Pré-Publicações do Departamento de Matemática Universidade de Coimbra Preprint Number 03–17

# AN EXPONENTIAL INEQUALITY FOR ASSOCIATED VARIABLES

### PAULO EDUARDO OLIVEIRA

ABSTRACT: Exponential inequalities have been an important tool in probability and statistics. Versions of Bernstein type inequalities have been proved for independent and for some dependence structure. We prove an exponential inequality for positively associated and strictly stationary random variables replacing an uniform boundedness assumption by the existence of Laplace transforms. As usual with this dependence structure we need some conditions on the covariances of the variables. These conditions are analogous to the ones assumed for version of the exponential inequality for bounded associated variables already known. The proof uses a truncation technique together with a block decomposition of the sums. The truncating sequence must be well tuned with the decrease rate of the covariances and the sizes of the blocks. A description of the behaviour of such sequences is given.

KEYWORDS: association, exponential inequalities. AMS SUBJECT CLASSIFICATION (2000): 60F10, 60E15.

# 1. Introduction

One of the main tools used for characterizing convergence rates in nonparametric estimation has been convenient versions of Bernstein type exponential inequalities. There exist several versions available in the literature for independent sequences of variables with assumptions of uniform boundedness or some, quite relaxed, control on their (centered or noncentered) moments. If the independent case is classical in the literature the treatment of dependent variables is more recent. The extension to dependent variables was first studied considering m-dependence or different mixing conditions. An exponential inequality for strong mixing variables eventually was proved (Carbon [3]) using the same type of assumptions on the variables, besides the strong mixing: uniformly bounded or some control on the moments. Naturally, these later extensions included some extra terms on the upper bounds depending on the mixing coefficients. An account of the main results briefly described before may be found in Bosq [4]. In another direction in controlling dependent variables, versions of exponential inequalities are also available for martingale

1

Author partially supported by Centro de Matemática da Universidade de Coimbra, Fundação para a Ciência e Tecnologia (FCT) and POCTI.

differences supposing the variables to be uniformly bounded (Azuma [1]) and, more recently, only the existence of Laplace transforms (Lesigne, Volný [11]). Another dependence structure that has attracted the interest of probabilists and statisticians is association, as introduced by Esary, Proschan, Walkup [7]. For this dependence structure the idea of asymptotic independence is not so explicitly stated as in mixing structures. For associated random variables Birkel [2] seems to have been the first author to prove some moment inequalities. An exponential type inequality appeared much later in Ioannides, Roussas [10] under the assumption of uniform boundedness and some convenient behaviour on the covariance structure of the variables. This result inspired a method that produced almost sure consistency results in nonparametric distribution function estimation, based on associated samples, with description of rates (Henriques, Oliveira [8]). In another direction, with a somewhat different method, exponential decay rates for nonparametric density estimation, also based on associated samples, were proved in Henriques, Oliveira [9]. The present article presents an extension of the Ioannides, Roussas's [10] inequality dropping the boundedness assumption, which is replaced by the existence of Laplace transforms.

The article is organized as follows: section 2 describes some auxiliary results and introduces the truncated variables used to approximate the original variables, the corresponding tails and the block decomposition of the sums; section 3 studies the truncated part giving conditions on the truncating sequence to enable the proof of an exponential inequality for these terms; section 4 treats the tails left aside from the truncation and, finally, section 5 summarizes the partial results into a final theorem. As indicated, the proof technique consists on a truncation which is then treated using a blocking decomposition of the sums, together with a control on the tails of the distribution, achieved assuming the existence of Laplace transforms.

## 2. Definitions, preliminary results and notation

We say that the variables  $X_1, X_2, \ldots$  are associated if, for every  $n \in \mathbb{N}$  and  $f, g: \mathbb{R}^n \longrightarrow \mathbb{R}$  coordinatewise increasing,

$$\operatorname{Cov}(f(X_1,\ldots,X_n), g(X_1,\ldots,X_n)) \ge 0,$$

whenever this covariance exists.

For associated variables there exit some general inequalities justifying the use of assumptions on the covariance structure. One of such inequalities, useful in the sequel, appears in Dewan, Prakasa Rao [6] and is a generalization of an earlier result by Newman [13]. This inequality states a version for generating functions of Newman's [12] inequality for characteristic functions.

**Lemma 2.1.** Let  $X_1, \ldots, X_n$  be associated random variables bounded by a constant M. Then, for every  $\theta > 0$ ,

$$\left| \mathbb{E} \left( e^{\theta \sum_{i=1}^{n} X_{i}} \right) - \prod_{i=1}^{n} \mathbb{E} \left( e^{\theta X_{i}} \right) \right| \leq \theta^{2} e^{n\theta M} \sum_{1 \leq i < j \leq n} \operatorname{Cov}(X_{i}, X_{j}).$$

This inequality was used in Dewan, Prakasa Rao [6] to prove an exponential convergence rate for the nonparametric estimator of the density, but the method used to control all the terms involved forced the authors to assume a condition that is unattainable for associated variables. The same inequality was later re-used in Henriques, Oliveira [9] to prove another version of such exponential rate.

We quote next a general lemma used to control some of the terms appearing in course of proof.

**Lemma 2.2** (Devroye [5]). Let X be a centered random variable. If there exist  $a, b \in \mathbb{R}$  such that  $P(a \leq X \leq b) = 1$ , then, for every  $\lambda > 0$ ,

$$I\!\!E(e^{\lambda X}) \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

Next we introduce the notations that will be used throughout the text. Let  $c_n, n \ge 1$ , be a sequence of nonnegative real numbers such that  $c_n \longrightarrow +\infty$  and, given the random variables  $X_n, n \ge 1$ , define, for each  $i, n \ge 1$ ,

$$X_{1,i,n} = -c_n I_{(-\infty,-c_n)}(X_i) + X_i I_{[-c_n,c_n]}(X_i) + c_n I_{(c_n,+\infty)}(X_i),$$

$$X_{2,i,n} = (X_i - c_n) I_{(c_n,+\infty)}(X_i), \qquad X_{3,i,n} = (X_i + c_n) I_{(-\infty,-c_n)}(X_i),$$
(1)

where  $I_A$  represents the characteristic function of the set A. For each  $n \ge 1$  fixed, the variables  $X_{1,1,n}, \ldots, X_{1,n,n}$  are uniformly bounded, thus they may be treated using Lemma 2.1. Note that, for each  $n \ge 1$  fixed, all these variables are monotone functions of the initial variables  $X_n$ . This implies that an association assumption is kept by this construction.

The proof of an exponential inequality will use, besides the truncation introduced before, a convenient decomposition into blocks of the sums. For this purpose consider a sequence of natural numbers  $p_n$  such that, for each

 $n \geq 1$ ,  $p_n < \frac{n}{2}$  and define  $r_n$  as the greatest integer less or equal to  $\frac{n}{2p_n}$ . Define then, for q = 1, 2, 3, and  $j = 1, \ldots, 2r_n$ ,

$$Y_{q,j,n} = \sum_{l=(j-1)p_n+1}^{jp_n} \left( X_{q,l,n} - I\!\!E(X_{q,l,n}) \right).$$
(2)

Finally, for each q = 1, 2, 3, and  $n \ge 1$ , define

$$Z_{q,n,od} = \sum_{j=1}^{r_n} Y_{q,2j-1,n}, \qquad Z_{q,n,ev} = \sum_{j=1}^{r_n} Y_{q,2j,n},$$

$$R_{q,n} = \sum_{l=2r_n p_n+1}^n \left( X_{q,l,n} - I\!\!E(X_{q,l,n}) \right).$$
(3)

The proof of the main result is now divided into the control of the bounded terms, corresponding to the index q = 1, and the control of the nonbounded terms that correspond to the indexes q = 2 and q = 3.

# 3. Controlling the bounded terms

Given the definitions (1) and (2) it is obvious that  $|Y_{1,j,n}| \leq 2p_n c_n$ ,  $j = 1, \ldots, r_n$ . This enables the use of Lemma 2.2 to control the Laplace transform of these variables. A straightforward application of this lemma produces the following upper bounds.

**Lemma 3.1.** Let  $X_1, X_2, \ldots$  be random variables. If  $Y_{1,j,n}$ ,  $j = 1, \ldots, 2r_n$  are defined by (2) then, for every  $\lambda > 0$ ,

$$\prod_{j=1}^{r_n} I\!\!E\left(e^{\lambda Y_{1,2j-1,n}}\right) \le \exp\left(2\lambda^2 r_n p_n^2 c_n^2\right),$$
$$\prod_{j=1}^{r_n} I\!\!E\left(e^{\lambda Y_{1,2j,n}}\right) \le \exp\left(2\lambda^2 r_n p_n^2 c_n^2\right).$$

As it was done in Ioannides, Roussas [10] and Henriques, Oliveira [8, 9] we will be interested in controlling the differences between the Laplace transform of a sum of variables and what we would find if the variables were independent, which are the terms appearing in the left side of the inequalities stated in the previous lemma. This control is achieved by summing the odd indexed

terms on one side and the even indexed terms on the other side, as was done in Henriques, Oliveira [9].

**Lemma 3.2.** Let  $X_1, X_2, \ldots$  be strictly stationary and associated random variables. With the definitions (1), (2) and (3), for every  $\lambda > 0$ 

and analogously for the term corresponding to  $Z_{1,n,ev}$ .

*Proof*: According to (3) and the fact that the variables defined in (1) are associated we find, from a direct application of Lemma 2.1,

Using the stationarity of the variables it follows that

$$\sum_{1 \le j < j' \le r_n} \operatorname{Cov}(Y_{1,2j-1,n}, Y_{1,2j'-1,n}) = \sum_{j=1}^{r_n-1} (r_n - j) \operatorname{Cov}(Y_{1,1,n}, Y_{1,2j-1,n}).$$

A further invocation of the stationarity implies that

$$\begin{aligned} \operatorname{Cov}(Y_{1,1,n},Y_{1,2j-1,n}) &= \\ &= \sum_{l=0}^{p_n-1} (p_n-l) \operatorname{Cov}(X_{1,1,n},X_{1,2jp_n+l+1,n}) + \\ &\quad + \sum_{l=1}^{p_n-1} (p_n-l) \operatorname{Cov}(X_{1,l+1,n},X_{1,2jp_n+1,n}) \leq \\ &\leq p_n \sum_{l=(2j-1)p_n+2}^{(2j+1)p_n} \operatorname{Cov}(X_{1,1,n},X_{1,l,n}) \,. \end{aligned}$$

We now analyze the term with the covariances using the Hoeffding formula:

$$Cov(X_{1,i,n}, X_{1,j,n}) = \int_{\mathbb{R}^2} P(X_{1,i,n} > u, X_{1,j,n} > v) - P(X_{1,i,n} > u) P(X_{1,j,n} > v) du dv.$$
(5)

According to the truncation made, it easily follows that the integrand function vanishes outside the square  $[-c_n, c_n]^2$ . Moreover, for  $u, v \in [-c_n, c_n]$  we may replace, in the integrand function, the variables  $X_{1,i,n}, X_{1,j,n}$  by  $X_i, X_j$ , respectively, so that

$$Cov(X_{1,i,n}, X_{1,j,n}) = \\ = \int_{[-c_n, c_n]^2} P(X_i > u, X_j > v) - P(X_i > u) P(X_j > v) du dv \leq \\ \leq \int_{\mathbb{R}^2} P(X_i > u, X_j > v) - P(X_i > u) P(X_j > v) du dv = Cov(X_i, X_j),$$

due to the nonnegativity of the later integrand function, as follows from the association of the original variables. Inserting this into the inequalities stated before the lemma follows.

We may now prove an exponential inequality for the sum of odd indexed or even indexed terms.

**Lemma 3.3.** Let  $X_1, X_2, \ldots$  be strictly stationary and associated random variables. Suppose that

$$\frac{n}{p_n^2 c_n^4} \exp\left(\frac{2n}{p_n c_n}\right) \sum_{j=p_n+2}^{\infty} \operatorname{Cov}(X_1, X_j) \le C_0 < \infty.$$
(6)

Then, for every  $\varepsilon \in (0, 1)$ ,

$$P\left(\frac{1}{n}\left|Z_{1,n,od}\right| > \varepsilon\right) \le \left(1 + 4C_0\right) \exp\left(-\frac{n^2\varepsilon^2}{8r_n p_n^2 c_n^2}\right),\tag{7}$$

and analogously for  $Z_{1,n,ev}$ .

7

*Proof*: Applying Markov's inequality and using the previous lemma we find that, for every  $\lambda > 0$ ,

$$P\left(\frac{1}{n} |Z_{1,n,od}| > \varepsilon\right) \leq \\
 \leq \lambda^2 r_n p_n e^{\lambda(2r_n p_n c_n - n\varepsilon)} \sum_{j=p_n+2}^{(2r_n - 1)p_n} \operatorname{Cov}(X_1, X_l) + \exp\left(2\lambda^2 r_n p_n^2 c_n^2 - \lambda\varepsilon\right).$$

Optimizing the exponent in the last term of this upper bound we find  $\lambda = \frac{n\varepsilon}{4r_n p_n^2 c_n^2}$ , so that this exponent becomes equal to  $-\frac{n^2\varepsilon^2}{8r_n p_n^2 c_n^2}$ . Replacing this choice of  $\lambda$  into the first term of the upper bound and taking into account (6) it follows that

$$P\left(\frac{1}{n} |Z_{1,n,od}| > \varepsilon\right) \leq \\
 \leq 4C_0 \exp\left(-\frac{n^2 \varepsilon^2}{4r_n p_n^2 c_n^2}\right) + \exp\left(-\frac{n^2 \varepsilon^2}{8r_n p_n^2 c_n^2}\right) \leq \\
 \leq (1+4C_0) \exp\left(-\frac{n^2 \varepsilon^2}{8r_n p_n^2 c_n^2}\right).$$

To complete the treatment of the bounded terms it remains to control the sum corresponding to the indexes after  $2r_np_n$ , that is,  $R_{1,n}$ .

**Lemma 3.4.** Let  $X_1, X_2, \ldots$  be strictly stationary associated variables and suppose that

$$\frac{n}{p_n c_n} \longrightarrow +\infty \,. \tag{8}$$

Then, with the definitions made in (3), for n large enough and every  $\varepsilon > 0$ ,  $P(|R_{1,n}| > n\varepsilon) = 0$ .

*Proof*: As  $R_{1,n} = \sum_{l=2r_np_n+1}^n \left( X_{1,l,n} - I\!\!E(X_{1,l,n}) \right)$  it follows that  $|R_{1,n}| \leq 2(n-2r_np_n)c_n \leq 4c_n$ , according to the construction of the sequences  $r_n$  and  $p_n$ . Now  $P(|R_{1,n}| > n\varepsilon) \leq P\left(4 > \frac{n\varepsilon}{p_nc_n}\right)$  and, using (8), this is zero for n large enough.

We may now state a theorem summarizing the partial results described in the lemmas of this section.

**Theorem 3.5.** Let  $X_1, X_2, \ldots$  be strictly stationary and associated variables satisfying (6) and (8). With the definitions (1), (2) and (3) it follows that, for every  $\varepsilon \in (0, 1)$  and n large enough,

$$\left| P\left(\frac{1}{n} \left| \sum_{i=1}^{n} \left( X_{1,i,n} - I\!\!E(X_{1,i,n}) \right) \right| > \varepsilon \right) \le 2(1 + 4C_0) \exp\left(-\frac{n^2 \varepsilon^2}{72r_n p_n^2 c_n^2}\right). \quad (9)$$

*Proof*: It suffices to write

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} \left(X_{1,i,n} - I\!\!E(X_{1,i,n})\right)\right| > \varepsilon\right) \leq \\
 \leq \mathbb{P}\left(\frac{1}{n}\left|Z_{1,n,od}\right| > \frac{\varepsilon}{3}\right) + \mathbb{P}\left(\frac{1}{n}\left|Z_{1,n,ev}\right| > \frac{\varepsilon}{3}\right) + \mathbb{P}\left(|R_{1,n}| > \frac{n\varepsilon}{3}\right)$$

and apply the previous lemmas.

Note that (8) does not imply the convergence to zero of the upper bound in (9). In fact, for this inequality to be really useful for proving almost sure convergence we need to assume that

$$\frac{n}{p_n c_n^2} \longrightarrow +\infty, \qquad (10)$$

as  $2r_n p_n \approx n$ . This assumption will then imply (8) as we have chosen  $c_n \longrightarrow +\infty$ . Nevertheless, (8) is sufficient to derive inequality (9).

## 4. Controlling the unbounded terms

The variables  $X_{2,i,n}$  and  $X_{3,i,n}$  are associated but not bounded, even for fixed n. This means that Lemma 2.1 may not be applied to the sum of such terms. But we may note that these variables depend only on the tails of the distribution of the original variables. So, by controlling the decrease rate of these tails we may prove an exponential inequality for sums of  $X_{2,i,n}$  or  $X_{3,i,n}$ . For this control we will not make use of the block decomposition of the sums  $\sum_{i=1}^{n} \left( X_{q,i,n} - \mathbb{E}(X_{q,i,n}) \right)$  as the condition derived would be exactly the same as the one obtained with a direct treatment (the upper bound derived would be the same, up to the multiplication by a constant). We have, for q = 2, 3, recalling that the variables are identically distributed,

$$P\left(\left|\sum_{i=1}^{n} \left(X_{q,i,n} - I\!\!E(X_{q,i,n})\right)\right| > n\varepsilon\right) \le \\
 \le n P\left(|X_{q,1,n} - I\!\!E(X_{q,1,n})| > \varepsilon\right) \le \frac{n}{\varepsilon^2} \operatorname{Var}(X_{q,1,n}) \le \frac{n}{\varepsilon^2} I\!\!E(X_{q,1,n}^2)$$

**Lemma 4.1.** Let  $X_1, X_2, \ldots$  be strictly stationary random variables such that there exists  $\delta > 0$  satisfying  $\sup_{|t| \leq \delta} I\!\!E(e^{tX_1}) \leq M_{\delta} < +\infty$ . Then, with the definitions (1), for  $t \in (0, \delta]$ ,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \left(X_{q,i,n} - I\!\!E(X_{q,i,n})\right)\right| > n\varepsilon\right) \le \frac{2M_{\delta}ne^{-tc_n}}{t^2\varepsilon^2}, \qquad q = 2, 3. \tag{11}$$

*Proof*: According to the inequality stated before this lemma it remains to control  $\mathbb{E}(X_{q,1,n}^2)$ . Let us fix q = 2, the other possible choice being treated analogously. We will denote  $\overline{F}(x) = P(X_1 > x)$ . Now, using Markov's inequality it follows that, for  $t \in (0, \delta)$ ,  $\overline{F}(x) \leq e^{-tx}\mathbb{E}(e^{tX_1}) \leq M_{\delta}e^{-tx}$ . Writing the mathematical expectation as a Stieltjes integral and integrating by parts we find

$$I\!\!E(X_{2,1,n}^2) = -\int_{(c_n,+\infty)} (x-c_n)^2 \overline{F}(dx) = \int_{c_n}^{+\infty} 2(x-c_n)\overline{F}(x) \, dx \le 2M_\delta \frac{e^{-tc_n}}{t^2},$$

from which the lemma follows.

Note that for this step the association of the variables is irrelevant.

# 5. Main result

This last section only summarizes the results obtained. Besides the assumptions already used we need an extra condition on the truncating sequence requiring a minimal increasing rate.

**Theorem 5.1.** Let  $X_1, X_2, \ldots$  be strictly stationary and associated random variables satisfying (6), (8) with  $c_n^3 > c' \frac{n}{p_n}$ , for some c' > 0, and there exists  $\delta > 0$  satisfying  $\sup_{|t| \le \delta} \mathbb{E}(e^{tX_1}) \le M_{\delta} < +\infty$ . Then, with the definitions

(1), (2) and (3), for  $\varepsilon \in (0, 1)$  and n large enough,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \left(X_{i} - I\!\!E(X_{i})\right)\right| > \varepsilon\right) \leq \left(2(1+4C_{0}) + \frac{186\,642M_{\delta}p_{n}c_{n}^{6}}{\varepsilon^{2}}\right) \exp\left(-\frac{n^{2}\varepsilon^{2}}{72r_{n}p_{n}^{2}c_{n}^{2}}\right).$$

$$(12)$$

Proof: Separate the sum in the left of (12) into three terms, apply (9) and (11) with  $\frac{\varepsilon}{3}$  in place of  $\varepsilon$  and choose  $t = \frac{n^2 \varepsilon^2}{288 r_n p_n^2 c_n^3}$  in (11) so that the exponents are equal. Note that we must have  $t = \frac{n}{p_n c_n} \frac{n\varepsilon}{288 r_n p_n c_n^2} \leq \delta$ , thus  $\frac{n}{r_n p_n c_n^2}$  should compensate the unboundedness of  $\frac{n}{p_n c_n}$ . The construction of the sequences  $r_n$  and  $p_n$  implies that  $\frac{n}{r_n p_n} \leq 4$  so it suffices that  $\frac{72\varepsilon}{c_n^3} \leq \frac{\delta p_n}{n}$ , which is achieved for n large enough according to the assumptions made.

Note that, as already remarked at the end of section 3, in order to make (12) really useful for proving almost sure convergence results, we should strengthen (8), assuming (10) instead.

## References

- K. Azuma (1967). Weighted sums of certain dependent random variables. Tôhoku Math. J. 19, 357–367.
- [2] T. Birkel (1988). The invariance principle for associated processes. *Stochastic Process. Appl.* 27, 57–71.
- [3] M. Carbon (1993). Une nouvelle inégalité de grandes deviations. Applications. Publ. IRMA Lille 32, XI.
- [4] D. Bosq (1996). Nonparametric statistics for stochastic processes. Lecture Notes in Statistics 110, Springer.
- [5] L. Devroye (1991). Exponential inequalities in nonparametric estimation. In: Nonparametric Functional Estimation and Related Topics, G. Roussas, ed., Kluwer Academic Publishers, Dordrecht, 31–44.
- [6] I. Dewan and B. L. S. Prakasa Rao (1999), A general method of density estimation for associated random variables. J. Nonparametr. Statist. 10, 405–420.
- [7] J. D. Esary, F. Proschan and D. W. Walkup (1967). Association of random variables, with applications. Ann. Math. Statist. 38, 1466–1474.
- [8] C. Henriques, P. E. Oliveira (2002). Convergence rates for the estimation of two-dimensional distribution functions under association and estimation of the covariance of the limit empirical process. Preprint. Pré-Publicações do Departamento de Matemática da Universidade de Coimbra 02–05.
- [9] C. Henriques, P. E. Oliveira (2002). Exponential rates for kernel density estimation under association. Preprint. Pré-Publicações do Departamento de Matemática da Universidade de Coimbra 02–23.

10

- [10] D. A. Ioannides, G. G. Roussas (1998). Exponencial inequality for associated random variables. Statist. Probab. Lett. 42, 423–431.
- [11] E. Lesigne, D. Volný (2001). Large deviations for martingales. Stochastic Process. Appl. 96, 143–159.
- [12] C. Newman (1984). Asymptotic independence and limit theorems for positively and negatively dependent random variables. In Y. L. Tong (ed) Inequalities in Statistics and Probability, IMS Lecture Notes - Monograph Series Vol. 5, 127–140.
- [13] C. Newman (1980). Normal fluctuation and the FKG inequalities. Comm. Math. Phys. 74, 119–128.

Paulo Eduardo Oliveira

Dep. Matemática, Univ. Coimbra, Apartado 3008, 3001 - 454 Coimbra, Portugal