Estimation and Inference with Deep Neural Networks Under Dependent Data

by

Chad Brown

B.G.S., University of Kansas, 2017

M.A., University of Colorado, 2021

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Economics

2024

Committee Members:

Carlos Martins-Filho, Chair

Yu-Jui Huang

Xiaodong Liu

Adam McCloskey

Alessandro Peri

Brown, Chad (Ph.D., Economics)

Estimation and Inference with Deep Neural Networks Under Dependent Data

Thesis directed by Prof. Carlos Martins-Filho

Abstract

This dissertation studies nonparametric estimation with deep neural networks (DNNs) under dependent data. The first chapter introduces my work, describes its contribution to the literature, and defines the mathematical notation used throughout.

The second chapter establishes statistical properties of deep neural network (DNN) estimators under dependent data. Two general results for nonparametric sieve estimators directly applicable to DNN estimators are given. The first establishes rates for convergence in probability under nonstationary data. The second provides non-asymptotic probability bounds on $\mathcal{L}^2$-errors under stationary β-mixing data. I apply these results to DNN estimators in both regression and classification contexts imposing only a standard Hölder smoothness assumption. The DNN architectures considered are common in applications, featuring fully connected feedforward networks with any continuous piecewise linear activation function, unbounded weights, and a width and depth that grows with sample size. The framework provided also offers potential for research into other DNN architectures and time-series applications.

The third chapter demonstrates the practical implications of these results in a partially linear regression model under stationary β-mixing data. In this setting, I obtain $\sqrt{n}$-asymptotic normality of the finite dimensional parameter after first-stage DNN estimation of infinite dimensional parameters.

Acknowledgements

I am extremely grateful to my advisor, Carlos Martins-Filho for his guidance and mentorship in my doctoral work. His careful and methodical approach to research taught me to think rigorously about problems, and helped me produce work that I can be confident in. I could not have asked for a better advisor, his passion for mathematical rigor is something I will carry with me throughout my research career.

I am deeply indebted to Alessandro Peri, whose mentorship early in my Ph.D. career was pivotal in my development as a researcher. I cannot thank him enough for his patience and dedication during the many hours of video calls and meetings over the years.

Thank you to my committee for their careful reading of this manuscript and helpful suggestions. A special thanks to Adam McCloskey, who served as my advisor at the outset of my research in econometrics, and helped me start the project that ultimately became this dissertation.

I extend my heartfelt gratitude to my parents, Craig and June Brown, for their endless support and encouragement throughout my academic career. I would also like to thank my friends, including Greg Pach and Sam Tankel, for the many welcome distractions and the countless evenings spent playing music together over the years.

Contents

Chapter

1 Introduction and Mathematical Notation **1**

1.1 Introduction . . . 1

1.2 Notation . . . 6

2 Statistical Properties of Deep Neural Networks with Dependent Data **8**

2.1 General sieve estimators . . . 8

2.1.1 Sieve extremum estimation and measurability . . . 10

2.1.2 Convergence rates without stationarity . . . 12

2.1.3 Nonasymptotic error bounds with stationarity and β-mixing . . . 15

2.2 DNN estimators . . . 19

2.2.1 DNN sieve spaces ($\mathcal{N}_n$) . . . 20

2.2.2 DNNs for nonparametric regression . . . 23

2.2.3 DNNs for binomial autoregressions with covariates . . . 26

2.2.4 Extensions to alternative DNN architectures . . . 28

2.3 Summary and extensions . . . 31

3 Inference in Partially Linear Models under Dependent Data with Deep Neural Networks **33**

3.1 Estimation procedure and results . . . 34

3.2 Summary and extensions . . . 37

Bibliography **38**

Appendix

A Appendix for Chapter 2 **44**

A.1 Measurability of Extrema of Random Functions 44

A.2 Proof of Theorem 2.1.1 47

A.3 Proof of Theorem 2.1.2 53

A.3.1 Main Decomposition 53

A.3.2 Truncation Term 54

A.3.3 Bias Term 55

A.3.4 Independent Blocks 56

A.3.5 Localization Analysis 57

A.3.6 Empirical Error Bound 70

A.3.7 Supporting Lemmas 72

A.4 Independent Block Construction 76

A.5 Proofs for Section 2.2 76

A.5.1 Proof of Theorem 2.2.1 79

A.5.2 Proof of Theorem 2.2.2 82

A.5.3 Proof of Theorem 2.2.3 85

A.5.4 Supporting Lemmas 88

B Appendix for Chapter 3 **92**

B.1 Preliminary Lemmas 92

B.2 Proof of Theorem 3.1.1 101

Figures

Figure

2.1 Example of $\mathcal{N}_n$ architecture graph structure where $L_n = 2$, $H_{n,1} = 3$, $H_{n,2} = 2$, $W_n = 20$, and $d = 2$. 21

2.2 Example of $\mathcal{N}_{n,\varphi}^{\text{FFN}}$ architecture graph structure where $L_n = 2$, $W_n = 17$, and $d = 2$. . 30

Chapter 1

Introduction and Mathematical Notation

1.1 Introduction

Deep neural networks (DNNs) have proven useful in the analysis of time series in economics and finance (e.g., Gu et al., 2020; Bucci, 2020; Criado-Ramón 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 applications with DNN estimators often lack a theoretical foundation.

Chapter 2 aims to address this deficiency by first providing general results for nonparametric sieve estimators that offer a framework that is flexible enough for studying DNN estimators under dependent data. These results are then applied to both nonparametric regression and classification contexts, yielding theoretical properties for a class of DNN architectures commonly used in applications. Chapter 3 demonstrates the practical implications of these results in a partially linear regression model with dependent data by obtaining $\sqrt{n}$-asymptotic normality of the estimator for the finite dimensional parameter after first-stage DNN estimation of infinite dimensional parameters.

DNN estimators can be viewed as adaptive linear sieve estimators, where inputs are passed

through hidden layers that 'learn' basis functions from the data by optimizing over compositions of simpler functions.[1] Some general conditions that are sufficient to obtain statistical properties of certain nonparametric estimators have been studied under independent and identically distributed data (i.i.d.) (e.g., Shen and Wong, 1994; Chen, 2007), and dependent data (Wooldridge and White, 1991; Chen and Shen, 1998; Chen and Christensen, 2015). Different from the extant literature, I provide two results for general sieve estimators that apply to DNN estimators in settings with dependent data that take values on unbounded sets. Theorem 2.1.1 provides rates of convergence in probability in a setting similar to that of Wooldridge and White (1991), which differs from previous results, such as those in Chen and Shen (1998) and Chen and Christensen (2015), by not requiring stationarity. Theorem 2.1.2 extends Farrell et al. (2021, Theorem 2) beyond DNNs and i.i.d. settings, to provide non-asymptotic probability bounds on both the theoretical and empirical $\mathcal{L}^2$-errors of general sieve estimators under stationary β-mixing data. These results are well suited to the study of DNN estimators for two key reasons. First, they accommodate general sieve extremum estimation and are not restricted to series methods treated by Chen and Christensen (2015), as verifying basis function properties is impractical with DNNs' adaptive structure. Second, they avoid conditions on the sieve spaces, relying on entropy with bracketing or interpolation between $\mathcal{L}^\infty$ and $\mathcal{L}^2$ norms (e.g., Chen and Shen, 1998, Conditions A.3 and A.4), which are not feasible for DNNs when network depth diverges with sample size.

Using these general results, I derive statistical properties for DNN estimators with architectures that reflect modern applications: (i) fully connected feedforward networks with continuous piece-wise linear activation functions; (ii) no parameter constraints; and (iii) depth and width that grow with sample size.[2] While early research focused on shallow, often single-layer networks with smooth activation functions (e.g., White and Gallant, 1992; Makovoz, 1998; Anthony and Bartlett,

[1] See Subsection 2.2.1 for a description of DNN architectures. See Chen (2007) for a treatment of sieve estimators, and Farrell et al. (2021) for more discussion on framing DNN estimators in the context of more familiar nonparametric estimation procedures.

[2] Fully connected feedforward neural networks with more than three hidden layers are often referred to as multilayer perceptrons in the DNN literature.

1999), modern applications favor deep networks with many hidden layers (Szegedy et al., 2016; Schmidt-Hieber, 2020). To mitigate the increased computational demands of deep networks, modern implementations do not impose parameter constraints and often use non-smooth activation functions (e.g., Glorot et al., 2011). Among DNN architectures, fully connected feedforward DNNs are standard in practice (Almeida, 2020; Criado-Ramón et al., 2022), and are frequently applied in time-series settings (e.g., Dudek, 2016; Borghi et al., 2021; AlShafeey and Csáki, 2021). Recently, the most popular activation function has been the rectified linear unit (ReLU), $\phi(x) = \max\{0, x\}$ (LeCun et al., 2015), which will be the main focus of this chapter's DNN results.[3] However, Subsection 2.2.4 shows that my results apply to DNNs with 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 2.2.1 applies Theorem 2.1.1, to obtain convergence rates for the $\mathcal{L}^2$-error with nonstationary α-mixing data, and Theorem 2.2.2 applies Theorem 2.1.2 to obtain error bounds with stationary β-mixing data. When the regressors are bounded, Theorem 2.2.2 implies a convergence rate differing 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.

A third DNN result pertains to classification, one of the most common applications for neural networks. I apply Theorem 2.1.1 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 on DNN estimators in classification contexts have considered their empirical performance with dependent data (see Fawaz et al., 2019 for a review) or their statistical properties in i.i.d. settings (e.g., Kim et al., 2021; Yara

[3] Compared to smooth activation functions, ReLU activation functions have also been shown to offer improved properties both empirically (e.g., Sadhwani et al., 2021) and theoretically (e.g., Glorot et al., 2011).

[4] Recurrent neural networks are also an important class of DNNs for time series settings not considered here. See Subsection 2.3 for a brief discussion of these architectures.

and Terada, 2024; and citations therein). To the best of my knowledge, this work is the first to derive their statistical properties with dependent data.

Much of the recent theoretical work for DNN estimators focuses on estimating regression functions with additive or hierarchical structures under i.i.d. data (see, Kohler and Krzyżak, 2017; Bauer and Kohler, 2019; Schmidt-Hieber, 2020; Kohler and Langer, 2021) and under dependent data (see, Kohler and Langer, 2021; Kurisu et al., 2024). The compositional functional form of neural networks makes them well-suited for estimating these restricted function classes. These studies use these restrictions to obtain fast, near minimax convergence rates, in some cases surpassing traditional nonparametric estimators (Schmidt-Hieber, 2020). While this literature offers possible theoretical insights into why DNNs have outperformed traditional estimators in some empirical work (e.g., Gu et al., 2020; Bucci, 2020), my approach differs by establishing a more flexible 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 classification, under a Hölder smoothness condition, which allows for greater generality relative to 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 β-mixing data. My work adds generality since they impose parameter constraints, focus only on regression settings, and impose structural assumptions on the regression function when applying their findings, although their general results do not explicitly require 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 2.2.4. Theorem 2.2.1 of this work also offers some added generality by allowing nonstationary α-mixing data, rather than β-mixing. One advantage of Kurisu et al.

(2024) is that they offer some consideration of β-mixing coefficients with polynomial decay, whereas my DNN results require exponential decay.

Finally, Chapter 3 demonstrates how my results enable valid large-sample inference in partially linear regression models with dependent data (see Robinson, 1988). I establish $\sqrt{n}$ asymptotic normality of the finite-dimensional parameter following first-stage DNN estimation of infinite-dimensional components. Using an approach similar to Chen et al. (2022), I do this without sample-splitting which is often impractical with dependent data. This exercise demonstrates the practical implications of the DNN results from Chapter 2 and suggests potential applications to more complex econometric models. In addition, it contributes to two growing bodies of literature. First, the literature on inference after machine learning (e.g., Chernozhukov et al., 2018; Farrell et al., 2021; Chernozhukov et al., 2022) by considering inference after DNN estimation in dependent data settings. Second, the work that studies partially linear time series models (see Gao, 2007; and Härdle et al., 2000; for a review) by incorporating DNNs to estimate nonparametric components.

1.2 Notation

MISC:

$\lesssim, \asymp$: 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 $x_t \gtrsim y_t$.

$\mathrm{cl}(A)$: 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 $A \subseteq \mathbb{R}^n$, then $\mathrm{cl}(A)$ 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 $\mathcal{A} \otimes \mathcal{B} \coloneqq \sigma\big(\{a \times b : a \in \mathcal{A}, b \in \mathcal{B}\}\big)$.

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

SETS:

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

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

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

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

MEASURES:

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

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., $P^*(B) = \inf\Big\{P(A) : \ A \supset B, \ A \in \mathcal{A}\Big\}$, for an arbitrary set $B \subset \Omega$.

NORMS:

$\|\boldsymbol{x}\|_{r,a}$: For any $a \in \mathbb{N}$ and $\boldsymbol{x} \in \mathbb{R}^a$ define the norm

$$\|\boldsymbol{x}\|_{r,a} = \begin{cases} \left(\frac{1}{a}\sum_{t=1}^{a} |x_t|^r\right)^{1/r} & r \in [1,\infty), \\ \max_{t\in\{1,\ldots,a\}} |x_t| & r = \infty. \end{cases}$$

$\|f\|_{\mathcal{L}^r}$: Let $\mathcal{L}^r(\Omega, \mathcal{A}, P)$ denote the space of functions $f : \Omega \to \mathbb{R}$ that are measurable-$\mathcal{A}/\mathcal{B}(\mathbb{R})$, such that $\|f\|_{\mathcal{L}^r(\Omega,\mathcal{A},P)} < \infty$, for the (pseudo-) norms

$$\|f\|_{\mathcal{L}^r(\Omega,\mathcal{A},P)} \coloneqq \begin{cases} \left(\int_\Omega |f|^r dP\right)^{1/r}, & \text{for } 1 \le r < \infty, \\ \inf\left\{C \ge 0 : P(\{\omega \in \Omega : |f(\omega)| \ge C\}) = 0\right\}, & \text{for } r = \infty. \end{cases}$$

We write $\mathcal{L}^r(P)$ or $\|f\|_{\mathcal{L}^r(P)}$ when no confusion may arise.

$\|f\|_\infty$: For a function $f : \mathbb{X} \to \mathbb{R}$ define $\|f\|_\infty \coloneqq \sup_{\boldsymbol{x}\in\mathbb{X}} |f(\boldsymbol{x})| \in \overline{\mathbb{R}}$.

COMPLEXITY MEASUREMENTS:

$\mathrm{N}(\delta, \mathcal{G}, \|\cdot\|)$: See Definition 2.1.2 for the definition of covering number.

$\mathrm{Pdim}(\mathcal{S})$: See Definition 2.1.4 for the definition of pseudo-dimension.

$\mathrm{D}(\delta, \mathcal{G}, \|\cdot\|)$: See Definition A.2.1 for the definition of packing number.

$\mathfrak{R}_n\mathcal{S}$: See Definition A.3.1 for the definition of Rademacher complexity.

MIXING COEFFICIENTS

$\beta(j), \alpha(j)$: See Definitions 2.1.3 and 2.2.1 for β and α mixing, respectively (also see Dehling and Philipp, 2002, Definition 3.1). For a random sequence $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ we write the mixing coefficients as $\beta_{\boldsymbol{Z}}(j), \alpha_{\boldsymbol{Z}}(j)$, or simply $\beta(j), \alpha(j)$ when no confusion may arise.

Chapter 2

Statistical Properties of Deep Neural Networks with Dependent Data

This chapter provides my general results for nonparametric sieve estimators that offer a framework that is flexible enough for studying DNN estimators under dependent data. These results are then applied to both nonparametric regression and classification contexts, yielding theoretical properties for a class of DNN architectures commonly used in applications. This work can also be found in Brown (2024b).

Section 2.1 considers a general nonparametric estimation setting and gives the main results for general sieve estimators. Section 2.2 describes the class of DNN architectures considered in this dissertation and applies the results of Section 2.1 to derive properties of DNN estimators in nonparametric regression and classification contexts. A discussion of extensions to alternative architectures is also provided. Section 2.3 concludes and discusses avenues for future research. Appendix A provides general measurability results for sieve estimation settings and technical proofs for all results in this chapter.

2.1 General sieve estimators

Let $(\Omega, \mathcal{A}, P)$ be a complete probability space, and $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ be a stochastic sequence on $(\Omega, \mathcal{A}, P)$, with coordinates given by random vectors $\boldsymbol{Z}_t : \Omega \to \boldsymbol{\mathcal{Z}} \subseteq \mathbb{R}^{d_Z}$ for some $d_Z \in \mathbb{N}$ and each t. The parameter space, $\mathcal{F}$, is a space of functions with elements $f : \boldsymbol{\mathcal{Z}} \to \mathbb{R}$ measurable-$\mathcal{B}(\boldsymbol{\mathcal{Z}})/\mathcal{B}(\mathbb{R})$. Let $q : \boldsymbol{\mathcal{Z}} \times \mathbb{R} \to \mathbb{R}$ be the single observation criterion function which is assumed to

be measurable-$\big(\mathcal{B}(\boldsymbol{\mathcal{Z}}) \otimes \mathcal{B}(\mathbb{R})\big)/\mathcal{B}(\mathbb{R})$, and $q\big(\boldsymbol{Z}_t(\cdot), f(\boldsymbol{Z}_t(\cdot))\big) \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$, for each t and $f \in \mathcal{F}$.[1] The empirical criterion function is

$$Q_n(f) = Q\big(\{\boldsymbol{Z}_t\}_{t=1}^n, f\big) \coloneqq \frac{1}{n}\sum_{t=1}^{n} q(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)), \qquad \text{for } \ n \in \mathbb{N},\ f \in \mathcal{F}.$$

The true parameter $f_0 \in \mathcal{F}$ is defined by $\mathbb{E}[Q_n(f_0)] \leq \mathbb{E}[Q_n(f)]$, for all $f \in \mathcal{F}$.

This setting covers a wide range of non/semi-parametric models. I give two examples that illustrate it's breadth. Example 2.1.1 is the regression model from Kurisu et al. (2024), and Example 2.1.2 considers a classification problem in a logistic binomial autoregression model. Generalizations of these examples will be considered in Section 2.2 as applications of this section's results. To consider these examples with the notation used above, note that when $\boldsymbol{Z}_t = (Y_t, \boldsymbol{X}_t)$ any mapping $\boldsymbol{X}_t \mapsto f(\boldsymbol{X}_t)$ can trivially be defined on $\boldsymbol{\mathcal{Z}}$ using the coordinate projection $\pi_{\boldsymbol{X}}(\boldsymbol{Z}_t) = \boldsymbol{X}_t$.

Example 2.1.1 (Nonparametric time-series regression)**.** For all $t \in \mathbb{N}$, let $\boldsymbol{Z}_t \coloneqq (Y_t, \boldsymbol{X}_t) \in \boldsymbol{\mathcal{Z}} \subseteq \mathbb{R} \times \mathbb{R}^d$ for some $d \in \mathbb{N}$, such that $Y_t \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ and

$$Y_t = f_0(\boldsymbol{X}_t) + \eta(\boldsymbol{X}_t)\upsilon_t \tag{2.1}$$

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

Example 2.1.2 (Logistic autoregression)**.** For all $t \in \mathbb{N}$, let $\boldsymbol{Z}_t \coloneqq (Y_t, \boldsymbol{X}_t)$ such that $Y_t \in \{0, 1\}$ and $\boldsymbol{X}_t = (\boldsymbol{V}_{t-1}, Y_{t-1}, \ldots Y_{t-r})$ for some random vector of covariates $\boldsymbol{V}_t \in \mathbb{R}^{d-r}$ for $d > r$. Suppose for any $y \in \{0, 1\}$, $t \in \mathbb{N}$,

$$P\Big(Y_t = y \,\Big|\, \{\boldsymbol{V}_t\}_{t=0}^{\infty}, Y_{t-1}, Y_{t-2}, \ldots\Big) = P\Big(Y_t = y \,\Big|\, \boldsymbol{X}_t\Big),$$

[1] Requiring that q be defined on $\boldsymbol{\mathcal{Z}} \times \mathbb{R}$ instead of the subset $\boldsymbol{\mathcal{Z}} \times \mathcal{F}(\boldsymbol{\mathcal{Z}})$, where $\mathcal{F}(\boldsymbol{\mathcal{Z}}) \coloneqq \cup_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}}\{f(\boldsymbol{z}) : f \in \mathcal{F}\} \subseteq \mathbb{R}$, is only for notational convenience later on and is without loss of generality by Stinchcombe and White (1992, Lemma 2.14).

and $\mathbb{E}[Y_t|\boldsymbol{X}_t] = e^{f_0}\left[1+e^{f_0}\right]^{-1}$ where $|f_0(\boldsymbol{x})| < \infty$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Then,

$$f_0 = \log\left(\frac{\mathbb{E}[Y_t|\boldsymbol{X}_t]}{1-\mathbb{E}[Y_t|\boldsymbol{X}_t]}\right), \quad \text{and} \quad q(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)) = -Y_t f(\boldsymbol{X}_t) + \log\left(1+e^{f(\boldsymbol{X}_t)}\right).$$

Note that $\boldsymbol{V}_{t-2}, \dots \boldsymbol{V}_{t-r}$ can trivially be included by replacing $\boldsymbol{V}_{t-1}$ with $\boldsymbol{V}^*_{t-1} \coloneqq (\boldsymbol{V}_{t-1}, \dots \boldsymbol{V}_{t-r})$.

2.1.1 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.

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

$$Q_n(\hat{f}_n) \leq \inf_{f\in\mathcal{F}_n} Q_n(f) + \theta_n. \tag{2.2}$$

θ_n is often referred to as the "plug-in" error, and whenever feasible $\theta_n \coloneqq 0$. 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 $\omega \mapsto \hat{f}_n$ from Ω to $\mathcal{F}_n$, may not be measurable when $\mathcal{F}_n$ is uncountable. Definition 2.1.1, and Propositions 2.1.1 and 2.1.2, provide easily verifiable conditions that assure measurability in many sieve estimation settings (see Section 2.2 and Remark 2.2.1 for example). Appendix A.1 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 2.1.1 (Pointwise-separable)**.** Let $\mathcal{G}$ be a set of functions with elements $g : \boldsymbol{\mathcal{Z}} \to \mathbb{R}$ that are measurable-$\mathcal{B}(\boldsymbol{\mathcal{Z}})/\mathcal{B}(\mathbb{R})$. The set $\mathcal{G}$ is pointwise-separable if there is a countable subset $\{g_j\}_{j\in\mathbb{N}} \subseteq \mathcal{G}$ where for every $g \in \mathcal{G}$, $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, and $\delta > 0$, there exists $j = j(\boldsymbol{z}, \delta, g) \in \mathbb{N}$ such that $|g_j(\boldsymbol{z}) - g(\boldsymbol{z})| < \delta$.

The next result uses Definition 2.1.1, to provide conditions that ensure the measurability of infimum over $\mathcal{F}_n$. Clearly, Proposition 2.1.1 will similarly apply to suprema, such as those in the proofs of this work's results.

Proposition 2.1.1. *Let $\mathcal{G}$ be a pointwise-separable class of functions. Then, for any $n \in \mathbb{N}$ and $U_n : \Omega \times \mathbb{R}^n \to \mathbb{R}$ that is measurable-$(\mathcal{A} \otimes \mathcal{B}(\mathbb{R}^n))/\mathcal{B}(\mathbb{R})$, the mappings*

$$\omega \mapsto \inf_{g \in \mathcal{H}} U_n\Big(\omega\,, \big\{g(\boldsymbol{Z}_t(\omega))\big\}_{t=1}^n\Big), \quad \forall\, \mathcal{H} \subseteq \mathcal{G},\ \mathcal{H} \neq \emptyset,$$

from Ω 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 $\mathcal{H} \subseteq \mathcal{F}_n$. This follows by letting

$$\inf_{f \in \mathcal{H}} U_n\Big(\cdot\,, \big\{f(\boldsymbol{Z}_t(\cdot))\big\}_{t=1}^n\Big) \coloneqq \inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{t=1}^n q\big(\boldsymbol{Z}_t(\cdot), f(\boldsymbol{Z}_t(\cdot))\big) : \Omega \to \overline{\mathbb{R}},$$

in Proposition 2.1.1, since q is measurable-$\big(\mathcal{B}(\boldsymbol{\mathcal{Z}}) \otimes \mathcal{B}(\mathbb{R})\big)/\mathcal{B}(\mathbb{R})$.[2] However, for $\hat{f}_n$ as in (2.2), this does not ensure that the mapping $\omega \mapsto \hat{f}_n$ is measurable. This will be dealt with using outer integrals/probability in Subsection 2.1.2 (defined therein). Everywhere else in this work, the following proposition will be used which adds structure to $\mathcal{F}$ and $\mathcal{F}_n$ to ensure the existence of a measurable mapping for $\hat{f}_n$.

Proposition 2.1.2. *Let $1 \leq r < \infty$ and $n \in \mathbb{N}$. Suppose $\mathcal{G} \subset \mathcal{L}^r(P_{\{\boldsymbol{Z}_t\}_{t=1}^n})$ is a pointwise-separable class of functions such that $\{g(\boldsymbol{z}) : g \in \mathcal{G}\} \subset \mathbb{R}$ is compact for each $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$. Let $U_n : \Omega \times \mathbb{R}^n \to \mathbb{R}$ be such that for each $\boldsymbol{x} \in \mathbb{R}^n$ the function $U_n(\cdot, \boldsymbol{x}) : \Omega \to \mathbb{R}$ is measurable-$\mathcal{A}/\mathcal{B}(\mathbb{R})$, and for each $\omega \in \Omega$ the function $U_n(\omega, \cdot) : \mathbb{R}^n \to \mathbb{R}$ is continuous. Then, there exists a function $s : \Omega \to \mathcal{G}$ such that for each $\omega \in \Omega$*

$$s(\omega) \in \left\{ g \in \mathcal{G} :\ U_n\Big(\omega\,, \big\{g(\boldsymbol{Z}_t(\omega))\big\}_{t=1}^n\Big) = \inf_{g \in \mathcal{G}} U_n\Big(\omega\,, \big\{g(\boldsymbol{Z}_t(\omega))\big\}_{t=1}^n\Big) \right\} \neq \emptyset,$$

and s is measurable-$\mathcal{A}/\mathcal{B}(\mathcal{G})$, where $\mathcal{B}(\mathcal{G})$ uses the topology on $\mathcal{G}$ generated by $\|\cdot\|_{\mathcal{L}^r(P_{\{\boldsymbol{Z}_t\}_{t=1}^n})}$.

[2] To see this, note that, for any t, the mapping $(\omega, x) \mapsto (\boldsymbol{Z}_t(\omega), x)$, from $\Omega \times \mathbb{R}$ to $\boldsymbol{\mathcal{Z}} \times \mathbb{R}$, is measurable-$\big(\mathcal{A} \otimes \mathcal{B}(\mathbb{R})\big)/\big(\mathcal{B}(\boldsymbol{\mathcal{Z}}) \otimes \mathcal{B}(\mathbb{R})\big)$ (Aliprantis and Border, 2006, Lemma 4.49). Consequently, the function $(\omega, x) \mapsto q(\boldsymbol{Z}_t(\omega), x)$, from $\Omega \times \mathbb{R}$ to $\mathbb{R}$, is measurable-$\big(\mathcal{A} \otimes \mathcal{B}(\mathbb{R})\big)/\mathcal{B}(\mathbb{R})$ since q is measurable-$\big(\mathcal{B}(\boldsymbol{\mathcal{Z}}) \otimes \mathcal{B}(\mathbb{R})\big)/\mathcal{B}(\mathbb{R})$. Then, by Aliprantis and Border (2006, Lemma 4.52), the mapping $(\omega, x_1, \ldots, x_n) \mapsto U_n(\omega, x_1, \ldots, x_n) \coloneqq \frac{1}{n}\sum_{t=1}^n q(\boldsymbol{Z}_t(\omega), x_t)$, from $\Omega \times \mathbb{R}^n$ to $\mathbb{R}$, is measurable-$(\mathcal{A} \otimes \mathcal{B}(\mathbb{R}^n))/\mathcal{B}(\mathbb{R})$, and Proposition 2.1.1 can be applied.

The conditions on U_n in Proposition 2.1.2 are somewhat stronger than those used in Proposition 2.1.1 since they imply that U_n is measurable-$(\mathcal{A}\otimes\mathcal{B}(\mathbb{R}^n))/\mathcal{B}(\mathbb{R})$ by using Aliprantis and Border (2006, Lemma 4.51). Notably, Q_n will satisfy this stronger condition for many common choices of q, such as least squares or logistic loss.

Proposition 2.1.2 implies that a mapping $\omega \mapsto \hat{f}_n$ from $\Omega \to \mathcal{F}_n$ exists and is measurable whenever $q(\boldsymbol{Z}_t(\omega), \cdot) : \mathbb{R} \to \mathbb{R}$ is continuous for all $\omega \in \Omega$, and $\mathcal{F} \subseteq \mathcal{L}^r(P_{\{\boldsymbol{Z}_t\}_{t=1}^n})$ is a pointwise-separable class of functions such that $\{f(\boldsymbol{z}) : f \in \mathcal{F}_n\} \subset \mathbb{R}$ is compact for each $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$. Proposition 2.1.2 is closely related to Wooldridge and White (1991, Theorem 2.2), which is commonly used to ensure the existence and measurability of sieve estimators (e.g., Chen, 2007). Although Wooldridge and White (1991, Theorem 2.2) is applicable to metric spaces $(\mathcal{F}_n, \rho)$ where ρ may not be induced by an $\mathcal{L}^r$-norm, they require that $(\mathcal{F}_n, \rho)$ be a compact metric space. Therefore, Proposition 2.1.2 adds generality by only requiring $\mathcal{F}_n \subset \mathcal{L}^r(P_{\{\boldsymbol{Z}_t\}_{t=1}^n})$ such that $\{f(\boldsymbol{z}) : f \in \mathcal{F}_n\} \subset \mathbb{R}$ is compact for all $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$, which does not imply that $(\mathcal{F}_n, \|\cdot\|_{\mathcal{L}^r(P_{\{\boldsymbol{Z}_t\}_{t=1}^n})})$ is a compact metric space. This will be crucial when the sieve spaces are constructed using DNNs with unbounded weights as in Section 2.2.

2.1.2 Convergence rates without stationarity

This subsection gives a convergence rate result for sieve estimators applicable to very general estimation settings where $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ is possibly non-stationary. The complexity of $\mathcal{F}_n$ will be controlled using a covering number as defined below. $(\mathbb{M}, \|\cdot\|)$ denotes a semi-metric space, induced by the norm $\|\cdot\|$.

Definition 2.1.2 (Covering number)**.** Let $\delta > 0$, and let $(\mathbb{M}, \|\cdot\|)$ be a semi-metric space.

(*i*) For $\mathcal{G} \subset \mathbb{M}$, the **$\delta$-covering number** of $\mathcal{G}$, denoted as $\mathrm{N}(\delta, \mathcal{G}, \|\cdot\|)$, is the smallest $J \in \mathbb{N}$ such that there exists a collection $\{m_j\}_{j=1}^J \subseteq \mathbb{M}$, where

$$\mathcal{G} \subseteq \bigcup_{j=1}^{J} \Big\{ g \in \mathcal{G} : \|g - m_j\| < \delta \Big\},$$

and if no such $J \in \mathbb{N}$ exists let $\mathrm{N}(\delta, \mathcal{G}, \|\cdot\|) = \infty$.

(*ii*) When $\mathbb{M}$ is a space of functions with elements $f : \boldsymbol{\mathcal{Z}} \to \mathbb{R}$, then, for any $a \in \mathbb{N}$, define $\mathbb{M}|_{\{\boldsymbol{Z}_t\}_{t=1}^a} \coloneqq \big\{\big(f(\boldsymbol{Z}_1), f(\boldsymbol{Z}_2), \ldots, f(\boldsymbol{Z}_a)\big) : f \in \mathbb{M}\big\}$. For any $r \geq 1$ let $\mathrm{N}_r^{(\infty)}(\delta, \mathbb{M}, a) \coloneqq \sup\Big\{\mathrm{N}\big(\delta, \mathbb{M}|_{\{\boldsymbol{Z}_t(\omega)\}_{t=1}^a}, \|\cdot\|_{r,a}\big) : \omega \in \Omega\Big\}$.

For $\mathcal{G} \subset \mathcal{F}$ and a sample $\{\boldsymbol{Z}_t\}_{t=1}^a$, note that $\mathrm{N}\big(\delta, \mathcal{G}|_{\{\boldsymbol{Z}_t\}_{t=1}^a}, \|\cdot\|_{r,a}\big)$ depends on $\omega \in \Omega$, but $\mathrm{N}_r^{(\infty)}(\delta, \mathcal{G}, a)$ does not. I adopt the usual convention, if $\mathcal{G} = \emptyset$ then $\mathrm{N}(\delta, \mathcal{G}, \|\cdot\|) = 1$, for any $\delta > 0$.

Theorem 2.1.1 is the first main result of this section. It provides a rate of convergence in probability using an approach similar to the consistency results in Wooldridge and White (1991). For generality, the following will not impose the conditions of Proposition 2.1.2 to ensure the measurability of $\hat{f}_n$. Instead, I will use outer probability, P^*, as defined in van der Vaart and Wellner (1996, §1.2), i.e., for an arbitrary set $B \subseteq \Omega$

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

Theorem 2.1.1. *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 $\{\epsilon_n\}_{n\in\mathbb{N}}$ such that $\theta_n = \mathrm{O}_P(\epsilon_n^2)$. Then, $\rho_n(\hat{f}_n, f_0) = \mathrm{O}_{P^*}(\epsilon_n)$, 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 $\rho_n(\tilde{f}_n, f_0) \leq \epsilon_n$, for all n.*

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

$$C_1\, \rho_n(f, f_0)^2 \;\leq\; \mathbb{E}\big[Q_n(f)\big] - \mathbb{E}\big[Q_n(f_0)\big] \;\leq\; C_2\, \rho_n(f, f_0)^2.$$

(a.3) *There exist $m_n : \boldsymbol{\mathcal{Z}} \to [0, \infty)$, measurable-$\mathcal{B}(\boldsymbol{\mathcal{Z}})/\mathcal{B}([0,\infty))$, 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 $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$*

$$\big|q\big(\boldsymbol{z}, f(\boldsymbol{z})\big) - q\big(\boldsymbol{z}, f'(\boldsymbol{z})\big)\big| \leq m_n(\boldsymbol{z})\big|f(\boldsymbol{z}) - f'(\boldsymbol{z})\big|;$$

(ii) $\lim_{n\to\infty} P\left(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) \geq M_n\right) = 0$; *and*

(iii) *for some $C_3 > 0$ not depending on n,*

$$\sup\left\{\mathbb{E}\Big[\big|q\big(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)\big)\big|\mathbb{1}_{\{m_n(\boldsymbol{Z}_t)\geq M_n\}}\Big] : \; f \in \mathcal{F}_n, \, t \in \{1, ..., n\}\right\} \leq C_3\, \epsilon_n^2.$$

(a.4) *There exists* $\lambda_n^{(q)} : (0,\infty) \to (0,\infty)$ *such that,*

(i) *for any* $\delta > 0$*, and* $f \in \mathcal{F}_n$*,*

$$P\left(\frac{1}{n}\left|\sum_{t=1}^{n}\left(q\big(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)\big)\mathbb{1}_{\{m_n(\boldsymbol{Z}_t)<M_n\}} - \mathbb{E}\big[q\big(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)\big)\mathbb{1}_{\{m_n(\boldsymbol{Z}_t)<M_n\}}\big]\right)\right| \geq \delta\right) \lesssim \lambda_n^{(q)}(\delta);$$

(ii) *and, for any fixed* $\delta > 0$ *sufficiently large,*[3]

$$\lim_{n\to\infty}\left\{\lambda_n^{(q)}\big(\delta\,\epsilon_n^2\big)\cdot \mathrm{N}_1^{(\infty)}\big(\delta\,\epsilon_n^2/M_n,\, \mathcal{F}_n,\, n\big)\right\} = 0.$$

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 satisfied when the extrema of $\{\boldsymbol{Z}_t\}_{t=1}^n$, and $\mathcal{F}_n$, grow sufficiently slowly with n. In many cases (a.3)(ii) will imply (a.3)(iii). This will be demonstrated in Subsection 2.2.2, with the use of Lemma 2.2.1.

The regularity conditions imposed by (a.3) are more general than those typically used in the nonparametric sieve estimation literature. For instance, Farrell et al. (2021) requires that $\boldsymbol{\mathcal{Z}}$ be compact, and q satisfy a Lipschitz condition like (a.3)(i), except m_n must be a constant. Condition A.4 of Chen and Shen (1998) requires that there exist $s \in (0,2)$, $\gamma > 4$, and $C > 0$ such that, for any $\delta > 0$, $\sup_{\{f\in\mathcal{F}_n:\rho_n(f,g)\leq\delta\}} |q\big(\boldsymbol{z}, f(\boldsymbol{z})\big) - q\big(\boldsymbol{z}, h(\boldsymbol{z})\big)| \leq \delta^s\, m_n(\boldsymbol{Z}_t)$, and $\sup_n \mathbb{E}[m_n(\boldsymbol{Z}_t)^\gamma] < C$. In either case, these conditions imply (a.3).

Letting ρ_n vary with n is often necessary in settings where $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ is non-stationary. For instance, when ρ_n is the metric induced by $\|\cdot\|_{\mathcal{L}^2(P^*_{\{\boldsymbol{Z}_t\}_{t=1}^n})}$ condition (a.2) is easily verified in many estimation settings (see Appendix A.5.1). Note that condition (a.2) will often not hold for fixed $\rho_n = \rho$, such as $\rho_n = \rho = \|\cdot\|_{\mathcal{L}^2(P^*_{\{\boldsymbol{Z}_t\}_{t=1}^\infty})}$ for all n, when $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ is non-stationary.

Condition (a.4)(i) can often be met by using an exponential inequality to obtain $\lambda_n^{(q)}$, such as Wooldridge and White (1991, Theorem 3.4), or Merlevède et al. (2009, Theorem 1). These

[3] $\mathcal{F}_n$ in the covering number may be replaced with $\left\{f \in \mathcal{F}_n : \epsilon_n\sqrt{(\delta^2 + C_2 + 4C_3)/C_1} \leq \rho_n(f, f_0)\right\} \subseteq \mathcal{F}_n$.

inequalities can be applied without requiring q to be bounded, due to (a.3)(i) and the inclusion of $\mathbb{1}_{\{m_n(\boldsymbol{Z}_t)<M_n\}}$ in (a.4)(i), provided that $\mathcal{F}_n$ satisfies a boundedness condition, such as $\sup_{f\in\mathcal{F}_n}\|f\|_\infty \lesssim M_n$.

2.1.3 Nonasymptotic error bounds with stationarity and β-mixing

This subsection gives finite sample error bounds for sieve estimators in a setting where $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ is strictly stationary and β-mixing. When $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ is stationary I write $P_{\boldsymbol{Z}} = P_{\boldsymbol{Z}_t}$ for all $t \in \mathbb{N}$. The following definition for the β-mixing coefficient is from Dehling and Philipp (2002, Definition 3.1, p.19), and is equivalent to many standard definitions.

Definition 2.1.3 (β-mixing)**.** The β-mixing coefficient is defined as,

$$\beta(j) \coloneqq \mathbb{E}\left[\sup\left\{\left|P\big(B|\,\sigma\big(\{\boldsymbol{Z}_t\}_1^k\big)\big) - P\big(B\big)\right| : B \in \sigma\left(\{\boldsymbol{Z}_t\}_{k+j}^\infty\right),\, k\in\mathbb{N}\right\}\right] \quad \text{for } j \in \mathbb{N}.$$

$\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ is β-mixing (or absolutely regular) if $\lim_{j\to\infty}\beta(j) = 0$.

Conditions that ensure β-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 β-mixing with exponentially decreasing coefficients, also referred to as geometric β-mixing. For instance, conditions to ensure stationarity and geometric β-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 2.1.1 (Chen and Chen (2000))**.** *Consider model (2.1) with* $\boldsymbol{X}_t = (Y_{t-1},\ldots,Y_{t-d})^\mathsf{T}$. *Given an initial condition* $\boldsymbol{X}_0 = (Y_0,\ldots,Y_{1-d})^\mathsf{T} \in \mathbb{R}^d$, *if*

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

(ii) the function s_0 *is uniformly bounded;*

(iii) there exists a constant c such that for all $\delta \geq 0$,

$$0 < c \leq \inf_{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_{2,d} \leq \delta} \eta(\boldsymbol{x}) \leq \sup_{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_{2,d} \leq \delta} \eta(\boldsymbol{x}) < \infty;$$

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

$$|s_0(\boldsymbol{x})| \leq c_0^{(s)} + \sum_{j=1}^{d} c_j^{(s)} |x_j|, \quad \eta(\boldsymbol{x}) \leq c_0^{(\eta)} + \sum_{j=1}^{d} c_j^{(\eta)} |x_j|, \quad \forall \boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_{2,d} \leq C,$$

and $\sum_{j=1}^{d} \{c_j^{(s)} + c_j^{(\eta)} |\upsilon_j|\} < 1$;

then, $\{\boldsymbol{Z}_t\}_{t \in \mathbb{N}}$ is strictly stationary and β-mixing such that there exist constants $C'_\beta, C_\beta > 0$ where $\beta(j) \leq C'_\beta e^{-C_\beta j}$.

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

Definition 2.1.4 (Pseudo-dimension)**.** Let $\mathcal{S}$ be a class of functions from $\mathbb{X} \to \mathbb{R}$. The pseudo-dimension of $\mathcal{S}$, denoted as $\text{Pdim}(\mathcal{S})$, is the largest $p \in \mathbb{N}$ for which there exists $(x_1, \ldots, x_p, y_1, \ldots, y_p) \in \mathbb{X}^p \times \mathbb{R}^p$ such that for any $(b_1, \ldots, b_p) \in \{0, 1\}^p$ there exists $s \in \mathcal{S}$ such that $\mathbb{1}\{s(x_i) - y_i > 0\} = b_i$, for all $i = 1, \ldots, p$.

Pseudo-dimension, along with related complexity measures like Vapnik–Chervonenkis dimension, (see e.g. Bartlett et al., 2019, Definition 1) is often described as 'scale insensitive' because, unlike the covering number, it does not depend on a specific threshold $\delta > 0$. Scale-insensitive complexity measures are particularly well-suited for 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 A.3.4 herein).

Theorem 2.1.2 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 Merlevède et al. (2009). Following Farrell et al. (2021), the proof

of Theorem 2.1.2 differs 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 offer straightforward applicability to DNN estimators that have unbounded parameters (see Farrell et al., 2021 for more discussion).

Theorem 2.1.2. *Let* $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ *be strictly stationary,* $(\mathcal{F}, \|\cdot\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})})$ *be a (semi-) metric space,* $q(\boldsymbol{Z}_t(\omega),\cdot) : \mathbb{R} \to \mathbb{R}$ *be continuous for all* $\omega \in \Omega$*, and* $\mathcal{F}_n$ *be a pointwise-separable class such that* $\{f(\boldsymbol{z}) : f \in \mathcal{F}_n\} \subset \mathbb{R}$ *is compact for each* $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ *and* $n \in \mathbb{N}$*. Suppose* $\|f_0\|_\infty \leq 1$*, and the following conditions hold:*

(b.1) $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ *is* β*-mixing with* $\beta(j) \leq C'_\beta e^{-C_\beta j}$ *for some* $C_\beta, C'_\beta > 0$.

(b.2) *For each* $n \in \mathbb{N}$*, there exists a non-stochastic* $\tilde{f}_n \in \mathcal{F}_n$ *where* $\tilde{\epsilon}_n \coloneqq \|\tilde{f}_n - f_0\|_\infty$ *is such that* $\lim_{n\to\infty} \tilde{\epsilon}_n(\log n)(\log\log n) = 0$.

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

$$C_1\|f - f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \;\leq\; \mathbb{E}\big[Q_n(f)\big] - \mathbb{E}\big[Q_n(f_0)\big] \;\leq\; C_2\|f - f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})}.$$

(b.4) $\mathcal{F}_n$ *is such that* $\mathrm{Pdim}(\mathcal{F}_n) \geq 1$ *for all* n*, and for some non-decreasing sequence* $\{B_n\}_{n\in\mathbb{N}}$ *such that* $B_1 \geq 2$*,* $\sup_{f\in\mathcal{F}_n}\|f\|_\infty \leq B_n < \infty$ *for each* n*, and*

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

(b.5) *There exists* $m_n : \boldsymbol{\mathcal{Z}} \to [0,\infty)$*, measurable-*$\mathcal{B}(\boldsymbol{\mathcal{Z}})/\mathcal{B}([0,\infty))$*, such that, for each* n*,*

(i) for any $f, f' \in \mathcal{F}_n$*,* $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$

$$\big|q\big(\boldsymbol{z}, f(\boldsymbol{z})\big) - q\big(\boldsymbol{z}, f'(\boldsymbol{z})\big)\big| \leq m_n(\boldsymbol{z})\big|f(\boldsymbol{z}) - f'(\boldsymbol{z})\big|;$$

and

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

$$\min\left\{B_n\mathbb{E}\big[m_n(\boldsymbol{Z}_t)\mathbb{1}_{\{m_n(\boldsymbol{Z}_t)>C_4B_n\}}\big], \sup_{f\in\{\mathcal{F}_n\cup\{f_0\}\}}\mathbb{E}\Big[\big|q\big(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)\big)\big|\mathbb{1}_{\{m_n(\boldsymbol{Z}_t)\geq C_4B_n\}}\Big]\right\} \leq \mu_n,$$

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

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

$$\max\left\{2\,,\ \frac{\tilde{\epsilon}_n(\log n)(\log\log n)}{B_n}\right\} < a \leq n/2, \quad \textit{and} \quad \sqrt{\delta} \geq \frac{\tilde{\epsilon}_n\sqrt{n}}{B_n a - \tilde{\epsilon}_n(\log n)(\log\log n)},$$

the following holds

$$P\Big(\|\hat{f}_n - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq C\,\epsilon_n(\delta, a)\Big) \geq 1 - e^{-\delta} - 2\log(n)\left[\frac{n\,\beta(a)}{a} + 2P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) \geq C_4 B_n\Big)\right], \quad \textit{and}$$

$$P\left(\|\hat{f}_n - f_0\|_{2,n} \leq C\,\epsilon_n(\delta, a)\right) \geq 1 - 4e^{-\delta} - 6\log(n)\left[\frac{n\,\beta(a)}{a} + 2P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big)\right]$$

for $\epsilon_n(\delta, a) := B_n\sqrt{\frac{a}{n}}\left[\sqrt{\text{Pdim}(\mathcal{F}_n)\log(n)} + \sqrt{\log\log(n) + \delta}\right] + \sqrt{\tilde{\epsilon}_n^2 + \mu_n + \theta_n}$.

The existence of a measurable mapping $\omega \to \hat{f}_n$ is not an extra assumption, as it will follow from Proposition 2.1.2 and the assumptions in the statement of Theorem 2.1.2. Therefore it is not necessary to use outer probabilities/integrals when dealing with $\hat{f}_n$.

The requirement that $a > \frac{\tilde{\epsilon}_n(\log n)(\log\log n)}{B_n}$ will not be binding for large n. This follows because (b.2) and (b.4) imply $\lim_{n\to\infty} \frac{\tilde{\epsilon}_n(\log n)(\log\log n)}{B_n} = 0$.

Note that $n \geq 16B_n^2/\log(n)$ and $n \geq 2\text{Pdim}(\mathcal{F}_n)$ for all n sufficiently large is implied by (b.4). Requiring that $C_4 \geq 1$ and $C_1 \leq 1$ is also without loss of generality. To see this, if (b.5)(ii), or (b.3) hold with $C_4 < 1$ or $C_1 > 1$, respectively, then they will also hold with the constant replaced by one. However, this requirement can be dropped if n is large enough or $(4 \cdot 288)C_4/C_1 \geq \sqrt{38}$ (see first paragraph of Appendix A.3.5.4 for more detials).

Conditions (b.3) and (b.5)(i) are analogous to Conditions (a.2) (a.3)(i), and the discussion following Theorem 2.1.1 still applies. Condition (b.5)(ii) is related to (a.3)(iii), but Condition (b.5)(ii) is somewhat more general since the uniform integrability type requirement on q can be replaced with $\lim_{n\to\infty} B_n m_n(\boldsymbol{Z}_t)\mathbb{1}_{\{m_n(\boldsymbol{Z}_t) > C_4 B_n\}} = 0$.

Although Theorem 2.1.2 does not explicitly require $\lim_{n\to\infty} P\big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\big) = 0$, like Condition (a.3)(iii), this will be necessary for the error bounds to hold with probability approaching one. In light of this, the requirement that $q(\boldsymbol{Z}_t(\omega), \cdot) : \mathbb{R} \to \mathbb{R}$ be continuous for

all $\omega \in \Omega$, imposes very little additional structure, since Condition (b.5)(i) will imply this with probability approaching one whenever $\lim_{n\to\infty} P\big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\big) = 0$.

Theorem 2.1.2 may also imply a rate of convergence by choosing $\delta = \delta_n$ and $a = a_n$ such that δ_n and a_n go to infinity with n at an appropriate rate. For this, note that for any $\{a_n\}_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} a_n/\log(n) = \infty$ then $\log(n)\frac{n\,\beta(a_n)}{a_n} = 0$. To see this, choose $b_n = a_n C_\beta \log^{-1}(n)$, so by Condition (b.1)

$$\frac{\log(n) n\,\beta(a_n)}{a_n} \lesssim \frac{\log(n) n e^{-C_\beta a_n}}{a_n} = \frac{\log(n) n^{1-b_n}}{a_n} = \frac{\log(n) n^{1-\frac{a_n C_\beta}{\log(n)}}}{a_n}.$$

2.2 DNN estimators

This section applies the results of Section 2.1 to obtain theoretical properties for a class of DNN estimators commonly used in applications. All proofs are included in Appendix A.5. As before, the setting from Section 2.1 can be applied with the understanding that, for any $\boldsymbol{z} = (y, \boldsymbol{x})$, and mapping $\boldsymbol{x} \to f(\boldsymbol{x})$, one can define $f(\boldsymbol{z}) \coloneqq f(\pi_{\boldsymbol{X}}(\boldsymbol{z})) = f(\boldsymbol{x})$ for the coordinate projection $\pi_{\boldsymbol{X}}(\boldsymbol{z}) = \boldsymbol{x}$.

Throughout this section, $\{\boldsymbol{Z}_t = (Y_t, \boldsymbol{X}_t^\intercal)^\intercal\}_{t\in\mathbb{N}}$ will be a stochastic sequence on the complete probability space $(\Omega, \mathcal{A}, P)$, such that, $Y_t \in \mathbb{R}$ and $\boldsymbol{X}_t \in [0,1]^d$ for each t. The object to be estimated is $f_0 : [0,1]^d \to [-1,1]$, which satisfies the following Hölder smoothness assumption.

Assumption 2.2.1. (Smoothness) *For a smoothness parameter $p \in \mathbb{N}$, and each multi-index $\boldsymbol{k} \in \{\mathbb{N}\cup\{0\}\}^d$ with $\sum_{j=1}^d k_j \leq p-1$, the regression function $f_0 : [0,1]^d \to \mathbb{R}$ is such that $D^{\boldsymbol{k}} f_0$ is continuous, and $\|D^{\boldsymbol{k}} f_0\|_\infty \leq 1$.*

This type of smoothness assumption is standard in the nonparametric estimation literature (e.g. Stone, 1982; Chen and Shen, 1998; Chen, 2007; Chen and Christensen, 2015; Farrell et al., 2021), and Yarotsky (2017) shows that DNNs approximate these functions well. Note that this assumption is weaker than the structural assumptions imposed in much of the DNN literature (e.g., Kohler and Krzyżak, 2017; Schmidt-Hieber, 2020; Kohler and Langer, 2021).

For most of the results in this section, β-mixing will be stronger than necessary. In these cases, I use α-mixing, as defined by Dehling and Philipp, 2002, Definition 3.1.

Definition 2.2.1 (α-mixing)**.** The α-mixing coefficient is defined as,

$$\alpha(j) \coloneqq \sup\left\{\left|P\big(A\cap B\big) - P(A)P(B)\right| : A \in \sigma\big(\{\boldsymbol{Z}_t\}_1^k\big),\, B \in \sigma\left(\{\boldsymbol{Z}_t\}_{k+j}^{\infty}\right),\, k \in \mathbb{N}\right\} \quad \text{for } j \in \mathbb{N}.$$

$\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ is α-mixing (or strong mixing) if $\lim_{j\to\infty}\alpha(j) = 0$.

Sufficient conditions for a process to be α-mixing have been heavily studied (e.g. Doukhan, 1994, §2.4; or Bradley, 2005), and an example is included in Corollary 2.2.1 of Subsection 2.2.3. It is well known that $\alpha(j) \leq \beta(j)$, so any β-mixing process is also α-mixing with a rate at least as fast. Consequently, Corollary 2.1.1 can also be used to imply α-mixing.

2.2.1 DNN sieve spaces ($\mathcal{N}_n$)

The sieve spaces used here, $\mathcal{N}_n$, will be constructed using fully connected feedforward DNNs. For concreteness, I focus on DNNs equipped with the ReLU activation function $\phi(x) = \max\{0, x\}$. However, Corollary 2.2.2 shows that all results in the following subsections apply for any continuous piecewise-linear activation function. Considerations for other feedforward architectures and activation functions are also discussed in Subsection 2.2.4.

For $n \in \mathbb{N}$ the graph structure of the architecture $\mathcal{N}_n$ is characterized by its depth $L_n \in \mathbb{N}$, and width vector $\boldsymbol{H}_n = [H_{n,0}, H_{n,1}, \ldots, H_{n,L_n}] \in \mathbb{N}^{L_n+1}$ where $H_{n,0} = d$ for all n. Each hidden layer, $l \in \{1, \ldots, L_n\}$, is comprised of $H_{n,l}$ "hidden" computational units, referred to as *nodes*, and denoted as $u_{l,h}$. The d inputs $\boldsymbol{x} = [x_1, \ldots, x_d]' \in [0,1]^d$ are fed into each node in the first hidden layer $l = 1$. Then each node in layer $l = 1$ feeds into each node of the next layer $l = 2$. This process repeats with each node in layer $l-1$ passing its output into each node in layer l up to the last hidden layer $l = L_n$. Finally, each node in the last hidden layer, $l = L_n$, feeds into the output layer of the network, $l = L_n + 1$, which consists of only one computation unit. Note that nodes in layer l receive inputs only from nodes in layer $l-1$, and none from layers $l-2$ or earlier. See Figure 2.1 for an example of the architecture $\mathcal{N}_n$.

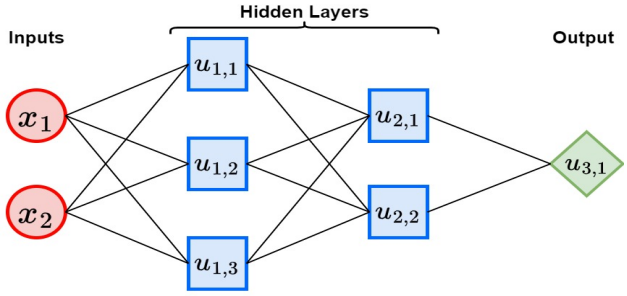

Figure 2.1: Example of $\mathcal{N}_n$ architecture graph structure where $L_n = 2$, $H_{n,1} = 3$, $H_{n,2} = 2$, $W_n = 20$, and $d = 2$.

Each node is a function taking values in $\mathbb{R}$ which depends on a real-valued vector of parameters $\boldsymbol{\gamma}_{l,h}$, and takes as arguments the outputs of all nodes in the previous layer, which ultimately depend on the original input $\boldsymbol{x}$. The parameters for each node consist of a scalar intercept term $\gamma_{l,h,0}$,[4] and weights $\gamma_{l,h,1}, \ldots, \gamma_{l,h,H_{n,l-1}}$ such that

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

The parameters for the entire network are collected together as

$$\overline{\boldsymbol{\gamma}}_{W_n} \coloneqq \{\gamma_{l,h,k}\}_{\forall\, l,h,k} \in \mathbb{R}^{W_n},$$

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 $u_{l,h}(\boldsymbol{x} \mid \overline{\boldsymbol{\gamma}}_{W_n})$ and the single node in the last layer ($l = L_n + 1$) is denoted as $u_{L_n+1,1}(\boldsymbol{x} \mid \overline{\boldsymbol{\gamma}}_{W_n})$ where the dependence on $\overline{\boldsymbol{\gamma}}_{W_n}$ will often be suppressed. Then each node can be written as

$$u_{l,h}\left(\boldsymbol{x} \mid \overline{\boldsymbol{\gamma}}_{W_n}\right) \coloneqq \begin{cases} \phi\left(\sum_{i=1}^{d} \gamma_{1,h,i} \cdot x_i + \gamma_{1,h,0}\right), & l = 1, \\ \phi\left(\sum_{i=1}^{H_{n,l-1}} \gamma_{l,h,i} \cdot u_{l-1,i}(\boldsymbol{x}) + \gamma_{l,h,0}\right), & 2 \le l \le L_n, \\ \sum_{i=1}^{H_{n,L_n}} \gamma_{L_n+1,1,i} \cdot u_{L_n,i}(\boldsymbol{x}) + \gamma_{L_n+1,1,0}, & l = L_n + 1. \end{cases} \tag{2.3}$$

[4] 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.

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

$$\mathcal{N}_n = \mathcal{N}(L_n, \boldsymbol{H}_n, B_n) \coloneqq \Big\{ f = u_{L_n+1,1}\big(\,\cdot \mid \overline{\gamma}_{W_n}\big) : \ \overline{\gamma}_{W_n} \in \mathbb{R}^{W_n}, \ \sup_{\boldsymbol{x}\in[0,1]^d} |u_{L_n+1,1}\big(\boldsymbol{x} \mid \overline{\gamma}_{W_n}\big)| \leq B_n \Big\}. \quad (2.4)$$

A particular network $f \in \mathcal{N}_n$ is determined by $\overline{\gamma}_{W_n}$ and can be written as $f(\cdot) = u_{L_n+1,1}\left(\cdot \mid \overline{\gamma}_{W_n}\right) : [0,1]^d \to [-B_n, B_n]$.

Remark 2.2.1. $\mathcal{N}_n$ as in (2.4) 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 (ϕ) and linear combinations of $\boldsymbol{\gamma}_{l,h}$ and $\{u_{l-1,h}\}_{h=1}^{H_{n,l-1}}$ (or $\boldsymbol{x} \in [0,1]^d$ when $l = 1$).

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

(*iii*) Note that for each n and $\boldsymbol{x} \in \boldsymbol{\mathcal{Z}}$

$$\{f(\boldsymbol{x}) : f \in \mathcal{N}_n\} = [-B_n, B_n],$$

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.

$$\mathcal{N}_n^* \coloneqq \Big\{ f \in \mathcal{N}_n : \gamma_{L_n+1,1,0} \in [-B_n, B_n], \quad \text{and} \quad \gamma_{l,h,k} = 0 \ \ \forall \{l,h,k\} \neq \{L_n + 1, 1, 0\} \Big\}.$$

Clearly $f^*(\boldsymbol{x}) = \gamma_{L_n+1,1,0}$ for any $f^* \in \mathcal{N}_n^*$, $\boldsymbol{x} \in [0,1]^d$. Thus, $\{f^*(\boldsymbol{x}) : f^* \in \mathcal{N}_n^*\} = [-B_n, B_n]$ for each $\boldsymbol{x} \in [0,1]^d$. Then the result follows since $\mathcal{N}_n^* \subset \mathcal{N}_n$ and $\sup_{f\in\mathcal{N}_n} \|f\|_\infty = B_n$ by (2.4).

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 2.2.1 since Proposition 2.1.2 will be applicable in the estimation settings considered throughout this section.

2.2.2 DNNs for nonparametric regression

I consider a general nonparametric regression estimation setting, that includes Example 2.1.1 as a special case. Theorem 2.2.1 will apply Theorem 2.1.1 to show the convergence of DNN estimators in a very general setting without assuming stationarity. Theorem 2.2.2 will apply Theorem 2.1.2 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 $f_0 = \mathbb{E}[Y_t|\boldsymbol{X}_t]$, with a DNN sieve estimator $\hat{f}_n$ as in (2.2), using the least squares criterion

$$q(\boldsymbol{Z}_t, f) \coloneqq (Y_t - f(\boldsymbol{X}_t))^2,$$

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

Assumption 2.2.2. *For all $t \in \mathbb{N}$, suppose $Y_t \in \mathcal{L}^2(\Omega, \mathcal{A}, P)$ such that there exists a non-decreasing sequence $\{B_n\}_{n\in\mathbb{N}}$ with $B_1 \geq 2$, where*

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

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 $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ 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 $\mathbb{E}[Y_t^2|\{\boldsymbol{X}_t, Y_{t-1}\}_{t=1}^n]$ being uniformly bounded or $\mathbb{E}[Y_t^{2+\delta}] < \infty$ for some $\delta > 0$ (Chen and Christensen, 2015, Assumption 2).

Assumption 2.2.2 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).

Proposition 2.2.1. *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* $\lim_{n\to\infty} nP(|Y_1| > B_n) = 0$;[5] *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*

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

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

The following lemma shows that Assumption 2.2.2 implies a form of uniform integrability that will be used to verify (b.5)(ii) and (a.3)(iii) for this subsection's main results.

Lemma 2.2.1. *If Assumption 2.2.2 holds then* $\lim_{n\to\infty} \left\{ \max_{t\in\{1,\ldots,n\}} \mathbb{E}\big[Y_t^2 \, \mathbb{1}_{|Y_t|\geq B_n}\big] \right\} = 0$.

The first main result of this section is Theorem 2.2.1, which applies Theorem 2.1.1, to obtain a rate of convergence in probability for DNN estimators in general nonparametric regression settings with nonstationary α-mixing data.

Theorem 2.2.1. *Suppose Assumptions 2.2.1 and 2.2.2 hold with* $B_n \lesssim n^{K_B}$ *for some* $K_B \in [0, 1/4)$. *Let* $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ *be an* α*-mixing process with* $\alpha(j) \leq C'_\alpha e^{-C_\alpha j}$ *for some* $C_\alpha, C'_\alpha > 0$. *Let* $\mathcal{N}_n = \mathcal{N}(L_n, \boldsymbol{H}_n, B_n)$ *be defined as in* (2.4) *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,* $H_{n,l} \asymp H_n$, *and*

$$L_n \asymp \log(n), \qquad H_n \asymp n^{\left(\frac{d}{p+d/2}\right)(1/4-K_B)} \log^2(n). \tag{2.5}$$

Suppose $\{\hat{f}_n\}_{n\in\mathbb{N}}$ *satisfies* (2.2) *and there exists* $\{\epsilon_n\}_{n\in\mathbb{N}}$ *such that* $\theta_n = \mathrm{O}_P(\epsilon_n^2)$, *and for some* $\upsilon > 0$,

$$\epsilon_n \gtrsim \max \left\{ n^{-\left(\frac{p}{p+d/2}\right)(1/4-K_B)} \log^{2+\upsilon}(n), \max_{t\in\{1,\ldots,n\}} \sqrt{\mathbb{E}\big[Y_t^2 \, \mathbb{1}_{|Y_t|\geq B_n}\big]} \right\}.$$

Then, $\|\hat{f}_n - f_0\|_{\mathcal{L}^2(P_{\{\boldsymbol{X}_t\}_{t=1}^n})} = \mathrm{O}_P(\epsilon_n)$.

[5] Note that α-mixing is stronger than necessary and could be replaced with Leadbetter et al. (1983, Condition $D(u_n)$, p.53) which is implied by α-mixing (see discussion on p.54 therein).

Theorem 2.2.1 shows that DNN estimators are consistent in very general settings, however the convergence rate is strictly slower than $n^{-1/4}$. The next result uses Theorem 2.1.2 to obtain nonasymptotic error bounds which can imply faster rates of convergence in probability when $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ is stationary and β-mixing.

Theorem 2.2.2. *Suppose Assumptions 2.2.1 and 2.2.2 hold with $B_n \asymp n^{K_B}$ for some $K_B \in [0, 1/2)$. Let $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ be a strictly stationary β-mixing process with $\beta(j) \leq C'_\beta e^{-C_\beta j}$ for some $C_\beta, C'_\beta > 0$. Let $\mathcal{N}_n = \mathcal{N}(L_n, \boldsymbol{H}_n, B_n)$ be defined as in (2.4) 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, $H_{n,l} \asymp H_n$, and*

$$L_n \asymp \log(n), \qquad H_n \asymp n^{\left(\frac{d}{p+d}\right)(1/2-K_B)} \log^2(n).$$

Then, for $\{\hat{f}_n\}_{n\in\mathbb{N}}$ satisfying (2.2), and

$$\epsilon_n = n^{-\left(\frac{p}{p+d}\right)(1/2-K_B)} \log^6(n) + \sqrt{\mathbb{E}\left[Y_t^2 \, \mathbb{1}_{\{|Y_t|\geq B_n\}}\right] + \theta_n},$$

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

$$P\left(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} \leq C\,\epsilon_n\right) \geq 1 - e^{-n^{\left(\frac{p}{p+d}\right)(1/2-K_B)}} - \frac{2C'_\beta n^{1-C_\beta \log(n)}}{\log(n)} - 4\log(n) P\Big(\max_{t\in\{1,\ldots,n\}} |Y_t| \geq B_n\Big),$$

$$P\left(\|\hat{f}_n - f_0\|_{2,n} \leq C\,\epsilon_n\right) \geq 1 - 4e^{-n^{\left(\frac{p}{p+d}\right)(1/2-K_B)}} - \frac{12C'_\beta n^{1-C_\beta \log(n)}}{\log(n)} - 24\log(n) P\Big(\max_{t\in\{1,\ldots,n\}} |Y_t| \geq B_n\Big).$$

This result can imply convergence rates as fast as $n^{-\left(\frac{p}{2(p+d)}\right)} \log^6(n)$ when Y_t is almost surely bounded and θ_n converges sufficiently quickly. In this case, the rate only differs by a factor of $\log^2(n)$ from the rate implied by Farrell et al. (2021, Theorem 1) for settings where $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ forms an i.i.d. sequence and Y_t takes values in a compact set. Chapter 3 demonstrates that the rates implied by Theorem 2.2.2 can be sufficiently fast to obtain $\sqrt{n}$-asymptotic normality of a finite-dimensional parameter following first stage nonparametric estimation of conditional expectations with DNN estimators in a partially linear regression under stationary β-mixing data with unbounded regressands.

The convergence rates from Theorem 2.2.1 and the rate implied by Theorem 2.2.2 will not depend on $\alpha(j)$ and $\beta(j)$ respectively, provided they are geometrically mixing. Although, the probability bound from Theorem 2.2.2 will depend on C_β, C'_β.

2.2.3 DNNs for binomial autoregressions with covariates

This section considers logistic binomial autoregression models, building upon Example 2.1.2. The main result of this section is Theorem 2.2.3, which demonstrates the applicability of Theorem 2.1.1 for DNN estimators in categorical autoregressive settings. Such settings are of particular interest for DNNs since they are frequently employed for classification problems. Although this binomial logistic setting is a simple case of the classification problem, similar results for multinomial and non-logistic settings could be attained with the methods employed here. Also, see Farrell et al. (2021, Lemma 9) for an example of how the criterion functions for Poisson, Gamma, and multinomial logistic models can be shown to satisfy the requirements of Theorem 2.1.1.

The setting will be a special case of the models considered in Truquet (2021). For all $t \in \mathbb{N}$, let $\boldsymbol{Z}_t \coloneqq (Y_t, \boldsymbol{X}_t)$ such that $Y_t \in \{0,1\}$ and $\boldsymbol{X}_t = (\boldsymbol{V}_{t-1}, Y_{t-1}, \dots Y_{t-r}) \in [0,1]^{d-r} \times \{0,1\}^r \subset [0,1]^d$ for some random vector of covariates $\boldsymbol{V}_t \in [0,1]^{d-r}$, where $d > r$. Suppose

$$Y_t = s\big(\boldsymbol{V}_{t-1}, Y_{t-1}, \dots, Y_{1-r}, \upsilon_t\big) \coloneqq s\big(\boldsymbol{X}_t, \upsilon_t\big), \quad \forall t \in \mathbb{N}, \tag{2.6}$$

where $\upsilon_t \in \mathbb{R}$ is some random noise, and the function $s : [0,1]^{d-r} \times \{0,1\}^r \times \mathbb{R} \to \{0,1\}$ is measurable. The estimation target will be a scaled version of the function $\log\left(\frac{\mathbb{E}[Y_t|\boldsymbol{X}_t]}{1-\mathbb{E}[Y_t|\boldsymbol{X}_t]}\right)$ under the following logistic regression assumption.

Assumption 2.2.3. *For all $t \in \mathbb{N}$, let $\mathbb{E}[Y_t|\boldsymbol{X}_t] = e^{Bf_0}\big[1 + e^{Bf_0}\big]^{-1}$ for some $B \geq 2$, and $f_0 : [0,1]^d \to [-1,1]$.*

It will be convenient to write the assumption this way since $\|f_0\|_\infty \leq 1$ under Assumption 2.2.1. Indeed, Assumption 2.2.3 is equivalent to the usual assumption $\mathbb{E}[Y_t|\boldsymbol{X}_t] = e^{s_0}\big[1 + e^{s_0}\big]^{-1}$ where it is also assumed that $\|s_0\|_\infty \leq B$, by setting $f_0 = s_0/B$. In what follows, the particular value of B will play no role in the convergence rate of the estimator. Also note that Assumption 2.2.1 will require f_0 to be defined on $[0,1]^d$, which imposes some additional structure on f_0 since $\boldsymbol{X}_t \in [0,1]^{d-r} \times \{0,1\}^r \subset [0,1]^d$.

The goal is to estimate $f_0 = B^{-1}\log\left(\frac{\mathbb{E}[Y_t|\boldsymbol{X}_t]}{1-\mathbb{E}[Y_t|\boldsymbol{X}_t]}\right)$, using a DNN sieve estimator $\hat{f}_n$ as in (2.2)

where the criterion is

$$q(\boldsymbol{Z}_t, f) \coloneqq -Y_t B f(\boldsymbol{X}_t) + \log\left(1 + e^{Bf(\boldsymbol{X}_t)}\right),$$

and the DNN sieve spaces are $\mathcal{F}_n = \left\{f(\pi_{\boldsymbol{X}}(\cdot)) : f \in \mathcal{N}_n\right\}$.

Theorem 2.2.3. *Suppose Assumptions 2.2.1 and 2.2.3 hold. Let $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ be an α-mixing process with $\alpha(j) \leq C'_\alpha e^{-C_\alpha j}$ for some $C_\alpha, C'_\alpha > 0$. Let $\mathcal{N}_n = \mathcal{N}(L_n, \boldsymbol{H}_n, 2)$ be defined as in (2.4) 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, $H_{n,l} \asymp H_n$, and*

$$L_n \asymp \log(n), \qquad H_n \asymp n^{\frac{1}{2}\left(\frac{d}{p+d}\right)} \log^2(n). \tag{2.7}$$

For $\{\hat{f}_n\}_{n\in\mathbb{N}}$ satisfying (2.2) if there exists $\{\epsilon_n\}_{n\in\mathbb{N}}$ such that $\theta_n = \mathrm{O}_P(\epsilon_n^2)$, and

$$\epsilon_n \gtrsim n^{-\frac{1}{2}\left(\frac{p}{p+d}\right)} \log^5(n),$$

then $\|\hat{f}_n - f_0\|_{\mathcal{L}^2(P_{\{\boldsymbol{X}_t\}_{t=1}^n})} = \mathrm{O}_P(\epsilon_n)$.

To the best of my knowledge, this is the first result providing a convergence rate for DNN estimators in classification settings with dependent data. Theorem 2.2.3 provides a convergence rate in settings with nonstationary α-mixing data that is identical, up to a logarithmic factor, to the rate implied by Farrell et al. (2021, Theorem 1) under i.i.d. data. In addition, this convergence rate is unaffected by the rate of decay of the α-mixing coefficient, provided it is geometric mixing.

Theorem 2.2.3 allows for very general forms of dependence, and includes many interesting examples. The following corollary provides two examples of settings in which Theorem 2.2.3 can be applied without directly assuming the mixing condition for $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$. Corollary 2.2.1(i) follows from Truquet (2021, Theorem 1) and point 2 of the discussion following their result, since Assumption 2.2.3 implies $P(Y_t = 1|\boldsymbol{X}_t) \in \left[e^{-B}/\left(1+e^{-B}\right), e^{B}/\left(1+e^{B}\right)\right]$. Corollary 2.2.1(ii) follows from Truquet (2021, Theorem 3, Proposition 1).[6] Using these results, one can also obtain similar sufficient conditions for mixing properties of $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ in the multinomial case, $Y_t \in \{0, 1, \ldots, N\}$. Let the mixing coefficient for $\{\boldsymbol{V}_t\}_{t=0}^{\infty}$ and $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ be denoted as $\alpha_{\boldsymbol{V}}$ and $\alpha_{\boldsymbol{Z}}$ respectively.

[6] To apply Truquet (2021, Proposition 1) note that ergodicity of $\{\boldsymbol{V}_t\}$ is implied by stationarity and α-mixing (see e.g. Bradley, 2005).

Corollary 2.2.1 (Truquet (2021)). *Consider the model from (2.6). Suppose* $\{\boldsymbol{V}_t\}_{t=0}^{\infty}$ *is strictly stationary and one of the following holds:*

(i) $\{\boldsymbol{V}_t\}_{t=0}^{\infty}$ *is* α*-mixing such that* $\alpha_{\boldsymbol{V}}(j) = \mathrm{O}\big(e^{-Cj}\big)$ *for some* $C > 0$*, and* $\{\upsilon_t\}_{t\in\mathbb{N}}$ *is an i.i.d. sequence, independent of* $\{\boldsymbol{V}_t\}_{t=0}^{\infty}$*, such that for any* $y \in \{0, 1\}$

$$P\Big(Y_t = y \,\Big|\, \{\boldsymbol{V}_t\}_{t=0}^{\infty}, Y_{t-1}, Y_{t-2}, \dots\Big) = P\Big(Y_t = y \,\Big|\, \boldsymbol{V}_{t-1}, Y_{t-1}, \dots, Y_{t-r}\Big) \equiv P\Big(Y_t = y \,\Big|\, \boldsymbol{X}_t\Big);$$

or

(ii) $\{\boldsymbol{V}_t\}_{t=0}^{\infty}$ *is* α*-mixing such that* $\alpha_{\boldsymbol{V}}(j) = \mathrm{O}\Big(e^{-Cj^2}\Big)$ *for some* $C > 0$*, and* $\{\upsilon_t\}_{t=0}^{\infty}$ *is uniformly distributed on* $(0, 1)$ *such that, for each* $t \in \mathbb{N}$*,* υ_t *is independent of* $\sigma\big(\{\boldsymbol{V}_{t-j}, \upsilon_{t-j}\}_{j=1}^{t}\big)$ *and for any* $y \in \{0, 1\}$

$$P\Big(Y_t = y \,\Big|\, \boldsymbol{V}_{t-1}, Y_{t-1}, Y_{t-2}, \dots\Big) = P\Big(Y_t = y \,\Big|\, \boldsymbol{V}_{t-1}, Y_{t-1}, \dots, Y_{t-r}\Big) \equiv P\Big(Y_t = y \,\Big|\, \boldsymbol{X}_t\Big);$$

and the model (2.6) satisfies

$$s\big(\boldsymbol{X}_t, \upsilon_t\big) = 0 \quad \Longleftrightarrow \quad 0 < \upsilon_t \leq \mathbb{E}[Y_t|\boldsymbol{X}_t].$$

Then, $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ *is strictly stationary and* α*-mixing with* $\alpha_{\boldsymbol{Z}}(j) = \mathrm{O}\big(e^{-C_\alpha j}\big)$*, for some* $C_\alpha > 0$*.*

Corollary 2.2.1(i) allows general distributions for υ_t but imposes a strict exogeneity assumption with respect to $\{\boldsymbol{V}_t\}_{t=0}^{\infty}$. Corollary 2.2.1(ii) requires $\{\upsilon_t\}_{t=0}^{\infty}$ to be i.i.d. with a uniform distribution, and $\{\boldsymbol{V}_t\}$ to be mixing at a faster rate, but allows for some endogeneity, since $\boldsymbol{V}_t$ can depend on υ_t, provided $\boldsymbol{V}_t$ is independent of υ_{t-j}, for $j = 1, \dots, t$. These results are somewhat stronger than needed for Theorem 2.2.3, since the stationarity of $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$, or $\{\boldsymbol{V}_t\}_{t=0}^{\infty}$ is not required. Truquet (2021) suggests these results could generalize to the non-stationary case, but further work is needed to verify this.

2.2.4 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 chapter's results to alternative feedforward DNN architectures.

Remark 2.2.2. In this section, and in the proofs for Section 2.2's results, I will use bounds on the the Vapnik Chervonenkis dimension (see Bartlett et al., 2019, Definition 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, Definition 2).

The first result uses Yarotsky (2017, Proposition 1) to show that Theorems 2.2.1, 2.2.2, and 2.2.3 are directly applicable to fully connected feedforward networks with any continuous piecewise-linear activation function, φ, that has a fixed number of breakpoints, $b \in \mathbb{N}$. One important example of this is the 'leaky' ReLU (LReLU) activation function

$$\varphi^{\text{LReLU}}(x) = \begin{cases} cx, & x < 0, \\ x, & x \geq 0, \end{cases}$$

for some small constant $c > 0$ (often set to $c = 0.01$) that is predetermined before optimizing over the parameters $\overline{\gamma}_{W_n}$. The LReLU is often used in practice to address the vanishing gradient problem that arises with the ReLU where certain computation units may never have non-zero outputs.

Corollary 2.2.2. *Let $b \in \mathbb{N}$ be a constant and $\varphi : \mathbb{R} \to \mathbb{R}$ be any continuous piece-wise linear function with b breakpoints. Let $\mathcal{N}_{n,\varphi}$ be defined as in (2.4) except with the ReLU activation function replaced by φ. Then, Theorems 2.2.1, 2.2.2, and 2.2.3 also apply to $\mathcal{N}_{n,\varphi}$.*

The proofs of Theorems 2.2.1, 2.2.2, and 2.2.3 follow similarly when $\mathcal{N}_n$ is replaced with $\mathcal{N}_{n,\varphi}$. To see this, note that with Yarotsky (2017, Proposition 1) the approximation result of Lemma A.5.1 can be extended to DNNs with a continuous piecewise-linear activation function by only increasing H_n by a constant factor; and the complexity results of Lemmas A.5.4 and A.5.5 are applications of Bartlett et al. (2019) and Anthony and Bartlett (1999, Theorem 12.2) which also apply to $\mathcal{N}_{n,\varphi}$.

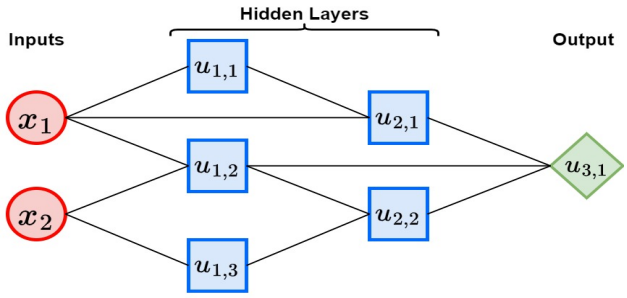

Figure 2.2: Example of $\mathcal{N}_{n,\varphi}^{\text{FFN}}$ architecture graph structure where $L_n = 2$, $W_n = 17$, and $d = 2$.

The next result provides a framework to obtain theoretical properties of DNN estimators with a wide variety of activation functions and any feedforward graph structure, provided an approximation result like Lemma A.5.1 exists for the DNN under consideration. These DNNs will be denoted as $\mathcal{N}_{n,\varphi}^{\text{FFN}}$, which allow for any graph structure where units in layer $l \in \{1, \ldots, L_n + 1\}$ takes inputs from any of the units in layers $l' \in \{0, \ldots, l-1\}$. See Figure 2.2 for an example of $\mathcal{N}_{n,\varphi}^{\text{FFN}}$'s graph structure.

Corollary 2.2.3. *Let $\mathcal{N}_{n,\varphi}^{\text{FFN}}$ be any feedforward neural network with L_n layers, W_n parameters, U_n computation units and some continuous activation function $\varphi : \mathbb{R} \to \mathbb{R}$, such that $\sup_{f \in \mathcal{N}_{n,\varphi}^{\text{FFN}}} \|f\|_\infty \leq B_n$. Define the complexity bound $\Xi_{n,\varphi}$ for the following three classes of activation functions:*

(i) if φ is piecewise-linear with $b \in \mathbb{N}$ breakpoints let $\Xi_{n,\varphi} \coloneqq W_n L_n \log(W_n)$;

(ii) if φ is piecewise-polynomial with $b \in \mathbb{N}$ breakpoints where each piece is a polynomial with degree $\leq p \in \mathbb{N}$, let $\Xi_{n,\varphi} \coloneqq W_n U_n \log\big((p+1)b\big)$;

(iii) if φ is the sigmoid function, $\varphi(x) = \frac{1}{1+e^x}$, let

$$\Xi_{n,\varphi} \coloneqq \big((W_n + 2)U_n\big)^2 + (W_n + 2)U_n \log_2\big(18(W_n + 2)U_n^2\big).$$

Suppose φ is described by one of the three above classes, then there exists $C \geq 1$ such that, for any $n \in \mathbb{N}$, $\mathrm{N}_\infty^{(\infty)}\big(\delta, \mathcal{N}_n, a\big) \leq \left(\frac{2eB_n a}{\delta\, \Xi_{n,\varphi}}\right)^{C\, \Xi_{n,\varphi}}$

$$\mathrm{Pdim}(\mathcal{N}_{n,\varphi}^{\text{FFN}}) \;\leq\; C\, \Xi_{n,\varphi}, \quad \textit{and} \quad \mathrm{N}_\infty^{(\infty)}\big(\delta, \mathcal{N}_n, n\big) \;\leq\; \left(\frac{2eB_n n}{\delta\, \Xi_{n,\varphi}}\right)^{C\, \Xi_{n,\varphi}},$$

for all $\delta > 0$*, and* $a \geq C\,\Xi_{n,\varphi}$*. Consequently, Theorems 2.1.1 or 2.1.2 can be applied with* $\mathcal{F}_n = \left\{f(\pi_{\boldsymbol{X}}(\cdot)) : f \in \mathcal{N}_{n,\varphi}^{\mathrm{FFN}}\right\}$*, whenever the conditions of these theorems hold using the complexity bounds in the previous display.*

The pseudo-dimension bounds follow from Bartlett et al. (2019, pp. 5,6) for the piecewise-linear case, Bartlett et al. (2019, Theorem 10) for the piecewise-polynomial case, and Anthony and Bartlett (1999, Theorem 8.13) for the sigmoid case. The covering number bound uses Anthony and Bartlett (1999, Theorem 12.2). To apply Theorems 2.1.1 or 2.1.2 with Corollary 2.2.3, an approximation result, like Lemma A.5.1, will be needed for $\mathcal{N}_{n,\varphi}^{\mathrm{FFN}}$ to verify Conditions (a.1) or (b.2) respectively.

2.3 Summary and extensions

This chapter addresses the lack of statistical foundation for empirical work using deep neural network (DNN) estimators under dependent data. By establishing general results for sieve estimators, I provide a flexible framework that applies to various DNN estimators in a wide range of dependent data settings. These results extend existing work to more complex and realistic scenarios, allowing for non-i.i.d. data with very general forms of dependence and taking values in unbounded sets. I apply this framework to derive properties for DNN estimators in both nonparametric regression and classification contexts, focusing on architectures that reflect modern applications–featuring ReLU activation functions, unbounded parameters, fully connected feedforward structures, and depth and width that grow with sample size. Notably, Corollary 2.2.2 shows that these results also apply when the ReLU activation function is replaced with any continuous piecewise-linear activation function, such as the leaky ReLU.

While this work only considers standard DNN architectures, Subsection 2.2.4 demonstrates how the general sieve estimator results presented here offer a pathway for extending the analysis to more complex architectures. Perhaps the most important avenue for future research on DNN estimators under dependent data is recurrent neural networks (RNNs). RNNs are a class of DNN

architectures that are specifically designed for time series settings, due to a recursive feedback loop that gives the network a form of 'memory' for past events. While the empirical results from Lazcano et al. (2024) indicate that the DNN architectures studied here outperform RNNs in simpler time-series models, RNNs have demonstrated superior empirical performance in more complex settings, such as in the work by Bucci (2020) on forecasting stock market volatility. Thus far, very little work has been done on the theoretical properties of RNNs. While Kohler and Krzyzak (2020) provides an initial study of a specific recurrent architecture in time-series nonparametric regression, much remains to be understood, particularly for more general recurrent architectures, and dependence settings. Following the ideas of Subsection 2.2.4, the sieve estimator framework introduced in this chapter 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 efficiency 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.

Chapter 3

Inference in Partially Linear Models under Dependent Data with Deep Neural Networks

This chapter applies the results of Chapter 2 in a partially linear regression model, to obtain valid asymptotic inference on finite-dimensional parameters following first-stage DNN estimation of infinite-dimensional components. The proofs for this chapter are included in Appendix B. This work can also be found in Brown (2024a).

Partially linear models were first considered by Engle et al. (1986) and Robinson (1988). Since then, these models have been widely used for empirical work in time series settings (see e.g., Gao, 2007; and Härdle et al., 2000; for a review). For instance, partially linear models have been employed by Engle et al. (1986) to study electricity sales, since the impact of temperature on electricity consumption is nonlinear, as both high and low temperatures lead to increased electricity demand; Li et al. (2024) to study the forward premium anomaly; and Gao and Yee (2000) to study the number of lynx trapped in the MacKenzie River district in the Canadian North-West Territories.

In particular, consider the partially linear model,

$$Y_t = D_t\zeta_0 + \eta_0(\boldsymbol{X}_t) + u_t, \qquad \text{where} \quad \mathbb{E}[u_t|D_t, \boldsymbol{X}_t] = 0. \tag{3.1}$$

Let $\{\boldsymbol{Z}_t \coloneqq (Y_t, D_t, \boldsymbol{X}_t)\}_{t\in\mathbb{N}}$ be a stochastic sequence on the probability space $(\Omega, \mathcal{A}, P)$. The data is $\{\boldsymbol{Z}_t\}_{t=1}^n$, where Y_t is the outcome, D_t is the policy or treatment variable, and $\boldsymbol{X}_t$ are additional covariates which affect Y_t through an unknown nuisance function η_0 that is measurable. Note that this model can apply to nonlinear auto-regressive settings by letting $\boldsymbol{X}_t$ include past values

Y_{t-j}, $j = 1, \ldots, r$ for $r \in \mathbb{N}$. For notational simplicity, we consider the case where D_t is a scalar treatment, although the results obtained here could easily generalize to the vector-valued case. The goal is to estimate and perform inference on the parameter $\zeta_0 \in \mathbb{R}$.

Using a procedure due to Robinson (1988), the estimator for ζ_0 will be constructed using DNN-estimated nuisance components. With the DNN results from Chapter 2, I show that this estimator will obtain $\sqrt{n}$-consistency and asymptotic normality in settings with stationary β-mixing data under mild regularity conditions. By employing the ideas of Chen et al. (2022, Theorem 1), I do this without sample splitting, which is particularly important in dependent data settings, where it can be difficult to construct independent validation sets.

3.1 Estimation procedure and results

I estimate ζ_0 following a procedure due to Robinson (1988). Note that

$$\mathbb{E}[Y_t|\boldsymbol{X}_t] = \mathbb{E}[D_t|\boldsymbol{X}_t]\zeta_0 + \eta_0(\boldsymbol{X}_t) + \mathbb{E}[u_t|\boldsymbol{X}_t].$$

Let $s_0 \coloneqq \mathbb{E}[Y_t|\boldsymbol{X}_t]$ and $w_0 \coloneqq \mathbb{E}[D_t|\boldsymbol{X}_t]$. Then, taking the difference of (3.1) and the previous display,

$$Y_t - s_0(\boldsymbol{X}_t) = \big(D_t - w_0(\boldsymbol{X}_t)\big)\zeta_0 + \upsilon_t, \qquad \text{where} \quad \mathbb{E}[\upsilon_t|D_t, \boldsymbol{X}_t] = 0.$$

This provides the moment condition $\mathbb{E}\big[\psi\big(\boldsymbol{Z}_t; \zeta_0, w_0, s_0\big)\big] = 0$, for the linear moment function

$$\begin{aligned}\psi\big(\boldsymbol{Z}_t; \zeta, w, s\big) &\coloneqq \big(D_t - w(\boldsymbol{X}_t)\big)\Big[\big(D_t - w(\boldsymbol{X}_t)\big)\zeta - Y_t - s(\boldsymbol{X}_t)\Big] \\ &= \mathsf{A}\big(\boldsymbol{Z}_t; w\big)\,\zeta - \mathsf{v}\big(\boldsymbol{Z}_t; w, s\big),\end{aligned}$$

where

$$\mathsf{A}\big(\boldsymbol{Z}_t; w\big) \coloneqq \big(D_t - w(\boldsymbol{X}_t)\big)^2, \quad \text{and} \quad \mathsf{v}\big(\boldsymbol{Z}_t; w, s\big) \coloneqq \big(D_t - w(\boldsymbol{X}_t)\big)\big(Y_t - s(\boldsymbol{X}_t)\big).$$

Now, I construct a two-stage method of moments estimator for ζ_0. In the first stage, I estimate the conditional expectations s_0 and w_0 with DNN estimators $\hat{s}_n$ and $\hat{w}_n$ constructed using the

framework of Section 2.2. Then, I have the estimator

$$\hat{\zeta} = \left[\sum_{t=1}^{n} \mathsf{A}\big(\boldsymbol{Z}_t; \hat{w}_n\big)\right]^{-1} \left[\sum_{t=1}^{n} \mathsf{v}\big(\boldsymbol{Z}_t; \hat{w}_n, \hat{s}_n\big)\right],$$

whenever $\sum_{t=1}^{n} \mathsf{A}\big(\boldsymbol{Z}_t; \hat{w}_n\big) \neq 0$. The following assumption will ensure that $\hat{\zeta}$ exists with probability approaching one and that Theorem 2.2.2 can be used to obtain DNN estimators of w_0, s_0 that converge faster than $n^{-1/4}$.

Assumption 3.1.1. *(i)* $\{\boldsymbol{Z}_t \coloneqq (Y_t, D_t, \boldsymbol{X}_t)\}_{t\in\mathbb{N}}$ *is a strictly stationary* β*-mixing process on the complete probability space* $(\Omega, \mathcal{A}, P)$ *with* $\beta(j) \leq C'_\beta e^{-C_\beta j}$ *for some* $C_\beta, C'_\beta > 0$.

(ii) $D_t \in [0,1]$, $\boldsymbol{X}_t \in [0,1]^d$, $\mathbb{E}\big[\mathsf{A}\big(\boldsymbol{Z}_t; w_0\big)\big] = \mathbb{E}\big[\big(D_t - w_0(\boldsymbol{X}_t)\big)^2\big] > 0$, *and* $\boldsymbol{X}_t$ *is continuously distributed.*

(iii) There exists $p^{(w)} \in \mathbb{N}$ *and* $p^{(s)} \in \mathbb{N}$ *such that* w_0, s_0 *satisfy Assumption 2.2.1 with smoothness* $p^{(w)}$, $p^{(s)}$ *respectively, and*

$$\frac{1}{4} < \min\left\{\frac{1}{2}\left(\frac{p^{(w)}}{p^{(w)}+d}\right), \left(\frac{p^{(s)}}{p^{(s)}+d}\right)(1/2 - K_B)\right\},$$

for some constant $0 \leq K_B < \min\left\{\frac{1}{4}\left(\frac{p^{(w)}-d}{p^{(w)}+d}\right), \frac{1}{4}\left(\frac{p^{(s)}}{p^{(s)}+d/2}\right)\right\}$.

(iv) $\mathbb{E}[\,|Y_t|^{2+\delta}\,] < \infty$ *for some* $\delta > 0$, *and there exists a non-decreasing sequence* $\{B_n\}_{n\in\mathbb{N}}$ *with* $B_1 > 2$, *and* $B_n \asymp n^{K_B}$ *such that*

$$\lim_{n\to\infty} P\left(\max_{t\in\{1,\ldots,n\}} |Y_t| \geq B_n\right) = 0, \quad \textit{and} \quad \max_{t\in\{1,\ldots,n\}} \mathbb{E}\big[Y_t^2\, \mathbb{1}_{|Y_t|\geq B_n}\big] = \mathrm{o}(n^{-1}).$$

The following corollary will use Assumption 3.1.1 and Theorem 2.2.2 to obtain a rate ϵ_n such that $\|\hat{w}_n - w_0\|_{\mathcal{L}^2(P_{\boldsymbol{X}})}$ and $\|\hat{s}_n - s_0\|_{\mathcal{L}^2(P_{\boldsymbol{X}})}$ are $\mathrm{O}_P(\epsilon_n)$. The smoothness requirement in Assumption 3.1.1(iii) is the key to obtaining ϵ_n that converges to zero fast enough to obtain the desired asymptotic properties of $\hat{\zeta}$. The upper bound on K_B is equivalent to requiring $B_n\epsilon_n = \mathrm{o}(n^{-1/4})$, for ϵ_n defined in Corollary 3.1.1. The feasibility of the bounds on K_B is not an additional assumption since the first part of Assumption 3.1.1(iii) implies $p^{(w)} > d$. Note that Assumption 3.1.1(iv) is stronger than necessary for Corollary 3.1.1, but will be needed for Thereom 3.1.1 below.

Corollary 3.1.1 (Theorem 2.2.2)**.** *For the DNN architecture defined as in* (2.4)*, let* $\mathcal{N}_n^{(w)} = \mathcal{N}(L_n^{(w)}, \boldsymbol{H}_n^{(w)}, 2)$ *and* $\mathcal{N}_n^{(s)} = \mathcal{N}(L_n^{(s)}, \boldsymbol{H}_n^{(s)}, B_n)$ *for the non-decreasing sequences* $L_n^{(w)} \asymp L_n^{(s)} \asymp \log(n)$*, and*

$$H_{n,l}^{(w)} \asymp n^{\frac{1}{2}\left(\frac{d}{p^{(w)}+d}\right)} \log^2(n), \qquad H_{n,l}^{(s)} \asymp n^{\left(\frac{d}{p^{(s)}+d}\right)(1/2-K_B)} \log^2(n), \quad \forall\, l \in \mathbb{N}.$$

Let $\hat{w}_n \in \mathcal{N}_n^{(w)}$*, and* $\hat{s}_n \in \mathcal{N}_n^{(s)}$ *be DNN sieve estimators defined as in Section 2.2 that both satisfy* (2.2) *with* $\theta_n = \mathrm{o}_P(n^{-1/2})$*. If Assumption 3.1.1 holds, then for*

$$\epsilon_n = \log^6(n) \cdot \max\left\{ n^{-\frac{1}{2}\left(\frac{p^{(w)}}{p^{(w)}+d}\right)}, \; n^{-\left(\frac{p^{(s)}}{p^{(s)}+d}\right)(1/2-K_B)} \right\}$$

it follows that $\|\hat{w}_n - w_0\|_{\mathcal{L}^2(P_{\boldsymbol{X}})} = \mathrm{O}_P(\epsilon_n)$*, and* $\|\hat{s}_n - s_0\|_{\mathcal{L}^2(P_{\boldsymbol{X}})} = \mathrm{O}_P(\epsilon_n)$*.*

The following theorem will use the properties of $\hat{w}_n$, and $\hat{s}_n$ from Corollary 3.1.1 to obtain $\sqrt{n}$-convergence and asymptotic normality of $\hat{\zeta}$.

Theorem 3.1.1. *Suppose Assumption 3.1.1 holds, and* $\hat{w}_n$*,* $\hat{s}_n$ *are constructed as in Corollary 3.1.1. Then, we have the following:*

(i) $\|\hat{\zeta} - \zeta_0\| = \mathrm{O}_P\big(n^{-1/2}\big)$*;*

(ii) there exists a constant $\sigma \geq 0$ *such that*

$$\lim_{n\to\infty} \mathrm{Var}\left[\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \psi_t\big(\zeta_0, w_0, s_0\big)\right] = \sigma^2 < \infty,$$

and if $\sigma > 0$*, then* $\sqrt{n}\big(\hat{\zeta} - \zeta_0\big) \xrightarrow{d} N\Big(0,\, \mathbb{E}\big[\mathsf{A}\big(\boldsymbol{Z}_t; w_0\big)\big]^{-2}\sigma^2\Big)$.[1]

The existence of $\sigma \geq 0$ is not a new result and follows from Bosq (1998, Theorem 1.5). In most settings with $\mathbb{E}[v_t^2|\boldsymbol{X}_t, D_t] > 0$ the requirement that $\sigma > 0$ is fairly mild. To see this, note that $\sigma > 0$ whenever the covariances of $\big\{v_t/\big(D_t - w(\boldsymbol{X}_t)\big)\big\}_{t\in\mathbb{N}}$ don't sum to zero, since by the

[1] As usual, for a sequence of real valued random variables $\{W_n\}_{n\in\mathbb{N}}$, we write $W_n \xrightarrow{d} N(\mu, \sigma^2)$ if W_n converges in distribution to a normally distributed random variable that has expected value μ and variance σ^2 as $n \to \infty$.

definition of ψ_t and Bienaymé's identity

$$\operatorname{Var}\left[\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_t\big(\zeta_0, w_0, s_0\big)\right] = \mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{v_t}{D_t - w_0(\boldsymbol{X}_t)}\right)^2\right], \quad \text{and}$$
$$\operatorname{Var}\left[\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_t\big(\zeta_0, w_0, s_0\big)\right] = \frac{1}{n}\sum_{i,j=1}^{n}\operatorname{Cov}\big[\psi_t\big(\zeta_0, w_0, s_0\big)\psi_j\big(\zeta_0, w_0, s_0\big)\big].$$

3.2 Summary and extensions

This chapter shows how the results from Chapter 2 can be applied to a partially linear regression model with dependent data. Using the DNN results from Section 2.2.2, I show that the estimator for the finite-dimensional parameter, constructed using DNN-estimated nuisance components, achieves $\sqrt{n}$-consistency and asymptotic normality. By avoiding sample splitting, I address one of the key challenges in applying machine learning techniques to econometric models with dependent data.

These results not only demonstrate the practical implications of Chapter 2, but also provide techniques that could be extended to address more complex econometric models, such as instrumental variable models, or more efficient estimation procedures (see e.g., Newey, 1990). This offers many promising avenues for future research that I plan to incorporate in a later version of this work.

Bibliography

Charalambos D. Aliprantis and Kim C. Border. 2006. Infinite Dimensional Analysis. Springer-Verlag, Berlin/Heidelberg. https://doi.org/10.1007/3-540-29587-9

Luis B Almeida. 2020. Multilayer perceptrons. In Handbook of Neural Computation. CRC Press, C1–2.

Mutaz AlShafeey and Csaba Csáki. 2021. Evaluating neural network and linear regression photovoltaic power forecasting models based on different input methods. Energy Reports 7 (Nov. 2021), 7601–7614. https://doi.org/10.1016/j.egyr.2021.10.125

Martin Anthony and Peter L. Bartlett. 1999. Neural Network Learning: Theoretical Foundations. Cambridge University Press. Google-Books-ID: UH6XRoEQ4h8C.

Peter Bartlett. 2013. Theoretical Statistics. Lecture 14. https://www.stat.berkeley.edu/~bartlett/courses/2013spring-stat210b/notes/14notes.pdf

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. 2005. Local Rademacher complexities. The Annals of Statistics 33, 4 (Aug. 2005). https://doi.org/10.1214/009053605000000282 arXiv:math/0508275.

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 João 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, Vol. 110. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-1718-3

Richard C. Bradley. 2005. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. Probability Surveys 2 (Jan. 2005). https://doi.org/10.1214/154957805100000104 arXiv:math/0511078.

Chad Brown. 2024a. Inference in Partially Linear Models under Dependent Data with Deep Neural Networks. https://doi.org/10.48550/arXiv.2410.22574

Chad Brown. 2024b. Statistical Properties of Deep Neural Networks with Dependent Data. https://doi.org/10.48550/arXiv.2410.11113 arXiv:2410.11113.

Wlodzimierz Bryc. 1982. ON THE APPROXIMATION THEOREM OF I. BERKES AND W. PHILIPP. Demonstratio Mathematica 15, 3 (July 1982), 807–816. https://doi.org/10.1515/dema-1982-0319 Publisher: De Gruyter Open Access.

Andrea Bucci. 2020. Realized Volatility Forecasting with Neural Networks. Journal of Financial Econometrics 18, 3 (June 2020), 502–531. https://doi.org/10.1093/jjfinec/nbaa008

Marine Carrasco and Xiaohong Chen. 2002. MIXING AND MOMENT PROPERTIES OF VARIOUS GARCH AND STOCHASTIC VOLATILITY MODELS. Econometric Theory 18, 1 (Feb. 2002), 17–39. https://doi.org/10.1017/S0266466602181023

Min Chen and Gemai Chen. 2000. Geometric Ergodicity of Nonlinear Autoregressive Models with Changing Conditional Variances. The Canadian Journal of Statistics / La Revue Canadienne de Statistique 28, 3 (2000), 605–613. https://doi.org/10.2307/3315968 Publisher: [Statistical Society of Canada, Wiley].

Qizhao Chen, Vasilis Syrgkanis, and Morgane Austern. 2022. Debiased Machine Learning without Sample-Splitting for Stable Estimators. http://arxiv.org/abs/2206.01825 arXiv:2206.01825 [cs, econ, math, stat].

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].

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21, 1 (Feb. 2018), C1–C68. https://doi.org/10.1111/ectj.12097

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. 2022. Automatic Debiased Machine Learning of Causal and Structural Effects. Econometrica 90, 3 (2022), 967–1027. https://doi.org/10.3982/ECTA18515 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18515.

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, Herold Dehling, Thomas Mikosch, and Michael Sørensen (Eds.). Birkhäuser, Boston, MA, 3–113. https://doi.org/10.1007/978-1-4612-0099-4_1

Pham Dinh Tuan. 1986. The mixing property of bilinear and generalised random coefficient autoregressive models. Stochastic Processes and their Applications 23, 2 (Dec. 1986), 291–300. https://doi.org/10.1016/0304-4149(86)90042-6

Paul Doukhan. 1994. Mixing. Lecture Notes in Statistics, Vol. 85. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4612-2642-0

Grzegorz Dudek. 2016. Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting. International Journal of Forecasting 32, 3 (July 2016), 1057–1060. https://doi.org/10.1016/j.ijforecast.2015.11.009

Ernst Eberlein. 1984. Weak convergence of partial sums of absolutely regular sequences. Statistics & Probability Letters 2, 5 (Oct. 1984), 291–293. https://doi.org/10.1016/0167-7152(84)90067-1

Robert F. Engle, C. W. J. Granger, John Rice, and Andrew Weiss. 1986. Semiparametric Estimates of the Relation Between Weather and Electricity Sales. J. Amer. Statist. Assoc. 81, 394 (1986), 310–320. https://doi.org/10.2307/2289218 Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. 2021. Deep Neural Networks for Estimation and Inference. Econometrica 89, 1 (2021), 181–213. https://doi.org/10.3982/ECTA16901 arXiv: 1809.09953.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. Data Mining and Knowledge Discovery 33, 4 (July 2019), 917–963. https://doi.org/10.1007/s10618-019-00619-1 arXiv:1809.04356.

Jiti Gao. 2007. Nonlinear Time Series: Semiparametric and Nonparametric Methods. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781420011210

Jiti Gao and Thomas Yee. 2000. Adaptive Estimation in Partially Linear Autoregressive Models. The Canadian Journal of Statistics / La Revue Canadienne de Statistique 28, 3 (2000), 571–586. https://doi.org/10.2307/3315966 Publisher: [Statistical Society of Canada, Wiley].

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, 315–323. https://proceedings.mlr.press/v15/glorot11a.html ISSN: 1938-7228.

Ulf Grenander. 1981. Abstract Inference. Wiley. Google-Books-ID: ng2oAAAAIAAJ.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical Asset Pricing via Machine Learning. The Review of Financial Studies 33, 5 (May 2020), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Wassily Hoeffding. 1963. Probability Inequalities for Sums of Bounded Random Variables. J. Amer. Statist. Assoc. 58, 301 (1963), 13–30. https://doi.org/10.2307/2282952 Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Wolfgang Härdle, Hua LIang, and Jiti Gao. 2000. Partially linear models. MPRA Paper. University Library of Munich, Germany. https://econpapers.repec.org/paper/pramprapa/39562.htm

Masaaki Imaizumi and Kenji Fukumizu. 2019. Deep Neural Networks Learn Non-Smooth Functions Effectively. In Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89), Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, 869–878. https://proceedings.mlr.press/v89/imaizumi19a.html

Yongdai Kim, Ilsang Ohn, and Dongha Kim. 2021. Fast convergence rates of deep neural networks for classification. Neural Networks 138 (June 2021), 179–197. https://doi.org/10.1016/j.neunet.2021.02.012

Michael Kohler and Adam Krzyzak. 2020. On the rate of convergence of a deep recurrent neural network estimate in a regression problem with dependent data. http://arxiv.org/abs/2011.00328 arXiv:2011.00328 [cs, stat].

Michael Kohler and Adam Krzyżak. 2017. Nonparametric Regression Based on Hierarchical Interaction Models. IEEE Transactions on Information Theory 63, 3 (March 2017), 1620–1630. https://doi.org/10.1109/TIT.2016.2634401 Conference Name: IEEE Transactions on Information Theory.

Michael Kohler and Sophie Langer. 2021. On the rate of convergence of fully connected deep neural network regression estimates. The Annals of Statistics 49, 4 (Aug. 2021), 2231–2249. https://doi.org/10.1214/20-AOS2034 Publisher: Institute of Mathematical Statistics.

Michael R. Kosorok. 2008. Introduction to Empirical Processes and Semiparametric Inference. Springer-Verlag. https://doi.org/10.1007/978-0-387-74978-5

Daisuke Kurisu, Riku Fukami, and Yuta Koike. 2024. Adaptive deep learning for nonlinear time series models. http://arxiv.org/abs/2207.02546 arXiv:2207.02546 [math, stat].

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

Jiaqi Li, Likai Chen, Kun Ho Kim, and Tianwei Zhou. 2024. Simultaneous inference of a partially linear model in time series. Journal of Time Series Analysis (2024). https://doi.org/10.1111/jtsa.12781

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

Whitney K. Newey. 1990. Efficient Instrumental Variables Estimation of Nonlinear Models. Econometrica 58, 4 (1990), 809–837. https://doi.org/10.2307/2938351 Publisher: [Wiley, Econometric Society].

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

P. M. Robinson. 1988. Root-N-Consistent Semiparametric Regression. Econometrica 56, 4 (1988), 931–954. https://doi.org/10.2307/1912705 Publisher: [Wiley, Econometric Society].

Apaar Sadhwani, Kay Giesecke, and Justin Sirignano. 2021. Deep Learning for Mortgage Risk*. Journal of Financial Econometrics 19, 2 (Aug. 2021), 313–368. https://doi.org/10.1093/jjfinec/nbaa025

Johannes Schmidt-Hieber. 2020. Nonparametric regression using deep neural networks with ReLU activation function. The Annals of Statistics 48, 4 (Aug. 2020). https://doi.org/10.1214/19-AOS1875

Xiaotong Shen and Wing Hung Wong. 1994. Convergence Rate of Sieve Estimates. The Annals of Statistics 22, 2 (June 1994), 580–615. https://doi.org/10.1214/aos/1176325486 Publisher: Institute of Mathematical Statistics.

Maxwell B. Stinchcombe and Halbert White. 1992. Some Measurability Results for Extrema of Random Functions Over Random Sets. The Review of Economic Studies 59, 3 (July 1992), 495–514. https://doi.org/10.2307/2297861

Charles J. Stone. 1982. Optimal Global Rates of Convergence for Nonparametric Regression. The Annals of Statistics 10, 4 (1982), 1040–1053. http://www.jstor.org/stable/2240707 Publisher: Institute of Mathematical Statistics.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. http://arxiv.org/abs/1602.07261 arXiv:1602.07261 [cs].

Lionel Truquet. 2021. Strong mixing properties of discrete-valued time series with exogenous covariates. https://doi.org/10.48550/arXiv.2112.03121

A. W. van der Vaart and Jon Wellner. 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4757-2545-2

Halbert White and A. Ronald Gallant. 1992. Artificial Neural Networks: Approximation and Learning Theory. Blackwell. Google-Books-ID: Xwd0QgAACAAJ.

Jeffrey M Wooldridge and Halbert White. 1991. Some results on sieve estimation with dependent observations. In Non-parametric and Semi-parametric Methods in Econometrics and Statistics, William A Barnett, James Powell, and George EugeneEditors Tauchen (Eds.). Cambridge University Press, 459–493. https://econpapers.repec.org/bookchap/cupcbooks/9780521370905.htm

Atsutomo Yara and Yoshikazu Terada. 2024. Nonparametric logistic regression with deep learning. https://doi.org/10.48550/arXiv.2401.12482 arXiv:2401.12482 [math, stat].

Dmitry Yarotsky. 2017. Error bounds for approximations with deep ReLU networks. Neural Networks 94 (2017), 103–114. https://doi.org/10.1016/j.neunet.2017.07.002

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.

Appendix A

Appendix for Chapter 2

A.1 Measurability of Extrema of Random Functions

This section provides proofs for the measurability results in Subsection 2.1.1. The findings presented here build upon previous work that has addressed similar problems (e.g. Stinchcombe and White, 1992; and van der Vaart and Wellner, 1996) by providing general results that offer straightforward applicability to sieve extremum estimation. Define a metrizable space as in Aliprantis and Border (2006, Example 2.2-3), i.e., for a topological space $(\mathbb{X}, \mathcal{O}_{\mathbb{X}})$ the space $\mathbb{X}$ is metrizable if there exists a metric ρ on $\mathbb{X}$ that generates the topology $\mathcal{O}_{\mathbb{X}}$.

Lemma A.1.1. *Let $(\Omega, \mathcal{A}, P)$ be a complete probability space, and let $\mathbb{M}$ be a complete and separable metrizable space. For $\mathbb{H} \subseteq \mathbb{M}$, suppose the function $U : \Omega \times \mathbb{H} \to \overline{\mathbb{R}}$ is measurable-$(\mathcal{A} \otimes \mathcal{B}(\mathbb{H}))/\mathcal{B}(\overline{\mathbb{R}})$, and the correspondence $\Psi : \Omega \rightrightarrows \mathbb{H}$ is such that*

$$\text{graph}(\Psi) \coloneqq \{(\omega, h) \in \Omega \times \mathbb{H} : h \in \Psi(\omega)\} \in \mathcal{A} \otimes \mathcal{B}(\mathbb{H}).$$

Let $v(\omega) \coloneqq \sup_{h \in \Psi(\omega)} U(\omega, h)$, then $v : \Omega \to \overline{\mathbb{R}}$ is measurable-$\mathcal{A}/\mathcal{B}(\overline{\mathbb{R}})$.

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

$$B_c = \big\{(\omega, h) : U(\omega, h) > c,\ \omega \in \Omega, h \in \Psi(\omega)\big\} \in \mathcal{A} \otimes \mathcal{B}(\mathbb{H}).$$

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

$$v^{-1}\big((c, \infty]\big) \coloneqq \big\{\omega : U(\omega, h) > c,\ h \in \Psi(\omega)\big\} = \big\{\omega : v(\omega) > c\big\}.$$

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

Lemma A.1.1 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 $(\mathbb{H}, \mathcal{B}(\mathbb{H}))$ is not required to be Souslin.[1] Note that requiring $\mathbb{H} \subseteq \mathbb{M}$, instead of $\mathbb{H} = \mathbb{M}$, allows for cases where $\mathbb{H}$ may not be complete.

To apply Lemma A.1.1 in nonparametric sieve estimation settings, the correspondence Ψ is often defined by the sieve spaces, e.g. $\Psi(\omega) = \{(f(\boldsymbol{Z}_1(\omega)), \ldots, f(\boldsymbol{Z}_n(\omega))) : f \in \mathcal{F}_n\}$. In such a setting, Proposition 2.1.1 shows that if $\mathcal{F}_n$ is pointwise-separable, then Lemma A.1.1 can be applied.

Proof of Proposition 2.1.1. First, we show the result holds when the supremum is over $\mathcal{G}$. Let $\Psi_n : \Omega \rightrightarrows \mathbb{R}^n$ be the correspondence $\Psi_n(\omega) \coloneqq \big\{\big(g(\boldsymbol{Z}_1(\omega)), \ldots, g(\boldsymbol{Z}_n(\omega))\big) : g \in \mathcal{G}\big\}$. By assumption, $\mathcal{G}$ is a pointwise-separable class, so there exists $\{g_j\}_{j\in\mathbb{N}} \subseteq \mathcal{G}$, such that $\{g(\boldsymbol{z}) : g \in \mathcal{G}\} = \mathrm{cl}(\{g_j(\boldsymbol{z})\}_{j\in\mathbb{N}}) \subseteq \mathbb{R}$, for each $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ (see Aliprantis and Border, 2006, §2.3, p.28). Thus, for each $\omega \in \Omega$,

$$\Psi_n(\omega) = \mathrm{cl}\big(\big\{\big(g_j(\boldsymbol{Z}_1(\omega)), \ldots, g_j(\boldsymbol{Z}_n(\omega))\big) : j \in \mathbb{N}\big\}\big).$$

This has two implications: first, for all $\omega \in \Omega$ the correspondence $\Psi_n(\omega)$ is closed and consequently equal to its closure; and second by Aliprantis and Border (2006, Corollary 18.14, p.601), Ψ_n is a weakly measurable correspondence. With this, Aliprantis and Border (2006, Theorem 18.6, p.596) implies that $\mathrm{graph}(\mathrm{cl}(\Psi_n)) \in \mathcal{A} \otimes \mathcal{B}(\mathbb{R}^n)$. Hence, $\mathrm{graph}(\Psi_n) \in \mathcal{A} \otimes \mathcal{B}(\mathbb{R}^n)$, since $\mathrm{graph}(\Psi_n) = \mathrm{graph}(\mathrm{cl}(\Psi_n))$. Then the result follows from Lemma A.1.1.

[1] At the expense of additional notation, Lemma A.1.1 can easily be generalized to a 'measure-free' version using Corbae et al. (2009, Theorem 7.9.12). Additionally, requiring $\mathbb{H}$ 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.

Now, we show the result holds when the supremum is over $\mathcal{H} \subset \mathcal{G}$, with $\mathcal{H} \neq \emptyset$. Let $\Psi'_n : \Omega \rightrightarrows \mathbb{R}^n$ be the correspondence $\Psi'_n(\omega) \coloneqq \big\{\big(g(\boldsymbol{Z}_1(\omega)), \ldots, g(\boldsymbol{Z}_n(\omega))\big) : g \in \mathcal{H}\big\}$. Since $\{g_j\}_{j\in\mathbb{N}}$ is a countable dense subset of $\mathcal{G}$, and $\mathcal{H} \subset \mathcal{G}$, then $\{g_j\}_{j\in\mathbb{N}} \cap \mathcal{H}$ is a countable dense subset of $\mathcal{H}$. Hence, for all $\omega \in \Omega$,

$$\Psi'_n(\omega) = \mathrm{cl}\Big(\Big\{\big(g_j(\boldsymbol{Z}_1(\omega)), \ldots, g_j(\boldsymbol{Z}_n(\omega))\big) : g_j \in \big\{\{g_j\}_{j\in\mathbb{N}} \cap \mathcal{H}\big\}\Big\}\Big).$$

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

Proof of Proposition 2.1.2. Throughout the proof consider arbitrary $n \in \mathbb{N}$. Let $\Psi_n : \Omega \rightrightarrows \mathbb{R}^n$ be the correspondence $\Psi_n(\omega) \coloneqq \big\{g(\boldsymbol{Z}_1(\omega)), \ldots, g(\boldsymbol{Z}_n(\omega)) : g \in \mathcal{G}\big\}$. Define the function $\upsilon_n : \Omega \to \mathbb{R}$

$$\upsilon_n(\omega) \coloneqq \min_{\boldsymbol{x}\in\Psi_n(\omega)} U_n(\omega, \boldsymbol{x}) = \inf_{g\in\mathcal{G}} U_n\Big(\omega, \big\{g(\boldsymbol{Z}_t(\omega))\big\}_{t=1}^n\Big),$$

which exists because for each $\omega \in \Omega$, the function $U_n(\omega, \cdot) : \mathbb{R}^n \to \mathbb{R}$ is continuous and $\Psi_n(\omega) \subset \mathbb{R}^n$ is compact, since $\{g(\boldsymbol{z}) : g \in \mathcal{G}\}$ is compact for each $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$. Note that Ψ_n is a weakly measurable correspondence by the argument from the proof of Proposition 2.1.1 since $\mathcal{G}$ is pointwise-separable and U_n is measurable-$(\mathcal{A}\otimes\mathcal{B}(\mathbb{R}^n))/\mathcal{B}(\mathbb{R})$ by Aliprantis and Border (2006, Lemma 4.51). With this, and Aliprantis and Border (2006, Theorem 18.19),

$$\begin{aligned}\emptyset &\neq \Big\{\boldsymbol{x} \in \Psi_n(\omega) : U_n(\omega, \boldsymbol{x}) = \upsilon_n(\omega)\Big\} \\ &= \Big\{\big\{g(\boldsymbol{Z}_t(\omega))\big\}_{t=1}^n : g \in \mathcal{G},\, U_n\Big(\omega, \big\{g(\boldsymbol{Z}_t(\omega))\big\}_{t=1}^n\Big) = \upsilon_n(\omega)\Big\},\end{aligned} \tag{A.1}$$

and there exists a function $h_n : \Omega \to \mathbb{R}^n$, measurable-$\mathcal{A}/\mathcal{B}(\mathbb{R}^n)$, such that for all $\omega \in \Omega$

$$h_n(\omega) \in \Big\{\big\{g(\boldsymbol{Z}_t(\omega))\big\}_{t=1}^n : g \in \mathcal{G},\, U_n\Big(\omega, \big\{g(\boldsymbol{Z}_t(\omega))\big\}_{t=1}^n\Big) = \upsilon_n(\omega)\Big\} \subset \mathbb{R}^n. \tag{A.2}$$

Note that $\big(\mathcal{L}^r(\boldsymbol{\mathcal{Z}}, \mathcal{B}(\boldsymbol{\mathcal{Z}}), P_{\{\boldsymbol{Z}_t\}_{t=1}^n}), \|\cdot\|_{\mathcal{L}^r(P_{\{\boldsymbol{Z}_t\}_{t=1}^n})}\big)$ is a complete and separable metric space, i.e., a Polish space. This follows because $1 \leq r < \infty$, so $\mathcal{L}^r(\boldsymbol{\mathcal{Z}}, \mathcal{B}(\boldsymbol{\mathcal{Z}}), P_{\{\boldsymbol{Z}_t\}_{t=1}^n})$ is complete by the Riez-Fisher Theorem, and is separable by Cohn (2013, Proposition 3.4.5) since $\boldsymbol{\mathcal{Z}} \subseteq \mathbb{R}^{d_Z}$ implies $\mathcal{B}(\boldsymbol{\mathcal{Z}})$ is countably generated.[2] Then, $(\mathcal{G}, \|\cdot\|_{\mathcal{L}^r(P_{\{\boldsymbol{Z}_t\}_{t=1}^n})})$ is also a separable metric space

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

since $\mathcal{G} \subset \mathcal{L}^r(\boldsymbol{\mathcal{Z}}, \mathcal{B}(\boldsymbol{\mathcal{Z}}), P_{\{\boldsymbol{Z}_t\}_{t=1}^n})$. Hence, there exists a countable subset $\{g_j\}_{j\in\mathbb{N}} \subseteq \mathcal{G}$ such that $\mathcal{G} = \mathrm{cl}(\{g_j\}_{j\in\mathbb{N}})$. Then the correspondence $\Phi_n : \Omega \Rightarrow \mathcal{G}$ such that

$$\Phi_n(\omega) \coloneqq \Big\{ g \in \mathcal{G} : \{g(\boldsymbol{Z}_t(\omega))\}_{t=1}^n = h_n(\omega) \Big\} = \mathrm{cl}\Big(\Big\{ g \in \{g_j\}_{j\in\mathbb{N}} : \{g(\boldsymbol{Z}_t(\omega))\}_{t=1}^n = h_n(\omega) \Big\} \Big),$$

is closed since it is equal to the closure of a set, and $\Phi_n(\omega) \neq \emptyset$ by (A.1) and (A.2). Thus, by Aliprantis and Border (2006, Corollary 18.14-1, p.601), Φ_n is a weakly measurable correspondence. Note that Φ_n takes values in a Polish space since $\mathcal{G} \subset \mathcal{L}^r(\boldsymbol{\mathcal{Z}}, \mathcal{B}(\boldsymbol{\mathcal{Z}}), P_{\{\boldsymbol{Z}_t\}_{t=1}^n})$. Then, by Aliprantis and Border (2006, Theorem 18.13) the correspondence $\Phi_n(\omega)$ admits a measurable selector, i.e., there exists a function $s_n : \Omega \to \mathcal{G}$ that is measurable-$\mathcal{A}/\mathcal{B}(\mathcal{G})$, such that $s_n(\omega) \in \Phi_n(\omega)$ for all $\omega \in \Omega$. This implies the desired result since for all $\omega \in \Omega$

$$s_n(\omega) \in \Phi_n(\omega) \subseteq \Big\{ g \in \mathcal{G} : U_n\Big(\omega, \{g(\boldsymbol{Z}_t(\omega))\}_{t=1}^n\Big) = v_n(\omega) \Big\}$$

by the definition of h_n. ■

A.2 Proof of Theorem 2.1.1

For brevity, we write $q_t(f) \coloneqq q\big(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)\big)$, and $m_{nt} \coloneqq m_n(\boldsymbol{Z}_t)$. First, we present an ancillary lemma for the proof of Theorem 2.1.1.

Lemma A.2.1. *For some $c \geq 2$, let $\mathcal{H}_n \subset \mathcal{F}_n$ be such that $c\,\epsilon_n \leq \rho_n(f, \tilde{f}_n)$ for each $n \in \mathbb{N}$ and all $f \in \mathcal{H}_n$. Suppose (a.1), (a.2), and (a.3)(iii) hold. Then, for any $\delta \geq 0$ and $n \in \mathbb{N}$,*

$$\sup_{f\in\mathcal{H}_n} \left\{ \frac{1}{n} \sum_{t=1}^n \big[q_t(\tilde{f}_n) - q_t(f)\big] \mathbb{1}_{nt} \right\}$$
$$\leq \sup_{f\in\mathcal{H}_n} \frac{1}{n} \sum_{t=1}^n \left\{ \big[q_t(\tilde{f}_n) - q_t(f)\big] \mathbb{1}_{nt} - \mathbb{E}\Big[\big(q_t(\tilde{f}_n) - q_t(f)\big) \mathbb{1}_{nt} \Big] \right\} - \big(C_1 c^2/4 - C_2 - 2C_3\big)\epsilon_n^2,$$

where $\mathbb{1}_{nt} \coloneqq \mathbb{1}_{\{m_n(\boldsymbol{Z}_t) < M_n\}}$.

Proof. Recall $\tilde{f}_n \in \mathcal{F}_n$ by (a.1). Then, for any $f \in \mathcal{F}_n$ by (a.2)

$$\mathbb{E}\big[Q_n(\tilde{f}_n) - Q_n(f)\big] = \mathbb{E}\big[Q_n(\tilde{f}_n)\big] - \mathbb{E}\big[Q_n(f_0)\big] + \mathbb{E}\big[Q_n(f_0)\big] - \mathbb{E}\big[Q_n(f)\big]$$
$$\leq C_2\, \rho_n(\tilde{f}_n, f_0)^2 - C_1\, \rho_n(f, f_0)^2.$$

By (a.1), $\rho_n(\tilde{f}_n, f_0) \le \epsilon_n \le c\,\epsilon_n/2$, since $c \ge 2$ by assumption. With this, and the triangle inequality, for any $f \in \mathcal{F}_n$ such that $c\,\epsilon_n \le \rho_n(f, \tilde{f}_n)$,

$$\rho_n(f, f_0) \;\ge\; \rho_n(f, \tilde{f}_n) - \rho_n(\tilde{f}_n, f_0) \;\ge\; c\,\epsilon_n - c\,\epsilon_n/2 \;\ge\; c\,\epsilon_n/2 \;>\; 0.$$

Combining the previous two displays,

$$\sup_{f \in \mathcal{H}_n} \mathbb{E}\big[Q_n(\tilde{f}_n) - Q_n(f)\big] \le C_2\,\epsilon_n^2 - C_1 c^2 \epsilon_n^2/4 = \big(C_2 - C_1 c^2/4\big)\epsilon_n^2. \tag{A.3}$$

Next, let $\mathbb{1}_{nt}^{\mathrm{c}} \coloneqq \mathbb{1}_{\{m_n(\boldsymbol{Z}_t) \ge M_n\}}$, so we have

$$\begin{aligned}
-\mathbb{E}\big[Q_n(\tilde{f}_n) - Q_n(f)\big] &= \frac{1}{n}\sum_{t=1}^{n}\Big\{ -\mathbb{E}\Big[\big(q_t(\tilde{f}_n) - q_t(f)\big)\mathbb{1}_{nt}^{\mathrm{c}}\Big] - \mathbb{E}\Big[\big(q_t(\tilde{f}_n) - q_t(f)\big)\mathbb{1}_{nt}\Big]\Big\} \\
&\le \frac{1}{n}\sum_{t=1}^{n}\Big\{\mathbb{E}\Big[\big|q_t(\tilde{f}_n) - q_t(f)\big|\mathbb{1}_{nt}^{\mathrm{c}}\Big] - \mathbb{E}\Big[\big(q_t(\tilde{f}_n) - q_t(f)\big)\mathbb{1}_{nt}\Big]\Big\}.
\end{aligned}$$

By (a.3)(iii),

$$\begin{aligned}
\sup_{f \in \mathcal{H}_n} \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\Big[\big|q_t(\tilde{f}_n) - q_t(f)\big|\mathbb{1}_{nt}^{\mathrm{c}}\Big] &\le \sup_{f \in \mathcal{H}_n} \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\Big[\big(|q_t(\tilde{f}_n)| + |q_t(f)|\big)\mathbb{1}_{nt}^{\mathrm{c}}\Big] \\
&\le \sup_{f \in \mathcal{F}_n} \frac{2}{n}\sum_{t=1}^{n}\mathbb{E}\Big[\big|q_t(f)\big|\mathbb{1}_{nt}^{\mathrm{c}}\Big] \;\le\; 2C_3\,\epsilon_n^2.
\end{aligned}$$

Combining the previous two displays, for any $f \in \mathcal{H}_n$,

$$-\mathbb{E}\big[Q_n(\tilde{f}_n) - Q_n(f)\big] \le 2C_3\,\epsilon_n^2 - \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\Big[\big(q_t(\tilde{f}_n) - q_t(f)\big)\mathbb{1}_{nt}\Big]. \tag{A.4}$$

With (A.3) and (A.4),

$$\begin{aligned}
&\sup_{f \in \mathcal{H}_n}\Big\{\frac{1}{n}\sum_{t=1}^{n}\big[q_t(\tilde{f}_n) - q_t(f)\big]\mathbb{1}_{nt}\Big\} \\
&\quad= \sup_{f \in \mathcal{H}_n}\frac{1}{n}\sum_{t=1}^{n}\Big\{\big[q_t(\tilde{f}_n) - q_t(f)\big]\mathbb{1}_{nt} + \mathbb{E}\big[Q_n(\tilde{f}_n) - Q_n(f)\big] - \mathbb{E}\big[Q_n(\tilde{f}_n) - Q_n(f)\big]\Big\} \\
&\quad\le \sup_{f \in \mathcal{H}_n}\frac{1}{n}\sum_{t=1}^{n}\Big\{\big[q_t(\tilde{f}_n) - q_t(f)\big]\mathbb{1}_{nt} - \mathbb{E}\Big[\big(q_t(\tilde{f}_n) - q_t(f)\big)\mathbb{1}_{nt}\Big]\Big\} + \big(C_2 - C_1 c^2/4\big)\epsilon_n^2 + 2C_3\,\epsilon_n^2.
\end{aligned}$$

■

For the proof of Theorem 2.1.1 it will be convenient to use the packing number, as defined below.

Definition A.2.1. (Packing Number) Let $\delta > 0$, and let $(\mathbb{M}, \|\cdot\|)$ be a semi-metric space.

(*i*) A set $\mathcal{G} \subseteq \mathbb{M}$ is **δ-separated** if $\|g - g'\| \geq \delta$ for any $g, g' \in \mathcal{G}$ with $g \neq g'$. The **δ-packing number**, is the maximum number of δ-separated points in $\mathbb{M}$.

(*ii*) When $\mathbb{M}$ is a space of functions, $f : \boldsymbol{\mathcal{Z}} \to \mathbb{R}$, for any $r \geq 1$, and $a \in \mathbb{N}$ define $\mathrm{D}_r^{(\infty)}(\delta, \mathbb{M}, a) \coloneqq \sup\Big\{\mathrm{D}\big(\delta, \mathbb{M}|_{\{\boldsymbol{Z}_t(\omega)\}_{t=1}^a}, \|\cdot\|_{r,a}\big) : \omega \in \Omega\Big\}$.

For a metric space $(\mathbb{M}, \|\cdot\|)$ the following will be used to create a δ-cover of $\mathbb{M}$ using a δ-separated subset of maximum size. Note that this also implies $\mathrm{N}(\delta, \mathbb{M}, \|\cdot\|) \leq \mathrm{D}(\delta, \mathbb{M}, \|\cdot\|)$ for any $\delta > 0$ and metric space $(\mathbb{M}, \|\cdot\|)$.

Lemma A.2.2. *For a semi-metric space $(\mathbb{M}, \|\cdot\|)$, and any $\delta > 0$, if $\{f_j\}_{j=1}^J \subseteq \mathbb{M}$ is δ-separated and $J = \mathrm{D}(\delta, \mathbb{M}, \|\cdot\|) \in \mathbb{N}$, then*

$$\mathbb{M} \subseteq \bigcup_{j=1}^{J} \Big\{f \in \mathbb{M} : \|f - f_j\| < \delta\Big\}.$$

Proof. See Kosorok (2008, §8.1, p.132). ∎

Lemma A.2.3. *For any nonstochastic $f^* \in \mathcal{F}$ and $\mathcal{H}_n \subseteq \mathcal{F}_n$, we have*

$$P\left(\sup_{f \in \mathcal{H}_n} \left\{\frac{1}{n}\sum_{t=1}^{n} \big[q_t(f^*) - q_t(f)\big] \mathbb{1}_{\{m_n(\boldsymbol{Z}_t) \geq M_n\}}\right\} > 0\right) \leq P\Big(\max_{t \in \{1,\ldots,n\}} m_{nt} \geq M_n\Big).$$

Proof. Define $\mathbb{1}_{nt}^{\mathrm{c}} : \Omega \to \{0, 1\}$ where $\mathbb{1}_{nt}^{\mathrm{c}}(\omega) \coloneqq \mathbb{1}_{\{m_n(\boldsymbol{Z}_t(\omega)) \geq M_n\}}$. Note that $\sum_{t=1}^n \big[q_t(f^*) - q_t(f)\big] \mathbb{1}_{nt}^{\mathrm{c}} > 0$ implies $\big[q_t(f^*) - q_t(f)\big] \mathbb{1}_{nt}^{\mathrm{c}} > 0$ for at least one $t \in \{1, \ldots, n\}$. Thus,

$$\begin{aligned}
&\left\{\omega : \sup_{f \in \mathcal{H}_n} \sum_{t=1}^{n} \Big[q\big(\boldsymbol{Z}_t(\omega), f^*(\boldsymbol{Z}_t(\omega)\big) - q\big(\boldsymbol{Z}_t(\omega), f(\boldsymbol{Z}_t(\omega))\big)\Big] \mathbb{1}_{nt}^{\mathrm{c}}(\omega) > 0\right\} \\
&\subseteq \left\{\omega : \max_{t \in \{1,\ldots,n\}} \left\{\sup_{f \in \mathcal{H}_n} \Big[q\big(\boldsymbol{Z}_t(\omega), f^*(\boldsymbol{Z}_t(\omega)\big) - q\big(\boldsymbol{Z}_t(\omega), f(\boldsymbol{Z}_t(\omega))\big)\Big] \mathbb{1}_{nt}^{\mathrm{c}}(\omega)\right\} > 0\right\}.
\end{aligned}$$

Next, $\max_{t \in \{1,\ldots,n\}} \big[q_t(f^*) - q_t(f)\big] \mathbb{1}_{nt}^{\mathrm{c}} > 0$ implies

$$\max_{t \in \{1,\ldots,n\}} \big[q_t(f^*) - q_t(f)\big] > 0, \quad \text{and} \quad \max_{t \in \{1,\ldots,n\}} \mathbb{1}_{nt}^{\mathrm{c}} > 0.$$

Hence,

$$\Big\{\omega : \max_{t\in\{1,\dots,n\}} \Big\{ \sup_{f\in\mathcal{H}_n} \Big[q\big(\boldsymbol{Z}_t(\omega), f^*(\boldsymbol{Z}_t(\omega)\big) - q\big(\boldsymbol{Z}_t(\omega), f(\boldsymbol{Z}_t(\omega))\big)\Big]\mathbb{1}^{\mathrm{c}}_{nt}(\omega)\Big\} > 0\Big\}$$
$$\subseteq \Big\{\omega : \max_{t\in\{1,\dots,n\}} \Big\{ \sup_{f\in\mathcal{H}_n} \Big[q\big(\boldsymbol{Z}_t(\omega), f^*(\boldsymbol{Z}_t(\omega)\big) - q\big(\boldsymbol{Z}_t(\omega), f(\boldsymbol{Z}_t(\omega))\big)\Big]\Big\} > 0\Big\} \bigcap \Big\{\omega : \max_{t\in\{1,\dots,n\}} \mathbb{1}^{\mathrm{c}}_{nt}(\omega) > 0\Big\}$$
$$\subseteq \Big\{\omega : \max_{t\in\{1,\dots,n\}} \mathbb{1}^{\mathrm{c}}_{nt}(\omega) > 0\Big\}.$$

By definition, $\mathbb{1}^{\mathrm{c}}_{nt} = 1$ if $m_n(\boldsymbol{Z}_t) \geq M_n$, otherwise $\mathbb{1}^{\mathrm{c}}_{nt} = 0$. Therefore,

$$\Big\{\omega : \max_{t\in\{1,\dots,n\}} \mathbb{1}^{\mathrm{c}}_{nt}(\omega) > 0\Big\} = \Big\{\omega : \max_{t\in\{1,\dots,n\}} m_n\big(\boldsymbol{Z}_t(\omega)\big) \geq M_n.\Big\}$$

Combining the previous three displays implies the desired result. ■

Proof of Theorem 2.1.1. Let c be a constant such that $c > \max\Big\{2, \sqrt{4(C_2 + 2C_3)/C_1}\Big\}$. By the triangle inequality

$$\begin{aligned} P^*\Big(\rho_n(\hat{f}_n, f_0) \geq 2c\,\epsilon_n\Big) &\leq P^*\Big(\rho_n(\hat{f}_n, \tilde{f}_n) \geq c\,\epsilon_n\Big) + P\Big(\rho_n(\tilde{f}_n, f_0) \geq c\,\epsilon_n\Big) \\ &= P^*\Big(\rho_n(\hat{f}_n, \tilde{f}_n) \geq c\,\epsilon_n\Big). \end{aligned} \tag{A.5}$$

since $c > 1$ and $\rho_n(\tilde{f}_n, f_0) \leq \epsilon_n$ by (a.1). Define

$$\mathcal{H}_n(c\,\epsilon_n) := \Big\{f \in \mathcal{F}_n : c\,\epsilon_n \leq \rho_n(f, \tilde{f}_n)\Big\}.$$

By (2.2),

$$\hat{f}_n \in \mathcal{H}_n(c\,\epsilon_n) \implies \inf_{f\in\mathcal{H}_n(c\,\epsilon_n)} Q_n(f) \leq Q_n(\hat{f}_n) \leq \inf_{f\in\mathcal{F}_n} Q_n(f) + \theta_n.$$

Using $\tilde{f}_n \in \mathcal{F}_n$, and $-\inf_{f\in\mathcal{F}_n} Q_n(f) = \sup_{f\in\mathcal{F}_n} -Q_n(f)$, the previous display implies

$$\begin{aligned} P^*\Big(\rho_n(\hat{f}_n, \tilde{f}_n) \geq c\,\epsilon_n\Big) &\leq P\Big(Q_n(\tilde{f}_n) + \theta_n \geq \inf_{f\in\mathcal{H}_n(c\,\epsilon_n)} Q_n(f)\Big) \\ &= P\Big(\sup_{f\in\mathcal{H}_n(c\,\epsilon_n)} \Big\{Q_n(\tilde{f}_n) - Q_n(f)\Big\} + \theta_n \geq 0\Big). \end{aligned}$$

Let $\mathbb{1}_{nt} := \mathbb{1}_{\{m_n(\boldsymbol{Z}_t) < M_n\}}$, and $\mathbb{1}^{\mathrm{c}}_{nt} := \mathbb{1}_{\{m_n(\boldsymbol{Z}_t) \geq M_n\}}$. With this

$$\begin{aligned} P^*\Big(\rho_n(\hat{f}_n, \tilde{f}_n) \geq c\,\epsilon_n\Big) &\leq P\Big(\sup_{f\in\mathcal{H}_n(c\,\epsilon_n)} \frac{1}{n}\sum_{t=1}^{n} \Big\{\big[q_t(\tilde{f}_n) - q_t(f)\big]\mathbb{1}_{nt} + \big[q_t(\tilde{f}_n) - q_t(f)\big]\mathbb{1}^{\mathrm{c}}_{nt}\Big\} + \theta_n \geq 0\Big) \\ &\leq P\Big(\sup_{f\in\mathcal{H}_n(c\,\epsilon_n)} \Big\{\frac{1}{n}\sum_{t=1}^{n} \big[q_t(\tilde{f}_n) - q_t(f)\big]\mathbb{1}_{nt}\Big\} + \theta_n \geq 0\Big) \\ &\quad + P\Big(\sup_{f\in\mathcal{H}_n(c\,\epsilon_n)} \Big\{\frac{1}{n}\sum_{t=1}^{n} \big[q_t(\tilde{f}_n) - q_t(f)\big]\mathbb{1}^{\mathrm{c}}_{nt}\Big\} > 0\Big). \end{aligned} \tag{A.6}$$

By Lemma A.2.3 and (a.3)(ii), for any $\upsilon > 0$ there exists $N_\upsilon \in \mathbb{N}$, such that for all $n \geq N_\upsilon$

$$P\left(\sup_{f \in \mathcal{H}_n(c\,\epsilon_n)} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left[q_t(\tilde{f}_n) - q_t(f) \right] \mathbb{1}_{nt}^{\mathrm{c}} \right\} > 0 \right) \leq P\Big(\max_{t \in \{1,\ldots,n\}} m_{nt} \geq M_n \Big) \leq \upsilon.$$

Henceforth let $n \geq N_\upsilon$. Then, combining the previous display, (A.6), and (A.5),

$$P^*\Big(\rho_n(\hat{f}_n, f_0) \geq 2c\,\epsilon_n\Big) \leq P\left(\sup_{f \in \mathcal{H}_n(c\,\epsilon_n)} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left[q_t(\tilde{f}_n) - q_t(f) \right] \mathbb{1}_{\{m_n(\boldsymbol{Z}_t) < M_n\}} \right\} + \theta_n \geq 0 \right) + \upsilon. \tag{A.7}$$

Let

$$U_{nt}^{(q)}(f) \coloneqq q_t(f) \mathbb{1}_{\{m_n(\boldsymbol{Z}_t) < M_n\}} - \mathbb{E}\big[q_t(f) \mathbb{1}_{\{m_n(\boldsymbol{Z}_t) < M_n\}} \big], \quad \text{and}$$
$$\bar{c} \coloneqq \sqrt{(C_1 c^2/4 - C_2 - 2C_3)/2},$$

where $\bar{c} > 0$ since $c > \sqrt{4(C_2 + 2C_3)/C_1}$. Then, by Lemma A.2.1, and (A.7),

$$\begin{aligned}
P^*\Big(\rho_n(\hat{f}_n, f_0) \geq 2c\,\epsilon_n\Big) &\leq P\left(\sup_{f \in \mathcal{H}_n(c\,\epsilon_n)} \frac{1}{n} \sum_{t=1}^{n} \left\{ U_{nt}^{(q)}(\tilde{f}_n) - U_{nt}^{(q)}(f) \right\} + \theta_n \geq 2(\bar{c}\,\epsilon_n)^2 \right) + \upsilon \\
&\leq P\left(\sup_{f \in \mathcal{H}_n(c\,\epsilon_n)} \frac{1}{n} \sum_{t=1}^{n} \left\{ U_{nt}^{(q)}(\tilde{f}_n) - U_{nt}^{(q)}(f) \right\} \geq (\bar{c}\,\epsilon_n)^2 \right) + P\Big(\theta_n \geq (\bar{c}\,\epsilon_n)^2 \Big) + \upsilon \\
&\leq P\left(\sup_{f \in \mathcal{H}_n(c\,\epsilon_n)} \left| \frac{1}{n} \sum_{t=1}^{n} \left\{ U_{nt}^{(q)}(\tilde{f}_n) - U_{nt}^{(q)}(f) \right\} \right| \geq (\bar{c}\,\epsilon_n)^2 \right) + 2\upsilon,
\end{aligned} \tag{A.8}$$

where the last inequality follows because $\theta_n = \mathrm{O}_P(\epsilon_n^2)$ by assumption, and $\bar{c}$ increases with c, so $P\big(\theta_n \geq C_1(c\,\epsilon_n)^2\big) \leq \upsilon$ for any $\upsilon > 0$, when n, and c are sufficiently large.

For $\kappa_n \coloneqq (\bar{c}\,\epsilon_n)^2/(12M_n)$, define

$$\mathrm{D}_n(\omega) \coloneqq \mathrm{D}\big(\kappa_n,\ \mathcal{H}_n(c\,\epsilon_n)|_{\{\boldsymbol{Z}_t(\omega)\}_{t=1}^n},\ \|\cdot\|_{1,n}\big),$$

and let $\{f_{nj}\}_{j=1}^{\mathrm{D}_n(\omega)}$ be a κ_n-separated set in $\mathcal{H}_n(c\,\epsilon_n)$ with respect to $\|\cdot\|_{1,n}$. Define

$$\mathcal{G}_n^j \coloneqq \Big\{ f \in \mathcal{H}_n(c\,\epsilon_n) : \|f - f_{nj}\|_{1,n} < \kappa_n \Big\}, \qquad j = 1, \ldots, \mathrm{D}_n(\omega),$$

and by Lemma A.2.2 we have $\mathcal{H}_n(c\,\epsilon_n) \subseteq \bigcup_{j=1}^{\mathrm{D}_n(\omega)} \mathcal{G}_n^j$. With this,

$$\begin{aligned}
&P\left(\sup_{f \in \mathcal{H}_n(c\,\epsilon_n)} \frac{1}{n} \left| \sum_{t=1}^{n} \left[U_{nt}^{(q)}(\tilde{f}_n) - U_{nt}^{(q)}(f) \right] \right| \geq (\bar{c}\,\epsilon_n)^2 \right) \\
&\qquad = P\left(\bigcup_{j=1}^{\mathrm{D}_n(\omega)} \left\{ \sup_{f \in \mathcal{G}_n^j} \left| \frac{1}{n} \sum_{t=1}^{n} \left[U_{nt}^{(q)}(\tilde{f}_n) - U_{nt}^{(q)}(f) \right] \right| \geq (\bar{c}\,\epsilon_n)^2 \right\} \right).
\end{aligned} \tag{A.9}$$

By (a.3)(i),

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(f)\right| \\
&= \left|\frac{1}{n}\sum_{t=1}^{n}\Big\{\big(q_t(f)-q_t(f_{nj})\big)\mathbb{1}_{nt} + q_t(f_{nj})\mathbb{1}_{nt} - \mathbb{E}\big[q_t(f_{nj})\mathbb{1}_{nt}\big] + \mathbb{E}\Big[\big(q_t(f_{nj})-q_t(f)\big)\mathbb{1}_{nt}\Big]\Big\}\right| \\
&\le M_n\|f-f_{nj}\|_{1,n} + \left|\frac{1}{n}\sum_{t=1}^{n}\Big\{q_t(f_{nj})\mathbb{1}_{nt} - \mathbb{E}\big[q_t(f_{nj})\mathbb{1}_{nt}\big]\Big\}\right| + M_n\mathbb{E}\Big[\|f-f_{nj}\|_{1,n}\Big].
\end{aligned}$$

Then, for any $f \in \mathcal{G}_n^j$,

$$\left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(f)\right| \le 2M_n\kappa_n + \left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(f_{nj})\right|.$$

Therefore,

$$\begin{aligned}
\sup_{f\in\mathcal{G}_n^j}\left|\frac{1}{n}\sum_{t=1}^{n}\Big[U_{nt}^{(q)}(\tilde{f}_n) - U_{nt}^{(q)}(f)\Big]\right| &\le \left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(\tilde{f}_n)\right| + \sup_{f\in\mathcal{G}_n^j}\left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(f)\right| \\
&\le 2M_n\kappa_n + \left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(\tilde{f}_n)\right| + \left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(f_{nj})\right|.
\end{aligned}$$

From this, we obtain

$$P\left(\bigcup_{j=1}^{\mathrm{D}_n(\omega)}\left\{\sup_{f\in\mathcal{G}_n^j}\left|\frac{1}{n}\sum_{t=1}^{n}\Big[U_{nt}^{(q)}(\tilde{f}_n) - U_{nt}^{(q)}(f)\Big]\right| \ge (\bar{c}\,\epsilon_n)^2\right\}\right) \le P\left(\bigcup_{j=1}^{\mathrm{D}_n(\omega)} E_{nj}\right), \tag{A.10}$$

where

$$E_{nj} := \left\{2M_n\kappa_n + \left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(\tilde{f}_n)\right| + \left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(f_{nj})\right| \ge (\bar{c}\,\epsilon_n)^2\right\}$$

To make E_{nj} defined for $j \ge n$, we append the sequence $\{f_{nj}\}_{j=1}^{\mathrm{D}_n(\omega)}$ by setting $f_{nj} := \tilde{f}_n$ for all $j > \mathrm{D}_n(\omega)$. Let $\overline{\mathrm{D}}_n := \mathrm{D}_1^{(\infty)}\big(\kappa_n, \mathcal{H}_n(c\,\epsilon_n), n\big)$. Since $\overline{\mathrm{D}}_n$ is non-stochastic we have

$$P\left(\bigcup_{j=1}^{\mathrm{D}_n(\omega)} E_{nj}\right) \le P\left(\bigcup_{j=1}^{\overline{\mathrm{D}}_n} E_{nj}\right) \le \sum_{j=1}^{\overline{\mathrm{D}}_n} P\Big(E_{nj}\Big) \le \overline{\mathrm{D}}_n \max_{j\in\{1,\ldots,\overline{\mathrm{D}}_n\}} P\big(E_{nj}\big). \tag{A.11}$$

Recall, $\kappa_n := (\bar{c}\,\epsilon_n)^2/(12M_n)$, so

$$2M_n\kappa_n = \frac{(\bar{c}\,\epsilon_n)^2}{6} < \frac{(\bar{c}\,\epsilon_n)^2}{3},$$

which, together with (a.4)(i), implies that for any $n, j \in \mathbb{N}$,

$$\begin{aligned}
P\big(E_{nj}\big) &\le P\left(\left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(\tilde{f}_n)\right| \ge \frac{(\bar{c}\,\epsilon_n)^2}{3}\right) + P\left(\left|\frac{1}{n}\sum_{t=1}^{n} U_{nt}^{(q)}(f_{nj})\right| \ge \frac{(\bar{c}\,\epsilon_n)^2}{3}\right) \\
&\le 2\lambda_n^{(q)}\left(\frac{(\bar{c}\,\epsilon_n)^2}{3}\right).
\end{aligned} \tag{A.12}$$

Combining (A.8), (A.9), (A.10), (A.11), and (A.12),

$$P^*\Big(\rho_n(\hat{f}_n, f_0) \geq 2c\,\epsilon_n\Big) \lesssim \overline{\mathrm{D}}_n \cdot \lambda_n^{(q)}\left(\frac{(\overline{c}\,\epsilon_n)^2}{3}\right) + 2\upsilon. \tag{A.13}$$

Note that

$$\overline{\mathrm{D}}_n \leq \mathrm{D}_1^{(\infty)}\big(\kappa_n, \mathcal{F}_n, n\big) \leq \mathrm{N}_1^{(\infty)}\big(\kappa_n/2, \mathcal{F}_n, n\big),$$

where the first inequality uses $\mathcal{H}_n(c\,\epsilon_n) \subseteq \mathcal{F}_n$, and the second inequality uses $\mathrm{D}_1^{(\infty)}\big(\delta, \mathcal{F}_n, n\big) \leq \mathrm{N}_1^{(\infty)}\big(\delta/2, \mathcal{F}_n, n\big)$ for any $\delta > 0$ (van der Vaart and Wellner, 1996, p.98). With this, and (A.13)

$$P^*\Big(\rho_n(\hat{f}_n, f_0) \geq 2c\,\epsilon_n\Big) \lesssim \mathrm{N}_1^{(\infty)}\left(\frac{(\overline{c}\,\epsilon_n)^2}{24M_n}, \mathcal{F}_n, n\right) \cdot \lambda_n^{(q)}\left(\frac{(\overline{c}\,\epsilon_n)^2}{3}\right) + 2\upsilon,$$

for any $\upsilon > 0$ and n, c sufficiently large. Thus, (a.4)(ii) completes the proof by setting $\eta = 3\,\overline{c}^{\,2} = 3(C_1c^2 - C_2 - 4C_3)/2$ therein, and choosing c sufficiently large. ■

A.3 Proof of Theorem 2.1.2

The following proof applies a localization analysis technique to obtain a nonasymptotic bound on the $\mathcal{L}^2$ error of sieve estimators. The steps used here follow those used by Farrell et al. (2021), and are named similarly. However, the results obtained here apply to a wider variety of estimators in more general estimation settings.

Appendix A.3.7 lists the ancillary lemmas used in this section. As before, we write $q_t(f) \coloneqq q\big(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)\big)$, and $m_{nt} \coloneqq m_n(\boldsymbol{Z}_t)$.

A.3.1 Main Decomposition

Let $\mathbb{1}_{nt} \coloneqq \mathbb{1}\{m_n(\boldsymbol{Z}_t) \leq C_4B_n\}$, and $\mathbb{1}_{nt}^{\mathrm{c}} \coloneqq \mathbb{1}\{m_n(\boldsymbol{Z}_t) > C_4B_n\}$, then define

$$g_f(\boldsymbol{Z}_t) \coloneqq \big[q_t(f) - q_t(f_0)\big]\mathbb{1}_{nt}, \quad \text{and} \quad g_f^{\mathrm{c}}(\boldsymbol{Z}_t) \coloneqq \big[q_t(f) - q_t(f_0)\big]\mathbb{1}_{nt}^{\mathrm{c}}.$$

By (2.2),

$$0 \leq -Q_n(\hat{f}_n) + Q_n(\tilde{f}_n) + \theta_n.$$

With this, and (b.3),

$$
\begin{aligned}
&C_1\|\hat{f}_n - f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq \mathbb{E}[Q_n(\hat{f}_n)] - \mathbb{E}[Q_n(f_0)] \\
&\leq \mathbb{E}\big[Q_n(\hat{f}_n)\big] - \mathbb{E}[Q_n(f_0)\big] - Q_n(\hat{f}_n) + Q_n(\tilde{f}_n) + \theta_n \\
&= \mathbb{E}\big[\big(Q_n(\hat{f}_n) - Q_n(f_0)\big)\big] - \big[Q_n(\hat{f}_n) - Q_n(f_0)\big] + Q_n(\tilde{f}_n) - Q_n(f_0) + \theta_n \\
&= \underbrace{\frac{1}{n}\sum_{t=1}^{n}\Big\{\mathbb{E}\big[g_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g_{\hat{f}_n}(\boldsymbol{Z}_t)\Big\}}_{\text{Empirical Process Term}} + \underbrace{\big[Q_n(\tilde{f}_n) - Q_n(f_0)\big]}_{\text{Bias Term}} + \underbrace{\frac{1}{n}\sum_{t=1}^{n}\Big\{\mathbb{E}\big[g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\Big\}}_{\text{Truncation Term}} + \theta_n.
\end{aligned}
\tag{A.14}
$$

A.3.2 Truncation Term

By stationarity, and the triangle inequality

$$
\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n}\Big\{\mathbb{E}\big[g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\Big\} &= \mathbb{E}\big[g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - \frac{1}{n}\sum_{t=1}^{n} g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t) \\
&\leq \mathbb{E}\big[\,|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)|\,\big] + \frac{1}{n}\sum_{t=1}^{n}\big|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\big|.
\end{aligned}
\tag{A.15}
$$

First, by Lemma A.2.3

$$
\begin{aligned}
P\left(\frac{1}{n}\sum_{t=1}^{n}\big|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\big| > 0\right) &\leq P\left(\sup_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{t=1}^{n}\big|g^{\mathrm{c}}_{f}(\boldsymbol{Z}_t)\big| > 0\right) \\
&= P\left(\sup_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{t=1}^{n}\big|q_t(f) - q_t(f_0)\big|\mathbb{1}^{\mathrm{c}}_{nt} > 0\right) \\
&\leq P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big).
\end{aligned}
\tag{A.16}
$$

Next, note that $\mathbb{E}\big[|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)|\big] \leq 2\mu_n$ for μ_n defined as in (b.5)(ii). To see this, by the triangle inequality

$$
\mathbb{E}\big[|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)|\big] \leq \mathbb{E}\Big[\big|q_t(\hat{f}_n) - q_t(f_0)\big|\mathbb{1}^{\mathrm{c}}_{nt}\Big] \leq 2\sup_{f\in\{\mathcal{F}_n\cup\{f_0\}\}}\mathbb{E}\Big[\big|q_t(f)\big|\mathbb{1}^{\mathrm{c}}_{nt}\Big];
$$

also, using (b.5)(i), $\sup_{f\in\mathcal{F}_n}\|f\|_\infty \leq B_n$ and $B_n \geq 2$ by (b.4), with $\|f_0\|_\infty \leq 1$,

$$
\mathbb{E}\big[|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)|\big] \leq \|\hat{f}_n - f_0\|_\infty\,\mathbb{E}\big[m_{nt}\mathbb{1}^{\mathrm{c}}_{nt}\big] \leq (B_n+1)\mathbb{E}\big[m_{nt}\mathbb{1}^{\mathrm{c}}_{nt}\big] \leq 2B_n\mathbb{E}\big[m_{nt}\mathbb{1}^{\mathrm{c}}_{nt}\big].
$$

Hence, $\mathbb{E}\big[|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)|\big] < 3\mu_n$, since $3\mu_n > 2\mu_n \geq g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)$ becuase $\{\mu_n\}_{n\in\mathbb{N}}$ is strictly positive. With this, (A.15) and (A.16),

$$\begin{aligned} P\left(\frac{1}{n}\sum_{t=1}^{n}\Big\{\mathbb{E}\big[g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\Big\} \geq 3\mu_n\right) &\leq P\left(\mathbb{E}\big[\,|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)|\,\big] + \frac{1}{n}\sum_{t=1}^{n}\big|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\big| \geq 3\mu_n\right) \\ &\leq P\left(3\mu_n + \frac{1}{n}\sum_{t=1}^{n}\big|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\big| > 3\mu_n\right) \\ &= P\left(\frac{1}{n}\sum_{t=1}^{n}\big|g^{\mathrm{c}}_{\hat{f}_n}(\boldsymbol{Z}_t)\big| > 0\right) \\ &\leq P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big). \end{aligned} \tag{A.17}$$

A.3.3 Bias Term

By (A.16), for any $A > 0$,

$$\begin{aligned} P\left(Q_n(\tilde{f}_n) - Q_n(f_0) \geq A\right) &= P\left(\frac{1}{n}\sum_{t=1}^{n}\big\{g_{\tilde{f}_n}(\boldsymbol{Z}_t) + g^{\mathrm{c}}_{\tilde{f}_n}(\boldsymbol{Z}_t)\big\} \geq A\right) \\ &\leq P\left(\frac{1}{n}\sum_{t=1}^{n} g_{\tilde{f}_n}(\boldsymbol{Z}_t) \geq A\right) + P\left(\frac{1}{n}\sum_{t=1}^{n}\big|g^{\mathrm{c}}_{f}(\boldsymbol{Z}_t)\big| > 0\right) \\ &\leq P\left(\frac{1}{n}\sum_{t=1}^{n} g_{\tilde{f}_n}(\boldsymbol{Z}_t) \geq A\right) + P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big), \end{aligned}$$

By (b.5)(i) and (b.2), $\Big|\mathbb{E}\big[g_{\tilde{f}_n}(\boldsymbol{Z}_t)\big]\Big| \leq \|g_{\tilde{f}_n}\|_\infty \leq C_4 B_n \|\tilde{f}_n - f_0\|_\infty \leq C_4 B_n \tilde{\epsilon}_n$. Hence,

$$\Big\|g_{\tilde{f}_n}(\boldsymbol{Z}_t) - \mathbb{E}\big[g_{\tilde{f}_n}(\boldsymbol{Z}_t)\big]\Big\|_\infty \leq \|g_{\tilde{f}_n}\|_\infty + \Big|\mathbb{E}\big[g_{\tilde{f}_n}(\boldsymbol{Z}_t)\big]\Big| \leq 2C_4 B_n \tilde{\epsilon}_n.$$

Then,

$$\begin{aligned} \Big|\mathbb{E}\big[g_{\tilde{f}_n}(\boldsymbol{Z}_t)\big]\Big| &= \Big|\mathbb{E}\big[q_t(\tilde{f}_n) - q_t(f_0)\big] - \mathbb{E}\big[g^{\mathrm{c}}_{\tilde{f}_n}(\boldsymbol{Z}_t)\big]\Big| \\ &\leq \Big|\mathbb{E}\big[q_t(\tilde{f}_n) - q_t(f_0)\big]\Big| + \Big|\mathbb{E}\big[g^{\mathrm{c}}_{\tilde{f}_n}(\boldsymbol{Z}_t)\big]\Big| \\ &= \mathbb{E}\big[q_t(\tilde{f}_n) - q_t(f_0)\big] + \Big|\mathbb{E}\big[g^{\mathrm{c}}_{\tilde{f}_n}(\boldsymbol{Z}_t)\big]\Big| \\ &\leq C_2\|\tilde{f}_n - f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})} + \mathbb{E}\big[|g^{\mathrm{c}}_{\tilde{f}_n}(\boldsymbol{Z}_t)|\big] \\ &\leq C_2\,\tilde{\epsilon}_n^2 + 2\mu_n, \end{aligned}$$

where the first line uses $\mathbb{E}\big[q_t(\tilde{f}_n) - q_t(f_0)\big] = \mathbb{E}\big[g_{\tilde{f}_n}(\boldsymbol{Z}_t)\big] + \mathbb{E}\big[g^c_{\tilde{f}_n}(\boldsymbol{Z}_t)\big]$; the second line uses the triangle inequality; the third line uses stationarity with $\mathbb{E}[Q_n(f_0)] \leq \mathbb{E}[Q_n(\tilde{f}_n)]$ by the definition of f_0 and $\tilde{f}_n \in \mathcal{F}_n \subseteq \mathcal{F}$; the fourth line uses (b.3); and the last line uses (b.2) and $\mathbb{E}\big[|g^c_{\tilde{f}_n}(\boldsymbol{Z}_t)|\big] \leq 2\mu_n$ which was shown in the last section. Then, by Lemma A.3.2, for $C_5 = 2C_4/\min\{C_6, 1\}$, and any $\delta' > 0$,

$$
\begin{aligned}
e^{-\delta'} &\geq P\left(\frac{1}{n}\sum_{t=1}^{n} g_{\tilde{f}_n}(\boldsymbol{Z}_t) - \mathbb{E}\big[g_{\tilde{f}_n}(\boldsymbol{Z}_t)\big] \geq C_5 B_n \tilde{\epsilon}_n \left[\frac{\delta'(\log n)(\log\log n)}{n} + \sqrt{\frac{\delta'}{n}}\right]\right) \\
&\geq P\left(\frac{1}{n}\sum_{t=1}^{n} g_{\tilde{f}_n}(\boldsymbol{Z}_t) \geq C_2\,\tilde{\epsilon}_n^2 + 2\mu_n + C_5 B_n \tilde{\epsilon}_n \left[\frac{\delta'(\log n)(\log\log n)}{n} + \sqrt{\frac{\delta'}{n}}\right]\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&e^{-\delta'} + P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big) \\
&\qquad \geq P\left(Q_n(\tilde{f}_n) - Q_n(f_0) \geq C_2\,\tilde{\epsilon}_n^2 + 2\mu_n + C_5 B_n \tilde{\epsilon}_n \left[\frac{\delta'(\log n)(\log\log n)}{n} + \sqrt{\frac{\delta'}{n}}\right]\right). \qquad \text{(A.18)}
\end{aligned}
$$

A.3.4 Independent Blocks

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

$$
\begin{aligned}
T_{1,j} &\coloneqq \Big\{t \in \mathbb{N} :\ 2(j-1)a + 1 \leq t \leq (2j-1)a\Big\}, \quad j = 1, \ldots, b; \\
T_{2,j} &\coloneqq \Big\{t \in \mathbb{N} :\ (2j-1)a + 1 \leq t \leq 2ja\Big\}, \quad j = 1, \ldots, b; \\
T_R &\coloneqq \Big\{t \in \mathbb{N} :\ 2ba + 1 \leq t \leq n\Big\}.
\end{aligned}
$$

As described in Appendix A.4, we use Berbee's Lemma to redefine $\{\boldsymbol{Z}_t\}_{t=1}^n$ on a richer probability space[3] where there exists a random sequence $\{\overline{\boldsymbol{Z}}_t\}_{t=1}^n$ with the following two properties: let $T, T' \in \{T_{1,1}, T_{2,1}, T_{1,2}, \ldots, T_{2,b}, T_R\}$, then (i) the block $\{\overline{\boldsymbol{Z}}_t\}_{t\in T}$ is independent from the blocks $\{\overline{\boldsymbol{Z}}_t\}_{t\in T'}$, $\{\boldsymbol{Z}_t\}_{t\in T'}$ for any $T' \neq T$; and (ii) $\{\overline{\boldsymbol{Z}}_t\}_{t\in T}$ has the same distribution as $\{\boldsymbol{Z}_t\}_{t\in T}$, i.e. $P_{\{\overline{\boldsymbol{Z}}_t\}_{t\in T}} =$

[3] 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 A.4 for details.

$P_{\{\boldsymbol{Z}_t\}_{t\in T}}$. By stationarity, all blocks of length a are identically distributed so the sequence of blocks $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,1}}, \{\overline{\boldsymbol{Z}}_t\}_{t\in T_{2,1}}, \{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,2}}, \ldots, \{\overline{\boldsymbol{Z}}_t\}_{t\in T_{2,b}}$ is i.i.d.,[4] and we have

$$\begin{aligned} P_{\{\overline{\boldsymbol{Z}}_t:\, t\in\cup_{j=1}^{b} T_{1,j}\}} &= P_{\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,1}}} \times P_{\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,2}}} \times \cdots \times P_{\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,b}}} \\ &= P_{\{\boldsymbol{Z}_t\}_{t\in T_{1,1}}} \times P_{\{\boldsymbol{Z}_t\}_{t\in T_{1,2}}} \times \cdots \times P_{\{\boldsymbol{Z}_t\}_{t\in T_{1,b}}}. \end{aligned} \tag{A.19}$$

Next, the usual β-mixing coefficient (e.g. Dehling and Philipp, 2002, Definition 3.1, p.19) can be equivalently written as (see Eberlein, 1984)

$$\beta(m) = \sup_{A\times B\in\sigma(\{\boldsymbol{Z}_t\}_{t=1}^{k})\otimes\sigma(\{\boldsymbol{Z}_t\}_{t=k+m+1}^{\infty})} |P(A\times B) - P(A)P(B)|.$$

Hence, for $j \in \{1, ..., b-1\}$

$$\begin{aligned} \beta(a) \geq \sup\Big\{ &\Big| P_{\{\boldsymbol{Z}_t:\, t\in\cup_{j=1}^{b} T_{1,j}\}}(A\times B) - P_{\{\boldsymbol{Z}_t:\, t\in\cup_{j=1}^{k} T_{1,j}\}}(A)\; P_{\{\boldsymbol{Z}_t:\, t\in\cup_{j=k+1}^{b} T_{1,j}\}}(B)\Big| : \\ &A\times B \in \sigma\Big(\Big\{\boldsymbol{Z}_t : t\in\cup_{j=1}^{k} T_{1,j}\Big\}\Big)\otimes\sigma\Big(\Big\{\boldsymbol{Z}_t : t\in\cup_{j=1}^{k} T_{1,j}\Big\}\Big)\Big\}. \end{aligned} \tag{A.20}$$

By (A.20) the conditions for Eberlein (1984) Lemma 2 are satisfied. So we apply this result, and use (A.19), to obtain, for any measurable set E,

$$\Big|P\Big(\Big\{\boldsymbol{Z}_t : t\in\cup_{j=1}^{b} T_{1,j}\Big\}\in E\Big) - P\Big(\Big\{\overline{\boldsymbol{Z}}_t : t\in\cup_{j=1}^{b} T_{1,j}\Big\}\in E\Big)\Big| \;\leq\; (b-1)\beta(a).$$

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

$$P\Big(\Big\{\boldsymbol{Z}_t : t\in\cup_{j=1}^{b} T_{1,j}\Big\}\in E\Big) \leq P\Big(\Big\{\overline{\boldsymbol{Z}}_t : t\in\cup_{j=1}^{b} T_{1,j}\Big\}\in E\Big) + \frac{n\,\beta(a)}{2a}. \tag{A.21}$$

A.3.5 Localization Analysis

We begin with some definitions that will be used throughout the rest of this proof. Define the following norms,

$$\begin{aligned} \|f\|_{\overline{T}_{1,j}} &:= \left(\frac{1}{a}\sum_{t\in T_{1,j}} \left|f(\overline{\boldsymbol{Z}}_t)\right|^2\right)^{1/2}, \quad \text{for } j\in\{1,\ldots,b\}, \\ \|f\|_{\overline{T}_1} &:= \left(\frac{1}{b}\sum_{j=1}^{b}\|f\|^2_{\overline{T}_{1,j}}\right)^{1/2} = \left(\frac{1}{ba}\sum_{j=1}^{b}\sum_{t\in T_{1,j}} \left|f(\overline{\boldsymbol{Z}}_t)\right|^2\right)^{1/2}. \end{aligned} \tag{A.22}$$

[4] All blocks except $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_R}$ are of length a, and therefore i.i.d. However, $\{\overline{\boldsymbol{Z}}_t\}_{t=1}^{n}$ is not an independent sequence since elements within a single block, $\{\overline{\boldsymbol{Z}}_t\}_{t\in T}$, may be correlated. For more details see Appendix A.4.

The following definition for the Rademacher complexity of a function class is from Bartlett et al. (2005).

Definition A.3.1. (Rademacher Complexity) For $n \in \mathbb{N}$, let $\{\boldsymbol{W}_t\}_{t=1}^n$ be random variables on $(\Omega, \mathcal{A}, P)$ taking values in $\mathbb{R}^{d_W}$ for $d_W \in \mathbb{N}$. The Rademacher random variables, $\{\xi_t\}_{t=1}^n$, are i.i.d. random variables that are independent of $\{\boldsymbol{W}_t\}_{t=1}^n$, and $\xi_t \in \{-1, 1\}$ where $P(\xi_t = 1) = P(\xi_t = -1) = 1/2$. For a pointwise-separable class of functions $\mathcal{S}$ with elements $s : \mathbb{R}^{d_{\boldsymbol{W}}} \to \mathbb{R}$ that are measurable-$\mathcal{B}(\mathbb{R}^{d_{\boldsymbol{W}}})/\mathcal{B}(\mathbb{R})$, define

$$\mathfrak{R}_n\mathcal{S} \coloneqq \sup_{s\in\mathcal{S}} \frac{1}{n}\sum_{t=1}^{n} \xi_t\, s(\boldsymbol{W}_t).$$

The Rademacher complexity is $\mathbb{E}[\mathfrak{R}_n\mathcal{S}]$, and the empirical Rademacher complexity is $\mathbb{E}_\xi[\mathfrak{R}_n\mathcal{S}] \coloneqq \mathbb{E}\big[\mathfrak{R}_n\mathcal{S} \,\big|\, \{\boldsymbol{W}_t\}_{t=1}^n\big]$.[5]

We will write $\mathfrak{R}_{ab}\mathcal{S} = \sup_{s\in\mathcal{S}} \sum_{j=1}^{b} \sum_{t\in T_{1,j}} \xi_t\, s(\boldsymbol{W}_t)$, since the double sum is of length $a \cdot b$.

A.3.5.1 Step I: Quadratic Process Bound

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

$$f \in \left\{ f \in \mathcal{F}_n : \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq r \right\},$$

throughout this section. This step will show that this implies $\|f - f_0\|_{\overline{T}_1} \leq 2r$ with probability greater than $1 - e^{-\delta'}$ for $\delta' > 0$ when r satisfies certain conditions.

First note that, $\mathbb{E}\Big[\|f - f_0\|_{\overline{T}_1}^2 - \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}^2\Big] = 0$, since by stationarity and (A.19)

$$\|f - f_0\|_{\overline{T}_1}^2 - \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}^2 = \frac{1}{b}\sum_{t=1}^{b} \left\{ \|f - f_0\|_{\overline{T}_{1,j}}^2 - \mathbb{E}\big[\|f - f_0\|_{\overline{T}_{1,j}}^2\big] \right\}. \tag{A.23}$$

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

$$\|f - f_0\|_{\overline{T}_{1,j}}^2 \leq (B_n + 1)\|f - f_0\|_{\overline{T}_{1,j}} \leq 2B_n\|f - f_0\|_{\overline{T}_{1,j}} \leq 4B_n^2,$$

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

by (b.4), with Assumptions 2.2.2(iii), 2.2.1; and

$$\mathrm{Var}\Big[\|f-f_0\|^2_{\overline{T}_{1,j}}\Big] \le \mathbb{E}\Big[\|f-f_0\|^4_{\overline{T}_{1,j}}\Big] \le (2B_n)^2\|f-f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \le 4B_n^2 r^2,$$

since

$$\mathbb{E}_{P_{\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,j}}}}\big[f-\tilde{f}_n\big] = \mathbb{E}_{P_{\{\boldsymbol{Z}_t\}_{t\in T_{1,j}}}}\big[f-\tilde{f}_n\big] = \mathbb{E}_{P_{\boldsymbol{Z}}}\big[f-\tilde{f}_n\big].$$

Recall from the previous section that $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,1}}, \{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,2}}, \ldots, \{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,b}}$ is an i.i.d. sequence, and consequently so is $\big\{\|f-f_0\|_{\overline{T}_{1,j}}\big\}_{j=1}^b$. Then, by the symmetrization inequality Bartlett et al. (2005, Theorem 2.1) (with $\alpha = 1/2$ therein) for any $\delta' > 0$

$$\begin{aligned} e^{-\delta'} \ge P\Bigg(&\sup_{\left\{f\in\mathcal{F}_n:\|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r\right\}} \frac{1}{b}\sum_{t=1}^b \Big\{\|f-f_0\|^2_{\overline{T}_{1,j}} - \mathbb{E}\big[\|f-f_0\|^2_{\overline{T}_{1,j}}\big]\Big\} \\ &\ge 3\mathbb{E}\Big[\mathfrak{R}_b\Big\{\|f-f_0\|^2_{\overline{T}_{1,j}} : f\in\mathcal{F}_n, \|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r\Big\}\Big] + 2B_n r\sqrt{\frac{2\delta'}{b}} + \frac{28B_n^2\delta'}{3b}\Bigg) \end{aligned} \tag{A.24}$$

By $a^2-b^2 = (a+b)(a-b)$ and the reverse triangle inequality

$$\begin{aligned} (f-f_0)^2 - (f'-f_0)^2 &= \Big((f-f_0)+(f'-f_0)\Big)\cdot\Big((f-f_0)-(f'-f_0)\Big) \\ &\le 4B_n\Big|(f-f_0)-(f'-f_0)\Big|. \end{aligned}$$

Then, by Lemma A.3.3

$$\begin{aligned} &\mathbb{E}\Big[\mathfrak{R}_b\Big\{\|f-f_0\|^2_{\overline{T}_{1,j}} : f\in\mathcal{F}_n, \|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r\Big\}\Big] \\ &= \frac{1}{b}\mathbb{E}\left[\sup\left\{\sum_{j=1}^b \xi_j\cdot\left(\frac{1}{a}\sum_{t\in T_{1,j}}\big(f(\boldsymbol{Z}_t)-f_0(\boldsymbol{Z}_t)\big)^2\right) : f\in\mathcal{F}_n, \|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r\right\}\right] \\ &\le \frac{4B_n\sqrt{2/a}}{b}\mathbb{E}\left[\sup\left\{\sum_{j=1}^b\sum_{t\in T_{1,j}}\xi_t\big(f(\boldsymbol{Z}_t)-f_0(\boldsymbol{Z}_t)\big) : f\in\mathcal{F}_n, \|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r\right\}\right] \\ &= 4B_n\sqrt{2a}\,\mathbb{E}\Big[\mathfrak{R}_{ab}\Big\{f-f_0 : f\in\mathcal{F}_n, \|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r\Big\}\Big] \end{aligned}$$

Applying this Rademacher complexity bound, (A.23), and $b := \lfloor n/(2a)\rfloor > n/(4a)$, to (A.24) we obtain

$$\begin{aligned} e^{-\delta'} \ge P\Bigg(&\sup_{\{f\in\mathcal{F}_n:\|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r\}} \Big\{\|f-f_0\|^2_{\overline{T}_1} - \|f-f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\Big\} \\ &\ge 12B_n\sqrt{2a}\,\mathbb{E}\Big[\mathfrak{R}_{ab}\Big\{f-f_0 : f\in\mathcal{F}_n, \|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r\Big\}\Big] + r\sqrt{\frac{16\,aB_n^2\delta'}{n}} + \frac{112\,aB_n^2\delta'}{3n}\Bigg). \end{aligned} \tag{A.25}$$

Next suppose

$$r^2 \geq 12B_n\sqrt{2a}\,\mathbb{E}\left[\mathfrak{R}_{ab}\left\{f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq r\right\}\right], \tag{A.26}$$

and

$$r^2 \geq \frac{38\,aB_n^2\delta'}{n}. \tag{A.27}$$

Note that if (A.27) holds then $2r^2 \geq r\sqrt{\frac{16\,aB_n^2\delta'}{n}} + \frac{112\,aB_n^2\delta'}{(3n)}$. Therefore, (A.25) implies that for all r such that (A.26) and (A.27) hold

$$\begin{aligned} e^{-\delta'} &\geq P\left(\sup_{\{f\in\mathcal{F}_n:\|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\leq r\}} \left\{\|f - f_0\|_{\overline{T}_1}^2 - \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}^2\right\} \geq 3r^2\right) \\ &\geq P\left(\sup_{\{f\in\mathcal{F}_n:\|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\leq r\}} \|f - f_0\|_{\overline{T}_1}^2 \geq 4r^2\right) \\ &= P\left(\sup_{\{f\in\mathcal{F}_n:\|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\leq r\}} \|f - f_0\|_{\overline{T}_1} \geq 2r\right). \end{aligned} \tag{A.28}$$

A.3.5.2 Step II: Radius One Step Tightening

Given some initial radius $r_0 \geq \|\hat{f}_n - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}$ and $\delta' \geq 1$ such that (A.26) and (A.27) hold, this step will show that we may use r_0 to obtain a tighter bound on $\|\hat{f}_n - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}$ with high probability, whenever the radius r_0 is sufficiently loose. The notion of 'sufficiently loose' will be specified at the end of this step.

For $m \in \{1, 2\}$, and $j \in \{1, ..., b\}$, define

$$G_{j,f}^{(m)} \coloneqq \frac{1}{a}\sum_{t\in T_{m,j}} g_f(\boldsymbol{Z}_t), \quad \text{and} \quad \overline{G}_{j,f}^{(m)} \coloneqq \frac{1}{a}\sum_{t\in T_{m,j}} g_f(\overline{\boldsymbol{Z}}_t).$$

With this, the empirical process term from (A.14) can be written as

$$\begin{aligned} &\frac{1}{n}\sum_{t=1}^{n}\left\{\mathbb{E}\big[g_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g_{\hat{f}_n}(\boldsymbol{Z}_t)\right\} \\ &\quad = \left(\frac{ab}{n}\right)\frac{1}{b}\sum_{j=1}^{b}\left\{\mathbb{E}\big[G_{j,\hat{f}_n}^{(1)}\big] - G_{j,\hat{f}_n}^{(1)} + \mathbb{E}\big[G_{j,\hat{f}_n}^{(2)}\big] - G_{j,\hat{f}_n}^{(2)}\right\} + \frac{1}{n}\sum_{t\in T_R}\left\{\mathbb{E}\big[g_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g_{\hat{f}_n}(\boldsymbol{Z}_t)\right\}. \end{aligned}$$

Then, by stationarity, for $A_1, A_2 > 0$ to be specified later,

$$
\begin{aligned}
&P\left(\frac{1}{n}\sum_{t=1}^{n}\left\{\mathbb{E}\big[g_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g_{\hat{f}_n}(\boldsymbol{Z}_t)\right\} \geq 2A_1 + A_2\right) \\
&\leq 2P\left(\left(\frac{ab}{n}\right)\frac{1}{b}\sum_{j=1}^{b}\left\{\mathbb{E}\big[G^{(1)}_{j,\hat{f}_n}\big] - G^{(1)}_{j,\hat{f}_n}\right\} \geq A_1\right) + P\left(\frac{1}{n}\sum_{t\in T_R}\left\{\mathbb{E}\big[g_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g_{\hat{f}_n}(\boldsymbol{Z}_t)\right\} \geq A_2\right) \\
&\coloneqq 2P_1 + P_2.
\end{aligned}
\tag{A.29}
$$

Consider P_2. By (b.5)(i), (b.4), and $\|f_0\|_\infty \leq 1$,

$$
\|g_{\hat{f}_n}\|_\infty \leq C_4 B_n \|\hat{f}_n - f_0\|_\infty \leq C_4 B_n (B_n + 1) \leq 2C_4 B_n^2,
$$

hence $\left\|\mathbb{E}\big[g_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g_{\hat{f}_n}(\boldsymbol{Z}_t)\right\|_\infty \leq 4C_4 B_n^2$. Denote the cardinality of T_R as $(\#T_R) = n - 2ab$, and note $(\#T_R) < 2a$, since $b \coloneqq \lfloor n/(2a) \rfloor$ implies $b > n/(2a) - 1$. With this,

$$
\frac{1}{n}\sum_{t\in T_R}\left\{\mathbb{E}\big[g_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g_{\hat{f}_n}(\boldsymbol{Z}_t)\right\} \leq \frac{2a}{n}\left\|\mathbb{E}\big[g_{\hat{f}_n}(\boldsymbol{Z}_t)\big] - g_{\hat{f}_n}(\boldsymbol{Z}_t)\right\|_\infty \leq \frac{8C_4 B_n^2 a}{n}.
$$

Thus,

$$
P_2 = 0, \quad \text{for} \quad A_2 = \frac{9C_4 B_n^2 a}{n}. \tag{A.30}
$$

To bound P_1 we first apply (A.21) with

$$
E = \left\{\left(\frac{ab}{n}\right)\frac{1}{b}\sum_{j=1}^{b}\left\{\mathbb{E}\big[G^{(1)}_{j,\hat{f}_n}\big] - G^{(1)}_{j,\hat{f}_n}\right\} \geq A_1\right\},
$$

to obtain

$$
\begin{aligned}
P_1 &\leq P\left(\left(\frac{ab}{n}\right)\frac{1}{b}\sum_{j=1}^{b}\left\{\mathbb{E}\big[\overline{G}^{(1)}_{j,\hat{f}_n}\big] - \overline{G}^{(1)}_{j,\hat{f}_n}\right\} \geq A_1\right) + \frac{n\,\beta(a)}{2a} \\
&\leq P\left(\frac{1}{b}\sum_{j=1}^{b}\left\{\mathbb{E}\big[\overline{G}^{(1)}_{j,\hat{f}_n}\big] - \overline{G}^{(1)}_{j,\hat{f}_n}\right\} \geq 2A_1\right) + \frac{n\,\beta(a)}{2a},
\end{aligned}
\tag{A.31}
$$

since $b \leq n/(2a)$. We bound the first term on the right side with Bartlett et al. (2005, Theorem

2.1). For any $f \in \mathcal{F}_n$, recall $\|g_f\|_\infty \le 2C_4B_n^2$, and note that

$$\begin{aligned}
\mathrm{Var}\big[\,\overline{G}^{(1)}_{j,f}\big] &\le \mathbb{E}\left[\left(\frac{1}{a}\sum_{t\in T_{1,j}} g_f(\overline{\boldsymbol{Z}}_t)\right)^2\right] \le \mathbb{E}\left[\left(\frac{1}{a}\sum_{t\in T_{1,j}} C_4B_n\big|f(\overline{\boldsymbol{Z}}_t) - f_0(\overline{\boldsymbol{Z}}_t)\big|\right)^2\right] \\
&\le \mathbb{E}\left[\frac{1}{a}\sum_{t\in T_{1,j}} (C_4B_n)^2\big|f(\overline{\boldsymbol{Z}}_t) - f_0(\overline{\boldsymbol{Z}}_t)\big|^2\right] = \mathbb{E}\left[\frac{1}{a}\sum_{t\in T_{1,j}} (C_4B_n)^2\big|f(\boldsymbol{Z}_t) - f_0(\boldsymbol{Z}_t)\big|^2\right] \\
&= (C_4B_n)^2\,\|f - f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \le (C_4B_n)^2\, r_0^2
\end{aligned}$$

where the third inequality uses

$$\sum_{j=1}^J |x_j| = \left(\Big[\sum_{j=1}^J 1|x_j|\Big]^2\right)^{1/2} \le \left(\Big[\sum_{j=1}^J 1^2\Big]\Big[\sum_{j=1}^J |x_j|^2\Big]\right)^{1/2} = \left(J\Big[\sum_{j=1}^J |x_j|^2\Big]\right)^{1/2}, \tag{A.32}$$

by the Cauchy-Schwarz inequality, the first equality uses $P_{\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,j}}} = P_{\{\boldsymbol{Z}_t\}_{t\in T_{1,j}}}$ for any $j \in \{1,\dots,b\}$, and the second equality uses stationarity. With this, since $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,1}}$, $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,2}}$, $\dots, \{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,b}}$ is an i.i.d. sequence, we can apply Bartlett et al. (2005, Theorem 2.1) (with $\alpha = 1/2$ therein) to obtain,

$$\begin{aligned}
1 - e^{-\delta'} &\le P\left(\frac{1}{b}\sum_{j=1}^b \Big\{\mathbb{E}\big[\overline{G}^{(1)}_{j,\hat{f}_n}\big] - \overline{G}^{(1)}_{j,\hat{f}_n}\Big\} \le 6\mathbb{E}_\xi\Big[\mathfrak{R}_b\Big\{\overline{G}^{(1)}_{j,f} : f\in\mathcal{F}_n, \|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \le r_0\Big\}\Big]\right. \\
&\qquad \left. + C_4B_nr_0\sqrt{\frac{2\delta'}{b}} + \frac{64C_4B_n^2\delta'}{3b}\right) \\
&\le P\left(\frac{1}{b}\sum_{j=1}^b \Big\{\mathbb{E}\big[\overline{G}^{(1)}_{j,\hat{f}_n}\big] - \overline{G}^{(1)}_{j,\hat{f}_n}\Big\} \le 6\mathbb{E}_\xi\Big[\mathfrak{R}_b\Big\{\overline{G}^{(1)}_{j,f} : f\in\mathcal{F}_n, \|f-f_0\|_{\overline{T}_1} \le 2r_0\Big\}\Big]\right. \\
&\qquad \left. + C_4B_nr_0\sqrt{\frac{2\delta'}{b}} + \frac{64C_4B_n^2\delta'}{3b}\right) + P\left(\sup_{\{f\in\mathcal{F}_n:\|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r_0\}} \|f-f_0\|_{\overline{T}_1} > 2r_0\right).
\end{aligned}$$

Thus by (A.28), since r_0 satisfies (A.26) and (A.27),

$$\begin{aligned}
2e^{-\delta'} \ge P\left(\frac{1}{b}\sum_{j=1}^b \Big\{\mathbb{E}\big[\overline{G}^{(1)}_{j,\hat{f}_n}\big] - \overline{G}^{(1)}_{j,\hat{f}_n}\Big\} \ge 6\mathbb{E}_\xi\Big[\mathfrak{R}_b\Big\{\overline{G}^{(1)}_{j,f} : f\in\mathcal{F}_n, \|f-f_0\|_{\overline{T}_1} \le 2r_0\Big\}\Big]\right. \\
\left. + C_4B_nr_0\sqrt{\frac{2\delta'}{b}} + \frac{64C_4B_n^2\delta'}{3b}\right).
\end{aligned} \tag{A.33}$$

Now we address the Rademacher complexity term above. Note that for any $f, f' \in \mathcal{F}_n$, by (b.5)

and $M_n = 4B_n$

$$\begin{aligned}\left|g_f(\overline{\boldsymbol{Z}}_t) - g_{f'}(\overline{\boldsymbol{Z}}_t)\right| &= \left|q(\overline{\boldsymbol{Z}}_t, f) - q(\overline{\boldsymbol{Z}}_t, f')\right| \leq C_4 B_n \left|f(\overline{\boldsymbol{Z}}_t) - f'(\overline{\boldsymbol{Z}}_t)\right| \\ &= C_4 B_n \Big|\big(f(\overline{\boldsymbol{Z}}_t) - f_0(\overline{\boldsymbol{Z}}_t)\big) - \big(f'(\overline{\boldsymbol{Z}}_t) - f_0(\overline{\boldsymbol{Z}}_t)\big)\Big|,\end{aligned}$$

so we can apply Lemma A.3.3 to obtain

$$\begin{aligned}&6\mathbb{E}_\xi\left[\mathfrak{R}_b\left\{\overline{G}^{(1)}_{j,f} : f \in \mathcal{F}_n, \|f - f_0\|_{\overline{T}_1} \leq 2r_0\right\}\right] \\ &\leq \frac{6C_4B_n\sqrt{2}}{b\sqrt{a}}\mathbb{E}_\xi\left[\sup\left\{\sum_{j=1}^{b}\sum_{t\in T_{1,j}} \xi_{j,t}\left(f(\overline{\boldsymbol{Z}}_t) - f_0(\overline{\boldsymbol{Z}}_t)\right) : f \in \mathcal{F}_n, \|f - f_0\|_{\overline{T}_1} \leq 2r_0\right\}\right] \\ &= 6C_4B_n\sqrt{2a}\,\mathbb{E}_\xi\left[\mathfrak{R}_{ab}\left\{f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\overline{T}_1} \leq 2r_0\right\}\right].\end{aligned} \tag{A.34}$$

Let $D_n \coloneqq \min\{2B_n, 2r_0\}$, so $\left\{f \in \mathcal{F}_n : \|f - f_0\|_{\overline{T}_1} \leq 2r_0\right\} = \left\{f \in \mathcal{F}_n : \|f - f_0\|_{\overline{T}_1} \leq D_n\right\}$, since $\sup_{f\in\mathcal{F}_n}\|f - f_0\|_\infty \leq B_n + 1 \leq 2B_n$. With this, and Lemma A.3.5

$$\begin{aligned}&\mathbb{E}_\xi\left[\mathfrak{R}_{ab}\left\{f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\overline{T}_1} \leq 2r_0\right\}\right] \\ &\qquad = \mathbb{E}_\xi\left[\mathfrak{R}_{ab}\left\{f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\overline{T}_1} \leq D_n\right\}\right] \\ &\qquad \leq \inf_{0<\alpha<D_n}\left\{4\alpha + 8\sqrt{\frac{2}{ab}}\int_\alpha^{D_n}\sqrt{\log \mathrm{N}(\upsilon, \mathcal{F}_n, \|\cdot\|_{\overline{T}_1})}\,d\upsilon\right\} \\ &\qquad \leq \inf_{0<\alpha<D_n}\left\{4\alpha + 8\sqrt{\frac{8}{n}}\int_\alpha^{D_n}\sqrt{\log \mathrm{N}_2^{(\infty)}(\upsilon, \mathcal{F}_n, n/2)}\,d\upsilon\right\}\end{aligned}$$

since $\|\cdot\|_{\overline{T}_1}$ is a sum of ab terms and $\frac{n}{4a} < b \leq \frac{n}{2a}$. By assumption $n/2 > \mathrm{Pdim}(\mathcal{F}_n)$ so $D_n < e^2B_nn/\mathrm{Pdim}(\mathcal{F}_n)$. Then, $\mathrm{N}_2^{(\infty)}(\upsilon, \mathcal{F}_n, n/2) \leq \left(\frac{eB_nn}{\upsilon\cdot\mathrm{Pdim}(\mathcal{F}_n)}\right)^{\mathrm{Pdim}(\mathcal{F}_n)}$ by Lemma A.3.4, and $\log\left(\frac{e^2B_nn}{\upsilon\mathrm{Pdim}(\mathcal{F}_n)}\right) > 0$ for $0 < \upsilon \leq D_n$. With this, and the Cauchy-Schwarz inequality,[6] the previous

[6] For any $f : \mathbb{R} \to \mathbb{R}$ and $a, b \in \mathbb{R}$ such that $a < b$ and $f(x) \geq 0$ for all $x \in [a, b]$, by the Cauchy-Schwarz inequality

$$\textstyle\int_a^b \sqrt{f(x)}\,dx \leq \left(\int_a^b f(x)\,dx\right)^{\frac{1}{2}}\left(\int_a^b 1\,dx\right)^{\frac{1}{2}} = \sqrt{b-a}\left(\int_a^b f(x)\,dx\right)^{\frac{1}{2}}.$$

display implies,

$$
\begin{aligned}
&\mathbb{E}_\xi\Big[\mathfrak{R}_{ab}\Big\{f-f_0 : f\in\mathcal{F}_n, \|f-f_0\|_{\overline{T}_1}\le 2r_0\Big\}\Big] \\
&\quad\le \inf_{0<\alpha<D_n}\left\{4\alpha + 8\sqrt{\frac{8}{n}}\int_\alpha^{D_n}\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log\left(\frac{eB_n n}{\upsilon \mathrm{Pdim}(\mathcal{F}_n)}\right)}\,d\upsilon\right\} \\
&\quad\le \inf_{0<\alpha<D_n}\left\{4\alpha + 8\sqrt{\frac{D_n\mathrm{Pdim}(\mathcal{F}_n)}{n}\int_\alpha^{D_n}\log\left(\frac{eB_n n}{\upsilon \mathrm{Pdim}(\mathcal{F}_n)}\right)\,d\upsilon}\right\} \\
&\quad= \inf_{0<\alpha<D_n}\left\{4\alpha + 8\sqrt{\frac{D_n\,\mathrm{Pdim}(\mathcal{F}_n)}{n}\left[\upsilon\cdot\log\left(\frac{e^2B_n n}{\upsilon \mathrm{Pdim}(\mathcal{F}_n)}\right)\right]_{\upsilon=\alpha}^{D_n}}\right\} \\
&\quad\le 4D_n\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_n)}{n}} + 8\sqrt{\frac{D_n\,\mathrm{Pdim}(\mathcal{F}_n)}{n}\left[\upsilon\cdot\log\left(\frac{e^2B_n n}{\upsilon \mathrm{Pdim}(\mathcal{F}_n)}\right)\right]_{\upsilon=D_n\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_n)}{n}}}^{D_n}} \\
&\quad\le 4D_n\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_n)}{n}} + 8D_n\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_n)}{n}\cdot\log\left(\frac{e^2B_n n}{D_n\mathrm{Pdim}(\mathcal{F}_n)}\right)}
\end{aligned}
$$

where the third inequality chooses $\alpha = D_n\sqrt{\mathrm{Pdim}(\mathcal{F}_n)/n} \in (0, D_n)$. Note that $D_n > e^2B_n/n$. To see this, recall $D_n := \min\{2B_n, 2r_0\}$, then, by assumption, $n \ge 4 > e^2/2$ which implies $e^2B_n/n < 2B_n$, and (A.27) implies $2r_0 \ge 2\sqrt{38\,aB_n^2\delta'/n} > e^2B_n/n$, since $a\in\mathbb{N}$ and $\delta'\ge 1$. Hence, $\log\left(\frac{e^2B_n n}{D_n\mathrm{Pdim}(\mathcal{F}_n)}\right) < \log(n)$, since $\mathrm{Pdim}(\mathcal{F}_n)\ge 1$ by definition. With this, the previous display becomes

$$
\mathbb{E}_\xi\Big[\mathfrak{R}_{ab}\Big\{f-f_0 : f\in\mathcal{F}_n, \|f-f_0\|_{\overline{T}_1}\le 2r_0\Big\}\Big] \le 24r_0\sqrt{\frac{2\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)}. \tag{A.35}
$$

Combining (A.34) and (A.35),

$$
\mathbb{E}_\xi\left[\mathfrak{R}_b\left\{\overline{G}_{j,f}^{(1)} : f\in\mathcal{F}_n, \|f-f_0\|_{\overline{T}_1}\le 2r_0\right\}\right] \le r_0\,(6\cdot 24\cdot 2)C_4B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)}.
$$

With this and (A.33), then using $b > \frac{n}{4a}$,

$$
\begin{aligned}
2e^{-\delta'} &\ge P\left(\frac{1}{b}\sum_{j=1}^b\left\{\mathbb{E}[\overline{G}_{j,\hat{f}_n}^{(1)}] - \overline{G}_{j,\hat{f}_n}^{(1)}\right\} \ge r_0\,288C_4B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)} + r_0\,C_4B_n\sqrt{\frac{2\delta'}{b}} + \frac{64C_4B_n^2\delta'}{3b}\right) \\
&\ge P\left(\frac{1}{b}\sum_{j=1}^b\left\{\mathbb{E}[\overline{G}_{j,\hat{f}_n}^{(1)}] - \overline{G}_{j,\hat{f}_n}^{(1)}\right\} \ge r_0\,288C_4B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)} + r_0\,C_4B_n\sqrt{\frac{8a\delta'}{n}} + \frac{256C_4B_n^2a\delta'}{3n}\right).
\end{aligned}
$$

Applying this to (A.31)

$$P_1 \;\le\; P\left(\frac{1}{b}\sum_{j=1}^{b}\left\{\mathbb{E}\big[\overline{G}^{(1)}_{j,\hat{f}_n}\big]-\overline{G}^{(1)}_{j,\hat{f}_n}\right\}\ge 2A_1\right)+\frac{n\,\beta(a)}{2a} \;\le\; 2e^{-\delta'}+\frac{n\,\beta(a)}{2a},$$
$$\text{for}\quad 2A_1 = r_0\,288C_4B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)}+r_0\,C_4B_n\sqrt{\frac{8a\delta'}{n}}+\frac{256C_4B_n^2a\delta'}{3n}. \tag{A.36}$$

Now, we can update (A.29) with the bounds on P_2 and P_1 from (A.30) and (A.36),

$$\begin{aligned}
&P\Bigg(\frac{1}{n}\sum_{t=1}^{n}\left\{\mathbb{E}\big[g_{\hat{f}_n}(\boldsymbol{Z}_t)\big]-g_{\hat{f}_n}(\boldsymbol{Z}_t)\right\}\ge \\
&\qquad r_0\,288C_4B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)}+r_0\,C_4B_n\sqrt{\frac{8a\delta'}{n}}+\frac{95C_4B_n^2a\delta'}{n}\Bigg) \\
&\le 2P_1+P_2 \\
&\le 4e^{-\delta'}+\frac{n\,\beta(a)}{a}
\end{aligned} \tag{A.37}$$

since $\delta'\ge 1$ implies $\frac{256C_4B_n^2a\delta'}{3n}+\frac{9C_4B_n^2a}{n}<\frac{95C_4B_n^2a\delta'}{n}$. Returning to the main decomposition (A.14), and applying (A.18), (A.17), and (A.37) yields

$$\begin{aligned}
&5e^{-\delta'}+\frac{n\,\beta(a)}{a}+2P\Big(\max_{t\in\{1,\dots,n\}}m_n(\boldsymbol{Z}_t)>C_4B_n\Big) \\
&\ge P\Bigg(C_1\|\hat{f}_n-f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\ge r_0\cdot\left(\frac{288C_4B_n\sqrt{a}}{\sqrt{n}}\right)\left[\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log(n)}+\sqrt{\delta'}\right]+\frac{95C_4B_n^2a\delta'}{n} \\
&\qquad +C_2\,\tilde{\epsilon}_n^2+C_5B_n\tilde{\epsilon}_n\left[\frac{\delta'(\log n)(\log\log n)}{n}+\sqrt{\frac{\delta'}{n}}\right]+5\mu_n+\theta_n\Bigg).
\end{aligned} \tag{A.38}$$

Therefore, if $r_0\ge\|\hat{f}_n-f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})}$ is given, and r_0 is sufficiently larger than

$$\max\left\{\left(\frac{B_n\sqrt{a}}{\sqrt{n}}\right)\left[\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log(n)}+\sqrt{\delta}\right],\ \frac{B_n^2a\delta'}{n},\ \tilde{\epsilon}_n^2,\ \frac{B_n\tilde{\epsilon}_n}{n},\ \mu_n\right\}$$

then (A.38) implies that there exists $r_1<r_0$ such that $\|\hat{f}_n-f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\le r_1$ with probability greater than $1-5e^{-\delta'}-n\,\beta(a)/a-2P\big(\max_{t\in\{1,\dots,n\}}m_n(\boldsymbol{Z}_t)>C_4B_n\big)$. This can be done repeatedly as long as the new bound satisfies the conditions on r_0 given at the beginning of this step.

A.3.5.3 Step III: Radius Tightening Lower Bound

This step obtains a critical radius, $\overline{r}$, that is a reasonably tight lower-bound for radii such that the tightening of the last step can be applied. Let $x \vee y = \max\{x, y\}$. Define

$$\overline{r} \coloneqq \left(\frac{1}{2}\sqrt{\frac{38\, a B_n^2 \log(n)}{n}} \;\vee\; \inf\left\{ r > 0 : \forall s \geq r,\, s^2 > 12 B_n \sqrt{2a}\, \mathbb{E}\big[\mathfrak{R}_{ab}\big\{ f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2} \leq s \big\}\big] \right\} \right).$$

The definition of $\overline{r}$ implies

$$2\overline{r} \geq \sqrt{\frac{38\, a B_n^2 \log(n)}{n}},$$

so $2\overline{r}$ satisfies (A.27) for $\delta' = \log(n)$, and by construction $2\overline{r}$ satisfies (A.26). Hence, for the event

$$E \coloneqq \left\{ \sup_{\{f \in \mathcal{F}_n : \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq 2\overline{r}\}} \|f - f_0\|_{\overline{T}_1} \leq 4\overline{r} \right\} \subseteq \Omega,$$

we can apply (A.28) to obtain $P(E) \geq 1 - 1/n$.

Now, consider the case where

$$\frac{1}{2}\sqrt{\frac{38\, a B_n^2 \log(n)}{n}} \leq \inf\left\{ r > 0 : \forall s \geq r,\, s^2 > 12 B_n \sqrt{2a}\, \mathbb{E}\big[\mathfrak{R}_{ab}\big\{ f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2} \leq s \big\}\big] \right\},$$

or equivalently

$$\overline{r} = \inf\left\{ r > 0 : \forall s \geq r,\; s^2 > 12 B_n \sqrt{2a}\, \mathbb{E}\left[\mathfrak{R}_{ab}\left\{ f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq s \right\}\right] \right\}.$$

In this case, note that[7]

$$\overline{r}^2 \leq 12 B_n \sqrt{2a}\, \mathbb{E}\left[\mathfrak{R}_{ab}\left\{ f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq \overline{r} \right\}\right].$$

In addition, recall $\sup_{f \in \mathcal{F}_n} \|f - f_0\|_\infty \leq 2B_n$, which implies $\mathfrak{R}_{ab}\{f - f_0 : f \in \mathcal{F}_n\} \leq 2B_n$. Then,

[7] To see this, suppose $\overline{r}^2 > 12 B_n \sqrt{2a}\, \mathbb{E}\left[\mathfrak{R}_{ab}\left\{ f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq \overline{r} \right\}\right]$. Then there exists $\upsilon > 0$ such that

$$(\overline{r} - \upsilon)^2 > 12 B_n \sqrt{2a}\, \mathbb{E}\left[\mathfrak{R}_{ab}\left\{ f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq \overline{r} \right\}\right].$$

However, $\{f \in \mathcal{F}_n : \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq \overline{r} - \upsilon\} \subseteq \{f \in \mathcal{F}_n : \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq \overline{r}\}$, so we have $(\overline{r} - \upsilon)^2 > 12 B_n \sqrt{2a}\, \mathbb{E}\left[\mathfrak{R}_{ab}\left\{ f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq \overline{r} - \upsilon \right\}\right]$, which is a contradiction.

using these two facts with $P(E) \geq 1 - 1/n$, we obtain

$$\begin{aligned}
\overline{r}^2 &\leq 12B_n\sqrt{2a}\,\mathbb{E}\left[\mathfrak{R}_{ab}\left\{f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq \overline{r}\right\}\right] \\
&\leq 12B_n\sqrt{2a}\,\mathbb{E}\left[\mathfrak{R}_{ab}\left\{f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \leq 2\overline{r}\right\}\right] \\
&\leq 12B_n\sqrt{2a}\,\mathbb{E}\left[\mathbb{E}_{P^\xi}\left[\mathfrak{R}_{ab}\left\{f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\overline{T}_1} \leq 4\overline{r}\right\}\right]\mathbb{1}_E + 2B_n(1 - \mathbb{1}_E)\right] \\
&\leq 12B_n\sqrt{2a}\,\mathbb{E}\left[\mathbb{E}_{P^\xi}\left[\mathfrak{R}_{ab}\left\{f - f_0 : f \in \mathcal{F}_n, \|f - f_0\|_{\overline{T}_1} \leq 4\overline{r}\right\}\right]\mathbb{1}_E\right] + \frac{24B_n^2}{n}.
\end{aligned}$$

Clearly $2\overline{r} \geq \sqrt{\frac{38\,aB_n^2\log(n)}{n}}$ implies $\overline{r} > e^2/n$, and it follows from the same reasoning used at the beginning of Step II that $\overline{r} \leq e^2 B_n n/(2\mathrm{Pdim}(\mathcal{F}_n))$. Therefore, we can apply (A.35) to the above, to obtain

$$\begin{aligned}
\overline{r}^2 &\leq 2\overline{r} \cdot 384B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)} + \frac{24B_n^2}{n} \\
&\leq 2\overline{r} \cdot 408B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)},
\end{aligned}$$

since $\overline{r} \geq \sqrt{\frac{38\,aB_n^2\log(n)}{n}}$. Thus, $\overline{r} \leq 816B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)}$.

Now, returning to the general case, note that

$$\frac{1}{2}\sqrt{\frac{38\,aB_n^2\log(n)}{n}} \leq 816B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)},$$

since $\mathrm{Pdim}(\mathcal{F}_n) \geq 1$ by definition, and $B_n \geq 1$ by (b.4). Therefore, in either case, we have

$$\overline{r} \leq 816B_n\sqrt{\frac{a\mathrm{Pdim}(\mathcal{F}_n)}{n}\log(n)}. \tag{A.39}$$

A.3.5.4 Step IV: Localization

Now we obtain a final bound on $\|\hat{f}_n - f_0\|_{\mathcal{L}^2}$. Define

$$\begin{aligned}
r_* \;\coloneqq\; &\overline{r} + \frac{4}{C_1}\left(\frac{288C_4B_n\sqrt{a}}{\sqrt{n}}\right)\left[\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log(n)} + \sqrt{\delta'}\right] \\
&+ \sqrt{\frac{2}{C_1}}\left[\frac{95C_4B_n^2a\delta'}{n} + C_2\,\tilde{\epsilon}_n^2 + C_5B_n\tilde{\epsilon}_n\left(\frac{\delta'(\log n)(\log\log n)}{n} + \sqrt{\frac{\delta'}{n}}\right) + 5\mu_n + \theta_n\right]^{1/2},
\end{aligned} \tag{A.40}$$

Choose

$$\delta' = \delta + \log(5J), \qquad \text{where} \qquad J \coloneqq \left\lfloor \log_2\left(\frac{4B_n\sqrt{n}}{\sqrt{\log(n)}}\right)\right\rfloor.$$

To apply (A.38) the requirements of Step II must be met. Clearly, $r_* > \overline{r}$ so (A.26) is met. From (A.40), if $(4 \cdot 288)C_4/C_1 \geq \sqrt{38}$ then (A.27) is met, which is the case since $C_1 \leq 1$ and $C_4 \geq 1$ by assumption. Note that $J \geq 2$ so $\delta' > \delta + \log(10) > 1$ for any $\delta > 0$. Hence, if it is given that $\|\hat{f}_n - f_0\|_{\mathcal{L}^2} \leq 2^j r_*$ for some $j \in \{1, \ldots, J\}$, then by (A.38), with probability greater than $1 - 5e^{-\delta'} - n\,\beta(a)/a - 2P\big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\big)$,

$$
\begin{aligned}
&\|\hat{f}_n - f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{Z}})} \\
&\leq 2^j r_* \cdot \frac{1}{C_1}\left(\frac{288 C_4 B_n \sqrt{a}}{\sqrt{n}}\right)\left[\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log(n)} + \sqrt{\delta}\right] \\
&\quad + \frac{1}{C_1}\left[\frac{95 C_4 B_n^2 a\delta'}{n} + C_2\,\tilde{\epsilon}_n^2 + C_5 B_n \tilde{\epsilon}_n\left(\frac{\delta'(\log n)(\log\log n)}{n} + \sqrt{\frac{\delta'}{n}}\right) + 5\mu_n + \theta_n\right] \\
&\leq 2^j r_* \cdot \left(\frac{2^j r_*}{8}\right) + \frac{(2^j r_*)^2}{8} \;=\; \frac{(2^j r_*)^2}{4} \;=\; (2^{j-1} r_*)^2,
\end{aligned}
$$

where the second inequality follows because, for any $j \in \{1, \ldots, J\}$, (A.40) implies

$$
\frac{1}{C_1}\left(\frac{288 C_4 B_n \sqrt{a}}{\sqrt{n}}\right)\left[\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log(n)} + \sqrt{\delta}\right] \;\leq\; \frac{r_*}{4} \;\leq\; \frac{2^j r_*}{8}, \quad \text{and}
$$

$$
\frac{1}{C_1}\left[\frac{146 C_5 B_n^2 a\delta'}{3n} + C_2\,\tilde{\epsilon}_n^2 + C_5 B_n \tilde{\epsilon}_n\left(\frac{\delta'(\log n)(\log\log n)}{n} + \sqrt{\frac{\delta'}{n}}\right) + 5\mu_n + \theta_n\right] \;\leq\; \frac{r_*^2}{2} \;\leq\; \frac{(2^j r_*)^2}{8}.
$$

In other words, for this choice of r_*, if $P\big(\|\hat{f}_n - f_0\|^2_{\mathcal{L}^2} \leq 2^j r_*\big) > 0$, then

$$
P\Big(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} > 2^{j-1} r_* \;\Big|\; \|\hat{f}_n - f_0\|_{\mathcal{L}^2} \leq 2^j r_*\Big) \leq 5e^{-\delta'} + \frac{n\,\beta(a)}{a} + 2P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big),
$$

and consequently

$$
\begin{aligned}
&P\Big(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} > 2^{j-1} r_*\Big) \\
&\quad \leq P\Big(\big\{\|\hat{f}_n - f_0\|_{\mathcal{L}^2} > 2^{j-1} r_*\big\} \cap \big\{\|\hat{f}_n - f_0\|_{\mathcal{L}^2} \leq 2^j r_*\big\}\Big) + P\Big(\big\{\|\hat{f}_n - f_0\|_{\mathcal{L}^2} > 2^j r_*\big\}\Big) \\
&\quad \leq 5e^{-\delta'} + \frac{n\,\beta(a)}{a} + 2P\left(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\right) + P\Big(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} > 2^j r_*\Big). \qquad \text{(A.41)}
\end{aligned}
$$

Note that $r_* \geq \sqrt{\log(n)/n}$, and

$$
J = \left\lfloor \log_2\left(\frac{4B_n\sqrt{n}}{\sqrt{\log(n)}}\right)\right\rfloor = \left\lfloor \log_2\left(\frac{2B_n\sqrt{n}}{\sqrt{\log(n)}}\right) + 1\right\rfloor \geq \log_2\left(\frac{2B_n\sqrt{n}}{\sqrt{\log(n)}}\right),
$$

which implies

$$2^J r_* \geq \left(\frac{2B_n\sqrt{n}}{\sqrt{\log(n)}}\right) r_* \geq 2B_n \geq \sup_{f\in\mathcal{F}_n} \|f - f_0\|_\infty \geq \|\hat{f}_n - f_0\|_{\mathcal{L}^2},$$

by (b.4) and $\|f_0\|_\infty \leq 1$. Hence, $P\big(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} \leq 2^J r_*\big) = 1$, and with (A.41)

$$P\Big(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} > 2^{J-1} r_*\Big) \leq 5e^{-\delta'} + \frac{n\,\beta(a)}{a} + 2P\left(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\right).$$

Then, it follows via induction that $P\big(\|\hat{f}_n - f_0\|^2_{\mathcal{L}^2} \leq 2^j r_*\big) > 0$ for all j, and

$$\begin{aligned} P\Big(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} > r_*\Big) &\leq J \cdot \left(5e^{-\delta'} + \frac{n\,\beta(a)}{a} + 2P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big)\right) \\ &\leq e^{-\delta} + 2\log(n)\left[\frac{n\,\beta(a)}{a} + 2P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big)\right], \end{aligned} \tag{A.42}$$

where the first term used $5Je^{-\delta'} = e^{-\delta - \log(5J)} = e^{-\delta}$, and the second term follows since by assumption $n \geq 16B_n^2/\log(n)$, which implies

$$J = \left\lfloor \log_2\left(\frac{4B_n\sqrt{n}}{\sqrt{\log(n)}}\right)\right\rfloor \leq \log_2\left(\frac{4B_n\sqrt{n}}{\sqrt{\log(n)}}\right) \leq \log_2(n) \leq 2\log(n).$$

With (A.42), the proof will be complete by showing $r_* \leq C\,\epsilon_n(\delta, a)$, for a suitable constant $C > 0$, and ϵ_n defined as in the statement of Theorem 2.1.2. Note that $aB_n - \tilde{\epsilon}_n(\log n)(\log\log n) > 0$ by the assumption on a in the statement of the theorem, with this

$$\begin{aligned} \frac{B_n^2 a\delta'}{n} \geq B_n\tilde{\epsilon}_n\left[\frac{\delta'(\log n)(\log\log n)}{n} + \sqrt{\frac{\delta'}{n}}\right] \quad &\Longleftrightarrow \quad B_n a\sqrt{\delta'} \geq \tilde{\epsilon}_n(\log n)(\log\log n)\sqrt{\delta'} + \tilde{\epsilon}_n\sqrt{n} \\ &\Longleftrightarrow \quad \sqrt{\delta'} \geq \frac{\tilde{\epsilon}_n\sqrt{n}}{B_n a - \tilde{\epsilon}_n(\log n)(\log\log n)}, \end{aligned}$$

which holds by the assumption on δ in the statement of the theorem, since $\delta' > \delta$. Therefore,

$$95C_4\left(\frac{B_n^2 a\delta'}{n}\right) + C_5 B_n\tilde{\epsilon}_n\left[\frac{\delta'(\log n)(\log\log n)}{n} + \sqrt{\frac{\delta'}{n}}\right] \leq 96C_5\left(\frac{B_n^2 a\delta'}{n}\right),$$

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

$$\sqrt{\delta'} \leq \sqrt{\log\big(10\log(n)\big) + \delta} < \sqrt{9\log\log(n) + \delta} \leq 3\sqrt{\log\log(n) + \delta}$$

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

$$\begin{aligned}
r_* &= \overline{r} + \frac{4}{C_1}\left(\frac{288C_4B_n\sqrt{a}}{\sqrt{n}}\right)\left[\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log(n)} + \sqrt{\delta'}\right] \\
&\quad + \sqrt{\frac{2}{C_1}}\left[95C_4\left(\frac{B_n^2 a\delta'}{n}\right) + C_2\,\tilde{\epsilon}_n^2 + C_5 B_n\tilde{\epsilon}_n\left[\frac{\delta'(\log n)(\log\log n)}{n} + \sqrt{\frac{\delta'}{n}}\right] + 5\mu_n + \theta_n\right]^{1/2} \\
&\leq \overline{r} + \frac{4(288C_4)}{C_1}\left(B_n\sqrt{\frac{a}{n}}\right)\left[\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log(n)} + \sqrt{\delta'}\right] \\
&\quad + \sqrt{\frac{192C_5}{C_1}}\left(B_n\sqrt{\frac{a}{n}}\right)\sqrt{\delta'} + \sqrt{\frac{2(C_2 \vee 5)}{C_1}}\sqrt{\tilde{\epsilon}_n^2 + \mu_n + \theta_n} \\
&\leq C\left(B_n\sqrt{\frac{a}{n}}\left[\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log(n)} + \sqrt{\log\log(n) + \delta}\right] + \sqrt{\tilde{\epsilon}_n^2 + \mu_n + \theta_n}\right) \\
&=: C\,\epsilon_n(\delta, a), \qquad \text{(A.43)}
\end{aligned}$$

where

$$C = \left(816 + \frac{4(3)(288C_4)}{C_1} + \sqrt{\frac{192C_5}{C_1}}\right) \vee \sqrt{\frac{2(C_2 \vee 5)}{C_1}}.$$

Recall from Section A.3.3, $C_5 = 2C_4/\min\{C_6, 1\}$, and C_6 from Lemma A.3.2 only depends on C_β and C'_β. Therefore, C only depends on C_1, C_2, C_4, C_β, and C'_β. This proves the first result of Theorem 2.1.2.

A.3.6 Empirical Error Bound

Note that

$$\begin{aligned}
\|\hat{f}_n - f_0\|_{2,n}^2 = &\left(\frac{1}{n}\sum_{j=1}^{b}\sum_{t\in T_{1,j}}\left[\hat{f}_n(\boldsymbol{Z}_t) - f_0(\boldsymbol{Z}_t)\right]^2\right) + \left(\frac{1}{n}\sum_{j=1}^{b}\sum_{t\in T_{2,j}}\left[\hat{f}_n(\boldsymbol{Z}_t) - f_0(\boldsymbol{Z}_t)\right]^2\right) \\
&+ \left(\frac{1}{n}\sum_{t\in T_R}\left[\hat{f}_n(\boldsymbol{Z}_t) - f_0(\boldsymbol{Z}_t)\right]^2\right).
\end{aligned}$$

Then, by (A.43) and stationarity,

$$
\begin{aligned}
P\Big(\|\hat{f}_n - f_0\|_{2,n} \geq \sqrt{3}\, C\, \epsilon_n(\delta, a)\Big) &\leq P\Big(\|\hat{f}_n - f_0\|_{2,n}^2 \geq 3r_*^2\Big) \\
&\leq 2P\left(\frac{1}{n}\sum_{j=1}^{b}\sum_{t\in T_{1,j}}\big[\hat{f}_n(\boldsymbol{Z}_t) - f_0(\boldsymbol{Z}_t)\big]^2 \geq 2r_*^2\right) + P\left(\frac{1}{n}\sum_{t\in T_R}\big[\hat{f}_n(\boldsymbol{Z}_t) - f_0(\boldsymbol{Z}_t)\big]^2 \geq r_*^2\right) \\
&\coloneqq 2P_3 + P_4
\end{aligned}
\tag{A.44}
$$

To bound P_3 we first apply (A.21) with

$$
E = \left\{\frac{1}{n}\sum_{j=1}^{b}\sum_{t\in T_{1,j}}\big[\hat{f}_n(\boldsymbol{Z}_t) - f_0(\boldsymbol{Z}_t)\big]^2 \geq 2r_*^2\right\},
$$

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

$$
\begin{aligned}
P_3 &\leq P\Big((ab/n)\|\hat{f}_n - f_0\|_{\overline{T}_1}^2 \geq 2r_*^2\Big) + \frac{n\,\beta(a)}{2a} \\
&\leq P\Big(\|\hat{f}_n - f_0\|_{\overline{T}_1}^2 \geq 4r_*^2\Big) + \frac{n\,\beta(a)}{2a} \\
&\leq P\Big(\Big\{\|\hat{f}_n - f_0\|_{\overline{T}_1}^2 \geq 4r_*^2\Big\} \cap \Big\{\|\hat{f}_n - f_0\|_{\mathcal{L}^2} \leq r_*\Big\}\Big) + P\Big(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} > r_*\Big) + \frac{n\,\beta(a)}{2a} \\
&\leq P\left(\sup_{\{f\in\mathcal{F}_n:\|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\leq r_*\}}\|f - f_0\|_{\overline{T}_1}^2 \geq 4r_*^2\right) + P\Big(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} > r_*\Big) + \frac{n\,\beta(a)}{2a} \\
&\leq P\left(\sup_{\{f\in\mathcal{F}_n:\|f-f_0\|_{\mathcal{L}^2(P_{\boldsymbol{Z}})}\leq r_*\}}\|f - f_0\|_{\overline{T}_1}^2 \geq 4r_*^2\right) \\
&\qquad + e^{-\delta} + 2\log(n)\left[\frac{n\,\beta(a)}{a} + 2P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big)\right] + \frac{n\,\beta(a)}{2a},
\end{aligned}
$$

by (A.42). In Section A.3.5.4 it was shown r_* and δ' satisfy (A.26) and (A.27). They are clearly still met by δ (although $\delta \geq 1$ may not hold). Hence, by applying (A.28) with δ in place of δ', the previous display becomes

$$
P_3 \leq 2e^{-\delta} + 3\log(n)\left[\frac{n\,\beta(a)}{a} + 2P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big)\right], \tag{A.45}
$$

Consider P_4. Recall $(\#T_R) \coloneqq n - 2ab < 2a$, from Section A.3.5.2. Then, by (b.4) and $\|f_0\|_\infty \leq 1$,

$$
\frac{1}{n}\sum_{t\in T_R}\big[\hat{f}_n(\boldsymbol{Z}_t) - f_0(\boldsymbol{Z}_t)\big]^2 \leq \frac{2a}{n}\|\hat{f}_n - f_0\|_\infty^2 < \frac{8B_n^2 a}{n} < r_*^2
$$

where the last inequality has used the definition of r_* from (A.40), and $C_4 \geq 1$ by (b.5)(ii). Therefore,

$$P_4 = P\left(\frac{1}{n}\sum_{t\in T_R}\left[\hat{f}_n(\boldsymbol{Z}_t) - f_0(\boldsymbol{Z}_t)\right]^2 \geq r_*^2\right) = 0. \tag{A.46}$$

Applying (A.45) and (A.46) to (A.44),

$$P\Big(\|\hat{f}_n - f_0\|_{2,n} \geq \sqrt{3}\,C\,\epsilon_n(\delta, a)\Big) \leq 4e^{-\delta} + 6\log(n)\left[\frac{n\,\beta(a)}{a} + 2P\Big(\max_{t\in\{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) > C_4 B_n\Big)\right].$$

This completes the proof of Theorem 2.1.2.

A.3.7 Supporting Lemmas

This section provides the ancillary lemmas used in Section A.3. These are simple modifications of existing results for more direct application to the setting used here.

Lemma A.3.1 is a simplified version of Davidson (2022, Theorem 15.1). For any random sequence $\{X_t\}_{t\in\mathbb{N}}$ let α_X and β_X be the mixing coefficients associated with $\{X_t\}_{t\in\mathbb{N}}$.

Lemma A.3.1. *Let $U : \boldsymbol{\mathcal{Z}} \to \mathbb{R}$ be measurable-$\mathcal{A}/\mathcal{B}(\mathbb{R})$ and define $W_t := U(\boldsymbol{Z}_t)$. Then $\alpha_W(j) \leq \alpha_Z(j)$ and $\beta_W(j) \leq \beta_Z(j)$ for any $j \in \mathbb{N}$.*

Proof. Note that Y_t is measurable-$\sigma(\boldsymbol{Z}_t)/\mathcal{B}(\mathbb{R})$ for each $t \in \mathbb{N}$. Consequently, $\sigma\big(\{Y_t\}_1^k\big) \subseteq \sigma\big(\{\boldsymbol{Z}_t\}_1^k\big)$ and $\sigma\left(\{Y_t\}_{k+j}^{\infty}\right) \subseteq \sigma\left(\{\boldsymbol{Z}_t\}_{k+j}^{\infty}\right)$, for any $k, j \in \mathbb{N}$. With this, the desired follows immediately from Definitions 2.1.3 and 2.2.1. ■

Lemma A.3.2 follows from the exponential inequality for α-mixing processes in Merlevède et al. (2009, Theorem 1). Note that stationarity is not required. See Definition 2.2.1 for the definition of α-mixing. This result is also applicable to β-mixing processes since $\beta(j) \geq \alpha(j)$.

Lemma A.3.2. *Let $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ be an α-mixing process with $\alpha(j) \leq C'_\alpha e^{-C_\alpha j}$ for some $C_\alpha, C'_\alpha > 0$. Let $f : \boldsymbol{\mathcal{Z}} \to \mathbb{R}$ be measurable-$\mathcal{B}(\boldsymbol{\mathcal{Z}})/\mathcal{B}(\mathbb{R})$, such that $\mathbb{E}[f(\boldsymbol{Z}_t)] = 0$ for each t and $\|f\|_{\mathcal{L}^\infty} < B$. Then, there exists $C_6 > 0$ depending only on C_α, C'_α, such that for any $a \geq 3$, and $n, \delta > 0$*

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

Proof. By Lemma A.3.1, $\{f(\boldsymbol{Z}_t)\}_{t=1}^a$ has an α-mixing coefficient that is less than or equal to $\alpha(j)$ from the statement of this lemma. Then, by Merlevède et al. (2009, Theorem 1),[8] there exists $C_6 > 0$ depending only on C_α, C'_α such that for any $\gamma > 0$

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{a} f(\boldsymbol{Z}_t)\right| \geq \gamma\right) \leq \exp\left[-\frac{C_6\gamma^2 n^2}{aB^2 + \gamma n B(\log a)(\log\log a)}\right].$$

Setting $\delta = C_6\gamma^2 n^2\big[aB^2 + \gamma n B(\log a)(\log\log a)\big]^{-1}$ implies

$$0 = C_6\gamma^2 n^2 - \delta\gamma n B(\log a)(\log\log a) - \delta a B^2.$$

Hence, by the quadratic formula,[9]

$$\begin{aligned}\gamma &= \frac{\delta n B(\log a)(\log\log a) + \sqrt{\big[\delta n B(\log a)(\log\log a)\big]^2 + 4C_6 n^2 \delta a B^2}}{2C_6 n^2} \\ &\leq \frac{2\delta n B(\log a)(\log\log a)}{2C_6 n^2} + \frac{2nB\sqrt{C_6\delta a}}{2C_6 n^2} = \frac{\delta B(\log a)(\log\log a)}{C_6 n} + \frac{B\sqrt{\delta a}}{\sqrt{C_6}\, n},\end{aligned}$$

since $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for any $x, y > 0$. ∎

Lemma A.3.3 is a contraction inequality for Rademacher complexities of sums that follows from Maurer (2016, Theorem 3).

Lemma A.3.3. *For $a, b \in \mathbb{N}$, and let $\{T_j\}_{j=1}^b$ form a partition of $\{1, \ldots, ab\}$ such that $\#T_j = a$ for each j. Let $\mathcal{S}$ be a pointwise separable space. For $t \in \{1, \ldots, ab\}$, let $g_t : \mathcal{S} \to \mathbb{R}$ and $h_t : \mathcal{S} \to \mathbb{R}$ be such that there exists a constant $L > 0$, where $g_t(f) - g_t(f') \leq L|h_t(f) - h_t(f')|$ for each t and $f, f' \in \mathcal{S}$. Then,*

$$\mathbb{E}_\xi\left[\sup_{f\in\mathcal{S}} \sum_{j=1}^{b} \xi_j \cdot \left(\frac{1}{a}\sum_{t\in T_j} g_t(f)\right)\right] \leq \sqrt{2/a}\, L\, \mathbb{E}_\xi\left[\sup_{f\in\mathcal{S}} \sum_{j=1}^{b}\sum_{t\in T_j} \xi_t\, h_t(f)\right],$$

where $\{\xi_j\}_{j=1}^b$ and $\{\xi_t\}_{t=1}^{ab}$ are sequences of i.i.d. Rademacher random variables.

[8] Note that Merlevède et al. (2009, Theorem 1) is for $\alpha(j) \leq e^{-C_\alpha j}$. However, this can be generalized to $\alpha(j) \leq C'_\alpha e^{-C_\alpha j}$ for $C'_\alpha > 0$, by adjusting the constants in their results. See (4.13), Lemma 8, and Corollary 11 therein.

[9] Note that we are only interested in $\gamma > 0$ and

$$\delta n B(\log a)(\log\log a) - \sqrt{\big[\delta n B(\log a)(\log\log a)\big]^2 + 4C_6 n^2 \delta a B^2} \;<\; \delta n B(\log a)(\log\log a) - \sqrt{\big[\delta n B(\log a)(\log\log a)\big]^2} \;=\; 0.$$

Proof. For $\boldsymbol{x} \in \mathbb{R}^a$ let $\|\boldsymbol{x}\|_E$ denote the Euclidean norm on $\mathbb{R}^a$. By Maurer (2016, Theorem 3)[10] for countable set $\mathcal{H}$ and functions $\psi_j : \mathcal{H} \to \mathbb{R}$, $\boldsymbol{\phi}_j : \mathcal{H} \to \mathbb{R}^a$, $j \in \{1, \ldots, b\}$ such that

$$\psi_j(f) - \psi_j(f') \leq \left\|\boldsymbol{\phi}_j(f) - \boldsymbol{\phi}_j(f')\right\|_E, \quad \forall f, f' \in \mathcal{H},\ j \in \{1, \ldots, b\},$$

we have, for $\boldsymbol{\phi}_j = \{\phi_t\}_{t \in T_j}$,

$$\mathbb{E}_\xi \left[\sup_{f \in \mathcal{H}} \sum_{j=1}^{b} \xi_j \psi_j(f) \right] \leq \sqrt{2}\, \mathbb{E}_\xi \left[\sup_{f \in \mathcal{H}} \sum_{j=1}^{b} \sum_{t \in T_j} \xi_t\, \phi_t(f) \right].$$

By assumption $\mathcal{S}$ is pointwise separable so choose $\mathcal{H}$ to be the countable dense subset of $\mathcal{S}$. Let

$$\psi_j(f) = \frac{1}{a} \sum_{t \in T_j} g_t(f), \quad \text{and} \quad \boldsymbol{\phi}_j(f) = \left\{ L h_t(f) / \sqrt{a} \right\}_{t \in T_j}.$$

With this,

$$\begin{aligned} \psi_j(f) - \psi_j(f') &= \frac{1}{a} \sum_{t \in T_j} \left\{ g_t(f) - g_t(f') \right\} \leq \frac{1}{a} \sum_{t \in T_j} L |h_t(f) - h_t(f')| \\ &\leq \left(\frac{1}{a} \sum_{t \in T_j} \left| L h_t(f) - L h_t(f') \right|^2 \right)^{1/2} = \left\| \boldsymbol{\phi}_j(f) - \boldsymbol{\phi}_j(f') \right\|_E. \end{aligned}$$

Hence, Maurer (2016, Theorem 3) implies

$$\mathbb{E}_\xi \left[\sup_{f \in \mathcal{H}} \sum_{j=1}^{b} \xi_j \left(\frac{1}{a} \sum_{t \in T_j} g_t(f) \right) \right] \leq \sqrt{2/a}\, L\, \mathbb{E}_\xi \left[\sup_{f \in \mathcal{H}} \sum_{j=1}^{b} \sum_{t \in T_j} \xi_{j,t}\, h_t(f) \right].$$

This completes the proof because the supremum is unchanged when $\mathcal{H}$ is replaced by $\mathcal{S}$. ■

Lemma A.3.4. *Let $\mathcal{G}$ be a set of real-valued functions such that $\sup_{f \in \mathcal{G}} \|f\|_\infty \leq B$. Then, for any $r \in [1, \infty]$, $\delta \in (0, 2B]$, and $n \in \mathbb{N}$ such that $n \geq \mathrm{Pdim}(\mathcal{G})$,*

$$\mathrm{N}_r^{(\infty)}(\delta,\, \mathcal{G},\, n) \leq \left(\frac{2eBn}{\delta \cdot \mathrm{Pdim}(\mathcal{G})} \right)^{\mathrm{Pdim}(\mathcal{G})}.$$

Proof. By Anthony and Bartlett (1999, Theorem 12.2)

$$\mathrm{N}_\infty^{(\infty)}(\delta,\, \mathcal{G},\, n) \leq \left(\frac{2eBn}{\delta \cdot \mathrm{Pdim}(\mathcal{G})} \right)^{\mathrm{Pdim}(\mathcal{G})}.$$

[10] There appears to be a typo in the statement of Theorem 3 in Maurer (2016), the term $\psi(s_i')$ in the contraction condition should be $\psi_i(s')$, as in Lemma 7 therein.

For any $r \in [1,\infty]$, $n \in \mathbb{N}$, and $\boldsymbol{x} \in \mathbb{R}^n$, we have $\|\boldsymbol{x}\|_{r,n} \leq \|\boldsymbol{x}\|_{\infty,n}$; so any δ-cover with respect to $\|\cdot\|_{\infty,n}$ is also a cover for $\|\cdot\|_{r,n}$. Thus, $\mathrm{N}_r^{(\infty)}(\delta, \mathcal{G}, n) \leq \mathrm{N}_\infty^{(\infty)}(\delta, \mathcal{G}, n)$. ■

Lemma A.3.5. *Let $\mathcal{G}$ be a pointwise separable set of functions with elements $f : \boldsymbol{\mathcal{Z}} \to \mathbb{R}$ such that $\|f\|_\infty < \infty$ for each $f \in \mathcal{G}$. Then, for any $n \in \mathbb{N}$ we have*

$$\mathbb{E}_\xi\Big[\mathfrak{R}_n\Big\{f : f \in \mathcal{G},\, \|f\| \leq r\Big\}\Big] \leq \inf_{0<\alpha<r}\left\{4\alpha + 8\sqrt{\frac{2}{n}}\int_\alpha^r \sqrt{\log \mathrm{N}(\upsilon, \mathcal{G}, \|\cdot\|)}\, d\upsilon\right\},$$

where $\|f\| = \left(\frac{1}{n}\sum_{t=1}^n f(\boldsymbol{Z}_t)^2\right)^{1/2}$.

Proof. Consider the case where $\{\boldsymbol{Z}_t\}_{t=1}^n = \{\boldsymbol{z}_t\}_{t=1}^n$ is an arbitrary fixed element of $\boldsymbol{\mathcal{Z}}^n$. Then $\left\{\frac{1}{\sqrt{n}}\sum_{t=1}^n \xi_t f(\boldsymbol{z}_t) : f \in \mathcal{G},\ \|f\| \leq r\right\}$ is a zero mean sub-Gaussian empirical process; since Hoeffding's inequality for Rademacher random variables implies,[11]

$$P\left(\frac{1}{\sqrt{n}}\sum_{t=1}^n \xi_t f(\boldsymbol{z}_t) \geq \upsilon\right) \leq 2\exp\left[\frac{-\upsilon^2 n}{2\sum_{t=1}^n f(\boldsymbol{z}_t)^2}\right] \leq 2\exp\left[\frac{-\upsilon^2}{2\|f\|^2}\right],$$

for any $f \in \{f \in \mathcal{G}, \|f\| \leq r\}$ and $\upsilon > 0$. Hence, Dudley's entropy integral can be applied[12] to obtain, for each $\alpha \in (0, r)$

$$\begin{aligned}
\mathbb{E}_\xi\Big[\mathfrak{R}_n\Big\{f : f \in \mathcal{G}|_{\{\boldsymbol{Z}_t\}_{t=1}^n},\, \|f\| \leq r\Big\}\Big] &= \frac{1}{\sqrt{n}}\,\mathbb{E}_\xi\left[\sup_{\{f\in\mathcal{G},\|f\|\leq r\}} \frac{1}{\sqrt{n}}\sum_{t=1}^n \xi_t f(\boldsymbol{z}_t)\right] \\
&\leq \frac{1}{\sqrt{n}}\left(2\mathbb{E}_\xi\left[\sup_{\{f\in\mathcal{G},\|f\|\leq 2\alpha\}} \frac{1}{\sqrt{n}}\sum_{t=1}^n \xi_t\big[f(\boldsymbol{z}_t) - f'(\boldsymbol{z}_t)\big]\right] + 8\sqrt{2}\int_\alpha^r \sqrt{\log \mathrm{N}(\upsilon, \mathcal{G}, \|\cdot\|)}\, d\upsilon\right) \\
&\leq 2\sup_{\{f\in\mathcal{G},\|f\|\leq 2\alpha\}}\left\{\frac{1}{n}\sum_{t=1}^n \big|f(\boldsymbol{z}_t) - f'(\boldsymbol{z}_t)\big|\right\} + 8\sqrt{\frac{2}{n}}\int_\alpha^r \sqrt{\log \mathrm{N}(\upsilon, \mathcal{G}, \|\cdot\|)}\, d\upsilon \\
&\leq 4\alpha + 8\sqrt{\frac{2}{n}}\int_\alpha^r \sqrt{\log \mathrm{N}(\upsilon, \mathcal{G}, \|\cdot\|)}\, d\upsilon.
\end{aligned}$$

The desired result follows since this holds for any $\{\boldsymbol{z}_t\}_{t=1}^n \in \boldsymbol{\mathcal{Z}}^n$ and $\alpha \in (0, r)$. ■

[11] This result is a simple modification of Hoeffding (1963, Theorem 2), and can be found in van der Vaart and Wellner (1996, Lemma 2.2.7).

[12] This is a well-known result with many formulations. This version, and it's proof can be found in Bartlett (2013, p. 11), or in the proof of van der Vaart and Wellner (1996, Theorem 2.2.4). This result is sometimes referred to as Dudley's chaining (e.g. Farrell et al., 2021, Lemma 3).

A.4 Independent Block Construction

This section more rigorously describes the process used to construct the sequence $\{\overline{\boldsymbol{Z}}_t\}_{t=1}^n$ from Appendix A.3.4. Define $\boldsymbol{Y}_1 \coloneqq \{\boldsymbol{Z}_t\}_{t\in T_{1,1}}$, and $\boldsymbol{X}_1 \coloneqq \{\boldsymbol{Z}_t\}_{t\in\{\{1,\dots,n\}\setminus T_{1,1}\}}$. Clearly, $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ are random variables on $(\Omega,\mathcal{A},P)$ taking values in $\mathcal{X} \coloneqq \boldsymbol{\mathcal{Z}}^{n-a}$ and $\mathcal{Y} \coloneqq \boldsymbol{\mathcal{Z}}^a$, respectively. Now, let λ be the Lebesgue measure, and consider the product probability space

$$(\Omega',\mathcal{A}',P') \;=\; (\Omega,\mathcal{A},P)\;\times\;\big([0,1],\mathcal{B}_{[0,1]},\lambda\big) \;\coloneqq\; \big(\Omega\times[0,1],\,\mathcal{A}\otimes\mathcal{B}_{[0,1]},P\times\lambda\big).$$

We can extend $(\Omega,\mathcal{A},P)$ to this richer space with the extension[13] $\pi:\Omega'\to\Omega$, where π denotes coordinate projection onto Ω, i.e. $\pi(\omega_1,\omega_2)=\omega_1\in\Omega$ for any $(\omega_1,\omega_2)\in\Omega'$. Thus, $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ on $(\Omega,\mathcal{A},P)$ can be redefined as $\boldsymbol{X}_1\circ\pi$ and $\boldsymbol{Y}_1\circ\pi$ on $(\Omega',\mathcal{A}',P')$, without changing their distribution. We may refer to $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ as random variables on $(\Omega',\mathcal{A}',P')$ with the understanding that this means $\boldsymbol{X}_1\circ\pi$ and $\boldsymbol{Y}_1\circ\pi$. Then, by Berbee's Lemma[14] there exists a random variable $\overline{\boldsymbol{Y}}_1$ on $(\Omega',\mathcal{A}',P')$ that has the same distribution as $\boldsymbol{Y}_1$ and is independent of $\boldsymbol{X}_1$.

Set $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,1}} \coloneqq \overline{\boldsymbol{Y}}_1$, then let $\boldsymbol{Y}_2 \coloneqq \{\boldsymbol{Z}_t\}_{t\in T_{2,1}}$, and $\boldsymbol{X}_2 \coloneqq \{\boldsymbol{Z}_t\}_{t\in\{\{1,\dots,n\}\setminus T_{2,1}\}}\cup\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,1}}$. Then, using the same process as before, we can construct $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{2,1}} \coloneqq \overline{\boldsymbol{Y}}_2$ independent of $\boldsymbol{X}_2$ and distributed as $\boldsymbol{Y}_2$. This process can be repeated until the sequence $\{\overline{\boldsymbol{Z}}_t\}_{t=1}^n$ is constructed with the desired properties.

A.5 Proofs for Section 2.2

First, we provide proofs for Lemma 2.2.1 and Proposition 2.2.1, followed by the proofs for the main theorems of Section 2.2. Appendix A.5.4 lists the additional ancillary lemmas used in this section. As before, we write $q_t(f) \coloneqq q\big(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)\big)$, and $m_{nt} \coloneqq m_n(\boldsymbol{Z}_t)$.

[13] Here extension of $(\Omega,\mathcal{A},P)$ to $(\Omega',\mathcal{A}',P')$ refers to a measurable map $\pi:\Omega'\to\Omega$ such that $P'(\pi^{-1}A)=P(A)$ for any $A\in\mathcal{A}$. It is easy to show that coordinate projections are extensions for product probability spaces (e.g. see Davidson (2022) discussion in first paragraph of §3.5 pp.70,71).

[14] This refers to Berbee (1979) Corollary 4.2.5 with proof on pages 91-95 therein. Similar results can also be found in the following: Bosq (1998) §1.2 Lemma 1.1; Doukhan (1994) §1.2.1 Theorem 1; Bryc (1982) Theorem 3.1; or Merlevède and Peligrad (2002).

Proof of Lemma 2.2.1. Note that $\sup_{t\in\mathbb{N}} Y_t^2$ is measurable-$\mathcal{A}/\mathcal{B}(\overline{\mathbb{R}})$. Then,

$$\begin{aligned}\max_{t\in\{1,\ldots,n\}} \mathbb{E}\big[Y_t^2\,\mathbb{1}_{|Y_t|\geq B_n}\big] &\leq \mathbb{E}\Big[\max_{t\in\{1,\ldots,n\}}\{Y_t^2\}\cdot\max_{t\in\{1,\ldots,n\}}\{\mathbb{1}_{|Y_t|\geq B_n}\}\Big] \\ &\leq \int_{\{\omega:\max_{1\leq t\leq n}|Y_t(\omega)|\geq B_n\}} \sup_{t\in\mathbb{N}} Y_t^2\,dP. \end{aligned} \tag{A.47}$$

For any $\delta>0$, there exists a simple function,[15] $s_\delta:\Omega\to\mathbb{R}$, such that $0\leq s_\delta\leq\sup_{t\in\mathbb{N}} Y_t^2$, and

$$\int_\Omega \sup_{t\in\mathbb{N}} Y_t^2\,dP - \int_\Omega s_\delta\,dP \leq \delta/2. \tag{A.48}$$

We may choose $C_\delta>0$ such that $\sup_{\omega\in\Omega} s_\delta(\omega)\leq C_\delta$, since a simple function takes on finitely many values. By Assumption 2.2.2 for any constants $\delta, C_\delta>0$ there exists $N_\delta\in\mathbb{N}$ such that, for all $n\geq N_\delta$,

$$P\left(\max_{t\in\{1,\ldots,n\}}|Y_t|\geq B_n\right)\leq \delta/(2C_\delta). \tag{A.49}$$

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

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

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

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

by (A.49) and (A.48). Applying this to (A.47) completes the proof since δ is arbitrary, N_δ depends only on δ, C_δ and s_δ, C_δ are independent of n. ■

Proof of Proposition 2.2.1. For condition (i), note that $\{|Y_t|\}_{t\in\mathbb{N}}$ is stationary and α-mixing by Lemma A.3.1. Then, $\{|Y_t|\}_{t\in\mathbb{N}}$ satisfies Leadbetter et al. (1983, Condition $D(u_n)$, p.53) (see

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

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., $\overline{Y}_t := \frac{Y_t - \mathbb{E}(Y_t)}{\sqrt{\text{Var}(Y_t)}}$. Note that $\mathbb{E}\big[\overline{Y}_t^4\big] = 3$ for all $t \in \mathbb{N}$, by Papoulis (1991, p.110). By Lemma A.3.1, $\alpha_{\overline{Y}}(j) \leq \alpha(j)$ 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 $i, k \in \mathbb{N}$, $i \neq k$,

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

and $|\text{Cor}(-Y_i, -Y_k)| \leq \upsilon_{|i-k|}$ by a similar argument. Then, by the assumptions on $\alpha(j)$ we have $\upsilon_j < 1$ for all $j \in \mathbb{N}$ and $\lim_{j\to\infty} \upsilon_j \log(j) = 0$. Next

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

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

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

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

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

as $n \to \infty$. ■

A.5.1 Proof of Theorem 2.2.1

Theorem 2.2.1 will follow by showing the conditions of Theorem 2.1.1 are met with the following setting:

- $q_t(f) = q(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)) \coloneqq \big(\pi_Y(\boldsymbol{Z}_t) - f(\pi_{\boldsymbol{X}}(\boldsymbol{Z}_t))\big)^2 = (Y_t - f(\boldsymbol{X}_t))^2$;
- $\rho_n(f, f') \coloneqq \|f - f'\|_{\mathcal{L}^2(P_{\{\boldsymbol{X}_t\}_{t=1}^n})} = \Big(\int_{[0,1]^d} \big[f(\boldsymbol{x}) - f'(\boldsymbol{x})\big]^2 dP_{\{\boldsymbol{X}_t\}_{t=1}^n}\Big)^{1/2}$;
- $M_n \coloneqq 4B_n$; and
- $m_n(\boldsymbol{Z}_t) \coloneqq 2\sup_{f\in\mathcal{N}_n} \big|Y_t - f(\boldsymbol{X}_t)\big|$.

Verification of (a.1). By Lemma A.5.1, and (2.5), there exists $\tilde{f}_n \in \mathcal{N}_n$ such that $\|\tilde{f}_n - f_0\|_\infty \lesssim n^{-\left(\frac{p}{p+d/2}\right)(1/4-K_B)}$. By assumption $\epsilon_n \gtrsim n^{\left(\frac{p}{p+d/2}\right)(1/4-K_B)}$. Thus, there exist constants $C, C' > 0$ such that $\|\tilde{f}_n - f_0\|_\infty \leq Cn^{\left(\frac{p}{p+d/2}\right)(1/4-K_B)} \leq C'\epsilon_n$, (a.1) is met by $C'\epsilon_n$. The result follows since $\rho_n(\hat{f}_n, f_0) = \mathrm{O}_P(\epsilon_n)$ is equivalent to $\rho_n(\hat{f}_n, f_0) = \mathrm{O}_P(C'\epsilon_n)$ so scaling ϵ_n by a constant has no impact on the final rate. ■

Verification of (a.2). For any $n \in \mathbb{N}$, and $f \in \mathcal{L}^2\big(P_{\{\boldsymbol{X}_t\}_{t=1}^n}\big)$, by iterated expectations,

$$\begin{aligned}
\mathbb{E}[Q_n(f)] - \mathbb{E}[Q_n(f_0)] &= \frac{1}{n}\sum_{t=1}^n \mathbb{E}\Big[f(\boldsymbol{X}_t)^2 - f_0(\boldsymbol{X}_t)^2 - 2f(\boldsymbol{X}_t)Y_t + 2f_0(\boldsymbol{X}_t)Y_t\Big] \\
&= \frac{1}{n}\sum_{t=1}^n \mathbb{E}\Big[f(\boldsymbol{X}_t)^2 - f_0(\boldsymbol{X}_t)^2 - 2f(\boldsymbol{X}_t)f_0(\boldsymbol{X}_t) + 2f_0(\boldsymbol{X}_t)^2\Big] \\
&= \frac{1}{n}\sum_{t=1}^n \mathbb{E}\Big[(f(\boldsymbol{X}_t) - f_0(\boldsymbol{X}_t))^2\Big] \\
&= \int_\Omega \frac{1}{n}\sum_{t=1}^n \Big[f(\boldsymbol{X}_t(\omega)) - f_0(\boldsymbol{X}_t(\omega))\Big]^2 dP \\
&= \int_{[0,1]^d} \big[f(\boldsymbol{x}) - f_0(\boldsymbol{x})\big]^2 dP_{\{\boldsymbol{X}_t\}_{t=1}^n} = \|\hat{f}_n - f_0\|^2_{\mathcal{L}^2(P_{\{\boldsymbol{X}_t\}_{t=1}^n})}.
\end{aligned}$$

The desired result follows because $\mathcal{N}_n \subset \mathcal{L}^2\big(P_{\{\boldsymbol{X}_t\}_{t=1}^n}\big)$ for any n, since $\sup_{f\in\mathcal{N}_n}\|f\|_\infty = B_n$. ■

Verification of (a.3). Let $m_n(\boldsymbol{Z}_t) \coloneqq 2\sup_{f\in\mathcal{N}_n} \big|Y_t - f(\boldsymbol{X}_t)\big|$, which is measurable-$\mathcal{B}(\boldsymbol{\mathcal{Z}})/\mathcal{B}([0,\infty))$

because $\mathcal{N}_n$ is pointwise-separable. Consider (a.3)(i). For any $f, g \in \mathcal{N}_n$ we have

$$\begin{aligned}
\big|q\big(\boldsymbol{z}, f\big) - q\big(\boldsymbol{z}, g\big)\big| &= \Big|\big(f(\boldsymbol{x}) + g(\boldsymbol{x})\big)\big(f(\boldsymbol{x}) - g(\boldsymbol{x})\big) - 2y\big(f(\boldsymbol{x}) - g(\boldsymbol{x})\big)\Big| \\
&= \Big|\big(f(\boldsymbol{x}) + g(\boldsymbol{x}) - 2y\big)\big(f(\boldsymbol{x}) - g(\boldsymbol{x})\big)\Big| \\
&\leq \Big(2 \sup_{f \in \mathcal{N}_n} \big|y - f(\boldsymbol{x})\big|\Big)\big|f(\boldsymbol{x}) - g(\boldsymbol{x})\big| \\
&= m_n(\boldsymbol{z})\big|f(\boldsymbol{x}) - g(\boldsymbol{x})\big|. \qquad \text{(A.50)}
\end{aligned}$$

Consider (a.3)(ii). Recall $m_n(\boldsymbol{Z}_t) \coloneqq 2\sup_{f \in \mathcal{N}_n} \big|Y_t - f(\boldsymbol{X}_t)\big|$, and $\sup_{f \in \mathcal{N}_n} \|f\|_\infty \leq B_n$. Then, for any $\boldsymbol{z} = (y, \boldsymbol{x}) \in \mathbb{R} \times [0,1]^d$, we have $m_n(\boldsymbol{z}) \leq 2\big(|y| + B_n\big)$, by the triangle inequality. Hence, for any $\omega \in \Omega$, $n \in \mathbb{N}$ and $t \in \{1, \ldots, n\}$,

$$m_n\big(\boldsymbol{Z}_t(\omega)\big) \leq 2\big(|Y_t(\omega)| + B_n\big),$$

since $\boldsymbol{Z}_t(\omega) \coloneqq \big(Y_t(\omega),\, \boldsymbol{X}_t(\omega)\big) \in \mathbb{R} \times [0,1]^d$. Thus, for any $\omega \in \Omega$, $n \in \mathbb{N}$,

$$\max_{t \in \{1,\ldots,n\}} m_n\big(\boldsymbol{Z}_t(\omega)\big) \leq \max_{t \in \{1,\ldots,n\}} 2\big(|Y_t(\omega)| + B_n\big),$$

which implies

$$\begin{aligned}
\Big\{\omega : \max_{t \in \{1,\ldots,n\}} m_n\big(\boldsymbol{Z}_t(\omega)\big) \geq M_n\Big\} &\subseteq \Big\{\omega : \max_{t \in \{1,\ldots,n\}} 2\big(|Y_t(\omega)| + B_n\big) \geq M_n\Big\} \\
&= \Big\{\omega : \max_{t \in \{1,\ldots,n\}} |Y_t(\omega)| \geq M_n/2 - B_n\Big\} \\
&= \Big\{\omega : \max_{t \in \{1,\ldots,n\}} |Y_t(\omega)| \geq B_n\Big\}, \qquad \text{(A.51)}
\end{aligned}$$

since $M_n \coloneqq 4B_n$. With this, and Assumption 2.2.2

$$P\Big(\max_{t \in \{1,\ldots,n\}} m_n(\boldsymbol{Z}_t) \geq M_n\Big) \leq P\Big(\max_{t \in \{1,\ldots,n\}} |Y_t| \geq B_n\Big) \to 0, \qquad \text{as } n \to \infty.$$

Consider (a.3)(iii). Note that (A.51) implies $\mathbb{1}_{\{m_n(\boldsymbol{Z}_t) \geq M_n\}} \leq \mathbb{1}_{\{|Y_t| \geq B_n\}}$, for all $n \in \mathbb{N}$ and $t \in \{1, \ldots, n\}$. Hence, for any $f \in \mathcal{N}_n$,

$$\begin{aligned}
\mathbb{E}\Big[\big|q_t(f)\big|\mathbb{1}_{\{m_n(\boldsymbol{Z}_t) \geq M_n\}}\Big] &\leq \mathbb{E}\Big[\big|Y_t^2 + f(\boldsymbol{X}_t)^2 - 2f(\boldsymbol{X}_t)Y_t\big|\mathbb{1}_{\{|Y_t| \geq B_n\}}\Big] \\
&\leq \mathbb{E}\big[Y_t^2\, \mathbb{1}_{\{|Y_t| \geq B_n\}}\big] + \mathbb{E}\big[f(\boldsymbol{X}_t)^2 \mathbb{1}_{\{|Y_t| \geq B_n\}}\big] + 2\mathbb{E}\big[\big|f(\boldsymbol{X}_t)Y_t\big|\mathbb{1}_{\{|Y_t| \geq B_n\}}\big] \\
&\leq \mathbb{E}\big[Y_t^2\, \mathbb{1}_{\{|Y_t| \geq B_n\}}\big] + B_n^2 P\big(|Y_t| \geq B_n\big) + 2B_n \mathbb{E}\big[|Y_t| \mathbb{1}_{\{|Y_t| \geq B_n\}}\big], \qquad \text{(A.52)}
\end{aligned}$$

where the last line has used $\sup_{f\in\mathcal{N}_n}\|f\|_\infty \leq B_n$ by (2.4). Note that, with $B_n > 0$ and Markov's inequality,

$$B_n^2\, P\big(|Y_t| \geq B_n\big) = B_n^2\, P\big(|Y_t|\mathbb{1}_{\{|Y_t|\geq B_n\}} \geq B_n\big) = B_n^2\, P\big(Y_t^2\, \mathbb{1}_{\{|Y_t|\geq B_n\}} \geq B_n^2\big) \leq \mathbb{E}\big[Y_t^2\, \mathbb{1}_{\{|Y_t|\geq B_n\}}\big],$$

and,

$$B_n\mathbb{E}\big[|Y_t|\, \mathbb{1}_{\{|Y_t|\geq B_n\}}\big] \leq \mathbb{E}\big[Y_t^2\, \mathbb{1}_{\{|Y_t|\geq B_n\}}\big],$$

since $B_n|Y_t|\, \mathbb{1}_{\{|Y_t|\geq B_n\}} \leq Y_t^2$ when $|Y_t| \geq B_n$, and $B_n|Y_t|\, \mathbb{1}_{\{|Y_t|\geq B_n\}} = 0$ when $|Y_t| < B_n$. Using the previous two displays with (A.52),

$$\max_{t\in\{1,\ldots,n\}} \left\{ \sup_{f\in\mathcal{N}_n} \mathbb{E}\Big[\big|q\big(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)\big)\big|\mathbb{1}_{\{m_n(\boldsymbol{Z}_t)\geq M_n\}}\Big] \right\} \leq 4\Big(\max_{t\in\{1,\ldots,n\}} \mathbb{E}\big[Y_t^2\, \mathbb{1}_{\{|Y_t|\geq B_n\}}\big] \Big) \lesssim \epsilon_n^2, \tag{A.53}$$

by the assumptions on ϵ_n. Then, $\lim_{n\to\infty}\epsilon_n = 0$ under Assumption 2.2.2 by Lemma 2.2.1 and the assumptions on ϵ_n and K_B. ■

Verification of (a.4). First, $m_n(\boldsymbol{Z}_t) \coloneqq 2\sup_{f\in\mathcal{N}_n}\big|Y_t - f(\boldsymbol{X}_t)\big|$, implies

$$\big|q_t(f)\mathbb{1}_{\{m_{nt}<M_n\}}\big| = \big(Y_t - f(\boldsymbol{X}_t)\big)^2\mathbb{1}_{\{m_{nt}<M_n\}} \leq \big(m_n(\boldsymbol{Z}_t)/2\big)^2\mathbb{1}_{\{m_{nt}<M_n\}} \leq M_n^2/4 = B_n^2$$

For each $n\in\mathbb{N}$, by Lemma A.3.1 $\big\{q_t(f)\mathbb{1}_{\{m_{nt}<M_n\}}\big\}_{t=1}^n$ has an α-mixing coefficient that is bounded above by the α-mixing coefficient for $\{\boldsymbol{Z}_t\}_{t=1}^n$. Then, using the same reasoning as the proof of Lemma A.3.2, by Merlevède et al. (2009, Theorem 1) there exists a constant $C' > 0$ depending only on C_α, C'_α such that such that for $\delta > 0$ and all $n \geq 4$,

$$\begin{aligned} P\left(\frac{1}{n}\left|\sum_{t=1}^n \Big(q_t(f)\mathbb{1}_{\{m_{nt}<M_n\}} - \mathbb{E}\big[q_t(f)\mathbb{1}_{\{m_{nt}<M_n\}}\big]\Big)\right| \geq \delta\right) &\leq \exp\left[\frac{-C'\,\delta^2\, n^2}{8nB_n^4 + 2\delta n B_n^2(\log n)(\log\log n)}\right] \\ &\leq \exp\left[\frac{-C\,\delta^2\, n}{n^{4K_B} + \delta\, n^{2K_B}(\log n)(\log\log n)}\right] \\ &=: \lambda_n^{(q)}(\delta), \end{aligned}$$

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

Consider (a.4)(ii). Note that

$$\lambda_n^{(q)}\big(\delta\,\epsilon_n^2\big) = \exp\left[\frac{-C\,\delta^2\epsilon_n^4\, n}{n^{4K_B} + \delta^2\epsilon_n^2\, n^{2K_B}(\log n)(\log\log n)}\right].$$

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

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

since $1 - \frac{p}{p+d/2} > 0$ and $K_B < 1/4$ by assumption. Then, Lemma A.5.2 can be applied, with $4M_n = B_n$ and the definition of $\lambda_n^{(q)}$, to obtain the following sufficient condition for (a.4)(ii),

$$\begin{aligned}&\lim_{n\to\infty}\left\{\left[n^{(1-4K_B)}\epsilon_n^4\right]^{-1} \cdot \log \mathrm{N}_1^{(\infty)}\big(4\delta\,\epsilon_n^2/B_n,\, \mathcal{N}_n,\, n\big)\right\} = 0, \\ \implies \quad &\lim_{n\to\infty}\left\{\lambda_n^{(q)}\big(\delta\,\epsilon_n^2\big) \cdot \mathrm{N}_1^{(\infty)}\big(\delta\,\epsilon_n^2/M_n,\, \mathcal{N}_n,\, n\big)\right\} = 0.\end{aligned} \tag{A.54}$$

Henceforth, let n be large enough such that $\epsilon_n < 1$. By (2.5) and Lemma A.5.4, we have $\mathrm{Pdim}(\mathcal{N}_n) \asymp n^{2\left(\frac{d}{p+d/2}\right)(1/4-K_B)} \log^7(n)$, which implies $\lim_{n\to\infty}\{n/\mathrm{Pdim}(\mathcal{N}_n)\} = \infty$. Then, we can apply Lemma A.5.5, with $\eta = \left(\frac{d}{p+d/2}\right)(1/4 - K_B)$ therein, to obtain

$$\begin{aligned}\log \mathrm{N}_1^{(\infty)}\left(\frac{4\delta\,\epsilon_n^2}{B_n},\, \mathcal{N}_n,\, n\right) &\lesssim n^{2\left(\frac{d}{p+d/2}\right)(1/4-K_B)} \log^7(n)\Big[\log(n) + \log\left(B_n/\epsilon_n^2\right)\Big] \\ &\lesssim n^{2\left(\frac{d}{p+d/2}\right)(1/4-K_B)} \log^8(n) = n^{\left(\frac{d/2}{p+d/2}\right)(1-4K_B)} \log^8(n).\end{aligned}$$

With this, and again using $\frac{\epsilon_n^4\, n}{n^{4K_B}+n^{2K_B}(\log n)(\log\log n)} \gtrsim n^{(1-4K_B)}\epsilon_n^4$,

$$\begin{aligned}&\left[\frac{\epsilon_n^4\, n}{n^{4K_B} + \delta\, n^{2K_B}(\log n)(\log\log n)}\right]^{-1} \cdot \log \mathrm{N}_1^{(\infty)}\left(\frac{4\delta\,\epsilon_n^2}{B_n},\, \mathcal{N}_n,\, n\right) \\ &\quad\lesssim \left[n^{-(1-4K_B)}\epsilon_n^{-4}\right] n^{\left(\frac{d/2}{p+d/2}\right)(1-4K_B)} \log^8(n) \\ &\quad= n^{-\left(1-\frac{d/2}{p+d/2}\right)(1-4K_B)}\epsilon_n^{-4} \log^8(n) \\ &\quad= n^{-\left(\frac{p}{p+d/2}\right)(1-4K_B)}\epsilon_n^{-4} \log^8(n) \\ &\quad\lesssim \log^{4\upsilon}(n) \to 0, \quad \text{as } n \to \infty,\end{aligned}$$

where the last line used $\epsilon_n \gtrsim n^{-\left(\frac{p}{p+d/2}\right)(1/4-K_B)} \log^{2+\upsilon}(n)$, and $\upsilon > 0$ by assumption. ■

A.5.2 Proof of Theorem 2.2.2

Theorem 2.2.2 will be a consequence of the following proposition.

Proposition A.5.1. *Suppose Assumptions 2.2.1 and 2.2.2 hold with $B_n \asymp n^{K_B}$ for some $K_B \in [0, 1/2)$. Let $\{\boldsymbol{Z}_t\}_{t\in\mathbb{N}}$ be a strictly stationary β-mixing process with $\beta(j) \leq C'_\beta e^{-C_\beta j}$ for some $C_\beta, C'_\beta > 0$. Let $\mathcal{N}_n = \mathcal{N}(L_n, \boldsymbol{H}_n, B_n)$ be defined as in (2.4) 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 $H_{n,l} \asymp H_n$. For any $\upsilon \in [0, 1/2 - K_B)$ if*

$$L_n \asymp \log(n), \qquad H_n \asymp n^{\left(\frac{d}{p+d}\right)(1/2-K_B-\upsilon)} \log^2(n), \tag{A.55}$$

then for $\{\hat{f}_n\}_{n\in\mathbb{N}}$ satisfying (2.2), and

$$\epsilon_n = n^{-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)} \log^6(n) + \sqrt{\mathbb{E}\big[Y_t^2 \, \mathbb{1}_{\{|Y_t|\geq B_n\}}\big] + \theta_n},$$

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

$$P\left(\|\hat{f}_n - f_0\|_{\mathcal{L}^2} \leq C\,\epsilon_n\right) \geq 1 - e^{-n^{\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}} - \frac{2C'_\beta n^{1-C_\beta n^{2\upsilon}\log(n)-2\upsilon}}{\log(n)} - 4\log(n) P\Big(\max_{t\in\{1,\ldots,n\}} |Y_t| \geq B_n\Big),$$

$$P\left(\|\hat{f}_n - f_0\|_{2,n} \leq C\,\epsilon_n\right) \geq 1 - 6e^{-n^{\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}} - \frac{12C'_\beta n^{1-C_\beta n^{2\upsilon}\log(n)-2\upsilon}}{\log(n)} - 24\log(n) P\Big(\max_{t\in\{1,\ldots,n\}} |Y_t| \geq B_n\Big).$$

Note that Theorem 2.2.2 follows directly from Proposition A.5.1 by choosing $\upsilon = 0$. Proposition A.5.1 will follow by applying Theorem 2.1.2. We begin by verifying conditions (b.1)-(b.5) hold.

Verification of (b.1). This is assumed directly in Proposition A.5.1. ■

Verification of (b.2). By Lemma A.5.1, and (A.55), there exists $\tilde{f}_n \in \mathcal{N}_n$ such that $\|\tilde{f}_n - f_0\|_\infty \lesssim n^{-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}$. The desired result follows since $1/2 - K_B - \upsilon > 0$. ■

Verification of (b.3). For any $n \in \mathbb{N}$, and $f \in \mathcal{L}^2\big(P_{\{\boldsymbol{X}_t\}_{t=1}^n}\big)$, by iterated expectations,

$$\begin{aligned}
\mathbb{E}[Q_n(f)] - \mathbb{E}[Q_n(f_0)] &= \mathbb{E}\Big[f(\boldsymbol{X}_t)^2 - f_0(\boldsymbol{X}_t)^2 - 2f(\boldsymbol{X}_t)Y_t + 2f_0(\boldsymbol{X}_t)Y_t\Big] \\
&= \mathbb{E}\Big[f(\boldsymbol{X}_t)^2 - f_0(\boldsymbol{X}_t)^2 - 2f(\boldsymbol{X}_t)f_0(\boldsymbol{X}_t) + 2f_0(\boldsymbol{X}_t)^2\Big] \\
&= \mathbb{E}\Big[(f(\boldsymbol{X}_t) - f_0(\boldsymbol{X}_t))^2\Big] \\
&= \|\hat{f}_n - f_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{X}})}.
\end{aligned}$$

The desired result follows because $\mathcal{N}_n \subset \mathcal{L}^2\big(P_{\{\boldsymbol{X}_t\}_{t=1}^n}\big)$ for any n, since $\sup_{f\in\mathcal{N}_n}\|f\|_\infty = B_n$. ■

Verification of (b.4). First, by (2.4) $\sup_{f\in\mathcal{F}_n}\|f\|_\infty \le B_n < \infty$ for each n. Next, by (A.55) and Lemma A.5.4, $\mathrm{Pdim}(\mathcal{N}_n) \asymp n^{2\left(\frac{d}{p+d}\right)(1/2-K_B-\upsilon)}\log^7(n)$. Hence, $\mathrm{Pdim}(\mathcal{N}_n) \gtrsim \log\log(n)$, and

$$\frac{B_n}{\sqrt{n}}\left[\sqrt{\mathrm{Pdim}(\mathcal{F}_n)\log(n)} + \sqrt{\log\log(n)}\right] \lesssim n^{\left(\frac{d}{p+d}\right)(1/2-K_B-\upsilon)-(1/2-K_B)}\log^4(n) \to 0, \quad \text{as} \quad n\to\infty,$$

since $K_B < 1/4$, $\upsilon \ge 0$ and $d/(p+d) \in (0,1)$. ■

Verification of (b.5). Choose $m_n(\boldsymbol{Z}_t) := 2\sup_{f\in\mathcal{N}_n}\left|Y_t - f(\boldsymbol{X}_t)\right|$. Then (b.5)(i) follows from (A.50). For (b.5)(ii) first choose

$$\mu_n := \max\left\{4\mathbb{E}\left[Y_t^2\,\mathbb{1}_{\{|Y_t|\ge B_n\}}\right],\ n^{-1}\right\}.$$

Then by Assumption 2.2.2 and Lemma 2.2.1, we have $\lim_{n\to\infty}\mu_n = 0$. By (A.53) and stationarity, $\mathbb{E}\left[|q_t(f)|\mathbb{1}_{\{m_n(\boldsymbol{Z}_t)\ge 4B_n\}}\right] \le 4\mathbb{E}\left[Y_t^2\,\mathbb{1}_{\{|Y_t|\ge B_n\}}\right] \le \mu_n$. Hence, (b.5)(ii) holds for $C_4 = 4$. ■

Final Steps for Proposition A.5.1. Now, we verify the remaining requirements for Theorem 2.1.2. By Remark 2.2.1(ii) and (iii), $\mathcal{N}_n$ is pointwise-separable, and $\{f(\boldsymbol{x}) : f\in\mathcal{N}_n\} = [-B_n, B_n] \subset \mathbb{R}$ is compact for each $\boldsymbol{x}\in[0,1]^d$. Under Assumption 2.2.1, $\|f_0\|_\infty \le 1$.

Choose

$$\delta := C_\delta\, n^{2\left(\frac{d}{p+d}\right)(1/2-K_B-\upsilon)}\log^8(n), \quad \text{and} \quad a := \left\lceil n^{2\upsilon}\log^2(n)\right\rceil.$$

for some $C_\delta > 0$. To apply Theorem 2.1.2 all that remains is to verify $\sqrt{\delta} \ge \frac{\tilde{\epsilon}_n\sqrt{n}}{B_n a - \tilde{\epsilon}_n(\log n)(\log\log n)}$. Note that by (b.2) $\lim_{n\to\infty}\tilde{\epsilon}_n(\log n)(\log\log n) = 0$ and by (b.4) $B_n a \ge 3$ for all n, so we have $B_n a - \tilde{\epsilon}_n(\log n)(\log\log n) \gtrsim B_n a$. Using this with $\tilde{\epsilon}_n \lesssim n^{-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}$ from the proof for (b.2), and $B_n \asymp n^{-K_B}$ by assumption, we have

$$\begin{aligned}\frac{\tilde{\epsilon}_n\sqrt{n}}{B_n a - \tilde{\epsilon}_n(\log n)(\log\log n)} &\lesssim \frac{n^{1/2-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}}{B_n a} \le \frac{n^{1/2-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}}{B_n n^{\upsilon}} \\ &\asymp n^{(1/2-K_B-\upsilon)-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)} = n^{\left(\frac{d}{p+d}\right)(1/2-K_B-\upsilon)}.\end{aligned}$$

Therefore,

$$\sqrt{\delta} \asymp n^{\left(\frac{d}{p+d}\right)(1/2-K_B-\upsilon)}\log^4(n) > n^{\left(\frac{d}{p+d}\right)(1/2-K_B-\upsilon)} \gtrsim \frac{\tilde{\epsilon}_n\sqrt{n}}{B_n a - \tilde{\epsilon}_n(\log n)(\log\log n)},$$

and the desired result follows by choosing C_δ sufficiently large.

Thus, Theorem 2.1.2 can be applied. To obtain the rate from Proposition A.5.1 note that $\text{Pdim}(\mathcal{N}_n)\log(n) \asymp n^{2\left(\frac{d}{p+d}\right)(1/2-K_B-\upsilon)}\log^8(n) \asymp \delta$, by (A.55) and Lemma A.5.4. Hence,

$$\begin{aligned}
\epsilon_n(\delta, a) &:= B_n\sqrt{\frac{a}{n}}\left[\sqrt{\text{Pdim}(\mathcal{F}_n)\log(n)} + \sqrt{\log\log(n)+\delta}\right] + \sqrt{\tilde{\epsilon}_n^2 + \mu_n + \theta_n} \\
&\lesssim B_n\sqrt{\frac{a}{n}}\, n^{-\left(\frac{d}{p+d}\right)(1/2-K_B-\upsilon)}\log^4(n) + \sqrt{\tilde{\epsilon}_n^2 + \mu_n + \theta_n} \\
&\lesssim n^{-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}\log^5(n) + \sqrt{\mu_n + \theta_n} \\
&\lesssim n^{-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}\log^5(n) + \sqrt{\mathbb{E}\left[Y_t^2\,\mathbb{1}_{\{|Y_t|\geq B_n\}}\right] + \theta_n},
\end{aligned}$$

where the third line has used $\tilde{\epsilon}_n \lesssim n^{-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}$; and the last line follows from $\mu_n := \max\left\{\mathbb{E}\left[Y_t^2\,\mathbb{1}_{\{|Y_t|\geq B_n\}}\right], n^{-1}\right\}$ with $\sqrt{n^{-1}} \leq n^{-\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}$. Then, by Theorem 2.1.2, there exists a constant $C > 0$ independent of n such that for all n sufficiently large

$$\|\hat{f}_n - f_0\|_{\mathcal{L}^2} \leq C\left[n^{\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}\log^5(n) + \sqrt{\mathbb{E}\left[Y_t^2\,\mathbb{1}_{\{|Y_t|\geq B_n\}}\right] + \theta_n}\right]$$

with probability greater than

$$\begin{aligned}
&1-e^{-n^{\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}} - 2\log(n)\left[\frac{n\,C'_\beta e^{-C_\beta\, n^{2\upsilon}\log^2(n)}}{n^{2\upsilon}\log^2(n)} + 2P\Big(\max_{t\in\{1,\ldots,n\}}|Y_t| \geq B_n\Big)\right] \\
&= 1 - e^{-n^{\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}} - 2\log(n)\left[\frac{n\,C'_\beta n^{-C_\beta\, n^{2\upsilon}\log(n)}}{n^{2\upsilon}\log^2(n)} + 2P\Big(\max_{t\in\{1,\ldots,n\}}|Y_t| \geq B_n\Big)\right] \\
&= 1 - e^{-n^{\left(\frac{p}{p+d}\right)(1/2-K_B-\upsilon)}} - \frac{2C'_\beta n^{1-C_\beta n^{2\upsilon}\log(n)-2\upsilon}}{\log(n)} - 4\log(n)P\Big(\max_{t\in\{1,\ldots,n\}}|Y_t| \geq B_n\Big).
\end{aligned}$$

The result for $\|\hat{f}_n - f_0\|_{2,n}$ follows via the same reasoning. ■

A.5.3 Proof of Theorem 2.2.3

Theorem 2.2.3 will follow by showing the conditions for Theorem 2.1.1 hold with the following setting:

- $q_t(f) = q(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t)) := -Y_t B f(\boldsymbol{X}_t) + \log\left(1 + e^{Bf(\boldsymbol{X}_t)}\right)$,
- $\rho_n(f, f') := \|f - f'\|_{\mathcal{L}^2(P_{\{\boldsymbol{X}_t\}_{t=1}^n})} = \left(\int_{[0,1]^d}\left[f(\boldsymbol{x}) - f'(\boldsymbol{x})\right]^2 dP_{\{\boldsymbol{X}_t\}_{t=1}^n}\right)^{1/2}$;
- $C_1 := \frac{1}{2}\left(\frac{1}{e^{B^2}+e^{-B^2}+2}\right)$ and $C_2 := 1/4$;

- $M_n \coloneqq 3B$ for all $n \in \mathbb{N}$; and
- $m_n(\boldsymbol{z}) \coloneqq 2B$, for all $n \in \mathbb{N}$, $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$.

Verification of (a.1). This follows by the same reasoning used in the proof for (a.1) in Appendix A.5.1. ■

Verification of (a.2). For any $f \in \mathcal{N}_n$, by iterated expectations

$$\mathbb{E}[Q_n(f)] - \mathbb{E}[Q_n(f_0)] = \mathbb{E}\left[\frac{e^{f_0(\boldsymbol{X})}}{1+e^{f_0(\boldsymbol{X})}} B\Big(f_0(\boldsymbol{X}_t) - f(\boldsymbol{X}_t)\Big) + \log\left(\frac{1+e^{Bf(\boldsymbol{X}_t)}}{1+e^{Bf_0(\boldsymbol{X}_t)}}\right)\right].$$

Let $g_a(b) \coloneqq -\frac{e^a}{1+e^a}\big(b-a\big) + \log\left(\frac{1+e^b}{1+e^a}\right)$, for arbitrary $a, b \in [-B^2, B^2]$. With this, $g_a(a) = 0$,

$$\frac{d}{db}\, g_a(b) = \frac{e^b}{1+e^b} - \frac{e^a}{1+e^a}, \quad \text{and} \quad \frac{d^2}{db^2}\, g_a(b) = \frac{e^b}{(1+e^b)^2} = \frac{1}{e^b + e^{-b} + 2}.$$

By Taylor's Theorem, with the Lagrange form of the remainder for some $\lambda \in (0,1)$

$$\begin{aligned} g_a(b) &= g_a(a) + (b-a)\left[\frac{d}{dx}\, g_a(x)\right]_{x=a} + \frac{(b-a)^2}{2}\left[\frac{d^2}{dx^2}\, g_a(x)\right]_{x=\lambda a+(1-\lambda)b} \\ &= \frac{(b-a)^2}{2}\left[\frac{1}{e^x + e^{-x} + 2}\right]_{x=\lambda a+(1-\lambda)b}. \end{aligned}$$

Note that $\frac{1}{2}\left(\frac{1}{e^{B^2}+e^{-B^2}+2}\right) \leq \frac{1}{2}\left[\frac{1}{e^x+e^{-x}+2}\right]_{x=\lambda a+(1-\lambda)b} \leq 1/4$ for any $\lambda \in (0,1)$. Clearly for all $\boldsymbol{x} \in [0,1]^d$, $Bf(\boldsymbol{x}) \in [-B^2, B^2]$ and $Bf_0(\boldsymbol{x}) \in [-B, B] \subset [-B^2, B^2]$. Thus, (a.2) holds with $C_1 \coloneqq \frac{1}{2}\left(\frac{1}{e^{B^2}+e^{-B^2}+2}\right)$ and $C_2 \coloneqq 1/4$, since $a, b \in [-B^2, B^2]$ are arbitrary. ■

Verification of (a.3). For any $f, f' \in \mathcal{N}_n$,

$$\begin{aligned} |q_t(f) - q_t(f')| &= \left|Y_t B\Big(f'(\boldsymbol{X}_t) - f(\boldsymbol{X}_t)\Big) + \log\left(\frac{1+e^{Bf(\boldsymbol{X}_t)}}{1+e^{Bf'(\boldsymbol{X}_t)}}\right)\right| \\ &\leq \left|Y_t B\Big(f'(\boldsymbol{X}_t) - f(\boldsymbol{X}_t)\Big)\right| + \left|\log\left(\frac{1+e^{Bf(\boldsymbol{X}_t)}}{1+e^{Bf'(\boldsymbol{X}_t)}}\right)\right| \leq 2B\big|f'(\boldsymbol{X}_t) - f(\boldsymbol{X}_t)\big|, \end{aligned}$$

since $Y_t \in \{0,1\}$ and

$$\begin{aligned} \left|\log\left(\frac{1+e^{Bf(\boldsymbol{X}_t)}}{1+e^{Bf'(\boldsymbol{X}_t)}}\right)\right| &= \log\left(\frac{1+e^{Bf(\boldsymbol{X}_t)}}{1+e^{Bf'(\boldsymbol{X}_t)}}\right)\mathbb{1}_{f(\boldsymbol{X}_t)>f'(\boldsymbol{X}_t)} + \log\left(\frac{1+e^{Bf'(\boldsymbol{X}_t)}}{1+e^{Bf(\boldsymbol{X}_t)}}\right)\mathbb{1}_{f(\boldsymbol{X}_t)<f'(\boldsymbol{X}_t)} \\ &\leq \left|\log\left(\frac{e^{Bf'(\boldsymbol{X}_t)}}{e^{Bf(\boldsymbol{X}_t)}}\right)\right| = B\big|f'(\boldsymbol{X}_t) - f(\boldsymbol{X}_t)\big|. \end{aligned}$$

Thus (a.2)(i) holds for $m_n \coloneqq 2B$. Then, (a.2)(ii) and (iii) hold trivially by setting $M_n = 3B$ for all $n \in \mathbb{N}$. ■

Verification of (a.4). Note that, $\mathbb{1}_{\{m_{nt}<M_n\}} = 1$ for all $n \in \mathbb{N}$, since $m_n = 2$ and $M_n = 3$. For any $f \in \mathcal{N}_n$,

$$|q(\boldsymbol{Z}_t, f(\boldsymbol{Z}_t))| \leq B\big|Y_t f(\boldsymbol{X}_t)\big| + \left|\log\left(1 + e^{Bf(\boldsymbol{X}_t)}\right)\right| \leq B\big|f(\boldsymbol{X}_t)\big| + \left|2\log\left(e^{B^2}\right)\right| \leq 3B^2$$

since $Y_t \in \{0,1\}$, $\|f\|_\infty \leq B$, and $B \geq 2$. Hence, $\big|q_t(f) - \mathbb{E}\big[q_t(f)\big]\big| \leq 6B^2$.

By Lemma A.3.1 $\{q_t(f)\}$ inherits the mixing properties of $\{\boldsymbol{Z}_t\}_{t=1}^n$. Then, by Merlevède et al. (2009, Theorem 1) there exists a constant $C' > 0$ depending only on C_α, C'_α such that for any $\delta > 0$ and all $n \geq 4$,

$$\begin{aligned} P\left(\frac{1}{n}\left|\sum_{t=1}^{n}\Big\{q_t(f) - \mathbb{E}\big[q_t(f)\big]\Big\}\right| \geq \delta\right) &\leq \exp\left[\frac{-C'\,\delta^2\,n^2}{6B^2 n + 6B^2\delta n(\log n)(\log\log n)}\right] \\ &\leq \exp\left[\frac{-C\,\delta\,n}{B^2(\log n)(\log\log n)}\right] =: \lambda_n^{(q)}(\delta), \end{aligned}$$

for some constant $C > 0$ not depending on n or δ since $(\log n)(\log\log n) > 0$ for $n \geq 4$.

Consider (a.4)(ii). Note that

$$\lambda_n^{(q)}\big(\delta\,\epsilon_n^2\big) = \exp\left[\frac{-C\,\delta\,\epsilon_n^2\,n}{B^2(\log n)(\log\log n)}\right].$$

With this, and $\epsilon_n \gtrsim n^{-\frac{1}{2}\left(\frac{p}{p+d}\right)}\log^5(n)$

$$\frac{\epsilon_n^2\,n}{(\log n)(\log\log n)} \gtrsim \frac{n^{1-\frac{1}{2}\left(\frac{p}{p+d}\right)}\log^5(n)}{(\log n)(\log\log n)} \to 0, \quad \text{as} \quad n \to \infty.$$

Then, Lemma A.5.2 can be applied, to obtain the following sufficient condition for (a.4)(ii),

$$\begin{aligned} &\lim_{n\to\infty}\left\{\left[\frac{\epsilon_n^2\,n}{(\log n)(\log\log n)}\right]^{-1} \cdot \log \mathrm{N}_1^{(\infty)}\big(\delta\,\epsilon_n^2/3,\, \mathcal{N}_n,\, n\big)\right\} = 0, \\ \implies \quad &\lim_{n\to\infty}\left\{\lambda_n^{(q)}\big(\delta\,\epsilon_n^2\big)\cdot \mathrm{N}_1^{(\infty)}\big(\delta\,\epsilon_n^2/M_n,\, \mathcal{N}_n,\, n\big)\right\} = 0 \end{aligned} \tag{A.56}$$

since $M_n = 3$ for all n. Henceforth, let n be large enough such that $\epsilon_n < 1$. By (2.7) and Lemma A.5.4, we have $\mathrm{Pdim}(\mathcal{N}_n) \asymp n^{\left(\frac{d}{p+d}\right)}\log^7(n)$, which implies $\lim_{n\to\infty}\{n/\mathrm{Pdim}(\mathcal{N}_n)\} = \infty$. Then, we can apply Lemma A.5.5. with $\eta = \frac{1}{2}\left(\frac{d}{p+d}\right)$, therein, to obtain

$$\log \mathrm{N}_1^{(\infty)}\big(\delta\,\epsilon_n^2/3,\, \mathcal{N}_n,\, n\big) \lesssim n^{\left(\frac{d}{p+d}\right)}\log^7(n)\Big[\log(n) + \log\left(\epsilon_n^{-2}\right)\Big] \lesssim n^{\left(\frac{d}{p+d}\right)}\log^8(n).$$

With this,

$$\left[\frac{\epsilon_n^2\, n}{(\log n)(\log\log n)}\right]^{-1} \log \mathrm{N}_1^{(\infty)}\big(\delta\, \epsilon_n^2/3,\, \mathcal{N}_n,\, n\big) \lesssim n^{\left(\frac{d}{p+d}\right)-1} \epsilon_n^{-2}\, \log^9(n)\, (\log\log n)$$
$$\lesssim n^{\left(\frac{d}{p+d}\right)-1+\left(\frac{p}{p+d}\right)}\, \log^{-1}(n)\, (\log\log n)$$
$$= \log^{-1}(n)\, (\log\log n) \to 0, \quad \text{as} \quad n \to \infty.$$

Thus, the desired result follows from (A.56). ■

A.5.4 Supporting Lemmas

Lemma A.5.1. *Suppose Assumption 2.2.1 holds and let $\mathcal{N}_n = \mathcal{N}(L_n, \boldsymbol{H}_n, B_n)$ be defined as in (2.4). There exists a constant $C_7 > 0$ depending only on d and p, such that for any $\eta \in (0,1)$ and $n \geq 3$, if*

$$L_n \geq \lceil C_7\, \log(n) \rceil, \quad \text{and} \quad \min_{l \in \{1,2,\ldots,L_n\}} H_{n,l} \geq \left\lceil C_7\, n^{\eta} \log^2(n) \right\rceil,$$

then there exists $\tilde{f}_n \in \mathcal{N}_n$ where $\|\tilde{f}_n - f_0\|_\infty \leq n^{-\eta \frac{p}{d}}$.

Proof. By Yarotsky (2017, Theorem 1), for any $\delta \in (0,1)$, there exists a feed-forward ReLU DNN architecture, denoted as $\mathcal{G}$, such that: there exists $g \in \mathcal{G}$ with $\|g - f_0\|_\infty \leq \delta$; and $\mathcal{G}$ has L^* hidden layers, U^* hidden nodes, and W^* parameters, where

$$L^*(\delta) \leq C_8\, \log(e/\delta), \quad \text{and} \quad W^*(\delta), U^*(\delta) \leq C_8\, \delta^{-\frac{d}{p}} \log(e/\delta), \tag{A.57}$$

for a constant C_8 independent of δ, depending only on d and p. By Farrell et al. (2021, Lemma 1),

$$L_n \geq L^*(\delta) \quad \text{and} \quad \min_{l \in \{1,2,\ldots,L_n\}} H_{n,l} \geq L^*(\delta) \cdot W^*(\delta) + U^*(\delta), \quad \Longrightarrow \quad \exists\, \tilde{f}_n \in \mathcal{N}_n,\ \ni g = \tilde{f}_n. \tag{A.58}$$

Note that $g = \tilde{f}_n \in \mathcal{N}_n$ is feasible with (2.4) since $\|g\|_\infty \leq \|f_0\|_\infty + \delta \leq B_n$ follows from $\|f_0\|_\infty \leq 1$, $B_n \geq 2$, and $\|g - f_0\|_\infty \leq \delta < 1$. For $\eta \in (0,1)$ set $\delta = n^{-\eta \frac{p}{d}}$. With this, it follows from (A.57) and (A.58) that if

$$L_n \geq \frac{p}{d}\, C_8\, \log(e\, n) \qquad \text{and} \qquad H_n^{(min)} \geq 2\left(1 \vee \frac{p\, C_8}{d}\right)^2 n^{\eta} \log^2(e\, n),$$

then there exists $\tilde{f}_n \in \mathcal{N}_n$ such that $\|\tilde{f}_n - f_0\|_\infty \leq n^{-\eta \frac{p}{d}}$. ■

Lemma A.5.2. *Let* $\{a_n\}_{n\in\mathbb{N}}$ *and* $\{b_n\}_{n\in\mathbb{N}}$ *be strictly positive sequences. If* $\lim_{n\to\infty} b_n = \infty$. *and* $\lim_{n\to\infty}\{\log(a_n)/b_n\} = 0$, *then* $\lim_{n\to\infty}\{a_n/e^{b_n}\} = 0$.

Proof. Note that

$$\frac{\log(a_n)}{b_n} = \frac{1}{b_n}\log\left(\frac{a_n}{e^{b_n}}\, e^{b_n}\right) = \frac{1}{b_n}\log\left(\frac{a_n}{e^{b_n}}\right) + 1.$$

Then, by assumption,

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

which implies

$$\lim_{n\to\infty}\left\{\frac{1}{b_n}\log\left(\frac{a_n}{e^{b_n}}\right)\right\} = -1.$$

However, $b_n > 0$ for all n, and $b_n \to \infty$, then $\lim_{n\to\infty} 1/b_n = 0$, so it must be the case that[16]

$$\lim_{n\to\infty}\left\{\log\left(\frac{a_n}{e^{b_n}}\right)\right\} = -\infty,$$

hence, $\lim_{n\to\infty}\{a_n/e^{b_n}\} = 0$. ■

Lemma A.5.3. *Let* $\mathcal{N}_n = \mathcal{N}(L_n, \boldsymbol{H}_n, B_n)$ *be defined as in* (2.4) *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* $H_{n,l} \asymp H_n$ *for all* $l \in \mathbb{N}$. *Then* $W_n \asymp H_n^2 L_n$.

Proof. First, consider an MLP architecture with W parameters and L hidden layers that each have H nodes, then

$$W = H^2(L-1) + H(d+L+1) + 1.$$

Therefore, for the architecture $\mathcal{N}_n$ where the number of nodes may vary between layers, since $L_n \geq 1$, it follows that

$$W_n \leq \left(H_n^{(max)}\right)^2(L_n - 1) + H_n^{(max)}(d + L_n + 1) + 1,$$

[16] To see this, let $\{f_n\}_{n\in\mathbb{N}}$, $\{g_n\}_{n\in\mathbb{N}}$ be real-valued sequences such that $\lim_{n\to\infty} f_n g_n = -1$, $\lim_{n\to\infty} f_n = 0$ and $f_n > 0$ for all n. Then, for any $\delta \in (0,1)$ there exists $N \in \mathbb{N}$ such that $|f_m g_m + 1| < \delta$ and $0 < f_m < \delta(1-\delta)$, for all $m \geq N$. With this, $g_m < (\delta - 1)/f_m < 0$, and $(\delta-1)/f_m < (\delta-1)/\big[\delta(1-\delta)\big] = -1/\delta$. This implies, $g_m < -1/\delta$. Since $\delta \in (0,1)$ is arbitrary, this implies $\lim_{n\to\infty} g_n = -\infty$.

where $H_n^{(max)} := \max_{l \in \{1,2,\dots,L_n\}} H_{n,l}$. By assumption $H_n \asymp H_n^{(max)}$, so we have

$$W_n \;\asymp\; H_n^2(L_n - 1) + H_n(d + L_n + 1) + 1 \;\asymp\; H_n^2 L_n + H_n L_n \;\asymp\; H_n^2 L_n.$$

■

Recall the definition of pseudo-dimension from Definition 2.1.4.

Lemma A.5.4. *Let $\mathcal{N}_n = \mathcal{N}(L_n, \boldsymbol{H}_n, B_n)$ be defined as in (2.4) 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, $H_{n,l} \asymp H_n$ for all $l \in \mathbb{N}$, and*

$$L_n \asymp \log(n), \quad H_n \asymp n^\eta \log^2(n), \quad \text{for some } \eta > 0.$$

Then, $\mathrm{Pdim}(\mathcal{N}_n) \asymp n^{2\eta} \log^7(n)$.

Proof. By Bartlett et al. (2019, Theorems 3 and 7),[17] there exist constants $c, C > 0$ such that, for all $n \in \mathbb{N}$,

$$c\, W_n L_n \log(W_n/L_n) \le \mathrm{Pdim}(\mathcal{N}_n) \le C\, W_n L_n \log(W_n).$$

Using this, with $W_n \asymp H_n^2 L_n$ by Lemma A.5.3, and $L_n \asymp \log(n)$, $H_n \asymp n^\eta \log^2(n)$ by assumption, we obtain

$$\begin{aligned}\mathrm{Pdim}(\mathcal{N}_n) \;&\gtrsim\; W_n L_n \log(W_n/L_n) \;\asymp\; H_n^2 L_n^2 \log(H_n) \;\asymp\; n^{2\eta} \log^6(n) \log\Big(n^\eta \log^2(n)\Big) \\ &\asymp\; n^{2\eta} \log^6(n) \Big(\log(n) + \log\log(n)\Big) \;\asymp\; n^{2\eta} \log^7(n),\end{aligned}$$

and

$$\begin{aligned}\mathrm{Pdim}(\mathcal{N}_n) \;&\lesssim\; W_n L_n \log(W_n) \;\asymp\; H_n^2 L_n^2 \log(H_n^2 L_n) \;\asymp\; n^{2\eta} \log^6(n) \log\Big(n^{2\eta} \log^5(n)\Big) \\ &\asymp\; n^{2\eta} \log^6(n) \Big(\log(n) + \log\log(n)\Big) \;\asymp\; n^{2\eta} \log^7(n).\end{aligned}$$

■

[17] These bounds are written explicitly in display (2) of Bartlett et al. (2019). Display (2) uses the Vapnik-Chervonenkis dimension instead of Pseudo-Dimension, however, these are equivalent for function classes generated by a neural network with fixed architecture and fixed activation functions. For details see the discussion following Bartlett et al. (2019, Definition 2), and Anthony and Bartlett (1999, Theorem 14.1).

Lemma A.5.5. *Let $\mathcal{N}_n = \mathcal{N}(L_n, \boldsymbol{H}_n, B_n)$ be defined as in (2.4), where $\{B_n\}_{n\in\mathbb{N}}$ is non decreasing, $B_1 \geq 2$ and $B_n \lesssim n^{K_B}$ for some $K_B > 0$. and the sequences $\{L_n\}_{n\in\mathbb{N}}$, $\{H_{n,l}\}_{n\in\mathbb{N}}$ for each $l \in \mathbb{N}$, are non-decreasing, $H_{n,l} \asymp H_n$ for all $l \in \mathbb{N}$, and*

$$L_n \asymp \log(n), \quad H_n \asymp n^{\eta}\log^2(n), \quad \text{for some } \eta > 0.$$

Let $\{\delta_n\}_{n\in\mathbb{N}}$ be a positive sequence such that $\delta_n \leq 1$ for all n sufficiently large. Let $\{a_n\}_{n\in\mathbb{N}}$ be such that $a_n \in \mathbb{N}$ for all n, and $a_n \geq \mathrm{Pdim}(\mathcal{N}_n)$ for all n sufficiently large. Then, for any $r \in [1,\infty]$

$$\log \mathrm{N}_r^{(\infty)}\big(\delta_n, \mathcal{N}_n, a_n\big) \lesssim n^{2\eta}\log^7(n)\Big[\log(n) + \log(a_n) + \log\big(\delta_n^{-1}\big)\Big]$$

Proof. By assumption, $a_n \geq \mathrm{Pdim}(\mathcal{N}_n)$ for all n sufficiently large. Hence, Lemma A.3.4 can be applied to obtain, for some $C > 0$,

$$\mathrm{N}_\infty^{(\infty)}\big(\delta_n, \mathcal{N}_n, a_n\big) \lesssim \left(\frac{2eB_na_n}{\delta_n \cdot \mathrm{Pdim}(\mathcal{N}_n)}\right)^{\mathrm{Pdim}(\mathcal{N}_n)} \leq \left(\frac{2eB_na_n}{\delta_n \cdot Cn^{2\eta}\log^7(n)}\right)^{Cn^{2\eta}\log^7(n)}, \tag{A.59}$$

where the last bound follows from $\mathrm{Pdim}(\mathcal{N}_n) \lesssim n^{2\eta}\log^7(n)$ by Lemma A.5.4, with

$$\frac{2eB_na_n}{\delta_n \cdot x} > e \quad \Longrightarrow \quad \frac{\partial}{\partial x}\left[\left(\frac{2eB_na_n}{\delta_n \cdot x}\right)^x\right] = \left(\log\left(\frac{2eB_na_n}{\delta_n \cdot x}\right) - 1\right)\left(\frac{2eB_na_n}{\delta_n \cdot x}\right)^x > 0,$$

and

$$\frac{2eB_na_n}{\delta_n \cdot \mathrm{Pdim}(\mathcal{N}_n)} > e, \quad \forall\, n \text{ sufficiently large},$$

since B_n is non-decreasing, $\delta_n \leq 1$ for all n sufficiently large, and $\lim_{n\to\infty} a_n/\mathrm{Pdim}(\mathcal{N}_n) = \infty$. By (A.59)

$$\begin{aligned}
\log\ \mathrm{N}_\infty^{(\infty)}\big(\delta_n, \mathcal{N}_n, a_n\big) &\lesssim n^{2\eta}\log^7(n) \cdot \log\left(\frac{2eB_na_n}{\delta_n \cdot n^{2\eta}\log^7(n)}\right) \\
&\lesssim n^{2\eta}\log^7(n)\Big[\log(B_n) + \log(a_n) + \log\big(\delta_n^{-1}\big) - (2\eta d/p)\log(n)\Big] \\
&\lesssim n^{2\eta}\log^7(n)\Big[\log(n) + \log(a_n) + \log\big(\delta_n^{-1}\big)\Big],
\end{aligned}$$

since $B_n \lesssim n^{K_B}$ by assumption. ■

Appendix B

Appendix for Chapter 3

This appendix provides the proof of Theorem 3.1.1. First, Appendix B.1 presents three preliminary lemmas that are analogous to the conditions from Chen et al. (2022, Theorem 1) and are labeled accordingly. Then, Appendix B.2 uses these to prove Theorem 3.1.1.

Throughout this section, we write

$$\mathsf{v}_t(w,s) \coloneqq \mathsf{v}(\boldsymbol{Z}_t; w, s), \quad \text{and} \quad \mathsf{A}_t(w) \coloneqq \mathsf{A}(\boldsymbol{Z}_t; w).$$

B.1 Preliminary Lemmas

Lemma B.1.1. (Stable Estimator)

$$\frac{1}{n}\sum_{t=1}^{n} \psi_t(\hat{\zeta}, \hat{w}_n, \hat{s}_n) = o_P(n^{-1/2})$$

Proof. Using the definition of $\hat{\zeta}$, whenever $\sum_{t=1}^{n} \mathsf{A}(\boldsymbol{Z}_t; \hat{w}_n) \neq 0$,

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} \psi_t(\hat{\zeta}, \hat{w}_n, \hat{s}_n) &= \frac{1}{n}\sum_{t=1}^{n} \mathsf{A}(\boldsymbol{Z}_t; \hat{w}_n)\,\hat{\zeta} - \frac{1}{n}\sum_{t=1}^{n} \mathsf{v}(\boldsymbol{Z}_t; \hat{w}_n, \hat{s}_n) \\
&= \left[\frac{1}{n}\sum_{t=1}^{n} \mathsf{A}(\boldsymbol{Z}_t; \hat{w}_n)\right] \left[\frac{1}{n}\sum_{t=1}^{n} \mathsf{A}(\boldsymbol{Z}_t; \hat{w}_n)\right]^{-1} \left[\frac{1}{n}\sum_{t=1}^{n} \mathsf{v}(\boldsymbol{Z}_t; \hat{w}_n, \hat{s}_n)\right] - \frac{1}{n}\sum_{t=1}^{n} \mathsf{v}(\boldsymbol{Z}_t; \hat{w}_n, \hat{s}_n) \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathsf{v}(\boldsymbol{Z}_t; \hat{w}_n, \hat{s}_n) - \frac{1}{n}\sum_{t=1}^{n} \mathsf{v}(\boldsymbol{Z}_t; \hat{w}_n, \hat{s}_n) = 0.
\end{aligned}$$

This completes the proof since $\sum_{t=1}^{n} \mathsf{A}(\boldsymbol{Z}_t; \hat{w}_n) \neq 0$ with probability approaching one under Assumption 3.1.1(ii). ■

Lemma B.1.2. (Stochastic Equicontinuity)

$$\sqrt{n}\left|\mathbb{E}\Big[\mathsf{A}_t\big(\hat{w}_n\big)-\mathsf{A}_t\big(w_0\big)\Big]-\frac{1}{n}\sum_{t=1}^{n}\Big\{\mathsf{A}_t\big(\hat{w}_n\big)-\mathsf{A}_t\big(w_0\big)\Big\}\right|=\mathrm{o}_P(1)$$

$$\sqrt{n}\left|\mathbb{E}\Big[\mathsf{v}_t\big(\hat{w}_n,\hat{s}_n\big)-\mathsf{v}_t\big(w_0,s_0\big)\Big]-\frac{1}{n}\sum_{t=1}^{n}\Big\{\mathsf{v}_t\big(\hat{w}_n,\hat{s}_n\big)-\mathsf{v}_t\big(w_0,s_0\big)\Big\}\right|=\mathrm{o}_P(1)$$

Proof. This result will follow using the ideas from Appendix A.3. Note that v depends on Y_t and $s\in\mathcal{N}_n^{(s)}$ which are not uniformly bounded. Therefore, we will show the result for v, and the result for A will follow via a similar, but somewhat simpler argument.

Let $\mathbb{1}_{nt}\coloneqq\mathbb{1}\{|Y_t|\leq B_n\}$, and $\mathbb{1}_{nt}^{\mathrm{c}}\coloneqq\mathbb{1}\{|Y_t|>B_n\}$, then define

$$g_{w,s}(\boldsymbol{Z}_t)\coloneqq\big[\mathsf{v}_t\big(w,s\big)-\mathsf{v}_t\big(w_0,s_0\big)\big]\mathbb{1}_{nt},\quad\text{and}\quad g_{w,s}^{\mathrm{c}}(\boldsymbol{Z}_t)\coloneqq\big[\mathsf{v}_t\big(w,s\big)-\mathsf{v}_t\big(w_0,s_0\big)\big]\mathbb{1}_{nt}^{\mathrm{c}}.$$

With this, and the triangle inequality,

$$\begin{aligned}&\sqrt{n}\left|\frac{1}{n}\sum_{t=1}^{n}\Big\{\mathbb{E}\big[\mathsf{v}_t\big(\hat{w}_n,\hat{s}_n\big)-\mathsf{v}_t\big(w_0,s_0\big)\big]-\big[\mathsf{v}_t\big(\hat{w}_n,\hat{s}_n\big)-\mathsf{v}_t\big(w_0,s_0\big)\big]\Big\}\right|\\&\leq\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\Big\{\mathbb{E}\big[g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]-\big[g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]\Big\}\right|+\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\Big\{\mathbb{E}\big[g_{\hat{w}_n,\hat{s}_n}^{\mathrm{c}}(\boldsymbol{Z}_t)\big]-\big[g_{\hat{w}_n,\hat{s}_n}^{\mathrm{c}}(\boldsymbol{Z}_t)\big]\Big\}\right|\end{aligned}\tag{B.1}$$

Consider the second term of (B.1). By stationarity, for any $\delta>0$

$$\begin{aligned}&P\left(\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\Big\{\mathbb{E}\big[g_{\hat{w}_n,\hat{s}_n}^{\mathrm{c}}(\boldsymbol{Z}_t)\big]-\big[g_{\hat{w}_n,\hat{s}_n}^{\mathrm{c}}(\boldsymbol{Z}_t)\big]\Big\}\right|\geq\delta\right)\\&\qquad\leq P\Big(\sqrt{n}\,\mathbb{E}\big[\,\big|g_{\hat{w}_n,\hat{s}_n}^{\mathrm{c}}(\boldsymbol{Z}_t)\big|\,\big]\geq\delta\Big)+P\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\big|g_{w,s}^{\mathrm{c}}(\boldsymbol{Z}_t)\big|>0\right)\end{aligned}$$

First, note that $|D_t-\hat{w}_n(\boldsymbol{X})|\leq 3$, $|D_t-w_0(\boldsymbol{X})|\leq 2<3$, $\|s\|_\infty\leq B_n$, and $\|s_0\|_\infty\leq 1<B_n$. Hence,

$$\begin{aligned}\mathbb{E}\big[\,\big|g_{\hat{w}_n,\hat{s}_n}^{\mathrm{c}}(\boldsymbol{Z}_t)\big|\,\big]&\leq 3\mathbb{E}\Big[\big|Y_t-\hat{s}_n(\boldsymbol{X}_t)\big|\mathbb{1}_{nt}^{\mathrm{c}}+\big|Y_t-s_0(\boldsymbol{X}_t)\big|\mathbb{1}_{nt}^{\mathrm{c}}\Big]\\&\leq 6\mathbb{E}\Big[\big(|Y_t|+B_n\big)\mathbb{1}_{nt}^{\mathrm{c}}\Big]\leq 12\mathbb{E}\big[|Y_t|\mathbb{1}_{nt}^{\mathrm{c}}\big]=\mathrm{o}(\sqrt{n}),\end{aligned}$$

where the third inequality uses $\big(|Y_t|+B_n\big)\mathbb{1}_{nt}^{\mathrm{c}}<2|Y_t|$ if $|Y_t|>B_n$, and $\big(|Y_t|+B_n\big)\mathbb{1}_{nt}^{\mathrm{c}}=0$ else; and the last equality follows from Assumption 3.1.1(iv) with Jensen's inequality. Second, for any $w\in\mathcal{N}_n^{(w)}$, $s\in\mathcal{N}_n^{(s)}$

$$\begin{aligned}P\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\big|g_{w,s}^{\mathrm{c}}(\boldsymbol{Z}_t)\big|>0\right)&=P\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\big|\mathsf{v}_t\big(w,s\big)-\mathsf{v}_t\big(w_0,s_0\big)\big|\mathbb{1}_{nt}^{\mathrm{c}}>0\right)\\&\leq P\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\mathbb{1}_{nt}^{\mathrm{c}}>0\right)=P\Big(\max_{t\in\{1,\ldots,n\}}|Y_t|>B_n\Big)\to 0,\ \text{ as } n\to\infty,\end{aligned}$$

by Assumption 3.1.1(iv). Combining the previous three displays

$$P\left(\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left\{\mathbb{E}\big[g^{\mathrm{c}}_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]-\big[g^{\mathrm{c}}_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]\right\}\right|\geq\delta\right)=\mathrm{o}(1). \tag{B.2}$$

Thus, all that remains is to address the first term of (B.1). We do this using similar techniques as Appendix A.3.5.2. Choose $a \coloneqq \lceil 2\log(n)\rceil$, so $1 \leq a \leq n/2$, and $b \coloneqq \lfloor n/(2a)\rfloor$ is well defined. Then, construct the random sequence $\{\overline{\boldsymbol{Z}}_t\}_{t=1}^n$ with the procedure described in Appendix A.3.4 by dividing $\{\boldsymbol{Z}_t\}_{t=1}^n$ into $2b$ blocks of length a, and the remainder into a block of length $n-2ba$, using the index sets T_R, $T_{1,j}$, $T_{2,j}$, for $j=1,\ldots,b$, defined therein. For $m\in\{1,2\}$, and $j\in\{1,...,b\}$, define

$$G^{(m)}_{j,w,s}\coloneqq\frac{1}{a}\sum_{t\in T_{m,j}}g_{w,s}(\boldsymbol{Z}_t),\quad\text{and}\quad\overline{G}^{(m)}_{j,w,s}\coloneqq\frac{1}{a}\sum_{t\in T_{m,j}}g_{w,s}(\overline{\boldsymbol{Z}}_t).$$

With this, the first term of (B.1) can be written as

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left\{\mathbb{E}\big[g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]-g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\right\}\\&\quad=\left(\frac{ab}{\sqrt{n}}\right)\frac{1}{b}\sum_{j=1}^{b}\left\{\mathbb{E}\big[G^{(1)}_{j,\hat{w}_n,\hat{s}_n}\big]-G^{(1)}_{j,\hat{w}_n,\hat{s}_n}+\mathbb{E}\big[G^{(2)}_{j,\hat{w}_n,\hat{s}_n}\big]-G^{(2)}_{j,\hat{w}_n,\hat{s}_n}\right\}+\frac{1}{\sqrt{n}}\sum_{t\in T_R}\left\{\mathbb{E}\big[g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]-g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\right\}.\end{aligned}$$

Then, by stationarity, for any $\delta>0$

$$\begin{aligned}&P\left(\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left\{\mathbb{E}\big[g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]-g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\right\}\right|\geq 3\delta\right)\\&\leq 2P\left(\left|\left(\frac{ab}{\sqrt{n}}\right)\frac{1}{b}\sum_{j=1}^{b}\left\{\mathbb{E}\big[G^{(1)}_{j,\hat{w}_n,\hat{s}_n}\big]-G^{(1)}_{j,\hat{w}_n,\hat{s}_n}\right\}\right|\geq\delta\right)+P\left(\left|\frac{1}{\sqrt{n}}\sum_{t\in T_R}\left\{\mathbb{E}\big[g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]-g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\right\}\right|\geq\delta\right)\\&\coloneqq 2P_1+P_2.\end{aligned} \tag{B.3}$$

Consider P_2. First, for any $(w,s)\in\mathcal{N}_n^{(w)}\times\mathcal{N}_n^{(s)}$ and all $\boldsymbol{Z}_t$

$$\big|g_{w,s}(\boldsymbol{Z}_t)\big|\leq 3\Big[\big|Y_t-s(\boldsymbol{X}_t)\big|\mathbb{1}_{nt}+\big|Y_t-s_0(\boldsymbol{X}_t)\big|\mathbb{1}_{nt}\Big]\leq 6\big(|Y_t|+B_n\big)\mathbb{1}_{nt}\leq 12B_n,$$

where the first inequality used $|D_t-w(\boldsymbol{X})|\leq 3$ and $|D_t-w_0(\boldsymbol{X})|\leq 2<3$; then the second inequality used $\|s\|_\infty\leq B_n$, and $\|s_0\|_\infty\leq 1<B_n$. With this $\big\|\mathbb{E}\big[g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]-g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big\|_\infty\leq 24B_n$ The cardinality of T_R is $(\#T_R)=n-2ab<2a$, since $b\coloneqq\lfloor n/(2a)\rfloor$ implies $b>n/(2a)-1$. Hence,

$$\left|\frac{1}{\sqrt{n}}\sum_{t\in T_R}\left\{\mathbb{E}\big[g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\big]-g_{\hat{w}_n,\hat{s}_n}(\boldsymbol{Z}_t)\right\}\right|\leq\frac{48a}{\sqrt{n}}=\mathrm{o}(1),$$

as n grows since $a \coloneqq \lceil 2\log(n) \rceil$. Thus, for any $\delta > 0$

$$P_2 = P\left(\left| \frac{1}{\sqrt{n}} \sum_{t \in T_R} \left\{ \mathbb{E}\big[g_{\hat{w}_n, \hat{s}_n}(\boldsymbol{Z}_t) \big] - g_{\hat{w}_n, \hat{s}_n}(\boldsymbol{Z}_t) \right\} \right| \geq \delta \right) = \mathrm{o}(1). \tag{B.4}$$

Now, we address P_1. For ϵ_n as in Corollary 3.1.1, and $\|\cdot\|_{\overline{T}_1}$ as in (A.22) define

$$\begin{aligned} \mathcal{N}_n^{(w,s)}(\epsilon_n) \coloneqq \Big\{ (w,s) \in \mathcal{N}_n^{(w)} \times \mathcal{N}_n^{(s)} \,:\, & \|w - w_0\|_{\mathcal{L}^2(P_{\boldsymbol{X}})} \leq C\epsilon_n,\ \|s - s_0\|_{\mathcal{L}^2(P_{\boldsymbol{X}})} \leq C\epsilon_n, \\ & \|w - w_0\|_{\overline{T}_1} \leq C\epsilon_n,\ \|s - s_0\|_{\overline{T}_1} \leq C\epsilon_n \Big\} \end{aligned}$$

for some $C > 0$ sufficiently large such that $P\big((\hat{w}_n, \hat{s}_n) \notin \mathcal{N}_n^{(w,s)}(\epsilon_n)\big) = \mathrm{o}(1)$.[1] Recall $b \leq n/(2a)$, which implies $\left(\frac{ab}{\sqrt{n}}\right) \leq \left(\frac{\sqrt{n}}{2}\right)$. Then, apply (A.21) with

$$E = \left\{ \left| \left(\frac{\sqrt{n}}{2}\right) \frac{1}{b} \sum_{j=1}^{b} \left\{ \mathbb{E}\big[G^{(1)}_{j,\hat{w}_n,\hat{s}_n} \big] - G^{(1)}_{j,\hat{w}_n,\hat{s}_n} \right\} \right| \geq \delta \right\},$$

to obtain

$$\begin{aligned} P_1 &\leq P\left(\sup_{(w,s) \in \mathcal{N}_n^{(w,s)}(\epsilon_n)} \left| \left(\frac{\sqrt{n}}{2}\right) \frac{1}{b} \sum_{j=1}^{b} \left\{ \mathbb{E}\big[G^{(1)}_{j,w,s} \big] - G^{(1)}_{j,w,s} \right\} \right| \geq \delta \right) + P\Big((\hat{w}_n, \hat{s}_n) \notin \mathcal{N}_n^{(w,s)}(\epsilon_n) \Big) \\ &\leq P\left(\sup_{(w,s) \in \mathcal{N}_n^{(w,s)}(\epsilon_n)} \left| \left(\frac{\sqrt{n}}{2}\right) \frac{1}{b} \sum_{j=1}^{b} \left\{ \mathbb{E}\big[\overline{G}^{(1)}_{j,w,s} \big] - \overline{G}^{(1)}_{j,w,s} \right\} \right| \geq \delta \right) + \frac{n\,\beta(a)}{2a} + \mathrm{o}(1) \\ &\leq P\left(\sup_{(w,s) \in \mathcal{N}_n^{(w,s)}(\epsilon_n)} \left| \left(\frac{\sqrt{n}}{2}\right) \frac{1}{b} \sum_{j=1}^{b} \left\{ \mathbb{E}\big[\overline{G}^{(1)}_{j,w,s} \big] - \overline{G}^{(1)}_{j,w,s} \right\} \right| \geq \delta \right) + \mathrm{o}(1), \end{aligned} \tag{B.5}$$

by Assumption 3.1.1(i) and $a \coloneqq \lceil 2\log(n) \rceil$ imply $\frac{n\,\beta(a)}{2a} = \mathrm{o}(1)$.[2] We bound the first term on the right side with Bartlett et al. (2005, Theorem 2.1). Recall $\big| g_{w,s}(\boldsymbol{Z}_t) \big| \leq 12 B_n$ so we have

$$\max_{j \in \{1,\ldots,b\}} \left(\frac{\sqrt{n}}{2}\right) \left| \mathbb{E}\big[\overline{G}^{(1)}_{j,w,s} \big] - \overline{G}^{(1)}_{j,w,s} \right| \leq \sqrt{n} \big\| g_{w,s}(\boldsymbol{Z}_t) \big\|_{\infty} \leq 12\sqrt{n} B_n.$$

[1] The existence of such a C sufficiently large follows from Corollary 3.1.1 and the discussion preceding (A.45) in Appendix A.3.6. To see that the reasoning used in Appendix A.3.6 applies here note that the conditions for (A.28) are met, since $\epsilon_n \gtrsim r_*$ for r_* defined as in (A.40) and some appropriate choice of δ therein such that $\delta \to \infty$ as $n \to \infty$.

[2] To see this holds for any $a \gtrsim \log(n)$ let $\delta \coloneqq a/\log(n)$, so $a = \delta \log(n)$. Then, by Assumption 3.1.1(i),

$$\frac{n\,\beta(a)}{a} \leq \frac{n\,C'_\beta e^{-C_\beta\, a}}{a} = \big(C'_\beta\, e^{C_\beta}\big) \frac{n\, e^{-\delta \log(n)}}{a} = \big(C'_\beta\, e^{C_\beta}\big) \frac{n^{1-\delta}}{a} = \big(C'_\beta\, e^{C_\beta}\big) \frac{n^{1 - a/\log(n)}}{a}.$$

Recall from Appendix A.3.4 that $P_{\{\boldsymbol{Z}_t\}_{t\in T_{1,j}}} = P_{\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,j}}}$, $\forall j$, so for any $(w,s)\in\mathcal{N}_n^{(w,s)}(\epsilon_n)$,

$$\begin{aligned}
\mathrm{Var}\left[\left(\frac{\sqrt{n}}{2}\right)\overline{G}^{(1)}_{j,w,s}\right] &\le \left(\frac{\sqrt{n}}{2}\right)^2\mathbb{E}\left[\left(\frac{1}{a}\sum_{t\in T_{1,j}} g_{w,s}(\boldsymbol{Z}_t)\right)^2\right] \le \left(\frac{\sqrt{n}}{2}\right)^2\mathbb{E}\left[\frac{1}{a}\sum_{t\in T_{1,j}} g_{w,s}(\boldsymbol{Z}_t)^2\right] \\
&= \left(\frac{9n}{4}\right)\mathbb{E}\left[\Big[\big(D_t - w(\boldsymbol{X}_t)\big)\big(Y_t - s(\boldsymbol{X}_t)\big) - \big(D_t - w_0(\boldsymbol{X}_t)\big)\big(Y_t - s_0(\boldsymbol{X}_t)\big)\Big]^2\mathbb{1}_{nt}\right] \\
&= \left(\frac{9n}{4}\right)\mathbb{E}\Bigg[\Big[\big(D_t - w(\boldsymbol{X}_t)\big)\big(Y_t - s(\boldsymbol{X}_t)\big) - \big(D_t - w(\boldsymbol{X}_t)\big)\big(Y_t - s_0(\boldsymbol{X}_t)\big) \\
&\qquad + \big(D_t - w(\boldsymbol{X}_t)\big)\big(Y_t - s_0(\boldsymbol{X}_t)\big) - \big(D_t - w_0(\boldsymbol{X}_t)\big)\big(Y_t - s_0(\boldsymbol{X}_t)\big)\Big]^2\mathbb{1}_{nt}\Bigg] \\
&= \left(\frac{9n}{4}\right)\mathbb{E}\left[\Big[\big(D_t - w(\boldsymbol{X}_t)\big)\big(s_0(\boldsymbol{X}_t) - s(\boldsymbol{X}_t)\big) + \big(w_0(\boldsymbol{X}_t) - w(\boldsymbol{X}_t)\big)\big(Y_t - s_0(\boldsymbol{X}_t)\big)\Big]^2\mathbb{1}_{nt}\right] \\
&\le \left(\frac{9n}{4}\right)\mathbb{E}\left[\Big[3\big|s_0(\boldsymbol{X}_t) - s(\boldsymbol{X}_t)\big| + 2B_n\big|w_0(\boldsymbol{X}_t) - w(\boldsymbol{X}_t)\big|\Big]^2\right],
\end{aligned}$$

where the second inequality used the Cauchy-Schwarz inequality, the first equality used stationarity, and the last inequality used $|D_t - w(\boldsymbol{X}_t)| \le 3$, and $|Y_t - s_0(\boldsymbol{X}_t)|\mathbb{1}_{nt} \le B_n + 1 < 2B_n$. Note that $3 < 2B_n$, so we have

$$\begin{aligned}
\mathrm{Var}\left[\left(\frac{\sqrt{n}}{2}\right)\overline{G}^{(1)}_{j,w,s}\right] &\le \big(9B_n^2 n\big)\,\mathbb{E}\left[\Big[\big|s_0(\boldsymbol{X}_t) - s(\boldsymbol{X}_t)\big| + \big|w_0(\boldsymbol{X}_t) - w(\boldsymbol{X}_t)\big|\Big]^2\right] \\
&= \big(9B_n^2 n\big)\,\mathbb{E}\Big[\big|s_0(\boldsymbol{X}_t) - s(\boldsymbol{X}_t)\big|^2 + \big|w_0(\boldsymbol{X}_t) - w(\boldsymbol{X}_t)\big|^2 + 2\big|s_0(\boldsymbol{X}_t) - s(\boldsymbol{X}_t)\big|\big|w_0(\boldsymbol{X}_t) - w(\boldsymbol{X}_t)\big|\Big] \\
&= \big(9B_n^2 n\big)\left[\|s_0 - s\|^2_{\mathcal{L}^2(P_{\boldsymbol{X}})} + \|w_0 - w\|^2_{\mathcal{L}^2(P_{\boldsymbol{X}})} + 2\big\|(w_0 - w)(s_0 - s)\big\|_{\mathcal{L}^1(P_{\boldsymbol{X}})}\right] \\
&\le \big(9B_n^2 n\big)\left[\|s_0 - s\|^2_{\mathcal{L}^2(P_{\boldsymbol{X}})} + \|w_0 - w\|^2_{\mathcal{L}^2(P_{\boldsymbol{X}})} + 2\|w_0 - w\|_{\mathcal{L}^2(P_{\boldsymbol{X}})}\|s_0 - s\|_{\mathcal{L}^2(P_{\boldsymbol{X}})}\right] \\
&\le 36B_n^2\, n\,\epsilon_n^2,
\end{aligned}$$

by Hölder's inequality. With this, since $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,1}}$, $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,2}}$, $\ldots$, $\{\overline{\boldsymbol{Z}}_t\}_{t\in T_{1,b}}$ is an i.i.d. sequence, we can apply Bartlett et al. (2005, Theorem 2.1) (with $\alpha = 1/2$, and $x = \log(n)$ therein) to obtain,

$$\begin{aligned}
e^{-\log(n)} \ge P\Bigg(\left(\frac{\sqrt{n}}{2}\right)\frac{1}{b}\sum_{j=1}^{b}\Big\{\mathbb{E}\big[\overline{G}^{(1)}_{j,\hat{w}_n,\hat{s}_n}\big] - \overline{G}^{(1)}_{j,\hat{w}_n,\hat{s}_n}\Big\} \ge \left(\frac{\sqrt{n}}{2}\right)6\mathbb{E}_{P^{(\xi)}}\Big[\mathfrak{R}_b\Big\{\overline{G}^{(1)}_{j,w,s} : (w,s)\in\mathcal{N}_n^{(w,s)}(\epsilon_n)\Big\}\Big] \\
+ 6B_n\epsilon_n\sqrt{\frac{2n\log(n)}{b}} + \frac{128\sqrt{n}B_n\log(n)}{b}\Bigg).
\end{aligned} \tag{B.6}$$

Now we show each of the terms in the probability bound converge to zero as n grows. First,

$b \geq n/(4a) = n/(4\log(n))$ for n sufficiently large, so we have

$$6B_n\epsilon_n\sqrt{\frac{2n\log(n)}{b}} + \frac{128\sqrt{n}B_n\log(n)}{b} \lesssim B_n\epsilon_n\log(n) + \frac{B_n\log^2(n)}{\sqrt{n}} = \mathrm{o}(1). \tag{B.7}$$

Now we bound the Rademacher complexity term. For any $(w,s) \in \mathcal{N}_n^{(w)} \times \mathcal{N}_n^{(s)}$ and $(w',s') \in \mathcal{N}_n^{(w)} \times \mathcal{N}_n^{(s)}$,

$$\begin{aligned}
\left|\overline{G}^{(1)}_{j,w,s} - \overline{G}^{(1)}_{j,w',s'}\right| &\leq \frac{1}{a}\sum_{t\in T_{1,j}} \Big|\big(D_t - w(\boldsymbol{X}_t)\big)\big(Y_t - s(\boldsymbol{X}_t)\big) - \big(D_t - w'(\boldsymbol{X}_t)\big)\big(Y_t - s'(\boldsymbol{X}_t)\big)\Big|\mathbb{1}_{nt} \\
&\leq \frac{1}{a}\sum_{t\in T_{1,j}} \Big|\big(D_t - w(\boldsymbol{X}_t)\big)\big(Y_t - s(\boldsymbol{X}_t)\big) - \big(D_t - w(\boldsymbol{X}_t)\big)\big(Y_t - s'(\boldsymbol{X}_t)\big) \\
&\qquad + \big(D_t - w(\boldsymbol{X}_t)\big)\big(Y_t - s'(\boldsymbol{X}_t)\big) - \big(D_t - w'(\boldsymbol{X}_t)\big)\big(Y_t - s'(\boldsymbol{X}_t)\big)\Big|\mathbb{1}_{nt} \\
&= \frac{1}{a}\sum_{t\in T_{1,j}} \Big|\big(D_t - w(\boldsymbol{X}_t)\big)\big(s(\boldsymbol{X}_t)' - s(\boldsymbol{X}_t)\big) + \big(w'(\boldsymbol{X}_t) - w(\boldsymbol{X}_t)\big)\big(Y_t - s'(\boldsymbol{X}_t)\big)\Big|\mathbb{1}_{nt} \\
&\leq \frac{1}{a}\sum_{t\in T_{1,j}} \Big\{3\big|s(\boldsymbol{X}_t)' - s(\boldsymbol{X}_t)\big| + 2B_n\big|w'(\boldsymbol{X}_t) - w(\boldsymbol{X}_t)\big|\Big\} \\
&\leq 2B_n\frac{1}{a}\sum_{t\in T_{1,j}} \Big\{\big|s(\boldsymbol{X}_t)' - s(\boldsymbol{X}_t)\big| + \big|w'(\boldsymbol{X}_t) - w(\boldsymbol{X}_t)\big|\Big\} \\
&\leq 2B_n\sqrt{\frac{1}{a}\sum_{t\in T_{1,j}} \Big\{\big|s(\boldsymbol{X}_t)' - s(\boldsymbol{X}_t)\big|^2 + \big|w'(\boldsymbol{X}_t) - w(\boldsymbol{X}_t)\big|^2\Big\}} \\
&= \frac{2B_n}{\sqrt{a}}\Bigg[\sum_{t\in T_{1,j}} \Big\{\big|\big(s(\boldsymbol{X}_t)' - s_0(\boldsymbol{X}_t)\big) - \big(s(\boldsymbol{X}_t) - s_0(\boldsymbol{X}_t)\big)\big|^2 \\
&\qquad + \big|\big(w(\boldsymbol{X}_t)' - w_0(\boldsymbol{X}_t)\big) - \big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\big|^2\Big\}\Bigg]^{1/2},
\end{aligned}$$

where the fourth inequality used $|D_t - w(\boldsymbol{X}_t)| \leq 3$, and $|Y_t - s_0(\boldsymbol{X}_t)|\mathbb{1}_{nt} \leq B_n + 1 < 2B_n$; then the fifth inequality used $2B_n > 3$; and the last inequality follows from the Cauchy-Schwarz inequality

as in (A.32). Now we can apply Maurer (2016, Theorem 3) (as in the proof of Lemma A.3.3)

$$\begin{aligned}
\mathbb{E}_{P(\xi)}\left[\mathfrak{R}_b\left\{\overline{G}^{(1)}_{j,w,s} : (w,s) \in \mathcal{N}_n^{(w,s)}(\epsilon_n)\right\}\right] &= \mathbb{E}_{P(\xi)}\left[\sup_{(w,s)\in\mathcal{N}_n^{(w,s)}(\epsilon_n)} \frac{1}{b}\sum_{j=1}^{b} \xi_j \overline{G}^{(1)}_{j,w,s}\right] \\
&\leq \frac{2B_n}{\sqrt{a}\,b}\,\mathbb{E}_{P(\xi)}\left[\sup_{(w,s)\in\mathcal{N}_n^{(w,s)}(\epsilon_n)} \left\{\sum_{j=1}^{b}\sum_{t\in T_{1,j}} \xi_t\big(w(\overline{\boldsymbol{X}}_t) - w_0(\overline{\boldsymbol{X}}_t)\big)\right.\right. \\
&\qquad\qquad \left.\left. + \sum_{j=1}^{b}\sum_{t\in T_{1,j}} \xi'_t\big(w(\overline{\boldsymbol{X}}_t) - w_0(\overline{\boldsymbol{X}}_t)\big)\right\}\right] \\
&\leq 2B_n\sqrt{a}\left(\mathbb{E}_{P(\xi)}\left[\mathfrak{R}_{ab}\left\{w - w_0 : w \in \mathcal{N}_n^{(w,s)}(\epsilon_n)\right\}\right] + \mathbb{E}_{P(\xi)}\left[\mathfrak{R}_{ab}\left\{s - s_0 : s \in \mathcal{N}_n^{(w,s)}(\epsilon_n)\right\}\right]\right).
\end{aligned}$$

Then, using the same reasoning as (A.35),

$$\mathbb{E}_{P(\xi)}\left[\mathfrak{R}_b\left\{\overline{G}^{(1)}_{j,w,s} : (w,s) \in \mathcal{N}_n^{(w,s)}(\epsilon_n)\right\}\right] \leq \big(24B_n\sqrt{a}\,C\epsilon_n\big)\sqrt{\frac{2\log(n)}{n}}\left(\sqrt{\mathrm{Pdim}(\mathcal{N}_n^{(w)})} + \sqrt{\mathrm{Pdim}(\mathcal{N}_n^{(s)})}\right).$$

Now, by Lemma (A.5.4) for some $C' > 0$,

$$\begin{aligned}
\mathbb{E}_{P(\xi)}\left[\mathfrak{R}_b\left\{\overline{G}^{(1)}_{j,w,s} : (w,s) \in \mathcal{N}_n^{(w,s)}(\epsilon_n)\right\}\right] &\leq C'\cdot\big(B_n\sqrt{a}\,\epsilon_n\big)\sqrt{\frac{\log^8(n)}{n}}\left(n^{\frac{1}{2}\left(\frac{d}{p^{(w)}+d}\right)} + n^{\left(\frac{d}{p^{(s)}+d}\right)(1/2-K_B)}\right) \\
&= C'\cdot\big(B_n\sqrt{a}\,\epsilon_n\big)\log^4(n)\left(n^{-\frac{1}{2}\left(\frac{p^{(w)}}{p^{(w)}+d}\right)} + n^{-\frac{1}{2}\left(\frac{p^{(s)}}{p^{(s)}+d}\right)-K_B\left(\frac{d}{p^{(s)}+d}\right)}\right) \\
&\leq 2C'\cdot\big(B_n\sqrt{a}\,\epsilon_n\big)\log^4(n)\left(n^{-\frac{1}{4}}\right) \\
&= \mathrm{o}\left(n^{-1/2}\right),
\end{aligned} \tag{B.8}$$

where the second inequality uses $1/4 < \frac{1}{2}\left(\frac{p^{(w)}}{p^{(w)}+d}\right)$ and $1/4 < \left(\frac{p^{(s)}}{p^{(s)}+d}\right)$ by Assumption 3.1.1(iii), and the last line uses $B_n\epsilon_n = \mathrm{o}\left(n^{-1/4}\right)$, by Assumption 3.1.1(iii),(iv) and the definition of ϵ_n in Corollary 3.1.1. Thus, combining (B.6), (B.7), and (B.8),

$$e^{-\log(n)} \geq P\left(\left(\frac{\sqrt{n}}{2}\right)\frac{1}{b}\sum_{j=1}^{b}\left\{\mathbb{E}\big[\overline{G}^{(1)}_{j,\hat{w}_n,\hat{s}_n}\big] - \overline{G}^{(1)}_{j,\hat{w}_n,\hat{s}_n}\right\} \geq \mathrm{o}(1)\right).$$

With (B.5) this implies $P_1 = \mathrm{o}(1)$. Combining this with (B.1), (B.2), (B.3), and (B.4), implies

$$\sqrt{n}\left|\mathbb{E}\Big[\mathsf{v}_t\big(\hat{w}_n,\hat{s}_n\big) - \mathsf{v}_t\big(w_0,s_0\big)\Big] - \frac{1}{n}\sum_{t=1}^{n}\Big\{\mathsf{v}_t\big(\hat{w}_n,\hat{s}_n\big) - \mathsf{v}_t\big(w_0,s_0\big)\Big\}\right| = \mathrm{o}_P(1).$$

The proof is complete since $\sqrt{n}\left|\mathbb{E}\Big[\mathsf{A}_t\big(\hat{w}_n\big) - \mathsf{A}_t\big(w_0\big)\Big] - \frac{1}{n}\sum_{t=1}^{n}\Big\{\mathsf{A}_t\big(\hat{w}_n\big) - \mathsf{A}_t\big(w_0\big)\Big\}\right| = \mathrm{o}_P(1)$, follows by a similar argument. ■

Lemma B.1.3. (Neyman Orthogonality and Smoothness) *For any $w, s \in \mathcal{L}^\infty(P_{\boldsymbol{X}})$ such that $\|w\|_\infty \leq 2$, $\|s\|_\infty \leq B_n$*

$$\mathbb{E}\Big[\psi\big(\zeta_0, w_0, s_0\big)\Big] - \mathbb{E}\Big[\psi\big(\zeta_0, \hat{w}_n, \hat{s}_n\big)\Big] = \mathrm{o}_P(n^{-1/2}).$$

Proof. For arbitrary $w, s \in \mathcal{L}^\infty(P_{\boldsymbol{X}})$ such that $\|w\|_\infty \leq 2$, $\|s\|_\infty \leq B_n$ and $\lambda \in \mathbb{R}$ define

$$F(\lambda) \coloneqq (w_0, s_0)^\mathsf{T} + \lambda\big[(w, s)^\mathsf{T} - (w_0, s_0)^\mathsf{T}\big].$$

With this, we write

$$\mathbb{E}\Big[\psi\big(\zeta_0, F(0)\big)\Big] = \mathbb{E}\Big[\psi\big(\zeta_0, w_0, s_0\big)\Big], \quad \text{and} \quad \mathbb{E}\Big[\psi\big(\zeta_0, F(1)\big)\Big] = \mathbb{E}\Big[\psi\big(\zeta_0, w, s\big)\Big].$$

Applying Taylor's Theorem to $\mathbb{E}\Big[\psi\big(\zeta_0, F(\lambda)\big)\Big]$ centered at $\lambda = 0$

$$\mathbb{E}\Big[\psi\big(\zeta_0, w, s\big)\Big] = \mathbb{E}\Big[\psi\big(\zeta_0, F(0)\big)\Big] + \left[\frac{d}{d\lambda}\mathbb{E}\Big[\psi\big(\zeta_0, F(\lambda)\big)\Big]\right]_{\lambda=0} + \frac{1}{2}\int_0^1 \left[\frac{d^2}{d\lambda^2}\mathbb{E}\Big[\psi\big(\zeta_0, F(\lambda)\big)\Big]\right] d\lambda$$

For the first-order derivatives of ψ

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\mathsf{A}\big(\boldsymbol{Z}_t; w_0 + \lambda(w - w_0)\big) &= 2\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\Big(D_t - w_0(\boldsymbol{X}_t) + \lambda\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\Big), \\
\frac{\partial}{\partial\lambda}\mathsf{v}\big(\boldsymbol{Z}_t; F(\lambda)\big) &= \big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\Big(Y_t - s_0(\boldsymbol{X}_t) + \lambda\big(s(\boldsymbol{X}_t) - s_0(\boldsymbol{X}_t)\big)\Big) \\
&\quad + \big(s(\boldsymbol{X}_t) - s_0(\boldsymbol{X}_t)\big)\Big(D_t - w_0(\boldsymbol{X}_t) + \lambda\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\Big).
\end{aligned}$$

For the second-order derivatives of ψ

$$\begin{aligned}
\frac{\partial^2}{\partial\lambda^2}\mathsf{A}\big(\boldsymbol{Z}_t; w_0 + \lambda(w - w_0)\big) &= 2\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)^2, \\
\frac{\partial^2}{\partial\lambda^2}\mathsf{v}\big(\boldsymbol{Z}_t; F(\lambda)\big) &= 2\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\big(s(\boldsymbol{X}_t) - s_0(\boldsymbol{X}_t)\big),
\end{aligned}$$

Hence

$$\begin{aligned}
\left[\frac{\partial}{\partial\lambda}\psi\big(\boldsymbol{Z}_t; \zeta_0, F(\lambda)\big)\right]_{\lambda=0} &= 2\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\big(D_t - w_0(\boldsymbol{X}_t)\big) \\
&\quad - \big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\big(Y_t - s_0(\boldsymbol{X}_t)\big) - \big(s(\boldsymbol{X}_t) - s_0(\boldsymbol{X}_t)\big)\big(D_t - w_0(\boldsymbol{X}_t)\big).
\end{aligned}$$

Note that for any $\delta > 0$, and any $\lambda \in (-\delta, 1 + \delta)$ the function $\mathbb{E}\Big[\psi\big(\zeta_0, F(\lambda)\big)\Big]$ is continuously differentiable infinitely many times and its derivatives are elements of $\mathcal{L}^2(P_{\boldsymbol{Z}})$ since w, s, w_0, s_0 are

bounded and $Y_t \in \mathcal{L}^2$. Therefore, a measure-theoreic version of Leibniz's Rule (e.g. Corbae et al., 2009, Theorem 7.5.19, p.405) can be applied to show ψ satisfies a Neyman Orthogonality condition,

$$\begin{aligned}
&\left[\frac{d}{d\lambda}\mathbb{E}\Big[\psi\big(\zeta_0, F(\lambda)\big)\Big]\right]_{\lambda=0} \\
&= 2\mathbb{E}\Big[\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\big(D_t - w_0(\boldsymbol{X}_t)\big)\Big] - \mathbb{E}\Big[\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\big(Y_t - s_0(\boldsymbol{X}_t)\big)\Big] \\
&\quad - \mathbb{E}\Big[\big(s(\boldsymbol{X}_t) - s_0(\boldsymbol{X}_t)\big)\big(D_t - w_0(\boldsymbol{X}_t)\big)\Big] \\
&= 2\mathbb{E}\Big[\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\big(\mathbb{E}[D_t|\boldsymbol{X}_t] - w_0(\boldsymbol{X}_t)\big)\Big] - \mathbb{E}\Big[\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\big(\mathbb{E}[Y_t|\boldsymbol{X}_t] - s_0(\boldsymbol{X}_t)\big)\Big] \\
&\quad - \mathbb{E}\Big[\big(s(\boldsymbol{X}_t) - s_0(\boldsymbol{X}_t)\big)\big(\mathbb{E}[D_t|\boldsymbol{X}_t] - w_0(\boldsymbol{X}_t)\big)\Big] \\
&= 0,
\end{aligned}$$

by iterated expectations and the definitions of w_0, s_0. Applying this to the Taylor series expansion from before, and by similar arguments we can apply Leibnez's rule again and Fubini's theorem to the remainder term to obtain

$$\begin{aligned}
\mathbb{E}\Big[\psi\big(\zeta_0, w, s\big)\Big] &= \mathbb{E}\Big[\psi\big(\zeta_0, F(0)\big)\Big] + \frac{1}{2}\int_0^1 \left[\frac{d^2}{d\lambda^2}\mathbb{E}\Big[\psi\big(\zeta_0, F(\lambda)\big)\Big]\right] d\lambda \\
&= \mathbb{E}\Big[\psi\big(\zeta_0, F(0)\big)\Big] + \mathbb{E}\Big[\big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)^2 - \big(w(\boldsymbol{X}_t) - w_0(\boldsymbol{X}_t)\big)\big(s(\boldsymbol{X}_t) - s_0(\boldsymbol{X}_t)\big)\Big] \\
&\leq \mathbb{E}\Big[\psi\big(\zeta_0, F(0)\big)\Big] + \|w - w_0\|^2_{\mathcal{L}^2(P_{\boldsymbol{X}})} + \|w - w_0\|_{\mathcal{L}^2(P_{\boldsymbol{X}})}\|s - s_0\|_{\mathcal{L}^2(P_{\boldsymbol{X}})} \\
&= \mathbb{E}\Big[\psi\big(\zeta_0, F(0)\big)\Big] + \mathrm{O}_P(\epsilon_n^2).
\end{aligned}$$

The desired result follows because $\mathbb{E}\Big[\psi\big(\zeta_0, F(0)\big)\Big] = \mathbb{E}\Big[\psi\big(\zeta_0, w_0, s_0\big)\Big]$ by definition, and $\epsilon_n = \mathrm{o}\big(n^{-1/4}\big)$ by Assumption 3.1.1(iii). ■

B.2 Proof of Theorem 3.1.1

Proof. This proof is similar to Chen et al. (2022, Theorem 1), but specifically tailored to settings with dependent data. First,

$$\begin{aligned}
\mathbb{E}\Big[\big(\mathsf{A}_t(\hat{w}_n) - \mathsf{A}_t(w_0)\big)\Big] &= \mathbb{E}\Big[\big(D_t - \hat{w}_n(\boldsymbol{X}_t)\big)^2 - \big(D_t - w_0(\boldsymbol{X}_t)\big)^2\Big] \\
&= \mathbb{E}\Big[\hat{w}_n(\boldsymbol{X}_t)^2 - w_0(\boldsymbol{X}_t)^2 + 2D_t\big(w_0(\boldsymbol{X}_t) - \hat{w}_n(\boldsymbol{X}_t)\big)\Big] \\
&= \mathbb{E}\Big[\big(w_0(\boldsymbol{X}_t) - \hat{w}_n(\boldsymbol{X}_t)\big)\big(w_0(\boldsymbol{X}_t) + \hat{w}_n(\boldsymbol{X}_t) + 2D_t\big)\Big] \\
&\leq 5\,\mathbb{E}\Big[\big|w_0(\boldsymbol{X}_t) - \hat{w}_n(\boldsymbol{X}_t)\big|\Big] \leq 5\,\|\hat{w}_n - w_0\|_{\mathcal{L}^2(P_{\boldsymbol{X}})} = \mathrm{O}_P(\epsilon_n).
\end{aligned}$$

With this,

$$\begin{aligned}
\mathbb{E}\big[\mathsf{A}_t(\hat{w}_n)\big]\big(\hat{\zeta} - \zeta_0\big) &= \mathbb{E}\big[\mathsf{A}_t(w_0)\big]\big(\hat{\zeta} - \zeta_0\big) + \mathbb{E}\big[\mathsf{A}_t(\hat{w}_n) - \mathsf{A}_t(w_0)\big]\big(\hat{\zeta} - \zeta_0\big) \\
&= \mathbb{E}\big[\mathsf{A}_t(w_0)\big]\big(\hat{\zeta} - \zeta_0\big) + \mathrm{O}_P(\epsilon_n)\mathrm{O}\big(|\hat{\zeta} - \zeta_0|\big) \\
&= \mathbb{E}\big[\mathsf{A}_t(w_0)\big]\big(\hat{\zeta} - \zeta_0\big) + \mathrm{o}_P\big(|\hat{\zeta} - \zeta_0|\big). \qquad \text{(B.9)}
\end{aligned}$$

Define the following notation

$$\Psi(\zeta, w, s) \coloneqq \mathbb{E}\big[\mathsf{A}_t(w)\big]\zeta - \mathbb{E}\big[\mathsf{v}_t(w, s)\big], \qquad \Psi_n(\zeta, w, s) \coloneqq \frac{1}{n}\sum_{t=1}^{n}\big\{\mathsf{A}_t(w)\zeta - \mathsf{v}_t(w, s)\big\},$$

$$\text{and} \quad U_n(\zeta, w, s) \coloneqq \Psi(\zeta, w, s) - \Psi_n(\zeta, w, s).$$

Then,

$$\begin{aligned}
\mathbb{E}\big[\mathsf{A}_t(\hat{w}_n)\big]\big(\hat{\zeta} - \zeta_0\big) &= \Psi\big(\hat{\zeta}, \hat{w}_n, \hat{s}_n\big) - \Psi\big(\zeta_0, \hat{w}_n, \hat{s}_n\big) \\
&= U_n\big(\hat{\zeta}, \hat{w}_n, \hat{s}_n\big) - \Psi\big(\zeta_0, \hat{w}_n, \hat{s}_n\big) + \Psi_n\big(\hat{\zeta}, \hat{w}_n, \hat{s}_n\big) \\
&= U_n\big(\hat{\zeta}, \hat{w}_n, \hat{s}_n\big) + \Psi\big(\zeta_0, w_0, s_0\big) - \Psi\big(\zeta_0, \hat{w}_n, \hat{s}_n\big) + \Psi_n\big(\hat{\zeta}, \hat{w}_n, \hat{s}_n\big) \\
&= U_n\big(\hat{\zeta}, \hat{w}_n, \hat{s}_n\big) + \mathrm{o}_P(n^{-1/2}), \qquad \text{(B.10)}
\end{aligned}$$

where the third equality uses $\Psi\big(\zeta_0, w_0, s_0\big) = \mathbb{E}\big[\mathsf{A}_t(w)\zeta_0 - \mathsf{v}_t\big(w, s\big)\big] = 0$, and the last equality uses Lemma B.1.1 and Lemma B.1.3. Combining (B.9) and (B.10)

$$\mathbb{E}\big[\mathsf{A}_t(w_0)\big]\big(\hat{\zeta} - \zeta_0\big) + \mathrm{o}_P\big(|\hat{\zeta} - \zeta_0|\big) = U_n\big(\hat{\zeta}, \hat{w}_n, \hat{s}_n\big) + \mathrm{o}_P(n^{-1/2}). \qquad \text{(B.11)}$$

Now we decompose $U_n(\hat{\zeta}, \hat{w}_n, \hat{s}_n)$ into an asymptotically normal component and components that converge to zero in probability at a suitable rate,

$$\begin{aligned} &U_n(\hat{\zeta}, \hat{w}_n, \hat{s}_n) \\ &= U_n(\zeta_0, w_0, s_0) + \Big[U_n(\hat{\zeta}, \hat{w}_n, \hat{s}_n) - U_n(\zeta_0, \hat{w}_n, \hat{s}_n)\Big] + \Big[U_n(\zeta_0, \hat{w}_n, \hat{s}_n) - U_n(\zeta_0, w_0, s_0)\Big]. \end{aligned} \tag{B.12}$$

For the middle term of (B.12)

$$\begin{aligned} U_n(\hat{\zeta}, \hat{w}_n, \hat{s}_n) - U_n(\zeta_0, \hat{w}_n, \hat{s}_n) &= \Psi(\hat{\zeta}, \hat{w}_n, \hat{s}_n) - \Psi_n(\hat{\zeta}, \hat{w}_n, \hat{s}_n) - \Psi(\zeta_0, \hat{w}_n, \hat{s}_n) + \Psi_n(\zeta_0, \hat{w}_n, \hat{s}_n) \\ &= (\hat{\zeta} - \zeta_0)\left(\mathbb{E}\big[\mathsf{A}_t(\hat{w}_n)\big] - \frac{1}{n}\sum_{t=1}^{n} \mathsf{A}_t(\hat{w}_n)\right). \end{aligned}$$

Then, by the triangle inequality

$$\begin{aligned} &\left|\mathbb{E}\big[\mathsf{A}_t(\hat{w}_n)\big] - \frac{1}{n}\sum_{t=1}^{n} \mathsf{A}_t(\hat{w}_n)\right| \\ &\leq \left|\mathbb{E}\big[\mathsf{A}_t(w_0)\big] - \frac{1}{n}\sum_{t=1}^{n} \mathsf{A}_t(w_0)\right| + \left|\mathbb{E}\Big[\mathsf{A}_t(\hat{w}_n) - \mathsf{A}_t(w_0)\Big] - \frac{1}{n}\sum_{t=1}^{n} \Big\{\mathsf{A}_t(\hat{w}_n) - \mathsf{A}_t(w_0)\Big\}\right| \\ &= \mathrm{o}_P\Big(n^{-1/2}\log_2(n)\log(n)\Big) + \mathrm{o}_P(n^{-1/2}), \end{aligned}$$

where the last line follows by applying Lemma B.1.2 to the second term, and Bosq (1998, Theorem 1.6, p.35) to the first term, which can be applied since $|\mathsf{A}_t(w_0)| = \big|D_t - w_0(\boldsymbol{X}_t)\big|^2 \leq 4$, and $\mathsf{A}_t(w_0)$ inherits the mixing properties of $\{\boldsymbol{Z}_t\}_{t=1}^n$ by Davidson (2022, Theorem 15.1). Combining the previous two displays,

$$U_n(\hat{\zeta}, \hat{w}_n, \hat{s}_n) - U_n(\zeta_0, \hat{w}_n, \hat{s}_n) = \mathrm{o}_P(|\hat{\zeta} - \zeta_0|).$$

For the last term of (B.12),

$$\begin{aligned} U_n(\zeta_0, \hat{w}_n, \hat{s}_n) - U_n(\zeta_0, w_0, s_0) &\leq \left|\mathbb{E}\Big[\mathsf{A}_t(\hat{w}_n) - \mathsf{A}_t(w_0)\Big] - \frac{1}{n}\sum_{t=1}^{n} \Big\{\mathsf{A}_t(\hat{w}_n) - \mathsf{A}_t(w_0)\Big\}\right|\zeta_0 \\ &\quad + \left|\mathbb{E}\Big[\mathsf{v}_t(\hat{w}_n, \hat{s}_n) - \mathsf{v}_t(w_0, s_0)\Big] - \frac{1}{n}\sum_{t=1}^{n} \Big\{\mathsf{v}_t(\hat{w}_n, \hat{s}_n) - \mathsf{v}_t(w_0, s_0)\Big\}\right| \\ &= \mathrm{o}_P(n^{-1/2}), \end{aligned}$$

by Lemma B.1.2. Applying the previous two displays to (B.12) and plugging this into (B.11)

$$\mathbb{E}\big[\mathsf{A}_t(w_0)\big](\hat{\zeta} - \zeta_0) + \mathrm{o}_P(|\hat{\zeta} - \zeta_0|) = U_n(\zeta_0, w_0, s_0) + \mathrm{o}_P(n^{-1/2}). \tag{B.13}$$

Note that $U_n(\zeta_0, w_0, s_0)$ is a zero-mean process, and $U_n(\zeta_0, w_0, s_0) = -\frac{1}{n}\sum_{t=1}^{n} \psi_t(\zeta_0, w_0, s_0)$. Then, by Bosq (1998, Theorem 1.5, p.34), for some constant $\sigma^2 \geq 0$

$$\begin{aligned}\operatorname{Var}\left[\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_t(\zeta_0, w_0, s_0)\right] &= \operatorname{Var}\Big[\sqrt{n}\,U_n(\zeta_0, w_0, s_0)\Big] = \mathbb{E}\Big[\Big(\sqrt{n}\,U_n(\zeta_0, w_0, s_0)\Big)^2\Big] \\ &= \Big\|\sqrt{n}\,U_n(\zeta_0, w_0, s_0)\Big\|_{\mathcal{L}^2}^2 \to \sigma^2,\end{aligned}$$

as $n \to \infty$. By Markov's inequality, and Hölder's inequality, for any $c > 0$,

$$\begin{aligned}P\Big(\big|\sqrt{n}\,U_n(\zeta_0, w_0, s_0)\big| \geq c\Big) &\leq \frac{1}{c}\Big\|\sqrt{n}\,U_n(\zeta_0, w_0, s_0)\Big\|_{\mathcal{L}^1} \leq \frac{1}{c}\Big\|\sqrt{n}\,U_n(\zeta_0, w_0, s_0)\Big\|_{\mathcal{L}^2} \\ &= \frac{1}{c}\sqrt{\operatorname{Var}\Big[\sqrt{n}\,U_n(\zeta_0, w_0, s_0)\Big]} \to \frac{\sigma}{c},\end{aligned}$$

as $n \to \infty$. Thus, $U_n(\zeta_0, w_0, s_0) = \mathrm{O}_P(n^{-1/2})$. Applying this to (B.13), since $\mathbb{E}\big[\mathsf{A}_t(w_0)\big] > 0$ is a constant,

$$\mathbb{E}\big[\mathsf{A}_t(w_0)\big](\hat{\zeta} - \zeta_0) + \mathrm{o}_P(|\hat{\zeta} - \zeta_0|) = \mathrm{O}_P(n^{-1/2}) \quad \Longrightarrow \quad (\hat{\zeta} - \zeta_0) = \mathrm{O}_P(n^{-1/2}),$$

which proves result (i).

With this, we can write (B.13) as

$$\sqrt{n}(\hat{\zeta} - \zeta_0) = \sqrt{n}\,\mathbb{E}\big[\mathsf{A}_t(w_0)\big]^{-1} U_n(\zeta_0, w_0, s_0) + \mathrm{o}_P(1).$$

Result (ii) follows by applying a central limit theorem to the first term. If $\sigma > 0$, then the conditions for Bosq (1998, Theorem 1.7) are met by Assumption 3.1.1(i)(ii), since $U_n(\zeta_0, w_0, s_0)$ inherits the mixing properties of $\{\boldsymbol{Z}_t\}_{t=1}^n$ by Davidson (2022, Theorem 15.1). Then, this result implies

$$\frac{\sqrt{n}\,U_n(\zeta_0, w_0, s_0)}{\sigma} \xrightarrow{d} N(0, 1).$$

■