PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 22, Fasc. 2 (2002), pp. 355–379

ADAPTIVE ESTIMATION OF HAZARD FUNCTIONS

BY

SEBASTIAN DÖHLER AND LUDGER RÜSCHENDORF (FREIBURG)

Abstract. In this paper we obtain convergence rates for sieved maximum-likelihood estimators of the log-hazard function in a censoring model. We also establish convergence results for an adaptive version of the estimator based on the method of structural risk-minimization. Applications are discussed to tensor product spline estimators as well as to neural net and radial basis function sieves. We obtain simplified bounds in comparison to the known literature. This allows us to derive several new classes of estimators and to obtain improved estimation rates. Our results extend to a more general class of estimation problems and estimation methods (minimum contrast estimators).

Key words and phrases: Adaptive estimation, sieved maximum likelihood, neural nets, structural risk minimization, hazard functions.

1. INTRODUCTION

In this paper we establish convergence rates for sieved maximum-likelihood estimators (sieved ML-estimators) for the log-hazard function in a censoring model. We also establish an adaptive version of the estimator based on the method of structural risk minimization (complexity regularization) as introduced in Vapnik [24]. Our results are obtained for general sieves and then are applied to some special types of sieves like tensor product splines or neural nets. We also state extensions of these results to more general estimation procedures (minimum contrast estimators) and to other types of estimation problems like regression problems comparable to those considered in Birgé and Massart [5] or in Barron et al. [4]. For related results see also Krzyzak and Linder [16], Lugosi and Zeger [18], Wong and Shen [25], Yang and Barron [26], and Kohler [12], [13].

Sieved ML-estimators are defined in the general framework of empirical risk-minimization. The main tools for their analysis are from empirical process theory. The main part of the proof of convergence properties is to establish an exponential maximal inequality for the log-likelihood functional and to obtain

estimates for the covering numbers and Vapnik–Červonenkis dimension of the involved function classes. In comparison to a similar maximal inequality in Birgé and Massart [5] we avoid the somewhat complicated condition M2 on control of fluctuations in the L_∞-metric and replace it by some more handy growth condition on L^1-covering numbers. Our L^1-covering condition is related to the condition $M_{1,[\ ]}$ (L^1-metric with bracketing) in Barron et al. [4], which is used in that paper to deal with model selection in a general framework and applied to several examples (see Sections 4.1.5 and 4.1.6). In comparison our covering condition seems to be particularly simple and well suited for the examples considered in this paper. Our proof is based on an exponential maximal inequality in Lee et al. [17]. In several examples we obtain improved convergence rates in comparison to the literature and some of them are established for the first time in this paper.

In the case of tensor product splines we obtain up to a logarithmic factor the optimal convergence rate in the minimax sense in smoothness classes as derived in Kooperberg et al. [15], the only paper on convergence rates in this context so far. For general background on censoring models and reference to martingale based estimation methods we refer to Andersen et al. [1]. Related consistency results for kernel type estimators and further references on non-parametric functional estimation of hazard functions can be found in van Keilegom and Veraverbeke [23]. In comparison to Kooperberg et al. [15] we consider the stronger MISE (mean integrated square error). The convergence rate obtained in this paper depends on the smoothness parameter p of the underlying class of hazard functions as well as on the dimension of the covariables. Some empirical study of an adaptive estimator ('HARE') has been given in Kooperberg et al. [14]. The related complexity regularized estimator introduced in Section 4 of this paper is proved to be adaptive up to a logarithmic order and, therefore, approximatively minimax adaptive. We also discuss applications to general net sieves assuming that the log-hazard function allows an integral representation. In particular, we consider neural nets, radial basis-function nets and wavelet nets. For further details related to this paper we refer to the thesis by Döhler [7]. Some related consistency results (without rates) have been given in Döhler [6].

The paper is organized as follows: In Section 2 we establish an exponential inequality for the log-likelihood functional in the case of right censored data and indicate how similar exponential inequalities can be derived in a general framework. We use this result to obtain general error bounds for sieved ML-estimators (Section 3) and their complexity regularized versions (Section 4). In Section 5 we apply these results to tensor product splines and neural net type sieves. We conclude the paper with a short outlook.

The framework of hazard function estimation is as in Kooperberg et al. [15] where however also additive models are considered. Let $(\Omega, \mathcal{A}, P)$ be the underlying probability space, $T: \Omega \to \boldsymbol{R}_+$ a survival (failure) time, $C: \Omega \to \mathcal{T}$

a bounded censoring time, $X: \Omega \to \mathscr{X} = [0, 1]^k$ a vector of covariates, and $Y = T \wedge C$ the observable time. By normalization we assume without loss of generality that $\mathscr{T} = [0, 1]$. With the censoring indicator $\delta = 1_{(T \leqslant C)}$ (right censoring) the observation vector is $Z = (X, Y, \delta)$. We assume the existence of a conditional density $f_0(t \mid x)$ and denote by $F_0(t \mid x)$ the conditional distribution function of T given $X = x$. Further, we define the conditional hazard function

$$\lambda_0(t \mid x) = \frac{f_0(t \mid x)}{\bar{F}_0(t \mid x)}$$

with conditional survival function $\bar{F}_0(t \mid x) = 1 - F_0(t \mid x)$, and finally the conditional log-hazard function $\alpha_0(t \mid x) = \log \lambda_0(t \mid x)$. Based on i.i.d. data $(T_1, C_1, X_1), \ldots, (T_n, C_n, X_n)$, respectively, the corresponding observed data $Z_i = (X_i, Y_i, \delta_i)$, $1 \leqslant i \leqslant n$, our aim is to estimate the underlying conditional log-hazard function α_0.

According to Kooperberg et al. [15] the conditional log-likelihood of a sample $z_1, \ldots, z_n$ is given by

$$L_n(\alpha) = \sum_{i=1}^{n} l(z_i, \alpha), \tag{1.1}$$

where $l((x, y, \delta), \alpha) = \delta\alpha(y, x) - \int_0^y \exp \alpha(u, x)\, du$.

The underlying log-hazard function is assumed to be in a class $\mathscr{F}$ of functions on $\mathscr{T} \times \mathscr{X}$ to be specified later. Generally, we assume that α is bounded on $\mathscr{T} \times \mathscr{X}$ and that T and C are conditionally independent given X.

Let

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

denote the expected conditional log-likelihood function. Then Λ is maximized at the underlying conditional log-hazard functional α_0. The sieved maximum-likelihood estimator $\hat{\alpha}_n$ will be defined by

$$\hat{\alpha}_n = \arg\max_{\alpha \in \mathscr{F}_n} L_n(\alpha) \tag{1.3}$$

over some net (sieve) $\mathscr{F}_n \subset \mathscr{F}$ depending on the number n of observations.

For the 'Λ-distance' between an arbitrary element $\alpha \in \mathscr{F}$ and the underlying true α_0 the following representation is useful (see Döhler [6]):

$$|\Lambda(\alpha) - \Lambda(\alpha_0)| = \Lambda(\alpha_0) - \Lambda(\alpha) = \int_{\mathscr{T} \times \mathscr{X}} \bar{F}_{C|X}\, G(\alpha - \alpha_0)\, dP^{(T,X)}, \tag{1.4}$$

where $\bar{F}_{C|X}$ is the conditional survival-function of the censoring time C and $G(y) = \exp(y) - (1 + y)$. A standard argument leads to the decomposition of the estimation error of the ML-estimator $\hat{\alpha}_n$ (in Λ-distance) in an approximation

error and a stochastic error:

$$|\Lambda(\hat{\alpha}_n)-\Lambda(\alpha_0)| \leqslant \inf_{\alpha\in\mathscr{F}_n} |\Lambda(\alpha)-\Lambda(\alpha_0)| + 2\sup_{\alpha\in\mathscr{F}_n} |n^{-1}L_n(\alpha)-\Lambda(\alpha)|. \tag{1.5}$$

The main tool for proving convergence rates for the stochastic error of $\hat{\alpha}_n$ will be an exponential maximal inequality derived in Section 2. As in Kooperberg et al. [15] we introduce the L^p-distance on $\mathscr{F}$ modified by the conditional survival function:

$$\|\alpha-\beta\|_p^p = \int_{\mathscr{T}\times\mathscr{X}} \bar{F}_{C|X}\,|\alpha-\beta|^p\,dP^{(T,X)}. \tag{1.6}$$

From the representation in (1.4) we obtain (see Döhler [6]):

(1.7) Λ-convergence of $\alpha_n \to \alpha_0$ implies $\|\alpha_n-\alpha_0\|_1 \to 0$.

Also, for $\alpha, \beta\in\mathscr{F}$, $|\alpha| \leqslant M$, $|\beta| \leqslant M$ we have

$$k\,\|\alpha-\alpha_0\|_2^2 \leqslant |\Lambda(\alpha)-\Lambda(\beta)| \leqslant k'\,\|\alpha-\beta\|_2^2, \tag{1.8}$$

where $k = k(M) = 1/4M$, $k' = k'(M) = [\exp(2M)]/4M^2$.

For the proof of (1.8) define

$$F(y) = \begin{cases} G(y)/y^2 & \text{if } y \neq 0, \\ \frac{1}{2} & \text{if } y = 0, \end{cases}$$

where G is as in (1.4). Then it is easy to establish that F is strictly increasing on $\boldsymbol{R}$ and $F(2M) \leqslant k'(M)$, $F(-2M) \geqslant k(M)$ for $M \geqslant 1$. Therefore, $k(M)y^2 \leqslant G(y) \leqslant k'(M)y^2$, which implies $k(M)(\alpha-\beta)^2 \leqslant G(\alpha-\beta) \leqslant k'(M)(\alpha-\beta)^2$, and the result follows.

Finally, we note that for $\beta, \alpha\in\mathscr{F}$, $|\beta|, |\alpha| \leqslant M$

$$E\big(l(Z,\alpha)-l(Z,\beta)\big)^2 \leqslant \big(B_0\exp(M)+1\big)^2\,\|\alpha-\beta\|_2^2, \tag{1.9}$$

where $B_0 = \exp(M)\exp\big(\exp(M)\big)$. For the proof see Döhler [7], Proposition 2.9. So the L_2-norm allows us to control the expected squared loss in the likelihood.

2. EXPONENTIAL MAXIMAL INEQUALITY FOR THE LOG-LIKELIHOOD FUNCTIONAL

In this section we derive an exponential maximal inequality for the log-likelihood functional $L_n(\alpha)$. The proof is based on the following exponential inequality of Lee et al. [17] which was used in their paper and also in Kohler [11] and Krzyzak and Linder [16] for regression estimation by minimum L^2-empirical risk estimators. Using the error decomposition in (1.5) and relations (1.6)–(1.8) we will apply this result to obtain convergence rates of ML-estimators for right censored data.

Let $N(\varepsilon, \mathscr{F}, d)$ denote the ε-covering number of $\mathscr{F}$ with respect to a metric d. In the following we will use L^p-metrics denoted by $d_{L^p(\mu)}$ on certain L^p-spaces. The notion of permissibility of $\mathscr{F}$ means a weak measurability condition on $\mathscr{F}$ allowing to measure sets involving suprema over $f \in \mathscr{F}$. For a formal definition see Pollard [20], p. 196. For this and related notions and some basic results on VC-classes we refer to van der Vaart and Wellner [22] and Pollard [21].

THEOREM 2.1 (Lee, Bartlett, and Williamson [17]). *Let $\mathscr{F}$ be a permissible class of functions on $\mathscr{Z}$ with $|f| \leqslant K_1$, $Ef \geqslant 0$ and $Ef^2 \leqslant K_2 Ef$ for all $f \in \mathscr{F}$. Let $v, v_c > 0$, $0 < \alpha \leqslant \frac{1}{2}$. Then for*

$$m \geqslant \max\{4(K_1+K_2)/\alpha^2(v+v_c), K_1^2/\alpha^2(v+v_c)\}$$

the following holds:

$$(2.1)\quad P\left(\sup_{f\in\mathscr{F}} \frac{Ef - m^{-1}\sum_{i=1}^m f(z_i)}{v+v_c+Ef} \geqslant \alpha\right) \leqslant \sup_{\bar{z}\in\mathscr{Z}^{2m}} 2N\left(\frac{\alpha v_c}{4}, \mathscr{F}, d_{L^1(\nu_{\bar{z}})}\right)\exp\left(\frac{-3\alpha^2 vm}{4K_1+162K_2}\right) + \sup_{\bar{z}\in\mathscr{Z}^{2m}} 4N\left(\frac{\alpha v_c}{4K_1}, \mathscr{F}, d_{L^1(\nu_{\bar{z}})}\right)\exp\left(\frac{-\alpha^2 vm}{2K_1^2}\right),$$

where $\nu_{\bar{z}} = (2m)^{-1}\sum_{i=1}^{2m}\delta_{\bar{z}_i}$.

Let now $\mathscr{Z} = \mathscr{X}\times\mathscr{T}\times\{0,1\}$ and for $z_i = (x_i, y_i, \delta_i) \in \mathscr{Z}$, $1 \leqslant i \leqslant n$, and $\bar{z} = (z_1, \ldots, z_n)$ let $\nu_{\bar{z}} = n^{-1}\sum_{i=1}^n \delta_{z_i}$, $\tilde{\nu}_{\bar{z}} = n^{-1}\sum_{i=1}^n \delta_{x_i}$ and let $U[0,1]$ be the uniform distribution on $[0,1]$.

THEOREM 2.2 (maximal inequality for the log-likelihood). *There exists $B_0 = B_0(\|\alpha_0\|) > 0$ such that for all $M \geqslant M_0 := \|\alpha_0\|_\infty$, for all admissible $\mathscr{F} \subset \{\alpha\colon \mathscr{T}\times\mathscr{X} \to [-M, M]\}$, for any $v, v_c > 0$, $0 < \gamma \leqslant \frac{1}{2}$ and*

$$n \geqslant \frac{24M(B_0\exp(M)+1)^2}{\gamma^2(v+v_c)}$$

the following holds:

$$(2.2)\quad P\left(\sup_{\alpha\in\mathscr{F}} \frac{\Lambda(\alpha_0)-\Lambda(\alpha)-n^{-1}(L_n(\alpha_0)-L_n(\alpha))}{v+v_c+\Lambda(\alpha_0)-\Lambda(\alpha)} \geqslant \gamma\right) \leqslant \kappa(v_c, \mathscr{F})\exp\left(-\frac{\gamma^2 vn}{\kappa_0 B_0^2 M\exp(2M)}\right),$$

where

$$\kappa(v_c, \mathscr{F}) = 6\sup_{\bar{z}\in\mathscr{Z}^{2n}}\left[N\left(\frac{\gamma v_c}{64\exp(M)}, \mathscr{F}, d_{L^1(\nu_{\bar{z}})}\right)N\left(\frac{\gamma v_c}{64\exp(2M)}, \mathscr{F}, d_{L^1(\tilde{\nu}_{\bar{z}}\otimes U[0,1])}\right)\right]$$

and $\gamma_0 = 2608/3$.

Proof. Without loss of generality assume that $M_0, B_0 \geq 1$. Define

$$F = \{f_\alpha = l(\cdot, \alpha_0) - l(\cdot, \alpha); \alpha \in \mathscr{F}\}.$$

Then, by (1.4), $|f_\alpha| \leq 2(M + \exp(M))$. Also by (1.4) we have $Ef_\alpha \geq 0$ and by the application of (1.9) and (1.8) we obtain

$$Ef_\alpha^2 = E[l(\cdot, \alpha_0) - l(\cdot, \alpha)]^2 \leq (B_0 \exp(M) + 1)^2 \|\alpha - \alpha_0\|_2^2 \leq 4M(B_0 \exp(M) + 1)^2 Ef_\alpha.$$

This implies that the conditions of Theorem 2.1 are fulfilled with

$$K_1 = 2(M + \exp(M)), \quad K_2 = 4M(B_0 \exp(M) + 1)^2.$$

Therefore, for

$$n \geq \max\left\{4\frac{K_1 + K_2}{\gamma^2(v + v_c)}, \frac{K_1^2}{\gamma^2(v + v_c)}\right\} = 4\frac{K_1 + K_2}{\gamma^2(v + v_c)}$$

the following holds:

$$(2.3) \quad P\left(\sup_{\alpha \in \mathscr{F}} \frac{\Lambda(\alpha_0) - \Lambda(\alpha) - n^{-1}(L_n(\alpha_0) - L_n(\alpha))}{v + v_c + \Lambda(\alpha_0) - \Lambda(\alpha)} \geq \gamma\right)$$
$$\leq \sup_{\bar{z} \in \mathscr{Z}^{2n}} 2N\left(\frac{\gamma v_c}{4}, F, d_{L^1(\nu_{\bar{z}})}\right) \exp\left(-\frac{-3\gamma^2 vn}{4K_1 + 162K_2}\right)$$
$$+ \sup_{\bar{z} \in \mathscr{Z}^{2n}} 4N\left(\frac{\gamma v_c}{4K_1}, F, d_{L^1(\nu_{\bar{z}})}\right) \exp\left(-\frac{\gamma^2 vn}{2K_1^2}\right).$$

By easy calculations,

$$\max\left\{2K_1^2, \frac{4K_1 + 162K_2}{3}\right\} \leq \kappa_0 B_0^2 M \exp(2M)$$

and

$$4\frac{K_1 + K_2}{\gamma^2(v + v_c)} \leq n_0 := \frac{24M(B_0 \exp(M) + 1)^2}{\gamma^2(v + v_c)}.$$

Therefore, using $4\exp(M) \geq K_1 \geq 1$ we infer that for $n \geq n_0$ the right-hand side of (2.3) is bounded above by

$$6 \sup_{\bar{z} \in \mathscr{Z}^{2n}} N\left(\frac{\gamma v_c}{4K_1}, F, d_{L^1(\nu_{\bar{z}})}\right) \exp\left(-\frac{\gamma^2 vn}{\kappa_0 B_0^2 M \exp(2M)}\right)$$
$$\leq 6 \sup_{\bar{z} \in \mathscr{Z}^{2n}} N\left(\frac{\gamma v_c}{16\exp(M)}, F, d_{L^1(\nu_{\bar{z}})}\right) \exp\left(-\frac{\gamma^2 vn}{\kappa_0 B_0^2 M \exp(2M)}\right).$$

Now Theorem 2.2 will be a consequence of the following estimate for $\varepsilon > 0$:

$$(2.4) \quad N(\varepsilon, \mathscr{F}, d_{L^1(\nu_{\bar{z}})}) \leq N\left(\frac{\varepsilon}{4}, \mathscr{F}, d_{L^1(\nu_{\bar{z}})}\right) N\left(\frac{\varepsilon}{4\exp(M)}, \mathscr{F}, d_{L^1(\tilde{\nu}_{\bar{z}} \otimes U[0,1])}\right).$$

For the proof of (2.4) let us introduce $\tilde{F} = \{\tilde{f}_\alpha(\cdot) = l(\cdot, \alpha);\ \alpha \in \mathscr{F}\}$. Then

$$N(\varepsilon, F, d_{L^1(\nu_{\tilde{z}})}) \leqslant N\left(\frac{\varepsilon}{2}, \tilde{F}, d_{L^1(\nu_{\tilde{z}})}\right) N\left(\frac{\varepsilon}{2}, \{l(\alpha_0)\}, d_{L^1(\nu_{\tilde{z}})}\right) = N\left(\frac{\varepsilon}{2}, \tilde{F}, d_{L^1(\nu_{\tilde{z}})}\right).$$

Let us define

$$\mathscr{H} = \{g_\alpha(x, y, \delta) = \delta\alpha(y, x);\ \alpha \in \mathscr{F}\}$$

and

$$\mathscr{K} = \{k_\alpha(x, y, \delta) = \int_0^y \exp\alpha(u, x)\,du;\ \alpha \in \mathscr{F}\}.$$

Then, obviously, $N(\varepsilon, \mathscr{H}, d_{L^1(\nu_{\tilde{z}})}) \leqslant N(\varepsilon, \mathscr{F}, d_{L^1(\nu_{\tilde{z}})})$. Further,

$$\begin{aligned} d_{L^1(\nu_{\tilde{z}})}(k_{\alpha_1}, k_{\alpha_2}) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{y_i} |\exp\alpha_1(u, x_i) - \exp\alpha_2(u, x_i)|\,du \\ &\leqslant \frac{1}{n}\sum_{i=1}^{n}\int_0^{1} |\exp\alpha_1(u, x_i) - \exp\alpha_2(u, x_i)|\,du \\ &= d_{L^1(\tilde{\nu}_{\tilde{z}}\otimes U[0,1])}(\exp\circ\alpha_1, \exp\circ\alpha_2), \end{aligned}$$

which implies

$$N(\varepsilon, \mathscr{K}, d_{L^1(\nu_{\tilde{z}})}) \leqslant N(\varepsilon, \exp\circ\mathscr{F}, d_{L^1(\tilde{\nu}_{\tilde{z}}\otimes U[0,1])}) \leqslant N\left(\frac{\varepsilon}{\exp(M)}, \mathscr{F}, d_{L^1(\tilde{\nu}_{\tilde{z}}\otimes U[0,1])}\right)$$

by using the fact that for $\mathscr{F}$ with $|f| \leqslant K$ for $f \in \mathscr{F}$ and Lipschitz functions $\varphi\colon [-K, K] \to \boldsymbol{R}$

$$N(\varepsilon, \varphi\circ\mathscr{F}, d_{L^p(\mu)}) \leqslant N\left(\frac{\varepsilon}{\operatorname{Lip}\varphi}, \mathscr{F}, d_{L^p(\mu)}\right). \tag{2.5}$$

Consequently, using a well-known upper bound for the covering number of the sum of two function classes we obtain

$$\begin{aligned} N(\varepsilon, \tilde{F}, d_{L^1(\nu_{\tilde{z}})}) &= N(\varepsilon, \mathscr{H}\ominus\mathscr{K}, d_{L^1(\nu_{\tilde{z}})}) \\ &\leqslant N\left(\frac{\varepsilon}{2}, \mathscr{F}, d_{L^1(\nu_{\tilde{z}})}\right) N\left(\frac{\varepsilon}{2\exp(M)}, \mathscr{F}, d_{L^1(\tilde{\nu}_{\tilde{z}}\otimes U[0,1])}\right). \end{aligned}$$

Thus, combining the above estimates we get the statement of Theorem 2.2. ∎

Remark 2.3 (more general loss functions and estimation problems). From the proof of Theorem 2.2 one obtains a similar maximal inequality for more general loss functions l (i.e. for more general estimation problems and

(minimum contrast) estimation methods) satisfying the following three conditions:

$$|l(\alpha_0)-l(\alpha)| \leqslant K_1, \quad El(\alpha_0) \geqslant El(\alpha),$$

$$E\big(l(\alpha_0)-l(\alpha)\big)^2 \leqslant K_2 E\big(l(\alpha_0)-l(\alpha)\big). \tag{2.6}$$

For (2.6) the following two conditions corresponding to (1.8) and (1.9) are sufficient:

$$E\big(l(\alpha_0)-l(\alpha)\big) \geqslant k\|\alpha-\alpha_0\|_2^2, \tag{2.7}$$

$$E\big(l(\alpha_0)-l(\alpha)\big)^2 \leqslant \tilde{k}\|\alpha-\alpha_0\|_2^2. \tag{2.8}$$

Therefore, under condition (2.6) we obtain exponential inequalities with

$$N\left(\frac{\gamma v_c}{4K_1}, F, d_{L^1(v_{\bar{z}})}\right) \tag{2.9}$$

replacing the capacity term

$$N\left(\frac{\gamma v_c}{64\exp(M)}, \mathscr{F}, d_{L^1(v_{\bar{z}})}\right) N\left(\frac{\gamma v_c}{64\exp(2M)}, \mathscr{F}, d_{L^1(\tilde{v}_{\bar{z}}\otimes U[0,1])}\right)$$

in (2.2), where $F = \{f_\alpha = l(\alpha_0)-l(\alpha);\ \alpha\in\mathscr{F}\}$ is defined as in the proof of Theorem 2.2. This exponential inequality can be applied to prove convergence rates for the corresponding empirical minimum risk estimators. Condition (2.8) corresponds roughly to condition M1 in Birgé and Massart [5]. Condition (2.7) together with an upper bound as in (1.8) corresponds to condition C in Birgé and Massart [5]. Their growth condition M2 involving also the L_∞-metric is replaced in our approach by corresponding growth conditions on the L^1-covering numbers $N(\cdot, F, d_{L^1(v_{\bar{z}})})$ which then is closer related to the L^1-metric condition with bracketing $M_{1,[\ ]}$ in Barron et al. [4].

3. ERROR BOUNDS FOR MAXIMUM-LIKELIHOOD ESTIMATORS FOR CONDITIONAL LOG-HAZARD FUNCTIONS

As a measure of complexity of a model $\mathscr{F}$ we define

$$\mathscr{C}_n(\mathscr{F}) = 6 \sup_{\bar{z}\in\mathscr{X}^{2n}} N(1/n, \mathscr{F}, d_{L^1(v_{\bar{z}})}) N(1/n, \mathscr{F}, d_{L^1(\tilde{v}_{\bar{z}}\otimes U[0,1])}), \tag{3.1}$$

which arises from the first part of the estimate in (2.2). The following theorem estimates the mean Λ-error and the MISE of the ML-estimator in a model $\mathscr{F}$. The admissibility of $\mathscr{F}$ is a weak measurability condition (cf. Lee et al. [17]) which is satisfied for the examples considered in this paper.

THEOREM 3.1. *Let* $\mathscr{F} \subset \{\alpha\colon \mathscr{T}\times\mathscr{X} \to [-M, M]\}$ *be admissible where* $M \geqslant M_0 = \|\alpha_0\|_\infty$ *and* B_0, κ_0 *are as in Theorem* 2.2. *Assume that* $\mathscr{C}_n(\mathscr{F}) < \infty$.

Then for the ML-estimator $\hat{\alpha}_n = \arg\max_{\alpha \in \mathscr{F}} L_n(\alpha)$ *the following error estimates hold*:

(3.2) $$E\,|\Lambda(\hat{\alpha}_n) - \Lambda(\alpha_0)| \leqslant 2 \inf_{\alpha \in \mathscr{F}} |\Lambda(\alpha) - \Lambda(\alpha_0)| + 8\kappa_0 B_0^2 M \exp(2M) \frac{\log \mathscr{C}_n(\mathscr{F}) + 1}{n}$$

and

(3.3) $$E\,\|\hat{\alpha}_n - \alpha_0\|_2^2 \leqslant 2\exp(2M) \inf_{\alpha \in \mathscr{F}} \|\alpha - \alpha_0\|_2^2 + 32\kappa_0 B_0^2 M^2 \exp(2M) \frac{\log \mathscr{C}_n(\mathscr{F}) + 1}{n}.$$

Proof. In our proof we use a similar technique to that in the context of the regression estimation in Kohler [11]–[13]. We decompose the Λ-error into two parts:

(3.4) $$|\Lambda(\hat{\alpha}_n) - \Lambda(\alpha_0)| = T_{1,n} + T_{2,n}$$

with

$$T_{1,n} = \Lambda(\alpha_0) - \Lambda(\hat{\alpha}_n) - \frac{2}{n}\big(L_n(\alpha_0) - L_n(\hat{\alpha}_n)\big) \quad \text{and} \quad T_{2,n} = \frac{2}{n}\big(L_n(\alpha_0) - L_n(\hat{\alpha}_n)\big).$$

From the definition of $\hat{\alpha}_n$ we obtain, by a standard argument,

$$ET_{2,n} \leqslant 2 \inf_{\alpha \in \mathscr{F}} |\Lambda(\alpha) - \Lambda(\alpha_0)|.$$

It remains to establish the inequality

(3.5) $$ET_{1,n} \leqslant 8\kappa_0 B_0^2 M \exp(2M) \frac{\log \mathscr{C}_n(\mathscr{F}) + 1}{n}.$$

For $t \geqslant t_0 = [96M(B_0 \exp(M) + 1)^2]/n$ we obtain from Theorem 2.2 with $\gamma = \frac{1}{2}$ and $v = v_c = t/2$ the relation

$$P(T_{1,n} \geqslant t) \leqslant P\left(\sup_{\alpha \in \mathscr{F}} \frac{\Lambda(\alpha_0) - \Lambda(\alpha) - n^{-1}\big(L_n(\alpha_0) - L_n(\alpha)\big)}{t/2 + t/2 + \Lambda(\alpha_0) - \Lambda(\alpha)} \geqslant \frac{1}{2}\right)$$

$$\leqslant 6 \sup_{\vec{z} \in \mathscr{F}^{2n}} \left[N\left(\frac{\frac{1}{2}(t_0/2)}{64\exp(M)}, \mathscr{F}, d_{L^1(\nu_{\vec{z}})}\right) N\left(\frac{\frac{1}{2}(t_0/2)}{64\exp(2M)}, \mathscr{F}, d_{L^1(\vec{\nu}_{\vec{z}} \otimes U[0,1])}\right)\right]$$

$$\times \exp\left(-\frac{tn}{8\kappa_0 B_0^2 M \exp(2M)}\right).$$

For $M \geqslant 1$ we have

$$\frac{\frac{1}{2}(t_0/2)}{64\exp(2M)} \geqslant \frac{1}{n}$$

and, therefore,

$$P(T_{1,n} \geqslant t) \leqslant \mathscr{C}_n(\mathscr{F}) \exp\left(-\frac{tn}{8\kappa_0 B_0^2 M \exp(2M)}\right).$$

This implies for $u \geqslant t_0$

$$(3.6) \quad ET_{1,n} \leqslant \int_0^u 1\,dt + \int_u^\infty P(T_{1,n} \geqslant t)\,dt$$

$$\leqslant u + \mathscr{C}_n(\mathscr{F}) \frac{8\kappa_0 B_0^2 M \exp(2M)}{n} \exp\left(-\frac{un}{8\kappa_0 B_0^2 M \exp(2M)}\right).$$

The right-hand side of (3.6) is minimized by

$$u_0 = \frac{1}{n} 8\kappa_0 B_0^2 M \exp(2M) \log \mathscr{C}_n(\mathscr{F}).$$

It is easy to see that $u_0 \geqslant t_0$. With this u_0 inserted in (3.6) we obtain the inequality (3.5), and so the statement (3.2).

From (1.8) we then conclude

$$E\|\hat{\alpha}_n - \alpha_0\|_2^2 \leqslant 4M\, E|A(\hat{\alpha}_n) - A(\alpha_0)|$$

$$\leqslant 8M \inf_{\alpha \in \mathscr{F}} |A(\alpha) - A(\alpha_0)| + 32\kappa_0 B_0^2 M^2 \exp(2M) \frac{\log \mathscr{C}_n(\mathscr{F}) + 1}{n}$$

$$\leqslant 8M \frac{\exp(2M)}{4M^2} \inf_{\alpha \in \mathscr{F}} \|\alpha - \alpha_0\|_2^2 + 32\kappa_0 B_0^2 M^2 \exp(2M) \frac{\log \mathscr{C}_n(\mathscr{F}) + 1}{n}. \quad \blacksquare$$

Remark 3.2 (general estimation problem). The error estimates in Theorem 3.1 decompose the error as usual into an approximation error and a stochastic error of order $[\log \mathscr{C}_n(\mathscr{F}) + 1]/n$. As in Remark 2.3 (see (2.9)) we obtain a similar estimate for general loss functions l by replacing the model complexity term $\mathscr{C}_n(\mathscr{F})$ by

$$(3.7) \qquad \mathscr{C}_n(F) = 6 \sup_{\bar{z} \in \mathscr{Z}^{2n}} N(1/n, F, d_{L^1(\nu_{\bar{z}})})$$

with $F = \{f_\alpha = l(\cdot, \alpha_0) - l(\cdot, \alpha);\ \alpha \in \mathscr{F}\}$. In comparison to a related result in Birgé and Massart ([5], Corollary 11, Section 5) which uses in the condition M2 assumptions on the L^2- and L^∞-covering numbers of $\mathscr{F}$ our estimate uses only L^1-covering numbers in the model complexity term $\mathscr{C}_n(\mathscr{F})$, respectively, $\mathscr{C}_n(F)$. Our condition is closer to the L^1-condition with bracketing $M_{1,[\ ]}$ in Barron et al. [4], Section 6.

By Pollard's estimate for bounded VC-classes $\mathscr{F}$, $d = \dim_{VC} \mathscr{F}$, with majorant H stating that for $\varepsilon > 0$

$$(3.8) \qquad N(\varepsilon \|H\|_{L^p(\mu)}, \mathscr{F}, d_{L^p(\mu)}) \leqslant \kappa d (16e)^d (1/\varepsilon)^{p(d-1)}$$

(see van der Vaart and Wellner [22], Theorem 2.6.7), we obtain from our estimate in (3.3) a direct connection of convergence rates to the VC-dimension of the class $\mathscr{F}$.

As a consequence of this remark we obtain

COROLLARY 3.3. *Under the conditions of Theorem* 3.1, *where* $\mathscr{F}$ *is a bounded VC-class, we obtain*

$$E\|\hat{\alpha}_n-\alpha_0\|_2^2 \leqslant C_1(M)\inf_{\alpha\in\mathscr{F}}\|\alpha-\alpha_0\|_2^2+C_2(M, B_0)\dim_{\mathrm{VC}}(\mathscr{F})\frac{\log n}{n}. \tag{3.9}$$

A similar convergence rate result holds for general estimation problems as in Remarks 2.3 and 3.2.

Remark 3.4 (sieve estimators). Let $(\mathscr{F}_K)_{K\in\mathbb{N}}$ be a sieve of VC-classes in the underlying model $\mathscr{F}$ with $D_K=\dim_{\mathrm{VC}}\mathscr{F}_K$ and approximation rate $b_K=\inf_{\alpha\in\mathscr{F}_K}\|\alpha-\alpha_0\|_2^2$. Assume that for some $r, s>0$

$$b_K=O(K^{-r}), \quad D_K=O(K^s). \tag{3.10}$$

There are two well-studied types of sieves, linear sieves, i.e. finite-dimensional vector spaces which approximate typically smooth function classes and, secondly, nets (like neural nets, radial basis function nets, etc.). Under assumption (3.10) we obtain from the estimate in (3.9), when choosing the optimal parameter K_n in the bias-variance decomposition (3.9), an estimate for the MISE of $\hat{\alpha}_n$ of the form

$$E\|\hat{\alpha}_n-\alpha_0\|_2^2=O\big(((\log n)/n)^{r/(r+s)}\big). \tag{3.11}$$

Here r determines the approximation rate of the sieve which is usually for splines, wavelets, polynomials related to smoothness of the parameter, and s determines the complexity of the net.

If $\mathscr{F}_K$ is a subset of a K-dimensional vector space, then $s=1$, and if $r=2p/d$ (where p is a degree of smoothness, and d a dimension of space), in some examples we will obtain optimal convergence rates up to logarithmic terms.

Polynomial rates (i.e. b_K are as in (3.10)) can also be obtained for function classes which are derived from VC-classes by some operations like transformations, sums, etc.

4. STRUCTURAL RISK-MINIMIZATION

From the maximal inequality in Theorem 2.2, for $\eta\in(0, 1)$ and any data dependent estimator $\alpha_n\in\mathscr{F}$ we infer that with probability $1-\eta$

$$|\Lambda(\alpha_n)-\Lambda(\alpha_0)| \leqslant \frac{1}{n}8\kappa_0 B_0^2 M\exp(2M)\log\frac{\mathscr{C}_n(\mathscr{F})}{\eta}+\frac{2}{n}\big(L_n(\alpha_0)-L_n(\alpha_n)\big). \tag{4.1}$$

The idea of structural risk minimization (complexity regularization) due to Vapnik [24] is to construct an estimator minimizing approximatively the right-hand side of (4.1), i.e. minimizing

$$c_n \frac{\log \mathscr{C}_n(\mathscr{F})}{n} - \frac{2}{n} L_n(\alpha), \tag{4.2}$$

where c_n is a slowly increasing function independent of the unknown parameters, which asymptotically majorizes the corresponding constant in (4.1). The minimization is carried out not only over α in one fixed class $\mathscr{F} = \mathscr{F}_n$ but also allows us to choose α within a finite set of model classes $\{\mathscr{F}_{n,p};\ p \in \mathscr{P}_n\}$, p typically describing some smoothness or network complexity. The error term

$$c_n \frac{\log \mathscr{C}_n(\mathscr{F}_{n,p})}{n}$$

can be interpreted as a penalization term for the complexity of the model.

A detailed and general description of this approach with several applications has been given in Barron et al. [4], based on the error estimates in Birgé and Massart [5] as well as on new tools. In that paper one also finds several references to this method. In our paper we use some technical ideas from Kohler ([11], the proof of Theorem 4.2), concerning regression estimates which minimize the empirical penalized squared loss in that paper.

Let $M_0 = \|\alpha_0\|_\infty > 0$ and $B_0 = B_0(\|\alpha_0\|_\infty) > 0$ be as in Theorem 2.2, let $\mathscr{P}_n$ be finite sets for $n \in \mathbb{N}$, and for $p \in \mathscr{P}_n$ let $\mathscr{F}_{n,p} \subset \{\alpha\colon \mathscr{T} \times \mathscr{X} \to [-M, M]\}$ be admissible models, $M \geq M_0$ with $\mathscr{C}_n(\mathscr{F}_{n,p}) < \infty$ for all $p \in \mathscr{P}_n$. Then the complexity regularized estimator α_n^* is defined in two steps:

Step 1. Let

$$p_n^* = \arg\min_{p \in \mathscr{P}_n} \left(-n^{-1} \sup_{\alpha \in \mathscr{F}_{n,p}} L_n(\alpha) + \mathrm{pen}_n(p)\right), \tag{4.3}$$

where $\mathrm{pen}_n(p)$ is a penalization term for complexity of the model $\mathscr{F}_{n,p}$ satisfying asymptotically as $n \to \infty$

$$\mathrm{pen}_n(p) \geq 4\kappa_0 B_0^2 M \exp(2M) \frac{\log \mathscr{C}_n(\mathscr{F}_{n,p})}{n}. \tag{4.4}$$

Step 2.

$$\alpha_n^* = \arg\max_{\alpha \in \mathscr{F}_{n,p_n^*}} L_n(\alpha). \tag{4.5}$$

It is important to note that the right-hand side of (4.4) is not supposed to be the actual penalty term used in application since it depends on the unknown M_0 and B_0. This expression represents a lower bound for the penalty, sufficient for Theorem 4.1 to hold (cf. also (4.2)). For asymptotic results the actual penal-

ty term should be chosen independently of M_0 and B_0, majorizing the right-hand side of (4.4) for large sample sizes. An example of how this can be done is given in Theorem 5.3. The following theorem gives an error bound for complexity regularized sieve estimators based on the maximal inequality in Theorem 2.2. A general related error bound is given in Barron et al. ([4], Theorem 8) under some alternative conditions on the L_2–L_∞-covering, respectively, the L_1-covering, with bracketing.

THEOREM 4.1. *For the complexity regularized ML-estimator α_n^* the following error estimates hold*:

$$E|\Lambda(\alpha_n^*)-\Lambda(\alpha_0)| \leqslant 2 \inf_{p\in\mathscr{P}_n}\left(\mathrm{pen}_n(p)+\inf_{\alpha\in\mathscr{F}_{n,p}}|\Lambda(\alpha)-\Lambda(\alpha_0)|\right) + \frac{4\kappa_0 B_0^2 M\exp(2M)}{n}(1+\log|\mathscr{P}_n|) \tag{4.6}$$

and

$$E\|\alpha_n^*-\alpha_0\|_2^2 \leqslant 2\inf_{p\in\mathscr{P}_n}\left(4M\,\mathrm{pen}_n(p)+\exp(2M)\inf_{\alpha\in\mathscr{F}_{n,p}}\|\alpha-\alpha_0\|_2^2\right) + \frac{16\kappa_0 B_0^2 M^2\exp(2M)}{n}(1+\log|\mathscr{P}_n|). \tag{4.7}$$

Proof. As in the proof of Theorem 3.1 we consider the decomposition of the error into two terms:

$$\begin{aligned} T_{1,n} &:= \Lambda(\alpha_0)-\Lambda(\alpha_n^*)-2n^{-1}\left(L_n(\alpha_0)-L_n(\alpha_n^*)\right)-2\mathrm{pen}_n(p_n^*),\\ T_{2,n} &:= 2n^{-1}\left(L_n(\alpha_0)-L_n(\alpha_n^*)\right)+2\mathrm{pen}_n(p_n^*). \end{aligned} \tag{4.8}$$

Our first aim is to prove

$$ET_{1,n} \leqslant \frac{4\kappa_0 B_0^2 M\exp(2M)}{n}(1+\log|\mathscr{P}_n|). \tag{4.9}$$

For the proof we obtain, as in Kohler ([11], p. 85),

$$P(T_{1,n}>t) \leqslant \sum_{p\in\mathscr{P}_n} P\left(\sup_{\alpha\in\mathscr{F}_{n,p}}\frac{\Lambda(\alpha_0)-\Lambda(\alpha)-n^{-1}\left(L_n(\alpha_0)-L_n(\alpha)\right)}{t+2\mathrm{pen}_n(p)+\Lambda(\alpha_0)-\Lambda(\alpha)}\geqslant\frac{1}{2}\right).$$

Then for

$$n \geqslant \frac{24M\left(B_0\exp(M)+1\right)^2}{\frac{1}{4}t}$$

we obtain from Theorem 2.2 with $\gamma=\frac{1}{2}$, $v=t+\mathrm{pen}_n(p)$ and $v_c=\mathrm{pen}_n(p)$, observing that the condition

$$n \geqslant \frac{24M\left(B_0\exp(M)+1\right)^2}{\frac{1}{4}\left(t+2\mathrm{pen}_n(p)\right)}$$

is fulfilled for any $p \in \mathscr{P}_n$,

$$P(T_{1,n} > t)$$

$$\leqslant \sum_{p \in \mathscr{P}_n} \underbrace{\left[6 \sup_{\tilde{z} \in \mathscr{Z}^{2n}} N\left(\frac{\frac{1}{2}\mathrm{pen}_n(p)}{64\exp(M)}, \mathscr{F}_{n,p}, d_{L^1(\nu_{\tilde{z}})}\right) N\left(\frac{\frac{1}{2}\mathrm{pen}_n(p)}{64\exp(2M)}, \mathscr{F}_{n,p}, d_{L^1(\tilde{\nu}_{\tilde{z}} \otimes U[0,1])}\right)\right.}_{=:s_n(p)}$$

$$\left. \times \exp\left(-\frac{\mathrm{pen}_n(p)\, n}{4\kappa_0 B_0^2 M \exp(2M)}\right)\right] \exp\left(-\frac{tn}{4\kappa_0 B_0^2 M \exp(2M)}\right).$$

Since $\log \mathscr{C}_n(\mathscr{F}_{n,p}) \geqslant 1$, and hence

$$\frac{1}{n} \leqslant \frac{\frac{1}{2}\mathrm{pen}_n(p)}{64\exp(M)}$$

for any $p \in \mathscr{P}_n$, we obtain $s_n(p) \leqslant \mathscr{C}_n(\mathscr{F}_{n,p})$. Further,

$$s_n(p) \exp\left(-\frac{\mathrm{pen}_n(p)}{4\kappa_0 B_0^2 M \exp(2M)} n\right)$$

$$\leqslant \exp\left(\log \mathscr{C}_n(\mathscr{F}_{n,p}) - \frac{\mathrm{pen}_n(p)}{4\kappa_0 B_0^2 M \exp(2M)} n\right) \leqslant \exp(0) = 1$$

by the definition of $\mathrm{pen}_n(p)$, and, therefore, for

$$t \geqslant t_0 := \frac{96M(B_0 \exp(M)+1)^2}{n}$$

we have

$$P(T_{1,n} > t) \leqslant |\mathscr{P}_n| \exp\left(-\frac{t}{4\kappa_0 B_0^2 M \exp(2M)} n\right). \tag{4.10}$$

This implies for $u \geqslant t_0$

$$ET_{1,n} \leqslant \int_0^u 1\,dt + \int_u^\infty P(T_{1,n} \geqslant t)\,dt \tag{4.11}$$

$$\leqslant u + |\mathscr{P}_n| \frac{4\kappa_0 B_0^2 M \exp(2M)}{n} \exp\left(-\frac{u}{4\kappa_0 B_0^2 M \exp(2M)} n\right).$$

The right-hand side of this inequality is minimized by

$$u = u_0 := \frac{4\kappa_0 B_0^2 M \exp(2M)}{n} \log |\mathscr{P}_n|,$$

and without loss of generality for

$$\kappa_0 \cdot \log |\mathscr{P}_n| \geqslant 24\left(1+\frac{1}{B_0 \exp(M)}\right)^2$$

it follows that $u_0 \geqslant t_0$. This choice of u leads to (4.9).

From the definition of p_n^* and α_n^* we obtain

$$T_{2,n} = 2\left[n^{-1} L_n(\alpha_0) - n^{-1} \sup_{\alpha \in \mathscr{F}_{n,p_n^*}} L_n(\alpha) + \mathrm{pen}_n(p_n^*)\right] \tag{4.12}$$

$$= 2\left[n^{-1} L_n(\alpha_0) + \inf_{p \in \mathscr{P}_n}\left(-n^{-1} \sup_{\alpha \in \mathscr{F}_{n,p}} L_n(\alpha) + \mathrm{pen}_n(p)\right)\right]$$

$$= 2 \inf_{p \in \mathscr{P}_n}\left[\inf_{\alpha \in \mathscr{F}_{n,p}} n^{-1}\left(L_n(\alpha_0) - L_n(\alpha)\right) + \mathrm{pen}_n(p)\right].$$

Therefore, since $\mathrm{pen}_n(p)$ is deterministic, we get

$$ET_{2,n} \leqslant 2 \inf_{p \in \mathscr{P}_n} E\left[\inf_{\alpha \in \mathscr{F}_{n,p}} n^{-1}\left(L_n(\alpha_0) - L_n(\alpha)\right) + \mathrm{pen}_n(p)\right] \tag{4.13}$$

$$\leqslant 2 \inf_{p \in \mathscr{P}_n}\left[\inf_{\alpha \in \mathscr{F}_{n,p}} En^{-1}\left(L_n(\alpha_0) - L_n(\alpha)\right) + \mathrm{pen}_n(p)\right]$$

$$\leqslant 2 \inf_{p \in \mathscr{P}_n}\left[\inf_{\alpha \in \mathscr{F}_{n,p}} |\Lambda(\alpha) - \Lambda(\alpha_0)| + \mathrm{pen}_n(p)\right].$$

The relations (4.11) and (4.13) imply (4.6). The estimate (4.7) then follows from (1.8). ∎

5. ADAPTIVE SIEVE ESTIMATES FOR THE CONDITIONAL LOG-HAZARD FUNCTION

In this section we apply the results of Sections 3 and 4 to several types of sieves. In the first part we show that the complexity regularized spline estimate is approximatively optimal even with unknown degree of smoothness, i.e. it has up to a logarithmic term the same optimal convergence rate as the estimator of Kooperberg et al. [15] in the case with known degree of smoothness. In the second part we obtain convergence results for net estimates under the assumption that the conditional log-hazards have a certain representation property. Some applications to Sobolev class models will be considered in a subsequent paper.

5.1. Tensor product splines. In this section we consider tensor product splines. For a general background of this class of functions we refer to Kohler [11] and the references given therein.

Let $V_{h,M}$ denote the class of tensor product splines of $[-hM, 1+hM]^{k+1}$ of degree $M \in N_0$ in each coordinate and of grid width $h > 0$. Let $\Phi(L, V_{h,M})$ (for $L > 0$) denote the class of truncated functions $T_L \circ g$, $g \in V_{h,M}$, where

$$T_L \circ g = \begin{cases} L & \text{if } g \geqslant L, \\ g & \text{if } -L \leqslant g \leqslant L, \\ -L & \text{if } g \leqslant -L. \end{cases} \tag{5.1}$$

We consider for $p = r+\beta$, $r \in N_0$, $\beta \in (0, 1)$, the smoothness classes $\Sigma(p, L)$ of bounded conditional hazard functions $\alpha(t, x)$ on $[0, 1]^{k+1}$ satisfying for all $z_1, z_2 \in [0, 1]^{k+1}$ the Hölder condition of order p:

$$\|\alpha\|_\infty \leqslant L \quad \text{and} \quad |D^r \alpha(z_1) - D^r \alpha(z_2)| \leqslant L \|z_1 - z_2\|_2^\beta. \tag{5.2}$$

For classes with known degree of smoothness we obtain the following result.

THEOREM 5.1 (known smoothness class). *Let* $1 \leqslant p < \infty$, $L > 0$, $\tilde{M} \in N$, $\tilde{M} \geqslant p-1$ *and* $h_n = \big((\log n)/n\big)^{1/(2p+k+1)}$. *Then the spline ML-estimator*

$$\hat{\alpha}_n = \operatorname*{arg\,max}_{\alpha \in \Phi(L, V_{h_n, \tilde{M}})} L_n(\alpha) \tag{5.3}$$

satisfies

$$\sup_{\alpha \in \Sigma(p,L)} E \|\hat{\alpha}_n - \alpha_0\|_2^2 = O\Big(\big((\log n)/n\big)^{2p/(2p+k+1)}\Big) \tag{5.4}$$

and

$$\sup_{\alpha \in \Sigma(p,L)} E\, |\Lambda(\hat{\alpha}_n) - \Lambda(\alpha_0)| = O\Big(\big((\log n)/n\big)^{2p/(2p+k+1)}\Big). \tag{5.5}$$

Proof. From the definition of the truncation operator T_L it follows that

$$\inf_{\alpha \in \Phi(L, V_{h_n, \tilde{M}})} \|\alpha_0 - \alpha\|_\infty \leqslant \inf_{\alpha \in V_{h_n, \tilde{M}}} \|\alpha_0 - \alpha\|_\infty$$

and, therefore, from the approximation result in Kohler ([11], Lemma 1.5) for the approximation of Hölder-continuous functions by tensor product splines (which is in sup-norm) we obtain

$$\inf_{\alpha \in \Phi(L, V_{h_n, \tilde{M}})} \|\alpha_0 - \alpha\|_2^2 \leqslant C h_n^{2p} \leqslant C\big((\log n)/n\big)^{2p/(2p+k+1)} \tag{5.6}$$

with $C = C(p, L)$ independent of α_0.

To estimate the stochastic error in Theorem 3.1 note that $V_{h_n, \tilde{M}}$ is a vector space of dimension less than or equal to $(\lceil 1/h_n \rceil + \tilde{M})^{k+1}$ (see

Kohler [11], p. 79), and therefore (cf. van der Vaart and Wellner [22], Lemma 2.6.18)

$$\dim_{VC} \Phi(L, V_{h_n,\tilde{M}}) \leqslant \dim_{VC} V_{h_n,\tilde{M}} \leqslant (\lceil 1/h_n \rceil + \tilde{M})^{k+1} + 2. \tag{5.7}$$

Therefore, from Theorem 3.1 we obtain

$$E\|\hat{a}_n - a_0\|_2^2 \leqslant C_1 \left(\frac{\log n}{n}\right)^{2p/(2p+k+1)} + C_2 \left[\left(\left\lceil \left(\frac{n}{\log n}\right)^{1/(2p+k+1)} \right\rceil + \tilde{M}\right)^{k+1} + 2\right] \frac{\log n}{n}. \tag{5.8}$$

Consequently, with $M := L$, $\mathcal{F} := \Phi(L, V_{h_n,\tilde{M}})$ and $B_0 = B_0(L)$ we get (5.4).

For the proof of (5.5) we next establish the approximation rate

$$\inf_{\alpha \in \Phi(L, V_{h_n,\tilde{M}})} |\Lambda(\alpha) - \Lambda(\alpha_0)| \leqslant C h_n^{2p} \tag{5.9}$$

for the Λ-distance.

From the representation (1.4) and by some elementary properties of the function G we obtain

$$\begin{aligned} \inf_{\alpha \in \Phi(L, V_{h_n,\tilde{M}})} |\Lambda(\alpha) - \Lambda(\alpha_0)| &= \inf_{\alpha \in \Phi(L, V_{h_n,\tilde{M}})} \int_{\mathcal{T} \times \mathcal{X}} \bar{F}_{C|X} G(\alpha - \alpha_0)\, dP^{(T,X)} \\ &\leqslant \inf_{\alpha \in \Phi(L, V_{h_n,\tilde{M}})} \int_{\mathcal{T} \times \mathcal{X}} \bar{F}_{C|X} G(\|\alpha - \alpha_0\|_\infty)\, dP^{(T,X)} \leqslant \inf_{\alpha \in \Phi(L, V_{h_n,\tilde{M}})} G(\|\alpha - \alpha_0\|_\infty) \\ &\leqslant G\Big(\inf_{\alpha \in \Phi(L, V_{h_n,\tilde{M}})} \|\alpha - \alpha_0\|_\infty \Big). \end{aligned}$$

For the last inequality we observe that $\inf_{x \in A} G(x) = G(\inf A)$ for $A \subset \mathbf{R}_+$. Since for $x_n \downarrow 0$, $G(x_n) = O(x_n^2)$, we obtain

$$\begin{aligned} \inf_{\alpha \in \Phi(L, V_{h_n,\tilde{M}})} |\Lambda(\alpha) - \Lambda(\alpha_0)| &= O\Big(\Big(\inf_{\alpha \in \Phi(L, V_{h_n,\tilde{M}})} \|\alpha - \alpha_0\|_\infty \Big)^2 \Big) \\ &= O\Big(\Big(\inf_{\alpha \in V_{h_n,\tilde{M}}} \|\alpha_0 - \alpha\|_\infty \Big)^2 \Big) = O(h_n^{2p}) \end{aligned}$$

as in (5.6). ∎

Remark 5.2. The convergence rate in (5.4) for the MISE is up to a logarithmic factor optimal in the minimax sense (see Kooperberg et al. [15], Remark to Corollary 1; however, note that the convergence in MISE is stronger), i.e.

$$\liminf_{n \to \infty} n^{2p/(2p+k+1)} \inf_{\hat{\alpha}_n} \sup_{\alpha_0 \in \Sigma(p,L)} E\|\hat{\alpha}_n - \alpha_0\|_2^2 > 0 \tag{5.10}$$

for any $p \geqslant 1$ and $L > 0$.

If we do not know the smoothness parameter, i.e. assume that

$$\alpha \in \Sigma := \bigcup_{1 \leqslant p < \infty; L < \infty} \Sigma(p, L), \tag{5.11}$$

then we will infer that our penalized spline ML-estimator defined in (4.5) adapts up to a logarithmic factor to the unknown smoothness and is up to $(\log n)^2$ minimax-adapted in the sense of Barron et al. [4]. An *adaptive* estimation method ('HARE') had been introduced in Kooperberg et al. [15] and empirically investigated there, however no adaptation result was proved. We show that the complexity regularized estimator α_n^* from (4.5) is approximatively adaptive.

THEOREM 5.3 (unknown smoothness degree, adaptation). *For* $n \in N$, $q_{\max}(n)$, $K_{\max}(n) \in N$ *let*

$$\mathscr{P}_n := \{(K, q) \in N \times N \mid K \leqslant K_{\max}(n),\ q \leqslant q_{\max}(n)\}$$

and for $(K, q) \in \mathscr{P}_n$ *and* $\beta_n := \frac{1}{5} \log\log n$ *define the models*

$$\mathscr{F}_{n,(K,q)} := \Phi(\beta_n, V_{1/K, q-1}).$$

Define the complexity regularized estimate α_n^* *as in* (4.5) *with penalization term*

$$\mathrm{pen}_n((K, q)) := \frac{(\log n)^{8/5}}{n} [(K+q-1)^{k+1} + 2].$$

For $K_{\max}(n) := n$ *and* $q = q_{\max}(n) \to \infty$ *such that* $q_{\max}(n)/n \to 0$ *we obtain for* $p \geqslant 1$ *and* $L > 0$

$$\sup_{\alpha_0 \in \Sigma(p,L)} E \|\alpha_n^* - \alpha_0\|_2^2 = O\left(\log n \left((\log n)/n\right)^{2p/(2p+k+1)}\right) \tag{5.12}$$

and

$$\sup_{\alpha_0 \in \Sigma(p,L)} E |\Lambda(\alpha_n^*) - \Lambda(\alpha_0)| = O\left(\log n \left((\log n)/n\right)^{2p/(2p+k+1)}\right). \tag{5.13}$$

Proof. For the proof of (5.12) and (5.13) we establish first the following more general estimates:

For $p \geqslant 1$, $L > 0$ there exists $N_0 = N_0(L) \in N$ such that for any $n \geqslant N_0$:

$$\begin{aligned} &\sup_{\alpha_0 \in \Sigma(p,L)} E \|\alpha_n^* - \alpha_0\|_2^2 \\ &\leqslant 2 \inf_{(K,q) \in \mathscr{P}_n} \left(\frac{4 \log\log n (\log n)^{8/5}}{5} \frac{}{n} [(K+q-1)^{k+1} + 2] \right. \\ &\quad \left. + (\log n)^{2/5} \inf_{\alpha \in \mathscr{F}_{n,(K,q)}} \|\alpha - \alpha_0\|_2^2 \right) \\ &\quad + \frac{16 \kappa_0 B_0^2}{25} \frac{(\log\log n)^2 (\log n)^{2/5}}{n} \left(1 + \log\left(K_{\max}(n)\, q_{\max}(n)\right)\right) \end{aligned} \tag{5.14}$$

and

$$\sup_{\alpha_0 \in \Sigma(p,L)} E\,|\Lambda(\alpha_n^*) - \Lambda(\alpha_0)| \tag{5.15}$$

$$\leqslant 2 \inf_{(K,q)\in\mathscr{P}_n} \left(\frac{(\log n)^{8/5}}{n} [(K+q-1)^{k+1}+2] + \inf_{\alpha \in \mathscr{F}_{n,(K,q)}} |\Lambda(\alpha_n^*) - \Lambda(\alpha_0)| \right)$$

$$+ \frac{16\kappa_0 B_0^2}{5} \frac{\log\log n (\log n)^{2/5}}{n} \left(1 + \log\left(K_{\max}(n)\, q_{\max}(n)\right)\right),$$

where κ_0 and B_0 are as in Theorem 3.1.

The statements (5.14) and (5.15) follow from Theorem 4.1 with $M := \beta_n > M_0 := L$, $n \geqslant n_0$ and the estimate

$$\mathrm{pen}_n((K, q)) \geqslant 4\kappa_0 B_0^2 \beta_n \exp(2\beta_n) \frac{\log \mathscr{C}_n(\mathscr{F}_{n,(K,q)})}{n}. \tag{5.16}$$

For the proof of (5.16) we use the fact that for a K-dimensional vector space V of functions and $\beta > 0$ the following estimate holds for any probability measure μ on $\mathscr{T} \times \mathscr{X}$ and $\varepsilon > 0$:

$$N(\varepsilon, \Phi(\beta, V), d_{L^p(\mu)}) \leqslant \kappa(K+2)(16e)^{K+2} \beta^{p(K+1)} (1/\varepsilon)^{p(K+1)} \tag{5.17}$$

with some universal constant κ. This implies that

$$N(\varepsilon, \mathscr{F}_{n,(K,q)}, d_{L^1(\mu)}) \tag{5.18}$$

$$\leqslant \kappa\left((K+q-1)^{k+1}+2\right)(16e)^{(K+q-1)^{k+1}+2} (\beta_n/\varepsilon)^{(K+q-1)^{k+1}+1},$$

since $V_{1/K,q-1}$ has dimension less than or equal to $(K+q-1)^{k+1}$. Therefore

$$\begin{aligned}
\log \mathscr{C}_n(\mathscr{F}_{n,(K,q)}) &\leqslant \log\kappa^2 + 2\log\left((K+q-1)^{k+1}+2\right) \\
&\quad + [(K+q-1)^{k+1}+2]\,[\log(n\beta_n) + 2\log 16e] \\
&\leqslant \log\kappa^2 + [(K+q-1)^{k+1}+2]\,[\log(n\beta_n) + 2\log 16e + 2] \\
&\leqslant 2\,[(K+q-1)^{k+1}+2]\log(n\beta_n)
\end{aligned}$$

for $n \geqslant N_0$, where N_0 is independent of K, q, L, p. Consequently,

$$\begin{aligned}
\frac{4\kappa_0 B_0^2 \beta_n \exp(2\beta_n) n^{-1} \log \mathscr{C}_n(\mathscr{F}_{n,(K,q)})}{\mathrm{pen}_n((K, q))} &\leqslant \frac{4\kappa_0 B_0^2 \beta_n \exp(2\beta_n) 2\log(n\beta_n)}{(\log n)^{8/5}} \\
&\leqslant \frac{4\kappa_0 B_0^2 \beta_n \exp(2\beta_n) 4}{(\log n)^{3/5}} \leqslant 1
\end{aligned}$$

for $n \geqslant N_0(\kappa_0, B_0(L))$, and so the result follows.

For the proof of (5.12) let $K_n := \lceil (n/\log n)^{1/(2p+k+1)} \rceil$. Then by the approximation result in (5.2) for $K_n \leqslant K_{\max}(n)$, and for $q_{\max}(n) \geqslant p$ we obtain

$$\inf_{(K,q)\in\mathscr{P}_n} \left(4\beta_n \operatorname{pen}_n((K,q)) + \exp(2\beta_n) \inf_{\alpha\in\mathscr{F}_{n,(K,q)}} \|\alpha-\alpha_0\|_2^2\right)$$

$$\leqslant 4\beta_n \operatorname{pen}_n((K_n,p)) + \exp(2\beta_n) \inf_{\alpha\in\mathscr{F}_{n,(K_n,p)}} \|\alpha-\alpha_0\|_2^2$$

$$\leqslant C_1 \frac{\log\log n (\log n)^{8/5}}{n} [(K_n+p-1)^{k+1}+2] + C_2 (\log n)^{2/5} \left(\frac{1}{K_n}\right)^{2p}$$

$$= O\left(\frac{\log\log n (\log n)^{8/5}}{n} \left(\frac{n}{\log n}\right)^{(k+1)/(2p+k+1)}\right) + O\left((\log n)^{2/5} \left(\frac{\log n}{n}\right)^{2p/(2p+k+1)}\right)$$

$$= O\left(\log n \left(\frac{\log n}{n}\right)^{2p/(2p+k+1)}\right).$$

This yields an estimate for the first term in (5.14). For the second term we use the assumptions on $K_{\max}(n)$ and $q_{\max}(n)$ to obtain the estimate

$$\frac{16\kappa_0 B_0^2}{25} \frac{(\log\log n)^2 (\log n)^{2/5}}{n} (1+\log(K_{\max}(n) q_{\max}(n)))$$

$$= O\left(\log n \left(\frac{\log n}{n}\right)^{2p/(2p+k+1)}\right),$$

which implies (5.12). We can prove (5.13) similarly observing that, as in the proof of the approximation error in (5.9),

$$\inf_{\alpha\in\mathscr{F}_{n,(K_n,p)}} |\Lambda(\alpha)-\Lambda(\alpha_0)| = O((1/K_n)^{2p}). \quad \blacksquare \tag{5.19}$$

The truncation constants $\beta_n = \frac{1}{5}\log\log n$ are not meant as proposals for practical examples. Note that the same estimates in the second part of the theorem hold for β_n of the form cl_n, where c is some bigger constant and l_n grows slower than $\log\log n$.

5.2. Net sieves. In this section we apply our results to obtain convergence rates for the conditional log-hazard function for net sieves under the assumption that the underlying conditional log-hazard function α_0 has an integral representation of the form

$$\alpha_0(t,x) = \int_\Theta \Psi(\alpha_\vartheta(t,x))\, dv(\vartheta), \tag{5.20}$$

where a_ϑ, $\vartheta\in\Theta\subset\mathbb{R}^m$, $x\in\mathscr{X}\subset\mathbb{R}^k$ is a set of sieve defining functions, $\Psi\circ a_\vartheta$ is the continuous net and v is a signed measure of bounded variation on Θ. This kind

of representation is typically related to some smoothness classes (see Yukich et al. [27]). Some approximation results by finite nets with rates of approximation are given in Döhler and Rüschendorf [9] and applied in the following. Let $\mathscr{X}_{k+1}(\mathscr{F}_0)$ denote the class of all functions satisfying (5.20).

Define the basis of the net $\mathscr{F}_0 = \{\Psi \circ a_\vartheta;\ \vartheta \in \Theta\}$ and, for $\beta > 0$, $K \in \mathbf{N}$, the finite approximation net

$$\mathscr{F}(\beta, K) = \{\alpha:\ \mathscr{T} \times \mathscr{X} \to \mathbf{R};\ \alpha(t, x) = \sum_{i=1}^{K} c_i f_i(t, x),\ f_i \in \mathscr{F}_0,\ \sum_{i=1}^{K} |c_i| \leqslant \beta\}. \tag{5.21}$$

The following conditions were introduced in Döhler and Rüschendorf [9] to prove approximation rates by finite nets. Let μ be a probability measure on $\mathbf{R}^{d+1}$.

(A1) There exists a $D > 1$ such that

$$N(\delta, \mathscr{F}_0, d_{L^2(\mu)}) = O\big((1/\delta)^{2(D-1)}\big). \tag{5.22}$$

(A2) Define $b_z(\vartheta) = a_\vartheta(z)$, $z \in \mathscr{T} \times \mathscr{X}$. Then the class $\{\Psi \circ b_z,\ z \in \mathbf{R}^{k+1}\}$ is a P-Donsker class for any probability measure P on Θ.

THEOREM 5.4. *Assume the conditions* (A1) *and* (A2) *are satisfied. Let*

$$K_n := n^{\frac{1}{2+1/(D-1)}}, \qquad \beta_n := \tfrac{1}{5}\log\log n,$$

and consider the net ML-estimator

$$\hat{\alpha}_n := \operatorname*{arg\,max}_{\alpha \in \mathscr{F}(\beta_n, K_n)} L_n(\alpha).$$

Then for any $\alpha_0 \in \mathscr{X}_{k+1}(\Psi, \mathfrak{T}, \mathscr{T} \times \mathscr{X})$ *the following holds*:

$$E\,\|\hat{\alpha}_n - \alpha_0\|_2^2 = O\left(\frac{(\log n)^2}{n^{1/2+1/(4D-2)}}\right) \tag{5.23}$$

and

$$E\,|\Lambda(\hat{\alpha}_n) - \Lambda(\alpha_0)| = O\left(\frac{(\log n)^2}{n^{1/2+1/(4D-2)}}\right). \tag{5.24}$$

Proof. We apply Theorem 4.1. The approximation error was estimated in Döhler and Rüschendorf [9]. Let ν_{α_0} be the signed measure representing α_0. Then for n with $\beta_n \geqslant 2|\nu_{\alpha_0}|$ we have

$$\inf_{\alpha \in \mathscr{F}(\beta_n, K_n)} \|\alpha - \alpha_0\|_2^2 = O\big((1/K_n)^{1+1/(D-1)}\big). \tag{5.25}$$

Next we prove that for $\beta > 0$, $K \in N$ the following inequality holds:

$$\mathscr{C}_n(\mathscr{F}(\beta, K)) \leqslant C(D)^K (\beta K)^{2K(2D-1)} n^{2K(2D-1)}. \tag{5.26}$$

Define

$$\mathscr{F}'(\beta, K) := \{\alpha\colon \mathscr{T} \times \mathscr{X} \to \mathbf{R},\ (t, x) \mapsto \sum_{i=1}^{K} c_i f_i(t, x) \mid f_i \in \mathscr{F}_0,\ |c_i| \leqslant \beta\}.$$

Then $\mathscr{F}(\beta, K) \subset \mathscr{F}'(\beta, K)$, and for any probability measure ν on $\mathscr{T} \times \mathscr{X}$ and $\delta > 0$ using some well-known rules for covering numbers (cf. van der Vaart and Wellner [22]) we obtain

$$\begin{aligned}
N(\delta, \mathscr{F}'(\beta, K), d_{L^1(\nu)}) &\leqslant N(\delta, \mathscr{F}'(\beta, K), d_{L^2(\nu)}) \\
&\leqslant N\Big(\delta, \bigoplus_{i=1}^{K} [-\beta, \beta] \odot \mathscr{F}_0, d_{L^2(\nu)}\Big) \leqslant N(\delta/K, [-\beta, \beta] \odot \mathscr{F}_0, d_{L^2(\nu)})^K \\
&\leqslant \left[N\left(\frac{\delta}{2\beta K}, \mathscr{F}_0, d_{L^2(\nu)}\right)\frac{4\beta K}{\delta}\right]^K \leqslant \left[C\left(\frac{\delta}{2\beta K}\right)^{-2(D-1)}\frac{4\beta K}{\delta}\right]^K \\
&= C(D)^K (\beta K)^{K(2D-1)} (1/\delta)^{K(2D-1)}
\end{aligned}$$

independent of ν as in (5.26).

From Theorem 4.1 with $M = \beta_n$ we infer that

$$\begin{aligned}
E\|\hat{\alpha}_n - \alpha_0\|_2^2 &= O\left(\exp(3\beta_n)\left[\left(\frac{1}{K_n}\right)^{1+1/(D-1)} + \frac{K_n}{n}\log\left(C(D)(n\beta_n K_n)^{2(2D-1)}\right)\right]\right) \\
&= O\left(\log n\left[(1/n)^{\frac{1+1/(D-1)}{2+1/(D-1)}} + (1/n)^{1-\frac{1}{2+1/(D-1)}}\log n\right]\right) \\
&= O\left((\log n)^2 (1/n)^{1-\frac{1}{2+1/(D-1)}}\right) = O\left((\log n)^2 (1/n)^{1/2+\frac{1}{4D-2}}\right),
\end{aligned}$$

and the result follows.

The proof of (5.24) is analogous by using the approximation estimate

$$\inf_{\alpha \in \mathscr{F}(\beta_n, K_n)} |\Lambda(\alpha) - \Lambda(\alpha_0)| = O\left((1/K_n)^{1+1/(D-1)}\right). \tag{5.27}$$

For the proof note that for n with $\beta_n \geqslant L := \max\{2|\nu_{\alpha_0}|, \|\alpha_0\|_\infty\}$ it follows by (1.8) that

$$\begin{aligned}
\inf_{\alpha \in \mathscr{F}(\beta_n, K_n)} |\Lambda(\alpha) - \Lambda(\alpha_0)| &\leqslant \inf_{\alpha \in \mathscr{F}(L, K_n)} |\Lambda(\alpha) - \Lambda(\alpha_0)| \\
&\leqslant k'(L) \inf_{\alpha \in \mathscr{F}(L, K_n)} \|\alpha - \alpha_0\|_2^2 = O\left((1/K_n)^{1+1/(D-1)}\right);
\end{aligned}$$

the last estimate is from Döhler and Rüschendorf [9]. ∎

As in Section 5.1, alternative choices of the truncation constants β_n are possible.

We consider special classes of neural nets, radial basis-function nets, and wavelet nets. In the following examples we use some approximation results from Döhler and Rüschendorf [9].

Neural nets. Here

$$\mathscr{F}_0 = \{f_0\colon \mathscr{T}\times\mathscr{X}\to[0,1],\ z\mapsto \Psi(\gamma z+\delta)\mid \gamma\in\mathbf{R}^{k+1},\ \delta\in\mathbf{R}\},$$

where $\Psi\colon \mathbf{R}\to[0,1]$ is of bounded variation. Then the conditions (A1) and (A2) are fulfilled with $D=k+4$ and Theorem 5.4 implies

$$E\,\|\hat{\alpha}_n-\alpha_0\|_2^2 = O\left(\frac{(\log n)^2}{n^{1/2+1/(4k+14)}}\right). \tag{5.28}$$

The same rate holds if the representation property of α_0 is replaced by Barron's [2] finiteness condition on the Fourier transform

$$C_f = \int |w|_1\,|\tilde{f}(w)|\,dw < \infty, \tag{5.29}$$

where $\tilde{f}$ is the Fourier transform of f.

If $P^{(T,X)}$ has a density with bounded support, the convergence rate can be improved to

$$E\,\|\hat{\alpha}_n-\alpha_0\|_2^2 = O\left(\frac{(\log n)^2}{n^{1/2+1/(4k+6)}}\right). \tag{5.30}$$

Similar rates with $1/2$ instead of $1/2+1/(4k+6)$ in the exponent were obtained previously for regression estimation in Barron [3], and for density estimation in Modha and Masry [19].

Radial basis-function nets. Here

$$\mathscr{F}_0 = \{f_0\colon \mathscr{T}\times\mathscr{X}\to[0,1],\ z\mapsto \varrho(\|\gamma(z-\delta)\|)\mid \gamma\in\mathbf{R}^{k+1},\ \delta\in\mathbf{R}\},$$

where $\varrho\colon \mathbf{R}^+\to[0,1]$ is monotonically non-increasing. Then the conditions (A1) and (A2) are fulfilled with $D=k+5$ and from Theorem 5.4 we obtain

$$E\,\|\hat{\alpha}_n-\alpha_0\|_2^2 = O\left(\frac{(\log n)^2}{n^{1/2+1/(4k+18)}}\right). \tag{5.31}$$

Wavelet nets. Here

$$\mathscr{F}_0 = \{f_0\colon \mathscr{T}\times\mathscr{X}\to[0,1],\ z\mapsto \Psi(\gamma(z-\delta))\mid \gamma\in\mathbf{R}^{k+1},\ \delta\in\mathbf{R}\},$$

where $\Psi\colon \mathbf{R}^{d+1}\to[0,1]$ is Lipschitz with bounded support. Then by Theorem 5.4 with $D=3k+4$ we obtain

$$E\,\|\hat{\alpha}_n-\alpha_0\|_2^2 = O\left(\frac{(\log n)^2}{n^{1/2+1/(12k+14)}}\right). \tag{5.32}$$

Note that in all the three cases a corresponding convergence result also holds in terms of the Λ-distance.

RÉSUMÉ

In conclusion, this paper gives quite general results on the convergence rates for sieved minimum contrast estimators and also for the related adaptive versions of these estimators. The results are formulated in detail for the example of estimating the log-hazard function in censoring models. In comparison to the related general approach in Birgé and Massart [5] and Barron et al. [4] we use some simpler conditions concerning the covering numbers. The results in this paper are illustrated with examples of sieves such as neural nets, wavelet nets, radial basis-function nets and tensor product splines. Some further applications of the method in this paper to more general type of censorings as well as to a more detailed study of neural net estimators are given in the forthcoming papers by Döhler and Rüschendorf [8], [10].

REFERENCES

[1] P. K. Andersen, O. Borgan, R. Gill and N. Keiding, *Statistical Models Based on Counting Processes*, Springer, 1993.

[2] A. R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inform. Theory 39 (3) (1993), pp. 930–945.

[3] A. R. Barron, *Approximation and estimation bounds for artificial neural networks*, Machine Learning 14 (1994), pp. 115–133.

[4] A. R. Barron, L. Birgé and P. Massart, *Risk bounds for model selection via penalization*, Probab. Theory Related Fields 113 (3) (1999), pp. 301–413.

[5] L. Birgé and P. Massart, *Minimum contrast estimators on sieves: Exponential bounds and rates of convergence*, Bernoulli 4 (3) (1998), pp. 329–375.

[6] S. Döhler, *Consistent hazard regression estimation by sieved maximum likelihood estimators*, in: *Proceedings of Conference on Limit Theorems in Balatonlelle*, 2000.

[7] S. Döhler, *Empirische Risiko-Minimierung bei zensierten Daten*, Ph. D. Thesis, Universität Freiburg 2000, http://webdoc.sub.gwdg.de/ebook/e/2001/freidok/69.dpf.

[8] S. Döhler and L. Rüschendorf, *A consistency result in general censoring models*, Statistics (2000).

[9] S. Döhler and L. Rüschendorf, *An approximation result for nets in functional estimation*, Statist. Probab. Lett. 52 (2001), pp. 373–380.

[10] S. Döhler and L. Rüschendorf, *On adaptive estimation by neural net type estimators*, in: *Nonlinear Estimation and Classification*, D. Denison, M. Hausen, C. Holmes, B. Mallick and B. Yu (Eds.), Lecture Notes in Statist., Springer 2002.

[11] M. Kohler, *Nichtparametrische Regressionsschätzung mit Splines*, Ph. D. Thesis, Universität Stuttgart, 1997, http://www.mathematik.unistuttgart.de/mathA/lst3/kohler/papers!html.

[12] M. Kohler, *Nonparametric estimation of piecewise smooth regression functions*, Statist. Probab. Lett. 43 (1999), pp. 49–55.

[13] M. Kohler, *Universally consistent regression function estimation using hierarchical B-splines*, J. Multivariate Anal. 68 (1999), pp. 138–164.

[14] C. Kooperberg, C. J. Stone and Y. K. Truong, *Hazard regression*, J. Amer. Statist. Assoc. 90 (1995), pp. 78–94.

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

[16] A. Krzyzak and T. Linder, *Radial basis function networks and computational regularization in function learning*, IEEE Trans. Inform. Theory 9 (1998), pp. 247–256.
[17] W. Lee, P. Bartlett and R. Williamson, *Efficient agnostic learning of neural networks with bounded fan-in*, IEEE Trans. Inform. Theory 42 (6) (1996), pp. 2118–2132.
[18] G. Lugosi and K. Zeger, *Nonparametric estimation via empirical risk minimization*, IEEE Trans. Inform. Theory 41 (3) (1995), pp. 677–687.
[19] D. Modha and E. Masry, *Rate of convergence in density estimation using neural networks*, Neural Computation 8 (1996), pp. 1107–1122.
[20] D. Pollard, *Convergence of Stochastic Processes*, Series in Statistics, Vol. 14, Springer, 1984.
[21] D. Pollard, *Empirical Processes: Theory and Applications*, Institute of Mathematical Statistics, Hayward, 1990.
[22] A. van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes*, Springer, New York 1996.
[23] I. van Keilegom and N. Veraverbeke, *Hazard rate estimation in non-parametric regression with censored data*, Ann. Inst. Statist. Math. 53 (2001), pp. 730–745.
[24] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York 1995.
[25] W. Wong and X. Shen, *Probability inequalities for likelihood ratios and convergence rates of sieve mles*, Ann. Statist. 23 (1995), pp. 339–362.
[26] Y. Yang and A. Barron, *An asymptotic property of model selection criteria*, IEEE Trans. Inform. Theory 44 (1998), pp. 95–116.
[27] J. Yukich, M. Stichcombe and H. White, *Sup-norm approximation bounds for networks through probabilistic methods*, IEEE Trans. Inform. Theory 41 (4) (1995), pp. 1021–1027.

Corresponding author:
Ludger Rüschendorf
University of Freiburg
Institute for Mathematical Stochastics
Eckerstr. 1
79104 Freiburg, Germany

Tel.: +49-761-2035665
Fax.: +49-761-2035661
E-mail: ruschen@stochastik.uni-freiburg.de

Received on 17.1.2002;
revised version on 14.10.2002