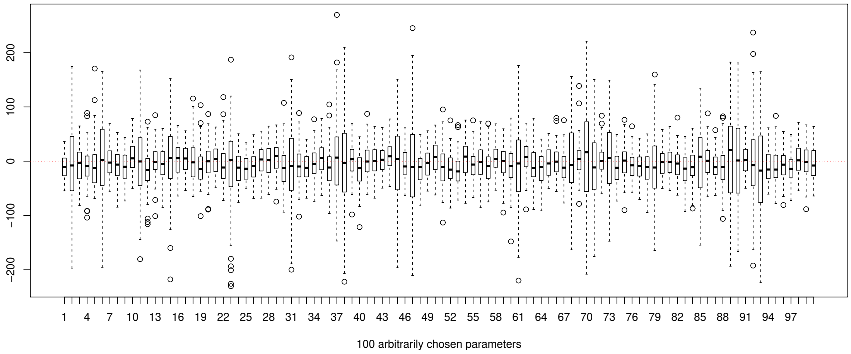


Fig. 2: Boxplots of 100 coordinates of η_n.

Table 1: Multivariate normality test results.

Test	Test Statistic	p-value
Mardia		
Skewness	687.5626	0.4120106
Kurtosis	-0.4461589	0.6554825
Henze-Zirkler	0.9931136	0.7983517
Royston	22.35297	0.0987472

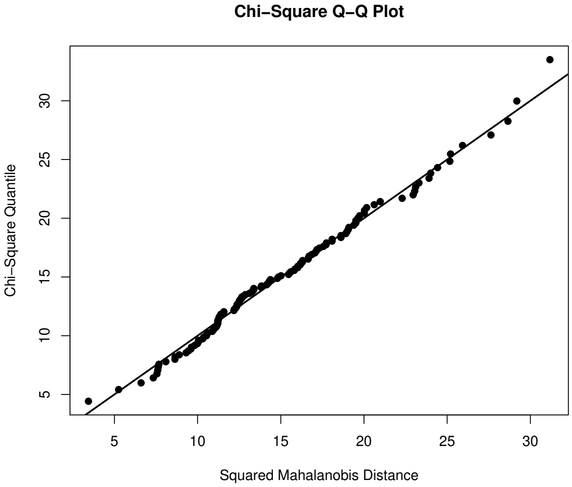


Fig. 3: Chi-Square Q-Q plot: empirical quantiles of squared Mahalanobis distance from $\eta_n|_B$ to $\vec{0}$ vs chi-square quantiles.

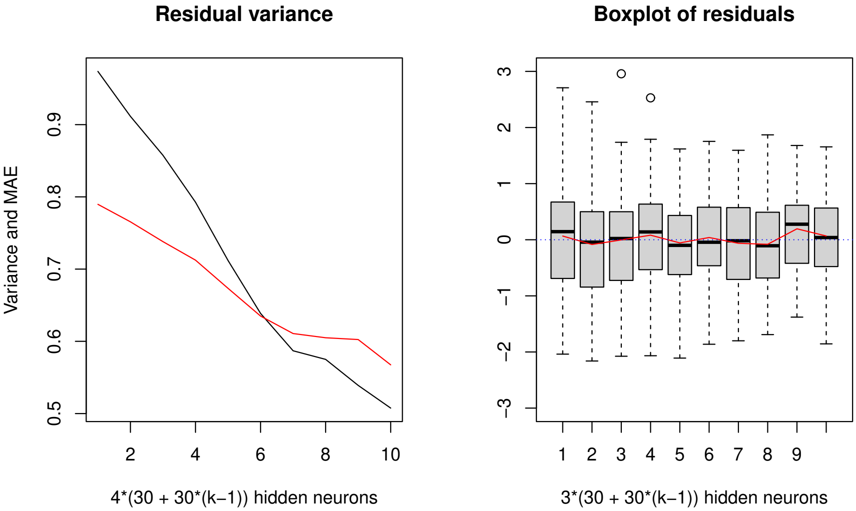


Fig. 4: On the left side: variance of estimated residuals (black solid line) and MAE (red solid line) of Experiment 5. On the right side: boxplots of estimated residuals (100 observations per boxplot) for $30, 60, \ldots, 300$ neurons per layer. Here, the red solid line is the estimated residuals mean.

Experiment 5 (Link between the number of neurons and the variability of the estimated residuals) In this experiment we study the relationship between the number of neurons (for a fixed number of layers) and the estimated residuals variance. For this, we have generated a dataset of $n = 10^5$ from the process:

$$X_t = \sin\left(0.15X_{t-1} + 0.2X_{t-2} + 0.15X_{t-3} + 0.1X_{t-4} + 0.3X_{t-5}\right) + \epsilon_t, \qquad t \in \mathbb{Z},$$

with $\epsilon_t \overset{iid}{\sim} \mathcal{N}(0, 1)$. Once the dataset is generated, we fit a feedforward NNs model with $L = 4$ hidden layers and we study the variance (and the MAE) of estimated residuals by increasing the number of neurons per layer.
For the estimation/training step, we consider the quadratic loss function defined in (11) and fit a feedforward NN with $L = 4$ hidden layers and $30k$ neurons per layer. For each $k \in [10]$, we perform this adjustment and we calculate the estimated residuals variance. Here, we have run 500 iterations of the SGD algorithm with learning rate 0.01.

On the left side in Figure 4, we can observe that when the number of neurons increases per layer (at a rate of 30 neurons per layer), the variance and the minimum absolute error (MAE) decreases. On the right side of the same figure, we show the boxplots (100 observations per boxplot) of the estimated residuals for each $k \in [10]$, that is, for each $30k = 30, 60, \ldots, 300$ neurons per layer of the adjusted NN.

8 Proofs

8.1 Proof of Theorem 3.1

(i) Note that the CHARME(∞) model defined in (2) with $p=\infty$, can be written as a Markov process:

$$X_t = F(X_{t-1}, X_{t-2}, \dots;\ \tilde{\xi}_t), \quad t \in \mathbb{Z}, \tag{23}$$

by taking the function

$$F(x;\ (\xi^{(0)}, \dots, \xi^{(K)})) = \sum_{k=1}^{K} \xi^{(k)} \left(f_k(x) + g_k(x)\xi^{(0)} \right) \tag{24}$$

with innovations $\tilde{\xi}_t := (\epsilon_t, \xi_t^{(1)}, \dots, \xi_t^{(K)}) = (\epsilon_t, \xi_t) \in E \times \{e_1, \dots, e_K\}$. Therefore, verifying [16, Conditions (3.1) and (3.3)], we will obtain the result by [16, Theorem 3.1]. Note that Condition (3.2) of that paper is already assumed.

Indeed, since the sequences $(\epsilon_t)_{t\in\mathbb{Z}}$ and $(R_t)_{t\in\mathbb{Z}}$ are independent and $\xi_0 \in \{e_1, \dots, e_K\}$, denoting $\mathbb{E}_\epsilon$ the expectation with respect to the distribution of ϵ, we obtain for $x = (x_1, x_2, \dots)$ and $y = (y_1, y_2, \dots)$, that

$$\begin{aligned}
\|F(x;\ \tilde{\xi}_0) - F(y;\ \tilde{\xi}_0)\|_1 &= \mathbb{E}\left[\|F(x;\ \tilde{\xi}_0) - F(y;\ \tilde{\xi}_0)\| \right] \\
&= \mathbb{E}\left[\left\| \sum_{k=1}^{K} \xi_0^{(k)} \left(f_k(x) - f_k(y) + (g_k(x) - g_k(y))\epsilon_0 \right) \right\| \right] \\
= \mathbb{E}_\epsilon \Bigg[\sum_{j=1}^{K} &\left\| \sum_{k=1}^{K} e_j^{(k)} \left(f_k(x) - f_k(y) + (g_k(x) - g_k(y))\epsilon_0 \right) \right\| \mathbb{P}(\xi_0 = e_j) \Bigg] \\
&= \mathbb{E}_\epsilon \left[\sum_{k=1}^{K} \pi_k \left\| \left(f_k(x) - f_k(y) + (g_k(x) - g_k(y))\epsilon_0 \right) \right\| \right] \\
&= \sum_{k=1}^{K} \pi_k \mathbb{E}_\epsilon \left\| \left(f_k(x) - f_k(y) + (g_k(x) - g_k(y))\epsilon_0 \right) \right\| \\
&\leq \sum_{i=1}^{\infty} \left(\sum_{k=1}^{K} \pi_k \left(a_i^{(k)} + b_i^{(k)} \|\epsilon_0\|_1 \right) \right) \|x_i - y_i\|,
\end{aligned}$$

by the Minkowski inequality and the Lipschitz-type assumptions (4) on f_k and g_k. So, this verifies (3.1) of [16].

On the other hand, using the same arguments as above, we can establish that

$$\tilde{\mu}_1 = \|F(0;\ \tilde{\xi}_0)\|_1 \leq \sum_{k=1}^{K} \pi_k \left(\|f_k(0)\| + |g_k(0)|\,\|\epsilon_0\|_1\right),$$

which is finite because $\epsilon_0 \in \mathbb{L}^1$. The first part of the theorem is proven.

(ii) Suppose now that $C(m) < 1$ for some $m \in \mathbb{N} \cap (1, \infty)$. Let $x = (x_1, \ldots)$ and rewrite $f_k(x) = f_k(x) - f_k(0) + f_k(0)$. Then, from (4) and the Minkowski inequality, we have

$$\|f_k(x)\| \leq \sum_{i=1}^{\infty} a_i^{(k)} \|x_i\| + o_k = w_k(x) + o_k, \qquad (25)$$

where $o_k = \|f_k(0)\|$ and $w_k(x) = \sum_{i=1}^{\infty} a_i^{(k)} \|x_i\|$. Thus,

$$\|f_k(x)\|^m \leq \sum_{j=0}^{m-1} \binom{m}{j} w_k^j(x)\ o_k^{m-j} + w_k^m(x). \qquad (26)$$

Taking the probability weights $\lambda_i = a_i^{(k)}/A_k$ (recall that $A_k = \sum_{i=1}^{\infty} a_i^{(k)}$), we can apply Jensen's inequality for any $s \geq 1$ as follows:

$$w_k^s(x) = A_k^s \left(\sum_{i=1}^{\infty} \frac{a_i^{(k)}}{A_k} \|x_i\| \right)^s \leq A_k^{s-1} \sum_{i=1}^{\infty} a_i^{(k)} \|x_i\|^s. \qquad (27)$$

Let us denote $Y_{t-1} = (X_{t-1}, X_{t-2}, \ldots)$. From the stationarity of $(X_t)_{t \in \mathbb{Z}}$, for $s \geq 1$, we obtain

$$\mathbb{E}\left[w_k^s(Y_{t-1})\right] \leq A_k^{s-1} \sum_{i=1}^{\infty} a_i^{(k)} \mathbb{E}\|X_{t-i}\|^s = A_k^s \mathbb{E}\|X_0\|^s \qquad (28)$$

and therefore

$$\mathbb{E}\|f_k(Y_{t-1})\|^m \leq A_k^m \mathbb{E}\|X_0\|^m + \mathbb{E}\left[R_{k,m}(\|X_0\|)\right], \qquad (29)$$

where $R_{k,s}(x) := \sum_{j=0}^{s-1} \binom{s}{j} A_k^j\ o_k^{s-j} x^j$.

Similarly, with the same steps, we can prove that

$$\mathbb{E}\left|g_k(Y_{t-1})\right|^m \leq B_k^m \mathbb{E}\|X_0\|^m + \mathbb{E}\left[\bar{R}_{k,m}(\|X_0\|)\right], \qquad (30)$$

where $\bar{R}_{k,s}(x) := \sum_{j=0}^{s-1} \binom{s}{j} B_k^j O_k^{s-j} x^j$, with $B_k = \sum_{i=1}^{\infty} b_i^{(k)}$ and $O_k = |g_k(0)|$.

Since $(\xi_t^{(1)}, \dots, \xi_t^{(K)}) \in \{e_1, \dots, e_K\}$, for $m \in \mathbb{N}^*$,

$$\begin{aligned}\|X_t\|^m &= \sum_{k=1}^{K} \xi_t^{(k)} \|f_k(Y_{t-1}) + g_k(Y_{t-1})\epsilon_t\|^m \\ &\leq 2^{m-1} \sum_{k=1}^{K} \xi_t^{(k)} \left(\|f_k(Y_{t-1})\|^m + |g_k(Y_{t-1})|^m \, \|\epsilon_t\|^m \right), \qquad (31)\end{aligned}$$

where the last line is due to Jensen's inequality. On the other hand, as R_t is independent of the random vector (ϵ_t, Y_{t-1}) and ϵ_t is independent of Y_{t-1}, then, under the invariant measure (the existence of this measure is from the stationarity of $(X_t)_{t\in\mathbb{Z}}$), we obtain that

$$\begin{aligned}\mathbb{E}\|X_0\|^m &= \mathbb{E}\|X_t\|^m \\ &\leq 2^{m-1} \sum_{k=1}^{K} \pi_k \left(\mathbb{E}\|f_k(Y_{t-1})\|^m + \|\epsilon_0\|_m^m \mathbb{E}\,|g_k(Y_{t-1})|^m \right) \\ &\leq 2^{m-1} \sum_{k=1}^{K} \pi_k (A_k^m + B_k^m \|\epsilon_0\|_m^m) \mathbb{E}\|X_0\|^m + C, \qquad (32)\end{aligned}$$

where $C = 2^{m-1} \sum_{k=1}^{K} \pi_k \left(\mathbb{E}\left[R_{k,m}(\|X_0\|)\right] + \|\epsilon_0\|_m^m \mathbb{E}\left[\bar{R}_{k,m}(\|X_0\|)\right] \right) < \infty$ since from recursion $\mathbb{E}\|X_0\|^{m-1} < \infty$. Therefore, by taking

$$D = \sum_{k=1}^{K} \pi_k \left(A_k^m + B_k^m \|\epsilon_0\|_m^m \right) < \frac{1}{2^{m-1}}, \qquad (33)$$

we conclude that

$$E\|X_0\|^m < \frac{C}{1 - 2^{m-1}D} < \infty.$$

For the case $m \in (1, \infty) \setminus \mathbb{N}$, we write $m = n + \delta$, where $n = \lfloor m \rfloor$ and $\delta \in (0, 1)$. Then, by using the expression (25), we have that

$$\begin{aligned}\|f_k(x)\|^m &= \|f_k(x)\|^\delta \; \|f_k(x)\|^n \leq (w_k(x) + o_k)^\delta \sum_{j=0}^{n} \binom{n}{j} w_k^j(x) \; o_k^{n-j} \\ &\leq \sum_{j=0}^{n} \binom{n}{j} w_k^{j+\delta}(x) \; o_k^{n-j} + \sum_{j=0}^{n} \binom{n}{j} w_k^j(x) \; o_k^{n+\delta-j} \\ &= w_k^m(x) + o_k^\delta w_k^n(x) + \sum_{j=0}^{n-1} \binom{n}{j} w_k^{j+\delta}(x) \; o_k^{n-j} + \sum_{j=0}^{n-1} \binom{n}{j} w_k^j(x) \; o_k^{n+\delta-j}.\end{aligned}$$

As in the previous case, using (27) and (28), we get that

$$\mathbb{E}\|f_k(Y_{t-1})\|^m \leq A_k^m \mathbb{E}\|X_0\|^m + o_k^\delta A_k^n \mathbb{E}\|X_0\|^n + \mathbb{E}\left[R_{k,m}^*(\|X_0\|)\right],$$

where

$$R_{k,s}^*(x) := \sum_{j=0}^{\lfloor s\rfloor-1} \binom{\lfloor s\rfloor}{j} A_k^{j+s-\lfloor s\rfloor}\ o_k^{\lfloor s\rfloor-j} x^{j+s-\lfloor s\rfloor} + \sum_{j=0}^{\lfloor s\rfloor-1} \binom{\lfloor s\rfloor}{j} A_k^j o_k^{s-j} x^j.$$

Similarly, with the same steps, we can prove that

$$\mathbb{E}\left|g_k(Y_{t-1})\right|^m \leq B_k^m \mathbb{E}\|X_0\|^m + O_k^\delta B_k^n \mathbb{E}\|X_0\|^n + \mathbb{E}\left[\bar{R}_{k,m}^*(\|X_0\|)\right],$$

where

$$\bar{R}_{k,s}^*(x) := \sum_{j=0}^{\lfloor s\rfloor-1} \binom{\lfloor s\rfloor}{j} B_k^{j+s-\lfloor s\rfloor}\ O_k^{\lfloor s\rfloor-j} x^{j+s-\lfloor s\rfloor} + \sum_{j=0}^{\lfloor s\rfloor-1} \binom{\lfloor s\rfloor}{j} B_k^j O_k^{s-j} x^j,$$

with $B_k = \sum_{i=1}^\infty b_i^{(k)}$ and $O_k = \left|g_k(0, \lambda_k^0)\right|$.

Using the same arguments to prove (32), we arrive at

$$\begin{aligned}\mathbb{E}\|X_0\|^m &= \mathbb{E}\|X_t\|^m \\ &\leq 2^{m-1} \sum_{k=1}^K \pi_k \left(\mathbb{E}\|f_k(Y_{t-1})\|^m + \|\epsilon_0\|_m^m \mathbb{E}\left|g_k(Y_{t-1})\right|^m\right) \\ &\qquad \leq 2^{m-1} \sum_{k=1}^K \pi_k (A_k^m + B_k^m \|\epsilon_0\|_m^m) \mathbb{E}\|X_0\|^m + C^*,\end{aligned}$$

$$\text{where } \ C^* = 2^{m-1} \sum_{k=1}^K \pi_k \left((o_k^\delta A_k^n + O_k^\delta B_k^n \|\epsilon_0\|_m^m) \mathbb{E}\|X_0\|^n \right. \\ \left. + \mathbb{E}\left[R_{k,m}^*(\|X_0\|) + \|\epsilon_0\|_m^m \bar{R}_{k,m}^*(\|X_0\|)\right]\right) \quad (34)$$

which is finite by recursion, because $\mathbb{E}\|X_0\|^{m-1} < \left(\mathbb{E}\|X_0\|^n\right)^{\frac{m-1}{n}} < \infty$. Therefore,

$$E\|X_0\|^m < \frac{C^*}{1 - 2^{m-1}D} < \infty,$$

which completes the proof of the theorem. □

8.2 Proof of Theorem 4.1

The proof consists in showing that all conditions of [34, Theorem 1.1] are in force under our assumptions, and to combine this with epi-convergence arguments; see [3, 10, 46] for more about epi-convergence theory and applications.

By virtue of **(A.3)** and **(A.4)**, it follows from the composition rule in [46, Proposition 14.45(a)] that

$$(Y_t, (\lambda_k, \theta_k)) \mapsto \ell\big(X_t, f_k(X_{t-1}, \ldots, X_{t-p}, \theta_k), g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k)\big)$$

is random lsc. This entails that

$$\xi_t^{(k)} \ell\big(X_t, f_k(X_{t-1}, \ldots, X_{t-p}, \theta_k), g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k)\big)$$

is also random lsc thanks to [46, Corollary 14.46]. In turn, h (see (9)), which is the sum of such K random lsc, is also random lsc in view of [46, Proposition 14.44(c)].

It remains to show that $\inf_{\Theta\times\Lambda} h(Y_t, \xi_t, \cdot, \cdot) \in \mathbb{L}^1$. We have

$$\begin{aligned}
0 &\underset{\textbf{(A.5)}}{\leq} \mathbb{E}\left[\inf_{\theta,\lambda} h(Y_t, \xi_t, \theta, \lambda)\right] \\
&= \mathbb{E}\left[\inf_{\theta,\lambda} \sum_{k=1}^{K} \xi_t^{(k)} \ell\big(X_t, f_k(X_{t-1}, \ldots, X_{t-p}, \theta_k), g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k)\big)\right] \\
&\underset{\text{Separability}}{=} \mathbb{E}\left[\sum_{k=1}^{K} \xi_t^{(k)} \inf_{\theta_k,\lambda_k} \ell\big(X_t, f_k(X_{t-1}, \ldots, X_{t-p}, \theta_k), g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k)\big)\right] \\
&\underset{\text{Optimality}}{\leq} \mathbb{E}\left[\sum_{k=1}^{K} \xi_t^{(k)} \inf_{\lambda_k} \ell\big(X_t, f_k(X_{t-1}, \ldots, X_{t-p}, \bar{\theta}_k), g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k)\big)\right] \\
&\underset{\textbf{(A.6)}}{=} \mathbb{E}\left[\sum_{k=1}^{K} \xi_t^{(k)} \inf_{\lambda_k} \ell\big(X_t, 0, g_k(X_{t-1}, \ldots, X_{t-p}, \lambda_k)\big)\right] \\
&\underset{\textbf{(A.7)}}{\leq} \left(\sum_{k=1}^{K} \pi_k\right) (C\mathbb{E}\|X_t\|^\gamma + c) = C\mathbb{E}\|X_t\|^\gamma + c.
\end{aligned}$$

Using the fact that $\gamma = m$ and $\mathbb{E}\|X_t\|^m < +\infty$ by Theorem 3.1, we deduce that $\inf_{\Theta\times\Lambda} h(Y_t, \xi_t, \cdot, \cdot) \in \mathbb{L}^1$.

Now, by **(A.2)**, $\Theta \times \Lambda$, as a product space of Polish spaces is also Polish. Thus combining this with **(A.1)**, that h is random lsc, and the summability property we have just shown, as well as the stationarity and ergodicity of Y_t which are inherited from those of X_t, it follows from [34, Theorem 1.1] that Q_n epi-converges to $\mathbb{E}h(Y, \xi, \cdot, \cdot)$ a.s. It remains now to invoke standard epi-convergence arguments that entail the convergence of the minimizers of Q_n to those of $\mathbb{E}h(Y, \xi, \cdot, \cdot)$.

(i) Apply [10, Corollary 7.20].
(ii) Apply [10, Corollary 7.24].
This completes the proof. □

8.3 Proof of Theorem 5.1

The proof consists in showing that the conditions (A1)-(A4) of [56, Theorem 3.2.23] are fulfilled.
Indeed, let us denote $Y_{t-1} = (X_{t-1}, \dots, X_{t-p})$. Then, from strict stationarity and ergodicity, the ergodic theorem and **(B.2)**, it follows that

$$\begin{aligned}\frac{1}{n}\frac{\partial Q_n(\theta^0)}{\partial \theta_{k,i}} &= -\frac{2}{n}\sum_{t=1}^{n}\xi_t^{(k)}\left(X_t - f_k(Y_{t-1},\theta_k^0)\right)^\top \frac{\partial f_k(Y_{t-1},\theta_k^0)}{\partial \theta_{k,i}} \\ &\xrightarrow{a.s.} -2\pi_k \mathbb{E}\left[\left(X_{p+1} - f_k(Y_p,\theta_k^0)\right)^\top \frac{\partial f_k(Y_p,\theta_k^0)}{\partial \theta_{k,i}}\right] = 0,\end{aligned}$$

for all $k \in [K]$ and all $i \in [d_k]$. Hence, condition (A1) of [56, Theorem 3.2.23] is satisfied.
Similarly, using again **(B.2)** and the ergodic theorem, we have that

$$\begin{aligned}\frac{1}{n}\frac{\partial^2 Q_n(\theta^0)}{\partial \theta_{l,j}\partial \theta_{k,i}} = \ & \frac{2}{n}\sum_{t=1}^{n}\xi_t^{(k)}\left[\left(\frac{\partial f_k(Y_{t-1},\theta_k^0)}{\partial \theta_{k,j}}\right)^\top \frac{\partial f_k(Y_{t-1},\theta_k^0)}{\partial \theta_{k,i}}\right. \\ & \left. - \left(X_t - f_k(Y_{t-1},\theta_k^0)\right)^\top \frac{\partial^2 f_k(Y_{t-1},\theta_k^0)}{\partial \theta_{k,j}\partial \theta_{k,i}}\right]\mathbb{I}_{\{l=k\}} \\ \xrightarrow{a.s.} \ & 2\pi_k \mathbb{E}\left[\left(\frac{\partial f_k(Y_p,\theta_k^0)}{\partial \theta_{k,j}}\right)^\top \frac{\partial f_k(Y_p,\theta_k^0)}{\partial \theta_{k,i}}\right]\mathbb{I}_{\{l=k\}} = 2(V_{kl})_{ij}\ , \qquad (35)\end{aligned}$$

for all $k, l \in [K]$ and all $(i,j) \in [d_k] \times [d_l]$, because

$$n^{-1}\sum_{t=1}^{n}\xi_t^{(k)}\left(X_t - f_k(Y_{t-1},\theta_k^0)\right)^\top \frac{\partial^2 f_k(Y_{t-1},\theta_k^0)}{\partial \theta_{k,j}\partial \theta_{k,i}}\mathbb{I}_{\{l=k\}} \xrightarrow{a.s.} 0,$$

for all $k, l \in [K]$ and all $(i,j) \in [d_k] \times [d_l]$; see [53]. In the expression (35), $(V_{kl})_{ij}$ denotes the (i,j)-th entry of the matrix V_{kl} defined in (13). From **(B.3)**, the Gram matrix of each Jacobian $J[f_k(X_p, \dots, X_1, \cdot)](\theta_k^0)$ is invertible for any $k \in [K]$, whence we deduce that V is positive definite since it is block-diagonal whose diagonal blocks are those Gram matrices (up to multiplication by $\pi_k > 0$). Thus, assumption (A2) of [56, Theorem 3.2.23] is also satisfied.

Now, let $\theta \in \mathcal{V}$, and $\delta > 0$ such that the ball $\|\theta - \theta^0\| < \delta$ is contained in $\mathcal{V}$ (δ can be chosen arbitratily small for this to hold). Let the closed segment $[\theta_0, \theta] = \{\rho\theta + (1-\rho)\theta_0 :\ \rho \in [0,1]\}$ and the open segment $]\theta_0, \theta[= \{\rho\theta + (1-\rho)\theta_0 :\ \rho \in]0,1[\}$. Then, for $\bar{\theta} \in [\theta_0, \theta]$, and any $k, l \in [K]$

and $(i,j) \in [d_k] \times [d_l]$, we have from the mean value theorem that

$$\big(T_n(\bar{\theta})\big)_{kl,ij} := \begin{cases} \dfrac{\partial^2 Q_n(\bar{\theta})}{\partial \theta_{k,j}\partial \theta_{k,i}} - \dfrac{\partial^2 Q_n(\theta^0)}{\partial \theta_{k,j}\partial \theta_{k,i}} & \text{if } l = k \\ 0 & \text{if } l \neq k \end{cases}$$
$$= (\bar{\theta} - \theta^0)^\top \nabla \left(\frac{\partial^2 Q_n(\bar{\bar{\theta}})}{\partial \theta_{k,j}\partial \theta_{k,i}} \right) \mathbb{I}_{\{l=k\}}, \quad \text{for some } \bar{\bar{\theta}} \in]\theta^0, \bar{\theta}[.$$

Since by definition $\|\bar{\theta} - \theta^0\| < \delta$, we have $\|\bar{\bar{\theta}} - \theta^0\| < \delta$ and thus $\bar{\bar{\theta}} \in \mathcal{V}$. Hence from continuity of the norm and that of the derivatives of Q_n up to third-order on $\mathcal{V}$, we get, upon using Cauchy-Scwartz inequality, that

$$\begin{aligned}
&\sup_{\delta \to 0} \frac{1}{n\delta} \left| \big(T_n(\bar{\theta})\big)_{kl,ij} \right| \\
&\leq \liminf_{\delta \to 0} \frac{1}{n} \left\| \nabla \left(\frac{\partial^2 Q_n(\bar{\bar{\theta}})}{\partial \theta_{k,j}\partial \theta_{k,i}} \right) \mathbb{I}_{\{l=k\}} \right\| = \lim_{\delta \to 0} \frac{1}{n} \left\| \nabla \left(\frac{\partial^2 Q_n(\bar{\bar{\theta}})}{\partial \theta_{k,j}\partial \theta_{k,i}} \right) \mathbb{I}_{\{l=k\}} \right\| \\
&\leq \lim_{\delta \to 0} \frac{2}{n} \mathbb{I}_{\{l=k\}} \sum_{r=1}^{d_k} \sum_{t=1}^{n} \xi_t^{(k)} \left(\left| \left(\frac{\partial f_k(Y_{t-1}, \bar{\bar{\theta}}_k)}{\partial \theta_{k,i}} \right)^\top \frac{\partial^2 f_k(Y_{t-1}, \bar{\bar{\theta}}_k)}{\partial \theta_{k,j}\partial \theta_{k,r}} \right| \right. \\
&+ \left| \left(\frac{\partial f_k(Y_{t-1}, \bar{\bar{\theta}}_k)}{\partial \theta_{k,j}} \right)^\top \frac{\partial^2 f_k(Y_{t-1}, \bar{\bar{\theta}}_k)}{\partial \theta_{k,i}\partial \theta_{k,r}} \right| + \left| \left(\frac{\partial f_k(Y_{t-1}, \bar{\bar{\theta}}_k)}{\partial \theta_{k,r}} \right)^\top \frac{\partial^2 f_k(Y_{t-1}, \bar{\bar{\theta}}_k)}{\partial \theta_{k,i}\partial \theta_{k,j}} \right| \\
&\left. + \left| \left(X_t - f_k(Y_{t-1}, \bar{\bar{\theta}}_k) \right)^\top \frac{\partial^3 f_k(Y_{t-1}, \bar{\bar{\theta}}_k)}{\partial \theta_{k,i}\partial \theta_{k,j}\partial \theta_{k,r}} \right| \right). \\
&= \frac{2}{n} \mathbb{I}_{\{l=k\}} \sum_{r=1}^{d_k} \sum_{t=1}^{n} \xi_t^{(k)} \left(\left| \left(\frac{\partial f_k(Y_{t-1}, \theta_k^0)}{\partial \theta_{k,i}} \right)^\top \frac{\partial^2 f_k(Y_{t-1}, \theta_k^0)}{\partial \theta_{k,j}\partial \theta_{k,r}} \right| \right. \\
&+ \left| \left(\frac{\partial f_k(Y_{t-1}, \theta_k^0)}{\partial \theta_{k,j}} \right)^\top \frac{\partial^2 f_k(Y_{t-1}, \theta_k^0)}{\partial \theta_{k,i}\partial \theta_{k,r}} \right| + \left| \left(\frac{\partial f_k(Y_{t-1}, \theta_k^0)}{\partial \theta_{k,r}} \right)^\top \frac{\partial^2 f_k(Y_{t-1}, \theta_k^0)}{\partial \theta_{k,i}\partial \theta_{k,j}} \right| \\
&\left. + \left| \left(X_t - f_k(Y_{t-1}, \theta_k^0) \right)^\top \frac{\partial^3 f_k(Y_{t-1}, \theta_k^0)}{\partial \theta_{k,i}\partial \theta_{k,j}\partial \theta_{k,r}} \right| \right).
\end{aligned}$$

From strict stationarity, ergodicity and the condition (B.4), by using the ergodic theorem again, it follows that

$$\lim_{n \to \infty} \sup_{\delta \to 0} \frac{1}{n\delta} \left| \big(T_n(\bar{\theta})\big)_{kl,ij} \right| \leq 2\pi_k \mathbb{I}_{\{l=k\}} \sum_{r=1}^{d_k} \left(G_k^{ijr} + G_k^{jir} + G_k^{rij} + H_k^{ijr} \right) < \infty.$$

With this we have shown that condition (A3) of [56, Theorem 3.2.23] holds. Finally, by using [56, Theorem 1.3.3], the vector process $(Z_t)_{t\in\mathbb{Z}}$ defined by

$$Z_t = -2\left(\xi_t^{(1)}(X_t - f_1(Y_{t-1}, \theta_1^0))^\top \frac{\partial f_1(Y_{t-1}, \theta_1^0)}{\partial \theta_{1,1}}, \dots \right.$$
$$\left. \dots, \xi_t^{(K)}(X_t - f_1(Y_{t-1}, \theta_1^0))^\top \frac{\partial f_K(Y_{t-1}, \theta_K^0)}{\partial \theta_{K,d_K}}\right)$$

is strictly stationary and ergodic. Therefore, condition (A4) of [56, Theorem 3.2.23] follows by combining **(B.5)** and [56, Theorem A.2.14]. This completes the proof. □

Appendix A Derivatives with respect to NN parameters

Let $\theta = \big((W^{(1)}, \beta^{(1)}), \dots, (W^{(L)}, b^{(L)})\big)$ be an architecture of a NN $f : (x, \theta) \in \mathbb{R}^d \longrightarrow \mathbb{R}^{N_L}$ and denote $W^{(l)} = (w^{(l)}_{j_l j_{l-1}})_{(j_l, j_{l-1}) \in [N_l] \times [N_{l-1}]}$ and $\beta^{(l)} = (\beta^{(l)}_{j_l})_{j_l \in [N_l]}$, with $l \in [L]$. We denote $D[f]_{W^{(l)}}(x, \theta)$ the Fréchet derivatives of f wrt to $W^{(l)}$ evaluated at (x, θ). Recalling the recursion in Definition 2.2, and by the standard chain rule, $D[f]_{W^{(l)}}$ acting in the direction $H^{(l)} \in \mathbb{R}^{N_l \times N_{l-1}}$ reads, for $l \in [L]$,

$$D[f]_{W^{(l)}}(x, \theta)(H^{(l)}) = \left(\prod_{i=L-1}^{l} W^{(i+1)} J[\varphi]\left(W^{(i)} x^{(i-1)} + b^{(i)}\right)\right) H^{(l)} x^{(l-1)}. \tag{A1}$$

Similarly, we have

$$J[f]_{b^{(l)}}(x, \theta) = \prod_{i=L-1}^{l} W^{(i+1)} J[\varphi]\left(W^{(i)} x^{(i-1)} + b^{(i)}\right). \tag{A2}$$

As usual, the partial derivatives $\dfrac{\partial f(x, \theta)}{\partial w^{(l)}_{ij}}(x, \theta)$ (resp. $\dfrac{\partial f(x, \theta)}{\partial \beta^{(l)}_i}(\theta)$) is nothing but (A1) (resp. (A2)) evaluated in the direction $H^{(l)}$ (resp. i-th standard basis vector of $\mathbb{R}^{L_l}$) such that $H^{(l)}_{ij} = 1$ and 0 otherwise.

A similar calculation can be carried out to get the second- and third-order derivatives that we leave to the reader.

Acknowledgments. The first author was funded by the Normandy Region RIN program. The second author was partly supported by Institut Universitaire de France.

Declarations

- Funding: Not applicable.
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use): The authors declare no conflict of interest.

- Ethics approval: Not applicable.
- Consent to participate: Not applicable.
- Consent for publication: The authors agree with the publication of this paper.
- Availability of data and materials: Not applicable.
- Code availability: The code is available upon reasonable request to the corresponding author.
- Authors' contributions: The authors contributed equally to this work.

References

[1] Andrews, D.W.K. Non strong mixing autoregressive processes. J. Appl. Prob., 21 (1984) 930–934.

[2] Artstein, Z. and Wets, R. J-B. Consistency of minimizers and the SLLN for stochastic programs. J. Convex Anal., 2 (1995) 1–17.

[3] Attouch, H. Variational convergence for functions and operators. Applicable mathematics series. Pitman Advanced Publishing Program, 1984.

[4] Attouch, H. and Wets, R. J-B. Epigraphical processes: laws of large numbers for random lsc unctions. Sém. Anal. Convexe, Montpellier, 13 (1990) 1–29.

[5] Bartlett, P. L., Foster D. J. and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 30 (2017) 6241–6250

[6] Bolte, J. and Pauwels, E. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. Mathematical Programming, 2020.

[7] Castera, C., Bolte, C., Févotte, C. and Pauwels, E. An inertial newton algorithm for deep learning. https://arxiv.org/pdf/1905.12278.pdf, 2019.

[8] Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y. and Usunier, N. Parseval networks: improving robustness to adversarial examples. ICML, 70 (2017) 854–863

[9] Dahlhaus, R. A likelihood approximation for locally stationary processes. Annals of Statistics, 28 (2000)1762–1794.

[10] Dal Maso, G. An introduction to Γ-convergence, volume 8. Springer Science & Business Media, 2012.

[11] Daubechies, I., DeVore, R., Foucart, S., Hanin, B.and Petrova, G. Nonlinear approximation and (deep) ReLU networks. arxiv preprint

arxiv:1905.02199, 2019.

[12] Dedecker, J. and Prieur, C. Coupling for τ-dependent sequences and applications. J. Theoret. Probab., 17(4) (2004) 861–885.

[13] Douc, R. Fokianos, K. and Moulines, E. Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. Electronic Journal of Statistic, 11 (2017) 2707–2740.

[14] Doukhan, P., Fokianos, K. and Li, X. On weak dependence conditions: The case of discrete valued processes. Statistics and Probability Letters, 82 (2012) 1941–1948.

[15] Doukhan, P., Fokianos, K. and Tjøstheim, D. On weak dependence conditions for Poisson autoregressions. Statistics and Probability Letters, 82 (2012) 942–948.

[16] Doukhan, P. and Wintenberger, O. Weakly dependent chains with infinite memory. Stochastic Processes and their Applications, 118 (2008) 1997–2013.

[17] Falbel, D., Allaire, JJ., Chollet, F., RStudio, Google, Tang, Y., Van Der Bijl, W., Studer, M. and Keydana, S. Package "keras". https://cran.r-project.org/web/packages/keras/keras.pdf, 2019.

[18] Ferland, R., Latour, A. and Oraichi, D. Integer-valued GARCH processes. Journal of Time Series Analysis, 27 (2006) 923–942.

[19] Fokianos, K. and Fied, R. Interventions in INGARCH processes. Journal of Time Series Analysis, 31 (2010) 210–225.

[20] Fokianos, K., Rahbek, A. and Tjøstheim, D. Poisson autoregression. Journal of the American Statistical Association, 104 (2009) 1430–1439.

[21] Fokianos, K. and Tjøstheim, D. Nonlinear poisson autoregression. Annals of the Institute of Statistical Mathematics, 64 (2012) 1205–1225.

[22] Franke, J., Hardle, W. and Hafner, C. Statistics of Financial Markets: An introduction. Springer, 5 edition, 2019.

[23] Griewank, A. and Walther, A. Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, Second Edition. SIAM, 2008.

[24] Haeffele, B. and Vidal, R. Global optimality in neural network training. In CVPR, 2017.

[25] Hafner, C. Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility. Contributions to Economics. Springer-Verlag

Berlin Heidelberg GmbH, 1998.

[26] Hanin, B. and Sellke, M. Approximating continuous functions by ReLU nets of minimal width. arxiv preprint arXiv:1710.11278, 2017.

[27] Hebb, D. The organization of behavior: A neuropsychological theory. Wiley, 1949.

[28] Henze, N. and Zirkler, B. A class of invariant consistent tests for multivariate normality. Commun Stat Theory Methods, 19(10) (1990) 3595–3617.

[29] Hess, C. Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator. Annals of Statistics, 24(3) (1996) 1298–1315.

[30] Hornik, K., Stinchcombe, M. and White, H. Multilayer feedforward networks are universal approximators. Neural networks, 2(5) (1989) 359–366.

[31] Karmakar, S., Richter, S. and Wu, W.B. Simultaneous inference for time-varying models. Journal of Econometrics 2022.

[32] Kirch, C. and Kamgaing, T. Testing for parameter stability in nonlinear autoregressive models. Journal of Time Series Analysis, 33(3) (2012) 365–385.

[33] Klimko, L.A. and Nelson, P.I. On conditional least squares estimation for stochastic processes. The Annals of Statistics, 6(3) (1978) 629–642.

[34] Korf; L.A. and Wets, R.J.B. Random lsc functions: An ergodic theorem. Mathematics of Operations Research, 26(2) (2001) 421–445.

[35] Korf, L.A. and Wets, R.J.B. An ergodic theorem for stochastic programming problems. In V.H. Nguyen, J.J. Strodiot, and P. Tossings, editors, *Proceedings of the 9th Belgian-French-German Conference on Optimization*, volume 481 of *Lecture Notes in Economics and Mathematical Sciences, pages*, pages 203–217. Springer, 2000.

[36] Korf, L.A. and Wets, R.J.B. Random lsc functions: An ergodic theorem. In Stochastic Programming E-print Series (SPEPS). Humboldt-Universität, 2000.

[37] Korkmaz, S., Goksuluk, D. and Zarasiz, G. An R package for assessing multivariate normality. R Journal, 6(2) (2014) 151–162.

[38] Liehr, S., Pawelzik, K., Kohlmorgen, J. and Moler, K.R. Hidden markov mixtures of experts with an application to eeg recordings from sleep. Theory of Biosciences, 118 (1999) 246–260.

[39] Lo, M.T., Tsai, P.H., Lin, P.F., Lin, C. and Hsin, Y.L. The nonlinear and nonstationary properties in eeg signals: probing the complex fluctuations by hilbert-huang transform. Advances in Adaptive Data Analysis, 1(3) (2009) 461–482.

[40] Von Luxburg, U. and Bousquet, O. Distance-Based Classification with Lipschitz Functions. J. Mach. Learn. Res. 5 (2004) 669–695.

[41] Mardia, K.V. Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3) (1970) 519–530.

[42] Meyn, S.P. and Tweedie; R.L. Markov Chain and Stochastic Stability. Springer-Verlag, 1993.

[43] Miyato, T., Kataoka, T., Koyama, M. and Yoshida, Y. Spectral normalization for generative adversarial networks. Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[44] Neyshabur, B., Wu, Y., Salakhutdinov, R. and Srebro, N. Path-normalized optimization of recurrent neural networks with relu activations. In NIPS, 2016.

[45] Rockafellar, R.T. Integral functionals, normal integrands and measurable selections. In J. Gossez and L. Waelbroeck, editors, *Nonlinear Operators nd the Calculus of Variations*, number 543 in Lecture Notes in Mathematics, pages 157–207. Springer, 1976.

[46] Rockafellar, R.T. and Wets, R.J.B. Variational Analysis. Springer, 1998.

[47] Rojas, I. Pomares, H. and Valenzuela, O. editors. Advances in Time Series Analysis and Forecasting: Selected Contributions from ITISE 2016. Contributions to Statistics. Springer International Publishing, 2017.

[48] Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6) (1958) 386.

[49] Royston, P. Approximating the shapiro-will W test for non-normality. Statistics and Computing, 2(3) (1992) 117–119.

[50] Scaman, K. and Virmaux, A. Lipschitz regularity of deep neural networks: analysis and efficient estimation. Advances in Neural Information Processing Systems (NeurIPS), 31 (2018) 3835–3844.

[51] Shalev-Shwartz, S. and Ben-David, S. Understanding machine learning: from theory to algorithms. Cambridge University Press, 2014.

[52] Stockis, J-P., Franke, J. and Tadjuidje Kamgaing, J. On geometric ergodicity of charme models. Journal of Time Series Analysis, 31 (2010) 141–152.

[53] Stout, W.F. Almost Sure Convergence. Academic Press, New York, 1974.

[54] Szegedy, C., Zaremba, W., Sutskever, I. Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. Intriguing properties of neural networks. ICLR, 2014.

[55] Tadjuidje-Kamgaing, J. Competing neural networks as model for non-stationary financial time series. PhD thesis, University of Kaiserslautern, 2005.

[56] Taniguchi, M and Kakizawa, Y. Asymptotic Theory of Statistical Inference for Time Series. Springer-Verlag, New York, 2000.

[57] Telgarsky, M. Representation benefits of deep feedforward networks. arXiv preprint arXiv:1509.08101, 2015.

[58] Vidal, R., Bruna, J., Giryes, R. and Soatto, S. Mathematics of deep learning. In IEEE CDC, 2017.

[59] Weigend, A.S. and Shi, S. Predicting daily probability distributions of s&p500 returns. Journal of Forecasting, 19(4) (2000) 375–392.

[60] Xu, H. and Mannor, S. Robustness and generalization. Machine Learning 86, 3 (2012) 391–423.

[61] Yarotsky, D. Error bounds for approximations with deep relu networks. Neural Networks, 94 (2017) 103–114.

[62] Yarotsky, D. Quantified advantage of discontinuous weight selection in approximations with deep neural networks. arXiv preprint arXiv:1705.01365, 2017.

[63] Yoshida, Y. and Miyato, T. Spectral norm regularization for improving the generalizability of deep learning. arXiv preprint arXiv:1705.10941, 2017.

[64] Yun, C., Sra, C. and Jadbabaie, A. Global optimality conditions for deep neural newtorks. In ICLR, 2018.