The Annals of Statistics
1995, Vol. 23, No. 5, 1779–1801

EXPONENTIAL INEQUALITIES FOR MARTINGALES, WITH APPLICATION TO MAXIMUM LIKELIHOOD ESTIMATION FOR COUNTING PROCESSES

BY SARA VAN DE GEER

University of Leiden

We obtain an exponential probability inequality for martingales and a uniform probability inequality for the process $\int g\,dN$, where N is a counting process and where g varies within a class of predictable functions $\mathscr{G}$. For the latter, we use techniques from empirical process theory. The uniform inequality is shown to hold under certain entropy conditions on $\mathscr{G}$. As an application, we consider rates of convergence for (nonparametric) maximum likelihood estimators for counting processes. A similar result for discrete time observations is also presented.

1. Introduction. There are many results in the literature on empirical processes indexed by functions. These include uniform central limit theorems, asymptotic equicontinuity and uniform probability inequalities [see, e.g., Ossiander (1987), Pollard (1990) and the references therein]. Statistical applications are, for example, the derivation of the asymptotic distribution of an estimator [Pollard (1989), Kim and Pollard (1990)] or a rate of convergence [van de Geer (1990, 1993a, b, 1995), Birgé and Massart (1993) and Wong and Shen (1995)].

One could say that Bernstein's (or Hoeffding's) inequality lies at the root of these results. We present some extensions of Bernstein's exponential probability inequality to general martingales. Given such extensions, one can think of proceeding to results analogous to those in empirical process theory and applications, but now for dependent observations. We shall not attempt to provide an exhaustive treatment of this idea here, but restrict ourselves mainly to counting processes.

The paper is organized as follows. In Section 2, we review some exponential inequalities for martingales and also present a new one. Section 3 contains a uniform inequality for the counting process. This will be applied in Section 4 to obtain a rate of convergence in Hellinger distance for the maximum likelihood estimator. Some examples on censored observations are given in Section 5. In Section 6, we state without proof a similar result for the case of discrete time observations. Section 7 contains concluding remarks. The proofs of the two main results (Lemma 2.2 and Theorem 3.1) are given in the Appendix.

Received March 1993; revised September 1994.

AMS 1991 *subject classifications*. Primary 60E15; secondary 62G05.

Key words and phrases. Counting process, entropy, exponential inequality, Hellinger process, martingale, maximum likelihood, rate of convergence.

2. Martingale inequalities.

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability triple and let $\{M_t\}_{t \geq 0}$ be a locally square integrable martingale w.r.t. the filtration $\{\mathscr{F}_t\}_{t \geq 0}$. Throughout, we take $M_0 = 0$. We assume that $\{\mathscr{F}_t\}$ satisfies the *usual conditions* [see, e.g., Lipster and Shiryayev (1989)]. Denote the predictable variation of $\{M_t\}$ by $V_t = \langle M, M \rangle_t$, $t \geq 0$, and its jumps by $\Delta M_t = M_t - M_{t-}$, $t > 0$.

The following inequality can be found in Shorack and Wellner (1986).

LEMMA 2.1. *Suppose that* $|\Delta M_t| \leq K$ *for all* $t > 0$ *and some* $0 \leq K < \infty$. *Then for each* $a > 0$, $b > 0$,

$$\mathbb{P}\big(M_t \geq a \text{ and } V_t \leq b^2 \text{ for some } t\big) \leq \exp\left[-\frac{a^2}{2(aK + b^2)}\right]. \tag{2.1}$$

Note that if $\{M_t\}$ is continuous, one can take $K = 0$ in (2.1).

Recall now that Bernstein's inequality for independent observations, as stated in Shorack and Wellner (1986), only imposes certain moment conditions. Indeed, the assumption $|\Delta M_t| \leq K$ for all t in Lemma 2.1 above can be relaxed. For this purpose, we introduce the higher order variation process $\{\Sigma_{s \leq t} |\Delta M_s|^m\}$, with compensator $\{V_{m,t}\}$, $m \in \{3, 4, \dots\}$. Moreover, we write $V_{2,t} = V_t$, $t \geq 0$.

LEMMA 2.2. *Suppose that, for all* $t \geq 0$ *and some* $0 < K < \infty$,

$$V_{m,t} \leq \frac{m!}{2} K^{m-2} R_t, \qquad m = 2, 3, \dots, \tag{2.2}$$

where $\{R_t\}$ *is a predictable process. Then for each* $a > 0$, $b > 0$,

$$\mathbb{P}\big(M_t \geq a \text{ and } R_t \leq b^2 \text{ for some } t\big) < \exp\left[-\frac{a^2}{2(aK + b^2)}\right]. \tag{2.3}$$

For the proof, see the Appendix.

In subsequent sections, we shall encounter martingales of the form $\{\int_0^t g\, dM\}$, with $\{g_t\}$ a predictable process bounded from below, and $\{M_t\}$ as before a martingale with variation processes $\{V_{m,t}\}$. Observe now that if $g_t \geq -L$, then

$$|g_t|^m \leq \frac{m!}{2} c_L^2 \frac{1}{2} (e^{g_t} - 1)^2, \qquad m \in \{2, 3, \dots\},$$

where

$$c_L^2 = \frac{4(e^L - 1 - L)}{(e^{-L} - 1)^2} \tag{2.4}$$

[see Wong and Shen (1995)]. If, moreover, $|\Delta M_t| \le K$ for all $t > 0$, then

$$\int_0^t |g|^m \, dV_m \le K^{m-2} \int_0^t |g|^m \, dV$$
$$\le \frac{m!}{2} K^{m-2} c_L^2 \frac{1}{2} \int_0^t (e^g - 1)^2 \, dV, \qquad m \in \{2, 3, \dots\}.$$

Let us write

$$d_t^2(g, 0) = \tfrac{1}{2} \int_0^t (e^g - 1)^2 \, dV, \qquad t > 0.$$

COROLLARY 2.3. *Suppose $|\Delta M_t| \le K$ for all $t > 0$ and some $0 < K < \infty$. Let $\{g_t\}$ be a predictable process satisfying $g_t \ge -L$ for all $t \ge 0$. Then for each $a > 0$, $b > 0$,*

$$\mathbb{P}\left(\left| \int_0^t g \, dM \right| \ge a \textit{ and } d_t^2(g, 0) \le b^2 \textit{ for some } t \right) \le 2 \exp\left[-\frac{a^2}{2(aK + c_L^2 b^2)} \right], \tag{2.5}$$

with c_L^2 given in (2.4).

3. A uniform inequality for counting processes.

Consider a counting process $\{N_t\}_{t \ge 0}$ with compensator $\{A_t\}_{t \ge 0}$. We assume $\{A_t\}$ to be continuous. Let $0 < T \le \infty$ be a fixed time and let $\{g_t\}$ be a predictable process with $g_t \ge -L$ for all t. By Corollary 2.3,

$$\mathbb{P}\left(\left| \int_0^T g d(N - A) \right| \ge a \text{ and } d_T^2(g, 0) \le b^2 \right) \le 2 \exp\left[-\frac{a^2}{2(a + c_L^2 b^2)} \right].$$

We shall extend this to a uniform inequality, where $\{g_t\}$ is allowed to vary within a class $\mathscr{G} \subseteq \Lambda$, with Λ the class of all predictable processes $\{g_t\}$ with $g_t \ge -L$ for all t.

As in empirical process theory, we shall need entropy conditions on $\mathscr{G}$ in order to make this possible. The metric we shall use here is

$$d_T(g, \tilde{g}) = \left(\tfrac{1}{2} \int_0^T (e^g - e^{\tilde{g}})^2 \, dA \right)^{1/2} \tag{3.1}$$

DEFINITION 3.1 (Entropy with bracketing). Given $b > 0$, $\delta > 0$ and a measurable set $B \subset \Omega$, let $\{[g_j^L, g_j^U]\}_{j=1}^m \subset \Lambda \times \Lambda$ be a collection of pairs of predictable functions, such that for each $g \in \mathscr{G}$ there exists a $j = j(g) \in \{1, \dots, m\}$ such that:

(i) the map $g \mapsto j(g)$ is nonrandom;
(ii) $g_{jt}^L \le g \le g_{jt}^U$ for all t and $\omega \in B$;
(iii) $d_T(g_j^L, g_j^U) \le \delta$ on $\{d_T(g, 0) \le b\} \cap B$.

Let $N(\delta, b, B)$ be the smallest value of m for which such a bracketing set $\{[g_j^L, g_j^U]\}_{j=1}^m$ exists and let $H(\delta, b, B) \geq \log N(\delta, b, B) \vee 1$ be a continuous majorant in $\delta > 0$. We call $H(\delta, b, B)$ an upper bound for δ-entropy with bracketing of $\mathscr{G}$, locally at a ball with radius b around the origin, on the set B. Usually we shall refer to $H(\delta, b, B)$ as the *entropy* of $\mathscr{G}$.

REMARK 3.1. Condition (i) can sometimes be taken care of by choosing an appropriate parametrization [see, e.g., (5.12) in the example for Case (ii)].

REMARK 3.2. If no finite bracketing set in the sense of Definition 3.1 exists, we take $H(\delta, b, B) = \infty$. It can in general be quite difficult (if at all possible) to calculate entropy. However, in (e.g.) the example of Section 5, one can use minor modifications of entropy results for spaces with nonrandom metric. For convenience, we do not require that $H(\delta, b, B)$ is the best possible bound satisfying the conditions of Definition 3.1.

REMARK 3.3. The set B in the definition is primarily introduced to make it possible to restrict oneself to $\omega \in \Omega$ for which $A_T \leq \sigma_T^2$, with σ_T^2 playing the same role as the number of observations n in discrete time.

THEOREM 3.1. *Let $\mathscr{G}$ be a class of predictable functions with $g_t \geq -L$ for all $t \geq 0$, $g \in \mathscr{G}$, and let $H(\delta, b, B)$ be entropy of $\mathscr{G}$, where $B \subset \{A_T \leq \sigma_T^2\}$. There exist constants C_1, C_2, C_3, C_4, depending on L, such that for $0 \leq \varepsilon \leq 1$ and*

$$\frac{\varepsilon b^2}{C_1} \geq \int_{\varepsilon b^2/(C_2\sigma_T)\wedge b/8}^{b} \sqrt{H(x, b, B)}\, dx \vee b, \tag{3.2}$$

we have

$$\begin{aligned} &\mathbb{P}\left(\left\{\left|\int_0^T g\,d(N - A)\right| \geq \varepsilon b^2 \text{ and } d_T^2(g, 0) \leq b^2 \text{ for some } g \in \mathscr{G}\right\} \cap B\right) \\ &\qquad \leq C_3 \exp\left[-\frac{\varepsilon^2 b^2}{C_4}\right]. \end{aligned} \tag{3.3}$$

For the proof, see the Appendix. The proof uses a chaining argument with adaptive truncation, of exactly the same nature as the one used in empirical process theory [see, e.g., Bass (1985), Ossiander (1987) and Andersen, Giné, Ossiander and Zinn (1988)]. It relies heavily on Corollary 2.3, and—for the truncated processes—on Lemma 2.1 (which for $K > 0$ is a special case of Corollary 2.3). For the case of independent and uniformly bounded random variables, see also Birgé and Massart (1993).

4. Maximum likelihood for counting processes.

As in the previous section, $\{N_t\}$ is a counting process with continuous compensator $\{A_t\}$. Assume now that

$$a_0 = \frac{dA}{d\mu} \in \mathscr{A},$$

where μ is a given (possibly random) dominating measure, and where $\mathscr{A}$ is a given set of intensities with respect to μ. We consider the case where the log-likelihood ratio for the process $\{N_t\}$, observed up to time T, is equal to

$$L_T(a, a_0) = \int_0^T \log\left(\frac{a}{a_0}\right) dN - \int_0^T (a - a_0)\, d\mu, \qquad a \in \mathscr{A}.$$

A maximum likelihood estimator $\hat{a}$ is given by

$$L_T(\hat{a}, a_0) = \max_{a \in \mathscr{A}} L_T(a, a_0).$$

Throughout, we shall assume that such an estimator $\hat{a}$ exists, but we do not require it to be unique.

The Hellinger process is

$$h_T^2(a, \tilde{a}) = \tfrac{1}{2}\int_0^T \left(\sqrt{a} - \sqrt{\tilde{a}}\right)^2 d\mu.$$

Observe now that for

$$g_{a_i} = \frac{1}{2}\log\left(\frac{a_i}{a_0}\right), \qquad i = 1, 2,$$

we have the equality

$$d_T^2(g_{a_1}, g_{a_2}) = h_T^2(a_1, a_2). \tag{4.1}$$

It is moreover easy to see that (4.1) implies

$$\frac{1}{2}\int_0^T \log\left(\frac{\hat{a}}{a_0}\right) d(N - A) - h_T^2(\hat{a}, a_0) \geq 0 \tag{4.2}$$

(see also the proof of Lemma 4.1 below). Thus, the theory of the previous section seems well suited to establish a rate of convergence for $h_T(\hat{a}, a_0)$. The only problem is that, in general, the log-ratio $\log(a/a_0)$, $a \in \mathscr{A}$, is not bounded from below. However, there is a simple way to overcome this by considering the log-ratio $\log(\frac{1}{2}(a + a_0)/a_0)$, $a \in \mathscr{A}$, instead. Two lemmas will show that this approach works.

The first lemma says that for the convex combination $\frac{1}{2}(\hat{a} + a_0)$ the inequality (4.2) is also valid.

LEMMA 4.1. *For* $\bar{g}_{\hat{a}} = \frac{1}{2}\log(\frac{1}{2}(\hat{a} + a_0)$, *we have*

$$\int_0^T \bar{g}_{\hat{a}}\, d(N - A) - h_T^2\left(\tfrac{1}{2}(\hat{a} + a_0) . a_0\right) \geq 0. \tag{4.3}$$

PROOF. Clearly, $L_T(\hat{a}, a_0) \geq 0$. The concavity of the log function yields that also

$$\begin{aligned}
L_T\left(\frac{1}{2}(\hat{a} + a_0), a_0\right) &= \int_0^T \log\left(\frac{\frac{1}{2}(\hat{a} + a_0)}{a_0}\right) dN \\
&\quad - \int_0^T \left(\frac{1}{2}(\hat{a} + a_0) - a_0\right) d\mu \\
&\geq \frac{1}{2}\int_0^T \log\left(\frac{\hat{a}}{a_0}\right) dN - \frac{1}{2}\int_0^T (\hat{a} - a_0)\, d\mu \\
&= \frac{1}{2} L_T(\hat{a}, a_0) \geq 0.
\end{aligned} \tag{4.4}$$

Now, for any $a \geq 0$, and for $g_a = \frac{1}{2}\log(a/a_0)$,

$$\tfrac{1}{2}L_T(a, a_0) = \int_0^T g_a\, d(N - A) + \int_0^T g_a\, dA - \tfrac{1}{2}\int_0^T (a - a_0)\, d\mu,$$

and

$$\begin{aligned}
\int_0^T g_a\, dA - \frac{1}{2}\int_0^T (a - a_0)\, d\mu &= \int_0^T \log\left(\sqrt{\frac{a}{a_0}}\right) dA - \frac{1}{2}\int_0^T (a - a_0)\, d\mu \\
&\leq \int_0^T \left(\sqrt{\frac{a}{a_0}} - 1\right) dA - \frac{1}{2}\int_0^T (a - a_0)\, d\mu \\
&= \int_0^T \sqrt{a a_0}\, d\mu - \frac{1}{2}\int_0^T a\, d\mu - \frac{1}{2}\int_0^T a_0\, d\mu \\
&= -h_T^2(a, a_0).
\end{aligned}$$

Thus

$$\tfrac{1}{2}L_T(a, a_0) \leq \int_0^T g_a\, d(N - A) - h_T^2(a, a_0).$$

This inequality with $a = \frac{1}{2}(\hat{a} + a_0)$ combined with inequality (4.4) completes the proof. □

Note that inequality (4.3) involves the Hellinger process evaluated at the convex combination $\frac{1}{2}(\hat{a} + a_0)$, instead of at $\hat{a}$. However, such Hellinger processes behave in an equivalent way, in the sense of the next lemma.

LEMMA 4.2. *For any nonnegative a,*

$$2h_T^2\left(\tfrac{1}{2}(a + a_0), a_0\right) \leq h_T^2(a, a_0) \leq 16 h_T^2\left(\tfrac{1}{2}(a + a_0), a_0\right).$$

For the proof, see van de Geer (1993a) and for better constants Birgé and Massart (1993).

It was already pointed out by Birgé and Massart (1995) that, in the context of density estimation for i.i.d. observations, the concavity of the log function ensures inequality (4.3), and that this combined with Lemma 4.2 shows that one has in a sense not to be afraid of likelihood ratios that are not bounded away from zero. In van de Geer (1993a), also a device with convex combinations was employed, but here the purpose was to construct bounded likelihood ratios. This device can be useful here too, but to avoid digressions we shall not present it here. [This means that in Section 5, the example for Case (ii), some details are left to the reader.]

Let us now return to uniform probability inequalities. We specify Definition 3.1 for a special choice of $\mathscr{G}$. Namely, in what follows, $H(\delta, b, B)$ is defined as entropy of the class $\mathscr{G} = \{\frac{1}{2}\log(\frac{1}{2}(a + a_0)/a_0)\colon a \in \mathscr{A}\}$. Because (4.1) holds true, $H(\delta, b, B)$ can be seen as entropy of $\mathscr{A}$ for the Hellinger metric.

THEOREM 4.3. *Let*

$$B \subset \{A_T \le \sigma_T^2\}. \tag{4.5}$$

There exist universal constants C_1, C_2, C_3, C_4 such that under the condition that $\phi(b)/b$ is nonincreasing, with

$$\phi(b) = \int_{b^2/(C_2\sigma_T)\wedge b/8}^{b} \sqrt{H(x, b, T)}\, dx \vee b,$$

we have, for

$$\frac{b_*^2}{C_1} \ge \phi(b_*), \tag{4.6}$$

that

$$\mathbb{P}(h_T(\hat{a}, a_0) > b_*) > C_4 \exp\left(-\frac{b_*^2}{C_3}\right) + \mathbb{P}(B^c). \tag{4.7}$$

PROOF. From Lemma 4.1 and 4.2, we see that it suffices to prove that

$$\mathbb{P}\left(\int_0^T \bar{g}_{\hat{a}}\, d(N - A) \ge h_T^2\left(\frac{1}{2}(\hat{a} + a_0), a_0\right) \text{ and } h_T\left(\frac{1}{2}(\hat{a} + a_0), a_0\right) > \frac{b_*}{4}\right)$$
$$\le C_4 \exp\left(-\frac{b_*^2}{C_3}\right) + \mathbb{P}(B^c).$$

Write $\bar{a} = \frac{1}{2}(a + a_0)$ and $\bar{g}_a = \frac{1}{2}\log(\bar{a}/a_0) = g_{\bar{a}}$, $a \in \mathscr{A}$. Now, the argument is as in, for example, Alexander (1985). We have

$$\mathbb{P}\left(\int_0^T \bar{g}_a d(N-A) \geq h_T^2(\bar{a}, a_0) \text{ and } h_T(\bar{a}, a_0) > \frac{b_*}{4} \text{ for some } a \in \mathscr{A}\right)$$

$$\begin{aligned}&\leq \sum_{j=1}^{\infty} \mathbb{P}\left(\left\{\int_0^T \bar{g}_a d(N-A) \geq \left(2^{j-1}\frac{b_*}{4}\right)^2 \text{ and}\right.\right. \\ &\qquad\left.\left. h_T(\bar{a}, a_0) \leq 2^j \frac{b_*}{4} \text{ for some } a \in \mathscr{A}\right\} \cap B\right) + \mathbb{P}(B^c) \qquad (4.8)\\ &= \sum_{j=1}^{\infty} \mathbb{P}_j + \mathbb{P}(B^c).\end{aligned}$$

Because $\phi(b)/b$ is assumed to be nonincreasing and (4.6) holds, the inequality

$$\frac{(2^j b_*)^2}{C_1} \geq \phi(2^j b_*)$$

is valid. Therefore, we may apply Theorem 3.1 to each $\mathbb{P}_j$. The result inserted in (4.8) leads to inequality (4.7) by choosing C_4 and C_3 appropriately. □

5. An example: censored observations.

Let $\tilde{X}_1, \dots, \tilde{X}_n$ be i.i.d. failure times and $U_1, \dots, U_n$ be censoring times. We only observe (X_i, Δ_i) for $i = 1, \dots, n$, where $X_i = \min(\tilde{X}_i, U_i)$ and $\Delta_i = 1\{\tilde{X}_i \leq U_i\}$, $i = 1, \dots, n$. Consider the counting processes

$$N_{it} = 1\{X_i \leq t, \Delta_i = 1\}, \qquad i = 1, \dots, n.$$

The number of observed failures at time t is

$$N_t = \sum_{i=1}^{n} N_{it}$$

and the number of individuals at risk immediately before time t is

$$R_n(t-) = \sum_{i=1}^{n} 1\{X_i \geq t\}.$$

Let A_{it} be the compensator of N_{it}, $i = 1, \dots, n$. We assume a multiplicative intensity model, with

$$\frac{dA_{it}}{dt} = \beta_{0t} 1\{X_i \geq t\}, \qquad t \geq 0,\ i = 1, \dots, n, \tag{5.1}$$

where $\beta_0 \in \mathscr{B}$ and $\mathscr{B}$ is a given class of hazard rates. In fact, we shall assume that the log-likelihood ratio $L_\infty(\beta, \beta_0)$ is

$$L_\infty(\beta, \beta_0) = \int_0^\infty \log\left(\frac{\beta}{\beta_0}\right) dN - \int_0^\infty (\beta_t - \beta_{0t}) R_n(t-)\, dt. \tag{5.2}$$

In Jacobsen (1989), one can find conditions such that, given the structure (5.1), (5.2) is indeed valid. See also Gill (1980) for several examples.

The maximum likelihood estimator $\hat{\beta}$ is defined as the maximizer over all $\beta \in \mathscr{B}$ of $L_\infty(\beta, \beta_0)$. We shall establish a rate of convergence for the Hellinger process $h_\infty(\hat{\beta}, \beta_0)$, where

$$h_\infty^2(\beta, \beta_0) = \tfrac{1}{2}\int_0^\infty \left(\sqrt{\beta_t} - \sqrt{\beta_{0t}}\right)^2 R_n(t-)\,dt.$$

We renormalize this to

$$\bar{h}^2(\beta, \beta_0) = (1/n)h_\infty^2(\beta, \beta_0).$$

We shall assume throughout that β_0 is bounded, say $\beta_0 \le 1$, and that

$$(5.3) \qquad \rho_{0n} = \frac{1}{n}\int_0^\infty R_n(t-)\,dt \to_{\mathbb{P}} \rho_0, \qquad 0 < \rho_0 < \infty.$$

Then for $\{A_t\} = \{\Sigma_{i=1}^n A_{it}\}$ being the compensator of $\{N_t\}$, we have

$$(5.4) \qquad A_\infty = \int_0^\infty \beta_{0t} R_n(t-)\,dt \le n\rho_{0n},$$

so that

$$(5.5) \qquad \mathbb{P}(A_\infty > 2n\rho_0) \to 0.$$

In Cases (i)–(iv) below, we shall take

$$(5.6) \qquad B = \left\{\tfrac{1}{2}\rho_0 \le \rho_{0n} \le 2\rho_0\right\}.$$

Condition (4.5) of Theorem 4.3 is then met, with $\sigma_\infty^2 = 2n\rho_0$, that is,

$$B \subset \{A_\infty \le 2n\rho_0\}.$$

In Case (v), we assume in addition that

$$(5.7) \quad \rho_{in} = \frac{1}{n}\int_0^\infty t^i R_n(t-)\,dt \to_{\mathbb{P}} \rho_i, \qquad 0 < \rho_i < \infty,\ i = 1,\dots,m-1,$$

and reduce B to

$$(5.8) \qquad B = \left\{\tfrac{1}{2}\rho_i \le \rho_{in} \le 2\rho_i,\ i = 0,\dots,m-1\right\}.$$

Furthermore, we consider in Case (v) the case where $X_1,\dots,X_n$ have common bounded support, say,

$$(5.9) \qquad X_i \in [0,1], \qquad i = 1,\dots,n.$$

Here are the convergence results for the particular Cases (i)–(v).

CASE (i). $\mathscr{B} = \{\beta \equiv \text{const.}\}$; $\bar{h}(\hat{\beta}, \beta_0) = O_{\mathbb{P}}(n^{-1/2})$.

CASE (ii). $\mathscr{B} = \{\beta \text{ is increasing}\}$; $\bar{h}(\hat{\beta}, \beta_0) = O_{\mathbb{P}}(n^{-1/3})$.

CASE (iii). $\mathscr{B} = \{\beta \text{ is unimodal}, \beta \le C\}$; $\bar{h}(\hat{\beta}, \beta_0) = O_{\mathbb{P}}(n^{-1/3})$.

CASE (iv). $\mathscr{B} = \{\beta = \alpha^2,\ \alpha \text{ is of variation bounded by } C\}$; $\bar{h}(\hat{\beta}, \beta_0) = O_{\mathbb{P}}(n^{-1/3})$.

CASE (v). $\mathscr{B} = \{\beta = \alpha^2,\ \int|\alpha^{(m)}|^2 \leq C\}$; $\bar{h}(\hat{\beta}, \beta_0) = O_{\mathbb{P}}(n^{-m/(2m+1)})$.

The study of computational aspects (including the existence of the maximum likelihood estimator) is not within the scope of this paper. However, $L_\infty(\beta, \beta_0)$ is of similar form as the log-likelihood ratio in the i.i.d. situation, so that the problem is closely related to density estimation. Now, a genuine estimator of a density should integrate to 1. The counterpart of this restriction for hazard rules is

$$\int_0^\infty \hat{\beta}_t R_n(t-)\, dt = N_\infty. \tag{5.10}$$

Equality (5.10) will hold if

$$r\beta \in \mathscr{B} \quad \text{for all } r > 0,\ \beta \in \mathscr{B}. \tag{5.11}$$

In other words, if (5.11) is true, computational results can often be derived along the lines of those in the literature on density estimation. Observe that (5.11) is true in Cases (i) and (ii), but not in Cases (iii)–(v). In Cases (iv) and (v), however, one may also choose to restrict the log-hazard rates, instead of the square root of the hazard rates.

Let us briefly discuss Cases (i)–(v). To avoid a repetition of entropy calculations [see van de Geer (1990, 1991, 1993a, b)], we only present the basic ideas.

Case (i). The maximum likelihood estimator is $\hat{\beta} = N_\infty/(n\rho_{0n})$. Of course, such an explicit expression can be exploited to establish the rate. However, we want to illustrate here that our general approach yields the usual rate in regular finite-dimensional models.

We only have to do the calculations on the set $\{h_\infty(\beta, \beta_0) \leq b\} \cap B$ [see (iii) in Definition 3.1]. On B,

$$\begin{aligned}\tfrac{1}{4}n\rho_0\left(\sqrt{\beta_1} - \sqrt{\beta_2}\right)^2 &\leq \tfrac{1}{2}n\rho_{0n}\left(\sqrt{\beta_1} - \sqrt{\beta_2}\right)^2\\ &= h_\infty^2(\beta_1, \beta_2) \leq n\rho_0\left(\sqrt{\beta_1} - \sqrt{\beta_2}\right)^2\end{aligned}$$

Using the fact that an interval of length b can be covered by const.(b/δ) intervals of length δ, one sees that

$$H(\delta, b, B) \leq \text{const.} \log(b/\delta).$$

Thus, $\phi(b) \leq \text{const.}\, b$, so that one can take $b_* = \text{const.}$, that is, $h_\infty(\hat{\beta}, \beta_0) = O_{\mathbb{P}}(1)$.

Case (ii). The estimator $\hat{\beta}$ is closely related to the Grenander estimator of a density [see, e.g., Groeneboom (1985)]. An explicit expression for $\hat{\beta}$ can be found in Huang and Wellner (1995), who show that, under certain conditions, $\hat{\beta}$ is asymptotically equivalent to the estimator obtained by differentiating the greatest convex minorant of the Nelson–Aalen estimator.

It suffices to calculate the entropy of $\mathscr{B}_1 = \{\beta \in \mathscr{B},\ \beta \leq 1\}$ (say). This follows from the same arguments as in van de Geer (1993a, b), replacing β_0 (instead of β) by the convex combination $\frac{1}{2}(\beta + \beta_0)$.

It is known that the set $\{g: [0,1] \to [0,1],\ g \text{ increasing}\}$ has δ-entropy with bracketing w.r.t. the L_2 (Lebesgue measure) norm of order $1/\delta$ [see Birman and Solomjak (1967) and van de Geer (1991)]. We are now facing the problem of calculating entropy of $\mathscr{B}_1 = \{\beta\colon [0,\infty) \to [0,1],\ \beta \text{ increasing}\}$ w.r.t. the (random) $L_2(nQ_n)$ norm, where

$$Q_n(t) = \frac{1}{n}\int_0^t R_n(t-)\,dt.$$

The following step may be thought of as a reparametrization (see Remark 3.1 following Definition 3.1). Writing

$$\beta_t 1\{R_n(t-) > 0\} = \tilde{\beta}_{Q_n(t)} 1\{R_n(t-) > 0\}, \qquad t > 0, \tag{5.12}$$

with

$$\tilde{\beta}_s = \beta_{Q_n^{-1}(s)}, \qquad 0 \leq s \leq Q_n(\infty),\ \beta \in \mathscr{B}_1,$$

we see that we may deal with the set of increasing functions with support in $[0, 2\rho_0]$, endowed with L_2 (Lebesgue measure). Hence, we may take

$$H(\delta, b, B) = \text{const.}\sqrt{n}/\delta, \qquad \delta > 0.$$

This gives

$$\phi(b_*) = \text{const.}\, n^{1/4}\sqrt{b_*}$$

and from (4.6),

$$b_* \geq \text{const.}\, n^{1/6}.$$

From Theorem 4.3, we now conclude that $h_\infty(\hat{\beta}, \beta_0) = O_{\mathbb{P}}(n^{1/6})$ and hence $\bar{h}(\hat{\beta}, \beta_0) = (1/\sqrt{n})h_\infty(\hat{\beta}, \beta_0) = O_{\mathbb{P}}(n^{-1/3})$.

Case (iii). Consider the classes

$$\mathscr{G}_r = \{g\colon [0,1] \to [0,C],\ g(t)\uparrow \text{ for } t \leq \tau,\ g(t)\downarrow \text{ for } t > \tau\}$$

and

$$\mathscr{G} = \bigcup_{\tau \in [0,1]} \mathscr{G}_\tau.$$

Let us write $H(\delta, \mathscr{G}_\tau)$ ($H(\delta, \mathscr{G})$) for the δ-entropy with bracketing of $\mathscr{G}_\tau$ ($\mathscr{G}$) endowed with L_2 (Lebesgue measure) norm. Using the assertion in Case (i) on increasing functions, one obtains

$$H(\delta, \mathscr{G}_\tau) \leq \text{const.}(1/\delta).$$

Now, take $\tau_j = j\delta$, $j = 0, 1, \dots, m$, $m \leq \text{const.}(1/\delta)$. For $g_\tau \in \mathscr{G}_\tau$, $\tau \in [\tau_{j-1}, \tau_j]$, there is a $g_{\tau_j} \in \mathscr{G}_{\tau_j}$, such that g_τ and g_{τ_j} only differ on the set $[\tau_{j-1}, \tau_j]$. In fact, one can take $g_{\tau_j} \geq g_\tau$ or $g_{\tau_j} \leq g_\tau$. This implies

$$\exp(H(2\delta, \mathscr{G})) \leq \sum_{j=1}^{m} \exp\bigl(H(\delta, \mathscr{G}_{\tau_j})\bigr),$$

so that also

$$H(\delta, \mathscr{G}) \leq \text{const.}(1/\delta).$$

Using the same reparametrization as in Case (ii), one finds therefore

$$H(\delta, b, B) \leq \text{const.}(\sqrt{n}/\delta).$$

Case (iv). This problem can be related to the estimation of the regression function of bounded variation, where the least squares estimator is a piecewise constant with data-dependent knots [see Mammen and van de Geer (1993)].

We have, for some constant C_0,

$$\left\{\sqrt{\frac{\beta + \beta_0}{2}} : \beta \in \mathscr{B}\right\}$$
$$\subset \mathscr{G} = \left\{g\colon [0, \infty) \to [0, \infty);\, g \text{ of variation bounded by } C_0\right\}.$$

[Indeed, we may take the convex combinations $(\beta + \beta_0)/2$: recall the definition of $H(\delta, b, B)$, just above Theorem 4.3.]

For $\sqrt{(\beta + \beta_0)/2} \in \mathscr{G}$, we can write

$$\sqrt{\frac{\beta + \beta_0}{2}} = c + g_1 - g_2,$$

with c a constant and g_i increasing, $|g_i| \leq C_0$, $i = 1, 2$. Using the same argument as in Case (v) below, one obtains that on $\{h_\infty(\beta, \beta_0) \leq b\} \cap B$, the constant c is uniformly bounded. The derivation is now the same as in Case (ii).

Case (v). This problem is similar to the density estimation problem in Silverman (1982). Alternative estimators of the (cumulative) hazard are, for example, smoothed versions of the Nelson–Aalen estimator [Ramlau-Hansen (1983)].

Similar to Case (iv),

$$\left\{\sqrt{\frac{\beta + \beta_0}{2}} : \beta \in \mathscr{B}\right\} \subset \mathscr{G} = \left\{g\colon [0, 1] \to [0, \infty),\, \int |g^{(m)}|^2 \leq C_0\right\}.$$

We use the line of reasoning of van de Geer [(1990), Example 2.1 (ii)]. Birman and Solomjak (1967) showed that the set

$$\mathscr{G}_1 = \{g \in \mathscr{G}\colon g \leq C_1\}$$

can be covered by exp[const. $(1/\delta)^{1/m}$] balls with radius δ for the supremum norm. Now, use the Sobolev embedding theorem. Write

$$\sqrt{\frac{\beta + \beta_0}{2}} = g_1 + g_2,$$

with $g_1 \in \mathscr{G}_1$ (C_1 appropriately chosen) and g_2 a polynomial of degree $(m-1)$. On the set $\{h_\infty(\beta, \beta_0)\} \cap B$, with B now defined in (5.8), we have

$$\begin{aligned} h_\infty(g_2^2, g_{02}^2) &\le b + h_\infty(g_1^2, g_{01}^2) \\ &\le b + 2C_1\sqrt{n\rho_0} \\ &\le 4C_1\sqrt{n\rho_0}, \end{aligned} \tag{5.13}$$

provided $b \le 2\sqrt{n\rho_0}$. Because on B, $\rho_{ni} \ge \frac{1}{2}\rho_i > 0$, $i = 0, \dots, m-1$, (5.13) implies that the coefficients of the polynomials g_2 are also uniformly bounded. In other words, for $b/\sqrt{n} \to 0$, one can indeed use a member of the covering set of $\mathscr{G}_1$ (C_1 appropriately chosen) to approximate $\beta \in \mathscr{B}$ on the set $\{h_\infty(\beta, \beta_0) \le b\} \cap B$.

Since the entropy with bracketing can be bounded by the entropy for the supremum norm, we may conclude that

$$H(\delta, b, B) \le \text{const.}(\sqrt{n}/\delta)^{1/m},$$

so that $\phi(b) \le \text{const.}\, n^{1/(4m)} b^{(2m-1)/(2m)}$.

6. Maximum likelihood for discrete time. In this section, we specify a version of Theorem 4.3 for n observations. Consider observations $X_1, \dots, X_n$ on $(\mathscr{X}, \mathscr{A})$. For definiteness, let us take the filtration $\mathscr{F}_t = \{\sigma(X_1, \dots, X_t)\}$, $t = 1, \dots, n$. Suppose that X_t has conditional density $f_{\theta_0 t}$ given $\mathscr{F}_{t-1}$, w.r.t. a fixed dominating measure μ, $t = 1, \dots, n$. Here θ_0 is assumed to be a member of a given parameter space Θ.

The log-likelihood ratio is

$$L_n(\theta, \theta_0) = \sum_{t=1}^n \log\left(\frac{f_{\theta t}(X_t)}{f_{\theta_0 t}(X_t)}\right), \qquad \theta \in \Theta, \tag{6.1}$$

and the maximum likelihood estimator $\hat{\theta} \in \Theta$ is defined as the maximizer over $\theta \in \Theta$ of (6.1) (assumed to exist, but not necessarily uniquely defined). The Hellinger process is now

$$h_n^2(\theta, \tilde{\theta}) = \tfrac{1}{2}\sum_{t=1}^n \int \left(\sqrt{f_{\theta t}} - \sqrt{f_{\tilde{\theta}, t}}\right)^2 d\mu.$$

Definition 6.1. We say that a function $f_t(x)$, $t = 1, 2, \dots, n$, $x \in \mathscr{X}$, is predictable if it is of the form

$$f_t(x) = f(x; X_1, \dots, X_{t-1}),$$

with $f\colon \mathscr{X}^t \to \mathbb{R}$ a nonrandom measurable function.

Definition 6.2 (Entropy with bracketing). Given $b > 0$, $\delta > 0$ and a measurable set $B \subset \Omega$, let $\{[f_{\theta_j^L t}, f_{\theta_j^U t}],\ t = 1, \dots, n\}_{j=1}^m$ be a collection of pairs of predictable functions, such that for each $\theta \in \Theta$ there is a $j = j(\theta) \in$

$\{1,\dots,m\}$, such that:

(i) The map $\theta \mapsto j(\theta)$ is nonrandom.
(ii) $f_{\theta_j^L t}(x) \le f_{\theta t}(x) \le f_{\theta_j^U t}(x)$ μ-a.e. for all t and $\omega \in B$.
(iii) $h_n(\theta_j^L, \theta_j^U) \le \delta$ on $\{h_n(\theta, \theta_0) \le b\} \cap B$.

Let $N(\delta, b, B)$ be the smallest value of m for which such a bracketing set $\{[f_{\theta_j^L t}, f_{\theta_j^U t}],\ t = 1,\dots,n\}_{j=1}^m$ exists and let $H(\delta, b, B)$ be a continuous majorant of $\log N(\delta, b, B) \vee 1$.

THEOREM 6.1. *There exist universal constants C_1, C_2, C_3 and C_4, such that under the condition that $\phi(b)/b$ is nonincreasing, with*

$$\phi(b) = \int_{b^2/(C_2\sqrt{n}) \wedge b/8}^{b} \sqrt{H(x, b, B)}\, dx \vee b,$$

we have, for

$$\frac{b_*^2}{C_1} \ge \phi(b_*), \tag{6.2}$$

that

$$\mathbb{P}\big(h_n(\hat{\theta}, \theta_0) > b_*\big) \le C_4 \exp\left(-\frac{b_*^2}{C_3}\right) + \mathbb{P}(B^c). \tag{6.3}$$

PROOF. The proof is completely analogous to that of Theorem 4.3. □

REMARK. For the case that $X_1,\dots,X_n$ are i.i.d., Theorem 6.1 can be found in Wong and Shen (1995).

EXAMPLE. Consider the k-dependent case, that is, for $t > k$, $f_{\theta_0 t}$ depends only on t through $\{X_{t-1},\dots,X_{t-k}\}$. We illustrate the result for a very simple case, where μ is the Lebesgue measure, $f_{\theta t}(x) = f_{\theta_0 t}$ is fixed (known) for $t \le k$ and

$$f_{\theta t}(x) = \theta(x, X_{t-1},\dots,X_{t-k}), \qquad t > k,$$

with

$$\theta \in \Theta = \big\{\theta = \alpha^2 \colon \alpha \colon [0,1]^{k+1} \to [0, C],$$
$$\text{all partial derivatives of order } l \le m \text{ of } \alpha \text{ are bounded by } C\big\}.$$

From Kolmogorov and Tihomirov (1959), we know that $\{\sqrt{\theta} \colon \theta \in \Theta\}$ can be covered by $\exp[\text{const.}(1/\delta)^{(k+1)/m}]$ balls with radius δ for the supremum norm. So clearly

$$H(\delta, b, B) \le \text{const.}\big(\sqrt{n}/\delta\big)^{(k+1)/m}$$

Theorem 6.1 now yields the rate

$$\frac{1}{\sqrt{n}} h_n\big(\hat{\theta}, \theta_0\big) = O_{\mathbb{P}}\big(n^{-m/(2m+k+1)}\big),$$

for $m > (k+1)/2$.

7. Concluding remarks. Uniform (exponential) probability inequalities have many statistical applications. In this paper, we concentrated on the derivation of rates of convergence for (nonparametric) maximum likelihood estimators. However, the inequality of Theorem 3.1 can also be used to obtain uniform local asymptotic normality in certain cases. This in turn may be used to establish the limiting distribution of estimators.

Careful reading of the proof of Theorem 3.1 reveals that (A.10) and (A.11) do not have obvious generalizations for arbitrary martingales. We believed it to be more transparent at this stage to consider only counting processes, instead of formulating conditions for martingales under which the proof of Theorem 3.1 goes through without too many modifications. It should be noted, however, that in the case of continuous martingales, conditions on the entropy without bracketing seem to suffice. The lower integrand in the entropy integral of (3.2) would then be zero. (Moreover, the process can be transformed into a Gaussian one by a random time transformation.)

We also remark that in the case of discrete time martingales, Lemma 2.2 could be replaced by Hoeffding's inequality. This also leads to conditions on a version of entropy without bracketing.

Finally, it is interesting to note that we could avoid explicit moment assumptions on the Hellinger process. However, it may be the case that in several applications these moment assumptions occur in the entropy calculations.

APPENDIX

PROOF OF LEMMA 2.2. We may write $M_t = M_t^c + M_t^d$, $t \geq 0$, where $\{M_t^c\}$ is a continuous martingale and $\{M_t^d\}$ is a purely discontinuous martingale. Consider now for $0 < \lambda < 1/K$ the process

$$Z_t = \lambda M_t - S_t,$$

where S_t is the compensator of the process

$$W_t = \tfrac{1}{2}\lambda^2 \langle M^c, M^c \rangle_t + \sum_{s \leq t} \bigl(\exp(\lambda |\Delta M_s|) - 1 - \lambda |\Delta M_s|\bigr).$$

We shall first show that $\{\exp Z_t\}$ is a supermartingale. For $t_1 < t_2$, we have by the stochastic integration formula,

$$\begin{aligned}\exp Z_{t_2} - \exp Z_{t_1} &= \int_{t_1}^{t_2} \exp Z_{u-}\, dZ_u + \tfrac{1}{2}\int_{t_1}^{t_2} \exp Z_{u-}\, d[Z,Z]_u^c \\ &\quad + \sum_{t_1 < u \leq t_2} \exp Z_{u-}\bigl[\exp(\Delta Z_u) - 1 - \Delta Z_u\bigr] \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

Clearly,

$$\mathrm{I} = \int_{t_1}^{t_2} \exp Z_{u-}\, dZ_u = \lambda \int_{t_1}^{t_2} \exp Z_{u-}\, dM_u - \int_{t_1}^{t_2} \exp Z_{u-}\, dS_u$$

and

$$\mathrm{II} = \tfrac{1}{2}\int_{t_1}^{t_2} \exp Z_{u-}\, d[Z,Z]_u^c = \tfrac{1}{2}\lambda^2 \int_{t_1}^{t_2} \exp Z_{u-}\, d\langle M^c, M^c\rangle_u.$$

Thus, for $\bar{S}_t = S_t - \frac{1}{2}\lambda^2\langle M^c, M^c\rangle_t$, $t \geq 0$, we find

$$\begin{aligned}\mathrm{I} + \mathrm{II} &= \lambda\int_{t_1}^{t_2} \exp Z_{u-}\, dM_u - \int_{t_1}^{t_2} \exp Z_{u-}\, d\bar{S}_u \\ &= \text{(i)} - \text{(ii)}.\end{aligned}$$

To handle III, we use

$$\begin{aligned}&\exp(\Delta Z_u) - 1 - \Delta Z_u \\ &\quad = \exp(\lambda\,\Delta M_u - \Delta S_u) - 1 - \lambda\,\Delta M_u + \Delta S_u \\ &\quad \leq \frac{\exp(\lambda\,\Delta M_u)}{1 + \Delta S_u} - 1 - \lambda\,\Delta M_u + \Delta S_u \\ &\quad = \frac{\exp(\lambda\,\Delta M_u) - 1 - \lambda\,\Delta M_u - \Delta S_u - \lambda\,\Delta M_u\,\Delta S_u}{1 + \Delta S_u} + \Delta S_u \\ &\quad \leq \frac{\Delta W_u - \Delta S_u - \lambda\,\Delta M_u\,\Delta S_u}{1 + \Delta S_u} + \Delta S_u.\end{aligned}$$

This yields

$$\begin{aligned}\mathrm{III} &= \sum_{t_1 < u \leq t_2} \exp Z_{u-}\left[\exp(\Delta Z_u) - 1 - \Delta Z_u\right] \\ &\leq \sum_{t_1 < u \leq t_2} \exp Z_{u-}\left[\frac{\Delta W_u - \Delta S_u - \lambda\,\Delta M_u\,\Delta S_u}{1 + \Delta S_u}\right] \\ &\quad + \sum_{t_1 < u \leq t_2} \exp Z_{u-}\,\Delta S_u = \text{(iii)} + \text{(iv)}.\end{aligned}$$

Define now

$$S_t^c = \bar{S}_t - \sum_{u \leq t} \Delta S_u, \qquad t \geq 0.$$

Then

$$\begin{aligned}-\text{(ii)} + \text{(iv)} &= -\int_{t_1}^{t_2} \exp Z_{u-}\, d\bar{S}_u + \sum_{t_1 < u \leq t_2} \exp Z_{u-}\,\Delta S_u \\ &= -\int_{t_1}^{t_2} \exp Z_{u-}\, d\bar{S}_u^c \\ &\leq -\int_{t_1}^{t_2} \frac{\exp Z_{u-}}{1 + \Delta S_u}\, d\bar{S}_u^c.\end{aligned}$$

Collecting the results so far gives

$$\begin{aligned}\exp Z_{t_2} - \exp Z_{t_1} &\le \text{(i)} - \text{(ii)} + \text{(iii)} + \text{(iv)}\\ &\le \lambda\int_{t_2}^{t_1}\exp Z_{u-}\, dM_u\\ &\quad + \sum_{t_1<u\le t_2}\frac{\exp Z_{u-}}{1+\Delta S_u}[\Delta W_u - \Delta S_u - \lambda\,\Delta M_u\,\Delta S_u]\\ &\quad - \int_{t_1}^{t_2}\frac{\exp Z_{u-}}{1+\Delta S_u}\,d\overline{S}_u^c\\ &= \lambda\int_{t_2}^{t_1}\exp Z_{u-}\, dM_u + \int_{t_1}^{t_2}\frac{\exp Z_{u-}}{1+\Delta S_u}\,d\big(\overline{W}_u - \overline{S}_u\big)\\ &\quad - \lambda\int_{t_1}^{t_2}\frac{\exp Z_{u-}}{1+\Delta S_u}\,d[M,S]_u,\end{aligned}$$

where $\overline{W}_t = \Sigma_{s\le t}\,\Delta W_s$, $t \ge 0$. Since $\{\overline{W}_t - \overline{S}_t\}$ as well as $\{[M,S]_t\}$ are martingales, we thus have

$$\mathbb{E}\big(\exp Z_{t_2} - \exp Z_{t_1}|\mathscr{F}_{t_1}\big) \le 0, \qquad t_1 < t_2,$$

that is, $\{\exp Z_t\}$ is a supermartingale.

Recall now that if $\{X_t\}$ is a supermartingale, then for any stopping time σ,

$$\text{(A.1)} \qquad \mathbb{E}X_\sigma\{\sigma < \infty\} \le 1.$$

Let $A = \{M_t \ge a \text{ and } R_t \ge b^2 \text{ for some } t\}$. Apply (A.1) with $X_t = \exp Z_t$, $t \ge 0$, and $\sigma = \inf\{t\colon M_t \ge a\}$. Because $\sigma < \infty$ on A, we have

$$\text{(A.2)} \qquad \int_A X_\sigma\, d\mathbb{P} \le 1.$$

On the other hand, using condition (2.2), we see that, for $0 < \lambda < 1/K$,

$$\text{(A.3)} \qquad S_t = \sum_{m=2}^{\infty}\frac{\lambda^m}{m!}V_{m,t} \le \frac{\lambda^2}{2(1-\lambda K)}R_t, \qquad t \ge 0.$$

So on A,

$$\text{(A.4)} \qquad X_\sigma \ge \exp\left(\lambda a - \frac{\lambda^2}{2(1-\lambda K)}b^2\right).$$

Combination of (A.2) and (A.4) yields

$$\text{(A.5)} \qquad \mathbb{P}(A) \le \exp\left(-\lambda a + \frac{\lambda^2}{2(1-\lambda K)}b^2\right).$$

This inequality is valid for all $0 < \lambda < 1/K$. Now, choose

$$\lambda = \frac{a}{b^2}\bigg/\left(1 + \frac{Ka}{b^2}\right).$$

Then (A.5) reads

$$\mathbb{P}(A) \leq \exp\left(-\frac{a^2}{2(aK+b^2)} \right). \qquad \square$$

PROOF OF THEOREM 3.1. In the proof, c_i, $i = 1, \dots, 16$, will be constants depending on L and $C_1, \dots, C_4$. Moreover, $M = N - A$. Let

$$I = \min\left\{ i \geq 1 : 2^{-i} \leq \frac{\varepsilon b}{2^5 \sigma_T} \right\}.$$

Write $H_i = H(2^{-(i+1)}b, b, B)$, $i = 0, \dots, I$. It is easy to see that, for proper choice of C_2,

$$\sum_{i=0}^{I} 2^{-i} H_i^{1/2} \leq \frac{c_1}{b} \int_{\varepsilon b^2/(C_2\sigma_T)\wedge b/8}^{b} \sqrt{H(x,b,B)}\, dx \leq \frac{c_1 \varepsilon b}{C_1}$$

and

$$\text{(A.6)} \qquad \sum_{i=1}^{I} 2^{-i} \left(\sum_{k=0}^{i} H_k \right)^{1/2} \leq c_2 \sum_{i=0}^{I} 2^{-i} H_i^{1/2} \leq \frac{c_3 \varepsilon b}{C_1}.$$

If we choose

$$\text{(A.7)} \qquad \eta_i = \max\left\{ \frac{2^{-(i+1)} \left(\sum_{k=0}^{i} H_k\right)^{1/2} C_1}{c_3 \varepsilon b}, 2^{-(i+2)}\sqrt{i} \right\}, \qquad i = 1, \dots, I,$$

then, by (A.6),

$$\sum_{i=1}^{I} \eta_i \leq 1.$$

Let, for $i = 0, \dots, I$, $\{[\tilde{g}_i^L, \tilde{g}_i^U]\}$ be a $(2^{-(i+1)}b)$-bracketing set for $\mathscr{G}$ as defined in Definition 3.1. The subscript refers to the bracketing set and not to the member of a bracketing set. Thus $\log|\{[\tilde{g}_i^L, \tilde{g}_i^U]\}| \leq H_i$, and for each $g \in \mathscr{G}$ there is a pair $[\tilde{g}_i^L, \tilde{g}_i^U]$ such that $\tilde{g}_i^L \leq g \leq \tilde{g}_i^U$ and $d_T(\tilde{g}_i^L, \tilde{g}_i^U) \leq 2^{-(i+1)}b$ on $\{d_T(g,0) \leq b\} \cap B$, $i = 0, \dots, I$. Write

$$\text{(A.8)} \qquad g_i^U = \min_{k \leq i} \tilde{g}_k^U, \qquad g_i^L = \max_{k \leq i} \tilde{g}_k^L, \qquad \Delta_i = g_i^U - g_i^L, \qquad i = 0, \dots, I,$$

and

$$\text{(A.9)} \qquad \nu = \min\{ i \geq 0 : \Delta_i \geq K_i \} \wedge I, \qquad K_i = \frac{2^5 2^{-2i}}{\varepsilon \eta_{i+1}}, \qquad i = 0, \dots, I-1.$$

Note that the pairs $[g_i^L, g_i^U]$ and Δ_i, $i = 0, \dots, I$, as well as ν depend on $g \in \mathscr{G}$, although we do not express this in our notation. Moreover, $[g_i^L, g_i^U]$, Δ_i, $i = 0, \dots, I$, and ν are predictable functions.

We may write

$$\begin{aligned} g &= g_0^L + \sum_{i=0}^{I} (g - g_0^L) 1\{\nu = i\} + \sum_{i=1}^{I} (g_i^L - g_{i-1}^L) 1\{\nu \geq i\} \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}, \end{aligned}$$

and

$$\begin{aligned} &\mathbb{P}\left(\left\{\left|\int_0^T g\,dM\right| \geq \varepsilon b^2 \text{ and } d_T(g,0) \leq b \text{ for some } g \in \mathscr{G}\right\} \cap B\right) \\ &\quad \leq \mathbb{P}\left(\left\{\left|\int_0^T g_0^L\,dM\right| \geq \frac{\varepsilon b^2}{2} \text{ and } d_T(g,0) \leq b \text{ for some } g \in \mathscr{G}\right\} \cap B\right) \\ &\qquad + \mathbb{P}\Bigg(\Bigg\{\left|\int_0^T \sum_{i=0}^{I} (g - g_0^L) 1\{\nu = i\}\,dM\right| \geq \frac{\varepsilon b^2}{4} \text{ and} \\ &\qquad\qquad d_T(g,0) \leq b \text{ for some } g \in \mathscr{G}\Bigg\} \cap B\Bigg) \\ &\qquad + \mathbb{P}\Bigg(\Bigg\{\left|\int_0^T \sum_{i=0}^{I} (g_i^L - g_{i-1}^L) 1\{\nu \geq i\}\,dM\right| \geq \frac{\varepsilon b^2}{4} \text{ and} \\ &\qquad\qquad d_T(g,0) \leq g \text{ for some } g \in \mathscr{G}\Bigg\} \cap B\Bigg) \\ &\quad = \mathbb{P}_{\mathrm{I}} + \mathbb{P}_{\mathrm{II}} + \mathbb{P}_{\mathrm{III}}. \end{aligned}$$

On $\{d_T(g,0) \leq b\} \cap B$, also $d_T(g_0^L, 0) \leq 2b$. Therefore, using Corollary 2.3, for some c_4,

$$\begin{aligned} \mathbb{P}_{\mathrm{I}} &\leq \mathbb{P}\left(\left|\int_0^T g_0^L\,dM\right| \geq \frac{\varepsilon b^2}{2} \text{ and } d_T(g_0^L, 0) \leq 2b \text{ for some } g_0^L\right) \\ &\leq 2\exp\left(H_0 - \frac{\varepsilon b^2}{c_4}\right). \end{aligned}$$

Hence, we see by condition (3.2) that, for C_1 sufficiently large,

$$\mathbb{P}_{\mathrm{I}} \leq 2\exp\left(-\frac{\varepsilon b^2}{c_5}\right).$$

For $i = 1,\dots, I-1$, $\Delta_i 1\{\nu = i\} \geq K_i$ by definition (A.9). Thus, on $\{d_T(g,0) \leq b\} \cap B$,

$$\begin{aligned}\text{(A.10)}\qquad \int_0^T \Delta_i 1\{\nu = i\}\, dA &\leq \frac{\int_0^T \Delta_i^2\, dA}{K_i} \leq \frac{4 d_T^2(g_i^L, g_i^U)}{K_i} \\ &\leq \frac{\varepsilon b^2 \eta_{i+1}}{2^5}, \qquad i = 1,\dots, I-1.\end{aligned}$$

For $i = I$, we have on $\{d_T(g,0) \leq b\} \cap B$,

$$\text{(A.11)}\qquad \int_0^T \Delta_I 1\{\nu = I\}\, dA \leq \sigma_T \left(\int_0^T \Delta_i^2\, dA \right)^{1/2} \leq 2\sigma_T d_T(g_I^L, g_I^U) \leq \frac{\varepsilon b^2}{2^5},$$

where we used the assumption that $B \subset \{A_T \leq \sigma_T^2\}$.

It follows from (A.10) and (A.11) that

$$\sum_{i=0}^{I} \int_0^T \Delta_i 1\{\nu = i\}\, dA \leq \frac{\varepsilon b^2}{2^3},$$

since $\sum_{i=1}^I \eta_i \leq 1$. This implies

$$\begin{aligned}\text{(A.12)}\qquad \mathbb{P}_{\mathrm{II}} \leq \mathbb{P}\Bigg(\Bigg\{ \sum_{i=0}^{I} \left| \int_0^T \Delta_i 1\{\nu = i\}\, dM \right| \geq \frac{\varepsilon b^2}{2^3} \text{ and} \\ d_T(g,0) \leq b \text{ for some } g \in \mathscr{G} \Bigg\} \cap B \Bigg).\end{aligned}$$

Now it is clear that (A.12) implies

$$\begin{aligned}\mathbb{P}_{\mathrm{II}} \leq \mathbb{P}\Bigg(\Bigg\{ \left| \int_0^T \Delta_0 1\{\nu = 0\}\, dM \right| \geq \frac{\varepsilon b^2}{2^4} \text{ and} \\ d_T(g,0) \leq b \text{ for some } g \in \mathscr{G} \Bigg\} \cap B \Bigg) \\ + \mathbb{P}\Bigg(\Bigg\{ \sum_{i=1}^{I} \left| \int_0^T \Delta_i 1\{\nu = i\}\, dM \right| \geq \frac{\varepsilon b^2}{2^4} \text{ and} \\ d_T(g,0) \leq b \text{ for some } g \in \mathscr{G} \Bigg\} \cap B \Bigg) \\ = \mathbb{P}_{\mathrm{II}}^{(1)} + \mathbb{P}_{\mathrm{II}}^{(2)}.\end{aligned}$$

To handle $\mathbb{P}_{\mathrm{II}}^{(1)}$, we use the fact that on $\{d_T(g,0) \leq b\} \cap B$, and for some constant c_6 depending on L,

$$d_T(\Delta_0, 0) \leq c_6 d_T(g_0^L, g_0^U) \leq \tfrac{1}{2} c_6 b^2,$$

so that, from Corollary 2.3,

$$\begin{aligned}\mathbb{P}_{\mathrm{II}}^{(1)} &\le \mathbb{P}\left(\left|\int_0^T \Delta_0 1\{\nu = 0\}\, dM\right| \ge \frac{\varepsilon b^2}{2^4} \text{ and}\right.\\ &\qquad\left. d_T(\Delta_0, 0) \le \frac{1}{2} c_6 b^2 \text{ for some } \Delta_0\right)\\ &\le 2\exp\left(H_0 - \frac{\varepsilon b^2}{c_7}\right) \le 2\exp\left(-\frac{\varepsilon b^2}{c_8}\right),\end{aligned}\tag{A.13}$$

for C_1 sufficiently large.

To handle $\mathbb{P}_{\mathrm{II}}^{(2)}$, we use

$$\left(\sum_{k \le i} H_k\right)^{1/2} \le \frac{c_3 \varepsilon b \eta_i 2^{i+1}}{C_1}$$

and $\Delta_i 1\{\nu = i\} \le K_{i-1}$, $i = 1, \dots, I$, by (A.9). Also, we know that on $\{d_T(g, 0) \le b\} \cap B$,

$$\int_0^T \Delta_i^2\, dA \le 4 d_T^2(g_i^L, g_i^U) \le (2^{-i} b)^2.$$

Application of Lemma 2.1 now yields

$$\begin{aligned}\mathbb{P}_{II}^{(2)} &\le \sum_{i=1}^{I} \mathbb{P}\left(\left|\int_0^T \Delta_i 1\{\nu = i\}\, dM\right| \ge \frac{\varepsilon b^2}{2^4}\eta_i \text{ and}\right.\\ &\qquad\left.\int_0^T \Delta_i^2\, dA \le (2^{-i} b)^2 \text{ for some } \{\Delta_k\}_{k=0}^{i}\right)\\ &\le \sum_{i=1}^{I} 2\exp\left(\sum_{k \le i} H_k - \frac{\varepsilon b^2 \eta_i^2 2^{2i}}{c_9}\right) \le \sum_{i=1}^{I} 2\exp\left(-\frac{\varepsilon b^2 \eta_i^2 2^{2i}}{c_{10}}\right)\\ &\le \sum_{i=1}^{I} 2\exp\left(-\frac{\varepsilon b^2 i}{c_{11}}\right) \le c_{12}\exp\left(-\frac{\varepsilon b^2}{c_{13}}\right).\end{aligned}\tag{A.14}$$

Observe now that $|g_i^L - g_{i-1}^L| \le \Delta_{i-1}$, $i = 1, \dots, I$. We may write

$$\begin{aligned}\mathbb{P}_{\mathrm{III}} &\le \mathbb{P}\left(\left\{\left|\int_0^T (g_1^L - g_0^L) 1\{\nu \ge 1\}\, dM\right| \ge \frac{\varepsilon b^2}{2^3} \text{ and}\right.\right.\\ &\qquad\qquad\left.\left. d_T(g, 0) \le b \text{ for some } g \in \mathscr{G}\right\} \cap B\right)\\ &\quad + \mathbb{P}\left(\left\{\sum_{i=2}^{I}\left|\int_0^T (g_i^L - g_{i-1}^2) 1\{\nu \ge i\}\, dM\right| \ge \frac{\varepsilon b^2}{2^3} \text{ and}\right.\right.\\ &\qquad\qquad\left.\left. d_T(g, 0) \le b \text{ for some } g \in \mathscr{G}\right\} \cap B\right)\\ &= \mathbb{P}_{\mathrm{III}}^{(1)} + \mathbb{P}_{\mathrm{III}}^{(2)}.\end{aligned}$$

It follows by similar arguments as in (A.13) that Corollary 2.3 implies

$$\mathbb{P}_{\mathrm{III}}^{(1)} \leq 2\exp\left(-\frac{\varepsilon b^2}{c_{14}}\right),$$

and similar arguments as in (A.14) yield that Lemma 2.1 implies

$$\mathbb{P}_{\mathrm{III}}^{(2)} \leq c_{15}\exp\left(-\frac{\varepsilon b^2}{c_{16}}\right). \qquad \square$$

REFERENCES

Alexander, K. S. (1985). Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. A. Olshen, eds.) **2** 475–493. Wadsworth, Belmont, CA.

Andersen, N. T., Giné, E., Ossiander, M. and Zinn, J. (1988). The central limit theorem and the law of the iterated logarithm for empirical processes under local conditions. *Probab. Theory Related Fields* **77** 271–305.

Bass, R. F. (1985). Law of the iterated logarithm for set-indexed partial sum processes with finite variance. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

Birman, M. Š. and Solomjak, M. Z. (1967). Piece-wise polynomial approximations of functions of the classes W_p^α. *Mat. Sb.* **73** 295–317.

Gill, R. D. (1980). *Censoring and Stochastic Integrals*. Mathematical Centre Tracts, Math. Centrum, Amsterdam.

Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. A. Olshen, eds.) **2** 539–555. Wadsworth, Belmont, CA.

Huang, J. and Wellner, J. A. (1995). Estimation of a monotone density or monotone hazard under random censoring. *Scand. J. Statist.* **22** 3–34.

Jacobsen, M. (1989). Right censoring and martingale methods for failure time data. *Ann. Statist.* **17** 1133–1156.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191–219.

Kolmogorov, A. N. and Tikhomirov, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk* **14** 3–86 [English transl. *Amer. Math. Soc. Transl. Ser. 2* **17** (1961) 277–364].

Lipster, R. S. and Shiryayev, A. N. (1989). *Theory of Martingales*. Kluwer, Dordrecht.

Mammen, E. and van de Geer, S. (1993). Locally adaptive regression splines. Technical report, Humboldt Univ., Berlin.

Ossiander, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15** 897–919.

Pollard, D. (1989). Asymptotics via empirical processes (with discussion). *Statist. Sci.* **4** 341–366.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. IMS, Hayward, CA.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11** 453–466.

Shorack, G. R. and Wellner, J. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924.

van de Geer, S. (1991). The entropy bound for monotone functions. Technical Report TW 91-10, Univ. Leiden.

VAN DE GEER, S. (1993a). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44.

VAN DE GEER, S. (1993b). Rates of convergence for the maximum likelihood estimator in mixture models. Technical Report TW 93-09, Univ. Leiden.

VAN DE GEER, S. (1995). The method of sieves and minimum contrast estimators. *Mathematical Methods of Statistics* **4** 20–38.

WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLE's. *Ann. Statist.* **23** 339–362.

MATHEMATICAL INSTITUTE
UNIVERSITY OF LEIDEN
P.O. BOX 9512
2300 RA LEIDEN
THE NETHERLANDS