Journal of Inequalities in Pure and
Applied Mathematics

http://jipam.vu.edu.au/

Volume 4, Issue 1, Article 15, 2003

A BOUND ON THE DEVIATION PROBABILITY FOR SUMS OF NON-NEGATIVE RANDOM VARIABLES

ANDREAS MAURER

ADALBERTSTR. 55
D-80799 MUNICH, GERMANY.
andreasmaurer@compuserve.com

Received 13 December, 2002; accepted 8 January, 2003
Communicated by T. Mills

ABSTRACT. A simple bound is presented for the probability that the sum of nonnegative independent random variables is exceeded by its expectation by more than a positive number t. If the variables have the same expectation the bound is slightly weaker than the Bennett and Bernstein inequalities, otherwise it can be significantly stronger. The inequality extends to one-sidedly bounded martingale difference sequences.

Key words and phrases: Deviation bounds, Bernstein's inequality, Hoeffdings inequality.

2000 *Mathematics Subject Classification.* Primary 60G50, 60F10.

1. INTRODUCTION

Suppose that the $\{X_i\}_{i=1}^m$ are independent random variables with finite first and second moments and use the notation $S := \sum_i X_i$. Let $t > 0$. This note discusses the inequality

$$\Pr\left\{E[S] - S \geq t\right\} \leq \exp\left(\frac{-t^2}{2\sum_i E\left[X_i^2\right]}\right), \tag{1.1}$$

valid under the assumption that the X_i are non-negative.

Similar bounds have a history beginning in the nineteenth century with the results of Bienaymé and Chebyshev ([3]). Set $\sigma^2 = (1/m)\sum_i \left(E\left[X_i^2\right] - \left(E\left[X_i\right]\right)^2\right)$. The inequality

$$\Pr\left\{\left|E[S] - S\right| \geq m\epsilon\right\} \leq \frac{\sigma^2}{m\epsilon^2}$$

requires minimal assumptions on the distributions of the individual variables and, if applied to identically distributed variables, establishes the consistency of the theory of probability: If the X_i represent the numerical results of independent repetitions of some experiment, then the probability that the average result deviates from its expectation by more than a value of ϵ decreases to zero as as $\sigma^2/\left(m\epsilon^2\right)$, where σ^2 is the average variance of the X_i.

ISSN (electronic): 1443-5756

I would like to thank David McAllester, Colin McDiarmid, Terry Mills and Erhard Seiler for encouragement and help.

145-02

If the X_i satisfy some additional boundedness conditions the deviation probabilities can be shown to decrease exponentially. Corresponding results were obtained in the middle of the twentieth century by Bernstein [2], Cramér, Chernoff [4], Bennett [1] and Hoeffding [7]. Their results, summarized in [7], have since found important applications in statistics, operations research and computer science (see [6]). A general method of proof, sometimes called the exponential moment method, is explained in [10] and [8].

Inequality (1.1) is of a similar nature and can be directly compared to one-sided versions of Bernstein's and Bennett's inequalities (see Theorem 3 in [7]) which also require the X_i to be bounded on only one side. It turns out that, once reformulated for non-negative variables, the classical inequalities are stronger than (1.1) if the X_i are similar in the sense that their expectations are uniformly concentrated. If the expectations of the individual variables are very scattered and/or for large deviations t our inequality (1.1) becomes stronger.

Apart from being stronger than Bernstein's theorem under perhaps somewhat extreme circumstances, the new inequality (1.1) appears attractive because of its simplicity. The proof (suggested by Colin McDiarmid) is very easy and direct and the method also gives a concentration inequality for martingales of one-sidedly bounded differences.

In Section 2 we give a first proof of (1.1) and list some simple consequences. In Section 3 our result is compared to Bernstein's inequality, in Section 4 it is extended to martingales. All random variables below are assumed to be members of the algebra of measurable functions defined on some probability space (Ω, Σ, μ). Order and equality in this algebra are assumed to hold only almost everywhere w.r.t. μ, i.e. $X \geq 0$ means $X \geq 0$ almost everywhere w.r.t. μ on Ω.

2. Statement and Proof of the Main Result

Theorem 2.1. *Let the $\{X_i\}_{i=1}^m$ be independent random variables, $E\left[X_i^2\right] < \infty$, $X_i \geq 0$. Set $S = \sum_i X_i$ and let $t > 0$. Then*

$$\Pr\left\{E\left[S\right] - S \geq t\right\} \leq \exp\left(\frac{-t^2}{2\sum_i E\left[X_i^2\right]}\right). \tag{2.1}$$

Proof. We first claim that for $x \geq 0$

$$e^{-x} \leq 1 - x + \frac{1}{2}x^2.$$

To see this let $f(x) = e^{-x}$ and $g\left(x\right) = 1 - x + \left(1/2\right)x^2$ and recall that for every real x

$$e^x \geq 1 + x \tag{2.2}$$

so that $f'(x) = -e^{-x} \leq -1 + x = g'\left(x\right)$. Since $f\left(0\right) = 1 = g\left(0\right)$ this implies $f\left(x\right) \leq g\left(x\right)$ for all $x \geq 0$, as claimed.

It follows that for any $i \in \{1, \ldots, m\}$ and any $\beta \geq 0$ we have

$$E\left[e^{-\beta X_i}\right] \leq 1 - \beta E\left[X_i\right] + \frac{\beta^2}{2}E\left[X_i^2\right] \leq \exp\left(-\beta E\left[X_i\right] + \frac{\beta^2}{2}E\left[X_i^2\right]\right),$$

where (2.2) was used again in the second inequality. This establishes the bound

$$\ln E\left[e^{-\beta X_i}\right] \leq -\beta E\left[X_i\right] + \frac{\beta^2}{2}E\left[X_i^2\right]. \tag{2.3}$$

Using the independence of the X_i this implies

$$\begin{aligned}\ln E\left[e^{-\beta S}\right] &= \ln \prod_i E\left[e^{-\beta X_i}\right]\\ &= \sum_i \ln E\left[e^{-\beta X_i}\right]\\ &\le -\beta E\left[S\right] + \frac{\beta^2}{2}\sum_i E\left[X_i^2\right]. \qquad (2.4)\end{aligned}$$

Let χ be the characteristic function of $[0,\infty)$. Then for any $\beta \ge 0$, $x \in \mathbb{R}$ we must have $\chi(x) \le \exp(\beta x)$ so, using (2.4),

$$\begin{aligned}\ln \Pr\left\{E\left[S\right] - S \ge t\right\} &= \ln E\left[\chi\left(-t + E\left[S\right] - S\right)\right]\\ &\le \ln E\left[\exp\left(\beta\left(-t + E\left[S\right] - S\right)\right)\right]\\ &= -\beta t + \beta E\left[S\right] + \ln E\left[e^{-\beta S}\right]\\ &\le -\beta t + \frac{\beta^2}{2}\sum_i E\left[X_i^2\right].\end{aligned}$$

We minimize the last expression with $\beta = t/\sum_i E\left[X_i^2\right] \ge 0$ to obtain

$$\ln \Pr\left\{E\left[S\right] - S \ge t\right\} \le \frac{-t^2}{2\sum_i E\left[X_i^2\right]},$$

which implies (2.1). □

Some immediate and obvious consequences are given in

Corollary 2.2. *Let the $\{X_i\}_{i=1}^m$ be independent random variables, $E\left[X_i^2\right] < \infty$. Set $S = \sum_i X_i$ and let $t > 0$.*

(1) *If $X_i \le b_i$ and set $\sigma_i^2 = E\left[X_i^2\right] - \left(E\left[X_i\right]\right)^2$ then*

$$\Pr\left\{S - E\left[S\right] \ge t\right\} \le \exp\left(\frac{-t^2}{2\sum_i \sigma_i^2 + 2\sum_i \left(b_i - E\left[X_i\right]\right)^2}\right).$$

(2) *If $0 \le X_i \le b_i$ then*

$$\Pr\left\{E\left[S\right] - S \ge t\right\} \le \exp\left(\frac{-t^2}{2\sum_i b_i E\left[X_i\right]}\right).$$

(3) *If $0 \le X_i \le b_i$ then*

$$\Pr\left\{E\left[S\right] - S \ge t\right\} \le \exp\left(\frac{-t^2}{2\sum_i b_i^2}\right)$$

Proof. (1) follows from application of Theorem 2.1 to the random variables $Y_i = b_i - X_i$ since

$$\begin{aligned}2\sum E\left[Y_i^2\right] &= 2\sum\left(E\left[X_i^2\right] - E\left[X_i\right]^2 + E\left[X_i\right]^2 - 2b_i E\left[X_i\right] + b_i^2\right)\\ &= 2\sum_i \sigma_i^2 + 2\sum_i \left(b_i - E\left[X_i\right]\right)^2,\end{aligned}$$

while (2) is immediate from Theorem 2.1 and (3) follows trivially from (2). □

3. COMPARISON TO OTHER BOUNDS

Observe that part (3) of Corollary 2.2 is similar to the familiar Hoeffding inequality (Theorem 2 in [7]) but weaker by a factor of 4 in the exponent. If there is information on the expectations of the X_i and $E\left[X_i\right] \leq b_i/4$ then (2) of Corollary 2.2 becomes stronger than Hoeffding's inequality. If the b_i are all equal then (2) is weaker than what we get from the relative-entropy Chernoff bound (Theorem 1 in [7]).

It is natural to compare our result to Bernstein's theorem which also requires only one-sided boundedness. We state a corresponding version of the theorem (see [1] or [10] or [9])

Theorem 3.1 (Bernstein's Inequality). *Let $\{X_i\}_{i=1}^m$ be independent random variables with $X_i - E\left[X_i\right] \leq d$ for all $i \in \{1, \ldots, m\}$. Let $S = \sum X_i$ and $t > 0$. Then, with $\sigma_i^2 = E\left[X_i^2\right] - E\left[X_i\right]^2$ we have*

$$\Pr\left\{S - E\left[S\right] \geq t\right\} \leq \exp\left(\frac{-t^2}{2\sum_i \sigma_i^2 + 2td/3}\right). \tag{3.1}$$

Now suppose we know $X_i \leq b_i$ for all i. In this case we can apply part (1) of Corollary 2.2. On the other hand if we set $d = \max_i\left(b_i - E\left[X_i\right]\right)$ then $X_i - E\left[X_i\right] \leq d$ for all i and we can apply Bernstein's theorem as well. The latter is evidently tighter than part (1) of Corollary 2.2 if and only if

$$\frac{t}{3}\max_i\left(b_i - E\left[X_i\right]\right) < \sum_i\left(b_i - E\left[X_i\right]\right)^2.$$

We introduce the abbreviations $B_\infty = \max_i\left(b_i - E\left[X_i\right]\right)$, $B_1 = \sum_i\left(b_i - E\left[X_i\right]\right)$ and $B_2 = \sum_i\left(b_i - E\left[X_i\right]\right)^2$. Both results are trivial unless $t < B_1$. Assume $t = \epsilon B_1$, where $0 < \epsilon < 1$, then Bernstein's theorem is stronger in the interval

$$0 < \epsilon < \frac{3B_2}{B_1 B_\infty},$$

which is never empty. The new inequality is stronger in the interval

$$\frac{3B_2}{B_1 B_\infty} < \epsilon < 1.$$

The latter interval may be empty, in which case Bernstein's inequality is stronger for all nontrivial deviations ϵ. This is clearly the case if all the $b_i - E\left[X_i\right]$ are equal, for then $B_2/\left(B_1 B_\infty\right) = 1$. This happens, for example, if the X_i are identically distributed. The fact that the new inequality can be stronger in a significant range of deviations may be seen if we set $E\left[X_i\right] = 0$ and $b_i = 1/i$ for $i \in \{1, \ldots, m\}$, then

$$\frac{3B_2}{B_1 B_\infty} < \frac{\pi^2}{2\sum_{i=1}^m (1/i)} \to 0 \text{ as } m \to \infty.$$

In this case, for every given deviation ϵ, the new inequality becomes stronger for sufficiently large m.

To summarize this comparison: If the deviation is small and/or the individual variables have a rather uniform behaviour, then Bernstein's inequality is stronger, otherwise weaker than the new result. A similar analysis applies to the stronger Bennett inequality and the yet stronger Theorem 3 in [7]. In all these cases a single uniform bound on the variables $X_i - E\left[X_i\right]$ enters into the bound on the deviation probability.

4. MARTINGALES

The key to the proof of Theorem 2.1 lies in inequality (2.3):

$$X \geq 0, \beta \geq 0 \Longrightarrow \ln E\left[e^{-\beta X}\right] \leq -\beta E\left[X\right] + \frac{\beta^2}{2} E\left[X^2\right].$$

Apart from the inequality $e^{-x} \leq 1 - x + (1/2)\, x^2$ (for non-negative x) its derivation uses only monotonicity, linearity and normalization of the expectation value. It therefore also applies to conditional expectations.

Lemma 4.1. *Let X, W be random variables, W not necessarily real valued, $\beta \geq 0$.*

(1) *If $X \geq 0$ then*

$$\ln E\left[e^{-\beta X}|W\right] \leq -\beta E\left[X|W\right] + \frac{\beta^2}{2} E\left[X^2|W\right].$$

(2) *If $X \leq b$ and $E\left[X|W\right] = 0$ and $E\left[X^2|W\right] \leq \sigma^2$ then*

$$\ln E\left[e^{\beta X}|W\right] \leq \frac{\beta^2}{2}\left(\sigma^2 + b^2\right).$$

Proof. To see part 1 retrace the first part of the proof of Theorem 2.1. Part 2 follows from applying part 1 to $Y = b - X$ to get

$$\begin{aligned} \ln E\left[e^{\beta X}|W\right] &= \beta b + \ln E\left[e^{-\beta Y}|W\right] \\ &\leq \beta b - \beta E\left[Y|W\right] + \frac{\beta^2}{2} E\left[Y^2|W\right] \\ &= \frac{\beta^2}{2} E\left[Y^2|W\right] = \frac{\beta^2}{2}\left(E\left[X^2|W\right] + b^2\right). \end{aligned}$$

□

Part (2) of this lemma gives a concentration inequality for martingales of one-sidedly bounded differences, with less restrictive assumptions than [5], Corollary 2.4.7.

Theorem 4.2. *Let X_i be random variables , $S_n = \sum_{i=1}^n X_i$, $S_0 = 0$. Suppose that $b_i, \sigma_i > 0$ and that $E\left[X_n|S_{n-1}\right] = 0$, $E\left[X_n^2|S_{n-1}\right] \leq \sigma_n^2$ and $X_n \leq b_n$, then, for $\beta \geq 0$,*

$$\ln E\left[e^{\beta S_n}\right] \leq \frac{\beta^2}{2} \sum_{i=1}^n \left(\sigma_i^2 + b_i^2\right) \tag{4.1}$$

and for $t > 0$,

$$\Pr\left\{S_n \geq t\right\} \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^n \left(\sigma_i^2 + b_i^2\right)}\right). \tag{4.2}$$

Proof. We prove (4.1) by induction on n. The case $n = 1$ is just part (2) of the lemma with $W = 0$. Assume that (4.1) holds for a given value of n. If Σ_n is the σ-algebra generated by S_n then $e^{\beta S_n}$ is Σ_n-measurable, so

$$E\left[e^{\beta S_{n+1}}|S_n\right] = E\left[e^{\beta S_n} e^{\beta X_{n+1}}|S_n\right] = e^{\beta S_n} E\left[e^{\beta X_{n+1}}|S_n\right]$$

almost surely. Thus,

$$\begin{aligned}
\ln E\left[e^{\beta S_{n+1}}\right] &= \ln E\left[E\left[e^{\beta S_{n+1}}|S_n\right]\right] \\
&= \ln E\left[e^{\beta S_n} E\left[e^{\beta X_{n+1}}|S_n\right]\right] \\
&\leq \ln E\left[e^{\beta S_n}\right] + \frac{\beta^2}{2}\left(\sigma_{n+1}^2 + b_{n+1}^2\right) \qquad (4.3) \\
&\leq \frac{\beta^2}{2}\sum_{i=1}^{n+1}\left(\sigma_i^2 + b_i^2\right), \qquad (4.4)
\end{aligned}$$

where Lemma 4.1, part 2 was used to get (4.3) and the induction hypothesis was used for (4.4).

To get (4.2), we proceed as in the proof of Theorem 2.1: For $\beta \geq 0$,

$$\ln \Pr\{S_n \geq t\} \leq \ln E\left[e^{\beta(S_n - t)}\right] \leq -\beta t + \frac{\beta^2}{2}\sum_{i=1}^{n}\left(\sigma_i^2 + b_i^2\right).$$

Minimizing the last expression with $\beta = t/\sum\left(\sigma_i^2 + b_i^2\right)$ gives (4.2). □

5. Conclusion

It remains to be seen if our inequality has any interesting practical implications. In view of the comparison to Bernstein's theorem this would have to be in a situation where the random variables considered have a highly non-uniform behaviour and the deviations to which the result is applied are large. Apart from its potential utility the new inequality may have some didactical value due to its simplicity.

References

[1] G. BENNETT, Probability inequalities for the sum of independent random variables, *J. Amer. Statist. Assoc.*, **57** (1962), 33–45.

[2] S. BERNSTEIN, *Theory of Probability*, Moscow, 1927.

[3] P. CHEBYCHEV, Sur les valeurs limites des intégrales, *J. Math. Pures Appl., Ser. 2*, **19** (1874), 157–160.

[4] H. CHERNOFF, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Annals of Mathematical Statistics*, **23** (1952), 493–507.

[5] A. DEMBO AND O. ZEITOUMI, *Large Deviation Techniques and Applications*, Springer 1998.

[6] L. DEVROYE, L. GYÖRFI AND G. LUGOSI, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[7] W. HOEFFDING, Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.*, **58** (1963), 13–30.

[8] D. McALLESTER AND L. ORTIZ, Concentration inequalities for the missing mass and for histogram rule error, *NIPS*, 2002.

[9] C. McDIARMID, Concentration, in *Probabilistic Methods of Algorithmic Discrete Mathematics*, Springer, Berlin, 1998, p. 195–248.

[10] H. WITTING AND U. MÜLLER–FUNK, *Mathematische Statistik*, Teubner Stuttgart, 1995.