1

On adaptive estimation by neural net type estimators

Sebastian Döhler and Ludger Rüschendorf*

Institute for Mathematical Stochastics, University of Freiburg, Eckerstr. 1, D-79104 Freiburg, Germany

28. June 2001

Abstract

We investigate convergence properties of neural net based estimators. We consider in particular sieved maximum likelihood estimators for the conditional log-hazard function in censoring models based on neural nets. A complexity regularized version of the estimator attains approximatively the same rate of convergence without knowledge of the degree of smoothness of the hazard functions. Our method gives a suboptimal convergence rate for sigmoidal networks. The minimax optimal convergence rate is attained approximatively for some other activation functions as e.g. for the threshold function and for piecewise polynomial functions.

Keywords: neural net estimators, censoring model, hazard function, maximum-likelihood estimator

1 Introduction

Kooperberg, Stone, and Truong (1995) proved that sieved maximum likelihood estimators based on tensor product splines attain the minimax rate in probability for estimation of the conditional log-hazard function in smoothness classes with known degree of smoothness. The same rate (up to a log-factor) was also derived for the MISE from a general approach to sieved estimation in Doehler (1999/2000) and in Doehler and Rueschendorf (2000) (denoted in the following by DR (2000)). In that paper also a

* Corresponding author. Tel.: +49-761-2035665
E-mail address: ruschen@stochastik.uni-freiburg.de

complexity regularized version of the estimator was introduced and proved to be approximatively adaptive. For theory and application of the method of functional estimation based on sieves we refer to Devroye, Györfi, and Lugosi (1996) and to Lugosi and Zeger (1995).

In this paper based on the approach in DR (2000) we investigate neural net based ML-estimators. We consider in detail the standard sigmoid neural net and obtain for the corresponding sieved ML-estimator a convergence rate w.r.t. MISE of the order $O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+4(k+1)}}\right)$ where p is the degree of smoothness and k is the dimension of the covariates. We also establish an adaptive complexity regularized version of the estimator. In comparison to the approximative minimax optimal rate $O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+k+1}}\right)$ of the tensor product spline estimator this rate is only suboptimal.

For some other activation functions like threshold functions and for piecewise polynomial functions the approximatively optimal rate is attained by the corresponding neural net estimators. In the final section we give an upper bound of the stochastic error for general net sieve estimators in terms of covering numbers only which do not need finite VC-dimension as typically assumed in the first sections. We also indicate that our estimation method applies in a similar way to further estimators like minimum contrast estimation and to other estimation problems like density estimation or regression estimation. For example for regression estimation the neural net minimum squared loss estimate of the regression function attains the approximatively optimal rate $E\|m_n - m\|_2^2 = O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+4d}}\right)$ where p is the degree of smoothness and d the dimension of the regressor.

We concentrate in this paper on the particular example of estimation of the conditional log-hazard function in a right censoring model as introduced in Kooperberg, Stone, and Truong (1995). Let T denote a survival (failure) time, $C : \Omega \to \mathcal{T} = [0,1]$ a censoring time, $X : \Omega \to \mathcal{X} = [0,1]^k$ a vector of covariables, $Y = T \wedge C$ the observable time and $\delta = 1_{(T \le C)}$ the censoring indicator. Assume the conditional density $f_0(t|x)$ of the distribution of T given $X = x$ to exist and let $\alpha_0(t|x) = \log \frac{f_0(t|x)}{\overline{F}_0(t|x)}$ denote the conditional log-hazard function, where $\overline{F}_0(t|x) = 1 - F_0(t|x)$ is the conditional survival function. We assume throughout that α_0 is bounded on $\mathcal{T} \times \mathcal{X}$ and that T and C are conditionally independent given X.

Let $(t_i, c_i, x_i), 1 \le i \le n$, be an *iid* sample of (T, C, X). The observed sample is (y_i, δ_i, x_i), $1 \le i \le n$, and the conditional log-likelihood functional is

$$L_n(\alpha) = \sum_{i=1}^{n} \left(\delta_i \alpha(y_i|x_i) - \int_0^{y_i} \exp \alpha(u|x_i) du \right) \tag{1.1}$$

Denote the expected conditional log-likelihood by

$$\Lambda(\alpha) = EL_1(\alpha) \tag{1.2}$$

Then $|\Lambda(\alpha) - \Lambda(\alpha_0)|$ defines a statistical distance of α to the true underlying α_0. It has the following representation

$$|\Lambda(\alpha) - \Lambda(\alpha_0)| = \Lambda(\alpha_0) - \Lambda(\alpha) = \int_{\mathcal{T}\times\mathcal{X}} \overline{F}_C G(\alpha - \alpha_0) dP^{(T,X)}, \tag{1.3}$$

where $\overline{F}_C$ is the conditional survival function of C given X and $G(y) = \exp(y) - (1+y)$. Based on a sieve $(\mathcal{F}_n)$ depending on the number of observations the maximum likelihood estimator $\widehat{\alpha}_n$ is defined as

$$\widehat{\alpha}_n = \underset{\alpha\in\mathcal{F}_n}{\arg\max}\, L_n(\alpha). \tag{1.4}$$

For some monotonically nondecreasing activation function $\sigma : \mathbb{R} \to [0,1]$ with $\lim\limits_{x\to\infty} \sigma(x) = 1$, $\lim\limits_{x\to-\infty} \sigma(x) = 0$, and $K \in \mathbb{N}$ let

$$\mathcal{F}(K) = \Big\{ f : [0,1]^{k+1} \to \mathbb{R}, f(z) = \sum_{i=1}^{K} c_i \sigma\left(a_i^T z + b_i\right), \tag{1.5}$$
$$a_i \in \mathbb{R}^{k+1}, b_i, c_i \in \mathbb{R}, 1 \le i \le K \Big\}$$

denote the neural net generated by σ. In particular $\sigma_s(x) = \frac{1}{1+\exp(-x)}$ denotes the standard sigmoid activation function and $\sigma_t = 1_{[0,\infty]}$ the threshold activation function. We consider neural net sieves of the form

$$\mathcal{F}_n = T_L \circ \mathcal{F}(K_n), \tag{1.6}$$

where T_L denotes the truncation at level L,

$$T_L(x) = \begin{cases} L & \text{if } x > L \\ x & \text{if } -L \le x \le L \\ -L & \text{if } x < -L. \end{cases} \tag{1.7}$$

L is an upper bound for the L^∞-norm of α_0 and $K_n \uparrow \infty$ a sequence of internal nodes to be chosen independent of the distribution.

2 Convergence rates for the neural net ML-estimator

For sieved ML-estimators a decomposition result for the estimation error was given in DR (2000). Let $M \ge \|\alpha_0\|_\infty$, $B_0 = \exp(\|\alpha_0\|_\infty) \cdot \exp(\exp\|\alpha_0\|_\infty)$, and consider a model $\mathcal{F} \subset \{\alpha : \mathcal{T}\times\mathcal{X} \to [-M,M]\}$. Then for the MISE of the ML-estimator (given by (1.4) with $\mathcal{F}_n = \mathcal{F}$)

holds:

$$
\begin{aligned}
E\,\|\widehat{\alpha}_n-\alpha_0\|_2^2 \quad \le \quad & 2\exp(2M)\inf_{\alpha\in\mathcal{F}}\|\alpha-\alpha_0\|_2^2 \qquad (2.1)\\
& + 32\kappa_0 B_0^2 M^2\exp(2M)\frac{\log \mathcal{C}_n(\mathcal{F})+1}{n},
\end{aligned}
$$

where $\kappa_0 = \frac{2608}{3}$ and where $\mathcal{C}_n(\mathcal{F})$ is a measure for the complexity of $\mathcal{F}$ given by

$$
\mathcal{C}_n(\mathcal{F}) = 6\sup_{\bar{z}\in Z^{2n}} N\left(\frac{1}{n},\mathcal{F},d_{L^1(\nu_{\bar{z}})}\right) N\left(\frac{1}{n},\mathcal{F},d_{L^1\left(\widetilde{\nu}_z\otimes U[0,1]\right)}\right) \qquad (2.2)
$$

where $Z = \mathcal{X}\times\mathcal{T}\times\{0,1\}$, $\nu_{\bar{z}} = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$, for $z=(z_1,\dots,z_n)$, $z_i = (x_i,y_i,\delta_i)$, $\widetilde{\nu}_z = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $N(\varepsilon,\mathcal{F},d)$ is the ε-covering number of $\mathcal{F}$ w.r.t. d and $U[0,1]$ is the uniform distribution on $[0,1]$. Similarly, for the Λ-error holds:

$$
\begin{aligned}
E\,|\Lambda(\widehat{\alpha}_n)-\Lambda(\alpha_0)| \quad \le \quad & 2\inf_{\alpha\in\mathcal{F}}|\Lambda(\alpha)-\Lambda(\alpha_0)| \qquad (2.3)\\
& + 8\kappa_0 B_0^2 M\exp(2M)\frac{\log \mathcal{C}_n(\mathcal{F})+1}{n}.
\end{aligned}
$$

For VC-classes $\mathcal{F}$ the estimate (2.1) simplifies to

$$
\begin{aligned}
E\|\widehat{\alpha}_n-\alpha_0\|_2^2 \quad \le \quad & C_1(M)\inf_{\alpha\in\mathcal{F}}\|\alpha-\alpha_0\|_2^2 \qquad (2.4)\\
& + C_2(M,B_0)\dim_{VC}(\mathcal{F})\frac{\log n}{n}\,;
\end{aligned}
$$

where $\dim_{VC}(\mathcal{F})$ is the Vapnik-Cervonenkis dimension of $\mathcal{F}$; a corresponding result holding true for the Λ-error. A related general upper bound was given in Birge and Massart (1998) where however a more complicated term including L^2- and L^∞-covering numbers is used.

Further recent related results are in Wong and Shen(1995) for density estimation, in Kohler (1997) for regression estimation based on minimal L^2-empirical risk in Yang and Barron (1998) for density estimation and in others. The estimates in (2.1), (2.3), (2.4) are extended to minimum contrast estimation for a general class of loss functions and also to a general class of estimation problems with a modified definition of the complexity $\mathcal{C}_n(\mathcal{F})$ (see DR (2000)).

We next consider in detail the case of standard sigmoid neural net estimators. Assume that

$$
\alpha_0\in W^q_{p,k+1},\ \|\alpha_0\|_\infty\le L \qquad (2.5)
$$

i.e. $\alpha\in W^q_{p,k+1}(L)$ the Sobolev-class of functions on $[0,1]^{k+1}$, whose p-th partial derivatives exist a.s. and are in $L^q([0,1]^{k+1})$ and α has L^∞-bound L.

Theorem 2.1 (sigmoid neural net estimator)
Consider the standard sigmoid neural net ML-estimator $\widehat{\alpha}_n$ of α_0,

$$\widehat{\alpha}_n = \arg\max_{\alpha \in \mathcal{F}_n} L_n(\alpha), \mathcal{F}_n = T_L \circ \mathcal{F}(K_n) \tag{2.6}$$

with $K_n = \left\lceil \left(\frac{n}{\log n}\right)^{\frac{1}{4+\frac{2p}{k+1}}} \right\rceil$. *Then for* $2 \leq q \leq \infty$ *and* $\alpha_0 \in W^q_{p,k+1}(L)$ *holds*

$$E\|\widehat{\alpha}_n - \alpha_0\|_2^2 = O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+4(k+1)}}\right). \tag{2.7}$$

If $\alpha_0 \in W^\infty_{p,k+1}$ *then*

$$E|\Lambda(\widehat{\alpha}_n) - \Lambda(\alpha_0)| = O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+4(k+1)}}\right). \tag{2.8}$$

Proof: For the proof we at first estimate the approximation error. Observe that $\|\alpha_0 - T_L\alpha\|_\infty = \|T_L\alpha_0 - T_L\alpha\|_\infty \leq \|\alpha_0 - \alpha\|_\infty$. Therefore, from the approximation result for neural nets in Mhaskar (Theorem 2.1, 1996) we obtain

$$\inf_{\alpha \in \mathcal{F}_n} \|\alpha_0 - \alpha\|_q^2 \leq C\left(\frac{1}{K_n}\right)^{\frac{2p}{k+1}} \|\alpha_0\|_{W^q_{p,k+1}}. \tag{2.9}$$

For the stochastic error we use the estimate of the VC-dimension of $\mathcal{F}(K)$, in Karpinski and Macintyre (1997) for the sigmoid net with $\sigma = \sigma_s$

$$\dim_{VC} \mathcal{F}(K) = O(K^4). \tag{2.10}$$

Therefore, we obtain

$$\dim_{VC} \mathcal{F}_n \;=\; \dim_{VC} T_L \circ \mathcal{F}(K_n) \;\leq\; \dim_{VC} \mathcal{F}(K_n) \;\leq\; CK_n^4. \tag{2.11}$$

The estimate in (2.1) resp. (2.4) then implies that

$$\begin{aligned} E\|\widehat{\alpha}_n - \alpha_0\|_2^2 &\leq C_1\left(\frac{1}{K_n}\right)^{\frac{2p}{k+1}} + C_2K_n^4\frac{\log n}{n} \\ &= O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+4(k+1)}}\right). \end{aligned} \tag{2.12}$$

For the proof of (2.8) we use (2.3) and establish for the approximation error

$$\inf_{\alpha \in \mathcal{F}_n} |\Lambda(\alpha) - \Lambda(\alpha_0)| \leq C\left(\frac{1}{K_n}\right)^{\frac{2p}{k+1}}. \tag{2.13}$$

From the representation of the approximation error in (1.3) and using some properties of the function G we obtain as in DR (2000)

$$\begin{aligned}
&\inf_{\alpha\in\mathcal{F}_n} |\Lambda(\alpha)-\Lambda(\alpha_0)| \\
&\quad \le \inf_{\alpha\in\mathcal{F}_n} G(\|\alpha-\alpha_0\|_\infty) \le G\left(\inf_{\alpha\in\mathcal{F}_n}\|\alpha-\alpha_0\|_\infty\right) \\
&\quad = O\left(\left(\inf_{\alpha\in\mathcal{F}_n}\|\alpha-\alpha_0\|_\infty\right)^2\right) = O\left(\left(\inf_{\alpha\in\mathcal{F}(K_n)}\|\alpha-\alpha_0\|_\infty\right)^2\right) \\
&\quad = O\left(\left(\tfrac{1}{K_n}\right)^{\frac{2p}{k+1}}\right) \quad cp.\ (2.9).
\end{aligned}$$

The stochastic error part then is dealt with as in the first part of the proof. □

3 Complexity regularized neural net estimator

For the construction of the neural net ML-estimator in section 2 knowledge of the degree of smoothness p and of L was supposed, i.e. the assumption that $\alpha \in W^q_{p,k+1}(L)$ resp. $W^q_{p,k+1}$ where $q \ge 2$. The aim of this section is to construct an adaptive estimator without knowledge of the smoothness class i.e. in the model $W = \cup\left\{W^q_{p,k+1}(L); 1 \le p < \infty, 0 < L, 2 \le q \le \infty\right\}$. A general construction of this type was given in DR (2000) based on the method of complexity regularization.

Consider for any $n \in \mathbb{N}$ a class of models $\{\mathcal{F}_{n,p}; p \in \mathcal{P}_n\} \subset \{\alpha = \mathcal{T} \times \mathcal{X} \to [-M, M]\}$ where $\mathcal{P}_n$ is a finite set, $M = M_n$, and assume that $\mathcal{C}_n(\mathcal{F}_{n,p}) < \infty$. Then define the complexity regularized estimator α_n^* in two steps:

1. For $n \in \mathbb{N}$, let $p_n^* \in \mathcal{P}_n$ be defined by

$$p_n^* = \arg\min_{p\in\mathcal{P}_n}\left(-\frac{1}{n}\sup_{\alpha\in\mathcal{F}_{n,p}} L_n(\alpha) + \mathrm{pen}_n(p)\right), \tag{3.1}$$

   where $\mathrm{pen}_n(p)$ is a penalization term with

$$\mathrm{pen}_n(p) \ge 4\kappa_0 B_0^2 M \exp(2M)\frac{\log \mathcal{C}_n(\mathcal{F}_{n,p})}{n}$$

   and $\mathcal{C}_n$ is given by (2.2).

2. $$\alpha_n^* = \arg\max_{\alpha\in\mathcal{F}_{n,p_n^*}} L_n(\alpha) \in \mathcal{F}_{n,p_n^*}. \tag{3.2}$$

In DR (2000) the following error estimates were proved for α_n^* in general sieved models

$$E\|\alpha_n^* - \alpha_0\|_2^2 \tag{3.3}$$

$$\leq 2 \inf_{p \in \mathcal{P}_n} \left(4M \operatorname{pen}_n(p) + \exp(2M) + \inf_{\alpha \in \mathcal{F}_{n,p}} \|\alpha - \alpha_0\|_2^2 \right) + \frac{16\kappa_0 B_0^2 M^2 \exp(2M)}{n}(1 + \log|\mathcal{P}_n|)$$

and similarly for the Λ-error

$$E|\Lambda(\alpha_n^*) - \Lambda(\alpha_0)| \leq 2 \inf_{p \in \mathcal{P}_n} \left(\operatorname{pen}_n(p) + \inf_{\alpha \in \mathcal{F}_{n,p}} |\Lambda(\alpha) - \Lambda(\alpha_0)| \right) + \frac{4\kappa_0 B_0^2 M \exp(2M)}{n}(1 + \log|\mathcal{P}_n|). \tag{3.4}$$

Some general theory of complexity regularized estimation theory has been developed in Barron, Birge, and Massart(1999). The upper bounds in their paper use some more involved L^2-L^∞-covering numbers in comparison to our bounds. The estimates (3.3), (3.4) hold true more generally for minimum contrast estimation in more general estimation problems; see the remarks in DR (2000). There are many related results on adaptive estimation for various types of estimation problems in the literature; see in particular Lee, Bartlett, and Williamson (1996), Wong and Shen(1995), Krzyzak and Linder (1998), Vapnik (1982), Barron and Cover (1991), and Lugosi and Zeger (1995). Typically empirical L^2-risk minimizing estimators are considered in these papers. For estimation of the conditional *log* hazard function let for $n \in \mathbb{N}$, $r_n = n$, $\beta_n = \frac{1}{5}\log\log n$, $\mathcal{P}_n = \{r \in \mathbb{N}; r \leq r_n\}$,and $\mathcal{F}_{n,r} = T_{\beta_n} \circ \mathcal{F}(K_{n,r})$, with

$$K_{n,r} = \left\lceil \left(\frac{n}{\log n} \right)^{\frac{1}{4 + \frac{2r}{k+1}}} \right\rceil. \tag{3.5}$$

Let α_n^* denote the complexity regularized neural net ML-estimator as defined in (3.1), (3.2) with penalization term $\operatorname{pen}_n(r) = \frac{(\log n)^{\frac{8}{5}}}{n} K_{n,r}^4$. Then we obtain

Theorem 3.1
If $\alpha_0 \in W_{p,k+1}^q(L)$, $p \geq 1$, $q \geq 2$, $L > 0$ then the complexity regularized sigmoid neural net ML-estimator α_n^* satisfies

$$E\|\alpha_n^* - \alpha_0\|_2^2 = O\left(\log n \left(\frac{\log n}{n} \right)^{\frac{2p}{2p+4(k+1)}} \right). \tag{3.6}$$

In the case $q = \infty$ we also obtain

$$E|\Lambda(\alpha_n^*) - \Lambda(\alpha_0)| = O\left(\log n \left(\frac{\log n}{n} \right)^{\frac{2p}{2p+4(k+1)}} \right). \tag{3.7}$$

Proof: To apply the general result in (3.3), (3.4) we first have to prove that

$$\mathrm{pen}_n(r) \geq 4\kappa_0 B_0^2 \beta_n \exp(2\beta_n) \frac{\log \mathcal{C}_n(\mathcal{F}_{n,r})}{n}. \tag{3.8}$$

From the bound $\dim_{VC} \mathcal{F}_{n,r} \leq CK_{n,r}^4$ (cp. (2.11)) we obtain as in the proof of Theorem 5.3 in DR (2000) the estimate $\log \mathcal{C}_n(\mathcal{F}_{n,r}) \leq 2CK_{n,r}^4 \log(n \log \beta_n)$ which implies (3.8). From the application of (3.3), (3.4) we then obtain

$$\begin{aligned} E\,\|\alpha_n^* - \alpha_0\|_2^2 \;\leq\; & 2 \inf_{r\in\mathcal{P}_n} \left(\tfrac{4}{5} \tfrac{\log\log n (\log n)^{\frac{8}{5}}}{n} \left(\left\lceil \left(\tfrac{n}{\log n} \right)^{\frac{1}{4+\frac{2r}{k+1}}} \right\rceil \right)^4 \right. \\ & \left. + (\log n)^{\frac{2}{5}} \inf_{\alpha\in\mathcal{F}_{n,r}} \|\alpha - \alpha_0\|_2^2 \right) + \tfrac{16\kappa_0 B_0^2}{25} \tfrac{(\log\log n)^2 (\log n)^{\frac{2}{5}}}{n} (1 + \log r_n) \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} E\,|\Lambda(\widetilde{\alpha}_n) - \Lambda(\alpha_0)| \;\leq\; & 2 \inf_{r\in\mathcal{P}_n} \left(\frac{(\log n)^{\frac{8}{5}}}{n} \left(\left\lceil \left(\frac{n}{\log n} \right)^{\frac{1}{4+\frac{2r}{k+1}}} \right\rceil \right)^4 \right. \\ & \left. + \inf_{\alpha\in\mathcal{F}_{n,r}} |\Lambda(\alpha) - \Lambda(\alpha_0)| \right) + \frac{16\kappa_0 B_0^2}{5} \frac{\log\log n (\log n)^{\frac{2}{5}}}{n} (1 + \log r_n). \end{aligned} \tag{3.10}$$

For $r = r_n = n \geq p$ and $\beta_n > B_0 = \|\alpha_0\|_\infty$ we obtain from the approximation result (2.9) for $q \geq 2$.

$$\begin{aligned} & 2 \inf_{r\in\mathcal{P}_n} \left(\tfrac{4}{5} \tfrac{\log\log n (\log n)^{\frac{8}{5}}}{n} \left(\left\lceil \left(\tfrac{n}{\log n} \right)^{\frac{1}{4+\frac{2r}{k+1}}} \right\rceil \right)^4 + (\log n)^{\frac{2}{5}} \inf_{\alpha\in\mathcal{F}_{n,r}} \left\|\alpha - \alpha_0\right\|_2^2 \right) \\ \leq\; & 2 \left(\tfrac{4}{5} \tfrac{\log\log n (\log n)^{\frac{8}{5}}}{n} \left(\left\lceil \left(\tfrac{n}{\log n} \right)^{\frac{1}{4+\frac{2p}{k+1}}} \right\rceil \right)^4 + (\log n)^{\frac{2}{5}} \inf_{\alpha\in\mathcal{F}_{n,p}} \|\alpha - \alpha_0\|_2^2 \right) \\ =\; & O\left(\tfrac{\log\log n (\log n)^{\frac{8}{5}}}{n} \left(\tfrac{n}{\log n} \right)^{\frac{4(k+1)}{2p+4(k+1)}} \right) + O\left((\log n)^{\frac{2}{5}} \left(\tfrac{\log n}{n} \right)^{\frac{2p}{2p+4(k+1)}} \right) \\ =\; & O\left(\log n \left(\tfrac{\log n}{n} \right)^{\frac{2p}{2p+4(k+1)}} \right). \end{aligned}$$

This gives an estimate for the first term in (3.9) and leads to the estimate (3.6) as in DR (2000). The proof of (3.7) is similar. □

4 Further neural net estimators, optimal rates and regression estimation

The convergence results in Theorems 2.1 and 3.1 yield a suboptimal estimation rate of the order $O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+4(k+1)}}\right)$ for the sigmoid neural net estimator in comparison to the minimax-optimal rate $O\left(\left(\frac{1}{n}\right)^{\frac{2p}{2p+k+1}}\right)$ attained by the product spline estimator in the related but smaller Hölder classes (see Kooperberg, Stone, and Truong (1995) resp. DR (2000) for the adaptive estimator).

From the proof of our result we find that for both sieve-types the approximation error as function of the number of basis functions is of the same order. The stochastic error however is for tensor product splines of linear order while it is for the class of sigmoid neural nets of the order K^4 which leads to the term $4(k+1)$ in the exponent in comparison to $k+1$ for splines (cp. (2.10)). As stated in the paper of Sontag (1998) the upper bound for the VC-dimension $\dim_{VC}(\mathcal{F}(K)) = O(K^4)$ for sigmoid neural nets possibly could be improved to some order $O(K^s)$ with $2 < s \leq 4$ but $O(K^2)$ is a strict lower bound and so the approximation order could be improved at most to some order of the form

$$O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+s(k+1)}}\right) \tag{4.1}$$

with $2 < s \leq 4$. So the optimal order is not attainable for sigmoidal neural by our method. The exact convergence rate of sigmoid net estimators in this problem is still unknown.

For some other activation functions with similar approximation rate and complexity growth rate as the tensor product splines the corresponding neural net estimator however can attain the approximative optimal minimax rates. For the threshold activation function $\sigma_t = 1_{[0,\infty]}$ it holds that the VC-dimension is approximatively linear, $dim_{VC}\mathcal{F}(K) = O(K \log K)$ (see Sontag(1998)). Furthermore, it is proved in Petrushev(1998) that the threshold neural nets have the approximation order $O((\frac{1}{K})^{\frac{2p}{k+1}})$. Combining these results we obtain from our upper estimates that the threshold neural net ML-estimator as well as the complexity regularized version attain the approximative optimal minimax rates.

Corollary 4.1 (Threshold neural net estimator)
Let $\widehat{\alpha}_n$ (resp. α_n^*) denote the neural net based (adaptive) ML-estimator with threshold function σ_t based on the sieve $\mathcal{F}_n$ as in sections 2, 3 with $K_n = \left\lceil \left(\frac{n}{\log n}\right)^{\frac{1}{4+\frac{2p}{k+1}}} \right\rceil$ $\left(\textit{resp. } K_{n,r} = \left\lceil \left(\frac{n}{\log n}\right)^{\frac{1}{4+\frac{2r}{k+1}}} \right\rceil\right)$. *Then $\widehat{\alpha}_n$ (resp.*

α_n^*) *attain the approximatively optimal minimax rates*

$$E\|\widehat{\alpha}_n - \alpha_0\|^2 = O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+k+1}}\right) \quad \textit{and} \tag{4.2}$$

$$E\|\alpha_n^* - \alpha_0\|^2 = O\left(\log n \left(\frac{\log n}{n}\right)^{\frac{2p}{2p+k+1}}\right). \tag{4.3}$$

Some general polynomial bounds on the VC-dimension of neural nets are given in Karpinski and Macintyre (1997). Approximatively linear bounds of the order $O(K \log K)$ are obtained in BartlettMaiorovMeir (1998) and Sakurai(1999) for piecewise polynomial networks with a fixed number L of layers. Also several optimal approximation rate results for neural net type functions have been established by Mhaskar (1996) (smooth nets), Petrushev (1998) (sigmoid nets and $p \geq \frac{d-1}{2}$), Delyon, Juditsky, and Benveniste (1995) (wavelet nets), Maiorov and Meir(2000) and Meir and Maiorov (2000) (neural nets in smoothness classes). For a recent survey see Pinkus (1999). As a result also further neural net type estimators, e.g. those with piecewise polynomial activation functions attain the approximative optimal minimax bounds.

Improved resp. approximatively optimal convergence rates for neural net type estimators can in some problems be obtained even without having nearly linear or even finite VC-dimension as in the previous examples. Consider general net sieves of the form

$$\mathcal{F}(\beta, K) = \left\{\alpha = \sum_{i=1}^{K} c_i f_i(t, x);\;\; f_i \in \mathcal{F}_0, \sum_{i=1}^{K} |c_i| \leq \beta\right\} \tag{4.4}$$

where $\mathcal{F}_0$ is a class of basis functions, $\mathcal{F}_0 = \{\Psi(a_\vartheta(x)); \vartheta \in \Theta\}$, parametrized by Θ and transformed by Ψ. Then the $L^1(\nu)$-covering number for any probability measure ν of $\mathcal{F}(\beta, K)$ can be estimated from above

$$N(\delta, \mathcal{F}(\beta, K), d_{L^1(\nu)}) \leq C^K (\beta K)^{K(2D-1)} \left(\frac{1}{\delta}\right)^{K(2D-1)} \tag{4.5}$$

where

$$N(\delta, \mathcal{F}_0, d_{L^2(\nu)}) = O\left(\left(\frac{1}{\delta}\right)^{2(D-1)}\right). \tag{4.6}$$

The proof of (4.5) is based on some well-known rules for covering numbers. The bound in (4.6) holds true by Pollard's estimate in particular if $dim_{VC}\mathcal{F}_0 = D$ but may be true even in cases with infinite VC-dimension. As consequence, one obtains an estimate for the complexity of $\mathcal{F}(\beta, K)$.

Proposition 4.2 *Assume that*

$$N(\delta, \mathcal{F}_0, d_{L^2(\nu)}) = O\left(\left(\frac{1}{\delta}\right)^{2(D-1)}\right) \tag{4.7}$$

then for some constant $C = C(D)$

$$\mathcal{C}_n(\mathcal{F}(\beta, K)) \leq C^K (\beta K)^{2K(2D-1)} n^{2K(2D-1)}. \tag{4.8}$$

Combining this bound for the complexity with an "optimal" approximation rate of α_0 by a sieve net $\mathcal{F}(\beta_n, K_n)$ one obtains as consequence of the upper bounds in (2.1), (2.3), (3.3), (3.4) convergence rates for the corresponding sieve type estimator.

As application of this method some improved convergence rates have been derived in DR (2000) for various net sieve estimators like neural net, radial basis function, and wavelet net estimators under the assumption that α_0 has an integral representation of the form

$$\alpha_0(t, x) = \int_\Theta \Psi(a_\vartheta(t, x)) d\nu(\vartheta) \tag{4.9}$$

for some measure ν of finite variation. In this context an interesting problem is whether the approximation results in Maiorov and Meir(2000) and Meir and Maiorov(2000) which are based on an integral representation of the form (4.9) for smoothness classes are still valid under the additional assumption that approximation is considered in the classes $\mathcal{F}(\beta_n, K_n)$ with slowly increasing bounds β_n for the coefficients (like $\beta_n = \log\log n$). Maiorov and Meir(2000) and Meir and Maiorov(2000) derive nearly optimal approximation results in the class $\mathcal{F}(\infty, K_n)$ without restrictions on the coefficients. If true this result would imply improved consistency results of neural net type estimators in general smoothness classes.

The convergence rates obtained in this paper for neural net estimators use only weak growth and approximation properties of the nets and are applicable also to further estimation problems like density estimation or regression problems. For the estimation in the regression problem e.g. it was proved in Kohler(1997) that the sieved minimum squared error regression estimate m_n w.r.t. a general VC-class $\mathcal{F}_n$ satisfies

$$E\|m_n - m\|_2^2 = O\left(\inf_{f \in \mathcal{F}_n} \|f - m\|_2^2 + \dim_{VC}(\mathcal{F}_n) \frac{\log n}{n}\right) \tag{4.10}$$

where m is the underlying regression function $m(x) = E(Y|X = x)$. As in the proof of Theorems 2.1 and 3.1 or of Corollary 4.2 in Kohler(1997) we obtain from a corresponding choice of the standard neural net sieve $\mathcal{F}_n$ the result

$$E\|m_n - m\|_2^2 = O\left(\left(\frac{\log n}{n}\right)^{\frac{2p}{2p+4d}}\right), \tag{4.11}$$

where d is the dimension of the regressor X and p is the degree of smoothness. A similar rate of convergence also holds true for the complexity regularized version of the estimator.

As in Corollary 4.1 approximatively optimal estimation rates are obtained for threshold nets.

References

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

[2] Barron, A.R., L. Birgé, and P. Massart (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields 113*, 301–413.

[3] Barron, A.R. and T. M. Cover (1991). A bound on the financial value of information. *IEEE Transactions on Information Theory 34*, 1097–1100.

[4] Bartlett, P.L., V. Maiorov, and R. Meir (1998). Almost linear VC dimension bounds for piecewise polynomial networks. *Neural computation 10*, 2159–2173.

[5] Birgé, L. and P. Massart (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli 4*, 329–375.

[6] Delyon, B., A. Juditsky, and A. Benveniste (1998). Accuracy analysis for wavelet approximations. *IEEE Transactions on Information Theory 6*, 332–348.

[7] Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

[8] Döhler, S. (1999). *Consistent hazard regression estimation by sieved maximum likelihood estimators.* Preprint, Universität Freiburg.

[9] Döhler, S. (2000). *Empirische Risiko-Minimierung bei zensierten Daten.* Dissertation, Universität Freiburg.

[10] Döhler, S. and L. Rüschendorf (2000). *Adaptive estimation of hazard functions.* Preprint, Universität Freiburg.

[11] Karpinski, M. and A. Macintyre (1997). Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *Journal of Computer and System Sciences 54*, 169–176.

[12] Kohler, M. (1997). *Nichtparametrische Regressionsschätzung mit Splines.* Dissertation, Universität Stuttgart.

[13] Kohler, M. (1999a). Nonparametric estimation of piecewise smooth regression functions. *Statistics & Probability Letters 43*, 49–55.

[14] Kohler, M. (1999b). Universally consistent regression function estimation using hierarchial B-splines. *Journal of Multivariate Analysis 68*, 138–164.

[15] Kooperberg, C., C.J. Stone, and Y.K. Truong (1995). The L_2 rate of convergence for hazard regression. *Scandinavian Journal of Statistics 22*, 143–157.

[16] Krzyzak, A. and T. Linder (1998). Radial basis function networks and computational regularization in function learning. *IEEE Transactions on Information Theory 9*, 247–256.

[17] Lee, W.S., P.L. Bartlett, and R. Williamson (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory 42*, 2118–2132.

[18] Lugosi, G. and K. Zeger (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory 41*, 677–687.

[19] Maiorov, V.E. and R. Meir (2000). On the near optimality of the stochastic approximation of smooth functions by neural networks. *Advances in Computational Mathematics 13*, 79–103.

[20] Makovoz, Y. (1996). Random approximants and neural networks. *Journal of Approximation Theory 85*, 98–109.

[21] Meir, R. and V. E. Maiorov (2000). On the optimality of neural networks. Preprint, Technion, Haifa.

[22] Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytical function. *Neural Computation 8*, 164–177.

[23] Modha, D. S. and E. Masry (1996). Rate of convergence in density estimation using neural networks. *Neural Computation 8*, 1107–1122.

[24] Petrushev, P.P. (1998). Approximation by ridge functions and neural networks. *SIAM Journal on Mathematical Analysis 20*, 155–189.

[25] Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica 8*, 143–195.

[26] Sakurai, A. (1999). Tight bonds for the VC-dimension of piecewise polynmial networks. *Advances in Neural Information Processing Systems 11*, 324–329.

[27] Shen, X. and W.H. Wong (1995). Convergence rate of sieve estimates. *Annals of Statistics 22*, 580–615.

[28] Sontag, E.D. (1998). *VC* dimension of neural networks. In C. Bishop (Ed.), *Neural Networks and Machine Learning*, pp. 69–95. Springer Berlin.

[29] Vapnik, V. (1982). *Estimation of dependencies based on empirical data.* Series in Statistics. Springer.

[30] Wong, W.H. and X. Shen (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLES. *Annals of Statistics 23*, 339–362.

[31] Yang, Y. and A. Barron (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory 44*, 95–116.