Chapter 12
Appendix A: Technical Results on Networks

The reader may have noticed that for GLMs we have developed an asymptotic theory that allowed us to assess the quality of predictors as well as it allowed us to validate the fitted models. For networks there does not exist such a theory, yet, and the purpose of this appendix is to present more technical results on the asymptotic behavior of FN networks and their estimators that may lead to an asymptotic theory. This appendix hopefully stimulates further research in this field of statistical modeling.

12.1 Universality Theorems

We present a specific version of the universality theorems for shallow FN networks; we refer to the discussion in Sect. 7.2.2. This section follows Hornik et al. [192]. Choose an input dimension $q_0 \in \mathbb{N}$ and consider the set of all affine functions

$$\mathcal{A}^{q_0} = \Big\{ A : \{1\} \times \mathbb{R}^{q_0} \to \mathbb{R}; \quad \boldsymbol{x} \mapsto A(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle, \ \boldsymbol{w} \in \mathbb{R}^{q_0+1} \Big\},$$

we add a 0th component in feature $\boldsymbol{x} = (x_0 = 1, x_1, \ldots, x_{q_0})^\top \in \{1\} \times \mathbb{R}^{q_0}$ for the intercept. Choose a measurable (activation) function $\phi : \mathbb{R} \to \mathbb{R}$ and define

$$\Sigma^{q_0}(\phi) = \left\{ f : \{1\} \times \mathbb{R}^{q_0} \to \mathbb{R}; \ \boldsymbol{x} \mapsto f(\boldsymbol{x}) = \sum_{j=0}^{q_1} \beta_j \phi(A_j(\boldsymbol{x})), \ A_j \in \mathcal{A}^{q_0}, \beta_j \in \mathbb{R}, q_1 \in \mathbb{N} \right\}.$$

M. V. Wüthrich, M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, Springer Actuarial, https://doi.org/10.1007/978-3-031-12409-9_12

This is the set of all shallow FN networks $f(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{z}^{(1:1)}(\boldsymbol{x})\rangle$ with activation function ϕ and the linear output activation, see (7.8); the intercept component of the output is integrated into the 0th component $j = 0$. Moreover, we define the networks

$$\Sigma\Pi^{q_0}(\phi) = \Bigg\{ f : \{1\} \times \mathbb{R}^{q_0} \to \mathbb{R};\mapsto f(\boldsymbol{x}) = \sum_{j=0}^{q_1} \beta_j \prod_{k=1}^{l_j} \phi(A_{j,k}(\boldsymbol{x})),$$

$$A_{j,k} \in \mathcal{A}^{q_0}, \beta_j \in \mathbb{R}, l_j \in \mathbb{N}, q_1 \in \mathbb{N} \Bigg\}.$$

The latter networks contain the former $\Sigma^{q_0}(\phi) \subset \Sigma\Pi^{q_0}(\phi)$, by setting $l_j = 1$ for all $0 \le j \le q_1$. We are going to prove a universality theorem first for the networks $\Sigma\Pi^{q_0}(\phi)$, and afterwards for the shallow FN networks $\Sigma^{q_0}(\phi)$.

Definition 12.1 The function $\phi : \mathbb{R} \to [0, 1]$ is called a squashing function if it is non-decreasing with $\lim_{x\to-\infty} \phi(x) = 0$ and $\lim_{x\to\infty} \phi(x) = 1$.

Since squashing functions can have at most countably many discontinuities, they are measurable; a continuous and a non-continuous example are given by the sigmoid and by the step function activation, respectively, see Table 7.1.

Lemma 12.2 *The sigmoid activation function is Lipschitz with constant* $1/4$.

Proof The derivative of the sigmoid function is given by $\phi' = \phi(1 - \phi)$. This provides for the second derivative $\phi'' = \phi' - 2\phi\phi' = \phi'(1 - 2\phi)$. The latter is zero for $\phi(x) = 1/2$. This says that the maximal slope of ϕ is attained for $x = 0$ and it is $\phi'(0) = 1/4$. □

We denote by $\mathcal{C}(\mathbb{R}^{q_0})$ the set of all continuous functions from $\{1\} \times \mathbb{R}^{q_0}$ to $\mathbb{R}$, and by $\mathcal{M}(\mathbb{R}^{q_0})$ the set of all measurable functions from $\{1\} \times \mathbb{R}^{q_0}$ to $\mathbb{R}$. If the measurable activation function ϕ is continuous, we have $\Sigma\Pi^{q_0}(\phi) \subset \mathcal{C}(\mathbb{R}^{q_0})$, otherwise $\Sigma\Pi^{q_0}(\phi) \subset \mathcal{M}(\mathbb{R}^{q_0})$.

Definition 12.3 A subset $S \subset \mathcal{M}(\mathbb{R}^{q_0})$ is said to be uniformly dense on compacta in $\mathcal{C}(\mathbb{R}^{q_0})$ if for every compact subset $K \subset \{1\} \times \mathbb{R}^{q_0}$ the set S is ρ_K-dense in $\mathcal{C}(\mathbb{R}^{q_0})$ meaning that for all $\epsilon > 0$ and all $g \in \mathcal{C}(\mathbb{R}^{q_0})$ there exists $f \in S$ such that

$$\rho_K(g, f) = \sup_{\boldsymbol{x}\in K} |g(\boldsymbol{x}) - f(\boldsymbol{x})| < \epsilon.$$

Theorem 12.4 (Theorem 2.1 in Hornik et al. [192]) *Assume ϕ is a non-constant and continuous activation function. $\Sigma\Pi^{q_0}(\phi) \subset \mathcal{C}(\mathbb{R}^{q_0})$ is uniformly dense on compacta in $\mathcal{C}(\mathbb{R}^{q_0})$.*

Proof The proof is based on the Stone–Weierstrass theorem. We briefly recall the Stone–Weierstrass theorem. Assume $\mathcal{A}$ is a family of real functions defined on a set E. $\mathcal{A}$ is called an *algebra* if it is closed under addition, multiplication and scalar multiplication. A family $\mathcal{A}$ *separates points* in E, if for every $x, z \in E$ with $x \neq z$ there exists a function $A \in \mathcal{A}$ with $A(x) \neq A(z)$. The family $\mathcal{A}$ does *not vanish at any point* of E if for all $x \in E$ there exists a function $A \in \mathcal{A}$ such that $A(x) \neq 0$.

Let $\mathcal{A}$ be an algebra of continuous real functions on a compact set K. The Stone–Weierstrass theorem says that if $\mathcal{A}$ separates points in K and if it does not vanish at any point of K, then $\mathcal{A}$ is ρ_K-dense in the space of all continuous real functions on K.

Choose any compact set $K \subset \{1\} \times \mathbb{R}^{q_0}$. For any activation function ϕ, $\Sigma\Pi^{q_0}(\phi)$ is obviously an algebra. So there remains to prove that this algebra separates points and does not vanish at any point. Firstly, choose $\boldsymbol{x}, \boldsymbol{z} \in K$ such that $\boldsymbol{x} \neq \boldsymbol{z}$. Since ϕ is non-constant we can choose $a, b \in \mathbb{R}$ such that $\phi(a) \neq \phi(b)$. Next choose $A \in \mathcal{A}^{q_0}$ such that $A(\boldsymbol{x}) = a$ and $A(\boldsymbol{z}) = b$. Then, $\phi(A(\boldsymbol{x})) \neq \phi(A(\boldsymbol{z}))$ and $\Sigma\Pi^{q_0}(\phi)$ separates points. Secondly, since ϕ is non-constant, we can choose $a \in \mathbb{R}$ such that $\phi(a) \neq 0$. Moreover, choose weight $\boldsymbol{w} = (a, 0, \ldots, 0)^\top \in \mathbb{R}^{q_0+1}$. Then for this $A \in \mathcal{A}^{q_0}$, $A(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle = a$ for any $\boldsymbol{x} \in K$. Henceforth, $\phi(A(\boldsymbol{x})) \neq 0$, therefore $\Sigma\Pi^{q_0}(\phi)$ does not vanish at any point of K. The claim then follows from the Stone–Weierstrass theorem and using that ϕ is continuous by assumption. □

For Theorem 12.4 to hold, the activation function ϕ can be any continuous and non-constant function, i.e., it does not need to be a squashing function. This is fairly general, but it rules out the step function activation as it is not continuous. However, for squashing functions continuity is not needed and one still receives the uniformly dense on compacta property of $\Sigma\Pi^{q_0}(\phi)$ in $\mathcal{C}(\mathbb{R}^{q_0})$, this has been proved in Theorem 2.3 of Hornik et al. [192]. The following theorem also does not need continuity, i.e., we do not require $\Sigma^{q_0}(\phi) \subset \mathcal{C}(\mathbb{R}^{q_0})$ as ϕ only needs to be measurable (and squashing).

Theorem 12.5 (Universality, Theorem 2.4 in Hornik et al. [192]) *Assume ϕ is a squashing activation function. $\Sigma^{q_0}(\phi)$ is uniformly dense on compacta in $\mathcal{C}(\mathbb{R}^{q_0})$.*

Sketch of Proof For the (continuous) cosine activation function choice $\cos(\cdot)$, Theorem 12.4 applies to $\Sigma\Pi^{q_0}(\cos)$. Repeatedly applying the trigonometric identity $\cos(a)\cos(b) = \cos(a+b) - \cos(a-b)$ allows us to rewrite any trigonometric polynomial $\prod_{k=1}^{l_j} \cos(A_{j,k}(\boldsymbol{x}))$ as $\sum_{t=1}^{T} \alpha_t \cos(A_t(\boldsymbol{x}))$ for suitable $A_t \in \mathcal{A}^{q_0}$, $\alpha_t \in \mathbb{R}$ and $T \in \mathbb{N}$. This allows us to identify $\Sigma^{q_0}(\cos) = \Sigma\Pi^{q_0}(\cos)$. As a consequence of Theorem 12.4, shallow FN networks $\Sigma^{q_0}(\cos)$ are uniformly dense on compacta in $\mathcal{C}(\mathbb{R}^{q_0})$.

The remaining part relies on approximating the cosine activation function. Firstly, Lemma A.2 of Hornik et al. [192] says that for any continuous squashing function ψ and any $\epsilon > 0$ there exists $H_\epsilon(x) = \sum_{j=1}^{q_1} \beta_j \phi(w_0^j + w_1^j x) \in \Sigma^1(\phi)$, $x \in \mathbb{R}$, such that

$$\sup_{x \in \mathbb{R}} |\psi(x) - H_\epsilon(x)| < \epsilon. \tag{12.1}$$

For the proof we refer to Lemma A.2 of Hornik et al. [192], it uses that ψ is a continuous squashing function, implying that for every $\delta \in (0, 1)$ there exists $m > 0$

such that $\psi(-m) < \delta$ and $\psi(m) > 1 - \delta$. Approximation $H_\epsilon \in \Sigma^1(\phi)$ of ψ is then constructed on $(-m, m)$ so that the error bound holds (and for δ sufficiently small).

Secondly, choose $\epsilon > 0$ and $M > 0$, there exists $\cos_{M,\epsilon} \in \Sigma^1(\phi)$ such that

$$\sup_{x \in [-M,M]} \left| \cos(x) - \cos_{M,\epsilon}(x) \right| < \epsilon. \tag{12.2}$$

This is Lemma A.3 of Hornik et al. [192]; to prove this, we consider the cosine squasher of Gallant–White [150], for $x \in \mathbb{R}$

$$\chi(x) = \frac{1}{2}\left(1 + \cos\left(x + \frac{3\pi}{2}\right)\right) \mathbb{1}_{\{-\pi/2 \le x \le \pi/2\}} + \mathbb{1}_{\{x > \pi/2\}} \in [0, 1].$$

This is a continuous squashing function. Adding, subtracting and scaling a *finite* number of affinely shifted versions of the cosine squasher χ can exactly replicate the cosine on $[-M, M]$. Claim (12.2) then follows from the fact that we need a finite number of cosine squashers χ to replicate the cosine on $[-M, M]$, the triangle equality, and the fact that the (continuous) cosine squasher can be approximated arbitrarily well in $\Sigma^1(\phi)$ using (12.1).

The final step is to patch everything together. Consider $\sum_{t=1}^T \alpha_t \cos(A_t(\boldsymbol{x}))$ which approximates on the compact set $K \subset \{1\} \times \mathbb{R}^{q_0}$ a given continuous function $g \in \mathcal{C}(\mathbb{R}^{q_0})$ with a given tolerance $\epsilon/2$. Choose $M > 0$ such that $A_t(K) \subset [-M, M]$ for all $1 \le t \le T$. Note that this M can be found because K is compact, A_t are continuous and T is finite. Define $T' = T \sum_{t=1}^T |\alpha_t| < \infty$. By (12.2) we can then choose $\cos_{M,\epsilon/(2T')} \in \Sigma^1(\phi)$ such that

$$\sup_{\boldsymbol{x} \in K} \left| \sum_{t=1}^T \alpha_t \cos(A_t(\boldsymbol{x})) - \sum_{t=1}^T \alpha_t \cos_{M,\epsilon/(2T')}(A_t(\boldsymbol{x})) \right| < \epsilon/2.$$

This completes the proof. □

12.2 Consistency and Asymptotic Normality

Universality Theorem 12.5 tells us that we can approximate any compactly supported continuous function arbitrarily well by a sufficiently large shallow FN network, say, with sigmoid activation function ϕ. The next natural question is whether we can *learn* these approximations from data $(Y_i, \boldsymbol{x}_i)_{i \ge 1}$ that follow the true but unknown regression function $\boldsymbol{x} \mapsto \mu_0(\boldsymbol{x})$, or in other words whether we have consistency for a certain class of learning methods. This is the question addressed, e.g., in White [379, 380], Barron [26], Chen–Shen [73], Döhler–Rüschendorf [109] and Shen et al. [336]. This turns the algebraic universality question into a statistical question about consistency.

Assume that the true data model satisfies

$$Y = \mu_0(\boldsymbol{x}) + \varepsilon = \mathbb{E}[Y|\boldsymbol{x}] + \varepsilon, \tag{12.3}$$

for a continuous regression function $\mu_0 : \mathcal{X} \to \mathbb{R}$ on a compact set $\mathcal{X} \subset \{1\} \times \mathbb{R}^{q_0}$, and with a centered error ε satisfying $\mathbb{E}[|\varepsilon|^{2+\delta}] < \infty$ for some $\delta > 0$ and being independent of $\boldsymbol{x}$. The question now is whether we can learn this (true) regression function μ_0 from independent data $(Y_i, \boldsymbol{x}_i)$, $1 \le i \le n$, obeying (12.3). Throughout this section we use the square error loss function $L(y, a) = (y - a)^2$. For given data, this results in solving

$$\widetilde{\mu}_n = \underset{\mu \in \mathcal{C}(\mathcal{X})}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_i, \mu(\boldsymbol{x}_i)\right) = \underset{\mu \in \mathcal{C}(\mathcal{X})}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mu(\boldsymbol{x}_i)\right)^2, \tag{12.4}$$

where $\mathcal{C}(\mathcal{X})$ denotes the set of continuous functions on the compact set $\mathcal{X} \subset \{1\} \times \mathbb{R}^{q_0}$. The main question is whether estimator $\widetilde{\mu}_n$ approaches the true regression function μ_0 for increasing sample size n.

Typically, the family of continuous functions $\mathcal{C}(\mathcal{X})$ is much too rich to be able to solve optimization problem (12.4), and the solution may have undesired properties. In particular, the solution to (12.4) will over-fit to the data for any sample size n, and consistency will not hold, see, e.g., Section 2.2.1 in Chen [72]. Therefore, the optimization needs to be done over (well-chosen) smaller sets $\mathcal{S}_n \subset \mathcal{C}(\mathcal{X})$. For instance, $\mathcal{S}_n$ can be the set of shallow FN networks having a maximal width $q_1 = q_1(n)$, depending on the sample size n of the data. Considering this regression problem in a non-parametric sense, we let grow these sets $\mathcal{S}_n$ with the sample size n. This idea is attributed to Grenander [172] and it is called the *method of sieve estimators* of μ_0. We define for $d \in \mathbb{N}$, $\Delta > 0$, $\widetilde{\Delta} > 0$ and activation function ϕ

$$\mathcal{S}(d, \Delta, \widetilde{\Delta}, \phi) = \left\{ f \in \Sigma^{q_0}(\phi);\ q_1 = d,\ \sum_{j=0}^{q_1} |\beta_j| \le \Delta,\ \max_{1 \le j \le q_1} \sum_{l=0}^{q_0} |w_{l,j}| \le \widetilde{\Delta} \right\}.$$

These sets $\mathcal{S}(d, \Delta, \widetilde{\Delta}, \phi)$ are shallow FN networks of a given width $q_1 = d$ and with some restrictions on the network parameters.[1] We then choose increasing sequences

[1] The bound $\sum_{j=0}^{q_1} |\beta_j| \le \Delta$ in $\mathcal{S}(d, \Delta, \widetilde{\Delta}, \phi)$ allows us to view this set of shallow FN networks as a symmetric convex hull of the family of functions $\mathcal{S}_0(\phi) = \{\boldsymbol{x} \mapsto \phi(A(\boldsymbol{x}));\ A \in \mathcal{A}^{q_0}\}$, see Sect. 2.6.3 in Van der Vaart–Wellner [364]. If we choose an increasing activation function ϕ, this family of functions $\phi \circ A$ is a composition of a fixed increasing function ϕ and a finite dimensional vector space $\mathcal{A}^{q_0}$ of functions A. This implies that $\mathcal{S}_0(\phi)$ is a VC-class saying that it has a finite Vapnik–Chervonenkis (VC) dimension [365]; see also Condition A and Theorem 2.1 in Döhler–Rüschendorf [109]. This VC-class is an important property in many proofs as it leads to a finite covering (metric entropy) of function spaces, and this allows to apply limit theorems to point processes, we refer to Van der Vaart–Wellner [364].

$(d_n)_{n\geq 1}$, $(\Delta_n)_{n\geq 1}$ and $(\widetilde{\Delta}_n)_{n\geq 1}$ which provides us with an increasing sequence of sieves (becoming finer as n increases)

$$\ldots \subseteq \mathcal{S}_n(\phi) \overset{\text{def.}}{=} \mathcal{S}(d_n, \Delta_n, \widetilde{\Delta}_n, \phi) \subseteq \mathcal{S}_{n+1}(\phi) \overset{\text{def.}}{=} \mathcal{S}(d_{n+1}, \Delta_{n+1}, \widetilde{\Delta}_{n+1}, \phi) \subseteq \ldots .$$

The following corollary is a simple consequence of Theorem 12.5.

Corollary 12.6 *Assume ϕ is a squashing activation function, and let the increasing sequences $(d_n)_{n\geq 1}$, $(\Delta_n)_{n\geq 1}$ and $(\widetilde{\Delta}_n)_{n\geq 1}$ tend to infinity for $n \to \infty$. Then $\bigcup_{n\geq 1} \mathcal{S}_n(\phi)$ is uniformly dense in $\mathcal{C}(\mathcal{X})$.*

This corollary says that for any regression function $\mu_0 \in \mathcal{C}(\mathcal{X})$ we can find $n \in \mathbb{N}$ and $\mu_n \in \mathcal{S}_n(\phi)$ such that μ_n is arbitrarily close to μ_0; remark that all functions are continuous on the compact set $\mathcal{X}$, and uniformly dense means $\rho_{\mathcal{X}}$-dense in that case. Corollary 12.6 does not hold true if $\Delta_n \equiv \Delta > 0$, for all n. In that case we can only approximate the smaller function class $\overline{\bigcup_{n\geq 1} \mathcal{S}_n(\phi)} \subset \mathcal{C}(\mathcal{X})$. This is going to be used in one of the cases, below.

For increasing sequences $(d_n)_{n\geq 1}$, $(\Delta_n)_{n\geq 1}$ and $(\widetilde{\Delta}_n)_{n\geq 1}$ we define the sieve estimator $(\widehat{\mu}_n)_{n\geq 1}$ by

$$\widehat{\mu}_n = \underset{\mu \in \mathcal{S}_n(\phi)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_i, \mu(\boldsymbol{x}_i)\right). \tag{12.5}$$

Under the following assumptions one can prove a consistency theorem.

Assumption 12.7 *Choose a complete probability space* $(\Omega, \mathcal{A}, \mathbb{P})$[2] *and* $\mathcal{X} = \{1\} \times [0, 1]^{q_0}$.

(1) *Assume $\mu_0 \in \mathcal{C}(\mathcal{X})$. Assume $(Y_i, \boldsymbol{X}_i)_{i\geq 1}$ are i.i.d. on $(\Omega, \mathcal{A}, \mathbb{P})$ following the regression structure (12.3) with ε_i being centered, having $\mathbb{E}[|\varepsilon_i|^{2+\delta}] < \infty$ for some $\delta > 0$ and being independent of $\boldsymbol{X}_i$. Set $\sigma^2 = \mathrm{Var}(\varepsilon_i) < \infty$.*
(2) *The activation function ϕ is the sigmoid function.*
(3) *The sequences $(d_n)_{n\geq 1}$, $(\Delta_n)_{n\geq 1}$ and $(\widetilde{\Delta}_n)_{n\geq 1}$ are increasing and tending to infinity as $n \to \infty$ with $d_n \Delta_n^2 \log(d_n \Delta_n) = o(n)$.*

Most results that we are going to present below hold for activation functions that are Lipschitz. The sigmoid activation function is Lipschitz, see Lemma 12.2.

The following considerations are based on the pseudo-norm, given $(\boldsymbol{X}_i)_{1\leq i\leq n}$,

$$\|\mu\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\mu(\boldsymbol{X}_i))^2} \qquad \text{for } \mu \in \mathcal{C}(\mathcal{X}).$$

[2] A probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is complete if for any $\mathbb{P}$-null set $B \in \mathcal{A}$ with $\mathbb{P}[B] = 0$ and every subset $A \subset B$ it follows that $A \in \mathcal{A}$.

This is a pseudo-norm because it is positive $\|\mu\|_n \geq 0$, absolutely homogeneous $\|a\mu\|_n = |a| \, \|\mu\|_n$ and the triangle inequality holds, but it is not definite because $\|\mu\|_n = 0$ does not imply that μ is the zero function (i.e. it is not point-separating). This pseudo-norm $\|\cdot\|_n$ depends on the (random) features $(\boldsymbol{X}_i)_{1\leq i\leq n}$ and, therefore, the subsequent statements involving this pseudo-norm hold in probability. The following result provides consistency, and that the true regression function μ_0, indeed, can be learned from i.i.d. data.

Theorem 12.8 (Consistency, Theorem 3.1 of Shen et al. [336]) *Under Assumption 12.7, the sieve estimator $(\widehat{\mu}_n)_{n\geq 1}$ in (12.5) exists. We have consistency $\|\widehat{\mu}_n - \mu_0\|_n \to 0$ in probability as $n \to \infty$, i.e., for all $\epsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}\left[\|\widehat{\mu}_n - \mu_0\|_n > \epsilon\right] = 0.$$

Remarks 12.9

- Such a consistency result for FN networks has first been proved in Theorem 3.3 of White [380], however, on slightly different spaces and under slightly different assumptions. Similar consistency results have been obtained for related point process situations by Döhler–Rüschendorf [109] and for time-series in White [380] and Chen–Shen [73].
- Item (3) of Assumption 12.7 gives upper complexity bounds on shallow FN networks as a function of the sample size n of the data, so that asymptotically they do not over-fit to the data. These bounds allow for much freedom in the choice of the growth rates, and different choices may lead to different speeds of convergence. The conditions of Assumption 12.7 are, e.g., satisfied for $\Delta_n = O(\log n)$ and $d_n = O(n^{1-\delta'})$, for any small $\delta' > 0$. Under these choices, the complexity d_n of the shallow FN network grows rather quickly. Table 1 of White [380] gives some examples, for instance, if for $n = 100$ data points we have a shallow FN network with 5 neurons, then these magnitudes support 477 neurons for $n = 10'000$ and 45'600 neurons for $n = 1'000'000$ data points (for the specific choice $\delta' = 0.01$). Of course, these numbers do not provide any practical guidance on the selection of the (shallow) FN network size.
- Theorem 12.8 requires that we can explicitly calculate the sieve estimator $\widehat{\mu}_n$, i.e., the global minimizer of the objective function in (12.5). In practical applications, relying on gradient descent algorithms, typically, this is not the case. Therefore, Theorem 12.8 is mainly of theoretical value saying that learning the true regression function μ_0 is possible within FN networks.

Sketch of Proof of Theorem 12.8 The proof of this theorem is based on a theorem in White–Woolridge [381] which states that if we have a sequence $(\mathcal{S}_n(\phi))_{n\geq 1}$ of compact subsets of $\mathcal{C}(\mathcal{X})$, and if $L_n : \Omega \times \mathcal{S}_n(\phi) \to \overline{\mathbb{R}}$ is a $\mathcal{A} \otimes \mathcal{B}(\mathcal{S}_n(\phi))/\mathcal{B}(\overline{\mathbb{R}})$-measurable sequence, $n \geq 1$, with $L_n(\omega, \cdot)$ being lower-semicontinuous on $\mathcal{S}_n(\phi)$ for all $\omega \in \Omega$. Then, there exists $\widehat{\mu}_n : \Omega \to \mathcal{S}_n(\phi)$ being $\mathcal{A}/\mathcal{B}(\mathcal{S}_n(\phi))$-measurable such that for each $\omega \in \Omega$, $L_n(\omega, \widehat{\mu}_n(\omega)) = \min_{\mu\in\mathcal{S}_n(\phi)} L_n(\omega, \mu)$. For the proof of the compactness of $\mathcal{S}_n(\phi)$ in $\mathcal{C}(\mathcal{X})$ we need that d_n and Δ_n are finite for any n. This then provides the existence of the sieve estimator, for details we refer Lemma 2.1 and Corollary 2.1 in Shen et al. [336]. The proof of the consistency result then uses the growth rates on $(d_n)_{n\geq 1}$ and $(\Delta_n)_{n\geq 1}$, for the details of the proof we refer to Theorem 3.1 in Shen et al. [336]. □

The next step is to analyze the rates of convergence of the sieve estimator $\widehat{\mu}_n \to \mu_0$, as $n \to \infty$. These rates heavily depend on (additional) regularity assumptions on the true regression function $\mu_0 \in \mathcal{C}(\mathcal{X})$; we refer to Remark 3 in Sect. 5 of Chen–Shen [73]. Here, we present some results of Shen et al. [336]. From the proof of Theorem 12.8 we know that $\mathcal{S}_n(\phi)$ is a compact set in $\mathcal{C}(\mathcal{X})$. This motivates to consider the closest approximation $\pi_n\mu \in \mathcal{S}_n(\phi)$ to $\mu \in \mathcal{C}(\mathcal{X})$. The uniform denseness of $\bigcup_{n\geq 1}\mathcal{S}_n(\phi)$ in $\mathcal{C}(\mathcal{X})$ implies that $\pi_n\mu$ converges to μ. The aforementioned rates of convergence of the sieve estimators will depend on how fast $\pi_n\mu_0 \in \mathcal{S}_n(\phi)$ converges to the true regression function $\mu_0 \in \mathcal{C}(\mathcal{X})$.

If one cannot determine the global minimum of (12.5), then often an accurate approximation is sufficient. For this one introduces an approximate sieve estimator. A sequence $(\widehat{\mu}_n)_{n\geq 1}$ is called an *approximate sieve estimator* if

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_n(\boldsymbol{X}_i))^2 \;\leq\; \inf_{\mu\in\mathcal{S}_n(\phi)} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(\boldsymbol{X}_i))^2 + O_P(\eta_n), \tag{12.6}$$

where $(\eta_n)_{n\geq 1}$ is a positive sequence converging to 0 as $n \to \infty$. The last term $O_P(\eta_n)$ denotes stochastic boundedness meaning that for all $\epsilon > 0$ there exits $K_\epsilon > 0$ such that for all $n \geq 1$

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_n(\boldsymbol{X}_i))^2 - \inf_{\mu\in\mathcal{S}_n(\phi)} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(\boldsymbol{X}_i))^2 > K_\epsilon \eta_n\right] < \epsilon.$$

Theorem 12.10 (Theorem 4.1 of Shen et al. [336], Without Proof) *Set Assumption 12.7. If*

$$\eta_n = O\left(\min\left\{\|\pi_n\mu_0 - \mu_0\|_n^2,\ \frac{d_n \log(d_n\Delta_n)}{n},\ \frac{d_n \log n}{n}\right\}\right),$$

the following stochastic boundedness holds for $n \geq 1$

$$\|\widehat{\mu}_n - \mu_0\|_n = O_P\left(\max\left\{\|\pi_n\mu_0 - \mu_0\|_n, \sqrt{\frac{d_n \log n}{n}}\right\}\right).$$

Remarks 12.11

- Assumption 12.7 implies that $d_n \log(d_n\Delta_n) = o(n)$ as $n \to \infty$. Therefore, $\eta_n \to 0$ as $n \to \infty$.

- The statement in Theorem 4.1 of Shen et al. [336] is more involved because it is stated under slightly different assumptions. Our assumptions are sufficient for having consistency of the sieve estimator, see Theorem 12.8, and making these assumptions implies that the rate of convergence in Theorem 12.10 is determined by the rate of convergence of $\|\pi_n\mu_0-\mu_0\|_n$ and $(n^{-1}d_n\log n)^{1/2}$, see Remark 4.1 in Shen et al. [336].
- The rate of convergence in Theorem 12.10 crucially depends on the rate $\|\pi_n\mu_0-\mu_0\|_n$, as $n\to\infty$. If μ_0 lies in the (sub-)space of functions with finite first absolute moments of the Fourier magnitude distributions, denoted by $\mathcal{F}(\mathcal{X})\subset\mathcal{C}(\mathcal{X})$, Makavoz [262] has shown that $\|\pi_n\mu_0-\mu_0\|_n$ decays at least as $d_n^{-(q_0+1)/(2q_0)}=d_n^{-1/2-1/(2q_0)}$, this has improved the rate of $d_n^{-1/2}$ obtained by Barron [25]. This space $\mathcal{F}(\mathcal{X})$ allows for the choices $d_n=(n/\log n)^{q_0/(2+q_0)}$, $\Delta_n\equiv\Delta>0$ and $\widetilde{\Delta}_n\equiv\widetilde{\Delta}>0$ to receive consistency and the following rate of convergence, see Chen–Shen [73] and Remark 4.1 in Shen et al. [336],

$$\|\widehat{\mu}_n-\mu_0\|_n=O_P(r_n^{-1}),$$

for

$$r_n=\left(\frac{n}{\log n}\right)^{(q_0+1)/(4q_0+2)}\qquad n\geq 2. \tag{12.7}$$

Note that $1/4\leq(q_0+1)/(4q_0+2)\leq 1/2$. Thus, this is a slower rate than the square root rule of typical asymptotic normality, for instance, for $q_0=1$ we get $1/3$. Interestingly, Barron [26] proposes the choice $d_n\sim(n/\log n)^{1/2}$ to receive an approximation rate of $(n/\log n)^{-1/4}$.

Also note that the space $\mathcal{F}(\mathcal{X})$ allows us to choose a finite $\Delta_n\equiv\Delta>0$ in the sieves, thus, here we do not receive denseness of the sieves in the space of continuous functions $\mathcal{C}(\mathcal{X})$, but only in the space of functions with finite first absolute moments of the Fourier magnitude distributions $\mathcal{F}(\mathcal{X})$.

The last step is to establish the asymptotic normality. For this we have to define perturbations of shallow FN networks $\mu\in\mathcal{S}_n(\phi)$. Choose $\eta_n\in(0,1)$ and define the function

$$\widetilde{\mu}_n(\mu)=(1-\eta_n^{1/2})\mu+\eta_n^{1/2}(\mu_0+1).$$

This allows us to state the following asymptotic normality result.

Theorem 12.12 (Theorem 5.1 of Shen et al. [336], Without Proof) *Set Assumption 12.7. We make the following additional assumptions: suppose $\eta_n=o(n^{-1})$ and choose ϱ_n such that we have stochastic boundedness $\varrho_n\|\widehat{\mu}_n-\mu_0\|_n=O_P(1)$. Let the following conditions hold:*

(C1) $d_n\Delta_n\log(d_n\Delta_n)=o(n^{1/4})$;
(C2) $n\varrho_n^{-2}/\Delta_n^{\delta}=o(1)$;

(C3) $\sup_{\mu \in \mathcal{S}_n(\phi): \|\mu-\mu_0\|_n \le \varrho_n^{-1}} \|\pi_n \widetilde{\mu}_n(\mu) - \widetilde{\mu}_n(\mu)\|_n = O_P(\varrho_n \eta_n)$;

(C4) $\sup_{\mu \in \mathcal{S}_n(\phi): \|\mu-\mu_0\|_n \le \varrho_n^{-1}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(\pi_n \widetilde{\mu}_n(\mu)(\boldsymbol{X}_i) - \widetilde{\mu}_n(\mu)(\boldsymbol{X}_i)\right) = O_P(\eta_n)$.

We have the following asymptotic normality for $n \to \infty$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\widehat{\mu}_n(\boldsymbol{X}_i) - \mu_0(\boldsymbol{X}_i)\right) \;\Rightarrow\; \mathcal{N}\left(0, \sigma^2\right).$$

The assumptions of Theorem 12.12 require a slower growth rate d_n on the shallow FN network compared to the consistency results. Shen et al. [336] bring forward the argument that for the asymptotic normality result to hold, the shallow FN network should grow slower in order to get the Gaussian property, otherwise the sieve estimator may skew towards the true function μ_0. Conditions (C3)–(C4) on the other side give lower growth rates on the networks such that the approximation error decreases sufficiently fast.

If the variance parameter $\sigma^2 = \mathrm{Var}(\varepsilon_i)$ is not known, we can empirically estimate it

$$\widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \widehat{\mu}_n(\boldsymbol{X}_i)\right)^2.$$

Theorem 5.2 in Shen et al. [336] proves that this estimator is consistent for σ^2, and the asymptotic normality result also holds true under this estimated variance parameter (using Slutsky's theorem), and under the same assumptions as in Theorem 12.12.

12.3 Functional Limit Theorem

Horel–Giesecke [190] push the above asymptotic results even one step further. Note that the asymptotic normality of Theorem 12.12 is not directly useful for variable selection, since the asymptotic result integrates over the feature space $\mathcal{X}$. Horel–Giesecke [190] prove a functional limit theorem which we briefly review in this section.

A q_0-tuple $\alpha = (\alpha_1, \ldots, \alpha_{q_0})^\top \in \mathbb{N}_0^{q_0}$ is called a multi-index, and we set $|\alpha| = \alpha_1 + \ldots + \alpha_{q_0}$. Define the derivative operator

$$\nabla^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_{q_0}^{\alpha_{q_0}}}.$$

Consider the compact feature space $\mathcal{X} = \{1\} \times [0,1]^{q_0}$ with $q_0 \geq 3$. Choose a distribution ν on this feature space $\mathcal{X}$ and define the L^2-space

$$L^2(\mathcal{X}, \nu) = \left\{ \mu : \mathcal{X} \to \mathbb{R} \text{ measurable; } \mathbb{E}_\nu[\mu(\boldsymbol{X})^2] = \int_{\mathcal{X}} \mu(\boldsymbol{x})^2 d\nu(\boldsymbol{x}) < \infty \right\}.$$

Next, define the Sobolev space for $k \in \mathbb{N}$

$$W^{k,2}(\mathcal{X}, \nu) = \left\{ \mu \in L^2(\mathcal{X}, \nu);\ \nabla^\alpha \mu \in L^2(\mathcal{X}, \nu) \text{ for all } \alpha \in \mathbb{N}_0^{q_0} \text{ with } |\alpha| \leq k \right\},$$

where $\nabla^\alpha \mu$ is the weak derivative of μ. The motivation for studying Sobolev spaces is that for sufficiently large k and the existence of weak derivatives $\nabla^\alpha \mu \in L^2(\mathcal{X}, \nu)$, $|\alpha| \leq k$, we eventually receive a classical derivative of μ, see below. We define the Sobolev norm for $\mu \in W^{k,2}(\mathcal{X}, \nu)$ by

$$\|\mu\|_{k,2} = \left(\sum_{|\alpha| \leq k} \mathbb{E}_\nu \left[\left(\nabla^\alpha \mu(\boldsymbol{X}) \right)^2 \right] \right)^{1/2}.$$

The normed Sobolev space $(W^{k,2}(\mathcal{X}, p), \|\cdot\|_{k,2})$ is a Hilbert space. Since we would like to consider gradient-based methods, we consider the following space

$$\mathcal{C}_B^1(\mathcal{X}, \nu) = \left\{ \mu : \mathcal{X} \to \mathbb{R} \text{ continuously differentiable; } \|\mu\|_{\lfloor q_0/2 \rfloor + 2, 2} \leq B \right\}, \tag{12.8}$$

for some positive constant $B < \infty$. We will assume that the true regression function $\mu_0 \in \mathcal{C}_B^1(\mathcal{X}, \nu)$, thus, the true regression function has a bounded Sobolev norm $\|\cdot\|_{\lfloor q_0/2 \rfloor + 2,2}$ of maximal size B. Assume that $\mathring{\mathcal{X}} \subset \mathbb{R}^{q_0}$ is the open interior of $\mathcal{X}$ (excluding the intercept component), and that ν is absolutely continuous w.r.t. the Lebesgue measure with a strictly positive and bounded density on $\mathcal{X}$ (excluding the intercept component). The Sobolev number of the space $W^{\lfloor q_0/2 \rfloor + 2,2}(\mathring{\mathcal{X}}, \nu)$ is given by $m = \lfloor q_0/2 \rfloor + 2 - q_0/2 \geq 1.5 > 1$. The Sobolev embedding theorem then tells us that for any function $\mu \in W^{\lfloor q_0/2 \rfloor + 2,2}(\mathring{\mathcal{X}}, \nu)$, there exists an $\lfloor m \rfloor$-times continuously differentiable function on $\mathring{\mathcal{X}}$ that is equal to μ a.e., thus, the class of equivalent functions $\mu \in W^{\lfloor q_0/2 \rfloor + 2,2}(\mathring{\mathcal{X}}, \nu)$ has a representative in $\mathcal{C}^1(\mathring{\mathcal{X}})$, $\lfloor m \rfloor = 1$, this motivates the consideration of the space in (12.8).

In practice, the bound B needs a careful consideration because the true μ_0 is unknown. Therefore, B should be sufficiently large so that μ_0 is contained in the space $\mathcal{C}_B^1(\mathcal{X}, \nu)$ and, on the other hand, it should not be too large as this will weaken the power of the tests, below.

We choose the sigmoid activation function for ϕ and we consider the approximate sieve estimators $(\widehat{\mu}_n)_{n\geq 1}$ for given data $(Y_i, \boldsymbol{X}_i)_i$ obtained by a solution to

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_n(\boldsymbol{X}_i))^2 \;\leq\; \inf_{\mu\in\mathcal{S}_n(\phi)} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(\boldsymbol{X}_i))^2 + o_P(1), \tag{12.9}$$

where we allow for an error term $o_P(1)$ that converges in probability to zero as $n \to \infty$. In contrast to (12.6) we do not specify the error rate, here.

Assumption 12.13 *Choose a complete probability space* $(\Omega, \mathcal{A}, \mathbb{P})$ *and* $\mathcal{X} = \{1\} \times [0, 1]^{q_0}$.

(1) *Assume* $\mu_0 \in \mathcal{C}^1_B(\mathcal{X}, \nu)$ *for some* $B > 0$, *and* $(Y_i, \boldsymbol{X}_i)_{i\geq 1}$ *are i.i.d. on* $(\Omega, \mathcal{A}, \mathbb{P})$ *following regression structure* (12.3) *with* ε_i *being centered, having* $\mathbb{E}[|\varepsilon_i|^{2+\delta}] < \infty$ *for some* $\delta > 0$, *being absolutely continuous w.r.t. the Lebesgue measure, and being independent of* $\boldsymbol{X}_i$*; the features* $\boldsymbol{X}_i \sim \nu$ *are absolutely continuous w.r.t. the Lebesgue measure having a bounded and strictly positive density on* $\mathcal{X}$ *(excluding the intercept component). Set* $\sigma^2 = \mathrm{Var}(\varepsilon_i) < \infty$.

(2) *The activation function* ϕ *is the sigmoid function.*

(3) *The sequence* $(d_n)_{n\geq 1}$ *is increasing and going to infinity satisfying* $d_n^{2+1/q_0}\log(d_n) = O(n)$ *as* $n \to \infty$, *and* $\Delta_n \equiv \Delta > 0$, $\widetilde{\Delta}_n \equiv \widetilde{\Delta} > 0$ *for* $n \geq 1$.

(4) *Define* $L_\mu(\boldsymbol{X}, \varepsilon) = -2\varepsilon(\mu(\boldsymbol{X}) - \mu_0(\boldsymbol{X})) + (\mu(\boldsymbol{X}) - \mu_0(\boldsymbol{X}))^2$, *and it holds for* $n \geq 2$

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(L_{\widehat{\mu}_n}(\boldsymbol{X}_i, \varepsilon_i) - \mathbb{E}_\nu\left[L_{\widehat{\mu}_n}(\boldsymbol{X}_1, \varepsilon_1)\right]\right)\\ &\qquad\leq \inf_{h\in\mathcal{C}^1_B(\mathcal{X},\nu)} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(L_{\mu_0+h/r_n}(\boldsymbol{X}_i, \varepsilon_i) - \mathbb{E}_\nu\left[L_{\mu_0+h/r_n}(\boldsymbol{X}_1, \varepsilon_1)\right]\right) + o_P(r_n^{-1}),\end{aligned}$$

for r_n *being the rate defined in* (12.7).

The first three items of this assumption are rather similar to Assumption 12.7 which provides consistency in Theorem 12.8 and the rates of convergence in Theorem 12.10. Item (4) of Assumption 12.13 needs to be compared to (C3)–(C4) of Theorem 12.12 which is used for getting the asymptotic normality. $(r_n)_n$ is the rate that provides convergence in probability of the sieve estimator to the true regression function, and this magnitude is used for the perturbation, see also (C3)–(C4) in Theorem 12.12.

Theorem 12.14 (Asymptotics, Theorem 1 of Horel–Gisecke [190], Without Proof) *Under Assumption 12.13 the approximate sieve estimator* $(\widehat{\mu}_n)_{n\geq 1}$ (12.9) *converges weakly in the metric space* $(\mathcal{C}^1_B(\mathcal{X}, \nu), d_\nu)$ *with* $d_\nu(\mu, \mu') = \mathbb{E}_\nu[(\mu(\boldsymbol{X}) - \mu'(\boldsymbol{X}))^2]$*:*

$$r_n\left(\widehat{\mu}_n - \mu_0\right) \;\Rightarrow\; \mu^\star \qquad \textit{as } n \to \infty,$$

where $\mu^\star$ *is the* arg max *of the Gaussian process* $\{G_\mu;\ \mu \in \mathcal{C}^1_B(\mathcal{X}, \nu)\}$ *with mean zero and covariance function* $\mathrm{Cov}(G_\mu, G_{\mu'}) = 4\sigma^2\mathbb{E}_\nu[\mu(\boldsymbol{X})\mu'(\boldsymbol{X})]$.

Remarks 12.15 We highlight the differences between Theorems 12.12 and 12.14.

- Theorem 12.12 provides a convergence in distribution to a Gaussian random variable, whereas the limit in Theorem 12.14 is a random function $\boldsymbol{x} \mapsto \mu^\star(\boldsymbol{x}) = \mu^\star_\omega(\boldsymbol{x})$, $\omega \in \Omega$, thus, the former convergence result integrates over the (empirical) feature distribution, whereas the latter also allows for a point-wise consideration in feature $\boldsymbol{x}$.
- The former theorem does not allow for variable selection in $\boldsymbol{X}$ whereas the latter does because the limiting function still discriminates different feature values.
- For the proof of Theorem 12.14 we refer to Horel–Giesecke [190]. It is based on asymptotic results on empirical point processes; we refer to Van der Vaart–Wellner [364]. The Gaussian process $\{G_\mu;\ \mu \in \mathcal{C}^1_B(\mathcal{X}, \nu)\}$ is parametrized by the (totally bounded) space $\mathcal{C}^1_B(\mathcal{X}, \nu)$, and it is continuous over this compact index space. This implies that it takes its maximum. Uniqueness of the maximum then gives us the random function $\mu^\star$ which exactly describes the limiting distribution of $r_n(\widehat{\mu}_n - \mu_0)$ as $n \to \infty$.

12.4 Hypothesis Testing

Theorem 12.14 can be used to provide a significance test for feature component selection, similarly to the LRT and the Wald test presented in Sect. 5.3.2 on GLMs. We define gradient-based test statistics, for $1 \le j \le q_0$, and w.r.t. the approximate sieve estimator $\widehat{\mu}_n \in \mathcal{S}_n(\phi)$ given in (12.9),

$$\Lambda_j^{(n)} = \int_{\mathcal{X}} \left(\frac{\partial \widehat{\mu}_n(\boldsymbol{x})}{\partial x_j}\right)^2 d\nu(\boldsymbol{x}) \qquad \text{and} \qquad \widehat{\Lambda}_j^{(n)} = \frac{1}{n}\sum_{i=1}^n \left(\frac{\partial \widehat{\mu}_n(\boldsymbol{X}_i)}{\partial x_j}\right)^2.$$

The test statistics $\Lambda_j^{(n)}$ integrates the squared partial derivative of the sieve estimator $\widehat{\mu}_n$ w.r.t. the distribution ν, whereas $\widehat{\Lambda}_j^{(n)}$ can be considered as its empirical counterpart if $\boldsymbol{X} \sim \nu$. Note that both test statistics depend on the data $(Y_i, \boldsymbol{X}_i)_{1 \le i \le n}$ determining the sieve estimator $\widehat{\mu}_n$, see (12.9). These test statistics are used to test the following null hypothesis H_0 against the alternative hypothesis H_1 for the true regression function $\mu_0 \in \mathcal{C}^1_B(\mathcal{X}, \nu)$

$$H_0 : \lambda_j = \mathbb{E}_\nu\left[\left(\frac{\partial \mu_0(\boldsymbol{X})}{\partial x_j}\right)^2\right] = 0 \qquad \text{against} \qquad H_1 : \lambda_j \neq 0. \tag{12.10}$$

We emphasize that the expression λ_j in (12.10) is a deterministic number, for this reason we use the expected value notation $\mathbb{E}_\nu[\cdot]$. This in contrast to $\Lambda_j^{(n)}$, which is only a conditional expectation, conditionally given the data $(Y_i, \boldsymbol{X}_i)_{1\le i\le n}$.

Proposition 12.16 (Theorem 2 and Proposition 3 of Horel–Giesecke [190], Without Proof) *Under Assumption 12.13 and under the null hypothesis H_0 we have for $n \to \infty$*

$$r_n^2 \Lambda_j^{(n)}, r_n^2 \widehat{\Lambda}_j^{(n)} \;\Rightarrow\; \Psi_j \overset{\text{def.}}{=} \int_{\mathcal{X}} \left(\frac{\partial \mu^\star(\boldsymbol{x})}{\partial x_j} \right)^2 d\nu(\boldsymbol{x}). \tag{12.11}$$

In order to use this proposition we need to be able to calculate the limiting distribution characterized by random variable Ψ_j. The maximal argument $\mu^\star$ of the Gaussian process $\{G_\mu;\ \mu \in \mathcal{C}_B^1(\mathcal{X}, \nu)\}$ is given by a random function such that for all $\omega \in \Omega$, $\mu_\omega^\star(\cdot)$ fulfills

$$G_{\mu_\omega^\star(\cdot)}(\omega) \ge G_\mu(\omega) \qquad \text{for all } \mu \in \mathcal{C}_B^1(\mathcal{X}, \nu).$$

A discretization and simulation approach can be explored to approximate this maximal argument $\mu^\star$ for different $\omega \in \Omega$, see Section 5.7 in Horel–Giesecke [190].

1. Sample random functions f_k from $\mathcal{C}_B^1(\mathcal{X}, \nu)$, $k \ge 1$. The universality theorems suggest that we sample these random functions f_k from the sieves $(\mathcal{S}_n \cap \mathcal{C}_B^1(\mathcal{X}, \nu))_{n\ge 1}$. This requires sampling dimension q_1 of the shallow FN network and the corresponding network weights. This provides us with candidate functions $f_1, \ldots, f_K \in \mathcal{C}_B^1(\mathcal{X}, \nu)$, these candidate functions can be understood as a random covering of the (totally bounded) index space $\mathcal{C}_B^1(\mathcal{X}, \nu)$.
2. Simulate K-dimensional multivariate Gaussian random variables $\boldsymbol{G}^{(t)}$ (i.i.d.) with mean zero and (empirical) covariance matrix

$$\widehat{\Sigma} = \left(\frac{1}{n} \sum_{i=1}^n f_k(\boldsymbol{X}_i) f_l(\boldsymbol{X}_i) \right)_{1\le k,l\le K}.$$

   These random variables $\boldsymbol{G}^{(1)}, \ldots, \boldsymbol{G}^{(T)}$ play the role of discretized random samples of the Gaussian process $\{G_\mu;\ \mu \in \mathcal{C}_B^1(\mathcal{X}, \nu)\}$.
3. The empirical arg max of the sample $G^{(t)}$, $1 \le t \le T$, is obtained by

$$\widehat{\mu}_t^\star \;=\; \underset{f_k:\ 1\le k\le K}{\arg\max}\; G_{f_k}^{(t)},$$

   where $G_{f_k}^{(t)}$ is the k-th component of $\boldsymbol{G}^{(t)}$.

4. The empirical distribution of the following sample $\widehat{\Psi}_j^{(t)}$, $1 \leq t \leq T$, gives us an approximation to the limiting distribution in Proposition 12.16

$$\widehat{\Psi}_j^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \widehat{\mu}_t^{\star}(\boldsymbol{X}_i)}{\partial x_j} \right)^2,$$

i.e., under the null hypothesis H_0 we approximate the right-hand side of (12.11) by the empirical distribution of $(\widehat{\Psi}_j^{(t)})_{1 \leq t \leq T}$.

We close this section we some remarks.

Remarks 12.17

- The quality of the empirical approximation $(\widehat{\Psi}_j^{(t)})_{1 \leq t \leq T}$ to the limiting distribution of Ψ_j will depend on how well we cover the index set $\mathcal{C}_B^1(\mathcal{X}, \nu)$. We could try to use covering theorems to control the accuracy. However, this is often too challenging. The simulation approach presented above suffers from not giving us any control on the quality of this covering, nor is it clear how the Sobolev norm condition for B in (12.8) can efficiently be checked during the simulation approach. We highlight that this Sobolev norm bound $\| f_k \|_{\lfloor q_0/2 \rfloor + 2, 2} \leq B$ is crucial when we want to empirically estimate the distribution of Ψ_j; under special assumptions Horel–Giesecke [190] prove in their Theorem 4 that Ψ_j scales as B^2. Thus, if we do not have any control over the Sobolev norm of the sampled shallow FN networks f_k, the above simulation algorithm is not useful to approximate the limiting distribution in Proposition 12.16.
- The assumptions of Proposition 12.16 require that $\boldsymbol{X} \sim \nu$ has a strictly positive density over the entire feature space $\mathcal{X}$ (excluding the intercept component). This is necessary to be able to capture any non-zero partial derivative $\partial \mu_0(\boldsymbol{x})/\partial x_j$ over the entire feature space $\mathcal{X}$. In practical applications, where we rely on a finite sample $(\boldsymbol{X}_i)_{1 \leq i \leq n}$, this may be problematic and needs some care. For instance, there may be the situation where the samples cluster in two disjoint regions, say $C_1 \subset \mathcal{X}$ and $C_2 \subset \mathcal{X}$, because we may have $\nu(C_1 \cup C_2) \approx 1$. That is, in that case we rarely have observations $\boldsymbol{X}_i$ not lying in one of these two clusters. If $\partial \mu_0(\boldsymbol{x})/\partial x_j = 0$ on these two clusters $\boldsymbol{x} \in C_1 \cup C_2$, but if μ_0 has a very steep slope between the two clusters (i.e., if they are really different in terms of μ_0), then the test on this finite sample will not find the significant slope.
- The distribution $\boldsymbol{X} \sim \nu$ of the features is assumed to be absolutely continuous on the hypercube $[0, 1]^{q_0}$, this is not fulfilled for binary and categorical features.
- Another question is how the test of Proposition 12.16 is affected by collinearity in feature components. Note that we only test one component at a time. Moreover, we would like to highlight the j-dependency in the limiting random variable Ψ_j. This dependency is induced by the properties of the feature distribution ν that may not be exchangeable in the components of $\boldsymbol{x}$.