!

DISCUSSION PAPERS IN ECONOMICS

Working Paper No. 24-01

! "#"$%"$&#'( ) *+ , -*"$-%(+.(/ - - , (0 -1*#'(0 -" 2 +*3%(2 $"4((/ - , -56-5"(/ #"#(

(

""""" . /01"2)&34"
54%7()8%9"&:". &;&)01&"2&<;1()"

"
"
"
"

#$%&' ()"*+,"*-*+"

"

"
"
"
"
"

Department of Economics

"

University of Colorado Boulder
Boulder, Colorado 80309

! "#$%&' ()"*+,"*-*+,". /01"2)&34"

1 Introduction

Deep neural networks (DNNs) have proven useful in the analysis of time series in economics and nance (e.g., Gu et al., 2020; Bucci, 2020; Criado-Ramon et al., 2022; Lazcano et al., 2024) and have become increasingly popular in empirical modeling (e.g., Sadhwani et al., 2021; Maliar et al., 2021; Leippold et al., 2022; Murray et al., 2024). However, the statistical properties of DNN estimators with dependent data are largely unknown, and existing results for general nonparametric estimators are often inapplicable to DNN estimators. As a result, empirical use

any continuous piecewise-linear activation function, and discusses how similar results could be obtained for alternative DNN architectures, including those with sigmoid activation functions[4].

Two results are obtained for these DNN estimators in nonparametric regression settings with mixing processes and unbounded regressors. Theorem 3 applies Theorem 1, to obtain convergence rates for the L^2-error with non-stationary -mixing data, and Theorem 4 applies Theorem 2 to obtain error bounds with stationary -mixing data. When the regressors are bounded, Theorem 4 implies a convergence rate di ering from the rate of that obtained by Farrell et al. (2021, Theorem 1) under i.i.d. data by only a poly-logarithmic factor, making this result useful for inference in some semiparametric settings, as discussed below.

The third DNN result pertains to classi cation, one of the most common applications for neural networks. I apply Theorem 1 to obtain convergence rates in logistic binomial autoregression models with covariates. A similar approach could also yield results for multinomial or non-logistic models using the ideas from Farrell et al. (2021, Lemma 9). Previous studies

Bauer and Kohler (2019), Schmidt-Hieber (2020), Kohler and Langer (2021), and dependent data Kohler and Langer (2021), Kurisu et al. (2024). The compositional nature of neural networks makes them well-suited for estimating these restricted function classes. These studies use this to obtain rapid, near minimax convergence rates, in some cases surpassing traditional nonparametric estimators (Schmidt-Hieber, 2020). While this literature o ers possible theoretical insights into why DNNs have outperformed traditional estimators in some empirical work (e.g., Gu et al., 2020; Bucci, 2020), my approach di ers by establishing a more exible framework for studying various DNN estimators in more general settings under time series data. I provide statistical properties for a common class of DNN architectures, considering both nonparametric regression and classi cation, under a Hölder smoothness condition, which is more general than these restricted classes. My work also adds generality by not placing bounds on the parameters, which can be critical for feasible implementation (see Farrell et al., 2021 for discussion).

Recently, Kurisu et al. (2024) provided closely related general results for DNN estimators and sparse-penalized adaptive DNN estimators for nonparametric regression under nonstationary -mixing data. My work adds generality to their results since they require parameter constraints, focus only on regression settings, and impose structural assumptions on the regression function when applying their ndings, although their general results do not explicitly make this restriction. While I do not address adaptive network architectures, empirical gains from sparsity penalties or other regularization techniques are often unclear (Zhang et al., 2017), and in many cases my general results could apply to similar adaptive DNNs in contexts beyond nonparametric regression, following the ideas discussed in Subsection 3.4. Theorem 3 of this paper also o ers some added generality by allowing nonstationary-mixing data, rather than -mixing. One advantage of Kurisu et al. (2024) is that they o er some consideration of -mixing coe cients with polynomial decay, whereas my DNN results only focus on mixing coe cients with exponential decay.

The rest of the paper is organized as follows. Section 2 considers a general nonparametric estimation setting and gives the main results for general sieve estimators. Section 3 describes

the class of DNN architectures considered in this paper and applies the results of Section 2 to derive properties of DNN estimators in nonparametric regression and classi cation contexts. A discussion of extensions to alternative architectures is also provided. Section 4 concludes and discusses avenues for future research. An appendix provides general measurability results for sieve estimation settings, technical proofs for all of the results presented in the paper, and a complete description of the notation used in the paper.

2 General sieve estimators

This section considers the problem of non-parametric sieve extremum estimation with depen-

$Z_t = (Y_t; X_t)$ any mapping $X_t \mapsto f(X_t)$ can trivially be defined on Z using the coordinate projection $_X(Z_t) = X_t$.

Example 1. (Nonparametric Time-Series Regression) For all $t \in \mathbb{N}$, let $Z_t := (Y_t; X_t) \in \mathcal{Z} \subseteq \mathbb{R} \times \mathbb{R}^d$ for some $d \in \mathbb{N}$, such that $Y_t \in L^2(\Omega; \mathcal{A}; P)$ and

$$Y_t = f_0(X_t) + \quad (X_t) \quad _t \tag{2.1}$$

where $\in L^2 P_{\{X_t\}_{t \in \mathbb{N}}}$, $_t \in L^2(\Omega; \mathcal{A}; P)$, and $E[\ _t | X_t] = 0$. Then, $\mathcal{F} = L^2 P_{\{X_t\}_{t \in \mathbb{N}}}$; $f_0 = E[Y_t | X_t]$; and $q(Z_t; f) = \quad Y_t \quad f(X_t)^2$. This nonparametric location-scale model includes many popular models, such as a nonlinear AR(p)-ARCH(r) model by letting $1 \quad p; r \quad d$ and $X_t = (Y_{t \quad 1}; \ldots; Y_{t \quad d})^T$, given initial conditions $Y_0; \ldots; Y_{1 \quad d}$. See Kurisu et al. (2024) for other special cases of (2.1).

Example 2. (Logistic Autoregression) For all $t \in \mathbb{N}$, let $Z_t := (Y_t; X_t)$ such that $Y_t \in \{0; 1\}$ and $X_t = (V_{t \quad 1}; Y_{t \quad 1}; \ldots Y_{t \quad r})$ for some random vector of covariates $V_t \in \mathbb{R}^{d \quad r}$ for $d > r$. Suppose for any $y \in \{0; 1\}$, $t \in \mathbb{N}$,

$$P \quad Y_t = y \quad \{V_t\}_{t=0}^{1}; Y_{t \quad 1}; Y_{t \quad 2}; \ldots \quad = P \quad Y_t = y \quad X_t \quad ;$$

and $E[Y_t | X_t] = e^{f_0} \quad 1 + e^{f_0} \quad ^1$ where $|f_0(x)| < 1$ for all $x \in \mathbb{R}^d$. Then,

$$f_0 = \log \quad \frac{E[Y_t | X_t]}{1 \quad E[Y_t | X_t]} \quad ; \quad \text{and} \quad q(Z_t; f(Z_t)) = \quad Y_t f(X_t) + \log \quad 1 + e^{f(X_t)} \quad :$$

Note that $V_{t \quad 2}; \ldots V_{t \quad r}$ can trivially be included by replacing $V_{t \quad 1}$ with $V_{t \quad 1} := (V_{t \quad 1}; \ldots V_{t \quad r})$.

2.1 Sieve extremum estimation and measurability

The target function f_0 can be estimated using the method of sieves (Grenander, 1981 Chen, 2007). This approach covers a wide variety of nonparametric estimation methods. For a textbook treatment of sieve estimation see Chen (2007).

Let $\{\mathcal{F}_n\}_{n \in \mathbb{N}}$ be a sequence of sieve spaces such that $\mathcal{F}_n \quad \mathcal{F}$, and $\sup_{f \in \mathcal{F}_n} \|f\|_1 < 1$. If $_n : \quad \to [0; 1)$ is a random variable, such that $_n = o_P(1)$, then $\hat{f}_n \in \mathcal{F}_n$ is an approximate

sieve estimator if

$$Q_n(\hat{f}_n) \quad \inf_{f \in \mathcal{F}_n} Q_n(f) + \quad {}_n: \tag{2.2}$$

${}_n$ is often referred to as the "plug-in" error, and whenever feasible ${}_n := 0$. In this case $\hat{f}_n$ is referred to as an *exact sieve estimator*.

In general, the infimum over $\mathcal{F}_n$ in (2.2), and the mapping $! \mapsto \hat{f}_n$ from to $\mathcal{F}_n$, may not be measurable when $\mathcal{F}_n$ is uncountable. Definition 1, and Propositions 1 and 2, provide easily verifiable conditions that assure measurability in many sieve estimation settings (see Section 3 and Remark 1 for example). Appendix B includes the proofs and a discussion of similar results. The definition of pointwise-separability that follows has many similar forms, see e.g., van der Vaart and Wellner (1996, Example 2.3.4).

Definition 1. (Pointwise-Separable) Let $\mathcal{G}$ be a set of functions with elements $g : \mathcal{Z} \to \mathbb{R}$ that are measurable $\mathcal{B}(\mathcal{Z})/\mathcal{B}(\mathbb{R})$. We say $\mathcal{G}$ is pointwise-separable if there is a countable subset $\{g_j\}_{j \in \mathbb{N}} \subset \mathcal{G}$ where for every $g \in \mathcal{G}$, $z \in \mathcal{Z}$, and > 0, there exists $j = j(z; ; g) \in \mathbb{N}$ such that $|g_j(z) \quad g(z)| < \quad$.

The next result uses Definition 1, to provide conditions that ensure the measurability of infimum over $\mathcal{F}_n$. Clearly, Proposition 1 will similarly apply to suprema, such as those in the proofs of this paper's results.

Proposition 1. Let $\mathcal{G}$ be a pointwise-separable class of functions. Then, for any $\in \mathbb{N}$ and $U_n : \mathbb{R}^n \to \mathbb{R}$ that is measurable $(\mathcal{A} \quad \mathcal{B}(\mathbb{R}^n))/\mathcal{B}(\mathbb{R})$, the mappings

$$! \mapsto \inf_{g \in \mathcal{H}} U_n \; !; \quad g(Z_t(!))_{t=1}^{n} \; ; \quad \forall \mathcal{H} \quad \mathcal{G} \; ; \mathcal{H} \text{ finite} \; ;$$

from to $\overline{\mathbb{R}}$, are measurable $\mathcal{A}/\mathcal{B}(\overline{\mathbb{R}})$.

With some abuse of notation, if $\mathcal{F}_n$ is pointwise-separable, then $\sup_{f \in \mathcal{H}} Q_n(f)$ is measurable, for any $\mathcal{H} \quad \mathcal{F}_n$. This follows by applying Proposition 1 to the functions

$$\inf_{f \in \mathcal{H}} U_n \quad ; \; f(Z_t(\,))_{t=1}^{n} \quad := \inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{t=1}^{n} q\,Z_t(\,); f(Z_t(\,)) \quad : \quad ! \quad \overline{\mathbb{R}}:$$

since q is measurable-$\mathcal{B}(Z) \quad \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R})$.[6]

where $\quad$ may not be induced by an L^r-norm, they require that $(\mathcal{F}_n; \quad)$ be a compact metric space. Therefore, Proposition 2 adds generality by only requiring $\mathcal{F}_n \quad \| \cdot \|_{L^r(P_{\{Z_t\}_{t=1}^n})}$ such that $\{f(z) : f \in \mathcal{F}_n\} \quad \mathbb{R}$ is compact for all $z \in \mathcal{Z}$, which does not imply that $(\mathcal{F}_n; \| \cdot \|_{L^r(P_{\{Z_t\}_{t=1}^n})})$ is a compact metric space. This will be crucial when the sieve spaces are constructed using DNNs with unbounded weights as in Section 3.

2.2 Convergence rates without stationarity

This section gives a convergence rate result for sieve estimators applicable to very general estimation settings where $\{Z_t\}_{t \in \mathbb{N}}$ is possibly non-stationary. To control the complexity of $\mathcal{F}_n$

measurability of $\hat{f}_n$. Instead, we will use outer probability, P^*, defined as in van der Vaart and Wellner (1996, §1.2), i.e., for an arbitrary set B

$$P^*(B) = \inf\left\{ P(A) : A \supseteq B;\ A \in \mathcal{A} \right\}.$$

Theorem 1. For each $n \in \mathbb{N}$ let $(\mathcal{F}, \rho_n)$ be a (semi-) metric space. Let $\mathcal{F}_n$ be a pointwise-separable class. Suppose there exist $\{\hat{f}_n\}_{n \in \mathbb{N}}$ satisfying (2.2), and $\{\delta_n\}_{n \in \mathbb{N}}$ such that $\delta_n = O_P(\varepsilon_n^2)$. Then, $\rho_n(\hat{f}_n, f_0) = O_P^*(\delta_n)$, if the following conditions hold:

(a:1) There exists a non-stochastic sequence $\{\tilde{f}_n\}_{n \in \mathbb{N}}$ such that $\tilde{f}_n \in \mathcal{F}_n$ and $\rho_n(\tilde{f}_n, f_0) \le \delta_n$, for all n.

(a:2) There exist constants $C_1, C_2 > 0$ such that, for any $n \in \mathbb{N}$ and $f \in \mathcal{F}_n$,

$$C_1 \rho_n(f, f_0)^2 \le E\, Q_n(f) - E\, Q_n(f_0) \le C_2 \rho_n(f, f_0)^2.$$

(a:3) There exist $m_n : \mathcal{Z} \to [0, 1)$, measurable $\mathcal{B}(\mathcal{Z}) = \mathcal{B}([0, 1))$, and a positive, non-decreasing sequence $\{M_n\}_{n \in \mathbb{N}}$, such that, for each $n \in \mathbb{N}$,

(i) for any $f, f' \in \mathcal{F}_n$ and $z \in \mathcal{Z}$ we have

$$q(z, f(z)) - q(z, f'(z)) \le m_n(z)\, |f(z) - f'(z)|;$$

(ii) $\lim_{n \to \infty} P^*\left(\max_{t \in \{1, \ldots, n\}} m_n(Z_t) > 0 \right)$

Condition (a:2) is a curvature condition on q near f_0 that is standard in nonparametric estimation (see e.g., Chen and Shen, 1998; Chen, 2007; Farrell et al., 2021). Condition (a:3)(i) is a Lipschitz condition on q. In many estimation settings, these requirements are met by common choices of q, such as least squares or logistic regression (Farrell et al., 2021). Conditions (a:3)(ii) and (a:3)(iii) are requirements on the tail behavior of m_n and q, which are often satis ed when

De nition 3.1, p.19), and is equivalent to many standard de nitions.

De nition 3. (-Mixing) The -mixing coe cient is de ned as,

$$(j) := E \sup \left\{ P(B \mid \{Z_t\}_1^k) - P(B) : B \in \{Z_t\}_{k+j}^1 ; k \in \mathbb{N} \right\} ; \quad \text{for } j \in \mathbb{N}:$$

We say $\{Z_t\}_{t \in \mathbb{N}}$ is -mixing (or absolutely regular) if $\lim_{j \to 1} (j) = 0$.

Conditions to ensure -mixing properties of stochastic sequences have been heavily studied (e.g., Doukhan, 1994; Bradley, 2005), and there are many examples of interesting processes that are -mixing with exponentially decreasing coe cients, also referred to as geometric-mixing. For instance, conditions to ensure stationarity and geometric-mixing for ARMA processes are given by Mokkadem (1988), and for GARCH processes by Chen and Chen (2000), and Carrasco and Chen (2002). The following corollary provides one such example, using Chen and Chen (2000, Theorem 1) with the discussion from Dinh Tuan (1986, pp. 292,293).

Corollary 1 (Chen and Chen, 2000) Consider model (2.1) with $X_t = (Y_{t-1}, \ldots, Y_{t-d})^T$. Given an initial condition $X_0 = (Y_0, \ldots, Y_{1-d})^T \in \mathbb{R}^d$, if

(i) $\{\ _t\}$ is i.i.d. with a strictly positive and continuous density, such that $E[\ _t] = 0$ and $\ _t$ is independent of X_{t-j} for all $j \in \mathbb{N}$;

(ii) the function s_0 is uniformly bounded;

(iii) there exists a constant c such that for all 0,

$$0 < c \inf_{x \in \mathbb{R}^d : \|x\|_{2;d}} (x) \sup_{x \in \mathbb{R}^d : \|x\|_{2;d}} (x) < 1 ;$$

(iv) there exist constants $C > 0$, $c_j^{(s)}$ 0, and $c_j^{(\ )}$ 0, for $j = 0, \ldots, d$

This section's main result will use the pseudo-dimension to control the complexity of the sieve spaces $\mathcal{N}_n$. The following de nition is from Bartlett et al. (2019, De nition 2).

De nition 4. (Pseudo-dimension) Let $\mathcal{S}$ be a class of functions from $\mathcal{X} \to \mathbb{R}$. The pseudo-dimension of $\mathcal{S}$, denoted as $\mathrm{Pdim}(\mathcal{S})$, is the largest $p \in \mathbb{N}$ for which there exists $(x_1, \ldots, x_p, y_1, \ldots, y_p) \in \mathcal{X}^p \times \mathbb{R}^p$ such that for any $(b_1, \ldots, b_p) \in \{0, 1\}^p$ there exists $s \in \mathcal{S}$ such that $1\{s(x_i) - y_i > 0\} = b_i$, for all $i = 1, \ldots, p$.

Pseudo-dimension, along with related complexity measures like Vapnik–Chervonenkis dimension, (see e.g. Bartlett et al., 2019, De nition 1) is often described as `scale insensitive' because, unlike the covering number, it does not depend on a speci c threshold $\delta > 0$. Scale-insensitive complexity measures are particularly well-suited with function classes constructed with DNNs that have unbounded parameters, for which Bartlett et al. (2019) provides nearly tight pseudo-dimension bounds. Pseudo-dimension can also be used to bound the covering number with Anthony and Bartlett (1999, Theorem 12.2) (also see Lemma 9 herein).

Theorem 2 extends Farrell et al. (2021, Theorem 2) to general nonparametric sieve estimators in settings with dependent data that may take values in unbounded sets. This extension relies on stationarity and β-mixing to apply a standard independent blocking technique, and an exponential inequality from Merlevede et al. (2009). Following Farrell et al. (2021), the proof of Theorem 2 di ers from classic sieve analysis (e.g., Wooldridge and White, 1991; van der Vaart and Wellner, 1996, §3.4; Chen and Shen, 1998) by using a localization approach with scale-insensitive measures of complexity to o er straightforward applicability to DNN estimators that have unbounded parameters (see Farrell et al., 2021 for more discussion).

Theorem 2. Let $\{Z_t\}_{t \in \mathbb{N}}$ be strictly stationary, $(\mathcal{F}, \|\cdot\|_{L^2(P_Z)})$ be a (semi-) metric space, $q(Z_t(\omega), \cdot) : \mathbb{R} \to \mathbb{R}$ be continuous for all $\omega \in \Omega$, and $\mathcal{F}_n$ be a pointwise-separable class such that $\{f(z) : f \in \mathcal{F}_n\} \subset \mathbb{R}$ is compact for each $z \in \mathcal{Z}$ and $n \in \mathbb{N}$. Suppose $\|f_0\|_1 < 1$

(b:2) For each $n \in \mathbb{N}$, there exists a non-stochastic $\tilde{f}_n \in \mathcal{F}_n$ where $\tilde{\varepsilon}_n := \|\tilde{f}_n - f_0\|_1$ is such that

$$\lim_{n\to\infty} \tilde{\varepsilon}_n (\log n)(\log\log n) = 0.$$

(b:3) There exist constants $C_1, C_2 > 0$ such that, $C_1 \le 1$, and for any n, $f \in \mathcal{F}_n$,

$$C_1 \|f - f_0\|^2_{L^2(P_Z)} \le E\, Q_n(f) - E\, Q_n(f_0) \le C_2 \|f - f_0\|^2_{L^2(P_Z)}.$$

(b:4) $\mathcal{F}_n$ is such that $\mathrm{Pdim}(\mathcal{F}_n) \ge 1$ for all n, and for some non-decreasing sequence $\{B_n\}_{n\in\mathbb{N}}$ such that $B_1 \ge 2$, and $\sup_{f\in\mathcal{F}_n} \|f\|_1 \le B_n < 1$ for each n, we have

$$\lim_{n\to\infty} \frac{B_n}{\sqrt{n}} \left[\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_n)\log(n)}{\ }} + \sqrt{\frac{\ }{\log\log(n)}} \right] = 0.$$

(b:5) There exists $m_n : \mathcal{Z} \to [0, 1)$, measurable $\mathcal{B}(\mathcal{Z}) = \mathcal{B}([0, 1))$, such that, for each n,

(i) for any $f, f' \in \mathcal{F}_n$, $z \in \mathcal{Z}$ we have $|q(z, f(z)) - q(z, f'(z))| \le m_n(z)\, |f(z) - f'(z)|$; and

(ii) there exists a constant $C_4 \ge 1$ and a strictly positive sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ such that

$$\min\left(B_n E\, m_n(Z_t) 1_{\{m_n(Z_t) > C_4 B_n\}} ;\ \sup_{f\in\mathcal{F}_n \cup \{f_0\}} E\left[|q(Z_t, f(Z_t)) - 1_{\{m_n(Z_t) \le C_4 B_n\}}| \right] \right) \le \varepsilon_n;$$

and $\lim_{n\to\infty} \varepsilon_n = 0$.

Then, there exists a constant $C > 0$, depending only on C_1, C_2, C_4, C, and C^0, such that for any $n \ge \max\{5, 2\mathrm{Pdim}(\mathcal{F}_n), 16B_n^2/\log(n)\}$; measurable mapping $\{\mathcal{Z}\} \mapsto \hat{f}_n$ satisfying (2.2), and constants $a \in \mathbb{N}$, $\ > 0$ where

$$\max\left\{2, \frac{\tilde{\varepsilon}_n (\log n)(\log\log n)}{B_n}\right\} < a \le n/2, \quad \text{and} \quad \sqrt{\ } \ge \frac{\tilde{\varepsilon}_n \sqrt{n}}{B_n a - \tilde{\varepsilon}_n (\log n)(\log\log n)};$$

we have

$$P\left\{ \|\hat{f}_n - f_0\|_{L^2(P_Z)} \ge C\ _n(\ ; a) \right\} \le 1 - e^{-2\log(n)} - \frac{n\ (a)}{a} + 2P\left\{ \max_{t\in\{1,\dots,n\}} m_n(Z_t) \ge C_4 B_n \right\}; \text{ and}$$

$$P\left\{ \|\hat{f}_n - f_0\|_{2;n} \ge C\ _n(\ ; a) \right\} \le 1 - 4e^{-6\log(n)} - \frac{n\ (a)}{a} + 2P\left\{ \max_{t\in\{1,\dots,n\}} m_n(Z_t) > C_4 B_n \right\}$$

where $_n(\ ; a) := B_n \sqrt{\frac{a}{n}} \left[\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_n)\log(n)}{\ }} + \sqrt{\frac{\ }{\log\log(n)}} \right] + \sqrt{\frac{\ }{\tilde{\varepsilon}_n^2}} + \ _n + \ _n.$

The existence of a measurable mapping $\{\mathcal{Z}\} \mapsto \hat{f}_n$ is not be an extra assumptim776 e06TJ/F63 113.36

The requirement that $a > \frac{\tilde{\ }_n (\log n)(\log \log n)}{B_n}$

Throughout this section, $\{Z_t = (Y_t; X_t^T)^T\}_{t \in N}$ will be a stochastic sequence on the complete probability space $(; A; P)$, such that, $Y_t \in R$ and $X_t \in [0;1]^d$ for each t. The object to be estimated is $f_0[0X$

intercept term $\gamma_{l;h;0}$,[8] and weights $\gamma_{l;h;1}, \ldots, \gamma_{l;h;H_{n;l-1}}$ such that

$$\gamma_{l;h} = [\gamma_{l;h;0}, \gamma_{l;h;1}, \ldots, \gamma_{l;h;H_{n;l-1}}]^T \in \mathbb{R}^{H_{n;l-1}+1}.$$

The parameters for the entire network are collected together as

$$\bar{\gamma}_{W_n} := \{\gamma_{l;h;k}\}_{\forall l;h;k} \in \mathbb{R}^{W_n};$$

where W_n is the total number of parameters in the network. For a set of parameters, the output of node h, in layer l, is denoted as $u_{l;h}(x \mid \bar{\gamma}_{W_n})$ and the single node in the last layer $l(= L_n + 1)$ is denoted as $u_{L_n+1;1}(x \mid \bar{\gamma}_{W_n})$ where the dependence on $\bar{\gamma}_{W_n}$ will often be suppressed. Then each node can be written as

$$u_{l;h}(x \mid \bar{\gamma}_{W_n}) := \begin{cases} \sum_{i=1}^{d} \gamma_{1;h;i} \cdot x_i + \gamma_{1;h;0}; & l = 1; \\ \sum_{i=1}^{H_{l-1}} \gamma_{l;h;i} \cdot u_{l-1;i}(x) + \gamma_{l;h;0}; & 2 \le l \le L_n; \\ \sum_{i=1}^{H_{L_n}} \gamma_{L_n+1;1;i} \cdot u_{L_n;i}(x) + \gamma_{L_n+1;1;0}; & l = L_n + 1. \end{cases} \tag{3.1}$$

With this notation, $u_{L_n+1;1} : [0,1]^d \times \mathbb{R}^{W_n} \to \mathbb{R}$, and we define the architecture $\mathcal{N}_n$ as

$$\mathcal{N}_n = \mathcal{N}(L_n; H_n; B_n) := \left\{ f = u_{L_n+1;1}(\cdot \mid \bar{\gamma}_{W_n}) : \bar{\gamma}_{W_n} \in \mathbb{R}^{W_n}; \sup_{x \in [0,1]^d} |u_{L_n+1;1}(x \mid \bar{\gamma}_{W_n})| \le B_n \right\}. \tag{3.2}$$

A particular network $f \in \mathcal{N}_n$ can be written as $f(x) = u_{L_n+1;1}(x \mid \bar{\gamma}_{W_n})$.

Remark 1. $\mathcal{N}_n$ as in (3.2) has the following useful properties:

(i) The functions $u_{L_n+1;1} : [0,1]^d \times \mathbb{R}^{W_n} \to \mathbb{R}$ are continuous. This follows because for each $l \in \{1, \ldots, L_n + 1\}$, $h \in \{1, \ldots, H_{n;l}\}$ the function $u_{l;h}$ is composed of compositions of the ReLU activation function $(\cdot)_+$ and linear combinations of $\gamma_{l;h}$ and $\{u_{l-1;h}\}_{h=1}^{H_{n;l-1}}$ (or $x \in [0,1]^d$ when $l = 1$).

(ii) $\mathcal{N}_n$ is a pointwise separable class. To see this, let $\mathbb{Q}$ denote the rationals and define $\mathcal{N}_n^{\mathbb{Q}} := \{f \in \mathcal{N}_n : \bar{\gamma}_{W_n} \in \mathbb{Q}^{W_n}\}$. Then, $\mathcal{N}_n^{\mathbb{Q}}$ is a countable dense subset of $\mathcal{N}_n$, since $u_{L_n+1;1}(x \mid \cdot) : \mathbb{R}^{W_n} \to \mathbb{R}$ is continuous for each $x \in [0,1]^d$, and $\mathbb{Q}$ is a countable dense

[8] The machine learning literature often refers to $\gamma_{l;h;0}$ as the bias, or threshold. To avoid confusion with similarly named objects in the econometrics literature I refer to this as the intercept term.

subset of $\mathbb{R}$ (for the standard topology on $\mathbb{R}$).

(iii) Note that for each n and $x \in \mathcal{Z}$ we have $\{f(x) : f \in \mathcal{N}_n\} = [-B_n; B_n]$; which is compact. To see this, consider the subset of $\mathcal{N}_n$ where all parameters are equal to zero except for the intercept term in the output node, i.e.

$$\mathcal{N}_n := \left\{ f \in \mathcal{N}_n : \theta_{L_n+1;1;0} \in [-B_n; B_n], \quad \text{and} \quad \theta_{l;h;k} = 0 \quad \forall \{l; h; k\} \in \{L_n + 1; 1; 0\} \right\}.$$

Clearly $f(x) = \theta_{L_n+1;1;0}$ for any $f \in \mathcal{N}_n$, $x \in [0; 1]^d$. Thus, $\{f(x) : f \in \mathcal{N}_n\} = [-B_n; B_n]$ for each $x \in [0; 1]^d$. Then the result follows since $\mathcal{N}_n \subset \mathcal{N}_n$ and $\sup_{f \in \mathcal{N}_n} \|f\|_1 = B_n$ by (3.2).

In what follows, the mappings $\omega \mapsto \hat{f}_n$ from Ω to $\mathcal{N}_n$ will be considered measurable. This will follow using Remark 1 since Proposition 2 will be applicable in the estimation settings considered throughout this section.

3.2 DNNs for nonparametric regression

I consider a general nonparametric regression estimation setting, that includes Example 1 as a special case. Theorem 3 will apply Theorem 1 to show the convergence of DNN estimators in a very general setting without assuming stationarity. Theorem 4 will apply Theorem 2 to obtain non-asymptotic probability bounds on the empirical and theoretical error of DNN estimators with stationary β-mixing data.

The goal is to estimate the function $f_0 = E[Y_t | X_t]$, with a DNN sieve estimator $\hat{f}_n$ as in (2.2), using the least squares criterion

$$q(Z_t; f) := (Y_t - f(X_t))^2;$$

and the DNN sieve space $\mathcal{F}_n = \{f(\theta_X(\cdot)) : f \in \mathcal{N}_n\}$. The regressands $\{Y_t\}_{t \in \mathbb{N}}$ will be assumed to satisfy the following conditions.

Assumption 2. For all $t \in \mathbb{N}$, suppose $Y_t \in L^2(\Omega; \mathcal{A}; P)$ such that there exists a non-decreasing

sequence $\{B_n\}_{n \in \mathbb{N}}$ with $B_1 \geq 2$; where

$$\lim_{n \to \infty} P\left(\max_{t \in \{1, \ldots, n\}} |Y_t| \geq B_n \right) = 0.$$

This assumption will be the key to ensuring Conditions (a:3)(ii)(iii) and (b:5)(ii) hold in this setting. The two main results of this section will also specify some additional conditions to control the dependence of $\{Z_t\}_{t \in \mathbb{N}}$ and ensure B_n grows sufficiently slowly. In both cases, these requirements will be quite general for nonparametric estimation as they will hold without Y_t taking values on a compact set (Farrell et al., 2021, Assumption 1), and without $E[Y_t^2 | \{X_t, Y_{t-1}\}_{t=1}^n]$ being uniformly bounded or $E[Y_t^{2+\delta}] < \infty$ for some $\delta > 0$ (Chen and Christensen, 2015, Assumption 2).

Assumption 2 can be verified using results from extreme value theory for a wide variety of $\{Y_t\}_{t \in \mathbb{N}}$ that are not uniformly bounded almost surely. The following proposition gives two examples that use Leadbetter et al. (1983, Theorem 3.4.1), and Leadbetter et al. (1983, Theorem 6.3.4). See Appendix E for a proof showing these results can be applied.

Proposition 3. Suppose $\{Y_t\}_{t \in \mathbb{N}}$ is α-mixing and one of the following holds:

(i) $\{Y_t\}_{t \in \mathbb{N}}$ is stationary, and $\lim_{n \to \infty} nP(|Y_1| > B_n) = 0$;[9] or

(ii) $\{Y_t\}_{t \in \mathbb{N}}$ is (possibly nonstationary) such that for each $n \in \mathbb{N}$ the joint distribution of $\{Y_t\}_{t=1}^n$ is an n-dimensional normal distribution; $\alpha(j) < (36\sqrt{2})^{-1}$ for any $j \in \mathbb{N}$; $\lim_{j \to \infty} \alpha(j) \log^2(j) = 0$; $\lim_{n \to \infty} \sum_{t=1}^n P(|Y_t| \geq B_n) = 0$; and

$$\lim_{n \to \infty} \min_{t \in \{1, \ldots, n\}} \min\left(\frac{B_n - E(-Y_t)}{\sqrt{Var(Y_t)}}, \frac{B_n - E(Y_t)}{\sqrt{Var(Y_t)}} \right) = \infty.$$

Then, in either case, $\lim_{n \to \infty} P\left(\max_{t \in \{1, \ldots, n\}} |Y_t| \geq B_n \right) = 0$.

The following lemma shows that Assumption 2 implies a form of uniform integrability that will be used to verify (b:5)(ii) and (a:3)(iii). These results are [illegible]

Lemma 1. If Assumption 2 holds then lim

there exists a constant $C > 0$ independent of n, such that for all n sufciently large

$$P \quad k\hat{f}_n - f_0 k_{L^2} \leq C \quad n \quad \geq 1 - e^{-n\left(\frac{p}{p+d}\right)(1=}$$

and for any $y \in \{0,1\}$

$$P(Y_t = y \mid V_{t-1}, Y_{t-1}, Y_{t-2}, \ldots) = P(Y_t = y \mid V_{t-1}, Y_{t-1}, \ldots, Y_{t-r}) \equiv P(Y_t = y \mid X_t);$$

and the model (3.4) satisfies

$$s(X_t, \varepsilon_t) = 0 \iff 0 < \varepsilon_t \leq E[Y_t \mid X_t].$$

Then, $\{Z_t\}_{t \in \mathbb{N}}$ is strictly stationary and α-mixing with $\alpha_Z(j) = O(e^{-C_\alpha j})$, for some $C_\alpha > 0$.

Corollary 2(i) allows general distributions for ε_t but imposes a strict exogeneity assumption with respect to $\{V_t\}_{t=0}^{1}$. Corollary 2(ii) requires $\{\varepsilon_t\}_{t=0}^{1}$ to be i.i.d. with a uniform distribution, and $\{V_t\}$ to be mixing at a faster rate, but allows for some endogeneity, since V_t can depend on ε_t, provided V_t is independent of ε_{t-j}, for $j = 1, \ldots, t$. These results are somewhat stronger than needed for Theorem 5, since the stationarity of $\{Z_t\}_{t \in \mathbb{N}}$, or $\{V_t\}_{t=0}^{1}$ is not required. Truquet (2021) suggests these results could generalize to the non-stationary case, but further work is needed to verify this.

3.4 Extensions to alternative DNN architectures

While the networks considered in the previous sections are standard, similar results can be obtained for other architectures. The key to this will be obtaining results for the complexity{ either covering number or pseudo-dimension{and approximation power of the DNN sieve spaces under consideration. This section demonstrates that this is possible and provides extensions of this paper's results to alternative feedforward DNN architectures.

Remark 2. In this section, and in the proofs for Section 3's results, we will use bounds on the the Vapnik Chervonenkis dimension (see Bartlett et al., 2019, De nition 1) of DNNs that take values in $\{0,1\}$ from Anthony and Bartlett (1999) and Bartlett et al. (2019). This is without loss of generality since these results can be directly applied to the pseudo-dimension of real valued DNNs using Anthony and Bartlett (1999, Theorem 14.1) (also see discussion following Bartlett et al., 2019, De nition 2).

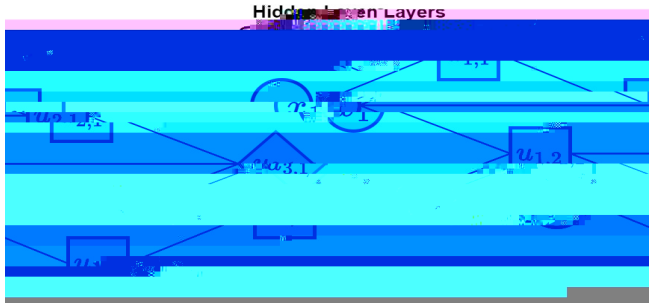

Figure 2: Example of $N_{n;'}^{FFN}$ architecture graph structure where $L_n = 2$, $W_n = 17$, and $d = 2$.

The rst result uses Yarotsky (2017, Proposition 1) to show that Theorems 3, 4, and 5 = 2.

4 Conclusion

This paper addresses the lack of statistical foundation for empirical work using deep neural

in time-series nonparametric regression, much remains to be understood, particularly for more general recurrent architectures, and dependence settings. Following the ideas of Subsection 3.4, the sieve estimator framework introduced in this paper could facilitate future research for RNNs once their approximation power and complexity are better understood.

Many other important aspects of DNNs are also not considered here, such as computational e ciency or potential gains from alternative architectures and regularization techniques. Another important class of DNNs not considered here are convolutional neural networks, which are standard in many important DNN applications, such as image recognition. The results given here could also be adapted for classes of functions beyond the standard Hölder smoothness condition, using approximation results such as Imaizumi and Fukumizu (2019). These considerations are left for future research.

Appendices

A Notation

MISC:

$\lesssim$; $\gtrsim$: For two sequences of non-negative real numbers $\{x_t\}_{t \in \mathbb{N}}$ and $\{y_t\}_{t \in \mathbb{N}}$, the notation $x_t \lesssim y_t$ means there exists a constant $0 < C < \infty$ such that $x_t \leq C y_t$ for all t sufficiently large. We write $x_t \asymp y_t$ if $x_t \lesssim y_t$ and $x_t \gtrsim y_t$.

cl(A): For a set A in a topological space, let $\mathrm{cl}(A)$ denote the closure of A; i.e. the intersection of all closed sets containing A. If $A \subset \mathbb{R}^n$, then cl(A) is with respect to the standard topology on $\mathbb{R}^n$.

$\mathcal{A} \otimes \mathcal{B}$: For two σ-algebras $\mathcal{A}, \mathcal{B}$ the product σ-algebra is $\mathcal{A} \otimes \mathcal{B} := \sigma\{a \times b : a \in \mathcal{A}, b \in \mathcal{B}\}$:

1_A : For some set X, and $A \subset X$, the indicator function is denoted as $1_A : X \to \{0, 1\}$, where $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 1$ if $x \in X \setminus A$.

SETS:

$\mathbb{N}$: The natural numbers are denoted as $\mathbb{N} = \{1, 2, \ldots\}$.

$\overline{\mathbb{R}}$: The extended real line is denoted as $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$.

$\sigma(\{X_t\}_k^n)$: For a random sequence $\{X_t\}_{t \in \mathbb{N}}$ let $\sigma(\{X_t\}_k^n)$ denote the σ-algebra generated by $\{X_t\}_{t=k}^n$.

$\mathcal{B}(X)$: For a topological space $(X, \mathcal{O}_X)$, let $\mathcal{B}(X) := \sigma(\mathcal{O}_X)$ denote the Borel σ-algebra associated with X.

MEASURES:

P_X : For a measurable space $(X, \mathcal{B}(X))$, and $X : \Omega \to X$, measurable $\mathcal{A} = \mathcal{B}(X)$, define the measure $P_X(B) = P(X^{-1}(B))$ for any $B \in \mathcal{B}(X)$.

P^* : Given a probability space $(\Omega, \mathcal{A}, P)$, define outer probability, P^*, as in van der Vaart and Wellner (1996) §1.2., i.e. for an arbitrary set B

$$P^*(B) = \inf\left\{P(A) : A \supset B, A \in \mathcal{A}\right\} :$$

NORMS:

$\|x\|_{r;a}$: For any $a \in \mathbb{N}$ and $x \in \mathbb{R}^a$ de ne the norm

$$\|x\|_{r;a} = \begin{cases} \left(\frac{1}{a} \sum_{t=1}^{a} |x_t|^r \right)^{1/r} & r \in [1;1); \\ \max_{t \in \{1;\dots;a\}} |x_t| & r = 1: \end{cases}$$

$\|f\|_{L^r}$: Let $L^r(\ ;A;P$

and White, 1992; and van der Vaart and Wellner, 1996) by providing general results that o er straightforward applicability to sieve extremum estimation. De ne a metrizable space as in Aliprantis and Border (2006, Example 2.2-3), i.e., for a topological space $(X, \mathcal{O}_X)$ the space X is metrizable if there exists a metric on X that generates the topology $\mathcal{O}_X$.

Lemma 2. Let $(\ ; \mathcal{A}; P)$ be a complete probability space, and let $\mathcal{M}$ be a complete and separable metrizable space. For $H \quad \mathcal{M}$, suppose the function $U : \quad \times H \ ! \ \overline{R}$ is measurable $(\mathcal{A} \quad \mathcal{B}(H)) = \mathcal{B}(\overline{R})$, and the correspondence $: \quad) \ H$ is such that

$$\text{graph}(\ ) := f(!; h) \ 2 \quad \times H : h \ 2 \ (\ !\ )g \ 2 \ \mathcal{A} \quad \mathcal{B}(H):$$

Let $v(!\ ) := \sup_{h2(\ !\ )} U(!; h)$; then $v : \quad ! \ \overline{R}$ is measurable $\mathcal{A} = \mathcal{B}(\overline{R})$.

Proof. Measurability of U and the assumptions on graph( ) imply, for any $c \ 2 \ \overline{R}$,

$$B_c = \quad (!; h) : U(!; h) > c; \ ! \ 2 \quad ; h \ 2 \ (\ !\ ) \ 2 \ \mathcal{A} \quad \mathcal{B}(H):$$

Then, as in Davidson (2022, p.472) equation (22.4), the projection of B_c onto is

$$v^{-1}(c; 1] := \ ! : U(!; h) > c; \ h \ 2 \ (\ !\ ) = \ ! : v(!\ ) > c :$$

If $\mathcal{A}(\mathcal{A})$ denotes the collection of all $\mathcal{A}$-analytic sets (see Corbae et al., 2009, De nition 7.9.11, p.433), then $v^{-1}(c; 1] \ 2 \ \mathcal{A}(\mathcal{A})$ by de nition, because H is a subset of a complete and separable metrizable space. Since $(; \mathcal{A}; P)$ is complete (w.r.t. P), Corbae et al. (2009, Theorem 7.9.12) implies $\mathcal{A}(\mathcal{A}) = \mathcal{A}$. Hence, $v^{-1}(c; 1] \ 2 \ \mathcal{A}$, which gives the measurability v. ■

Lemma 2 is a generalization of Stinchcombe and White (1992, Theorem 2.17-a) (also see Corbae et al., 2009, Theorem 7.9.19-1), since the measurable space $(H, \mathcal{B}(H))$ is not required to be Souslin.[11] Note that requiring $H \quad \mathcal{M}$, instead of $H = \mathcal{M}$, allows for cases where H may not be complete.

[11] At the expense of additional notation, Lemma 2 can easily be generalized to a 'measure-free' version using Corbae et al. (2009, Theorem 7.9.12). Additionally, requiring H to be a subset of a metric space, rather than measurably isomorphic to one, is without loss of generality; see the discussion following Stinchcombe and White (1992, Fact 2.6) for details.

To apply Lemma 2 in nonparametric sieve estimation settings, the correspondence is often de ned by the sieve spaces, e.g. $!() = f(f(Z_1(!)); \dots; f(Z_n(!))) : f \in F_n g$. In such a setting, Proposition 1 shows that if F_n is pointwise-separable, then Lemma 2 can be applied.

Proof of Proposition 1. First, we show the result holds when the supremum is over G. Let $_n : \ ) \ R^n$ be the correspondence $_n(!) := \ g(Z_1(!)); \dots; g(Z_n(!)) \ : g \in G \ :$ By assumption, G is a pointwise-separable class, so there exists $\{g_j\}_{j \in N} \ G$, such that $f g(z) : g \in G g = cl(f g_j(z)g_{j \in N}) \ R$, for each $z \in Z$ (see Aliprantis and Border, 2006, §2.3, p.28). Thus, for each $! \in$,

$$_n(!) = cl \ g_j(Z_1(!)); \dots; g_j(Z_n(!)) \ : j \in N \ :$$

This has two implications: rst, for all $! \in$ the correspondence $_n(!)$ is closed and consequently equal to its closure; and second by Aliprantis and Border (2006, Corollary 18.14, p.601), $_n$ is a weakly measurable correspondence. With this, Aliprantis and Border (2006, Theorem 18.6, p.596) implies that graph(cl($_n$)) $\in A \ B \ (R^n)$. Hence, graph($_n$) $\in A \ B \ (R^n)$, since graph($_n$) = graph(cl($_n$)). Then the result follows from Lemma 2.

Now, we show the result holds when the supremum is over $H \ G$, with $H \neq ;$. Let $^0_n : \ ) \ R^n$ be the correspondence $^0_n(!) := \ g(Z_1(!)); \dots; g(Z_n(!)) \ : g \in H \ :$ Since $f g_j g_{j \in N}$ is a countable dense subset of G, and $H \ G$, then $f g_j g_{j \in N} \setminus H$ is a countable dense subset of H. Hence, for all $! \in$,

$$^0_n(!) = cl \ g_j(Z_1(!)); \dots; g_j(Z_n(!)) \ : g_j \in f g_j g_{j \in N} \setminus H \ o$$

With this, the result follows using the same argument as before. ■

Proof of Proposition 2. Throughout the proof consider arbitrary $n \in N$. Let) R^n be the correspondence $_n(!) := \ g(Z_1(!)); \dots; g(Z_n(!)) : g \in G \ :$ De ne the on $_n :$ $! \ R$

$$_n(!) := \min_{x \in \ _n(!)} U_n(!; x) = \inf_{g \in G} U_n \ !; \ g(Z_t(!)) \ ^{nn \ _n}$$

$\in N$

$_n(!\ )\quad R^n$ is compact, since $f\,g(z) : g\ 2\ G\,g$ is compact for each $z\ 2\ Z$. Note that $_n$ is a weakly measurable correspondence by the argument from the proof of Proposition 1 since G is pointwise-separable and U_n is measurable-$(A\ \ B\ \ (R^n))=B(R)$ by Aliprantis and Border (2006, Lemma 4.51). With this, and Aliprantis and Border (2006, Theorem 18.19),

$$;\ 6= \quad x\ 2\ \ _n(!\ ) : U_n(!;\ x) = \ _n(!\ )$$

$$= \quad g(Z_t(!\ ))\ _{t=1}^{n} : g\ 2\ G;\ U_n\ \ !;\ \ g(Z_t(!\ ))\ _{t=1}^{n} \ = \ _n(!\ )\ \ ; \tag{B.1}$$

and there exists a function $h_n : \ \ !\ \ R^n$, measurable $A=B(R^n)$, such that for all $!\ \ 2$

$$h_n(!\ )\ 2 \quad g(Z_t(!\ ))\ _{t=1}^{n} : g\ 2\ G;\ U_n\ \ !;\ \ g(Z_t(!\ ))\ _{t=1}^{n} \ = \ _n(!\ ) \qquad R^n: \tag{B.2}$$

Note that $L^r(Z\ ;B(Z\ );P_{f\,Z_t\,g_{t=1}^n});k\ k_{L^r(P_{f\,Z_t\,g_{t=1}^n})}$ is a complete and separable metric space, i.e., a Polish space. This follows because $1\,r < \ 1$, so $L^r(Z\ ;B(Z\ );P_{f\,Z_t\,g_{t=1}^n})$ is complete by the Riez-Fisher Theorem, and is separable by Cohn (2013, Proposition 3.4.5) since $Z\ \ R^{d_Z}$ implies $B(Z\ )$ is countably generated.[12] Then, $(G;k\ k_{L^r(P_{f\,Z_t\,g_{t=1}^n})})$ is also a separable metric space since $G\ \ L^r(Z\ ;B(Z\ );P_{f\,Z_t\,g_{t=1}^n})$. Hence, there exists a countable subset $f\,g_j\,g_{j\ 2\,N}\ \ G$ such that $G = cl(\,f\,g_j\,g_{j\ 2\,N})$. Then the correspondence $_n : \ \ )\ G$ such that

$$_n(!\ ) := \ \ \ g\ 2\ G : \ \ g(Z_t(!\ ))\ _{t=1}^{n} \ = \ h_n(!\ )\ \ \ = cl\ \ \ g\ 2\ f\,g_j\,g_{j\ 2\,N} : \ \ g(Z_t(!\ ))\ _{t=1}^{n} \ = \ h_n(!\ )\ \ \ ;$$

is closed since it is equal to the closure of a set, and $_n(!\ )\ 6=\ ;$ by (B.1) and (B.2). Thus, by Aliprantis and Border (2006, Corollary 18.14-1, p.601), $_n$ is a weakly measurable correspondence. Note that $_n$ takes values in a Polish space since $G\ \ L^r(Z\ ;B(Z\ );P_{f\,Z_t\,g_{t=1}^n})$. Then, by Aliprantis and Border (2006, Theorem 18.13) the correspondence $_n(!\ )$ admits a measurable selector, i.e., there exists a function $s_n : \ \ !\ G$ that is measurable $A=B(G)$, such that $s_n(!\ )\ 2\ \ _n(!\ )$ for all $!\ \ 2\ \ $. This implies the desired result since for all $!\ \ 2$

$$s_n(!\ )\ 2\ \ _n(!\ ) \qquad g\ 2\ G : U_n\ \ !;\ \ g(Z_t(!\ ))\ _{t=1}^{n} \ = \ _n(!\ )$$

by the de nition of h_n. ■

[12] From Cohn (2013, p.102), $B(R)$ is countably generated, where we say a -algebra A is countably generated if there exists a countable subcollection C of A such that $\ \ (C) = \ A$.

C Proof of Theorem 1

By (a:3)(iii),

$$\sup_{f \in H_n} \frac{1}{n} \sum_{t=1}^{n} E\left[q_t(\tilde{f}_n) - q_t(f) \, 1_{nt}^c \right] \le \sup_{f \in H_n} \frac{1}{n} \sum_{t=1}^{n} E\left[|q_t(\tilde{f}_n)| + |q_t(f)| \, 1_{nt}^c \right]$$

$$\sup_{f \in F_n} \frac{2}{n} \sum_{t=1}^{n} E\left[q_t(f) \, 1_{nt}^c \right] \le 2C_3 \, \delta_n^2 .$$

Combining the previous two displays, for any $f \in H_n$,

$$E\left[Q_n(\tilde{f}_n) - Q_n(f) \right] \le 2C_3 \, \delta_n^2 + \frac{1}{n} \sum_{t=1}^{n} E\left[q_t(\tilde{f}_n) - q_t(f) \, 1_{nt} \right] . \tag{C.2}$$

With (C.1) and (C.2),

$$\sup_{f \in H_n} \frac{1}{n} \sum_{t=1}^{n} q_t(\tilde{f}_n) - q_t(f) \, 1_{nt}$$

$$= \sup_{f \in H_n} \frac{1}{n} \sum_{t=1}^{n} q_t(\tilde{f}_n) - q_t(f) \, 1_{nt} + E\left[Q_n(\tilde{f}_n) - Q_n(f) \right] - E\left[Q_n(\tilde{f}_n) - Q_n(f) \right]$$

$$\le \sup_{f \in H_n} \frac{1}{n} \sum_{t=1}^{n} q_t(\tilde{f}_n) - q_t(f) \, 1_{nt} - E\left[q_t(\tilde{f}_n) - q_t(f) \, 1_{nt} \right] + C_2 \le C_1 c^2 / 4 \, \delta_n^2 + 2C_3 \, \delta_n^2 .$$

■

For the proof of Theorem 1 it will be convenient to use the packing number, as defined below.

Definition 6. (Packing Number) Let $\varepsilon > 0$, and let $(M; \|\cdot\|)$ be a semi-metric space.

(i) A set $G \subset M$ is ε-separated if $\|g - g'\| \ge \varepsilon$ for any $g, g' \in G$ with $g \ne g'$. The ε-packing number, is the maximum number of ε-separated points in M.

(ii) When M is a space of functions, $f : Z \to \mathbb{R}$, for any $r \ge 1$, and $a \in \mathbb{N}$ define $D_r^{(1)}(\varepsilon; M; a) := \sup \left\{ D(\varepsilon; M | \{Z_t(\omega)\}_{t=1}^{a}; \|\cdot\|_{r;a}) : \omega \in \Omega \right\}$.

For a metric space $(M; \|\cdot\|)$ the following will be used to create a ε-cover of M using a ε-separated subset of maximum size. Note that this also implies $N(\varepsilon; M; \|\cdot\|) \le D(\varepsilon; M; \|\cdot\|)$ for any $\varepsilon > 0$ and metric space $(M; \|\cdot\|)$.

Lemma 4. For a semi-metric space $(M; \|\cdot\|)$, and any $\varepsilon > 0$, if $\{f_j\}_{j=1}^{J} \subset M$ is ε-separated

and $J = D(\ ; M; k\ k) \ 2\ N$, then

$$M \quad \bigcup_{j=1}^{J} \ n \ f \ 2 \ M : kf \quad f_j k <$$

since $c > 1$ and $\ _n(\tilde{f}_n; f_0) \quad _n$ by (a:1). De ne

$$H_n(c\ _n) := \left\{ f\ 2\ F_n : c\ _n \quad _n(f; \tilde{f}_n) \right\} :$$

By (2.2),

$$\hat{f}_n\ 2\ H_n(c\ _n) \quad =) \quad \inf_{f\ 2H_n(c\ _n)} Q_n(f) \quad Q_n(\hat{f}_n) \quad \inf_{f\ 2F_n} Q_n(f) + \ _n :$$

Using $\tilde{f}_n\ 2\ F_n$, and $\inf_{f\ 2F_n} Q_n(f) = \sup_{f\ 2F_n} Q_n(f)$; the previous display implies

$$\begin{aligned} P \quad _n(\hat{f}_n; \tilde{f}_n) \quad c\ _n \quad & P \quad Q_n(\tilde{f}_n) + \ _n \quad \inf_{f\ 2H_n(c\ _n)} Q_n(f) \\ &= P \left\{ \sup_{f\ 2H_n(c\ _n)} Q_n(\tilde{f}_n) \quad Q_n(f) + \ _n \quad 0 \right\} : \end{aligned}$$

Let $1_{nt} := 1_{\{m_n(Z_t) < M_n\}}$; and $1^c_{nt} := 1_{\{m_n(Z_t) \quad M_n\}}$: With this

where $\bar{c} > 0$ since $c > \sqrt{4(C_2 + 2C_3)/C_1}$. Then, by Lemma 3, and (C.5),

$$
\begin{aligned}
P\left({}_n(\hat{f}_n; f_0) \quad 2c_n \right) \quad & P\left(\sup_{f \in H_n(c_n)} \frac{1}{n} \sum_{t=1}^{n} \left(U_{nt}^{(q)}(\tilde{f}_n) \quad U_{nt}^{(q)}(f) \right) + {}_n \quad 2(\bar{c}_n)^2 \right) + \\
& P\left(\sup_{f \in H_n(c_n)} \frac{1}{n} \sum_{t=1}^{n} \left(U_{nt}^{(q)}(\tilde{f}_n) \quad U_{nt}^{(q)}(f) \right) \quad (\bar{c}_n)^2 \right) + P\left({}_n \quad (\bar{c}_n)^2 \right) + \\
& P\left(\sup_{f \in H_n(c_n)} \frac{1}{n} \sum_{t=1}^{n} \left(U_{nt}^{(q)}(\tilde{f}_n) \quad U_{nt}^{(q)}(f) \right) \quad (\bar{c}_n)^2 \right) + 2 \; ; \qquad \text{(C.6)}
\end{aligned}
$$

where the last inequality follows because ${}_n = O_P(\frac{2}{n})$ by assumption, and $\bar{c}$ increases with c, so $P\left({}_n \quad C_1(c_n)^2 \right)$ for any > 0, when n, and c are su ciently large.

For ${}_n := (\bar{c}_n)^2/(12M_n)$; de ne

$$
D_n(\omega) := D\left({}_n; H_n(c_n) \right)_{\{Z_t(\omega)\}_{t=1}^{n}} ; \; \| \cdot \|_{1;n} \; ;
$$

and let $\{f_{nj}\}_{j=1}^{D_n(\omega)}$ be a ${}_n$-separated set in $H_n(c_n)$ with respect to $\| \cdot \|_{1;n}$. De ne

$$
G_n^j := \left\{ f \in H_n(c_n) \;\middle|\; \|f \quad f_{nj}\|_{1;n} < {}_n \right\}; \qquad j = 1; \ldots; D_n(\omega);
$$

and by Lemma 4 we have $H_n(c_n) \subset \bigcup_{j=1}^{D_n(\omega)} G_n^j$. With this

9.9952 6o61 Tf -41067 0110 Td [(5-5458 524.0)]T [(F)82(or)]TJ/F77 11 [(;)-1]TJ9701 Tf -20.409 -9.299 Td [(j)]TJ/F
nj8 [-3d)(0h)T)8.565 0 Td [(k k)]T5H)]TJ/F61 7.9701 Tf 10.096 -1.0 Td [(5-5458 524.0)]T [(F)82(or)]TJ/F77 11 [(;)-1]TJ9701 Tf -20.409 -9.299 Td

for any

D.2 Truncation term

By stationarity, and the triangle inequality

$$\frac{1}{n}\sum_{t=1}^{n} E\left\{ g^c_{\hat{f}_n}(Z_t) - g^c_{\hat{f}_n}(Z_t) \right\} = E\left\{ g^c_{\hat{f}_n}(Z_t) - \frac{1}{n}\sum_{t=1}^{n} g^c_{\hat{f}_n}(Z_t) \right.$$

$$\leq E\left| g^c_{\hat{f}_n}(Z_t) \right| + \frac{1}{n}\sum_{t=1}^{n} g^c_{\hat{f}_n}(Z$$

D.3 Bias term

By (D.3), for any $A > 0$,

$$
\begin{aligned}
P\left(Q_n(\tilde{f}_n) - Q_n(f_0) \geq A\right) &= P\left(\frac{1}{n}\sum_{t=1}^{n} g_{\tilde{f}_n}(Z_t) + g^c_{\tilde{f}_n}(Z_t) \geq A\right) \\
&\leq P\left(\frac{1}{n}\sum_{t=1}^{n} g_{\tilde{f}_n}(Z_t) \geq A\right) + P\left(\frac{1}{n}\sum_{t=1}^{n} g^c_{\tilde{f}}(Z_t) > 0\right) \\
&\leq P\left(\frac{1}{n}\sum_{t=1}^{n} g_{\tilde{f}_n}(Z_t) \geq A\right) + P\left(\max_{t \in \{1,\ldots,n\}} m_n(Z_t) > C_4 B_n\right);
\end{aligned}
$$

By (b:5)(i) and (b:2), $E\, g_{\tilde{f}_n}(Z_t) \leq \| g_{\tilde{f}_n} \|_1 \leq C_4 B_n \| \tilde{f}_n - f_0 \|_1 \leq C_4 B_n \tilde{\varepsilon}_n$: Hence,

$$
\left\| g_{\tilde{f}_n}(Z_t) - E\, g_{\tilde{f}_n}(Z_t) \right\|_1 \leq \| g_{\tilde{f}_n} \|_1 + E\, g_{\tilde{f}_n}(Z_t) \leq 2 C_4 B_n \tilde{\varepsilon}_n:
$$

Then,

$$
\begin{aligned}
E\, g_{\tilde{f}_n}(Z_t) &= E\, q_t(\tilde{f}_n) - q_t(f_0) - E\, g^c_{\tilde{f}_n}(Z_t) \\
&\leq E\, q_t(\tilde{f}_n) - q_t(f_0) + E\, g^c_{\tilde{f}_n}(Z_t) \\
&= E\, q_t(\tilde{f}_n) - q_t(f_0) + E\, g^c_{\tilde{f}_n}(Z_t) \\
&\leq C_2 \| \tilde{f}_n - f_0 \|^2_{L^2(P_Z)} + E\, | g^c_{\tilde{f}_n}(Z_t) | \\
&\leq C_2 \tilde{\varepsilon}_n^2 + 2\, \varepsilon_n;
\end{aligned}
$$

Therefore,

$$e^{-\delta^0} + P\left(\max_{t \in \{1,\dots,n\}} m_n(Z_t) > C_4 B_n\right)$$

$$\geq P\left(Q_n(\tilde{f}_n) - Q_n(f_0) \geq C_2 \tilde{\epsilon}_n^2 + 2\epsilon_n + C_5 B_n \tilde{\epsilon}_n \left(\frac{\sqrt{(\log n)(\log\log n)}}{n} + \sqrt{\frac{\delta^0}{n}}\right)\right). \quad \text{(D.5)}$$

D.4 Independent blocks

This step constructs independent 'blocks' commonly used when dealing with α-mixing processes (e.g. Chen and Shen, 1998). By assumption, $1 \leq a \leq n/2$, so $b := \lfloor n/(2a) \rfloor$ is well defined. Then, we can divide $\{Z_t\}_{t=1}^n$ into $2b$ blocks of length a, and the remainder into a block of length $n - 2ba$, using the index sets

$$T_{1;j} := \{t \in \mathbb{N} : 2(j-1)a + 1 \leq t \leq (2j-1)a\}, \quad j = 1, \dots, b,$$
$$T_{2;j} := \{t \in \mathbb{N} : (2j-1)a + 1 \leq t \leq 2ja\}, \quad j = 1, \dots, b,$$
$$T_R := \{t \in \mathbb{N} : 2ba + 1 \leq t \leq n\}.$$

As described in Appendix F, we use Berbee's Lemma to redefine $\{Z_t\}_{t=1}^n$ on a richer probability space[13] where there exists a random sequence $\{\bar{Z}_t\}_{t=1}^n$ with the following two properties: let $T, T' \in \{T_{1;1}, T_{2;1}, T_{1;2}, \dots, T_{2;b}, T_R\}$, then (i) the block $\{\bar{Z}_t\}_{t \in T}$ is independent from the blocks $\{\bar{Z}_t\}_{t \in T'}$, $\{Z_t\}_{t \in T'}$ for any $T' \neq T$; and (ii) $\{\bar{Z}_t\}_{t \in T}$ has the same distribution as $\{Z_t\}_{t \in T}$, i.e. $P_{\{\bar{Z}_t\}_{t \in T}} = P_{\{Z_t\}_{t \in T}}$. By stationarity, all blocks of length a are identically distributed so the sequence of blocks $\{\bar{Z}_t\}_{t \in T_{1;1}}, \{\bar{Z}_t\}_{t \in T_{2;1}}, \{\bar{Z}_t\}_{t \in T_{1;2}}, \dots, \{\bar{Z}_t\}_{t \in T_{2;b}}$ is i.i.d.,[14] and we have

$$\begin{aligned} P_{\{\bar{Z}_t : t \in [\cup_{j=1}^b T_{1;j}]\}} &= P_{\{\bar{Z}_t\}_{t \in T_{1;1}}} \times P_{\{\bar{Z}_t\}_{t \in T_{1;2}}} \times \cdots \times P_{\{\bar{Z}_t\}_{t \in T_{1;b}}} \\ &= P_{\{Z_t\}_{t \in T_{1;1}}} \times P_{\{Z_t\}_{t \in T_{1;2}}} \times \cdots \times P_{\{Z_t\}_{t \in T_{1;b}}}. \end{aligned} \quad \text{(D.6)}$$

[13] We will continue to refer to the richer probability space as $(\Omega, \mathcal{A}, P)$ since the extension preserves the distribution of random variables defined on the original space. See Appendix F for details.

[14] All blocks except $\{\bar{Z}_t\}_{t \in T_R}$ are of length a, and therefore i.i.d. However, $\{\bar{Z}_t\}_{t=1}^n$ is not an independent sequence since elements within a single block $\{\bar{Z}_t\}_{t \in T}$, may be correlated. For more details see Appendix F.

Next, the usual -mixing coe cient (e.g. Dehling and Philipp, 2002, De nition 3.1, p.19) can be equivalently written as (see Eberlein, 1984)

$$(m) = \sup_{A \times B \, 2 \, \sigma(\{Z_t\}_{t=1}^{k}) \quad \sigma(\{Z_t\}_{t=k+m+1}^{1})} jP(A \times B) - P(A)P(B)j:$$

Hence, for $j \; 2 \{1; :::; b-1\}$

$$\begin{aligned} (a) \geq \sup \quad & P_{\{Z_t : t \, 2 [\, {}_{j=1}^{b} T_{1;j}\}}(A \times B) - P_{\{Z_t : t \, 2 [\, {}_{j=1}^{k} T_{1;j}\}}(A) \, P_{\{Z_t : t \, 2 [\, {}_{j=k+1}^{b} T_{1;j}\}}(B) \; : \\ & A \times B \; 2 \qquad Z_t : t \, 2 [\, {}_{j=1}^{k} T_{1;j} \qquad\qquad Z_t : t \, 2 [\, {}_{j=1}^{k} T_{1;j} \qquad : \end{aligned} \tag{D.7}$$

By (D.7) the conditions for Eberlein (1984) Lemma 2 are satis ed. So we apply this result, and use (D.6), to obtain, for any measurable set E,

$$P\left\{ Z_t : t \, 2 [\, {}_{j=1}^{b} T_{1;j} \right\} 2 \, E \; - P\left\{ \overline{Z}_t : t \, 2 [\, {}_{j=1}^{b} T_{1;j} \right\} 2 \, E \quad \leq \; (b-1) \; (a):$$

Then, by the triangle inequality and $b := \; n=(2a) \; < n=(2a) + 1$;

$$P\left\{ Z_t : t \, 2 [\, {}_{j=1}^{b} T_{1;j} \right\} 2 \, E \; \leq P\left\{ \overline{Z}_t : t \, 2 [\, {}_{j=1}^{b} T_{1;j} \right\} 2 \, E \; + \frac{n \;\; (a)}{2a}: \tag{D.8}$$

D.5 Localization analysis

We begin with some de nitions that will be used throughout the rest of this proof. De ne the following norms,

$$kf \, k_{\overline{T}_{1;j}} := \frac{1}{a} \sum_{t 2 T_{1;j}} f \;\; \overline{Z}_t \;\; {}^{2} \; {}^{1=2}; \quad \text{for } j \; 2 \{1; : : : ; b\};$$

$$kf \, k_{\overline{T}_1} := \frac{1}{b} \sum_{j=1}^{b} kf \, k^2_{\overline{T}_{1;j}} \; {}^{1=2} = \frac{1}{ba} \sum_{j=1}^{b} \sum_{t 2 T_{1;j}} f \, (\overline{Z}_t) \; {}^{2} \; {}^{1=}$$

(

1

(2

that are measurable $\mathcal{B}(\mathbb{R}^{d_W}) = \mathcal{B}(\mathbb{R})$, define

$$R_n S := \sup_{s \in S} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t s(W_t).$$

The Rademacher complexity is $E[R_n S]$; and the empirical Rademacher complexity is $E_\xi[R_n S] := E\left[R_n S \mid \{W_t\}_{t=1}^{n}\right]$.[15]

We will write $R_{ab} S = \sup_{s \in S} \sum_{j=1}^{b} \sum_{t \in T_{1;j}} \epsilon_t s(W_t)$, since the double sum is of length $a \cdot b$.

D.5.1 Step I: Quadratic process bound

Given some radius $r > 0$, to be specified later, let

$$\left\{ f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_Z)} \le r \right\};$$

throughout this section. This step will show that this implies $\|f - f_0\|_{\bar{T}_1} \le 2r$ with probability greater than $1 - e^{-\delta_0}$ for $\delta_0 > 0$ when r satisfies certain conditions.

First note that, $E\left[\|f - f_0\|^2_{\bar{T}_1} - \|f - f_0\|^2_{L^2(P_Z)}\right] = 0$; since by stationarity and (D.6)

$$\|f - f_0\|^2_{\bar{T}_1} - \|f - f_0\|^2_{L^2(P_Z)} = \frac{1}{b} \sum_{t=1}^{b} \left\{ \|f - f_0\|^2_{\bar{T}_{1;j}} - E\left[\|f - f_0\|^2_{\bar{T}_{1;j}}\right] \right\}. \tag{D.10}$$

For all $j \in \{1, \ldots, b\}$, we have

$$\|f - f_0\|^2_{\bar{T}_{1;j}} \le (B_n + 1)\|f - f_0\|_{\bar{T}_{1;j}} \le 2B_n \|f - f_0\|_{\bar{T}_{1;j}} \le 4B_n^2;$$

by (b:4), with Assumptions 2(iii), 1; and

$$\mathrm{Var}\left[\|f - f_0\|^2_{\bar{T}_{1;j}}\right] \le E\left[\|f - f_0\|^4_{\bar{T}_{1;j}}\right] \le (2B_n)^2 \|f - f_0\|^2_{L^2(P_Z)} \le 4B_n^2 r^2;$$

since

$$E_{P_{\{\bar{Z}_t\}_{t \in T_{1;j}}}}\left[f - \tilde{f}_n\right] = E_{P_{\{Z_t\}_{t \in T_{1;j}}}}\left[f - \tilde{f}_n\right] = E_{P_Z}\left[f - \tilde{f}_n\right].$$

Recall from the previous section that $\{\bar{Z}_t\}_{t \in T_{1;1}}, \{\bar{Z}_t\}_{t \in T_{1;2}}, \ldots, \{\bar{Z}_t\}_{t \in T_{1;b}}$ is an i.i.d. sequence, and consequently so is $\|f - f_0\|_{\bar{T}_{1;j}}\Big|_{j=1}^{b}$. Then, by the symmetrization inequality Bartlett et al.

[15] Note that $E[R_n S]$ is well defined by letting $\{\epsilon_t\}_{t=1}^{n}$ be defined on $(\Omega, \mathcal{A}, P)$ whenever $(\Omega, \mathcal{A}, P)$ is rich enough, otherwise we can define $\{\epsilon_t\}_{t=1}^{n}$ on an auxiliary probability space $(\Omega^{(\xi)}, \mathcal{A}^{(\xi)}, P^{(\xi)})$, and take the expectation over the product probability space $(\Omega, \mathcal{A}, P) \times (\Omega^{(\xi)}, \mathcal{A}^{(\xi)}, P^{(\xi)}) := (\Omega \times S, \mathcal{A} \otimes \mathcal{S}, P \times P^{(\xi)})$.

Note that if (D.14) holds then $2r^2 \quad r \sqrt{\frac{16aB_n^2 \quad {}^0}{n} + \frac{112aB_n^2 \quad {}^0}{(3n)}}$: Therefore, (D.12) implies that for all r such that (D.13) and (D.14) hold

$$e^{\quad 0} \quad P \left(\sup_{f f \in F_n : k f \quad f_0 k_{L^2(P_Z)} \quad r g} \left\{ k f \quad f_0 k_{\bar{T}_1}^2 \quad k f \quad f_0 k_{L^2(P_Z)}^2 \right\} \quad 3r^2 \right)^{!}$$

$$P \left(\sup_{f f \in F_n : k f \quad f_0 k_{L^2(P_Z)} \quad r g} f \right.$$

Consider P_2. By (b:5)(i), (b:4), and $\|f_0\|_1 \quad 1$,

$$\|g_{\hat{f}_n}\|_1 \quad C_4 B_n \|\hat{f}_n \quad f_0\|_1 \quad C_4 B_n (B_n + 1) \quad 2C_4 B_n^2;$$

hence $E\, g_{\hat{f}_n}(Z_t) \quad g_{\hat{f}_n}(Z_t)\,_1 \quad 4C_4 B_n^2$: Denote the cardinality of T_R as $(\# T_R) = n \quad 2ab$, and note $(\# T_R) < 2a^2 3^{-1}$ since

since $\sup_{f \in 2F_n} \|f - f_0\|_1 \le B_n + 1 \le 2B_n$. With this, and Lemma 10

$$
\begin{aligned}
E\left[R_{ab}\left(\{f - f_0 : f \in 2F_n; \|f - f_0\|_{\bar{T}_1} \le 2r_0\}\right)\right]
&= E\left[R_{ab}\left(\{f - f_0 : f \in 2F_n; \|f - f_0\|_{\bar{T}_1} \le D_n\}\right)\right] \\
&\le \inf_{0 < \delta < D_n} \left(4\delta + 8\sqrt{\frac{2}{ab}} \int_\delta^{D_n} \sqrt{\frac{\log N(\varepsilon; F_n; \|\cdot\|_{\bar{T}_1})}{\varepsilon}}\, d\varepsilon\right) \\
&\le \inf_{0 < \delta < D_n} \left(4\delta + 8\sqrt{\frac{8}{n}} \int_\delta^{D_n} \sqrt{\frac{\log N_2^{(1)}(\varepsilon; F_n; n=2)}{\varepsilon}}\, d\varepsilon\right)
\end{aligned}
$$

since $\|\cdot\|_{\bar{T}_1}$ is a sum of ab terms and $\frac{n}{4a} < b \le \frac{n}{2a}$: By assumption $n=2 > \mathrm{Pdim}(F_n)$ so $D_n < e^2 B_n n / \mathrm{Pdim}(F_n)$. Then, $N_2^{(1)}(\varepsilon; F_n; n=2) \le \left(\frac{eB_n n}{\mathrm{Pdim}(F_n)}\right)^{\mathrm{Pdim}(F_n)}$ by Lemma 9, and $\log \frac{e^2 B_n n}{\mathrm{Pdim}(F_n)} > 0$ for $0 < \varepsilon \le D_n$. With this, and the Cauchy-Schwarz inequality,[16] the previous display implies,

$$
\begin{aligned}
E\left[R_{ab}\left(\{f - f_0 : f \in 2F_n; \|f - f_0\|_{\bar{T}_1} \le 2r_0\}\right)\right]
&\le \inf_{0 < \delta < D_n} \left(4\delta + 8\sqrt{\frac{8}{n}} \int_\delta^{D_n} \sqrt{\mathrm{Pdim}(F_n) \log \frac{eB_n n}{\mathrm{Pdim}(F_n)}}\, d\varepsilon\right) \\
&\le \inf_{0 < \delta < D_n} \left(4\delta + 8\sqrt{\frac{D_n \mathrm{Pdim}(F_n)}{n}} \int_\delta^{D_n} \sqrt{\log \frac{eB_n n}{\mathrm{Pdim}(F_n)}}\, d\varepsilon\right) \\
&= \inf_{0 < \delta < D_n} \left(4\delta + 8\sqrt{\frac{D_n \mathrm{Pdim}(F_n)}{n}} \int_\delta^{D_n} \sqrt{\log \frac{e^2 B_n n}{\mathrm{Pdim}(F_n)}}\, d\varepsilon\right) \\
&= 4D_n \sqrt{\frac{\mathrm{Pdim}(F_n)}{n}} + 8\sqrt{}
\end{aligned}
$$

Hence, $\log \frac{e^2 B_n n}{D_n \mathrm{Pdim}(F_n)} < \log(n)$; since $\mathrm{Pdim}(F_n) \quad 1$ by de nition. With this, the previous display becomes

$$E \quad \overset{h}{R_{ab}} \overset{n}{f} \quad f_0 : f \quad 2 F_n; kf \quad f_0 k_{\overline{T}_1} \quad 2r_0 \quad \overset{oi}{24r_0} \quad \overset{r}{\frac{2\mathrm{Pdim}(F_n)}{n}} \log(n): \tag{D.22}$$

Combining (D.21) and (D.22),

$$E \quad R_b \quad \overline{G}^{(1)}_{j;f} : f \quad 2 F_n; kf \quad f_0 k_{\overline{T}_1} \quad 2r_0 \quad r_0 (6 \quad 24 \quad 2) C_4 B_n \quad \overset{r}{\frac{a\mathrm{Pdim}(F_n)}{n}} \log(n):$$

With this and (D.20), then using $b > \ ^{n}$

and applying (D.5), (D.4), and (D.24) yields

$$
\begin{aligned}
&5e^{\ 0} + \frac{n\ \ (a)}{a} + 2P \max_{t2f\ 1;::::;n\,g} m_n(Z_t) > C_4 B_n \\
&\quad P\ \ C_1 k\hat{f}_n\ \ f_0 k^2_{L^2(P_Z)}\ \ r_0\ \ \frac{288 C_4 B_n}{p\,\bar{n}}\ \ {}^{p\,\bar{a}}\ \ {}^{h}p\ \frac{}{P\dim(F_n)\log(n)} + p\ \frac{i}{\ 0} + \frac{95 C_4 B_n^2 a^{\ 0}}{n} \\
&\quad + C_2 \tilde{\ }_n^2 + C_5 B_n \tilde{\ }_n\ \ \frac{{}^{0}(\log n)(\log\log n)}{n} + \ \ r\ \ \frac{\#}{\ 0}\ \ \ \

Now, consider the case where

$$\frac{1}{2}\sqrt{\frac{38aB_n^2\log(n)}{n}} \quad \inf \quad r > 0 : 8s \quad r;\ s^2 > 12B_n \sqrt{2a}\, E\ R_{ab}\ f \quad f_0 : f\ 2\ F_n ; kf \quad f_0 k_{L^2} \quad s \quad ;$$

or equivalently

$$\bar{r} = \inf \quad r > 0 : 8s \quad r;\ s^2 > 12B_n \sqrt{2a}\, E\ R_{ab}\ f \quad f_0 : f\ 2\ F_n ; kf \quad f_0 k_{L^2(P_Z)} \quad s \quad :$$

In this case, note that

Now, returning to the general case, note that

$$\frac{1}{2}\sqrt{\frac{38aB_n^2\log(n)}{n}} \quad 816B_n\sqrt{\frac{aPdim(F_n)}{n}\log(n)}\,Pdim(\quad {}_n \qquad {}_n$$

by the assumption on a_n in the statement of the theorem, with this

$$\frac{B_n^2 a_0}{n} \geq B_n \tilde{B}_n \left(\frac{\alpha_0 (\log n)(\log\log n)}{n} + \sqrt{\frac{\alpha_0}{n}} \right) \iff B_n a_0^{\sqrt{\alpha_0}} \geq \tilde{\alpha}_n (\log n)(\log\log n)^{\sqrt{\alpha_0}} + \tilde{\alpha}_n \sqrt{\frac{\alpha_0}{n}}$$

$$\iff \sqrt{\alpha_0} \geq \frac{\tilde{\alpha}_n \sqrt{\frac{\alpha_0}{n}}}{B_n a_0 - \tilde{\alpha}_n (\log n)(\log\log n)};$$

which holds by the assumption on α_0 in the statement of the theorem, since $\alpha_0 > \quad$. Therefore,

$$95 C_4 \frac{B_n^2 a_0}{n} + C_5 B_n \tilde{\alpha}_n \left(\frac{\alpha_0 (\log n)(\log\log n)}{n} + \sqrt{\frac{\alpha_0}{n}} \right) \leq 96 C_5 \frac{B_n^2 a_0}{n};$$

since $C_4 < C_5 = 2C_4 = \min\{C_6; 1\}$. Next note that

$$\sqrt{\alpha_0} \leq \sqrt{\frac{\quad}{\log\ 10 \log(n) + \quad}} < \sqrt{\frac{\quad}{9 \log\log(n) + \quad}} \leq 3 \sqrt{\frac{\quad}{\log\log(n) + \quad}}$$

since $n \geq 4$ by assumption, and $J \leq 2\log(n)$ was shown previously. Using the previous two displays, we obtain

$$r = \bar{r} + \frac{4}{C_1} \left(\frac{288 C_4 B_n \sqrt{\bar{a}}}{\sqrt{\bar{n}}} \left[\sqrt{\frac{\quad}{\dim(F_n) \log(n) + \quad}} + \sqrt{\alpha_0} \right] \right.$$

$$+ \sqrt{\frac{2}{C_1}} \left(95 C_4 \quad B \right.$$

n

D.6 Empirical error bound

Note that

$$\|\hat{f}_n - f_0\|_{2;n}^2 = \left(\frac{1}{n}\sum_{j=1}^{b}\sum_{t\in T_{1;j}} \left(\hat{f}_n(Z_t) - f_0(Z_t)\right)^2\right) + \left(\frac{1}{n}\sum_{j=1}^{b}\sum_{t\in T_{2;j}} \left(\hat{f}_n(Z_t) - f_0(Z_t)\right)^2\right)$$
$$+ \frac{1}{n}\sum_{t\in T_R} \left(\hat{f}_n(Z_t) - f_0(Z_t)\right)^2 :$$

Then, by (D.30) and stationarity,

$$P\left(\|\hat{f}_n - f_0\|_{2;n} \geq \sqrt{3}C_n(;a)\right) \leq P\left(\|\hat{f}_n - f_0\|_{2;n}^2 \geq 3r^2\right)$$
$$\leq 2P\left(\frac{1}{n}\sum_{j=1}^{b}\sum_{t\in T_{1;j}} \left(\hat{f}_n(Z_t) - f_0(Z_t)\right)^2 \geq 2r^2\right) + P\left(\frac{1}{n}\sum_{t\in T_R} \left(\hat{f}_n(Z_t) - f_0(Z_t)\right)^2 \geq r^2\right)$$
$$:= 2P_3 + P_4 \tag{D.31}$$

To bound P_3 we first apply (D.8) with

$$E = \left\{\frac{1}{n}\sum_{j=1}^{b}\sum_{t\in T_{1;j}} \left(\hat{f}_n(Z_t) - f_0(Z_t)\right)^2 \geq 2r^2\right\};$$

and then $b \leq n/(2a)$; to obtain

$$P_3 \leq P\left((ab/n)\|\hat{f}_n - f_0\|_{\bar{T}_1}^2 \geq 2r^2 + \frac{n(a)}{2a}\right)$$
$$\leq P\left(\|\hat{f}_n - f_0\|_{\bar{T}_1}^2 \geq 4r^2 + \frac{n(a)}{2a}\right)$$
$$\leq P\left(\left\{\|\hat{f}_n - f_0\|_{\bar{T}_1}^2 \geq 4r^2\right\} \setminus \left\{\|\hat{f}_n - f_0\|_{L^2} \leq r\right\}\right) + P\left(\|\hat{f}_n - f_0\|_{L^2} > r + \frac{n(a)}{2a}\right)$$
$$\leq P$$

the previous display becomes

$$P_3 \leq 2e^{-\delta} + 3\log(n) \quad \frac{n \quad (a)}{a} + 2P \quad \max_{t2f\ 1;::::;n\,g} m_n($$

of -mixing. This result is also applicable to -mixing processes since $(j) \geq (j)$.

Lemma 7. Let $\{Z_t\}_{t \in \mathbb{N}}$ be an -mixing process with $(j) \leq C_\alpha^0 e^{-C_\alpha j}$ for some $C_\alpha; C_\alpha^0 > 0$.

Let $f : \mathcal{Z} \to \mathbb{R}$ be measurable, $B(\mathcal{Z}) = B(\mathbb{R})$, such that $E[f(Z_t)] = 0$ for each t and $\|f\|_\infty < B$.

Then, there exists $C_6 > 0$ depending only on $C_\alpha; C_\alpha^0$, such that for any $a \geq 3$, and $n; > 0$

Setting

$$P\left(\frac{1}{n}\sum_{t=1}^{a} f(Z_t) \geq \frac{B(\log a)(\log\log a)}{C_6 n} + B\sqrt{\frac{a}{C_6 n}}\right) \leq e^{-\delta}.$$

Proof. By Lemma 6, $\{f(Z_t)\}_{t=1}^{a}$ has an -mixing coefficient that is less than or equal to (j) from the statement of this lemma. Then, by Merlevede et al. (2009, Theorem 1),[18] there exists $C_6 > 0$ depending only on $C_\alpha; C_\alpha^0$ such that for any > 0

$$P\left(\frac{1}{n}\sum_{t=1}^{a} f(Z_t) \geq \right) \leq \exp\left(-\frac{C_6 \ ^2 n^2}{aB^2 + nB(\log a)(\log\log a)}\right).$$

Setting $= C_6 \ ^2 n^2 \left(aB^2 + nB(\log a)(\log\log a)\right)^{-1}$ implies

$$0 = C_6 \ ^2 n^2 - nB(\log a)(\log\log a) - aB^2.$$

Hence, by the quadratic formula,[19]

$$= \frac{nB(\log a)(\log\log a) + \sqrt{nB(\log a)(\log\log a)^2 + 4C_6 n^2 aB^2}}{2C_6 n^2}$$

$$\leq \frac{2nB(\log a)(\log\log a)}{2C_6 n^2} + \frac{2nB}{C_6 n^2}$$

r

obtain, for each $\in (0;r)$

$$E_\xi\left[R_n\left(\{f : f \in G\}_{\{Z_t\}_{t=1}^n}; \|f\| \le r\right)\right] = \frac{1}{\sqrt{n}} E_\xi\left[\sup_{\{f \in G; \|f\| \le r\}} \frac{1}{\sqrt{n}} \sum_{t=1}^n \xi_t f(z_t)\right]$$

$$\le \frac{1}{\sqrt{n}} 2E_\xi\left[\sup_{\{f \in G; \|f\| \le 2\alpha\}} \frac{1}{\sqrt{n}} \sum_{t=1}^n \xi_t \left(f(z_t) - f^0(z_t)\right)\right] + 8\sqrt{\frac{2}{n}} \int_\alpha^r \sqrt{\log N(\varepsilon; G; \|\cdot\|)}\, d\varepsilon$$

$$\le 2 \sup_{\{f \in G; \|f\| \le 2\alpha\}} \left(\frac{1}{n} \sum_{t=1}^n \left(f(z_t) - f^0(z_t)\right)\right) + 8\sqrt{\frac{2}{n}} \int_\alpha^r \sqrt{\log N(\varepsilon; G; \|\cdot\|)}\, d\varepsilon$$

$$\le 4 \quad + 8\sqrt{\frac{2}{n}} \int_\alpha^r \sqrt{\log N(\varepsilon; G; \|\cdot\|)}\, d\varepsilon :$$

The desired result follows since this holds for any $\{z_t\}_{t=1}^n \in Z^n$ and $\in (0;r)$. ■

E Proofs for Section 3

First, we provide proofs for Lemma 1 and Proposition 3, followed by the proofs for the main theorems of Section 3. Appendix E.4 lists the additional ancillary lemmas used in this section. As before, we write $q_t(f) := q(Z_t; f(Z_t))$; and $m_{nt} := m_n(Z_t)$:

Proof of Lemma 1. Note that $\sup_{t \in \mathbb{N}} Y_t^2$ is measurable $\mathcal{A} = \mathcal{B}(\overline{\mathbb{R}})$. Then,

$$\max_{t \in \{1;\dots;n\}} E\left[Y_t^2 1_{\{|Y_t| \ge B_n\}}\right] \le E\left[\max_{t \in \{1;\dots;n\}} \{Y_t^2\} \cdot \max_{t \in \{1;\dots;n\}} \{1_{\{|Y_t| \ge B_n\}}\}\right]$$

$$\le \int_{\{\omega : \max_{1 \le t \le n} |Y_t(\omega)| \ge B_n\}} \sup_{t \in \mathbb{N}} Y_t^2\, dP: \tag{E.1}$$

For any > 0, there exists a simple function,[23] $s_\delta : \to \mathbb{R}$, such that $0 \le s_\delta \le \sup_{t \in \mathbb{N}} Y_t^2$, and

$$\int \sup_{t \in \mathbb{N}} Y_t^2\, dP - \int s_\delta\, dP \le /2: \tag{E.2}$$

We may choose $C_\delta > 0$ such that $\sup_{\omega \in} s_\delta(\omega) \le C_\delta$, since a simple function takes on nitely many values. By Assumption 2 for any constants; $C_\delta > 0$ there exists $N_\delta \in \mathbb{N}$ such that, for all

[23] As usual, a simple function is a function that can be represented as $s(\omega) = \sum_{j=1}^J c_j 1_{A_j}$; for some $J \in \mathbb{N}$, and $A_j \in \mathcal{A}$, $c_j \in \mathbb{R}$ for $j = 1; \dots; J$.

$n \geq N_\delta$,

$$P \max_{t \in \{1,\dots,n\}} |Y_t| \geq B_n \leq \eta/(2C_\delta). \tag{E.3}$$

By construction, $s_\delta \leq \sup_{t \in \mathbb{N}} Y_t^2$ so

$$\int_{\{\omega : \max_{1 \leq t \leq n} |Y_t(\omega)| \geq B_n\}} \sup_{t \in \mathbb{N}} Y_t^2 - s_\delta \, dP \leq \int \sup_{t \in \mathbb{N}} Y_t^2 - s_\delta \, dP.$$

Hence, for all $n \geq N_\delta$,

$$\begin{aligned}
\int_{\{\omega : \max_{1 \leq t \leq n} |Y_t(\omega)| \geq B_n\}} \sup_{t \in \mathbb{N}} Y_t^2 \, dP &\leq \int_{\{\omega : \max_{1 \leq t \leq n} |Y_t(\omega)| \geq B_n\}} s_\delta \, dP + \int \sup_{t \in \mathbb{N}} Y_t^2 \, dP - \int s_\delta \, dP \\
&\leq C_\delta P \max_{t \in \{1,\dots,n\}} |Y_t| \geq B_n + \int \sup_{t \in \mathbb{N}} Y_t^2 \, dP - \int s_\delta \, dP \\
&\leq \eta/2 + \int Y_t^2 \, dP - \int s_\delta \, dP \\
&\leq \eta;
\end{aligned}$$

by (E.3) and (E.2). Applying this to (E.1) completes the proof since η is arbitrary, N_δ depends only on η; C_δ and s_δ; C_δ are independent of n. ■

Proof of Proposition 3. For condition (i), note that $\{Y_t\}_{t \in \mathbb{N}}$ is stationary and α-mixing by Lemma 6. Then, $\{Y_t\}_{t \in \mathbb{N}}$ satisfies Leadbetter et al. (1983, Condition $D(u_n)$, p.53) (see discussion on p.54 therein). With this, the desired result follows from the proof of Leadbetter et al. (1983, Theorem 3.4.1) (also see discussion on p.58 therein).

For condition (ii), we first verify that the conditions of Leadbetter et al. (1983, Theorem 6.3.4, p.132) are met. Let $\{\overline{Y}_t\}_{t \in \mathbb{N}}$ be the standardized version of $\{Y_t\}_{t \in \mathbb{N}}$, i.e., $\overline{Y}_t := \frac{Y_t - E(Y_t)}{\sqrt{Var(Y_t)}}$. Note that $E\,\overline{Y}_t^4 = 3$ for all $t \in \mathbb{N}$, by Papoulis (1991, p.110). By Lemma 6, $\alpha_{\overline{Y}}(j) \leq \alpha(j)$ since $Y_t \mapsto \overline{Y}_t$ is measurable $\mathcal{A} = \mathcal{B}(\mathbb{R})$. With this, and since $\overline{Y}_t$ has a standard normal distribution, we apply Bosq (1998, Corollary 1.1) to obtain for any $i, k \in \mathbb{N}$, $i \neq k$,

$$|Cor(\overline{Y}_i; \overline{Y}_k)| = |Cov(\overline{Y}_i; \overline{Y}_k)| \leq 4\sqrt{2\alpha(|i-k|)}\,\|\overline{Y}_i\|_{L^4}\|\overline{Y}_k\|_{L^4} \leq 36\sqrt{2\alpha(|i-k|)} := \rho_{|i-k|};$$

and $|Cor(-Y_i; -Y_k)| \leq \rho_{|i-k|}$ by a similar argument. Then, by the assumptions on $\alpha(j)$ we have

$_j < 1$ for all $j \geq N$ and $\lim_{j \to 1} {}_j \log(j) = 0$. Next

$$\sum_{t=1}^{n} P\left(\overline{Y}_t \geq \frac{B_n - E(Y_t)}{\sqrt{Var(Y_t)}}\right) = \sum_{t=1}^{n} P(Y_t \geq B_n) \leq \sum_{t=1}^{n} P(|Y_t| \geq B_n) \to 0; \quad \text{and}$$

$$\sum_{t=1}^{n} P\left(-\overline{Y}_t \geq \frac{B_n - E(-Y_t)}{\sqrt{Var(Y_t)}}\right) = \sum_{t=1}^{n} P(-Y_t \geq B_n) \leq \sum_{t=1}^{n} P(|Y_t| \geq B_n) \to 0;$$

as $n \to 1$ by assumption. Thus, Leadbetter et al. (1983, Theorem 6.3.4, p.132) can be applied to $\{\overline{Y}_t\}_{t \in N}$ and $\{-\overline{Y}_t\}_{t \in N}$ to obtain

$$P\left(\max_{t \in \{1,\ldots,n\}} Y_t \geq B_n\right) = P\left(\bigcup_{t=1}^{n} \{Y_t \geq B_n\}\right) = P\left(\bigcup_{t=1}^{n} \left(\overline{Y}_t \geq \frac{B_n - E(Y_t)}{\sqrt{Var(Y_t)}}\right)\right) \to 0; \quad \text{and}$$

$$P\left(\max_{t \in \{1,\ldots,n\}} -Y_t \geq B_n\right) = P\left(\bigcup_{t=1}^{n} \{-Y_t \geq B_n\}\right) = P\left(\bigcup_{t=1}^{n} \left(-\overline{Y}_t \geq \frac{B_n - E(-Y_t)}{\sqrt{Var(Y_t)}}\right)\right) \to 0;$$

as $n \to 1$. Which gives the desired result since

$$\begin{aligned}
P\left(\max_{t \in \{1,\ldots,n\}} |Y_t| \geq B_n\right) &= P\left(\left\{\max_{t \in \{1,\ldots,n\}} Y_t \geq B_n\right\} \cup \left\{\min_{t \in \{1,\ldots,n\}} Y_t \leq -B_n\right\}\right) \\
&= P\left(\left\{\max_{t \in \{1,\ldots,n\}} Y_t \geq B_n\right\} \cup \left\{-\max_{t \in \{1,\ldots,n\}} -Y_t \leq -B_n\right\}\right) \\
&= P\left(\left\{\max_{t \in \{1,\ldots,n\}} Y_t \geq B_n\right\} \cup \left\{\max_{t \in \{1,\ldots,n\}} -Y_t \geq B_n\right\}\right) \\
&\leq P\left(\max_{t \in \{1,\ldots,n\}} Y_t \geq B_n\right) + P\left(\max_{t \in \{1,\ldots,n\}} -Y_t \geq B_n\right) \to 0;
\end{aligned}$$

as $n \to 1$. ■

E.1 Proof of Theorem 3

Theorem 3 will follow by showing the conditions of Theorem 1 are met with the following setting:

- $q_t(f) = q(Z_t; f(Z_t)) := \left({}_Y(Z_t) - f({}_X(Z_t))\right)^2 = (Y_t - f(X_t))^2$;
- ${}_n(f; f^0) := \|f - f^0\|_{L^2(P_{\{X_t\}_{t=1}^n})} = \left(\int_{[0;1]^d} |f(x) - f^0(x)|^2 dP_{\{X_t\}_{t=1}^n}\right)^{1/2}$;
- $M_n := 4B_n$; and
- $m_n(Z_t) := 2 \sup_{f \in N_n} |Y_t - f(X_t)|$:

Veri cation of (a

$n^{-\left(\frac{p}{p+d=2}\right)(1=4-K_B)}$: By assumption $_n$ & $n^{\left(\frac{p}{p+d=2}\right)(1=4-K_B)}$: Thus, there exist constants $C; C^0 > 0$ such that $kf\tilde{}_n$

since $Z_t(\omega) := (Y_t(\omega), X_t(\omega)) \in \mathbb{R} \times [0,1]^d$. Thus, for any $\omega \in \Omega$, $n \in \mathbb{N}$,

$$\max_{t \in \{1,\dots,n\}} m_n(Z_t(\omega)) \le \max_{t \in \{1,\dots,n\}} 2|Y_t(\omega)| + B_n,$$

which implies

$$\begin{aligned}\{\omega : \max_{t \in \{1,\dots,n\}} m_n(Z_t(\omega)) > M_n\} &\subset \{\omega : \max_{t \in \{1,\dots,n\}} 2|Y_t(\omega)| + B_n) > M_n\} \\ &= \{\omega : \max_{t \in \{1,\dots,n\}} |Y_t(\omega)| > M_n/2 - B_n\} \\ &= \{\omega : \max_{t \in \{1,\dots,n\}} |Y_t(\omega)| > B_n\}, \qquad \text{(E.5)}\end{aligned}$$

since $M_n := 4B_n$. With this, and Assumption 2

$$P(\max_{t \in \{1,\dots,n\}} m_n(Z_t) > M_n) \le P(\max_{t \in \{1,\dots,n\}} |Y_t| > B_n) \to 0, \quad \text{as } n \to \infty.$$

Consider (a3)(iii). Note that (E.5) implies $1_{\{m_n(Z_t) > M_n\}} \le 1_{\{|Y_t| > B_n\}}$, for all $n \in \mathbb{N}$ and $t \in \{1,\dots,n\}$. Hence, for any $f \in \mathcal{N}_n$,

$$\begin{aligned}E[q_t(f) 1_{\{m_n(Z_t) > M_n\}}] &\le E[(Y_t^2 + f(X_t)^2 - 2f(X_t)Y_t) 1_{\{|Y_t| > B_n\}}] \\ &\le E[Y_t^2 1_{\{|Y_t| > B_n\}}] + E[f(X_t)^2 1_{\{|Y_t| > B_n\}}] + 2E[f(X_t)Y_t 1_{\{|Y_t| > B_n\}}] \\ &\le E[Y_t^2 1_{\{|Y_t| > B_n\}}] + B_n^2 P(|Y_t| > B_n) + 2B_n E[|Y_t| 1_{\{|Y_t| > B_n\}}], \qquad \text{(E.6)}\end{aligned}$$

where the last line has used $\sup_{f \in \mathcal{N}_n} \|f\|_\infty \le k_1 B_n$ by (3.2). Note that, with $B_n > 0$ and Markov's inequality,

$$B_n^2 P(|Y_t| > B_n) = B_n^2 P(|Y_t| 1_{\{|Y_t| > B_n\}} > B_n) \le B_n^2 P(Y_t^2 1_{\{|Y_t| > B_n\}} > B_n^2) \le E[Y_t^2 1_{\{|Y_t| > B_n\}}],$$

and,

$$B_n E[|Y_t| 1_{\{|Y_t| > B_n\}}] \le E[Y_t^2 1_{\{|Y_t| > B_n\}}],$$

since $B_n|Y_t| 1_{\{|Y_t| > B_n\}} \le Y_t^2$ when $|Y_t| \ge B_n$, and $B_n|Y_t| 1_{\{|Y_t| > B_n\}} = 0$ when $|Y_t| < B_n$. Using the previous two displays with (E.6),

$$\max_{t \in \{1,\dots,n\}} \sup_{f \in \mathcal{N}_n} E[q(Z_t; f(Z_t)) 1_{\{m_n(Z_t) > M_n\}}] \le 4 \max_{t \in \{1,\dots,n\}} E[Y_t^2 1_{\{|Y_t| > B_n\}}] \le \varepsilon_n^2, \qquad \text{(E.7)}$$

by the assumptions on ε_n. Then, $\lim_{n \to \infty} \varepsilon_n = 0$ under Assumption 2 by Lemma 1 and the assumptions on ε_n and K_B. ■

Verification of (a:4). First, $m_n(Z_t) := 2 \sup_{f \in \mathcal{N}_n} |Y_t - f(X_t)|$; implies

$$q_t(f)1_{\{|m_{nt}| < M_n\}} = |Y_t - f(X_t)|^2 1_{\{|m_{nt}| < M_n\}} \le m_n(Z_t)^2 1_{\{|m_{nt}| < M_n\}} \le M_n^2 = 4 = B_n^2$$

For each $n \in \mathbb{N}$, by Lemma 6 $\{q_t(f)1_{\{|m_{nt}| < M_n\}}\}_{t=1}^n$ has an α-mixing coefficient that is bounded above by the α-mixing coefficient for $\{Z_t\}_{t=1}^n$. Then, using the same reasoning as the proof of Lemma 7, by Merlevede et al. (2009, Theorem 1) there exists a constant $C^0 > 0$ depending only on C, C^0 such that such that for $\varepsilon > 0$ and all $n \ge 4$,

$$\begin{aligned} P\left(\frac{1}{n}\sum_{t=1}^n \left\{q_t(f)1_{\{|m_{nt}| < M_n\}} - E\, q_t(f)1_{\{|m_{nt}| < M_n\}}\right\} > \varepsilon\right) &\le \exp\left(-\frac{C^0 \varepsilon^2 n^2}{8nB_n^4 + 2\varepsilon nB_n^2(\log n)(\log\log n)}\right) \\ &\le \exp\left(-\frac{C \varepsilon^2 n}{n^{4K_B} + \varepsilon n^{2K_B}(\log n)(\log\log n)}\right) \\ &=: \Phi_n^{(q)}(\varepsilon); \end{aligned}$$

for some constant $C > 0$ not depending on n or ε since $B_n \asymp n^{K_B}$ by assumption.

Consider (a4)(ii). Note that

$$\Phi_n^{(q)}(\varepsilon^2 B_n^2) = \exp\left(-\frac{C \varepsilon^2 B_n^4 n}{n^{4K_B} + \varepsilon^2 B_n^2 n^{2K_B}(\log n)(\log\log n)}\right):$$

With this, and $B_n \asymp n^{\left(\frac{p}{p+d/2}\right)(1/4 - K_B)}\log^{2+\varepsilon}(n)$ by assumption,

$$\begin{aligned} \frac{B_n^4 n}{n^{4K_B} + n^{2K_B}(\log n)(\log\log n)} &= \left[n^{(1-4K_B)} B_n^4 + n^{(1-2K_B)} B_n^2(\log n)(\log\log n)\right]^{-1} \\ &\asymp n^{(1-4K_B)} B_n^4 \\ &\asymp n^{\left(1-\frac{p}{p+d/2}\right)(1/4 - K_B)}\log^{4(2+\varepsilon)}(n) \to 1 \quad ; \quad \text{as } n \to \infty \quad ; \end{aligned}$$

since $1 - \frac{p}{p+d/2} > 0$ and $K_B < 1/4$ by assumption. Then, Lemma 12 can be applied, with $4M_n = B_n$ and the definition of $\Phi_n^{(q)}$, to obtain the following sufficient condition for (a:4)(ii),

$$\begin{aligned} &\lim_{n\to\infty} \left\{n^{(1-4K_B)} B_n^{-4} \log N_1^{(1-\varepsilon)}\left(4\varepsilon^2 B_n^2/B_n; N_n; n\right)\right\} = 0; \\ \Rightarrow\ &\lim_{n\to\infty} \left\{\Phi_n^{(q)}(\varepsilon^2 B_n^2)\, N_1^{(1-\varepsilon)}\left(\varepsilon^2 B_n^2/M_n; N_n; n\right)\right\} = 0: \end{aligned} \tag{E.8}$$

Henceforth, let n be large enough such that $B_n < 1$. By (3.3) and Lemma 14, we have $\mathrm{Pdim}(N_n) \lesssim n^{2\left(\frac{d}{p+d/2}\right)(1/4 - K_B)}\log^7(n)$; which implies $\lim_{n\to\infty}\{\mathrm{Pdim}(N_n)/n\} = 1$. Then, we can

apply Lemma 15, with $= \frac{d}{p + d=2}$ (1=4

Then by Assumption 2 and Lemma 1, we have $\lim_{n\to\infty} \varepsilon_n = 0$. By (E.7) and stationarity, $E\,|q_t f|\,1_{\{|m_n(Z_t)| \le 4B_n\}} \le 4E\,Y_t^2 1_{\{|Y_t| > B_n\}} \le \varepsilon_n$. Hence, (b5)(ii) holds for $C_4 = 4$. ■

Final Steps for Proposition 4. Now, we verify the remaining requirements for Theorem 2. By Remark 1(ii) and (iii), $\mathcal{N}_n$ is pointwise-separable, and $\{f(x) : f \in \mathcal{N}_n\} = [-B_n; B_n] \subset \mathbb{R}$ is compact for each $x \in [0;1]^d$. Under Assumption 1, $\|f_0\|_1 \le 1$.

Choose

$$:= C\, n^{2\left(\frac{d}{p+d}\right)(1/2-\kappa_B)} \log^8(n); \quad \text{and} \quad a := n^2 \log^2(n) :$$

for some $C > 0$. To apply Theorem 2 all that remains is to verify $\rho - \dfrac{\tilde{\varepsilon}_n \sqrt{n}}{B_n a \tilde{\varepsilon}_n (\log n)(\log\log n)}$:

Note that by (b:2) $\lim_{n\to\infty} \tilde{\varepsilon}_n (\log n)(\log\log n) = 0$ and by (b:4) $B_n a \ge 3$ for all n, so we have $B_n a \tilde{\varepsilon}_n (\log n)(\log\log n) \le B_n a$: Using this with $\tilde{\varepsilon}_n \asymp n^{\left(\frac{p}{p+d}\right)(1/2-\kappa_B)}$ from the proof for (b:2), and $B_n \ge n^{\kappa_B}$ by assumption, we have

$$\frac{\tilde{\varepsilon}_n \sqrt{n}}{B_n a \tilde{\varepsilon}_n (\log n)(\log\log n)} \asymp \frac{n^{1/2-\left(\frac{p}{p+d}\right)(1/2-\kappa_B)}}{B_n a} \ge \frac{n^{1/2-\left(\frac{p}{p+d}\right)(1/2-\kappa_B)}}{B_n n}$$

$$\ge n^{(1/2-\kappa_B)-\left(\frac{p}{p+d}\right)(1/2-\kappa_B)} = n^{\left(\frac{d}{p+d}\right)(1/2-\kappa_B)}:$$

Therefore,

$$\rho - n^{\left(\frac{d}{p+d}\right)(1/2-\kappa_B)} \log^4(n) > n^{\left(\frac{d}{p+d}\right)(1/2-\kappa_B)} \ge \frac{\tilde{\varepsilon}_n \sqrt{n}}{B_n a \tilde{\varepsilon}_n (\log n)(\log\log n)};$$

and the desired result follows by choosing C su ciently large.

Thus, Theorem 2 can be applied. To obtain the rate from Proposition 4 note that $\mathrm{Pdim}(\mathcal{N}_n) \log(n) \le n^{2\left(\frac{d}{p+d}\right)(1/2-\kappa_B)} \log^8(n)$; by (E.9) and Lemma 14. Hence,

$$\begin{aligned}
\varphi_n(\varepsilon; a) &:= B_n \sqrt{\frac{a}{n}} \left[\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_n)\log(n)}{\varepsilon}} + \sqrt{\frac{\log\log(n)}{\varepsilon^2}} + \sqrt{\frac{\tilde{\varepsilon}_n + \varepsilon_n + \varepsilon_n}{\varepsilon^2}} \right] \\
&\le B_n \sqrt{\frac{a}{n}}\, n^{\left(\frac{d}{p+d}\right)(1/2-\kappa_B)} \log^4(n) + \sqrt{\frac{\tilde{\varepsilon}_n + \varepsilon_n + \varepsilon_n}{\varepsilon^2}} \\
&\le n^{\left(\frac{p}{p+d}\right)(1/2-\kappa_B)} \log^5(n) + \sqrt{\frac{\tilde{\varepsilon}_n + \varepsilon_n}{\varepsilon^2}} \\
&\le n^{\left(\frac{p}{p+d}\right)(1/2-\kappa_B)} \log^5(n) + \sqrt{\frac{E\,Y_t^2 1_{\{|Y_t| > B_n\}}}{\varepsilon^2}}
\end{aligned}$$

max E

(2009

E.4 Supporting lemmas

Lemma 11. Suppose Assumption 1 holds and let $\mathcal{N}_n = N(L_n; H_n; B_n)$ bting

[redacted]

The[redacted]y assumption,

$$\lim_{n\to\infty} \frac{1}{b_n} \log \frac{a_n}{e^{b_n}} + 1 = \lim_{n\to\infty} \frac{\log(a_n)}{b_n} = 0;$$

wh[redacted]nplies

$$\lim_{n\to\infty} \frac{1}{b_n} \log \frac{a_n}{e^{b_n}} = 1$$

Ho[redacted]r, $b_n > 0$ for all n, and $b_n \to \infty$, then $\lim_{n\to\infty} 1/b_n =$ [redacted] that[24]

$$\lim_{n\to\infty} \left\{ \log \frac{a_n}{e^{b_n}} \right\} = 1;$$

her[redacted]$e^{b_n} g = 0$: ■

Lemma 13. Let $N_n = N(L_n; H_n; B_n)$ be de ned as in (3.2) where the sequences $\{L_n\}_{n\in\mathbb{N}}$, $\{H_{n,l}\}_{n\in\mathbb{N}}$ for each $l \in \mathbb{N}$, are non-decreasing, and $H_{n,l} \le H_n$ for all $l \in \mathbb{N}$, and [redacted]

Lemma 14. Let $N_n = N(L_n; H_n; B_n)$ be defned as in (3.2) where the sequences $\{L_n\}_{n \geq N}$, $\{H_{n,l}\}_{n \geq N}$ for each $l \geq N$, are non-decreasing, $H_{n,l} \quad H_n$ for all $l \geq N$; and

$$L_n \quad \log(n); \quad H_n \quad n^{\eta} \log^2(n); \quad \text{for some} \quad > 0:$$

Then, $\mathrm{Pdim}(N_n) \quad n^{2\eta} \log^7(n)$:

Proof. By Bartlett et al. (2019, Theorems 3 and 7),[25] there exist constants $c; C > 0$ such that, for all $n \geq N$,

$$c\, W_n L_n \log(W_n/L_n) \leq \mathrm{Pdim}(N_n) \leq C\, W_n L_n \log(W_n):$$

Using this, with $W_n \quad H_n^2 L_n$ by Lemma 13, and $L_n \quad \log(n)$; $H_n \quad n^{\eta} \log^2(n)$ by assumption, we obtain

$$\mathrm{Pdim}(N_n) \;\&\; W_n L_n \log(W_n/L_n) \quad H_n^2 L_n^2 \log(H_n) \quad n^{2\eta} \log^6(n) \log \quad n^{\eta} \log^2(n)$$

$$n^{2\eta} \log^6(n) \quad \log(n) + \log\log(n) \quad n \log(W \;($$

r 2 [1; 1]

$$\log N_r^{(1\ )}\ \ {}_n;\ N_n;\ a_n\ \ .\ \ n^2\ \ \log^7(n)\ \ \Big[\log(n) + \log(a_n) + \log\ \ {}_n^{\ 1}\Big]$$

Proof. By assumption, a_n Pdim(N_n) for all n su ciently large. Hence, Lemma 9 can be applied to obtain, for some $C > 0$,

$$N_1^{(1\ )}\ \ {}_n;\ N_n;\ a_n\ \ .\ \ \frac{2eB_n a_n}{{}_n\ \ \mathrm{Pdim}(N_n)}^{\mathrm{Pdim}(N_n)}\ \ \frac{2eB_n a_n}{{}_n\ \ Cn^2\ \ \log^7(n)}^{Cn^2\ \ \log^7(n)}; \qquad \text{(E.13)}$$

where the last bound follows from Pdim(N_n) . n^2 $\log^7(n)$ by Lemma 14, with

$$\frac{2eB_n a_n}{{}_n\ \ x} > e \quad =) \quad \frac{@}{@x}\ \ \frac{2eB_n a_n}{{}_n\ \ x}^{x} = \log\ \ \frac{2eB_n a_n}{{}_n\ \ x}\ \ 1\ \ \frac{2eB_n a_n}{{}_n\ \ x}^{x} > 0;$$

and

$$\frac{2eB_n a_n}{{}_n\ \ \mathrm{Pdim}(N_n)} > e; \quad 8\, n \text{ su ciently large};$$

since B_n is non-decreasing, ${}_n$ 1 for all n su ciently large, and $\lim_{n!1} a_n = \mathrm{Pdim}(N_n) = 1$: By (E.13)

$$\begin{aligned}
\log\ N_1^{(1\ )}\ \ {}_n;\ N_n;\ a_n\ \ &.\ \ n^2\ \ \log^7(n)\ \ \log\ \ \frac{2eB_n a_n}{{}_n\ \ n^2\ \ \log^7(n)} \\
&.\ \ n^2\ \ \log^7(n)\ \ \Big[\log(B_n) + \log(a_n) + \log\ \ {}_n^{\ 1}\Big]\ \ (2\ d=p)\log(n) \\
&.\ \ n^2\ \ \log^7(n)\ \ \Big[\log(n) + \log(a_n) + \log\ \ {}_n^{\ 1}\Big];
\end{aligned}$$

since B_n . n^{K_B} by assumption. ■

F Independent block construction

This section more rigorously describes the process used to construct the sequence $\{\bar{Z}_t\}_{t=1}^n$ from Appendix D.4. De ne $Y_1 := \{Z_t\}_{t2T_{1;1}}$, and $X_1 := \{Z_t\}_{t2ff\ 1;::::;n\,gnT_{1;1}g}$. Clearly, X_1 and Y_1 are random variables on $(\ ; A; P)$ taking values in $X := Z^{n\ \ a}$ and $Y := Z^{a}$, respectively. Now, let be the Lebesgue measure, and consider the product probability space

$$(\ \ {}^0; A^0; P^0) = (\ \ ; A; P) \qquad [0;1]; B_{[0;1]}; \qquad := \qquad [0;1]; A\ \ B_{[0;1]}; P \qquad :$$

We can extend $(; A ; P)$ to this richer space with the extension[26] $: {}^{0} !$, where denotes coordinate projection onto , i.e. $(! _{1} ; ! _{2}) = ! _{1} 2$ for any $(! _{1} ; ! _{2}) 2 {}^{0}$. Thus, $X _{1}$ and $Y _{1}$ on $(; A ; P)$ can be redefined as $X _{1}$ and $Y _{1}$ on $({}^{0} ; A^{0} ; P^{0})$, without changing their distribution. We may refer to $X _{1}$ and $Y _{1}$ as random variables on $(; A ; P)$ as random variables.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. 2019. Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks. Journal of Machine Learning Research 20, 63 (2019), 1{17. http://jmlr.org/papers/v20/17-612.html

Benedikt Bauer and Michael Kohler. 2019. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. The Annals of Statistics 47, 4 (Aug. 2019), 2261{2285. https://doi.org/10.1214/18-AOS1747 Publisher: Institute of Mathematical Statistics.

Henry C. P. Berbee. 1979. Random walks with stationary increments and renewal theory Number 112 in Mathematical Centre tracts. Mathematisch Centrum, Amsterdam. https://core.ac.uk/download/301669772.pdf

Pedro Henrique Borghi, Oleksandr Zakordonets, and Joao Paulo Teixeira. 2021. A COVID-19 time series forecasting model based on MLP ANN. Procedia Computer Science 181 (Jan. 2021), 940{947. https://doi.org/10.1016/j.procs.2021.01.250

D. Bosq. 1998. Nonparametric Statistics for Stochastic Processes. Lecture Notes in Statistics,

Xiaohong Chen. 2007. Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models. In *Handbook of Econometrics*, James J. Heckman and Edward E. Leamer (Eds.). Vol. 6. Elsevier, 5549{5632. https://doi.org/10.1016/S1573-4412(07)06076-X

Xiaohong Chen and Timothy M. Christensen. 2015. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics* 188, 2 (Oct. 2015), 447{465. https://doi.org/10.1016/j.jeconom.2015.03.010

Xiaohong Chen and Xiaotong Shen. 1998. Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica* 66, 2 (1998), 289{314. https://doi.org/10.2307/2998559 Publisher: [Wiley, Econometric Society].

Donald L. Cohn. 2013. *Measure Theory: Second Edition*. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6956-8

Dean Corbae, Maxwell B. Stinchcombe, and Juraj Zeman. 2009. *An Introduction to Mathematical Analysis for Economic Theory and Econometrics*. Princeton University Press.

D. Criado-Ramón, L.G.B. Ruiz, and M.C. Pegalajar. 2022. Electric demand forecasting with neural networks and symbolic time series representations. *Applied Soft Computing* 122 (June 2022), 108871. https://doi.org/10.1016/j.asoc.2022.108871

James Davidson. 2022. *Stochastic Limit Theory: An Introduction for Econometricians* (second edition, second edition ed.). Oxford University Press, Oxford, New York.

Herold Dehling and Walter Philipp. 2002. Empirical Process Techniques for Dependent Data. In *Empirical Process Techniques for Dependent Data*

Ana Lazcano, Miguel A. Jaramillo-Morán, and Julio E. Sandubete. 2024. Back to Basics: The Power of the Multilayer Perceptron in Financial Time Series Forecasting. Mathematics 12, 12 (Jan. 2024), 1920. https://doi.org/10.3390/math12121920 Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

M. R. Leadbetter, Georg Lindgren, and Holger Rootzén. 1983. Extremes and Related Properties of Random Sequences and Processes. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-5449-2

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. Nature 521, 7553 (May 2015), 436–444. https://doi.org/10.1038/nature14539 Publisher: Nature Publishing Group.

Markus Leippold, Qian Wang, and Wenyu Zhou. 2022. Machine learning in the Chinese stock market. Journal of Financial Economics 145, 2, Part A (Aug. 2022), 64–82. https://doi.org/10.1016/j.jfineco.2021.08.017

Y. Makovoz. 1998. Uniform Approximation by Neural Networks. Journal of Approximation Theory 95, 2 (Nov. 1998), 215–228. https://doi.org/10.1006/jath.1997.3217

Lilia Maliar, Serguei Maliar, and Pablo Winant. 2021. Deep learning for solving dynamic economic models. Journal of Monetary Economics 122 (Sept. 2021), 76–101. https://doi.org/10.1016/j.jmoneco.2021.07.004

Andreas Maurer. 2016. A vector-contraction inequality for Rademacher complexities. http://arxiv.org/abs/1605.00251 arXiv:1605.00251 [cs, stat].

Florence Merlevède and Magda Peligrad. 2002. On the Coupling of Dependent Random Variables and Applications. In Empirical Process Techniques for Dependent Data, Herold Dehling, Thomas Mikosch, and Michael Sørensen (Eds.). Birkhäuser, Boston, MA, 171–193. https://doi.org/10.1007/978-1-4612-0099-4_5

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. 2009. Bernstein inequality and moderate deviations under strong mixing conditions. In Institute of Mathematical Statistics Collections. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 273–292. https://doi.org/10.1214/09-IMSCOLL518

Abdelkader Mokkadem. 1988. Mixing properties of ARMA processes. Stochastic Processes and their Applications 29, 2 (Sept. 1988), 309–315. https://doi.org/10.1016/0304-4149(88)90045-2

Scott Murray, Yusen Xia, and Houping Xiao. 2024. Charting by machines. Journal of Financial Economics 153 (March 2024), 103791. https://doi.org/10.1016/j.jfineco.2024.103791

A. Papoulis. 1991. Probability, Random Variables, and Stochastic Processes (3 ed.). McGraw-Hill.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. http://arxiv.org/abs/1611.03530 arXiv:1611.03530.