Electronic Journal of Probability

Vol. 14 (2009), Paper no. 64, pages 1884–1899.

Journal URL
http://www.math.washington.edu/~ejpecp/

On concentration of self-bounding functions

Stéphane Boucheron*
LPMA
Université Paris-Diderot,
Paris
stephane.boucheron@math.jussieu.fr

Gábor Lugosi†
Icrea and
Pompeu Fabra University
Barcelona
gabor.lugosi@gmail.com

Pacal Massart
Université Paris-Sud
Orsay
pascal.massart@gmail.com

Abstract

We prove some new concentration inequalities for self-bounding functions using the entropy method. As an application, we recover Talagrand's convex distance inequality. The new Bernstein-like inequalities for self-bounding functions are derived thanks to a careful analysis of the so-called Herbst argument. The latter involves comparison results between solutions of differential inequalities that may be interesting in their own right.

Key words: concentration inequality, convex distance, self-bounding function.

AMS 2000 Subject Classification: Primary 60E15, 60C05, 28A3; Secondary: 05C80.

Submitted to EJP on June 12, 2009, final version accepted August 10, 2009.

*Supported by ANR grant TAMIS

†Supported by the Spanish Ministry of Science and Technology grant MTM2006-05650. and by the PASCAL Network of Excellence under EC grant no. 506778.

1 Introduction

Let $X_1,\ldots,X_n$ be independent random variables, taking values in some measurable space $\mathscr{X}$ and let $f:\mathscr{X}^n\to\mathbb{R}$ be a real-valued function of n variables. We are interested in concentration of the random variable $Z=f(X_1,\ldots,X_n)$ around its expected value. Well-known concentration inequalities establish quantitative bounds for the probability that Z deviates significantly from its mean under smoothness conditions on the function f, see, for example, Ledoux [9], McDiarmid [12] for surveys.

However, some simple conditions different from smoothness have been shown to guarantee concentration. Throughout the text, for each $i\le n$, f_i denotes a measurable function from $\mathscr{X}^{n-1}$ to $\mathbb{R}$. The following condition used by Boucheron, Lugosi, and Massart [2] generalizes the notion of a *configuration function* introduced by Talagrand [21].

Definition 1. *A function $f:\mathscr{X}^n\to\mathbb{R}$ is called* self-bounding *if for all $x=(x_1,\ldots,x_n)\in\mathscr{X}^n$,*

$$0\le f(x)-f_i(x^{(i)})\le 1\,,$$

and

$$\sum_{i=1}^{n}\left(f(x)-f_i(x^{(i)})\right)\le f(x)$$

where $x^{(i)}=(x_1,\ldots,x_{i-1},x_{i-1},\ldots,x_n)\in\mathscr{X}^{n-1}$ is obtained by dropping the i-th component of x.

It is shown in [2] that if f is self-bounding then Z satisfies, for all $\lambda\in\mathbb{R}$, the sub-Poissonian inequality

$$\log\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right]\le\left(e^{\lambda}-\lambda-1\right)\mathbb{E}Z$$

which implies that for every $t\ge 0$,

$$\mathbb{P}\{Z\ge\mathbb{E}Z+t\}\le\exp\left(\frac{-t^2}{2(\mathbb{E}Z+t/3)}\right)$$

and for all $0<t<\mathbb{E}Z$,

$$\mathbb{P}\{Z\le\mathbb{E}Z-t\}\le\exp\left(\frac{-t^2}{2\mathbb{E}Z}\right)\ .$$

An often convenient choice for f_i is

$$f_i(x^{(i)})=\inf_{x_i'\in\mathscr{X}}f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)\,. \tag{1}$$

Throughout the paper we implicitly assume that f_i is measurable. (McDiarmid and Reed [13, Lemma 5] point out that this is not a restrictive assumption).

Several generalizations of such inequalities have been proposed in the literature, see Boucheron, Lugosi, Massart [3], Boucheron, Bousquet, Lugosi, Massart [1], Devroye [5], Maurer [11], McDiarmid and Reed [13]. McDiarmid and Reed further generalize the notion of self-bounding functions.

Definition 2. *A function $f:\mathscr{X}^n\to\mathbb{R}$ is called* (a,b)-self-bounding *if for some $a,b>0$, for all $i=1,\ldots,n$ and all $x\in\mathscr{X}^n$,*

$$0\le f(x)-f_i(x^{(i)})\le 1\,,$$

and

$$\sum_{i=1}^{n}\left(f(x)-f_i(x^{(i)})\right)\le af(x)+b\,.$$

McDiarmid and Reed [13] show that under this condition, for all $t>0$,

$$\mathbb{P}\{Z\ge\mathbb{E}Z+t\}\le\exp\left(\frac{-t^2}{2(a\mathbb{E}Z+b+at)}\right)$$

and

$$\mathbb{P}\{Z\le\mathbb{E}Z-t\}\le\exp\left(\frac{-t^2}{2(a\mathbb{E}Z+b+t/3)}\right)\,.$$

Maurer [11] considers a even weaker notion.

Definition 3. *A function $f:\mathcal{X}^n\to\mathbb{R}$ is called* weakly (a,b)-self-bounding *if all $x\in\mathcal{X}^n$,*

$$\sum_{i=1}^{n}\left(f(x)-f_i(x^{(i)})\right)^2\le af(x)+b\,.$$

Of course, if f is (a,b)-self-bounding then it is also weakly (a,b)-self-bounding. Maurer [11, Theorem 13] proves that if f is weakly $(a,0)$ self-bounding, then

$$\mathbb{P}\{Z\ge\mathbb{E}Z+t\}\le\exp\left(\frac{-t^2}{2a\mathbb{E}Z+at}\right)\,.$$

If, in addition, $0\le f(x)-f_i(x^{(i)})\le 1$, for each $i\le n$ and each $x\in\mathcal{X}^n$ then

$$\mathbb{P}\{Z\le\mathbb{E}Z-t\}\le\exp\left(\frac{-t^2}{2\max(a,1)\mathbb{E}Z}\right)\,. \tag{2}$$

The purpose of this paper is to further sharpen these results. The proofs, just like for the above-mentioned inequalities, is based on the *entropy method* pioneered by Ledoux [8] and further developed, among others, by Boucheron, Lugosi, Massart [3], Boucheron, Bousquet, Lugosi, Massart [1], Bousquet [4], Klein [6], Massart [10], Rio [16], Klein and Rio [7]. We present some applications. In particular, we are able to recover Talagrand's celebrated convex distance inequality [21] for which no complete proof based on the entropy method has been available.

For any real number $a\in\mathbb{R}$, we denote by $a_+=\max(a,0)$ and $a_-=\max(-a,0)$ the positive and negative parts of a. The main result of the paper is the following.

Theorem 1. *Let $X=(X_1,\ldots,X_n)$ be a vector of independent random variables, each taking values in a measurable set $\mathcal{X}$ and let $f:\mathcal{X}^n\to\mathbb{R}$ be a non-negative measurable function such that $Z=f(X)$ has finite mean.*

For $a,b\ge 0$, define $c=(3a-1)/6$.

If f is (a,b)-self-bounding, *then for all $\lambda\ge 0$,*

$$\log\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right]\le\frac{(a\mathbb{E}Z+b)\lambda^2}{2(1-c_+\lambda)}$$

and for all $t > 0$*,*

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq \exp\left(-\frac{t^2}{2\left(a\mathbb{E}Z + b + c_+ t\right)}\right) .$$

If f *is* weakly (a,b)-self-bounding *and for all* $i \leq n$*, all* $x \in \mathcal{X}$*,* $f_i(x^{(i)}) \leq f(x)$*, then for all* $0 \leq \lambda \leq 2/a$*,*

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{(a\mathbb{E}Z + b)\lambda^2}{2(1 - a\lambda/2)}$$

and for all $t > 0$*,*

$$\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq \exp\left(-\frac{t^2}{2\left(a\mathbb{E}Z + b + at/2\right)}\right) .$$

If f *is* weakly (a,b)-self-bounding *and* $f(x) - f_i(x^{(i)}) \leq 1$ *for each* $i \leq n$ *and* $x \in \mathcal{X}^n$*, then for* $0 < t \leq \mathbb{E}Z$*,*

$$\mathbb{P}\{Z \leq \mathbb{E}Z - t\} \leq \exp\left(-\frac{t^2}{2\left(a\mathbb{E}Z + b + c_- t\right)}\right) .$$

The bounds of the theorem reflect an interesting asymmetry between the upper and lower tail estimates. If $a \geq 1/3$, the left tail is sub-Gaussian with variance proxy $a\mathbb{E}Z + b$. If $a \leq 1/3$, then the upper tail is sub-Gaussian. If $a = 1/3$ then we get purely sub-Gaussian estimates of both sides. Of course, if f is (a,b)-self-bounding for some $a \leq 1/3$ then it is also $(1/3,b)$-self-bounding, so for all values of $a \leq 1/3$, we obtain sub-Gaussian bounds, though for $a < 1/3$ the theorem does yield optimal constants in the denominator of the exponent. If $a \leq 1/3$ and f is weakly (a,b)-self-bounding, we thus have

$$\mathbb{P}\{Z \leq \mathbb{E}Z - t\} \leq \min\left(\exp\left(-\frac{t^2}{2(a\mathbb{E}Z + b + t(1-3a)/6)}\right), \exp\left(-\frac{t^2}{2((1/3)\mathbb{E}Z + b)}\right)\right) .$$

This type of phenomenon appears already in Maurer's bound (2) but the critical value of a is now improved from 1 to 1/3. We have no special reason to believe that the threshold value 1/3 is optimal but this is the best we get by our analysis.

Note that the bounds for the upper tail for weakly self-bounded random variables are due to Maurer [11]. They are recalled here for the sake of self-reference.

2 The convex distance inequality

In a remarkable series of papers (see [21],[19],[20]), Talagrand developed an induction method to prove powerful concentration results. Perhaps the most widely used of these is the so-called "convex-distance inequality." Recall first the definition of the *convex distance:*

In the sequel, $\|\cdot\|_2$ denotes the Euclidean norm. For any $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, let

$$d_T(x,A) = \sup_{\alpha \in [0,\infty)^n : \|\alpha\|_2 = 1} d_\alpha(x,A)$$

denote the convex distance of x from the set A where

$$d_\alpha(x,A) = \inf_{y \in A} d_\alpha(x,y) = \inf_{y \in A} \sum_{i : x_i \neq y_i} |\alpha_i|$$

is a weighted Hamming distance of x to the set A. Talagrand's convex distance inequality states that if X is an $\mathscr{X}^n$-valued vector of independent random variables, then for any set $A \subset \mathscr{X}$,

$$\mathbb{E}\left[e^{d_T(X,A)^2/4}\right] \leq \frac{1}{\mathbb{P}\{X \in A\}}$$

which implies, by Markov's inequality, that for any $t > 0$,

$$\mathbb{P}\{d_T(X,A) \geq t\} \cdot \mathbb{P}\{X \in A\} \leq e^{-t^2/4} .$$

Even though at the first sight it is not obvious how Talagrand's result can be used to prove concentration for general functions f of X, with relatively little work, the theorem may be converted into very useful inequalities. Talagrand [19], Steele [18], and Molloy and Reed [14] survey a large variety of applications. Pollard [15] revisits Talagrand's orginal proof in order to make it more transparent.

Several attempts have been made to recover Talagrand's convex distance inequality using the entropy method (see [3; 11; 13]). However, these attempts have only been able to partially recover Talagrand's result. In [3] we pointed out that the Efron-Stein inequality may be used to show that for all X and $A \subset \mathscr{X}^n$,

$$\mathsf{Var}\left(d_T(X,A)\right) \leq 1 .$$

The same argument was used to show that Talagrand's inequality holds (with slightly different constants) for sets A with $\mathbb{P}\{X \in A\} \geq 1/2$. Maurer [11] improved the constants but still fell short of proving it for all sets.

Here we show how Theorem 1 may be used to recover the convex distance inequality with a somewhat worse constant (10 instead of 4) in the exponent. Note that we do not use the full power of Theorem 1. In fact, Maurer's results may also be applied together with the argument below.

The main observation is that the square of the convex distance is self-bounding:

Lemma 1. *For any $A \in \mathscr{X}^n$ and $x \in \mathscr{X}^n$, the function $f(x) = d_T(x,A)^2$ satisfies $0 \leq f(x) - f_i(x^{(i)}) \leq 1$ where f_i is defined as in (1). Moreover, f is weakly $(4,0)$-self-bounding.*

Proof. The proof is based on different formulations of the convex distance. Let $\mathscr{M}(A)$ denote the set of probability measures on A. Then, using Sion's minimax theorem, we may re-write d_T as

$$d_T(x,A) = \inf_{\nu \in \mathscr{M}(A)} \sup_{\alpha : \|\alpha\|_2 \leq 1} \sum_{j=1}^{n} \alpha_j \mathbb{E}_\nu\left[\mathbb{1}_{x_j \neq Y_j}\right] \tag{3}$$

where $Y = (Y_1, \ldots, Y_n)$ is distributed according to ν. By the Cauchy-Schwarz inequality,

$$d_T(x,A)^2 = \inf_{\nu \in \mathscr{M}(A)} \sum_{j=1}^{n} \left(\mathbb{E}_\nu\left[\mathbb{1}_{x_j \neq Y_j}\right]\right)^2 .$$

Rather than minimizing in the large space $\mathscr{M}(A)$, we may as well perform minimization on the convex compact set of probability measures on $\{0,1\}^n$ by mapping $y \in A$ on $(\mathbb{1}_{y_j \neq X_j})_{1 \leq j \leq n}$. Denote this mapping by χ. Note that the mapping depends on x but we omit this dependence to lighten notation. The set $\mathscr{M}(A) \circ \chi^{-1}$ of probability measures on $\{0,1\}^n$ coincides with $\mathscr{M}(\chi(A))$. It is convex and compact and therefore the infimum in the last display is achieved at some $\widehat{\nu}$. Then

$d_T(X,A)$ is just the Euclidean norm of the vector $\left(\mathbb{E}_{\widehat{\nu}}[\mathbb{1}_{x_j \neq Y_j}]\right)_{j \leq n}$, and therefore the supremum in (3) is achieved by the vector $\widehat{\alpha}$ of components

$$\widehat{\alpha}_i = \frac{\mathbb{E}_{\widehat{\nu}}[\mathbb{1}_{x_i \neq Y_i}]}{\sqrt{\sum_{j=1}^n \left(\mathbb{E}_{\widehat{\nu}}[\mathbb{1}_{x_j \neq Y_j}]\right)^2}} .$$

For simplicity, assume that the infimum in the definition of $f_i(x^{(i)})$ in (1) is achieved by a proper choice of the i-th coordinate.

Clearly, $f(x) - f_i(x^{(i)}) \geq 0$ for all i. On the other hand let $x_i^{(i)}$ and $\widehat{\nu}_i$ denote the coordinate value and the probability distribution on A that witness the value of $f_i(x^{(i)})$, that is,

$$f_i(x^{(i)}) = \sum_{j \neq i} \left(\mathbb{E}_{\widehat{\nu}_i} \mathbb{1}_{x_j \neq Y_j}\right)^2 + \left(\mathbb{E}_{\widehat{\nu}_i} \mathbb{1}_{x_i^{(i)} \neq Y_i}\right)^2 .$$

As $f(x) \leq \sum_{j \neq i} \left(\mathbb{E}_{\widehat{\nu}_i}\left[\mathbb{1}_{x_j \neq Y_j}\right]\right)^2 + \left(\mathbb{E}_{\widehat{\nu}_i}\left[\mathbb{1}_{x_i^{(i)} \neq Y_i}\right]\right)^2$, we have

$$f(x) - f_i(x^{(i)}) \leq \left(\mathbb{E}_{\widehat{\nu}_i}\left[\mathbb{1}_{x_i \neq Y_i}\right]\right)^2 - \left(\mathbb{E}_{\widehat{\nu}_i}\left[\mathbb{1}_{x_i^{(i)} \neq Y_i}\right]\right)^2 \leq 1 .$$

It remains to prove that f is weakly $(4,0)$-self-bounding. To this end, we may use once again Sion's minimax theorem, as in [3], to write the convex distance as

$$\begin{aligned} d_T(x,A) &= \inf_{\nu \in \mathcal{M}(A)} \sup_{\alpha : \|\alpha\|_2 \leq 1} \sum_{j=1}^n \alpha_j \mathbb{E}_\nu [\mathbb{1}_{x_j \neq Y_j}] \\ &= \sup_{\alpha : \|\alpha\|_2 \leq 1} \inf_{\nu \in \mathcal{M}(A)} \sum_{j=1}^n \alpha_j \mathbb{E}_\nu [\mathbb{1}_{x_j \neq Y_j}] . \end{aligned}$$

Denote the pair (ν, α) at which the saddle point is achieved by $(\widehat{\nu}, \widehat{\alpha})$. In [3] it is shown that for all x,

$$\sum_{i=1}^n \left(\sqrt{f(x)} - \sqrt{f_i(x^{(i)})}\right)^2 \leq 1 . \tag{4}$$

For completeness, we recall the argument:

$$\sqrt{f_i(x^{(i)})} = \inf_{\nu \in \mathcal{M}(A)} \sup_{\alpha : \|\alpha\|_2 \leq 1} \sum_{j=1}^n \alpha_j \mathbb{E}_\nu [\mathbb{1}_{x_j^{(i)} \neq Y_j}] \geq \inf_{\nu \in \mathcal{M}(A)} \sum_{j=1}^n \widehat{\alpha}_j \mathbb{E}_\nu [\mathbb{1}_{x_j^{(i)} \neq Y_j}] .$$

Let $\tilde{\nu}$ denote the distribution on A that achieves the infimum in the latter expression. Then we have

$$\sqrt{f(x)} = \inf_\nu \sum_{j=1}^n \widehat{\alpha}_j \mathbb{E}_\nu \left[\mathbb{1}_{x_j \neq Y_j}\right] \leq \sum_{j=1}^n \widehat{\alpha}_j \mathbb{E}_{\tilde{\nu}} \left[\mathbb{1}_{x_j \neq Y_j}\right] .$$

Hence,

$$\sqrt{f(x)} - \sqrt{f_i(x^{(i)})} \leq \sum_{j=1}^n \widehat{\alpha}_j \mathbb{E}_{\tilde{\nu}} [\mathbb{1}_{x_j \neq Y_j} - \mathbb{1}_{x_j^{(i)} \neq Y_j}] = \widehat{\alpha}_i \mathbb{E}_{\tilde{\nu}} \left[\mathbb{1}_{x_i \neq Y_i} - \mathbb{1}_{x_i^{(i)} \neq Y_i}\right] \leq \widehat{\alpha}_i ,$$

so

$$\left(\sqrt{f(x)} - \sqrt{f_i(x^{(i)})}\right)^2 \le \widehat{\alpha}_i^2 ,$$

from which (4) follows. Finally,

$$\begin{aligned}
\sum_{i=1}^n \left(f(x) - f_i(x^{(i)})\right)^2 &= \sum_{i=1}^n \left(\sqrt{f(x)} - \sqrt{f_i(x^{(i)})}\right)^2 \left(\sqrt{f(x)} + \sqrt{f_i(x^{(i)})}\right)^2 \\
&\le \sum_{i=1}^n \widehat{\alpha}_i^2 4f(x) \\
&\le 4f(x) .
\end{aligned}$$

□

Now the convex distance inequality follows easily:

Corollary 1.

$$\mathbb{P}\{X \in A\}\mathbb{E}\left[e^{d_T(X,A)^2/10}\right] \le 1 .$$

Proof. First recall that $A = \{X : d_T(X,A) = 0\}$. Observe now that combining Lemma 1 and Theorem 1, and choosing $t = \mathbb{E}\left[d_T^2(X,A)\right]$, we have

$$\mathbb{P}\{X \in A\} = \mathbb{P}\left\{d_T(X,A)^2 \le \mathbb{E}\left[d_T^2(X,A)\right] - t\right\} \le \exp\left(-\frac{\mathbb{E}[d_T(X,A)^2]}{8}\right) .$$

On the other hand, for $0 \le \lambda \le 1/2$, from Theorem 1 again,

$$\log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}Z)}\right] \le \frac{\lambda^2 2\mathbb{E}Z}{1 - 2\lambda} .$$

Choosing $\lambda = 1/10$ leads to the desired result. □

3 The square of a regular function

Let $g : \mathscr{X}^n \to \mathbb{R}^+$ be a function of n variables and assume that there exists a constant $v > 0$ and there are measurable functions $g_i : \mathscr{X}^{n-1} \to \mathbb{R}^+$ such that for all $x \in \mathscr{X}^n$, $g(x) \ge g_i(x^{(i)})$,

$$\sum_{i=1}^n \left(g(x) - g_i(x^{(i)})\right)^2 \le v .$$

We call such a function *v-regular.* If $X = (X_1, \ldots, X_n) \in \mathscr{X}^n$ is a vector of independent $\mathscr{X}$-valued random variables, then by the Efron-Stein inequality, $\mathsf{Var}(g(X)) \le v$. Also, it is shown in [8; 3] that for all $t > 0$,

$$\mathbb{P}\left\{g(X) \ge \mathbb{E}g(X) + t\right\} \le e^{-t^2/(2v)} .$$

For the lower tail, Maurer [11] showed that if, in addition, $g(x) - g_i(x^{(i)}) \le 1$ for all i and x, then

$$\mathbb{P}\left\{g(X) \le \mathbb{E}g(X) - t\right\} \le e^{-t^2/(2(v+t/3))} .$$

However, in many situations one expects a purely sub-Gaussian behavior of the lower tail, something Maurer's inequality fails to capture. Here we show how Theorem 1 may be used to derive purely sub-Gaussian lower-tail bounds under an additional "bounded differences" condition for the *square* of g.

Corollary 2. *Let $g : \mathscr{X}^n \to \mathbb{R}^+$ be a v-regular function such that for all $x \in \mathscr{X}^n$ and $i = 1, \ldots, n$, $g(x)^2 - g_i(x^{(i)})^2 \le 1$. Then*

$$\mathbb{P}\left\{g(X)^2 \le \mathbb{E}\left[g(X)^2\right] - t\right\} \le \exp\left(\frac{-t^2}{8v\mathbb{E}\left[g(X)^2\right] + t(4v - 1/3)_-}\right) .$$

In particular, if $v \ge 1/12$,

$$\mathbb{P}\left\{g(X) \le \mathbb{E}g(X) - t\right\} \le \exp\left(\frac{-t^2}{8v}\right) .$$

Proof. Introduce $f(x) = g(x)^2$ and $f_i(x^{(i)}) = g_i(x^{(i)})^2$. Then

$$0 \le f(x) - f_i(x^{(i)}) \le 1 .$$

Moreover,

$$\begin{aligned}
\sum_{i=1}^n \left(f(x) - f_i(x^{(i)})\right)^2 &= \sum_{i=1}^n \left(g(x) - g_i(x^{(i)})\right)^2 \left(g(x) + g_i(x^{(i)})\right)^2 \\
&\le 4g(x)^2 \sum_{i=1}^n \left(g(x) - g_i(x^{(i)})\right)^2 \\
&\le 4vf(x)
\end{aligned}$$

and therefore f is $(4v, 0)$ self-bounding. This means that the third inequality of Theorem 1 is applicable and this is how the first inequality is obtained.

The second inequality follows from the first by noting that as $\mathbb{E}g(X) \le \sqrt{\mathbb{E}g(X)^2}$,

$$\begin{aligned}
\mathbb{P}\left\{g(X) \le \mathbb{E}g(X) - t\right\} &\le \mathbb{P}\left\{g(X)\sqrt{\mathbb{E}g(X)^2} \le \mathbb{E}g(X)^2 - t\sqrt{\mathbb{E}g(X)^2}\right\} \\
&\le \mathbb{P}\left\{g(X)^2 \le \mathbb{E}g(X)^2 - t\sqrt{\mathbb{E}g(X)^2}\right\}
\end{aligned}$$

and now the first inequality may be applied. $\square$

For a more concrete class of applications, consider a convex function g defined on a bounded hyper-rectangle, say $[0,1]^n$. If $X = (X_1, \ldots, X_n)$ are independent random variables taking values in $[0,1]$, then Talagrand [19] shows that

$$\mathbb{P}\{|g(X) - \mathbb{M}g(X)| > t\} \le 4e^{-t^2/(4L^2)}$$

where $\mathbb{M}g(X)$ denotes the median of the random variable $g(X)$ and L is the Lipschitz constant of g. (In fact, this inequality holds under the weaker assumption that the level sets $\{x : g(x) \le t\}$ are convex.) Ledoux [8] used the entropy method to prove the one-sided inequality

$$\mathbb{P}\{g(X) - \mathbb{E}g(X) > t\} \le e^{-t^2/(2L^2)}$$

under the condition that g is *separately convex*, that is, it is convex in any of its variables when the rest of the variables are fixed at an arbitrary value. We may use Ledoux's argument in combination with the corollary above.

Let $g : [0,1]^n \to \mathbb{R}$ be a non-negative separately convex function. Without loss of generality we may assume that g is differentiable on $[0,1]^n$ because otherwise one may approximate g by a smooth function in a standard way. Then, denoting

$$g_i(x^{(i)}) = \inf_{x_i' \in \mathscr{X}} g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) ,$$

by separate convexity,

$$g(x) - g_i(x^{(i)}) \quad \le \quad \left| \frac{\partial g}{\partial x_i}(x) \right| .$$

Thus, for every $x \in [0,1]^n$,

$$\sum_{i=1}^{n} \left(g(x) - g_i(x^{(i)}) \right)^2 \le L^2 .$$

This means that g is L^2-regular and therefore Corollary 2 applies as long as $g(x)^2 - g_i(x^{(i)})^2$ takes its values in an interval of length 1.

4 Proofs

Our starting point is a so-called modified logarithmic Sobolev inequality that goes back (at least) to [10]. This inequality is at the basis of several concentration inequalities proved by the entropy method, see [2; 3; 17; 16; 4; 11; 13].

Theorem 2. (A MODIFIED LOGARITHMIC SOBOLEV INEQUALITY.) *Let $X = (X_1, X_2, \ldots, X_n)$ be a vector of independent random variables, each taking values in some measurable space $\mathscr{X}$. Let $f : \mathscr{X}^n \to \mathbb{R}$ be measurable and let $Z = f(X)$. Let $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and let Z_i denote a measurable function of $X^{(i)}$. Introduce $\psi(x) = e^x - x - 1$. Then for any $\lambda \in \mathbb{R}$,*

$$\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] \le \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z} \psi\left(-\lambda(Z - Z_i)\right)\right] .$$

The entropy method converts the modified logarithmic Sobolev inequality into a differential inequality involving the logarithm of the moment generating function of Z. A more-or-less standard way of proceeding is as follows.

If $\lambda \geq 0$ and f is (a,b)-self-bounding, then, using $Z - Z_i \leq 1$ and the fact that for all $x \in [0,1]$, $\psi(-\lambda x) \leq x\psi(-\lambda)$,

$$\begin{aligned}\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right]\log \mathbb{E}\left[e^{\lambda Z}\right] &\leq \psi(-\lambda)\mathbb{E}\left[e^{\lambda Z}\sum_{i=1}^{n}(Z - Z_i)\right] \\ &\leq \psi(-\lambda)\mathbb{E}\left[(aZ+b)e^{\lambda Z}\right].\end{aligned}$$

For any $\lambda \in \mathbb{R}$, define $G(\lambda) = \log \mathbb{E}\left[e^{(\lambda Z - \mathbb{E}Z)}\right]$. Then the previous inequality may be written as the differential inequality

$$\left[\lambda - a\psi(-\lambda)\right] G'(\lambda) - G(\lambda) \leq v\psi(-\lambda), \tag{5}$$

where $v = a\mathbb{E}Z + b$.

On the other hand, if $\lambda \leq 0$ and f is weakly (a,b)-self-bounding, then since $\psi(x)/x^2$ is non-decreasing over $\mathbb{R}^+$, $\psi(-\lambda(Z - Z_i)) \leq \psi(-\lambda)(Z - Z_i)^2$ so

$$\begin{aligned}\lambda \mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right]\log \mathbb{E}\left[e^{\lambda Z}\right] &\leq \psi(-\lambda)\mathbb{E}\left[e^{\lambda Z}\sum_{i=1}^{n}(Z - Z_i)^2\right] \\ &\leq \psi(-\lambda)\mathbb{E}\left[(aZ+b)e^{\lambda Z}\right].\end{aligned}$$

This again leads to the differential inequality (5) but this time for $\lambda \leq 0$.

When $a = 1$, this differential inequality can be solved exactly (see [2]), and one obtains

$$G(\lambda) \leq v\psi(\lambda).$$

The right-hand side is just the logarithm of the moment generating function of a Poisson(v) random variable.

However, when $a \neq 1$, it is not obvious what kind of bounds for G should be expected. If $a > 1$, then $\lambda - a\psi(-\lambda)$ becomes negative when λ is large enough. Since both $G'(\lambda)$ and $G(\lambda)$ are non-negative when λ is non-negative, (5) becomes trivial for large values of λ. Hence, at least when $a > 1$, there is no hope to derive Poissonian bounds from (5) for positive values of λ (i.e., for the upper tail).

Note that using the fact that $\psi(-\lambda) \leq \lambda^2/2$ for $\lambda \geq 0$, (5) implies that for $\lambda \in [0, 2/a)$,

$$\left(\frac{1}{\lambda} - \frac{a}{2}\right) G'(\lambda) - \frac{1}{\lambda^2}G(\lambda) \leq \frac{v}{2}.$$

Observe that the left-hand side is just the derivative of $(1/\lambda - a/2)\,G(\lambda)$. Using the fact that $G(0) = G'(0) = 0$, and that $G'(\lambda) \geq 0$ for $\lambda > 0$, integrating this differential inequality leads to

$$G(\lambda) \leq vG_{a/2}(\lambda) = \frac{v\lambda^2}{2(1 - a\lambda/2)} \text{ for } \lambda \in [0, 2/a),$$

which, by Markov's inequality and optimization of λ, leads to a first Bernstein-like upper tail inequality. Note that this is enough to derive the bounds for the upper tail of weakly self-bounded random variables. But we want to prove something more.

The following lemma is the key in the proof of the theorem. It shows that if f satisfies a self-bounding property, then on the relevant interval, the logarithmic moment generating function of $Z - \mathbb{E}Z$ is upper bounded by v times a function G_γ defined by

$$G_\gamma(\lambda) = \frac{\lambda^2}{2(1-\gamma\lambda)} \quad \text{for every } \lambda \text{ such that } \gamma\lambda < 1$$

where $\gamma \in \mathbb{R}$ is a real-valued parameter. In the lemma below we mean $c_+^{-1} = \infty$ (resp. $c_-^{-1} = \infty$) when $c_+ = 0$ (resp. $c_- = 0$).

Lemma 2. *Let $a, v > 0$ and let G be a solution of the differential inequality*

$$\left[\lambda - a\psi(-\lambda)\right] H'(\lambda) - H(\lambda) \le v\psi(-\lambda).$$

Define $c = (a - 1/3)/2$. Then, for every $\lambda \in \left(0, c_+^{-1}\right)$

$$G(\lambda) \le vG_{c_+}(\lambda)$$

and for every $\lambda \in (-\theta, 0)$

$$G(\lambda) \le vG_{-c_-}(\lambda)$$

where $\theta = c_-^{-1}\left(1 - \sqrt{1 - 6c_-}\right)$ if $c_- > 0$ and $\theta = a^{-1}$ whenever $c_- = 0$.

This lemma is proved in the next section. First we show how it implies out main result:

Proof of Theorem 1. The upper-tail inequality for (a, b)-self-bounding functions follows from Lemma 2 and Markov's inequality by routine calculations, exactly as in the proof of Bernstein's inequality when $c_+ > 0$ and it is straightforward when $c_+ = 0$.

The bound for the upper tail of weakly (a, b)-self-bounding functions is due to Maurer [11]. The derivation of the bound for the lower tail requires some more care. Indeed, we have to check that the condition $\lambda > -\theta$ is harmless. Since $\theta < c_-^{-1}$, by continuity, for every positive t,

$$\sup_{u \in (0,\theta)} \left(tu - \frac{u^2 v}{2\left(1 - c_- u\right)}\right) = \sup_{u \in (0,\theta]} \left(tu - \frac{u^2 v}{2\left(1 - c_- u\right)}\right).$$

Note that we are only interested in values of t that are smaller than $\mathbb{E}Z \le v/a$. Now the supremum of

$$u \to tu - \frac{u^2 v}{2\left(1 - c_- u\right)}$$

on the interval $\left(0, c_-^{-1}\right)$ is achieved either at $u_t = t/v$ (if $c_- = 0$) or at $u_t = c_-^{-1}\left(1 - \left(1 + \left(2tc_-/v\right)\right)^{-1/2}\right)$ (if $c_- > 0$).

It is time to take into account the restriction $t \le v/a$. In the first case, when $u_t = t/v$, this implies that $u_t \le a^{-1} = \theta$, while in the second case, since $a = \left(1 - 6c_-\right)/3$ it implies that $1 + \left(2tc_-/v\right) \le \left(1 - 6c_-\right)^{-1}$ and therefore $u_t \le c_-^{-1}\left(1 - \sqrt{1 - 6c_-}\right) = \theta$. In both cases $u_t \le \theta$ which means that for every $t \le v/a$

$$\sup_{u \in (0,\theta]} \left(tu - \frac{u^2 v}{2\left(1 - c_- u\right)}\right) = \sup_{u \in \left(0, c_-^{-1}\right)} \left(tu - \frac{u^2 v}{2\left(1 - c_- u\right)}\right)$$

and the result follows. □

5 Proof of Lemma 2

The entropy method consists in deriving differential inequalities for the logarithmic moment generating functions and solving those differential inequalities. In many circumstances, the differential inequality can be solved exactly as in [10; 2]. The next lemma allows one to deal with a large family of solvable differential inequalities. Lemma 4 will allow us to use this lemma to cope with more difficult cases and this will lead to the proof of Lemma2.

Lemma 3. *Let f be a non-decreasing continuously differentiable function on some interval I containing 0 such that $f(0) = 0$, $f'(0) > 0$ and $f(x) \neq 0$ for every $x \neq 0$. Let g be a continuous function on I and consider an infinitely many times differentiable function G on I such that $G(0) = G'(0) = 0$ and for every $\lambda \in I$,*

$$f(\lambda)G'(\lambda) - f'(\lambda)G(\lambda) \leq f^2(\lambda)g(\lambda)\ .$$

Then, for every $\lambda \in I$, $G(\lambda) \leq f(\lambda)\int_0^\lambda g(x)dx$.

Note that the special case when $f(\lambda) = \lambda$, and $g(\lambda) = L^2/2$ is the differential inequality obtained by the Gaussian logarithmic Sobolev inequality via Herbst's argument (see, e.g., Ledoux [9]) and is used to obtain Gaussian concentration inequalities. If we choose $f(\lambda) = e^\lambda - 1$ and $g(\lambda) = -d(\lambda/e^\lambda - 1)/d\lambda$, we recover the differential inequality used to prove concentration of $(1,0)$-self-bounding functions in [3].

Proof. Define $\rho(\lambda) = G(\lambda)/f(\lambda)$, for every $\lambda \neq 0$ and $\rho(0) = 0$. Using the assumptions on G and f, we see that ρ is continuously differentiable on I with

$$\rho'(\lambda) = \frac{f(\lambda)G'(\lambda) - f'(\lambda)G(\lambda)}{f^2(\lambda)} \quad \text{for} \quad \lambda \neq 0 \quad \text{and} \quad \rho'(0) = \frac{G''(0)}{2f'(0)}\ .$$

Hence $f(\lambda)G'(\lambda) - f'(\lambda)G(\lambda) \leq f^2(\lambda)g(\lambda)$ implies that

$$\rho'(\lambda) \leq g(\lambda)$$

and therefore that the function $\Delta(\lambda) = \int_0^\lambda g(x)dx - \rho(\lambda)$ is nondecreasing on I. Since $\Delta(0) = 0$, Δ and f have the same sign on I, which means that $\Delta(\lambda)f(\lambda) \geq 0$ for $\lambda \in I$ and the result follows. □

Except when $a = 1$, the differential inequality (5) cannot be solved exactly. A roundabout is provided by the following lemma that compares the solutions of a possibly difficult differential inequality with solutions of a differential equation.

Lemma 4. *Let I be an interval containing 0 and let ρ be continuous on I. Let $a \geq 0$ and $v > 0$. Let $H : I \to \mathbb{R}$, be an infinitely many times differentiable function satisfying*

$$\lambda H'(\lambda) - H(\lambda) \leq \rho(\lambda)\left(aH'(\lambda) + v\right)$$

with

$$aH'(\lambda) + v > 0 \quad \textit{for every} \quad \lambda \in I \ \ \textit{and} \ \ H'(0) = H(0) = 0\ .$$

Let $\rho_0 : I \to \mathbb{R}$ be a function. Assume that $G_0 : I \to \mathbb{R}$ is infinitely many times differentiable such that for every $\lambda \in I$,

$$aG_0'(\lambda) + 1 > 0 \quad \text{and} \quad G_0'(0) = G_0(0) = 0 \quad \text{and} \quad G_0''(0) = 1 \; .$$

Assume also that G_0 solves the differential equation

$$\lambda G_0'(\lambda) - G_0(\lambda) = \rho_0(\lambda)\left(aG_0'(\lambda) + 1\right) \; .$$

If $\rho(\lambda) \le \rho_0(\lambda)$ for every $\lambda \in I$, then $H \le vG_0$.

Proof. Let $I, \rho, a, v, H, G_0, \rho_0$ be defined as in the statement of the lemma. Combining the assumptions on H, ρ_0, ρ and G_0,

$$\lambda H'(\lambda) - H(\lambda) \le \frac{\left(\lambda G_0'(\lambda) - G_0(\lambda)\right)\left(aH'(\lambda) + v\right)}{aG_0'(\lambda) + 1}$$

for every $\lambda \in I$, or equivalently,

$$\left(\lambda + aG_0(\lambda)\right) H'(\lambda) - \left(1 + aG_0'(\lambda)\right) H(\lambda) \le v\left(\lambda G_0'(\lambda) - G_0(\lambda)\right) \; .$$

Setting $f(\lambda) = \lambda + aG_0(\lambda)$ for every $\lambda \in I$ and defining $g : I \to \mathbb{R}$ by

$$g(\lambda) = \frac{v\left(\lambda G_0'(\lambda) - G_0(\lambda)\right)}{\left(\lambda + aG_0(\lambda)\right)^2} \quad \text{if} \quad \lambda \neq 0 \text{ and } g(0) = \frac{v}{2} \; ,$$

our assumptions on G_0 imply that g is continuous on the whole interval I so that we may apply Lemma 3. Hence, for every $\lambda \in I$

$$H(\lambda) \le f(\lambda) \int_0^\lambda g(x)dx = vf(\lambda) \int_0^\lambda \left(\frac{G_0(x)}{f(x)}\right)' dx$$

and the conclusion follows since $G_0(x)/f(x)$ tends to 0 when x tends to 0. □

Observe that the differential inequality in the statement of Lemma 2 has the same form as the inequalities considered in Lemma 4 where ψ replaces ρ. Note also that for any $\gamma \ge 0$,

$$2G_\gamma(\lambda) = \frac{\lambda^2}{1 - \gamma\lambda}$$

solves the differential inequality

$$\lambda H'(\lambda) - H(\lambda) = \lambda^2(\gamma H'(\lambda) + 1) \, . \tag{6}$$

So choosing $\gamma = a$ and recalling that for $\lambda \ge 0$, $\psi(-\lambda) \le \frac{\lambda^2}{2}$, it follows immediately from Lemma 4, that

$$G(\lambda) \le \frac{\lambda^2 v}{2(1 - a\lambda)} \quad \text{for} \quad \lambda \in (0, 1/a) \; .$$

Since G is the logarithmic moment generating function of $Z - \mathbb{E}Z$, this can be used to derive a Bernstein-type inequality for the left tail of Z. However, the obtained constants are not optimal, so proving Lemma 2 requires some more care.

Proof of Lemma 2. The function $2G_\gamma$ may be the unique solution of equation (6) but this is not the only equation G_γ is the solution of. Define

$$\rho_\gamma(\lambda) = \frac{\lambda G'_\gamma(\lambda) - G_\gamma(\lambda)}{1 + aG'_\gamma(\lambda)} \ .$$

Then, on some interval I, G_γ is the solution of the differential equation

$$\lambda H'(\lambda) - H(\lambda) = \rho_\gamma(\lambda)(1 + aH'(\lambda)) \, ,$$

provided $1 + aG'_\gamma$ remains positive on I.

Thus, we have to look for the smallest $\gamma \geq 0$ such that, on the relevant interval I (with $0 \in I$), we have both $\psi(-\lambda) \leq \rho_\gamma(\lambda)$ and $1 + aG'_\gamma(\lambda) > 0$ for $\lambda \in I$.

Introduce

$$D_\gamma(\lambda) = (1-\gamma\lambda)^2(1 + aG'_\gamma(\lambda)) = (1-\gamma\lambda)^2 + a\lambda\left(1 - \frac{\gamma\lambda}{2}\right) = 1 + 2(a/2-\gamma)\lambda - \gamma(a/2-\gamma)\lambda^2 \ .$$

Observe that $\rho_\gamma(\lambda) = \lambda^2/(2D_\gamma(\lambda))$.

For any interval I, $1 + aG'_\gamma(\lambda) > 0$ for $\lambda \in I$ holds if and only if $D_\gamma(\lambda) > 0$ for $\lambda \in I$. Hence, if $D_\gamma(\lambda) > 0$ and $\psi(-\lambda) \leq \rho_\gamma(\lambda)$, then it follows from Lemma 4 that for every $\lambda \in I$, we have $G(\lambda) \leq vG_\gamma(\lambda)$.

We first deal with intervals of the form $I = [0, c_+^{-1})$ (with $c_+^{-1} = \infty$ when $c_+ = 0$). If $a \leq 1/3$, that is, $c_+ = 0$, $D_{c_+}(\lambda) = 1 + a\lambda > 0$ and $\rho_{c_+}(\lambda) \geq \frac{\lambda^2}{2(1+\lambda/3)} \geq \psi(-\lambda)$ for $\lambda \in I = [0, +\infty)$.

If $a > 1/3$, then $D_{c_+}(\lambda) = 1 + \lambda/3 - c_+\lambda^2/6$ satisfies $0 < 1 + \lambda/6 \leq D_{c_+}(\lambda) \leq 1 + \lambda/3$ on an interval I containing $[0, c_+^{-1})$, and therefore $\rho_{c_+}(\lambda) \geq \psi(-\lambda)$ on I.

Next we deal with intervals of the form $I = (-\theta, 0]$ where $\theta = a^{-1}$ if $c_- = 0$, and $\theta = c_-^{-1}(1 - \sqrt{1-6c_-})$ otherwise. Recall that for any $\lambda \in (-3, 0]$, $\psi(-\lambda) \leq \frac{\lambda^2}{2(1+\lambda/3)}$.

If $a \geq 1/3$, that is, $c_- = 0$, $D_{-c_-}(\lambda) = 1 + a\lambda > 0$ for $\lambda \in (a^{-1}, 0]$, while

$$\rho_{-c_-}(\lambda) = \frac{\lambda^2}{2(1+a\lambda)} \geq \frac{\lambda^2}{2(1+\lambda/3)} \ .$$

For $a \in (0, 1/3)$, note first that $0 < c_- \leq 1/6$, and that

$$0 < D_{-c_-}(\lambda) \leq 1 + \frac{\lambda}{3} + \frac{\lambda^2}{36} \leq \left(1 + \frac{\lambda}{6}\right)^2$$

for every $\lambda \in (-\theta, 0]$. This also entails that $\rho_{-c_-}(\lambda) \geq \psi(-\lambda)$ for $\lambda \in (-\theta, 0]$. $\square$

Acknowledgment. We thank Andreas Maurer for his insightful comments.

References

[1] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *The Annals Probability*, 33:514–560, 2005. MR2123200

[2] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000. MR1749290

[3] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals Probability*, 31:1583–1614, 2003. MR1989444

[4] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris*, 334:495–500, 2002. MR1890640

[5] L. Devroye. Laws of large numbers and tail inequalities for random tries and Patricia trees. *Journal of Computational and Applied Mathematics*, 142:27–37, 2002. MR1910516

[6] T. Klein. Une inégalité de concentration à gauche pour les processus empiriques. *C. R. Math. Acad. Sci. Paris*, 334(6):501–504, 2002. MR1890641

[7] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33:1060–1077, 2005. MR2135312

[8] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1997. `http://www.emath.fr/ps/`. MR1399224

[9] M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, Providence, RI, 2001. MR1849347

[10] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28:863–884, 2000. MR1782276

[11] A. Maurer. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 29:121–138, 2006. MR2245497

[12] C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, New York, 1998. MR1678578

[13] C. McDiarmid and B. Reed. Concentration for self-bounding functions and an inequality of talagrand. *Random Structures and Algorithms*, 29:549–557, 2006. MR2268235

[14] M. Molloy and B. Reed. *Graph colouring and the probabilistic method*, volume 23 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 2002. MR1869439

[15] D. Pollard. A note on Talagrand's convex hull concentration inequality. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 196–203. Inst. Math. Statist., Beachwood, OH, 2007. MR2459940

[16] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, 119:163–175, 2001. MR1818244

[17] Emmanuel Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probab. Theory Related Fields*, 119(2):163–175, 2001. MR1818244

[18] J.M. Steele. *Probability Theory and Combinatorial Optimization*. SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics 69, 3600 University City Science Center, Phila, PA 19104, 1996. MR1422018

[19] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.*, 81:73–205, 1995. MR1361756

[20] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996. MR1419006

[21] M. Talagrand. A new look at independence. *Annals of Probability*, 24:1–34, 1996. (Special Invited Paper). MR1387624