C. R. Acad. Sci. Paris, Ser. I 334 (2002) 495–500

Probabilités/*Probability Theory*

A Bennett concentration inequality and its application to suprema of empirical processes

Olivier Bousquet

CMAP, École polytechnique, 91128 Palaiseau, France

Received 17 January 2002; accepted 18 January 2002

Note presented by Michel Talagrand.

Abstract We introduce new concentration inequalities for functions on product spaces. They allow to obtain a Bennett type deviation bound for suprema of empirical processes indexed by upper bounded functions. The result is an improvement on Rio's version [6] of Talagrand's inequality [7] for equidistributed variables. ***To cite this article: O. Bousquet, C. R. Acad. Sci. Paris, Ser. I 334 (2002) 495–500.*** © 2002 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Une inégalité de concentration de type Bennett et son application aux maxima de processus empiriques

Résumé Nous proposons deux inégalités de concentration pour des fonctions de n variables indépendantes. L'une d'elles permet d'obtenir une inégalité de déviation de type Bennett pour les processus empiriques indexés par des classes de fonctions bornées à droite. Cela améliore la version donnée par Rio [6] de l'inégalité de Talagrand [7] pour des observations équidistribuées. ***Pour citer cet article : O. Bousquet, C. R. Acad. Sci. Paris, Ser. I 334 (2002) 495–500.*** © 2002 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Version française abrégée

1. Introduction

Soient $X_1, \ldots, X_n$ n variables indépendantes de loi jointe P à valeurs dans un espace polonais $\mathcal{X}$. Soit F une fonction mesurable de $\mathcal{X}^n$ dans $\mathbb{R}$. Nous étudions la concentration de la variable aléatoire $Z := F(X_1, \ldots, X_n)$ par rapport à sa moyenne. Nous donnons des conditions sur Z qui permettent d'obtenir une inégalité exponentielle de type Bennett ainsi que des conditions plus générales qui permettent d'obtenir un contrôle de la transformée de Laplace de type Bernstein. Nous montrons ensuite que ces résultats s'appliquent aux maxima de processus empiriques, c'est-à-dire aux variables aléatoires $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$ telles que $\mathcal{F}$ soit un ensemble de fonctions mesurables de $\mathcal{X}$ dans $\mathbb{R}$ de carré intégrable sous P.

$\mathcal{A}$ est la σ-algèbre engendrée par $(X_1, \ldots, X_n)$ et pour tout $k \in \{1, \ldots, n\}$, $\mathcal{A}_n^k$ est la σ-algèbre engendrée par $(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)$. $\mathbb{E}_n^k[\cdot]$ est l'éspérance par rapport à $\mathcal{A}_n^k$.

E-mail address: bousquet@cmapx.polytechnique.fr (O. Bousquet).

S1631-073X(02)02292-6/FLA

2. Résultats principaux

Le premier résultat donne une inégalité de type Bennett qui peut être considérée comme une généralisation d'un théorème de Boucheron, Lugosi et Massart [1].

THÉORÈME 2.1. – *Soient* $(Z, Z'_1, \ldots, Z'_n)$ *une séquence de v.a.* $\mathcal{A}$*-mesurables et* $(Z_k)_{k=1,\ldots,n}$ *une séquence de v.a. respectivement* $\mathcal{A}_n^k$*-mesurables. Supposons qu'il existe un réel positif* u *tel que, pour tout* $k = 1, \ldots, n$ *les inégalités suivantes sont vraies* :

$$Z'_k \leqslant Z - Z_k \leqslant 1 \quad p.s., \quad \mathbb{E}_n^k\big[Z'_k\big] \geqslant 0 \quad et \quad Z'_k \leqslant u \quad p.s. \tag{1}$$

Soit alors σ *un réel tel que* $\sigma^2 \geqslant \frac{1}{n}\sum_{k=1}^n \mathbb{E}_n^k[(Z'_k)^2]$ *presque sûrement. Posons* $v = (1+u)\mathbb{E}[Z] + n\sigma^2$, $\psi(\lambda) = \exp(-\lambda) - 1 + \lambda$ *et* $h(x) = (1+x)\log(1+x) - x$. *Si la condition suivante est vérifiée* :

$$\sum_{k=1}^n Z - Z_k \leqslant Z \quad p.s.,$$

nous obtenons pour tout $\lambda \geqslant 0$,

$$\log \mathbb{E}\big[e^{\lambda(Z-\mathbb{E}[Z])}\big] \leqslant \psi(-\lambda)v,$$

ce qui donne, pour tout $x \geqslant 0$

$$\mathbb{P}\big[Z \geqslant \mathbb{E}[Z] + x\big] \leqslant \exp\left(-vh\left(\frac{x}{v}\right)\right).$$

et

$$\mathbb{P}\left[Z \geqslant \mathbb{E}[Z] + \sqrt{2vx} + \frac{x}{3}\right] \leqslant e^{-x}.$$

Le second résultat donne un contrôle de type Bernstein de la transformée de Laplace de Z à partir de conditions moins restrictives que précédemment.

THÉORÈME 2.2. – *Avec les notations du Théorème* 2.1*, et sous la condition* (1)*, lorsqu'il existe deux variable aléatoires* $\mathcal{A}$*-mesurables* V *et* W *telles que*

$$\sum_{k=1}^n Z - Z_k \leqslant V \quad p.s. \quad et \quad \sum_{k=1}^n \mathbb{E}\big[\big(Z'_k\big)^2\big] \leqslant W \quad p.s.,$$

alors, pour tout $\theta > 0$*, et tout* $\lambda \in [0, (1+u)/\theta)$,

$$\log \mathbb{E}\big[e^{\lambda(Z-\mathbb{E}[Z])}\big] \leqslant \frac{\lambda}{1 - \lambda\theta/(1+u)}\left(\log \mathbb{E}\big[e^{\lambda V}\big] + \frac{\theta}{1+u}\log \mathbb{E}\big[e^{\lambda W/\theta}\big]\right).$$

Enfin nous appliquons le Théorème 2.1 pour obtenir un contrôle des déviations au dessus de sa moyenne du supremum d'un processus empirique indexé par une classe de fonctions bornées ou simplement bornées à droite.

THÉORÈME 2.3. – *Supposons les* X_i *equidistribués selon* P. *Soit* $\mathcal{F}$ *un ensemble dénombrable de fonctions de* $\mathcal{X}$ *dans* $\mathbb{R}$ *de carré intégrable et d'espérance nulle sous* P. *Si* $\sup_{f\in\mathcal{F}} \operatorname{ess\,sup} f \leqslant 1$ *alors on définit*

$$Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i),$$

et si $\sup_{f\in\mathcal{F}} \|f\|_\infty \leqslant 1$ *on définit*

$$Z = \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n f(X_i)\right|.$$

Soit alors σ *un réel tel que* $\sigma^2 \geqslant \sup_{f\in\mathcal{F}} \mathrm{Var}[f(X_1)]$ *presque sûrement, alors pour tout* $x \geqslant 0$*, on a*

$$\mathbb{P}\big[Z \geqslant \mathbb{E}[Z] + x\big] \leqslant \exp\left(-vh\left(\frac{x}{v}\right)\right),$$

où $v = n\sigma^2 + 2\mathbb{E}[Z]$ *et aussi*

$$\mathbb{P}\left[Z \geqslant \mathbb{E}[Z] + \sqrt{2xv} + \frac{x}{3}\right] \leqslant \mathrm{e}^{-x}.$$

1. Introduction

We consider a sequence of independent random variables $X_1, \ldots, X_n$ with values in some polish space $\mathcal{X}$ and distributed according to P. Let F be a P-measurable function from $\mathcal{X}^n$ to $\mathbb{R}$. We are interested in the conditions that the random variable $Z = F(X_1, \ldots, X_n)$ should satisfy in order to be concentrated around its expectation. We provide two theorems that give upper bounds on the Laplace transform of Z under general conditions. We prove that these conditions are satisfied in particular by suprema of empirical processes, i.e., $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$ with $\mathcal{F}$ a countable family of P-measurable functions. This allows us to get a Bennett inequality for such random variables which improves on Rio's version of Talagrand's inequality.

For all $k = 1, \ldots, n$, let $\mathcal{A}_k$ be the sigma field generated by $(X_1, \ldots, X_k)$ and let $\mathcal{A}_n^k$ be the sigma field generated by $(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)$. We denote by $\mathbb{E}_n^k[\cdot]$ the expectation taken conditionally to $\mathcal{A}_n^k$. Let $h(x) = (1+x)\log(1+x) - x$, $\psi(x) = \mathrm{e}^{-x} - 1 + x$ and $\phi(x) = 1 - (1+x)\mathrm{e}^{-x}$.

2. Main results

The first result can be considered as a generalization of a result of Boucheron, Lugosi and Massart [1] since it gives a Bennett type concentration inequality for Z under less restrictive conditions.

THEOREM 2.1. – *Let* $(Z, Z'_1, \ldots, Z'_n)$ *be a sequence of* $\mathcal{A}$*-measurable r.v. and let* $(Z_k)_{k=1,\ldots,n}$ *be a sequence of r.v. respectively* $\mathcal{A}_n^k$*-measurable. Assume that there exists* $u > 0$ *such that for all* $k = 1, \ldots, n$ *the following inequalities are satisfied*

$$Z'_k \leqslant Z - Z_k \leqslant 1 \quad a.s., \quad \mathbb{E}_n^k\big[Z'_k\big] \geqslant 0 \quad and \quad Z'_k \leqslant u \quad a.s. \tag{2}$$

Let σ *be a real satisfying* $\sigma^2 \geqslant \frac{1}{n}\sum_{k=1}^n \mathbb{E}_n^k[(Z'_k)^2]$ *almost surely and let* $v = (1+u)\mathbb{E}[Z] + n\sigma^2$*. If the following condition holds*

$$\sum_{k=1}^n Z - Z_k \leqslant Z \quad a.s., \tag{3}$$

we obtain, for all $\lambda \geqslant 0$*,*

$$\log \mathbb{E}\big[\mathrm{e}^{\lambda(Z-\mathbb{E}[Z])}\big] \leqslant \psi(-\lambda)v,$$

which gives the following bounds for all $x > 0$*,*

$$\mathbb{P}\big[Z \geqslant \mathbb{E}[Z] + x\big] \leqslant \exp\left(-vh\left(\frac{x}{v}\right)\right) \quad and \quad \mathbb{P}\left[Z \geqslant \mathbb{E}[Z] + \sqrt{2vx} + \frac{x}{3}\right] \leqslant \mathrm{e}^{-x}.$$

The second result relaxes further the conditions on Z. This allows us to obtain upper bounds on the Laplace transform of Z of Bernstein type, provided one controls the Laplace transform of two quantities, the first being the sum of first order finite differences and the second being the sum of the squares of these differences. This result can be considered as a refinement of one of the results in [2].

THEOREM 2.2. – *With the notations of Theorem* 2.1*, when the condition* (2) *is satisfied, denoting by* V *and* W *two* $\mathcal{A}$*-measurable random variables such that*

$$\sum_{k=1}^{n} Z - Z_k \leqslant V \quad \textit{a.s.} \quad \textit{and} \quad \sum_{k=1}^{n} \mathbb{E}_n^k\left[\left(Z_k'\right)^2\right] \leqslant W \quad \textit{a.s.},$$

then for all $\theta > 0$*, and all* $\lambda \in [0, (1+u)/\theta)$ *we have*

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(Z-\mathbb{E}[Z])}\right] \leqslant \frac{\lambda}{1-\lambda\theta/(1+u)}\left(\log \mathbb{E}\left[\mathrm{e}^{\lambda V}\right] + \frac{\theta}{1+u}\log \mathbb{E}\left[\mathrm{e}^{\lambda W/\theta}\right]\right).$$

The next result is an application of Theorem 2.1 which gives a functional generalization of Bennett's inequality. More precisely, it gives a bound on the deviation above its mean of the supremum of an empirical process indexed by a class of upper bounded or bounded functions. The bound we obtain reduces to the classical Bennett's inequality for sums of i.i.d. random variables when the index set is a singleton.

THEOREM 2.3. – *Assume the* X_i *are identically distributed according to* P*. Let* $\mathcal{F}$ *be a countable set of functions from* $\mathcal{X}$ *to* $\mathbb{R}$ *and assume that all functions* f *in* $\mathcal{F}$ *are* P*-measurable, square-integrable and satisfy* $\mathbb{E}[f] = 0$*. If* $\sup_{f\in\mathcal{F}} \operatorname{ess\,sup} f \leqslant 1$ *then we denote*

$$Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i),$$

and if $\sup_{f\in\mathcal{F}} \|f\|_\infty \leqslant 1$ *we denote*

$$Z = \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^{n} f(X_i)\right|.$$

Let σ *be a positive real number such that* $\sigma^2 \geqslant \sup_{f\in\mathcal{F}} \operatorname{Var}[f(X_1)]$ *almost surely, then for all* $x \geqslant 0$*, we have*

$$\mathbb{P}\left[Z \geqslant \mathbb{E}[Z] + x\right] \leqslant \exp\left(-vh\left(\frac{x}{v}\right)\right),$$

with $v = n\sigma^2 + 2\mathbb{E}[Z]$ *and also*

$$\mathbb{P}\left[Z \geqslant \mathbb{E}[Z] + \sqrt{2xv} + \frac{x}{3}\right] \leqslant \mathrm{e}^{-x}.$$

This result improves the main result in [6] where the exponential rate is $-\frac{x}{2}\log(1+\frac{x}{v})$ in the first inequality and the factor of x in the second inequality is $1/2$ instead of $1/3$. It also provides a positive answer to the question raised in [4] about the possibility of obtaining a functional version of Bennett's inequality with optimal constants.

3. Sketch of the proofs

In order to derive concentration inequalities around the expectation of the random variables, we use the so-called *entropy method* introduced by Ledoux [3], and further refined by Massart [4] and Rio [5] among others. In particular, many applications of this method have been exposed in [1,2].

This method consists in obtaining bounds on the logarithmic Laplace transform of a random function on a product space from bounds on the first-order finite differences of this function. Two main steps are necessary. The first one consists in using the so-called tensorization property of entropy which allows to decompose the entropy of a function of n independent random variables into a sum of entropies with respect to each individual random variable. The second step uses a variational principle for the entropy to bound the entropy with respect to one variable in terms of a first-order partial difference. The result is a differential inequality involving the Laplace transform of the random function. Once integrated, this gives an upper bound on the log-Laplace transform which can be turned into a deviation inequality via the classical Markov's inequality.

The main tool in the entropy method is the following inequality by Massart [4].

LEMMA 3.1. – *Let Z be a $\mathcal{A}$-measurable random variable, then we have for all λ, and for any sequence $(Z_k)_{1\leqslant k\leqslant n}$ of respectively $\mathcal{A}_n^k$-measurable random variables,*

$$\lambda\mathbb{E}\left[Z\mathrm{e}^{\lambda Z}\right]-\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right]\log\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right]\leqslant\mathbb{E}\left[\sum_{k=1}^{n}\psi\left(\lambda(Z-Z_k)\right)\mathrm{e}^{\lambda Z}\right].\tag{4}$$

We will also use the following lemma (*see*, e.g., [4]) as a decoupling device.

LEMMA 3.2. – *If V and Z are two $\mathcal{A}$-measurable random variables, we have for any λ and any $\theta>0$,*

$$\lambda\mathbb{E}\left[V\,\mathrm{e}^{\lambda Z}\right]\leqslant\lambda\theta\mathbb{E}\left[Z\,\mathrm{e}^{\lambda Z}\right]-\theta\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right]\log\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right]+\theta\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right]\log\mathbb{E}\left[\mathrm{e}^{\lambda V/\theta}\right].$$

Using the above two lemmas, we can prove that if V is a $\mathcal{A}$-measurable r.v. such that $\sum_{k=1}^{n}Z-Z_k\leqslant V$ then we have for all $\lambda>0$,

$$\sum_{k=1}^{n}\mathbb{E}\left[\mathrm{e}^{\lambda Z}-\mathrm{e}^{\lambda Z_k}\right]\leqslant\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right]\log\mathbb{E}\left[\mathrm{e}^{\lambda V}\right].\tag{5}$$

Proof of Theorem 2.1. – We essentially refine an argument introduced by Rio [6]. We start with inequality (4) and upper bound the summands in the right-hand side term. A careful analysis of the properties of the functions ϕ and ψ leads to the following upper bound

$$\psi\left(\lambda(Z-Z_k)\right)\mathrm{e}^{\lambda Z}\leqslant\frac{\phi(-\lambda)}{\psi(-\lambda)+\lambda/(1+u)}\left(\mathrm{e}^{\lambda Z}-\mathrm{e}^{\lambda Z_k}+\lambda\,\mathrm{e}^{\lambda Z_k}\left((1+u)^{-1}\left(Z_k'\right)^2-Z_k'\right)\right).\tag{6}$$

Now, because of conditions (2) we have for any $\mathcal{A}$-measurable random variable U

$$\mathbb{E}\left[\mathrm{e}^{\lambda Z_k}U\right]\leqslant\mathbb{E}\left[\mathrm{e}^{\lambda Z}\mathbb{E}_n^k[U]\right].$$

Using this inequality and the fact that $\mathbb{E}[Z_k']\geqslant 0$ and summing up (6) for $k=1,\dots,n$ we obtain

$$\mathbb{E}\left[\sum_{k=1}^{n}\psi\left(\lambda(Z-Z_k)\right)\mathrm{e}^{\lambda Z}\right]\leqslant\frac{\phi(-\lambda)}{\psi(-\lambda)+\lambda/(1+u)}\mathbb{E}\left[\sum_{k=1}^{n}\left(\mathrm{e}^{\lambda Z}-\mathrm{e}^{\lambda Z_k}\right)+\frac{n\sigma^2\lambda}{1+u}\,\mathrm{e}^{\lambda Z}\right].$$

Using inequality (5) and plugging the result in inequality (4) we thus obtain a differential inequality that has to be satisfied by $F(\lambda)=\mathbb{E}[\mathrm{e}^{\lambda Z}]$, the Laplace transform of Z.

Integrating this inequality gives the upper bound on the Laplace transform and standard calculus using Markov's inequality gives the deviation bounds. □

Proof of Theorem 2.2. – Using a similar reasoning as in the proof of Theorem 2.1 we obtain

$$\mathbb{E}\left[\sum_{k=1}^{n}\psi\left(\lambda(Z-Z_k)\right)\mathrm{e}^{\lambda Z}\right]\leqslant\frac{\phi(-\lambda)}{\psi(-\lambda)+\lambda/(1+u)}\mathbb{E}\left[V\,\mathrm{e}^{\lambda Z}+\frac{\lambda}{1+u}W\,\mathrm{e}^{\lambda Z}\right].$$

Then we use Lemma 3.2 to decouple $\mathbb{E}[W\,\mathrm{e}^{\lambda Z}]$ and use proof techniques inspired by [2] to integrate the resulting differential inequality. In particular we use the properties of the log-Laplace transforms (convexity, 0 in 0) to obtain a simple upper bound. □

Proof of Theorem 2.3. – Theorem 2.3 is an easy consequence of Theorem 2.1. The proof simply consists in proving that the random variable Z satisfies conditions (2) and (3). For this purpose, considering the case where Z is defined with absolute values (the other case is treated in the same way), we define the following auxiliary random variables for all $k = 1, \ldots, n$,

$$Z_k = \sup_{f \in \mathcal{F}} \left| \sum_{i \neq k} f(X_i) \right| \quad \text{and} \quad Z'_k = \left| \sum_{i=1}^{n} f_k(X_i) \right| - Z_k,$$

where f_k denotes the function for which the supremum is reached in Z_k (we use f_0 for the function in Z). We then get

$$Z'_k \leqslant Z - Z_k \leqslant \left| \sum_{i=1}^{n} f_0(X_i) \right| - \left| \sum_{i \neq k} f_0(X_i) \right| \leqslant \left| f_0(X_k) \right| \leqslant 1 \quad \text{a.s.}$$

Moreover, we have

$$\mathbb{E}_n^k \left[Z'_k \right] \geqslant \left| \mathbb{E}_n^k \left[\sum_{i=1}^{n} f_k(X_i) \right] \right| - Z_k = 0,$$

which concludes the proof of (2) with $u = 1$. Also,

$$(n-1)Z = \left| \sum_{k=1}^{n} \sum_{i \neq k} f_0(X_i) \right| \leqslant \sum_{k=1}^{n} \left| \sum_{i \neq k} f_0(X_i) \right| \leqslant \sum_{k=1}^{n} Z_k,$$

which gives (3), and finally, since

$$\sum_{k=1}^{n} \mathbb{E}_n^k \left[\left(Z'_k \right)^2 \right] \leqslant \sum_{k=1}^{n} \mathbb{E}_n^k \left[f_k^2(X_k) \right] \leqslant n \sup_{f \in \mathcal{F}} \operatorname{Var} \left[f(X_1) \right],$$

we can choose σ as proposed. □

Acknowledgements. The author is grateful to Emmanuel Rio whose work motivated and inspired this study and for his helpful comments.

References

[1] S. Boucheron, G. Lugosi, P. Massart, A sharp concentration inequality with applications, Random Structures Algorithms 16 (3) (2000) 277–292.
[2] S. Boucheron, G. Lugosi, P. Massart, Concentration of measure based on logarithmic Sobolev inequalities, 2001 (submitted).
[3] M. Ledoux, On Talagrand's deviation inequalities for product measures, Probab. Statist. 1 (1996) 63–87.
[4] P. Massart, About the constants in Talagrand's inequality for empirical processes, Ann. Probab. 29 (2) (2000) 863–884.
[5] E. Rio, Inégalités de concentration pour les processus empiriques de classes de parties, Probab. Theory Related Fields 119 (2000) 163–175.
[6] E. Rio, Une inegalité de Bennett pour les maxima de processus empiriques, Colloque en l'honneur de J. Bretagnolle, D. Dacunha-Castelle et I. Ibragimov, 2001 (to appear).
[7] M. Talagrand, New concentration inequalities in product spaces, Invent. Math. 126 (1996) 503–563.