Statistical Inference with Ill-Conditioned Random Weight Neural Network Sieves

Giovanni Ballarin [*] Lukas Gonon [†] Lyudmila Grigoryeva [‡]

November 20, 2026

Link to most recent version

Abstract

Random weight neural networks (RWNNs) are shallow networks whose hidden-layer parameters are randomly generated and then kept fixed, while only the output coefficients are estimated. Conditional on the sampled parameters, RWNN regression is a linear sieve problem with random approximation space. We show that such random sieves cannot remain uniformly well-conditioned as their dimension grows. For parametrized feature families satisfying polynomial covering and Lipschitz conditions, the eigenvalues of the population covariance matrix are proven to have polynomial upper bounds determined by the covering geometry of the parameter space, leading to a vanishing smallest eigenvalue. We develop approximation, estimation, and inference theory that explicitly treats the ill-conditioning. Under a β-Hölder source condition, we obtain both faster-than-Monte-Carlo L^2 approximation rates and new L^∞ approximation bounds based on stratified sampling. We then derive improved consistency guarantees for ridge-regularized least squares driven by ridge effective dimension, spectral, and leverage-based stability conditions. For inference, we establish asymptotic linearity and Gaussian limits for linear functional directions while allowing the smallest eigenvalue of the population covariance matrix to vanish. Nondegeneracy is directional: we characterize it with a ridge-normalized Rayleigh quotient and extend it to smooth functionals via Riesz representation. Consistency of ridge-adjusted variance estimators is proved, yielding feasible studentized inference. Numerical experiments illustrate the deterioration of random-feature conditioning and the finite-sample behavior of the resulting estimation and inference procedures.

Keywords: random weight neural networks, random feature sieves, nonparametric estimation, spectral ill-conditioning, ridge regression, functional inference

[*]Corresponding author – University of St. Gallen, Division of Mathematics and Statistics, Rosenbergstrasse 22, 9000 St. Gallen, Switzerland. `Giovanni.Ballarin@unisg.ch`

[†]University of St. Gallen, School of Computer Science, Rosenbergstrasse 30, 9000 St. Gallen, Switzerland. `Lukas.Gonon@unisg.ch`. Honorary Senior Lecturer. Imperial College London, United Kingdom.

[‡]University of St. Gallen, Division of Mathematics and Statistics, Rosenbergstrasse 22, 9000 St. Gallen, Switzerland. `Lyudmila.Grigoryeva@unisg.ch`. Honorary Associate Professor. University of Warwick, Department of Statistics, Coventry, CV4 7AL, United Kingdom.

Acknowledgements: GB is thankful to the Great Minds Postdoctoral Fellowship program of the University of St. Gallen for financial support. We used generative AI to polish the manuscript, proofread, and assist with code development, but not to formulate theoretical models, conceptual frameworks, hypotheses, or mathematical claims. We manually reviewed all AI-assisted work and take full responsibility for the final content.

Content

1 Introduction **2**
- 1.1 Preview of contributions . . . 4
- 1.2 Notation . . . 6

2 Model and random weight neural network sieves **6**
- 2.1 Nonparametric regression model . . . 6
- 2.2 Shallow neural networks and random weight neural networks . . . 7
- 2.3 RWNN least squares estimator . . . 8

3 Approximation properties of RWNNs **9**
- 3.1 L^2 approximation under a β-Hölder source condition . . . 10
- 3.2 L^2 approximation on integral Sobolev classes . . . 13
- 3.3 Sup-norm approximation via stratified sampling . . . 14

4 Ridge-regularized RWNN estimation and inference **18**
- 4.1 Spectral deterioration and effective dimension . . . 19
- 4.2 L^2 estimation . . . 21
- 4.3 Directional inference for linear functionals . . . 24
- 4.4 Feasible inference . . . 27
- 4.5 Inference for smooth functionals . . . 30
- 4.6 Sup-norm analysis . . . 32

5 Numerical experiments **40**
- 5.1 Sup-norm approximation . . . 41
- 5.2 Weight normalization, conditioning, and L^2 stability . . . 42
- 5.3 Sup-norm stability . . . 42
- 5.4 Asymptotic inference . . . 42
- 5.5 Bootstrap inference . . . 45

6 Conclusion **47**

Appendix **54**

1 Introduction

Random weight neural networks (RWNNs) are shallow neural networks whose hidden-layer parameters are randomly drawn and subsequently kept fixed, so that only the coefficients of the linear output layer are estimated from the data (Gonon et al., 2023). Neural networks of this type have been applied very successfully in a variety of settings; we refer in particular to the seminal works on random feature models (Rahimi and Recht, 2007) and extreme learning machines (Huang et al., 2006). Conditional on the sampled parameters, estimation under squared loss is a linear least-squares problem over the random finite-dimensional space spanned by the K hidden units. When the same sequence of hidden-layer parameters is used as K increases, these spaces are nested and form a random sieve, which we study within the framework of nonparametric sieve regression (Grenander, 1981; Györfi et al., 2006; Chen, 2007; Tsybakov, 2009). However, the population covariance matrix of the randomly generated features need not remain well-conditioned as K increases: additional hidden units may improve approximation while introducing substantial redundancy in the feature representation. Understanding the statistical consequences of such ill-conditioning, particularly for inference on functionals of the regression function, is the central problem addressed in this paper.

Approximation properties of random neural networks have been studied under different assumptions on the target function and the distribution of the hidden-layer parameters (Rahimi and Recht, 2008; Gonon et al., 2023; Neufeld and Schmocker, 2023; Liu et al., 2025; De Ryck et al., 2025; Mehmood and Gonon, 2026). In the context of positive definite kernels, Bach (2017) established approximation guarantees for random feature expansions in terms of the spectral properties of the associated integral operator, including results in L^2 and L^∞. Related quantitative approximation results for shallow neural networks include those for general activation functions (Siegel and Xu, 2020), for Sobolev classes with ReLU activations (Mao et al., 2026), and for non-polynomial activations on noncompact domains (Neufeld and Schmocker, 2026). Statistical convergence rates for random feature estimators have been established in several settings (Rudi and Rosasco, 2017; Carratino et al., 2018; Gonon, 2023; Lanthaler and Nelsen, 2023). In particular, Braun et al. (2024) establish an expected squared $L^2(X)$ error bound of order $O((\log n)^4 n^{-1/2})$ for a shallow neural network estimator learned by gradient descent, under suitable Fourier-decay assumptions on the regression function. They obtain the same rate for a truncated linear least-squares estimator with randomly sampled hidden-layer parameters. Mei and Montanari (2022) derive precise asymptotics for random features ridge regression. Conditional on the sampled features, a ridge-regularized RWNN estimator is a finite-rank kernel ridge regressor, for which Cheng et al. (2023) give sharp non-asymptotic bounds on the test error.

From the kernel point of view—the original setting of the random features approach of Rahimi and Recht (2007)—statistical inference for kernel ridge regression has been studied in the recent literature, including uniform confidence bands based on Gaussian and bootstrap approximations (Singh and Vijaykumar, 2023) and asymptotic normality for linear functionals, including point evaluations and derivatives (Tuo and Zou, 2024). These results concern regularization in a prescribed reproducing kernel Hilbert space, and existing inferential results involving sampled features address particular estimation targets and statistical models; see Shen and Bellec (2020); Steland (2021); Guignard et al. (2021) for recent discussions. In contrast, we develop a theory of functional inference for growing RWNN sieves that accommodates increasingly ill-conditioned or singular population feature covariance matrices.

Randomly generated features present a fundamental and distinctive structural obstacle: a uniform positive lower bound on the population covariance eigenvalues cannot be sustained as the number of features increases. In fact, we show that, for parameterized feature families satisfying polynomial covering and Lipschitz conditions, the eigenvalues of the population feature covariance matrix satisfy polynomial decay upper bounds determined by the covering dimension of the parameter space, uniformly over feature realizations. In particular, its smallest eigenvalue necessarily converges to zero as the number of features increases. Related spectral ill-conditioning has been investigated by Ming and Yu (2026), who establish rapid singular-value decay for random-feature discretization matrices associated with Fourier and hyperbolic tangent features. Our results instead concern the population regression covariance matrix under general covering and Lipschitz conditions.

Unlike an $L^2(X)$-orthonormal series sieve, whose population Gram matrix is I_K, an RWNN sieve does not generally have an orthogonal feature representation. Standard sieve inference typically relies on suitably normalized bases and control of their population Gram matrices (Newey, 1997; Chen, 2007; Belloni et al., 2015; Chen and Christensen, 2015). Importantly, our spectral deterioration results are not restricted to randomly generated features: they hold uniformly over arbitrary choices of hidden-layer parameters within the prescribed compact feature family and therefore also apply to fully trainable shallow networks satisfying these restrictions. Our subsequent statistical analysis focuses on RWNNs. In this setting, both spectral deterioration and the resulting regularization bias must be accounted for explicitly.

We develop a framework for statistical inference with random sieves and RWNNs based on ridge normalization, in which the relevant quantities are measured relative to the regularized

population covariance matrix. Two key results guide our analysis. First, the same covering geometry that induces spectral decay also controls the *ridge effective dimension*, which governs the stochastic estimation error and is sublinear in the nominal sieve dimension under suitable ridge scaling. This effective dimension plays a role analogous to that of its counterpart in kernel ridge regression (Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017). Second, and more importantly for inference, global spectral nondegeneracy is not required. We formulate a sufficient nondegeneracy condition *along the functional direction*, relative to the ridge level, through a ridge-normalized Rayleigh quotient. Together with conditions controlling empirical covariance stability, approximation, and ridge bias, we derive asymptotic normality conditional on the sampled hidden-layer parameters, in P_θ-probability, even when the smallest eigenvalue of the population feature covariance matrix converges to zero. Under additional conditions ensuring consistency of the ridge-adjusted variance estimator, we obtain feasible studentized inference. The framework also accommodates irregular sieve functionals, such as point evaluation, whose normalizing variance may vary with the sieve dimension and need not remain bounded.

1.1 Preview of contributions

The main contributions of this paper concern four interconnected aspects of growing RWNN sieves: approximation by randomly generated features; spectral properties and $L^2(X)$ estimation; asymptotic inference for linear and smooth functionals; and uniform estimation under prediction-space stability conditions. Throughout this subsection, P_θ denotes the law of the sampled hidden-layer parameters, O_{P_θ} refers to this law, and O_P refers to the joint law of the observations and the hidden-layer parameters (see Subsection 1.2 below). Table 1 summarizes selected approximation benchmarks from the literature and the principal approximation, spectral, estimation, and inference results established in this paper.

First, we establish L^2 approximation rates for RWNNs under a β-Hölder direct source condition. Adapting the control-neighbors Monte Carlo construction of Leluc et al. (2025) to the RWNN integral representation yields an approximation error of order $O_{P_\theta}(K^{-1/2-\beta/(d+1)})$, improving upon the classical Monte Carlo order $K^{-1/2}$. Under the additional conditions of the generic sieve theory developed in Appendix E, this improvement yields an $o_P(n^{-1/4})$ convergence rate for the constrained unregularized sieve estimator, ensuring that quadratic remainder terms are negligible at the $\sqrt{n}$ scale in smooth-functional inference. We also derive approximation guarantees for integral Sobolev classes (see Subsection 3.2). Furthermore, we establish sup-norm approximation bounds using stratified sampling of the hidden-layer parameters. Under the β-Hölder source condition, the expected sup-norm approximation error is bounded by a constant multiple of $(\sqrt{d}+\sqrt{\log K})K^{-1/2-\beta/(d+1)}$. For sign-valued source weights and ReLU^m activations, $m \in \{1,2\}$, the corresponding bound is $(\sqrt{d}+\sqrt{\log K})K^{-1/2-1/(d+1)}$. In the latter case, the construction uses a partition respecting the sign sets of the source function and is therefore oracle unless this partition is known. The proof relies on a single centered empirical-process representation of the approximation error. The construction also provides deterministic ℓ_1- and ℓ_2-bounds on the output coefficients, used in the subsequent statistical analysis.

Second, we investigate the spectral properties of the population feature covariance matrix and their implications for statistical estimation. Let g_K denote the vector of K randomly generated features, and define $\Gamma_K := \mathbb{E}_X[g_K(X)g_K(X)^\top]$ and $\Gamma_{K,\lambda} := \Gamma_K + \lambda I_K$, where $\mathbb{E}_X$ denotes integration with respect to the distribution of X for fixed hidden-layer parameters. Under polynomial covering and Lipschitz conditions on the feature family, with covering exponent $q_\Pi > 0$, we establish spectral-tail bounds implying $\lambda_{\min}(\Gamma_K) \lesssim K^{-2/q_\Pi}$, uniformly over feature realizations. The same bounds control the ridge effective dimension,

$$\mathcal{N}_{K,\lambda} := \mathrm{tr}(\Gamma_K \Gamma_{K,\lambda}^{-1}) \lesssim 1 + \left(\frac{K}{\lambda}\right)^{q_\Pi/(q_\Pi+2)}.$$

In particular, $\mathcal{N}_{K_n,\lambda} = o(K_n)$ under the scaling conditions $K_n/\lambda_{K_n,n} \to \infty$ and $\lambda_{K_n,n} K_n^{2/q_\Pi} \to \infty$.

Thus, the same covering geometry that rules out uniform lower spectral conditioning also provides bounds on the effective dimension governing ridge estimation. Building on this spectral analysis, we establish $L^2(X)$ convergence rates for ridge-regularized RWNN least squares estimators. Under the β-Hölder source condition and suitable ridge-normalized covariance stability, the ridge estimator $\widehat{h}_{n,\lambda}$ satisfies

$$\|\widehat{h}_{n,\lambda} - h\|_{L^2(X)} = O_P\left(\sqrt{\frac{\mathcal{N}_{K_n,\lambda}}{n}} + K_n^{-1/2-\beta/(d+1)} + \sqrt{\frac{\lambda_{K_n,n}\log K_n}{K_n}}\right).$$

The three terms correspond to stochastic estimation, sieve approximation, and regularization bias. In particular, when $\mathcal{N}_{K_n,\lambda} = o(K_n)$, the stochastic term is of smaller order than the classical $\sqrt{K_n/n}$ sieve term, without requiring a uniform lower bound on the eigenvalues of the population feature covariance matrix. For comparison, spline and wavelet series estimators attain the minimax-optimal $L^2(X)$ rate $n^{-p/(2p+d)}$ for regression functions with Hölder smoothness $p > 0$ under suitable assumptions (Stone, 1982; Huang, 2003). Deep neural network estimators can also achieve this rate up to logarithmic factors on Hölder classes under appropriate architectural restrictions (Schmidt-Hieber, 2020). For shallow neural networks, Braun et al. (2024) obtain an expected squared $L^2(X)$ error bound of order $(\log n)^4 n^{-1/2}$ under Fourier-decay assumptions, including for an estimator with randomly sampled hidden-layer parameters. These benchmarks are not directly comparable with ours, since our β-Hölder condition concerns the source function in an integral representation of h, rather than the smoothness of h itself. The distinguishing feature of our bound is that its stochastic term is governed by the ridge effective dimension $\mathcal{N}_{K_n,\lambda}$, which can be $o(K_n)$, rather than by the nominal sieve dimension K_n, even when the population covariance matrix becomes increasingly ill-conditioned.

Third, we establish asymptotic normality for linear functionals of ridge-regularized RWNN estimators. For a prescribed unit direction α_{K_n}, measurable with respect to the sampled hidden-layer parameters, we characterize directional nondegeneracy through the ridge-normalized Rayleigh quotient

$$\mathcal{R}_{K_n,\lambda}(\alpha_{K_n}) := \frac{\alpha_{K_n}^\top \Gamma_{K_n,\lambda}^{-1}\Gamma_{K_n}\Gamma_{K_n,\lambda}^{-1}\alpha_{K_n}}{\alpha_{K_n}^\top \Gamma_{K_n,\lambda}^{-1}\alpha_{K_n}}.$$

Under suitable directional nondegeneracy, empirical covariance stability, approximation, and ridge-bias conditions, we obtain asymptotic normality for linear functionals of the estimated coefficients, centered at the corresponding $L^2(X)$ projection coefficients. The convergence holds conditionally on the sampled hidden-layer parameters, in P_θ-probability. Under additional conditions, we establish consistency of a residual-based ridge-adjusted variance estimator, yielding feasible studentized inference. An explicit ReLU example demonstrates that the inferential and variance-consistency conditions can hold even when the population feature covariance matrix is exactly singular. The directional theory also accommodates sieve-dependent normalizations, including those arising for irregular functionals such as point evaluation. Inference on the corresponding functional of h additionally requires control of the sieve approximation bias. We extend the analysis to smooth nonlinear functionals Ψ whose derivatives admit $L^2(X)$ Riesz representers. For quadratic remainders, an $o_P(n^{-1/4})$ convergence rate in $L^2(X)$ ensures negligibility at the $\sqrt{n}$ scale. Together with suitable functional bias and variance nondegeneracy conditions, this yields asymptotic normality for $\Psi(\widehat{h}_{n,\lambda}) - \Psi(h)$. For the integrated-square functional, we also verify the first-order sieve-bias condition under an explicit choice of sieve dimension.

Fourth, we establish sup-norm error bounds for ridge-regularized RWNN estimators. We first derive bounds under spectral stability conditions and then develop an alternative approach based on prediction-space stability, formulated through ridge smoothing weights, the empirical ridge Lebesgue constant, and leverage-type quantities. Under suitable conditions on these quantities, we obtain uniform convergence bounds without requiring a uniform lower bound on the smallest

eigenvalue of the empirical feature covariance matrix. In particular, for fixed d and under the conditions of Corollary 4.10, we obtain

$$\|\widehat{h}_{n,\lambda} - h\|_\infty = O_P\left(n^{-\frac{1+2\alpha}{4(1+\alpha)}}\sqrt{\log n}\right),$$

where $\alpha = \beta/(d+1)$ under the Hölder source condition and $\alpha = 1/(d+1)$ under the sign source condition. The prediction-space stability quantities entering these bounds are also examined numerically in Section 5.

The remainder of the paper is organized as follows. Section 2 introduces the regression model and RWNN estimators. Section 3 establishes L^2 and sup-norm approximation results. In Section 4, we develop the spectral analysis, estimation bounds, and inferential theory for ridge-regularized RWNNs, including uniform estimation. Section 5 presents numerical experiments, and Section 6 concludes. Proofs of the main results are provided in Appendices A, B, C, and D. Appendix E develops a general sieve framework for estimation and smooth-functional inference. Additional theoretical and numerical results are collected in Appendices F and G, respectively.

1.2 Notation

We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random elements are defined. We write $\mathbb{N} = \{0, 1, 2, \ldots\}$ and $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$. Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm for vectors and the operator 2-norm for matrices. For symmetric matrices, $A \preceq B$ means that $B - A$ is positive semidefinite. We denote by λ_m the Lebesgue measure on $\mathbb{R}^m$ and by V_m the volume of the Euclidean unit ball in $\mathbb{R}^m$. For measurable functions $f : \mathcal{X} \to \mathbb{R}$, we write $\|f\|_{L^2(X)} = (\mathbb{E}_X[|f(X)|^2])^{1/2}$ and $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$, where P_X denotes the marginal distribution of X and $\mathbb{E}_X$ denotes integration with respect to P_X, holding the realized hidden-layer parameters fixed. We denote by P_θ the law of the i.i.d. hidden-layer parameter sequence and write $\mathbb{P}_{A,B}$ and $\mathbb{E}_{A,B}$ (or $\mathbb{E}_\theta$) for probability and expectation with respect to this law. For functions f that may depend on the sampled hidden-layer parameters, we define $\|f\|_{L^2(X,A,B)} := \left(\mathbb{E}_{A,B}[\|f\|^2_{L^2(X)}]\right)^{1/2}$. The stochastic order symbols O_P and o_P refer to the joint law of the observations and random features, whereas O_{P_θ} and o_{P_θ} refer to the law of the random features alone. For positive sequences, $a_n \lesssim b_n$ means $a_n \le C b_n$ for a constant $C > 0$ that does not depend on n, K, or other running indices, and $a_n \asymp b_n$ means $a_n \lesssim b_n \lesssim a_n$. Convergence conditional on the hidden-layer parameters is stated explicitly.

2 Model and random weight neural network sieves

2.1 Nonparametric regression model

Let $\{(X_i, Y_i)\}_{i\ge1}$ be i.i.d. random variables, where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $Y_i \in \mathbb{R}$, with $\mathbb{E}[Y_i^2] < \infty$. We work in the standard nonparametric regression setting, cf. Györfi et al. (2006); Chen (2007); Tsybakov (2009); Belloni et al. (2015); Chen and Christensen (2015). Consider the regression model

$$Y_i = h(X_i) + \epsilon_i, \quad i \in \mathbb{N}^+, \tag{2.1}$$

where the regression function $h : \mathcal{X} \to \mathbb{R}$ is defined by $h(x) := \mathbb{E}[Y_i|X_i = x]$ and the errors $\epsilon_i := Y_i - h(X_i)$ satisfy $\mathbb{E}[\epsilon_i|X_i] = 0$ and $\mathbb{V}[\epsilon_i|X_i] = \sigma^2_\epsilon(X_i) < \infty$ a.s. Moreover, $\mathbb{E}[Y_i^2] < \infty$ implies $h \in L^2(X)$.

The object of interest is the conditional expectation function (CEF) h, which is the minimizer of the population quadratic risk

$$h = \arg\min_{f\in L^2(X)} \mathbb{E}\big[(Y_i - f(X_i))^2\big].$$

Table 1: Selected approximation, estimation, and inference results

Panel A: Approximation results

Reference/result	Setting	Error criterion	Rate
Selected existing results			
[B93]	Trainable features; Barron class	L^2	$K^{-1/2}$
[KB18]	Trainable ReLU features; sign source representation; weighted coefficients	L^∞	$\sqrt{d+\log(K)}\,K^{-1/2-1/d}$
[LPSZ25]	Control-neighbors integration; β-Hölder integrand on a $(d+1)$-dimensional domain	Root MSE	$K^{-1/2-\beta/(d+1)}$
[LMX25]	Random ReLU^m features; integral Sobolev class Λ^q	$L^2(\mathcal{X})$	$(K/\log(K))^{-1/2-(2m+1)/(2d)}$
This paper			
Theorem 3.1	I.i.d. random features; β-Hölder source condition	$L^2(X,A,B)$	$K^{-1/2-\beta/(d+1)}$
Theorem 3.5	Stratified random features; β-Hölder or sign source condition	$\mathbb{E}\|\cdot\|_\infty$	$a_K K^{-1/2-\alpha}$

Panel B: Spectral properties, estimation, and inference

Result	Main conditions	Object	Conclusion
Proposition 4.1	Polynomial covering; Lipschitz features	Spectrum	$\lambda_{\min}(\Gamma_K) \lesssim K^{-2/q_\Pi}$
Corollary 4.1	Spectral decay; suitable ridge scaling	Effective dimension	$\mathcal{N}_{K_n,\lambda} = o(K_n)$
Theorem 4.1	I.i.d. data; ridge-normalized stability	$L^2(X)$ estimation	$O_P\big(\sqrt{\mathcal{N}_{K_n,\lambda}/n} + \Delta_{K_n} + B^{\mathrm{pop}}_{K_n,\lambda}\big)$
Corollary 4.3	Directional nondegeneracy; stability and ridge-bias control	Linear functionals	Gaussian limit allowing $\lambda_{\min}(\Gamma_{K_n}) \to 0$
Corollaries 4.4, 4.5	Directional variance, moment, and growth conditions	Variance estimation	Relative consistency of the ridge-adjusted variance estimator
Corollary 4.7	Directional CLT; negligible functional bias and remainder	Smooth functionals	Gaussian limit under sieve-dependent normalization
Theorem 4.16	Prediction-space and leverage stability; stratified features	L^∞ estimation	$O_P\big(\sqrt{K_n\log(n)/n} + a_{K_n}K_n^{-1/2-\alpha} + \sqrt{\lambda_{K_n,n}}\big)$

In Panel A, [B93], [KB18], [LPSZ25], and [LMX25] refer to Barron (1993), Klusowski and Barron (2018), Leluc et al. (2025), and Liu et al. (2025), respectively. These results concern different function classes, sampling schemes, and error criteria and are not directly comparable.

The [KB18] row reports the weighted-coefficient bound for ReLU; sign-restricted coefficients yield $\sqrt{d+\log(K)}\,K^{-1/2-1/(d+2)}$. For squared ReLU, both bounds have prefactor $\sqrt{d}$. [KB18] uses a d-dimensional parameter set, unlike the $(d+1)$-dimensional set Π in Theorem 3.5; on the d-dimensional set Π^{sph}, the exponent $1/d$ is recovered; see Remark 6.

Root MSE denotes integration error rather than regression error. The [LMX25] rate holds with high probability for $q = (d+2m+1)/2$, $m \geq 1$. In Theorems 3.5 and 4.16, $a_K := \sqrt{d} + \sqrt{\log(K)}$, and $\alpha = \beta/(d+1)$ under the β-Hölder source condition, whereas $\alpha = 1/(d+1)$ under the sign source condition with a ReLU^m activation, $m \in \{1,2\}$.

Panel B lists principal conclusions, subject to the full theorem assumptions. Gaussian limits hold conditionally on $\mathcal{F}_{K_n}$, in P_θ-probability. A single ridge-adjusted variance estimator is relatively consistent for both directional variances $\sigma^{2,(\ell)}_{K_n,\lambda}(\alpha)$, $\ell \in \{0,1\}$, defined in (4.31).

2.2 Shallow neural networks and random weight neural networks

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a fixed measurable activation function. For $K \in \mathbb{N}^+$, the class of shallow neural networks of width K is

$$\mathcal{H}^{\mathrm{NN}}_K := \left\{ x \mapsto \sum_{k=1}^K W_k \sigma(A_k^\top x + B_k) \;\middle|\; W_k \in \mathbb{R},\ A_k \in \mathbb{R}^d,\ B_k \in \mathbb{R} \right\}, \tag{2.2}$$

where all parameters $\{(W_k, A_k, B_k)\}_{k=1}^K$ are trainable (Chen and White, 1999; Shen et al., 2023).

A *random weight neural network* (RWNN) is obtained from (2.2) by fixing the inner weights

$\{(A_k, B_k)\}_{k=1}^K$ at random draws and retaining only the outer weights $\{W_k\}_{k=1}^K$ as trainable parameters. Formally, let $\Pi \subseteq \mathbb{R}^{d+1}$ be a parameter space and let π_θ be a probability measure on Π. Let $\{(A_k, B_k)\}_{k\geq 1}$ be an i.i.d. sequence of random variables such that $\theta_k := (A_k^\top, B_k)^\top \sim \pi_\theta$, independent of the data sequence $\{(X_i, Y_i)\}_{i\geq 1}$. All random elements are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each realization $\omega \in \Omega$, the sequence $\{(A_k(\omega), B_k(\omega))\}_{k\geq 1}$ defines, for every $K \in \mathbb{N}^+$, the *towering RWNN class*

$$\mathcal{H}_K(\omega) := \left\{ x \mapsto \sum_{k=1}^K W_k \sigma(A_k(\omega)^\top x + B_k(\omega)) \;\middle|\; W_k \in \mathbb{R} \right\}. \tag{2.3}$$

Since the same weight sequence is reused, the classes are nested: for any $K_1 \leq K_2$ and all $\omega \in \Omega$

$$\mathcal{H}_{K_1}(\omega) \subseteq \mathcal{H}_{K_2}(\omega).$$

This nesting property aligns with the functional sieves framework of Grenander (1981), and allows direct comparison with nonparametric sieve estimators (Chen, 2007). To simplify notation, in the rest of the paper, we suppress the dependence on ω.

Remark 1 (Aleatory RWNN class)**.** In practice, RWNNs are typically implemented by drawing a new i.i.d. sample $\{(A_k^{(K)}, B_k^{(K)})\}_{k=1}^K$ independently from π_θ for each width K. The resulting *aleatory* class $\overline{\mathcal{H}}_K$ need not be nested in K and therefore does not fit the functional sieve framework of Grenander (1981). The towering construction provides the nesting required by sieve arguments that explicitly exploit this property. The ridge-regularized results developed in Section 4 do not generally require nesting of the feature spaces. For any fixed width K, however, $\mathcal{H}_K$ and $\overline{\mathcal{H}}_K$ have the same distribution under π_θ. Consequently, all approximation results stated for $\mathcal{H}_K$ in Subsection 3.1 also hold for the aleatory class $\overline{\mathcal{H}}_K$ at a fixed K.

We also consider RWNN classes generated by stratified sampling of the inner parameters in Subsections 3.3 and 4.6. For each $K \in \mathbb{N}^+$, let $\{\theta_k^{\mathrm{IS},K}\}_{k=1}^K$, with $\theta_k^{\mathrm{IS},K} = ((A_k^{\mathrm{IS},K})^\top, B_k^{\mathrm{IS},K})^\top$, be random inner parameters generated by a stratified sampling scheme on Π. The corresponding *stratified RWNN class* is

$$\mathcal{H}_K^{\mathrm{IS}} := \left\{ x \mapsto \sum_{k=1}^K W_k \sigma\big((A_k^{\mathrm{IS},K})^\top x + B_k^{\mathrm{IS},K}\big) \;\middle|\; W_k \in \mathbb{R} \right\}. \tag{2.4}$$

The superscript IS is retained to denote this stratified (importance) sampling construction.

Unlike the towering RWNN class $\mathcal{H}_K$, the distribution of the inner parameters in $\mathcal{H}_K^{\mathrm{IS}}$ may depend on K, and the resulting sequence of classes need not be nested. The stratified hidden parameters are generated independently of the regression observations. In case (ii) of Theorem 3.5, the stratification uses the sign partition $\Pi_\pm$ of the source function f in Assumption 9, so the inner parameters depend on the target. The estimation results of Subsection 4.6 are therefore oracle results in that case, unless this partition is known.

2.3 RWNN least squares estimator

Given a sample $\{(Y_i, X_i)\}_{i=1}^n$, we estimate the regression function h by *empirical risk minimization* (ERM) over the towering RWNN class $\mathcal{H}_K$ in (2.3), conditional on the realized inner parameters $\theta := \{(A_k, B_k)\}_{k=1}^K$. Let $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be a loss function. For a network $h_K^{\theta,W}(x) := \sum_{k=1}^K W_k \sigma(A_k^\top x + B_k)$, $W = (W_1, \ldots, W_K)^\top \in \mathbb{R}^K$, the empirical risk is defined as

$$L_n(h_K^{\theta,W}) = \frac{1}{n} \sum_{i=1}^n \ell\left(Y_i, h_K^{\theta,W}(X_i)\right).$$

We focus on the squared loss $\ell(y,t)=(y-t)^2$, while a more general formulation allowing loss functions beyond squared loss is considered in Appendix E. The readout weights are obtained as the empirical risk minimizer, which is the linear least squares solution

$$\widehat{W}_{K,n} \in \arg\min_{W\in\mathbb{R}^K} L_n(h_K^{\theta,W}) = \arg\min_{W\in\mathbb{R}^K} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{k=1}^{K} W_k\sigma(A_k^\top X_i + B_k)\right)^2. \tag{2.5}$$

Following the series (sieve) notation in Chen and Christensen (2015), introduce the random feature map

$$g_K(x) := \Big(\sigma(A_1^\top x + B_1), \ldots, \sigma(A_K^\top x + B_K)\Big)^\top \in \mathbb{R}^K, \tag{2.6}$$

the design matrix and the response vector

$$G_{K,n} := (g_K(X_1), \ldots, g_K(X_n))^\top \in \mathbb{R}^{n\times K}, \quad Y := (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n, \tag{2.7}$$

respectively. Then problem (2.5) admits the minimum-Euclidean-norm solution

$$\widehat{W}_{K,n} = (G_{K,n}^\top G_{K,n})^\dagger G_{K,n}^\top Y, \tag{2.8}$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse. The *RWNN least-squares estimator* $h_{K,n}^{\theta,\widehat{W}} \in \mathcal{H}_K$ of the regression function h, corresponding to the minimum-Euclidean-norm solution (2.8), is defined as

$$h_{K,n}^{\theta,\widehat{W}}(x) := \sum_{k=1}^{K}(\widehat{W}_{K,n})_k\sigma(A_k^\top x + B_k) = g_K(x)^\top \widehat{W}_{K,n} = g_K(x)^\top (G_{K,n}^\top G_{K,n})^\dagger G_{K,n}^\top Y, \ \ x\in\mathcal{X}. \tag{2.9}$$

The principal object of our statistical analysis is the ridge-regularized RWNN least-squares estimator, which we distinguish from the unregularized estimator $h_{K,n}^{\theta,\widehat{W}}$ defined in (2.9). For the ridge-regularized analysis, we use the coefficient notation β in place of W, following the conventional notation for linear sieve regression. For deterministic sequences $K_n \to \infty$ and $\lambda_{K_n,n} > 0$, define $\widehat{h}_{n,\lambda}(x) := g_{K_n}(x)^\top \widehat{\beta}_{K_n,n,\lambda}$, where

$$\widehat{\beta}_{K_n,n,\lambda} := \Big(n^{-1}G_{K_n,n}^\top G_{K_n,n} + \lambda_{K_n,n} I_{K_n}\Big)^{-1} n^{-1} G_{K_n,n}^\top Y. \tag{2.10}$$

The associated population covariance matrix and ridge-normalized quantities are introduced in Section 4.1.1.

3 Approximation properties of RWNNs

In this section we consider different approximation regimes for random weight neural networks under related, but distinct, regularity assumptions.

Subsections 3.1 and 3.3 both assume a direct source condition under which the target function admits an integral representation of the form

$$h(x) = \int_\Pi f(A,B)\sigma(A^\top x + B)\mathrm{d}\pi_\theta(A,B).$$

They differ in terms of how this representation is used. Subsection 3.1 assumes a β-Hölder product source condition (Assumption 3), combining a Lipschitz activation σ (Assumption 2) with a β-Hölder weight function f on a compact parameter set $\Pi \subset \mathbb{R}^{d+1}$ equipped with a metric ρ_Π. It derives L^2 approximation rates via the control-neighbors rule of Leluc et al. (2025). Subsection 3.3 derives sup-norm approximation rates via a stratified sampling construction,

based on partitioning the parameter space and controlling within-cell oscillations. In that setting two direct source conditions are considered: the same β-Hölder product source condition as in Subsection 3.1, and a ReLU^m activation with $m \in \{1, 2\}$ (Assumption 7) combined with a sign weight function f (Assumption 9).

Subsection 3.2 takes a different approach, assuming that h belongs to an integral Sobolev class Λ^q and working with ReLU^m activations for general $m \geq 1$.

Proofs of the L^2 approximation results under the β-Hölder source condition, including the control-neighbors weight bounds, are provided in Appendix A.1. The proof of the $L^2(X)$ approximation corollary for integral Sobolev classes is given in Appendix A.2. Proofs of the stratified sup-norm approximation results and the corresponding output-weight bounds are collected in Appendix A.3, while the uniform Monte Carlo approximation bound is established in Appendix A.4.

We consider shallow random weight neural networks on a compact domain.

Assumption 1. For some $M > 0$, the regressor space is $\mathcal{X} = [-M, M]^d$.

3.1 L^2 approximation under a β-Hölder source condition

Under a β-Hölder source condition, we show that the RWNN sieve admits an L^2 approximation error of order $K^{-1/2-\beta/(d+1)}$.

Whenever the source and regular-sampling assumptions below are invoked jointly, they are understood to hold for the same parameter space Π, metric ρ_Π, and sampling measure π_θ.

Assumption 2 (Lipschitz activation)**.** $\sigma : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant $L_\sigma > 0$.

Assumption 3 (Product source condition)**.** There exist $\beta \in (0, 1]$, a compact set $\Pi \subset \mathbb{R}^{d+1}$ equipped with a metric ρ_Π such that $D_\Pi := \sup_{\theta_1,\theta_2\in\Pi} \rho_\Pi(\theta_1, \theta_2) < \infty$, a probability measure π_θ on Π, constant $L_f > 0$, and a function $f : \Pi \to \mathbb{R}$ satisfying

$$|f(\theta_1) - f(\theta_2)| \leq L_f \rho_\Pi(\theta_1, \theta_2)^\beta, \quad \theta_1, \theta_2 \in \Pi,$$

such that

$$h(x) = \int_\Pi \phi_x(\theta) \mathrm{d}\pi_\theta(\theta), \qquad x \in \mathcal{X},$$

with $\theta = (A^\top, B)^\top$ and $\phi_x(\theta) = f(A, B)\sigma(A^\top x + B)$. We also define $M_f := \|f\|_\infty$, $M_\sigma := \sup_{x\in\mathcal{X},\theta\in\Pi} |\sigma(A^\top x + B)|$.

To establish the approximation result, we use a general integral representation with a Hölder continuous integrand, together with regularity conditions on the sampling distribution. We subsequently verify that the RWNN product source condition implies this abstract representation under a suitable compatibility condition on the metric.

Assumption 4 (Integral representation)**.** There exist a compact set $\Pi \subset \mathbb{R}^{d+1}$ equipped with a metric ρ_Π, a probability measure π_θ on Π, and $\beta \in (0, 1]$ such that for each $x \in \mathcal{X}$ there exists a measurable function $\phi_x : \Pi \to \mathbb{R}$ such that $\theta \mapsto \phi_x(\theta)$ is β-Hölder continuous on (Π, ρ_Π) with constant $L_{\phi,x} > 0$, and

$$h(x) = \int_\Pi \phi_x(\theta) \mathrm{d}\pi_\theta(\theta), \qquad x \in \mathcal{X}.$$

Assumption 5 (Regular sampling)**.** Let Π, ρ_Π, and π_θ be as in Assumption 4. Suppose that π_θ admits a density $\mu_{\pi,\theta}$ with respect to λ_{d+1} satisfying, for constants $c_b, c_u, c_\pi, r_0 > 0$:

(i) $c_b \leq \mu_{\pi,\theta}(\theta) \leq c_u$ for all $\theta \in \Pi$,

(ii) $\lambda_{d+1}(\Pi \cap B_r^{\|\cdot\|}(\theta)) \geq c_\pi \lambda_{d+1}(B_r^{\|\cdot\|}(\theta))$ for all $r \in (0, r_0]$, $\theta \in \Pi$, where $B_r^{\|\cdot\|}(\theta) \subset \mathbb{R}^{d+1}$ denotes the Euclidean ball of radius r centered at θ,

(iii) There exist constants $0 < m_\rho \le M_\rho < \infty$ such that $m_\rho\|\theta_1-\theta_2\| \le \rho_\Pi(\theta_1,\theta_2) \le M_\rho\|\theta_1-\theta_2\|$, $\theta_1,\theta_2 \in \Pi$.

Remark 2 (Choice of metric ρ_Π)**.** Two natural choices of ρ_Π arise in this paper.

(i) *Euclidean metric*: $\rho_\Pi(\theta_1,\theta_2) = \|\theta_1 - \theta_2\|$.

(ii) *Weighted ℓ_1 metric*: $\rho_\Pi(\theta_1,\theta_2) = d_\Pi(\theta_1,\theta_2) := M\|A_1 - A_2\|_1 + |B_1 - B_2|$, with M as in Assumption 1, used in Subsection 3.3. No regular sampling condition is required there, and, as we show later, this metric is a convenient choice in that setting.

The two metrics are equivalent on $\mathbb{R}^{d+1}$ with constants depending only on M and d, so a β-Hölder condition formulated with respect to one metric is equivalent to a β-Hölder condition with respect to the other, up to a change in the Hölder constant. Indeed, since $\|A_1 - A_2\| \le \|A_1 - A_2\|_1 \le \sqrt{d}\|A_1 - A_2\|$, one has

$$\begin{aligned}\min(M,1)\|\theta_1 - \theta_2\| \le d_\Pi(\theta_1,\theta_2) &\le M\sqrt{d}\|A_1 - A_2\| + |B_1 - B_2| \\ &\le \sqrt{2}\max(M\sqrt{d},1)\|\theta_1 - \theta_2\|, \quad \theta_1,\theta_2 \in \Pi,\end{aligned}$$

so the Hölder constant transforms by a factor of at most $\big(\sqrt{2}\,\max(M\sqrt{d},1)/\min(M,1)\big)^\beta$.

Notice that, under Assumption 3, $M_f < \infty$ by the β-Hölder continuity of f and the finite ρ_Π-diameter of Π, and $D_\Pi < \infty$ by definition. Moreover, under Assumptions 1, 2, and 3, $M_\sigma < \infty$ by compactness of $\mathcal{X} \times \Pi$ and continuity of σ.

Proposition 3.1. *Let Assumptions 1, 2, and 3 hold, and suppose there exists a constant $L_\rho > 0$ such that*

$$|(A_1 - A_2)^\top x + (B_1 - B_2)| \le L_\rho\,\rho_\Pi(\theta_1,\theta_2) \qquad \text{for all } \theta_1,\theta_2 \in \Pi,\ x \in \mathcal{X}. \tag{3.1}$$

Then Assumption 3 implies Assumption 4 with Hölder constant that can be chosen uniformly in $x \in \mathcal{X}$ by

$$L_{\phi,x} \le L_\phi := M_f L_\sigma L_\rho D_\Pi^{1-\beta} + M_\sigma L_f,$$

where D_Π and M_σ are as in Assumption 3.

Remark 3. For the two particular choices of ρ_Π in Remark 2, the compatibility condition (3.1) holds with explicit constants. For the Euclidean metric $\rho_\Pi(\theta_1,\theta_2) = \|\theta_1 - \theta_2\|$, one may take $L_\rho = \sup_{x\in\mathcal{X}} \|(x^\top,1)^\top\| = \sqrt{dM^2+1}$, while for the weighted ℓ_1 metric $\rho_\Pi = d_\Pi$, one may take $L_\rho = 1$.

To construct an approximation based on finitely many sampled features, we build upon the control-neighbors Monte Carlo rule of Leluc et al. (2025). This rule exploits the Hölder regularity of the integrand and the regularity of the sampling distribution to improve upon the standard Monte Carlo rate. The result used and the verification of its sampling conditions are given in Appendix A.1.

The control-neighbors rule assigns weights to the sampled parameters $\theta_1,\dots,\theta_K$ according to

$$\omega_{k,K} := \frac{1 + KV_{K,k} - \widehat{d}_{K,k}}{K}, \tag{3.2}$$

where $V_{K,k} := \pi_\theta(C_{K,k})$ is the probability mass of the Voronoi cell $C_{K,k}$ associated with θ_k, and $\widehat{d}_{K,k}$ is the number of times θ_k is the nearest neighbor of θ_j, $j \ne k$, with respect to ρ_Π.

Since the rule is linear in its integrand, applying it to the product source representation yields

$$h_K^{\theta,\omega\text{-cnn}}(x) := \sum_{k=1}^K \omega_{k,K}\big[f(A_k,B_k)\sigma(A_k^\top x + B_k)\big] = \sum_{k=1}^K W_k\sigma(A_k^\top x + B_k), \tag{3.3}$$

where $W_k := \omega_{k,K} f(A_k,B_k)$, $k = 1,\dots,K$. Hence, conditional on the sampled hidden parameters $\{\theta_k\}_{k=1}^K$, the control-neighbors approximation belongs to the RWNN class.

Theorem 3.1 (L^2 RWNN approximation)**.** *Let Assumptions 1, 2, 5, and 3 hold, and suppose the compatibility condition (3.1) is satisfied. Let $\theta_k = (A_k^\top, B_k)^\top$, $k = 1, \dots, K$, be i.i.d. from π_θ. Then, for $K \geq 4$, the control-neighbors weight vector $W \in \mathbb{R}^K$ defined in (3.3) is $\sigma(\theta_1, \dots, \theta_K)$-measurable, and the corresponding RWNN approximant $h_K^{\theta,W} = h_K^{\theta,\omega\text{-cnn}}$ satisfies*

$$\|h(X) - h_K^{\theta,W}(X)\|_{L^2(X,A,B)} \leq C_q K^{-1/2-\beta/(d+1)}, \tag{3.4}$$

where $C_q > 0$ depends only on d, β, M_f, L_f, L_σ, M_σ, L_ρ, c_b, c_u, c_π, r_0, m_ρ, M_ρ, and D_Π.

Remark 4 (Approximation rates and comparison with Monte Carlo integration)**.** A standard empirical-process argument based on symmetrization and Dudley's entropy integral applied to the centered process $x \mapsto h(x) - h_K^\theta(x)$, where $h_K^\theta(x) = K^{-1}\sum_{k=1}^K \phi_x(\theta_k)$, yields only the Monte Carlo rate

$$\mathbb{E}_\theta\left[\sup_{x\in\mathcal{X}} |h(x) - h_K^\theta(x)|\right] = \mathcal{O}(K^{-1/2});$$

see Appendix A.4 for a complete derivation. This rate does not exploit the regularity of $\theta \mapsto \phi_x(\theta)$ on Π.

The sharper rate in (3.4) arises from the β-Hölder regularity of $\theta \mapsto \phi_x(\theta)$ on (Π, ρ_Π), whose metric small-ball behavior is of order r^{d+1} under Assumption 5. The control-neighbors estimator leverages the local spacing of $\theta_1, \dots, \theta_K$ to construct weights $\omega_{k,K}$ that reduce the error by $K^{-\beta/(d+1)}$ relative to the Monte Carlo baseline. Since Assumption 5 (iii) makes ρ_Π bi-Lipschitz equivalent to the Euclidean metric, the corresponding β-Hölder classes are equivalent up to constants. Hence, for $(d+1)$-dimensional β-Hölder integrands, the minimax rate under i.i.d. sampling is $\mathcal{O}(K^{-1/2-\beta/(d+1)})$ (Novak, 2016), so Theorem 3.1 achieves the optimal rate (up to constants) for this class of integrands under i.i.d. sampling. In particular, for $\beta = 1$ the rate $K^{-1/2-1/(d+1)}$ is optimal for Lipschitz integrands on a $(d+1)$-dimensional parameter space.

Theorem 3.2 (High-probability L^2 approximation)**.** *Under the assumptions of Theorem 3.1, let $\theta_k = (A_k^\top, B_k)^\top$, $k = 1, \dots, K$, be i.i.d. from π_θ. Then for every $K \geq 4$, there exists a $\sigma(\theta_1, \dots, \theta_K)$-measurable weight vector $W \in \mathbb{R}^K$ such that for every $\varepsilon \in (0, 1)$,*

$$\mathbb{P}_{A,B}\left(\|h(X) - h_K^{\theta,W}(X)\|_{L^2(X)} \leq C_q K^{-1/2-\beta/(d+1)}\varepsilon^{-1/2}\right) \geq 1 - \varepsilon, \tag{3.5}$$

where $C_q > 0$ is the constant appearing in Theorem 3.1.

Corollary 3.1 (RWNN approximation in probability)**.** *Under the assumptions of Theorem 3.2,*

$$\inf_{g\in\mathcal{H}_K} \|h - g\|_{L^2(X)} = O_{P_\theta}\left(K^{-1/2-\beta/(d+1)}\right).$$

The preceding approximation results impose no constraint on the output weights. To establish the same approximation rate for the constrained RWNN sieve (Corollary E.1), which will be used in Section 4, we need to control the Euclidean norm of the control-neighbors weight vector. The following geometric assumption on nearest-neighbor indegrees allows us to derive the required bound.

Assumption 6. There exists a constant $\psi_\Pi < \infty$ such that for every finite point set $P \subset \Pi$ and every $x \in P$ the number of points in P for which x is a nearest neighbor under ρ_Π is at most ψ_Π.

We note that in the case of Euclidean metric ρ_Π, by Lemma F.1, Assumption 6 holds with $\psi_\Pi = \psi_{d+1}$, where ψ_{d+1} is the $(d+1)$-dimensional kissing number.

Lemma 3.3 (ℓ_2-norm bound for control-neighbors weights)**.** *Let Assumptions 1, 5, 3, and 6 hold, and let $\theta_k = (A_k^\top, B_k)^\top$, $k = 1, \dots, K$, be i.i.d. from π_θ with $K \geq 3$. Define the weight vector $W \in \mathbb{R}^K$ by*

$$W_k := \omega_{k,K} f(A_k, B_k), \quad k = 1, \dots, K,$$

where f is as in Assumption 3 and the control-neighbors weights $\omega_{k,K}$ are defined in (3.2). Then, for every $\alpha > 0$ there exist $C_\alpha, C'_\alpha > 0$ such that

$$\mathbb{P}_{A,B}\left(\|W\| \le C_\alpha M_f \sqrt{\frac{\log(K)}{K}}\right) \ge 1 - C'_\alpha K^{-\alpha},$$

where M_f is as in Assumption 3. In particular,

$$\|W\| = O_{P_\theta}\left(\sqrt{\frac{\log K}{K}}\right).$$

Moreover, it holds

$$\|W\|_1 \le M_f(2 + \psi_\Pi) \quad \text{and} \quad \|W\| \le M_f\sqrt{3(2 + \psi_\Pi^2)}, \tag{3.6}$$

where $\psi_\Pi < \infty$ denotes the uniform nearest-neighbor indegree bound corresponding to ρ_Π.

3.2 L^2 approximation on integral Sobolev classes

Recent work by Liu et al. (2025) derives approximation rates for shallow ReLU^m networks under additional smoothness assumptions. These results improve on the sampling-based argument used in Subsection 3.1 by working with Barron- and Sobolev-type classes of higher regularity. The underlying idea of relying on suitable integral representations is similar; such representations are closely related to ridgelet analysis (see, e.g., Sonoda and Murata, 2017). Moreover, Liu et al. (2025) establish approximation rates for RWNNs with parameters sampled uniformly from the sphere, which further motivates their use in our nonparametric regression setting. Using ridgelet representations, Mehmood and Gonon (2026) obtain RWNN approximation rates of order $K^{-1/2}$ in time-dependent Sobolev norms for smooth activations such as tanh and the sigmoid, with hidden-layer parameters drawn from a fixed heavy-tailed distribution; related Sobolev-norm bounds for randomized networks are given in De Ryck et al. (2025). The results in this subsection, as well as those in Subsection 3.3, rely on the following assumption on the activation function.

Assumption 7. For $m \in \mathbb{N}^+$, let $\sigma_m : \mathbb{R} \to \mathbb{R}$ be the ReLU^m activation function $\sigma_m(x) = \max(x, 0)^m$.

Subsection 3.3 uses this assumption with $m \in \{1, 2\}$.

For $m \in \mathbb{N}^+$, we denote by $\mathcal{H}_K^{(m)}$ the RWNN class (2.3) under Assumption 7, that is with activation $\sigma_m(x) = \max(x, 0)^m$. The case $m = 1$ coincides with $\mathcal{H}_K$ when the activation in (2.3) is chosen to be ReLU.

For $q := (d + 2m + 1)/2$, define the class

$$\Lambda^q(\mathcal{X}) := \left\{ x \mapsto \fint_{\mathbb{S}^d} \phi_x^m(\theta) f(\theta) d\theta \;\middle|\; f \in L^2(\mathbb{S}^d) \right\},$$

where $\fint_{\mathbb{S}^d} := |\mathbb{S}^d|^{-1} \int_{\mathbb{S}^d}$ denotes the surface-normalized integral and $\phi_x^m(\theta) := \sigma_m(\theta_{1:d}^\top x + \theta_0)$ with $\theta = (\theta_{1:d}^\top, \theta_0)^\top \in \mathbb{S}^d$. We equip $\Lambda^q(\mathcal{X})$ with the norm

$$\|h\|_{\Lambda^q(\mathcal{X})} := \inf\left\{ \|f\|_{L^2(\mathbb{S}^d)} \;\middle|\; h(x) = \fint_{\mathbb{S}^d} \phi_x^m(\theta) f(\theta)\, d\theta,\ x \in \mathcal{X} \right\}.$$

By Theorem 2.3 of Liu et al. (2025), $\Lambda^q(\mathcal{X})$ coincides with the classical Sobolev space $H^q(\mathcal{X})$, with equivalent norms. This representation is the key structural property that enables random-feature approximation.

Let $\mu_{\mathbb{S}^d}$ denote the uniform probability measure on $\mathbb{S}^d$, let $\theta_1, \dots, \theta_K \overset{\text{i.i.d.}}{\sim} \mu_{\mathbb{S}^d}$, and write $\theta := (\theta_1, \dots, \theta_K)$. Define the random feature map $\phi^m_{\theta,K}(x) := \big(\phi^m_x(\theta_1), \dots, \phi^m_x(\theta_K)\big)^\top \in \mathbb{R}^K$. Then any function in $\mathcal{H}^{(m)}_K$ can be written in the form $x \mapsto W^\top \phi^m_{\theta,K}(x)$ for some $W \in \mathbb{R}^K$.

The following result, due to Liu et al. (2025), establishes that ReLU^m RWNNs with parameters sampled uniformly from $\mathbb{S}^d$ achieve an $L^2(\mathcal{X})$ approximation rate of $(K/\log(K))^{-1/2-(2m+1)/(2d)}$ with high probability over the parameter draw. Corollary 3.2 below extends this to the $L^2(X)$ norm required for our statistical analysis.

Theorem 3.4 (Liu et al., 2025, Theorem 6.2)**.** *Let Assumption 1 hold, let $h \in \Lambda^q(\mathcal{X})$ with $q = (d+2m+1)/2$, and let $\theta_1, \dots, \theta_K \overset{\text{i.i.d.}}{\sim} \mu_{\mathbb{S}^d}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, there exists a weight vector $W \in \mathbb{R}^K$ such that*

$$\|W\| \lesssim \left(\frac{K}{\log(K/\delta)}\right)^{-\frac{1}{2}} \|h\|_{\Lambda^q(\mathcal{X})}$$

and

$$\|h - W^\top \phi^m_{\theta,K}\|_{L^2(\mathcal{X})} \lesssim \left(\frac{K}{\log(K/\delta)}\right)^{-\frac{1}{2}-\frac{2m+1}{2d}} \|h\|_{\Lambda^q(\mathcal{X})}.$$

The implicit constants depend only on d and m, and the domain $\mathcal{X}$.

Remark 5. Theorem 3.4 follows from an approximation bound expressed in terms of the covering radius of the sampled parameters on $\mathbb{S}^d$. To make this precise, define the covering radius

$$\widetilde{\rho}_{K,d} := \max_{\theta \in \mathbb{S}^d} \min_{1 \le k \le K} \rho(\theta, \theta_k),$$

where $\rho : \mathbb{S}^d \times \mathbb{S}^d \to [0, \pi]$ denotes the spherical geodesic metric. The points $\{\theta_k\}_{k=1}^K$ are said to be *well-distributed* on $\mathbb{S}^d$ if $\widetilde{\rho}_{K,d} \le C_d K^{-1/d}$ for some constant $C_d > 0$. The central approximation result of Liu et al. (2025), Theorem 2.2, provides a bound on the approximation error that depends explicitly on $\widetilde{\rho}_{K,d}$. For i.i.d. uniformly sampled parameters, the covering radius instead admits high-probability bounds of order $(\log(K/\delta)/K)^{1/d}$, which account for the logarithmic factor in Theorem 3.4.

Theorem 3.4 establishes the approximation rate in $L^2(\mathcal{X})$, that is, with respect to Lebesgue measure on $\mathcal{X}$. Choosing $\delta = K^{-\alpha}$ gives a polynomial failure probability for any $\alpha > 0$, and the following corollary transfers the resulting bound to $L^2(X)$ under a bounded density condition on P_X.

Corollary 3.2. *Under the assumptions of Theorem 3.4, suppose additionally that P_X admits a density p_X with respect to Lebesgue measure on $\mathcal{X}$ satisfying $\|p_X\|_\infty \le \overline{M} < \infty$. Then, for every $\alpha > 0$ and $K \ge 2$, with probability at least $1 - K^{-\alpha}$ with respect to the i.i.d. uniform sample $\theta_1, \dots, \theta_K$ on $\mathbb{S}^d$, there exists $W \in \mathbb{R}^K$ such that*

$$\|W\| \lesssim \sqrt{\frac{\log(K)}{K}} \|h\|_{\Lambda^q(\mathcal{X})} \quad \text{and} \quad \|h - W^\top \phi^m_{\theta,K}\|_{L^2(X)} \lesssim \left(\frac{K}{\log(K)}\right)^{-\frac{1}{2}-\frac{2m+1}{2d}} \|h\|_{\Lambda^q(\mathcal{X})},$$

where the implicit constants depend only on $\overline{M}$, α, d, m, and M.

3.3 Sup-norm approximation via stratified sampling

In this subsection, we derive sup-norm approximation bounds for RWNNs via a stratified sampling scheme for the hidden-layer parameters, building on Klusowski and Barron (2018). The results complement Theorem 3.1: both subsections assume an integral representation of h of the form $h(x) = \int_\Pi f(A,B)\sigma(A^\top x + B) d\pi_\theta$, but here we obtain L^∞ rather than L^2 bounds,

under different regularity assumptions on f (Hölder continuity or sign structure) and without requiring the regular sampling condition (Assumption 5).

The achieved rate is determined by the regularity of f with respect to θ. In all cases considered below, the approximation error admits a single centered empirical process representation, and the rate follows from bounding the within-cell oscillation of $\phi_x(\theta) := f(A,B)\sigma(A^\top x + B)$. When f is β-Hölder continuous (Assumption 3) with $\beta \in (0,1]$, the within-cell oscillation scales as r^β, yielding the rate $K^{-1/2-\beta/(d+1)}$ on Π. In particular, the Lipschitz case $\beta = 1$ achieves the exponent $1/(d+1)$. When f is sign-valued (Assumption 9), the partition is constructed to respect the sign boundary of f, so that f is constant on each cell $B_j \subset \Pi_\pm$. Within each cell, $\phi_x(\theta)$ is therefore Lipschitz in θ due to the Lipschitz continuity of σ, and the same one-term argument applies. This yields the rate $K^{-1/2-1/(d+1)}$.

Thus, the key structural feature of the analysis is the availability of a single centered empirical process representation of the error, which avoids the need for additional decomposition or balancing arguments.

Assumption 8 (Parameter space)**.** For constants $C_A, C_B > 0$, define

$$\Pi := \{A \in \mathbb{R}^d : \|A\|_1 \le C_A\} \times [-C_B, C_B] \subset \mathbb{R}^{d+1},$$

and let π_θ be a probability measure on Π.

Assumption 9 (Direct source condition, sign f)**.** Let Assumptions 1, 7 and 8 hold with $m \in \{1,2\}$. There exist a constant $M_f > 0$ and a measurable function $f : \Pi \to \{-M_f, +M_f\}$ such that

$$h(x) = \int_\Pi f(A,B)\sigma_m(A^\top x + B)\mathrm{d}\pi_\theta(A,B), \qquad x \in \mathcal{X}.$$

Under Assumption 9, the measure π_θ may be chosen as part of the source representation of h and may therefore depend on the target function. Moreover, the stratified construction in case (ii) of Theorem 3.5 requires a partition respecting the sign sets $\Pi_\pm := \{\theta \in \Pi : f(\theta) = \pm M_f\}$. Consequently, this construction is oracle unless the sampling measure and the sign sets are known.

We now describe the stratified sampling construction used in both settings. The idea is to partition the parameter space into strata of controlled diameter and sample independently within each stratum according to the corresponding conditional distribution. Weighting the contribution of each stratum by its probability mass preserves unbiasedness while allowing us to exploit the reduced within-stratum variation of the integrand.

For sufficiently large K, set $\widetilde{K} := \lfloor K/2 \rfloor$ and consider a measurable partition $\{B_j\}_{j=1}^J$ of Π into $J \le \widetilde{K}$ cells of controlled d_Π diameter. The reduced allocation budget allows us to ensure at least one sample per stratum while keeping the total number of sampled parameters at most K. Under the sign-valued source condition, the partition is additionally chosen so that each cell is contained in one of the sign sets $\Pi_\pm$. Let $P_j := \pi_\theta(B_j)$ and, for $P_j > 0$, define $\pi_{\theta,j}(E) := \pi_\theta(E \cap B_j)/P_j$. For cells of zero probability mass, the choice of the conditional sampling law is not important. Let $K_j \in \{\lfloor \widetilde{K} P_j \rfloor, \lceil \widetilde{K} P_j \rceil\}$ be independent random allocations satisfying $\mathbb{E}[K_j] = \widetilde{K} P_j$, and set $L_j := K_j + \mathbb{1}_{\{K_j = 0\}}$. Conditional on the allocations, draw $\theta_{j,1}, \dots, \theta_{j,L_j}$ independently from $\pi_{\theta,j}$, independently across cells. Writing $\theta_{j,\ell} = (A_{j,\ell}^\top, B_{j,\ell})^\top$ and $\overline{K} := \sum_{j=1}^J L_j$, define

$$h_{\overline{K}}^{\mathrm{IS}}(x) := \sum_{j=1}^J \frac{P_j}{L_j} \sum_{\ell=1}^{L_j} f(\theta_{j,\ell})\widetilde{\sigma}(A_{j,\ell}^\top x + B_{j,\ell}), \tag{3.7}$$

where $\widetilde{\sigma} = \sigma$ under the β-Hölder source condition and $\widetilde{\sigma} = \sigma_m$ under the sign-valued source condition. Hence, the corresponding output weights are $W_{j,\ell} := P_j f(\theta_{j,\ell})/L_j$. The partitions

are chosen so that $J \le \widetilde{K}$, which gives $\overline{K} \le \widetilde{K} + J \le K$. Appending neurons with zero output weights yields an element of $\mathcal{H}_K^{\mathrm{IS}}$. The finitely many smaller values of K are handled separately in the proof.

Theorem 3.5 (Sup-norm approximation via stratified sampling)**.** *Let Assumption 1 and 8 hold. Suppose one of the following conditions holds:*

(i) (β-Hölder continuous f, Lipschitz σ) *Assumptions 2 and 3 hold for the same* Π *and* π_θ *as in Assumption 8, with*

$$\rho_\Pi(\theta_1, \theta_2) = d_\Pi(\theta_1, \theta_2) := M\|A_1 - A_2\|_1 + |B_1 - B_2|, \qquad \theta_1, \theta_2 \in \Pi,$$

where M *is the constant from Assumption 1.*

(ii) (Sign f, ReLU^m σ) *Assumption 9 holds with* $m \in \{1, 2\}$.

Then, for every $K \in \mathbb{N}^+$, *there exist random inner parameters* $\{(A_k, B_k)\}_{k=1}^K$ *defining the stratified RWNN class* $\mathcal{H}_K^{\mathrm{IS}}$, *and random coefficients* $W = (W_1, \dots, W_K)^\top \in \mathbb{R}^K$, *measurable with respect to the random allocation variables and the sampled parameters, such that the stratified sampling approximation*

$$h_K^{\mathrm{IS}}(x) := \sum_{k=1}^K W_k \widetilde{\sigma}(A_k^\top x + B_k) \in \mathcal{H}_K^{\mathrm{IS}},$$

where $\widetilde{\sigma} = \sigma$ *in case (i), and* $\widetilde{\sigma} = \sigma_m$ *with* $m \in \{1, 2\}$ *in case (ii), satisfies*

$$\mathbb{E}\left[\sup_{x \in \mathcal{X}} \left|h_K^{\mathrm{IS}}(x) - h(x)\right|\right] \le C(\sqrt{d} + \sqrt{\log(K)})K^{-1/2-\alpha}, \tag{3.8}$$

where

$$\alpha = \begin{cases} \beta/(d+1), & \text{in case (i)}, \\ 1/(d+1), & \text{in case (ii)}. \end{cases} \tag{3.9}$$

The constant $C > 0$ *in (3.8) in case (i) depends only on* d, C_A, C_B, *and* M, *as well as on* M_f, M_σ, L_σ, L_f, *and* β *and is specified in (A.26) in the proof. In case (ii)* $C > 0$ *depends only on* d, C_A, C_B, M, m, *and* M_f *and is given in (A.31) in the proof.*

Remark 6 (Faster rates on the ℓ_1-sphere)**.** In both cases of Theorem 3.5, if the corresponding source representation and regularity assumptions hold on the ℓ_1-sphere $\Pi^{\mathrm{sph}} = \{\|A\|_1 = C_A\} \times [-C_B, C_B]$ instead of the ball $\Pi = \{\|A\|_1 \le C_A\} \times [-C_B, C_B]$, the effective dimension of the parameter space decreases from $d+1$ to d, leading to the following improvements of the rate exponent α:

Case	f	σ	α on ball	α on sphere
(i)	β-Hölder, $\beta \in (0, 1]$	Lipschitz	$\beta/(d+1)$	β/d
(ii)	Sign	ReLU^m, $m \in \{1, 2\}$	$1/(d+1)$	$1/d$

The improved exponents follow from the covering properties of Π^{sph}. Indeed, the ℓ_1-sphere $\{A \in \mathbb{R}^d : \|A\|_1 = C_A\}$ is the union of 2^d facets of dimension $d-1$. Including the bias coordinate, each facet of Π^{sph} admits a Lipschitz parametrization by a bounded subset of $\mathbb{R}^d$. Consequently, for sufficiently small $r > 0$, $\mathcal{N}(r, \Pi^{\mathrm{sph}}, d_\Pi) \le C_{\mathrm{sph}} r^{-d}$, where $C_{\mathrm{sph}} > 0$ depends only on d, M, C_A, and C_B. In case (ii), covering the two sign sets separately preserves this order. Substituting these covering bounds into the stratified approximation argument yields the exponents β/d and $1/d$, respectively.

For only bounded f without additional regularity, one should not expect an improvement beyond the classical Monte Carlo rate $K^{-1/2}$ from metric stratification alone: the within-cell

oscillation of $\phi_x(\theta) = f(A,B)\sigma(A^\top x + B)$ does not shrink with the cell diameter in the absence of regularity of f. Thus the improved exponents in case (i) rely essentially on the Hölder continuity assumption; if f is only bounded, the natural benchmark is the $K^{-1/2}$ rate both on the ball and on the sphere.

Remark 7 (Comparison with Klusowski-Barron Rates)**.** Klusowski and Barron (2018) consider approximation under a sign-valued representation analogous to Assumption 9, where the inner parameters satisfy $\|A\|_1 = 1$ and $B \in [-1, 0]$. For the ReLU^m activation with $m \in \{1, 2\}$, their Theorem 1 gives (up to a $\sqrt{d + \log(K)}$ factor) the rate $K^{-1/2-1/(d+2)}$ when the coefficients b_k in their finite-network representation are restricted to $b_k \in \{-1, +1\}$, and the rate $K^{-1/2-1/d}$ when $b_k \in [-1, 1]$. The latter is the appropriate comparison with our weighted construction and coincides with the exponent $1/d$ obtained here on Π^{sph}. For the $b_k \in \{-1, +1\}$ construction, their proof relies on a two-term decomposition of the error into an allocation (or rounding) component and a centered fluctuation term, and the rate follows from balancing these contributions. When $b_k \in [-1, 1]$, the allocation is chosen so that the former term vanishes. The weighted stratified estimator considered here admits an exact centered representation of the approximation error, so that no allocation term arises. As a consequence, the analysis reduces to a single empirical process bound, and the rate is driven by the within-cell oscillation of the integrand. Here, in the sign-valued case, our construction provides a different derivation of the same exponent $1/d$ under the present assumptions. More generally, Theorem 3.5 also covers the β-Hölder source condition of case (i), allows approximation on the full $(d+1)$-dimensional parameter space Π, where the corresponding exponent is $1/(d+1)$ in the sign-valued case, and records ℓ_1 and ℓ_2 bounds on the resulting RWNN output weights used in the subsequent statistical analysis.

Remark 8. For the equal weight Monte Carlo approximant based on i.i.d. sampling from π_θ, the classical rate $K^{-1/2}$ is obtained under the conditions of Proposition A.1. The stratified sampling scheme of Theorem 3.5 improves this to $K^{-1/2-\beta/(d+1)}$ in case (i) and $K^{-1/2-1/(d+1)}$ in case (ii), by allocating neurons proportionally to the cell masses $\pi_\theta(B_j)$, hence ensuring uniform coverage of the parameter space Π and reducing within-cell variability. In case (ii), the sign structure of f guarantees that it is constant on each cell, yielding Lipschitz behavior within cells regardless of the global regularity of f, and thus the full exponent $1/(d+1)$ is achieved. The gains of $K^{-\beta/(d+1)}$ and $K^{-1/(d+1)}$, respectively, over the equal weight Monte Carlo case reflect the reduction in within-cell variability achieved by stratified sampling, as also exploited in the approximation constructions of Klusowski and Barron (2018).

Theorem 3.6 (High-probability sup-norm approximation)**.** *Let the assumptions of Theorem 3.5 hold. Then for every $K \in \mathbb{N}^+$, there exist random inner parameters $\{(A_k, B_k)\}_{k=1}^K$ defining the stratified RWNN class $\mathcal{H}_K^{\mathrm{IS}}$ in (2.4) and a random vector $W = (W_1, \ldots, W_K)^\top \in \mathbb{R}^K$, measurable with respect to the random allocation variables and the sampled parameters, such that for every $\varepsilon \in (0, 1)$,*

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}} \left|h_K^{\mathrm{IS}}(x) - h(x)\right| \le C(\sqrt{d} + \sqrt{\log(K)})K^{-1/2-\alpha}\varepsilon^{-1}\right) \ge 1 - \varepsilon, \tag{3.10}$$

where $C > 0$ is the constant appearing in Theorem 3.5, and α is as in (3.9).

Remark 9. Compared to the L^2 high-probability bound in Theorem 3.2, where the probability penalty is $\varepsilon^{-1/2}$, the sup-norm bound provides a penalty of ε^{-1}. This reflects the fact that Markov's inequality is applied directly to the sup-norm rather than to its square, and is the standard cost of converting an expectation bound on a non-negative random variable into a high-probability statement without additional second-moment or tail assumptions.

Corollary 3.3 (Stratified RWNN approximation in probability)**.** *Under the assumptions of Theorem 3.6,*

$$\inf_{g\in\mathcal{H}_K^{\mathrm{IS}}} \|h - g\|_{L^\infty(\mathcal{X})} = O_P\left((\sqrt{d} + \sqrt{\log(K)})K^{-1/2-\alpha}\right),$$

where α is as in (3.9). The probability is taken with respect to the random allocation variables and the sampled parameters in the stratified construction.

The following lemma establishes bounds on the output weights of the stratified RWNN approximation. These bounds are subsequently used in the sup-norm analysis of ridge-regularized RWNN estimators in Subsection 4.6.

Lemma 3.7 (ℓ_1- and ℓ_2-norm bounds for stratified sampling weights)**.** *Let the assumptions of Theorem 3.5 hold, and let*

$$h_K^{\mathrm{IS}}(x) = \sum_{k=1}^{K} W_k \widetilde{\sigma}(A_k^\top x + B_k) \in \mathcal{H}_K^{\mathrm{IS}},$$

be the stratified RWNN approximation, where $\widetilde{\sigma} = \sigma$ in case (i) and $\widetilde{\sigma} = \sigma_m$ in case (ii). Then its output-weight vector $W \in \mathbb{R}^K$ satisfies

$$\|W\|_1 \le M_f, \tag{3.11}$$

and

$$\|W\| \le \sqrt{6}\, M_f K^{-1/2}. \tag{3.12}$$

In particular,

$$\|W\|_1 = O(1) \qquad \text{and} \qquad \|W\| = O(K^{-1/2}).$$

Remark 10 (Comparison with control-neighbors weights)**.** Lemma 3.7 and Lemma 3.3 together show that the stratified weights are better controlled than the control-neighbors weights in two respects. First, the stratified construction yields the deterministic ℓ_2 bound, while the control-neighbors bound holds with high probability. Second, the stratified weights satisfy $\|W\| \le \sqrt{6}\, M_f K^{-1/2}$, improving on the order $\|W\| = O_P(M_f\sqrt{\log(K)/K})$ of Lemma 3.3 by a factor of $\sqrt{\log(K)}$.

4 Ridge-regularized RWNN estimation and inference

This section develops the statistical theory of ridge-regularized least-squares estimation with RWNN sieves. We proceed from the spectral properties of random features to statistical estimation and inference. More explicitly, in Section 4.1 we show the spectral deterioration of RWNN sieves with growing sieve dimension, characterize the role played by ridge normalization and the spectrum of the population sieve covariance matrix in determining the effective dimension of the estimation problem by introducing the so-called ridge effective dimension and deriving a bound for it. Subsection 4.2 analyzes ridge-normalized covariance stability and uses it to obtain $L^2(X)$ estimation bounds. Subsection 4.3 develops directional inference through linearization and asymptotic normality, allowing the smallest population covariance eigenvalue to vanish. We then establish consistency of a residual-based ridge-adjusted variance estimator, enabling feasible studentized inference, before extending the results to smooth functionals in Subsection 4.5. Finally, Subsection 4.6 addresses uniform estimation, first through spectral stability and then through prediction-space and leverage-based conditions.

Complementing the ridge-regularized analysis, Appendix E develops a general framework for constrained sieve estimation and asymptotic inference on smooth functionals, accommodating loss functions *beyond squared-loss regression*. The constrained RWNN approximation results established there are also used in the ridge-regularized estimation and inference analysis of this section.

Proofs of the results on spectral properties and $L^2(X)$ estimation are collected in Appendix B, the ones that concern directional and smooth-functional inference and feasible variance estimation are provided in Appendix C, and, finally, the reader can find the proofs of our results on uniform estimation in Appendix D.

4.1 Spectral deterioration and effective dimension

We first introduce the ridge-normalized geometry of the population feature covariance matrix. We then establish spectral deterioration for parametrized random-feature sieves and derive corresponding bounds on ridge effective dimension. These results provide the spectral foundation for the estimation and inference theory developed below.

4.1.1 Ridge normalization and population covariance

We first introduce the ridge-normalized quantities underlying the $L^2(X)$ analysis. This normalization allows us to formulate the relevant stability and complexity measures without requiring the population sieve covariance matrix to remain uniformly well-conditioned as the sieve dimension K_n increases.

Recall the notation of Section 2.3, with $K = K_n$. In particular, the feature map g_{K_n} is defined in (2.6), the design matrix $G_{K_n,n}$, and the response vector Y are defined in (2.7). We also write $e := (\epsilon_1, \dots, \epsilon_n)^\top \in \mathbb{R}^n$. The ridge-regularized RWNN estimator $\widehat{h}_{n,\lambda}$ and its coefficient vector $\widehat{\beta}_{K_n,n,\lambda}$ are defined in (2.10). We write

$$\widehat{\Gamma}_{K_n,n,\lambda} := \widehat{\Gamma}_{K_n,n} + \lambda_{K_n,n} I_{K_n} \quad \text{and} \quad \widehat{\Gamma}_{K_n,n} := n^{-1} G_{K_n,n}^\top G_{K_n,n}. \tag{4.1}$$

Define the population sieve covariance matrix and its ridge-regularized counterpart by

$$\Gamma_{K_n} := \mathbb{E}_X[g_{K_n}(X) g_{K_n}(X)^\top] \quad \text{and} \quad \Gamma_{K_n,\lambda} := \Gamma_{K_n} + \lambda_{K_n,n} I_{K_n}, \tag{4.2}$$

respectively. The $L^2(X)$-projection coefficient of h onto the linear span of the sieve is given by

$$\beta_{K_n} \in \operatorname*{argmin}_{b \in \mathbb{R}^{K_n}} \mathbb{E}_X\big[(h(X) - g_{K_n}(X)^\top b)^2\big]. \tag{4.3}$$

If the minimizer is not unique, we choose β_{K_n} to be the minimum-Euclidean-norm solution. We denote the corresponding approximation error by $\Delta_{K_n} := \|h - g_{K_n}^\top \beta_{K_n}\|_{L^2(X)}$. We also introduce the ridge-normalized feature map

$$\widetilde{g}_{K_n}(x) := \Gamma_{K_n,\lambda}^{-1/2} g_{K_n}(x). \tag{4.4}$$

Additionally, define the uniform norms of the original and ridge-normalized feature maps by

$$\xi_{K_n} := \sup_{x \in \mathcal{X}} \|g_{K_n}(x)\| \quad \text{and} \quad \widetilde{\xi}_{K_n,\lambda} := \sup_{x \in \mathcal{X}} \|\widetilde{g}_{K_n}(x)\|, \tag{4.5}$$

respectively. Finally, let s_j^2, $j = 1, \dots, K_n$, be the eigenvalues of the population sieve covariance matrix Γ_{K_n} and define

$$\overline{s}_{K_n}^2 := \lambda_{\min}(\Gamma_{K_n}). \tag{4.6}$$

Since the ridge regularization is isotropic, $\lambda_{\min}(\Gamma_{K_n,\lambda}) = \overline{s}_{K_n}^2 + \lambda_{K_n,n}$.

4.1.2 Spectral deterioration of random feature sieves

The smallest population eigenvalue $\overline{s}_{K_n}^2$ defined in (4.6) need not remain bounded away from zero as the sieve dimension K_n increases. More generally, for parameterized sieve families, the covering and Lipschitz conditions below yield explicit control of the spectral tail of the population sieve covariance matrix. The following bound is closely related to classical connections between metric approximation, Kolmogorov widths, and spectral decay (see Ismagilov (1968) and the recent treatment in Takhanov (2026), in particular the proofs of Theorem 1 and Theorems 3-4). We give the direct finite-dimensional argument below since it applies immediately to the population sieve covariance matrix.

Proposition 4.1. *Let (Π, ρ_Π) be a compact metric space such that, for some constants $C_\Pi, q_\Pi > 0$, there exists a constant $\overline{r} > 0$ such that, for all $r < \overline{r}$, there exists an $r/2$-net $\mathcal{N}_{r/2}$ of Π under ρ_Π satisfying*

$$|\mathcal{N}_{r/2}| \le C_\Pi r^{-q_\Pi}. \tag{4.7}$$

Let $\{\phi_\theta \mid \theta \in \Pi\} \subset L^2(X)$ satisfy, for some $L_\phi > 0$,

$$\|\phi_{\theta_1} - \phi_{\theta_2}\|_{L^2(X)} \le L_\phi \rho_\Pi(\theta_1, \theta_2), \qquad \theta_1, \theta_2 \in \Pi. \tag{4.8}$$

For arbitrary $\theta_1, \dots, \theta_K \in \Pi$, define

$$g_K(x) := \big(\phi_{\theta_1}(x), \dots, \phi_{\theta_K}(x)\big)^\top, \qquad \Gamma_K := \mathbb{E}_X[g_K(X) g_K(X)^\top],$$

and let $s_1^2 \ge \cdots \ge s_K^2$ denote the eigenvalues of Γ_K.

Then, for all sufficiently large K and $m < K$,

$$\sum_{j=m+1}^{K} s_j^2 \le C_{\text{tail}} K m^{-2/q_\Pi}, \tag{4.9}$$

where $C_{\text{tail}} = L_\phi^2 C_\Pi^{2/q_\Pi}$ is independent of K and $\theta_1, \dots, \theta_K$. Moreover, for all sufficiently large $j \le K$,

$$s_j^2 \le C_{\text{eig}} K j^{-1-2/q_\Pi}, \tag{4.10}$$

where $C_{\text{eig}} = 2 \cdot 3^{2/q_\Pi} C_{\text{tail}}$. In particular, for all sufficiently large K,

$$\lambda_{\min}(\Gamma_K) \le C_{\text{eig}} K^{-2/q_\Pi}. \tag{4.11}$$

The bounds in Proposition 4.1 are deterministic and hold for arbitrary $\theta_1, \dots, \theta_K \in \Pi$ whenever K is large. The upper bound on $\lambda_{\min}(\Gamma_K)$ shows that uniform lower spectral conditioning cannot, in general, be maintained for growing parameterized sieves. This motivates formulating the subsequent analysis in terms of ridge-normalized quantities that depend on the spectrum beyond its smallest eigenvalue.

4.1.3 Ridge effective dimension

Using the ridge-normalized feature map $\widetilde{g}_{K_n}$ defined in (4.4), we introduce the *ridge effective dimension* of the sieve as

$$\mathcal{N}_{K_n,\lambda} := \mathbb{E}_X\left[\|\widetilde{g}_{K_n}(X)\|^2\right] = \operatorname{tr}\left(\Gamma_{K_n} \Gamma_{K_n,\lambda}^{-1}\right) = \sum_{j=1}^{K_n} \frac{s_j^2}{s_j^2 + \lambda_{K_n,n}}. \tag{4.12}$$

The spectral-tail bound in Proposition 4.1 then gives control of the effective dimension.

Corollary 4.1. *Under the conditions of Proposition 4.1, let $K_n \to \infty$ and $\lambda_{K_n,n} > 0$ satisfy*

$$\frac{K_n}{\lambda_{K_n,n}} \to \infty \qquad \text{and} \qquad \lambda_{K_n,n} K_n^{2/q_\Pi} \to \infty.$$

Then, for all sufficiently large n,

$$\mathcal{N}_{K_n,\lambda} \le \min\left\{K_n, 1 + C_{\mathcal{N}} \left(\frac{C_{\text{tail}} K_n}{\lambda_{K_n,n}}\right)^{q_\Pi/(q_\Pi+2)}\right\}, \tag{4.13}$$

where $C_{\mathcal{N}} = \left(\frac{2}{q_\Pi}\right)^{q_\Pi/(q_\Pi+2)} + \left(\frac{q_\Pi}{2}\right)^{2/(q_\Pi+2)}$.

Consequently,

$$\mathcal{N}_{K_n,\lambda} = o(K_n). \tag{4.14}$$

4.2 L^2 estimation

We now study the $L^2(X)$ estimation error of the ridge-regularized sieve estimator $\widehat{h}_{n,\lambda}$. We first establish ridge-normalized covariance stability and derive a general estimation bound in terms of the ridge effective dimension, sieve approximation error, and population ridge bias. We then specialize these results to RWNN sieves, obtaining explicit convergence rates under the approximation conditions developed in Section 3.

4.2.1 Ridge-normalized covariance stability

We start by imposing a condition on the ridge-normalized feature norm $\widetilde{\xi}_{K_n,\lambda}$ defined in (4.5), ensuring stability of the empirical covariance matrix in ridge-normalized coordinates.

Assumption 10 (Ridge-normalized covariance matrix stability)**.** It holds

$$\frac{\widetilde{\xi}^2_{K_n,\lambda}\log(K_n)}{n} = o(1).$$

We notice that, since

$$\widetilde{\xi}_{K_n,\lambda} \le \frac{\xi_{K_n}}{\sqrt{\overline{s}^2_{K_n} + \lambda_{K_n,n}}},$$

with ξ_{K_n} and $\overline{s}^2_{K_n}$ defined in (4.5) and (4.6), respectively, Assumption 10 is implied by

$$\frac{\xi^2_{K_n}\log(K_n)}{n\left(\overline{s}^2_{K_n} + \lambda_{K_n,n}\right)} = o(1). \tag{4.15}$$

Remark 11. Assumption 10 is a ridge-normalized analogue of the standard condition $\xi^2_{K_n}/n \to 0$ and its logarithmic variants commonly imposed in sieve least squares estimation (Newey, 1997). Widely used well-conditioned sieves—such as splines, wavelets, and other localized series bases—are such that $\overline{s}^2_{K_n} \ge \underline{c}_\Gamma > 0$ and $\xi_{K_n} \lesssim \sqrt{K_n}$ (see, for example, Chen, 2007; Chen and Christensen, 2015; Belloni et al., 2015). In such cases, (4.15) is implied by

$$\frac{\xi^2_{K_n}\log(K_n)}{n} = o(1),$$

which, when $\xi_{K_n} \lesssim \sqrt{K_n}$, is in turn implied by $K_n\log(K_n)/n = o(1)$.

More generally, when $\overline{s}^2_{K_n}$ is allowed to vanish, the sufficient condition (4.15) may be used to verify Assumption 10. This sufficient condition may be conservative as it controls all feature directions through the smallest eigenvalue of Γ_{K_n}. To see why Assumption 10 implies empirical covariance matrix stability, observe that

$$\Gamma^{-1/2}_{K_n,\lambda}\left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n}\right)\Gamma^{-1/2}_{K_n,\lambda} = \frac{1}{n}\sum_{i=1}^n\left(\widetilde{g}_{K_n}(X_i)\widetilde{g}_{K_n}(X_i)^\top - \mathbb{E}_X\left[\widetilde{g}_{K_n}(X_i)\widetilde{g}_{K_n}(X_i)^\top\right]\right).$$

A matrix-valued law of large numbers, such as Lemma F.2, together with

$$\left\|\Gamma^{-1/2}_{K_n,\lambda}\Gamma_{K_n}\Gamma^{-1/2}_{K_n,\lambda}\right\| \le 1$$

and

$$\left\|\widetilde{g}_{K_n}(X_i)\widetilde{g}_{K_n}(X_i)^\top\right\| = \|\widetilde{g}_{K_n}(X_i)\|^2 \le \widetilde{\xi}^2_{K_n,\lambda}$$

yields

$$\mathbb{E}\left[\left\|\Gamma^{-1/2}_{K_n,\lambda}\left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n}\right)\Gamma^{-1/2}_{K_n,\lambda}\right\|\right] \lesssim \frac{\widetilde{\xi}^2_{K_n,\lambda}\log(K_n)}{n} + \sqrt{\frac{\widetilde{\xi}^2_{K_n,\lambda}\log(K_n)}{n}}.$$

Therefore, Assumption 10 and Markov's inequality imply

$$\left\|\Gamma_{K_n,\lambda}^{-1/2}\left(\widehat{\Gamma}_{K_n,n}-\Gamma_{K_n}\right)\Gamma_{K_n,\lambda}^{-1/2}\right\| = O_P\left(\sqrt{\frac{\widetilde{\xi}_{K_n,\lambda}^2\log(K_n)}{n}}\right) = o_P(1). \tag{4.16}$$

This resulting ridge-normalized covariance matrix stability is the key perturbation condition used in Theorem 4.1.

For later use, define the ridge-normalized covariance perturbation matrix and its associated inverse by

$$\begin{aligned} A_{K_n,n,\lambda} &:= \Gamma_{K_n,\lambda}^{-1/2}(\widehat{\Gamma}_{K_n,n}-\Gamma_{K_n})\Gamma_{K_n,\lambda}^{-1/2}, \\ \widehat{D}_{K_n,n,\lambda} &:= \Gamma_{K_n,\lambda}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\Gamma_{K_n,\lambda}^{1/2} = (I_{K_n}+A_{K_n,n,\lambda})^{-1}. \end{aligned} \tag{4.17}$$

In particular, (4.16) implies $\|A_{K_n,n,\lambda}\| = o_P(1)$ and hence $\|\widehat{D}_{K_n,n,\lambda}\| = O_P(1)$.

4.2.2 L^2 estimation bounds

We now derive an $L^2(X)$ error bound for the ridge-regularized least-squares sieve estimator. We explicitly separate and analyze the sieve approximation error, the stochastic estimation error, and the effect of ridge regularization.

Assumption 11 (Regression errors)**.** The regression errors satisfy $\mathbb{E}[\epsilon_i \mid X_i] = 0$. Moreover, there exists a constant $\overline{\sigma}^2 < \infty$ such that $\mathbb{E}[\epsilon_i^2 \mid X_i] \le \overline{\sigma}^2$.

Theorem 4.1 (A refined $L^2(X)$ error bound)**.** *Let $\{Z_i\}_{i\in\mathbb{N}^+}$ be i.i.d. observations satisfying Assumption 11. Let $\{g_{K_n}:\mathcal{X}\to\mathbb{R}^{K_n}\}_{n\ge1}$ be a sequence of feature maps satisfying Assumption 10. Define the population ridge bias*

$$B_{K_n,\lambda}^{\text{pop}} := \lambda_{K_n,n}\left\|\Gamma_{K_n}^{1/2}\Gamma_{K_n,\lambda}^{-1}\beta_{K_n}\right\|. \tag{4.18}$$

Then

$$\left\|\widehat{h}_{n,\lambda}-h\right\|_{L^2(X)} = O_P\left(\Delta_{K_n}+\sqrt{\frac{\mathcal{N}_{K_n,\lambda}}{n}}+B_{K_n,\lambda}^{\text{pop}}\right). \tag{4.19}$$

The population ridge bias defined in (4.18) admits a direct spectral interpretation. Let $\overline{\beta}_{j,K_n}$ denote the coordinates of β_{K_n} with respect to an eigenbasis of Γ_{K_n}. Then

$$\left(B_{K_n,\lambda}^{\text{pop}}\right)^2 = \lambda_{K_n,n}^2\sum_{j=1}^{K_n}\frac{s_j^2}{\left(s_j^2+\lambda_{K_n,n}\right)^2}\overline{\beta}_{j,K_n}^2.$$

Hence, the population ridge bias is measured in the $L^2(X)$ prediction norm and weights each coefficient direction according to the corresponding eigenvalue of the population covariance matrix. Moreover,

$$B_{K_n,\lambda}^{\text{pop}} \le \frac{\lambda_{K_n,n}}{\sqrt{\overline{s}_{K_n}^2+\lambda_{K_n,n}}}\|\beta_{K_n}\|.$$

Therefore, a naïve smallest-eigenvalue bound provides a sufficient, but potentially conservative, upper bound on the population ridge bias.

Since $\mathcal{N}_{K_n,\lambda} \le K_n$, Theorem 4.1 implies

$$\|\widehat{h}_{n,\lambda}-h\|_{L^2(X)} = O_P\left(\Delta_{K_n}+\sqrt{\frac{K_n}{n}}+B_{K_n,\lambda}^{\text{pop}}\right)$$

and hence if $B^{\mathrm{pop}}_{K_n,\lambda} = O_P\left(\Delta_{K_n} + \sqrt{K_n/n}\right)$, the ridge estimator achieves the classical sieve estimation rate. Nonetheless, under the conditions of Corollary 4.1, $\mathcal{N}_{K_n,\lambda} = o(K_n)$ and, hence, the stochastic term in Theorem 4.1, $\sqrt{\mathcal{N}_{K_n,\lambda}/n}$, is asymptotically strictly smaller than its classical sieve counterpart, which has order $\sqrt{K_n/n}$.

We conclude this subsection by providing the result, which will be useful later in the analysis of the variance estimator. The following lemma quantifies an empirical prediction bound, which is derived from the same coefficient decomposition used in the proof of Theorem 4.1.

Lemma 4.2 (Empirical prediction bound for ridge sieves)**.** *Let the assumptions of Theorem 4.1 hold. Then*

$$\left\|\widehat{\Gamma}^{1/2}_{K_n,n}\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right)\right\| = O_P\left(\Delta_{K_n} + B^{\mathrm{pop}}_{K_n,\lambda} + \sqrt{\frac{\mathcal{N}_{K_n,\lambda}}{n}}\right).$$

4.2.3 RWNN specialization and primitive rates

We now specialize Theorem 4.1 to the RWNN regression setting. Although the ridge estimator is not constrained to belong to $\mathcal{H}_{K_n}(M_n)$, the constrained approximation constructed in Subsection E.2 can be used to control both the sieve approximation error Δ_{K_n} and the population ridge bias $B^{\mathrm{pop}}_{K_n,\lambda}$. Indeed, $\mathcal{H}_{K_n}(M_n)$ is contained in a full linear span $\left\{x \mapsto g_{K_n}(x)^\top \beta : \beta \in \mathbb{R}^{K_n}\right\}$ and hence

$$\Delta_{K_n} = \inf_{\beta \in \mathbb{R}^{K_n}} \|h - g_{K_n}^\top \beta\|_{L^2(X)} \;\le\; \inf_{\{W \in \mathbb{R}^{K_n} \mid \|W\| \le M_n\}} \|h - g_{K_n}^\top W\|_{L^2(X)}$$

Consequently, under the source condition of Corollary E.1, $\Delta_{K_n} = O_{P_\theta}\left(K_n^{-1/2-\beta/(d+1)}\right)$.

The covering properties of the parameter space underlying this approximation result also have a direct implication for the global spectral conditioning of the RWNN sieve. In particular, the polynomial covering bound used in Section 3 implies that, as the number of features grows, the smallest population covariance matrix eigenvalue cannot remain uniformly bounded away from zero.

Corollary 4.2 (Smallest population eigenvalue of the RWNN sieve)**.** *Let Assumptions 1 and 2 hold, let $\Pi \subset \mathbb{R}^{d+1}$ be compact, and suppose that (3.1) holds. For $\theta_k = (A_k^\top, B_k)^\top \in \Pi$, $k = 1, \dots, K$, define $g_K(x) := \left(\sigma(A_1^\top x + B_1), \dots, \sigma(A_K^\top x + B_K)\right)^\top$. Then*

$$\overline{s}^2_K \lesssim K^{-2/(d+1)}.$$

The bound holds for every realization of the hidden-layer parameters and hence, in particular, almost surely when the parameters are randomly sampled.

Corollary 4.2 shows that a uniform positive lower bound on the smallest population covariance eigenvalue cannot be maintained for the compact RWNN parameterizations considered here. This provides additional motivation for the ridge-normalized stability condition in Assumption 10, which does not require $\overline{s}^2_{K_n}$ to remain bounded away from zero. We next use the constrained RWNN approximant to control the population ridge bias $B^{\mathrm{pop}}_{K_n,\lambda}$ defined in (4.18).

Lemma 4.3 (Population ridge bias bound for RWNNs)**.** *Under the conditions of Corollary E.1,*

$$B^{\mathrm{pop}}_{K_n,\lambda} = O_{P_\theta}\left(K_n^{-1/2-\beta/(d+1)} + \sqrt{\frac{\lambda_{K_n,n}\log(K_n)}{K_n}}\right).$$

Combining the approximation bound for Δ_{K_n} with Theorem 4.1 and the preceding result for ridge bias in Lemma 4.3, together with the corresponding RWNN conditions ensuring ridge-normalized covariance matrix stability, yields the following rate for the ridge-regularized RWNN least-squares estimator.

Theorem 4.4. *Let $\{Z_i\}_{i\in\mathbb{N}^+}$ be i.i.d. observations satisfying $Y_i = h(X_i)+\epsilon_i$, $\mathbb{E}[\epsilon_i \mid X_i] = 0$, and $\mathbb{E}[\epsilon_i^2 \mid X_i] \le \bar{\sigma}^2 < \infty$. Let the conditions of Corollary E.1 hold and let $K_n \to \infty$ and $\lambda_{K_n,n} > 0$ be deterministic sequences. Assume moreover that the defining conditions of Assumption 10 hold in P_θ-probability along the chosen sequences K_n and $\lambda_{K_n,n}$. Then*

$$\|\widehat{h}_{n,\lambda} - h\|_{L^2(X)} = O_P\left(\sqrt{\frac{\mathcal{N}_{K_n,\lambda}}{n}} + K_n^{-1/2-\beta/(d+1)} + \sqrt{\frac{\lambda_{K_n,n}\log(K_n)}{K_n}}\right). \tag{4.20}$$

Furthermore, let $\beta \in (0,1)$ and suppose that

$$\lambda_{K_n,n} \asymp K_n^{-r}, \qquad \frac{2\beta}{d+1} < r < \frac{2}{d+1}. \tag{4.21}$$

Then

$$\mathcal{N}_{K_n,\lambda} = O\left(K_n^{\nu_r}\right), \qquad \nu_r := \frac{(d+1)(1+r)}{d+3} < 1, \tag{4.22}$$

and

$$\|\widehat{h}_{n,\lambda} - h\|_{L^2(X)} = O_P\left(\sqrt{\frac{K_n^{\nu_r}}{n}} + K_n^{-1/2-\beta/(d+1)}\right). \tag{4.23}$$

If

$$K_n \asymp n^{\frac{1}{1+\nu_r+2\beta/(d+1)}},$$

then

$$\|\widehat{h}_{n,\lambda} - h\|_{L^2(X)} = O_P\left(n^{-\frac{1/2+\beta/(d+1)}{1+\nu_r+2\beta/(d+1)}}\right). \tag{4.24}$$

4.3 Directional inference for linear functionals

In this subsection, we develop directional inference for ridge-regularized RWNN sieve estimators, including irregular functionals such as point evaluation. We begin by establishing a ridge-normalized linearization for fixed feature maps, separating the leading stochastic term from the remainders arising from empirical covariance perturbations, sieve approximation, and ridge regularization. We then derive directional asymptotic normality under conditions ensuring that these remainders are negligible relative to the corresponding standard deviation. Importantly, we show how this *general result can be specialized to RWNN sieves* by conditioning on the sampled hidden-layer parameters and formulating sufficient conditions in terms of RWNN approximation properties, ridge-normalized stability, and directional nondegeneracy characterized through a ridge-normalized Rayleigh quotient. The resulting Gaussian limits are centered at the corresponding sieve projection coefficients, while inference on functionals of h additionally requires control of the sieve approximation bias.

4.3.1 Ridge-normalized linearization

We start by extending the linearization result in Belloni et al. (2015) to the ridge-regularized RWNN least squares estimator. This provides a route to inference for *linear* sieve functionals, including point evaluation, and is complementary to the smooth-functional CLT developed in Section E.3. In contrast to the *regular* functionals considered there, the linear sieve functionals studied here need not be regular in the sense of Van der Vaart (1991), that is, with respect to the limiting $L^2(X)$ geometry. Point evaluation is a prime example of this case, as its sieve Riesz representer generally depends on K_n and need not be bounded.

The results below are first established for fixed feature maps and subsequently applied conditionally on the sampled hidden-layer parameters. Throughout this subsection we use the notation from Subsection 4.1, where β_{K_n} denotes the population projection coefficient defined

in (4.3). We define $g_i := g_{K_n}(X_i)$, $\widetilde{g}_i := \widetilde{g}_{K_n}(X_i) = \Gamma_{K_n,\lambda}^{-1/2} g_i$ using the definition of $\widetilde{g}_{K_n}$ in (4.4), and $\eta_i := \eta_{K_n}(X_i) = h(X_i) - g_{K_n}(X_i)^\top \beta_{K_n}$. For any random variable V_i, define

$$\mathbb{G}_n[V_i] := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (V_i - \mathbb{E}[V_i]).$$

For a unit vector $\alpha_{K_n} \in \mathbb{S}^{K_n-1}$, we consider the linear functional $v \mapsto \alpha_{K_n}^\top v$, $v \in \mathbb{R}^{K_n}$. Point evaluation at $x \in \mathcal{X}$ corresponds to $\alpha_{K_n} = g_{K_n}(x)$ and is obtained from the result below by applying it to the normalized direction $g_{K_n}(x)/\|g_{K_n}(x)\|$ and rescaling. We also define $\widetilde{\alpha}_{K_n,\lambda} := \Gamma_{K_n,\lambda}^{-1/2} \alpha_{K_n}$, which will be used to express the linearization in ridge-normalized coordinates. Whenever there is no possibility of confusion, we further streamline notation and write α and $\widetilde{\alpha}$ in place of α_{K_n} and $\widetilde{\alpha}_{K_n,\lambda}$, respectively.

Lemma 4.5 (Linearization). *Assume that the observations are i.i.d. and satisfy Assumption 11. Suppose further that*

$$\left\| \Gamma_{K_n,\lambda}^{-1/2} \left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n} \right) \Gamma_{K_n,\lambda}^{-1/2} \right\| = o_P(1). \tag{4.25}$$

Then, for every $\alpha \in \mathbb{S}^{K_n-1}$, the ridge estimator error admits the first-order expansion

$$\sqrt{n}\, \alpha^\top \left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n} \right) = \widetilde{\alpha}^\top \mathbb{G}_n[\widetilde{g}_i \epsilon_i] + R_{1,n}(\alpha) + R_{2,n}(\alpha) + R_{\lambda,n}(\alpha), \tag{4.26}$$

where

$$\begin{aligned} R_{1,n}(\alpha) &:= \widetilde{\alpha}^\top \left(\widehat{D}_{K_n,n,\lambda} - I_{K_n} \right) \mathbb{G}_n[\widetilde{g}_i(\epsilon_i + \eta_i)], \\ R_{2,n}(\alpha) &:= \widetilde{\alpha}^\top \mathbb{G}_n[\widetilde{g}_i \eta_i], \\ R_{\lambda,n}(\alpha) &:= -\sqrt{n}\, \lambda_{K_n,n} \widetilde{\alpha}^\top \widehat{D}_{K_n,n,\lambda} \Gamma_{K_n,\lambda}^{-1/2} \beta_{K_n}. \end{aligned} \tag{4.27}$$

Moreover, it holds

$$R_{1,n}(\alpha) = O_P\left(\|\widetilde{\alpha}\| \left\| \Gamma_{K_n,\lambda}^{-1/2} \left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n} \right) \Gamma_{K_n,\lambda}^{-1/2} \right\| \left(1 + \widetilde{\xi}_{K_n,\lambda} \Delta_{K_n} \right) \right), \tag{4.28}$$

$$R_{2,n}(\alpha) = O_P\left(\|\widetilde{\alpha}\| \widetilde{\xi}_{K_n,\lambda} \Delta_{K_n} \right), \tag{4.29}$$

and

$$R_{\lambda,n}(\alpha) = O_P\left(\sqrt{n} \lambda_{K_n,n} \|\widetilde{\alpha}\| \|\Gamma_{K_n,\lambda}^{-1/2} \beta_{K_n}\| \right). \tag{4.30}$$

4.3.2 Directional asymptotic normality

We can now provide an asymptotic normality result based on the ridge-normalized linearization derived in Lemma 4.5. For $\ell \in \{0, 1\}$, define

$$\Omega_{K_n,\lambda}^{(\ell)} := \Gamma_{K_n,\lambda}^{-1} \mathbb{E}\left[(\epsilon_i + \ell \eta_i)^2 g_i g_i^\top \right] \Gamma_{K_n,\lambda}^{-1},$$

and, for $\alpha \in \mathbb{R}^{K_n}$, define the corresponding directional variance by

$$\sigma_{K_n,\lambda}^{2,(\ell)}(\alpha) := \alpha^\top \Omega_{K_n,\lambda}^{(\ell)} \alpha = \widetilde{\alpha}^\top \mathbb{E}\left[(\epsilon_i + \ell \eta_i)^2 \widetilde{g}_i \widetilde{g}_i^\top \right] \widetilde{\alpha}. \tag{4.31}$$

Notice that the specification $\ell = 0$ involves the regression errors, while $\ell = 1$ also incorporates the sieve approximation error. The latter is particularly relevant for the residual-based variance estimator studied in Subsection 4.4.

Theorem 4.6 (Directional asymptotic normality for linear functionals). *Let the assumptions of Lemma 4.5 hold, and let $\alpha_{K_n} \in \mathbb{S}^{K_n-1}$ be a prescribed sequence of linear functional directions. Fix $\ell \in \{0, 1\}$ and suppose that $\sigma_{K_n,\lambda}^{2,(\ell)}(\alpha_{K_n}) > 0$ for all sufficiently large n and define*

$$\kappa_{K_n,\lambda} := \frac{\|\widetilde{\alpha}_{K_n,\lambda}\|}{\sigma_{K_n,\lambda}^{(\ell)}(\alpha_{K_n})}.$$

Assume, along this sequence, that:

(i) $n^{-1/2}\kappa_{K_n,\lambda}\widetilde{\xi}_{K_n,\lambda} = o(1)$.

(ii) $\kappa_{K_n,\lambda}\left\|\Gamma_{K_n,\lambda}^{-1/2}\left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n}\right)\Gamma_{K_n,\lambda}^{-1/2}\right\|\left(1+\widetilde{\xi}_{K_n,\lambda}\Delta_{K_n}\right) = o_P(1)$.

(iii) $(1-\ell)\kappa_{K_n,\lambda}\widetilde{\xi}_{K_n,\lambda}\Delta_{K_n} = o(1)$.

(iv) $\sqrt{n}\kappa_{K_n,\lambda}\lambda_{K_n,n}\|\Gamma_{K_n,\lambda}^{-1/2}\beta_{K_n}\| = o(1)$.

(v) $\mathbb{E}[\epsilon_i^2 \mid X_i] \geq \underline{\sigma}^2 > 0$ *almost surely.*

(vi) $\lim_{M\to\infty}\sup_{n\geq 1}\sup_{x\in\mathcal{X}}\mathbb{E}\big[(\epsilon_i + \ell\eta_{K_n}(x))^2\,\mathbb{1}\{|\epsilon_i+\ell\eta_{K_n}(x)| > M\} \mid X_i = x\big] = 0$.

Then

$$\sqrt{n}\frac{\alpha_{K_n}^\top\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right)}{\sigma_{K_n,\lambda}^{(\ell)}(\alpha_{K_n})} \xrightarrow{d} \mathcal{N}(0,1).$$

4.3.3 RWNN functional directions

For the i.i.d. RWNN construction, let $\mathcal{F}_{K_n} := \sigma(\theta_1,\ldots,\theta_{K_n})$ denote the σ-field generated by the sampled hidden-layer parameters. The prescribed functional directions $\alpha_{K_n} \in \mathbb{S}^{K_n-1}$ considered below are $\mathcal{F}_{K_n}$-measurable.

Define the *directional Rayleigh quotient* of the ridge-normalized population covariance matrix by

$$\mathcal{R}_{K_n,\lambda}(\alpha_{K_n}) := \frac{\alpha_{K_n}^\top\Gamma_{K_n,\lambda}^{-1}\Gamma_{K_n}\Gamma_{K_n,\lambda}^{-1}\alpha_{K_n}}{\alpha_{K_n}^\top\Gamma_{K_n,\lambda}^{-1}\alpha_{K_n}} = \frac{\widetilde{\alpha}_{K_n}^\top\Gamma_{K_n,\lambda}^{-1/2}\Gamma_{K_n}\Gamma_{K_n,\lambda}^{-1/2}\widetilde{\alpha}_{K_n}}{\|\widetilde{\alpha}_{K_n}\|^2}. \tag{4.32}$$

We observe that in (4.32) we can introduce $\mathfrak{a}_{K_n} := \widetilde{\alpha}_{K_n}/\|\widetilde{\alpha}_{K_n}\|$ with $\mathfrak{a}_{K_n} \in \mathbb{S}^{K_n-1}$, so that $\mathcal{R}_{K_n,\lambda}(\alpha_{K_n}) = \mathfrak{a}_{K_n}^\top\Gamma_{K_n,\lambda}^{-1/2}\Gamma_{K_n}\Gamma_{K_n,\lambda}^{-1/2}\mathfrak{a}_{K_n}$. Thus, letting $\bar{\mathfrak{a}}_{j,K_n}$ be the coefficients of $\mathfrak{a}_{K_n}$ in an eigenbasis of Γ_{K_n}, one has

$$\mathcal{R}_{K_n,\lambda}(\alpha_{K_n}) = \sum_{j=1}^{K_n}\frac{s_j^2}{s_j^2+\lambda_{K_n,n}}\bar{\mathfrak{a}}_{j,K_n}^2. \tag{4.33}$$

Corollary 4.3 (Asymptotic normality for RWNN linear functionals)**.** *Let $\{Z_i\}_{i\in\mathbb{N}^+}$ be i.i.d. observations satisfying Assumption 11 and let the conditions of Corollary E.1 hold for some $\beta > 0$, so that*

$$\Delta_{K_n} = O_{P_\theta}\left(K_n^{-1/2-\beta/(d+1)}\right). \tag{4.34}$$

Let $K_n \to \infty$ and $\lambda_{K_n,n} > 0$ be deterministic sequences, and suppose that the defining condition of Assumption 10 holds in P_θ-probability. Assume moreover that $\mathbb{E}[\epsilon_i^2 \mid X_i] \geq \underline{\sigma}^2 > 0$ almost surely and $\sup_{x\in\mathcal{X}}\mathbb{E}\left[\epsilon_i^2\mathbb{1}\{|\epsilon_i| > M\} \mid X_i = x\right] \to 0$ as $M\to\infty$. Suppose that $\mathcal{R}_{K_n,\lambda}(\alpha_{K_n}) > 0$ with P_θ-probability tending to one and that

$$\mathcal{R}_{K_n,\lambda}(\alpha_{K_n})^{-1/2} = O_{P_\theta}(1), \qquad \widetilde{\xi}_{K_n,\lambda} = o_{P_\theta}\left(K_n^{1/2+\beta/(d+1)}\right), \tag{4.35}$$

and

$$\sqrt{n}\lambda_{K_n,n}\left\|\Gamma_{K_n,\lambda}^{-1/2}\beta_{K_n}\right\| = o_{P_\theta}(1). \tag{4.36}$$

Suppose, in addition, that

$$K_n^{1+2\beta/(d+1)} = O(n). \tag{4.37}$$

Then, conditionally on $\mathcal{F}_{K_n}$, in P_θ-probability, for both $\ell \in \{0,1\}$

$$\sqrt{n}\frac{\alpha_{K_n}^\top\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right)}{\sigma_{K_n,\lambda}^{(\ell)}(\alpha_{K_n})} \xrightarrow{d} \mathcal{N}(0,1). \tag{4.38}$$

Remark 12 (Admissible choices of sieve dimension)**.** Condition (4.37) allows different choices of the sieve dimension. In particular, the choice

$$K_n \asymp n^{\frac{d+1}{2(\beta+d+1)}}$$

satisfies (4.37) for every $\beta > 0$. Moreover, under the polynomial-ridge regime of Theorem 4.4, with $\beta \in (0,1)$ and $\nu_r = \frac{(d+1)(1+r)}{d+3} < 1$, the effective-dimension choice

$$K_n \asymp n^{\frac{1}{1+\nu_r+2\beta/(d+1)}}$$

also satisfies (4.37), since

$$\frac{K_n^{1+2\beta/(d+1)}}{n} \asymp n^{-\frac{\nu_r}{1+\nu_r+2\beta/(d+1)}} \to 0.$$

Thus, the sieve dimension associated with the sharper effective-dimension $L^2(X)$ rate is admissible for directional inference as well.

In view of (4.33), the first condition in (4.35) requires that the directions $\mathfrak{a}_{K_n}$ are not asymptotically concentrated towards the eigenvectors of Γ_{K_n} associated with decreasing eigenvalues. To see this, assume, following Proposition 4.1 and Theorem 4.4 for the sake of example, that $\overline{s}^2_{K_n} \lesssim K_n^{-2/(d+1)}$ and $\lambda_{K_n} \asymp K_n^{-r}$ for $r < 2/(d+1)$, and that $\overline{\mathfrak{a}}_{j,K_n} = 0$ for all $j < K_n$ and $\overline{\mathfrak{a}}_{K_n,K_n} = 1$ for all K_n sufficiently large and P_θ almost surely. Then, (4.33) reduces to

$$\mathcal{R}_{K_n,\lambda}(\alpha_{K_n}) = \frac{\overline{s}^2_{K_n}}{\overline{s}^2_{K_n} + \lambda_{K_n,n}} \leq \frac{\overline{s}^2_{K_n}}{\lambda_{K_n,n}} \lesssim K_n^{r-2/(d+1)},$$

meaning P_θ almost surely $\mathcal{R}_{K_n,\lambda}(\alpha_{K_n})^{-1/2} \gtrsim K_n^{1/(d+1)-r/2} \to \infty$ as K_n grows, violating (4.35). Conversely, if $\mathfrak{a}_{K_n}$ is asymptotically concentrated on eigenspaces whose eigenvalues dominate $\lambda_{K_n,n}$ uniformly, then $\mathcal{R}_{K_n,\lambda}(\alpha_{K_n}) \to 1$. Thus, the asymptotic normality of the functional $\sqrt{n}\alpha_{K_n}^\top(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})/\sigma^{(\ell)}_{K_n,\lambda}(\alpha_{K_n})$ does not hinge only on the decay of the least eigenvalue, $\overline{s}^2_{K_n}$, but rather on the orientation of α_{K_n} with respect to the *subset of well-conditioned eigenvectors* of Γ_{K_n}.

The ridge-bias condition in (4.36) is sufficient but not necessary for a prescribed functional direction. Indeed, although Theorem 4.6 imposes the stronger norm-based condition (iv), its proof only requires the following condition on the directional ridge remainder:

$$\frac{\sqrt{n}\lambda_{K_n,n}\left|\widetilde{\alpha}^\top_{K_n,\lambda}\widehat{D}_{K_n,n,\lambda}\Gamma^{-1/2}_{K_n,\lambda}\beta_{K_n}\right|}{\sigma^{(0)}_{K_n,\lambda}(\alpha_{K_n})} = o_P(1).$$

Condition (4.36) provides a convenient direction-independent sufficient condition for this requirement, while sharper conditions may exploit the alignment between the functional direction and the ridge-bias direction.

4.4 Feasible inference

We now proceed by moving from asymptotic normality towards feasible inference. More specifically, we derive consistency of a residual-based ridge-adjusted estimator of the directional asymptotic variance. First, we analyze the effects of approximating the population quantities by their empirical counterparts. Next, we establish relative consistency under general conditions, and, finally, specialize these conditions to RWNNs, showing that the same estimator consistently estimates both directional variance specifications considered in Subsection 4.3. We conclude with an example demonstrating that feasible inference is possible even when the population feature covariance matrix is singular.

Lemma 4.7. *Let $\{Z_i\}_{i\in\mathbb{N}^+}$ be i.i.d. and satisfy Assumption 11, and let $\alpha = \alpha_{K_n} \in \mathbb{S}^{K_n-1}$. Set $\widehat{\epsilon}_i := Y_i - \widehat{h}_{n,\lambda}(X_i)$, $r_i := \alpha^\top \Gamma_{K_n,\lambda}^{-1} g_i$, and $\widehat{r}_i := \alpha^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} g_i$. Assume that there exist deterministic positive sequences $\overline{a}_{K_n,n,\lambda} = o(1)$ and $\overline{\mathfrak{M}}_{K_n,n,\lambda}$ such that*

$$\|A_{K_n,n,\lambda}\| = O_P\left(\overline{a}_{K_n,n,\lambda}\right) \quad \text{and} \quad \left\|\widehat{\Gamma}_{K_n,n}^{1/2}\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right)\right\| = O_P\left(\overline{\mathfrak{M}}_{K_n,n,\lambda}\right) \tag{4.39}$$

Then

(i) $\left|n^{-1}\sum_{i=1}^n \left[\widehat{\epsilon}_i^2 - (\epsilon_i+\eta_i)^2\right]\widehat{r}_i^2\right| = O_P\left(\|\widetilde{\alpha}\|^2 \widetilde{\xi}_{K_n,\lambda}^2 \overline{\mathfrak{M}}_{K_n,n,\lambda}\left(\overline{\mathfrak{M}}_{K_n,n,\lambda} + \sqrt{\overline{\sigma}^2 + \Delta_{K_n}^2}\right)\right)$.

(ii) $\left|n^{-1}\sum_{i=1}^n (\epsilon_i+\eta_i)^2\left[\widehat{r}_i^2 - r_i^2\right]\right| = O_P\left(\|\widetilde{\alpha}\|^2 \widetilde{\xi}_{K_n,\lambda}^2 \overline{a}_{K_n,n,\lambda}\left(\overline{\sigma}^2 + \Delta_{K_n}^2\right)\right)$.

Proposition 4.2 (Consistency of the ridge-adjusted variance estimator)**.** *Let the assumptions of Lemma 4.7 hold. Fix a possibly n-dependent direction $\alpha = \alpha_{K_n} \in \mathbb{S}^{K_n-1}$. Let $\sigma_{K_n,\lambda}^{2,(1)}(\alpha)$ be defined as in (4.31), and assume that $\sigma_{K_n,\lambda}^{2,(1)}(\alpha) > 0$ for all sufficiently large n. Define the directional Rayleigh quotient and the corresponding factor, respectively, as*

$$\mathcal{R}_{K_n,\lambda}^{(1)}(\alpha) := \frac{\sigma_{K_n,\lambda}^{2,(1)}(\alpha)}{\|\widetilde{\alpha}\|^2} \quad \text{and} \quad \mathcal{L}_{K_n,\lambda}^{(1)}(\alpha) := \frac{\widetilde{\xi}_{K_n,\lambda}^2}{\mathcal{R}_{K_n,\lambda}^{(1)}(\alpha)}. \tag{4.40}$$

Assume that:

(i) $\mathcal{L}_{K_n,\lambda}^{(1)}(\alpha)\overline{\mathfrak{M}}_{K_n,n,\lambda}\left(\overline{\mathfrak{M}}_{K_n,n,\lambda} + \sqrt{\overline{\sigma}^2 + \Delta_{K_n}^2}\right) = o(1)$.

(ii) $\mathcal{L}_{K_n,\lambda}^{(1)}(\alpha)\overline{a}_{K_n,n,\lambda}\left(\overline{\sigma}^2 + \Delta_{K_n}^2\right) = o(1)$.

(iii) $n^{-1/2}\mathcal{L}_{K_n,\lambda}^{(1)}(\alpha) = o(1)$

(iv) $\sup_{n\geq 1}\sup_{x\in\mathcal{X}} \mathbb{E}\left[(\epsilon_i + \eta_{K_n}(x))^4 \,|\, X_i = x, \mathcal{F}_{K_n}\right] \leq \overline{\sigma}_{\epsilon,\eta}^4 < \infty$.

Define

$$\widehat{\sigma}_{K_n,n,\lambda}^2(\alpha) := \alpha^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(\frac{1}{n}\sum_{i=1}^n \widehat{\epsilon}_i^2 g_i g_i^\top\right)\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\alpha$$

Then,

$$\frac{\widehat{\sigma}_{K_n,n,\lambda}^2(\alpha)}{\sigma_{K_n,\lambda}^{2,(1)}(\alpha)} \xrightarrow{p} 1.$$

Corollary 4.4 (RWNN consistency of the ridge-adjusted variance estimator)**.** *Let the assumptions of the β-Hölder RWNN approximation result in Corollary E.1 hold, and let $\{Z_i\}_{i\in\mathbb{N}^+}$ be i.i.d. observations satisfying Assumption 11. Suppose that Assumption 10 holds and that, for some $\beta \in (0,1)$,*

$$\lambda_{K_n,n} \asymp K_n^{-r}, \qquad \frac{2\beta}{d+1} < r < \frac{2}{d+1}. \tag{4.41}$$

Set

$$\nu_r := \frac{(d+1)(1+r)}{d+3}. \tag{4.42}$$

Define

$$\overline{\mathfrak{M}}_{K_n,n,\lambda} := \sqrt{\frac{K_n^{\nu_r}}{n}} + K_n^{-1/2-\beta/(d+1)} \quad \text{and} \quad \overline{a}_{K_n,n,\lambda} := \sqrt{\frac{K_n^{1+r}\log(K_n)}{n}}.$$

Fix a possibly n-dependent $\mathcal{F}_{K_n}$-measurable direction $\alpha = \alpha_{K_n} \in \mathbb{S}^{K_n-1}$. Let $\sigma^{2,(1)}_{K_n,\lambda}(\alpha)$ be defined as in (4.31), and assume that $\sigma^{2,(1)}_{K_n,\lambda}(\alpha) > 0$ for all sufficiently large n. Assume that, for some $0 \le \rho < \frac{1}{2} + \frac{\beta}{d+1}$,

$$\mathcal{L}^{(1)}_{K_n,\lambda}(\alpha) = O_{P_\theta}(K_n^{\rho}), \tag{4.43}$$

and that

$$\frac{K_n^{2\rho+1+r}\log(K_n)}{n} = o(1). \tag{4.44}$$

Finally, assume that

$$\sup_{n\ge 1}\sup_{x\in\mathcal{X}} \mathbb{E}\big[\big(\epsilon_i + \eta_{K_n}(x)\big)^4 \mid X_i = x, \mathcal{F}_{K_n}\big] \le \overline{\sigma}^4_{\epsilon,\eta} < \infty. \tag{4.45}$$

Then

$$\frac{\widehat{\sigma}^2_{K_n,n,\lambda}(\alpha)}{\sigma^{2,(1)}_{K_n,\lambda}(\alpha)} \xrightarrow{p} 1.$$

We next show that the same estimator is also consistent for $\sigma^{2,(0)}_{K_n,\lambda}(\alpha)$, since the contribution of the sieve approximation error to $\sigma^{2,(1)}_{K_n,\lambda}(\alpha)$ is asymptotically negligible.

Corollary 4.5 (RWNN consistency for the noise variance)**.** *Let the assumptions of Corollary 4.4 hold. Assume additionally that $\sigma^{2,(0)}_{K_n,\lambda}(\alpha) > 0$ for all sufficiently large n. Then*

$$\frac{\widehat{\sigma}^2_{K_n,n,\lambda}(\alpha)}{\sigma^{2,(0)}_{K_n,\lambda}(\alpha)} \xrightarrow{p} 1.$$

The following example demonstrates that singularity of the population covariance matrix is compatible with asymptotically valid inference and consistent variance estimation.

Example 1 (Feasible inference with singular population covariance)**.** Consider $d = 1$, $\mathcal{X} = [-1, 1]$, and the ReLU activation $\sigma(z) = z_+$. Let $X_i \sim \mathcal{U}([-1, 1])$ and $Y_i = X_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \tau^2)$, independently of X_i. Independently of the data, draw $A_k \sim \mathcal{U}([-1, 1])$ and $B_k \sim \mathcal{U}([2, 3])$, independently across k. Since $Ax + B \ge 1$ on the specified domains, the regression function admits the integral representation

$$h(x) = \mathbb{E}_{A,B}\big[3A(Ax + B)_+\big] = x, \qquad x \in [-1, 1].$$

It is easy to see that the product source condition holds with $f(A, B) = 3A$ and $\beta = 1/2$, and the regular-sampling assumptions are satisfied.

Let $u_K = (A_1, \dots, A_K)^\top$ and $v_K = (B_1, \dots, B_K)^\top$. The corresponding feature map and population covariance matrix satisfy $g_K(x) = xu_K + v_K$ and $\Gamma_K = \frac{1}{3}u_K u_K^\top + v_K v_K^\top$, respectively. Consequently, $\operatorname{rank}(\Gamma_K) = 2$ almost surely for every $K \ge 2$, and hence $\lambda_{\min}(\Gamma_K) = 0$ for every $K \ge 3$. Nevertheless, choosing $K_n = \lceil n^{2/5} \rceil$ and $\lambda_{K_n,n} = K_n^{-3/4}$, the conditions of Corollaries 4.3 and 4.4 are satisfied along the normalized point evaluation directions and the direction associated with $\Psi_2(h) = \int_{-1}^{1} h(x)^2\,\mathrm{d}x$. In particular, conditionally on $\mathcal{F}_{K_n}$, in P_θ-probability,

$$\sqrt{n}\big(\widehat{h}_{n,\lambda}(x_0) - x_0\big) \xrightarrow{d} \mathcal{N}\big(0, \tau^2(1 + 3x_0^2)\big),$$

for any fixed $x_0 \in [-1, 1]$ and

$$\sqrt{n}\left(\Psi_2(\widehat{h}_{n,\lambda}) - \frac{2}{3}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{16\tau^2}{3}\right).$$

Moreover, the corresponding ridge-adjusted directional variance estimators are consistent. We provide detailed verification of this example in Appendix C.4.

4.5 Inference for smooth functionals

In this subsection, we extend the preceding directional inference results to smooth nonlinear functionals of h. The argument uses differentiability and Riesz representation to identify a coefficient direction corresponding to the functional derivative, allowing us to apply the linear-functional theory of Subsection 4.3. The $L^2(X)$ estimation bounds control the nonlinear remainder, while a separate condition controls the first-order sieve approximation bias. Together, these ingredients yield asymptotic normality centered at the target functional $\Psi(h)$.

4.5.1 Smooth functionals and Riesz representation

Let $\Psi : \mathcal{H} \to \mathbb{R}$ be a known functional and assume that the object of interest for inference is $\Psi(h)$.

Assumption 12 (Smooth functional and Riesz representer)**.** Let $\mathcal{V}$ be a closed linear subspace of $L^2(X)$ containing $\mathcal{H} - h$ and the RWNN sieve spaces considered below. There exists a continuous linear functional $\Psi'_h : \mathcal{V} \to \mathbb{R}$ such that, as $\|\widetilde{h} - h\|_{L^2(X)} \to 0$,

$$\Psi(\widetilde{h}) - \Psi(h) = \Psi'_h(\widetilde{h} - h) + O(\|\widetilde{h} - h\|^{\omega}_{L^2(X)}) \tag{4.46}$$

for some $\omega > 1$. Let $\psi \in \mathcal{V}$ denote the Riesz representer of Ψ'_h, so that

$$\Psi'_h(v) = \langle v, \psi \rangle_{L^2(X)}, \qquad v \in \mathcal{V}. \tag{4.47}$$

For the RWNN sieve, we use the component-wise notation $\Psi'_h(g_{K_n}) := (\Psi'_h(g_{1,K_n}), \ldots, \Psi'_h(g_{K_n,K_n}))^\top \in \mathbb{R}^{K_n}$. For any $u \in \mathbb{R}^{K_n}$ define $u_g(x) := g_{K_n}(x)^\top u$, $x \in \mathcal{X}$. By the definition of Γ_{K_n} in (4.2), for any $u, v \in \mathbb{R}^{K_n}$, $\langle u_g, v_g \rangle_{L^2(X)} = u^\top \Gamma_{K_n} v$. By linearity of Ψ'_h,

$$\Psi'_h(u_g) = u^\top \Psi'_h(g_{K_n}). \tag{4.48}$$

Hence, on the RWNN sieve, the derivative of the smooth functional is a linear functional in the coefficient direction $\Psi'_h(g_{K_n})$. When this direction is nonzero, the directional results of Subsection 4.3.2 can be applied after normalization, provided their assumptions are satisfied.

We now notice that, by (4.47), $\Psi'_h(g_{K_n}) = \mathbb{E}_X[g_{K_n}(X)\psi(X)]$. For the ridge-regularized estimator, define the ridge-adjusted sieve Riesz representer associated with Ψ'_h by

$$\psi_{K_n,\lambda}(x) := g_{K_n}(x)^\top \Gamma^{-1}_{K_n,\lambda} \Psi'_h(g_{K_n}), \qquad x \in \mathcal{X}. \tag{4.49}$$

Equivalently, let $b_{\psi,K_n,\lambda} := \Gamma^{-1}_{K_n,\lambda} \Psi'_h(g_{K_n})$. Then, for every $u \in \mathbb{R}^{K_n}$, one has $u^\top \Gamma_{K_n,\lambda} b_{\psi,K_n,\lambda} = u^\top \Psi'_h(g_{K_n}) = \Psi'_h(u_g)$.

Hence $b_{\psi,K_n,\lambda}$ is the Riesz representer of the restriction of Ψ'_h to the RWNN coefficient space equipped with the $\Gamma_{K_n,\lambda}$ inner product, while $\psi_{K_n,\lambda}$ is the corresponding sieve function.

The regularity of Ψ'_h with respect to the $L^2(X)$ norm controls the associated ridge-normalized coefficient direction. Indeed, for every $u \in \mathbb{R}^{K_n}$,

$$\left| u^\top \Psi'_h(g_{K_n}) \right| = \left| \langle u_g, \psi \rangle_{L^2(X)} \right| \le (u^\top \Gamma_{K_n} u)^{1/2} \|\psi\|_{L^2(X)} \le (u^\top \Gamma_{K_n,\lambda} u)^{1/2} \|\psi\|_{L^2(X)},$$

and hence $\left\| \Gamma^{-1/2}_{K_n,\lambda} \Psi'_h(g_{K_n}) \right\| \le \|\psi\|_{L^2(X)}$. This construction is closely related to methods based on regularized sieves and semiparametric inference via Riesz representation (see, for example, Shen, 1997; Chen and Pouzo, 2009, 2012; Pouzo, 2016). We apply the empirical counterpart of $\psi_{K_n,\lambda}$ and the associated variance estimator in the simulations of Subsection 5.4.

4.5.2 Functional linearization and remainder control

We now combine the smooth-functional expansion in Assumption 12 with the $L^2(X)$ convergence results of Subsection 4.2. Applying (4.46) to $\widehat{h}_{n,\lambda}$ yields

$$\Psi(\widehat{h}_{n,\lambda}) - \Psi(h) = \Psi'_h(\widehat{h}_{n,\lambda} - h) + O_P\left(\|\widehat{h}_{n,\lambda} - h\|^{\omega}_{L^2(X)}\right). \tag{4.50}$$

Hence the functional remainder is negligible at $\sqrt{n}$ scale whenever $\sqrt{n}\,\|\widehat{h}_{n,\lambda} - h\|^{\omega}_{L^2(X)} = o_P(1)$.

Corollary 4.6 (Negligibility of the smooth-functional remainder for RWNNs)**.** *Let the conditions of Theorem 4.4 and Assumption 12 hold with $\omega \geq 2$. Suppose moreover that $\beta \in (0,1)$ and (4.21) holds, and choose*

$$K_n \asymp n^{\frac{d+1}{2(\beta+d+1)}}. \tag{4.51}$$

Then

$$\|\widehat{h}_{n,\lambda} - h\|_{L^2(X)} = o_P(n^{-1/4}), \tag{4.52}$$

and consequently

$$\sqrt{n}\big(\Psi(\widehat{h}_{n,\lambda}) - \Psi(h) - \Psi'_h(\widehat{h}_{n,\lambda} - h)\big) = o_P(1).$$

We now consider the first-order linear term in (4.50). Define $h_{K_n}(x) := g_{K_n}(x)^\top \beta_{K_n}$ and $a_{K_n} := \Psi'_h(g_{K_n})$. By linearity of Ψ'_h and (4.48),

$$\Psi'_h(\widehat{h}_{n,\lambda} - h) = \Psi'_h(\widehat{h}_{n,\lambda} - h_{K_n}) + \Psi'_h(h_{K_n} - h) = a_{K_n}^\top\big(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\big) + \Psi'_h(h_{K_n} - h). \tag{4.53}$$

Define

$$B_{\Psi,K_n} := \Psi'_h(h_{K_n} - h) = \langle h_{K_n} - h, \psi\rangle_{L^2(X)}. \tag{4.54}$$

Hence, the first-order linear term in (4.50) is a linear sieve functional of a type studied in Subsection 4.3.2, with coefficient direction a_{K_n}, while B_{Ψ,K_n} is the sieve approximation bias in the functional direction.

Combining (4.50), (4.53), and Corollary 4.6, we obtain

$$\sqrt{n}\big(\Psi(\widehat{h}_{n,\lambda}) - \Psi(h)\big) = \sqrt{n}\, a_{K_n}^\top\big(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\big) + \sqrt{n}\, B_{\Psi,K_n} + o_P(1). \tag{4.55}$$

4.5.3 Asymptotic normality for RWNN smooth functionals

Suppose that $a_{K_n} \neq 0$ and define the normalized functional direction

$$\mathfrak{a}^{\Psi}_{K_n} := \frac{a_{K_n}}{\|a_{K_n}\|} = \frac{\Psi'_h(g_{K_n})}{\|\Psi'_h(g_{K_n})\|} \in \mathbb{S}^{K_n-1}. \tag{4.56}$$

Since g_{K_n} is determined by the sampled random parameters, $\mathfrak{a}^{\Psi}_{K_n}$ is $\mathcal{F}_{K_n}$-measurable and hence is an admissible prescribed direction in Corollary 4.3. For $\ell \in \{0,1\}$, define the corresponding functional standard deviation by

$$\sigma^{(\ell)}_{\Psi,K_n,\lambda} := \|a_{K_n}\|\sigma^{(\ell)}_{K_n,\lambda}(\mathfrak{a}^{\Psi}_{K_n}). \tag{4.57}$$

Corollary 4.7 (Asymptotic normality for RWNN smooth functionals)**.** *Let the conditions of Corollary 4.6 hold and suppose that $a_{K_n} \neq 0$ with P_θ-probability tending to one. Assume that the conditions of Corollary 4.3 hold along the direction $\mathfrak{a}^{\Psi}_{K_n}$ defined in (4.56). Suppose moreover that*

$$\big(\sigma^{(0)}_{\Psi,K_n,\lambda}\big)^{-1} = O_{P_\theta}(1), \tag{4.58}$$

and that the sieve approximation bias in the functional direction satisfies

$$\sqrt{n}\,\big|\Psi'_h(h_{K_n} - h)\big| = o_{P_\theta}(1). \tag{4.59}$$

Then, conditionally on $\mathcal{F}_{K_n}$, in P_θ-probability, for both $\ell \in \{0,1\}$,

$$\sqrt{n}\frac{\Psi(\widehat{h}_{n,\lambda}) - \Psi(h)}{\sigma^{(\ell)}_{\Psi,K_n,\lambda}} \xrightarrow{d} \mathcal{N}(0,1). \tag{4.60}$$

Remark 13 (Sieve bias for the integrated-square functional)**.** Consider $\Psi_2(h) = \int_{-1}^{1} h(x)^2\,\mathrm{d}x$, with $X \sim \mathcal{U}([-1,1])$. Since h_{K_n} is the $L^2(X)$-orthogonal projection of h onto the RWNN sieve, the first-order sieve bias satisfies

$$B_{\Psi_2,K_n} = \Psi'_{2,h}(h_{K_n} - h) = 4\langle h, h_{K_n} - h\rangle_{L^2(X)} = -4\|h - h_{K_n}\|^2_{L^2(X)} = -4\Delta^2_{K_n}.$$

Under the assumptions of Corollary E.1, it follows that

$$\sqrt{n}\,|B_{\Psi_2,K_n}| = O_{P_\theta}\left(\sqrt{n}\,K_n^{-1-2\beta/(d+1)}\right).$$

Consequently, for K_n chosen as in (4.51), the right-hand side is $O_{P_\theta}\left(n^{-\beta/[2(\beta+d+1)]}\right) = o_{P_\theta}(1)$. Hence the sieve-bias condition (4.59) is automatically satisfied for this functional.

4.6 Sup-norm analysis

The sup-norm analysis we develop in this subsection follows two approaches to controlling the stochastic error and ridge-induced bias. First, we derive the bounds using the singular values of the empirical feature matrix. Importantly, these bounds can be conservative when the feature representation is poorly conditioned, which immediately motivates our second approach based on the empirical ridge pseudo-projection. This approach is inspired by the sup-norm stability concepts from sieve estimation (Chen and Christensen, 2015) and ridge leverage quantities studied in the kernel regression literature (Alaoui and Mahoney, 2015; Avron et al., 2017; Pauwels et al., 2018). By controlling the ridge smoothing weights, the empirical ridge Lebesgue constant, and leverage quantities, we obtain uniform estimation bounds under prediction-space stability conditions that do not require a positive lower bound on the smallest empirical covariance eigenvalue.

We start by recalling that $g_{K_n}(x) \in \mathbb{R}^{K_n}$ denotes the vector of random features evaluated at $x \in \mathcal{X}$, and the associated design matrix is

$$G_{K_n,n} = \big(g_{K_n}(X_1), \ldots, g_{K_n}(X_n)\big)^\top \in \mathbb{R}^{n\times K_n}.$$

For a deterministic ridge regularization parameter $\lambda_{K_n,n} > 0$, the ridge-regularized RWNN least-squares estimator is

$$\widehat{h}_{n,\lambda}(x) = g_{K_n}(x)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda}\big(n^{-1}G^\top_{K_n,n}Y\big), \tag{4.61}$$

where

$$\widehat{\Gamma}_{K_n,n,\lambda} := n^{-1}G^\top_{K_n,n}G_{K_n,n} + \lambda_{K_n,n}I_{K_n}, \quad \text{and} \quad Y := (Y_1,\ldots,Y_n)^\top \in \mathbb{R}^n.$$

To separate the stochastic and conditional bias components of the uniform error, define the ridge pseudo-projection

$$\widetilde{h}_{n,\lambda}(x) := g_{K_n}(x)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda}\big(n^{-1}G^\top_{K_n,n}H\big), \tag{4.62}$$

where

$$H := (h(X_1),\ldots,h(X_n))^\top \in \mathbb{R}^n.$$

The uniform estimation error satisfies the following bias-variance decomposition bound

$$\|\widehat{h}_{n,\lambda} - h\|_\infty \le \|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty + \|\widetilde{h}_{n,\lambda} - h\|_\infty. \tag{4.63}$$

The first term is the stochastic component of the estimation error (variance), whereas the second term (bias) contains the sieve approximation error and the additional shrinkage bias induced by ridge regularization.

Throughout this subsection, we work under Assumption 20 on the data, the squared-loss regression model, and Assumption 11 on its error term. Since the uniform stochastic bound proved below is established only for independent observations, we restrict attention to case (i) of Assumption 20 with $\{Z_i\}_{i\in\mathbb{N}^+}$ i.i.d. For the uniform stochastic analysis, we impose the following additional moment and sieve-regularity conditions.

Assumption 13. There exists $\delta > 0$ such that $\mathbb{E}[|\epsilon_i|^{2+\delta}] < \infty$.

Assumption 14. There exist deterministic positive sequences $\{\zeta_K\}_{K\geq 1}$ and $\{L_K\}_{K\geq 1}$ such that, almost surely,

(i) $\sup_{x\in\mathcal{X}} \|g_K(x)\| \leq \zeta_K$.

(ii) $\|g_K(x_1) - g_K(x_2)\| \leq L_K\|x_1 - x_2\|$ for all $x_1, x_2 \in \mathcal{X}$.

(iii) Along the sequence $K = K_n$, there exist constants $C > 0$ and $q > 0$ such that

$$\frac{L_{K_n}\sqrt{m_n}}{\zeta_{K_n}\sqrt{\log(n)}} \leq Cn^q \tag{4.64}$$

with $m_n := \min\{K_n, n\}$, for all sufficiently large n.

Remark 14. The rate in Assumption 14(iii) may alternatively be simplified to

$$\frac{L_{K_n}}{\zeta_{K_n}}\sqrt{m_n} \lesssim n^{q'},$$

for some $q' > 0$, since $\sqrt{\log(n)}\, n^q = o(n^{q'})$ for any $q' > q$.

4.6.1 Uniform stochastic error

We start by bounding the uniform variance term $\|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty$ in (4.63). Following the strategy of Lemma 2.3 in Chen and Christensen (2015), we combine a covering argument on $\mathcal{X}$ with concentration conditional on the regressors. Since ridge regularization does not generate an empirical projection, we use the singular value decomposition of the design matrix. The algebraic arguments below do not require $K_n \leq n$. We therefore set $m_n := \min\{K_n, n\}$ and consider the thin singular value decomposition

$$n^{-1/2}G_{K_n,n} = U_{m_n,n}S_{m_n,n}V_{m_n,n}^\top, \tag{4.65}$$

where $U_{m_n,n} \in \mathbb{R}^{n\times m_n}$, $V_{m_n,n} \in \mathbb{R}^{K_n\times m_n}$ and

$$S_{m_n,n} := \operatorname{diag}(s_{1,n}, \ldots, s_{m_n,n}) \in \mathbb{R}^{m_n\times m_n}, \qquad s_{1,n} \geq \cdots \geq s_{m_n,n} \geq 0.$$

It holds $V_{m_n,n}^\top V_{m_n,n} = I_{m_n}$ and $U_{m_n,n}^\top U_{m_n,n} = I_{m_n}$. The nonzero eigenvalues of matrix $\widehat{\Gamma}_{K_n,n} = n^{-1}G_{K_n,n}^\top G_{K_n,n}$ are then given by the nonzero elements of $\{s_{k,n}^2\}_{k=1}^{m_n}$. When $K_n > n$, this matrix has at least $K_n - n$ additional zero eigenvalues. Accordingly, we extend the sequence to K_n indices by setting $s_{k,n} := 0$, $k = m_n + 1, \ldots, K_n$ whenever $K_n > n$; zero singular values arising from rank deficiency are not excluded from $S_{m_n,n}$. Hence, throughout this section, $s_{1,n}^2 \geq \cdots \geq s_{K_n,n}^2 \geq 0$ are the eigenvalues of $\widehat{\Gamma}_{K_n,n}$ counted with multiplicity. In particular, when $K_n \leq n$, $s_{m_n,n}^2 = \lambda_{\min}(\widehat{\Gamma}_{K_n,n})$.

Define the empirical ridge stability factor

$$\psi_{K_n,n,\lambda} := \max_{1\leq k\leq m_n} \frac{s_{k,n}}{s_{k,n}^2 + \lambda_{K_n,n}}. \tag{4.66}$$

Assumption 15. There exists a deterministic positive sequence $\overline{\psi}_{K_n,n}$ such that $\psi_{K_n,n,\lambda} = O_P\big(\overline{\psi}_{K_n,n}\big)$ and

(i) $\zeta_{K_n}\overline{\psi}_{K_n,n}\sqrt{\log(n)/n} = o(1)$,

(ii) $\big(\zeta_{K_n}/(\overline{\psi}_{K_n,n}\lambda_{K_n,n})\big)^{(2+\delta)/\delta} \lesssim \sqrt{n/\log(n)}$,

(iii) $\big(\zeta_{K_n}/(\overline{\psi}_{K_n,n}\lambda_{K_n,n})\big)\sqrt{n/\log(n)} \to \infty$,

where $\delta > 0$ is the exponent in Assumption 13.

Remark 15. If the sieve is such that $\zeta_{K_n} \asymp \sqrt{K_n}$, Assumption 15(i) then requires $\overline{\psi}_{K_n,n}\sqrt{K_n\log(n)/n} = o(1)$. If $\overline{\psi}_{K_n,n} \asymp 1$, this implies $K_n = o(n/\log(n))$. If $\overline{\psi}_{K_n,n} \to \infty$, the restriction becomes more stringent, requiring

$$K_n = o\left(\frac{n}{\overline{\psi}^2_{K_n,n}\log(n)}\right).$$

Without a lower bound on $\overline{\psi}_{K_n,n}$, condition (i) alone does not imply $K_n/n \to 0$. Nevertheless, combining conditions (i) and (ii) yields

$$K_n = o\left(\lambda_{K_n,n}\left(\frac{n}{\log(n)}\right)^{\frac{1+\delta}{2+\delta}}\right).$$

Consequently, if $\lambda_{K_n,n} = O(1)$, then

$$K_n = o\left(\left(\frac{n}{\log(n)}\right)^{\frac{1+\delta}{2+\delta}}\right).$$

In particular, under bounded ridge regularization, Assumption 15 excludes high-dimensional regimes with $K_n/n \to \kappa \in (0,\infty)$ and ultra-high-dimensional regimes with $n = o(K_n)$.

Notice that since $\sup_{s\geq 0} s/(s^2+\lambda_{K_n,n}) = (2\sqrt{\lambda_{K_n,n}})^{-1}$, the deterministic choice $\overline{\psi}_{K_n,n} = (2\sqrt{\lambda_{K_n,n}})^{-1}$ always satisfies $\psi_{K_n,n,\lambda} \leq \overline{\psi}_{K_n,n}$. However, conditions (i)–(iii) of Assumption 15 must still be verified separately for this choice.

Since the domain $\mathcal{X}$ is fixed and compact under Assumption 1, we do not introduce a weighted sup-norm and work instead throughout with the ordinary sup-norm.

Lemma 4.8 (Uniform stochastic bound)**.** *Let Assumptions 1, 20, 11, 13, 14, and 15 hold. Suppose, in addition, that $\{Z_i\}_{i\in\mathbb{N}^+}$ is i.i.d. Then*

$$\|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty = O_P\left(\zeta_{K_n}\overline{\psi}_{K_n,n}\sqrt{\frac{\log(n)}{n}}\right).$$

Consequently,

$$\|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty = o_P(1).$$

We now specialize the generic stochastic analysis to the RWNN sieves considered in Section 3. More specifically, Lemma 4.9 shows that, under either of the two activation and parameter space conditions used in the approximation results, the RWNN feature map satisfies Assumption 14. Consequently, Corollary 4.8 applies Lemma 4.9 to obtain the corresponding variance bound for the ridge-regularized RWNN least squares estimator.

Lemma 4.9. *Let Assumption 1 hold and define*

$$g_K(x) := \Big(\widetilde{\sigma}(A_1^\top x + B_1), \ldots, \widetilde{\sigma}(A_K^\top x + B_K)\Big)^\top, \quad x \in \mathcal{X}.$$

Let one of the following conditions be satisfied:

(i) $\widetilde{\sigma} = \sigma$, *Assumption 2 holds and the parameter space* $\Pi \subset \mathbb{R}^{d+1}$ *is compact;*

(ii) $\widetilde{\sigma} = \sigma_m$, $m = 1, 2$, *and Assumption 8 holds.*

Then there exist constants $M_g, L_g > 0$, *independent of* K *and* n, *such that, almost surely,*

$$\sup_{x \in \mathcal{X}} \|g_K(x)\| \leq M_g \sqrt{K}$$

and

$$\|g_K(x_1) - g_K(x_2)\| \leq L_g \sqrt{K} \|x_1 - x_2\|, \text{ for all } x_1, x_2 \in \mathcal{X}.$$

Consequently, Assumption 14 holds with $\zeta_K = M_g\sqrt{K}$ *and* $L_K = L_g\sqrt{K}$ *and, in particular, Assumption 14 (iii) holds, for example, with* $q = 1/2$.

Corollary 4.8. *Let the assumptions of Lemma 4.8 hold, except that Assumption 14 is replaced by either condition (i) or condition (ii) of Lemma 4.9. Then*

$$\|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty = O_P\left(\overline{\psi}_{K_n,n}\sqrt{\frac{K_n \log(n)}{n}}\right).$$

If, in addition, $\overline{\psi}_{K_n,n}\sqrt{K_n \log(n)/n} = o(1)$, *then*

$$\|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty = o_P(1).$$

4.6.2 Sup-norm bias and the ridge pseudo-projection

We now derive a bound on the sup-norm bias term, $\|\widetilde{h}_{n,\lambda} - h\|_\infty$, for the ridge-regularized RWNN estimator. For an unregularized series least-squares estimator, the standard approach is to exploit the empirical projection structure of the estimator (see, for example, Chen and Christensen, 2015). In the present setting, this approach cannot be applied directly for two reasons:

(a) We do not impose a uniform lower bound on the smallest nonzero singular value of the empirical feature matrix. Consequently, the coefficient bounds involving the inverse empirical covariance may be arbitrarily conservative when the random features are nearly collinear.

(b) Ridge regularization does not preserve the projection property. More precisely, the corresponding projection operator is not idempotent and does not reproduce every element of the RWNN sieve. An additional ridge-shrinkage term therefore appears in the bias decomposition.

For any bounded function $f : \mathcal{X} \to \mathbb{R}$, define the empirical ridge pseudo-projection $\overline{P}_{K_n,n,\lambda}$ by

$$\left(\overline{P}_{K_n,n,\lambda} f\right)(x) := g_{K_n}(x)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} \left(n^{-1} G_{K_n,n}^\top F\right), \tag{4.67}$$

where $F := \big(f(X_1), \ldots, f(X_n)\big)^\top$ and $\widehat{\Gamma}_{K_n,n,\lambda} := n^{-1} G_{K_n,n}^\top G_{K_n,n} + \lambda_{K_n,n} I_{K_n}$. In particular, taking $H := \big(h(X_1), \ldots, h(X_n)\big)^\top$ gives $\overline{P}_{K_n,n,\lambda} h = \widetilde{h}_{n,\lambda}$. We view $\overline{P}_{K_n,n,\lambda}$ as an operator on the space of bounded measurable functions on $\mathcal{X}$, equipped with the supremum norm, and denote its induced operator norm by $\|\overline{P}_{K_n,n,\lambda}\|_\infty$.

We first provide a generic coefficient-space bound on this operator norm, which does not exploit any additional structure of the random-feature sieve.

Lemma 4.10. *Suppose Assumption 14(i) holds. Then, almost surely,* $\|\overline{P}_{K_n,n,\lambda}\|_\infty \leq \zeta_{K_n} \psi_{K_n,n,\lambda}$.

Remark 16. The bound in Lemma 4.10 is generic and may be conservative. It controls the pseudo-projection through the worst amplified direction of the empirical feature covariance and therefore ignores possibly redundant random features. For localized series bases, such as (tensor products of) splines and wavelets, sharper bounds on the sup-norm of the empirical projection can be obtained by exploiting the intrinsic structure of the basis (see, for example, Huang (2003)). For the RWNN sieves considered here, we instead investigate stability directly through the empirical ridge pseudo-projection.

A sharper analysis may instead study directly the condition

$$\|\overline{P}_{K_n,n,\lambda}\|_\infty = O_P(1),$$

and we pursue this direction in more detail below. In particular, this condition appears to be reasonable for some of the settings we explore in the simulations of Section 5.

We now combine Lemma 4.10 with the sup-norm approximation result of Theorem 3.5. We consider the estimator in (4.61) with the stratified random features constructed in Theorem 3.5, independently of the regression observations. In case (ii) of Theorem 3.5, the stratification depends on the sign partition of the source function f. Consequently, the corresponding estimation results provide oracle guarantees unless this partition is known. Let

$$h^{\mathrm{IS}}_{K_n}(x) = g_{K_n}(x)^\top W_{K_n},$$

denote the corresponding stratified RWNN approximant, where $W_{K_n} \in \mathbb{R}^{K_n}$ and g_{K_n} is the same feature map used in (4.61).

Lemma 4.11. *Let the assumptions of Theorem 3.5 and Assumption 14(i) hold. Let $h^{\mathrm{IS}}_{K_n}(x) = g_{K_n}(x)^\top W_{K_n}$ be the importance sampling RWNN approximant constructed in Theorem 3.5. Then*

$$\|\widetilde{h}_{n,\lambda} - h\|_\infty \le \left(1 + \left\|\overline{P}_{K_n,n,\lambda}\right\|_\infty\right) \left\|h - h^{\mathrm{IS}}_{K_n}\right\|_\infty + R_{K_n,n,\lambda}, \tag{4.68}$$

where

$$R_{K_n,n,\lambda} := \left\|\overline{P}_{K_n,n,\lambda} h^{\mathrm{IS}}_{K_n} - h^{\mathrm{IS}}_{K_n}\right\|_\infty. \tag{4.69}$$

Moreover,

$$R_{K_n,n,\lambda} \le \zeta_{K_n} \frac{\lambda_{K_n,n}}{s^2_{K_n,n} + \lambda_{K_n,n}} \|W_{K_n}\| \le \sqrt{6} M_f \zeta_{K_n} K_n^{-1/2} \frac{\lambda_{K_n,n}}{s^2_{K_n,n} + \lambda_{K_n,n}}, \tag{4.70}$$

with M_f as in Theorem 3.5.

Thus, in particular,

$$\|\widetilde{h}_{n,\lambda} - h\|_\infty \le (1 + \zeta_{K_n} \psi_{K_n,n,\lambda}) \left\|h - h^{\mathrm{IS}}_{K_n}\right\|_\infty + \sqrt{6}\, M_f\, \zeta_{K_n} K_n^{-1/2} \frac{\lambda_{K_n,n}}{s^2_{K_n,n} + \lambda_{K_n,n}}. \tag{4.71}$$

4.6.3 Sup-norm estimation bounds

We now combine the stochastic and bias bounds derived in the preceding subsections.

Theorem 4.12 (Sup-norm estimation bound)**.** *Suppose that the assumptions of Lemmas 4.8 and 4.11 hold. Then*

$$\begin{aligned}\|\widehat{h}_{n,\lambda} - h\|_\infty \le \|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty + (1 + \zeta_{K_n} \psi_{K_n,n,\lambda}) \left\|h - h^{\mathrm{IS}}_{K_n}\right\|_\infty \\ + \sqrt{6} M_f \zeta_{K_n} K_n^{-1/2} \frac{\lambda_{K_n,n}}{s^2_{K_n,n} + \lambda_{K_n,n}}.\end{aligned} \tag{4.72}$$

Moreover, let

$$\alpha := \begin{cases} \beta/(d+1), & \text{in case (i) of Theorem 3.5,} \\ 1/(d+1), & \text{in case (ii) of Theorem 3.5.} \end{cases}$$

Then

$$\begin{aligned} \|\widehat{h}_{n,\lambda} - h\|_\infty &= O_P\left(\zeta_{K_n}\overline{\psi}_{K_n,n}\sqrt{\frac{\log(n)}{n}}\right) \\ &\quad + O_P\left(\left(1+\zeta_{K_n}\overline{\psi}_{K_n,n}\right)\left(\sqrt{d}+\sqrt{\log(K_n)}\right)K_n^{-1/2-\alpha}\right) \\ &\quad + \sqrt{6}M_f\zeta_{K_n}K_n^{-1/2}\frac{\lambda_{K_n,n}}{s_{K_n,n}^2+\lambda_{K_n,n}}. \end{aligned} \tag{4.73}$$

Corollary 4.9 (Sup-norm rate for RWNN features)**.** *Suppose that the assumptions of Theorem 4.12 and Lemma 4.9 hold. Then*

$$\begin{aligned} \left\|\widehat{h}_{n,\lambda} - h\right\|_\infty &= O_P\left(\overline{\psi}_{K_n,n}\sqrt{\frac{K_n\log(n)}{n}}\right) \\ &\quad + O_P\left(\left(1+\sqrt{K_n}\overline{\psi}_{K_n,n}\right)\left(\sqrt{d}+\sqrt{\log(K_n)}\right)K_n^{-1/2-\alpha}\right) \\ &\quad + \sqrt{6}M_fM_g\frac{\lambda_{K_n,n}}{s_{K_n,n}^2+\lambda_{K_n,n}}, \end{aligned} \tag{4.74}$$

where α is as in Theorem 4.12.

In particular, if $\overline{\psi}_{K_n,n} = O(1)$, then

$$\begin{aligned} \left\|\widehat{h}_{n,\lambda} - h\right\|_\infty &= O_P\left(\sqrt{\frac{K_n\log(n)}{n}}\right) + O_P\left(\left(\sqrt{d}+\sqrt{\log(K_n)}\right)K_n^{-\alpha}\right) \\ &\quad + \sqrt{6}M_fM_g\frac{\lambda_{K_n,n}}{s_{K_n,n}^2+\lambda_{K_n,n}}. \end{aligned} \tag{4.75}$$

Remark 17. The special case in (4.75) shows that, when $\overline{\psi}_{K_n,n} = O(1)$, the variance term has the standard sup-norm order. At the same time, notice that the first term of the bias contribution loses a factor $K_n^{-1/2}$ relative to the original RWNN approximation rate. This originates from the spectral estimate $\left\|\overline{P}_{K_n,n,\lambda}\right\|_\infty \le \zeta_{K_n}\psi_{K_n,n,\lambda} = O_P\left(\sqrt{K_n}\right)$, and a sharper bound $\left\|\overline{P}_{K_n,n,\lambda}\right\|_\infty = O_P(1)$ would retain the original approximation rate.

A separate limitation concerns the ridge-shrinkage term. If $\widehat{\Gamma}_{K_n,n}$ is singular, then $s_{K_n,n} = 0$, and the upper bound in (4.74) does not vanish. Therefore, the smallest singular value argument alone cannot establish that the ridge bias vanishes in the overparametrized regime. A sharper analysis is needed to address this issue.

These limitations motivate the leverage-based analysis developed next, which controls prediction-space stability and ridge shrinkage without relying exclusively on the smallest empirical singular value.

4.6.4 Leverage-based sup-norm stability

The preceding spectral analysis controls the estimator through the empirical ridge stability factor $\psi_{K_n,n,\lambda}$ and the smallest empirical singular value $s_{K_n,n}$. Such coefficient-space quantities may be conservative for prediction, particularly when the random features are highly redundant. We therefore formulate alternative conditions directly in terms of the empirical ridge pseudo-projection $\overline{P}_{K_n,n,\lambda}$ defined in (4.67). This follows the conventional series-estimation perspective

in which boundedness of the sup-norm operator norm is referred to as sup-norm stability (see, for example, Chen and Christensen, 2015). For the ridge pseudo-projection, we use the empirical ridge Lebesgue constant as a sufficient measure of this stability.

For $x \in \mathcal{X}$ and $i = 1, \dots, n$, define the ridge sieve smoothing weights

$$p_{i,n,\lambda}(x) := n^{-1} g_{K_n}(x)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} g_{K_n}(X_i), \tag{4.76}$$

and let $p_{n,\lambda}(x) := (p_{1,n,\lambda}(x), \dots, p_{n,n,\lambda}(x))^\top$.

Lemma 4.13 (Properties of the ridge smoothing weights)**.** *For every bounded measurable function $f : \mathcal{X} \to \mathbb{R}$,*

$$\big(\overline{P}_{K_n,n,\lambda} f\big)(x) = \sum_{i=1}^{n} p_{i,n,\lambda}(x) f(X_i). \tag{4.77}$$

Define

$$\Lambda_{n,\lambda} := \sup_{x \in \mathcal{X}} \sum_{i=1}^{n} |p_{i,n,\lambda}(x)|, \quad \nu_{n,\lambda} := \sup_{x \in \mathcal{X}} \sum_{i=1}^{n} p_{i,n,\lambda}(x)^2, \tag{4.78}$$

as well as

$$\ell_{n,\lambda} := \sup_{x \in \mathcal{X}} g_{K_n}(x)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} g_{K_n}(x) \quad \textit{and} \quad \mathfrak{m}_{n,\lambda} := \sup_{x \in \mathcal{X}} \max_{1 \le i \le n} |p_{i,n,\lambda}(x)|. \tag{4.79}$$

Then

$$\|\overline{P}_{K_n,n,\lambda}\|_\infty \le \Lambda_{n,\lambda}, \tag{4.80}$$

and it holds that

$$\nu_{n,\lambda} \le \frac{\ell_{n,\lambda}}{n} \quad \textit{and} \quad \mathfrak{m}_{n,\lambda} \le \frac{\ell_{n,\lambda}}{n}. \tag{4.81}$$

The quantity $\ell_{n,\lambda}$ introduced in Lemma 4.13 is studied in random kernel regression literature (see, for example, Alaoui and Mahoney, 2015; Avron et al., 2017; Pauwels et al., 2018). In the following, we call $\ell_{n,\lambda}$ the uniform empirical ridge leverage score, and refer to $\Lambda_{n,\lambda}$ as the empirical ridge Lebesgue constant. The ridge leverage quantities are closely related to the effective dimension introduced in (4.12). Indeed, the empirical analogue of $\mathcal{N}_{K_n,\lambda}$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} g_{K_n}(X_i)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} g_{K_n}(X_i) = \operatorname{tr}\left(\widehat{\Gamma}_{K_n,n} \widehat{\Gamma}_{K_n,n,\lambda}^{-1} \right).$$

Thus, effective dimension measures average ridge leverage, whereas $\ell_{n,\lambda}$ controls its supremum over $\mathcal{X}$. Moreover, under Assumption 10, the ridge-normalized covariance stability in (4.16) implies

$$\ell_{n,\lambda} \le (1 + o_P(1)) \widetilde{\xi}_{K_n,\lambda}^2. \tag{4.82}$$

Indeed, by (4.17), on the event $\|A_{K_n,n,\lambda}\| < 1$, we have $\widehat{\Gamma}_{K_n,n,\lambda}^{-1} \preceq (1 - \|A_{K_n,n,\lambda}\|)^{-1} \Gamma_{K_n,\lambda}^{-1}$. Taking the supremum over $x \in \mathcal{X}$ and using $\|A_{K_n,n,\lambda}\| = o_P(1)$ from (4.16) yields (4.82). Consequently, the additional condition $\widetilde{\xi}_{K_n,\lambda}^2 = O_P(K_n)$ is sufficient for $\ell_{n,\lambda} = O_P(K_n)$. This does not, however, imply $\Lambda_{n,\lambda} = O_P(1)$, which requires separate control of the absolute smoothing weights.

We now note that using the singular value decomposition in (4.65), the vector of ridge sieve smoothing weights satisfies

$$p_{n,\lambda}(x) = \frac{1}{\sqrt{n}} U_{m_n,n} \operatorname{diag}\left(\left(\frac{s_{k,n}}{s_{k,n}^2 + \lambda_{K_n,n}} \right)_{k=1}^{m_n} \right) V_{m_n,n}^\top g_{K_n}(x),$$

which, together with Assumption 14(i), implies $\nu_{n,\lambda} \le \zeta_{K_n}^2 \psi_{K_n,n,\lambda}^2 / n$, while the Cauchy-Schwarz inequality gives $\Lambda_{n,\lambda} \le \sqrt{n \nu_{n,\lambda}} \le \zeta_{K_n} \psi_{K_n,n,\lambda}$. Hence, the SVD reproduces the feature-space conditioning bound in Lemma 4.10, though it does not exploit possible cancellations among the ridge sieve smoothing weights.

Assumption 16. The empirical ridge Lebesgue constant and the uniform empirical ridge leverage score satisfy $\Lambda_{n,\lambda} = O_P(1)$ and $\ell_{n,\lambda} = O_P(K_n)$, respectively.

Notice that, by Lemma 4.13, Assumption 16 implies $\|\overline{P}_{K_n,n,\lambda}\|_\infty = O_P(1)$, $\nu_{n,\lambda} = O_P(K_n/n)$, and $\mathfrak{m}_{n,\lambda} = O_P(K_n/n)$. Unlike a condition of the form $\psi_{K_n,n,\lambda} = O_P(1)$, Assumption 16 controls the fitted values directly and may therefore remain valid when the feature-space problem is poorly conditioned.

Lemma 4.14 (Uniform variance bound under sup-norm and leverage stability)**.** *Let Assumptions 1, 20, 11, 13, 14 (i)-(ii), and the condition $\ell_{n,\lambda} = O_P(K_n)$ in Assumption 16 hold. Suppose, in addition, that $\{Z_i\}_{i\in\mathbb{N}^+}$ is i.i.d. and that there exists a deterministic positive sequence $\overline{\psi}_{K_n,n}$ such that $\psi_{K_n,n,\lambda} = O_P(\overline{\psi}_{K_n,n})$. Assume that*

$$\frac{K_n^{1+2/\delta}\log(n)}{n} = O(1), \tag{4.83}$$

and that, for some constants $C, q > 0$,

$$\frac{L_{K_n}\overline{\psi}_{K_n,n}\sqrt{m_n}}{\sqrt{K_n\log(n)}} \le Cn^q \tag{4.84}$$

for all sufficiently large n. Then

$$\|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty = O_P\left(\sqrt{\frac{K_n\log(n)}{n}}\right). \tag{4.85}$$

Recall that the ridge remainder in (4.69) is defined by

$$R_{K_n,n,\lambda} := \left\|\overline{P}_{K_n,n,\lambda}h^{\mathrm{IS}}_{K_n} - h^{\mathrm{IS}}_{K_n}\right\|_\infty.$$

As shown in (4.70), this remainder admits a bound in terms of the smallest empirical singular value. However, its direct formulation allows us to derive an alternative bound that does not involve this quantity.

Lemma 4.15 (Ridge-remainder bound)**.** *It holds that*

$$R_{K_n,n,\lambda} \le \sqrt{\lambda_{K_n,n}\ell_{n,\lambda}}\|W_{K_n}\|. \tag{4.86}$$

Consequently, for the importance-sampling weights constructed in Theorem 3.5,

$$R_{K_n,n,\lambda} \le \sqrt{6}M_f\sqrt{\frac{\lambda_{K_n,n}\ell_{n,\lambda}}{K_n}}. \tag{4.87}$$

Moreover, under Assumption 16,

$$R_{K_n,n,\lambda} = O_P\left(\sqrt{\lambda_{K_n,n}}\right). \tag{4.88}$$

The bound in Lemma 4.15 does not involve the smallest empirical singular value and remains valid when $\widehat{\Gamma}_{K_n,n}$ is singular. Under Assumption 16, it yields $R_{K_n,n,\lambda} = O_P(\sqrt{\lambda_{K_n,n}})$, independently of the smallest empirical singular value.

Theorem 4.16 (Sup-norm estimation bound under sup-norm and leverage stability)**.** *Suppose that the assumptions of Lemmas 4.14, 4.11, and 4.15 hold, together with Assumption 16. Then*

$$\|\widehat{h}_{n,\lambda} - h\|_\infty \le \|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty + (1+\Lambda_{n,\lambda})\|h - h^{\mathrm{IS}}_{K_n}\|_\infty + \sqrt{6}M_f\sqrt{\frac{\lambda_{K_n,n}\ell_{n,\lambda}}{K_n}}. \tag{4.89}$$

Let

$$\alpha := \begin{cases} \beta/(d+1), & \text{in case (i) of Theorem 3.5,} \\ 1/(d+1), & \text{in case (ii) of Theorem 3.5.} \end{cases}$$

Then

$$\|\widehat{h}_{n,\lambda} - h\|_\infty = O_P\left(\sqrt{\frac{K_n \log(n)}{n}}\right) + O_P\left((\sqrt{d} + \sqrt{\log(K_n)})K_n^{-1/2-\alpha}\right) + O_P\left(\sqrt{\lambda_{K_n,n}}\right). \quad (4.90)$$

Corollary 4.10 (Explicit rate under sup-norm and leverage stability)**.** *Suppose that the assumptions of Theorem 4.16 hold, d is fixed, and*

$$K_n \asymp n^{1/(2(1+\alpha))}. \quad (4.91)$$

Suppose also that

$$\delta > \frac{2}{1+2\alpha} \quad (4.92)$$

and

$$\lambda_{K_n,n} = O\left(n^{-\frac{1+2\alpha}{2(1+\alpha)}} \log(n)\right). \quad (4.93)$$

Then

$$\|\widehat{h}_{n,\lambda} - h\|_\infty = O_P\left(n^{-\frac{1+2\alpha}{4(1+\alpha)}} \sqrt{\log(n)}\right). \quad (4.94)$$

Remark 18 (Comparison with Chen and Christensen)**.** Under sup-norm stability of the empirical projection, Chen and Christensen (2015) obtain the standard stochastic term $\sqrt{K_n \log(n)/n}$ and show that spline and wavelet series estimators attain the rate $(n/\log(n))^{-p/(2p+d)}$ when the regression function has smoothness p. To compare this benchmark with (4.94), set $p/d = 1/2 + \alpha$. The Chen-Christensen rate then becomes

$$\left(\frac{n}{\log(n)}\right)^{-p/(2p+d)} = n^{-\frac{1+2\alpha}{4(1+\alpha)}} (\log(n))^{\frac{1+2\alpha}{4(1+\alpha)}}.$$

Hence, the rate in (4.94) has the same polynomial exponent in n and is slower by the logarithmic factor $(\log(n))^{1/(4(1+\alpha))}$. This additional logarithmic factor originates from the $\sqrt{\log(K_n)}$ term in the RWNN sup-norm approximation bound.

5 Numerical experiments

In this section, we provide simulation evidence to highlight the different structural properties of random weight neural networks empirically. First, we provide simulation evidence for the sup-norm approximation rates we obtain by means of the importance sampling approximator derived in Section 3.3. When approximating a one-dimensional function with a known integral representation, we find that the IS approach achieves faster approximation rates than naïve random sampling. Then, we consider the effects of *scaling* on the conditioning of RWNNs in terms of the least-squares problem. We find that normalization can substantially improve the conditioning of RWNN sieves, in line with findings from the machine learning literature. Nonetheless, even without normalization, we show that the effective dimension $\mathcal{N}_{K,\lambda}$ is well-behaved and grows moderately with respect to K. We also study numerically the sup-norm stability conditions introduced in Section 4.6.4, and compare the spectral decay of the empirical sieve covariance matrix with stability factors $\Lambda_{n,\lambda}$ and $\ell_{n,\lambda}$. Finally, we consider inference for both smooth and non-smooth functionals of the regularized random weight neural network least squares estimator, following the theory developed in Section 4. We evaluate confidence intervals obtained both via the asymptotic normality of the estimated functionals, as well as bootstrapping.

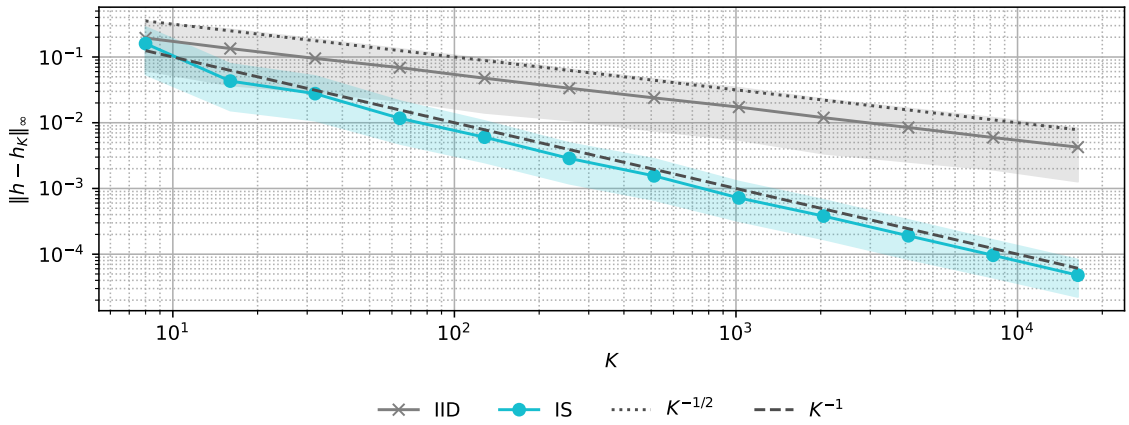


Figure 1: Sup-norm approximation rates in 1D for the importance sampling (IS) approximant (cyan) compared with the naïve i.i.d. sampling (IID) approximant (gray). Dashed and dotted lines are used to plot $K^{-1/2}$ and K^{-1} decay rates, respectively. For each K, we construct $\mathfrak{M} = 1000$ RWNN weight draws. Lines indicate mean values across replications; shaded areas display 90% quantile bands.

Throughout our simulations—save for the numerical analysis of L^∞ approximation via importance sampling—we consider both compactly-supported RWNN weight sampling distributions (uniform) and laws with unbounded support (Gaussian). Either type of distribution is widely employed in practice, depending on the setting; hence, we wish to provide numerical evidence for both, even though our formal theory only encompasses the former.

Additional numerical results are collected in Appendix G, including comparisons of residual, wild, and score-based bootstrap procedures for pointwise inference, as well as supplementary coverage, bias, and confidence interval results for the inferential experiments.

5.1 Sup-norm approximation

We start by providing numerical evidence towards the effectiveness of the stratified sampling approximation scheme that we derived in Section 3.3, cf. Theorem 3.5. Since the stratified sampling approach is adaptive in $h(x)$, we require explicit knowledge of its integral form in order to ensure that we perform stratified sampling correctly. We consider the one-dimensional function $h(x) = \mathbb{E}_{A,B}[(Ax + B)_+] = 1/4 + x^2/12$, where $(A, B)^\top \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1, 1]^2)$, for which it holds $f(A, B) = 1$ by construction. This simple structure allows us to design a straightforward importance sampling scheme: we construct a squared grid over $[-1, 1]^2$ of size $L = \lfloor \sqrt{K} \rfloor$, and we perform uniform sampling within each of the cells. As shown in the proof of Theorem 3.5, if $L^2 < K$, we can formally introduce $K - L^2$ additional sieve elements with zero weight to ensure that the RWNN has exactly K hidden units without impacting the approximation error rate. The sup-norm error is computed on a uniform grid over $[-1, 1]$ with 1000 points.

We provide a plot for the approximation performance of h_K^{IS} versus that of the equivalent i.i.d. sampling approximator h_K^{IID}, with $(A_k, B_k)^\top \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1, 1]^2)$, across $\mathfrak{M} = 1000$ independent draws in Figure 1. Our simulation indicates that indeed the sup-norm approximation rate achieved via stratified sampling is faster than the one that can be obtained through standard Monte Carlo sampling. This improvement, of course, rests on having prior knowledge regarding the regularity of the kernel $f(A, B)$ in the integral representation of $h(x)$. Hence, the design of efficient data-driven sampling schemes that are adaptive to the local smoothness, in the spirit of Bolager et al. (2023), appears to be an important avenue for future research on RWNNs.

5.2 Weight normalization, conditioning, and L^2 stability

Our consistency results from Section 4.1 highlight that the conditioning of the RWNN regression least squares problem is key in determining the quality of $\widehat{h}_{n,\lambda}$, as well as its consistency properties. Since the smallest eigenvalue of $\widehat{\Gamma}_{K_n,n}$ depends nonlinearly on both the specific draw of RWNN weights and the data, its decay as K_n grows is hard to characterize theoretically: we focus instead on studying the smallest singular value of the population matrix Γ_K for growing K, as well as the effective dimension factor $\mathcal{N}_{K,\lambda}$, with $\lambda \in \{10^{-6}, 10^{-4}, 10^{-2}, 1\}$. For the sake of clarity, we focus on two of the most commonly used activation functions: ReLU and the hyperbolic tangent (tanh). It is well-known that, in practice, weight normalization helps in improving training of neural networks, and we find that this also translates to improved conditioning for the population sieve covariance matrix Γ_K of RWNNs. The ReLU activation can be weight-normalized by using the mapping $A_k \mapsto A_k/\|A_k\|_2$; for the tanh activation, weight-normalization is achieved via the scaling $A_k \mapsto \sqrt{K}A_k$. These simulations allow us to directly investigate the spectral conditioning guarantee in Proposition 4.1 and the effective dimensionality bound obtained in Corollary 4.1. We use $n = 10^4$ draws from the uniform distribution over the square $[-1,1]^2$ to approximate the true population value of Γ_K, and we sample the random RWNN weights as $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1])$ and $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1])$.

In the top panel of Figure 2, we plot the empirical decay curves of $\overline{s}_K^2 = \lambda_{\min}(\Gamma_K)$ for each activation function, with both unnormalized and normalized weights, and include the floating-point numerical precision as a reference for numerical ill-conditioning. It is clear that normalization indeed helps substantially in improving the decay rate of $\overline{s}_K^2$ compared to the unnormalized baseline. Figure 2(b) and (c) nonetheless indicate that the growth of the effective dimension $\mathcal{N}_{K,\lambda}$ is moderate compared to K, across different choices of the regularization parameter λ.

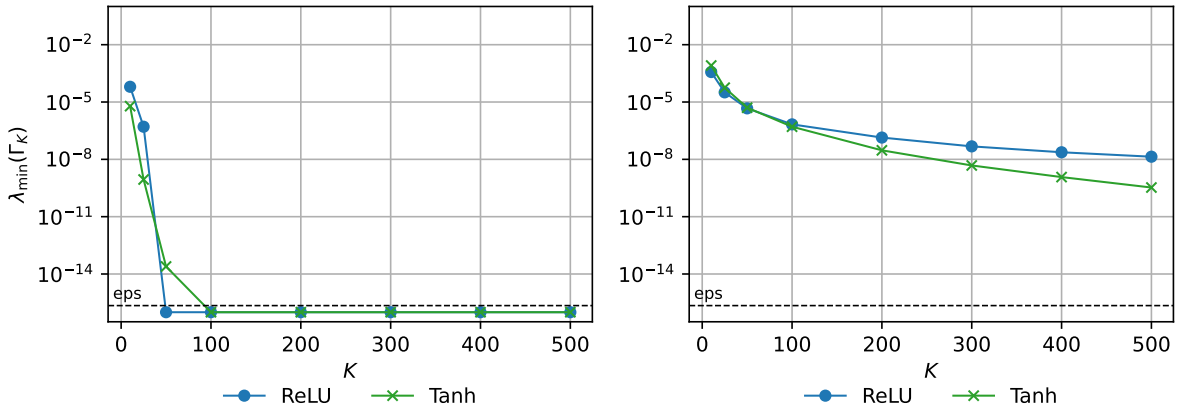
5.3 Sup-norm stability

As we discuss in detail in Section 4.6, less conservative sup-norm estimation rates can be achieved by introducing additional bounds on sup-norm stability and leverage quantities, as studied in Lemma 4.14. Since, in this setting, too, studying the empirical conditioning of the RWNN sieve in a theoretical manner is complex, we seek to numerically show that Assumption 16 is realistic. We draw an i.i.d. sample of $X_i \sim \mathcal{U}([-1,1])$ of size n, where $n \in \{100, 200, 400, 800, 1600, 3200, 6400\}$, and for each sample we compute the inner RWNN layer features for $\mathfrak{M} = 100$ draws of $\{A_k, B_k\}_{k=1}^K$, and these weights are not normalized. For our main specification, we let $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ element-wise, and set $K_n = \lfloor 3n^{1/3} \rfloor$, $\lambda_{K_n,n} = 10^{-5} n^{-2/3} \log(n)$. We then compute $\Lambda_{n,\lambda}$ and $\ell_{n,\lambda}$ as defined in (4.78) and (4.79), respectively, across MC = 100 replications.

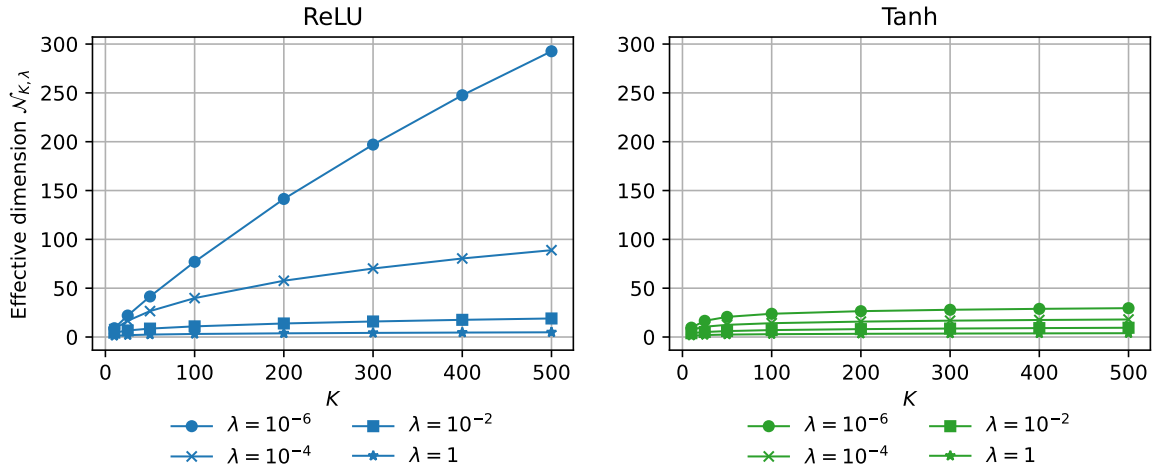
Figure 3 reports our simulation results. We again focus on the ReLU and hyperbolic tangent activation functions. The left panel confirms that the decay of the smallest eigenvalue of the empirical RWNN sieve covariance matrix $\lambda_{\min}(\widehat{\Gamma}_{K_n,n})$ is rapid when we do not explicitly threshold based on the numerical floating-point precision. For both maps, however, from the center and right panels we can see that both $\Lambda_{n,\lambda}$ and the ratio $\ell_{n,\lambda}/K_n$ are well-behaved as the sample size n grows, and sieve width K_n and ridge penalty $\lambda_{K_n,n}$ with it.

5.4 Asymptotic inference

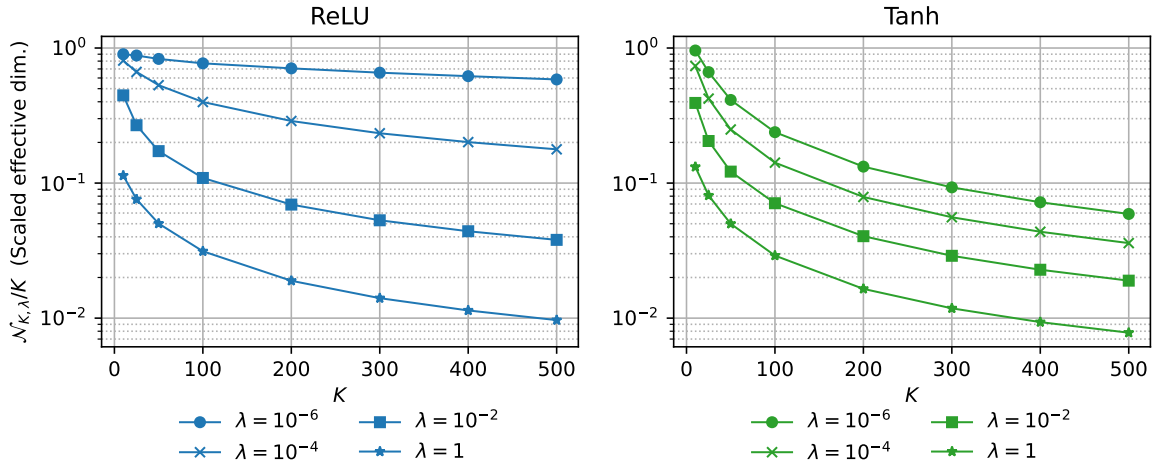
The L^2 RWNN inference results we derived in Section 4.5 provide direct theoretical justification for inference on smooth functionals. Although one could study many different functionals—for instance, partial mean and weighted integral functionals, or consumer surplus in demand analysis (see, for example, Newey, 1997)—we choose the integrated square functional $\Psi_2(h) = \int_{\mathcal{X}} h(x)^2 \omega(x) dx$; for simplicity, we employ a uniform weighting function, $\omega(x) = 1$, so that



(a) Least singular value of Γ_K with unnormalized (left) and normalized (right) activations.



(b) Effective dimension $\mathcal{N}_{K,\lambda}$ computed with ReLU (left) and tanh (right) activations.



(c) Scaled effective dimension $\mathcal{N}_{K,\lambda}/K$ computed with ReLU (left) and tanh (right) activations.

Figure 2: Conditioning of RWNN sieves at different sieve dimensions K and $\sigma \in \{\text{ReLU}, \tanh\}$, with $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1])$, $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1])$. Top panel plots the smallest eigenvalues of Γ_K computed using $n = 10^4$ i.i.d. draws from $\mathcal{U}([-1,1]^2)$, with and without normalization of activation functions. Middle panel shows the effective dimension $\mathcal{N}_{K,\lambda}$ across a range of ridge penalties $\lambda \in \{10^{-6}, 10^{-4}, 10^{-2}, 1\}$ without normalization. Bottom panel plots the associated scaled effective dimension $\mathcal{N}_{K,\lambda}/K$. For each K, we construct $\mathfrak{M} = 100$ RWNN weight draws. Lines indicate mean values across replications.

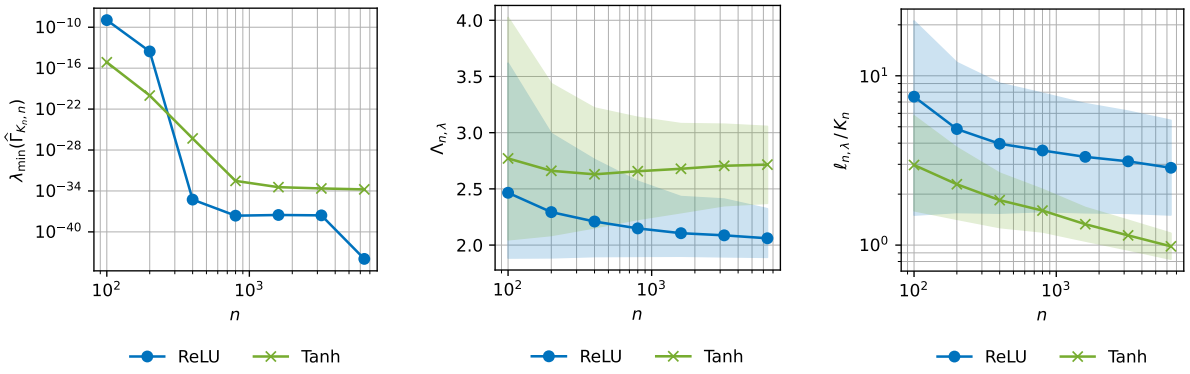


Figure 3: Sup-norm stability factors computed for $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1])$ and RWNNs with $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\sigma \in \{\text{ReLU}, \tanh\}$; activation functions are not normalized. For each n, we construct MC = 100 Monte Carlo data draws and $\mathfrak{M} = 100$ RWNN weight draws. Lines indicate mean values across replications; shaded areas display 90% quantile bands.

$\Psi_2(h) = \int_X h(x)^2 dx$. We study a one-dimensional regression problem with n i.i.d. draws from $Y_i = g_j(X_i) + \epsilon_i$, where $X_i \sim U([-1,1])$, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, $\sigma_\epsilon^2 = 0.04$, and $g_j : \mathbb{R} \to \mathbb{R}$, $j \in \{1,2,3\}$. Specifically, we study $g_1(x) := x + 5\varphi(10x)$, $g_2(x) := \sin(3\pi x/2)/[1 + 18x^2(\text{sgn}(x) + 1)]$ and $g_3(x) := \sin(\pi x/2)/[1 + 2x^2(\text{sgn}(x) + 1)]$.[1] For different choices of λ, we compute $\widehat{h}_{n,\lambda}$ and construct confidence intervals (CIs) for $\Psi(g_j)$ with a nominal coverage probability of 95%.

We can perform inference on Ψ_2 by leveraging the theory from Section 4.5. Let $\widehat{\epsilon}_i = Y_i - \widehat{h}_{n,\lambda}(X_i)$ be the RWNN regression residuals. Notice that

$$\Psi_2(\widehat{h}_{n,\lambda}) = \widehat{\beta}_{K_n,n,\lambda}^\top C_{g,K_n} \widehat{\beta}_{K_n,n,\lambda}, \quad \text{where} \quad C_{g,K_n} := \int_{-1}^{1} g_{K_n}(x) g_{K_n}(x)^\top \mathrm{d}x.$$

Since g_{K_n} is a known RWNN sieve basis, C_{g,K_n} can be computed numerically by an appropriate quadrature rule; we term this approach the "sandwich-type estimator" (SNW). Let

$$\Psi'_{\widehat{h}_{n,\lambda}}(g_{K_n}) := \left(\Psi'_{\widehat{h}_{n,\lambda}}(g_{1,K_n}), \ldots, \Psi'_{\widehat{h}_{n,\lambda}}(g_{K_n,K_n})\right)^\top,$$

so that for Ψ_2 we have $\Psi'_{\widehat{h}_{n,\lambda}}(g_{K_n}) = 2C_{g,K_n}\widehat{\beta}_{K_n,n,\lambda}$. The empirical counterpart of the ridge-adjusted sieve Riesz representer defined in (4.49) is hence

$$\widehat{\psi}_{K_n,\lambda}(x) \;:=\; g_{K_n}(x)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda} \Psi'_{\widehat{h}_{n,\lambda}}(g_{K_n}),$$

or, equivalently, $\widehat{\psi}_{K_n,\lambda}(x) := 2g_{K_n}(x)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda} C_{g,K_n} \widehat{\beta}_{K_n,n,\lambda}$. Let

$$\widehat{\Omega}_{K_n,n,\lambda} := \widehat{\Gamma}^{-1}_{K_n,n,\lambda} \left(n^{-1} \sum_{i=1}^{n} \widehat{\epsilon}_i^2 \, g_{K_n}(X_i) g_{K_n}(X_i)^\top \right) \widehat{\Gamma}^{-1}_{K_n,n,\lambda}.$$

The variance estimator associated with the empirical ridge-adjusted Riesz representer can be written as

$$\begin{aligned}
\widehat{\sigma}^2_{\Psi_2,\text{SNW}} \;&:=\; \frac{1}{n}\sum_{i=1}^{n} \widehat{\epsilon}_i^2 \widehat{\psi}_{\Psi_2,\lambda}(X_i)^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\epsilon}_i^2 \left(g_{K_n}(X_i)^\top \widehat{\Gamma}^{-1}_{K_n,\lambda} \Psi'_{\widehat{h}_{n,\lambda}}(g_{K_n})\right)^2 \\
&= \Psi'_{\widehat{h}_{n,\lambda}}(g_{K_n})^\top \widehat{\Omega}_{K_n,n,\lambda} \Psi'_{\widehat{h}_{n,\lambda}}(g_{K_n}) \;=\; 4\widehat{\beta}^\top_{K_n,n,\lambda} C_{g,K_n} \widehat{\Omega}_{K_n,n,\lambda} C_{g,K_n} \widehat{\beta}_{K_n,n,\lambda},
\end{aligned}$$

[1] Here φ denotes the standard Gaussian density and $\text{sgn}(x) \in \{-1, 0, 1\}$ denotes the sign function. These functions, plotted in Appendix G.3, Figure G.1, have been previously used by Hall and Horowitz (2013) and Calonico et al. (2018), among others, in Monte Carlo simulations of nonparametric inference procedures.

which is then the sandwich-type variance estimator.

Alternatively, since $X \sim U([-1,1])$, the Riesz representer of Ψ_h' with respect to $\|\cdot\|_{L^2(X)}$ inner product is, for $\Psi \equiv \Psi_2$, directly $\psi(x) = 4h(x)$. Indeed, for every direction $v \in L^2(X)$, $\Psi_h'(v) = 2\int_{-1}^{1} h(x)v(x)\mathrm{d}x = 4\mathbb{E}[h(X)v(X)] = \langle v, 4h\rangle_{L^2(X)}$. Thus, we may consider the direct plug-in functional and variance estimators based on the unregularized $L^2(X)$ Riesz representer $\psi = 4h$, given by, respectively,

$$\Psi_2(\widehat{h}_{n,\lambda}) = \int_{-1}^{1} \widehat{h}_{n,\lambda}(x)^2 dx \quad \text{and} \quad \widehat{\sigma}^2_{\Psi_2,\mathrm{RSZ}} := \frac{16}{n}\sum_{i=1}^{n}\left(\widehat{\epsilon}_i \widehat{h}_{n,\lambda}(X_i)\right)^2,$$

where the integral can again be computed by a suitable quadrature rule. We call these the Riesz-type functional and variance estimators (RSZ).

An $100(1-\alpha)\%$ confidence interval for $\Psi_2(h)$ can thus be constructed as

$$\mathrm{CI}_{1-\alpha,b}(\Psi_2) = \left[\Psi_2(\widehat{h}_{n,\lambda}) - \sqrt{\frac{\widehat{\sigma}^2_{\Psi_2,b}}{n}}\, z_{1-\alpha/2},\ \Psi_2(\widehat{h}_{n,\lambda}) + \sqrt{\frac{\widehat{\sigma}^2_{\Psi_2,b}}{n}}\, z_{1-\alpha/2}\right],$$

where $b \in \{\mathrm{SNW}, \mathrm{RSZ}\}$ and z_α denotes the α-quantile of the standard Gaussian distribution. Table 2 summarizes simulation results based on three different RWNN specifications with tanh activation function. We use $\mathrm{MC} = 1000$ replications of $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1])$ and four combinations of sample size n and RWNN sieve dimension K_n. The random weights $\{A_k, B_k\}_{k=1}^{K_n}$ are resampled at each Monte Carlo replication: for the Gaussian-Gaussian specification, $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_A^2)$ and $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_B^2)$, with $\sigma_A = 5$ and $\sigma_B = 1$; for the Gaussian-Uniform, $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_A^2)$ and $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-\sigma_B,\sigma_B])$, with $\sigma_A = 5$ and $\sigma_B = 2$; for the Uniform-Uniform, $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-\sigma_A,\sigma_A])$ and $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-\sigma_B,\sigma_B])$, with $\sigma_A = 10$ and $\sigma_B = 5$. In all settings, we use a fixed ridge regularization factor of $\lambda_{K_n,n} = 10^{-6}$ and a trapezoidal integration scheme to compute both the sandwich- and Riesz-type estimators.[2] Appendix G.2 contains results for simulations made with $\lambda \in \{10^{-8}, 10^{-4}, 10^{-2}\}$. For both variance estimators, the coverage probabilities approach the nominal level as K_n and n increase, and interval lengths are extremely similar throughout. In the small-sample regimes where $n = 100$ and $K_n = 15$ or $n = 300$ and $K_n = 20$, we note that $\widehat{\sigma}^2_{\Psi_2,\mathrm{SNW}}$ appears to be marginally more robust. At the largest sample-sieve combination, the two procedures produce very similar results, with coverage close to the nominal level for all three functions.

5.5 Bootstrap inference

While our theory applies to both smooth and linear functionals, it generally requires explicit computation of the ridge-adjusted Riesz representer. The bootstrap, by comparison, is an effective and flexible approach that does not generally require explicit computation of functional or directional quantities. It has been widely applied in nonparametric statistics; see, for example, Horowitz (2009); Hall and Horowitz (2013) and Cavaliere et al. (2024) for a recent discussion of bias-correction refinements. Although our inferential theory does not encompass bootstrap validity, we believe it is nonetheless interesting to showcase its empirical performance, especially for practitioners. In this section, we thus focus on a straightforward nonparametric residual bootstrap construction; in Appendix G we provide detailed simulations for the point evaluation functional and apply two additional bootstrapping schemes—a Gaussian wild bootstrap and the

[2] In either case, we integrate using an equally-spaced grid of 1001 points (so that the zero belongs to the grid) over $[-1,1]$. In our simulations, we find that this choice is numerically robust and produces no meaningful integration artifacts unless we set $\lambda = 10^{-2}$, a choice that yields substantial bias and poor CI coverage. Even then, the differences in bias are negligible.

Table 2: Asymptotic inference results for $\Psi_2(\widehat{h}_{n,\lambda})$ with $\lambda = 10^{-6}$.

			Gaussian–Gaussian			Gaussian–Uniform			Uniform–Uniform		
n	K_n	Method	Bias	Coverage	Length	Bias	Coverage	Length	Bias	Coverage	Length
$h = g_1$											
100	15	RSZ	0.089	0.891	0.242	0.086	0.896	0.244	-0.022	0.877	0.269
		SNW	0.089	0.901	0.254	0.086	0.898	0.258	-0.022	0.900	0.287
300	20	RSZ	0.004	0.930	0.146	0.002	0.929	0.146	-0.040	0.928	0.151
		SNW	0.004	0.936	0.149	0.002	0.939	0.150	-0.040	0.934	0.155
1000	30	RSZ	0.017	0.938	0.081	0.016	0.939	0.081	0.012	0.939	0.082
		SNW	0.017	0.935	0.082	0.016	0.941	0.082	0.012	0.937	0.082
3000	40	RSZ	-0.008	0.950	0.047	-0.008	0.951	0.047	-0.009	0.949	0.047
		SNW	-0.008	0.951	0.047	-0.008	0.954	0.047	-0.009	0.951	0.047
$h = g_2$											
100	15	RSZ	0.095	0.891	0.152	0.104	0.894	0.152	0.094	0.876	0.152
		SNW	0.095	0.914	0.160	0.104	0.920	0.160	0.094	0.911	0.161
300	20	RSZ	0.039	0.935	0.091	0.041	0.937	0.091	0.043	0.933	0.091
		SNW	0.039	0.939	0.093	0.041	0.943	0.093	0.043	0.942	0.093
1000	30	RSZ	0.010	0.943	0.051	0.012	0.942	0.051	0.016	0.939	0.051
		SNW	0.010	0.945	0.051	0.012	0.947	0.051	0.016	0.942	0.051
3000	40	RSZ	0.002	0.947	0.029	0.003	0.949	0.029	0.005	0.947	0.029
		SNW	0.002	0.948	0.030	0.003	0.947	0.030	0.005	0.951	0.029
$h = g_3$											
100	15	RSZ	0.085	0.913	0.159	0.094	0.907	0.159	0.105	0.891	0.157
		SNW	0.085	0.928	0.166	0.094	0.927	0.166	0.105	0.915	0.165
300	20	RSZ	0.034	0.923	0.095	0.034	0.922	0.095	0.039	0.923	0.095
		SNW	0.034	0.935	0.097	0.034	0.938	0.097	0.039	0.931	0.097
1000	30	RSZ	0.020	0.953	0.053	0.020	0.949	0.053	0.023	0.944	0.053
		SNW	0.020	0.950	0.053	0.020	0.949	0.053	0.023	0.945	0.053
3000	40	RSZ	0.001	0.943	0.031	0.001	0.943	0.031	0.003	0.941	0.031
		SNW	0.001	0.941	0.031	0.001	0.942	0.031	0.003	0.943	0.031

Notes: Bias, coverage rate, and interval length of asymptotic confidence intervals for $\Psi_2(h)$ estimated using MC = 1000 replications. The method column reports the type of functional and variance estimators (sandwich-type or Riesz-type) used to compute asymptotic confidence intervals. The Gaussian-Gaussian, Gaussian-Uniform, and Uniform-Uniform column groupings indicate random sampling laws of A_k and B_k, respectively. Bias is scaled by a factor of 10 to improve legibility.

score bootstrap of Chen and Christensen (2018). For the sake of space, we limit ourselves to two cases, both involving g_1, as defined above: we perform inference on the evaluation functional at $x = 0$, $\Psi_0(g_1) = g_1(0)$, and the expected interaction functional $\Psi_1(g_1) = \mathbb{E}[g_1(X)(1 + X)]$.

To implement the residual construct, we generate independent draws of the response variable as $Y_i^* := \widehat{h}_{n,\lambda}(X_i) + \epsilon_i^*$, $i = 1, \ldots, n$. Let $\mathbb{P}_{\widehat{\epsilon},n}$ denote the empirical distribution of the regression residuals. The innovations are sampled independently from the centered empirical residual distribution, meaning $\epsilon_i^* := \widehat{\epsilon}_i^* - \bar{\widehat{\epsilon}}$, with $\widehat{\epsilon}_i^* \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\widehat{\epsilon},n}$ and $\bar{\widehat{\epsilon}} := n^{-1}\sum_{i=1}^n \widehat{\epsilon}_i$. Let $\widehat{h}_{n,\lambda}^*$ denote the RWNN estimator obtained by refitting the model to the corresponding bootstrap sample (without changing the sieve or the ridge penalty). Let $z_\alpha^*(\Psi)$ denote the conditional α-quantile of the bootstrap distribution of $\Psi(\widehat{h}_{n,\lambda}^*)$, given the original sample. The corresponding residual percentile-bootstrap confidence interval is

$$\mathrm{CI}_{1-\alpha}(\Psi) \;:=\; \left[z_{\alpha/2}^*,\; z_{1-\alpha/2}^*\right], \quad \Psi \in \{\Psi_0, \Psi_1\}.$$

Since Ψ_0 is irregular with respect to the $L^2(X)$ norm, contrary to Ψ_1, we also wish to compare their empirical Rayleigh stability factors to ascertain if they also align with our directional inference theory. Recall from Section 4.3.3 that

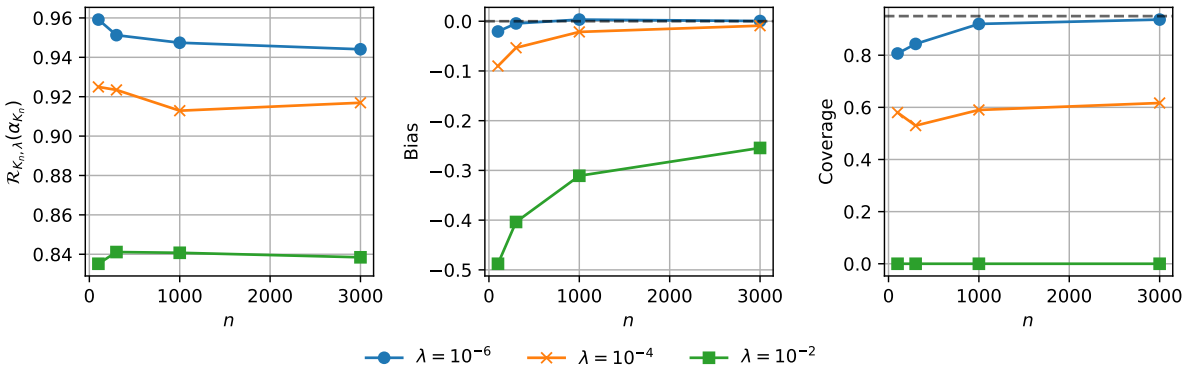
$$\mathcal{R}_{K_n,\lambda}(\alpha_{K_n}) = \frac{\alpha_{K_n}^\top \Gamma_{K_n,\lambda}^{-1}\Gamma_{K_n}\Gamma_{K_n,\lambda}^{-1}\alpha_{K_n}}{\alpha_{K_n}^\top \Gamma_{K_n,\lambda}^{-1}\alpha_{K_n}}.$$

In our setup, it is simple to show that $\alpha_{K_n}^{(0)} = g_{K_n}(0)$ for Ψ_0 and $\alpha_{K_n}^{(1)} = \mathbb{E}[g_{K_n}(X_i)(1 + X_i)]$ for Ψ_1. Therefore, the empirical counterpart of $\mathcal{R}_{K_n,\lambda}(\alpha_{K_n})$ can be readily computed by plug-in quantities, allowing us to also quantify the directional conditioning inherent to the inferential problem.

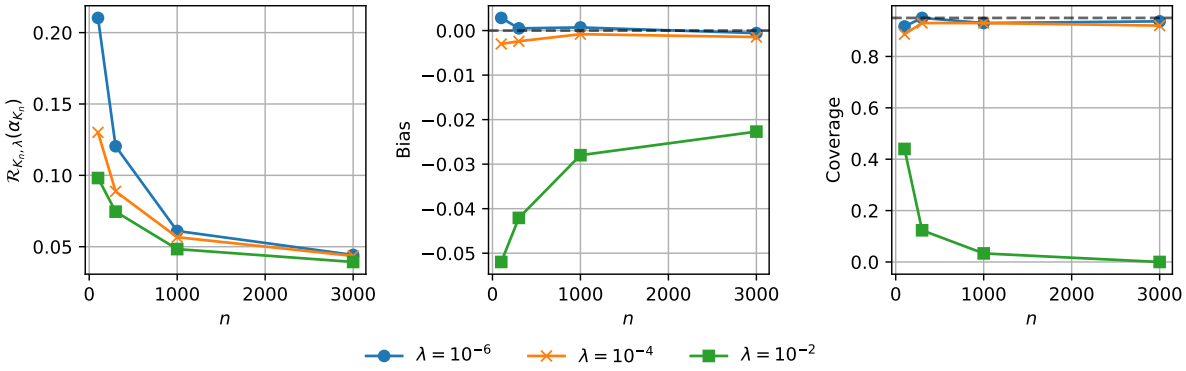
In the data generating process we use again $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1, 1])$, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, $\sigma_\epsilon^2 = 0.04$. The RWNN uses tanh activation, $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_A^2)$ and $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_B^2)$, with $\sigma_A = 5$ and $\sigma_B = 1$, and the weights are not normalized. In Appendix G.3, we provide additional figures for Gaussian-Uniform and Uniform-Uniform weight sampling design. We set the sample size to $n = \{100, 300, 1000, 3000\}$ and the sieve size to $K_n = \{15, 20, 30, 40\}$, following the simulation designs for the mean-squared functional in Section 5.4 above. For both functionals, MC = 300 and $\mathfrak{M} = 300$ Monte Carlo and bootstrap resamplings are used, and the RWNN weights are resampled at each new data draw. Similarly to the other simulations, we compare results at different fixed values for the ridge regularization parameter, and use $\lambda \in \{10^{-6}, 10^{-4}, 10^-2\}$. The results are shown in Figure 4, where we plot values averaged across all Monte Carlo replications. Panel (a) shows that the empirical Rayleigh quotients for the evaluation functional Ψ_0 are all well-behaved as n grows, and across different scalings of the ridge penalty. On the other hand, Panel (b) indicates that, even though the Rayleigh quotients appear to be decaying as n increases, inference is not adversely affected. Indeed, we can see that bias for $\Psi_1(g_1)$ is substantially smaller than that of the evaluation functional, and CIs retain good coverage across sample sizes with $\lambda = 10^{-6}$. Setting $\lambda = 10^{-4}$, shows that the point-evaluation functional is much more sensitive to regularization than Ψ_1, which is natural given the graph of g_1. Finally, in both panels, we note that the RWNN estimator shows large bias and very poor coverage when we set $\lambda = 10^{-2}$: this is expected, as our asymptotic theory indeed emphasizes that an appropriate choice of λ is key in obtaining valid RWNN inference.

6 Conclusion

We have developed approximation, estimation, and inference results for random weight neural network sieves that explicitly accommodate asymptotic ill-conditioning of their population covariance matrices. Under polynomial covering and Lipschitz conditions, we establish spectral



(a) $\Psi_0(g_1) = g_1(0)$



(b) $\Psi_1(g_1) = \mathbb{E}[g_1(X)(1+X)]$

Figure 4: Bootstrap inference results for functionals of g_1 with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1])$, $\epsilon_i \sim \mathcal{N}(0,\sigma_\epsilon^2)$, $\sigma_\epsilon^2 = 0.04$. RWNNs use tanh activation, $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_A^2)$ and $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_B^2)$, with $\sigma_A = 5$ and $\sigma_B = 1$. Ridge penalty is set as $\lambda \in \{10^{-6}, 10^{-4}, 10^{-2}\}$. Lines indicate averages across Monte Carlo replications.

tail bounds that imply vanishing smallest eigenvalues and control the ridge effective dimension. We also derive improved L^2 approximation rates using control neighbors integration, as well as sup-norm approximation bounds using stratified sampling. For ridge-regularized least squares, our results provide L^2 and uniform estimation bounds under appropriate stability conditions, together with conditional asymptotic normality for linear and smooth functionals and consistent ridge-adjusted directional variance estimation. In particular, directional inference can remain valid even when the population feature covariance matrix is singular. Beyond the squared loss setting, we also develop a generic sieve framework for estimation and smooth-functional inference: it accommodates other loss functions, including absolute, Huber, quantile, and logistic losses, provided the corresponding identification, curvature, and empirical process conditions are satisfied.

The results we obtain rely on explicit approximation, regularization bias, and stability conditions, whose verification may be challenging in general random feature designs. Some stratified constructions also require knowledge of the source representation. Developing data-driven sampling schemes and verifying the inferential conditions for broader feature distributions are natural directions for further research. As evidenced by our numerical results in Section 5, the choice of the ridge regularization parameter $\lambda \geq 0$ is a fundamental determinant of the inferential properties of the RWNN estimator. In practice, the choice of λ is non-trivial, and, even though the use of out-of-sample and cross-validation methods is natural (Yang, 2007), a comprehensive theoretical treatment of model selection procedures for RWNN regression models would be highly desirable as a next step. Another important extension concerns bias correction

procedures for inference, particularly in settings where ridge regularization or random feature approximation induces substantial finite sample bias. From a practical point of view, we note that our numerical experiments focus on bootstrap procedures without an explicit theory for their validity or an explicit bias correction. Since sieve approximation and ridge regularization can introduce bias whose magnitude also depends on the realized hidden-layer parameters, adapting bootstrap and bias correction methods to RWNN inference is a natural direction for future research (Hall and Horowitz, 2013; Calonico et al., 2018; Cavaliere et al., 2024, 2026). Another important direction we would like to pursue in the future is to establish Gaussian approximations for the supremum of the normalized estimation error, which could provide a theoretical basis for constructing asymptotically valid uniform confidence bands for RWNN regression functions.

References

Alaoui, A. and Mahoney, M. W. (2015). Fast Randomized Kernel Ridge Regression with Statistical Guarantees. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. (2017). Random Fourier Features for Kernel Ridge Regression: Approximation Bounds and Statistical Guarantees. In *Proceedings of the 34th International Conference on Machine Learning*, pages 253–262. PMLR.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38.

Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Bolager, E. L., Burak, I., Datar, C., Sun, Q., and Dietrich, F. (2023). Sampling weights of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 63075–63116. Curran Associates, Inc.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Braun, A., Kohler, M., Langer, S., and Walk, H. (2024). Convergence rates for shallow neural networks learned by gradient descent. *Bernoulli*, 30(1):475–502.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018). On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference. *Journal of the American Statistical Association*, 113(522):767–779.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368.

Carratino, L., Rudi, A., and Rosasco, L. (2018). Learning with SGD and random features. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 10192–10203.

Cavaliere, G., Gonçalves, S., Nielsen, M. Ø., and Zanelli, E. (2024). Bootstrap Inference in the Presence of Bias. *Journal of the American Statistical Association*, 119(548):2908–2918.

Cavaliere, G., Gonçalves, S., Nielsen, M. Ø., and Zanelli, E. (2026). Improved inference for nonparametric regression and regression-discontinuity designs. *Working Paper*.

Chen, X. (2007). Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6, pages 5549–5632. Elsevier.

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.

Chen, X. and Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84.

Chen, X. and Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46–60.

Chen, X. and Pouzo, D. (2012). Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals. *Econometrica*, 80(1):277–321.

Chen, X. and Shen, X. (1998). Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica*, 66(2):289.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691.

Cheng, T. S., Lucchi, A., Kratsios, A., Dokmanić, I., and Belius, D. (2023). A theoretical analysis of the test error of finite-rank kernel ridge regression. In *Advances in Neural Information Processing Systems*, volume 36, pages 4767–4798. Curran Associates, Inc.

De Ryck, T., Mishra, S., Shang, Y., and Wang, F. (2025). Approximation theory and applications of randomized neural networks for solving high-dimensional PDEs. *Working Paper [arXiv:2501.12145]*.

Fan, K. (1949). On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America*, 35(11):652–655.

Gonon, L. (2023). Random Feature Neural Networks Learn Black-Scholes Type PDEs Without Curse of Dimensionality. *Journal of Machine Learning Research*, 24:1–51.

Gonon, L., Grigoryeva, L., and Ortega, J.-P. (2023). Approximation bounds for random neural networks and reservoir systems. *The Annals of Applied Probability*, 33(1):28–69.

Grenander, U. (1981). *Abstract Inference*. New York: Wiley.

Guignard, F., Amato, F., and Kanevski, M. (2021). Uncertainty quantification in extreme learning machine: Analytical developments, variance estimates and confidence intervals. *Neurocomputing*, 456:436–449.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, NY, USA.

Hall, P. and Horowitz, J. (2013). A Simple Bootstrap Method for Constructing Nonparametric Confidence Bands for Functions. *The Annals of Statistics*, 41(4):1892–1921.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer, New York, NY, USA.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635.

Ismagilov, R. S. (1968). On n-dimensional diameters of compacts in a Hilbert space. *Functional Analysis and Its Applications*, 2(2):125–132.

Klusowski, J. M. and Barron, A. R. (2018). Approximation by Combinations of ReLU and Squared ReLU Ridge Functions With ell^1 and ell^0 Controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656.

Lanthaler, S. and Nelsen, N. (2023). Error bounds for learning with vector-valued random features. In *Proceedings of the Neural Information Processing Systems*.

Leluc, R., Portier, F., Segers, J., and Zhuman, A. (2025). Speeding up Monte Carlo integration: Control neighbors for optimal convergence. *Bernoulli*, 31(2):1160–1180.

Liu, X., Mao, T., and Xu, J. (2025). Integral Representations of Sobolev Spaces via ReLU^k Activation Function and Optimal Error Estimates for Linearized Networks. *Working Paper [arXiv:2505.00351]*.

Mao, T., Siegel, J. W., and Xu, J. (2026). Approximation Rates for Shallow ReLU^k Neural Networks on Sobolev Spaces via the Radon Transform. *SIAM Journal on Mathematical Analysis*, 58(2).

Mehmood, M. A. and Gonon, L. (2026). Random neural network expressivity for non-linear partial differential equations. In *The Fortieth Annual Conference on Neural Information Processing Systems*.

Mei, S. and Montanari, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766.

Ming, P. and Yu, H. (2026). Spectral Convergence of Random Feature Method in Multiple Dimensions. *Working Paper [arXiv:2609.03401]*.

National Institute of Standards and Technology (2010). *NIST Handbook of Mathematical Functions*. Cambridge University Press, Cambridge New York Melbourne.

Neufeld, A. and Schmocker, P. (2023). Universal approximation property of Banach space-valued random feature models including random neural networks. *Working Paper [arXiv:2312.08410]*.

Neufeld, A. and Schmocker, P. (2026). Universal approximation results for neural networks with non-polynomial activation function over non-compact domains. *Analysis and Applications*, 24(5):1123–1173.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168.

Novak, E. (2016). Some Results on the Complexity of Numerical Integration. In Cools, R. and Nuyens, D., editors, *Monte Carlo and Quasi-Monte Carlo Methods*, volume 163, pages 161–183. Springer International Publishing, Cham.

Pauwels, E., Bach, F., and Vert, J.-P. (2018). Relating Leverage Scores and Density using Regularized Christoffel Functions. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Pouzo, D. (2016). On the Non-Asymptotic Properties of Regularized M-estimators. *Working Paper [arXiv:1512.06290].*

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

Rahimi, A. and Recht, B. (2008). Uniform approximation of functions with random bases. In *46th Annual Allerton Conference on Communication, Control, and Computing.*

Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. In *Proceedings of the Neural Information Processing Systems*, pages 3218–3228.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897.

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics*, 25(6):2555–2591.

Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. (2023). Asymptotic properties of neural network sieve estimators. *Journal of Nonparametric Statistics*, 0(0):1–30.

Shen, X. and Wong, W. H. (1994). Convergence Rate of Sieve Estimates. *The Annals of Statistics*, 22(2):580 – 615.

Shen, Y. and Bellec, P. C. (2020). Asymptotic normality and confidence intervals for derivatives of 2-layers neural network in the random features model. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18625–18636. Curran Associates, Inc.

Siegel, J. W. and Xu, J. (2020). Approximation rates for neural networks with general activation functions. *Neural Networks*, 128:313–321.

Singh, R. and Vijaykumar, S. (2023). Kernel Ridge Regression Inference. *Working Paper [arXiv:2302.06578].*

Sonoda, S. and Murata, N. (2017). Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268.

Steland, A. (2021). Consistency of Extreme Learning Machines and Regression under Non-Stationarity and Dependence for ML-Enhanced Moving Objects. *Working Paper [arXiv:2005.11115].*

Stone, C. J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4).

Takhanov, R. (2026). On the Intrinsic Dimensions of Data in Kernel Learning. *Working Paper [arXiv:2601.16139].*

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer, New York ; London.

Tuo, R. and Zou, L. (2024). Asymptotic Theory for Linear Functionals of Kernel Ridge Regression. *Forthcoming in Statistica Sinica.*

Van der Vaart, A. (1991). On Differentiable Functionals. *The Annals of Statistics*, 19(1):178–204.

van der Vaart, A. W. and Wellner, J. A. (2023). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer International Publishing, Cham.

Vershynin, R. (2026). *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2 edition.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473.

Appendix

Statistical Inference with Ill-Conditioned Random Weight Neural Network Sieves

Giovanni Ballarin Lyudmila Grigoryeva Lukas Gonon

Appendices

A Proofs for RWNN approximation **55**
A.1 L^2 approximation and control-neighbors weights . . . 55
A.2 L^2 approximation on integral Sobolev classes . . . 58
A.3 Sup-norm approximation via stratified sampling . . . 58
A.4 Uniform Monte Carlo approximation via metric entropy . . . 67

B Proofs for spectral properties and L^2 ridge estimation **70**
B.1 Spectral deterioration and ridge effective dimension . . . 70
B.2 Ridge-normalized covariance stability and L^2 estimation . . . 71
B.3 RWNN specialization and primitive rates . . . 74

C Proofs for directional and smooth functional inference **76**
C.1 Ridge-normalized linearization and directional asymptotic normality . . . 76
C.2 Ridge-adjusted variance estimation and studentization . . . 80
C.3 Smooth-functional inference . . . 84
C.4 Proof of Example 1 . . . 85

D Proofs for sup-norm ridge estimation **88**
D.1 Uniform stochastic error . . . 88
D.2 Ridge pseudo-projection and spectral stability . . . 94
D.3 Prediction space and leverage based stability . . . 96

E Theory for generic sieve estimation and inference **100**
E.1 A baseline L^2 risk bound . . . 100
E.2 Generic sieve estimation and β-Hölder RWNN rates . . . 101
E.3 Asymptotic normality for smooth functionals . . . 117

F Auxiliary results **123**
F.1 Geometric and sampling inequalities . . . 123
F.2 Matrix concentration and perturbation inequalities . . . 123
F.3 Entropy and empirical-process inequalities . . . 123

G Additional numerical experiments and results **125**
G.1 Residual, wild and score bootstrap inference for evaluation functional . . . 125
G.2 Additional tables . . . 127
G.3 Additional figures . . . 131

A Proofs for RWNN approximation

A.1 L^2 approximation and control-neighbors weights

In this appendix, we recall the control-neighbors approximation theorem of Leluc et al. (2025), verify the conditions needed for its application, and present the proofs of the approximation results in Subsection 3.1, including the bounds on the control-neighbors weights.

Theorem A.1 (Leluc et al., 2025, Theorem 1)**.** *Let Assumptions 1, 4 and 5 hold, and let* $\theta_k = (A_k^\top, B_k)^\top$, $k = 1, \dots, K$, *be i.i.d. from* π_θ. *Then for* $K \geq 4$, *the control-neighbors estimate*

$$h_K^{\theta,\omega\text{-}cnn}(x) := \sum_{k=1}^{K} \omega_{k,K} \phi_x(\theta_k),$$

where $\{\omega_{k,K}\}_{k=1}^K$ *are the nearest-neighbor weights introduced in Section 2 in Leluc et al. (2025), satisfies*

$$\mathbb{E}_{A,B}\big[|h(x) - h_K^{\theta,\omega\text{-}cnn}(x)|^2\big]^{1/2} \leq C_{cnn}(x) K^{-1/2-\beta/(d+1)},$$

where

$$C_{cnn}(x) := L_{\phi,x} C_0^{-\beta/(d+1)} \sqrt{\Gamma\left(\frac{2\beta}{d+1} + 2\right)} \left(2\sqrt{\frac{\beta}{d+1} + 1} + 2^{\beta/(d+1)} \sqrt{10 \frac{C_1}{C_0}}\right),$$

where $C_0, C_1 > 0$ *are constants such that* $C_0 r^{d+1} \leq \pi_\theta\big(B_r^{\rho_\Pi}(\theta)\big) \leq C_1 r^{d+1}$ *for any* $\theta \in \Pi$ *and* $r \in (0, D_\Pi]$, *with* $D_\Pi = \sup_{\theta_1,\theta_2 \in \Pi} \rho_\Pi(\theta_1, \theta_2) < \infty$.

We note that, by Assumption 5 (iii), $B_{r/M_\rho}^{\|\cdot\|}(\theta) \subseteq B_r^{\rho_\Pi}(\theta) \subseteq B_{r/m_\rho}^{\|\cdot\|}(\theta)$. Hence Assumption 5 (i)-(ii) implies that, for all sufficiently small $r > 0$, $c_b c_\pi V_{d+1} M_\rho^{-(d+1)} r^{d+1} \leq \pi_\theta\big(B_r^{\rho_\Pi}(\theta)\big) \leq c_u V_{d+1} m_\rho^{-(d+1)} r^{d+1}$, uniformly in $\theta \in \Pi$. Since Π is bounded under ρ_Π, these local bounds imply, after adjusting the constants, the metric small-ball condition required in Theorem A.1.

Proof of Proposition 3.1. For $\theta_1, \theta_2 \in \Pi$ and $x \in \mathcal{X}$,

$$\begin{aligned}
|\phi_x(\theta_1) - \phi_x(\theta_2)| &\leq M_f |\sigma(A_1^\top x + B_1) - \sigma(A_2^\top x + B_2)| + M_\sigma |f(\theta_1) - f(\theta_2)| \\
&\leq M_f L_\sigma |(A_1 - A_2)^\top x + (B_1 - B_2)| + M_\sigma L_f \rho_\Pi(\theta_1, \theta_2)^\beta \\
&\leq M_f L_\sigma L_\rho \rho_\Pi(\theta_1, \theta_2) + M_\sigma L_f \rho_\Pi(\theta_1, \theta_2)^\beta \\
&\leq \big(M_f L_\sigma L_\rho D_\Pi^{1-\beta} + M_\sigma L_f\big) \rho_\Pi(\theta_1, \theta_2)^\beta = L_\phi\, \rho_\Pi(\theta_1, \theta_2)^\beta,
\end{aligned}$$

where the first inequality uses the triangle inequality together with Assumption 3, the second uses the same assumption and Assumption 2, the third uses (3.1), and the fourth uses $\rho_\Pi(\theta_1, \theta_2) \leq D_\Pi$ to bound $\rho_\Pi(\theta_1, \theta_2) \leq D_\Pi^{1-\beta} \rho_\Pi(\theta_1, \theta_2)^\beta$. This proves the claim. □

Proof of Theorem 3.1. By the hypotheses of the theorem, Proposition 3.1 yields that Assumption 3 implies Assumption 4 with the integrand $\phi_x(\theta) := f(A, B)\sigma(A^\top x + B)$ being β-Hölder in θ under ρ_Π, uniformly in $x \in \mathcal{X}$, with constant

$$L_{\phi,x} \leq L_\phi := M_f L_\sigma L_\rho D_\Pi^{1-\beta} + M_\sigma L_f,$$

where D_Π and M_σ are as in Assumption 3.

Assumption 5 provides the measure conditions of Theorem A.1. Applying Theorem A.1 and replacing $L_{\phi,x}$ by the uniform bound L_ϕ yields, for each $x \in \mathcal{X}$,

$$\mathbb{E}_{A,B}\big[|h(x) - h_K^{\theta,\omega\text{-}\mathrm{cnn}}(x)|^2\big] \leq C_{\mathrm{cnn}}^2 K^{-1-2\beta/(d+1)}, \tag{A.1}$$

where $C_{\texttt{cnn}} > 0$ depends only on d, β, L_ϕ, c_b, c_u, c_π, and hence only on d, β, M_f, L_f, L_σ, L_ρ, M_σ, c_b, c_u, c_π, r_0, m_ρ, M_ρ, D_Π.

Since the right-hand side of (A.1) does not depend on x, Fubini's theorem yields

$$\|h(X) - h_K^{\theta,\omega\text{-cnn}}(X)\|^2_{L^2(X,A,B)} = \mathbb{E}_{A,B}\left[\mathbb{E}_X\left[|h(X) - h_K^{\theta,\omega\text{-cnn}}(X)|^2\right]\right] \le C^2_{\texttt{cnn}} K^{-1-2\beta/(d+1)}.$$

Taking a square root yields the stated bound with $C_q = C_{\texttt{cnn}}$.

Finally, by (3.3), $h_K^{\theta,\omega\text{-cnn}} \in \mathcal{H}_K$, with output coefficients $W_k := \omega_{k,K} f(A_k, B_k)$, $k = 1, \dots, K$. The weights $\omega_{k,K}$ are measurable functions of $\{\theta_j\}_{j=1}^K$ by construction (Leluc et al., 2025), and f is measurable. Hence W is measurable with respect to $\sigma(\theta_1, \dots, \theta_K)$, completing the proof. □

Proof of Theorem 3.2. By Theorem 3.1,

$$\|h(X) - h_K^{\theta,W}(X)\|^2_{L^2(X,A,B)} \le C_q^2 K^{-1-2\beta/(d+1)}.$$

Since $\|\cdot\|^2_{L^2(X,A,B)} = \mathbb{E}_{A,B}[\|\cdot\|^2_{L^2(X)}]$, this gives

$$\mathbb{E}_{A,B}\left[\|h(X) - h_K^{\theta,W}(X)\|^2_{L^2(X)}\right] \le C_q^2 K^{-1-2\beta/(d+1)}.$$

Applying Markov's inequality yields

$$\mathbb{P}_{A,B}\left(\|h(X) - h_K^{\theta,W}(X)\|^2_{L^2(X)} > C_q^2 K^{-1-2\beta/(d+1)} \varepsilon^{-1}\right) \le \varepsilon.$$

Taking a square root on both sides of the inequality inside the probability proves (3.5). □

Proof of Corollary 3.1. Fix $\varepsilon \in (0,1)$. By Theorem 3.2, there exists a $\sigma(A,B)$-measurable weight vector W such that

$$\mathbb{P}_{A,B}\left(\mathcal{E}_{K,\varepsilon}\right) \ge 1 - \varepsilon,$$

where

$$\mathcal{E}_{K,\varepsilon} := \left\{\|h - h_K^{\theta,W}\|_{L^2(X)} \le C_q K^{-1/2-\beta/(d+1)} \varepsilon^{-1/2}\right\},$$

and $C_q > 0$ is the constant of Theorem 3.1. On $\mathcal{E}_{K,\varepsilon}$,

$$\inf_{g \in \mathcal{H}_K} \|h - g\|_{L^2(X)} \le \|h - h_K^{\theta,W}\|_{L^2(X)},$$

and hence

$$\mathbb{P}_{A,B}\left(\inf_{g \in \mathcal{H}_K} \|h - g\|_{L^2(X)} > C_q K^{-1/2-\beta/(d+1)} \varepsilon^{-1/2}\right) \le \varepsilon,$$

which proves the O_P rate. □

Proof of Lemma 3.3. Let $C_{K,k}$ denote the Voronoi cell of θ_k with $V_{K,k} := \pi_\theta(C_{K,k})$. Notice that since $\{C_{K,k}\}_{k=1}^K$ forms a partition of Π, one has $V_{K,k} \ge 0$ and $\sum_{k=1}^K V_{K,k} = 1$. Hence

$$\sum_{k=1}^K V_{K,k}^2 \le \max_{1 \le k \le K} V_{K,k}. \tag{A.2}$$

By Assumption 3, it holds that

$$\|W\|_1 = \sum_{k=1}^K |\omega_{k,K}||f(A_k, B_k)| \le M_f \sum_{k=1}^K |\omega_{k,K}|.$$

The definition of control-neighbor weights (3.2) yields

$$\sum_{k=1}^{K}|\omega_{k,K}| = \frac{1}{K}\sum_{k=1}^{K}\left|1 + KV_{K,k} - \widehat{d}_{K,k}\right| \le 1 + \sum_{k=1}^{K} V_{K,k} + \frac{1}{K}\sum_{k=1}^{K}\widehat{d}_{K,k} \le 2 + \psi_\Pi,$$

where the last inequality follows from Assumption 6, which gives $\widehat{d}_{K,k} \le \psi_\Pi$ for all k. This proves the first bound in (3.6).

Then, again by Assumption 3,

$$\|W\|^2 = \sum_{k=1}^{K}\omega_{k,K}^2 f(A_k, B_k)^2 \le M_f^2 \sum_{k=1}^{K}\omega_{k,K}^2.$$

Using the weight definition (3.2) and $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, one obtains that

$$\begin{aligned}\sum_{k=1}^{K}\omega_{k,K}^2 &= \frac{1}{K^2}\sum_{k=1}^{K}(1 + KV_{K,k} - \widehat{d}_{K,k})^2 \le \frac{3}{K^2}\sum_{k=1}^{K}\left(1 + K^2V_{K,k}^2 + \widehat{d}_{K,k}^2\right)\\ &\le \frac{3}{K} + 3\sum_{k=1}^{K}V_{K,k}^2 + \frac{3}{K^2}\sum_{k=1}^{K}\widehat{d}_{K,k}^2. \end{aligned} \tag{A.3}$$

Since $\sum_{k=1}^{K}V_{K,k}^2 \le 1$ and $\widehat{d}_{K,k} \le \psi_\Pi$, it follows from (A.3) that

$$\sum_{k=1}^{K}\omega_{k,K}^2 \le \frac{3}{K} + 3 + 3\frac{\psi_\Pi^2}{K} \le 3(2 + \psi_\Pi^2),$$

for $K \ge 1$. Hence

$$\|W\|_2^2 \le 3M_f^2(2 + \psi_\Pi^2),$$

which proves the second bound in (3.6).

Let $\rho_K := \sup_{\theta\in\Pi}\min_{1\le k\le K}\rho_\Pi(\theta,\theta_k)$ be the covering radius of $\{\theta_1,\dots,\theta_K\}$ in (Π,ρ_Π). By Assumption 5 (iii), $B^{\|\cdot\|}_{r/M_\rho}(\theta) \subseteq B^{\rho_\Pi}_r(\theta) \subseteq B^{\|\cdot\|}_{r/m_\rho}(\theta)$, $\theta\in\Pi$, $r>0$. Since $C_{K,k} \subseteq B^{\rho_\Pi}_{\rho_K}(\theta_k)$ by definition of the covering radius, Assumption 5 gives

$$V_{K,k} = \pi_\theta(C_{K,k}) \le c_u\lambda_{d+1}(C_{K,k}) \le c_u V_{d+1} m_\rho^{-(d+1)}\rho_K^{d+1}. \tag{A.4}$$

By a standard covering argument, for every $\alpha>0$ there exist constants $\bar{C}_\alpha, \bar{C}'_\alpha > 0$ such that

$$\mathbb{P}_{A,B}\left(\rho_K^{d+1} \le \bar{C}_\alpha\frac{\log(K)}{K}\right) \ge 1 - \bar{C}'_\alpha K^{-\alpha}. \tag{A.5}$$

Indeed, by Assumption 5 (iii), an $r/(2M_\rho)$-net of Π under the Euclidean metric is an $r/2$-net under ρ_Π. Moreover, since D_Π is as in Assumption 3, $\sup_{\theta_1,\theta_2\in\Pi}\|\theta_1-\theta_2\| \le D_\Pi/m_\rho$. Hence, by a standard volumetric bound (see, for example, Corollary 4.2.11 in Vershynin (2026)), there exists an $r/2$-net $\mathcal{N}_{r/2}$ of Π under ρ_Π such that, for $r \le 2M_\rho r_0$,

$$|\mathcal{N}_{r/2}| \le \left(1 + \frac{4M_\rho D_\Pi}{m_\rho r}\right)^{d+1} = r^{-(d+1)}\left(r + \frac{4M_\rho D_\Pi}{m_\rho}\right)^{d+1} \le C_\Pi r^{-(d+1)},$$

where $C_\Pi := (2M_\rho r_0 + (4M_\rho D_\Pi)/m_\rho)^{d+1}$. If $\rho_K > r$, then there exists $z\in\mathcal{N}_{r/2}$ such that $B^{\rho_\Pi}_{r/2}(z)\cap\Pi$ contains no sample point.

For $r \le 2M_\rho r_0$, Assumption 5 (i)-(iii) gives

$$\pi_\theta\big(B^{\rho_\Pi}_{r/2}(z)\big) \ge \pi_\theta\big(B^{\|\cdot\|}_{r/(2M_\rho)}(z)\cap\Pi\big) \ge c_b c_\pi V_{d+1}\big(r/(2M_\rho)\big)^{d+1},$$

and hence

$$\mathbb{P}_{A,B}(\rho_K > r) \le C_\Pi r^{-(d+1)} \exp\left(-K c_b c_\pi V_{d+1} \big(r/(2M_\rho)\big)^{d+1}\right).$$

Set $r^{d+1} = \bar{C}_\alpha \log(K)/K$ and choose $\bar{C}_\alpha = (2M_\rho)^{d+1}(\alpha+1)/(c_b c_\pi V_{d+1})$. Since $r = \big(\bar{C}_\alpha \log(K)/K\big)^{1/(d+1)} \to 0$, there exists $K_0 \ge 3$, depending only on $\bar{C}_\alpha$, M_ρ, and r_0, such that $r \le 2M_\rho r_0$ for all $K \ge K_0$. Hence, for $K \ge K_0$, using $1/\log(K) \le 1$, $\mathbb{P}_{A,B}(\rho_K > r) \le (C_\Pi/\bar{C}_\alpha)K^{-\alpha}$. By choosing $\bar{C}'_\alpha \ge \max\{C_\Pi/\bar{C}_\alpha, K_0^\alpha\}$ the same bound with right-hand side $\bar{C}'_\alpha K^{-\alpha}$ holds also for finitely many values $3 \le K < K_0$, and this yields that (A.5) holds for all $K \ge 3$.

On the event (A.5), substituting into (A.4), yields

$$\max_{1\le k\le K} V_{K,k} \le c_u V_{d+1} m_\rho^{-(d+1)} \bar{C}_\alpha \frac{\log(K)}{K} =: D_\alpha \frac{\log(K)}{K}. \tag{A.6}$$

Again, we note that $\widehat{d}_{K,k} \le \psi_\Pi$ and hence

$$\sum_{k=1}^{K} \widehat{d}_{K,k}^2 \le K\psi_\Pi^2. \tag{A.7}$$

Thus, substituting (A.2), (A.6) and (A.7) into (A.3), and once more using $1/K \le \log(K)/K$ for $K \ge e$,

$$\sum_{k=1}^{K} \omega_{k,K}^2 \le \frac{3D_\alpha \log(K)}{K} + \frac{3\log(K)}{K} + \frac{3\psi_\Pi^2 \log(K)}{K} = \widetilde{C}_\alpha \frac{\log(K)}{K},$$

with $\widetilde{C}_\alpha := 3D_\alpha + 3 + 3\psi_\Pi^2$ on an event of probability at least $1 - \bar{C}'_\alpha K^{-\alpha}$. Consequently,

$$\|W\|^2 \le M_f^2 \widetilde{C}_\alpha \frac{\log(K)}{K},$$

which yields the result with $C_\alpha := \sqrt{\widetilde{C}_\alpha}$ and $C'_\alpha := \bar{C}'_\alpha$. □

A.2 L^2 approximation on integral Sobolev classes

Proof of Corollary 3.2. Apply Theorem 3.4 with $\delta = K^{-\alpha}$. Since $\log(K/\delta) = (1+\alpha)\log K$, the theorem gives the asserted coefficient bound and the corresponding $L^2(\mathcal{X})$ approximation bound with probability at least $1 - K^{-\alpha}$. The conclusion follows from

$$\|h - W^\top \phi^m_{\theta,K}\|_{L^2(X)} \le \sqrt{M}\, \|h - W^\top \phi^m_{\theta,K}\|_{L^2(\mathcal{X})}.$$

□

A.3 Sup-norm approximation via stratified sampling

Proof of Theorem 3.5. **Proof of case (i).**

For $K \ge 2$, set $\widetilde{K} := \lfloor K/2 \rfloor$. We first carry out the stratified construction with budget $\widetilde{K}$. The proof proceeds in four steps. First, we construct a stratified sampling scheme on Π with allocations proportional to the cell masses, partitioning it into J cells and allocating neurons approximately proportionally to the cell masses $\pi_\theta(B_j)$ while ensuring that each cell receives at least one sample. Second, we express the error as a centered empirical process indexed by $x \in \mathcal{X}$ and symmetrize it. Third, we bound the expected supremum of the symmetrized process using Dudley's entropy integral. Finally, we choose the partition scale so that the resulting number $\overline{K}$ of active neurons satisfies $\overline{K} \le 2\widetilde{K} \le K$, augment the construction with $K - \overline{K}$ additional inner parameters, and assign to the remaining neurons zero output weights. We then obtain the rate $K^{-1/2-\beta/(d+1)}$ up to logarithmic factors.

STEP 1. Fix $\epsilon > 0$. For convenience, in this proof write $\phi(\theta)(x) := \phi_x(\theta) = f(A,B)\sigma(A^\top x + B)$ for $\theta = (A^\top, B)^\top \in \Pi$ and $x \in \mathcal{X}$.

Let $\mathcal{U} := \{A^\top x + B : x \in \mathcal{X},\ (A,B) \in \Pi\}$, which is compact as the continuous image of $\mathcal{X} \times \Pi$, and recall $M_\sigma = \sup_{u\in\mathcal{U}} |\sigma(u)| < \infty$ from Assumption 3. Since σ is Lipschitz with constant L_σ and $\|A\|_1 \le C_A$, for fixed $\theta \in \Pi$ and $x, \widetilde{x} \in \mathcal{X}$,

$$|\sigma(A^\top x + B) - \sigma(A^\top \widetilde{x} + B)| \le L_\sigma \|A\|_1 \|x - \widetilde{x}\|_\infty \le L_\sigma C_A \|x - \widetilde{x}\|_\infty,$$

and for $\theta, \widetilde{\theta} \in \Pi$ and $x \in \mathcal{X}$,

$$|\sigma(A^\top x + B) - \sigma(\widetilde{A}^\top x + \widetilde{B})| \le L_\sigma \big|(A - \widetilde{A})^\top x + (B - \widetilde{B})\big| \le L_\sigma d_\Pi(\theta, \widetilde{\theta}), \tag{A.8}$$

by Assumption 2 and the definition of d_Π, using $\|x\|_\infty \le M$.

For $\theta = (A^\top, B)^\top, \widetilde{\theta} = (\widetilde{A}^\top, \widetilde{B})^\top \in \Pi$ and $x \in \mathcal{X}$, it holds that

$$\begin{aligned}
|\phi(\theta)(x) - \phi(\widetilde{\theta})(x)| &\le |f(A,B)||\sigma(A^\top x + B) - \sigma(\widetilde{A}^\top x + \widetilde{B})| \\
&\quad + |\sigma(\widetilde{A}^\top x + \widetilde{B})||f(A,B) - f(\widetilde{A}, \widetilde{B})| \\
&\le M_f L_\sigma d_\Pi(\theta, \widetilde{\theta}) + M_\sigma L_f d_\Pi(\theta, \widetilde{\theta})^\beta \\
&\le M_f L_\sigma (2MC_A + 2C_B)^{1-\beta} d_\Pi(\theta, \widetilde{\theta})^\beta + M_\sigma L_f d_\Pi(\theta, \widetilde{\theta})^\beta,
\end{aligned}$$

where in the second inequality we used $|f(A,B)| \le M_f$ and(A.8) for the first term, and $|\sigma(\widetilde{A}^\top x + \widetilde{B})| \le M_\sigma$ together with Assumption 3 for the second one; in the last inequality we used $d_\Pi(\theta, \widetilde{\theta}) \le 2MC_A + 2C_B$ for $\theta, \widetilde{\theta} \in \Pi$.

Set $L_u := M_f L_\sigma (2MC_A + 2C_B)^{1-\beta} + M_\sigma L_f$. Then

$$\sup_{x\in\mathcal{X}} |\phi(\theta)(x) - \phi(\widetilde{\theta})(x)| \le L_u d_\Pi(\theta, \widetilde{\theta})^\beta. \tag{A.9}$$

Since $\Pi = \{A \in \mathbb{R}^d : \|A\|_1 \le C_A\} \times [-C_B, C_B]$ and $d_\Pi(\theta, \widetilde{\theta}) = M\|A - \widetilde{A}\|_1 + |B - \widetilde{B}|$ is additive across the two components, the covering number factorizes. Covering the ℓ_1-ball $\{A : \|A\|_1 \le C_A\}$ under the metric $M\|\cdot\|_1$ with radius $r/2$ is equivalent to covering it under $\|\cdot\|_1$ with radius $r/(2M)$, which yields at most $(1 + 4MC_A/r)^d$ by a standard volumetric argument (see, for example, Corollary 4.2.11 in Vershynin (2026)). Similarly, covering $[-C_B, C_B]$ under $|\cdot|$ with radius $r/2$ yields at most $(1 + 4C_B/r)$ points. Taking the product covering gives

$$\mathcal{N}(r, \Pi, d_\Pi) \le \left(1 + \frac{4MC_A}{r}\right)^d \left(1 + \frac{4C_B}{r}\right), \qquad r > 0. \tag{A.10}$$

Let $r_\epsilon := (\epsilon/L_u)^{1/\beta}$. Let $\{\bar{\theta}_j\}_{j=1}^J$ be an r_ϵ-net of Π under d_Π of minimal cardinality, so that $J = \mathcal{N}(r_\epsilon, \Pi, d_\Pi)$. Let $\{B_j\}_{j=1}^J$ be a measurable partition of Π such that $B_j \subset \{\theta \in \Pi : d_\Pi(\theta, \bar{\theta}_j) \le r_\epsilon\}$ for each $j = 1, \dots, J$. Such a partition exists since $\{\bar{\theta}_j\}$ is an r_ϵ-net. Therefore, by (A.9),

$$\sup_{\theta\in B_j} \sup_{x\in\mathcal{X}} |\phi(\theta)(x) - \phi(\bar{\theta}_j)(x)| \le \epsilon. \tag{A.11}$$

If $r_\epsilon \le \min(2MC_A, 2C_B)$, then

$$1 + \frac{4MC_A}{r_\epsilon} \le \frac{6MC_A}{r_\epsilon}, \qquad 1 + \frac{4C_B}{r_\epsilon} \le \frac{6C_B}{r_\epsilon}.$$

Then, by (A.10),

$$J = \mathcal{N}(r_\epsilon, \Pi, d_\Pi) \le \left(1 + \frac{4MC_A}{r_\epsilon}\right)^d \left(1 + \frac{4C_B}{r_\epsilon}\right).$$

Combining these bounds yields

$$J \le (6MC_A)^d (6C_B) r_\epsilon^{-(d+1)} = C_\Pi r_\epsilon^{-(d+1)}, \text{ with } C_\Pi := 6C_B(6MC_A)^d. \tag{A.12}$$

For each $j = 1, \dots, J$, define $P_j := \pi_\theta(B_j)$. If $P_j > 0$, define

$$\pi_{\theta,j}(E) := \frac{\pi_\theta(E \cap B_j)}{P_j}, \qquad E \subset \Pi \text{ measurable},$$

and if $P_j = 0$, let $\pi_{\theta,j}$ be any probability measure on B_j. Let $\{K_j\}_{j=1}^J$ be independent random variables satisfying $K_j \in \{\lfloor \widetilde{K} P_j \rfloor, \lceil \widetilde{K} P_j \rceil\}$, $\mathbb{E}[K_j] = \widetilde{K} P_j$, and define

$$L_j := K_j + \mathbb{1}_{\{K_j = 0\}}, \qquad \underline{K} := (K_1, \dots, K_J)^\top.$$

Conditional on $\underline{K}$, draw $\theta_{j,1}, \dots, \theta_{j,L_j} \overset{\text{i.i.d.}}{\sim} \pi_{\theta,j}$, independently across j. Define

$$\overline{K} := \sum_{j=1}^J L_j.$$

Then

$$\begin{aligned}
\overline{K} &= \sum_{j=1}^J K_j \mathbb{1}_{\{K_j > 0\}} + \sum_{j=1}^J \mathbb{1}_{\{K_j = 0\}} \le \sum_{j=1}^J (\widetilde{K} P_j + 1) \mathbb{1}_{\{K_j > 0\}} + \sum_{j=1}^J \mathbb{1}_{\{K_j = 0\}} \\
&= \widetilde{K} \sum_{j=1}^J P_j \mathbb{1}_{\{K_j > 0\}} + J \le \widetilde{K} + J,
\end{aligned} \tag{A.13}$$

where in the first inequality we used that $K_j \le \lceil \widetilde{K} P_j \rceil \le \widetilde{K} P_j + 1$.

Recall the stratified approximant $h_{\overline{K}}^{\text{IS}}$ defined in (3.7), where $\phi(\theta)(x) = f(\theta)\widetilde{\sigma}(A^\top x + B)$. Since $\theta_{j,\ell}$ are i.i.d. from $\pi_{\theta,j}$,

$$\mathbb{E}[h_{\overline{K}}^{\text{IS}}(x)] = \sum_{j=1}^J P_j \mathbb{E}_{\pi_{\theta,j}}[\phi(\theta)(x)] = h(x). \tag{A.14}$$

Step 2. Define the centered summands

$$\widetilde{\phi}_{j,\ell}(x) := \phi(\theta_{j,\ell})(x) - \mathbb{E}_{\pi_{\theta,j}}[\phi(\theta)(x)].$$

Since $\theta_{j,\ell} \sim \pi_{\theta,j}$, $\mathbb{E}_{\pi_{\theta,j}}[\widetilde{\phi}_{j,\ell}(x)] = 0$.

Let $\mu_j(x) := \phi(\theta_j)(x)$. Since $\theta_{j,\ell} \in B_j$ $\pi_{\theta,j}$-almost surely, (A.11) yields

$$|\phi(\theta_{j,\ell})(x) - \mu_j(x)| \le \epsilon, \qquad x \in \mathcal{X}.$$

Taking expectations and using Jensen's inequality,

$$\left|\mathbb{E}_{\pi_{\theta,j}}[\phi(\theta)(x)] - \mu_j(x)\right| = \left|\mathbb{E}_{\pi_{\theta,j}}[\phi(\theta)(x) - \mu_j(x)]\right| \le \mathbb{E}_{\pi_{\theta,j}}\left[|\phi(\theta)(x) - \mu_j(x)|\right] \le \epsilon.$$

Combining these bounds gives

$$\sup_{x \in \mathcal{X}} |\widetilde{\phi}_{j,\ell}(x)| \le \sup_{x \in \mathcal{X}} |\phi(\theta_{j,\ell})(x) - \mu_j(x)| + \sup_{x \in \mathcal{X}} \left|\mu_j(x) - \mathbb{E}_{\pi_{\theta,j}}[\phi(\theta)(x)]\right| \le 2\epsilon. \tag{A.15}$$

Moreover, by (3.7) and (A.14)

$$h_{\overline{K}}^{\text{IS}}(x) - h(x) = \sum_{j=1}^J \frac{P_j}{L_j} \sum_{\ell=1}^{L_j} \left(\phi(\theta_{j,\ell})(x) - \mathbb{E}_{\pi_{\theta,j}}[\phi(\theta)(x)]\right) = \sum_{j=1}^J \frac{P_j}{L_j} \sum_{\ell=1}^{L_j} \widetilde{\phi}_{j,\ell}(x). \tag{A.16}$$

Conditional on $\underline{K}$, the family $\{\widetilde{\phi}_{j,\ell}\}_{j,\ell}$ is independent and centered. By the Rademacher symmetrization inequality (van der Vaart and Wellner, 2023, Lemma 2.3.6),

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}|h^{\mathrm{IS}}_{\underline{K}}(x)-h(x)|\;\middle|\;\underline{K}\right]\le 2\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\sum_{j,\ell}\varepsilon_{j,\ell}\frac{P_j}{L_j}\widetilde{\phi}_{j,\ell}(x)\right|\;\middle|\;\underline{K}\right], \tag{A.17}$$

where $\{\varepsilon_{j,\ell}\}$ are i.i.d. Rademacher variables independent of $\underline{\theta}$ and $\underline{K}$.

Step 3. For $x\in\mathcal{X}$, define the Rademacher process

$$Z(x):=\sum_{j=1}^{J}\frac{P_j}{L_j}\sum_{\ell=1}^{L_j}\varepsilon_{j,\ell}\widetilde{\phi}_{j,\ell}(x),$$

and write $\underline{\theta}:=(\theta_{j,\ell})_{j,\ell}$. Conditional on $(\underline{\theta},\underline{K})$, the variables $\varepsilon_{j,\ell}$ are i.i.d. Rademacher and the coefficients $\widetilde{\phi}_{j,\ell}(x)$ are deterministic. Define the pseudometric

$$\rho_{\underline{\theta},\underline{K}}(x,\widetilde{x})^2:=\sum_{j=1}^{J}\frac{P_j^2}{L_j^2}\sum_{\ell=1}^{L_j}\left[\widetilde{\phi}_{j,\ell}(x)-\widetilde{\phi}_{j,\ell}(\widetilde{x})\right]^2,\qquad x,\widetilde{x}\in\mathcal{X}. \tag{A.18}$$

By Hoeffding's lemma, Z is sub-Gaussian with respect to $\rho_{\underline{\theta},\underline{K}}$. Fix $x_0\in\mathcal{X}$. Setting

$$A^2_{\underline{K}}:=\sum_{j=1}^{J}\frac{P_j^2}{L_j},$$

by Cauchy-Schwarz and (A.15), $\mathbb{E}_\varepsilon\left[|Z(x_0)|\mid\underline{\theta},\underline{K}\right]\le 2\epsilon A_{\underline{K}}$. Hence Dudley's entropy integral (van der Vaart and Wellner, 2023, Corollary 2.2.9) applied to $Z(x)-Z(x_0)$ yields, for each fixed realization of $(\underline{\theta},\underline{K})$,

$$\mathbb{E}_\varepsilon\left[\sup_{x\in\mathcal{X}}|Z(x)|\;\middle|\;\underline{\theta},\underline{K}\right]\le 2\epsilon A_{\underline{K}}+C_1\int_0^{\Delta_{\underline{\theta},\underline{K}}}\sqrt{\log\mathcal{N}(s,\mathcal{X},\rho_{\underline{\theta},\underline{K}})}\mathrm{d}s, \tag{A.19}$$

where $C_1>0$ is a universal constant and $\Delta_{\underline{\theta},\underline{K}}:=\sup_{x,\widetilde{x}\in\mathcal{X}}\rho_{\underline{\theta},\underline{K}}(x,\widetilde{x})$.

Since $|\widetilde{\phi}_{j,\ell}(x)|\le 2\epsilon$ by (A.15), the triangle inequality gives $|\widetilde{\phi}_{j,\ell}(x)-\widetilde{\phi}_{j,\ell}(\widetilde{x})|\le 4\epsilon$, and hence $\rho_{\underline{\theta},\underline{K}}(x,\widetilde{x})^2\le 16\epsilon^2A^2_{\underline{K}}$, so that

$$\Delta_{\underline{\theta},\underline{K}}\le 4\epsilon A_{\underline{K}}. \tag{A.20}$$

For fixed $\theta=(A^\top,B)^\top\in\Pi$ and $x,\widetilde{x}\in\mathcal{X}$,

$$\begin{aligned}|\phi(\theta)(x)-\phi(\theta)(\widetilde{x})|&=|f(A,B)||\sigma(A^\top x+B)-\sigma(A^\top\widetilde{x}+B)|\le M_fL_\sigma|A^\top(x-\widetilde{x})|\\&\le M_fL_\sigma\|A\|_1\|x-\widetilde{x}\|_\infty\le M_fL_\sigma C_A\|x-\widetilde{x}\|_\infty.\end{aligned}$$

Hence, by the triangle inequality and Jensen's inequality applied to $\mathbb{E}_{\pi_{\theta,j}}$,

$$\begin{aligned}|\widetilde{\phi}_{j,\ell}(x)-\widetilde{\phi}_{j,\ell}(\widetilde{x})|&\le|\phi(\theta_{j,\ell})(x)-\phi(\theta_{j,\ell})(\widetilde{x})|+\left|\mathbb{E}_{\pi_{\theta,j}}\left[\phi(\theta)(x)-\phi(\theta)(\widetilde{x})\right]\right|\\&\le 2M_fL_\sigma C_A\|x-\widetilde{x}\|_\infty.\end{aligned}$$

Substituting into the definition of $\rho_{\underline{\theta},\underline{K}}$ in (A.18) yields

$$\rho_{\underline{\theta},\underline{K}}(x,\widetilde{x})\le 2M_fL_\sigma C_AA_{\underline{K}}\|x-\widetilde{x}\|_\infty.$$

Therefore, since $\mathcal{X}=[-M,M]^d$, a standard volumetric argument (see, for example, Corollary 4.2.11 in Vershynin (2026)) gives

$$\mathcal{N}(s,\mathcal{X},\rho_{\underline{\theta},\underline{K}})\le\left(1+\frac{4MM_fL_\sigma C_AA_{\underline{K}}}{s}\right)^d. \tag{A.21}$$

Set $C_2 := MM_f L_\sigma C_A$. Next, combining (A.19), (A.20), and (A.21), and using the nonnegativity of the integrand, we obtain

$$\begin{aligned}\mathbb{E}_\varepsilon\left[\sup_{x\in\mathcal{X}}|Z(x)| \;\middle|\; \underline{\theta},\underline{K}\right] &\le 2\epsilon A_{\underline{K}} + C_1\int_0^{4\epsilon A_{\underline{K}}}\sqrt{d\log\left(1+\frac{4C_2A_{\underline{K}}}{s}\right)}\mathrm{d}s\\ &\le 2\epsilon A_{\underline{K}} + 4C_1\epsilon A_{\underline{K}}\sqrt{d}\int_0^1\sqrt{\log\left(1+\frac{C_2}{\epsilon v}\right)}\mathrm{d}v\\ &\le 2\epsilon A_{\underline{K}} + 4C_1\epsilon A_{\underline{K}}\sqrt{d}\int_0^1\left[\sqrt{\log\left(1+\frac{C_2}{\epsilon}\right)}+\sqrt{\log\left(\frac{1}{v}\right)}\right]\mathrm{d}v. \quad \text{(A.22)}\end{aligned}$$

Here, in the second line we applied the change of variables $v := s/(4\epsilon A_{\underline{K}})$ so that $s = 4\epsilon A_{\underline{K}}v$ and $\mathrm{d}s = 4\epsilon A_{\underline{K}}\mathrm{d}v$. In the third line we used that setting $a := C_2/\epsilon$, for every $v \in (0,1]$ it holds that $\log(1+a/v) \le \log(1+a)+\log(1/v)$ and that $\sqrt{u+w} \le \sqrt{u}+\sqrt{w}$ for $u, w \ge 0$. We proceed by considering the second summand in the integrand of the last expression. More explicitly, using the change of variables $u := \log(1/v)$, so that $v = e^{-u}$ and $\mathrm{d}v = -e^{-u}\mathrm{d}u$, we get that

$$\int_0^1\sqrt{\log(1/v)}\mathrm{d}v = \int_0^\infty u^{1/2}e^{-u}\mathrm{d}u = \Gamma(3/2) = \sqrt{\pi}/2 < 1.$$

We apply this inequality to (A.22), noting that $\sqrt{\log(1+C_2/\epsilon)}+1 \ge 1$ for $\sqrt{d} \ge 1$, yielding

$$\mathbb{E}_\varepsilon\left[\sup_{x\in\mathcal{X}}|Z(x)| \;\middle|\; \underline{\theta},\underline{K}\right] \le (4C_1+2)\epsilon A_{\underline{K}}\sqrt{d}\left(\sqrt{\log\left(1+\frac{C_2}{\epsilon}\right)}+1\right) =: F(\underline{K}), \quad \text{(A.23)}$$

where the right-hand side depends on $(\underline{\theta},\underline{K})$ only through $\underline{K}$.

If $\widetilde{K}P_j \le 2$, then $P_j^2/L_j \le 2P_j/\widetilde{K}$. If $\widetilde{K}P_j > 2$, then $L_j \ge K_j \ge \widetilde{K}P_j/2$, so again $P_j^2/L_j \le 2P_j/\widetilde{K}$. Summing over j and using $\sum_j P_j = 1$ gives

$$A_{\underline{K}}^2 \le \frac{2}{\widetilde{K}}. \quad \text{(A.24)}$$

Taking conditional expectation with respect to $\underline{\theta}$ given $\underline{K}$ in (A.23), we obtain

$$\mathbb{E}\left[\mathbb{E}_\varepsilon\left[\sup_{x\in\mathcal{X}}|Z(x)| \;\middle|\; \underline{\theta},\underline{K}\right] \;\middle|\; \underline{K}\right] \le F(\underline{K}).$$

Substituting (A.24) into $F(\underline{K})$, combining with (A.17), and taking expectation over $\underline{K}$ yields

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}|h_{\underline{K}}^{\mathrm{IS}}(x)-h(x)|\right] \le C_3\frac{\epsilon}{\sqrt{\widetilde{K}}}\sqrt{d}\left(\sqrt{\log\left(1+\frac{C_2}{\epsilon}\right)}+1\right), \quad \text{(A.25)}$$

where $C_3 := \sqrt{8}(4C_1+2)$ and is a universal constant.

Step 4. Set

$$\epsilon := L_u r_\epsilon^\beta = c_0^\beta L_u \widetilde{K}^{-\beta/(d+1)}.$$

and $r_\epsilon := c_0\widetilde{K}^{-1/(d+1)}$, where $c_0 := C_\Pi^{1/(d+1)}$. Since r_ϵ tends to zero as $\widetilde{K} \to \infty$, there exists $K_0 \in \mathbb{N}^+$, depending only on d, C_A, C_B, and M, such that $r_\epsilon \le \min\{2MC_A, 2C_B\}$ for all $\widetilde{K} \ge K_0$. Hence, for $\widetilde{K} \ge K_0$, (A.12) applies and gives $J \le C_\Pi r_\epsilon^{-(d+1)} = C_\Pi c_0^{-(d+1)}\widetilde{K} = \widetilde{K}$. Then (A.13) implies $\overline{K} \le 2\widetilde{K} \le K$. Augment the $\overline{K}$ active inner parameters by $K-\overline{K}$ additional inner parameters in Π and assign zero output weights to the corresponding neurons. Denoting the resulting network by $h_K^{\mathrm{IS}} \in \mathcal{H}_K^{\mathrm{IS}}$, we have $h_K^{\mathrm{IS}}(x) = h_{\overline{K}}^{\mathrm{IS}}(x)$, $x \in \mathcal{X}$.

Substituting this choice into (A.25) and using $h_K^{\mathrm{IS}} = h_{\widetilde{K}}^{\mathrm{IS}}$, we obtain

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}|h_K^{\mathrm{IS}}(x)-h(x)|\right] \le C_4\widetilde{K}^{-1/2-\beta/(d+1)}\sqrt{d}\left(\sqrt{\log\left(1+C_5\widetilde{K}^{\beta/(d+1)}\right)}+1\right),$$

where $C_4 := C_3c_0^{\beta}L_u$ and $C_5 := C_2/(c_0^{\beta}L_u)$. Since $\log(1+C_5\widetilde{K}^{\beta/(d+1)}) \le \log(1+C_5)+\frac{\beta}{d+1}\log\widetilde{K}$, it holds that

$$\begin{aligned}\sqrt{d}\left(\sqrt{\log\left(1+C_5\widetilde{K}^{\beta/(d+1)}\right)}+1\right) &\le \left(1+\sqrt{\log(1+C_5)}\right)\sqrt{d}+\sqrt{\frac{d\beta}{d+1}}\sqrt{\log(\widetilde{K})}\\ &\le \left(1+\sqrt{\log(1+C_5)}\right)\left(\sqrt{d}+\sqrt{\log(\widetilde{K})}\right),\end{aligned}$$

where the last inequality uses $\beta \le 1$. Hence for $K \ge K_* := \max\{2K_0, 2\}$, since $\widetilde{K} = \lfloor K/2\rfloor \ge K/3$ for every $K \ge 2$ and $\log(\widetilde{K}) \le \log(K)$,

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}|h_K^{\mathrm{IS}}(x)-h(x)|\right] \le C_{\mathrm{IS}}^{(0)}\left(\sqrt{d}+\sqrt{\log(K)}\right)K^{-1/2-\beta/(d+1)},$$

where $C_{\mathrm{IS}}^{(0)} := 3^{1/2+\beta/(d+1)}C_4\left(1+\sqrt{\log(1+C_5)}\right)$.

It remains to consider the finitely many $1 \le K < K_*$. For these values of K, take the trivial partition $J = 1$ with $B_1 = \Pi$, choose any $\bar{\theta}_1 \in \Pi$, and set $r_\epsilon := D_\Pi$, which is finite by Assumption 3. Since $\rho_\Pi = d_\Pi$ in case (i), the definition of D_Π gives $B_1 = \Pi \subseteq \left\{\theta \in \Pi \mid d_\Pi(\theta, \bar{\theta}_1) \le r_\epsilon\right\}$, so this is an admissible one-cell instance of the preceding construction, and taking $\epsilon := L_u D_\Pi^{\beta}$ ensures that (A.11) holds. Moreover, $P_1 = \pi_\theta(B_1) = 1$, and hence $K_1 \in \{\lfloor KP_1\rfloor, \lceil KP_1\rceil\} = \{K\}$, so that $K_1 = K$. Since $K \ge 1$, the definition $L_1 = K_1 + \mathbb{1}_{\{K_1=0\}}$ gives $L_1 = K$, and consequently $\overline{K} = L_1 = K \le 2K$. The corresponding estimator is $h_K^{\mathrm{IS}}(x) = 1/K\sum_{\ell=1}^{K}\phi(\theta_{1,\ell})(x)$, which is the one-cell instance of the same stratified sampling construction. In particular, its unbiasedness follows from (A.14). Next, by the integral representation of h and the bound $|\phi(\theta)(x)| \le M_f M_\sigma$,

$$|h(x)| \le \int_\Pi |\phi(\theta)(x)|\mathrm{d}\pi_\theta(\theta) \le M_f M_\sigma \quad \text{and} \quad \left|h_K^{\mathrm{IS}}(x)\right| \le \frac{1}{K}\sum_{\ell=1}^{K}|\phi(\theta_{1,\ell})(x)| \le M_f M_\sigma,$$

so that for $1 \le K < K_*$, we have $\mathbb{E}\left[\sup_{x\in\mathcal{X}}|h_K^{\mathrm{IS}}(x)-h(x)|\right] \le 2M_fM_\sigma$.

Since $K^{1/2+\beta/(d+1)} \le K_*^{1/2+\beta/(d+1)}$ and $\sqrt{d}+\sqrt{\log K} \ge 1$, it follows that (3.8) holds for $1 \le K < K_*$ with $C_{\mathrm{IS}}^{(1)} := 2M_fM_\sigma K_*^{1/2+\beta/(d+1)}$. Consequently, (3.8) holds for all $K \in \mathbb{N}^+$ with $C_{\mathrm{IS}} := \max\left\{C_{\mathrm{IS}}^{(0)}, C_{\mathrm{IS}}^{(1)}\right\}$. Equivalently,

$$C_{\mathrm{IS}} := \max\left\{3^{1/2+\beta/(d+1)}C_3c_0^{\beta}L_u\left(1+\sqrt{\log\left(1+\frac{C_2}{c_0^{\beta}L_u}\right)}\right), 2M_fM_\sigma K_*^{1/2+\beta/(d+1)}\right\}, \quad \text{(A.26)}$$

where we recall that $C_3 > 0$ is a universal constant, $C_\Pi = 6C_B(6MC_A)^d$, $c_0 = C_\Pi^{1/(d+1)}$, $C_2 = MM_fL_\sigma C_A$, $L_u = M_fL_\sigma(2MC_A+2C_B)^{1-\beta}+M_\sigma L_f$,

$$K_0 = \left\lceil (c_0/\min\{2MC_A, 2C_B\})^{d+1}\right\rceil,$$

and $K_* = \max\{2K_0, 2\}$.

For $K \ge K_*$, the active stratified construction uses $\overline{K} \le K$ neurons and the zero-appending described above yields an element of $\mathcal{H}_K^{\mathrm{IS}}$ without changing the approximating function. For

$1 \le K < K_*$, the one-cell construction already uses exactly K neurons. This completes the proof.

Proof of case (ii).

For $K \ge 2$, set $\widetilde{K} := \lfloor K/2 \rfloor$, as in the proof of case (i), and carry out the stratified construction with budget $\widetilde{K}$. The proof follows that of case (i), with the main modification in Step 1, where the Hölder continuity of f is replaced by its sign structure. Steps 2–4 are analogous, with the corresponding changes in the constants.

Step 1. Define $\phi(\theta)(x) := \phi_x(\theta) = f(A,B)\sigma_m(A^\top x + B)$ for $\theta = (A^\top, B)^\top \in \Pi$ and, as in case (i), equip Π with the metric d_Π defined in Remark 2(ii).

For $\theta, \widetilde{\theta} \in \Pi$ and $x \in \mathcal{X}$, set $s := A^\top x + B$ and $\widetilde{s} := \widetilde{A}^\top x + \widetilde{B}$, so that $|s|, |\widetilde{s}| \le C_A M + C_B =: R$. For $m = 1$, non-expansiveness of the ReLU activation function gives $|\sigma_1(s) - \sigma_1(\widetilde{s})| \le |s - \widetilde{s}|$. For $m = 2$, the mean value theorem applied to $\sigma_2(t) = \max(t,0)^2$, which has derivative $2\max(t,0)$, gives $|\sigma_2(s) - \sigma_2(\widetilde{s})| \le 2R|s - \widetilde{s}|$. In both cases,

$$|\sigma_m(s) - \sigma_m(\widetilde{s})| \le mR^{m-1}|s - \widetilde{s}| \le L_{\sigma,m} d_\Pi(\theta, \widetilde{\theta}), \tag{A.27}$$

where $L_{\sigma,m} := m(C_A M + C_B)^{m-1}$ and the second inequality uses

$$|s - \widetilde{s}| = |(A - \widetilde{A})^\top x + (B - \widetilde{B})| \le M\|A - \widetilde{A}\|_1 + |B - \widetilde{B}| = d_\Pi(\theta, \widetilde{\theta}),$$

by Hölder's inequality and $\|x\|_\infty \le M$ by Assumption 1.

Since $f : \Pi \to \{-M_f, +M_f\}$, partition Π into $\Pi_\pm := \{(A,B) \in \Pi : f(A,B) = \pm M_f\}$, so that f is constant on each part.

Set $L_u^* := M_f L_{\sigma,m}$. Fix $\epsilon > 0$ and $r_\epsilon := \epsilon / L_u^*$. Cover Π_+ and Π_- separately with r_ϵ-nets under d_Π, and choose measurable partitions subordinate to the corresponding d_Π-balls, yielding a measurable partition $\{B_j\}_{j=1}^J$ of Π with $B_j \subset \Pi_\pm$ and representatives $\bar{\theta}_j \in B_j$ satisfying $\sup_{\theta \in B_j} d_\Pi(\theta, \bar{\theta}_j) \le r_\epsilon$. Since f is constant on each B_j, for $\theta \in B_j$,

$$\begin{aligned} |\phi(\theta)(x) - \phi(\bar{\theta}_j)(x)| &= M_f |\sigma_m(A^\top x + B) - \sigma_m(\bar{A}_j^\top x + \bar{B}_j)| \\ &\le M_f L_{\sigma,m} d_\Pi(\theta, \bar{\theta}_j) \le L_u^* r_\epsilon = \epsilon, \end{aligned}$$

using (A.27). Hence

$$\sup_{\theta \in B_j} \sup_{x \in \mathcal{X}} |\phi(\theta)(x) - \phi(\bar{\theta}_j)(x)| \le \epsilon. \tag{A.28}$$

Covering Π_+ and Π_- separately by a standard volumetric argument (see, for example, Corollary 4.2.11 in Vershynin (2026)) gives, for $r_\epsilon \le \min(2MC_A, 2C_B)$, the following relation

$$J \le 2\left(1 + \frac{2(MC_A + C_B)}{r_\epsilon}\right)^{d+1} \le C_\Pi^* r_\epsilon^{-(d+1)}, \quad \text{with } C_\Pi^* := 2\big(4(MC_A + C_B)\big)^{d+1}. \tag{A.29}$$

The remainder of Step 1 follows precisely the proof of case (i), with $\widetilde{K}$ as the sampling budget, in terms of defining P_j, $\pi_{\theta,j}$, K_j, L_j, $\underline{K}$, the conditional draws, the estimator $h_{\overline{K}}^{\mathrm{IS}}$ in (3.7), and its unbiasedness in (A.14). In particular, (A.13) gives $\overline{K} \le \widetilde{K} + J$.

Step 2. Given (A.28) in place of (A.11), this step follows as Step 2 of case (i), yielding $\sup_x |\widetilde{\phi}_{j,\ell}(x)| \le 2\epsilon$, the error representation (A.16), and the symmetrization inequality (A.17).

Step 3. The Rademacher process $Z(x)$, pseudometric $\rho_{\underline{\theta},\underline{K}}$, sub-Gaussianity argument, Dudley bound (A.19), Δ bound (A.20), and tower property argument follow as in Step 3 of case (i). The only difference is the Lipschitz constant of $x \mapsto \phi_x(\theta)$. For fixed $\theta \in \Pi$ and $x, \widetilde{x} \in \mathcal{X}$, we write

$$\begin{aligned} |\phi(\theta)(x) - \phi(\theta)(\widetilde{x})| &= M_f |\sigma_m(A^\top x + B) - \sigma_m(A^\top \widetilde{x} + B)| \le M_f L_{\sigma,m} |A^\top (x - \widetilde{x})| \\ &\le M_f L_{\sigma,m} \|A\|_1 \|x - \widetilde{x}\|_\infty \le L_u^* C_A \|x - \widetilde{x}\|_\infty. \end{aligned}$$

Setting $C_2^* := ML_u^* C_A$ and applying the same integral evaluation as in Step 3 of case (i) gives

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}} |h_{\overline{K}}^{\mathrm{IS}}(x) - h(x)|\right] \le C_3 \frac{\epsilon}{\sqrt{\widetilde{K}}} \sqrt{d} \left(\sqrt{\log\left(1 + \frac{C_2^*}{\epsilon}\right)} + 1\right), \tag{A.30}$$

where $C_3 := \sqrt{8}(4C_1 + 2) > 0$ is the universal constant as in case (i).

Step 4. Set $\epsilon := c_0^* L_u^* \widetilde{K}^{-1/(d+1)}$ and $r_\epsilon := c_0^* \widetilde{K}^{-1/(d+1)}$, with $c_0^* := (C_\Pi^*)^{1/(d+1)}$. The condition $r_\epsilon \le \min(2MC_A, 2C_B)$ holds for all $\widetilde{K} \ge K_0$, with

$$K_0 := \left\lceil \left(\frac{c_0^*}{\min\{2MC_A, 2C_B\}} \right)^{d+1} \right\rceil.$$

For such $\widetilde{K}$, (A.29) applies and gives

$$J \le C_\Pi^* r_\epsilon^{-(d+1)} = C_\Pi^* (c_0^*)^{-(d+1)} \widetilde{K} = \widetilde{K}.$$

Hence, by (A.13), $\overline{K} \le \widetilde{K} + J \le 2\widetilde{K} \le K$. Augment the $\overline{K}$ active inner parameters by $K - \overline{K}$ additional inner parameters in Π and assign zero output weights to the corresponding neurons. Denoting the resulting network by $h_K^{\mathrm{IS}} \in \mathcal{H}_K^{\mathrm{IS}}$, we have $h_K^{\mathrm{IS}}(x) = h_{\overline{K}}^{\mathrm{IS}}(x)$, $x \in \mathcal{X}$. Substituting into (A.30) yields

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}} |h_K^{\mathrm{IS}}(x) - h(x)|\right] \le C_4^* \widetilde{K}^{-1/2-1/(d+1)} \sqrt{d} \left(\sqrt{\log\left(1 + C_5^* \widetilde{K}^{1/(d+1)}\right)} + 1\right),$$

where $C_4^* := C_3 c_0^* L_u^*$ and $C_5^* := C_2^*/(c_0^* L_u^*) = MC_A/c_0^*$.

Following the same argument as in Step 4 of case (i), for $K \ge K_* := \max\{2K_0, 2\}$ using $\widetilde{K} = \lfloor K/2 \rfloor \ge K/3$ and $\log(\widetilde{K}) \le \log K$, we obtain

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}} |h_K^{\mathrm{IS}}(x) - h(x)|\right] \le C_{\mathrm{IS}}^{*,(0)} \left(\sqrt{d} + \sqrt{\log K}\right) K^{-1/2-1/(d+1)},$$

where $C_{\mathrm{IS}}^{*,(0)} := 3^{1/2+1/(d+1)} C_4^* \left(1 + \sqrt{\log(1 + C_5^*)}\right)$.

It remains to consider the finitely many $1 \le K < K_*$. For these values of K, take the trivial one-cell stratification $B_1 = \Pi$, sample K inner parameters from π_θ, and set all output weights equal to zero. Then $h_K^{\mathrm{IS}} \in \mathcal{H}_K^{\mathrm{IS}}$ and $h_K^{\mathrm{IS}} \equiv 0$. Recalling that $R := MC_A + C_B$, we have $|h(x)| \le \int_\Pi |f(\theta)\sigma_m(A^\top x + B)| \mathrm{d}\pi_\theta(\theta) \le M_f R^m$ and hence $\mathbb{E}\left[\sup_{x\in\mathcal{X}} |h_K^{\mathrm{IS}}(x) - h(x)|\right] \le M_f R^m$.

We proceed analogously to the proof of case (i). Since $K^{1/2+1/(d+1)} \le K_*^{1/2+1/(d+1)}$ and $\sqrt{d} + \sqrt{\log(K)} \ge 1$, it follows that (3.8) holds for $1 \le K < K_*$ with $C_{\mathrm{IS}}^{*,(1)} := M_f R^m K_*^{1/2+1/(d+1)}$. Consequently, (3.8) holds for all $K \in \mathbb{N}^+$ with $C_{\mathrm{IS}}^* := \max\left\{C_{\mathrm{IS}}^{*,(0)}, C_{\mathrm{IS}}^{*,(1)}\right\}$. Equivalently,

$$C_{\mathrm{IS}}^* := \max\left\{3^{1/2+1/(d+1)} C_3 c_0^* L_u^* \left(1 + \sqrt{\log\left(1 + C_5^*\right)}\right), M_f R^m K_*^{1/2+1/(d+1)}\right\}, \tag{A.31}$$

where we recall that $C_3 > 0$ is a universal constant, $C_5^* = MC_A/c_0^*$, $c_0^* = (C_\Pi^*)^{1/(d+1)}$, $C_\Pi^* = 2\big(4(MC_A + C_B)\big)^{d+1}$, $L_u^* = M_f L_{\sigma,m}$, $R = MC_A + C_B$, and $K_* = \max\{2K_0, 2\}$, with $K_0 = \left\lceil (c_0^*/\min\{2MC_A, 2C_B\})^{d+1} \right\rceil$.

Combining the large- and small-K bounds as in the final paragraph of the proof of case (i) completes the proof. □

Proof of Theorem 3.6. By Theorem 3.5,

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|h_K^{\mathrm{IS}}(x)-h(x)\right|\right]\leq C(\sqrt{d}+\sqrt{\log(K)})K^{-1/2-\alpha}.$$

Applying Markov's inequality to $\sup_{x\in\mathcal{X}}\left|h_K^{\mathrm{IS}}(x)-h(x)\right|$ yields

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}\left|h_K^{\mathrm{IS}}(x)-h(x)\right|>C(\sqrt{d}+\sqrt{\log(K)})K^{-1/2-\alpha}\varepsilon^{-1}\right)\leq\varepsilon,$$

which proves (3.10) as required. □

Proof of Corollary 3.3. Fix $\varepsilon\in(0,1)$. By Theorem 3.6, there exists $h_K^{\mathrm{IS}}\in\mathcal{H}_K^{\mathrm{IS}}$ such that

$$\mathbb{P}\left(\mathcal{E}_{K,\varepsilon}^{\infty}\right)\geq 1-\varepsilon,$$

where

$$\mathcal{E}_{K,\varepsilon}^{\infty}:=\left\{\|h-h_K^{\mathrm{IS}}\|_{L^\infty(\mathcal{X})}\leq C(\sqrt{d}+\sqrt{\log(K)})K^{-1/2-\alpha}\varepsilon^{-1}\right\}.$$

On $\mathcal{E}_{K,\varepsilon}^{\infty}$, since $h_K^{\mathrm{IS}}\in\mathcal{H}_K^{\mathrm{IS}}$, one has

$$\inf_{g\in\mathcal{H}_K^{\mathrm{IS}}}\|h-g\|_{L^\infty(\mathcal{X})}\leq\|h-h_K^{\mathrm{IS}}\|_{L^\infty(\mathcal{X})}.$$

Hence

$$\mathbb{P}\left(\inf_{g\in\mathcal{H}_K^{\mathrm{IS}}}\|h-g\|_{L^\infty(\mathcal{X})}>C(\sqrt{d}+\sqrt{\log(K)})K^{-1/2-\alpha}\varepsilon^{-1}\right)\leq\varepsilon,$$

which proves the stated O_P rate. □

Proof of Lemma 3.7. For the large K construction in the proof of Theorem 3.5, by (3.7), after identifying k with (j,ℓ),

$$W_k=W_{j,\ell}:=\frac{P_j}{L_j}f(\theta_{j,\ell}).$$

The remaining $(K-\overline{K})$ output weights introduced by appending zeros are equal to zero.

In case (i), Assumption 3 gives $|f|\leq M_f$, while in case (ii), Assumption 9 gives $f\in\{-M_f,+M_f\}$. In both cases it holds that $|f|\leq M_f$ and one obtains

$$\|W\|_1\leq M_f\sum_{j=1}^{J}\sum_{\ell=1}^{L_j}\frac{P_j}{L_j}=M_f\sum_{j=1}^{J}P_j=M_f,$$

which yields (3.11). Additionally, we write that

$$\|W\|^2\leq M_f^2\sum_{j=1}^{J}\frac{P_j^2}{L_j}.$$

We now show that $P_j^2/L_j\leq 2P_j/\widetilde{K}$ for $j=1\ldots,J$. Recall that $K_j\in\{\lfloor\widetilde{K}P_j\rfloor,\lceil\widetilde{K}P_j\rceil\}$ and $L_j=K_j+\mathbb{1}_{\{K_j=0\}}$. If $\widetilde{K}P_j\leq 2$, then $L_j\geq 1$, so

$$\frac{P_j^2}{L_j}\leq P_j^2\leq\frac{2P_j}{\widetilde{K}}.$$

If $\widetilde{K}P_j>2$, then $K_j\geq\lfloor\widetilde{K}P_j\rfloor\geq\widetilde{K}P_j/2$, hence $L_j=K_j$ and $P_j^2/L_j\leq 2P_j/\widetilde{K}$. Therefore,

$$\sum_{j=1}^{J}\frac{P_j^2}{L_j}\leq\frac{2}{\widetilde{K}}\sum_{j=1}^{J}P_j=\frac{2}{\widetilde{K}},$$

which implies that $\|W\| \le \sqrt{2} M_f \widetilde{K}^{-1/2} \le \sqrt{6} M_f K^{-1/2}$, where the last inequality uses $\widetilde{K} = \lfloor K/2 \rfloor \ge K/3$ for $K \ge 2$. Hence (3.11) and (3.12) hold for the large K construction.

For the values $1 \le K < K_*$ considered separately at the end of Step 4 of the proof of Theorem 3.5, case (i) uses the one-cell construction with

$$W_k = \frac{1}{K} f(\theta_k), \qquad k = 1, \dots, K,$$

and hence

$$\|W\|_1 \le M_f, \qquad \|W\| \le M_f K^{-1/2} \le \sqrt{6} M_f K^{-1/2}.$$

In case (ii), all output weights are zero.

We conclude that (3.11) and (3.12) hold for all $K \in \mathbb{N}^+$. □

A.4 Uniform Monte Carlo approximation via metric entropy

We establish that the standard Monte Carlo estimator

$$h_K(x) := \frac{1}{K} \sum_{k=1}^{K} \phi_x(\theta_k), \qquad \theta_k \overset{\text{iid}}{\sim} \pi_\theta,$$

satisfies a uniform Monte Carlo bound over the compact covariate domain.

Proposition A.1. *Let Assumptions 1, 2 and 3 hold. Define $M_f := \sup_{\theta \in \Pi} |f(\theta)|$. Then there exists a constant $C_{\text{ent}} > 0$ such that*

$$\mathbb{E}_\theta \left[\sup_{x \in \mathcal{X}} |h(x) - h_K(x)| \right] \le C_{\text{ent}} K^{-1/2}.$$

The constant C_{ent} depends only on d, M, M_f, M_σ, L_σ, and the support Π.

Proof. By Assumption 3, $h(x) = \mathbb{E}_\theta[\phi_x(\theta)]$, and one can write that

$$h(x) - h_K(x) = \mathbb{E}_\theta[\phi_x(\theta)] - \frac{1}{K} \sum_{k=1}^{K} \phi_x(\theta_k).$$

Hence, it suffices to bound

$$\mathbb{E}_\theta \left[\sup_{x \in \mathcal{X}} \left| \frac{1}{K} \sum_{k=1}^{K} \phi_x(\theta_k) - \mathbb{E}_\theta[\phi_x(\theta)] \right| \right].$$

Next, using the symmetrization inequality (Boucheron et al., 2013, Lemma 11.4) one obtains

$$\mathbb{E}_\theta \left[\sup_{x \in \mathcal{X}} \left| \frac{1}{K} \sum_{k=1}^{K} \phi_x(\theta_k) - \mathbb{E}_\theta[\phi_x(\theta)] \right| \right] \le 2 \mathbb{E}_{\theta, \varepsilon} \left[\sup_{x \in \mathcal{X}} \left| \frac{1}{K} \sum_{k=1}^{K} \varepsilon_k \phi_x(\theta_k) \right| \right], \tag{A.32}$$

where $\varepsilon_1, \dots, \varepsilon_K$ are i.i.d. Rademacher random variables independent of $\theta_1, \dots, \theta_K$.

Conditional on $\theta_1, \dots, \theta_K$, we define

$$W_x := \frac{1}{K} \sum_{k=1}^{K} \varepsilon_k \phi_x(\theta_k), \qquad x \in \mathcal{X}. \tag{A.33}$$

Since Π is compact, define $M_A := \sup_{(A^\top, B)^\top \in \Pi} \|A\|$ and recall that $M_f := \sup_{\theta \in \Pi} |f(\theta)|$.

Under Assumption 2 and 3, for all $x_1, x_2 \in \mathcal{X}$ it holds that

$$\begin{aligned} |\phi_{x_1}(\theta) - \phi_{x_2}(\theta)| &= |f(A,B)||\sigma(A^\top x_1 + B) - \sigma(A^\top x_2 + B)| \\ &\leq M_f L_\sigma |A^\top (x_1 - x_2)| \leq M_f L_\sigma M_A \|x_1 - x_2\|. \end{aligned} \tag{A.34}$$

We proceed by noticing that for $x_1, x_2 \in \mathcal{X}$

$$W_{x_1} - W_{x_2} = \sum_{k=1}^{K} \varepsilon_k a_k(x_1, x_2), \quad \text{with} \quad a_k(x_1, x_2) := \frac{1}{K} \left(\phi_{x_1}(\theta_k) - \phi_{x_2}(\theta_k)\right)$$

and hence for every $\lambda \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\exp\left(\lambda(W_{x_1} - W_{x_2})\right)\right] &= \prod_{k=1}^{K} \mathbb{E}_\varepsilon \left[\exp\left(\lambda \varepsilon_k a_k(x_1, x_2)\right)\right] = \prod_{k=1}^{K} \cosh\left(\lambda a_k(x_1, x_2)\right) \\ &\leq \exp\left(\frac{\lambda^2}{2} \sum_{k=1}^{K} a_k(x_1, x_2)^2\right), \end{aligned}$$

where we used the definition of the moment generating function of Rademacher random variables and where the last step follows from $\cosh(t) \leq \exp(t^2/2)$ for all $t \in \mathbb{R}$.

Hence $(W_x)_{x \in \mathcal{X}}$ has conditionally sub-Gaussian increments with respect to the pseudometric defined as

$$d_K(x_1, x_2) := \left(\sum_{k=1}^{K} a_k(x_1, x_2)^2\right)^{1/2} = \left(\frac{1}{K^2} \sum_{k=1}^{K} |\phi_{x_1}(\theta_k) - \phi_{x_2}(\theta_k)|^2\right)^{1/2}, \quad x_1, x_2 \in \mathcal{X},$$

and for which (A.34) implies

$$d_K(x_1, x_2) \leq \frac{M_f L_\sigma M_A}{\sqrt{K}} \|x_1 - x_2\|. \tag{A.35}$$

We now apply Dudley's entropy bound (van der Vaart and Wellner, 2023, Corollary 2.2.9). For every $x_0 \in \mathcal{X}$, it holds

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\sup_{x \in \mathcal{X}} |W_x|\right] &\leq \mathbb{E}_\varepsilon \left[|W_{x_0}|\right] + C \int_0^\infty \sqrt{\log \mathcal{N}(r, \mathcal{X}, d_K)} dr \\ &= \mathbb{E}_\varepsilon \left[|W_{x_0}|\right] + C \int_0^{\operatorname{diam}(\mathcal{X}, d_K)} \sqrt{\log \mathcal{N}(r, \mathcal{X}, d_K)} dr, \end{aligned} \tag{A.36}$$

where $\mathcal{N}(r, \mathcal{X}, d_K)$ is the covering number of $\mathcal{X}$ with respect to the pseudometric d_K, $C > 0$ is a universal constant which absorbs factor 2 from the standard relation between covering and packing numbers, $\mathcal{N}(r, \mathcal{X}, d_K) \leq \mathcal{D}(r, \mathcal{X}, d_K) \leq \mathcal{N}(r/2, \mathcal{X}, d_K)$, with $\mathcal{D}$ denoting the packing number (see Definition 2.2.3 in van der Vaart and Wellner, 2023). The equality in (A.36) is obtained by noticing that $\mathcal{N}(r, \mathcal{X}, d_K) = 1$ for $r > \operatorname{diam}(\mathcal{X}, d_K)$. By (A.35), one has

$$\mathcal{N}(r, \mathcal{X}, d_K) \leq \mathcal{N}\left(\frac{r\sqrt{K}}{M_f L_\sigma M_A}, \mathcal{X}, \|\cdot\|\right). \tag{A.37}$$

By Assumption 1, $\mathcal{X} = [-M, M]^d$, and hence the volumetric covering bound (see, for example, Corollary 4.2.11 in Vershynin (2026)) gives for any $s > 0$,

$$\log \mathcal{N}(s, \mathcal{X}, \|\cdot\|) \leq d \log\left(1 + \frac{2M\sqrt{d}}{s}\right). \tag{A.38}$$

Notice that (A.35) implies that $\operatorname{diam}(\mathcal{X}, d_K) \leq \frac{M_f L_\sigma M_A}{\sqrt{K}} \operatorname{diam}(\mathcal{X}, \|\cdot\|) = \frac{2M\sqrt{d} M_f L_\sigma M_A}{\sqrt{K}}$. Additionally, by Cauchy-Schwarz, definition (A.33), and Assumption 3,

$$\mathbb{E}_\varepsilon \left[|W_{x_0}|\right] \leq \left(\mathbb{E}_\varepsilon \left[W_{x_0}^2\right]\right)^{1/2} = \frac{1}{K} \left(\sum_{k=1}^{K} |\phi_{x_0}(\theta_k)|^2 \right)^{1/2} \leq \frac{1}{\sqrt{K}} M_f M_\sigma. \tag{A.39}$$

Taking $s = r\sqrt{K}/(M_f L_\sigma M_A)$ in (A.38) and using (A.37) and (A.39) in (A.36) yields

$$\mathbb{E}_\varepsilon \left[\sup_{x \in \mathcal{X}} |W_x| \right] \leq \frac{1}{\sqrt{K}} \left(M_f M_\sigma + C M_f L_\sigma M_A \int_0^{2M\sqrt{d}} \sqrt{d \log\left(1 + \frac{2M\sqrt{d}}{s}\right)} ds \right),$$

which is finite.

Combining this expression with (A.32) yields the required result with

$$C_{\text{ent}} := 2 \left(M_f M_\sigma + C M_f L_\sigma M_A \int_0^{2M\sqrt{d}} \sqrt{d \log\left(1 + \frac{2M\sqrt{d}}{s}\right)} ds \right).$$

□

B Proofs for spectral properties and L^2 ridge estimation

B.1 Spectral deterioration and ridge effective dimension

Proof of Proposition 4.1. Fix $m < K$ and set $r_m := (C_\Pi/m)^{1/q_\Pi}$, with both m and K sufficiently large so that $r_m < \overline{r}$. By (4.7), there exists an $r_m/2$-net $\{\vartheta_1, \dots, \vartheta_{M_m}\}$ with $M_m \le m$. For each θ_k, choose $\vartheta_{c(k)}$ such that $\rho_\Pi(\theta_k, \vartheta_{c(k)}) \le r_m/2$, and set $\overline{g}_K(x) := \big(\phi_{\vartheta_{c(1)}}(x), \dots, \phi_{\vartheta_{c(K)}}(x)\big)^\top$. Since $\overline{g}_K(x)$ lies in a subspace of $\mathbb{R}^K$ of dimension at most m, Theorem 1 of Fan (1949) and (4.8) yield

$$\sum_{j=m+1}^{K} s_j^2 \le \mathbb{E}_X \|g_K(X) - \overline{g}_K(X)\|^2 = \sum_{k=1}^{K} \|\phi_{\theta_k} - \phi_{\vartheta_{c(k)}}\|_{L^2(X)}^2 \le K L_\phi^2 r_m^2 = K L_\phi^2 C_\Pi^{2/q_\Pi} m^{-2/q_\Pi},$$

which with $C_{\mathrm{tail}} := L_\phi^2 C_\Pi^{2/q_\Pi}$ proves (4.9). For all sufficiently large j, taking $m = \lfloor j/2 \rfloor$ and assuming $r_m < \overline{r}$ gives

$$(j-m) s_j^2 \le \sum_{\ell=m+1}^{K} s_\ell^2 \le C_{\mathrm{tail}} K m^{-2/q_\Pi}.$$

Since $j - m \ge j/2$ and $m \ge j/3$, taking $C_{\mathrm{eig}} = 2 \cdot 3^{2/q_\Pi} C_{\mathrm{tail}}$ yields (4.10). Finally, taking $j = K$ proves (4.11), as required. □

Proof of Corollary 4.1. We start by noticing that, since $0 \le \frac{s_j^2}{s_j^2 + \lambda_{K_n,n}} \le 1$, the definition of $\mathcal{N}_{K_n,\lambda}$ in (4.12) immediately gives the first term in the minimum in (4.13).

Next, again, using (4.12) and Proposition 4.1, we obtain

$$\mathcal{N}_{K_n,\lambda} \le m + \frac{1}{\lambda_{K_n,n}} \sum_{j=m+1}^{K_n} s_j^2 \le m + \frac{C_{\mathrm{tail}} K_n}{\lambda_{K_n,n}} m^{-2/q_\Pi}. \tag{B.1}$$

Minimizing the two terms on the right-hand side with respect to $m > 0$ gives

$$m_n^* := \left(\frac{2 C_{\mathrm{tail}} K_n}{q_\Pi \lambda_{K_n,n}}\right)^{q_\Pi/(q_\Pi+2)} \qquad \text{and} \qquad m_n := \lceil m_n^* \rceil. \tag{B.2}$$

Since $K_n/\lambda_{K_n,n} \to \infty$, we have $m_n \to \infty$, while $\lambda_{K_n,n} K_n^{2/q_\Pi} \to \infty$ implies $m_n/K_n \to 0$. Hence m_n is admissible for all sufficiently large n. Substituting m_n into (B.1) and using that, by (B.2), $m_n \le m_n^* + 1$ and $m_n^{-2/q_\Pi} \le (m_n^*)^{-2/q_\Pi}$ gives

$$\begin{aligned}
\mathcal{N}_{K_n,\lambda} &\le 1 + m_n^* + \frac{C_{\mathrm{tail}} K_n}{\lambda_{K_n,n}} (m_n^*)^{-2/q_\Pi} \\
&= 1 + \left(\frac{2 C_{\mathrm{tail}} K_n}{q_\Pi \lambda_{K_n,n}}\right)^{q_\Pi/(q_\Pi+2)} + \frac{C_{\mathrm{tail}} K_n}{\lambda_{K_n,n}} \left(\frac{2 C_{\mathrm{tail}} K_n}{q_\Pi \lambda_{K_n,n}}\right)^{-2/(q_\Pi+2)} \\
&= 1 + \left[\left(\frac{2}{q_\Pi}\right)^{q_\Pi/(q_\Pi+2)} + \left(\frac{q_\Pi}{2}\right)^{2/(q_\Pi+2)}\right] \left(\frac{C_{\mathrm{tail}} K_n}{\lambda_{K_n,n}}\right)^{q_\Pi/(q_\Pi+2)},
\end{aligned}$$

where defining $C_\mathcal{N}$ as a constant in front of the second summand gives the second argument in (4.13).

Finally, dividing both sides of the inequality above by K_n gives

$$\frac{\mathcal{N}_{K_n,\lambda}}{K_n} \le \frac{1}{K_n} + C_\mathcal{N} C_{\mathrm{tail}}^{q_\Pi/(q_\Pi+2)} \left(\lambda_{K_n,n} K_n^{2/q_\Pi}\right)^{-q_\Pi/(q_\Pi+2)} = o(1),$$

which proves (4.14), as required. □

B.2 Ridge-normalized covariance stability and L^2 estimation

Proof of Theorem 4.1. We follow the proof of Theorem 4.1 in Belloni et al. (2015), keeping track of the ridge normalization. First, by the triangle inequality,

$$\begin{aligned}\|\widehat{h}_{n,\lambda} - h\|_{L^2(X)} &\le \|g_{K_n}^\top(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})\|_{L^2(X)} + \|g_{K_n}^\top \beta_{K_n} - h\|_{L^2(X)} \\ &= \|\Gamma_{K_n}^{1/2}(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})\| + \Delta_{K_n},\end{aligned} \tag{B.3}$$

where the equality follows from $\|g_{K_n}^\top b\|_{L^2(X)}^2 = b^\top \Gamma_{K_n} b = \|\Gamma_{K_n}^{1/2} b\|^2$ for every $b \in \mathbb{R}^{K_n}$.

By (4.17) and (4.16), $\|\widehat{D}_{K_n,n,\lambda}\| = O_P(1)$.

Define the approximation residuals $\eta_i := h(X_i) - g_{K_n}(X_i)^\top \beta_{K_n}$, and $\eta := (\eta_1, \dots, \eta_n)^\top$. Then $Y = G_{K_n,n}\beta_{K_n} + \eta + e$. Using this together with the definition of $\widehat{\beta}_{K_n,n,\lambda}$ in (2.10), we obtain

$$\begin{aligned}\Gamma_{K_n}^{1/2}\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right) &= \Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top Y\right) - \Gamma_{K_n}^{1/2}\beta_{K_n} \\ &= \Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top e\right) + \Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top \eta\right) \\ &\quad + \Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top G_{K_n,n}\beta_{K_n}\right) - \Gamma_{K_n}^{1/2}\beta_{K_n} \\ &= \Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top e\right) + \Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top \eta\right) \\ &\quad + \Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(\widehat{\Gamma}_{K_n,n,\lambda} - \lambda_{K_n,n}I_{K_n}\right)\beta_{K_n} - \Gamma_{K_n}^{1/2}\beta_{K_n} \\ &= \Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top e\right) + \Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top \eta\right) \\ &\quad - \lambda_{K_n,n}\Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\beta_{K_n},\end{aligned}$$

where in the third equality we used that $n^{-1}G_{K_n,n}^\top G_{K_n,n} = \widehat{\Gamma}_{K_n,n,\lambda} - \lambda_{K_n,n}I_{K_n}$. Consequently, for the first term in the right-hand side of (B.3) it holds that

$$\begin{aligned}\|\Gamma_{K_n}^{1/2}\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right)\| &\le \|\Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top e\right)\| + \|\Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top \eta\right)\| \\ &\quad + \lambda_{K_n,n}\|\Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\beta_{K_n}\|.\end{aligned} \tag{B.4}$$

We now proceed in steps and study each term in the right-hand side of (B.4) separately.

Step 1. Using

$$\widehat{\Gamma}_{K_n,n,\lambda}^{-1} = \Gamma_{K_n,\lambda}^{-1/2}\widehat{D}_{K_n,n,\lambda}\Gamma_{K_n,\lambda}^{-1/2},$$

we obtain

$$\begin{aligned}\|\Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top e\right)\| &= \left\|\Gamma_{K_n}^{1/2}\Gamma_{K_n,\lambda}^{-1/2}\widehat{D}_{K_n,n,\lambda}\Gamma_{K_n,\lambda}^{-1/2}\left(n^{-1}G_{K_n,n}^\top e\right)\right\| \\ &\le \left\|\Gamma_{K_n}^{1/2}\Gamma_{K_n,\lambda}^{-1/2}\right\|\left\|\widehat{D}_{K_n,n,\lambda}\right\|\left\|\Gamma_{K_n,\lambda}^{-1/2}\left(n^{-1}G_{K_n,n}^\top e\right)\right\|.\end{aligned} \tag{B.5}$$

We recall that we have already established that $\|\widehat{D}_{K_n,n,\lambda}\| = O_P(1)$. Since $\Gamma_{K_n} \preceq \Gamma_{K_n,\lambda}$, it follows that $\left\|\Gamma_{K_n}^{1/2}\Gamma_{K_n,\lambda}^{-1/2}\right\| \le 1$ and hence it is enough to bound the last factor in (B.5), for which we write

$$\begin{aligned}\mathbb{E}\left[\|\Gamma_{K_n,\lambda}^{-1/2}\left(n^{-1}G_{K_n,n}^\top e\right)\|^2\right] &= \mathbb{E}\left[\left\|n^{-1}\sum_{i=1}^n \Gamma_{K_n,\lambda}^{-1/2}g_{K_n}(X_i)\epsilon_i\right\|^2\right] \\ &= n^{-1}\mathbb{E}\left[\epsilon_i^2 g_{K_n}(X_i)^\top \Gamma_{K_n,\lambda}^{-1}g_{K_n}(X_i)\right] \\ &\le n^{-1}\bar{\sigma}^2 \operatorname{tr}\left(\Gamma_{K_n}\Gamma_{K_n,\lambda}^{-1}\right),\end{aligned}$$

where we used that, under the hypotheses of the theorem, $\mathbb{E}[\epsilon_i|X_i] = 0$, the summands are centered, and the data are i.i.d.

Hence, by Markov's inequality, with (4.12), one has

$$\|\Gamma_{K_n,\lambda}^{-1/2}\left(n^{-1}G_{K_n,n}^{\top}e\right)\| = O_P\left(\sqrt{\frac{\mathcal{N}_{K_n,\lambda}}{n}}\right), \tag{B.6}$$

which gives the same bound for (B.5).

Step 2. We handle the second term in (B.4) in a similar way to the first. Using (4.17), we first notice that $\widehat{\Gamma}_{K_n,n,\lambda}^{-1} = \Gamma_{K_n,\lambda}^{-1/2}\widehat{D}_{K_n,n,\lambda}\Gamma_{K_n,\lambda}^{-1/2}$ and write

$$\|\Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^{\top}\eta\right)\| \le \|\Gamma_{K_n}^{1/2}\Gamma_{K_n,\lambda}^{-1/2}\|\|\widehat{D}_{K_n,n,\lambda}\|\|\Gamma_{K_n,\lambda}^{-1/2}\left(n^{-1}G_{K_n,n}^{\top}\eta\right)\|.$$

Since $\|\widehat{D}_{K_n,n,\lambda}\| = O_P(1)$ and $\|\Gamma_{K_n}^{1/2}\Gamma_{K_n,\lambda}^{-1/2}\| \le 1$, it remains to analyze the last factor.

Using the definition of η_i and the fact that β_{K_n} is an $L^2(X)$ projection coefficient and $\mathbb{E}\left[g_{K_n}(X_i)\eta_i\right] = 0$, we immediately have that

$$\mathbb{E}[g_{K_n}(X_i)\eta_i] = \mathbb{E}_X[g_{K_n}(X_i)(h(X_i) - g_{K_n}(X_i)^{\top}\beta_{K_n})] = 0,$$

which further implies $\mathbb{E}_X[\Gamma_{K_n,\lambda}^{-1/2}g_{K_n}(X)\eta(X)] = 0$. We can now write

$$\begin{aligned}\mathbb{E}\left[\|\Gamma_{K_n,\lambda}^{-1/2}\left(n^{-1}G_{K_n,n}^{\top}\eta\right)\|^2\right] &= n^{-2}\mathbb{E}\left[\left\|\sum_{i=1}^{n}\widetilde{g}_{K_n}(X_i)\eta_i\right\|^2\right] = n^{-1}\mathbb{E}\left[\eta_i^2\|\widetilde{g}_{K_n}(X_i)\|^2\right] \\ &\le \frac{\widetilde{\xi}_{K_n,\lambda}^2}{n}\mathbb{E}[\eta_i^2] = \frac{\widetilde{\xi}_{K_n,\lambda}^2}{n}\Delta_{K_n}^2,\end{aligned} \tag{B.7}$$

where in the first equality we applied again the fact that cross terms are zero by independence; in the inequality, we used $\|\widetilde{g}_{K_n}(X_i)\| \le \sup_{x\in\mathcal{X}}\|\widetilde{g}_{K_n}(x)\| \le \widetilde{\xi}_{K_n,\lambda}$, while the last equality follows from the definition $\mathbb{E}[\eta_i^2] = \|h - g_{K_n}^{\top}\beta_{K_n}\|_{L^2(X)}^2 = \Delta_{K_n}^2$.

By (B.7), Markov's inequality implies $\|\Gamma_{K_n,\lambda}^{-1/2}\left(n^{-1}G_{K_n,n}^{\top}\eta\right)\| = O_P(n^{-1/2}\widetilde{\xi}_{K_n,\lambda}\Delta_{K_n})$, while Assumption 10 yields $n^{-1/2}\widetilde{\xi}_{K_n,\lambda} = o(1)$ (see again Remark 11), and hence one gets

$$\|\Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^{\top}\eta\right)\| = O_P\left(\frac{\widetilde{\xi}_{K_n,\lambda}}{\sqrt{n}}\Delta_{K_n}\right) = O_P\left(\Delta_{K_n}\right). \tag{B.8}$$

Step 3. For the third and last term in (B.4), adding and subtracting $\lambda_{K_n,n}\Gamma_{K_n}^{1/2}\Gamma_{K_n,\lambda}^{-1}\beta_{K_n}$ yields

$$\lambda_{K_n,n}\|\Gamma_{K_n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\beta_{K_n}\| \le \lambda_{K_n,n}\|\Gamma_{K_n}^{1/2}\Gamma_{K_n,\lambda}^{-1}\beta_{K_n}\| + \lambda_{K_n,n}\|\Gamma_{K_n}^{1/2}\left(\widehat{\Gamma}_{K_n,n,\lambda}^{-1} - \Gamma_{K_n,\lambda}^{-1}\right)\beta_{K_n}\|, \tag{B.9}$$

where the first term on the right-hand side is the population ridge bias measured in the $L^2(X)$ prediction norm, $B_{K_n,\lambda}^{\text{pop}}$. For the second term, we use that $\widehat{\Gamma}_{K_n,n,\lambda}^{-1} - \Gamma_{K_n,\lambda}^{-1} = -\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n}\right)\Gamma_{K_n,\lambda}^{-1}$ and apply once more the identity for $\widehat{\Gamma}_{K_n,n,\lambda}^{-1}$ in the beginning of Step 1 to get

$$\begin{aligned}&\lambda_{K_n,n}\|\Gamma_{K_n}^{1/2}\left(\widehat{\Gamma}_{K_n,n,\lambda}^{-1} - \Gamma_{K_n,\lambda}^{-1}\right)\beta_{K_n}\| \\ &\quad\le \lambda_{K_n,n}\|\Gamma_{K_n}^{1/2}\Gamma_{K_n,\lambda}^{-1/2}\|\|\widehat{D}_{K_n,n,\lambda}\|\|\Gamma_{K_n,\lambda}^{-1/2}\left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n}\right)\Gamma_{K_n,\lambda}^{-1}\beta_{K_n}\| \\ &\quad\le \lambda_{K_n,n}\|\widehat{D}_{K_n,n,\lambda}\|\|\Gamma_{K_n,\lambda}^{-1/2}\left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n}\right)\Gamma_{K_n,\lambda}^{-1}\beta_{K_n}\|.\end{aligned}$$

Set $v_{K_n,\lambda} := \Gamma_{K_n,\lambda}^{-1}\beta_{K_n}$ and define

$$\zeta_{i,K_n,\lambda} := \Gamma_{K_n,\lambda}^{-1/2}\left(g_{K_n}(X_i)g_{K_n}(X_i)^\top - \Gamma_{K_n}\right)v_{K_n,\lambda}, \quad i = 1,\dots,n.$$

Then $\mathbb{E}[\zeta_{i,K_n,\lambda}] = 0$, since $\mathbb{E}\left[g_{K_n}(X_i)g_{K_n}(X_i)^\top\right] = \Gamma_{K_n}$. Additionally, it holds that $\Gamma_{K_n,\lambda}^{-1/2}(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n})v_{K_n,\lambda} = n^{-1}\sum_{i=1}^n \zeta_{i,K_n,\lambda}$. Moreover,

$$\begin{aligned}
\mathbb{E}\left[\|\zeta_{i,K_n,\lambda}\|^2\right] &= \mathbb{E}_X\left[\left\|\widetilde{g}_{K_n}(X)g_{K_n}(X)^\top v_{K_n,\lambda} - \Gamma_{K_n,\lambda}^{-1/2}\Gamma_{K_n}v_{K_n,\lambda}\right\|^2\right] \\
&= \mathbb{E}_X\left[\|\widetilde{g}_{K_n}(X)\|^2\left(g_{K_n}(X)^\top v_{K_n,\lambda}\right)^2\right] + v_{K_n,\lambda}^\top\Gamma_{K_n}\Gamma_{K_n,\lambda}^{-1}\Gamma_{K_n}v_{K_n,\lambda} \\
&\qquad - 2\,v_{K_n,\lambda}^\top\mathbb{E}_X\left[g_{K_n}(X)\widetilde{g}_{K_n}(X)^\top\right]\Gamma_{K_n,\lambda}^{-1/2}\Gamma_{K_n}v_{K_n,\lambda} \\
&\le \mathbb{E}_X\left[\|\widetilde{g}_{K_n}(X)\|^2\left(g_{K_n}(X)^\top v_{K_n,\lambda}\right)^2\right] \\
&\le \widetilde{\xi}_{K_n,\lambda}^2\left(v_{K_n,\lambda}^\top\Gamma_{K_n}v_{K_n,\lambda}\right) = \widetilde{\xi}_{K_n,\lambda}^2\left(\frac{B_{K_n,\lambda}^{\mathrm{pop}}}{\lambda_{K_n,n}}\right)^2,
\end{aligned}$$

where the first inequality follows from the fact that $\mathbb{E}_X\left[g_{K_n}(X_i)\widetilde{g}_{K_n}(X)^\top\right] = \Gamma_{K_n}\Gamma_{K_n,\lambda}^{-1/2}$. Hence, by independence and Markov's inequality,

$$\left\|\Gamma_{K_n,\lambda}^{-1/2}\left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n}\right)v_{K_n,\lambda}\right\| = O_P\left(\frac{\widetilde{\xi}_{K_n,\lambda}}{\sqrt{n}}\frac{B_{K_n,\lambda}^{\mathrm{pop}}}{\lambda_{K_n,n}}\right).$$

Since $\|\widehat{D}_{K_n,n,\lambda}\| = O_P(1)$, it follows that

$$\lambda_{K_n,n}\left\|\Gamma_{K_n}^{1/2}\left(\widehat{\Gamma}_{K_n,n,\lambda}^{-1} - \Gamma_{K_n,\lambda}^{-1}\right)\beta_{K_n}\right\| = O_P\left(B_{K_n,\lambda}^{\mathrm{pop}}\frac{\widetilde{\xi}_{K_n,\lambda}}{\sqrt{n}}\right).$$

Assumption 10 implies $\widetilde{\xi}_{K_n,\lambda}/\sqrt{n} = o(1)$, and therefore the third term in (B.4) is $O_P(B_{K_n,\lambda}^{\mathrm{pop}})$.

Combining the initial error bound (B.3) with our derivations for the three terms in (B.4), namely, (B.6), (B.8), and the third one immediately above, we obtain the required bound in (4.19), completing the proof. □

Proof of Lemma 4.2. Similarly to the decomposition (B.4) in the proof of Theorem 4.1, we write

$$\begin{aligned}
\widehat{\Gamma}_{K_n,n}^{1/2}\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right) &= \widehat{\Gamma}_{K_n,n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top e\right) + \widehat{\Gamma}_{K_n,n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1}G_{K_n,n}^\top\eta\right) \\
&\qquad - \lambda_{K_n,n}\widehat{\Gamma}_{K_n,n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\beta_{K_n}.
\end{aligned} \tag{B.10}$$

Next, since $\widehat{\Gamma}_{K_n,n} \preceq \widehat{\Gamma}_{K_n,n,\lambda}$, for every $v \in \mathbb{R}^{K_n}$ it holds that

$$\begin{aligned}
\left\|\widehat{\Gamma}_{K_n,n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}v\right\|^2 &\le v^\top\widehat{\Gamma}_{K_n,n,\lambda}^{-1}v = \left(\Gamma_{K_n,\lambda}^{-1/2}v\right)^\top\widehat{D}_{K_n,n,\lambda}\left(\Gamma_{K_n,\lambda}^{-1/2}v\right) \\
&\le \|\widehat{D}_{K_n,n,\lambda}\|\left\|\Gamma_{K_n,\lambda}^{-1/2}v\right\|^2.
\end{aligned} \tag{B.11}$$

Under Assumption 10, $\|\widehat{D}_{K_n,n,\lambda}\| = O_P(1)$, and applying (B.11) to the first term in (B.10), together with the corresponding bound (B.6) derived in Step 1 of the proof of Theorem 4.1, gives

$$\left\|\widehat{\Gamma}_{K_n,n}^{1/2}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(\frac{1}{n}G_{K_n,n}^\top e\right)\right\| = O_P\left(\sqrt{\frac{\mathcal{N}_{K_n,\lambda}}{n}}\right). \tag{B.12}$$

Similarly, for the second term in (B.10), we use the corresponding bound (B.8) derived in Step 2 of the proof of Theorem 4.1 and obtain

$$\left\|\widehat{\Gamma}^{1/2}_{K_n,n}\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\left(\frac{1}{n}G^{\top}_{K_n,n}\eta\right)\right\| = O_P\left(\frac{\widetilde{\xi}_{K_n,\lambda}\Delta_{K_n}}{\sqrt{n}}\right) = O_P\left(\Delta_{K_n}\right), \tag{B.13}$$

where in the last step we used that, by Assumption 10, $\widetilde{\xi}_{K_n,\lambda}/\sqrt{n} = O_P(1)$.

For the last term, again, similarly to (B.9), we write

$$\begin{aligned}\lambda_{K_n,n}\widehat{\Gamma}^{1/2}_{K_n,n}\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\beta_{K_n} &= \lambda_{K_n,n}\widehat{\Gamma}^{1/2}_{K_n,n}\Gamma^{-1}_{K_n,\lambda}\beta_{K_n} \\ &\quad + \lambda_{K_n,n}\widehat{\Gamma}^{1/2}_{K_n,n}\left(\widehat{\Gamma}^{-1}_{K_n,n,\lambda} - \Gamma^{-1}_{K_n,\lambda}\right)\beta_{K_n}.\end{aligned} \tag{B.14}$$

For the first term on the right-hand side of (B.14), we write

$$\mathbb{E}\left[\left\|\lambda_{K_n,n}\widehat{\Gamma}^{1/2}_{K_n,n}\Gamma^{-1}_{K_n,\lambda}\beta_{K_n}\right\|^2\right] = \left\|\lambda_{K_n,n}\Gamma^{1/2}_{K_n}\Gamma^{-1}_{K_n,\lambda}\beta_{K_n}\right\|^2 = \left(B^{\mathrm{pop}}_{K_n,\lambda}\right)^2,$$

where in the first equality we used that $\mathbb{E}\big[\widehat{\Gamma}_{K_n,n}\big] = \Gamma_{K_n}$. Consequently, by Markov's inequality,

$$\left\|\lambda_{K_n,n}\widehat{\Gamma}^{1/2}_{K_n,n}\Gamma^{-1}_{K_n,\lambda}\beta_{K_n}\right\| = O_P\left(B^{\mathrm{pop}}_{K_n,\lambda}\right). \tag{B.15}$$

For the second term in (B.14), again similarly to Step 3 of the proof of Theorem 4.1, we obtain

$$\begin{aligned}&\lambda_{K_n,n}\left\|\widehat{\Gamma}^{1/2}_{K_n,n}\left(\widehat{\Gamma}^{-1}_{K_n,n,\lambda} - \Gamma^{-1}_{K_n,\lambda}\right)\beta_{K_n}\right\| \\ &\quad\le \lambda_{K_n,n}\|\widehat{D}_{K_n,n,\lambda}\|^{1/2}\left\|\Gamma^{-1/2}_{K_n,\lambda}\left(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n}\right)\Gamma^{-1}_{K_n,\lambda}\beta_{K_n}\right\| \\ &\quad= O_P\left(\frac{\widetilde{\xi}_{K_n,\lambda}B^{\mathrm{pop}}_{K_n,\lambda}}{\sqrt{n}}\right) = O_P\left(B^{\mathrm{pop}}_{K_n,\lambda}\right),\end{aligned} \tag{B.16}$$

where the last step follows from noting that, under Assumption 10, $\widetilde{\xi}_{K_n,\lambda}/\sqrt{n} = O_P(1)$.

Combining (B.12), (B.13), (B.15), and (B.16) in (B.10) proves the claim, as required. □

B.3 RWNN specialization and primitive rates

Proof of Corollary 4.2. As in the proof of Lemma 3.3, compactness of $\Pi \subset \mathbb{R}^{d+1}$ and the equivalence of ρ_Π with the Euclidean metric established in Remark 2 gives $|\mathcal{N}_{r/2}| \lesssim r^{-(d+1)}$. Moreover, Assumption 2 and (3.1) imply $\|\phi_{\theta_1} - \phi_{\theta_2}\|_{L^2(X)} \le L_\sigma L_\rho \rho_\Pi(\theta_1,\theta_2)$. The claim follows from Proposition 4.1 with $q_\Pi = d+1$. □

Proof of Lemma 4.3. Let $W^*_{K_n}$ be the control-neighbors coefficient vector used in the proof of Corollary E.1, and set $\mathfrak{D}^*_{K_n} := \|h - g^{\top}_{K_n}W^*_{K_n}\|_{L^2(X)}$. Define $\beta^{\mathrm{pop}}_{K_n,\lambda} := \Gamma^{-1}_{K_n,\lambda}\Gamma_{K_n}\beta_{K_n}$. Then $\beta_{K_n} - \beta^{\mathrm{pop}}_{K_n,\lambda} = \lambda_{K_n,n}\Gamma^{-1}_{K_n,\lambda}\beta_{K_n}$, and hence $\|g^{\top}_{K_n}(\beta_{K_n} - \beta^{\mathrm{pop}}_{K_n,\lambda})\|_{L^2(X)} = B^{\mathrm{pop}}_{K_n,\lambda}$. Since β_{K_n} is the $L^2(X)$ projection coefficient,

$$\|h - g^{\top}_{K_n}\beta^{\mathrm{pop}}_{K_n,\lambda}\|^2_{L^2(X)} = \Delta^2_{K_n} + \left(B^{\mathrm{pop}}_{K_n,\lambda}\right)^2.$$

Moreover, by optimality of $\beta^{\mathrm{pop}}_{K_n,\lambda}$ for the population ridge criterion,

$$\|h - g^{\top}_{K_n}\beta^{\mathrm{pop}}_{K_n,\lambda}\|^2_{L^2(X)} + \lambda_{K_n,n}\|\beta^{\mathrm{pop}}_{K_n,\lambda}\|^2 \le \|h - g^{\top}_{K_n}b\|^2_{L^2(X)} + \lambda_{K_n,n}\|b\|^2 \quad \forall\, b \in \mathbb{R}^{K_n},$$

and thus

$$\Delta^2_{K_n} + \left(B^{\mathrm{pop}}_{K_n,\lambda}\right)^2 + \lambda_{K_n,n}\|\beta^{\mathrm{pop}}_{K_n,\lambda}\|^2 \le (\mathfrak{D}^*_{K_n})^2 + \lambda_{K_n,n}\|W^*_{K_n}\|^2.$$

As $\Delta_{K_n} \le \mathfrak{D}^*_{K_n}$, it follows, as required, that

$$B^{\mathrm{pop}}_{K_n,\lambda} \le \sqrt{2}\left(\mathfrak{D}^*_{K_n} + \sqrt{\lambda_{K_n,n}}\|W^*_{K_n}\|\right) = O_{P_\theta}\left(K_n^{-1/2-\beta/(d+1)} + \sqrt{\frac{\lambda_{K_n,n}\log(K_n)}{K_n}}\right),$$

where in the last equality we used the bounds for $W^*_{K_n}$ established in the proof of Corollary E.1, namely $\mathfrak{D}^*_{K_n} = O_{P_\theta}(K_n^{-1/2-\beta/(d+1)})$ and $\|W^*_{K_n}\| = O_{P_\theta}(\sqrt{\log(K_n)/K_n})$, where the latter follows by Lemma 3.3. □

Proof of Theorem 4.4. By Theorem 4.1, together with Corollary E.1, Lemma 4.3, and the assumed P_θ-probability conditions on the random feature maps, (4.20) immediately follows.

Suppose now that (4.21) holds. By using the effective-dimension bound in (4.13) with $q_\Pi = d+1$, (4.22) holds. Next, since $r > 2\beta/(d+1)$ by (4.21)

$$\sqrt{\frac{\lambda_{K_n,n}\log(K_n)}{K_n}} = o\left(K_n^{-1/2-\beta/(d+1)}\right).$$

Hence,

$$\|\widehat{h}_{n,\lambda} - h\|_{L^2(X)} = O_P\left(\sqrt{\frac{K_n^{\nu_r}}{n}} + K_n^{-1/2-\beta/(d+1)}\right). \tag{B.17}$$

We choose K_n by balancing the stochastic and approximation terms, so that $n^{-1/2}K_n^{\nu_r/2} \asymp K_n^{-1/2-\beta/(d+1)}$ or, equivalently,

$$K_n \asymp n^{\frac{1}{1+\nu_r+2\beta/(d+1)}}.$$

Substituting this choice into either term in (B.17) yields

$$K_n^{-1/2-\beta/(d+1)} \asymp \left(n^{\frac{1}{1+\nu_r+2\beta/(d+1)}}\right)^{-1/2-\beta/(d+1)} = n^{-\frac{1/2+\beta/(d+1)}{1+\nu_r+2\beta/(d+1)}}.$$

Noticing that, by construction, the stochastic term has the same order, this gives (4.24), completing the proof. □

C Proofs for directional and smooth functional inference

C.1 Ridge-normalized linearization and directional asymptotic normality

Proof of Lemma 4.5. We start by writing

$$\begin{aligned}
\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n} &= \widehat{\Gamma}^{-1}_{K_n,n,\lambda}(n^{-1}G^\top_{K_n,n}Y) - \beta_{K_n} \\
&= \widehat{\Gamma}^{-1}_{K_n,n,\lambda}\big[n^{-1}G^\top_{K_n,n}G_{K_n,n}\beta_{K_n} + n^{-1}G^\top_{K_n,n}(\eta+e)\big] - \beta_{K_n} \\
&= \widehat{\Gamma}^{-1}_{K_n,n,\lambda}\big[\widehat{\Gamma}_{K_n,n}\beta_{K_n} + n^{-1}G^\top_{K_n,n}(\eta+e)\big] - \beta_{K_n} \\
&= \widehat{\Gamma}^{-1}_{K_n,n,\lambda}\big[(\widehat{\Gamma}_{K_n,n,\lambda} - \lambda_{K_n,n}I_{K_n})\beta_{K_n} + n^{-1}G^\top_{K_n,n}(\eta+e)\big] - \beta_{K_n} \\
&= \widehat{\Gamma}^{-1}_{K_n,n,\lambda}\big(n^{-1}G^\top_{K_n,n}(\eta+e)\big) - \lambda_{K_n,n}\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\beta_{K_n},
\end{aligned}$$

where in the third and fourth equalities we used (4.1). Multiplying by $\sqrt{n}\alpha^\top$ gives

$$\sqrt{n}\alpha^\top\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right) = \alpha^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\mathbb{G}_n\left[g_i(\eta_i+\epsilon_i)\right] - \sqrt{n}\lambda_{K_n,n}\alpha^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\beta_{K_n},$$

where

$$\mathbb{G}_n[g_i(\epsilon_i+\eta_i)] = \frac{1}{\sqrt{n}}\sum_{i=1}^n\big(g_i(\epsilon_i+\eta_i) - \mathbb{E}[g_i(\epsilon_i+\eta_i)]\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^n g_i(\epsilon_i+\eta_i),$$

since $\mathbb{E}[g_i\epsilon_i] = \mathbb{E}\big[g_i\,\mathbb{E}[\epsilon_i \mid X_i]\big] = 0$ by the law of iterated expectations and by Assumption 11, and $\mathbb{E}[g_i\eta_i] = 0$ by the $L^2(X)$ projection property of β_{K_n}.

Next, note that

$$\alpha^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda} = \alpha^\top\Gamma^{-1/2}_{K_n,\lambda}\Gamma^{1/2}_{K_n,\lambda}\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\Gamma^{1/2}_{K_n,\lambda}\Gamma^{-1/2}_{K_n,\lambda} = \widetilde{\alpha}^\top\widehat{D}_{K_n,n,\lambda}\Gamma^{-1/2}_{K_n,\lambda}.$$

Similarly to the proof of Theorem 4.1, recall from (4.17) that $\widehat{D}_{K_n,n,\lambda} = (I_{K_n} + A_{K_n,n,\lambda})^{-1}$.

Hence,

$$\begin{aligned}
\sqrt{n}\alpha^\top\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right) &= \widetilde{\alpha}^\top\widehat{D}_{K_n,n,\lambda}\left(\Gamma^{-1/2}_{K_n,\lambda}\mathbb{G}_n\left[g_i(\eta_i+\epsilon_i)\right] - \sqrt{n}\lambda_{K_n,n}\Gamma^{-1/2}_{K_n,\lambda}\beta_{K_n}\right) \\
&= \widetilde{\alpha}^\top\widehat{D}_{K_n,n,\lambda}\left(\mathbb{G}_n\left[\widetilde{g}_i(\eta_i+\epsilon_i)\right] - \sqrt{n}\lambda_{K_n,n}\Gamma^{-1/2}_{K_n,\lambda}\beta_{K_n}\right) \\
&= \widetilde{\alpha}^\top\widehat{D}_{K_n,n,\lambda}\mathbb{G}_n\left[\widetilde{g}_i(\eta_i+\epsilon_i)\right] + \widetilde{\alpha}^\top\mathbb{G}_n\left[\widetilde{g}_i\epsilon_i\right] \\
&\quad - \widetilde{\alpha}^\top\mathbb{G}_n\left[\widetilde{g}_i\epsilon_i\right] - \sqrt{n}\lambda_{K_n,n}\widetilde{\alpha}^\top\widehat{D}_{K_n,n,\lambda}\Gamma^{-1/2}_{K_n,\lambda}\beta_{K_n} \\
&= \widetilde{\alpha}^\top\mathbb{G}_n\left[\widetilde{g}_i\epsilon_i\right] + \widetilde{\alpha}^\top\left(\widehat{D}_{K_n,n,\lambda} - I_{K_n}\right)\mathbb{G}_n\left[\widetilde{g}_i(\eta_i+\epsilon_i)\right] \\
&\quad + \widetilde{\alpha}^\top\mathbb{G}_n\left[\widetilde{g}_i\eta_i\right] - \sqrt{n}\lambda_{K_n,n}\widetilde{\alpha}^\top\widehat{D}_{K_n,n,\lambda}\Gamma^{-1/2}_{K_n,\lambda}\beta_{K_n},
\end{aligned}$$

where we used the linearity of $\mathbb{G}_n$ and the identity $\Gamma^{-1/2}_{K_n,\lambda}g_i = \widetilde{g}_i$. The last equality then gives the expansion (4.26), with the last three terms of the right-hand side corresponding to $R_{1,n}(\alpha)$, $R_{2,n}(\alpha)$, and $R_{\lambda,n}(\alpha)$ in (4.27), respectively. We now bound each residual term separately.

Step 1. We first bound

$$R_{1,n}(\alpha) = \widetilde{\alpha}^\top\left(\widehat{D}_{K_n,n,\lambda} - I_{K_n}\right)\mathbb{G}_n\left[\widetilde{g}_i(\eta_i+\epsilon_i)\right].$$

By (4.25), $\|A_{K_n,n,\lambda}\| = o_P(1)$, $\|\widehat{D}_{K_n,n,\lambda}\| = O_P(1)$ and $\|\widehat{D}_{K_n,n,\lambda} - I_{K_n}\| = O_P\left(\|A_{K_n,n,\lambda}\|\right)$. We now split the term $R_{1,n}(\alpha)$ as

$$R_{1,n}(\alpha) = R'_{1,n}(\alpha) + R''_{1,n}(\alpha),$$

where

$$R'_{1,n}(\alpha) := \widetilde{\alpha}^\top \left(\widehat{D}_{K_n,n,\lambda} - I_{K_n}\right) \mathbb{G}_n[\widetilde{g}_i \epsilon_i],$$
$$R''_{1,n}(\alpha) := \widetilde{\alpha}^\top \left(\widehat{D}_{K_n,n,\lambda} - I_{K_n}\right) \mathbb{G}_n[\widetilde{g}_i \eta_i].$$

Note that, conditional on the sample $X_{1:n}$,

$$\mathbb{E}\left[R'_{1,n}(\alpha) \,\middle|\, X_{1:n}\right] = \widetilde{\alpha}^\top \left(\widehat{D}_{K_n,n,\lambda} - I_{K_n}\right) \mathbb{G}_n\left[\widetilde{g}_i \,\mathbb{E}[\epsilon_i \mid X_i]\right] = 0.$$

Moreover, exactly as in the proof of Theorem 4.1,

$$\Gamma_{K_n,\lambda}^{-1/2} \widehat{\Gamma}_{K_n,n} \Gamma_{K_n,\lambda}^{-1/2} = \Gamma_{K_n,\lambda}^{-1/2} \Gamma_{K_n} \Gamma_{K_n,\lambda}^{-1/2} + A_{K_n,n,\lambda},$$

so that

$$\left\|\Gamma_{K_n,\lambda}^{-1/2} \widehat{\Gamma}_{K_n,n} \Gamma_{K_n,\lambda}^{-1/2}\right\| \le \left\|\Gamma_{K_n,\lambda}^{-1/2} \Gamma_{K_n} \Gamma_{K_n,\lambda}^{-1/2}\right\| + o_P(1) = O_P(1).$$

We write

$$\mathbb{E}\left[\mathbb{G}_n[\widetilde{g}_i \epsilon_i] \mathbb{G}_n[\epsilon_i \widetilde{g}_i^\top] \,\middle|\, X_{1:n}\right] = \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n \epsilon_i^2 \widetilde{g}_i \widetilde{g}_i^\top \,\middle|\, X_{1:n}\right] \le \overline{\sigma}^2 \Gamma_{K_n,\lambda}^{-1/2} \widehat{\Gamma}_{K_n,n} \Gamma_{K_n,\lambda}^{-1/2}.$$

and we thus obtain

$$\begin{aligned}\mathrm{Var}(R'_{1,n}(\alpha) \,|\, X_{1:n}) &\le \overline{\sigma}^2 \, \widetilde{\alpha}^\top \left(\widehat{D}_{K_n,n,\lambda} - I_{K_n}\right) \Gamma_{K_n,\lambda}^{-1/2} \widehat{\Gamma}_{K_n,n} \Gamma_{K_n,\lambda}^{-1/2} \left(\widehat{D}_{K_n,n,\lambda} - I_{K_n}\right) \widetilde{\alpha} \\ &\le \overline{\sigma}^2 \|\widetilde{\alpha}\|^2 \, \|\widehat{D}_{K_n,n,\lambda} - I_{K_n}\|^2 \left\|\Gamma_{K_n,\lambda}^{-1/2} \widehat{\Gamma}_{K_n,n} \Gamma_{K_n,\lambda}^{-1/2}\right\|.\end{aligned}$$

Hence, conditional Chebyshev's inequality, together with $\|\widehat{D}_{K_n,n,\lambda} - I_{K_n}\| = O_P(\|A_{K_n,n,\lambda}\|)$, yields

$$R'_{1,n}(\alpha) = O_P\left(\|\widetilde{\alpha}\| \, \|A_{K_n,n,\lambda}\|\right). \tag{C.1}$$

Exactly as in the proof of Theorem 4.1,

$$\|\mathbb{G}_n[\widetilde{g}_i \eta_i]\| = O_P\left(\widetilde{\xi}_{K_n,\lambda} \Delta_{K_n}\right).$$

Consequently,

$$R''_{1,n}(\alpha) = O_P\left(\|\widetilde{\alpha}\| \|A_{K_n,n,\lambda}\| \widetilde{\xi}_{K_n,\lambda} \Delta_{K_n}\right). \tag{C.2}$$

Combining (C.1) and (C.2) gives

$$R_{1,n}(\alpha) = O_P\left(\|\widetilde{\alpha}\| \|A_{K_n,n,\lambda}\| \left(1 + \widetilde{\xi}_{K_n,\lambda} \Delta_{K_n}\right)\right),$$

which proves the bound in (4.28).

Step 2. For the term $R_{2,n}(\alpha) = \widetilde{\alpha}^\top \mathbb{G}_n\left[\widetilde{g}_i \eta_i\right]$, by the Cauchy-Schwarz inequality, it holds that $|R_{2,n}(\alpha)| \le \|\widetilde{\alpha}\| \, \|\mathbb{G}_n[\widetilde{g}_i \eta_i]\|$. Hence, using $\|\mathbb{G}_n[\widetilde{g}_i \eta_i]\| = O_P\left(\widetilde{\xi}_{K_n,\lambda} \Delta_{K_n}\right)$ derived in Step 1, gives (4.29), as required.

Step 3. Finally, we apply the Cauchy-Schwarz inequality to

$$R_{\lambda,n}(\alpha) = -\sqrt{n} \lambda_{K_n,n} \widetilde{\alpha}^\top \widehat{D}_{K_n,n,\lambda} \Gamma_{K_n,\lambda}^{-1/2} \beta_{K_n}$$

and obtain

$$|R_{\lambda,n}(\alpha)| \le \sqrt{n} \, \lambda_{K_n,n} \|\widetilde{\alpha}\| \|\widehat{D}_{K_n,n,\lambda}\| \left\|\Gamma_{K_n,\lambda}^{-1/2} \beta_{K_n}\right\|.$$

Noticing that $\|\widehat{D}_{K_n,n,\lambda}\| = O_P(1)$ yields the bound in (4.30), as required. □

Proof of Theorem 4.6. Fix $\ell \in \{0,1\}$ and, for the sake of convenience, define

$$\zeta_{n,i}^{(\ell)} := \epsilon_i + \ell\eta_i \qquad \text{and} \qquad \sigma_{n,\ell} := \sqrt{\sigma^{2,(\ell)}_{K_n,\lambda}(\alpha_{K_n})}.$$

Since $\mathbb{E}[\widetilde{g}_i\epsilon_i] = 0$ and $\mathbb{E}[\widetilde{g}_i\eta_i] = 0$, we have $\mathbb{E}\left[\widetilde{g}_i\zeta_{n,i}^{(\ell)}\right] = 0$. By Lemma 4.5, and since

$$\widetilde{\alpha}_{K_n,\lambda}^\top \mathbb{G}_n[\widetilde{g}_i\epsilon_i] + R_{2,n}(\alpha_{K_n}) = \widetilde{\alpha}_{K_n,\lambda}^\top \mathbb{G}_n[\widetilde{g}_i\zeta_{n,i}^{(\ell)}] + (1-\ell)R_{2,n}(\alpha_{K_n}),$$

we obtain

$$\begin{aligned}\sqrt{n}\frac{\alpha_{K_n}^\top(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})}{\sigma_{n,\ell}} =& \frac{\widetilde{\alpha}_{K_n,\lambda}^\top}{\sigma_{n,\ell}}\mathbb{G}_n[\widetilde{g}_i\zeta_{n,i}^{(\ell)}] + \frac{R_{1,n}(\alpha_{K_n})}{\sigma_{n,\ell}} \\ &+ \frac{(1-\ell)R_{2,n}(\alpha_{K_n})}{\sigma_{n,\ell}} + \frac{R_{\lambda,n}(\alpha_{K_n})}{\sigma_{n,\ell}}. \end{aligned} \tag{C.3}$$

Using the bounds in Lemma 4.5, condition (ii) yields

$$\frac{R_{1,n}(\alpha_{K_n})}{\sigma_{n,\ell}} = O_P\left(\frac{\|\widetilde{\alpha}_{K_n,\lambda}\|}{\sigma_{n,\ell}}\left\|\Gamma_{K_n,\lambda}^{-1/2}(\widehat{\Gamma}_{K_n,n} - \Gamma_{K_n})\Gamma_{K_n,\lambda}^{-1/2}\right\|\left(1 + \widetilde{\xi}_{K_n,\lambda}\Delta_{K_n}\right)\right) = o_P(1).$$

Similarly, condition (iii) gives

$$\frac{(1-\ell)R_{2,n}(\alpha_{K_n})}{\sigma_{n,\ell}} = O_P\left(\frac{(1-\ell)\|\widetilde{\alpha}_{K_n,\lambda}\|\widetilde{\xi}_{K_n,\lambda}\Delta_{K_n}}{\sigma_{n,\ell}}\right) = o_P(1).$$

Notice that this condition is vacuous when $\ell = 1$. Finally, condition (iv) yields

$$\frac{R_{\lambda,n}(\alpha_{K_n})}{\sigma_{n,\ell}} = O_P\left(\frac{\sqrt{n}\lambda_{K_n,n}\|\widetilde{\alpha}_{K_n,\lambda}\|\|\Gamma_{K_n,\lambda}^{-1/2}\beta_{K_n}\|}{\sigma_{n,\ell}}\right) = o_P(1).$$

Consequently,

$$\sqrt{n}\frac{\alpha_{K_n}^\top(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})}{\sigma_{n,\ell}} = \frac{\widetilde{\alpha}_{K_n,\lambda}^\top}{\sigma_{n,\ell}}\mathbb{G}_n[\widetilde{g}_i\zeta_{n,i}^{(\ell)}] + o_P(1). \tag{C.4}$$

Define

$$\omega_{n,i}^{(\ell)} := \frac{\widetilde{\alpha}_{K_n,\lambda}^\top\widetilde{g}_i}{\sqrt{n}\,\sigma_{n,\ell}}. \tag{C.5}$$

Then

$$\frac{\widetilde{\alpha}_{K_n,\lambda}^\top}{\sigma_{n,\ell}}\mathbb{G}_n[\widetilde{g}_i\zeta_{n,i}^{(\ell)}] = \sum_{i=1}^n \omega_{n,i}^{(\ell)}\zeta_{n,i}^{(\ell)}. \tag{C.6}$$

The summands have mean zero. Moreover, by the definition of $\sigma_{n,\ell}$,

$$\mathrm{Var}\left(\sum_{i=1}^n \omega_{n,i}^{(\ell)}\zeta_{n,i}^{(\ell)}\right) = \sum_{i=1}^n \mathbb{E}\left[\omega_{n,i}^{(\ell)2}\zeta_{n,i}^{(\ell)2}\right] = 1.$$

Condition (i) implies

$$\max_{1\le i\le n}|\omega_{n,i}^{(\ell)}| \le \frac{\|\widetilde{\alpha}_{K_n,\lambda}\|\widetilde{\xi}_{K_n,\lambda}}{\sqrt{n}\,\sigma_{n,\ell}} =: a_{n,\ell} = o(1). \tag{C.7}$$

We now verify the Lindeberg condition. For any $\delta > 0$,

$$\sum_{i=1}^{n} \mathbb{E}\left[\omega_{n,i}^{(\ell)2}\zeta_{n,i}^{(\ell)2}\mathbb{1}\{|\omega_{n,i}^{(\ell)}\zeta_{n,i}^{(\ell)}| > \delta\}\right] \le \sum_{i=1}^{n} \mathbb{E}\left[\omega_{n,i}^{(\ell)2}\zeta_{n,i}^{(\ell)2}\mathbb{1}\{|\zeta_{n,i}^{(\ell)}| > \delta/a_{n,\ell}\}\right]$$
$$\le \left(\sup_{x\in\mathcal{X}} \mathbb{E}\left[\zeta_{n,i}^{(\ell)2}\mathbb{1}\{|\zeta_{n,i}^{(\ell)}| > \delta/a_{n,\ell}\} \mid X_i = x\right]\right) \sum_{i=1}^{n} \mathbb{E}\left[\omega_{n,i}^{(\ell)2}\right].$$

Since $\mathbb{E}[\epsilon_i \mid X_i] = 0$ and $\ell \in \{0, 1\}$,

$$\mathbb{E}\left[\zeta_{n,i}^{(\ell)2} \mid X_i\right] = \mathbb{E}[\epsilon_i^2 \mid X_i] + \ell\eta_i^2 \ge \underline{\sigma}^2.$$

It follows that

$$\sum_{i=1}^{n} \mathbb{E}\left[\omega_{n,i}^{(\ell)2}\right] = \frac{\mathbb{E}[(\widetilde{\alpha}_{K_n,\lambda}^{\top}\widetilde{g}_i)^2]}{\sigma_{n,\ell}^2} \le \frac{1}{\underline{\sigma}^2}.$$

Therefore,

$$\sum_{i=1}^{n} \mathbb{E}\left[\omega_{n,i}^{(\ell)2}\zeta_{n,i}^{(\ell)2}\mathbb{1}\{|\omega_{n,i}^{(\ell)}\zeta_{n,i}^{(\ell)}| > \delta\}\right] \le \frac{1}{\underline{\sigma}^2} \sup_{x\in\mathcal{X}} \mathbb{E}\left[\zeta_{n,i}^{(\ell)2}\mathbb{1}\{|\zeta_{n,i}^{(\ell)}| > \delta/a_{n,\ell}\} \mid X_i = x\right].$$

Since $a_{n,\ell} \to 0$, condition (vi) implies that the right-hand side converges to zero. Hence, the Lindeberg condition holds, and the Lindeberg-Feller central limit theorem gives

$$\sum_{i=1}^{n} \omega_{n,i}^{(\ell)}\zeta_{n,i}^{(\ell)} \xrightarrow{d} \mathcal{N}(0, 1).$$

Combining this with (C.4) and applying Slutsky's theorem yields

$$\sqrt{n}\frac{\alpha_{K_n}^{\top}(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})}{\sigma_{K_n,\lambda}^{(\ell)}(\alpha_{K_n})} \xrightarrow{d} \mathcal{N}(0, 1),$$

as required. □

Proof of Corollary 4.3. We first verify conditions (i)-(vi) of Theorem 4.6 for $\ell = 0$ along the prescribed sequence α_{K_n}. We write

$$\sigma_{K_n,\lambda}^{2,(0)}(\alpha_{K_n}) = \widetilde{\alpha}_{K_n,\lambda}^{\top}\mathbb{E}\left[\epsilon_i^2\widetilde{g}_i\widetilde{g}_i^{\top}\right]\widetilde{\alpha}_{K_n,\lambda} \ge \underline{\sigma}^2\widetilde{\alpha}_{K_n,\lambda}^{\top}\Gamma_{K_n,\lambda}^{-1/2}\Gamma_{K_n}\Gamma_{K_n,\lambda}^{-1/2}\widetilde{\alpha}_{K_n,\lambda}, \tag{C.8}$$

where in the inequality we used that, by hypotheses, $\mathbb{E}[\epsilon_i^2 \mid X_i] \ge \underline{\sigma}^2 > 0$ almost surely. Using the definition of $\mathcal{R}_{K_n,\lambda}(\alpha_{K_n})$ in (4.32), we obtain

$$\frac{\|\widetilde{\alpha}_{K_n,\lambda}\|}{\sigma_{K_n,\lambda}^{(0)}(\alpha_{K_n})} \le \underline{\sigma}^{-1}\mathcal{R}_{K_n,\lambda}(\alpha_{K_n})^{-1/2} = O_{P_\theta}(1). \tag{C.9}$$

By (4.37), $K_n^{1/2+\beta/(d+1)}/\sqrt{n} = O(1)$. Hence, the second condition in (4.35) implies

$$\frac{\widetilde{\xi}_{K_n,\lambda}}{\sqrt{n}} = o_{P_\theta}(1).$$

Combining this with (C.9), we obtain

$$\frac{\|\widetilde{\alpha}_{K_n,\lambda}\|\widetilde{\xi}_{K_n,\lambda}}{\sqrt{n}\,\sigma_{K_n,\lambda}^{(0)}(\alpha_{K_n})} = o_{P_\theta}(1),$$

which verifies condition (i) in P_θ-probability.

Moreover, the second condition in (4.35) and (4.34) imply

$$\widetilde{\xi}_{K_n,\lambda}\Delta_{K_n} = o_{P_\theta}(1).$$

Together with (C.9) and the ridge-normalized empirical covariance matrix stability implied by Assumption 10, this verifies condition (ii) in P_θ-probability. Furthermore,

$$\frac{\|\widetilde{\alpha}_{K_n,\lambda}\|\widetilde{\xi}_{K_n,\lambda}\Delta_{K_n}}{\sigma^{(0)}_{K_n,\lambda}(\alpha_{K_n})} = o_{P_\theta}(1),$$

so condition (iii) is also satisfied. Finally, (C.9) together with (4.36) verifies condition (iv). Conditions (v) and (vi) for $\ell = 0$ are imposed directly in the hypotheses of the corollary.

Thus, the conditions of Theorem 4.6 hold in P_θ-probability for $\ell = 0$ and hence Theorem 4.6 yields (4.38) for $\ell = 0$. It remains only to show that replacing $\sigma^{(0)}_{K_n,\lambda}(\alpha_{K_n})$ by $\sigma^{(1)}_{K_n,\lambda}(\alpha_{K_n})$ does not affect the limit. Since $\mathbb{E}[\epsilon_i \mid X_i] = 0$,

$$\sigma^{2,(1)}_{K_n,\lambda}(\alpha_{K_n}) - \sigma^{2,(0)}_{K_n,\lambda}(\alpha_{K_n}) = \mathbb{E}\left[\eta_i^2\left(\widetilde{\alpha}^\top_{K_n,\lambda}\widetilde{g}_i\right)^2\right].$$

Hence

$$0 \le \frac{\sigma^{2,(1)}_{K_n,\lambda}(\alpha_{K_n}) - \sigma^{2,(0)}_{K_n,\lambda}(\alpha_{K_n})}{\sigma^{2,(0)}_{K_n,\lambda}(\alpha_{K_n})} \le \underline{\sigma}^{-2}\mathcal{R}_{K_n,\lambda}(\alpha_{K_n})^{-1}\widetilde{\xi}^2_{K_n,\lambda}\Delta^2_{K_n} = o_{P_\theta}(1),$$

which implies that

$$\frac{\sigma^{(1)}_{K_n,\lambda}(\alpha_{K_n})}{\sigma^{(0)}_{K_n,\lambda}(\alpha_{K_n})} = 1 + o_{P_\theta}(1). \tag{C.10}$$

Combining this with (4.38) for $\ell = 0$ and Slutsky's theorem gives the same conditional limit with $\sigma^{(1)}_{K_n,\lambda}(\alpha_{K_n})$ in the denominator. Hence (4.38) holds for both $\ell \in \{0, 1\}$. □

C.2 Ridge-adjusted variance estimation and studentization

Proof of Lemma 4.7. To show (i) holds, we develop the sum as

$$\frac{1}{n}\sum_{i=1}^n [\widehat{\epsilon}_i^2 - (\epsilon_i + \eta_i)^2]\widehat{r}_i^2 = \frac{1}{n}\sum_{i=1}^n \left[g_i^\top(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})\right]^2\widehat{r}_i^2 - \frac{2}{n}\sum_{i=1}^n(\epsilon_i + \eta_i)g_i^\top(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})\widehat{r}_i^2,$$

where we have used that $\widehat{\epsilon}_i = (\eta_i + \epsilon_i) - g_i^\top(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})$. We have

$$\begin{aligned}
\sup_{1\le i\le n}\widehat{r}_i^2 &= \sup_{1\le i\le n}\left|\alpha^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda}g_i\right|^2 \\
&= \sup_{1\le i\le n}\left\|\alpha^\top\Gamma^{-1/2}_{K_n,\lambda}(\Gamma^{1/2}_{K_n,\lambda}\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\Gamma^{1/2}_{K_n,\lambda})\Gamma^{-1/2}_{K_n,\lambda}g_i\right\|^2 \\
&\le \|\widetilde{\alpha}\|^2\|\widehat{D}_{K_n,n,\lambda}\|^2\sup_{1\le i\le n}\|\widetilde{g}_i\|^2 \le \|\widetilde{\alpha}\|^2\|\widehat{D}_{K_n,n,\lambda}\|^2\widetilde{\xi}^2_{K_n,\lambda},
\end{aligned} \tag{C.11}$$

with $\widehat{D}_{K_n,n,\lambda}$ defined as in (4.17). By the second condition in (4.39),

$$\frac{1}{n}\sum_{i=1}^n\left[g_i^\top(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n})\right]^2 = O_P\left(\overline{\mathfrak{M}}^2_{K_n,n,\lambda}\right), \tag{C.12}$$

Moreover, by the first condition in (4.39) and $\overline{a}_{K_n,n,\lambda} = o(1)$, $\|A_{K_n,n,\lambda}\| = o_P(1)$. Since $\widehat{D}_{K_n,n,\lambda} = (I_{K_n} + A_{K_n,n,\lambda})^{-1}$ it follows that $\|\widehat{D}_{K_n,n,\lambda}\| = O_P(1)$ (see also the proof of Theorem 4.1). Thus, combining bounds (C.11) and (C.12), we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left[g_i^\top(\widehat{\beta}_{K_n,n,\lambda}-\beta_{K_n})\right]^2\widehat{r}_i^2 &\le \left[\frac{1}{n}\sum_{i=1}^{n}\left[g_i^\top(\widehat{\beta}_{K_n,n,\lambda}-\beta_{K_n})\right]^2\right]\sup_{1\le i\le n}\widehat{r}_i^2\\ &= O_P\left(\|\widetilde{\alpha}\|^2\widetilde{\xi}^2_{K_n,\lambda}\overline{\mathfrak{M}}^2_{K_n,n,\lambda}\right)\end{aligned}$$

For the cross term, we use Cauchy-Schwarz:

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i+\eta_i)g_i^\top(\widehat{\beta}_{K_n,n,\lambda}-\beta_{K_n})\widehat{r}_i^2\right|\\ &\quad\le\left[\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i+\eta_i)^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\left[g_i^\top(\widehat{\beta}_{K_n,n,\lambda}-\beta_{K_n})\right]^2\right]^{1/2}\sup_{1\le i\le n}\widehat{r}_i^2.\end{aligned}$$

We now use that

$$\mathbb{E}\big[(\epsilon_i+\eta_i)^2\big] = \mathbb{E}[\epsilon_i^2]+\mathbb{E}[\eta_i^2] \le \overline{\sigma}^2+\Delta^2_{K_n},$$

since $\mathbb{E}[\epsilon_i\eta_i] = \mathbb{E}\big[\eta_i\,\mathbb{E}[\epsilon_i\,|\,X_i]\big] = 0$, with Markov's inequality then yielding

$$\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i+\eta_i)^2 = O_P\left(\overline{\sigma}^2+\Delta^2_{K_n}\right). \tag{C.13}$$

Hence, (C.11), (C.12), and (C.13) give

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i+\eta_i)g_i^\top(\widehat{\beta}_{K_n,n,\lambda}-\beta_{K_n})\widehat{r}_i^2\right| = O_P\left(\|\widetilde{\alpha}\|^2\widetilde{\xi}^2_{K_n,\lambda}\overline{\mathfrak{M}}_{K_n,n,\lambda}\sqrt{\overline{\sigma}^2+\Delta^2_{K_n}}\right).$$

By merging bounds, we then find

$$\left|\frac{1}{n}\sum_{i=1}^{n}[\widehat{\epsilon}_i^2-(\epsilon_i+\eta_i)^2]\widehat{r}_i^2\right| = O_P\left(\|\widetilde{\alpha}\|^2\widetilde{\xi}^2_{K_n,\lambda}\overline{\mathfrak{M}}_{K_n,n,\lambda}\left(\overline{\mathfrak{M}}_{K_n,n,\lambda}+\sqrt{\overline{\sigma}^2+\Delta^2_{K_n}}\right)\right),$$

which establishes (i).

To prove (ii), we first write

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i+\eta_i)^2\left[\widehat{r}_i^2-r_i^2\right]\right| \le \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i+\eta_i)^2\left|\widehat{r}_i^2-r_i^2\right|,$$

and note

$$\begin{aligned}\widehat{r}_i^2 &= \big(\widetilde{\alpha}^\top\widehat{D}_{K_n,n,\lambda}\widetilde{g}_i\big)^2\\ &= \big(\widetilde{\alpha}^\top\widetilde{g}_i\big)^2+\big(\widetilde{\alpha}^\top\big(\widehat{D}_{K_n,n,\lambda}-I_{K_n}\big)\widetilde{g}_i\big)^2+2\widetilde{\alpha}^\top\big(\widehat{D}_{K_n,n,\lambda}-I_{K_n}\big)\widetilde{g}_i\,\widetilde{\alpha}^\top\widetilde{g}_i.\end{aligned}$$

Since $r_i^2 = \big(\widetilde{\alpha}^\top\widetilde{g}_i\big)^2$, it readily follows

$$\begin{aligned}\left|\widehat{r}_i^2-r_i^2\right| &\le \|\widetilde{\alpha}\|^2\|\widetilde{g}_i\|^2\|\widehat{D}_{K_n,n,\lambda}-I_{K_n}\|^2+2\|\widetilde{\alpha}\|^2\|\widetilde{g}_i\|^2\|\widehat{D}_{K_n,n,\lambda}-I_{K_n}\|\\ &\le \|\widetilde{\alpha}\|^2\widetilde{\xi}^2_{K_n,\lambda}\big(\|\widehat{D}_{K_n,n,\lambda}-I_{K_n}\|^2+2\|\widehat{D}_{K_n,n,\lambda}-I_{K_n}\|\big)\end{aligned} \tag{C.14}$$

Combining (C.14) and (C.13) we find

$$\left|\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i+\eta_i)^2\left[\widehat{r}_i^2-r_i^2\right]\right| = O_P\left(\|\widetilde{\alpha}\|^2\widetilde{\xi}^2_{K_n,\lambda}\bar{a}_{K_n,n,\lambda}\big(\overline{\sigma}^2+\Delta^2_{K_n}\big)\right).$$

In the last expression we used that, since $\|A_{K_n,n,\lambda}\| = O_P(\bar{a}_{K_n,n,\lambda})$ by (4.39), $\bar{a}_{K_n,n,\lambda} = o(1)$ by the hypotheses of the lemma, and $\widehat{D}_{K_n,n,\lambda} = (I_{K_n} + A_{K_n,n,\lambda})^{-1}$, we further have $\|\widehat{D}_{K_n,n,\lambda} - I_{K_n}\| \leq \|\widehat{D}_{K_n,n,\lambda}\|\|A_{K_n,n,\lambda}\| = O_P(\bar{a}_{K_n,n,\lambda})$ and hence $\|\widehat{D}_{K_n,n,\lambda} - I_{K_n}\|^2 + 2\|\widehat{D}_{K_n,n,\lambda} - I_{K_n}\| = O_P(\bar{a}_{K_n,n,\lambda})$. □

Proof of Proposition 4.2. To streamline notation, we drop some subscripts and write $\sigma_\lambda^2 := \sigma^{2,(1)}_{K_n,\lambda}(\alpha)$. By (4.31) and by the definition of r_i in Lemma 4.7, $\sigma_\lambda^2 = \mathbb{E}\left[(\epsilon_i + \eta_i)^2 r_i^2\right]$. Accordingly, by the definition of $\widehat{r}_i$ in Lemma 4.7, $\widehat{\sigma}_\lambda^2 = \widehat{\sigma}^2_{K_n,n,\lambda}(\alpha) = \frac{1}{n}\sum_{i=1}^n \widehat{\epsilon}_i^2 \widehat{r}_i^2$. Dividing by σ_λ^2 and then adding and subtracting, we can then find

$$\begin{aligned}\frac{\widehat{\sigma}_\lambda^2}{\sigma_\lambda^2} &= \frac{1}{n\sigma_\lambda^2}\sum_{i=1}^n \left[\widehat{\epsilon}_i^2 - (\epsilon_i + \eta_i)^2\right]\widehat{r}_i^2 + \frac{1}{n\sigma_\lambda^2}\sum_{i=1}^n (\epsilon_i + \eta_i)^2\left(\widehat{r}_i^2 - r_i^2\right) \\ &\quad + \frac{1}{n\sigma_\lambda^2}\sum_{i=1}^n (\epsilon_i + \eta_i)^2 r_i^2 \\ &=: E_{1,n}(\alpha) + E_{2,n}(\alpha) + E_{3,n}(\alpha).\end{aligned} \tag{C.15}$$

We now proceed to show that both $E_{1,n}(\alpha)$ and $E_{2,n}(\alpha)$ are asymptotically negligible, and that $E_{3,n}(\alpha) = 1 + o_P(1)$.

We first consider $E_{1,n}(\alpha)$. Using the definitions in (4.40), Lemma 4.7 (i), and hypothesis (i),

$$|E_{1,n}(\alpha)| = O_P\left(\mathcal{L}^{(1)}_{K_n,\lambda}(\alpha)\overline{\mathfrak{M}}_{K_n,n,\lambda}\left(\overline{\mathfrak{M}}_{K_n,n,\lambda} + \sqrt{\overline{\sigma}^2 + \Delta^2_{K_n}}\right)\right) = o_P(1), \tag{C.16}$$

by the first negligibility condition.

We next control $E_{2,n}(\alpha)$. Again, by the definitions in (4.40), Lemma 4.7 (ii), and hypothesis (ii),

$$|E_{2,n}(\alpha)| = O_P\left(\mathcal{L}^{(1)}_{K_n,\lambda}(\alpha)\overline{a}_{K_n,n,\lambda}\left(\overline{\sigma}^2 + \Delta^2_{K_n}\right)\right) = o_P(1). \tag{C.17}$$

It remains to control $E_{3,n}(\alpha)$. Define $q_i := (\epsilon_i + \eta_i)^2 r_i^2$. By the definition of σ_λ^2, $\mathbb{E}[q_i] = \sigma_\lambda^2$. Therefore,

$$E_{3,n}(\alpha) - 1 = \frac{1}{n\sigma_\lambda^2}\sum_{i=1}^n (q_i - \mathbb{E}[q_i]). \tag{C.18}$$

By independence,

$$\begin{aligned}\mathrm{Var}\left(E_{3,n}(\alpha) - 1\right) &= \frac{1}{n\sigma_\lambda^4}\mathrm{Var}(q_i) \leq \frac{1}{n\sigma_\lambda^4}\mathbb{E}\left[(\epsilon_i + \eta_i)^4 r_i^4\right] \\ &\leq \frac{1}{n\sigma_\lambda^4}\overline{\sigma}^4_{\epsilon,\eta}\|\widetilde{\alpha}\|^4\widetilde{\xi}^4_{K_n,\lambda} = \frac{\overline{\sigma}^4_{\epsilon,\eta}}{n}\left(\mathcal{L}^{(1)}_{K_n,\lambda}(\alpha)\right)^2,\end{aligned}$$

where in the last inequality we used hypothesis (iv) and in the last equality we applied the definition in (4.40).

By Chebyshev's inequality and hypothesis (iii),

$$E_{3,n}(\alpha) - 1 = O_P\left(\frac{\mathcal{L}^{(1)}_{K_n,\lambda}(\alpha)}{\sqrt{n}}\right) = o_P(1), \tag{C.19}$$

Combining the bounds (C.16), (C.17), and (C.19) in (C.15) concludes the proof. □

Proof of Corollary 4.4. We verify the conditions of Proposition 4.2.

By Corollary E.1,

$$\Delta_{K_n} = O_{P_\theta}\left(K_n^{-1/2-\beta/(d+1)}\right). \tag{C.20}$$

Moreover, under (4.41), Theorem 4.4 gives

$$\mathcal{N}_{K_n,\lambda} = O_{P_\theta}\left(K_n^{\nu_r}\right), \qquad \nu_r = \frac{(d+1)(1+r)}{d+3}. \tag{C.21}$$

Since $r < 2/(d+1)$, $1-\nu_r = \frac{2-(d+1)r}{d+3} > 0$, and hence $\nu_r < 1$.

Additionally, Lemma 4.3 yields

$$B^{\text{pop}}_{K_n,\lambda} = O_{P_\theta}\left(K_n^{-1/2-\beta/(d+1)} + \sqrt{\frac{\lambda_{K_n,n}\log(K_n)}{K_n}}\right).$$

Again, with the choice of $\lambda_{K_n,n}$ in (4.41),

$$\frac{\sqrt{\lambda_{K_n,n}\log(K_n)/K_n}}{K_n^{-1/2-\beta/(d+1)}} \asymp K_n^{\beta/(d+1)-r/2}\sqrt{\log(K_n)} \to 0,$$

using the lower bound $r > 2\beta/(d+1)$ in (4.41). Consequently,

$$B^{\text{pop}}_{K_n,\lambda} = O_{P_\theta}\left(K_n^{-1/2-\beta/(d+1)}\right). \tag{C.22}$$

By Lemma 4.2,

$$\left\|\widehat{\Gamma}^{1/2}_{K_n,n}\left(\widehat{\beta}_{K_n,n,\lambda} - \beta_{K_n}\right)\right\| = O_P\left(\Delta_{K_n} + B^{\text{pop}}_{K_n,\lambda} + \sqrt{\frac{\mathcal{N}_{K_n,\lambda}}{n}}\right) = O_P\left(\overline{\mathfrak{M}}_{K_n,n,\lambda}\right),$$

where in the second equality we used (C.20), (C.21), and (C.22). This renders the second condition in (4.39) in Lemma 4.7 satisfied.

Next, by (4.16) in Remark 11, Assumption 10 gives

$$\|A_{K_n,n,\lambda}\| = O_P\left(\sqrt{\frac{\widetilde{\xi}^2_{K_n,\lambda}\log(K_n)}{n}}\right).$$

Since

$$\widetilde{\xi}^2_{K_n,\lambda} \le \frac{\xi^2_{K_n}}{\overline{s}^2_{K_n} + \lambda_{K_n,n}} \le \frac{\xi^2_{K_n}}{\lambda_{K_n,n}},$$

with $\xi^2_{K_n} = O_{P_\theta}(K_n)$, and $\lambda_{K_n,n} \asymp K_n^{-r}$ by (4.41), we obtain

$$\widetilde{\xi}^2_{K_n,\lambda} = O_{P_\theta}(K_n^{1+r}).$$

Hence

$$\|A_{K_n,n,\lambda}\| = O_P\left(\sqrt{\frac{K_n^{1+r}\log(K_n)}{n}}\right) = O_P\left(\overline{a}_{K_n,n,\lambda}\right). \tag{C.23}$$

Moreover, since $\rho \ge 0$ by the hypotheses of the corollary, (4.44) implies $\overline{a}_{K_n,n,\lambda} = o(1)$. Hence, the first condition in (4.39) in Lemma 4.7 is also satisfied.

It remains to verify conditions (i)-(iii) of Proposition 4.2. By (4.43), it holds that

$$\mathcal{L}^{(1)}_{K_n,\lambda}(\alpha)\overline{a}_{K_n,n,\lambda} = O_{P_\theta}\left(\sqrt{\frac{K_n^{2\rho+1+r}\log(K_n)}{n}}\right) = o_{P_\theta}(1)$$

where the last equality follows from (4.44). Next,

$$\mathcal{L}^{(1)}_{K_n,\lambda}(\alpha)\overline{\mathfrak{M}}_{K_n,n,\lambda} = O_{P_\theta}\left(\sqrt{\frac{K_n^{2\rho+\nu_r}}{n}} + K_n^{\rho-1/2-\beta/(d+1)}\right) = o_{P_\theta}(1), \tag{C.24}$$

where to get the last equality we used that since $\nu_r < 1+r$ by (4.42), (4.44) implies $K_n^{2\rho+\nu_r}/n = o(1)$, while $\rho < 1/2+\beta/(d+1)$ implies $K_n^{\rho-1/2-\beta/(d+1)} \to 0$. Also, (4.44) implies $K_n^{2\rho}/n = o(1)$, and hence

$$\frac{\mathcal{L}^{(1)}_{K_n,\lambda}(\alpha)}{\sqrt{n}} = o_{P_\theta}(1). \tag{C.25}$$

Finally, (C.20) implies $\Delta_{K_n} = o_{P_\theta}(1)$, and therefore $\sqrt{\overline{\sigma}^2 + \Delta^2_{K_n}} = O_{P_\theta}(1)$ which yields $\overline{\sigma}^2 + \Delta^2_{K_n} = O_{P_\theta}(1)$.

Together with (C.24), this verifies condition (i) of Proposition 4.2. At the same time, (C.23) and (4.44) verify condition (ii), while (C.25) verifies condition (iii). Condition (iv) is exactly (4.45). Hence all the conditions of Proposition 4.2 are satisfied, which proves the claim, as required. □

Proof of Corollary 4.5. Since $\mathbb{E}[\epsilon_i \mid X_i] = 0$, $\sigma^{2,(1)}_{K_n,\lambda}(\alpha) - \sigma^{2,(0)}_{K_n,\lambda}(\alpha) = \mathbb{E}\left[\eta_i^2 r_i^2\right]$. Hence

$$0 \le \frac{\sigma^{2,(1)}_{K_n,\lambda}(\alpha) - \sigma^{2,(0)}_{K_n,\lambda}(\alpha)}{\sigma^{2,(1)}_{K_n,\lambda}(\alpha)} \le \frac{\|\widetilde{\alpha}\|^2 \widetilde{\xi}^2_{K_n,\lambda}\Delta^2_{K_n}}{\sigma^{2,(1)}_{K_n,\lambda}(\alpha)} = \mathcal{L}^{(1)}_{K_n,\lambda}(\alpha)\Delta^2_{K_n}.$$

By Corollary E.1 and (4.43), $\mathcal{L}^{(1)}_{K_n,\lambda}(\alpha)\Delta^2_{K_n} = O_{P_\theta}\left(K_n^{\rho-1-2\beta/(d+1)}\right) = o_{P_\theta}(1)$. Hence, $\sigma^{2,(1)}_{K_n,\lambda}(\alpha)/\sigma^{2,(0)}_{K_n,\lambda}(\alpha) \xrightarrow{p} 1$. Combining this with Corollary 4.4 and Slutsky's theorem proves the claim. □

C.3 Smooth-functional inference

Proof of Corollary 4.6. By (4.23), one has $\|\widehat{h}_{n,\lambda} - h\|_{L^2(X)} = O_P\big(\sqrt{K_n^{\nu_r}/n} + K_n^{-1/2-\beta/(d+1)}\big)$, where $\nu_r = \frac{(d+1)(1+r)}{d+3} < 1$. Let $b := \beta/(d+1) > 0$. Under (4.51), $K_n \asymp n^{1/(2(1+b))}$, and therefore

$$\sqrt{\frac{K_n^{\nu_r}}{n}} \asymp n^{-\frac{1}{2}+\frac{\nu_r}{4(1+b)}},$$
$$K_n^{-\frac{1}{2}-b} \asymp n^{-\frac{1+2b}{4(1+b)}}.$$

Since $\nu_r < 1$ and $b > 0$, $\frac{1}{2} - \frac{\nu_r}{4(1+b)} > \frac{1}{4}$, $\frac{1+2b}{4(1+b)} > \frac{1}{4}$. Hence both terms are $o(n^{-1/4})$, which proves (4.52). Since $\omega \ge 2$, it thus holds $\sqrt{n}\|\widehat{h}_{n,\lambda} - h\|^{\omega}_{L^2(X)} = o_P(1)$, and the conclusion follows from (4.50). □

Proof of Corollary 4.7. By (C.10) in the proof of Corollary 4.3 and by (4.57), it follows that

$$\frac{\sigma^{(1)}_{\Psi,K_n,\lambda}}{\sigma^{(0)}_{\Psi,K_n,\lambda}} = 1 + o_{P_\theta}(1).$$

Therefore, (4.58) implies

$$\left(\sigma^{(\ell)}_{\Psi,K_n,\lambda}\right)^{-1} = O_{P_\theta}(1), \qquad \ell \in \{0,1\}. \tag{C.26}$$

By (4.55), (4.54), (4.59), and (C.26),

$$\sqrt{n}\frac{\Psi(\widehat{h}_{n,\lambda})-\Psi(h)}{\sigma^{(\ell)}_{\Psi,K_n,\lambda}}=\sqrt{n}\frac{a_{K_n}^\top\big(\widehat{\beta}_{K_n,n,\lambda}-\beta_{K_n}\big)}{\sigma^{(\ell)}_{\Psi,K_n,\lambda}}+o_P(1),\qquad \ell\in\{0,1\}. \tag{C.27}$$

Then, (4.57) and (4.56) allow us to write

$$\sqrt{n}\frac{a_{K_n}^\top\big(\widehat{\beta}_{K_n,n,\lambda}-\beta_{K_n}\big)}{\sigma^{(\ell)}_{\Psi,K_n,\lambda}}=\sqrt{n}\frac{(\mathfrak{a}^\Psi_{K_n})^\top\big(\widehat{\beta}_{K_n,n,\lambda}-\beta_{K_n}\big)}{\sigma^{(\ell)}_{K_n,\lambda}(\mathfrak{a}^\Psi_{K_n})}.$$

Hence, applying Corollary 4.3, we find that the right-hand side of the last display converges in distribution to $\mathcal{N}(0,1)$, conditionally on $\mathcal{F}_{K_n}$, in P_θ-probability. Combining this with (C.27) and Slutsky's theorem proves (4.60). □

C.4 Proof of Example 1

Proof. We first verify the approximation assumptions. Since $Ax+B\geq 1$ for every $(A,B,x)\in[-1,1]\times[2,3]\times[-1,1]$, the ReLU activation satisfies $\sigma(Ax+B)=Ax+B$ throughout the relevant domains. Using independence of A and B, $\mathbb{E}[A]=0$, and $\mathbb{E}[A^2]=1/3$, we obtain

$$\mathbb{E}_{A,B}[3A(Ax+B)_+]=3x\mathbb{E}[A^2]+3\mathbb{E}[A]\mathbb{E}[B]=x=h(x).$$

The function $f(A,B)=3A$ is Lipschitz and hence β-Hölder with $\beta=1/2$ on the compact parameter space $\Pi=[-1,1]\times[2,3]$. Moreover, the uniform distribution on this rectangle satisfies the regular sampling assumptions for the Euclidean metric, and the nearest-neighbor indegree is uniformly bounded. Hence the assumptions of Corollary E.1 are satisfied.

Write $u_K=(A_1,\dots,A_K)^\top$, $v_K=(B_1,\dots,B_K)^\top$, $S_K=[u_K\ \ v_K]$, and $Q=\mathrm{diag}(1/3,1)$. Then $g_K(x)=S_K(x,1)^\top$ and $\Gamma_K=S_KQS_K^\top$. By the strong law of large numbers,

$$\frac{1}{K}S_K^\top S_K\longrightarrow\begin{pmatrix}1/3 & 0\\ 0 & 19/3\end{pmatrix},\qquad P_\theta\text{-a.s.}$$

Consequently, $\mathrm{rank}(S_K)=2$ a.s. for every $K\geq 2$, and Γ_K has exactly two positive eigenvalues. The nonzero eigenvalues of Γ_K coincide with those of $Q^{1/2}S_K^\top S_KQ^{1/2}$, so both are of order K, P_θ-almost surely. Therefore $\mathrm{rank}(\Gamma_K)=2$ and $\lambda_{\min}(\Gamma_K)=0$ for every $K\geq 3$, almost surely.

For $K\geq 2$, the regression function $h(x)=x$ belongs to the realized RWNN sieve. Its minimum-Euclidean-norm projection coefficient is

$$\beta_K=S_K(S_K^\top S_K)^{-1}(1,0)^\top,$$

and hence $\Delta_K=0$ and $\eta_K(x)=0$ for all $x\in\mathcal{X}$. Since $\|S_K\|=O_{P_\theta}(\sqrt{K})$ and $\|(S_K^\top S_K)^{-1}\|=O_{P_\theta}(K^{-1})$, it follows that $\|\beta_K\|=O_{P_\theta}(K^{-1/2})$. Moreover, $\beta_K\in\mathrm{Range}(\Gamma_K)$, and the positive eigenvalues of Γ_K are of order K. Thus, uniformly for $\lambda>0$,

$$\|\Gamma_{K,\lambda}^{-1/2}\beta_K\|=O_{P_\theta}(K^{-1}).$$

The same spectral bounds yield

$$\widetilde{\xi}_{K,\lambda}=\sup_{x\in[-1,1]}\|\Gamma_{K,\lambda}^{-1/2}g_K(x)\|=O_{P_\theta}(1),$$

since $g_K(x)\in\mathrm{Range}(\Gamma_K)$ and $\sup_{x\in[-1,1]}\|g_K(x)\|=O_{P_\theta}(\sqrt{K})$.

For every nonzero $\alpha_K\in\mathrm{Range}(\Gamma_K)$, the spectral representation (4.33) implies

$$\frac{\lambda^+_{\min}(\Gamma_K)}{\lambda^+_{\min}(\Gamma_K)+\lambda}\leq\mathcal{R}_{K,\lambda}(\alpha_K)\leq 1,$$

where $\lambda^+_{\min}(\Gamma_K)$ denotes the smallest strictly positive eigenvalue of Γ_K. Since $\lambda^+_{\min}(\Gamma_K) \asymp K$, the Rayleigh quotient converges to one whenever $\lambda/K \to 0$. In particular, this applies to the normalized point evaluation direction $\alpha_K = g_K(x_0)/\|g_K(x_0)\|$ for every fixed $x_0 \in [-1,1]$. It also applies to the normalized derivative direction associated with Ψ_2, since

$$\Psi'_{2,h}(g_K) = 2\int_{-1}^{1} x g_K(x)\,\mathrm{d}x = \frac{4}{3}u_K \;\in\; \mathrm{Range}(\Gamma_K).$$

Now choose $K_n = \lceil n^{2/5} \rceil$ and $\lambda_{K_n,n} = K_n^{-3/4}$. With $d = 1$, $\beta = 1/2$, and $r = 3/4$, the polynomial-ridge restriction (4.41) is satisfied since $1/2 < 3/4 < 1$. Furthermore,

$$\frac{K_n^{1+2\beta/(d+1)}}{n} = \frac{K_n^{3/2}}{n} = O(n^{-2/5}) = o(1) \quad \text{and} \quad \sqrt{n}\,\lambda_{K_n,n}\|\Gamma^{-1/2}_{K_n,\lambda}\beta_{K_n}\| = O_{P_\theta}(n^{-1/5}) = o_{P_\theta}(1).$$

The normalized feature condition in (4.35) follows from $\widetilde{\xi}_{K_n,\lambda} = O_{P_\theta}(1)$, while the directional Rayleigh condition follows from the preceding spectral argument. Moreover,

$$\frac{\widetilde{\xi}^2_{K_n,\lambda}\log K_n}{n} = O_{P_\theta}\left(\frac{\log K_n}{n}\right) = o_{P_\theta}(1).$$

Thus Assumption 10 holds in P_θ-probability. The regression errors satisfy the required second moment and uniform integrability conditions as they are independent Gaussian random variables with variance $\tau^2 > 0$. This verifies the conditions of Corollary 4.3 along both prescribed directions.

We next verify the conditions for feasible inference. Since $\eta_{K_n} = 0$ for all sufficiently large n and $\mathbb{E}[\epsilon_i^2 \mid X_i] = \tau^2$, for every nonzero direction $\alpha_K \in \mathrm{Range}(\Gamma_K)$ we have

$$\mathcal{R}^{(1)}_{K,\lambda}(\alpha_K) = \tau^2\mathcal{R}_{K,\lambda}(\alpha_K) \quad \text{and} \quad \mathcal{L}^{(1)}_{K,\lambda}(\alpha_K) = \frac{\widetilde{\xi}^2_{K,\lambda}}{\tau^2\mathcal{R}_{K,\lambda}(\alpha_K)} = O_{P_\theta}(1).$$

Hence (4.43) holds with $\rho = 0$. The corresponding rate condition is

$$\frac{K_n^{2\rho+1+r}\log K_n}{n} = \frac{K_n^{7/4}\log K_n}{n} = O(n^{-3/10}\log n) = o(1).$$

The conditional fourth moment condition (4.45) follows from Gaussianity and $\eta_{K_n} = 0$ for $K_n \geq 2$; the finitely many initial values are uniformly bounded since the feature parameters lie in the compact rectangle and the population projection is bounded. Consequently, Corollaries 4.4 and 4.5 apply to both normalized directions.

It remains to identify the limiting distributions. Let $q(x) = (x,1)^\top$, $\vartheta_0 = (1,0)^\top$, $T_K = S_K^\top S_K$, and $\widehat{Q}_n = n^{-1}\sum_{i=1}^n q(X_i)q(X_i)^\top$. Since every function in the realized sieve has the form $x \mapsto q(x)^\top\vartheta$, the minimum-Euclidean-norm coefficient vector representing ϑ is $S_K T_K^{-1}\vartheta$. Therefore the original ridge regression problem is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^n \big(Y_i - q(X_i)^\top\vartheta\big)^2 + \lambda_{K_n,n}\vartheta^\top T_{K_n}^{-1}\vartheta$$

over $\vartheta \in \mathbb{R}^2$. Its minimizer satisfies

$$\widehat{\vartheta}_n = (\widehat{Q}_n + \lambda_{K_n,n}T_{K_n}^{-1})^{-1}\left(\frac{1}{n}\sum_{i=1}^n q(X_i)Y_i\right),$$

and $\widehat{h}_{n,\lambda}(x) = q(x)^\top\widehat{\vartheta}_n$. Since $\|T_{K_n}^{-1}\| = O_{P_\theta}(K_n^{-1})$, we have

$$\sqrt{n}\,\|\lambda_{K_n,n}T_{K_n}^{-1}\| = O_{P_\theta}(n^{-1/5}) = o_{P_\theta}(1).$$

Using $Y_i = q(X_i)^\top \vartheta_0 + \epsilon_i$, we obtain the exact decomposition

$$\sqrt{n}(\widehat{\vartheta}_n - \vartheta_0) = (\widehat{Q}_n + \lambda_{K_n,n} T_{K_n}^{-1})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n q(X_i)\epsilon_i - \sqrt{n}(\widehat{Q}_n + \lambda_{K_n,n} T_{K_n}^{-1})^{-1} \lambda_{K_n,n} T_{K_n}^{-1} \vartheta_0.$$

By the law of large numbers, $\widehat{Q}_n \xrightarrow{p} Q = \mathrm{diag}(1/3, 1)$. The second term is $o_P(1)$ conditionally on the features, while the multivariate central limit theorem gives

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n q(X_i)\epsilon_i \xrightarrow{d} \mathcal{N}(0, \tau^2 Q).$$

Consequently, conditionally on $\mathcal{F}_{K_n}$, in P_θ-probability,

$$\sqrt{n}(\widehat{\vartheta}_n - \vartheta_0) \xrightarrow{d} \mathcal{N}(0, \tau^2 Q^{-1}), \qquad Q^{-1} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

Since $\widehat{h}_{n,\lambda}(x_0) = q(x_0)^\top \widehat{\vartheta}_n$, it follows that

$$\sqrt{n}\big(\widehat{h}_{n,\lambda}(x_0) - x_0\big) \xrightarrow{d} \mathcal{N}\big(0, \tau^2(1 + 3x_0^2)\big).$$

Finally, writing $H = \mathrm{diag}(2/3, 2)$, we have $\Psi_2(\widehat{h}_{n,\lambda}) = \widehat{\vartheta}_n^\top H \widehat{\vartheta}_n$ and $\Psi_2(h) = \vartheta_0^\top H \vartheta_0 = 2/3$. The derivative at ϑ_0 is $2H\vartheta_0 = (4/3, 0)^\top$. The delta method therefore gives

$$\sqrt{n}\left(\Psi_2(\widehat{h}_{n,\lambda}) - \frac{2}{3}\right) \xrightarrow{d} \mathcal{N}\left(0, (2H\vartheta_0)^\top \tau^2 Q^{-1} (2H\vartheta_0)\right) = \mathcal{N}\left(0, \frac{16\tau^2}{3}\right).$$

The same conclusion follows from Corollary 4.7: the quadratic functional satisfies Assumption 12 with $\omega = 2$, its sieve approximation bias vanishes, its nonlinear remainder is negligible since $\|\widehat{h}_{n,\lambda} - h\|_{L^2(X)} = O_P(n^{-1/2})$, and its asymptotic variance converges to $16\tau^2/3 > 0$.

For completeness, we note that the feasible variance estimators are also consistent for the unnormalized point evaluation and smooth functional limits. Indeed, let $\widehat{\epsilon}_i = Y_i - q(X_i)^\top \widehat{\vartheta}_n$ and $D_n = \lambda_{K_n,n} T_{K_n}^{-1}$. For either $c_n = q(x_0)$ or $c_n = 2H\widehat{\vartheta}_n$, the corresponding ridge-adjusted sandwich estimator can be written as

$$\widehat{V}_n(c_n) = \frac{1}{n} \sum_{i=1}^n \widehat{\epsilon}_i^2 \left[c_n^\top (\widehat{Q}_n + D_n)^{-1} q(X_i)\right]^2.$$

Since $\widehat{\vartheta}_n \xrightarrow{p} \vartheta_0$, $D_n = o_{P_\theta}(1)$, and $n^{-1}\sum_{i=1}^n \widehat{\epsilon}_i^2 q(X_i) q(X_i)^\top \xrightarrow{p} \tau^2 Q$, it follows that

$$\widehat{V}_n(q(x_0)) \xrightarrow{p} \tau^2(1 + 3x_0^2), \qquad \widehat{V}_n(2H\widehat{\vartheta}_n) \xrightarrow{p} \frac{16\tau^2}{3}.$$

All convergence statements above hold conditionally on $\mathcal{F}_{K_n}$, in P_θ-probability, as claimed. □

D Proofs for sup-norm ridge estimation

D.1 Uniform stochastic error

Proof of Lemma 4.8. Step 1. First, we define $\check{h}_{n,\lambda} := \widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}$ and observe

$$\begin{aligned}\check{h}_{n,\lambda}(x) &= g_{K_n}(x)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} \left(n^{-1} G_{K_n,n}^\top e\right) \\ &= g_{K_n}(x)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} \left(n^{-1/2} V_{m_n,n} S_{m_n,n} U_{m_n,n}^\top e\right) \\ &= n^{-1/2} g_{K_n}(x)^\top V_{m_n,n} \operatorname{diag}\left(\left(\frac{s_{k,n}}{s_{k,n}^2 + \lambda_{K_n,n}}\right)_{k=1}^{m_n}\right) U_{m_n,n}^\top e,\end{aligned} \tag{D.1}$$

where in the first equality we used the definitions of $\widehat{h}_{n,\lambda}$ and $\widetilde{h}_{n,\lambda}$ in (4.61) and (4.62), respectively, together with $Y - H = e$, while in the second one we applied the singular value decomposition in (4.65). From the second equality in (D.1), we see that

$$\left\|\widehat{\Gamma}_{K_n,n,\lambda}^{-1} \left(n^{-1} G_{K_n,n}^\top e\right)\right\| \le \psi_{K_n,n,\lambda} \left\|n^{-1/2} U_{m_n,n}^\top e\right\| = \psi_{K_n,n,\lambda} \widetilde{U}_n, \tag{D.2}$$

where $\widetilde{U}_n := \left\|n^{-1/2} U_{m_n,n}^\top e\right\|$. Notice that, $U_{m_n,n}$ is measurable with respect to $X_{1:n} := (X_1, \ldots, X_n)$ and $\theta_{1:K_n} = (\theta_1, \ldots, \theta_{K_n})$. Since the random features are independent of the observations, Assumption 11 and the i.i.d. hypothesis imply

$$\begin{aligned}\mathbb{E}\left[\widetilde{U}_n^2 \middle| X_{1:n}, \theta_{1:K_n}\right] &= \frac{1}{n} \operatorname{tr}\left(U_{m_n,n}^\top \mathbb{E}\left[ee^\top \mid X_{1:n}\right] U_{m_n,n}\right) \\ &\le \frac{\overline{\sigma}^2}{n} \operatorname{tr}\left(U_{m_n,n}^\top U_{m_n,n}\right) = \overline{\sigma}^2 \frac{m_n}{n},\end{aligned} \tag{D.3}$$

which, by conditional Markov inequality applied to random variable $\widetilde{U}_n^2$, followed by the tower property, together with using the fact that the columns of $U_{m_n,n}$ are orthonormal vectors, implies $\widetilde{U}_n = O_P\left(\sqrt{m_n/n}\right)$.

Further, for any $x_1, x_2 \in \mathcal{X}$, Assumption 14 (i)-(ii), together with (D.2), give

$$\begin{aligned}|\check{h}_{n,\lambda}(x_1) - \check{h}_{n,\lambda}(x_2)| &\le \|g_{K_n}(x_1) - g_{K_n}(x_2)\| \left\|\widehat{\Gamma}_{K_n,n,\lambda}^{-1} \left(n^{-1} G_{K_n,n}^\top e\right)\right\| \\ &\le L_{K_n} \|x_1 - x_2\| \psi_{K_n,n,\lambda} \widetilde{U}_n.\end{aligned} \tag{D.4}$$

This implies that the sample paths of $\check{h}_{n,\lambda}$ are Lipschitz on $\mathcal{X}$, with a random Lipschitz coefficient controlled by $\psi_{K_n,n,\lambda}\widetilde{U}_n$.

Step 2. We now define

$$t_n := \zeta_{K_n} \overline{\psi}_{K_n,n} \sqrt{\frac{\log(n)}{n}}. \tag{D.5}$$

Since $\psi_{K_n,n,\lambda} = O_P(\overline{\psi}_{K_n,n})$ by Assumption 15 and $\widetilde{U}_n = O_P\left(\sqrt{m_n/n}\right)$, for every $\eta > 0$ there exists a constant $M_\eta > 0$ such that, for all sufficiently large n,

$$\mathbb{P}\left(\psi_{K_n,n,\lambda} > M_\eta \overline{\psi}_{K_n,n}\right) \le \frac{\eta}{2} \quad \text{and} \quad \mathbb{P}\left(\widetilde{U}_n > M_\eta \sqrt{\frac{m_n}{n}}\right) \le \frac{\eta}{2}.$$

Next, define $\mathcal{A}_{n,\eta} := \left\{\psi_{K_n,n,\lambda} \le M_\eta \overline{\psi}_{K_n,n}\right\}$ and $\mathcal{B}_{n,\eta} := \mathcal{A}_{n,\eta} \cap \left\{\widetilde{U}_n \le M_\eta \sqrt{m_n/n}\right\}$. Then one gets that $\mathbb{P}\left(\mathcal{B}_{n,\eta}^\complement\right) \le \eta$ for all sufficiently large n. Moreover, on the event $\mathcal{B}_{n,\eta}$, (D.4) implies the following bound:

$$|\check{h}_{n,\lambda}(x_1) - \check{h}_{n,\lambda}(x_2)| \le M_\eta^2 L_{K_n} \overline{\psi}_{K_n,n} \sqrt{\frac{m_n}{n}} \|x_1 - x_2\|. \tag{D.6}$$

Let $\mathcal{C}_n = \{x_{1,n}, \ldots, x_{N_n,n}\}$ be a minimal ρ_n-net of $\mathcal{X}$ with respect to the Euclidean norm, with $N_n := |\mathcal{C}_n|$ and covering radius $\rho_n > 0$. We choose ρ_n such that the right-hand side of (D.6) is at most $t_n/4$ whenever $\|x_1 - x_2\| \leq \rho_n$, which gives

$$\rho_n := \frac{\zeta_{K_n}\sqrt{\log(n)}}{4M_\eta^2 L_{K_n}\sqrt{m_n}}.$$

Indeed, one can verify that $\rho_n M_\eta^2 L_{K_n}\overline{\psi}_{K_n,n}\sqrt{m_n/n} = t_n/4$ and hence, on $\mathcal{B}_{n,\eta}$, whenever $\|x_1 - x_2\| \leq \rho_n$ one has from (D.6) that $|\check{h}_{n,\lambda}(x_1) - \check{h}_{n,\lambda}(x_2)| \leq t_n/4$. Therefore, for every $x \in \mathcal{X}$ there exists an index $j(x) \in \{1, \ldots, N_n\}$ such that $\|x - x_{j(x),n}\| \leq \rho_n$. Since $\mathcal{X} = [-M, M]^d$ by Assumption 1, the same standard volumetric argument (see, for example, Corollary 4.2.11 in Vershynin (2026)) used in (A.21) gives

$$N_n = \mathcal{N}(\rho_n, \mathcal{X}, \|\cdot\|) \leq \left(1 + \frac{2M\sqrt{d}}{\rho_n}\right)^d, \tag{D.7}$$

and hence

$$\begin{aligned}\log(N_n) &\leq d\log\left(1 + \frac{2M\sqrt{d}}{\rho_n}\right) = d\log\left(1 + 8MM_\eta^2\sqrt{d}\frac{L_{K_n}\sqrt{m_n}}{\zeta_{K_n}\sqrt{\log(n)}}\right)\\ &\leq d\log\left(1 + 8CMM_\eta^2\sqrt{d}n^q\right) \leq d\log\left(1 + C_\eta n^q\right) \leq d\log\left((1 + C_\eta)n^q\right) \lesssim \log(n)\end{aligned}$$

where in the second inequality we used Assumption 14 (iii), while in the third one defined $C_\eta := 8CMM_\eta^2\sqrt{d}$. Therefore, by the construction of ρ_n, on $\mathcal{B}_{n,\eta}$, it holds

$$\sup_{x\in\mathcal{X}} |\check{h}_{n,\lambda}(x) - \check{h}_{n,\lambda}(x_{j(x),n})| \leq \frac{t_n}{4}.$$

Hence, on $\mathcal{B}_{n,\eta}$,

$$\|\check{h}_{n,\lambda}\|_\infty \leq \max_{x_{j,n}\in\mathcal{C}_n} |\check{h}_{n,\lambda}(x_{j,n})| + \frac{t_n}{4}. \tag{D.8}$$

Step 3. For each $x \in \mathcal{X}$ and $1 \leq i \leq n$, define $a_{i,n}(x) := g_{K_n}(x)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda} g_{K_n}(X_i)$. Then, by the first equality in (D.1), we have that

$$\check{h}_{n,\lambda}(x) = \frac{1}{n}\sum_{i=1}^n a_{i,n}(x)\epsilon_i.$$

Let $\{\tau_n\}_{n\geq 1}$ be a deterministic positive sequence satisfying $\tau_n \to \infty$, a precise choice of which we will specify below. Define

$$\epsilon_{i,n}^{(1)} := \epsilon_i \mathbb{1}\{|\epsilon_i| \leq \tau_n\} - \mathbb{E}\left[\epsilon_i \mathbb{1}\{|\epsilon_i| \leq \tau_n\} \mid X_i\right] \tag{D.9}$$

and

$$\epsilon_{i,n}^{(2)} := \epsilon_i - \epsilon_{i,n}^{(1)}. \tag{D.10}$$

We may hence write $\check{h}_{n,\lambda}(x) = \check{h}^{(1)}_{n,\lambda}(x) + \check{h}^{(2)}_{n,\lambda}(x)$, where, for $\ell \in \{1, 2\}$,

$$\check{h}^{(\ell)}_{n,\lambda}(x) := \frac{1}{n}\sum_{i=1}^n a_{i,n}(x)\epsilon_{i,n}^{(\ell)}. \tag{D.11}$$

We first analyze the random coefficients $a_{i,n}(x)$ in (D.11). Let $u_{i,n}^\top$ denote the ith row of $U_{m_n,n}$, $i \in \{1, \ldots, n\}$. Since the singular value decomposition in (4.65) gives

$$g_{K_n}(X_i) = \sqrt{n}V_{m_n,n}S_{m_n,n}u_{i,n},$$

we have

$$\frac{1}{n}|a_{i,n}(x)| = \frac{1}{\sqrt{n}}\left|g_{K_n}(x)^\top V_{m_n,n}\,\mathrm{diag}\left(\left(\frac{s_{k,n}}{s_{k,n}^2+\lambda_{K_n,n}}\right)_{k=1}^{m_n}\right)u_{i,n}\right|$$
$$\leq \frac{\|g_{K_n}(x)\|}{\sqrt{n}}\psi_{K_n,n,\lambda}\|u_{i,n}\| \leq \frac{\zeta_{K_n}\psi_{K_n,n,\lambda}}{\sqrt{n}},$$

where the first inequality follows from the Cauchy-Schwarz inequality and the definition of $\psi_{K_n,n,\lambda}$ in (4.66), while the second one uses $\|g_{K_n}(x)\| \leq \zeta_{K_n}$ for any $x \in \mathcal{X}$ due to Assumption 14 and that $\|u_{i,n}\| \leq 1$ since $U_{m_n,n}$ has orthonormal columns.

Moreover, we have immediately that

$$\frac{1}{n^2}\sum_{i=1}^n a_{i,n}(x)^2 = \frac{1}{n^2}\sum_{i=1}^n g_{K_n}(x)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} g_{K_n}(X_i) g_{K_n}(X_i)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} g_{K_n}(x)$$
$$= \frac{1}{n} g_{K_n}(x)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1}\widehat{\Gamma}_{K_n,n}\widehat{\Gamma}_{K_n,n,\lambda}^{-1} g_{K_n}(x)$$
$$\leq \frac{\zeta_{K_n}^2\psi_{K_n,n,\lambda}^2}{n}, \tag{D.12}$$

where we used that

$$\left\|\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\widehat{\Gamma}_{K_n,n}\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\right\| = \max_{1\leq k\leq m_n}\frac{s_{k,n}^2}{(s_{k,n}^2+\lambda_{K_n,n})^2} = \psi_{K_n,n,\lambda}^2$$

and that $\|g_{K_n}(x)\| \leq \zeta_{K_n}$ by Assumption 14.

We will use two different coefficient bounds below. The sharper bound in (D.12) controls the conditional variance in the application of the Bernstein inequality. For the maximal size of the remainder term $\epsilon_{i,n}^{(2)}$ in (D.10), we will derive a uniform bound. Since $\widehat{\Gamma}_{K_n,n} = n^{-1}G_{K_n,n}^\top G_{K_n,n}$ is positive semidefinite and $\lambda_{K_n,n} > 0$, we have $\lambda_{\min}(\widehat{\Gamma}_{K_n,n,\lambda}) = \lambda_{\min}(\widehat{\Gamma}_{K_n,n} + \lambda_{K_n,n}I_{K_n}) \geq \lambda_{K_n,n}$ and hence, $\left\|\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\right\| = \lambda_{\min}(\widehat{\Gamma}_{K_n,n,\lambda})^{-1} \leq \lambda_{K_n,n}^{-1}$. Recalling that by the definition $a_{i,n}(x) = g_{K_n}(x)^\top\widehat{\Gamma}_{K_n,n,\lambda}^{-1}g_{K_n}(X_i)$ in the beginning of Step 3, the Cauchy-Schwarz inequality gives, for all $x \in \mathcal{X}$ and $1 \leq i \leq n$,

$$|a_{i,n}(x)| \leq \|g_{K_n}(x)\|\left\|\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\right\|\|g_{K_n}(X_i)\| \leq \zeta_{K_n}\lambda_{K_n,n}^{-1}\zeta_{K_n} = \frac{\zeta_{K_n}^2}{\lambda_{K_n,n}}, \tag{D.13}$$

where we used $\|g_{K_n}(x)\| \leq \zeta_{K_n}$ and, since $X_i \in \mathcal{X}$, also $\|g_{K_n}(X_i)\| \leq \zeta_{K_n}$.

Step 4. We first analyze random variables $\epsilon_{i,n}^{(1)}$ in (D.9) which by construction are centered conditionally on X_i, that is $\mathbb{E}[\epsilon_{i,n}^{(1)} \mid X_i] = 0$. Moreover, by the conditional triangle inequality,

$$|\epsilon_{i,n}^{(1)}| \leq |\epsilon_i|\mathbb{1}\{|\epsilon_i| \leq \tau_n\} + |\mathbb{E}\left[\epsilon_i\mathbb{1}\{|\epsilon_i| \leq \tau_n\} \mid X_i\right]|$$
$$\leq \tau_n + \mathbb{E}\left[|\epsilon_i|\mathbb{1}\{|\epsilon_i| \leq \tau_n\} \mid X_i\right] \leq 2\tau_n. \tag{D.14}$$

Since the observations are i.i.d. and the random features are independent of the data, conditionally on $X_{1:n}$ and $\theta_{1:K_n}$, the variables $\epsilon_{1,n}^{(1)}, \ldots, \epsilon_{n,n}^{(1)}$ are independent. In addition, for every covering point $x_{j,n} \in \mathcal{C}_n$, the coefficients $a_{i,n}(x_{j,n})$ are measurable with respect to the same conditioning variables. Consequently, the random variables $n^{-1}a_{i,n}(x_{j,n})\epsilon_{i,n}^{(1)}$, $i = 1, \ldots, n$, are conditionally independent and centered, since

$$\mathbb{E}\left[\frac{1}{n}a_{i,n}(x_{j,n})\epsilon_{i,n}^{(1)}\middle| X_{1:n}, \theta_{1:K_n}\right] = \frac{1}{n}a_{i,n}(x_{j,n})\mathbb{E}[\epsilon_{i,n}^{(1)} \mid X_i] = 0.$$

Combining the uniform bound in (D.13) with (D.14) yields, uniformly over $x \in \mathcal{X}$ and $1 \le i \le n$,

$$\frac{1}{n}|a_{i,n}(x)\epsilon_{i,n}^{(1)}| \le \frac{2\zeta_{K_n}^2 \tau_n}{n\lambda_{K_n,n}}. \tag{D.15}$$

We now bound the conditional variance of $\epsilon_{i,n}^{(1)}$. Since $\epsilon_{i,n}^{(1)}$ is obtained by conditionally centering the truncated error, it holds that

$$\mathrm{Var}\left(\epsilon_{i,n}^{(1)} \mid X_i\right) = \mathbb{E}\big[(\epsilon_{i,n}^{(1)})^2 \mid X_i\big] \le \mathbb{E}\left[\epsilon_i^2 \mathbb{1}\{|\epsilon_i| \le \tau_n\} \mid X_i\right] \le \mathbb{E}[\epsilon_i^2 \mid X_i] \le \overline{\sigma}^2.$$

where the last inequality follows from Assumption 11. Since $a_{i,n}(x)$ is measurable with respect to $\sigma(X_{1:n}, \theta_{1:K_n})$, it follows from (D.12) that

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[\left(\frac{1}{n}a_{i,n}(x)\epsilon_{i,n}^{(1)}\right)^2 \middle| X_{1:n}, \theta_{1:K_n}\right] &= \frac{1}{n^2}\sum_{i=1}^n a_{i,n}(x)^2 \mathbb{E}\left[(\epsilon_{i,n}^{(1)})^2 \,\middle|\, X_{1:n}, \theta_{1:K_n}\right] \\ &\le \frac{\overline{\sigma}^2}{n^2}\sum_{i=1}^n a_{i,n}(x)^2 \le \frac{\overline{\sigma}^2 \zeta_{K_n}^2 \psi_{K_n,n,\lambda}^2}{n},\end{aligned} \tag{D.16}$$

where in the final inequality we used the bound from (D.12). Recall now that the event $\mathcal{A}_{n,\eta} := \{\psi_{K_n,n,\lambda} \le M_\eta \overline{\psi}_{K_n,n}\}$ is measurable with respect to $X_{1:n}$ and $\theta_{1:K_n}$. Therefore, on $\mathcal{A}_{n,\eta}$, the conditional variance bound in (D.16) becomes

$$\sum_{i=1}^n \mathbb{E}\left[\left(\frac{1}{n}a_{i,n}(x)\epsilon_{i,n}^{(1)}\right)^2 \middle| X_{1:n}, \theta_{1:K_n}\right] \le \frac{\overline{\sigma}^2 M_\eta^2 \zeta_{K_n}^2 \overline{\psi}_{K_n,n}^2}{n}. \tag{D.17}$$

We can now apply the conditional Bernstein inequality in Lemma F.4. Indeed, conditionally on $X_{1:n}$ and $\theta_{1:K_n}$, the random variables $n^{-1}a_{i,n}(x_{j,n})\epsilon_{i,n}^{(1)}$, $i = 1, \dots, n$, are independent and centered, their absolute values are bounded by (D.15), and their conditional variance sum is bounded by (D.17). Hence, for every $1 \le j \le N_n$ and every $c > 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\check{h}_{n,\lambda}^{(1)}(x_{j,n})\right| > ct_n \,\middle|\, X_{1:n}, \theta_{1:K_n}\right) \mathbb{1}_{\mathcal{A}_{n,\eta}} \\ &\quad\le 2\exp\left(-\frac{c^2 t_n^2}{2\left(\overline{\sigma}^2 M_\eta^2 \zeta_{K_n}^2 \overline{\psi}_{K_n,n}^2/n + 2c\zeta_{K_n}^2 \tau_n t_n/(3n\lambda_{K_n,n})\right)}\right) \\ &\quad= 2\exp\left(-\frac{c^2 \log(n)}{2\left(\overline{\sigma}^2 M_\eta^2 + \frac{2}{3}c\frac{\zeta_{K_n}\tau_n}{\lambda_{K_n,n}\overline{\psi}_{K_n,n}}\sqrt{\frac{\log(n)}{n}}\right)}\right)\end{aligned} \tag{D.18}$$

where in the last equality we used the definition of t_n in (D.5).

We choose the truncation τ_n as

$$\tau_n^{1+\delta} := \frac{\zeta_{K_n}}{\lambda_{K_n,n}\overline{\psi}_{K_n,n}}\sqrt{\frac{n}{\log(n)}}, \tag{D.19}$$

where $\delta > 0$ is the exponent in Assumption 13, and recall that, by Assumption 15, $\tau_n \to \infty$. Notice that with this choice it holds that

$$\frac{\zeta_{K_n}\tau_n}{\lambda_{K_n,n}\overline{\psi}_{K_n,n}}\sqrt{\frac{\log(n)}{n}} = \left[\frac{\left(\zeta_{K_n}/(\lambda_{K_n,n}\overline{\psi}_{K_n,n})\right)^{(2+\delta)/\delta}}{\sqrt{n/\log(n)}}\right]^{\delta/(1+\delta)} = O(1), \tag{D.20}$$

where the final equality follows from Assumption 15. It follows from (D.18) and (D.20), for every fixed $c > 0$, there exists a constant $C_{\eta,c} > 0$ such that

$$\mathbb{P}\left(\left|\check{h}^{(1)}_{n,\lambda}(x_{j,n})\right| > ct_n \middle| X_{1:n}\right) \mathbb{1}_{\mathcal{A}_{n,\eta}} \leq 2n^{-C_{\eta,c}},$$

and $C_{\eta,c} \to \infty$ as $c \to \infty$. Taking expectations and applying the union bound over the covering points in $\mathcal{C}_n$, we now get that

$$\mathbb{P}\left(\left\{\max_{1\leq j\leq N_n} |\check{h}^{(1)}_{n,\lambda}(x_{j,n})| > ct_n\right\} \cap \mathcal{A}_{n,\eta}\right) \leq 2N_n n^{-C_{\eta,c}}.$$

Using the fact that if ρ_n^{-1} grows at most polynomially in n, then $\log(N_n) \lesssim \log(n)$, as we noticed below (D.7), results in the right-hand side converging to zero for $c > 0$ sufficiently large, namely

$$\mathbb{P}\left(\left\{\max_{1\leq j\leq N_n} |\check{h}^{(1)}_{n,\lambda}(x_{j,n})| > ct_n\right\} \cap \mathcal{A}_{n,\eta}\right) = o(1). \tag{D.21}$$

Step 5. Next, we move to analyzing random variables $\epsilon^{(2)}_{i,n}$ defined in (D.10). Using their definition, we have that

$$\begin{aligned}\epsilon^{(2)}_{i,n} &= \epsilon_i - \epsilon_i \mathbb{1}\{|\epsilon_i| \leq \tau_n\} + \mathbb{E}\left[\epsilon_i \mathbb{1}\{|\epsilon_i| \leq \tau_n\} \mid X_i\right] \\ &= \epsilon_i \mathbb{1}\{|\epsilon_i| > \tau_n\} - \mathbb{E}\left[\epsilon_i \mathbb{1}\{|\epsilon_i| > \tau_n\} \mid X_i\right],\end{aligned}$$

where in the last equality we used that by Assumption 11, $\mathbb{E}[\epsilon_i \mid X_i] = 0$. We hence have that

$$\mathbb{E}\left[|\epsilon^{(2)}_{i,n}| \mid X_i\right] \leq 2\mathbb{E}\left[|\epsilon_i| \mathbb{1}\{|\epsilon_i| > \tau_n\} \mid X_i\right]. \tag{D.22}$$

We now can bound the maximum of the second summand $\check{h}^{(2)}_{n,\lambda}(x)$ in the decomposition of $\check{h}_{n,\lambda}(x)$ over $\mathcal{C}_n$ as follows

$$\begin{aligned}\max_{1\leq j\leq N_n} |\check{h}^{(2)}_{n,\lambda}(x_{j,n})| &= \max_{1\leq j\leq N_n} \left|\frac{1}{n}\sum_{i=1}^n a_{i,n}(x_{j,n})\epsilon^{(2)}_{i,n}\right| \\ &\leq \frac{1}{n}\sum_{i=1}^n \max_{1\leq j\leq N_n} |a_{i,n}(x_{j,n})| \left|\epsilon^{(2)}_{i,n}\right| \leq \frac{\zeta^2_{K_n}}{\lambda_{K_n,n}} \frac{1}{n}\sum_{i=1}^n \left|\epsilon^{(2)}_{i,n}\right|,\end{aligned}$$

where the first equality follows from the definition (D.11) with $\ell = 2$, while in the last inequality we used that, uniformly over $x \in \mathcal{X}$ and $1 \leq i \leq n$, the bound in (D.13) holds. Finally, by Markov's inequality, for any $c > 0$,

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq j\leq N_n} \left|\check{h}^{(2)}_{n,\lambda}(x_{j,n})\right| > ct_n\right) &\leq \frac{1}{n}\frac{\zeta^2_{K_n}}{c\lambda_{K_n,n}t_n}\sum_{i=1}^n \mathbb{E}\left[\left|\epsilon^{(2)}_{i,n}\right|\right] \\ &\leq \frac{2}{c}\frac{\zeta^2_{K_n}}{\lambda_{K_n,n}t_n}\mathbb{E}\left[|\epsilon_i|\mathbb{1}\{|\epsilon_i| > \tau_n\}\right] \leq \frac{2}{c}\frac{\zeta^2_{K_n}}{\lambda_{K_n,n}t_n}\mathbb{E}\left[\tau_n^{-(1+\delta)}|\epsilon_i|^{2+\delta}\mathbb{1}\{|\epsilon_i| > \tau_n\}\right] \\ &= \frac{2}{c}\frac{\zeta_{K_n}}{\lambda_{K_n,n}\overline{\psi}_{K_n,n}}\sqrt{\frac{n}{\log(n)}}\frac{1}{\tau_n^{1+\delta}}\mathbb{E}\left[|\epsilon_i|^{2+\delta}\mathbb{1}\{|\epsilon_i| > \tau_n\}\right] \\ &= \frac{2}{c}\mathbb{E}\left[|\epsilon_i|^{2+\delta}\mathbb{1}\{|\epsilon_i| > \tau_n\}\right] = o(1),\end{aligned} \tag{D.23}$$

where the second inequality follows from (D.22) and the tower property, in the third inequality under the expectation operator we used the higher-order moment bound on a truncated random variable, the first equality follows after using the definition of t_n in (D.5), and the second one is

a result of substituting the choice of truncation τ_n from (D.19), while the last one follows from the fact that $\tau_n \to \infty$ by construction, by Assumption 13 and by dominated convergence.

STEP 6. Choose $c > 0$ sufficiently large so that (D.21) holds. By the definition $\check{h}_{n,\lambda} = \check{h}^{(1)}_{n,\lambda} + \check{h}^{(2)}_{n,\lambda}$, the triangle inequality gives

$$\mathbb{P}\left(\left\{\max_{1\le j\le N_n} |\check{h}_{n,\lambda}(x_{j,n})| > 2ct_n\right\} \cap \mathcal{A}_{n,\eta}\right) \le \mathbb{P}\left(\left\{\max_{1\le j\le N_n} |\check{h}^{(1)}_{n,\lambda}(x_{j,n})| > ct_n\right\} \cap \mathcal{A}_{n,\eta}\right) + \mathbb{P}\left(\max_{1\le j\le N_n} |\check{h}^{(2)}_{n,\lambda}(x_{j,n})| > ct_n\right) = o(1), \quad \text{(D.24)}$$

where the last equality follows from (D.21) and (D.23).

Recall that $\mathcal{B}_{n,\eta} \subseteq \mathcal{A}_{n,\eta}$ and that, on $\mathcal{B}_{n,\eta}$, (D.8) gives

$$\|\check{h}_{n,\lambda}\|_\infty \le \max_{x_{j,n}\in\mathcal{C}_n} |\check{h}_{n,\lambda}(x_{j,n})| + \frac{t_n}{4}.$$

Hence we can write that

$$\begin{aligned}
\mathbb{P}\left(\|\check{h}_{n,\lambda}\|_\infty > \left(2c+\frac{1}{4}\right)t_n\right) &\le \mathbb{P}\left(\left\{\|\check{h}_{n,\lambda}\|_\infty > \left(2c+\frac{1}{4}\right)t_n\right\} \cap \mathcal{B}_{n,\eta}\right) + \mathbb{P}(\mathcal{B}^{\complement}_{n,\eta}) \\
&\le \mathbb{P}\left(\left\{\|\check{h}_{n,\lambda}\|_\infty > \left(2c+\frac{1}{4}\right)t_n\right\} \cap \mathcal{B}_{n,\eta}\right) + \eta \\
&\le \mathbb{P}\left(\left\{\max_{1\le j\le N_n} |\check{h}_{n,\lambda}(x_{j,n})| > 2ct_n\right\} \cap \mathcal{B}_{n,\eta}\right) + \eta \\
&\le \mathbb{P}\left(\left\{\max_{1\le j\le N_n} |\check{h}_{n,\lambda}(x_{j,n})| > 2ct_n\right\} \cap \mathcal{A}_{n,\eta}\right) + \eta \\
&= o(1) + \eta,
\end{aligned}$$

where the last equality follows from (D.24).

Since the choice of $\eta > 0$ is arbitrary, it follows that

$$\|\check{h}_{n,\lambda}\|_\infty = O_P(t_n) = O_P\left(\zeta_{K_n}\overline{\psi}_{K_n,n}\sqrt{\frac{\log(n)}{n}}\right),$$

which proves the first claim. We conclude by noticing that the first condition in Assumption 15 gives $t_n = o(1)$ and hence it also holds that

$$\|\check{h}_{n,\lambda}\|_\infty = o_P(1),$$

which proves the second claim, as required. □

Proof of Lemma 4.9. We consider the two cases separately. Suppose, first, that condition (i) holds. Since $\mathcal{X}$ and Π are compact and σ is Lipschitz, the constants $M_\sigma := \sup_{\substack{x\in\mathcal{X}\\(A,B)\in\Pi}} \left|\sigma(A^\top x + B)\right| < \infty$ and $C_{A,\Pi} := \sup_{(A,B)\in\Pi} \|A\| < \infty$ are well defined. For every $K \in \mathbb{N}^+$ and $x \in \mathcal{X}$,

$$\sup_{x\in\mathcal{X}} \|g_K(x)\|^2 = \sup_{x\in\mathcal{X}} \sum_{k=1}^{K} \left|\sigma(A_k^\top x + B_k)\right|^2 \le KM_\sigma^2.$$

Consequently, $\sup_{x\in\mathcal{X}} \|g_K(x)\| \le M_\sigma\sqrt{K}$.

Moreover, for any $x_1, x_2 \in \mathcal{X}$, by the Lipschitz continuity of σ and the Cauchy-Schwarz inequality, one has

$$\|g_K(x_1) - g_K(x_2)\|^2 \le \sum_{k=1}^{K} \left|\sigma(A_k^\top x_1 + B_k) - \sigma(A_k^\top x_2 + B_k)\right|^2 \le KL_\sigma^2 C_{A,\Pi}^2 \|x_1 - x_2\|^2.$$

Hence, in case (i) one may take $M_g = \max\{1, M_\sigma\}$ and $L_g = \max\{1, L_\sigma C_{A,\Pi}\}$.

Assume now that condition (ii) holds. Since $\mathcal{X} = [-M, M]^d$ and $\Pi = \{A \in \mathbb{R}^d \mid \|A\|_1 \le C_A\} \times [-C_B, C_B]$, then, for every $x \in \mathcal{X}$ and $(A, B) \in \Pi$,

$$\left|A^\top x + B\right| \le \|A\|_1 \|x\|_\infty + |B| \le MC_A + C_B.$$

Set $R := MC_A + C_B$ and notice that $|\sigma_m(u)| \le R^m$, $u \in [-R, R]$. Hence, $\sup_{x\in\mathcal{X}} \|g_K(x)\| \le R^m\sqrt{K}$. Moreover, σ_m is Lipschitz on $[-R, R]$ with Lipschitz constant mR^{m-1} and hence, for any $x_1, x_2 \in \mathcal{X}$,

$$\begin{aligned}
\|g_K(x_1) - g_K(x_2)\|^2 &\le \sum_{k=1}^K \left|\sigma_m(A_k^\top x_1 + B_k) - \sigma_m(A_k^\top x_2 + B_k)\right|^2 \\
&\le m^2 R^{2m-2} \sum_{k=1}^K |A_k^\top (x_1 - x_2)|^2 \le m^2 R^{2m-2} \sum_{k=1}^K \|A_k\|^2 \|x_1 - x_2\|^2 \\
&\le m^2 R^{2m-2} C_A^2 K \|x_1 - x_2\|^2.
\end{aligned}$$

Therefore, in case (ii), one may take $M_g = \max\{1, R^m\}$ and $L_g = \max\{1, mR^{m-1}C_A\}$.

In both cases, since the sampled parameters belong to Π almost surely, these bounds hold almost surely for every K. Consequently, Assumption 14 (i)-(ii) holds with $\zeta_K = M_g\sqrt{K}$ and $L_K = L_g\sqrt{K}$ as required.

Finally, it remains to check the statement in Assumption 14 (iii), namely (4.64). Indeed,

$$\frac{L_{K_n}\sqrt{m_n}}{\zeta_{K_n}\sqrt{\log(n)}} = \frac{L_g\sqrt{K_n}}{M_g\sqrt{K_n}} \frac{\sqrt{m_n}}{\sqrt{\log(n)}} \le \frac{L_g}{M_g}\sqrt{\frac{n}{\log(n)}} \le Cn^{1/2}$$

for all sufficiently large n and some constant $C > 0$. Hence (4.64) is satisfied, for example with $q = 1/2$, which completes the proof. □

Proof of Corollary 4.8. By Lemma 4.9, Assumption 14 holds and hence the claim follows from Lemma 4.8. □

D.2 Ridge pseudo-projection and spectral stability

Proof of Lemma 4.10. Let $f : \mathcal{X} \to \mathbb{R}$ be bounded. Then, for every $x \in \mathcal{X}$, we write

$$\begin{aligned}
\left|\left(\overline{P}_{K_n,n,\lambda} f\right)(x)\right| &\le \|g_{K_n}(x)\| \left\|\widehat{\Gamma}_{K_n,n,\lambda}^{-1}\left(n^{-1} G_{K_n,n}^\top F\right)\right\| \\
&= \|g_{K_n}(x)\| \left\|\frac{1}{\sqrt{n}} V_{m_n,n} \operatorname{diag}\left(\left(\frac{s_{k,n}}{s_{k,n}^2 + \lambda_{K_n,n}}\right)_{k=1}^{m_n}\right) U_{m_n,n}^\top F\right\| \\
&\le \|g_{K_n}(x)\| \frac{\psi_{K_n,n,\lambda}}{\sqrt{n}} \|U_{m_n,n}^\top F\| \le \|g_{K_n}(x)\| \frac{\psi_{K_n,n,\lambda}}{\sqrt{n}} \|F\|, \qquad \text{(D.25)}
\end{aligned}$$

where in the first equality we applied the singular value decomposition in (4.65), in the second inequality we used the definition of $\psi_{K_n,n,\lambda}$ in (4.66), while in the same and the subsequent inequality we also used that $U_{m_n,n}$ and $V_{K_n,n}$ are matrices with orthonormal columns.

Next, notice that

$$\frac{1}{\sqrt{n}}\|F\| = \left(\frac{1}{n}\sum_{i=1}^n f(X_i)^2\right)^{1/2} \le \|f\|_\infty.$$

Hence, taking in (D.25) the supremum over $x \in \mathcal{X}$ and then over all bounded f satisfying $\|f\|_\infty \le 1$ together with using Assumption 14 (i), yields

$$\|\overline{P}_{K_n,n,\lambda}\|_\infty \le \zeta_{K_n} \psi_{K_n,n,\lambda},$$

as required. □

Proof of Lemma 4.11. We start by writing that

$$\begin{aligned}\left\|\widetilde{h}_{n,\lambda}-h\right\|_\infty &= \left\|\overline{P}_{K_n,n,\lambda}\left(h-h^{\mathrm{IS}}_{K_n}\right)+\left(\overline{P}_{K_n,n,\lambda}h^{\mathrm{IS}}_{K_n}-h^{\mathrm{IS}}_{K_n}\right)+\left(h^{\mathrm{IS}}_{K_n}-h\right)\right\|_\infty \\ &\leq \left(1+\left\|\overline{P}_{K_n,n,\lambda}\right\|_\infty\right)\left\|h-h^{\mathrm{IS}}_{K_n}\right\|_\infty+\left\|\overline{P}_{K_n,n,\lambda}h^{\mathrm{IS}}_{K_n}-h^{\mathrm{IS}}_{K_n}\right\|_\infty. \qquad \text{(D.26)}\end{aligned}$$

where we used the definition of the pseudo-projection operator in the first equality and the definition of $\|\overline{P}_{K_n,n,\lambda}\|_\infty$ in the subsequent inequality.

It remains to control the second term. Let $H^{\mathrm{IS}}_{K_n,n} := \left(h^{\mathrm{IS}}_{K_n}(X_1),\ldots,h^{\mathrm{IS}}_{K_n}(X_n)\right)^\top = G_{K_n,n}W_{K_n}$. Then

$$\begin{aligned}\left(\overline{P}_{K_n,n,\lambda}h^{\mathrm{IS}}_{K_n}\right)(x) &= g_{K_n}(x)^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\left(n^{-1}G^\top_{K_n,n}H^{\mathrm{IS}}_{K_n,n}\right) \\ &= g_{K_n}(x)^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\left(n^{-1}G^\top_{K_n,n}G_{K_n,n}\right)W_{K_n} \\ &= g_{K_n}(x)^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\widehat{\Gamma}_{K_n,n}W_{K_n} \\ &= g_{K_n}(x)^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\left(\widehat{\Gamma}_{K_n,n,\lambda}-\lambda_{K_n,n}I_{K_n}\right)W_{K_n} \\ &= g_{K_n}(x)^\top W_{K_n}-\lambda_{K_n,n}g_{K_n}(x)^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda}W_{K_n}.\end{aligned}$$

Using this expression, one immediately gets that

$$\begin{aligned}\sup_{x\in\mathcal{X}}\left|\left(\overline{P}_{K_n,n,\lambda}h^{\mathrm{IS}}_{K_n}-h^{\mathrm{IS}}_{K_n}\right)(x)\right| &= \sup_{x\in\mathcal{X}}\left|-\lambda_{K_n,n}g_{K_n}(x)^\top\widehat{\Gamma}^{-1}_{K_n,n,\lambda}W_{K_n}\right| \\ &\leq \lambda_{K_n,n}\left\|\widehat{\Gamma}^{-1}_{K_n,n,\lambda}W_{K_n}\right\|\sup_{x\in\mathcal{X}}\|g_{K_n}(x)\| \\ &\leq \lambda_{K_n,n}\zeta_{K_n}\left\|\widehat{\Gamma}^{-1}_{K_n,n,\lambda}W_{K_n}\right\|, \qquad \text{(D.27)}\end{aligned}$$

where we used the Cauchy-Schwarz inequality and Assumption 14(i). Substituting this bound into (D.26) proves (4.68).

We next bound $R_{K_n,n,\lambda}$ as follows:

$$\begin{aligned}R_{K_n,n,\lambda} &\leq \lambda_{K_n,n}\zeta_{K_n}\left\|\widehat{\Gamma}^{-1}_{K_n,n,\lambda}\right\|\|W_{K_n}\| = \zeta_{K_n}\frac{\lambda_{K_n,n}}{s^2_{K_n,n}+\lambda_{K_n,n}}\|W_{K_n}\| \\ &\leq \sqrt{6}M_f\zeta_{K_n}K_n^{-1/2}\frac{\lambda_{K_n,n}}{s^2_{K_n,n}+\lambda_{K_n,n}},\end{aligned}$$

where in the first equality we used again the singular value decomposition in (4.65) and applied in the second inequality Lemma 3.7 for the bound of the importance sampling weights constructed in Theorem 3.5. This proves (4.70).

Finally, Lemma 4.10 gives $\left\|\overline{P}_{K_n,n,\lambda}\right\|_\infty \leq \zeta_{K_n}\psi_{K_n,n,\lambda}$ and combining it with (4.68) and (D.27) yields (4.71), as required. □

Proof of Theorem 4.12. We recall decomposition in (4.63) which follows from the application of the triangle inequality, namely,

$$\|\widehat{h}_{n,\lambda}-h\|_\infty \leq \|\widehat{h}_{n,\lambda}-\widetilde{h}_{n,\lambda}\|_\infty+\|\widetilde{h}_{n,\lambda}-h\|_\infty.$$

Applying (4.71) from Lemma 4.11 to the second term proves (4.72).

Finally, we apply Lemma 4.8 to the first term and notice that by (3.8) in Theorem 3.5 it holds, via Markov's inequality, that

$$\|h-h^{\mathrm{IS}}_{K_n}\|_\infty = O_P\left(\left(\sqrt{d}+\sqrt{\log(K_n)}\right)K_n^{-1/2-\alpha}\right).$$

Combining these bounds in (4.72) and using that by the first statement of Assumption 15, $\psi_{K_n,n,\lambda} = O_P\left(\overline{\psi}_{K_n,n}\right)$, proves (4.73), as required. □

Proof of Corollary 4.9. The proof is an immediate application of Lemma 4.9 in the statement of Theorem 4.12. □

D.3 Prediction space and leverage based stability

Proof of Lemma 4.13. Let $F := (f(X_1), \dots, f(X_n))^\top$. By (4.67), one has that

$$\big(\overline{P}_{K_n,n,\lambda} f\big)(x) = g_{K_n}(x)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda} \left(n^{-1} G^\top_{K_n,n} F\right) = \sum_{i=1}^n p_{i,n,\lambda}(x) f(X_i),$$

which proves (4.77). Hence,

$$\big|\big(\overline{P}_{K_n,n,\lambda} f\big)(x)\big| \le \|f\|_\infty \sum_{i=1}^n |p_{i,n,\lambda}(x)|,$$

and taking the supremum over $x \in \mathcal{X}$ and $\|f\|_\infty \le 1$ together with the first definition in (4.78) proves (4.80).

Next, we write that

$$\begin{aligned}\sum_{i=1}^n p_{i,n,\lambda}(x)^2 &= \frac{1}{n^2}\left(g_{K_n}(x)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda} G^\top_{K_n,n} G_{K_n,n} \widehat{\Gamma}^{-1}_{K_n,n,\lambda} g_{K_n}(x)\right) \\ &= \frac{1}{n}\left(g_{K_n}(x)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda} \widehat{\Gamma}_{K_n,n} \widehat{\Gamma}^{-1}_{K_n,n,\lambda} g_{K_n}(x)\right),\end{aligned}$$

where the first equality follows from definition (4.76) and in the second one we used $n^{-1} G^\top_{K_n,n} G_{K_n,n} = \widehat{\Gamma}_{K_n,n}$. Moreover, $\widehat{\Gamma}_{K_n,n} \preceq \widehat{\Gamma}_{K_n,n,\lambda}$ implies

$$\widehat{\Gamma}^{-1}_{K_n,n,\lambda} \widehat{\Gamma}_{K_n,n} \widehat{\Gamma}^{-1}_{K_n,n,\lambda} \preceq \widehat{\Gamma}^{-1}_{K_n,n,\lambda},$$

which proves the first inequality in (4.81).

Finally, the Cauchy-Schwarz inequality for the inner product induced by $\widehat{\Gamma}^{-1}_{K_n,n,\lambda}$ together with the definition of $\ell_{n,\lambda}$ in (4.79) gives

$$|p_{i,n,\lambda}(x)| \le \frac{1}{n}\left(g_{K_n}(x)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda} g_{K_n}(x)\right)^{1/2} \left(g_{K_n}(X_i)^\top \widehat{\Gamma}^{-1}_{K_n,n,\lambda} g_{K_n}(X_i)\right)^{1/2} \le \frac{\ell_{n,\lambda}}{n}.$$

Taking the supremum over $x \in \mathcal{X}$ and $i = 1, \dots, n$ proves the second inequality in (4.81). □

Proof of Lemma 4.14. Set $\check{h}_{n,\lambda} := \widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}$ and $e := (\epsilon_1, \dots, \epsilon_n)^\top$. By (4.77),

$$\check{h}_{n,\lambda}(x) = \sum_{i=1}^n p_{i,n,\lambda}(x) \epsilon_i.$$

We follow again the steps used in the proof of Lemma 4.8. Indeed, the same argument used to obtain (D.4) gives

$$|\check{h}_{n,\lambda}(x_1) - \check{h}_{n,\lambda}(x_2)| \le L_{K_n} \psi_{K_n,n,\lambda} \widetilde{U}_n \|x_1 - x_2\|, \tag{D.28}$$

where $\widetilde{U}_n := \|n^{-1/2} U^\top_{m_n,n} e\|$. As in (D.2), conditioning on $X_{1:n}$ and $\theta_{1:K_n}$ gives

$$\mathbb{E}[\widetilde{U}_n^2 \mid X_{1:n}, \theta_{1:K_n}] \le \overline{\sigma}^2 \frac{m_n}{n},$$

and hence $\widetilde{U}_n = O_P(\sqrt{m_n/n})$. Fix $\eta > 0$. By the hypothesis $\ell_{n,\lambda} = O_P(K_n)$, the condition $\psi_{K_n,n,\lambda} = O_P(\overline{\psi}_{K_n,n})$, and the preceding bound on $\widetilde{U}_n$, there exists $M_\eta > 0$ such that, for all sufficiently large n, the events

$$\begin{aligned}\mathcal{A}_{n,\eta} &:= \left\{\psi_{K_n,n,\lambda} \le M_\eta \overline{\psi}_{K_n,n},\ \ell_{n,\lambda} \le M_\eta K_n\right\}, \\ \mathcal{B}_{n,\eta} &:= \mathcal{A}_{n,\eta} \cap \left\{\widetilde{U}_n \le M_\eta \sqrt{\frac{m_n}{n}}\right\}\end{aligned}$$

satisfy $\mathbb{P}(\mathcal{B}_{n,\eta}^{\complement}) \le \eta$. Moreover, $\mathcal{A}_{n,\eta}$ is measurable with respect to $\sigma(X_{1:n}, \theta_{1:K_n})$. On $\mathcal{A}_{n,\eta}$, Lemma 4.13 implies

$$\nu_{n,\lambda} \le M_\eta \frac{K_n}{n} \quad \text{and} \quad \mathfrak{m}_{n,\lambda} \le M_\eta \frac{K_n}{n}. \tag{D.29}$$

Let $\mathcal{C}_n$ be a minimal ρ_n-net of $\mathcal{X}$, where

$$\rho_n := \frac{\sqrt{K_n \log(n)}}{4M_\eta^2 L_{K_n} \overline{\psi}_{K_n,n} \sqrt{m_n}}.$$

Let

$$t_n := \sqrt{\frac{K_n \log(n)}{n}}.$$

For each $x \in \mathcal{X}$, let $x_{j(x),n}$ be a closest element of $\mathcal{C}_n$. It follows from (D.28) that, on $\mathcal{B}_{n,\eta}$,

$$\sup_{x \in \mathcal{X}} |\check{h}_{n,\lambda}(x) - \check{h}_{n,\lambda}(x_{j(x),n})| \le \frac{t_n}{4}. \tag{D.30}$$

Since $\mathcal{X} = [-M, M]^d$, the volumetric covering bound (see, for example, Corollary 4.2.11 in Vershynin (2026))and (4.84) imply

$$\log(|\mathcal{C}_n|) = O(\log(n)). \tag{D.31}$$

Define

$$\tau_n := \left(\frac{K_n n}{\log(n)}\right)^{\frac{1}{2(1+\delta)}}$$

and decompose $\epsilon_i = \epsilon_{i,n}^{(1)} + \epsilon_{i,n}^{(2)}$, where

$$\epsilon_{i,n}^{(1)} := \epsilon_i \mathbb{1}\{|\epsilon_i| \le \tau_n\} - \mathbb{E}[\epsilon_i \mathbb{1}\{|\epsilon_i| \le \tau_n\} \mid X_i] \quad \text{and} \quad \epsilon_{i,n}^{(2)} := \epsilon_i - \epsilon_{i,n}^{(1)}.$$

Conditionally on $X_{1:n}$ and $\theta_{1:K_n}$, the random variables $p_{i,n,\lambda}(x_{j,n})\epsilon_{i,n}^{(1)}$, $i = 1, \dots, n$, are independent and centered. On $\mathcal{A}_{n,\eta}$, their conditional variance and maximal absolute value satisfy

$$\sum_{i=1}^n \mathbb{E}\left[p_{i,n,\lambda}(x_{j,n})^2 (\epsilon_{i,n}^{(1)})^2 \mid X_{1:n}, \theta_{1:K_n}\right] \le \overline{\sigma}^2 \nu_{n,\lambda} \le M_\eta \overline{\sigma}^2 \frac{K_n}{n}$$

and

$$\max_{1 \le i \le n} \left|p_{i,n,\lambda}(x_{j,n})\epsilon_{i,n}^{(1)}\right| \le 2\tau_n \mathfrak{m}_{n,\lambda} \le 2M_\eta \tau_n \frac{K_n}{n}.$$

Next,

$$(\tau_n t_n)^{2(1+\delta)} = \left(\frac{K_n^{1+2/\delta} \log(n)}{n}\right)^{\delta} = O(1)$$

by (4.83). Therefore, the conditional Bernstein inequality gives, for every sufficiently large constant $C_\eta > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n p_{i,n,\lambda}(x_{j,n})\epsilon_{i,n}^{(1)}\right| > C_\eta t_n \,\middle|\, X_{1:n}, \theta_{1:K_n}\right) \mathbb{1}_{\mathcal{A}_{n,\eta}} \le 2n^{-c_\eta C_\eta}$$

for some constant $c_\eta > 0$ independent of C_η. Combining this inequality with (D.31) and applying the union bound yields

$$\mathbb{P}\left(\left\{\max_{x_{j,n} \in \mathcal{C}_n} \left|\sum_{i=1}^n p_{i,n,\lambda}(x_{j,n})\epsilon_{i,n}^{(1)}\right| > C_\eta t_n\right\} \cap \mathcal{A}_{n,\eta}\right) = o(1). \tag{D.32}$$

It remains to control the remainder. Since $\mathbb{E}[\epsilon_i \mid X_i] = 0$,

$$\epsilon_{i,n}^{(2)} = \epsilon_i \mathbb{1}\{|\epsilon_i| > \tau_n\} - \mathbb{E}[\epsilon_i \mathbb{1}\{|\epsilon_i| > \tau_n\} \mid X_i].$$

On $\mathcal{A}_{n,\eta}$, (D.29) gives

$$\sup_{x\in\mathcal{X}} \left| \sum_{i=1}^n p_{i,n,\lambda}(x)\epsilon_{i,n}^{(2)} \right| \le \mathfrak{m}_{n,\lambda} \sum_{i=1}^n |\epsilon_{i,n}^{(2)}| \le M_\eta \frac{K_n}{n} \sum_{i=1}^n |\epsilon_{i,n}^{(2)}|.$$

Consequently, for every $c > 0$,

$$\mathbb{P}\left(\left\{ \sup_{x\in\mathcal{X}} \left| \sum_{i=1}^n p_{i,n,\lambda}(x)\epsilon_{i,n}^{(2)} \right| > ct_n \right\} \cap \mathcal{A}_{n,\eta} \right) \le \frac{2M_\eta K_n}{ct_n\tau_n^{1+\delta}} \mathbb{E}[|\epsilon_i|^{2+\delta}\mathbb{1}\{|\epsilon_i| > \tau_n\}] = o(1). \tag{D.33}$$

By the definitions of t_n and τ_n,

$$t_n\tau_n^{1+\delta} = K_n.$$

The preceding probability is therefore $o(1)$ by Assumption 13 and dominated convergence. Combining (D.30) with the decomposition $\epsilon_i = \epsilon_{i,n}^{(1)} + \epsilon_{i,n}^{(2)}$, we obtain, on $\mathcal{B}_{n,\eta}$,

$$\|\check{h}_{n,\lambda}\|_\infty \le \max_{x_{j,n}\in\mathcal{C}_n} \left| \sum_{i=1}^n p_{i,n,\lambda}(x_{j,n})\epsilon_{i,n}^{(1)} \right| + \sup_{x\in\mathcal{X}} \left| \sum_{i=1}^n p_{i,n,\lambda}(x)\epsilon_{i,n}^{(2)} \right| + \frac{t_n}{4}.$$

Consequently, for every $\eta > 0$ and a sufficiently large constant $C_\eta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\|\check{h}_{n,\lambda}\|_\infty > (C_\eta + \frac{5}{4})t_n \right) \\
&\quad \le \mathbb{P}(\mathcal{B}_{n,\eta}^{\complement}) + \mathbb{P}\left(\left\{ \max_{x_{j,n}\in\mathcal{C}_n} \left| \sum_{i=1}^n p_{i,n,\lambda}(x_{j,n})\epsilon_{i,n}^{(1)} \right| > C_\eta t_n \right\} \cap \mathcal{A}_{n,\eta} \right) \\
&\qquad + \mathbb{P}\left(\left\{ \sup_{x\in\mathcal{X}} \left| \sum_{i=1}^n p_{i,n,\lambda}(x)\epsilon_{i,n}^{(2)} \right| > t_n \right\} \cap \mathcal{A}_{n,\eta} \right) \le \eta + o(1),
\end{aligned}$$

where we used $\mathcal{B}_{n,\eta} \subseteq \mathcal{A}_{n,\eta}$, (D.32), and the remainder estimate in (D.33). Since $\eta > 0$ is arbitrary, this establishes $\|\check{h}_{n,\lambda}\|_\infty = O_P(t_n)$ and proves (4.85). □

Proof of Lemma 4.15. Since $h_{K_n}^{\mathrm{IS}}(x) = g_{K_n}(x)^\top W_{K_n}$, the definition of $\overline{P}_{K_n,n,\lambda}$ gives

$$(\overline{P}_{K_n,n,\lambda}h_{K_n}^{\mathrm{IS}} - h_{K_n}^{\mathrm{IS}})(x) = -\lambda_{K_n,n} g_{K_n}(x)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} W_{K_n}.$$

The Cauchy-Schwarz inequality for the inner product induced by $\widehat{\Gamma}_{K_n,n,\lambda}^{-1}$ yields

$$R_{K_n,n,\lambda} \le \lambda_{K_n,n}\ell_{n,\lambda}^{1/2} \left(W_{K_n}^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} W_{K_n} \right)^{1/2} \le \lambda_{K_n,n}\ell_{n,\lambda}^{1/2} \left(\lambda_{K_n,n}^{-1} \|W_{K_n}\|^2 \right)^{1/2},$$

where in the second inequality we used that $\widehat{\Gamma}_{K_n,n,\lambda} \succeq \lambda_{K_n,n} I_{K_n}$, and which proves (4.86). The bound (4.87) follows from $\|W_{K_n}\| \le \sqrt{6} M_f K_n^{-1/2}$. Finally, $\ell_{n,\lambda} = O_P(K_n)$ proves (4.88). □

Proof of Theorem 4.16. The decomposition in (4.63) and Lemma 4.11 give

$$\|\widehat{h}_{n,\lambda} - h\|_\infty \le \|\widehat{h}_{n,\lambda} - \widetilde{h}_{n,\lambda}\|_\infty + (1 + \|\overline{P}_{K_n,n,\lambda}\|_\infty)\|h - h_{K_n}^{\mathrm{IS}}\|_\infty + R_{K_n,n,\lambda}.$$

Applying (4.80) and (4.87) proves (4.89). The rate in (4.90) follows from Lemma 4.14, Assumption 16, Theorem 3.5, Markov's inequality, and (4.88). □

Proof of Corollary 4.10. Under (4.91), one has $\log(K_n) \asymp \log(n)$ and

$$\sqrt{\frac{K_n \log(n)}{n}} \asymp \sqrt{\log(K_n)} K_n^{-1/2-\alpha} \asymp n^{-\frac{1+2\alpha}{4(1+\alpha)}} \sqrt{\log(n)}.$$

Moreover, (4.92) implies (4.83), while (4.93) gives

$$\sqrt{\lambda_{K_n,n}} = O\left(n^{-\frac{1+2\alpha}{4(1+\alpha)}} \sqrt{\log(n)}\right).$$

The conclusion follows from (4.90). □

E Theory for generic sieve estimation and inference

This appendix develops complementary results for constrained RWNN sieve estimation under general loss functions. We first recall a baseline L^2 risk bound, then combine the approximation results of Section 3 with generic sieve arguments to obtain convergence rates under suitable empirical-process and dependence conditions. Finally, we establish asymptotic normality for smooth functionals using a local perturbation argument. Unlike the ridge-normalized theory of Section 4, these results concern constrained, unregularized sieve estimators, although the approximation bounds developed here also enter the ridge analysis.

E.1 A baseline L^2 risk bound

Consider the regression problem (2.1) and the associated RWNN least squares estimator $h_{K,n}^{\theta,\widehat{W}}$ in (2.9). As a baseline, we recall the L^2 risk bound of Gonon (2023) for the L-clipped RWNN least-squares estimator. More explicitly, for a clipping level $L > 0$, define the clipped RWNN least-squares estimator by

$$\widehat{h}_{K,n,L}(x) := \min\{\max\{\widehat{h}_{K,n}(x), -L\}, L\}, \qquad x \in \mathcal{X}.$$

Whenever this estimator is used, we assume that h is bounded and $L \geq \|h\|_\infty$. The L^2 error guarantees are derived in Gonon (2023) under the following assumptions.

Assumption 17. Function $h : \mathbb{R}^d \to \mathbb{R}$ is convolutional, meaning that a $\mathbb{R}^d$-valued random vector V and a function $\Phi \in L^1(\mathbb{R}^d)$ exist such that

$$h(x) = \mathbb{E}_V[\Phi(x + V)].$$

Assumption 18. There exists $C_V > 0$ such that

$$|\mathbb{E}[e^{i\xi^\top V}]| \leq \exp(-C_V\|\xi\|_2^2) \quad \text{for all } \xi \in \mathbb{R}^d.$$

Assumption 19. The weight density π_A is a multivariate t-distribution with $\eta > 1$ degrees of freedom, $t_\eta(0, \mathbb{I}_d)$. The bias density π_B satisfies a polynomial lower-tail bound; that is, there exists a polynomial function $p(z)$ such that $\inf_{z\in\mathbb{R}} p(z)\pi_B(z) \geq 1$.

The following theorem (Gonon 2023, Theorem 17) provides an L^2 error bound for the RWNN least squares L-clipped estimator $\widehat{h}_{K,n,L}$. Let (X, Y) have the same distribution as (X_1, Y_1) but be independent from $\{(Y_i, X_i)\}_{i=1}^n$, A, and B.

Theorem E.1 (Gonon 2023, Theorem 17)**.** *Let Assumptions 1, 17, 18 and 19 hold with $C_V > (2^{3/2}\pi)^{-1}$. Let $L > 0$ and assume $\varsigma^2 := \sup_{x\in\mathbb{R}^d} \mathbb{E}\left[(Y_1 - h(X_1))^2 | X_1 = x\right] < \infty$ and $\sup_{x\in\mathbb{R}^d} |h(x)| \leq L$. Then, there exists $q \in \mathbb{N}$ and $C_{app} > 0$ such that*

$$\mathbb{E}\big[|h(X) - \widehat{h}_{K,n,L}(X)|^2\big]^{1/2} \leq C_{app}\left(\max(\varsigma, L)\sqrt{\frac{K(\log(n)+1)}{n}} + \|\Phi\|_{L^1(\mathbb{R}^d)}\frac{(\eta+d)^{q+3}}{\sqrt{K}}\right).$$

The constant q depends only on π_B while C_{app} depends on C_V, η, π_B, M, but does not depend on d, n, or K.

Remark 19. The bound in Theorem E.1 implies the $\mathcal{O}(\sqrt{K(\log(n)+1)/n} + 1/\sqrt{K})$ error. Balancing the two terms yields the choice

$$K \asymp n^{1/2}(\log(n)+1)^{-1/2},$$

which leads to the convergence rate

$$\mathcal{O}\big(n^{-1/4}(\log(n)+1)^{1/4}\big).$$

Up to logarithmic factors, this rate coincides with the classical Barron approximation rate (Barron, 1993, 1994). However, it remains slower than the rates typically required for nonparametric regression estimators under smoothness assumptions. In particular, sieve estimators based on fully trainable shallow neural networks can achieve rates $o_P(n^{-1/4})$ under suitable regularity conditions (Chen and White, 1999). This rate is the upper rate bound needed to obtain standard asymptotic CLT results for the CEF estimators under weak dependence (Shen (1997)).

Consequently, improving upon the convergence rate in Theorem E.1 is important for obtaining sharper statistical guarantees for RWNN estimators.

E.2 Generic sieve estimation and β-Hölder RWNN rates

We now combine the L^2 approximation result under the β-Hölder source condition with a sieve-estimation argument for RWNN estimators. To derive our asymptotic sieve results, we use the framework of Shen and Wong (1994), its extension to weakly dependent data in Chen and Shen (1998), and its neural-network specialization in Chen and White (1999). We view h as a function-valued parameter belonging to a class $\mathcal{H}$, and collect outcomes and regressors in the random variables $Z_i := (Y_i, X_i^\top)^\top \in \mathcal{Z} \subseteq \mathbb{R}^{d+1}$, $i = 1, \dots, n$, where we assume that $X_i \in \mathcal{X} \subset \mathbb{R}^d$ almost surely, with $\mathcal{X}$ compact.

Assumption 20 (Data)**.** The sequence $\{Z_i\}_{i\in\mathbb{N}^+}$ is stationary, with common distribution P_Z, and satisfies one of the following conditions:

(i) $\{Z_i\}$ is i.i.d. or ν-dependent for some fixed $\nu \in \mathbb{N}^+$;

(ii) $\{Z_i\}$ is β-mixing with $\beta_{\text{mix}}(j) \le \beta_0 j^{-\zeta}$, $j \ge 1$, for some $\beta_0 > 0$ and $\zeta > 2$.

Let $\mathcal{V}$ be a linear space of functions on $\mathcal{X}$ containing $\mathcal{H}$ and let $\|\cdot\|_\circ$ be a seminorm on $\mathcal{V}$. Recall from Section 2 that $\mathcal{H}_K$ denotes the RWNN class of width K. In this subsection, $K_n = K(n)$ may depend on the sample size. Conditional on the sampled $\theta_k = (A_k^\top, B_k)^\top$, $k = 1, \dots, K_n$, we work with the ℓ_2-constrained RWNN sieve classes

$$\mathcal{H}_{K_n}(M_n) := \left\{ x \mapsto W^\top g_{K_n}(x) = \sum_{k=1}^{K_n} W_k \sigma(A_k^\top x + B_k) \,\middle|\, \|W\| \le M_n \right\} \subseteq \mathcal{H}, \tag{E.1}$$

where

$$g_{K_n}(x) = \left(\sigma(A_1^\top x + B_1), \dots, \sigma(A_{K_n}^\top x + B_{K_n})\right)^\top.$$

All entropy bounds below are understood conditionally on the hidden-layer parameters, while the final O_P statements are with respect to the joint law of the data and the random features.

The generic sieve rate in Theorem E.4 and the entropy bound in Lemma E.5 hold for any deterministic sequence $M_n > 0$ under their respective assumptions. For the explicit L^2 regression rate, we take $M_n = C_\alpha M_f \sqrt{\frac{\log(K_n)}{K_n}}$ with C_α as in Lemma 3.3. This choice is dictated by the constrained approximation result in Corollary E.1, since with this radius, the control-neighbors approximant belongs to $\mathcal{H}_{K_n}(M_n)$ on the high-probability event quantified in Lemma 3.3. This choice is also used subsequently in Theorem E.7, which delivers an explicit convergence rate.

We first state the identification, curvature and empirical-process conditions used in the generic sieve argument. The population identification condition (a) and the local separation condition (b) are similar to the corresponding conditions (a) and (b) in Assumption 3.2 in Chen and White (1999), with the local separation condition (b) also analogous to condition C1 of Shen and Wong (1994). The local variance (c), entropy (d), and concentration (e) conditions are analogous to conditions C2′-C4′ for unbounded criteria, respectively. Our specific formulation of the local variance and concentration conditions follows more closely Assumption 3.4 of Chen and White (1999). The entropy condition is stated explicitly, while in Chen and White (1999)

the corresponding entropy bound is derived from the ANN sieve structure in the proof of their Theorem 3.1. All conditions are adapted to the loss-minimization setting and to uniformity over the realized random features.

Assumption 21 (Generic loss conditions)**.** Let $\rho_\circ(g,h) := \|g-h\|_\circ$, $g,h \in \mathcal{H}$. There exists a constant $\eta_0 > 0$ such that the following conditions hold on

$$\mathcal{H}_{K_n}(M_n;\eta_0) := \{g \in \mathcal{H}_{K_n}(M_n) \mid \rho_\circ(g,h) \le \eta_0\}.$$

The constant $C_1 > 0$ and exponent $s \in (0,2)$ do not depend on n or on the realization of $\theta_1,\dots,\theta_{K_n}$. The variance and local concentration constants may depend on n but not on the realization of $\theta_1,\dots,\theta_{K_n}$, and are uniformly bounded by a deterministic sequence $\Lambda_n \ge 1$, with $\Lambda_n = O((\log(n))^m)$ for some fixed $m < \infty$.

(a) [Identification] $\mathbb{E}_Z[\ell(Z_i,h)] < \infty$ and

$$\mathbb{E}_Z[\ell(Z_i,h)] \le \mathbb{E}_Z[\ell(Z_i,\widetilde{h})]$$

for all $\widetilde{h} \in \mathcal{H}$ with $\mathbb{E}_Z[\ell(Z_i,\widetilde{h})] < \infty$.

(b) [Local quadratic separation] For all $g \in \mathcal{H}_{K_n}(M_n;\eta_0)$, conditionally on $\theta_1,\dots,\theta_{K_n}$,

$$\mathbb{E}_Z[\ell(Z_i,g) - \ell(Z_i,h)] \ge 2C_1\rho_\circ(g,h)^2.$$

(c) [Local variance] There exists $C_{2,n} \le \Lambda_n$ such that, conditionally on $\theta_1,\dots,\theta_{K_n}$, for all sufficiently small $\varepsilon > 0$,

$$\sup_{\{g \in \mathcal{H}_{K_n}(M_n;\eta_0) \mid \rho_\circ(g,h) \le \varepsilon\}} \mathrm{Var}_Z[\ell(Z_i,g) - \ell(Z_i,h)] \le C_{2,n}\varepsilon^2.$$

(d) [Local bracketing entropy] For $\delta > 0$, define

$$\mathcal{M}_n(\delta) := \{z \mapsto \ell(z,g) - \ell(z,h) \mid g \in \mathcal{H}_{K_n}(M_n;\eta_0),\ \rho_\circ(g,h) \le \delta\}.$$

Let $\mathcal{N}_{[]}(\varepsilon,\mathcal{F},L^2(P_Z))$ denote the $L^2(P_Z)$ bracketing number of $\mathcal{F}$ (see, for example, Definition 2.1.6 in van der Vaart and Wellner (2023)). There exists a deterministic sequence $\delta_n \to 0$ such that $\delta_n \ge 0$ is the smallest value for which it holds

$$\delta^{-2}\int_{\delta^2}^{\delta} \sqrt{\log \mathcal{N}_{[]}(w,\mathcal{M}_n(\delta),L^2(P_Z))}\,dw \lesssim \sqrt{n}, \tag{E.2}$$

for all $\delta \in [\delta_n,\eta_0]$, uniformly over the realization of $\theta_1,\dots,\theta_{K_n}$.

(e) [Local concentration] There exist constants $\kappa > 0$, a deterministic constant $C_{3,n} \le \Lambda_n$, and a measurable function $\psi_n : \mathcal{Z} \to [0,\infty)$ such that, conditionally on $\theta_1,\dots,\theta_{K_n}$, for every $\Delta > 0$,

$$\sup_{\{g \in \mathcal{H}_{K_n}(M_n;\eta_0) \mid \rho_\circ(g,h) \le \Delta\}} |\ell(Z_i,g) - \ell(Z_i,h)| \le \Delta^s\psi_n(Z_i),$$

with

$$\mathbb{E}_Z[|\psi_n(Z_i)|^\kappa] \le C_{3,n},$$

where the moment exponent satisfies

$$\begin{cases} \kappa > 2, & \text{under Assumption 20 (i)},\\ \kappa > 4, & \text{with } \zeta = \kappa - 2 \text{ under Assumption 20 (ii)}. \end{cases}$$

Define the empirical risk for the regression problem, for $\widetilde{h} \in \mathcal{H}$, by

$$L_n(\widetilde{h}) := \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, \widetilde{h}),$$

where $\ell : \mathcal{Z} \times \mathcal{H} \to [0, \infty)$ is a loss function. Additionally, for $g \in \mathcal{H}_{K_n}(M_n)$, define

$$q(z, g, h) := \ell(z, g) - \ell(z, h) \tag{E.3}$$

and the corresponding population and empirical risk differences as

$$Q(g, h) := \mathbb{E}_Z[q(Z_i, g, h)] \text{ and } Q_n(g, h) := \frac{1}{n}\sum_{i=1}^{n} q(Z_i, g, h) = L_n(g) - L_n(h), \tag{E.4}$$

respectively.

As in the sieve-extremum formulations of Shen and Wong (1994), Chen and Shen (1998), and the ANN specialization of Chen and White (1999), we allow for approximate empirical minimization.

Assumption 22 (Near-minimizer)**.** There exists a measurable sequence $\widehat{h}_n \equiv h_{K_n}^{\theta, \widehat{W}} \in \mathcal{H}_{K_n}(M_n)$ such that

$$L_n(\widehat{h}_n) \leq \inf_{g \in \mathcal{H}_{K_n}(M_n)} L_n(g) + R_n, \quad \text{where} \quad R_n = O_P(\varepsilon_n^2) \quad \text{and} \quad \varepsilon_n \to 0 \quad \text{as} \quad n \to \infty.$$

We now introduce a best sieve approximation of h with respect to the pseudo-metric. Assume that the infimum is attained by a measurable selection by $\pi_n h \in \mathcal{H}_{K_n}(M_n)$, and define it by

$$\rho_\circ(\pi_n h, h) := \inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_\circ. \tag{E.5}$$

Define the projection loss discrepancy by

$$\chi(h, \pi_n h) := \mathbb{E}_Z[q(Z_i, \pi_n h, h)] = \mathbb{E}_Z[\ell(Z_i, \pi_n h) - \ell(Z_i, h)].$$

Assumption 21(b)-(e) is imposed only on the local class $\mathcal{H}_{K_n}(M_n; \eta_0)$. In the subsequent proof of our result in Theorem E.4, these bounds can be applied to $\widehat{h}_n$ provided that $\widehat{h}_n \in \mathcal{H}_{K_n}(M_n; \eta_0)$ with probability tending to one. The uniformity of the entropy integral in Assumption 21(d) posed for δ_n allows us to derive the shell tail bound of Lemma E.3(ii) below, and is stronger than imposing that (E.2) hold only for $\delta = \delta_n$. Comparable uniform conditions are also imposed, e.g., in Shen and Wong (1994), where an explicit rate for the bracketing entropy is assumed (cf. their Assumption C3'). In fact, when considering RWNNs, we show in Lemma E.6 that such uniformity is not restrictive since the entropy of local loss difference classes does not depend explicitly on the chosen radius δ. The following lemma provides the localization result under a uniform law of large numbers, global well-separatedness of the population criterion away from h, and under vanishing population excess risk of the sieve projection.

Lemma E.2 (Preliminary localization)**.** *Let Assumption 22 hold. Assume that the following uniform law of large numbers holds:*

$$\sup_{g \in \mathcal{H}_{K_n}(M_n)} |Q_n(g, h) - Q(g, h)| = o_P(1). \tag{E.6}$$

Additionally, assume global well-separation condition: for every $\eta > 0$ there exists $\gamma(\eta) > 0$, independent of n and of the realization of $\theta_1, \ldots, \theta_{K_n}$, such that

$$\inf_{\{g \in \mathcal{H}_{K_n}(M_n) | \rho_\circ(g,h) \geq \eta\}} Q(g, h) \geq \gamma(\eta). \tag{E.7}$$

Finally, assume that

$$Q(\pi_n h, h) = \chi(h, \pi_n h) = o_P(1). \tag{E.8}$$

Then, for every fixed $\eta_0 > 0$,

$$\mathbb{P}(\rho_\circ(\widehat{h}_n, h) > \eta_0) \to 0.$$

Proof of Lemma E.2. By Assumption 22,

$$Q_n(\widehat{h}_n, h) \le Q_n(\pi_n h, h) + R_n,$$

where $R_n = O_P(\varepsilon_n^2) = o_P(1)$. Since $\pi_n h \in \mathcal{H}_{K_n}(M_n)$, (E.6) gives

$$Q_n(\pi_n h, h) = Q(\pi_n h, h) + o_P(1) = o_P(1),$$

where in the last equality we used hypothesis (E.8). Hence, for some $U_n = o_P(1)$,

$$Q_n(\widehat{h}_n, h) \le U_n.$$

Since $\widehat{h}_n \in \mathcal{H}_{K_n}(M_n)$, the uniform law of large numbers (E.6) implies

$$Q(\widehat{h}_n, h) \le U_n + \sup_{g \in \mathcal{H}_{K_n}(M_n)} |Q_n(g, h) - Q(g, h)| =: V_n,$$

where $V_n = o_P(1)$.

On the event $\{\rho_\circ(\widehat{h}_n, h) > \eta_0\}$, (E.7) gives $Q(\widehat{h}_n, h) \ge \gamma(\eta_0)$. Therefore

$$\mathbb{P}(\rho_\circ(\widehat{h}_n, h) > \eta_0) \le \mathbb{P}(V_n \ge \gamma(\eta_0)) \to 0,$$

since $\gamma(\eta_0) > 0$ is fixed. □

The following lemma isolates the two empirical process estimates corresponding to those used in the proof of Theorem 1 of Chen and Shen (1998), and hence also the sieve-rate argument in the proof of Theorem 3.1 of Chen and White (1999). Part (i) controls the empirical fluctuation at the sieve approximation $\pi_n h$, corresponding to the single point term treated separately in that proof. Part (ii) is the shell-restricted, two-sided version of the large deviation inequality in Theorem 3 in Chen and Shen (1998). All probabilities and stochastic order statements are understood conditionally on the realized hidden-layer parameters $\theta_1, \dots, \theta_{K_n}$, with the corresponding bounds and constants uniform over their realization.

Lemma E.3 (Local modulus and shell tail maximal inequality)**.** *Define*

$$\mu_n(g, h) := (Q_n - Q)(g, h) \quad \text{and} \quad a_n := \max\{\delta_n, \rho_\circ(\pi_n h, h)\}.$$

Let Assumptions 20 and 21 (c)–(e) hold on $\mathcal{H}_{K_n}(M_n; \eta_0)$*, and let* $\Lambda_n \ge 1$ *be as in Assumption 21 and chosen sufficiently large to dominate* $C_{2,n}$*,* $C_{3,n}$*, and the fixed algebraic combinations of these constants appearing below. Let* $a_n \to 0$*,*

$$n a_n^2 / \Lambda_n \to \infty, \quad \text{and} \quad \Lambda_n n^{-\kappa/2} a_n^{-(2-s)\kappa} \to 0. \tag{E.9}$$

Then the following hold:

(i)

$$|\mu_n(\pi_n h, h)| = O_P(a_n^2). \tag{E.10}$$

(ii) For every fixed $c > 0$*, there exist constants* $A_0, A_1, A_2 > 0$*, independent of the realization of* $\theta_1, \dots, \theta_{K_n}$*, such that, for all* $x \ge 1$ *and all sufficiently large* n*,*

$$\mathbb{P}\left(\sup_{\{g \in \mathcal{H}_{K_n}(M_n; \eta_0) \,|\, x a_n < \rho_\circ(g,h) \le 2x a_n\}} |\mu_n(g, h)| \ge c(x a_n)^2 \right) \le A_0 e^{-A_1 n x^{2-s} a_n^2 / \Lambda_n} + A_2 \Lambda_n n^{-\kappa/2} a_n^{-(2-s)\kappa} x^{-(2-s)\kappa}. \tag{E.11}$$

(iii) For every sequence $b_n \geq a_n$ and every $C > 0$, define

$$p_{j,n}(C) := A_0 \exp\left\{-A_1 \frac{n a_n^2}{\Lambda_n}\left(\frac{2^j C b_n}{a_n}\right)^{2-s}\right\} + A_2 \Lambda_n n^{-\kappa/2} a_n^{-(2-s)\kappa}\left(\frac{2^j C b_n}{a_n}\right)^{-(2-s)\kappa}.$$

Then

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sum_{j\geq 0} p_{j,n}(C) = 0. \tag{E.12}$$

Proof of Lemma E.3. For part (i), we proceed as in Theorem 1 of Chen and Shen (1998). Since $a_n \to 0$, and $\rho_\circ(\pi_n h, h) \leq a_n$, Assumption 21(e) yields that $|q(Z_i, \pi_n h, h)| \leq a_n^s \psi_n(Z_i)$. Consequently, $\mathbb{E}_Z\left[|q(Z_i, \pi_n h, h)|^\kappa\right] \leq a_n^{s\kappa}\mathbb{E}_Z\left[|\psi_n(Z_i)|^\kappa\right] \leq C_{3,n} a_n^{s\kappa}$. Recall that, by definition, $\mu_n(\pi_n h, h) = \frac{1}{n}\sum_{i=1}^n \{q(Z_i, \pi_n h, h) - \mathbb{E}_Z[q(Z_i, \pi_n h, h)]\}$. Applying the same partial-sum moment inequality as in the bound for P_2 in the proof of Theorem 1 of Chen and Shen (1998), there exists a constant $C_{\text{mom}} > 0$, independent of n and of the realization of the hidden-layer parameters, such that, for every $u > 0$,

$$\begin{aligned}\mathbb{P}\left(|\mu_n(\pi_n h, h)| > u a_n^2\right) &\leq 2^\kappa C_{\text{mom}} u^{-\kappa} C_{3,n} n^{-\kappa/2} a_n^{-(2-s)\kappa} \\ &\leq 2^\kappa C_{\text{mom}} u^{-\kappa} \Lambda_n n^{-\kappa/2} a_n^{-(2-s)\kappa} \to 0\end{aligned}$$

where we used the second condition in (E.9) and which proves (E.10).

In part (ii), under Assumption 20 (i) we follow the i.i.d. argument of Shen and Wong (1994), as used in the proof of Theorem 3.1 of Chen and White (1999), while under Assumption 20 (ii), we follow the proof of Theorem 3 of Chen and Shen (1998), applied separately to $q(z, g, h) = \ell(z, g) - \ell(z, h)$ and to $-q(z, g, h)$. Restricting the supremum to the shell and using the union bound gives

$$\begin{aligned}\mathbb{P}\left(\sup_{\{g\in\mathcal{H}_{K_n}(M_n;\eta_0)\,|\,x a_n < \rho_\circ(g,h)\leq 2x a_n\}} |\mu_n(g,h)| \geq c(x a_n)^2\right) \\ \leq 14 e^{-C_n n x^{2-s} a_n^2} + D_n n^{-\kappa/2} a_n^{-(2-s)\kappa} x^{-(2-s)\kappa},\end{aligned}$$

where C_n and D_n are deterministic and uniform over the realization of the hidden-layer parameters, and C_n^{-1} and D_n depend on $C_{2,n}$ and $C_{3,n}$ in Assumption 21(c) and (e), via their fixed powers. Hence, by the choice of Λ_n, there exist constants $A_1, A_2 > 0$, independent of n and of the realization of the hidden-layer parameters, such that $C_n \geq \frac{A_1}{\Lambda_n}$ and $D_n \leq A_2 \Lambda_n$. Substituting these in the expression above and taking $A_0 = 14$ yields (E.11) as required.

For part (iii), the polynomial term is summable since $b_n \geq a_n$ gives $2^j C b_n / a_n \geq 2^j C$ and $(2-s)\kappa > 0$. Indeed,

$$\begin{aligned}\sum_{j\geq 0} A_2 \Lambda_n n^{-\kappa/2} a_n^{-(2-s)\kappa}\left(\frac{2^j C b_n}{a_n}\right)^{-(2-s)\kappa} &= A_2 \Lambda_n n^{-\kappa/2} (C b_n)^{-(2-s)\kappa} \sum_{j\geq 0} 2^{-j(2-s)\kappa} \\ &\leq A_3 A_2 \Lambda_n C^{-(2-s)\kappa} n^{-\kappa/2} a_n^{-(2-s)\kappa},\end{aligned} \tag{E.13}$$

with $A_3 := \sum_{j\geq 0} 2^{-j(2-s)\kappa} < \infty$, where the last inequality follows from $b_n \geq a_n$. The right hand side of this inequality tends to zero under (E.9). For the exponential term,

$$\begin{aligned}\sum_{j\geq 0} A_0 \exp\left\{-A_1 \frac{n a_n^2}{\Lambda_n}\left(\frac{2^j C b_n}{a_n}\right)^{2-s}\right\} &\leq A_0 \sum_{j\geq 0} \exp\left\{-A_1 \frac{n a_n^2}{\Lambda_n} C^{2-s} 2^{j(2-s)}\right\} \\ &\leq A_0 \frac{\exp\{-A_1 C^{2-s} n a_n^2/\Lambda_n\}}{1 - \exp\{-A_1 C^{2-s}(2^{2-s}-1) n a_n^2/\Lambda_n\}},\end{aligned}$$

where in the last inequality we used $2-s>0$ and $2^{j(2-s)} \geq 1+j(2^{2-s}-1)$. This expression tends to zero for every fixed $C>0$ by $na_n^2/\Lambda_n \to \infty$, which together with (E.13) proves (E.12). □

The preliminary localization result in Lemma E.2 reduces the rate analysis to the local neighborhood $\mathcal{H}_{K_n}(M_n;\eta_0)$, while Lemma E.3 provides the result which controls the empirical deviations on dyadic shells within this neighborhood. These results yield a version of the generic sieve argument of Chen and Shen (1998), as extended to approximate sieve extremum estimators by Chen and White (1999). Here we keep the approximation error and the corresponding population criterion discrepancy separate, which will be useful for the RWNN sieves considered below.

Theorem E.4 (Generic sieve convergence rate)**.** *Suppose Assumption 21(b)-(e), Assumption 22, and the conditions of Lemmas E.2 and E.3 hold. Then*

$$\|h-\widehat{h}_n\|_\circ = O_P\left(\max\left\{\delta_n, \varepsilon_n, \|h-\pi_n h\|_\circ, [\chi(h,\pi_n h)]^{1/2}\right\}\right). \tag{E.14}$$

Proof of Theorem E.4. Fix a realization of $\theta_1,\dots,\theta_{K_n}$. Conditional on this realization, the sieve $\mathcal{H}_{K_n}(M_n)$ is deterministic. The constants and deterministic sequences in the assumptions are uniform over the realization of $\theta_1,\dots,\theta_{K_n}$. Hence the conditional bound obtained below also holds under the joint law of the data and random features.

Define

$$r_n := \max\{\delta_n, \varepsilon_n, \rho_\circ(\pi_n h, h), [\chi(h,\pi_n h)]^{1/2}\}, \tag{E.15}$$

where $\rho_\circ(g,h) := \|g-h\|_\circ$ as in Assumption 21 and recall the definitions introduced in (E.3) and (E.4). Then

$$\rho_\circ(\pi_n h, h) \leq r_n, \qquad Q(\pi_n h, h) = \chi(h,\pi_n h) \leq r_n^2. \tag{E.16}$$

It suffices to prove

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}(\rho_\circ(\widehat{h}_n, h) > C r_n) = 0.$$

By Assumption 22,

$$Q_n(\widehat{h}_n, h) \leq Q_n(\pi_n h, h) + R_n, \tag{E.17}$$

where $R_n = O_P(\varepsilon_n^2) = O_P(r_n^2)$. Set

$$a_n := \max\{\delta_n, \rho_\circ(\pi_n h, h)\}.$$

Then $a_n \leq r_n$. Since $\pi_n h \in \mathcal{H}_{K_n}(M_n)$ and $\rho_\circ(\pi_n h, h) \leq a_n \to 0$, we have $\pi_n h \in \mathcal{H}_{K_n}(M_n;\eta_0)$ for all sufficiently large n. Hence, (E.10) together with $a_n \leq r_n$ gives

$$|Q_n(\pi_n h, h) - Q(\pi_n h, h)| = O_P(a_n^2) = O_P(r_n^2).$$

Combining this with $R_n = O_P(r_n^2)$, for every $\eta>0$ there exists $C_\eta>0$ such that, for all sufficiently large n,

$$\mathbb{P}(A_{n,C_\eta}^c) \leq \eta, \tag{E.18}$$

where

$$A_{n,C} := \left\{ R_n + |Q_n(\pi_n h, h) - Q(\pi_n h, h)| \leq \frac{C_1}{4} C^2 r_n^2 \right\}.$$

Since $A_{n,C}$ is increasing in C, the same bound holds for any $C \geq C_\eta$.

Fix $C>0$ and define

$$\mathcal{S}_{j,n}^{\eta_0} := \{g \in \mathcal{H}_{K_n}(M_n;\eta_0) \mid 2^j C r_n < \rho_\circ(g,h) \leq 2^{j+1} C r_n\}, \quad j \geq 0,$$

where we recall that $\mathcal{H}_{K_n}(M_n;\eta_0) = \{g \in \mathcal{H}_{K_n}(M_n) \mid \rho_\circ(g,h) \leq \eta_0\}$.

$$J_n(C) := \max\{j \geq 0 \mid 2^j C r_n < \eta_0\}$$

so that

$$\{Cr_n < \rho_\circ(\widehat{h}_n, h) \le \eta_0\} \subseteq \bigcup_{j=0}^{J_n(C)} \{\widehat{h}_n \in \mathcal{S}_{j,n}^{\eta_0}\},$$

with the convention that the union over $j = 0, \dots, J_n(C)$ is empty if the indexing set is empty. If the set defining $J_n(C)$ is empty, then $Cr_n \ge \eta_0$ and $\{Cr_n < \rho_\circ(\widehat{h}_n, h) \le \eta_0\} = \varnothing$, meaning that the corresponding probability is zero.

Next, $0 \le j \le J_n(C)$, on $\{\widehat{h}_n \in \mathcal{S}_{j,n}^{\eta_0}\}$, Assumption 21(b) gives

$$Q(\widehat{h}_n, h) \ge 2C_1(2^j Cr_n)^2. \tag{E.19}$$

On the other hand, by (E.17),

$$Q(\widehat{h}_n, h) \le Q(\pi_n h, h) + |Q_n(\pi_n h, h) - Q(\pi_n h, h)| + R_n + |Q_n(\widehat{h}_n, h) - Q(\widehat{h}_n, h)|. \tag{E.20}$$

Using (E.16), (E.19), and (E.20), on $A_{n,C} \cap \{\widehat{h}_n \in \mathcal{S}_{j,n}^{\eta_0}\}$ we have

$$2C_1(2^j Cr_n)^2 \le r_n^2 + \frac{C_1}{4} C^2 r_n^2 + |Q_n(\widehat{h}_n, h) - Q(\widehat{h}_n, h)|. \tag{E.21}$$

Choose $C \ge \max\{1, 2/\sqrt{C_1}\}$. With this choice

$$r_n^2 + \frac{C_1}{4} C^2 r_n^2 \le \frac{C_1}{2} C^2 r_n^2 \le \frac{C_1}{2} (2^j Cr_n)^2.$$

Hence, on $A_{n,C} \cap \{\widehat{h}_n \in \mathcal{S}_{j,n}^{\eta_0}\}$, using this bound in (E.21) and rearranging the terms after substitution, one obtains the following expression:

$$|Q_n(\widehat{h}_n, h) - Q(\widehat{h}_n, h)| \ge \frac{3C_1}{2} (2^j Cr_n)^2.$$

In particular, taking any fixed constant $c_0 \in (0, 3C_1/2]$ (for example, $c_0 = C_1$) gives

$$A_{n,C} \cap \{\widehat{h}_n \in \mathcal{S}_{j,n}^{\eta_0}\} \subseteq \left\{ \sup_{g \in \mathcal{S}_{j,n}^{\eta_0}} |(Q_n - Q)(g, h)| \ge c_0 (2^j Cr_n)^2 \right\}.$$

Set

$$x_{j,n} := \frac{2^j Cr_n}{a_n} \ge 1,$$

so that, by the definition of $\mathcal{S}_{j,n}^{\eta_0}$,

$$\mathcal{S}_{j,n}^{\eta_0} = \{g \in \mathcal{H}_{K_n}(M_n; \eta_0) \mid x_{j,n} a_n < \rho_\circ(g, h) \le 2x_{j,n} a_n\}$$

is exactly the shell appearing in (E.11) at $x = x_{j,n}$. Applying the shell tail bound (E.11) with $x = x_{j,n}$ and $c = c_0$ gives, for $0 \le j \le J_n(C)$,

$$\mathbb{P}(A_{n,C} \cap \{\widehat{h}_n \in \mathcal{S}_{j,n}^{\eta_0}\}) \le p_{j,n}(C),$$

where $p_{j,n}(C)$ is as in Lemma E.3 (iii) with $b_n = r_n$. Therefore, by (E.12),

$$\lim_{C \to \infty} \limsup_{n \to \infty} \sum_{j=0}^{J_n(C)} p_{j,n}(C) = 0.$$

Hence

$$\mathbb{P}(Cr_n < \rho_\circ(\widehat{h}_n, h) \le \eta_0) \le \mathbb{P}\left(\bigcup_{j=0}^{J_n(C)} \{\widehat{h}_n \in \mathcal{S}_{j,n}^{\eta_0}\}\right)$$
$$\le \mathbb{P}(A_{n,C}^c) + \sum_{j=0}^{J_n(C)} \mathbb{P}(A_{n,C} \cap \{\widehat{h}_n \in \mathcal{S}_{j,n}^{\eta_0}\}) \le \mathbb{P}(A_{n,C}^c) + \sum_{j=0}^{J_n(C)} p_{j,n}(C).$$

Taking $\limsup_{n\to\infty}$, then $C \to \infty$, and finally using (E.18) with $\eta \downarrow 0$, gives

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}(Cr_n < \rho_\circ(\widehat{h}_n, h) \le \eta_0) = 0. \tag{E.22}$$

By Lemma E.2,

$$\mathbb{P}(\rho_\circ(\widehat{h}_n, h) > \eta_0) \to 0. \tag{E.23}$$

Combining (E.22) and (E.23), we get

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{P}(\rho_\circ(\widehat{h}_n, h) > Cr_n) = 0.$$

Therefore $\rho_\circ(\widehat{h}_n, h) = O_P(r_n)$. Since $\rho_\circ(g, h) = \|g - h\|_\circ$, the definition of r_n in (E.15) gives (E.14), which completes the proof. □

We note that the result in Theorem E.4 does not require $K_n < n$. Any restriction on the relative growth of K_n with respect to n enters through the verification of its assumptions for the particular RWNN sieve considered in our setting, in particular through the entropy and stochastic equicontinuity conditions.

We now verify the RWNN-specific ingredients entering Theorem E.4. More specifically, in Lemma E.5, we provide the bracketing entropy of $\mathcal{H}_{K_n}(M_n)$, which determines δ_n, Corollary E.1 quantifies the sieve approximation error $\|h - \pi_n h\|_\circ$, and we bound the projection loss discrepancy $\chi(h, \pi_n h)$.

Lemma E.5 (Entropy of RWNN sieve with ℓ_2-bounded readout weights)**.** *Let Assumption 1 hold, assume that $\Pi \subset \mathbb{R}^{d+1}$ is compact and define $M_\sigma := \sup_{x\in\mathcal{X}} \sup_{\theta=(A^\top,B)^\top\in\Pi} |\sigma(A^\top x + B)| < \infty$. For $M_n > 0$, let the constrained RWNN class $\mathcal{H}_{K_n}(M_n)$ be as in (E.1). Let $\theta_k = (A_k^\top, B_k)^\top \in \Pi$, $k = 1, \dots, K_n$. Then, for every realization of $\theta_1, \dots, \theta_{K_n}$ and every $r > 0$,*

$$\log \mathcal{N}_{[]}\left(r, \mathcal{H}_{K_n}(M_n), \|\cdot\|_{L^2(X)}\right) \le K_n \log\left(\frac{4M_n M_\sigma \sqrt{K_n}}{r} + 1\right). \tag{E.24}$$

Consequently, for every $\delta \in (0, 1)$,

$$\delta^{-2} \int_{\delta^2}^{\delta} \sqrt{\log \mathcal{N}_{[]}\left(r, \mathcal{H}_{K_n}(M_n), \|\cdot\|_{L^2(X)}\right)}\, dr \lesssim \delta^{-1} \sqrt{K_n \log\left(\frac{1 + 4M_n M_\sigma \sqrt{K_n}}{\delta^2}\right)}. \tag{E.25}$$

Let

$$\delta_{n,2} := \inf\left\{\delta > 0 \mid \delta^{-2} \int_{\delta^2}^{\delta} \sqrt{\log \mathcal{N}_{[]}\left(r, \mathcal{H}_{K_n}(M_n), \|\cdot\|_{L^2(X)}\right)}\, dr \lesssim \sqrt{n}\right\}. \tag{E.26}$$

If

$$M_n \asymp \sqrt{\frac{\log(K_n)}{K_n}}, \qquad \log(K_n) \asymp \log(n), \qquad K_n \log(K_n) = o(n),$$

then

$$\delta_{n,2} \lesssim \sqrt{\frac{K_n \log(K_n)}{n}}.$$

Proof of Lemma E.5. Fix a realization of $\theta_1, \ldots, \theta_{K_n} \in \Pi$. Conditional on this realization, $\mathcal{H}_{K_n}(M_n)$ is a K_n-dimensional parametrized class indexed by $W \in B_2(M_n) := \{W \in \mathbb{R}^{K_n} : \|W\| \le M_n\}$.

For $f_W(x) := W^\top g_{K_n}(x)$ and any $x \in \mathcal{X}$, the Cauchy-Schwarz inequality and $|\sigma(A_k^\top x + B_k)| \le M_\sigma$ give

$$|f_W(x)| \le \|W\| \|g_{K_n}(x)\| \le \|W\| \sqrt{K_n} M_\sigma.$$

Hence, for any $W, \widetilde{W} \in B_2(M_n)$,

$$\sup_{x \in \mathcal{X}} |f_W(x) - f_{\widetilde{W}}(x)| \le \|W - \widetilde{W}\| \sqrt{K_n}\, M_\sigma. \tag{E.27}$$

Let $\eta > 0$, and let $\mathcal{W}_\eta$ be an η-net of $B_2(M_n)$ under $\|\cdot\|$. By the standard volumetric bound (see, for example, Corollary 4.2.11 in Vershynin (2026)),

$$|\mathcal{W}_\eta| \le \left(\frac{2M_n}{\eta} + 1\right)^{K_n}.$$

For every $W \in B_2(M_n)$ there exists $\bar{W} \in \mathcal{W}_\eta$ such that $\|W - \bar{W}\| \le \eta$, so by (E.27), $\sup_{x \in \mathcal{X}} |f_W(x) - f_{\bar{W}}(x)| \le \eta \sqrt{K_n} M_\sigma$. Choosing $\eta = r/(2\sqrt{K_n} M_\sigma)$ gives $\sup_{x \in \mathcal{X}} |f_W(x) - f_{\bar{W}}(x)| \le \frac{r}{2}$. It follows that, for each $x \in \mathcal{X}$,

$$f_{\bar{W}}(x) - \frac{r}{2} \le f_W(x) \le f_{\bar{W}}(x) + \frac{r}{2}.$$

The brackets $\left[f_{\bar{W}} - \frac{r}{2},\ f_{\bar{W}} + \frac{r}{2}\right]$, $\bar{W} \in \mathcal{W}_\eta$, cover $\mathcal{H}_{K_n}(M_n)$ pointwise, and each has $L^2(X)$-diameter equal to r. Hence they form an r-bracketing cover of $\mathcal{H}_{K_n}(M_n)$ with respect to $\|\cdot\|_{L^2(X)}$.

Hence,

$$\mathcal{N}_{[]}\left(r, \mathcal{H}_{K_n}(M_n), \|\cdot\|_{L^2(X)}\right) \le |\mathcal{W}_\eta| \le \left(\frac{4M_n M_\sigma \sqrt{K_n}}{r} + 1\right)^{K_n},$$

which proves (E.24) after taking logarithms.

We next bound the entropy integral. Set $A_n := 4M_n M_\sigma \sqrt{K_n}$. By (E.24),

$$\delta^{-2} \int_{\delta^2}^{\delta} \sqrt{\log \mathcal{N}_{[]}\left(r, \mathcal{H}_{K_n}(M_n), \|\cdot\|_{L^2(X)}\right)}\, dr \le \delta^{-2} \int_{\delta^2}^{\delta} \sqrt{K_n \log\left(\frac{A_n}{r} + 1\right)}\, dr.$$

For $r \in [\delta^2, \delta]$ and $\delta \in (0,1)$, $\log(A_n/r + 1) \le \log((1 + A_n)/r) \le \log\left((1 + A_n)/\delta^2\right)$. Hence,

$$\delta^{-2} \int_{\delta^2}^{\delta} \sqrt{K_n \log\left(\frac{A_n}{r} + 1\right)}\, dr \le \delta^{-1} \sqrt{K_n \log\left(\frac{1 + A_n}{\delta^2}\right)},$$

which proves (E.25).

Next, take $\bar{\delta}_n := C\sqrt{(K_n \log(K_n))/n}$, with $C > 0$ sufficiently large. Since $M_n \asymp \sqrt{\log(K_n)/K_n}$, one has that $A_n = 4M_n M_\sigma \sqrt{K_n} \lesssim \sqrt{\log(K_n)}$. Moreover, $\log(K_n) \asymp \log(n)$ and $K_n \log(K_n) = o(n)$ imply $\bar{\delta}_n \to 0$ and $|\log \bar{\delta}_n| \lesssim \log(n) \asymp \log(K_n)$. Hence, it holds that $\log\left((1 + A_n)/\bar{\delta}_n^2\right) \lesssim \log(K_n)$. Substituting $\delta = \bar{\delta}_n$ in (E.25) gives

$$\bar{\delta}_n^{-2} \int_{\bar{\delta}_n^2}^{\bar{\delta}_n} \sqrt{\log \mathcal{N}_{[]}\left(r, \mathcal{H}_{K_n}(M_n), \|\cdot\|_{L^2(X)}\right)}\, dr \lesssim \sqrt{n}.$$

Thus the entropy condition defining $\delta_{n,2}$ in (E.26) is satisfied for $\delta = \bar{\delta}_n$, which yields

$$\delta_{n,2} \lesssim \sqrt{\frac{K_n \log(K_n)}{n}},$$

as required. □

The entropy condition in Theorem E.4 is stated for the localized class of loss differences while Lemma E.5 provides the entropy of the RWNN class itself. The following lemma allows us to connect the latter bound to the loss class under a Lipschitz condition on the loss. More explicitly, we show that a Lipschitz constant growing polylogarithmically does not change the resulting loss class critical radius order.

Lemma E.6 (Loss-entropy transfer)**.** *Assume that on the prediction range of $\mathcal{H}_{K_n}(M_n;\eta_0)\cup\{h\}$, ℓ is Lipschitz in its second argument with the constant $L_{\ell,n}>0$. Let*

$$\mathcal{M}_n(\delta) := \{z \mapsto \ell(z,g)-\ell(z,h) \mid g\in\mathcal{H}_{K_n}(M_n;\eta_0),\ \rho_\circ(g,h)\le\delta\}.$$

Then, for every $w>0$ and every $\delta>0$,

$$\log\mathcal{N}_{[]}\big(w,\mathcal{M}_n(\delta),L^2(P_Z)\big) \le \log\mathcal{N}_{[]}\left(\frac{w}{2L_{\ell,n}},\mathcal{H}_{K_n}(M_n;\eta_0),L^2(X)\right). \tag{E.28}$$

Consequently, if $L_{\ell,n}\le\Lambda_n$, $\Lambda_n=O((\log(n))^m)$, $M_n\asymp\sqrt{\log(K_n)/K_n}$, $\log(K_n)\asymp\log(n)$, and $K_n\log(K_n)=o(n)$, then the loss-class critical radius satisfies

$$\delta_n \lesssim \sqrt{\frac{K_n\log(K_n)}{n}}.$$

Proof of Lemma E.6. Fix $w>0$ and set $\eta:=w/(2L_{\ell,n})$. Let $\{[g_j^L,g_j^U]\}_{j=1}^N$ be an η-bracketing cover of $\mathcal{H}_{K_n}(M_n;\eta_0)$ in $L^2(X)$. Thus, for every $g\in\mathcal{H}_{K_n}(M_n;\eta_0)$, there is j such that $g_j^L(x)\le g(x)\le g_j^U(x)$ for P_X-a.e. x and $\|g_j^U-g_j^L\|_{L^2(X)}\le\eta$. Discard empty brackets and for each j choose $g_j\in\mathcal{H}_{K_n}(M_n;\eta_0)\cap[g_j^L,g_j^U]$. If $g\in[g_j^L,g_j^U]$, then $|g(x)-g_j(x)|\le g_j^U(x)-g_j^L(x)$ for P_X-a.e. x. Define $q_j^L(z):=\ell(z,g_j)-\ell(z,h)-L_{\ell,n}\big(g_j^U(x)-g_j^L(x)\big)$, $q_j^U(z):=\ell(z,g_j)-\ell(z,h)+L_{\ell,n}\big(g_j^U(x)-g_j^L(x)\big)$, with $z=(y,x)$.

By the Lipschitz assumption in the hypotheses of the lemma,

$$q_j^L(z)\le\ell(z,g)-\ell(z,h)\le q_j^U(z)\quad\text{for } P_Z\text{-a.e. } z.$$

Hence the brackets $\{[q_j^L,q_j^U]\}_{j=1}^N$ cover $\mathcal{M}_n(\delta)$. Their widths satisfy

$$\|q_j^U-q_j^L\|_{L^2(Z)}\le 2L_{\ell,n}\|g_j^U-g_j^L\|_{L^2(X)}\le w.$$

Thus

$$\mathcal{N}_{[]}\big(w,\mathcal{M}_n(\delta),L^2(Z)\big)\le\mathcal{N}_{[]}\left(\frac{w}{2L_{\ell,n}},\mathcal{H}_{K_n}(M_n;\eta_0),L^2(X)\right),$$

which proves (E.28).

Combining (E.28) with Lemma E.5, with $A_n:=4M_nM_\sigma\sqrt{K_n}$, gives

$$\log\mathcal{N}_{[]}\big(w,\mathcal{M}_n(\delta),L^2(Z)\big)\le K_n\log\left(\frac{2A_nL_{\ell,n}}{w}+1\right).$$

Therefore, for $\delta\in(0,1)$,

$$\delta^{-2}\int_{\delta^2}^{\delta}\sqrt{\log\mathcal{N}_{[]}\big(w,\mathcal{M}_n(\delta),L^2(Z)\big)}\,dw\lesssim\delta^{-1}\sqrt{K_n\log\left(\frac{1+2A_nL_{\ell,n}}{\delta^2}\right)}.$$

Under $M_n\asymp\sqrt{\log(K_n)/K_n}$, one has $A_n\lesssim\sqrt{\log(K_n)}$. If $L_{\ell,n}\le\Lambda_n$ with $\Lambda_n=O((\log(n))^m)$ and $\log(K_n)\asymp\log(n)$, then at $\bar\delta_n:=C\sqrt{K_n\log(K_n)/n}$, with $C>0$ sufficiently large, the logarithmic factor satisfies

$$\log\left(\frac{1+2A_nL_{\ell,n}}{\bar\delta_n^2}\right)\lesssim\log(K_n),$$

provided $K_n \log(K_n) = o(n)$. Substituting $\delta = \bar{\delta}_n$ gives

$$\bar{\delta}_n^{-2} \int_{\bar{\delta}_n^2}^{\bar{\delta}_n} \sqrt{\log \mathcal{N}_{[]}\big(w, \mathcal{M}_n(\bar{\delta}_n), L^2(Z)\big)}\, dw \lesssim \sqrt{n}.$$

Hence, by the definition of the loss-class critical radius δ_n,

$$\delta_n \lesssim \sqrt{\frac{K_n \log(K_n)}{n}},$$

as required. □

In the results below we will find it useful to impose the following domination condition.

Assumption 23 (Pseudo-metric dominated by $L^2(X)$)**.** There exists a constant $C_\circ > 0$, independent of n and of the realized $\theta_1, \dots, \theta_{K_n} \in \Pi$, such that

$$\|g_1 - g_2\|_\circ \le C_\circ \|g_1 - g_2\|_{L^2(X)}$$

for all $g_1, g_2 \in \operatorname{span}\big(\mathcal{H}_{K_n}(M_n) \cup \{h, \mathbf{1}\}\big)$, where $\mathbf{1}$ denotes the constant function on $\mathcal{X}$.

Assumption 23 ensures that the $L^2(X)$ approximation bounds derived for the RWNN approximation theorem also control the pseudo-metric error appearing in Theorem E.4. In particular,

$$\|h - \pi_n h\|_\circ = \inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_\circ \le C_\circ \inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_{L^2(X)}.$$

The approximation result of Theorem 3.2 is constructed without imposing the readout constraint defining the statistical sieve. Lemma 3.3 shows that the corresponding control-neighbors weights satisfy the required ℓ_2 bound with probability tending to one. Hence the same approximation rate is attained within the constrained RWNN class.

Corollary E.1 (Constrained RWNN approximation)**.** *Under the assumptions of Theorem 3.2 and Lemma 3.3, define*

$$M_n := C_\alpha M_f \sqrt{\frac{\log(K_n)}{K_n}}$$

for $\alpha > 0$ and $C_\alpha > 0$ as in Lemma 3.3. Then

$$\inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_{L^2(X)} = O_P\left(K_n^{-1/2-\beta/(d+1)}\right),$$

where the probability is with respect to $\theta_1, \dots, \theta_{K_n}$.

Proof of Corollary E.1. Let $W^* \in \mathbb{R}^{K_n}$ be the $\sigma(\theta_1, \dots, \theta_{K_n})$-measurable control-neighbors weight vector produced by Theorem 3.2, with components $W_k^* = \omega_{k,K_n} f(A_k, B_k)$. By Theorem 3.2,

$$\|h - h_{K_n}^{\theta, W^*}\|_{L^2(X)} = O_P\left(K_n^{-1/2-\beta/(d+1)}\right). \tag{E.29}$$

Moreover, by Lemma 3.3, for the chosen $\alpha > 0$ there exist constants $C_\alpha, C'_\alpha > 0$ such that

$$\mathbb{P}_{A,B}\left(\|W^*\| \le C_\alpha M_f \sqrt{\frac{\log(K_n)}{K_n}}\right) \ge 1 - C'_\alpha K_n^{-\alpha}. \tag{E.30}$$

By the definition of M_n, the event in (E.30) is $\{\|W^*\| \le M_n\}$. On this event, $h_{K_n}^{\theta, W^*} \in \mathcal{H}_{K_n}(M_n)$, and therefore

$$\inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_{L^2(X)} \le \|h - h_{K_n}^{\theta, W^*}\|_{L^2(X)}.$$

Since $\mathbb{P}_{A,B}(\|W^*\| \le M_n) \to 1$, combining this inequality with (E.29) yields

$$\inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_{L^2(X)} = O_P\left(K_n^{-1/2-\beta/(d+1)}\right),$$

as required. □

Combining Assumption 23 with Corollary E.1 allows us to write the constrained RWNN approximation rate in terms of the pseudo-metric $\|\cdot\|_\circ$, appropriate for Theorem E.4 to be applied.

Corollary E.2 (RWNN approximation in pseudo-metric)**.** *Under Assumption 23 and the assumptions of Corollary E.1,*

$$\inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_\circ = O_P\left(K_n^{-1/2-\beta/(d+1)}\right).$$

Hence, if additionally the assumptions of Theorem E.4 hold,

$$\|h - \widehat{h}_n\|_\circ = O_P\left(\max\left\{\delta_n, \varepsilon_n, K_n^{-1/2-\beta/(d+1)}, [\chi(h, \pi_n h)]^{1/2}\right\}\right).$$

Proof of Corollary E.2. By definition of the sieve projection in (E.5),

$$\|h - \pi_n h\|_\circ = \inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_\circ.$$

By Assumption 23,

$$\|h - \pi_n h\|_\circ \le C_\circ \inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_{L^2(X)}.$$

Applying Corollary E.1 yields

$$\inf_{g \in \mathcal{H}_{K_n}(M_n)} \|h - g\|_{L^2(X)} = O_P\left(K_n^{-1/2-\beta/(d+1)}\right),$$

and therefore

$$\|h - \pi_n h\|_\circ = O_P\left(K_n^{-1/2-\beta/(d+1)}\right).$$

The second claim follows by substitution into Theorem E.4. □

Similarly to Subsection E.2, we consider an asymptotically underparametrized RWNN sieve case. The choice of K_n satisfies $K_n = o(n)$ and therefore $K_n < n$ for all sufficiently large n.

Theorem E.7 (RWNN convergence rate in the pseudo-metric)**.** *Suppose that the conditions of Corollary E.2 and Lemmas E.5 and E.6 hold. Suppose that*

$$M_n \asymp \sqrt{\frac{\log(K_n)}{K_n}}$$

and that K_n is chosen so that

$$K_n^{2(1+\beta/(d+1))} \log(K_n) \asymp n,$$

meaning $K_n \asymp (n/\log(n))^{\frac{d+1}{2(\beta+d+1)}}$. Then

$$\|h - \widehat{h}_n\|_\circ = O_P\left(\max\left\{(n/\log(n))^{-\frac{1+2\beta/(d+1)}{4(1+\beta/(d+1))}}, \varepsilon_n, [\chi(h, \pi_n h)]^{1/2}\right\}\right).$$

Proof of Theorem E.7. By Corollary E.2,

$$\|h - \widehat{h}_n\|_\circ = O_P\left(\max\left\{\delta_n, \varepsilon_n, K_n^{-1/2-\beta/(d+1)}, [\chi(h, \pi_n h)]^{1/2}\right\}\right), \tag{E.31}$$

and, by Lemma E.5 and Lemma E.6,

$$\delta_n \lesssim \sqrt{\frac{K_n \log(K_n)}{n}}.$$

Using this bound in (E.31) then gives

$$\|h - \widehat{h}_n\|_\circ = O_P\left(\max\left\{\sqrt{\frac{K_n \log(K_n)}{n}}, \varepsilon_n, K_n^{-1/2-\beta/(d+1)}, [\chi(h, \pi_n h)]^{1/2}\right\}\right). \tag{E.32}$$

The choice of K_n is made by balancing the stochastic and approximation terms in (E.32). More precisely, after taking squares, we impose

$$\frac{K_n \log(K_n)}{n} \asymp K_n^{-1-2\beta/(d+1)},$$

or, equivalently,

$$K_n^{2(1+\beta/(d+1))} \log(K_n) \asymp n.$$

This gives

$$K_n \asymp \left(\frac{n}{\log(n)}\right)^{\frac{1}{2(1+\beta/(d+1))}}.$$

Indeed, for this choice one has that $\log(K_n) \asymp \log(n)$ and hence

$$K_n^{2(1+\beta/(d+1))} \log(K_n) \asymp \frac{n}{\log(n)} \log(n) \asymp n.$$

Finally, substituting this choice into the approximation term yields

$$K_n^{-1/2-\beta/(d+1)} \asymp (n/\log(n))^{-\frac{1/2+\beta/(d+1)}{2(1+\beta/(d+1))}} \asymp (n/\log(n))^{-\frac{1+2\beta/(d+1)}{4(1+\beta/(d+1))}}.$$

By the balancing relation, the stochastic term is of the same order

$$\sqrt{\frac{K_n \log(K_n)}{n}} \asymp (n/\log(n))^{-\frac{1+2\beta/(d+1)}{4(1+\beta/(d+1))}}.$$

Substituting this into (E.32) yields the required rate. □

Remark 20 (Choice of sieve radius)**.** The rate $\delta_n \lesssim \sqrt{K_n \log(K_n)/n}$ and the balancing condition $K_n^{2(1+\beta/(d+1))} \log(K_n) \asymp n$ in Theorem E.7 hold for any M_n satisfying $\sqrt{\log(K_n)/K_n} \lesssim M_n \lesssim K_n^C$ for some fixed $C > 0$. The lower bound ensures that the control-neighbors approximant with W^* belongs to $\mathcal{H}_{K_n}(M_n)$ with high probability by Lemma 3.3, while the upper bound implies $\log(K_n M_n) \asymp \log(K_n)$, so that Lemmas E.5 and E.6 yield $\delta_n \lesssim \sqrt{K_n \log(K_n)/n}$. The choice $M_n \asymp \sqrt{\log(K_n)/K_n}$ is the smallest admissible order.

We finally specialize the generic framework to squared-loss regression and provide other examples of practically useful loss functions that fit in this setting.

E.2.1 Squared-loss regression and other examples

Assumption 24 (Squared-loss local empirical-process condition)**.** For the squared-loss criterion with $\|g-h\|_\circ = \|g-h\|_{L^2(X)}$ and the corresponding local class

$$\mathcal{M}_n(\delta) := \{z \mapsto \ell(z,g) - \ell(z,h) \mid g \in \mathcal{H}_{K_n}(M_n;\eta_0),\ \|g-h\|_{L^2(X)} \le \delta\},$$

the conclusions of Lemma E.3 hold with Λ_n growing at most polylogarithmically in n.

Proposition E.1 (Squared-loss regression with conditional noise moments)**.** *Consider the regression model*

$$Y_i = h(X_i) + \epsilon_i$$

with squared loss

$$\ell(Z_i, g) = \frac{1}{2}\big(Y_i - g(X_i)\big)^2$$

and pseudo-metric $\|g-h\|_\circ = \|g-h\|_{L^2(X)}$. Suppose Assumption 20 holds, together with the assumptions of Theorem 3.1, Lemma 3.3, Assumption 22, and Assumption 24. Assume also that

(i) $\mathbb{E}[\epsilon_i \mid X_i] = 0$ *and* $\mathbb{E}[\epsilon_i^2 \mid X_i] \le \bar\sigma^2 < \infty$ *almost surely;*

(ii) $\mathbb{E}[|\epsilon_i|^\kappa \mid X_i] \le C_\epsilon < \infty$ *for the moment exponent κ in Assumption 21(e).*

Then Assumption 21(a)–(d) holds with $C_1 = 1/4$ and with variance and entropy constants growing at most polylogarithmically. Moreover,

$$\chi(h, \pi_n h) = \frac{1}{2}\|h - \pi_n h\|^2_{L^2(X)}.$$

Consequently, if

$$K_n^{2(1+\beta/(d+1))}\log(K_n) \asymp n$$

and

$$\varepsilon_n = O\left((n/\log(n))^{-\frac{1+2\beta/(d+1)}{4(1+\beta/(d+1))}}\right),$$

then

$$\|h - \widehat{h}_n\|_{L^2(X)} = O_P\left((n/\log(n))^{-\frac{1+2\beta/(d+1)}{4(1+\beta/(d+1))}}\right). \tag{E.33}$$

Proof of Proposition E.1. Let $v := g - h$. Since $Y_i = h(X_i) + \epsilon_i$,

$$q(Z_i, g, h) := \ell(Z_i, g) - \ell(Z_i, h) = -\epsilon_i v(X_i) + \frac{1}{2}v(X_i)^2.$$

Using $\mathbb{E}[\epsilon_i \mid X_i] = 0$, we get

$$Q(g,h) = \mathbb{E}_Z[q(Z_i, g, h)] = \frac{1}{2}\mathbb{E}[v(X_i)^2] - \mathbb{E}[\epsilon_i v(X_i)] = \frac{1}{2}\|g-h\|^2_{L^2(X)}.$$

Hence h minimizes the population risk, so Assumption 21(a) holds. Moreover,

$$Q(g,h) = \frac{1}{2}\|g-h\|^2_{L^2(X)} \ge 2C_1\|g-h\|^2_{L^2(X)}$$

for any $C_1 \le 1/4$, proving Assumption 21(b).

For $g \in \mathcal{H}_{K_n}(M_n)$, the sieve sup-bound gives

$$\|g\|_\infty \le M_n\sqrt{K_n}M_\sigma =: B_n.$$

Under $M_n \asymp \sqrt{\log(K_n)/K_n}$, $B_n \asymp \sqrt{\log(K_n)}$. Since h is bounded under the source condition, write $\|h\|_\infty \le H$. Then $\|v\|_\infty \le B_n + H$. Using $q(Z_i, g, h) = -\epsilon_i v(X_i) + \frac{1}{2}v(X_i)^2$,

$$\mathrm{Var}_Z(q(Z_i, g, h)) \le \mathbb{E}[q(Z_i, g, h)^2] \le 2\mathbb{E}[\epsilon_i^2 v(X_i)^2] + \frac{1}{2}\mathbb{E}[v(X_i)^4].$$

By $\mathbb{E}[\epsilon_i^2 \mid X_i] \le \bar{\sigma}^2$ and $\|v\|_\infty \le B_n + H$,

$$\mathrm{Var}_Z(q(Z_i, g, h)) \le \left(2\bar{\sigma}^2 + \frac{1}{2}(B_n + H)^2\right) \|g - h\|^2_{L^2(X)}.$$

Thus Assumption 21(c) holds with $C_{2,n} \lesssim 1 + B_n^2$, which is polylogarithmic in n under the stated calibration.

For the entropy condition, squared loss satisfies on the sieve prediction range

$$|\ell((y, x), u) - \ell((y, x), v)| \le (|\epsilon| + B_n + H)|u - v|.$$

Together with $\mathbb{E}[|\epsilon_i|^\kappa \mid X_i] \le C_\epsilon$, this gives the conditional-L^2 loss-entropy transfer bound

$$\log \mathcal{N}_{[]}\big(w, \mathcal{M}_n(\delta), L^2(P_Z)\big) \le K_n \log\left(\frac{\widetilde{L}_n A_n}{w} + 1\right),$$

where $A_n = 4M_n M_\sigma \sqrt{K_n}$ and $\widetilde{L}_n^2 \lesssim \bar{\sigma}^2 + (B_n + H)^2$. Since $\widetilde{L}_n A_n$ is polylogarithmic, the same calculation as in Lemma E.6 gives

$$\delta_n \lesssim \sqrt{\frac{K_n \log(K_n)}{n}},$$

so Assumption 21(d) holds.

The local empirical-process condition required for Lemma E.3 is imposed in Assumption 24. This condition replaces the pointwise local loss concentration verification (Assumption 21(e)), which is not implied by $L^2(X)$-localization alone.

Taking $g = \pi_n h$ in the identity for $Q(g, h)$ yields

$$\chi(h, \pi_n h) = Q(\pi_n h, h) = \frac{1}{2}\|\pi_n h - h\|^2_{L^2(X)}.$$

By Corollary E.1,

$$[\chi(h, \pi_n h)]^{1/2} = O_P\left(K_n^{-1/2-\beta/(d+1)}\right).$$

Under $K_n^{2(1+\beta/(d+1))} \log(K_n) \asymp n$,

$$K_n^{-1/2-\beta/(d+1)} \asymp (n/\log(n))^{-\frac{1+2\beta/(d+1)}{4(1+\beta/(d+1))}}.$$

We now check that $\mathbb{P}(\|\widehat{h}_n - h\|_{L^2(X)} > \eta_0) \to 0$ as needed to apply Theorem E.4. By Assumption 22, $Q_n(\widehat{h}_n, h) \le Q_n(\pi_n h, h) + o_P(1)$, cf. (E.17); by (E.10), we also have $Q_n(\pi_n h, h) = o_P(1)$. We proceed by contradiction: assume that $\|\widehat{h}_n - h\|_{L^2(X)} > \eta_0$ while $\|\pi_n h - h\|_{L^2(X)} < \eta_0/2$. By convexity of the squared loss, there exists an element $\breve{h} \in \mathcal{H}_{K_n}(M_n)$ on the segment with endpoints $\widehat{h}_n$ and $\pi_n h$ such that $\|\breve{h} - h\|_{L^2(X)} = \eta_0/2$, population excess loss $Q(\breve{h}, h) = \eta_0^2/8$, and $Q_n(\breve{h}, h) = o_P(1)$. We apply the local shell bound from Lemma E.3(ii) with $x = \eta_0/(4a_n)$ and $c = 1/4$, and find

$$\mathbb{P}\left(\sup_{\{g \in \mathcal{H}_{K_n}(M_n) \,|\, \eta_0/4 < \|g-h\|_{L^2(X)} \le \eta_0/2\}} |(Q_n - Q)(g, h)| \ge \frac{\eta_0^2}{64}\right)$$
$$\le A_0 \exp\left(-A_1 \frac{n a_n^2}{\Lambda_n}\left(\frac{\eta_0}{4a_n}^{2-s}\right)\right) + A_2(\eta_0/4)^{-(2-s)\kappa} \Lambda_n n^{-\kappa/2} = o(1). \tag{E.34}$$

However, with probability tending to one,

$$|(Q_n - Q)(\breve{h}, h)| \geq Q(\breve{h}, h) - Q_n(\breve{h}, h) > \frac{\eta_0^2}{16}$$

which contradicts (E.34) since $\breve{h} \in \{g \in \mathcal{H}_{K_n}(M_n) \mid \eta_0/4 < \|g - h\|_{L^2(X)} \leq \eta_0\}$.

Combining the preceding bounds with the rate argument in the proof of Theorem E.7 and using the assumed bound on ε_n gives (E.33). □

The following result shows that post-estimation clipping of the estimator preserves our established $L^2(X)$ convergence rate provided that the clipping threshold exceeds $\|h\|_\infty$.

Corollary E.3 (Post-estimation clipping)**.** *Let T_τ denote the clipping operator*

$$T_\tau u := \max\{-\tau, \min\{u, \tau\}\}.$$

Let the assumptions of Proposition E.1 hold and define the clipped estimator $T_{\tau_n}\widehat{h}_n$ by applying T_{τ_n} pointwise after estimation for a sequence $\tau_n > 0$ of clipping thresholds. Then

$$\|T_{\tau_n}\widehat{h}_n - h\|_{L^2(X)} \leq \|\widehat{h}_n - h\|_{L^2(X)} + \|T_{\tau_n}h - h\|_{L^2(X)}.$$

In particular, if $\|h\|_\infty \leq \tau_n$ for all sufficiently large n, then

$$\|T_{\tau_n}\widehat{h}_n - h\|_{L^2(X)} \leq \|\widehat{h}_n - h\|_{L^2(X)}.$$

Consequently, under the conditions of Proposition E.1,

$$\|T_{\tau_n}\widehat{h}_n - h\|_{L^2(X)} = O_P\left((n/\log(n))^{-\frac{1+2\beta/(d+1)}{4(1+\beta/(d+1))}}\right)$$

whenever $\|h\|_\infty \leq \tau_n$ for all sufficiently large n.

Proof of Corollary E.3. The clipping map $T_\tau : \mathbb{R} \to [-\tau, \tau]$ is nonexpansive, that is

$$|T_\tau a - T_\tau b| \leq |a - b|$$

for all $a, b \in \mathbb{R}$. Therefore,

$$\begin{aligned}|T_{\tau_n}\widehat{h}_n(X) - h(X)| &\leq |T_{\tau_n}\widehat{h}_n(X) - T_{\tau_n}h(X)| + |T_{\tau_n}h(X) - h(X)| \\ &\leq |\widehat{h}_n(X) - h(X)| + |T_{\tau_n}h(X) - h(X)|.\end{aligned}$$

Taking $L^2(X)$ norms gives the first claim. If $\|h\|_\infty \leq \tau_n$, then $T_{\tau_n}h = h$, and the contraction bound gives the second claim. The rate follows immediately from Proposition E.1. □

Remark 21 (Other losses)**.** The generic sieve convergence theorem, Theorem E.4, can also be applied to other losses once the corresponding local separation, variance, entropy, and Chen-Shen empirical-process conditions are verified. Losses that are Lipschitz in the second argument are especially convenient, since Lemma E.6 transfers the RWNN prediction-class entropy to the loss-difference class. Examples include:

(i) absolute loss $\ell((y, x), u) = |y - u|$, which is globally 1-Lipschitz in u;

(ii) Huber loss with fixed threshold, which is globally Lipschitz in u;

(iii) quantile or pinball loss, which is globally Lipschitz in u;

(iv) logistic loss $\ell((y, x), u) = \log(1 + e^u) - yu$ for $y \in \{0, 1\}$, whose derivative with respect to u is uniformly bounded;

(v) hinge loss $\ell((y,x),u) = (1-yu)_+$ for $y \in \{-1,1\}$, which is globally Lipschitz in u.

For smooth likelihood-type losses, the local quadratic separation condition follows from local positive curvature of the population risk. For nonsmooth losses, such as absolute or quantile loss, separation typically requires an additional condition on the conditional distribution of Y given X, for instance a density bounded away from zero near the conditional median or quantile.

Remark 22 (Comparison with the sieve rate in Chen and White (1999))**.** For certain integral Sobolev classes, Chen and White (1999) obtained the consistency rate

$$O_P\left((n/\log(n))^{-\frac{1+2\alpha/(d+1)}{4(1+\alpha/(d+1))}}\right),$$

where $\alpha \in (0,1]$ is related to the regularity of the activation function used to construct fully trainable shallow neural networks. The rate in Theorem E.7 has the same algebraic form, with the source regularity parameter β replacing α. In particular, when $\beta = \alpha$, the RWNN estimator achieves the same statistical rate as Chen and White (1999), although the underlying approximation classes and network constructions are different.

E.3 Asymptotic normality for smooth functionals

We now consider inference on smooth functionals of the target function. Let $\Psi : \mathcal{H} \to \mathbb{R}$ be a known functional and suppose that the object of interest is $\Psi(h)$. We follow the local-perturbation method of Shen (1997) and Chen and Shen (1998), and provide the proof in our notation.

We now define all the ingredients that we need in order to prove asymptotic normality of smooth functionals of RWNN estimators. For any function $u \in \mathcal{H}$ considered below, let $\pi_n u$ denote a measurable sieve projection of u onto $\mathcal{H}_{K_n}(M_n)$, that is $\pi_n u \in \arg\min_{g \in \mathcal{H}_{K_n}(M_n)} \|u-g\|_\circ$. For a direction v such that $h + v \in \mathcal{H}$, define the loss increment

$$q_h(v, Z_i) := \ell(Z_i, h+v) - \ell(Z_i, h). \tag{E.35}$$

Hence, for $g \in \mathcal{H}$, we can write the expression of its population excess risk and its empirical analogue, with respect to h , using the notation introduced in (E.4)-(E.3), as

$$Q(g,h) = \mathbb{E}_Z[q(Z_i,g,h)] = \mathbb{E}_Z[q_h(g-h,Z_i)]$$

and

$$Q_n(g,h) = \frac{1}{n}\sum_{i=1}^n q(Z_i,g,h) = L_n(g) - L_n(h) = \frac{1}{n}\sum_{i=1}^n q_h(g-h,Z_i),$$

respectively.

For an integrable measurable function f, define the corresponding centered empirical process as

$$\nu_n(f) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(f(Z_i) - \mathbb{E}_Z[f(Z_i)]\right).$$

and, finally, for $g_1, g_2 \in \mathcal{H}$, write, using the notation introduced in Lemma E.3,

$$\mu_n(h, g_1, g_2) = \mu_n(g_1, h) - \mu_n(g_2, h) = (Q_n(g_1,h) - Q(g_1,h)) - (Q_n(g_2,h) - Q(g_2,h)). \tag{E.36}$$

Assumption 25 (Smooth functional and Riesz representer)**.** Let $\mathcal{V} := \overline{\mathrm{span}(\mathcal{H} - \{h\})}^{\|\cdot\|_\circ}$. The norm $\|\cdot\|_\circ$ is induced by an inner product $\langle\cdot,\cdot\rangle_\circ$ on $\mathcal{V}$. The functional Ψ is differentiable at h in the sense that there exists a continuous linear functional $\Psi'_h : \mathcal{V} \to \mathbb{R}$ such that, as $\|\widetilde{h} - h\|_\circ \to 0$,

$$\Psi(\widetilde{h}) - \Psi(h) = \Psi'_h(\widetilde{h} - h) + O(\|\widetilde{h} - h\|_\circ^\omega) \tag{E.37}$$

for some $\omega > 0$. Let $\psi \in \mathcal{V}$ denote the Riesz representer of Ψ'_h, so that

$$\Psi'_h(v) = \langle v, \psi\rangle_\circ, \qquad v \in \mathcal{V}. \tag{E.38}$$

Assumption 26 (Local criterion linearization)**.** For every g in a neighborhood of h, write $v := g - h$ and suppose that there exists a path $t \mapsto h_t(v)$ in $\mathcal{H}$, $t \in [0,1]$, such that $h_0(v) = h$ and $h_1(v) = g$. Define the pathwise derivative of the loss difference at h in the direction v by

$$q'_h(v, Z_i) := \lim_{t \downarrow 0} \frac{\ell(Z_i, h_t(v)) - \ell(Z_i, h)}{t},$$

whenever the limit exists. The loss difference (E.35) admits the expansion

$$q_h(v, Z_i) = q'_h(v, Z_i) + r_h(v, Z_i), \tag{E.39}$$

where $r_h(v, Z_i)$ is the corresponding remainder. We assume that $q'_h(v, Z_i) - \mathbb{E}_Z[q'_h(v, Z_i)]$ is linear in v. The empirical contribution of the remainder is controlled by Assumption 27.

Use $\mathfrak{R}_n$ to denote the rate obtained via Theorem E.7, as in Shen (1997, p. 2567), so that $\|\widehat{h}_n - h\|_\circ = O_P(\mathfrak{R}_n)$, and let $\mathfrak{R}_n = o(r_n)$ so that

$$\|\widehat{h}_n - h\|_\circ = o_P(r_n).$$

Further, let $t_n > 0$ be a deterministic sequence satisfying

$$t_n = o(n^{-1/2}).$$

For $s \in \{-1, 1\}$, define the local perturbation of g in the Riesz direction $s\psi$ as

$$g^*_{n,s}(g) := (1 - t_n)g + t_n(h + s\psi).$$

Since $g^*_{n,s}(g)$ need not belong to $\mathcal{H}_{K_n}(M_n)$, let

$$\widetilde{g}_{n,s}(g) := \pi_n g^*_{n,s}(g) \in \mathcal{H}_{K_n}(M_n)$$

be its sieve projection, where π_n is the sieve projection introduced in (E.5). The following assumption is the analogue of Conditions (A)-(D) in Shen (1997), presented in our notation and in our loss-minimization setting.

Assumption 27. The following conditions hold conditionally on realized random parameters $\theta_1, \theta_2, \ldots, \theta_{K_n}$.

(a) [Stochastic equicontinuity of the remainder] For r_h defined in (E.39),

$$\sup_{\substack{\{g \in \mathcal{H}_{K_n}(M_n) \mid \|g-h\|_\circ \le r_n\} \\ s \in \{-1,1\}}} n^{-1/2} \left| \nu_n \left(r_h(g - h, Z_i) - r_h(\widetilde{g}_{n,s}(g) - h, Z_i) \right) \right| = O_P(t_n^2).$$

(b) [Expectation of the criterion difference]

$$\sup_{\substack{\{g \in \mathcal{H}_{K_n}(M_n) \mid 0 < \|g-h\|_\circ \le r_n\} \\ s \in \{-1,1\}}} \left| Q(\widetilde{g}_{n,s}(g), h) - Q(g, h) - \frac{1}{2} \left(\|\widetilde{g}_{n,s}(g) - h\|_\circ^2 - \|g - h\|_\circ^2 \right) \right| = O(t_n^2).$$

(c) [Projection error of the local perturbation]

$$\sup_{\substack{\{g \in \mathcal{H}_{K_n}(M_n) \mid 0 < \|g-h\|_\circ \le r_n\} \\ s \in \{-1,1\}}} \|g^*_{n,s}(g) - \widetilde{g}_{n,s}(g)\|_\circ = O(r_n^{-1} t_n^2),$$

and

$$\sup_{\substack{\{g \in \mathcal{H}_{K_n}(M_n) \mid \|g-h\|_\circ \le r_n\} \\ s \in \{-1,1\}}} n^{-1/2} \left| \nu_n \left(q'_h(g^*_{n,s}(g) - \widetilde{g}_{n,s}(g), Z_i) \right) \right| = O_P(t_n^2).$$

(d) [Local derivative bound]

$$\sup_{\{g\in\mathcal{H}_{K_n}(M_n)\,|\,\|g-h\|_\circ\le r_n\}} n^{-1/2}\left|\nu_n\left(q_h'(g-h,Z_i)\right)\right| = O_P(t_n).$$

Assumption 28 (Optimization error and local rate)**.** The estimator satisfies

$$Q_n(\widehat{h}_n,h) \le \inf_{g\in\mathcal{H}_{K_n}(M_n)} Q_n(g,h) + \rho_n, \tag{E.40}$$

where $\rho_n = o_P(t_n n^{-1/2})$. Moreover, (E.37) holds with some $\omega \ge 2$ and the rate r_n satisfies

$$\sqrt{n}r_n^2 \to 0 \quad \text{and} \quad n^{-1/2} = o(r_n).$$

For comparison's sake, given the rate obtained in Theorem E.7, from Proposition E.1 we can see that a valid choice of r_n with the squared loss is

$$r_n = \log(n)(n/\log(n))^{-\gamma}, \quad \text{with } \gamma = \frac{1+2\beta/(d+1)}{4(1+\beta/(d+1))} \in \left(\frac{1}{4},\frac{1}{2}\right).$$

It is easy to show that this choice fulfills both conditions of Assumption 28.

Assumption 29 (Finite variance and scalar CLT in the Riesz direction)**.** The linearized loss increment in the Riesz direction has finite positive asymptotic variance

$$0 < \sigma_\Psi^2 := \lim_{n\to\infty}\frac{1}{n}\mathrm{Var}\left[\sum_{i=1}^n q_h'(\psi,Z_i)\right] < \infty.$$

Moreover,

$$\nu_n(q_h'(\psi,Z_i)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(q_h'(\psi,Z_i) - \mathbb{E}_Z[q_h'(\psi,Z_i)]\right) \xrightarrow{d} \mathcal{N}(0,\sigma_\Psi^2). \tag{E.41}$$

In the i.i.d. case of Assumption 20, (E.41) follows from the usual scalar central limit theorem under the finite-variance condition. In the dependent cases of Assumption 20, it is the analogue of Condition B.5 in Chen and Shen (1998).

Theorem E.8 (Asymptotic normality of smooth RWNN functionals)**.** *Let Assumptions 25-29 hold. Then*

$$\sqrt{n}\{\Psi(\widehat{h}_n) - \Psi(h)\} \xrightarrow{d} \mathcal{N}(0,\sigma_\Psi^2).$$

Proof of Theorem E.8. We start by defining

$$v_n := \widehat{h}_n - h.$$

By Assumption 25, it holds that

$$\Psi(\widehat{h}_n) - \Psi(h) = \Psi_h'(v_n) + O_P(\|v_n\|_\circ^\omega) = \langle v_n,\psi\rangle_\circ + O_P(\|v_n\|_\circ^\omega),$$

where in the first equality we used (E.37), while in the second one used the Riesz representation (E.38). Next, since $\|v_n\|_\circ \equiv \|\widehat{h}_n - h\|_\circ = o_P(r_n)$, we observe that the event

$$\mathcal{A}_n := \left\{\|\widehat{h}_n - h\|_\circ > r_n\right\} \tag{E.42}$$

fulfills $\mathbb{P}_Z(\mathcal{A}_n) = o(1)$ as $n\to\infty$, and hence is asymptotically negligible. Thus, in the following we can work conditionally on the event $\mathcal{A}_n^\complement = \{\|\widehat{h}_n - h\|_\circ \le r_n\}$. Then, as $\omega \ge 2$ and $\sqrt{n}\, r_n^\omega \to 0$, by Assumption 28, we have

$$\sqrt{n}\{\Psi(\widehat{h}_n) - \Psi(h)\} = \sqrt{n}\langle v_n,\psi\rangle_\circ + o_P(1). \tag{E.43}$$

Hence, it remains to quantify the leading term $\sqrt{n}\langle v_n, \psi\rangle_\circ$.

Fix $s \in \{-1, 1\}$ and define

$$\widetilde{h}_{n,s} := \widetilde{g}_{n,s}(\widehat{h}_n) = \pi_n g^*_{n,s}(\widehat{h}_n) \in \mathcal{H}_{K_n}(M_n).$$

We now notice that (E.40) in Assumption 28 gives

$$Q_n(\widetilde{h}_{n,s}, h) - Q_n(\widehat{h}_n, h) \geq -\rho_n. \tag{E.44}$$

We now decompose the left-hand side into its population and empirical-process parts:

$$Q_n(\widetilde{h}_{n,s}, h) - Q_n(\widehat{h}_n, h) = Q(\widetilde{h}_{n,s}, h) - Q(\widehat{h}_n, h) + \mu_n(h, \widetilde{h}_{n,s}, \widehat{h}_n), \tag{E.45}$$

where, by definition (E.36),

$$\mu_n(h, \widetilde{h}_{n,s}, \widehat{h}_n) = (Q_n(\widetilde{h}_{n,s}, h) - Q(\widetilde{h}_{n,s}, h)) - (Q_n(\widehat{h}_n, h) - Q(\widehat{h}_n, h)).$$

We analyze first the empirical-process term in (E.45). More specifically, by Assumption 26,

$$\begin{aligned}\mu_n(h, \widetilde{h}_{n,s}, \widehat{h}_n) = n^{-1/2}\nu_n \left(q'_h(\widetilde{h}_{n,s} - h, Z_i) - q'_h(v_n, Z_i)\right) \\ + n^{-1/2}\nu_n \left(r_h(\widetilde{h}_{n,s} - h, Z_i) - r_h(v_n, Z_i)\right),\end{aligned} \tag{E.46}$$

where we applied (E.39) to each of the brackets in the previous expression.

Moreover, one can write down the following equality:

$$\widetilde{h}_{n,s} - h = (\widetilde{h}_{n,s} - g^*_{n,s}(\widehat{h}_n)) + (g^*_{n,s}(\widehat{h}_n) - \widehat{h}_n) + (\widehat{h}_n - h).$$

Since by Assumption 26, the centered derivative is linear in its direction, we obtain that the first term in (E.46) can be expressed as

$$\begin{aligned}&n^{-1/2}\nu_n \left(q'_h(\widetilde{h}_{n,s} - h, Z_i) - q'_h(v_n, Z_i)\right) \\ &= n^{-1/2}\nu_n \left(q'_h(\widetilde{h}_{n,s} - g^*_{n,s}(\widehat{h}_n), Z_i)\right) + n^{-1/2}\nu_n \left(q'_h(g^*_{n,s}(\widehat{h}_n) - \widehat{h}_n, Z_i)\right).\end{aligned}$$

By Assumption 27 (c), the first term on the right-hand side is $O_P(t_n^2)$. Next, since

$$g^*_{n,s}(\widehat{h}_n) - \widehat{h}_n = t_n(s\psi - v_n),$$

linearity of the centered derivative gives

$$n^{-1/2}\nu_n \left(q'_h(g^*_{n,s}(\widehat{h}_n) - \widehat{h}_n, Z_i)\right) = t_n n^{-1/2} s\nu_n \left(q'_h(\psi, Z_i)\right) - t_n n^{-1/2}\nu_n \left(q'_h(v_n, Z_i)\right). \tag{E.47}$$

By Assumption 27 (d),

$$n^{-1/2}\nu_n \left(q'_h(v_n, Z_i)\right) = O_P(t_n),$$

and hence

$$t_n n^{-1/2}\nu_n \left(q'_h(v_n, Z_i)\right) = O_P(t_n^2). \tag{E.48}$$

Using the fact that, by Assumption 27 (a), the remainder contribution in (E.46) is $O_P(t_n^2)$ and combining expressions (E.48) and (E.47) allows to write (E.46) as

$$\mu_n(h, \widetilde{h}_{n,s}, \widehat{h}_n) = t_n n^{-1/2} s\nu_n \left(q'_h(\psi, Z_i)\right) + O_P(t_n^2). \tag{E.49}$$

We next analyze the population term in (E.45). By Assumption 27 (b),

$$Q(\widetilde{h}_{n,s}, h) - Q(\widehat{h}_n, h) = \frac{1}{2}\left(\|\widetilde{h}_{n,s} - h\|_\circ^2 - \|v_n\|_\circ^2\right) + O_P(t_n^2).$$

By Assumption 27 (c), replacing $\widetilde{h}_{n,s}$ by $g^*_{n,s}(\widehat{h}_n)$ in the squared norm introduces only an $O_P(t_n^2)$ error. Indeed, by the reverse triangle inequality,

$$\left| \|\widetilde{h}_{n,s} - h\|_\circ^2 - \|g^*_{n,s}(\widehat{h}_n) - h\|_\circ^2 \right| \le \|\widetilde{h}_{n,s} - g^*_{n,s}(\widehat{h}_n)\|_\circ \left(\|\widetilde{h}_{n,s} - h\|_\circ + \|g^*_{n,s}(\widehat{h}_n) - h\|_\circ \right).$$

By Assumption 27 (c), the first factor is $O_P(r_n^{-1}t_n^2)$. Since $\|\widehat{h}_n - h\|_\circ = O_P(r_n)$, $t_n = o(n^{-1/2}) = o(r_n)$, and $\psi \in \mathcal{V}$ is fixed, $\|g^*_{n,s}(\widehat{h}_n) - h\|_\circ \le (1-t_n)\|v_n\|_\circ + t_n\|\psi\|_\circ = O_P(r_n)$; as $t_n^2/r_n = o(r_n)$, the same holds for $\|\widetilde{h}_{n,s} - h\|_\circ$. Hence the displayed difference is $O_P(t_n^2)$.

Hence,

$$Q(\widetilde{h}_{n,s}, h) - Q(\widehat{h}_n, h) = \frac{1}{2}\left(\|g^*_{n,s}(\widehat{h}_n) - h\|_\circ^2 - \|v_n\|_\circ^2 \right) + O_P(t_n^2). \tag{E.50}$$

Since

$$g^*_{n,s}(\widehat{h}_n) - h = (1-t_n)v_n + t_n s\psi = v_n + t_n(s\psi - v_n),$$

we have

$$\begin{aligned}\frac{1}{2}\left(\|g^*_{n,s}(\widehat{h}_n) - h\|_\circ^2 - \|v_n\|_\circ^2 \right) &= t_n\langle v_n, s\psi - v_n\rangle_\circ + \frac{1}{2}t_n^2\|s\psi - v_n\|_\circ^2 \\ &= t_n s\langle v_n, \psi\rangle_\circ - t_n\|v_n\|_\circ^2 + \frac{1}{2}t_n^2\|s\psi - v_n\|_\circ^2.\end{aligned}$$

Now as $\|v_n\|_\circ = O_P(r_n)$ and $\sqrt{n}r_n^2 \to 0$, one gets

$$t_n\|v_n\|_\circ^2 = o_P(t_n n^{-1/2}).$$

Also, since $t_n = o(n^{-1/2})$ and ψ is fixed in $\mathcal{V}$,

$$t_n^2\|s\psi - v_n\|_\circ^2 = o_P(t_n n^{-1/2}).$$

The $O_P(t_n^2)$ term is also $o_P(t_n n^{-1/2})$ since $t_n = o(n^{-1/2})$. Hence in (E.50) one gets

$$Q(\widetilde{h}_{n,s}, h) - Q(\widehat{h}_n, h) = t_n s\langle v_n, \psi\rangle_\circ + o_P(t_n n^{-1/2}). \tag{E.51}$$

Combining (E.44), (E.45), (E.49), and (E.51), we obtain

$$0 \le t_n s\langle v_n, \psi\rangle_\circ + t_n n^{-1/2} s\nu_n\left(q'_h(\psi, Z_i)\right) + O_P(t_n^2) + \rho_n. \tag{E.52}$$

Since $t_n = o(n^{-1/2})$, we have $t_n^2 = o(t_n n^{-1/2})$, and, by Assumption 28,

$$\rho_n = o_P(t_n n^{-1/2}).$$

Hence, we rewrite (E.52) as

$$0 \le t_n s\langle v_n, \psi\rangle_\circ + t_n n^{-1/2} s\nu_n\left(q'_h(\psi, Z_i)\right) + o_P(t_n n^{-1/2}).$$

Dividing by t_n and multiplying by $\sqrt{n}$ gives

$$s\sqrt{n}\langle v_n, \psi\rangle_\circ + s\nu_n\left(q'_h(\psi, Z_i)\right) \ge o_P(1). \tag{E.53}$$

Finally, $s = 1$ in (E.53) gives

$$\sqrt{n}\langle v_n, \psi\rangle_\circ \ge -\nu_n\left(q'_h(\psi, Z_i)\right) + o_P(1),$$

while taking $s = -1$ results in

$$\sqrt{n}\langle v_n, \psi\rangle_\circ \le -\nu_n\left(q'_h(\psi, Z_i)\right) + o_P(1),$$

which yield

$$\sqrt{n}\langle \widehat{h}_n - h, \psi \rangle_\circ = -\nu_n \left(q_h'(\psi, Z_i) \right) + o_P(1). \tag{E.54}$$

Combining (E.43) and (E.54), we obtain

$$\sqrt{n}\{\Psi(\widehat{h}_n) - \Psi(h)\} = -\nu_n \left(q_h'(\psi, Z_i) \right) + o_P(1).$$

By Assumption 29 and Slutsky's theorem, we obtained the required result

$$\sqrt{n}\{\Psi(\widehat{h}_n) - \Psi(h)\} \xrightarrow{d} \mathcal{N}(0, \sigma_\Psi^2),$$

since multiplication by -1 does not change the limiting variance. □

Remark 23 (Relation to Shen's local perturbation argument)**.** The proof does not rely on an empirical score equation at $\widehat{h}_n$. Instead, it compares the empirical loss contrast at $\widehat{h}_n$ with that at the projected local perturbations $\widetilde{h}_{n,s}$, $s \in \{-1, 1\}$. This is the device that makes the argument compatible with sieve and constrained estimators, where $\widehat{h}_n$ may lie on the boundary of the approximating space.

F Auxiliary results

F.1 Geometric and sampling inequalities

Lemma F.1 (Kabatianski, Zeger). *Given a point set $P \subset \mathbb{R}^d$ and $x \in P$, the maximum number of points in P that can have x as nearest neighbor is bounded by the d-dimensional kissing number ψ_d, where $\psi_d \leq 2^{0.401d(1+o(1))}$ as $d \to \infty$.*

F.2 Matrix concentration and perturbation inequalities

The following matrix-valued law of large numbers is needed for the L^2 consistency result in Section 4.2.

Lemma F.2 (Rudelson's Matrix LLN, Belloni et al., 2015, Lemma 6.2). *Let $Q_1, \ldots, Q_n$ be a sequence of independent symmetric non-negative $K \times K$-matrix valued random variables with $K \geq 2$ such that $Q = \mathbb{E}[Q_i]$ and $\|Q_i\| \leq M$ a.s., then for $\widehat{Q} = \mathbb{E}_n[Q_i]$*

$$\mathbb{E}\|\widehat{Q} - Q\| \lesssim \frac{M \log(K)}{n} + \sqrt{\frac{M\|Q\| \log(K)}{n}}.$$

In particular, if $Q_i = p_i p_i^\top$, with $\|p_i\| \leq \xi_K$ a.s., then

$$\mathbb{E}\|\widehat{Q} - Q\| \lesssim \frac{\xi_K^2 \log(K)}{n} + \sqrt{\frac{\xi_K^2 \|Q\| \log(K)}{n}}.$$

F.3 Entropy and empirical-process inequalities

The following statement of Bernstein's inequality, together with its conditional statement in Lemma F.4 are used in Subsection 4.6.1.

Theorem F.3 (Bernstein's inequality, Boucheron et al., 2013). *Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that there exist positive numbers v and c such that $\sum_{i=1}^n \mathbb{E}[X_i^2] \leq v$ and*

$$\sum_{i=1}^n \mathbb{E}\big[(X_i)_+^q\big] \leq \frac{q!}{2}\, v\, c^{q-2} \qquad \text{for all integers } q \geq 3,$$

where $x_+ := \max(x, 0)$.

If $S := \sum_{i=1}^n \big(X_i - \mathbb{E}X_i\big)$, then, for every $\lambda \in (0, 1/c)$ and every $t > 0$,

$$\psi_S(\lambda) \leq \frac{v\lambda^2}{2(1 - c\lambda)}$$

and

$$\psi_S^*(t) \geq \frac{v}{c^2} h_1\left(\frac{ct}{v}\right),$$

where $h_1(u) = 1 + u - \sqrt{1 + 2u}$ for $u > 0$. In particular, for every $t > 0$,

$$\mathbb{P}\left(S \geq \sqrt{2vt} + ct\right) \leq e^{-t}.$$

The following conditional formulation is a consequence of Theorem F.3 applied under the regular conditional distribution given $\mathcal{G}$, and then symmetrically applied to both ξ_i and $-\xi_i$.

Lemma F.4 (Conditional Bernstein inequality)**.** *Let $\mathcal{G}$ be a sub-σ-field, and let $\xi_1, \dots, \xi_n$ be random variables that are conditionally independent given $\mathcal{G}$. Suppose that, almost surely, $\mathbb{E}[\xi_i \mid \mathcal{G}] = 0$ and $|\xi_i| \le b$ for every $i = 1, \dots, n$. Define the conditional variance proxy*

$$v := \sum_{i=1}^{n} \mathbb{E}\left[\xi_i^2 \mid \mathcal{G}\right].$$

Then, for every $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \xi_i\right| \ge t \,\middle|\, \mathcal{G}\right) \le 2\exp\left(-\frac{t^2}{2\left(v + bt/3\right)}\right)$$

almost surely.

Lemma F.5. *For all $a \ge 1$, it holds*

$$\int_0^1 \sqrt{\log\left(\frac{a}{x}\right)}\, dx \le \sqrt{\log(a)} + 1.$$

Proof of Lemma F.5. The definite integral, given that a is strictly positive, is equal to

$$\int_0^1 \sqrt{\log\left(\frac{a}{x}\right)}\, dx = \sqrt{\log(a)} + a\frac{\sqrt{\pi}}{2}\operatorname{erfc}\left(\sqrt{\log(a)}\right),$$

where $\operatorname{erfc}(\cdot)$ is the complementary error function. Applying, for all $t > 0$, the bound

$$\operatorname{erfc}(t) \le \frac{2}{\sqrt{\pi}} \frac{e^{-t^2}}{t + \sqrt{t^2 + \frac{4}{\pi}}},$$

(see, e.g., *NIST Handbook of Mathematical Functions*, National Institute of Standards and Technology (2010), Section 7.8.2.) the latter term in the integral can be immediately controlled by

$$a\frac{\sqrt{\pi}}{2}\operatorname{erfc}\left(\sqrt{\log(a)}\right) \le \frac{1}{\sqrt{\log(a)} + \sqrt{\log(a) + \frac{4}{\pi}}} \le 1$$

since $a \ge 1$ and $\frac{4}{\pi} > 1$. □

G Additional numerical experiments and results

G.1 Residual, wild and score bootstrap inference for evaluation functional

We now consider only the pointwise evaluation functional $\Psi_x(h) = h(x)$, which is treated as an irregular sieve functional using the sieve-dependent studentization described below. We evaluate the RWNN estimate and construct pointwise confidence intervals on a uniformly spaced grid of 31 points over the interval $[-0.9, 0.9]$. Restricting the evaluation grid to the interior of $[-1, 1]$ avoids boundary estimation issues and allows us to analyze the local coverage properties of the inference procedures. We discuss two RWNN specifications: we consider both ReLU and tanh activations with the same weight distributions and specifications used in the simulations of Section 5.4. We also apply the same Monte Carlo design as in that setting, redrawing RWNN weights across different data draws. We construct pointwise confidence intervals over the grid by using three bootstrapping approaches: residual (RB), wild (WB), and score-based (SB) bootstrap.

We generate bootstrap response variables as described in Section 5.5 for the residual bootstrap. As before, in the RB the innovations are sampled independently from the centered empirical residual distribution, meaning $\epsilon_i^* := \widehat{\epsilon}_i^* - \overline{\widehat{\epsilon}}$, with $\widehat{\epsilon}_i^* \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\widehat{\epsilon},n}$ and $\overline{\widehat{\epsilon}} := n^{-1}\sum_{i=1}^n \widehat{\epsilon}_i$. For WB, we instead set $\epsilon_i^* := (\widehat{\epsilon}_i - \overline{\widehat{\epsilon}})\xi_i$, where $\{\xi_i\}_{i=1}^n$ are i.i.d. multipliers, independent of data and satisfying $\mathbb{E}[\xi_i] = 0$, $\mathrm{Var}(\xi_i) = 1$, and having a finite absolute third moment. In the simulations, we take $\xi_i \sim \mathcal{N}(0,1)$.[3]

For bootstrap schemes $b \in \{\mathrm{RB}, \mathrm{WB}\}$, let $\widehat{h}_{n,\lambda}^{*,b}$ denote the RWNN estimator obtained by refitting the model to the corresponding bootstrap sample, using the same realized random parameters, dimension K_n, and the same $\lambda_{K_n,n}$ as in the original sample. For each x in the evaluation grid, let $z_\alpha^{*,b}(x)$ denote the conditional α-quantile of the bootstrap distribution of $\widehat{h}_{n,\lambda}^{*,b}(x)$, given the original sample. The corresponding percentile-bootstrap confidence interval is

$$\mathrm{CI}_{1-\alpha,b}(\Psi; x) := \left[z_{\alpha/2}^{*,b}(x),\ z_{1-\alpha/2}^{*,b}(x)\right], \quad b \in \{\mathrm{RB}, \mathrm{WB}\}.$$

Both RB and WB schemes avoid direct estimation of the asymptotic sieve-dependent variance of the pointwise evaluation functional.

The score bootstrap procedure we implement is adapted from the scheme proposed by Chen and Christensen (2018) for uniform inference in nonparametric instrumental-variable models. For the evaluation functional $\Psi_x(h) = h(x)$, it is easy to show that $\Psi'_{\widehat{h}_{n,\lambda}}(v) = v(x)$ for every sieve direction v. Hence, $\Psi'_{\widehat{h}_{n,\lambda}}(g_{K_n}) = g_{K_n}(x)$. Similarly to (5.4), the empirical ridge-adjusted Riesz representer of Ψ_x is

$$\widehat{\psi}_{\Psi_x,\lambda}(z) := g_{K_n}(z)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} g_{K_n}(x), \quad z \in \mathcal{X},$$

Consequently, the sieve variance estimator for pointwise evaluation admits the equivalent representation

$$\widehat{\sigma}^2_{\Psi_x}(x) = g_{K_n}(x)^\top \widehat{\Omega}_{K_n,\lambda} g_{K_n}(x), \quad x \in \mathcal{X}.$$

Let $\{\xi_i\}_{i=1}^n$ be i.i.d. standard Gaussian multipliers, independent of the data. The pointwise score-bootstrap t-statistic is defined by

$$\mathfrak{z}_n^*(x) := \frac{1}{\widehat{\sigma}_{\Psi_x}(x)\sqrt{n}} \sum_{i=1}^n \widehat{\epsilon}_i \widehat{\psi}_{\Psi_x,\lambda}(X_i)\xi_i \ = \sqrt{\frac{n}{\widehat{\sigma}^2_{\Psi_x}(x)}} \sum_{i=1}^n p_{i,n,\lambda}(x)\, \widehat{\epsilon}_i\, \xi_i,$$

where $p_{i,n,\lambda}(x)$ denotes the ridge sieve smoothing weight defined in (4.76), and where the second equality follows from $\widehat{\psi}_{\Psi_x,\lambda}(X_i) = g_{K_n}(X_i)^\top \widehat{\Gamma}_{K_n,n,\lambda}^{-1} g_{K_n}(x) = n\, p_{i,n,\lambda}(x)$.[4]

[3] Alternative choices include, for example, Rademacher and recentered exponential multipliers.

[4] Note that, by construction, conditional on the data, the score bootstrap distribution is, in fact, Gaussian at each fixed x: Chen and Christensen (2018) take the supremum over $x \in \mathcal{X}$ to obtain uniform critical values for $\widehat{h}_{n,\lambda}$, while we skip this step.

Let $z^{\circ}_{\alpha}(x)$ denote the conditional α-quantile, given the original sample, of $|\mathfrak{z}^{*}_{n}(x)|$. The corresponding $100(1-\alpha)\%$ pointwise score bootstrap CI is

$$\mathrm{CI}_{1-\alpha,\mathrm{SB}}(\Psi_x; x) := \left[\widehat{h}_{n,\lambda}(x) - \frac{\widehat{\sigma}_{\Psi_x}(x)}{\sqrt{n}} z^{\circ}_{1-\alpha}(x),\ \widehat{h}_{n,\lambda}(x) + \frac{\widehat{\sigma}_{\Psi_x}(x)}{\sqrt{n}} z^{\circ}_{1-\alpha}(x)\right]$$

We use $\mathrm{MC} = 1000$ Monte Carlo replications and $\mathfrak{B} = 1000$ bootstrap resamplings. To match the pipeline of real-world applications of RWNN, we again independently resample the RWNN weights $\{A_k, B_k\}_{k=1}^{K_n}$ at each Monte Carlo data draw. The reported Monte Carlo coverage probabilities thus average over both the sampling variation in the data and the randomness in the hidden-layer parameters, although each bootstrap confidence interval is conditional on the realized random features. The exact model specifications used are summarized in Table G.4.

We report the results for both coverage and bias in the plots of Appendix G.3. We find that all bootstrap schemes perform comparatively well for both tanh and ReLU RWNN specifications in the simpler specification involving g_3. More pronounced differences arise for g_1 and g_2. For g_2, the ReLU-based confidence intervals tend to systematically undercover in regions where the regression function is steeper, while the tanh CIs are substantially better calibrated. In these regions, the RB procedure generally attains the highest coverage among the three bootstrap methods. For g_1, the ReLU RWNN confidence intervals perform particularly poorly near the origin, where the regression function has a sharp peak, with observed coverage at times falling below 50% on the interval $[-0.25, 0.25]$. The confidence intervals based on the tanh RWNN perform better over the entire evaluation region, and the RB procedure again generally yields the highest coverage probabilities.

G.2 Additional tables

Table G.1: Asymptotic inference results for $\Psi_2(\widehat{h}_{n,\lambda})$ with $\lambda = 10^{-2}$.

			Gaussian–Gaussian			Gaussian–Uniform			Uniform–Uniform		
n	K_n	Method	Bias	Coverage	Length	Bias	Coverage	Length	Bias	Coverage	Length
$h = g_1$											
100	15	RSZ	-1.565	0.582	0.356	-1.474	0.611	0.360	-1.780	0.575	0.395
		SNW	-1.565	0.570	0.348	-1.474	0.605	0.355	-1.780	0.559	0.385
300	20	RSZ	-1.294	0.318	0.195	-1.231	0.370	0.199	-1.495	0.297	0.218
		SNW	-1.294	0.293	0.189	-1.231	0.342	0.194	-1.495	0.273	0.210
1000	30	RSZ	-0.914	0.090	0.099	-0.849	0.132	0.100	-1.076	0.078	0.110
		SNW	-0.914	0.081	0.097	-0.849	0.124	0.098	-1.076	0.069	0.106
3000	40	RSZ	-0.746	0.003	0.055	-0.706	0.016	0.055	-0.890	0.002	0.060
		SNW	-0.746	0.003	0.053	-0.706	0.013	0.054	-0.890	0.002	0.058
$h = g_2$											
100	15	RSZ	-1.760	0.100	0.176	-1.561	0.151	0.173	-0.931	0.518	0.163
		SNW	-1.760	0.079	0.152	-1.561	0.115	0.152	-0.931	0.487	0.155
300	20	RSZ	-1.403	0.015	0.101	-1.214	0.040	0.098	-0.632	0.391	0.093
		SNW	-1.403	0.012	0.090	-1.214	0.033	0.088	-0.632	0.373	0.090
1000	30	RSZ	-1.036	0.001	0.054	-0.846	0.005	0.053	-0.391	0.278	0.051
		SNW	-1.036	0.001	0.049	-0.846	0.002	0.049	-0.391	0.263	0.049
3000	40	RSZ	-0.808	0.000	0.031	-0.655	0.000	0.030	-0.284	0.111	0.029
		SNW	-0.808	0.000	0.029	-0.655	0.000	0.028	-0.284	0.110	0.029
$h = g_3$											
100	15	RSZ	-0.190	0.892	0.168	-0.143	0.901	0.166	-0.070	0.909	0.162
		SNW	-0.190	0.899	0.170	-0.143	0.913	0.168	-0.070	0.922	0.165
300	20	RSZ	-0.133	0.887	0.097	-0.100	0.907	0.097	-0.050	0.927	0.096
		SNW	-0.133	0.886	0.097	-0.100	0.911	0.097	-0.050	0.924	0.096
1000	30	RSZ	-0.081	0.905	0.053	-0.059	0.928	0.053	-0.024	0.945	0.053
		SNW	-0.081	0.903	0.053	-0.059	0.928	0.053	-0.024	0.942	0.053
3000	40	RSZ	-0.070	0.846	0.031	-0.053	0.878	0.031	-0.027	0.929	0.031
		SNW	-0.070	0.841	0.031	-0.053	0.877	0.031	-0.027	0.929	0.031

Notes: Bias, coverage rate, and interval length of asymptotic confidence intervals for $\Psi_2(h)$ estimated using MC = 1000 replications. The method column reports the type of functional and variance estimators (sandwich-type or Riesz-type) used to compute asymptotic confidence intervals. The Gaussian-Gaussian, Gaussian-Uniform, and Uniform-Uniform column groupings indicate random sampling laws of A_k and B_k, respectively. Bias is scaled by a factor of 10 to improve legibility.

Table G.2: Asymptotic inference results for $\Psi_2(\widehat{h}_{n,\lambda})$ with $\lambda = 10^{-4}$.

			Gaussian–Gaussian			Gaussian–Uniform			Uniform–Uniform		
n	K_n	Method	Bias	Coverage	Length	Bias	Coverage	Length	Bias	Coverage	Length
$h = g_1$											
100	15	RSZ	-0.115	0.900	0.259	-0.120	0.896	0.265	-0.335	0.865	0.298
		SNW	-0.115	0.925	0.269	-0.120	0.925	0.277	-0.335	0.888	0.311
300	20	RSZ	-0.095	0.917	0.150	-0.118	0.910	0.152	-0.229	0.871	0.162
		SNW	-0.095	0.922	0.153	-0.118	0.921	0.155	-0.229	0.880	0.165
1000	30	RSZ	-0.032	0.936	0.082	-0.035	0.939	0.082	-0.084	0.897	0.084
		SNW	-0.032	0.934	0.083	-0.035	0.941	0.083	-0.084	0.901	0.085
3000	40	RSZ	-0.035	0.939	0.047	-0.038	0.925	0.048	-0.063	0.894	0.048
		SNW	-0.035	0.940	0.048	-0.038	0.930	0.048	-0.063	0.895	0.048
$h = g_2$											
100	15	RSZ	-0.085	0.875	0.157	-0.034	0.884	0.156	-0.027	0.886	0.156
		SNW	-0.085	0.890	0.162	-0.034	0.901	0.161	-0.027	0.907	0.163
300	20	RSZ	-0.050	0.906	0.093	-0.025	0.925	0.092	-0.011	0.928	0.092
		SNW	-0.050	0.907	0.094	-0.025	0.930	0.093	-0.011	0.930	0.093
1000	30	RSZ	-0.033	0.930	0.051	-0.020	0.932	0.051	-0.004	0.937	0.051
		SNW	-0.033	0.926	0.051	-0.020	0.934	0.051	-0.004	0.941	0.051
3000	40	RSZ	-0.026	0.926	0.030	-0.017	0.941	0.030	-0.005	0.949	0.029
		SNW	-0.026	0.926	0.030	-0.017	0.943	0.030	-0.005	0.948	0.030
$h = g_3$											
100	15	RSZ	0.037	0.916	0.161	0.050	0.916	0.161	0.065	0.907	0.159
		SNW	0.037	0.926	0.167	0.050	0.923	0.166	0.065	0.923	0.165
300	20	RSZ	0.018	0.936	0.096	0.021	0.933	0.096	0.026	0.927	0.095
		SNW	0.018	0.939	0.097	0.021	0.938	0.097	0.026	0.937	0.097
1000	30	RSZ	0.014	0.953	0.053	0.015	0.953	0.053	0.018	0.949	0.053
		SNW	0.014	0.949	0.053	0.015	0.950	0.053	0.018	0.947	0.053
3000	40	RSZ	-0.001	0.944	0.031	-0.001	0.944	0.031	0.001	0.943	0.031
		SNW	-0.001	0.941	0.031	-0.001	0.944	0.031	0.001	0.942	0.031

Notes: Bias, coverage rate, and interval length of asymptotic confidence intervals for $\Psi_2(h)$ estimated using MC = 1000 replications. The method column reports the type of functional and variance estimators (sandwich-type or Riesz-type) used to compute asymptotic confidence intervals. The Gaussian-Gaussian, Gaussian-Uniform, and Uniform-Uniform column groupings indicate random sampling laws of A_k and B_k, respectively. Bias is scaled by a factor of 10 to improve legibility.

Table G.3: Asymptotic inference results for $\Psi_2(\widehat{h}_{n,\lambda})$ with $\lambda = 10^{-8}$.

			Gaussian–Gaussian			Gaussian–Uniform			Uniform–Uniform		
n	K_n	Method	Bias	Coverage	Length	Bias	Coverage	Length	Bias	Coverage	Length
$h = g_1$											
100	15	RSZ	0.139	0.889	0.237	0.129	0.878	0.239	0.086	0.871	0.260
		SNW	0.139	0.898	0.251	0.129	0.902	0.253	0.086	0.902	0.285
300	20	RSZ	0.021	0.932	0.144	0.022	0.928	0.145	-0.000	0.930	0.148
		SNW	0.021	0.937	0.149	0.022	0.935	0.149	-0.000	0.936	0.152
1000	30	RSZ	0.021	0.936	0.081	0.021	0.937	0.081	0.023	0.936	0.081
		SNW	0.021	0.938	0.082	0.021	0.939	0.082	0.023	0.937	0.082
3000	40	RSZ	-0.006	0.948	0.047	-0.006	0.951	0.047	-0.005	0.950	0.047
		SNW	-0.006	0.953	0.047	-0.006	0.954	0.047	-0.005	0.952	0.047
$h = g_2$											
100	15	RSZ	0.129	0.885	0.150	0.138	0.875	0.150	0.131	0.866	0.150
		SNW	0.129	0.914	0.159	0.138	0.909	0.159	0.131	0.908	0.162
300	20	RSZ	0.052	0.932	0.090	0.052	0.936	0.090	0.056	0.926	0.090
		SNW	0.052	0.935	0.093	0.052	0.937	0.093	0.056	0.932	0.093
1000	30	RSZ	0.016	0.937	0.051	0.017	0.936	0.051	0.020	0.934	0.050
		SNW	0.016	0.943	0.051	0.017	0.937	0.051	0.020	0.937	0.051
3000	40	RSZ	0.006	0.949	0.029	0.006	0.949	0.029	0.007	0.948	0.029
		SNW	0.006	0.952	0.029	0.006	0.951	0.029	0.007	0.951	0.029
$h = g_3$											
100	15	RSZ	0.114	0.894	0.158	0.117	0.889	0.157	0.134	0.882	0.156
		SNW	0.114	0.927	0.166	0.117	0.913	0.167	0.134	0.902	0.167
300	20	RSZ	0.042	0.923	0.095	0.042	0.916	0.095	0.046	0.918	0.094
		SNW	0.042	0.933	0.097	0.042	0.936	0.097	0.046	0.930	0.097
1000	30	RSZ	0.023	0.948	0.053	0.023	0.948	0.053	0.026	0.947	0.053
		SNW	0.023	0.948	0.053	0.023	0.949	0.053	0.026	0.945	0.053
3000	40	RSZ	0.002	0.943	0.031	0.003	0.944	0.031	0.004	0.941	0.031
		SNW	0.002	0.943	0.031	0.003	0.944	0.031	0.004	0.945	0.031

Notes: Bias, coverage rate, and interval length of asymptotic confidence intervals for $\Psi_2(h)$ estimated using MC = 1000 replications. The method column reports the type of functional and variance estimators (sandwich-type or Riesz-type) used to compute asymptotic confidence intervals. The Gaussian-Gaussian, Gaussian-Uniform, and Uniform-Uniform groupings indicate random sampling laws of A_k and B_k, respectively. Bias is scaled by a factor of 10 to improve legibility.

Table G.4: RWNN model specifications for bootstrap inference

	Weight Sampling		Scaling		
Activation	A_k	B_k	A_k	B_k	λ
Tanh	Gaussian	Gaussian	5.0	1.0	10^{-6}
Tanh	Gaussian	Uniform	5.0	2.0	10^{-6}
Tanh	Uniform	Uniform	10.0	5.0	10^{-6}
ReLU	Gaussian	Gaussian	5.0	2.0	10^{-6}
ReLU	Gaussian	Uniform	5.0	2.0	10^{-6}
ReLU	Uniform	Uniform	1.5	0.7	10^{-6}

G.3 Additional figures

In Figure G.1 we provide a plot of the three functions—g_1, g_2 and g_3—from Hall and Horowitz (2013) that we use in our simulations.

Figures G.2 and G.3 present the results obtained when running the simulations of Section 5.5 with different sampling laws for the RWNN weights.

Figures G.4 to G.6 plot the pointwise coverage probabilities associated with the evaluation functional over a fixed grid for the residual, wild and score bootstrap CIs. Figures G.7 to G.9 plot the pointwise bias of the RWNN ridge estimator over the same evaluation grid.

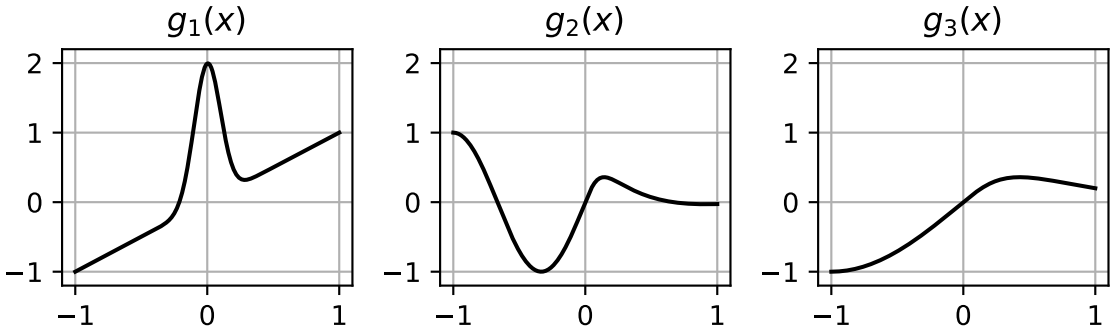
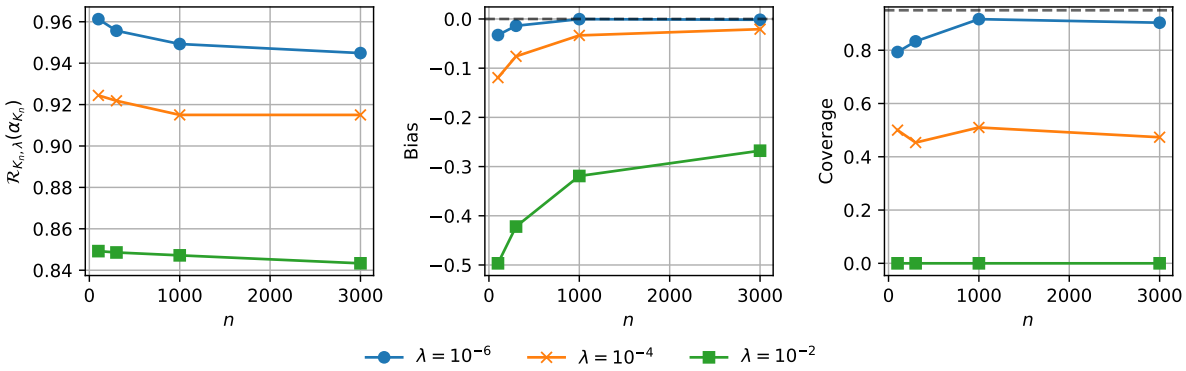
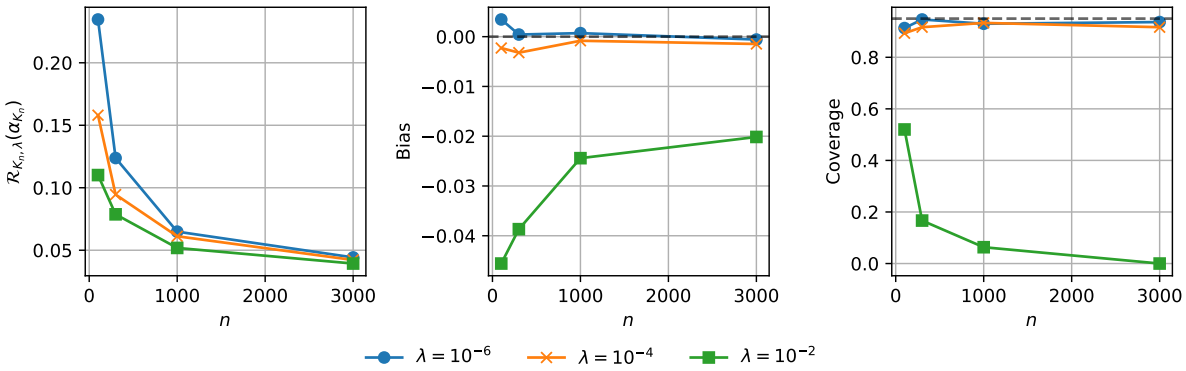


Figure G.1: Regression functions from Hall and Horowitz (2013).

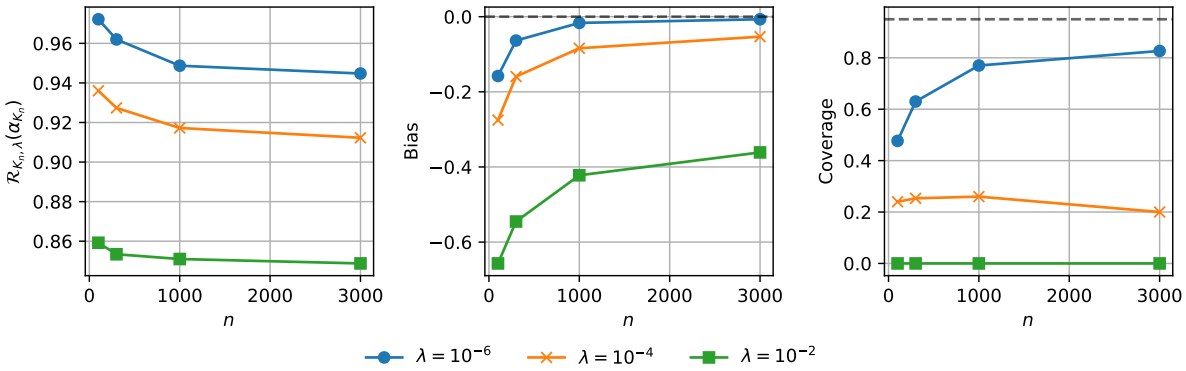


(a) $\Psi_0(g_1) = g_1(0)$

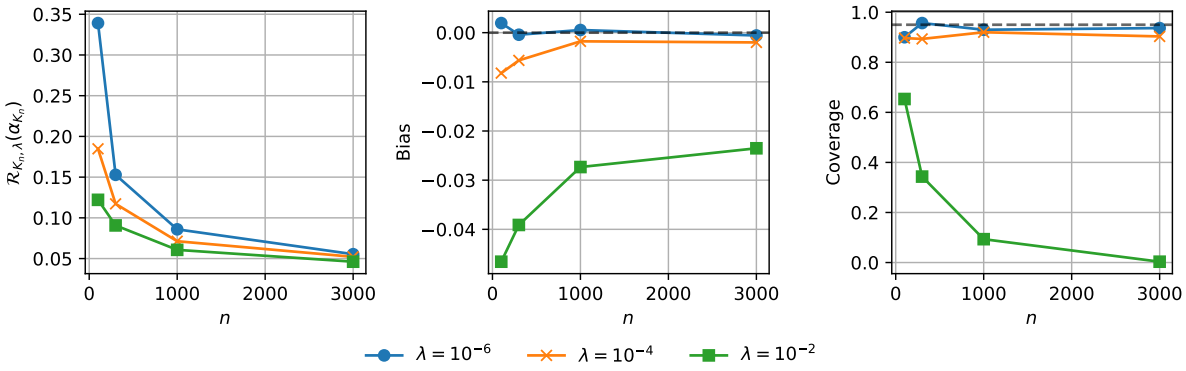


(b) $\Psi_1(g_1) = \mathbb{E}[g_1(X)(1+X)]$

Figure G.2: Bootstrap inference results for functionals of g_1 with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1])$, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, $\sigma_\epsilon^2 = 0.04$. RWNNs use tanh activation, $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_A^2)$ and $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-\sigma_B, \sigma_B])$, with $\sigma_A = 5$ and $\sigma_B = 2$. Ridge penalty is set as $\lambda \in \{10^{-6}, 10^{-4}, 10^{-2}\}$. Lines indicate averages across Monte Carlo replications.

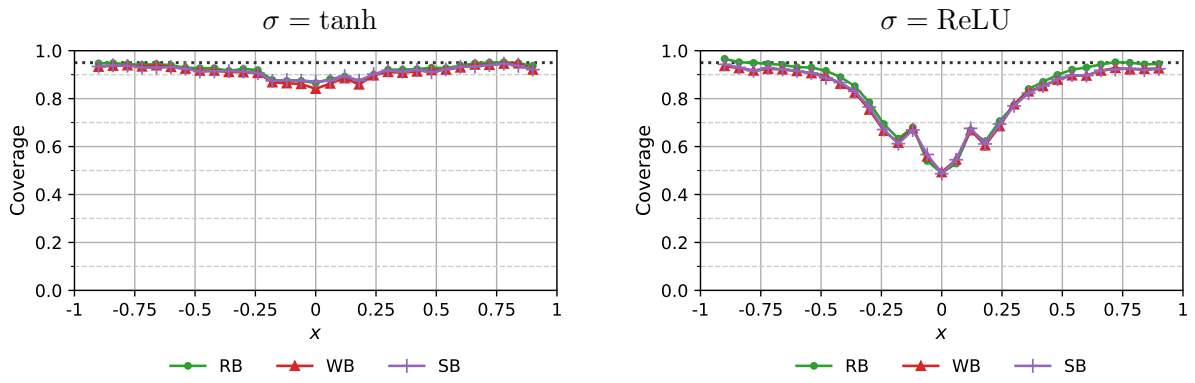


(a) $\Psi_0(g_1) = g_1(0)$

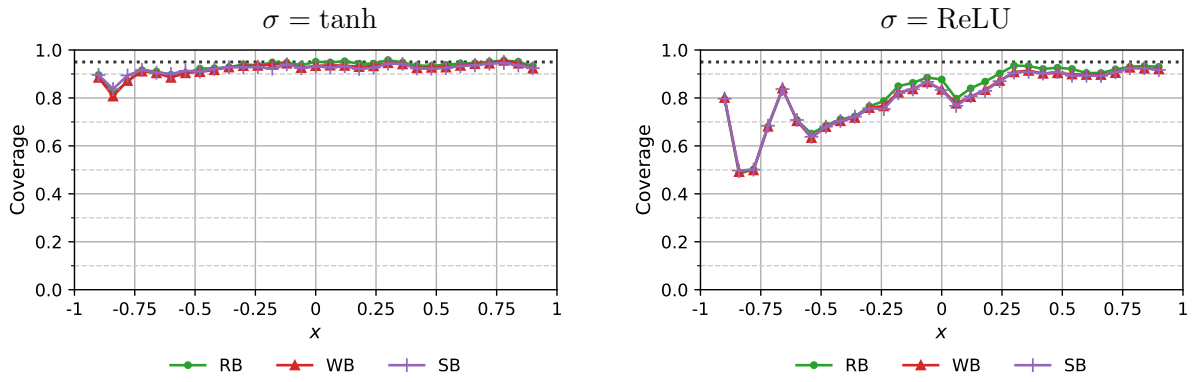


(b) $\Psi_1(g_1) = \mathbb{E}[g_1(X)(1+X)]$

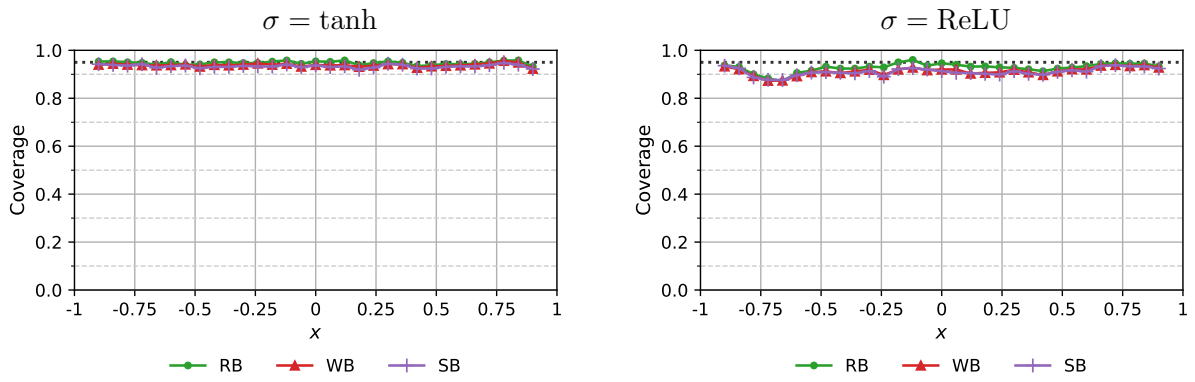
Figure G.3: Bootstrap inference results for functionals of g_1 with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-1,1])$, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, $\sigma_\epsilon^2 = 0.04$. RWNNs use tanh activation, $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-\sigma_A, \sigma_A])$ and $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}([-\sigma_B, \sigma_B])$, with $\sigma_A = 10$ and $\sigma_B = 5$. Ridge penalty is set as $\lambda \in \{10^{-6}, 10^{-4}, 10^{-2}\}$. Lines indicate averages across Monte Carlo replications.



(a) $g_1(x)$

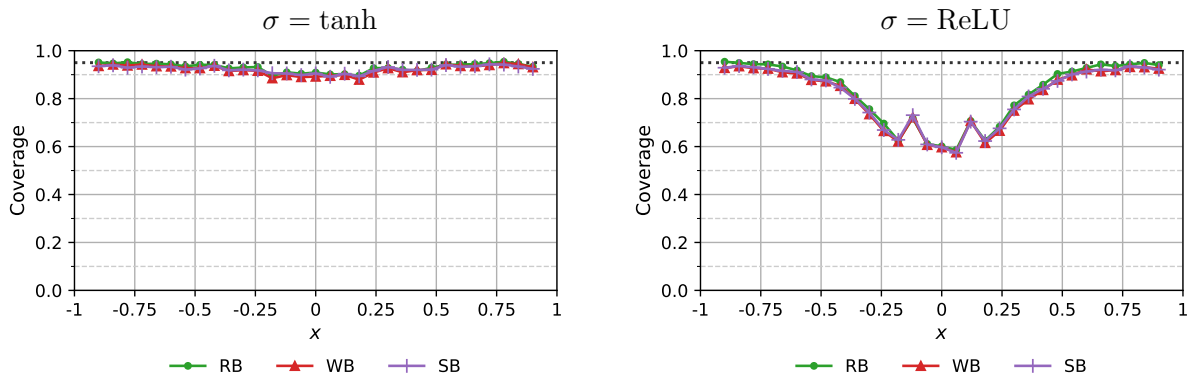


(b) $g_2(x)$

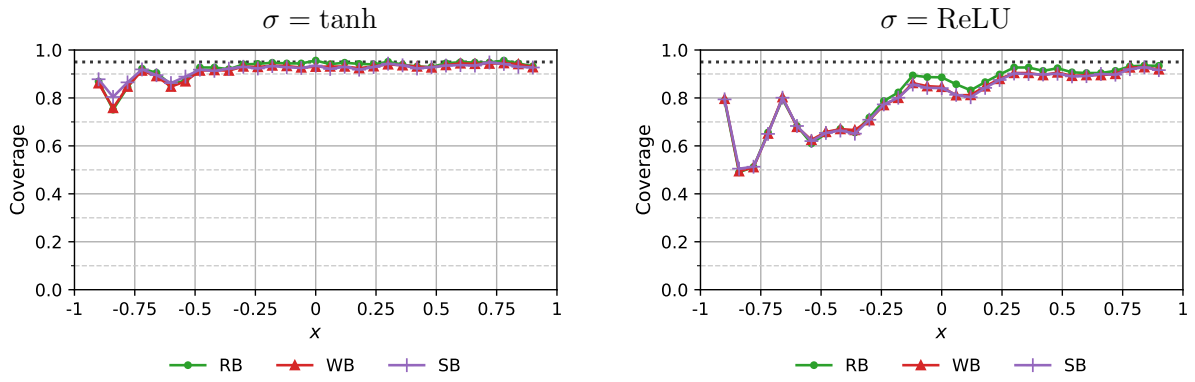


(c) $g_3(x)$

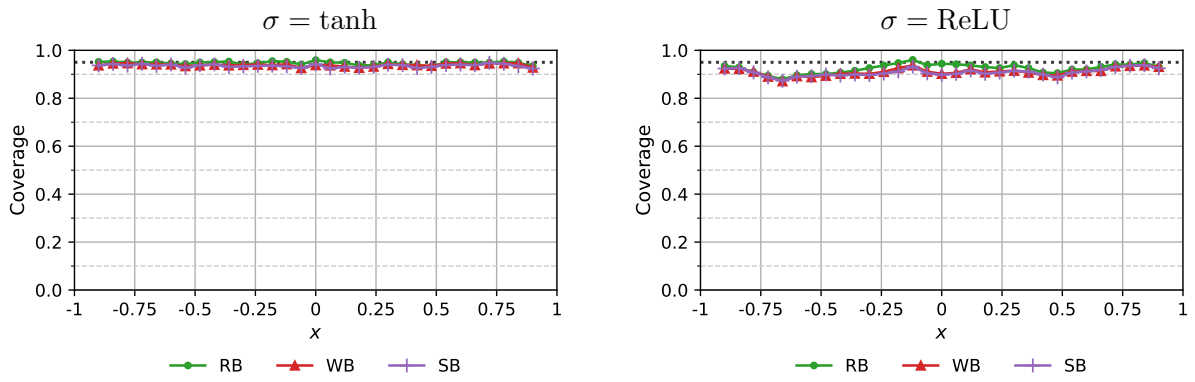
Figure G.4: Pointwise coverage probabilities of residual bootstrap (RB), wild bootstrap (WB), and score bootstrap (SB) confidence intervals for $\Psi_x(g_j) = g_j(x)$, $j \in \{1, 2, 3\}$ computed for $X_i \sim \mathcal{U}([-1, 1])$ and RWNNs with $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}$, $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}$ and $\sigma \in \{\tanh, \mathrm{ReLU}\}$. MC = 1000 Monte Carlo replications and $\mathfrak{B} = 1000$ bootstrap resamples.



(a) $g_1(x)$

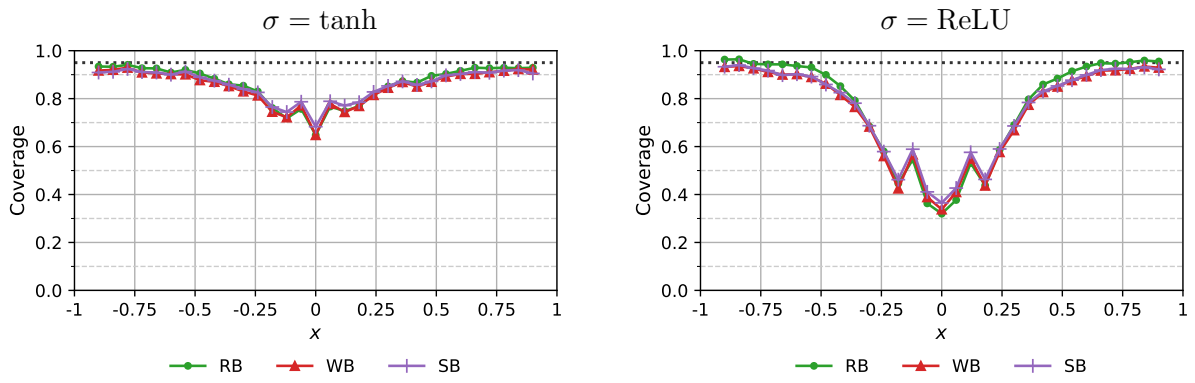


(b) $g_2(x)$

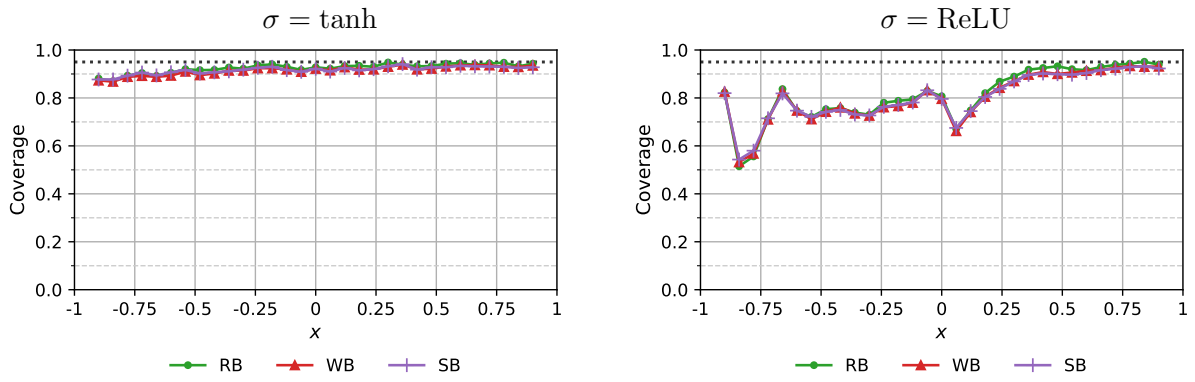


(c) $g_3(x)$

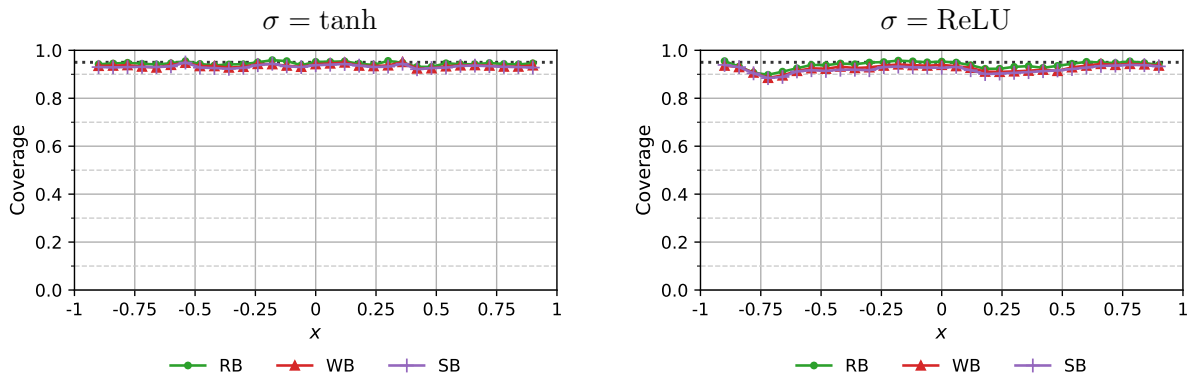
Figure G.5: Pointwise coverage probabilities of residual bootstrap (RB), wild bootstrap (WB), and score bootstrap (SB) confidence intervals for $\Psi_x(g_j) = g_j(x)$, $j \in \{1, 2, 3\}$ computed for $X_i \sim \mathcal{U}([-1, 1])$ and RWNNs with $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}$, $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}$, and $\sigma \in \{\tanh, \mathrm{ReLU}\}$. MC = 1000 Monte Carlo replications and $\mathfrak{B} = 1000$ bootstrap resamples.



(a) $g_1(x)$

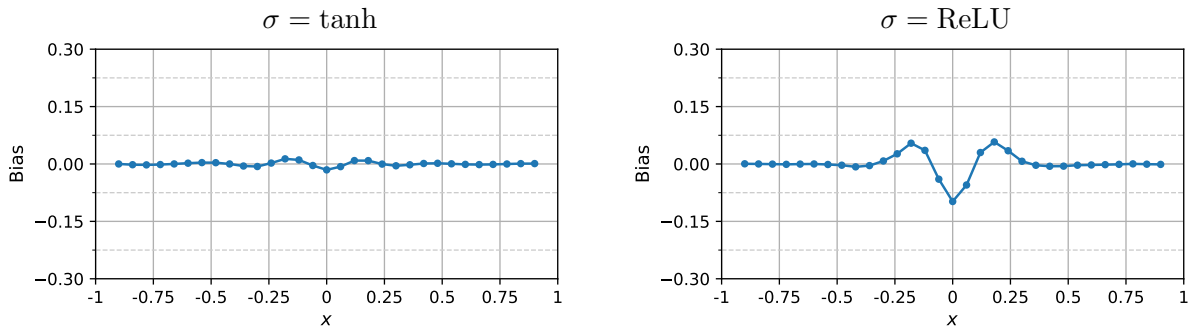


(b) $g_2(x)$

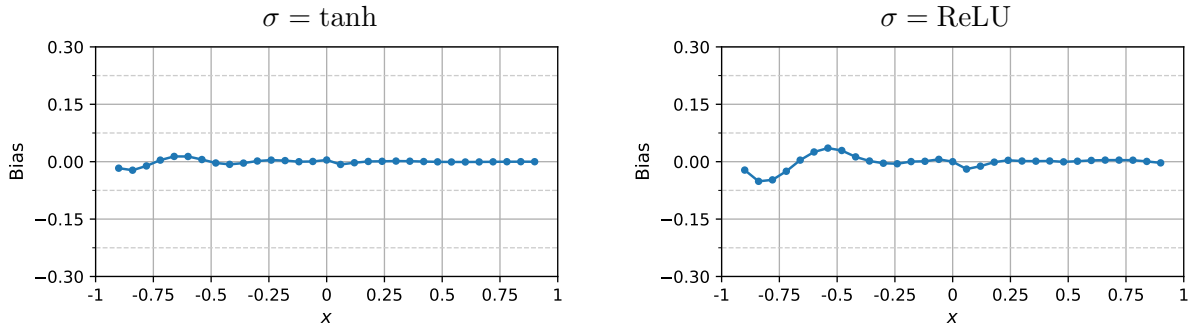


(c) $g_3(x)$

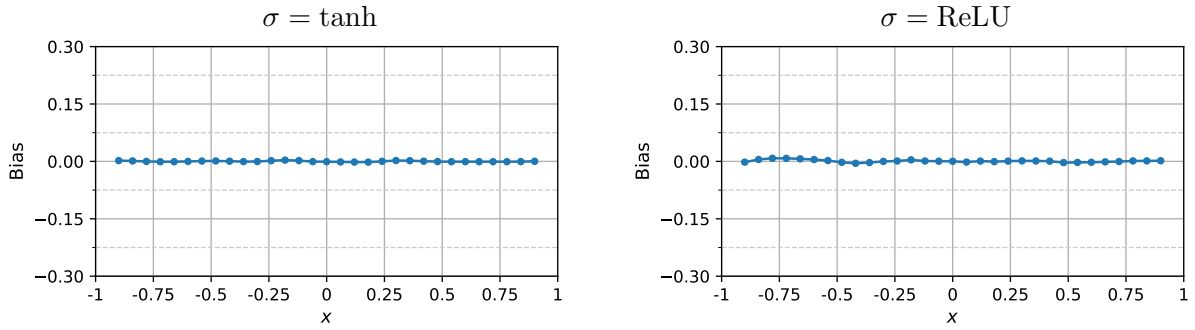
Figure G.6: Pointwise coverage probabilities of residual bootstrap (RB), wild bootstrap (WB), and score bootstrap (SB) confidence intervals for $\Psi_x(g_j) = g_j(x)$, $j \in \{1, 2, 3\}$ computed for $X_i \sim \mathcal{U}([-1, 1])$ and RWNNs with $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}$, $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}$ and $\sigma \in \{\tanh, \mathrm{ReLU}\}$. MC = 1000 Monte Carlo replications and $\mathfrak{B} = 1000$ bootstrap resamples.



(a) $g_1(x)$

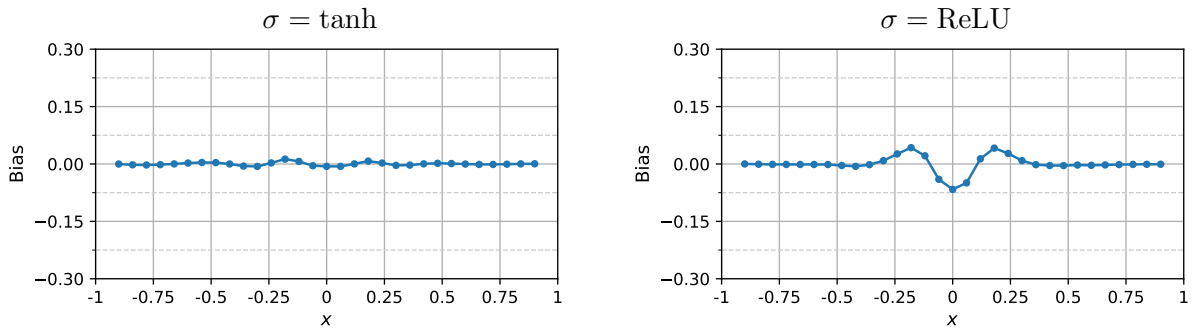


(b) $g_2(x)$

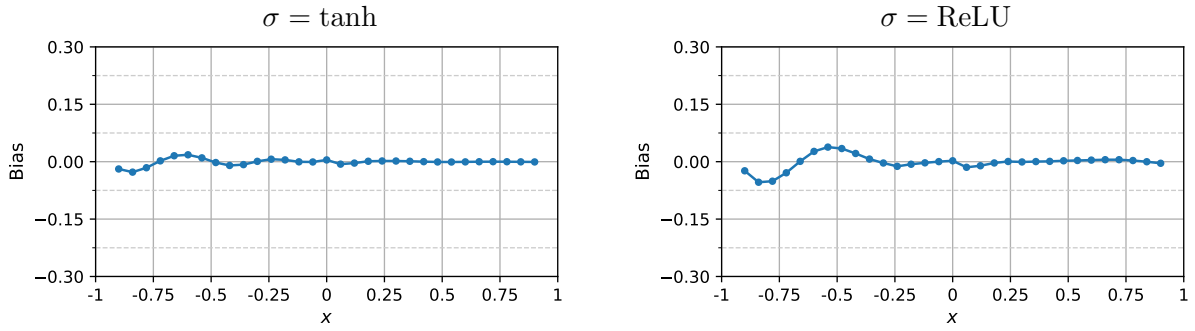


(c) $g_3(x)$

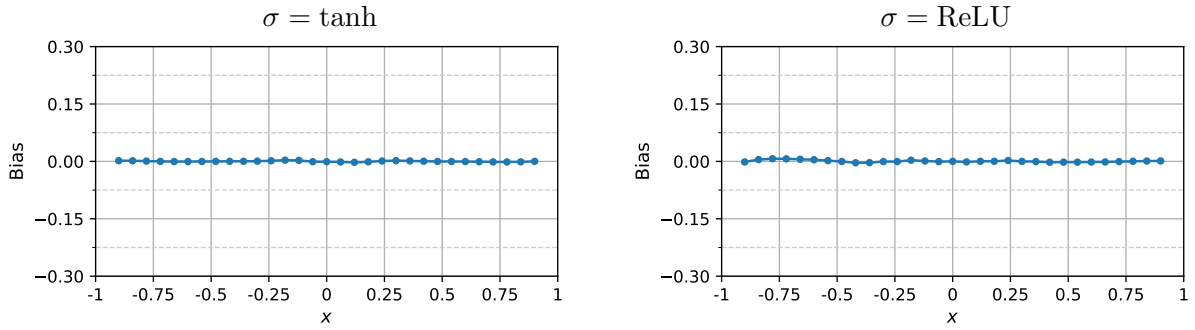
Figure G.7: Pointwise bias for $\Psi_x(g_j) = g_j(x)$, $j \in \{1, 2, 3\}$ computed for $X_i \sim \mathcal{U}([-1, 1])$ and RWNNs with $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}$, $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}$ and $\sigma \in \{\tanh, \mathrm{ReLU}\}$. MC = 1000 Monte Carlo replications.



(a) $g_1(x)$

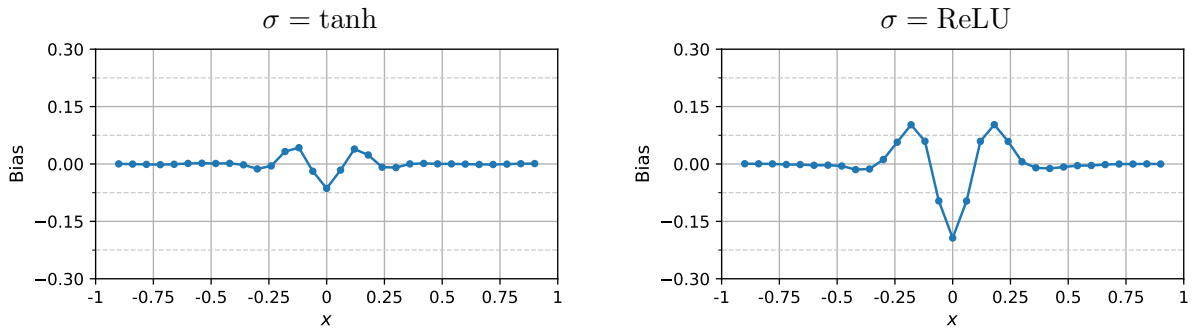


(b) $g_2(x)$

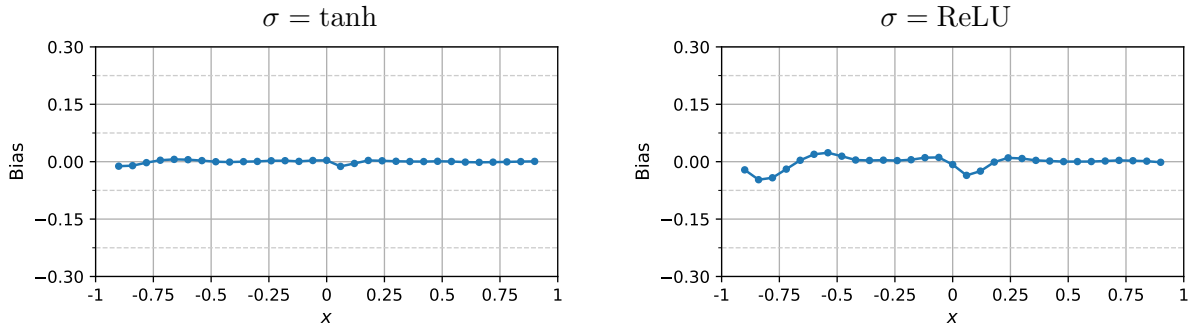


(c) $g_3(x)$

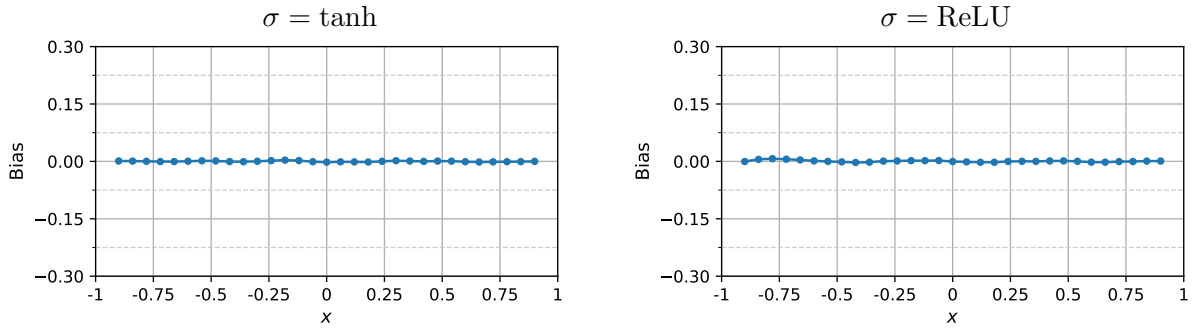
Figure G.8: Pointwise bias for $\Psi_x(g_j) = g_j(x)$, $j \in \{1, 2, 3\}$ computed for $X_i \sim \mathcal{U}([-1, 1])$ and RWNNs with $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}$, $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}$ and $\sigma \in \{\tanh, \mathrm{ReLU}\}$. MC = 1000 Monte Carlo replications.



(a) $g_1(x)$



(b) $g_2(x)$



(c) $g_3(x)$

Figure G.9: Pointwise bias for $\Psi_x(g_j) = g_j(x)$, $j \in \{1, 2, 3\}$ computed for $X_i \sim \mathcal{U}([-1, 1])$ and RWNNs with $A_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}$, $B_k \overset{\text{i.i.d.}}{\sim} \mathcal{U}$ and $\sigma \in \{\tanh, \mathrm{ReLU}\}$. MC = 1000 Monte Carlo replications.