Consistent Hazard Regression Estimation by Sieved Maximum Likelihood Estimators

Sebastian Döhler
Institut für Mathematische Stochastik, Universität Freiburg

Abstract

We consider maximum likelihood estimators in general sieve function classes for the estimation of the conditional log-hazard function of a survival time in a censored data model. We prove consistency of these types of estimators under the assumption that the conditional log-hazard function is continuous, by using results from empirical process theory. As special examples we consider feedforward neural network estimators and radial basis function network estimators.

1 Introduction

The aim of this paper is to show that sieved maximum likelihood estimators can be used to consistently estimate the logarithm of a continuous conditional hazard function of the survival time in a censoring model.
We consider random variables (rvs) T, C and X, where

$$\begin{aligned} T &: (\Omega, \mathcal{A}, P) \longrightarrow \mathbb{R}_+ && \text{is a survival (failure) time,} \\ C &: (\Omega, \mathcal{A}, P) \longrightarrow \mathcal{T} := [0,1] && \text{is a censoring time and} \\ X &: (\Omega, \mathcal{A}, P) \longrightarrow \mathcal{X} := [0,1]^k && \text{is a vector of covariates.} \end{aligned}$$

In survival analysis the survival and censoring time are not observed directly, but instead the "observable time"

$$Y := T \wedge C,$$

the "censoring indicator"

$$\delta := 1_{\{T \leq C\}} = \begin{cases} 1 & \text{failure takes place before censoring,} \\ 0 & \text{failure event is censored} \end{cases}$$

as well as the covariates X are observed. This censoring mechanism is called *right-censorship*. We are interested in inference on the conditional distribution of the lifetime

T given the vector of covariates X.
We assume that for the model described above there exists a conditional density function $f_0(t|x)$ and denote by $F_0(t|x)$ the conditional distribution function of T given $X = x$. Then

$$\begin{aligned}
\overline{F_0}(t|x) &:= 1 - F_0(t|x) && \text{is the conditional survival function,} \\
\lambda_0(t|x) &:= \frac{f_0(t|x)}{\overline{F_0}(t|x)} && \text{is the conditional hazard function and} \\
\alpha_0(t|x) &:= \log \lambda_0(t|x) && \text{is the conditional log-hazard function.}
\end{aligned}$$

The conditional hazard function has the following interpretation (cf. p. 127 in [FH]): For small Δt

$$\lambda_0(t|x)\Delta t \quad \approx \quad P(t \leq T < t + \Delta t | T \geq t, x)$$

is the approximate conditional probability of observing a failure in the time-interval $[t, t + \Delta t)$ given x and no failure before time t. Our goal is to find an estimator for λ_0 respectively α_0 based on iid censored data $(y_1, \delta_1, x_1), \ldots, (y_n, \delta_n, x_n)$.
One well-known way to model the conditional hazard function is the Cox (proportional hazard) model. Here it is assumed that the conditional log-hazard function is the sum of an unspecified function of t and a linear function in x

$$\alpha_0(t|x) = \widetilde{\alpha_0}(t) + x \cdot \beta,$$

where $\widetilde{\alpha_0}$ is called the baseline hazard function and $\beta \in \mathbb{R}^k$ is a vector of parameters; often one is interested only in estimating β. This model has the property that $\frac{\lambda_0(t|x_1)}{\lambda_0(t|x_2)}$ is independent of t. Partial-likelihood can be used to estimate the vector β.
The disadvantage of the Cox model is that it does not allow for interactions between the covariates and time. We want to deal with the problem of estimating α_0, where $\alpha_0 = \alpha_0(t, x)$ is known to belong to some (usually large) function space (e.g. $L_2(\mathcal{T} \times \mathcal{X})$). This means that we should use functions as estimators that are known to have some sort of universal approximation property. Since neural networks and radial basis function networks have this property they should be considered as candidates for these estimators.
A different approach by Kooperberg, Stone and Truong in [KST] is based on tensor product splines. Their main result is a L_2 rate of convergence for spline estimators under the assumption that the true conditional log-hazard function is sufficiently smooth. In our approach we dismiss the smoothness condition as well as an additional condition on the censoring distribution, and establish a general consistency result under the assumption that the true conditional log-hazard function is continuous.
Other related results could be expected in the context of density estimation for instance for neural networks. However, as far as we know there are no results available for the case of censored data.
Our approach is also different from these two approaches in the sense that we do not consider a special sieve like splines or neural networks, but a rather general type of

sieve that includes various estimators used in nonparametric statistics. Furthermore, the method established in this paper should also be applicable to further types of censoring as well as to other types of estimators.

We use two assumptions that are similar to two of the three assumptions introduced in [KST]. The first condition is the same as in [KST].

Condition 1
T and C are conditionally independent given X.

The second condition is weaker than the the corresponding condition in [KST]. It could be replaced by an even weaker integrability condition on α_0 but since we assume for our main results that α_0 is bounded we use this assumption.

Condition 2
α_0 is bounded on $\mathcal{T} \times \mathcal{X}$.

Let $(t_1, c_1, x_1), \cdots, (t_n, c_n, x_n)$ be an iid sample from the distribution of (T, C, X). We observe $(y_1, \delta_1, x_1), \ldots, (y_n, \delta_n, x_n)$, where

$$y_i := t_i \wedge c_i$$
$$\delta_i = \begin{cases} 1 & \text{i-th survival time observed,} \\ 0 & \text{i-th survival event censored.} \end{cases}$$

According to [KST] the conditional log-likelihood corresponding to this sample is [1]

$$L_n(\alpha) \;:=\; \sum_{i=1}^{n} [\delta_i \alpha(y_i|x_i) - \int_0^{y_i} \exp \alpha(u|x_i) du]$$

where $\alpha : \mathcal{T} \times \mathcal{X} \longrightarrow \mathbb{R}$. The expected conditional log-likelihood is

$$\Lambda(\alpha) \;:=\; \mathbb{E} L_1(\alpha).$$

Later we will define our estimator as the maximum likelihood estimator over some class of functions (sieve): $\hat{\alpha}_n := \text{argmax}_{\alpha \in \mathcal{F}_n} L_n(\alpha)$ where the sieve $\mathcal{F}_n$ depends on the number n of observations. Of course the crucial question with regard to consistency is how this sequence of function classes should be chosen.
The following lemma states a formula for the "distance" between an arbitrary function and the "true" conditional log-hazard function with respect to the expected conditional log-likelihood.

[1] Here and for the rest of the paper, when we use integrals etc., we always assume them to exist and be well-defined.

Lemma 1
Let $\alpha : \mathcal{T} \times \mathcal{X} \longrightarrow \mathbb{R}$, $G : \mathbb{R} \longrightarrow \mathbb{R}$, $y \mapsto \exp(y) - (1 + y)$. Let the rvs (T, C, X) satisfy conditions 1 and 2. Then we have

$$\begin{aligned} \left|\Lambda(\alpha_0) - \Lambda(\alpha)\right| &= \Lambda(\alpha_0) - \Lambda(\alpha) \\ &= \int_{\mathcal{T}\times\mathcal{X}} \overline{F_C} \cdot G(\alpha - \alpha_0)\ dP^{(T,X)} \end{aligned}$$

where α_0 is the "true" conditional log-hazard function, and $\overline{F_C}$ is the conditional survival function of the censoring time C given X.

For $y \longrightarrow -\infty$ the function $G(y)$ behaves like $-(1 + y)$ and for $y \longrightarrow \infty$ it behaves like an exponential function.

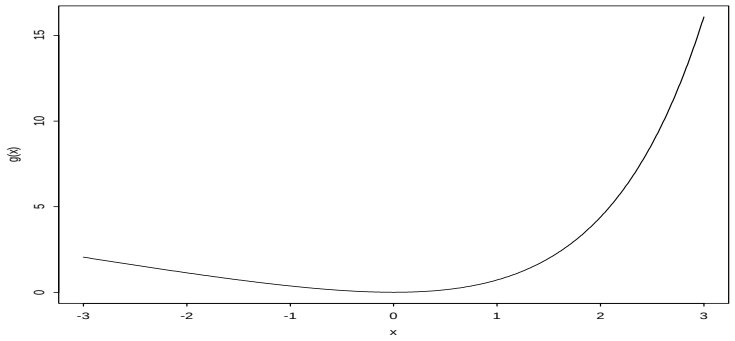

Figure 1: The function G

Proof We use the following representation from [KST]

$$\Lambda(\alpha) = \int_{\mathcal{X}} \int_{\mathcal{T}} \overline{F_C}(t|x) \Big[\alpha(t|x) f_0(t|x) - \overline{F_0}(t|x) \exp(\alpha(t|x))\Big]\ dt\ f_X(x)\ dx,$$

where $\alpha : \mathcal{X} \times \mathcal{T} \longrightarrow \mathbb{R}$ and f_X denotes the density function of X. The claim of the lemma then follows from some algebra and the definition of α_0.

2 Consistency of Sieve Estimators

Here we introduce a type of function class that satisfies some conditions sufficient to prove consistency of a sieved maximum likelihood estimator. We consider vector spaces of functions $\mathcal{F} = \{\sum_i a_i f_i \mid a_i \in \mathbb{R},\ f_i \in \mathcal{F}_0\}$ generated by a set of "basis functions" $\mathcal{F}_0$. The first condition we impose is that $\mathcal{F}$ is dense in a suitable space of functions. This is

a natural requirement, since it guarantees that it is possible to (asymptotically) remove the bias of estimators that are members of this class. The other condition we need is control of the "complexity" or "richness" of the set of basis functions $\mathcal{F}_0$. This will enable us to control the stochastic part of the overall error of the estimator. Examples of such classes are "feedforward neural networks with one hidden layer" and "radial basis function networks".

Condition 3
Let $\mathcal{F} \subset \{f : \mathcal{T} \times \mathcal{X} \longrightarrow \mathbb{R}\}$ be a class of functions with the following properties:

a) *$\mathcal{F}$ is dense in $(C(\mathcal{T} \times \mathcal{X}), \| \cdot \|_\infty)$, where $\| \cdot \|_\infty$ is the sup-norm on $\mathcal{T} \times \mathcal{X}$.*

b) *$\mathcal{F} = \{\sum_i a_i f_i \mid a_i \in \mathbb{R},\ f_i \in \mathcal{F}_0\}$ is a vector-space generated by a Vapnik-Cervonenkis-class $\mathcal{F}_0$ with $f : \mathcal{T} \times \mathcal{X} \longrightarrow [0,1]$ for $f \in \mathcal{F}_0$.*

Definition 4 (Neural networks)
For an increasing function $\sigma : \mathbb{R} \longrightarrow [0,1]$ satisfying $\lim_{x\to\infty} \sigma(x) = 1$ and $\lim_{x\to-\infty} \sigma(x) = 0$ let

$$\mathcal{F} = \mathcal{F}(\sigma) := \Big\{ f : \mathbb{R}^m \longrightarrow \mathbb{R},\ z \longmapsto \sum_{i=1}^{K} c_i \sigma(a_i^T \cdot z + b_i) + c_0 \ \Big|\ K \in \mathbb{N},\ a_i \in \mathbb{R}^m,\ b_i,\ c_i \in \mathbb{R} \Big\}.$$

Definition 5 (Radial basis function networks)
For a decreasing, continuous, nonzero function $h : \mathbb{R}_+ \longrightarrow [0,1]$ with $\int_{\mathbb{R}^m} h(\|x\|)\, d\lambda^m(x) < \infty$ let

$$\mathcal{F} = \mathcal{F}(h) := \Big\{ f : \mathbb{R}^m \longrightarrow \mathbb{R},\ z \longmapsto \sum_{i=1}^{K} c_i h(\|A_i \cdot z + b_i\|) + c_0 \ \Big|\ K \in \mathbb{N},\ A_i \in \mathbb{R}^{m\times m},\ b_i \in \mathbb{R}^m,\ c_i \in \mathbb{R} \Big\}.$$

Remark 6
Both neural networks and radial basis function networks satisfy condition 3 (as functions defined on $[0,1] \times [0,1]^k$).

Proof For the denseness part cf. Theorem 2.3 in [HSW] for the case of neural networks and Theorem 5 in [PS] for the case of radial basis function networks.
The statement about the finiteness of the VC-dimension follows from the fact that in both cases every member of $\mathcal{F}_0$ is the composition of a fixed increasing or decreasing function (in one case σ in the other case $h \circ \sqrt{}$) with a finite-dimensional vector space of functions.

Since finite-dimensional vector spaces of measurable functions are VC-classes (cf. lemma 2.6.15 in [vdV-W]) and since the composition of a fixed monotone function with a VC-class is again a VC-class (cf. lemma 2.6.18(viii) in [vdV-W]) condition b) is also satisfied. □

To establish "universal consistency" for a sieved maximum-likelihood estimator with respect to the Λ-error, we introduce bounds for the number of components and for the weights of elements of $\mathcal{F}$.

Definition 7
Let $\mathcal{F}$ satisfy condition 3. For $\beta > 0, K \in \mathbb{N}$ define

$$\mathcal{F}(\beta, K) := \Big\{ \alpha : \mathcal{T} \times \mathcal{X} \longrightarrow [-K\beta, K\beta],\ (t,x) \longmapsto \sum_{i=1}^{K} c_i f_i(t,x) \ \Big| \ f_i \in \mathcal{F}_0, \mid c_i \mid \leq \beta \Big\}.$$

In the case of neural networks this means that $\mathcal{F}(\beta, K)$ is a set of "feedforward neural networks" with one "hidden layer" consisting of K hidden units and the magnitude of the weights bounded by β.

Theorem 8 (Consistency with respect to the Λ-error)
Let $\mathcal{F}$ be a class of functions that satisfies condition 3. For a subset $\mathcal{F}_n \subset \mathcal{F}$ let $\hat{\alpha}_n$ be the maximum-likelihood estimator $\hat{\alpha}_n := \operatorname{argmax}_{\alpha \in \mathcal{F}_n} L_n(\alpha)$.
Then there are sequences $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ such that $\mathcal{F}_n := \mathcal{F}(\beta_n, K_n) \uparrow \mathcal{F}$ and for all rvs (T, C, X) that satisfy conditions 1 and 2 with $\alpha_0 \in C(\mathcal{T} \times \mathcal{X})$:

$$\Big| \Lambda(\hat{\alpha}_n) - \Lambda(\alpha_0) \Big| \longrightarrow 0 \qquad \text{a.s.}$$

Convergence holds for any $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ such that

$$K_n \exp(3\beta_n K_n) = O(n^{\frac{1}{4}}).$$

From the preceeding result we can easily deduce L^1-consistency of $\hat{\alpha}_n$.

Theorem 9 (Consistency with respect to the L^1-error)
Let the assumptions and terminology of Theorem 8 be given. Then

$$\int_{\mathcal{T} \times \mathcal{X}} \overline{F_C} \cdot |\hat{\alpha}_n - \alpha_0| dP^{(T,X)} \longrightarrow 0 \qquad \text{a.s.}$$

Proof An easy calculation shows that there exist $A, B > 0$ such that

$$\begin{aligned} G(y) &\geq Ay^2 && \text{for} && |y| \leq 1 \\ G(y) &\geq B|y| && \text{for} && |y| > 1. \end{aligned}$$

Let $E_n := \{(t,x) \in \mathcal{T} \times \mathcal{X} \mid |\hat{\alpha}_n(t,x) - \alpha_0(t,x)| \leq 1 \}$. Then we have

$$\int_{\mathcal{T}\times\mathcal{X}} \overline{F_C} \cdot G(\hat{\alpha}_n - \alpha_0)\, dP^{(T,X)} \geq A \int_{E_n} \overline{F_C} \cdot |\hat{\alpha}_n - \alpha_0|^2 dP^{(T,X)} + B \int_{E_n^c} \overline{F_C} \cdot |\hat{\alpha}_n - \alpha_0| dP^{(T,X)}.$$

Since

$$\int_{\mathcal{T}\times\mathcal{X}} \overline{F_C} \cdot G(\hat{\alpha}_n - \alpha_0)\, dP^{(T,X)} = \left| \Lambda(\hat{\alpha}_n) - \Lambda(\alpha_0) \right|$$

on account of Lemma 1, we can conclude with Theorem 8 that

$$\int_{E_n} \overline{F_C} \cdot |\hat{\alpha}_n - \alpha_0|^2 dP^{(T,X)} \longrightarrow 0 \qquad \text{and} \qquad \int_{E_n^c} \overline{F_C} \cdot |\hat{\alpha}_n - \alpha_0| dP^{(T,X)} \longrightarrow 0.$$

Since $\int_{E_n} \overline{F_C}^2 \cdot |\hat{\alpha}_n - \alpha_0|^2 dP^{(T,X)} \leq \int_{E_n} \overline{F_C} \cdot |\hat{\alpha}_n - \alpha_0|^2 dP^{(T,X)}$, application of Jensen's inequality leads to

$$\int_{E_n} \overline{F_C} \cdot |\hat{\alpha}_n - \alpha_0| dP^{(T,X)} \longrightarrow 0$$

which finishes the proof. □

These theorems show that the function classes considered here can be used to consistently estimate the conditional log-hazard function of a survival time. In addition they indicate how the constraints on the number of basis functions and the magnitude of the weights should be chosen to achieve consistency.
The idea of the proof of Theorem 8 is to decompose the overall error into an approximation error and an estimation error. The approximation error can then be removed asymptotically due to the denseness condition imposed on $\mathcal{F}$ and the estimation error is handled by a uniform strong law of large numbers.

Lemma 2 (Decomposition of the Λ-error)
Let $\mathcal{F}$ be a class of functions such that for $\alpha \in \mathcal{F}$ we have $\alpha : \mathcal{T} \times \mathcal{X} \longrightarrow \mathbb{R}$. Let $\hat{\alpha}_n := \operatorname{argmax}_{\alpha\in\mathcal{F}} L_n(\alpha)$. Then

$$\left| \Lambda(\hat{\alpha}_n) - \Lambda(\alpha_0) \right| \quad \leq \quad \inf_{\alpha\in\mathcal{F}} \left| \Lambda(\alpha) - \Lambda(\alpha_0) \right| + 2 \sup_{\alpha\in\mathcal{F}} \left| \frac{1}{n} L_n(\alpha) - \Lambda(\alpha) \right|$$

Proof Let $\alpha^* \in \mathcal{F}$ with $\left|\Lambda(\alpha^*) - \Lambda(\alpha_0)\right| = \inf_{\alpha\in\mathcal{F}} \left|\Lambda(\alpha) - \Lambda(\alpha_0)\right|$. Then

$$\begin{aligned}
\left|\Lambda(\hat{\alpha}_n) - \Lambda(\alpha_0)\right| &= \Lambda(\alpha_0) - \Lambda(\hat{\alpha}_n) && \text{since } \Lambda \text{ is maximal for } \alpha_0 \\
&= \Lambda(\alpha_0) - \Lambda(\alpha^*) + \Lambda(\alpha^*) - \Lambda(\hat{\alpha}_n) \\
&= \inf_{\alpha\in\mathcal{F}} \left|\Lambda(\alpha) - \Lambda(\alpha_0)\right| + \Lambda(\alpha^*) - \frac{1}{n} L_n(\alpha^*) \\
&\quad + \frac{1}{n} L_n(\alpha^*) - \frac{1}{n} L_n(\hat{\alpha}_n) + \frac{1}{n} L_n(\hat{\alpha}_n) - \Lambda(\hat{\alpha}_n) \\
&\leq \inf_{\alpha\in\mathcal{F}} \left|\Lambda(\alpha) - \Lambda(\alpha_0)\right| + 2 \sup_{\alpha\in\mathcal{F}} \left|\frac{1}{n} L_n(\alpha) - \Lambda(\alpha)\right|
\end{aligned}$$

since $\hat{\alpha}_n$ is defined as maximizing L_n. □

2.1 Proof of Theorem 8

By Lemma 2 we have

$$\left|\Lambda(\hat{\alpha}_n) - \Lambda(\alpha_0)\right| \quad \leq \quad \inf_{\alpha\in\mathcal{F}_n} \left|\Lambda(\alpha) - \Lambda(\alpha_0)\right| + 2 \sup_{\alpha\in\mathcal{F}_n} \left|\frac{1}{n} L_n(\alpha) - \Lambda(\alpha)\right|.$$

The second term converges to 0 (a.s.) because of the uniform law of large numbers for the log-likelihood functional, Theorem 10, which is given in the following section. Then it is left to show that the first term also converges to 0. Since $\alpha_0 \in C(\mathcal{T} \times \mathcal{X})$ and $\mathcal{F}$ is dense in $(C(\mathcal{T} \times \mathcal{X}), \| \, . \, \|_\infty)$ there exists, for every $\varepsilon > 0$, an $n \in \mathbb{N}$, $\alpha_n \in \mathcal{F}_n$ such that $\|\alpha_n - \alpha_0\|_\infty < \varepsilon$. With Lemma 1 we conclude that

$$\begin{aligned}
\left|\Lambda(\alpha_n) - \Lambda(\alpha_0)\right| &= \int_{\mathcal{T}\times\mathcal{X}} \overline{F_C} \cdot G(\alpha_n - \alpha_0) \, dP^{(T,X)} \\
&\leq \max(G(\varepsilon), G(-\varepsilon))
\end{aligned}$$

Since $G(x) \longrightarrow 0$ for $x \longrightarrow 0$ this finishes the proof. □

3 A uniform law of large numbers for the log-likelihood functional

The main part of the proof of Theorem 8 is the following uniform law of large numbers for the log-likelihood functional, which is also of interest in itself.

Theorem 10 (LLN for the log-likelihood functional)
Suppose condition 3 holds for $\mathcal{F}$. Then there are sequences $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ such that $\mathcal{F}_n := \mathcal{F}(\beta_n, K_n) \uparrow \mathcal{F}$ and for all rvs (T, C, X) that satisfy conditions 1 and 2:

$$\sup_{\alpha \in \mathcal{F}_n} \left| \frac{1}{n} L_n(\alpha) - \Lambda(\alpha) \right| \quad \longrightarrow \quad 0 \qquad \text{a.s.}$$

Convergence holds for any $\beta_n \uparrow \infty$, $K_n \uparrow \infty$ such that

$$K_n \exp(3\beta_n K_n) \quad = \quad O(n^{\frac{1}{4}}).$$

The idea of the proof of Theorem 10 is to derive an exponential maximal inequality for the likelihood functional over function classes. We accomplish this by using some results from empirical process and Vapnik-Cervonenkis theory that we recall here.

3.1 Some results from empirical process theory

Here we introduce some results and notation we need to derive probability bounds on the maximal deviation of empirical processes indexed by functions. For the first two definitions cf. [vdV-W].

Definition 11 (Covering Numbers, Packing Numbers)
Let (T, d) be a semimetric space. Then the covering number $N(\varepsilon, T, d)$ is defined as the minimal number of balls of radius ε needed to cover T. The packing number $N(\varepsilon, T, d)$ is defined as the maximal number of points in T such that the distance between each pair of points is strictly larger than ε.

Definition 12 (Vapnik-Cervonenkis-class)
Let $\mathcal{C}$ be a collection of subsets of some set $\mathcal{S}$. A set $s \subset \mathcal{S}$ with $s = \{s_1, \ldots, s_n\}$ is shattered by $\mathcal{C}$ if for any $s' \subset s$ there is a $c' \in \mathcal{C}$ such that $s' = s \cap c'$. The VC-index of $\mathcal{C}$ is the smallest $n \in \mathbb{N}$ such that for no set $s \in \mathcal{S}$ with $|s| = n$ s can be shattered by $\mathcal{C}$.
A collection $\mathcal{F}$ of measurable functions on some space $\mathcal{Y}$ is called a VC-class of functions (and the associated VC-index is called the VC-dimension) if the collection of all subgraphs of the functions (that is sets $\{(x,t) : t < f(x)\}$ for $f \in \mathcal{F}$) in $\mathcal{F}$ is a VC-class of sets in $\mathcal{Y} \times \mathbb{R}$.

Lemma 3 (Pollard)
Let $\mathcal{F}$ be a Vapnik-Cervonenkis-class with envelope function F and $d = \dim_{\mathrm{VC}}(\mathcal{F})$. Then there is a constant $C(d) \geq 0$ such that

$$N(\varepsilon \|\mathrm{F}\|_{L^2(\mu)}, \mathcal{F}, d_{L^2(\mu)}) \quad \leq \quad C(d) \varepsilon^{-2(d-1)}$$

for all $0 < \varepsilon \leq 1$ and all probability measures μ.

Proof cf. Theorem 2.6.7. in [vdV-W].

Lemma 4 (Stability properties of covering numbers)
Let μ be a probability measure and let $\mathcal{F}, \mathcal{G}$ function classes on $\mathbb{R}^m$. Then

a.

$$N(\varepsilon+\delta, \mathcal{F} \oplus \mathcal{G}, d_{L^2(\mu)}) \quad \leq \quad N(\varepsilon, \mathcal{F}, d_{L^2(\mu)}) \cdot N(\delta, \mathcal{G}, d_{L^2(\mu)})$$

b. *If $\forall f \in \mathcal{F} : |f| \leq K$ then for $a > 0$*

$$N(\varepsilon, [-a,a] \odot \mathcal{F}, d_{L^2(\mu)}) \quad \leq \quad N(\frac{\varepsilon}{2a}, \mathcal{F}, d_{L^2(\mu)}) \cdot \frac{4aK}{\varepsilon}$$

c. *Let $\forall f \in \mathcal{F} : |f| \leq K$ and $\varphi : [-K, K] \longrightarrow \mathbb{R}$ be a Lipschitz-function with Lipschitz-constant $Lip(\varphi)$, then:*

$$N(\varepsilon, \varphi \circ \mathcal{F}, d_{L^2(\mu)}) \quad \leq \quad N(\frac{\varepsilon}{\mathrm{Lip}(\varphi)}, \mathcal{F}, d_{L^2(\mu)})$$

d. *If $\forall f \in \mathcal{F} : |f| \leq K$ and $\forall g \in \mathcal{G} : |g| \leq L$, then:*

$$N(\varepsilon, \mathcal{F} \odot \mathcal{G}, d_{L^2(\mu)}) \quad \leq \quad N(\frac{\varepsilon}{2L}, \mathcal{F}, d_{L^2(\mu)}) \cdot N(\frac{\varepsilon}{2K}, \mathcal{G}, d_{L^2(\mu)})$$

Proof Similar to the calculations in Section 5 in [P] and Chapter 29 in [DGL]. Next we introduce some terminology from Section 7 in [P].

Definition 13
For a class of functions $\mathcal{F}$ let $\Phi_i : \mathcal{F} \times \Omega \longrightarrow \mathbb{R}$, $i = 1, ..., n$ be independent stochastic processes. Define the

Random set of points indexed by $\mathcal{F}$

$$\mathcal{Z}_n(\mathcal{F})(\omega) \quad := \quad \left\{ \Big(\Phi_1(\alpha, \omega), \ldots, \Phi_n(\alpha, \omega) \Big) \,\middle|\, \alpha \in \mathcal{F} \right\} \subset \mathbb{R}^n$$

and the

Random entropy integral

$$\begin{aligned} J_n(\mathcal{F})(\omega) \quad &:= \quad 9 \int_0^{\delta_n(\omega)} \sqrt{\log D_2(x, \mathcal{Z}_n(\mathcal{F})(\omega))}\, dx \\ \text{where } \delta_n(\omega) \quad &:= \quad \sup\{\, \|y\|_2 \mid y \in \mathcal{Z}_n(\mathcal{F})(\omega) \,\}. \end{aligned}$$

3.2 Maximal Inequality for L_n

Definition 14
For $\beta > 0, K \in \mathbb{N}$ let

$$\mathcal{G}(\beta, K) := \Big\{ g_\alpha : \{0,1\} \times \mathcal{T} \times \mathcal{X} \longrightarrow \mathbb{R},\ (\delta, y, x) \longmapsto \delta\alpha(y,x) - \int_0^y \exp\alpha(u,x)\, du \ \Big| \ \alpha \in \mathcal{F}(\beta, K) \Big\}.$$

and for $g_\alpha \in \mathcal{G}(\beta, K)$

$$\Phi_i(g_\alpha, \omega) := g_\alpha(\delta_i, y_i, x_i).$$

Proposition 15
Let the rvs (T, C, X) satisfy the assumptions 1 and 2. Let $d = \dim_{\mathrm{VC}}(\mathcal{F}_0)$ with $\mathcal{F}_0$ as in Condition 3. Then there exists a constant $C(d) > 0$ such that for all $\beta > 0, K \in \mathbb{N}$ and $t > 0$:

$$\begin{aligned} P\Big(\sup_{\alpha \in \mathcal{F}(\beta,K)} \Big| \frac{1}{n} L_n(\alpha) - \Lambda(\alpha) \Big| \geq t \Big) &= P\Big(\sup_{g \in \mathcal{G}(\beta,K)} \Big| \frac{1}{n} \sum_{i=1}^n g(\delta_i, y_i, x_i) - \mathbb{E}\Big[g(\delta, Y, X) \Big] \Big| \geq t \Big) \\ &\leq 5 \cdot \exp\Big(-\frac{1}{2} \frac{n}{C(d) K (\beta K + \exp(\beta K))^2}\, t^2 \Big) \end{aligned}$$

Proof For $\beta > 0, K \in \mathbb{N}$ let

$$\mathcal{H}(\beta, K) := \Big\{ h_\alpha : \{0,1\} \times \mathcal{T} \times \mathcal{X} \longrightarrow \mathbb{R},\ (\delta, y, x) \longmapsto \delta\alpha(y,x) \ \Big| \ \alpha \in \mathcal{F}(\beta, K) \Big\},$$

$$\mathcal{K}(\beta, K) := \Big\{ k_\alpha : \{0,1\} \times \mathcal{T} \times \mathcal{X} \longrightarrow \mathbb{R},\ (\delta, y, x) \longmapsto -\int_0^y \exp\alpha(u,x)\, du \ \Big| \ \alpha \in \mathcal{F}(\beta, K) \Big\},$$

then we have $\mathcal{G}(\beta, K) \subset \mathcal{H}(\beta, K) \oplus \mathcal{K}(\beta, K)$. Let $\nu_n := \frac{1}{n} \sum_{i=1}^n \delta_{(\delta_i, y_i, x_i)}$ be the empirical measure associated with the data $\{(\delta_1, y_1, x_1), ..., (\delta_n, y_n, x_n)\}$ and let $\tilde{\nu}_n := \frac{1}{n} \sum_{i=1}^n \delta_{(y_i, x_i)}$, $\tilde{\tilde{\nu}}_n := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. By Lemma 4 a.

$$(1) \qquad N(\varepsilon, \mathcal{G}(\beta, K), d_{L^2(\nu_n)}) \leq N\Big(\frac{\varepsilon}{2}, \mathcal{H}(\beta, K), d_{L^2(\nu_n)}\Big) \cdot N\Big(\frac{\varepsilon}{2}, \mathcal{K}(\beta, K), d_{L^2(\nu_n)}\Big).$$

For the covering numbers of $\mathcal{H}$ it is easily seen that

$$(2) \qquad N(\delta, \mathcal{H}(\beta, K), d_{L^2(\nu_n)}) \leq N(\delta, \mathcal{F}(\beta, K), d_{L^2(\tilde{\nu}_n)}).$$

For the covering numbers of $\mathcal{K}$ we show

$$(3) \qquad N(\delta, \mathcal{K}(\beta, K), d_{L^2(\nu_n)}) \leq N(\delta, \exp \circ \mathcal{F}(\beta, K), d_{L^2(\tilde{\tilde{\nu}}_n \otimes U[0,1])}),$$

where $U[0,1]$ is the uniform distribution on $[0,1]$. For $k_{\alpha_1}, k_{\alpha_2} \in \mathcal{K}(\beta, K)$ we have

$$\begin{aligned} d^2_{L^2(\nu_n)}(k_{\alpha_1}, k_{\alpha_2}) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\int_0^{y_i} \exp\alpha_1(u,x_i) - \exp\alpha_2(u,x_i)\,du\Big)^2 \\ &\le \frac{1}{n}\sum_{i=1}^{n}\int_0^{y_i}\Big(\exp\alpha_1(u,x_i) - \exp\alpha_2(u,x_i)\Big)^2\,du \quad \text{by Jensen's inequality} \\ &\le \frac{1}{n}\sum_{i=1}^{n}\int_0^{1}\Big(\exp\alpha_1(u,x_i) - \exp\alpha_2(u,x_i)\Big)^2\,du \\ &= d^2_{L^2(\tilde{\nu}_n\otimes U[0,1])}(\exp\circ\alpha_1, \exp\circ\alpha_2). \end{aligned}$$

Applying Lemma 4 c. to (3) then yields

$$N(\delta, \mathcal{K}(\beta,K), d_{L^2(\nu_n)}) \le N\Big(\frac{\delta}{\exp(\beta K)}, \mathcal{F}(\beta,K), d_{L^2(\tilde{\nu}_n\otimes U[0,1])}\Big)$$

and we conclude with (1) and (2) that

(4)

$$N(\varepsilon, \mathcal{G}(\beta,K), d_{L^2(\nu_n)}) \le N\Big(\frac{\varepsilon}{2}, \mathcal{F}(\beta,K), d_{L^2(\tilde{\nu}_n)}\Big)\cdot N\Big(\frac{\varepsilon}{2\exp\beta K}, \mathcal{F}(\beta,K), d_{L^2(\tilde{\nu}_n\otimes U[0,1])}\Big).$$

Now let ν be any probability measure on $\mathcal{T}\times\mathcal{X}$. Then for all $\delta > 0$

$$\begin{aligned} N(\delta, \mathcal{F}(\beta,K), d_{L^2(\nu)}) &\le N(\delta, \bigoplus_{i=1}^{K}[-\beta,\beta]\odot\mathcal{F}_0, d_{L^2(\nu)}) \\ &\le N\Big(\frac{\delta}{K}, [-\beta,\beta]\odot\mathcal{F}_0, d_{L^2(\nu)}\Big)^K && \text{Lemma 4 a.} \\ &\le \Big[N\Big(\frac{\delta}{2\beta K}, \mathcal{F}_0, d_{L^2(\nu)}\Big)\cdot\frac{4\beta K}{\delta}\Big]^K && \text{Lemma 4 b.} \\ &\le \Big[C(d)\cdot\Big(\frac{\delta}{2\beta K}\Big)^{-2(d-1)}\cdot\frac{4\beta K}{\delta}\Big]^K && \text{Lemma 3} \\ &= C(d)^K(\beta K)^{K(2d-1)}\Big(\frac{1}{\delta}\Big)^{K(2d-1)} && \text{with a new } C(d). \end{aligned}$$

Since

$$\begin{aligned} &D_2(\varepsilon, \mathcal{Z}_n(\mathcal{G}(\beta,K))(\omega)) \\ &\qquad \le N_2\Big(\frac{\varepsilon}{2}, \mathcal{Z}_n(\mathcal{G}(\beta,K))(\omega)\Big) \\ &\qquad = N\Big(\frac{\varepsilon}{2}, \mathcal{G}(\beta,K), \sqrt{n}\cdot d_{L^2(\nu_n)}\Big) \qquad \text{by def. of } N_2 \text{ in [P]} \\ &\qquad \le N\Big(\frac{\varepsilon}{2\sqrt{n}}, \mathcal{G}(\beta,K), d_{L^2(\nu_n)}\Big), \end{aligned}$$

we obtain

$$(5)\ D_2(\varepsilon, \mathcal{Z}_n(\mathcal{G}(\beta,K))(\omega)) \ \leq\ C(d)^K(\beta K)^{2K(2d-1)}\Big(\exp(\beta K)\Big)^{K(2d-1)}\left(\frac{\sqrt{n}}{\varepsilon}\right)^{2K(2d-1)}.$$

Now let $a_n := \sqrt{n}(\beta K + \exp \beta K)$. Then $\delta_n(\omega) \leq a_n$ (where $\delta_n(\omega)$ is defined in Definition 13). So

$$\begin{aligned} J_n(\mathcal{G}(\beta,K))(\omega) &\leq 9\int_0^{a_n}\sqrt{\log D_2(x, \mathcal{Z}_n(\mathcal{G}(\beta,K))(\omega))}\,dx \\ &= 9a_n\cdot\int_0^1\sqrt{\log D_2(a_n y, \mathcal{Z}_n(\mathcal{G}(\beta,K))(\omega))}\,dy \qquad (6) \end{aligned}$$

using the substitution $y := \frac{x}{a_n}$. Estimate (5) yields

$$\begin{aligned} D_2(a_n y, \mathcal{Z}_n(\mathcal{G}(\beta,K))(\omega)) &\leq C(d)^K(\beta K)^{2K(2d-1)}\Big(\exp(\beta K)\Big)^{K(2d-1)}\left(\frac{\sqrt{n}}{a_n y}\right)^{2K(2d-1)} \\ &= C(d)^K\left(\frac{(\beta K)^2\cdot\exp(\frac{1}{2}\beta K)}{\beta K+\exp(\beta K)}\right)^{2K(2d-1)}\left(\frac{1}{y}\right)^{2K(2d-1)} \end{aligned}$$

and, since the function $\mathbb{R}^+ \longrightarrow \mathbb{R}, x \mapsto \frac{x^2\cdot\exp(0.5x)}{x+\exp x}$ is bounded,

$$\leq C(d)^K\left(\frac{1}{y}\right)^{2K(2d-1)},$$

and so

$$\log D_2(a_n y, \mathcal{Z}_n(\mathcal{G}(\beta,K))(\omega)) \leq K\log C(d) + 2K(2d-1)\cdot\log\left(\frac{1}{y}\right).$$

Since $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ for $a,b>0$ this leads to

$$\sqrt{\log D_2(a_n y, \mathcal{Z}_n(\mathcal{G}(\beta,K))(\omega))} \leq \sqrt{K\log C(d)} + \sqrt{2K(2d-1)\cdot\log\left(\frac{1}{y}\right)}.$$

Because

$$\int_0^1\sqrt{\log\frac{1}{y}}\,dy < \infty$$

we can conclude that

$$\int_0^1\sqrt{\log D_2(a_n y, \mathcal{Z}_n(\mathcal{G}(\beta,K))(\omega))}\,dy \leq C(d)\sqrt{K}.$$

Therefore, inequality (6) implies

$$\begin{aligned} J_n(\mathcal{G}(\beta,K))(\omega) &\le 9\sqrt{n}(\beta K + \exp\beta K)C(d)\sqrt{K} \\ &= C(d)\sqrt{n}\sqrt{K}(\beta K + \exp\beta K). \end{aligned}$$

This means that the rv $J_n(\mathcal{G}(\beta,K))$ is bounded by the constant (rv) $C(d)\sqrt{n}\sqrt{K}(\beta K + \exp\beta K)$. Hence we can use equation (7.3) of [P] to obtain the maximal inequality

$$P\Big(\sup_{g\in\mathcal{G}(\beta,K)}\Big|\frac{1}{n}\sum_{i=1}^{n} g(\delta_i,y_i,x_i) - \mathbb{E}\Big[g(\delta,Y,X)\Big]\Big| \ge t\Big) \le 5\cdot\exp\Big(-\frac{1}{2}\frac{n}{(C(d)\sqrt{K}(\beta K+\exp\beta K))^2}\,t^2\Big)$$

(our $\sup_{g\in\mathcal{G}(\beta,K)}\Big|\frac{1}{n}\sum_{i=1}^{n} g(\delta_i,y_i,x_i) - \mathbb{E}\Big[g(\delta,Y,X)\Big]\Big|$ corresponds to $\frac{1}{n}\Delta_n$ in [P])

$$= 5\cdot\exp\Big(-\frac{1}{2}\frac{n}{C(d)K(\beta K+\exp(\beta K))^2}\,t^2\Big).$$

□

We note that the bound (4) in the proof holds true for any function class $\mathcal{F}$ that is uniformly bounded by some constant $M > 0$:

$$N(\varepsilon,\mathcal{G},d_{L^2(\nu_n)}) \le N\Big(\frac{\varepsilon}{2},\mathcal{F},d_{L^2(\tilde{\nu}_n)}\Big)\cdot N\Big(\frac{\varepsilon}{2\exp M},\mathcal{F},d_{L^2(\tilde{\nu}_n\otimes U[0,1])}\Big),$$

where $\mathcal{G} = \{\ g_\alpha \mid \alpha\in\mathcal{F}\ \}$ is defined similarly to Definition 14. Since the behavior of the random entropy integral is determined by the random covering numbers $N(\cdot,\mathcal{G},d_{L^2(\nu_n)})$, the key to proving the maximal inequality (and therefore the LLN for the log-likelihood functional) lies in obtaining suitable bounds for the random covering numbers $N(\cdot,\mathcal{F},d_{L^2(\tilde{\nu}_n)})$ and $N(\cdot,\mathcal{F},d_{L^2(\tilde{\nu}_n\otimes U[0,1])})$. This means a similar result to Proposition 15 (and therefore to Theorem 10) is true for any function class $\mathcal{F}$, for which $N(\varepsilon,\mathcal{F},d_{L^2(\nu)})$ increases sufficiently slowly (for instance polynomially) in $\frac{1}{\varepsilon}$ for any probability measure ν.

3.3 Proof of Theorem 10

We let $K_n\uparrow\infty, \beta_n\uparrow\infty$ so that

$$K_n\exp(3\beta_n K_n) \;=\; O(n^{\frac{1}{4}}).$$

Let $\mathcal{F}_n := \mathcal{F}(\beta_n,K_n)$ and

$$A_n \;:=\; \Big\{\sup_{\alpha\in\mathcal{F}_n}\Big|\frac{1}{n}L_n(\alpha)-\Lambda(\alpha)\Big| \ge \frac{1}{n^{\frac{1}{4}}}\Big\}.$$

From Proposition 15 we get for large $n\in\mathbb{N}$

$$P(A_n) \le 5\cdot\exp\Big(-Cn^{\frac{1}{4}}\Big).$$

Since for large $n \in \mathbb{N}$

$$\exp\Big(- Cn^{\frac{1}{4}}\Big) \leq \frac{1}{n^2},$$

we have

$$\sum_{n\in\mathbb{N}} P(A_n) < \infty$$

and we apply the Lemma of Borel-Cantelli to conclude the proof. □

Acknowledgement I thank Prof. Dr. Ludger Rüschendorf for a very careful reading of the manuscript and for numerous suggestions that greatly improved the presentation of the paper.

References

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

[FH] Fleming, T.R., Harrington, D.P., (1991), *Counting Processes and Survival Analysis*, Wiley & Sons, New York.

[HSW] Hornik, K., Stinchcombe, M., White, H., (1989), *Multilayer Feedforward Networks are Universal Approximators*, Neural Networks, **2**, 359-366.

[KST] Kooperberg, C., Stone, C.J. und Truong, Y.K., (1995), *The L_2 Rate of Convergence for Hazard Regression*, Scandinavian Journal of Statistics, **22**, 143-157.

[PS] Park, J., Sandberg, I. (1993), *Approximation and Radial-Basis-function Networks*, Neural Computation, **5**, 305-316.

[P] Pollard, D., (1990), *Empirical Processes: Theory and Applications*, Institute of Mathematical Statistics.

[vdV-W] van der Vaart, A., Wellner, J.A. (1996), *Weak convergence and empirical processes*, Springer.