Elect. Comm. in Probab. **6** (2001) 55–65

ELECTRONIC COMMUNICATIONS in PROBABILITY

A NOTE ON TALAGRAND'S CONCENTRATION INEQUALITY

DMITRIY PANCHENKO
Department of Mathematics and Statistics, The University of New Mexico, Albuquerque, NM, 87131, USA
email: `panchenk@math.unm.edu`

submitted February 7, 2001 *Final version accepted* April 24, 2001

AMS 2000 Subject classification: 60E15, 28A35
Concentration of measure, empirical processes

Abstract

In this paper we revisit Talagrand's proof of concentration inequality for empirical processes. We give a different proof of the main technical lemma that garantees the existence of a certain kernel. Moreover, we generalize the result of Talagrand to a family of kernels which in one particular case allows us to produce the Poissonian bound without using the truncation argument. In section 2 we give some examples of application of the abstract concentration inequality to empirical processes that demonstrate some interesting properties of Talagrand's kernel method.

1 Introduction and the proof of main lemma

This paper was motivated by the Section 4 of the "New concentration inequalities in product spaces" by Michel Talagrand. For the most part we will keep the same notations with possible minor changes. We slightly weaken the definition of the distance $m(A,x)$ introduced in [8], but, essentially, this is what is used in the proof of the concentration inequality for empirical processes. Moreover, we introduce a parametrized family of distances $m_\alpha(A,x)$ for $\alpha > 0$, which will allow us to produce one example of interest in Section 2. The case of $\alpha = 1$ essentially corresponds to the distance $m(A,x)$ in [8], and will also be used in several examples of Section 2. The Theorem 1 below is almost identical to Theorem 4.2 in [8] and we assume that the reader is familiar with the proof. The main technical step, Proposition 4.2 in [8], is proved differently and constitutes the statement of Lemma 1 below.

Let Ω^n be a measurable product space with a product measure μ^n. Consider a probability measure ν on Ω^n and $x \in \Omega^n$. If $\mathcal{C}_i = \{y \in \Omega^n : y_i \neq x_i\}$, we consider the image of the restriction of ν to $\mathcal{C}_i$ by the map $y \to y_i$, and its Radon-Nikodym derivative d_i with respect to μ. As in [8] we assume that Ω is finite and each point is measurable with a positive measure. Let m be a number of atoms in Ω and $p_1, \ldots, p_m$ be their probabilities. By the definition of

d_i we have

$$\int_{\mathcal{C}_i} g(y_i)d\nu(y) = \int_\Omega g(y_i)d_i(y_i)d\mu(y_i).$$

For $\alpha > 0$ we define a function $\psi_\alpha(x)$ by

$$\psi_\alpha(x) = \begin{cases} x^2/(4\alpha), & \text{when } x \leq 2\alpha, \\ x - \alpha, & \text{when } x \geq 2\alpha. \end{cases}$$

We set

$$m_\alpha(\nu, x) = \sum_{i \leq n} \int \psi_\alpha(d_i)d\mu \quad \text{and} \quad m_\alpha(A, x) = \inf\{m_\alpha(\nu, x) : \nu(A) = 1\}.$$

For each $\alpha > 0$ let L_α be any positive number satisfying the following inequality:

$$\frac{2L_\alpha(e^{1/L_\alpha} - 1)}{1 + 2L_\alpha} \leq \alpha. \tag{1.1}$$

The following theorem holds.

Theorem 1 *Let $\alpha > 0$ and L_α satisfy (1.1). Then for any n and $A \subseteq \Omega^n$ we have*

$$\int \exp \frac{1}{L_\alpha} m_\alpha(A, x)d\mu^n(x) \leq \frac{1}{\mu^n(A)}. \tag{1.2}$$

Proof. As we mentioned above the proof is identical to the proof of Theorem 4.2 in [8] where Proposition 4.2 is substituted by the Lemma 1 below, and the case of $n = 1$ must be adjusted to the new definition of ψ_α. Namely, one has to prove that

$$p + (1 - p)\exp\{\frac{1}{L_\alpha} p\psi_\alpha(\frac{1}{p})\} \leq \frac{1}{p}$$

for all $p \in [0, 1]$. This can be rewritten as

$$\exp\{\frac{1}{L_\alpha} p\psi_\alpha(\frac{1}{p})\} \leq \frac{1+p}{p} \quad \text{or} \quad \frac{p\psi_\alpha(\frac{1}{p})}{\log(1 + \frac{1}{p})} \leq L_\alpha.$$

By the definition of ψ_α one has to consider two separate cases:

1. If $p^{-1} \geq 2\alpha$ then $(1 - \alpha p)/\log(1 + \frac{1}{p}) \leq L_\alpha$,
2. If $p^{-1} \leq 2\alpha$ then $1/4\alpha p \log(1 + \frac{1}{p}) \leq L_\alpha$.

Taking into account (1.1) we must show that for all $L > 0$

$$1 - p\frac{2L(e^{1/L} - 1)}{1 + 2L} \leq L\log(1 + \frac{1}{p}), \quad \text{if} \quad \frac{1}{p} \geq \frac{4L(e^{1/L} - 1)}{1 + 2L},$$

$$\frac{1}{p\log(1 + \frac{1}{p})} \leq \frac{8L^2(e^{1/L} - 1)}{1 + 2L}, \quad \text{if} \quad \frac{1}{p} \leq \frac{4L(e^{1/L} - 1)}{1 + 2L}.$$

The proof of both of these inequalities constitutes a tedious exersize in calculus and is boring enough not to include it in this paper. □

Lemma 1 *Let $g_1 \geq g_2 \geq \ldots \geq g_m > 0$. For $\alpha > 0$ and L_α satistying (1.1) there exist $\{k_j^i : 1 \leq j < i \leq m\}$ such that*

$$k_j^i \geq 0, \text{ and } \sum_{j<i} k_j^i p_j \leq 1 \text{ for all } i \leq m, \tag{1.3}$$

and

$$\sum_{i\leq m} \frac{p_i}{g_i} \exp\Big\{\sum_{j<i}\Big(k_j^i \log\frac{g_i}{g_j} + \frac{1}{L_\alpha}\psi_\alpha(k_j^i)\Big)p_j\Big\} \leq \frac{1}{p_1 g_1 + \ldots + p_m g_m}. \tag{1.4}$$

Proof: The proof is by induction on the number of atoms m. The statement of lemma is trivial for $m = 1$. Note that in order to show the existence of $\{k_j^i\}$ in the statement of lemma one should try to minimize the left side of (1.4) with respect to $\{k_j^i\}$ under the constraints (1.3). Note also that each term on the left side of (1.4) has its own set of k_j^i, $j < i$ and, therefore, minimization can be performed for each term individually. From now on we assume that k_j^i are chosen in an optimal way minimizing the left side of (1.4) and it will be convenient to take among all such optimal choices the one maximizing $\sum_{j<i} k_j^i p_j$ for all $i \leq m$. To make the induction step we will start by proving the following statement, where we assume that k_j^i correspond to the *specific* optimal choice indicated above.

Claim 1.*For each $i \leq m$ we have*

$$\sum_{j<i} k_j^i p_j < 1 \Longleftrightarrow \log\frac{g_1}{g_i} < \frac{1}{L_\alpha} \text{ and } \sum_{j<i} 2L_\alpha \alpha p_j \log\frac{g_j}{g_i} < 1. \tag{1.5}$$

In this case $k_j^i = 2L_\alpha \alpha \log\frac{g_j}{g_i}$ for $1 \leq j < i$.

Proof: Let us fix i throughout the proof of the statement. We first assume that the left side of (1.5) holds. Suppose that $\log\frac{g_1}{g_i} \geq \frac{1}{L_\alpha}$. In this case, since $\sup\{\psi_\alpha'(x) : x \in \mathcal{R}\} = 1$, one would decrease

$$k_1^i \log\frac{g_i}{g_1} + \frac{1}{L_\alpha}\psi_\alpha(k_1^i)$$

by increasing k_1^i until $\sum_{j<i} k_j^i p_j = 1$, thus, decreasing the left side of (1.4) which contradicts the choice of k_j^i. On the other hand, $\log\frac{g_1}{g_i} < \frac{1}{L_\alpha}$ implies that $k_j^i \leq 2\alpha$ for $j < i$, since for $k \geq 2\alpha$, $\psi_\alpha(k) = k - \alpha$ and the choice of $k_j^i \geq 2\alpha$ would only increase the left side of (1.4). For $k \leq 2\alpha$, $\psi_\alpha(k) = k^2/(4\alpha)$ and

$$\text{argmin}_k\Big(k\log\frac{g_i}{g_j} + \frac{k^2}{4L_\alpha\alpha}\Big) = 2L_\alpha\alpha\log\frac{g_j}{g_i}.$$

Hence, if $\sum_{j<i} 2L_\alpha\alpha p_j \log\frac{g_j}{g_i} \geq 1$ then since $\sum_{j<i} k_j^i p_j < 1$ the set

$$\mathcal{J} := \{j : k_j^i \leq 2L_\alpha\alpha\log\frac{g_j}{g_i}\} \neq \emptyset$$

is not empty. But again this would imply that $\sum_{j<i} k_j^i p_j = 1$; otherwise, increasing k_j^i for $j \in \mathcal{J}$ would decrease the left side of (1.4). This completes the proof of the statement. □

The equivalence statement of Claim 1 implies that if for some $i \le m$, $\sum_{j<i} k_j^i p_j < 1$ then $\sum_{j<l} k_j^l p_j < 1$ for all $l \le i$. We first assume that the equality

$$\sum_{j<m-1} k_j^{m-1} p_j = 1 \tag{1.6}$$

holds. It implies that $\sum_{j<m} k_j^m p_j = 1$. Moreover, in this case we are able to prove an even stronger statement, namely: (1.6) implies that

$$k_j^m = k_j^{m-1} \text{ for } j < m-1, \text{ and } k_{m-1}^m = 0.$$

(Notice, that this step is meaningless for $m = 2$ and should simply be skipped). Indeed,

$$\inf_{\sum_{j<m} k_j p_j = 1} \sum_{j<m} \Big(k_j \log \frac{g_m}{g_j} + \frac{1}{L_\alpha}\psi_\alpha(k_j)\Big) p_j = \log \frac{g_m}{g_{m-1}}$$
$$+ \inf_{\sum_{j<m} k_j p_j = 1} \Big(\sum_{j<m-1} \Big(k_j \log \frac{g_{m-1}}{g_j} + \frac{1}{L_\alpha}\psi_\alpha(k_j)\Big) p_j + \frac{1}{L_\alpha}\psi_\alpha(k_{m-1}) p_{m-1}\Big). \tag{1.7}$$

The assumption (1.6) means that the optimal choice of the vector $\{k_j^{m-1}; j < m-1\}$ is such that $\sum_{j<m-1} k_j^{m-1} p_j = 1$, and, therefore, in (1.7) it is advantageous to set $k_{m-1}^m = 0$ and $k_j^m = k_j^{m-1}$, $j < m-1$. It implies that

$$\sum_{j<m} \Big(k_j^m \log \frac{g_m}{g_j} + \frac{1}{L_\alpha}\psi_\alpha(k_j^m)\Big) p_j = \log \frac{g_m}{g_{m-1}} + \sum_{j<m-1} \Big(k_j^{m-1} \log \frac{g_{m-1}}{g_j} + \frac{1}{L_\alpha}\psi_\alpha(k_j^{m-1})\Big) p_j$$

and, hence,

$$\frac{p_m}{g_m} \exp\Big\{\sum_{j<m} \Big(k_j^m \log \frac{g_m}{g_j} + \frac{1}{L_\alpha}\psi_\alpha(k_j^m)\Big) p_j\Big\} = \frac{p_m}{g_{m-1}} \exp\Big\{\sum_{j<m-1} \Big(k_j^{m-1} \log \frac{g_{m-1}}{g_j} + \frac{1}{L_\alpha}\psi_\alpha(k_j^{m-1})\Big) p_j\Big\}.$$

This allows us to combine the last two terms on the left side of (1.4) and apply the induction assumption to the sets $(g_1, \ldots, g_{m-1})$ and $(p_1, \ldots, p_{m-1} + p_m)$. Since $p_{m-1} g_{m-1} + p_m g_m \le (p_{m-1} + p_m) g_{m-1}$, it proves (1.4) for $(g_1, \ldots, g_m)$ and $(p_1, \ldots, p_m)$.
Now let us assume that $\sum_{j<m-1} k_j^{m-1} p_j < 1$ or, equivalently,

$$\log \frac{g_1}{g_{m-1}} < \frac{1}{L_\alpha} \quad \text{and} \quad \sum_{j<m-1} 2L_\alpha \alpha p_j \log \frac{g_j}{g_{m-1}} < 1.$$

By continuity, it should be obvious that there exist $g_0 < g_{m-1}$ such that

$$\log \frac{g_1}{g_m} < \frac{1}{L_\alpha} \quad \text{and} \quad \sum_{j<m} 2L_\alpha \alpha p_j \log \frac{g_j}{g_m} < 1 \quad \text{for} \quad g_m \in (g_0, g_{m-1}]. \tag{1.8}$$

holds and, therefore, $\sum_{j<m} k_j^m p_j < 1$. We assume that g_0 is the smallest number such that (1.8) holds. Let us show that for a fixed $g_1, \ldots, g_{m-1}$ in order to prove lemma for $g_m < g_0$ it is enough to prove it for $g_m = g_0$. Indeed, let us take $g_m < g_0$. Then by the definition of g_0 and Claim 1 we have $\sum_{j<m} k_j^m p_j = 1$. Then (1.7) still holds and implies in this case that k_j^m do not depend on g_m for $g_m < g_0$. Moreover,

$$\frac{p_m}{g_m} \exp\Big\{\sum_{j<m} \Big(\log \frac{g_m}{g_j} k_j^m + \frac{1}{L_\alpha}\psi_\alpha(k_j^m)\Big) p_j\Big\} = \frac{p_m}{g_{m-1}} \exp\Big\{\sum_{j<m} \Big(\log \frac{g_{m-1}}{g_j} k_j^m + \frac{1}{L_\alpha}\psi_\alpha(k_j^m)\Big) p_j\Big\},$$

which means that for $g_m < g_0$ the left side of the inequality (1.4) does not depend on g_m. Since $(p_1g_1 + \ldots + p_mg_m)^{-1}$ decreases with respect to g_m it is enough to prove the inequality for $g_m = g_0$.

Hence, we can finally assume that

$$\log\frac{g_1}{g_m} \le \frac{1}{L_\alpha}, \quad \sum_{j<m} 2L_\alpha\alpha\log\frac{g_j}{g_m}p_j \le 1 \quad \text{and} \quad k_j^i = 2L_\alpha\alpha\log\frac{g_j}{g_i}.$$

and rewrite (1.4) as

$$\sum_{i\le m}\frac{p_i}{g_i}\exp\Big\{-L_\alpha\alpha\sum_{j<i}\Big(\log\frac{g_j}{g_i}\Big)^2 p_j\Big\} \le \frac{1}{p_1g_1+\ldots+p_mg_m}. \tag{1.9}$$

By the induction hypothesis (1.9) holds for $g_m = g_{m-1}$. To prove it for $g_m < g_{m-1}$ we will compare the derivatives of both sides of (1.9) with respect to g_m. It is enough to have

$$\frac{p_m}{g_m}\exp\Big\{-L_\alpha\alpha\sum_{j<m}\Big(\log\frac{g_m}{g_j}\Big)^2 p_j\Big\}\Big(-\frac{1}{g_m} - 2L_\alpha\alpha\frac{1}{g_m}\sum_{j<m}\log\frac{g_m}{g_j}p_j\Big) \ge -\frac{p_m}{(p_1g_1+\ldots+p_mg_m)^2}$$

or, equivalently,

$$\exp\Big\{-L_\alpha\alpha\sum_{j<m}\Big(\log\frac{g_m}{g_j}\Big)^2 p_j\Big\}\Big(1 - 2L_\alpha\alpha\sum_{j<m}\log\frac{g_j}{g_m}p_j\Big) \le \Big(\frac{g_m}{p_1g_1+\ldots+p_mg_m}\Big)^2.$$

Since $1-x \le e^{-x}$ for $x \ge 0$ it is enough to show

$$\exp\Big\{-L_\alpha\alpha\sum_{j<m}p_j\Big(\Big(\log\frac{g_j}{g_m}\Big)^2 + 2\log\frac{g_j}{g_m}\Big)\Big\} \le \Big(\frac{g_m}{p_1g_1+\ldots+p_mg_m}\Big)^2.$$

One can check that $(\log x)^2 + 2\log x$ is concave for $x \ge 1$. If we express $g_j = \lambda_j g_1 + (1-\lambda_j)g_m$, $j = 1,\ldots,m-1$, then

$$\sum_{j<m}p_j\Big(\Big(\log\frac{g_j}{g_m}\Big)^2 + 2\log\frac{g_j}{g_m}\Big) \ge (\sum_{j<m}p_j\lambda_j)\Big(\Big(\log\frac{g_1}{g_m}\Big)^2 + 2\log\frac{g_1}{g_m}\Big)$$

$$p_1g_1+\ldots+p_mg_m = (\sum_{j<m}p_j\lambda_j)g_1 + (p_m + \sum_{j<m}(1-\lambda_j)p_j)g_m.$$

If we denote $p = \sum_{j<m}p_j\lambda_j$ and $t = \log\frac{g_1}{g_m}$ we have to prove

$$\exp\Big\{-L_\alpha\alpha p(t^2+2t)\Big\} \le \Big(\frac{1}{pe^t+1-p}\Big)^2, \ \ 0\le p\le 1,\ 0\le t\le\frac{1}{L_\alpha}. \tag{1.10}$$

Equivalently,

$$\varphi(p,t) = (pe^t+1-p)\exp\Big\{-\frac{1}{2}L_\alpha\alpha p(t^2+2t)\Big\} \le 1, \ \ 0\le p\le 1,\ 0\le t\le\frac{1}{L_\alpha}.$$

We have

$$\varphi_t'(p,t) = \varphi(p,t)\Big(\frac{pe^t}{pe^t+1-p} - L_\alpha\alpha p(t+1)\Big).$$

Since for all $p \geq 0$ $\varphi(p,0)=1$ we need $\varphi_t'(p,0)=p(1-L_\alpha\alpha)\leq 0$, which always holds provided (1.1). Indeed,

$$1-L_\alpha\alpha \leq 1-\frac{2L_\alpha^2(e^{1/L_\alpha}-1)}{1+2L_\alpha}<0,$$

since the last inequality is equivalent to

$$1+\frac{1}{L_\alpha}+\frac{1}{2L_\alpha^2}<e^{1/L_\alpha}.$$

It is easy to see that $\varphi_t'(p,t)=0$ in at most one point t. In combination with $\varphi_t'(p,0)\leq 0$ it implies that for a fixed p maximum of $\varphi(p,t)$ is attained at $t=0$ or $t=1/L_\alpha$. Therefore, we have to show $\varphi(p,1/L_\alpha)\leq 1$, $0\leq p\leq 1$. We have,

$$\varphi_p'(p,\frac{1}{L_\alpha})=\varphi(p,\frac{1}{L_\alpha})\Big(\frac{e^{\frac{1}{L_\alpha}}-1}{pe^{\frac{1}{L_\alpha}}+1-p}-\frac{L_\alpha\alpha}{2}\Big(\frac{1}{L_\alpha^2}+2\frac{1}{L_\alpha}\Big)\Big).$$

Since $\varphi(0,\frac{1}{L_\alpha})=1$ we should have $\varphi_p'(0,\frac{1}{L_\alpha})\leq 0$ which would also imply $\varphi_p'(p,\frac{1}{L_\alpha})\leq 0$, $p>0$. One can check that

$$\varphi_p'(0,\frac{1}{L_\alpha})=e^{\frac{1}{L_\alpha}}-1-\frac{\alpha}{2}\Big(\frac{1}{L_\alpha}+2\Big)\leq 0$$

by (1.1). This finishes the proof of the Lemma. □

Remark. If one defines L_α optimally by making (1.1) into an equality, then it is easy to show (by using L'Hôpital's rule) that

$$\lim_{\alpha\to 0}L_\alpha\alpha=1, \quad \text{and} \quad \lim_{\alpha\to\infty}L_\alpha\log\alpha=1. \tag{1.11}$$

In one special case of $\alpha=1$, which as we mentioned above essentially corresponds to the kernel introduced in [8], (1.1) gives us $L_\alpha\sim 1.12$. In this particular case we solved the optimization problem in the Lemma 1 numerically for $m=2$ to show that the optimal value of the constant L is approximately equal to 1.07, thus indicating that our proof produces a rather tight value L_α at least for $\alpha=1$.

2 Some applications: empirical processes

In this section we prove several results related to one specific example of application of empirical processes, which although is not the most general formulation possible, nevertheless, demonstrates the power and flexibility of Talagrand's kernel method most clearly. We consider a family of functions $\mathcal{F}=\{f:\Omega\to[0,1]\}$. Given a vector

$$x=(x_1,\ldots,x_n)\in\Omega^n$$

one picks a function $f_x\in\mathcal{F}$ according to an arbitrary algorithm, which means that the choice of function f_x can depend on x. As an example, one can consider an empirical risk minimization problem:

$$f_x=\mathrm{argmin}_{\mathcal{F}}\sum_{i\leq n}f(x_i). \tag{2.1}$$

In any case, the goal of the results below is to construct a good bound on the true mean $\mu f_x = \int f_x d\mu$ (that holds with high probability) given that the sample mean $\bar{f}_x = n^{-1}\sum_{i\le n} f_x(x_i)$ of f_x is small. Denote by

$$Z(x) = \sup_{\mathcal{F}} \sum_{i\le n} (\mu f - f(x_i)), \quad x \in \Omega^n.$$

The bound for μf_x is related to the process $Z(x)$ via the following inequality

$$n\mu f_x = \sum_{i\le n} f_x(x_i) + \sum_{i\le n} (\mu f_x - f_x(x_i)) \le \sum_{i\le n} f_x(x_i) + Z(x).$$

The distinct feature of Theorem 2, the first result we prove using Talagrand's kernel method in comparison, for instance, with the typical results of differential inequalities methods is that instead of using the uniform variance term in the bound one can use - in our version - the second moment of the (random) function f_x and, moreover, substitute it with a sample mean $\bar{f}_x$ if one so desires.

In our first result we will be using the distance $m_1(A,x)$, for which Theorem 1 holds with $L = L_1 = 1.12$

Theorem 2 *Let $L = 1.12$ and M be a median of Z. Then, for any $u > 0$,*

$$\mathbb{P}\Big(\sum_{i\le n} (\mu f_x - f_x(x_i)) \ge M + \inf_{\delta>1}\Big(\frac{1}{\delta} n(1-\bar{f}_x)\mu f_x^2 + \delta L u\Big)\Big) \le 2e^{-u} \tag{2.2}$$

and

$$\mathbb{P}\Big(\sum_{i\le n} (\mu f_x - f_x(x_i)) \ge \inf_{\delta>1}\Big(1-\frac{1}{\delta}\Big)^{-1}\Big(M + \frac{1}{\delta} n(1-\bar{f}_x)\bar{f}_x + \delta L u\Big)\Big) \le 2e^{-u}. \tag{2.3}$$

Proof. Without loss of generality we assume that $\mathcal{F}$ is finite. Let us consider the set

$$A = \{Z(x) \le M\}.$$

Clearly, $\mu(A) \ge 1/2$. Let us fix a point $x \in \Omega^n$ and then choose $f \in \mathcal{F}$. For any point $y \in A$ we have

$$\sum_{i\le n} (\mu f - f(y_i)) \le M.$$

Therefore, for the probability measure ν such that $\nu(A) = 1$ we will have

$$\begin{aligned}\sum_{i\le n} (\mu f - f(x_i)) - M &\le \int \Big(\sum_{i\le n} (\mu f - f(x_i)) - \sum_{i\le n} (\mu f - f(y_i))\Big) d\nu(y) \\ &= \sum_{i\le n} \int (f(y_i) - f(x_i)) d_i(y_i) d\mu(y_i).\end{aligned}$$

It is easy to observe that for $v \ge 0$, and $-1 \le u \le 1$,

$$uv \le u^2 I(u > 0) + \psi_1(v). \tag{2.4}$$

Therefore, for any $\delta > 1$

$$\begin{aligned}\sum_{i\le n}(\mu f - f(x_i)) - M &\le \delta \sum_{i\le n}\int \frac{f(y_i)-f(x_i)}{\delta} d_i(y_i)d\mu(y_i)\\ &\le \frac{1}{\delta}\sum_{i\le n}\int (f(y_i)-f(x_i))^2 I(f(y_i)>f(x_i))d\mu(y_i) + \delta\sum_{i\le n}\int \psi_1(d_i)d\mu\end{aligned}$$

Taking the infimum over ν we obtain that for any $\delta > 1$

$$\sum_{i\le n}(\mu f - f(x_i)) \le M + \frac{1}{\delta}\sum_{i\le n}\int (f(y_i)-f(x_i))^2 I(f(y_i)>f(x_i))d\mu(y_i) + \delta m_1(A,x).$$

Let us denote the random variable $\xi = f(y_1)$, $F_\xi(t)$ - the distribution function of ξ, and $c_i = f(x_i)$. For $c \in [0,1]$ define the function $h(c)$ as

$$h(c) = \int (f(y_1)-c)^2 I(f(y_1)>c)d\mu(y_1) = \int_c^1 (t-c)^2 dF_\xi(t).$$

One can check that $h(c)$ is decreasing, convex, $h(0) = \mu f^2$ and $h(1) = 0$. Therefore,

$$\frac{1}{n}\sum_{i\le n} h(c_i) \le \Bigl(\frac{1}{n}\sum_{i\le n} c_i\Bigr)h(1) + \Bigl(1 - \frac{1}{n}\sum_{i\le n} c_i\Bigr)h(0) = (1-\bar{f})\mu f^2.$$

Hence, we showed that

$$\sum_{i\le n}(\mu f - f(x_i) \le M + \frac{1}{\delta}n(1-\bar{f})\mu f^2 + \delta m_1(A,x).$$

Theorem 1 then implies via the application of Chebyshev's inequality that with probability at least $1 - 2e^{-u}$, $m_1(A,x) \le Lu$ and, hence (we now spell out the dependence of f on x),

$$\sum_{i\le n}(\mu f_x - f_x(x_i) \le M + \inf_{\delta>1}\Bigl(\frac{1}{\delta}n(1-\bar{f}_x)\mu f_x^2 + \delta Lu\Bigr),$$

thus proving the first statement of the theorem. To prove the second statement one has to bound μf_x^2 by μf_x and then move the term $n\mu f_x$ to the left side of the inequality. □

Remark. In the case when $\mathcal{F}$ is a family of indicators of sets, the term $(1-\bar{f}_x)\bar{f}_x$ in (2.3) is actually equal to a sample variance.

The fact that the bound in (2.2) traces the "variance" term for each function $f \in \mathcal{F}$ individually will allow us, for instance, to write a uniform bound for a subset of functions satisfying some specific condition. As an example of application of this theorem we consider the zero-error case for the risk minimization problem. For a given $x = (x_1, \ldots, x_n)$ let $\mathcal{F}_x$ be a subset of functions in $\mathcal{F}$ defined by (2.1). By zero-error case we understand the fact that for each $f \in \mathcal{F}_x$

$$\bar{f} = \frac{1}{n}\sum_{i\le n} f(x_i) = 0.$$

In this case Theorem 2 implies the following bound.

Corollary 1 *For any $u \geq 0$,*

$$\mathbb{P}(\exists f \in \mathcal{F}\ \ \bar{f} = 0,\ \ \mu f \geq M + u) \leq 2 \exp\Big\{-\frac{u^2}{4L(M+u)}\Big\}.$$

Proof. For any $t > 0$ (2.3) implies that with probability at least $1 - 2e^{-t}$, for any $f \in \mathcal{F}$ such that $\sum f(x_i) = 0$ we have

$$n\mu f \leq \inf_{\delta > 1}\Big(1 - \frac{1}{\delta}\Big)^{-1}(M + \delta L t).$$

Setting the right side of this inequality to $M + u$ and solving for t we get that

$$t = \frac{u^2}{4L(M+u)},$$

which completes the proof of the corollary.

$\square$

We will now prove the uniform bound for $Z(x)$, where the role of the "variance" term will be played by $\sigma^2 = n \sup \mu f^2$. The proof of Theorem 3 below will utilize the family of distances $m_\alpha(A, x)$ rather than $m_1(A, x)$, and as a result will automatically provide the Poissonian tail behavior for large values of u/σ^2, thus avoiding the necessity of using the truncation argument of Talagrand (see [2], [3], [8]).

Theorem 3 *Let $\sigma^2 = n \sup_{\mathcal{F}} \mu f^2$. For any $u \geq 0$,*

$$\mathbb{P}(Z \geq M + u) \leq 2 \exp\Big\{-\sup_{s>0}\Big(su - \frac{2s(e^s - 1)}{s+2}\sigma^2\Big)\Big\} \tag{2.5}$$

and

$$\mathbb{P}(Z \leq M - u) \leq 2 \exp\Big\{-\sup_{s>0}\Big(su - \frac{2s(e^s - 1)}{s+2}\sigma^2\Big)\Big\}. \tag{2.6}$$

Proof. For a fixed real number a consider a set

$$A = \{Z(x) \leq a\}.$$

The choice of a will be made below. Again, as in Theorem 2, for the probability measure ν such that $\nu(A) = 1$ we will have

$$\begin{aligned}\sum_{i\leq n}(\mu f - f(x_i)) - a &\leq \int\Big(\sum_{i\leq n}(\mu f - f(x_i)) - \sum_{i \leq n}(\mu f - f(y_i))\Big)d\nu(y)\\ &= \sum_{i\leq n}\int (f(y_i) - f(x_i))d_i(y_i)d\mu(y_i) \leq \sum_{i \leq n}\int f(y_i)d_i(y_i)d\mu(y_i).\end{aligned}$$

Instead of (2.4), we now use the following: for $v \geq 0$, $0 \leq u \leq 1$, and $\alpha > 0$ we have

$$uv \leq \alpha u^2 + \psi_\alpha(v), \tag{2.7}$$

which implies that

$$Z(x) = \sup_{\mathcal{F}}\sum_{i \leq n}(\mu f - f(x_i)) \leq a + \alpha n \sup_{\mathcal{F}} \mu f^2 + m_\alpha(A, x) = a + \alpha\sigma^2 + m_\alpha(A, x).$$

For a given $t > 0$ let α_0 be the one minimizing

$$\alpha_0 = \mathrm{argmin}(\alpha\sigma^2 + L_\alpha t),$$

where L_α is defined by (1.1). Theorem 1 applied to $m_{\alpha_0}(A, x)$ implies that

$$\mathbb{P}\{Z \le a + \alpha_0\sigma^2 + L_{\alpha_0} t\} \ge 1 - \frac{e^{-t}}{P(Z \le a)}.$$

Applied to $a = M$ and $a = M - \alpha_0\sigma^2 - L_{\alpha_0} t$, this inequality implies that

$$\mathbb{P}\{Z \ge M + \alpha_0\sigma^2 + L_{\alpha_0} t\} \le 2e^{-t}$$

and

$$\mathbb{P}\{Z \le M - \alpha_0\sigma^2 - L_{\alpha_0} t\} \le 2e^{-t}$$

(strictly speaking, one needs an approximation argument to be able to write weak inequalities). Setting

$$u = \alpha_0\sigma^2 + L_{\alpha_0} t = \inf_{\alpha > 0}(\alpha\sigma^2 + L_\alpha t)$$

and solving for t we get

$$t = \sup_{\alpha > 0} \frac{1}{L_\alpha}(u - \alpha\sigma^2).$$

Using the relationship between α and L_α

$$\alpha = \frac{2L_\alpha(e^{1/L_\alpha} - 1)}{1 + 2L_\alpha}$$

and rewriting the supremum in terms of $s = 1/L_\alpha$ we get the result.

□

It is very easy to show that for large values of u/σ^2 the bounds (2.5) and (2.6) have Poissonian behavior $\sim u \log \frac{u}{K\sigma^2}$, for instance, for $K = 2e$. In order to give a simple expression as an estimate of (2.5) and (2.6) and at the same time not to lose much accuracy we had to use the combination of some calculus and numerical computations. Basically, we found the condition for the supremum in (2.5), analyzed it outside of some bounded interval, and transformed an estimation problem to a robust numerical problem on this bounded interval, where we used numerical computations to preserve the accuracy. As a result we have the following corollary (we don't give the proof of it here).

Corollary 2 *Let $\sigma^2 = n \sup_{\mathcal{F}} \mu f^2$. For any $u \ge 0$,*

$$\mathbb{P}(Z \ge M + u) \le 2\exp\Big\{-0.98\, u \log\Big(1 + \frac{u}{4\sigma^2}\Big)\Big\} \tag{2.8}$$

and

$$\mathbb{P}(Z \le M - u) \le 2\exp\Big\{-0.98\, u \log\Big(1 + \frac{u}{4\sigma^2}\Big)\Big\}. \tag{2.9}$$

Remark. This result should be compared to the Theorem 5.1 of E.Rio [6] for the classes of sets. Ignoring the fact that concentration inequalities in [6] are around the mean rather than the median and comparing the tail behavior, one can show that Rio's inequalities are superior to (2.8) and (2.9); although, the right tail inequality in [6] is given for $u \le \sigma^2$ only.

Acknowledgments. We want to thank Michel Talagrand for an inspiring correspondence. We also want to thank Vladimir Koltchinskii and Jon Wellner for very helpful comments and suggestions.

References

[1] Dembo, A., Information inequalities and concentration of measure, *Ann. Probab.,* **25** (1997), 527-539.

[2] Ledoux, M., On Talagrand's deviation inequalities for product measures, *ESAIM: Probab. Statist.*, **1** (1996), 63 - 87.

[3] Massart, P., About the constants in Talagrand's concentration inequalities for empirical processes, *Ann. Probab.,* **28** (2000), 863 - 885.

[4] Boucheron, S., Lugosi, G., Massart, P., A sharp concentration inequality with applications, *Random Structures Algorithms,* **16** (2000), 277 - 292.

[5] Rio E., Inégalités exponentielles pour les processus empiriques, *C.R. Acad. Sci. Paris,* t.330, Série I (2000), 597-600.

[6] Rio E., Inégalités de concentration pour les processus empiriques de classes de parties, *Probab. Theory Relat. Fields,* 119 (2001), 163-175.

[7] Talagrand, M., Concentration of measure and isoperimetric inequalities in product spaces, *Publications Mathématiques de l'I.H.E.S.* **81** (1995), 73-205.

[8] Talagrand, M., New concentration inequalities in product spaces. *Invent. Math.,* **126** (1996) , 505-563.