On Hoeffding's inequality for dependent random variables

Sara van de Geer

Abstract. Let $\{Z_\theta : \theta \in \Theta\}$ be a random process indexed by a parameter θ in some (metric) space Θ. Given an appropriate moment or probability inequality for each fixed θ (a *pointwise* inequality), one can often derive an inequality that holds *uniformly* in $\theta \in \Theta$ by applying the chaining technique. Therefore, pointwise inequalities are (apart from being of interest in itself) quite relevant within the theory of stochastic processes. We present a generalization of Hoeffding's inequality, and the related bounded difference inequality of McDiarmid [7]. We also state the corresponding uniform inequality. As an application, we consider estimation in the auto-regression model.

1 Introduction

For a stochastic stochastic process $\{Z_\theta : \theta \in \Theta\}$, one is often interested in moment or probability inequalities, that hold *uniformly* in $\theta \in \Theta$. For instance, in some statistical applications, one may want to apply an inequality with θ randomly chosen (an estimator, say), which is possible indeed when the inequality is uniform. Uniform inequalities are in quite a few cases an immediate consequence of *pointwise* inequalities, that is, inequalities that hold for all $\theta \in \Theta$ fixed.

In Theorem 1.2 below we briefly review a result given in van der Vaart and Wellner [11] (their Corollary 2.2.5). This theorem serves as a major illustration that for many situations, it is enough to prove a pointwise inequality. The theorem is followed by a short discussion. We then establish in Section 2 a pointwise Hoeffding type inequality. The extension to a uniform inequality is presented in Section 3. Section 4 is a statistical application, namely on proving rates of convergence in auto-regression.

The inequalities of Theorem 1.2 concern the Orlicz norm of the random variables involved, which is defined as follows.

Definition 1.1. *Let ψ be a convex, nondecreasing, nonzero function on $[0, \infty)$, with $\psi(0) = 0$ (an Orlicz function). The Orlicz norm of the random variable X is defined as*

$$\|X\|_\psi = \inf\{K > 0 : \mathbf{E}\psi(|X|/K) \leq 1\},$$

(with the usual conventions if the expectation is infinite).

For example, when $\psi(x) = x^p$, with $p \geq 1$, the Orlicz norm $\|X\|_\psi$ is equal to the L_p norm $(E|X|^p)^{1/p}$.

Condition (1.1) below is a pointwise inequality for the Orlicz norm (albeit that there are two parameters θ and ϑ involved), and (1.2) is the resulting uniform inequality.

Theorem 1.2. *Let ψ be an Orlicz function satisfying $\limsup_{x,y\to\infty} \psi(x)\psi(y)/\psi(cxy) < \infty$, for some constant c. Suppose $\{Z_\theta : \theta \in \Theta\}$ is a separable stochastic process indexed by θ in the (pseudo-)metric space (Θ, τ). Assume that*

$$\|Z_\theta - Z_\vartheta\|_\psi \leq C\tau(\theta, \vartheta), \quad \textit{for every } \theta, \vartheta, \tag{1.1}$$

where C is some constant. Then there exists a constant C' depending only on ψ and C, such that

$$\|\sup_{\theta,\vartheta} |Z_\theta - Z_\vartheta|\|_\psi \leq C' \int_0^{\mathrm{diam}(\Theta)} \psi^{-1}(D(\delta))d\delta, \tag{1.2}$$

where $\mathrm{diam}(\Theta)$ *is the diameter of Θ and $D(\delta)$ is the δ-packing number of Θ (i.e. the maximum number of δ-separated points in Θ).*

When ψ is the L_p norm, the theorem may be applied to the case where Θ is a bounded subset of d-dimensional Euclidean space, with $d < p$. But for very large (infinite dimensional) Θ, Theorem 1.2 needs more than an L_p norm.

An important special case is

$$\psi(x) = \exp[x^2] - 1. \tag{1.3}$$

A random variable with finite Orlicz norm for this choice of ψ is called sub-Gaussian, because its tails behave like those of a Gaussian random variable, or are even slimmer. Theorem 1.2 with ψ given in (1.3) gives the uniform bound in terms of the square root of the log-packing number (or the *entropy*) of Θ. The resulting uniform bound can be used to prove e.g. tightness of empirical processes based on independent observations, (using symmetrization techniques) or equicontinuity of certain Gaussian processes indexed by functions.

One may also consider

$$\psi(x) = \exp[x] - 1. \tag{1.4}$$

A random variable with finite Orlicz norm for this choice of ψ is called sub-exponential, because its tails decrease at least exponentially fast. In Theorem 1.2, the uniform bound in the sub-exponential case is again in terms of the log-packing number, but the square root which appeared in the sub-Gaussian case is now lost.

We remark however that in special cases one can prove that a bound of the form (1.2) holds in the sub-exponential case, with on the left hand side ψ given by (1.4) and on the right hand side a different ψ, namely the one given in (1.3). This is for example true when $\{Z_\theta\}$ is the empirical process based on i.i.d. observations with distribution P. One may then take Θ as a subset of $L_2(P)$, in which case the log-packing number(entropy), should be replaced by the stronger entropy with bracketing. The result is closely intertwined with Bernstein's inequality, which is roughly speaking sub-Gaussian in the near tails, and sub-exponential in the far tails.

Theorem 1.2 can be applied to sums of random variables, say $Z_\theta = \sum_{i=1}^n X_{i,\theta}$. It is then not necessary to assume independence of the $X_{i,\theta}$, as long as (1.1) is met. Also more complicated objects based on dependent random variables can be studied with the

help of Theorem 1.2. See for example the paper of Dahlhaus and Polonik [4] (in this volume), where the empirical spectral process is examined.

In this paper, we consider an analogue of the sub-Gaussian case for dependent variables. Related results for the sub-exponential case have been treated in Nishiyama [8], and van de Geer [9].

Our main result is the derivation of a *pointwise* inequality, which will be a generalization of Hoeffding's inequality (and which will be further generalized to a bounded difference inequality, cf. [7]). The extension to a uniform inequality for weighted sums is straightforward. This extension makes use of a partitioning entropy condition, which relies on a generalization of the usual entropy concept.

2 Hoeffding's inequality and related results

Consider a probability triple $(\Omega, \mathcal{F}, \mathbf{P})$ and let $\emptyset = \mathcal{F}_0 \subset \mathcal{F}_1, \subset \ldots \subset \mathcal{F}$ be an increasing sequence of sigma-algebras. Let for each i, X_i be a real-valued $\mathcal{F}_i$-measurable random variable, satisfying

$$\mathbf{E}(X_i|\mathcal{F}_{i-1}) = 0, \text{ a.s.}$$

We will consider the martingale

$$S_n = \sum_{i=1}^{n} X_i, \ n \geq 1.$$

Throughout the rest of this paper, we let ψ be the Orlicz function

$$\psi(x) = \exp[x^2] - 1, \ x \geq 0.$$

Consider $\mathcal{F}_{i-1}$ measurable random variables $K_i > 0$, $i = 1, 2, \ldots$. Define $B_0^2 = 0$ and for $n \geq 1$,

$$B_n^2 = \sum_{i=1}^{n} K_i^2 \left(1 + \mathbf{E}(\psi(\frac{|X_i|}{K_i})|\mathcal{F}_{i-1})\right).$$

Lemma 2.1. *For all β,*

$$\{\xi_n(\beta) = \exp[\beta S_n - 2\beta^2 B_n^2]\}$$

is a supermartingale.

Proof. By a slight extension of Lemma 8.1 in van de Geer [10], we have that for all $i \geq 1$,

$$\mathbf{E}\left(\exp[\beta X_i - 2\beta^2 K_i^2(1 + \psi(\frac{|X_i|}{K_i}))]|F_{i-1}\right) \leq 1, \text{a.s.}$$

Thus,

$$\mathbf{E}(\exp[\beta S_n - 2\beta^2 B_n^2]|\mathcal{F}_{i-1}) \leq \exp[\beta S_{n-1} - 2\beta^2 B_{n-1}^2], \text{ a.s., } n \geq 1.$$

□

Theorem 2.2. *For all $a > 0$, $b > 0$,*

$$\mathbf{P}(S_n \geq a \text{ and } B_n^2 \leq b^2 \text{ for some } n) \leq \exp[-\frac{a^2}{8b^2}].$$

Proof. The argument is as in Freedman [5]. Let $\{\xi_n(\beta)\}$ be defined as in Lemma 2.1. Then, for any stopping time σ,

$$\mathbf{E}(\xi_\sigma(\beta)\{\sigma < \infty\}) \leq 1.$$

Take A as the set

$$A = \{S_n \geq a \text{ and } B_n^2 \leq b^2 \text{ for some } n\},$$

and take

$$\sigma = \inf\{n : \ S_n \geq a\}.$$

Because $A \subset \{\sigma < \infty\}$,

$$\mathbf{E}\xi_\sigma(\beta)1_A \leq 1.$$

On A,

$$\xi_\sigma(\beta) \geq \exp[\beta a - 2\beta^2 b^2].$$

So

$$\mathbf{P}(A) \leq \exp[-\beta a + 2\beta^2 b^2].$$

Now, take $\beta = a/(4b^2)$. □

An important special case is the one where S_n is a weighted sum of martingale differences, with predictable weights.

Corollary 2.3. *Let W_i be $\mathcal{F}_i$-measurable, and $\mathbf{E}(W_i|\mathcal{F}_{i-1}) = 0$, $i \geq 1$. Suppose that for some constant $c < \infty$,*

$$\mathbf{E}(\psi(\frac{|W_i|}{c})|\mathcal{F}_{i-1}) \leq 1, \text{ a.s. } i = 1, 2, \ldots.$$

Let g_i be $\mathcal{F}_{i-1}$-measurable weights, and take $X_i = g_i W_i$, $i = 1, 2, \ldots$. It follows that for all $a > 0$, $d > 0$

$$\mathbf{P}(\sum_{i=1}^n g_i W_i \geq a \text{ and } \sum_{i=1}^n g_i^2 \leq d^2 \text{ for some } n) \leq \exp[-\frac{a^2}{16c^2d^2}]. \tag{2.1}$$

Clearly, Theorem 2.2 can also be applied to the case where the X_i are bounded. In that case, one arrives at a Hoeffding-type of inequality. However, the constants that appear in the exponential bound are then larger than those of Hoeffding [6]. We therefore present the improvement with Hoeffding's constants for the bounded case.

Lemma 2.4. *Suppose that*

$$L_i \leq X_i \leq U_i, \text{ a.s. for all } i \geq 1,$$

where $L_i < U_i$ are $\mathcal{F}_{i-1}$-measurable random variables, $i \geq 1$. Define $C_0^2 = 0$ and

$$C_n^2 = \sum_{i=1}^n (U_i - L_i)^2, \ n \geq 1.$$

Then for all β,

$$\{\zeta_n(\beta) = \exp[\beta S_n - \beta^2 C_n^2/8]\}$$

is a supermartingale.

Proof. This follows from a slight extension of Hoeffding [6]:

$$\mathbf{E}(\exp[\beta X_i]|\mathcal{F}_{i-1}) \leq \exp[\beta^2(U_i - L_i)^2/8], \text{ a.s.}$$

□

Theorem 2.5. *(Hoeffding's inequality.) Suppose the condition of Lemmma 2.4 holds. Let $\{C_n\}$ be defined as there. Then for all $a > 0$, $c > 0$,*

$$\mathbf{P}(S_n \geq a \text{ and } C_n^2 \leq c^2 \text{ for some } n) \leq \exp[-\frac{2a^2}{c^2}].$$

Proof. In view of Lemma 2.4, this can be derived using the same arguments as in Theorem 2.2. □

Indeed, Theorem 2.5 is the inequality of Hoeffding [6], except that Hoeffding assumes non-random bounds $\{L_i, U_i\}$. The inequality is also known as Azuma's inequality (Azuma [1]).

Theorem 2.5 considers the $\mathcal{F}_n$-measurable random variables $S_n = \sum_{i=1}^n X_i$. We now examine general $\mathcal{F}_n$-measurable random variables. We state the result for fixed n.

Theorem 2.6. *(Bounded difference inequality.) Fix $n \geq 1$. Let Z_n be a $\mathcal{F}_n$-measurable random variable, satisfying for each $i = 1, \ldots, n$,*

$$L_i \leq \mathbf{E}(Z_n|\mathcal{F}_i) \leq U_i, \text{ a.s.}$$

where $L_i < U_i$ are $\mathcal{F}_{i-1}$-measurable, $i = 1, \ldots, n$. Define $C_n^2 = \sum_{i=1}^n (U_i - L_i)^2$. Then for all $a > 0$, $c > 0$,

$$\mathbf{P}(Z_n - \mathbf{E}Z_n \geq a \text{ and } C_n^2 \leq c^2) \leq \exp[-\frac{2a^2}{c^2}].$$

Proof. We may write

$$Z_n - \mathbf{E}Z_n = \sum_{i=1}^n \tilde{X}_i,$$

where $\tilde{X}_i = \mathbf{E}(Z_n|\mathcal{F}_i) - \mathbf{E}(Z_n|\mathcal{F}_{i-1})$, $i = 1, \ldots, n$. Clearly, $\tilde{X}_i$ is $\mathcal{F}_i$-measurable, and $\mathbf{E}(\tilde{X}_i|\mathcal{F}_{i-1}) = 0$, $i = 1, \ldots, n$. Moreover, for $i = 1, \ldots, n$,

$$\tilde{L}_i = L_i - \mathbf{E}(Z_n|\mathcal{F}_{i-1}) \leq \tilde{X}_i \leq U_i - \mathbf{E}(Z_n|\mathcal{F}_{i-1}) = \tilde{U}_i,$$

so that $\tilde{L}_i < \tilde{U}_i$ are $\mathcal{F}_{i-1}$-measurable random variables, with $\tilde{U}_i - \tilde{L}_i = U_i - L_i$. The result is thus a consequence of Theorem 2.5. □

Theorem 2.6 generalizes the bounded difference inequality of McDiarmid [7]. As an illustration, consider independent random variables $Y_1, \ldots, Y_n$, with values in some measurable space $\mathcal{Y}$. Let $\mathcal{F}_i = \sigma(Y_1, \ldots, Y_i)$, $i = 1, \ldots, n$. Let $g(y_1, \ldots, y_n)$ be some real-valued function on $\mathcal{Y}^n$ satisfying the bounded difference assumption

$$|g(y_1, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_n) - g(y_1, \ldots, y_{i-1}, y_i', y_{i+1}, \ldots, y_n)|$$

$$\leq k_i, \text{ for all } y_i, y_i', \text{ and all } i,$$

where $k_1, \ldots, k_n$ are constants. Take $Z_n = g(Y_1, \ldots, Y_n)$. Then one can take

$$L_i = \inf_y \mathbf{E}(g(Y_1, \ldots, Y_{i-1}, y, Y_{i+1}, Y_n)|\mathcal{F}_i),$$

$$U_i = \sup_y \mathbf{E}(g(Y_1, \ldots, Y_{i-1}, y, Y_{i+1}, \ldots, Y_n|\mathcal{F}_i),$$

which satisfy $U_i - L_i \leq k_i$, $i = 1, \ldots, n$. The bounded difference inequality now reads: for all $a > 0$,

$$\mathbf{P}(g(Y_1, \ldots, Y_n) - \mathbf{E}g(Y_1, \ldots, Y_n) \geq a) \leq \exp[-\frac{2a^2}{\sum_{i=1}^n k_i^2}].$$

In the case $\mathcal{Y} = \mathbf{R}$, and $g(Y_1, \ldots, Y_n) = \sum_{i=1}^n Y_i$, this is Hoeffding's inequality for independent random variables.

3 A uniform inequality for weighted sums

In this section, we present an analogue of Theorem 1.2, with Orlicz function $\psi(x) = \exp[x^2] - 1$. Consider some space of parameters Θ. Instead of using a metric τ on Θ, we will define a random quantity $D_n(\theta, \vartheta)$ to measure the closeness of two parameters θ and ϑ (see below). Consider $\mathcal{F}_{i-1}$-measurable real-valued random variables $g_{i,\theta}$, $\theta \in \Theta$. Let W_i be real-valued $\mathcal{F}_i$-measurable, with $\mathbf{E}(W_i|\mathcal{F}_{i-1}) = 0$, $i = 1, 2, \ldots$. Assume moreover that for all i, and some constant c,

$$\mathbf{E}(\psi(\frac{|W_i|}{c})|\mathcal{F}_{i-1}) \leq 1. \tag{3.1}$$

Write

$$D_n^2(\theta, \vartheta) = \frac{1}{n}\sum_{i=1}^n (g_{i,\theta} - g_{i,\vartheta})^2. \tag{3.2}$$

Take a fixed $\theta_0 \in \Theta$. Let $F \in \mathcal{F}$ be some measurable set. This will be the set where the partitioning entropy defined below is supposed to be well-behaved. It depends on the particular application, but generally one wants that $\mathbf{P}(F^c)$, where F^c is the complement of F, is small.

The log-packing number in (1.2) will be replaced by a partitioning entropy:

Assumption 3.1. (The partitioning entropy condition.) For $0 \leq \delta \leq d$, let $\{\theta_j\}_{j \in \mathcal{J}}$ be such that for all $\theta \in \Theta$, there is a $j = j(\theta) \in \mathcal{J}$ such that $D_n(\theta, \theta_j) \leq \delta$ on $\{D_n(\theta, \theta_0) \leq d\} \cap F$. We assume that $\mathcal{J}$ can be chosen as a finite set $\mathcal{J} = \{1, \ldots, J\}$ for each δ and d. We then write $N(\delta, d) = J$, and let $H(\delta, d)$ be a continuous majorant of $\log(1 + J)$. We call $H(\delta, d)$ a partitioning entropy.

Theorem 3.2. *Suppose that on F, $\sum_{i=1}^n W_i^2/n \leq \sigma^2$, a.s. There exists a constant c_0 depending only on c such that for*

$$\sqrt{n}a \geq c_0(\int_{\frac{a}{\sigma c_0}}^d H^{1/2}(\delta, d)d\delta \vee d), \tag{3.3}$$

we have

$$\mathbf{P}\left((\frac{1}{n}\sum_{i=1}^n (g_{i,\theta} - g_{i,\theta_0})W_i \geq a \text{ and } D_n(\theta, \theta_0) \leq d \text{ for some } \theta \in \Theta) \cap F\right)$$

$$\leq c_0 \exp[-\frac{na^2}{c_0^2 d^2}].$$

Proof. This follows by copying the proof of Lemma 3.2 in van de Geer [10], with its pointwise sub-Gaussian probability inequality replaced by the extension (2.1) of Corollary 2.3. □

4 Application to auto-regression

Now that Theorem 3.2 is available, one can start the engines for a theory on M-estimation in auto-regression, possibly using sieves and/or penalties. The results are very similar to the independent case. The only additional technicality is the handling of random metrics. In this section, we will briefly present a sample of the theory.

Consider n real-valued observations $Y_1, \ldots, Y_n$, with Y_i $\mathcal{F}_i$-measurable, $i \geq 1$. Let Θ be a given parameter space, and suppose that

$$\mathbf{E}(Y_i|\mathcal{F}_{i-1}) = g_{i,\theta_0} \text{ a.s., } i \geq 1,$$

where $\theta_0 \in \Theta$. Define $W_i = Y_i - g_{i,\theta_0}$, $i \geq 1$. We assume that $\{W_i\}$ satisfies the sub-Gaussianity condition (3.1).

We let $\hat{\theta}_n$ be the least squares estimator

$$\hat{\theta}_n = \arg\min_{\Theta} \sum_{i=1}^{n} (Y_i - g_{i,\theta})^2.$$

Let $D_n(\theta, \vartheta)$ be given in (3.2). Consider a partitioning entropy $H(\delta, d)$ as introduced in Assumption 3.1. A rate of convergence for $D_n(\hat{\theta}_n, \theta_0)$ can now be obtained in exactly the same way as in Theorem 9.1 of van de Geer [10]. To facilitate the exposition, we only consider the case where the entropy-integral exists.

Theorem 4.1. *Define*

$$\phi(d) = \int_0^d H^{1/2}(\delta, d) d\delta \vee d,$$

and suppose that $\phi(d)/d^2$ is a non-increasing function of d. Then there exists a constant c_0, such that for all $d \geq d_n$, where $\sqrt{n} d_n^2 \geq c_0 \phi(d_n)$,

$$\mathbf{P}(D_n(\hat{\theta}_n, \theta_0) \geq d) \leq c_0 \exp[-\frac{nd^2}{c_0^2}] + \mathbf{P}(F^c).$$

Example 4.2. (Linear auto-regression.) Take $\Theta = \mathbf{R}^r$ and

$$g_{i,\theta} = \theta_1 \psi_{i,1} + \cdots + \theta_r \psi_{i,r}, \ \theta \in \Theta,$$

where for each i, the variables $\psi_{i,1}, \ldots, \psi_{i,r}$ are $\mathcal{F}_{i-1}$-measurable. Note that now

$$D_n^2(\theta, \vartheta) = (\theta - \vartheta)' \Sigma_n (\theta - \vartheta),$$

with Σ_n a random $r \times r$ matrix. We assume that the range of Σ_n is non-random, and let $\Sigma \geq 0$ be a symmetric non-random $r \times r$ matrix with the same range. Take

$$F \subset \{\frac{1}{4} \leq \|\Sigma^{-1/2} \Sigma_n \Sigma^{-1/2}\| \leq 4\},$$

where $\Sigma^{-1/2}$ is the square root of the (generalized) inverse of Σ, and where $\|\cdot\|$ denotes the norm of a matrix as linear map. On F, the random metric D_n is equivalent to the non-random metric D given by

$$D^2(\theta,\vartheta) = (\theta-\vartheta)'\Sigma(\theta-\vartheta).$$

Therefore (invoke e.g. Corollary 2.6 of van de Geer [10]),

$$H(\delta,d) \leq r\log(\frac{20d}{\delta}).$$

It follows that

$$\phi(d) \leq Ad\sqrt{r},$$

with $A = \int_0^1 \sqrt{\log(\frac{20}{u})}du$. Thus, for $c_1 = c_0 A$ a constant depending only on c, and for all $T \geq c_1$,

$$\mathbf{P}(D(\hat{\theta}_n,\theta_0) \geq T\sqrt{\frac{r}{n}}) \leq c_1\exp[-\frac{r^2T^2}{c_1^2}] + \mathbf{P}(F^c).$$

For ways of handling $\mathbf{P}(F^c)$, see Baraud, Comte and Viennet ([2]). In fact, they study the much more involved problem of adaptation.

Example 4.3. (Smooth regression.) One can easily extend Theorem 4.1 to the case of penalized least squares. Let us briefly consider a simple example. Suppose that $Y_i \in [0,1]$ for all i, and that $g_{i,\theta} = \theta(Y_{i-1})$, $i \geq 1$ (with $Y_0 = 1/2$ (say)). Let $m \in \{1,2,\ldots\}$ be given, and let

$$\Theta = \{\theta : I(\theta) < \infty\},$$

where $I^2(\theta) = \int_0^1 |\theta^{(m)}(y)|^2 dy$. Consider the penalized least squares estimator

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta}\left\{\frac{1}{n}\sum_{i=1}^n (Y_i - \theta(Y_{i-1}))^2 + \mathrm{pen}^2(\theta)\right\},$$

with penalty

$$\mathrm{pen}^2(\theta) = \lambda_n^2 I^2(\theta).$$

Here $0 < \lambda_n^2 \leq 1$ is a smoothing parameter. The partitioning entropy $H(\delta,d)$ should now be taken on the set where also $\mathrm{pen}(\theta) \leq d$. Birman and Solomjak [3] show that the δ-entropy for the uniform norm of a set of functions θ, bounded by 1 and with $I(\theta) \leq L$, is of order $(L/\delta)^{1/m}$. Because in our case, $\mathrm{pen}(\theta) \leq d$ reads $I(\theta) \leq d/\lambda_n$, we find

$$H(\delta,d) \leq \text{const.}(\frac{d}{\lambda_n\delta})^{1/m}.$$

The condition $\sqrt{n}d_n^2 \geq c_0(\int_0^{d_n} H^{1/2}(\delta,d)d\delta \vee d_n)$ leads to $d_n \geq c_0' n^{-1/2}\lambda_n^{-1/2m}$. As a consequence, one finds the rate $n^{-1/2}\lambda_n^{-1/2m} \vee \lambda_n I(\theta_0)$ for $D_n(\hat{\theta}_n,\theta_0)$.

Bibliography

[1] K. Azuma: Weighted sums of certain dependent random variables. *Tôkuku Mathematical Journal* **19** 357–367, 1967.

[2] Y. Baraud, F. Comte and G. Viennet: Adaptive estimation in autoregression or β-mixing regression via model selection. Technical Report 566, Laboratoire de Probabilités et Modèles Aléatoires, Universités de Paris 6 et Paris 7, 2000. Available under www.dma.ens.fr/∼baraud.

[3] M.Š. Birman and M.Z. Solomjak: Piece-wise polynomial approximations of functions in the classes W_p^α. *Mathematics of the USSR Sbornik* **73** 295–317, 1967.

[4] R. Dahlhaus and W. Polonik: Empirical spectral process and nonparametric maximum likelihood estimation for time series (tentative title). In this volume: H. Dehling, T. Mikosch and M. Sørensen (Eds.) *Empirical Process Techniques for Dependent Data.* Birkhäuser, Boston, 2001.

[5] D. Freedman: On tail probabilities for martingales. *The Annals of Probability* **3** 100–118, 1975.

[6] W. Hoeffding: Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58** 13–30, 1963.

[7] C. McDiarmid: On the method of bounded differences. In *Surveys in Combinatorics* 148–188. Cambridge University Press, Cambridge, 1989.

[8] Y. Nishiyama: Weak convergence of some classes of martingales with jumps. *The Annals of Probability* **28** 685–712, 2000.

[9] S.A. van de Geer: Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics* **23** 1779–1801, 1995.

[10] S.A. van de Geer: *Empirical Processes in M-Estimation.* Cambridge University Press, Cambridge, 2000.

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

Sara van de Geer
Mathematical Institute
University of Leiden
P.O. Box 9512
2300 RA Leiden
The Netherlands
geer@math.leidenuniv.nl