JOURNAL OF NONPARAMETRIC STATISTICS
https://doi.org/10.1080/10485252.2023.2209218

Taylor &Francis
Taylor&Francis Group

Asymptotic properties of neural network sieve estimators

Xiaoxi Shen[a], Chang Jiang[b], Lyudmila Sakhanenko[c] and Qing Lu[b]

[a]TexasState University,San Marcos,TX, USA;[b]University of Florida, Gainesville, FL,USA[c]; Michigan State University, East Lansing, Ml,USA

ABSTRACT

Neural networks havebecome one of the most popularly used methods in machine learning and artificial intelligence. Due to the universal approximation theorem, a neural network with one hidden layer can approximate any continuous function on compact support as long as the number of hidden units is sufficiently large. Statistically, a neural network can be classified into a nonlinear regression framework. However, if we consider it parametrically, due to the unidentifiability of the parameters, it is difficult to derive its asymptotic properties. Instead, we consider the estimation problem in a nonparametric regression framework and use the results from sieve estimation to establish the consistency,the ratesofconvergence and the asymptotic normality of the neural network estimators. We also illustrate the validity of the theories via simulations.

ARTICLE HISTORY
Received 6 June 2020
Accepted 20 April 2023

KEYWORDS
Empirical processes ; entropy integral

1. Introduction

With the success of machine learning and artificial intelligence in research and industry, neural networks have become popularly used methods nowadays. Many newly developed machine learning methodsare basedon deep neural networksand have achieved greatclassification and prediction accuracy. We refer interested readers to Goodfellow et al. (2016) for more background and details. In classical statistical learning theory, the consistency and the rate of convergence of the empirical risk minimisation principle are of great interest. Many upper bounds have been established for the empirical risk and the sample complexity based on the growth function and the Vapnik- Chervonenkis dimension (see, e.g., Vapnik 1998; Anthony and Bartlett 2009; Devroye et al. 2013). However,few studies have focused on the asymptotic properties for neural networks. As Thomas J. Sargent said, 'artificial intelligence is actually statistics, but in a very gorgeous phrase, it is statistics '. So it is natural and worthwhile to explore whether neural networks possess nice asymptotic properties. As if they do, it may be possible to conduct statistical inference based on neural networks. Throughout this paper, we will focus on the asymptotic properties of neural networks with one hidden layer.

CONTACT Qing Lu lucienq@ufl.edu
This article has been corrected with minorchanges. These changes do not impact the academic content ofthe article.

Supplemental data forthisarticle can beaccessed here. https://doi.org/10.1080/10485252.2023.2209218

One of the most important properties of neural networks is that they are universal approximants (Hornik et al. 1989), which means any continuous function on a compact support can be approximated arbitrarily well bya neuralnetwork with one hidden layer.So it seems natural to consider it as a nonparametric regression problem and approximate the underlying function class through a class of neural networks with one hidden layer. Many other series-basedestimators such as splines and wavelets also possess similar approximation properties and have been extensivelystudied in the literature. Chen (2007) provides a comprehensive review for those methods. For nonparametric regression problems, random design and fixed design are the two main frameworks. Many existing literature on neural networks focus on random design (e.g. Chen and White 1999; Gyorfi et al. 2002). On the other hand, general theories on nonparametric regression under fixed design have been well studied in van de Geer (2000). Therefore, it is still worthwhile to study neural networks under fixed design.

Specifically, consider the following nonparametric regression mode l:

$$y_i = f_0(x_i) + \epsilon_i,$$

where Et, . . . , En are i.i.d. random variables defined on a complete probability space (Q, A,IJD) with IE[E] = 0, Var[E] = a^2 and IE[IEl^{2}H] < oo for some A > 0; *xi* , ... ,*Xn* E *X* C *]Rd* are vectors of covariates with *X* being a compact set in *]Rd* and *Jo* is an unknown function needed to be estimated. Weassume that/ $_0$ E *F,* whereFis the classofcontinuous functions with compact supports. Clearly,f$_0$ minimises the population criterion function

$$Q_n(f) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f_0(x_i))^2 + \sigma^2.$$

A least squares estimator of the regression function can be obtained by minimising the empirical squared error loss Q n(f):

$$\hat{f}_n = \mathrm{argmin}_{f\in\mathcal{F}}\mathbb{Q}_n(f) = \mathrm{argmin}_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2.$$

However, if the class of functions *F* is too rich, the resulting least squares estimator may have undesired properties, such as inconsistency (Shen and Wong 1994; Shen 1997; van de Geer 2000). Instead, we can optimise the squared error loss over some less complex function space *Fn,* which is an approximation of *F* while the approximation error tends to Oas the sample size increases. In the language of Grenander (1981), such a sequence of function classes is known as a sieve. More precisely, we consider a sequence of function classes,

Fi F2 · · · *Fn Fn+l* · · · *F,*

approximating *Fin* the sense that LJ $_1$ *Fn* is dense in *F.* In other words, for eachf E *F,* thereexists *rcnf* E *Fn* such that *d(f ,rcnf)* -+ 0 as *n* -+ oo, where d-(, •) is some pseudo-metric defined on *F.* With some abuse of notation, an approximate sieve estimator *fn* is defined to be

$$Qnlfn) :'.:= \inf_{jE:Fn} Qn<J) + Op(T/n), \tag{1}$$

where *TJn* - 0 as *n* - oo. We refer interested reader to Chen (2007) for a thorough discussion on sieve extremum estimators.

Throughout the rest of the paper, we focus on the sieve of neural networks with one hidden layer and sigmoid activation function. Specifically, we let

$$\mathcal{F}_{r_n} = \Bigg\{ \alpha_0 + \sum_{j=1}^{r_n} \alpha_j \sigma\left(\gamma_j^T x + \gamma_{0,j}\right) : \gamma_j \in \mathbb{R}^d, \alpha_j, \gamma_{0,j} \in \mathbb{R},$$

$$\sum_{j=0}^{r_n} |\alpha_j| \le V_n \text{ for some } V_n > 4 \text{ and } \max_{1\le j\le r_n} \sum_{i=0}^{d} |\gamma_{i,j}| \le M_n \text{ for some } M_n > 0 \Bigg\}, \tag{2}$$

where *rn, Vn, Mn* t oo as *n* - oo. For theoretical simplification, we impose boundedness assumption on the weights of the neural networks in *F,.,* which is related to the £1-regularisation when fitting a neural network. Such a method hasbeen discussed in previous literatures (e.g. White 1989, 1990). In those papers, the consistency of the neural network sieve estimators has been established under random designs. However, there are few results on the asymptotic distribution of the neural network sieve estimators, which will be established in this paper. In terms of the rate of convergence, Chen and Shen (1998) obtained rate of convergence Op((ll; 10) 4 <i+i(2dJ>) for neural network sieve estimators. Later on, Chen and White (1999) improved the convergence rate to *O p((10 :n-)* 4 l+2a/(d+l) Cl+a / (d+ in), where *a* relates to the smoothness of the true function *Jo* and in their paper, a central limit theorem for smooth functional of the estimated function is also provided. In Chen et al. (2001), rate of convergence was also obtained for stationary ,B- m ixing data. One important characteristic of the aformentioned rate of convergence is that as *d* - oo, both rates become Op((1 n -) l f 4)_ Similar results can also be found in Barron (1994). It is well known from Stone (1982) that local smoothing methods suffer from the curse of dimensionality. But such phenomenon seems to vanish for approximate sieve extremum estimators based on neural networks. Bauer and Kohler (2019) also developed general theory and conditions to justify that neural networks can be used to circumvent the issue of curse of dimensionality. Hornik et al. (1989) showed that Un *Frn* is dense in Funder thesup-norm.But when considering the asymptotic properties of the sieve estimators, we use the pseudo-norm 11/1 = n 1 *I:=7 1 f- (x ;)* (see Proposition 2.1 in the supplementary material) defined on *F* and *F,•.*

With the increasing popularity of deep learning, recent research also starts to focus on the statistical propert ies of deep neural networks. For example, Schmidt -Hieber (2020) provided the rate ofconvergence for sparse deep neural network with Rectified Linear Unit (ReLU) activation function under the assumption that the underlying function is a composition of functions in some Holder space. Farrell et al. (2020) and Farrell et al. (2021) established the rate of convergence for deep neural network estimators when the underlying function belongs to a unit ball in a Sobolev space. It is worth pointing out that in these two papers, no restrictions on the boundedness of weights in neural networks are imposed. The rate of convergence for deep ReLU neural networks has also been developed in Fabozzi et al. (2019) and Kohler and Langer (2021). It is worth mentioning that the results in Fabozzi et al. (2019) were developed based on the original version of this manuscript on ArXiv. We believe that the ultimate goal of developing these theories is to perform statistical inferencebasedon neural networksfor real-world problems and the results discussed in this paper may provide a starting point for further developments. For instance, Chen and White (1999) developed asymptotic normality for neural network sieve extremum estimators. However,to apply their result,a calculation of a seriesof covariance is essential, which may be hard to accomplish in practice. Recently, Horel and Giesecke (2020) developed a significance test based on neural networks. However, the theories in that paper are difficult to apply and verify in practice. Using similar techniques to be discussed in this paper, Shen et al. (2021) developed a goodness-of-fit test based on neural networks.

The remaining paper is organised as follows. In Section 2, we discuss the existence of neural network sieve estimato rs. The weak consistency and rateof convergence of the neural network sieve estimators will be established in Sections 3 and 4, respectively. Section 5 focuses on the asymptotic distribution of the neural network sieve estimators. Simulation results are presented in Section 6.

Notations: Throughout the rest of this paper, bold font alphabetic letters and Greek letters are vectors. C(X) is the set of continuous functions defined on *X*. The symbol means 'bounded above up to a universal constant' and *an* ~ *bn* means -+ 1 as *n* -+ oo. For a pseudo-metric space *(T, d)*, N(E, *T, d)* is its covering number, which is the minimum number of E-ballsneeded to cover *T*. Its natural logarithm is the entropy number and is denoted by H(E, *T, d)*.

2. Existe nce

A natural question to ask is whether the sieve estimator based on neural networks exists. Before addressing this question, we first study some properties of *Fnr*• Proposition 2.1 shows that the sigmoid function is a Lipschitz function with Lipschitz constant *L* = 1/ 4.

Proposition 2.1: *A sigmoid function a* (z) = *fiZ'/* (1 + f?') *is a Lipschizt function on* JR *with Lipschitz constant* 1/4.

The secondproposition provides an upper bound for the envelopefunction *sup* *1*E:F ,,. lf l.

Proposition 2.2: *For each fixed n,*

sup llf lloo :S *Vn.*
/ E:F,n

Now we quote a general result from White and Wooldridge (1991) for readers who are not familiarwith the theoriesof sieveextremum estimators. The theorem tells us that under some mild conditions, there exists a sieve approximate estimator and such an estimator is also measurable.

Theorem 2.1 (Theorem 2.2 in White and Wooldr idge (1991)): *Let* (Q, A, IP') *be a complete probability space and let* (8, *p) be a pseudo-metric space. Let* {8n} *be a sequence* *of compact subsets of* e. *Let* Qn: Q x *en-+* i *be A® B(en)/B(i.)-measurable, and suppose that for each w* E Q, *Qn(w,•) is lower semicontinuous on en, n*=1,2, *Then for each n*= 1, 2, . . ., *there exists 0n* : Q -+ *en, A/ B(en)-measurable such that for each w* E Q, *Qn(W,0n(w))* = infeEBn *Qn(w,0).*

Note that

$$\begin{aligned}\mathbb{Q}_n(f) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(f_0(x_i) + \epsilon_i - f(x_i))^2\\ &= \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f_0(x_i))^2 - 2\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(x_i) - f_0(x_i)) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2.\end{aligned}$$

Since the randomness onlycomes from E;'s, it is clear that *Qn* isa measurable function and for a fixed *w, Qn* is continuous inf Therefore, to show the existence of the sieve estimator, it suffices to show that *Fr"* is compact in C(X), which is proved in the following lemma.

Lemma 2.1: *Let X be a compact subset of d. Then for each fixed n, Frn is a compact set.*

Proof: For each fixed *n,* let *Bn* = *[ao,. . . ,a ,",Yo,1, . . . , yo,rn, Y [, . . . ,y]T* belong to *[- V n, Vnl'"+'* x [- Mn, Mn l' " (d+I) := e n. If *n* is fixed, *en* is a bounded closed set and hence it is a compact set in rn(d+2)_+l Consider a map

$$H : (\Theta_n, \|\cdot\|_2) \to (\mathcal{F}_{r_n}, \|\cdot\|_n)$$

Bn i---+ *H(Bn)* = *ao* + *Lap* *Tn* *j= I* (rJ *x* + *Yo,j)*

Note that *F,"* = H(en)- Therefore, to show that *F,"* is a compact set, it suffices to show that *H* is a continuous map due to the compactness of *en.* Let *8 l,n,8*$_2$*,n* E *en,* then

IIH (B1,n) - H (B2,n)II

;;E [al" + t aJ' a(,J"' x,+vJ)')-al"-t aJ' a(,!"'x,+ yJJ') r

" ;;E [lai" - ąi ' l + t H"a(,)"' x,+rJ;') - aJ' a(,!""'" + ,m]

;t [lal!- al"I+t 1aj1'1 la(r?"x, +,J,?) -a (rJ"x, +,ml+

J)-a J lla (r\J >Tx; + C 1)]^2 0,]

-<;t [t la/^0 - a/"I+ · t,l(rJ^0 r/')' + lr,' - r,T|]'

-< [t la)1' - aJ"I± (1 V llxllo0 t||rJ!- r/'t + lrJ)- ,JJ'lr

2

:S (n (1 V !lxl1oo)) [rn(d + 1)]1191 ,n - 92,n!I .

H ence, for any E > 0, we choose 8 = E/ ((1 V llxi1oo)J rn(d + 1)),. When 1191,n - 92,n112 < 8, we have

!1H (91 , n) - H (92,n) lln < E,

which implies that H is a continuous map and hence $Fr,.$ is a compact set for each fixed n. ■

Asa corollary of Lemma 2.1 and Theorem 2.1, we can easily obtain the existence of sieve estimator.

Corollary 2.1: *Based on the notations above, for each n* = 1, 2, . . ., *there exists fn* : Q - *Fr,., A/B(Frn)-measurable such that Q nifn(w)) =* inf Efn *Q n(/)* .

3. Consistency

In this section, we are going to show the consistency of the neural network sieve estimator. The consistency result leans heavily on the following Uniform Law of Large Numbers.

Lemma 3.1: *Under the assumption of*

[rn(d + 2) + 1]V log(V n[rn(d + 2) + 1] = o(n), as n - 00,

we have

$$\sup_{f\in\mathcal{F}_{r_n}} |\mathbb{Q}_n(f) - Q_n(f)| \xrightarrow{p^*} 0, \quad \text{as } n \to \infty.$$

Proof: For any *8* > 0, we have

$$\mathbb{P}^*\left(\sup_{f\in\mathcal{F}_{r_n}} |\mathbb{Q}_n(f) - Q_n(f)| > \delta\right)$$

= JP'*(sup VEFrn !._ t Ef - *a* ² - 2!.. t E; *{f(x;) - fo(x;))* > *8)*
n i=l *n* i=l

JP' (1_ *n* 1'=1 E²;- *a*² > *8* 2 *)* + JP' C sup ETFrn 1_ 1=1 *E;(j(x;) - fo(x;))* *n* | > $\frac{\delta}{4}$)

:= (I) + *(II).*

It follows from the Weak Law of Large Numbers that (I) 0. Now, we are going to evaluate *(II).* By using the Markov's inequality, (II) 0 holds if

E* [sup *fE:Frn* *!._* *n* t =i 1 *E;(f(x;) - fo(x;))*] 0, as *n* oo.

Note that IE[E] = 0 and each f E *Fr"* has its corresponding parameterisation *On.* Since *9n* is in a compact set, there exists a sequence *Onk,* *On* as k oo with *On,k* E Q(n(d+Z)+1n *([- V n, VnY"+l* x [- Mn,MnY'['d +l)). Each On,k Corresponds to a function fic E Fr . Based on continuity, we have *fic(x) f(x)* for each *x* E *X.* From Example 2.3.4 in van der Vaart and Wellner (1996), we know that *Fr.* is P-measurable. Based on symmetrisation inequality, we have

E* [sup *fEFrn* t =i 1 *E;(f* **(x;)** - *fo(x;))*] 2E.,E [sup *fE:F,.* !._ **n** t =i I ;E; *{f* **(x;)** - *fo(x;))*] ,

where 1, . . . , *n* are i.i.d. Rademacher random variables independent of EI, . . . , *En-* Based on the Strong Law of Large Numbers, there exists N1 > 0, such that for all *n* 2'.'. N1,

1 - *n* *L* i= I Ef < *a* ² + 1, a.s.

For fixed EI,. . . *,En,* I:=?l ;E;(j(x;) - *fo(x;))* is a sub-Gaussian process indexed by *f* E *Frn•* Suppose that (8 , C, µ,) is the probability space on which 1, . . . , *n* are defined and let *Y(f,w)* = I::=? 1 ;(w)E;(j(x;) - *fo(x;))* with f E *Fr"* and *w* E 8. As we have shown above, we have *fie f* and by continuity, *Y(/k,w) Y(f,w)* for any *w* E 8. This shows that *{Y (f,w),f* E F,.} is a separable sub-Gaussian process. Hence Corollary 2.2.8 in van der Vaart and Wellner (1996) implies that there exists a universal constant *K* and for any *J;* E *Frn* with *n* 2'.'. N1,

$$\mathbb{E}_{\xi}\left[\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\epsilon_i(f(\boldsymbol{x}_i)) - f_0(\boldsymbol{x}_i))\right|\right]$$

$$\leq \mathbb{E}_{\xi}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\epsilon_i(f_n^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right|\right] + K\int_0^{\infty}$$ logN *(½rJ,Fr.,d)* - - - - - *n* d17

$$\leq \mathbb{E}_\xi \left[\left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \epsilon_i (f_n^*(x_i) - f_0(x_i)) \right| \right]$$

$$+ K \int_0^{2V_n}$$ logN C :'+ TJ,Frn• ||· ||oo) n dTJ,

where forf,g E Frn•

d(f,g) = (n E;(j(x;) - g(x;))^2) 1/2

sothat the last inequality follows by noting that sup$_1$<ffin, ||f||$_{oo}$ V$_n$ and

d(f,g) ||f- g||oo (n Ef)1/2

We then evaluate these two terms. For the first term, for n N1, by Cauchy- Schwarz inequality, wehave

Ei [t S,,,<J:(x;) - /O(x;))] ,;(t,/r(t lf:(x;) - /o(x;))r 2

Ja2 ± 1 sup lf;(x) - fo(x)|, a.s.
xEX

By choosing J; = :rcrJo and using the universal approximation theorem introduced by Hornik et al. (1989), we know that supExX lf;(x;) - fo(xi)I ---+ 0 as n ---+ oo. Therefore, for any { > 0, there exists N2 > 0, such that for all n N2,

sup [f;(x;) - fo(x;)I < .k
xEX a + 1

By choosing n N1 V N2, we get

$$\mathbb{E}_\xi \left[\left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \epsilon_i (f_n^*(x_i) - f_0(x_i)) \right| \right] < \zeta \quad \text{a.s.}$$

For the second term, we use the same bound from Theorem 14.5 in Anthony and Bartlett (2009) as we did in the proof of Lemma 2:

$$N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty\right) \leq \left(\frac{8\sqrt{\sigma^2+1}e[r_n(d+2)+1]\left(\frac{1}{4}V_n\right)^2}{\eta\left(\frac{1}{4}V_n - 1\right)}\right)^{r_n(d+2)+1}$$

$$:= \tilde{B}_{r_n,d,V_n} \eta^{-[r_n(d+2)+1]},$$

where *Br ,d,Vn* = *(2-JaT'+Te[r n(d +* 2) + *l]V / (Vn -* 4)) '(d + $\frac{2}{?}$ +l. Let

Brn,dYn = *log Brn,d,Vn - [rn(d* + 2) + 1]

= [*rn(d* + 2) + 1 l(1 og $\frac{2 \; a^2 + 1 \, e[rn(d + 2) + 1] V}{n - 4}$) — 1

< 2[r *(d* + 2) + 1] lo g $\frac{[rn(d + \frac{2}{?} + 1]V}{V_n - 4}$ ' for all $n \geq N_1 \vee N_3$,

wh ere N3 is chosen to satisfy *rn(d* + 2) + 1 ::: 2 a ^{2} [+ (d + 1)e] l/2 . The last inequality then follows b noting that V^2 - V + 4 > 0 for al V so that log [rn(d + 2) + 1] V > log vu--ri (n) = Y n n - n - *V,.- 4* a - *Vn- 4*

log {2 a ^{2} + 1). We also have

H($\frac{1}{2 \; a \; +1}$ *.JdT'+117,F,,.,*‖ · ‖oo) = *logBr ,d,V,.* + *[rn(d* + 2) + 1] log $\frac{1}{17}$

:S *Bn, ,d,V,.* + *[r n(d* + 2) + 1] $\frac{1}{17}$

$$\leq B_{r_n,d,V_n}\left(1+\frac{1}{\eta}\right),$$

an d hence for all *n* :::. N1 v N3,

1 $^2_0$ H 112($\frac{1}{2\,a + 1}$ 17,F .... ‖·‖oo) d17

< B 1/2 {l,v. (1 +) 1/2 d 17
- *r,.,d,Vn* o*l* 17

= *B!:., ,v,.* [1 1 (1 + ¾ /2 d17 + *l* 2v ,. (1 + ¾ /2 d17]

:S *B!:., ,v.* [2 fo ^{1} -17 1/ ^{2} d17 + 2 *(2V n* - 1)]

:S 4 r,B 1/2 *V*
-v *L. Tn,d,V,. n,*

which implies that

H C :2+ 1 17, *Frn>*‖ · ‖oo) / *n* d17 :S 8 $\sqrt{\frac{[r n(d + 2) + 1] \quad \log r. (d \; t \; \frac{2}{"} \; 1] V}{n}}$

~ 8 $\sqrt{\frac{[rn(d+2)+1]VJ log(Vn[rn(d+2)+1])}{n}}$

where the last part follows by noting that log $\frac{v}{4}$ ~ log*Vn*. Under the assumption given in the Lemma, there exists N4 > 0, such that for all *n* :::_ N 4 , we have

$$\sqrt{\frac{[rn(d + 2) + 1] \quad \log(Vn[rn(d + 2) + 1])}{n}} \leq \frac{\{}{8}$$

Therefore, by choosing *n* 2::. N1 V N2 V N3 V N4, we get

$$\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\epsilon_i(f(x_i)-f_0(x_i))\right|\right]<2\zeta \text{ a.s.,}$$

i.e. IE [supfEF,. I¼ L7 *1* ;E;(j(x;) - *Jo(x;))I]* ---+ 0 a.s .. Moreove r, b ased on what we have shown, for a sufficiently large *n,* we have

IE :f.. *t* *;E;(f (xi)* - Jo(x;))] S #+I ilnr,Jo - Jolioo

+ 4 2 KB $^{112}_{rn\,,d,Vn}$ -*n* ^{112}*vn* ---+ 0, as ..

Since

$$\begin{aligned}&\mathbb{E}_\epsilon[\sqrt{\sigma^2+1}\|\pi_{r_n}f_0-f_0\|_\infty+4\sqrt{2}KB^{1/2}_{r_n,d,V_n}n^{-1/2}V_n]\\&\quad=\sqrt{\sigma^2+1}\|\pi_{r_n}f_0-f_0\|_\infty+4\sqrt{2}KB^{1/2}_{r_n,d,V_n}n^{-1/2}V_n\to 0<\infty,\end{aligned}$$

by using the Generalised Dom inated Convergence Theorem, we know that

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(x_i)-f_0(x_i))\right|\right]\le 2\mathbb{E}_\epsilon\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\epsilon_i\left(f(x_i)-f_0(x_i)\right)\right|\right]\to 0,$$

which completes the proof. ■

Based on the above lemmas, we are ready to state the theorem on the consistency of neural network sieve estimato rs.

Theorem 3.1: *Under the notation given above, if*

[rn(d + 2) + *l] V log(Vn[rn(d* + 2) + 1] = *o(n),* as **n---+** 00, (3)

then

Ilfn - *Jolin* $\overset{p}{---+}$ 0 .

Proof: Since Q is continuous at *Jo* E *F* and *Q(fo)* = *a* ^{2} < oo, for any E > 0, we have

$$\inf_{f:\|f-f_0\|_n\ge\epsilon}Q_n(f)-Q_n(f_0)=\inf_{f:\|f-f_0\|_n\ge\epsilon}\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-f_0(x_i))^2\ge\epsilon^2>0.$$

Hence, based on Lem ma 1, Lemma 3 and Corollary 2.6 in White and Wooldridge {1991), we have

Ilfn - *Jolin* $\overset{p}{---+}$ 0. ■

Remark 3.1: Wediscuss the condition (3) in Theorem 3.1 viasomesimple examples here. If <xj = 0(1) forj = 1, . . . , rn, then Vn = O(rn) and

[rn(d + 2) + 1)V log(Vn[rn(d + 2) + 1)) = O(r logrn)-

T herefore , a possible growth rate for the number of hidden units in a neural network is Tn = o((n/ log n)^{1}1^{3}). On the other hand, if we have a slow growth rate for the number of hidden units in the neural network, such as rn = log Vn, then we have

$$[r_n(d+2)+1]V_n^2\log(V_n[r_n(d+2)+1]) = \mathcal{O}((V_n\log V_n)^2).$$

Hence, a possible growth rate for the upper bound of the weights from the hidden layer to the output layer is Vn = o(n^{1} 1^{2} / log n).

4. Rate ofconvergence

To obtain the rateof convergence for neural network sieves, we applyTheorem 3.4.1 in van der Vaart and Wellner (1996).

Theorem 4.1: *Based on the above notations, if*

$$\eta_n = \mathcal{O}\left(\min\{\|\pi_{r_n}f_0 - f_0\|_n^2, r_n(d+2)\log(r_nV_n(d+2))/n, r_n(d+2)\log n/n\}\right),$$

then

llfn - folln

= Op (max llrr,,Jo - follm $\sqrt{\frac{rn(d + 2) \log[rnVn(d + 2))}{n}}$ $\sqrt{\frac{r_n(d+2)\log n}{n}}$ })).

Proof: Use the same bound from Theorem 14.5 in Anthony and Bartlett (2009), we have

$$\log N(\eta, \mathcal{F}_{r_n}, \|\cdot\|_n) \le \log N(\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty) \le \log\left(\frac{4e[r_n(d+2)+1]\left(\frac{1}{4}V_n\right)^2}{\eta\left(\frac{1}{4}V_n - 1\right)}\right)^{r_n(d+2)+1}$$
$$= [r_n(d+2)+1]\log\frac{\tilde{C}_{r_n,d,V_n}}{\eta},$$

where cη ,d Vn =e rn (d: 1 V > e. Then from Lemma 3.8 in Mendelson (2003), foro < 1,

$$\int_0^\delta \sqrt{\log N(\eta, \mathcal{F}_{r_n}, \|\cdot\|_n)}\,\mathrm{d}\eta \le [r_n(d+2)+1]^{1/2}\int_0^\delta \sqrt{\log\frac{\tilde{C}_{r_n,d,V_n}}{\eta}}\,\mathrm{d}\eta$$
$$\lesssim [r_n(d+2)+1]^{1/2}\delta\sqrt{\log\frac{\tilde{C}_{r_n,d,V_n}}{\delta}}$$
$$:= \phi_n(\delta).$$

Define h: *8* 1--+ *</Jn(o)/8°* = *[rn(d* + 2) + 1] 1/2 ... J 1ogt,"{v"*. Since for O < *8* < 1 and 1<a<2

h'(;) [,. (d+ 2) + I]>/> ((I - a)a- ... C,. d, v. 1 8² C,. d,v.I ... 1/2 C,. d, v ... log ... 8 ... 2 C,,d v. 1,² og ... 1,

[,. (d + 2)+ tJ' i' ((I - a)a- ... og c•d,v. 8 ... 2 og 112 C8,d,v).

< 0,

8 1--+ *</Jn(o)/8°* is decreasing on (0, oo). Let *Pn* ;S 1!11'r.fo - f oll;;-¹ . Note that

P</Jn (;n) = *Pn[rn(d* + 2) + 1]^112 log^112 *(PnCr,,,d,Vn)*

= *[rn(d* + 2) + 1]^112 *PnJlog Pn* + log*C,,,,d,v.*

and

$$\log \tilde{C}_{r_n,d,V_n} = 1 + \log \frac{[r_n(d+2)+1]V_n^2}{V_n - 4} \lesssim \log \frac{[r_n(d+2)+1]V_n^2}{V_n - 4}$$

~ log[rn*Vn(d* + 2)],

we have

p'fi</Jn (;n) ;S -v'n{} *rn(d* + 2)p'fi *(logpn* + log[rn*Vn(d* + 2)1) *;Sn.*

Therefore, for *Pn* mt'n

[(n / rn(d + 2) log[rnVn(d + 2)])^1/2 (n / rn(d + 2) logn)1/2)

we have *Pn² ¢ n(1/Pn)* ;S ..*jn.* Based on these observation, Lemma 1, Lemma 2 in the Supplementary Materials and Theorem 3.4.1 in van der Vaart and Wellner (1996) imply that

llfn - Jl',.f olln

= *Op* (max | l!Jl',.fo - *follm* *rn(d* + 2) *logn[rnVn(d* + 2)] *,;rn(d* +n2) *logn*)).

By using the triangle inequality, we can further get

llfn - Jolin :'.:: llf n - rr,Jolin + lirr,Jo - Jolin

= *Op* max lirr,Jo - *Jolin,* ([*rn(d* + 2) log[rn*Vn(d* +2)) / n *rn(d* +:) *logn*)).

■

Remark 4.1: Recall that a sufficient condition to ensure consistency is *rn(d* + 2) *V* log[rn*Vn(d* + 2)] = *o(n).* Under such a condition, *rn(d* + 2) log[rn*Vn(d* + 2)] _:s *n,* the r ate of convergence can be simplified to

llfn - *Jolin* = *Op* max lin,,Jo - Jolin, *r nd* + *n*2) *lo g n*)) ·

If we assume *Jo* E *F* where *F* is the space of functions with finite first absolute moments of the Fourier magnitude distributions, i.e.

F = {t: *Rd* - R :*J (x)* = *J* exp *{iaT* X} *dμ,f(a),* IILJ li1

:= *j* max(1iall1, 1) diμ,J l(a) :Sc}, (4)

where */.Lf* is a complex measure on *Rd*. IILJ I denotes the total variation of */.Lf,* i.e. [μ, I(*A)* = sup Z:::, 1 *I μ,(An)* I and the supremum is taken over all measurable partitions {An} 1 of *A.* 1iali1 = I:f=1 la il for *a* = [a1,... ,*adjY* E *Rd.* Theorem 3 in Makovoz (1996) shows that *On* := llfo- *n,,Jolin* ;Sr-;; 112 -1/ (2 d) . Therefore, if we let *d* fixed and *Pn* = O;;-1 and *V n* = *V* in the proof of Theorem 4.1, *On* must also satisfy the following inequality:

P </>(_n) ;S *Pnr* 12 log 11 2 *(PnCmd,,v ,.)* ;S *,In*

=} *p rn log pn* + *p rn* log r n *;S n*

=} *rn* 1 + 13 *rnlogr n ;S n.*

One poss ible choice of *rn* to satisfy such condition is *rn* ;,c::: *(n /* log *n)r+a.* d In such a case, we obtain

llfn - *Joli n* = *Op* (*n*)- 4(1+1/(2d)) logn ,

which is the same rate obtained in Chen and Shen (1998). It is interesting to note that in the case where *d* = 1, we have llfn - *Joli n* = *O p((n /* log n-) 113) . Such rate is close to the *Op(-n 113),* which is the convergence rate in nonparametric least square problems when the class of functions considered has bounded variation in R (see Example 9.3.3 in van de Geer (2000)). As shown in Proposition 3 in the supplementary material, *F,.* is a class of functions with bounded variation in R. Therefore, the convergence rate we obtained makes sense.

5. Asymptotic normality

To establish the asymptotic normality of sieve estimator for neural networks, we follow the idea in Shen (1997) and start by calculatingthe Gateaux derivative of the empirical criterion function *Qn(f)* = n 1 I=:7 1 *(y; - f(x ;))²,*

$$\mathbb{Q}'_{n,f_0}[f-f_0]=\lim_{t\to 0}\frac{1}{t}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-f_0(x_i)-t(f(x_i)-f_0(x_i)))^2-\frac{1}{n}\sum_{i=1}^{n}(y_i-f_0(x_i))^2\right]$$

= -- 2/n n L *E;(j(x;) - fo(x;)).* i= 1

Then the remainder of first-order functional Taylor series expansion is

Rn[f - fol= *Qn(f) - Qnlfo) - "nlJo [f -* fol = 1/n n L, i=1 *(j (xi) - fo(x;))²* = llf - follñ-

As will be seen in the proof of asymptotic normality, the rate of convergence for the empirical process {-n 112 I=::7 1 *E;(j(x ;) - fo(x;))* :f E *Frn}* plays an important role. Here we establish a lemma, which will be used to find the desired rate of convergence.

Lemma 5.1: *Let* Xi , . . . ,*Xn be independent random variables with Xi~ Pi. Define the empirical process* {*Vn*(f)} *as*

$$v_n(f)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[f(X_i)-P_if].$$

Let Fn =if: llflloo .'.:: *Vn},* E > 0 *and a supfEFn*-n 1 I=::7 1 Var[f(X;)l *be arbitrary. Define* to *by H(to, F n,* II · 1 00) = ¼1/l(M , *n,a), where* 1/l(M, *n,a)* = M²/[2a(1 +)I . *If*

$$H(u,\mathcal{F}_n,\|\cdot\|_\infty)\le A_n u^{-r}, \tag{5}$$

for some O < r < 2 and u E (0, al, *where a is a small positive number, and there exists a positive constant Ki = K (r ,* E), *i = 1, 2 such that*

$$M\ge K_1A_n^{\frac{2}{r+2}}V_n^{\frac{2-r}{r+2}}n^{\frac{r-2}{2(r+2)}}\vee K_2A_n^{1/2}\alpha^{\frac{2-r}{4}},$$

we have

IP* (;sup lvn(f)l > M) S Sexp{- (1 - E)1/l(M, *n,a)}*.
EFn

Proof: The proof of the lemma is similar to the proof of Corollary 2.2 in Alexander (1984) and the proof of Lemma 1 in Shen and Wong (1994). Since *H(u,Fn,*II · lloo) S *A n-ur* for some O < *r* < 2, we have

I(s,t) := ∫ ∫ *H*^112^*(u, Fn,*|| · ||oo) du S 2(2 - r-) ^1^ *AJ* t^1^- f

Based on the assumption of

Ant ^- r^ 2: *H(to,J*^r^*"m* 11 · 11) = ^E^ *ifr(M, n,a),*

0 00 4

we have *to* S [';4;)l r/ . Note that ifr(M, *n,a)* 2: M ^2^/(4a) if MS *3,Jria/Vn* and 2(,Jria + *MVn/3)* S *4MVn/3* if *M* 2: *3,Jria/Vn* and hence *ifr(M,n,a)* 2: *3,JriM/(4Vn)*. In summary,

1/t(M,n,a) 2:{ M 2/(4a) if *M* < *3,Jria/Vn,*
3,JriM/(4Vn) if *M* 2: *3,/na/Vn·*

Therefore, if *M* 2: *3,/na/Vn,*

2B - 3/2J (EM / 64,/n' 0 *t)* < 29E - 3/2(2 - r) - 1Al/2- 5:
n 0

$$
\leq 2^9\epsilon^{-3/2}(2-r)^{-1}\left(\frac{4}{\epsilon}\right)^{\frac{1}{r}-\frac{1}{2}} A_n^{1/r}\left(\frac{3}{4V_n}\sqrt{n}M\right)^{\frac{1}{2}-\frac{1}{r}}
$$

$$
= \tilde{K}_1 A_n^{1/r} V_n^{\frac{1}{r}-\frac{1}{2}} n^{\frac{1}{4}-\frac{1}{2r}} M^{\frac{1}{2}-\frac{1}{r}},
$$

where $\tilde{K}_1 = 2^9\epsilon^{-3/2}(2-r)^{-1}(\frac{4}{\epsilon})^{\frac{1}{r}-\frac{1}{2}}(\frac{3}{4})^{\frac{1}{2}-\frac{1}{r}}$. Hence

$$
2^8\epsilon^{-3/2} I\left(\frac{\epsilon M}{64\sqrt{n}}, t_0\right) < M \Leftrightarrow \tilde{K}_1 A_n^{1/r} V_n^{\frac{1}{r}-\frac{1}{2}} n^{\frac{1}{4}-\frac{1}{2r}} M^{\frac{1}{2}-\frac{1}{r}} < M
$$

$$
\Leftrightarrow \tilde{K}_1 A_n^{1/r} V_n^{\frac{1}{r}-\frac{1}{2}} n^{\frac{r-2}{4r}} < M^{\frac{1}{r}+\frac{1}{2}}
$$

$$
\Leftrightarrow M > K_1 A_n^{\frac{2}{r+2}} V_n^{\frac{2-r}{r+2}} n^{\frac{r-2}{2(r+2)}},
$$

where K1 = *k*^2r^*F*. On the other hand, if *M* < *3,/na/Vn,*

2 8E - 3/2r (EM / 64,/n' 0 *t)* < 2 9E - 3/2(2- r) - lA l/21 f
n 0

$$
\leq 2^9\epsilon^{-3/2}(2-r)^{-1}\left(\frac{4}{\epsilon}\right)^{\frac{1}{r}-\frac{1}{2}} A_n^{1/r}\left(\frac{M^2}{4\alpha}\right)^{\frac{1}{2}-\frac{1}{r}}
$$

$$
= \tilde{K}_2 A_n^{1/r} M^{1-\frac{2}{r}} \alpha^{\frac{1}{r}-\frac{1}{2}},
$$

where $\tilde{K}_2 = 2^9\epsilon^{-3/2}(2-r)^{-1}(\frac{4}{\epsilon})^{\frac{1}{r}-\frac{1}{2}}(\frac{1}{4})^{\frac{1}{2}-\frac{1}{r}}$. Hence

$$
2^8\epsilon^{-3/2} I\left(\frac{\epsilon M}{64\sqrt{n}}, t_0\right) < M \Leftrightarrow \tilde{K}_2 A_n^{1/r} M^{1-\frac{2}{r}} \alpha^{\frac{1}{r}-\frac{1}{2}} < M
$$

$$
\Leftrightarrow \tilde{K}_2 A_n^{1/r} \alpha^{\frac{2-r}{2r}} < M^{\frac{2}{r}}
$$

$$
\Leftrightarrow M > K_2 A_n^{1/2} \alpha^{\frac{2-r}{4}},
$$

where K $=Kl,^{r/2}$. In conclusion, if M 2: K A + $v;$+ •- 2 $n![r+2i$ V $K_2A^{1/2}$ 2-, aT, then 2^8E $\frac{3}{7}$ ^{2}J($\frac{2}{64}$fM$_{1''}$ t) < M By Theorem 2.1 in Alexander (1984), we have the desired reruk ■

As a Corollary to Lemm a 5.1, we can show that the supremum of the empirical process {n ^{112} I=:7 , $Ei(j(x;) - fo(xi)) :f$ E Fr.}converges to O in probabil ity.

Coro llary 5.1: *Let Pn satisfy Pnllfn -* fo lln = Op(1) *and Fr. be the class of neural network sieves as defined in* (2). *Then under the conditions*

(Cl) $rn(d$ + $2)Vn$ log[rn$Vn(d$ + 2)] = $o(n^{1}1^{4})$;
(C2) np -;;$^2/V$ = o{1),

we have

$$\sup_{\text{llf- fo lln:S.P;;}^1 JE:F,n} \text{,,/ii.}\overset{1 \ n}{\text{L,}} \_ \ E;(j - fo)(x;) \Big| = \text{op}\{1).$$

Proof To estab lish the desired result, we apply the truncation device.

$$\text{JP>* (} \underset{\text{llf-folln::'.Pn}^1 \text{JEJ",n}}{\text{su 1:\_}} \underset{\text{l= I}}{\_t} \ Ei(j - fo)(x;) \quad \mathbf{M)}$$

$$\leq \mathbb{P}^*\left(\sup_{\|f-f_0\|_n\leq\rho_n^{-1}, f\in\mathcal{F}_{r_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\mathbb{I}_{\{|\epsilon_i|\leq V_n\}}(f-f_0)(x_i)\right|\gtrsim M\right)$$

$$+ \text{JP>* (} \sup_{\text{llf- Jolln.9;;}^1 J\,E:F_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\mathbb{I}_{\{|\epsilon_i|>V_n\}}(f-f_0)(x_i)\right|\gtrsim M\right)$$

$$:= (\text{I}) + (I\,I)\,.$$

For (I), we can apply Lemma 6 directly. Note that IEll1f1 l:S.V.}if - fo)(x) I :S $Vn<Vn$ + llfolloo) ;S V since llfolloo < oo and for O < $T/$ < 1,

$$\text{log N (TJ,F r. , ll} \cdot \text{lloo) :S log}\left(\frac{4e\ [rn(d+2)+1]\ (\frac{1}{4}Vn)}{TJ\left(\frac{'}{4}Vn - 1\right)}\right)^{2)\ r\ ,,d\ +2)\ +1}$$

$$:S\ [rn(d + 2) + 1]\ (\log Cr.\ ,d,Vn + \mathbf{t}\text{-}1)$$

$$= Cr,,,d,Vn\,(1 + \mathbf{t})$$

$$:S\ 2Cr,,,d,Vn\,\frac{1}{T/},$$

w here $C_{rn,d\ ,V.,}$ = $\frac{e[rn(d+2)+1]V;}{Vn\ -\ 4}$ an d

$$C_{r_n,d,V_n} = [r_n(d+2)+1]\log \tilde{C}_{r_n,d,V_n} - [r_n(d+2)+1] \sim r_n(d+2)\log[r_n V_n(d+2)].$$

Therefore, Equation (5) in the main text is satisfied with $r = 1$ and $An = 2\mathrm{Cr}_{d\,v}$. Following from Lemma 6, for M $c2^{l}_{rn,\,,} v\ V^{13}_{n}$-$n$ 1 6 v C $^{11}_{rn,\,,\,n}v$ a| | 4, we have (I) t: s e; p{- (1 - E)i/r(M, n, a)} and hence

$$\sup_{\|f-f_0\|\le\rho_n^{-1}, f\in\mathcal{F}_{r_n}} \left|\frac{1}{\sqrt{n}}\epsilon_i \mathbb{I}_{\{|\epsilon_i|\le V_n\}}(f-f_0)(x_i)\right| = \mathcal{O}_p\left(\frac{C^{2/3}_{r_n,d,V_n}V_n^{2/3}}{n^{1/6}}\right).$$

From (Cl),

$$\frac{C^{2/3}_{r_n,d,V_n}V_n^{2/3}}{n^{1/6}} \sim \left(\frac{r_n(d+2)V_n\log[r_nV_n(d+2)]}{n^{1/4}}\right)^{2/3} = o_p(1).$$

For(II), by using the Cauchy- Schwarz inequality, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbb{I}_{\{|\epsilon_i|>V_n\}}(f-f_0)(x_i)\right| \le \left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\mathbb{I}_{\{|\epsilon_i|>V_n\}}\right)^{1/2}\|f-f_0\|_n.$$

Then it follows from the Markov inequality that

$$(II) \le \mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\mathbb{I}_{\{|\epsilon_i|>V_n\}}\right)^{1/2}\rho_n^{-1} \gtrsim Mn^{-1/2}\right) \lesssim M^{-2}n\rho_n^{-2}\mathbb{E}[\epsilon^2\mathbb{I}_{|\epsilon|>V_n}]$$
$$\lesssim M^{-2}n\rho_n^{-2}\frac{\mathbb{E}[|\epsilon|^{2+\lambda}]}{V_n^{\lambda}}.$$

Based on condition (C2), we have *(II)* - 0, and

$$\sup_{\|f-f_0\|_n\le\rho_n^{-1}, f\in\mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbb{I}_{\{|\epsilon_i|>V_n\}}(f-f_0)(x_i)\right| = o_p(1).$$

Combining the results we obtained above, we get

$$\sup_{\|f-f_0\|_n\le\rho_n^{-1}, f\in\mathcal{F}_{r_n}} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i(f-f_0)(x_i)\right| = o_p(1).$$

■

Remark 5.1 : Condition (C2) can be further simplifiedusing the resultsfrom Theorem 4.1. If

T/n =*0* (min{IIJrr,Jo - *foll ,rn(d* +2) log(rn*Vn(d* +*2))/n,rn(d* +2)logn/ n}) ,

then

$$\rho_n^{-1} \asymp \max\left\{\|\pi_{r_n}f_0 - f_0\|_n, \sqrt{r_n(d+2)\log[r_nV_n(d+2)]/n}, \sqrt{r_n(d+2)\log n/n}\right\}.$$

It followsfrom condition (Cl) that

$$\rho_n^{-1} \asymp \max\left\{\|\pi_{r_n}f_0 - f_0\|_n, \sqrt{r_n(d+2)\log n/n}\right\}.$$

For simplicity, we assume that *p;,*$^{\cdot 1}$ x *Jrn(d* +2)log*n/n,* which holds for functions having finite first absolute moments of the Fourier magnitude distributions as discussed at the end of Section 4.4. Then in this case,

$$n\rho_n^{-2}/V_n^{\lambda} \asymp r_n(d+2)\log n/V_n^{\lambda},$$

so that condition (C2) becomes *rn(d* + 2) *logn/V!-+* 0.

Now we are going to establish the asymptotic normality for neural network estimators. For/ E ff E *Frn* : llf - Jolin S *p;,*$^{\cdot 1}$}, we consider a local alternative

$$\tilde{f}_n(f) = (1-\delta_n)f + \delta_n(f_0 + \iota), \tag{6}$$

where O S *8n* = *TJ!* $^{1\,2}$ = o(-n $^{1}1^{2}$) is chosen such that *Pn8n* = o(l) and*t(x)* =1.

Theorem 5.1 (Asymptotic Normality): *Suppose that* O S *T/n* = *o (-n* ^{1} *) and conditions* (Cl) *and*(C2) *in Corollary* 5.1 *hold. We further assume that the following two conditions hold:*

(C3) $\sup_{f\in\mathcal{F}_{r_n}:\|f-f_0\|_n\le\rho_n^{-1}} \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n = \mathcal{O}_p(\rho_n\delta_n^2);$

(C4) $\sup_{f\in\mathcal{F}_{r_n}:\|f-f_0\|_n\le\rho_n^{-1}} \frac{1}{n}\sum_{i=1}^n \epsilon_i(\pi_{r_n}\tilde{f}_n(f)(x_i) - \tilde{f}_n(f)(x_i)) = \mathcal{O}_p(\delta_n^2),$

then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\hat{f}_n(x_i) - f_0(x_i)\right] \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

Proof The main ideaof the proof is to usethe functional Taylor series expansion for *Qn*(j) and then carefully bound each term in the expansion. For any *f* E ff E *Fr,,* : llf - *J0*11*n* S

Pn̄ 1},

$$\mathbb{Q}_n(f) = \mathbb{Q}_n(f_0) + \mathbb{Q}'_{n,f_0}[f - f_0] + R_n[f - f_0]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i(f(x_i) - f_0(x_i)) + \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f_0(x_i))^2. \qquad (7)$$

Note that

llfn(f) - Jolin= **11(1** *- on)}n* + *Onlfo* + *l) - Jolin*

= **11(1** *- On)ifn - Jo)+ Ontll n*

S (1 - *On)llfn - Joli n + On,*

and since *On= o(-n* $^{1}1^{2}$*),* we can know that with probability tending to 1 , lifn(f) - Jolin S *p;;!* Then replacing/ in (7) by*]n* and *n:,,Jn(f),* we get

<r:lnlfn)= *L* Ef - *L* *E;lfn(x;) -fo(x;))* + *llfn* - foll
n i= 1 *n* i= 1

$$\mathbb{Q}_n(\pi_{r_n}\tilde{f}_n(f)) = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i(\pi_{r_n}\tilde{f}_n(f)(x_i) - f_0(x_i)) + \|\pi_{r_n}\tilde{f}_n(f) - f_0\|_n^2.$$

Subtracting these two equations yields

<r:lnlfn) =*<r:ln(n:,,Jn(f))*+ *L* *Ei* (*n:,,Jn(f)(x;)* - *]n(x;))*
1= 1

+ *llfn - Joli n* - lirr,.fn(f) - Jolin-

Now note that

$$\|\pi_{r_n}\tilde{f}_n(f) - f_0\|_n^2 = \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f) + \tilde{f}_n(f) - f_0\|_n^2$$

$$= \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f) + (1 - \delta_n)(\hat{f}_n - f_0) + \delta_n\iota\|_n^2$$

S (l · *On)*^{2}*llfn - Jolin*+ 2(1 - *0n) lr yn - fo,Onl* + *on*

+ 2(1 - *On)lln:,,Jn(f)* - *]n(f)llnllfn* - *Jolin*

+ *2onlrrr,Jn(f)* - *fn(f)lln*+ *lln:,,Jn(f)*- *fn(f)ll*^{2}*n,*

where the last inequality follows from the Cauchy-Schwarzinequality.Since

E; (*n:,,Jn(f)(x;) - Jn(x;))* = *L* Ej(*n:,,Jnlf)(xi) - Jnlf)(xi)* +*Jn(f)Cxi) - Jn(x;))*
1= 1 1= 1

= *L* *Ei* (*n:,,Jn(f)(Xi) -]nlf)(Xi))*
1=1

- $\frac{2}{n}$ *OnL* *Ei{]n(x;) - fo(xi))* - $\frac{2}{n}$ *OnL* Ei , by the definition of*J**n,* we have

- *O p*(*8n*) S inf *Qn(J)* - *Qnlfn)* S *Qn(7rr,Jn(J))* - *Qnlfn)*
fE:F,,,

S (l · *8n)*² llfn - *Jolin* - llf n - *Jolin* + 2(1 - *8n)8n* lr yn - *fo,L*

+ 2(1 - 8n)llfn - *Jolinlin,Jnlf)* - *Jn(J)lin*

+ *28nli7rr,Jn(J)* - *Jn(J)lin* + li1rr,*Jn(J*) - *Jn(J)lin*

- LE;(*n,Jnlf)(x;)* - *Jnlf)(x;))* + *8n* E;*(! n(x ;)* - *Jo(x;))*

+ - *8n* LE;+ *Op(8)*

S 8nllfn - *Jolin* + 2(1 - *8n)8n* lr yn - *Jo,l*

+ 2(1 - 8n)llfn - *Jolinli1r,Jn(J)* - *]n(J)lin*

+ *28nll7r,,Jn(J)* - *Jn(J)lln* + *lin,,Jnlf)* - *Jnlf)lin*

- *t* E; (*1r,Jn(J)(x;)* - *]n(J)(x;))* + *8n* *t* E; *(!n(x;)* - *Jo(x;))*
z=l z=l

$$+ \frac{2}{n}\delta_n \sum_{i=1}^{n} \epsilon_i + \mathcal{O}_p(\delta_n^2), \tag{8}$$

where the last inequalityfollows bynoting that (1 - *8n)2* - 1 = *-28n* +*8* S *8* . From the condition (Cl), we can get

[rn(d + 2) + 1]V log[rn*Vn(d* + 2) + 1]

S *([rn(d* + 2) + 1]Vn log[rn*Vn(d* + 2) + 1])4 = *o(n).*

Combining with Theorem 2, we obtain that llf n - *Jo*11*n* = *Op*(1) and hence *8;*llfn - *Jo*11 = *op(8;).* From condition (C3), we have

2(1 - 8n) llfn - fo ll n lln ,J nlf) - Jn (J) lln .S 2 llfn - Jo lin lln ,J nlf) - Jn (J) lln

= *Op* (p;¹*Pn8*) = *O p(8).*

Similarly, since *Pn8n* = o(l), we have

*28n*lln „J n(J) - *Jn(J*)lln = *Op(8n* · *Pn8n)* = *Op(8n)*

lln „J n(J) - Jn lf) lln = *Op(Pn8n)* = o *p(8n).*

Based on condition (C4), we know that

$$\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\right) = \mathcal{O}_p(\delta_n^2),$$

and from Corollary 5.1, we also have

$$\frac{2}{n}\delta_n \sum_{i=1}^{n} \epsilon_i \left(\hat{f}_n(x_i) - f_0(x_i)\right) = o_p(\delta_n \cdot n^{-1/2}).$$

It follows from these observations that

$$-2(1-\delta_n)\left\langle \hat{f}_n - f_0, \delta_n \iota \right\rangle + \frac{2\delta_n}{n}\sum_{i=1}^{n} \epsilon_i \le \mathcal{O}_p(\delta_n^2) + o_p(\delta_n^2) + o_p(\delta_n \cdot n^{-1/2}),$$

which implies that

$$-(1-\delta_n)\left\langle \hat{f}_n - f_0, \iota \right\rangle + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \le \mathcal{O}_p(\delta_n) + o_p(n^{-1/2}) = o_p(n^{-1/2}).$$

By replacing t with $-t$, we can obtain the same result and hence

$$\begin{aligned} \left| \left\langle \hat{f}_n - f_0, \iota \right\rangle - \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \right| &\le \left| (1-\delta_n)\left\langle \hat{f}_n - f_0, \iota \right\rangle - \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \right| + \delta_n \left| \left\langle \hat{f}_n - f_0, \iota \right\rangle \right| \\ &\le o_p(n^{-1/2}) + \delta_n \|\hat{f}_n - f_0\|_n \\ &= o_p(n^{-1/2}). \end{aligned}$$

Therefore,

$$\left\langle \hat{f}_n - f_0, \iota \right\rangle = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i + o_p(n^{-1/2}),$$

and the desired result follows from the classical Central Limit Theorem. ■

Let us focus on the conditions given in the theorem. Note that if (C1) holds, we have

$$r_n(d+2)V_n^2 \log[r_n V_n(d+2)] \le [r_n(d+2)]^4 V_n^4 \left(\log[r_n V_n(d+2)]\right)^4 = o(n),$$

so it is a sufficient condition to ensure the consistency of the neural network sieve estimator. As in Remark 3.1, we consider some simple scenarios here. If $V_n = O(r_n)$, then $r_n(d+2)V_n \log[r_n V_n(d+2)] = O(\log r_n)$ so that a possible growth rate for r_n is $r_n = o(n^{1/8}/(\log n)^2)$. On the other hand, if $r_n = \log V_n$, then $r_n(d+2)V_n \log[r_n V_n(d+2)] = O(V_n(\log V_n)^2)$ and a possible growth rate for V_n is $V_n = o(n^{1/4}/(\log n)^2)$. Thus, in both cases, the growth rate required for the asymptotic normality of neural network sieve estimator is slower than the growth rate required for the consistency as given in Remark 3.1. One explanation is that due to the Universal Approximation Theorem, a neural network with one hidden layer can approximate a continuous function on compact support arbitrarily well if the number of hidden units is sufficiently large. Therefore, if the number of hidden units is too large, the neural network sieve estimator $\hat{f}_n$ may be very close to the best projector of the true function f_0 in $\mathcal{F}_{r_n}$ so that the error $\sum_{i=1}^{n}[\hat{f}_n(x_i) - f_0(x_i)]$ could be close to zero, resulting a small variation. By allowing slower growth rate of the number of hidden units can increasethe variations of I:=7 1 *[fn(x;) - fo(xi)]*, which makes the asymptotic normality more reasonable. On the otherhand,condition (C3) and condition (C4) are similar conditions as in Shen (1997), which are known for conditions on approximation error. These conditions indicate that the approximation rate of a single layer neural network cannot be too slow, otherwise it may require a huge number of samples to reach the desired approximation error. Therefore, the conditions in the theorem can be considered as a trade-off between bias and variance.

Theorem 5.1 can be used directly for hypothesis testing of neural network with one hidden layer if we know the variance of the random error a^2. In practice, this is rarely the case. To perform hypothesis testing when a^2 is unknown, it is natural to find a good estimator of a^2 and use a 'plug-in' test statistic. A natural estimator for a^2 is

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_n(x_i)\right)^2 = \mathbb{Q}_n\left(\hat{f}_n\right).$$

We then need to establish the asymptotic normality for the statistic *a-.Jn L=7 I [fn(x;) - fo(x;)]*.

Theorem 5.2 **(Asymptotic Normality for Plug-in Statistic):** *Suppose that Jo* E *C(X), where X* C *llld is a compact set and* O _:'S *T/n* = o(-n 1). *Then under the conditions as stated in Theorem* 5.1, *we have*

$$\frac{O-n\backslash}{} \sum^{n} [Jn(Xi) - fo(x;)] .\mathrm{i}\ N(0, 1).$$

Proof: Note that

$$a; = Qnlfn) = \sum_{t=1} (>,i - fn(X;)r = \sum_{t=1} (to(x;) + \mathrm{Ej} - fn(X;)r$$

$$= \frac{In}{n}\sum_{i=l} (!n(x;) - fo(xi))^2 - \frac{2n}{n}\sum_{i=l} E; (!n(x;) - fo(xi)) + \frac{I\ n}{n}\sum_{i=l} Et$$

$$= \frac{1}{;;} \sum_{t=1} E_i^2 - \frac{2}{;;} \sum_{t=I} E; \left(fn(x;) - fo(Xi)\right) + \mathrm{llfn} - Joli^2_n$$

Based on the rate of convergence of $\}n$ we obtained in Theorem 4.1 and condition (Cl), we know that

$$lvn - !o[= 0;\left(\max\left\{\mathrm{lln,,Jo - foll}, \frac{rn(d+:)log}{}\right\}\right).$$

Under (C3), llrr,,Jo - foll = $o(p\ o) = o(-n^{1\,1^2})$ and under (Cl), we have

$$\frac{cn(d+:)\ b\ g\ n)}{} .:'S\ o\left(n^{\frac{1\ 4}{L}} \frac{;og\ n}{}\right)$$

$$= o\ \frac{(\log n)}{n3/4} = o(n^{112}),$$

which implies that llfn - foll = op(-n 11^2). Moreover, by the same arguments as in the proof of Theorem 5.1, we can show that

2;; !--,Ei (fn(xi) - fo(x;)) -– op(n- $^{1/2}$).
l= 1

Therefore

$$\mathbb{Q}_n(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 + o_p(n^{-1/2}).$$

Based on the Weak Law of Large Numbers, we know that ¼ I:=7 $_1$ E/ = a ^{2} + op(l). Therefore ,

$$\hat{\sigma}_n^2 = \mathbb{Q}_n(\hat{f}_n) = \sigma^2 + o_p(1),$$

and it follows from the Slutsky's Theorem and Theorem 5.1in the main text, we obtain

$$\frac{1}{\hat{\sigma}_n\sqrt{n}}\sum_{i=1}^{n}\left[\hat{f}_n(x_i) - f_0(x_i)\right] = \frac{\sigma}{\hat{\sigma}_n}\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}\left[\hat{f}_n(x_i) - f_0(x_i)\right] \xrightarrow{d} \mathcal{N}(0,1). \qquad \blacksquare$$

6. Simulation studies

In this section, simulations were conducted to check the validity of the theoretical results obtained in the previous sections. The consistency of the neural network sieve estimators was examined under various simulation scenarios. Finally, we evaluated the asymptotic normality of the neural network sieve estimators . For illustration purpose, we only include the simulations where the dimension of the covariates is 1. More simulations for the multivariate cases are given in the supplementary materials.

6.1. Consistency for neural network sieve estimators

In this simulation, we are going to check the consistency result from Section 3 and the validity of the assumption made in Theorem 3.1. Based on our construction of the neural network sieve estimators, in each sieve space *Frn,* there is a constraint on the l1 norm for **a:** I: :o lad s *Vn.* So finding the nearly optimal function in *Fr,,* for *Qnlf)* is in fact a constrained optimisation problem. A classical way to conduct this optimisation is through introducing a Lagrange multiplier for each constraint. Nevertheless, it is usually hard to find an explicit connection between the Lagrange multiplier and the upper bound in the inequalityconstraint. Instead, we use the subgradient method as discussed in section 7 in Boyd and Mutapcic (2008). The basic idea is to update the parameter *ao, . . . ,ar,,* through

$$\alpha_i^{(k+1)} = \alpha_i^{(k)} - \delta_k g^{(k)}, \quad i = 0, \ldots, r_n,$$

whereok > 0 is a stepsize and ok is chosen to be 0.1/ log(e + *k)* throughout the simulation, which is known as a nonsummable diminishing step size rule. gCk is a subgradient of the objective or the constra int function $Lj\ O$ |a| · Vn a t a <k) . Mo r e specifically, we take

$$\begin{cases} \partial_{\alpha^{(k)}} \mathbb{Q}_n(f) & \text{if } \underset{j=0}{\overset{Tn}{L}} \text{lajl} \quad Vn \\ \partial_{\alpha^{(k)}} \sum_{j=0}^{r_n} |\alpha_j| & \text{if } \underset{j=0}{\overset{rn}{L}} \text{lajl} > Vn, \end{cases}$$

Table 1. Comparison of errors *llfn* - *fo*ll and theleast square *errorsQn(fn)* after20,000iterationsunder different sample sizes.

	Neural network		Sine		Piecewise continuous	
Sample sizes	llfn- foll	Q,,(fn)	llfn - foll	Qn(fn)	llfn- foll	Q n(fn)
50	3.33E- 2	0.519	6.04E- 2	0.513	6.20E- 1	1.124
100	2.79E- 2	0.552	3.04E- 2	0.587	3.20E- 1	0.920
200	6.0SE- 3	0.500	1.05E- 2	0.501	2.51E- 1	0.786
500	8.1 SE- 3	0.484	1.19E- 2	0.499	3.26E- 1	0.769
1000	3.02E- 3	0.475	1.54E- 2	0 .480	2.98E - 2	0.489
2000	2.88E- 3	0.500	9.72E- 3	0.506	1.69E- 2	0.515

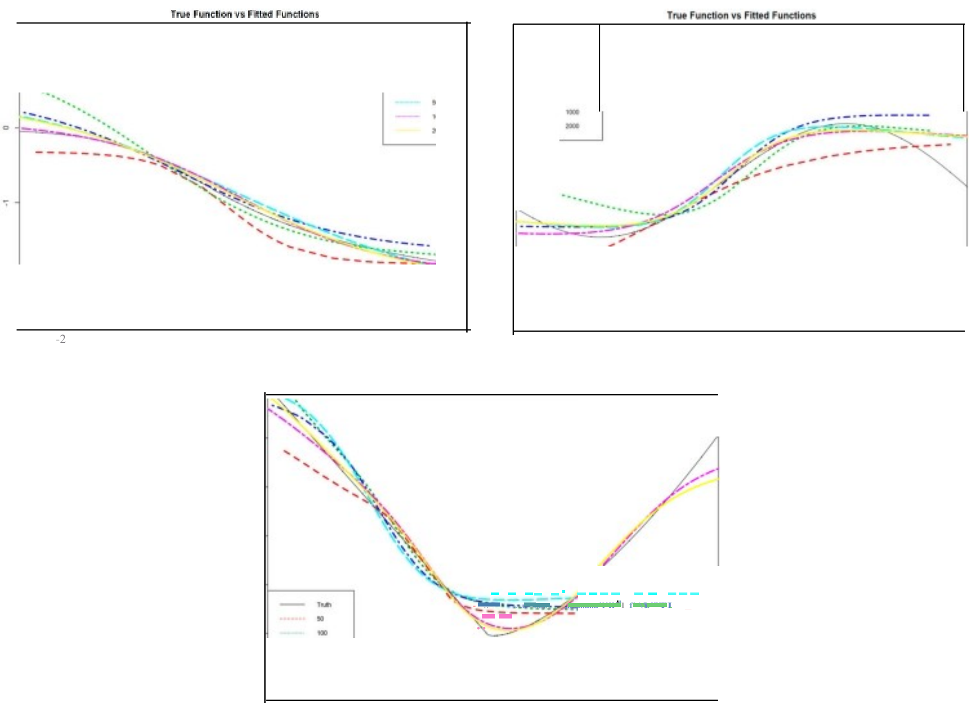

Figure 1. Comparison of the truefunction and thefitted functionfor threedifferent types of non-linear functions. The top panel showsthe scenariowhen the truefunction isa single layer neural network; the middle panel showsthe scenario when the true function isa sine function, and the bottom panel shows the scenario when the true function is a conti nuousfunction having a non-differentiable point. As we cansee fromall thecases, the fitt ed curve becomes closerto the truthasthesamplesize increases.

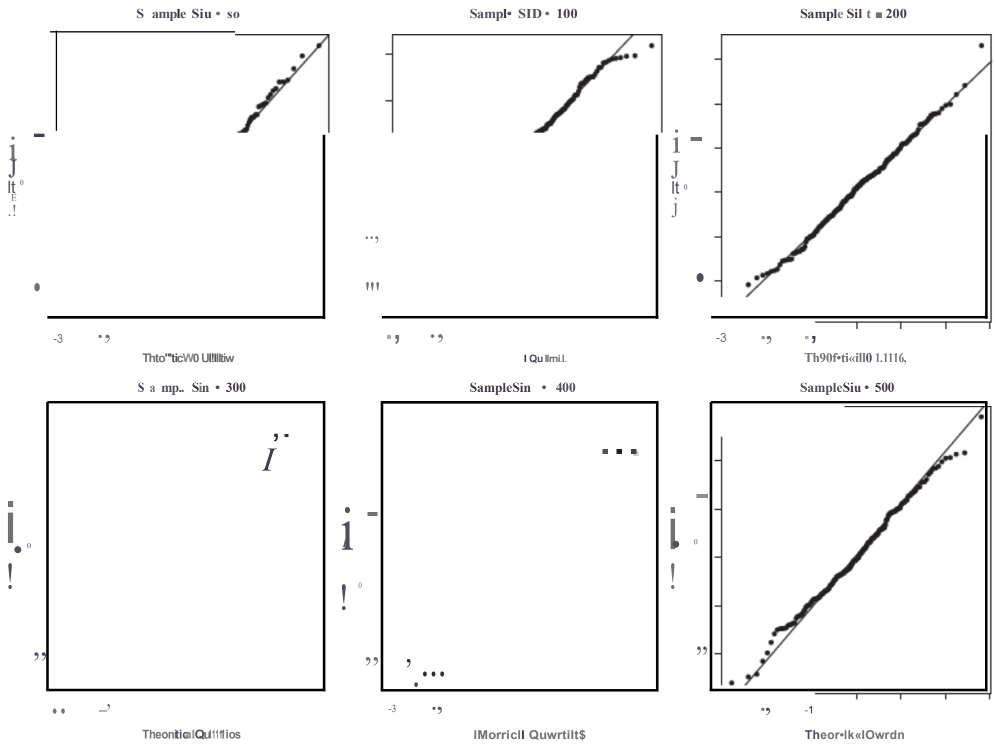

Figure 2. Normal Q-Q plot for $n^{-1/2}\sum_{i=1}$ [fn(x1) - fo(x1)I various sample sizes. Thetrue function fo is a single-layerneural network withtwohidden units asdefined in (10).

The updating equations of *y* 1, . . . *y* "', Yo,1, ... , *Yo,r,,* remain the same as those in the classical gradient descent algorithm.

We simulated the response through the following model:

$$y; = fo(x;) + E;, \qquad i = 1, \dots, n, \tag{9}$$

where the total sample size *n* varies from 50 to 2000, *Xt,* •.. *,Xn* ~ i.i.d. $N(0, 1)$, E1, • . . , *En* ~ i.i.d.N(0,0.7^2). For the truefunction*Jo(x),*weconsidered the following three functions:

(1) A neural network with a single hidden layer and two hidden units:

$$fo(x;) = -1 - a(2x; + 1) + a(- x; + 1).$$

(2) A trigonometric function:

$$fo(x) = \sin(;x) + i\cos(x+1) \tag{11}$$

(3) A continuous function havinga non-differential point

$$f_0(x) = \begin{cases} -2x & \text{if } x \,:\!.\!:\, 0 \\ \sqrt{x}\left(x - \frac{1}{4}\right) & \text{if } x > 0. \end{cases} \tag{12}$$

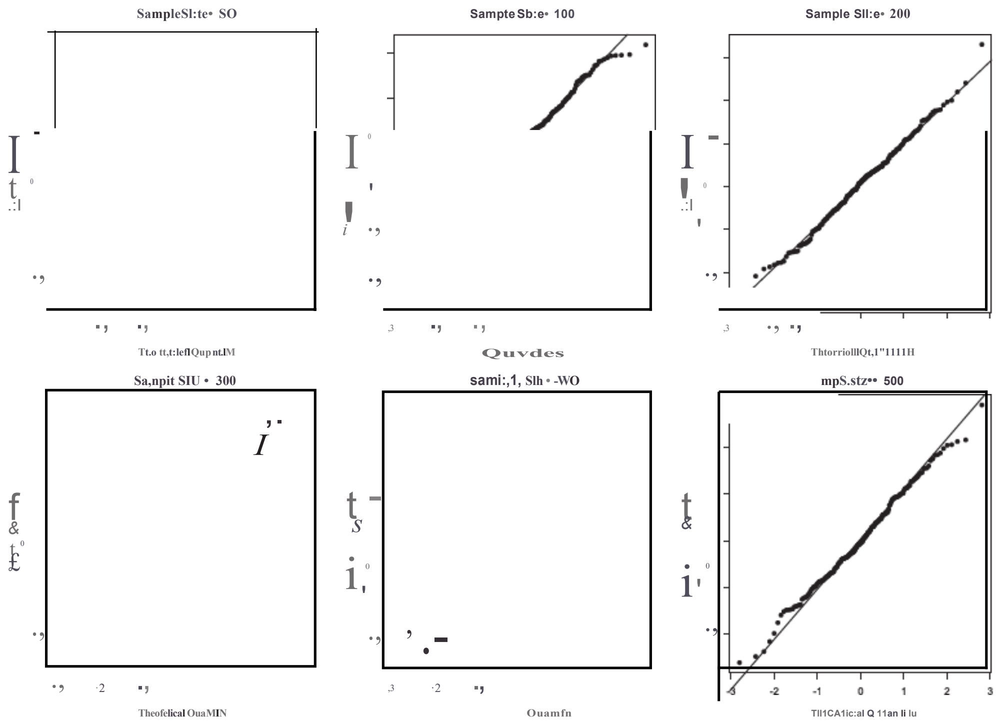

Figure 3. Normal Q- Q plot for-n $^{1}/^{2}$ I: $_{1}$*[fn(x1) - fo(x1)/* varioussamplesizes.Thetruefunctionfois a trigonometricfunction as defined in (11).

We then trained a neural network using the subgradient method mentioned above and set the number of iterations used for fitting as 20,000. We chose the growth rate on the numberof hidden units *rn* = $n^{1}1^{4}$ and the upper bound for.f.$_{1}$ norm of the weights and bias from the hidden layer to the output layer *Vn* = 10n^{1}1^{4}• Such choice satisfies the condition mentioned in Remark 3.1 and hence satisfies the condition in Theorem 3.1. We compared the errors llf n - fo ll and the least square errors *Qnifn)* under different sample sizes. The results are summarised in Table 1.

As we can see from Table 1, the errors llf n - foll overall has a decreasing pattern as the sample size increases. There are some cases where the error becomes a little bit larger when the sample sizes increases (e.g. the errors using 500 samples in all scenarios is larger than those errors using 200 sample). One explanation is that the number of hidden units increasesfrom 3 (for 200samples) to 4 (for 500 samples) underour simulation setup, which adds variation to the estimation performance. Overall, the table shows that the estimated function]n is indeed consistent in the sense that llfn - f oll n = *o;(l).* Figure 1 plots the fitted functions and the true function, from which we can straightforwardly visualise the result more and draw the conclusions.

6.2. Asymptotic normality forneural network sieve estimators

The last part of the simulation focuses on the asymptotic normalityderived in Theorem 5.1. Westillconsideredthesametypesoftruefunctionsasdescribed in Section 6.1 butsampled

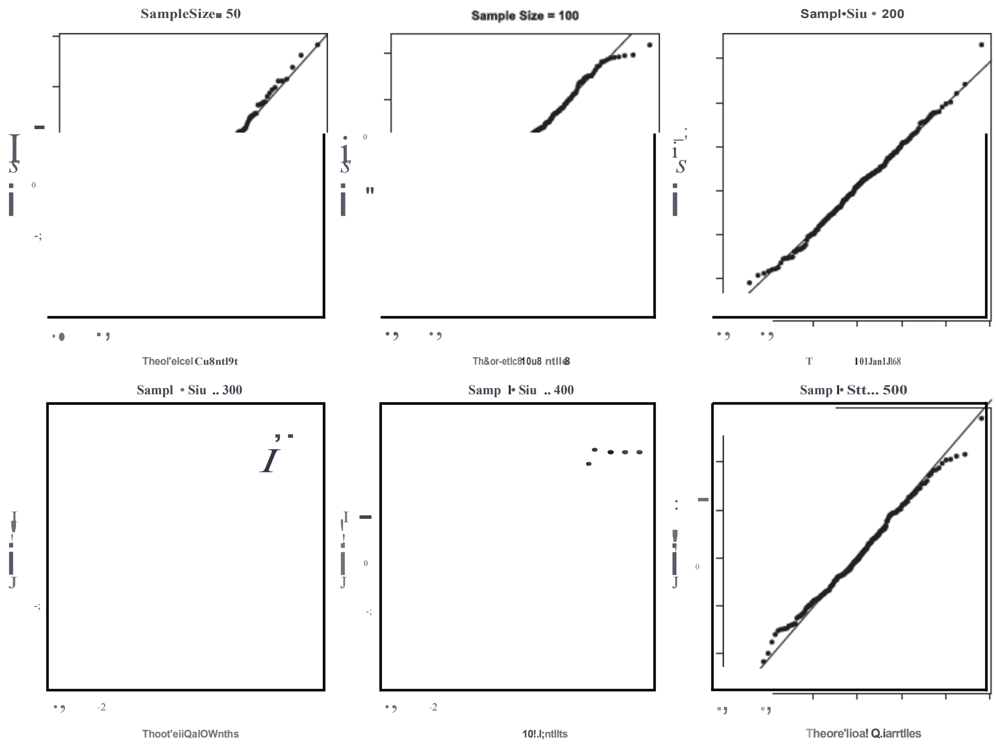

Figure 4. Normal Q- Q plot forn- $^{1}{}_{1}^{2}$ L =l [fn(x1) - fo(x1)l varioussamplesizes. Thetrue function fo is a continuous function having a non-differential point as definedin (12).

the random errors from the standard normal distribution. In this simulation, we still used the subgradient method to obtain the fitted model. The numberofiterationsused for fitting was set at 20,000. What isdifferent from Section 6.1 is the growth rates for *rn* and *Vn* set in this simulation. As mentioned in Section 5, the growth rates required for asymptotic normality are slower than those required for consistency. Therefore, we chose $r_n = n^{1}{}_{1}^{8}$ and *Vn* = 10n^{1}1^{10}. Such choice satisfies the condition (Cl) inTheorem 5.1. To get the normal Q- Q plot for-*n* ^{112} I=:7 $_1$ *lfn(Xi) - fo(xi)],* we repeated the simulation 200 times.

Figures 2 to 4 are the normal Q- Q plots under different nonlinear functions and various sample sizes. From the figures, we found that the statistic-*n* ^{112} I=:? *lfn(Xi) - fo(xi)]* fit the normal distribution pretty well under all simulation scenarios. It is also worth to note that the Q- Q plots looks similar under all simulation scenarios. This is what we would expect since the limiting distribution for the statistic-*n* $^{1}{}_{1}^{2}$ I=:? *lfn(Xi) - fo(xi)]* is $N(0, 1)$ under all scenarios. Another implication we can obtain from the Q-Q plots is that the statistic-*n* $^{1}{}_{1}{}^{2}$ *I: =? i lf n(Xi) - fo(xi)]* is robust to the choice *offo.* Therefore, as long as the true function *Jo* is continuous, *N(O,* 1) is a good asymptotic distribution for -*n* $^{1}{}_{1}^{2}$ I: ?= *[fn(xi) - Jo(xi)],* which facilitates hypothesis testing.

Besides the Q- Q plots, we also conducted the normality tests to check whether-*n* $^{1}{}_{1}{}^{2}$ I:=7 $_1$ *[fn(Xi) - fo(xi)]* follows the standard normal distribution. Specifically, we used the Shapiro- Wilks test and the Kolmogorov- Smim ov test to perform the normality test. Table 2 summarises the p-values for both normality tests. As we observed from Table 2, in

Table 2. Summary of results from the Shapiro- Wilks test and the Kolmogorov- Smirnov test. We use 'NN', 'TRI' and 'ND' to denote a neural network described in (10), a trigonometric function described in **(11)** and a continuous function having a non-differential point described in (12), respectively.

	Shapiro- Wilkstest			Kolmogorov- Smirnovtest		
Sample sizes	**NN**	TRI	ND	**NN**	TRI	ND
50	0.878	0.884	0.881	0.584	0.597	0.595
100	0.098	0.095	0.095	0.472	0.508	0.484
200	0.940	0.944	0.944	0.731	0.719	0.708
300	0.884	0.888	0.872	0.976	0.986	0.973
400	0.514	0.525	0.513	0.670	0.754	0.708
500	0.768	0.778	0.768	0.733	0.769	0.733

all cases, we failed to reject that-*n* $^1 I^2$ *I:*∓i ∦ *n(x;)* - J$_0$*(x ;)]* follows the standard normal distribution.

7. Discussion

We have investigated the asymptotic properties, including the consistency, rate of convergence and asymptotic normality for neural network sieve estimators with one hidden layer. While in practice, the number of hidden unites is often chosen ad hoc, it is important to note that the conditions in the theorems provide theoretical guidelines on choosing the number of hidden units for a neural network with one hidden layer to achieve the desired statistical properties. The validity of the conditions made in the theorems has also been checked through simulation results. Theorems 5.1and 5.2depend on the knowledge of the underlying function *Jo,* which is typically unknown in practice. Therefore, if we assume*Jo* has some certain form, the resultscan be applied and served as preliminary work for conducting hypothesis testing on *Ho* : *Jo* = *ho* for a fixed funct ion *ho*.On the other hand,since multiple functions can lead to the same valueof *n*- ^{1} ⊫:? $_1$ *_fo(x;),*thetest may not be power. The asymptotic normality results are crucial in developing more sophisticated significance test methods for neural networks (Shen et al. 2022).

The work conducted in this paper mainly focuses on sieve estimators based on neural networks with one hidden layer and standard sigmoid activation function. The work presented in this paper can be extended in several ways. The main theorems in this paper depend heavily on the covering number or the entropy number of the function class consisting of neural network with one hidden layer. Theorem 14.5 in Anthony and Bartlett (2009) provides a general upper bound for the covering number of a function class consisting of deep neural networks with Lipchitz continuous activation functions. Therefore, it is possible to extend our resultsdiscussed in this paper to a deep neuralnetwork with Lipchitz continuous activation functions. It is also worthwhile to investigate asymptotic properties of other commonly used deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

On the other hand, although homoscedasticity is assumed in the previous discussions, it can be relaxed to take heteroscedasticity into consideration. To see this, if lE[B] = ¢^{2} ^{2}(x;), then under the assumptions that l 0 l :S for some > 0and*L* Var[Ef l/ i^2 < oo, the proof of Lemma 3.1 can go through. The only modifications to be made are to use Kolmogorov Strong Law of Large Numbers to show (I) 0 and to change a^2 to ^{2}¢^{2}

later on. Therefore, the consistency result stillholds under heteroskedasticitywith the aformentioned two assumptions satisfied. Moreover, after a clear examination on the proof of Theorems 4.1, 5.1 and 5.2, it is easy to see that only the consistency part is involved with heteroskedasticity. Therefore, these results still hold under the aforementioned two assumptions.

When we train a deep neural network, we usually need to face an overfitting issue. In practice, regularisation is frequently used to reduce overfitting. Another natural extension of the work discussed in this paper is to modify the loss function by involving some regularisation terms. By taking regularisation into account, we believe the theories could have a much broader application in real-world scenarios.

Acknowledgments

The authors are grateful to the insights and suggestions from two referees to make this manuscript better.

Disclosure statement

No potential conflict of interest was reported by the author{s).

Funding

X. Shen, C. Jiang, and Q. Lu are supported by the National Institute on Drug Abuse (Award No. K01DA033346 and No. R01DA043501}. L. Sakhanenko is supported by NSF grant DMS-120823.8

ORCID

Qing Lu G http://orcid.org/0000-0002-7943-966X

References

Alexander, K.S. {1984), 'Probability Inequalities for Empirical Processes and a Law of the Iterated Logarithm', *The Annals of Probability,* 12(4), 1041- 1067.

Anthony, M., and Bartlett, P.L. {2009), *Neural Network Learning: Theoretical Foundations,* Cambridge University Press. https://www.stat.berkeley.edu/ bartlett/ nnl/index .html

Barron, A.R. {1994), 'Approximation and Estimation Bounds for Artificial Neural Networks', *Machine Learning,* 14(1), 115- 133.

Bauer, B., and Kohler, M.{2019), 'On Deep Learning As a Remedy for the Curse of Dimensionality in Nonparametric Regression', *The Annals of Statistics,* 47(4), 2261- 2285.

Boyd, S., and Mutapcic, A. (2008), 'Subgradient Methods (Notes for EE364B Winter 2006-07, Stanford University)'.

Chen, X. (2007), 'Large Sample Sieve Estimation of Semi-nonparametric Models', *Handbook of Econometrics,* 6, 5549- 5632.

Chen, X., Racine, J., and Swanson, N.R. (2001), 'Semiparametric Arx Neural-network Models with an Application to Forecasting Inflation', *IEEE Transactions on Neural Networks,* 12(4), 674- 683.

Chen, X., and Shen, X. (1998), 'Sieve Extremum Estimates for Weakly Dependent Data', *Econometrica,* 66(2), 289- 314.

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.

Devroye, L., Gyorfi, L., and Lugosi, G. (2013), *A Probabilistic Theory of Pattern Recognition,* Vol. 31. New York, NY: Springer Science & Business Media.

Fabozzi, F.J., Fallahgoul, H., Franstianto, V., and Loeper, G. (2019), 'Towards Explaining Deep Learning: Asymptotic Properties of relu ffn Sieve Estimators', Available at SSRN 3499324.

Farrell, **M.H .,** Liang, T., and Misra, S. (2020), 'Deep Learning for Individual Heterogeneity: an Automatic Inference Framework', Preprint arXiv:2010.14694

Farrell, M.H., Liang, T., and Misra, S. (2021), 'Deep Neural Networks for Estimation and Inference', *Econometrica,* 89(1), 181- 213.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016), *Deep Learning,* Vol. 1, MIT Press Cambridge. https://www.deeplearningbook.org/

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

Gyorfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002), *A Distribution-FreeTheory of Nonparametric Regression,* Vol. 1. New York, NY: Springer .

Hore!, E., and Giesecke., K. (2020), 'Significan ce Tests for Neural Networks', *Journal of Machine Learning Research,* 21(227), 1- 29.

Hornik,K.,Stinchcombe, M., and White, H. (1989),'Multilayer Feedforward Networks Are Universal Approximators',*Neural Networks,* 2(5), 359- 366.

Kohler, M., and Langer, S. (2021), 'On the Rate of Convergence of Fully Connected Deep Neural Network Regression Estimates', *The Annals of Statistics,* 49(4), 2231- 2249.

Makovoz, Y. (1996), 'Random Approximants and Neural Networks', *Journal of Approximation Theory,*85(1),98- 109.

Mendelson, S. (2003), 'A few noteson statistical learning theory', in *Advanced Lectures on Machine Learning,* Springer, pp. 1- 40. https://link.springer.com/chapter/10.1007/3-540-36434-X_1

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. (2021), 'A Goodness-of-fit Test Based on Neural Network Sieve Estimators', *Statistics & Probability Letters,* 174, 109100.

Shen,X.,Jiang,C.,Sakhanenko,L.,and Lu,Q. (2022),'A SieveQuasi-likelihood Ratio Test for Neural Networks with Applications to Genetic Association Studies', Preprint arXiv:2212.08255.

Shen, X., and Wong, W H. (1994), 'Convergence Rate of Sieve Estimates', *The Annals of Statistics,* 580 - 615. htt ps:// www.jstor .org/ stable/2242281

Stone, C.J.(1982), 'Optimal Global Rates of Convergence for Nonparametric Regression', *TheAnnals of Statistics,* 1040 - 1053. http s://www .jstor.o rg/stable/2240707

van de Geer, S. (2000), *Empirical Processes in M-estimation,* Vol. 6, Cambridge University Press. https://www.cambridge.org/ us/academic/ subjects/statistics-probability/statistical-theory-and -methods/empirical-processes-mestimation?format=PB&isbn=9780521123259

van der Vaart,A.W., and Wellner, J.A. (1996), *Weak Convergence and Empirical Processes.* New York: Springer.

Vapnik, V. (1998), *Statistical Learning Theory,*Vol. 3, New York: Wiley.

White, H. (1989), 'Learning in Artificial Neural Networks: a Statistical Perspective', *Neural Compu - tation,* 1(4), 425- 464.

White, H. (1990),'ConnectionistNonparametricRegression: MultilayerFeedforward NetworksCan Learn Arbitrary Mappings', *Neural Networks,* 3(5), 535- 549.

White, H., and Wooldridge, J. (1991), 'Some results on sieve estimation with dependent observations', in *Nonparametric and SemiparametricMethods in Economics,* ed. W Barnett, J. Powell, and G. Tauchen, New York: Cambridge University Press, pp. 459- 493.