Notes on Exponential Bounds via Azuma's Inequality

Jon A. Wellner

April 17, 2004

Abstract

1. Azuma's Inequality

Inequality 1. (Azuma (1967)). Suppose that $f \in L_1(P)$ and $f - Ef = \sum_{i=1}^n d_i$ where $\{d_i, \mathcal{A}_i\}_{i=1}^n$ is a martingale difference sequence:

$$E(d_i|\mathcal{A}_{i-1}) = 0, \qquad i = 1, \ldots, n.$$

[Question: What is $\mathcal{A}_0$?]
Assume that $\|d_i\|_\infty < \infty$, and set $a^2 = \sum_{i=1}^n \|d_i\|_\infty^2$. Then, for every $t > 0$,

$$Pr(f - Ef > t) \le \exp\left(-\frac{t^2}{2a^2}\right),$$

$$Pr(-(f - Ef) > t) \le \exp\left(-\frac{t^2}{2a^2}\right),$$

and

$$Pr(|f - Ef| > t) \le 2\exp\left(-\frac{t^2}{2a^2}\right).$$

Proof. Note that if Y is a random variable such that $|Y| \le 1$ a.s. and $E(Y) = 0$, then for any real number r,

$$E\exp(rY) \le \exp(r^2/2).$$

This is proved as follows: for any $|x| \le 1$,

$$\begin{aligned} \exp(rx) &\le \cosh(r) + x\sinh(r) \\ &\le \exp(r^2/2) + x\sinh(r), \end{aligned}$$

(note that $e^r = \cosh(r) + \sinh(r)$ and $e^{-r} = \cosh(r) - \sinh(r)$ so equality holds from the first to fourth lines at the endpoints $x = \pm 1$) so

$$E(\exp(rY)) \le \exp(r^2/2).$$

It follows that

$$E(\exp(rd_i)|\mathcal{A}_{i-1}) \le \exp(r^2\|d_i\|_\infty^2/2).$$

Hence we find, by iterating this inequality,

$$\begin{aligned} E\exp(r(f - Ef)) &= E\exp(r\sum_{i=1}^n d_i) \\ &= E\left\{\exp(r\sum_{i=1}^{n-1} d_i)E\left(\exp(rd_n)|\mathcal{A}_{n-1}\right)\right\} \\ &\le E\left\{\exp(r\sum_{i=1}^{n-1} d_i)\exp\left(r^2\|d_n\|_\infty^2/2\right)\right\} \\ \ldots \\ &\le \exp(r^2a^2/2). \end{aligned}$$

Thus Markov's inequality yields

$$\begin{aligned} Pr(f - Ef > t) &\le \exp(-rt + r^2a^2/2) \qquad \text{for all } \; r > 0 \\ &= \exp(-t^2/2a^2) \end{aligned}$$

by choosing $r = t/a^2$. $\square$

This form of Azuma's inequality is ideally suited for situations in which $-c_i \le d_i \le c_i$ almost surely; that is, the martingale differences take values in symmetric intervals of length $2c_i$ about 0. A slightly different formulation is as follows:

Theorem. (Hoeffding (1963), Azuma (1967)). Let $X_1, \ldots, X_n$ be a sequence of random vectors. Let $\mathcal{A}_i \equiv \sigma\{X_1, \ldots, X_i\}$, $i = 1, \ldots, n$. Suppose that $\{V_i, \mathcal{A}_i\}_{1 \le i \le n}$ is a martingale - difference sequence, and that there exist random variables $Z_1, Z_2, \ldots$ and non-negative constants $c_1, c_2, \ldots$ such that for every $i = 1, \ldots, n$, Z_i is a function of $X_1, \ldots, X_{i-1}$ and

$$Z_i \le V_i \le Z_i + c_i$$

almost surely. Then, for every $\epsilon > 0$ and n,

$$P(\sum_{i=1}^n V_i \ge \epsilon) \le \exp\left(-2\epsilon^2 / \sum_{i=1}^n c_i^2\right)$$

and

$$P(-\sum_{i=1}^n V_i \ge \epsilon) \le \exp\left(-2\epsilon^2 / \sum_{i=1}^n c_i^2\right).$$

The proof of the theorem is based on the following lemma:

Lemma. Assume that the random variables V and Z satisfy $E(V|Z) = 0$ a.s. and for some function f and constant $c > 0$,

$$f(Z) \le V \le f(Z) + c\,.$$

Then, for every $r > 0$,

$$E\left(e^{rV}|Z\right) \le \exp(r^2c^2/8)\,.$$

Proof. Recopy the proof of the lemma used to prove Hoeffding's inequality, and compute conditionally. $\square$

Proof of the Theorem: For any $r > 0$ we have

$$\begin{aligned} P(\sum_{i=1}^n V_i \ge \epsilon) &\le E\exp\left(r\sum_{i=1}^n V_i\right) \\ &= e^{-r\epsilon}E\left\{\exp\left(r\sum_{i=1}^{n-1} V_i\right) E(e^{rV_n}|\mathcal{A}_{n-1})\right\} \end{aligned}$$

$$
\begin{aligned}
&\le \quad e^{-r\epsilon} E\left\{ \exp\left(r \sum_{i=1}^{n-1} V_i \right) \exp(r^2 c_n^2/8) \right\} \\
&\le \quad \ldots \le e^{-r\epsilon} \exp(r^2 \sum_{i=1}^{n} c_i^2/8) \\
&= \quad \exp\left(-2\epsilon^2 / \sum_1^n c_i^2 \right)
\end{aligned}
$$

by choosing $r = 4\epsilon / \sum_1^n c_i^2$. □

This leads naturally to the following theorem of McDiarmid (1989) in which the hypotheses are formulated in a very convenient form for our particular applications:

Theorem. (McDiarmid (1989)). Let $X_1, \ldots, X_n$ be independent random vectors with values in A, and assume that $g : A^n \to R$ satisfies

$$
\sup_{x_1,\ldots,x_n \in A, x_i' \in A} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \le c_i \tag{1.1}
$$

for $i = 1, \ldots, n$. Then for all $\epsilon > 0$,

$$
Pr(g(X_1, \ldots, X_n) - Eg(X_1, \ldots, X_n) \ge \epsilon) \le \exp\left(-2\epsilon^2 / \sum_1^n c_i^2 \right),
$$

and

$$
Pr(-(g(X_1, \ldots, X_n) - Eg(X_1, \ldots, X_n)) \ge \epsilon) \le \exp\left(-2\epsilon^2 / \sum_1^n c_i^2 \right).
$$

Proof. Let $V \equiv g(X_1, \ldots, X_n) - Eg(X_1, \ldots, X_n)$, and set

$$
V_i \equiv \left(E^{\mathcal{A}_i} - E^{\mathcal{A}_{i-1}} \right)(V), \qquad i = 1, \ldots, n,
$$

so that $V = \sum_1^n V_i$. (Here $E^{\mathcal{A}_0} \equiv E$.) Clearly $\{V_i, \mathcal{A}_i\}$ is a martingale difference sequence. Note that

$$
\begin{aligned}
V_k &= E\{g(X_1, \ldots, X_n) | \mathcal{A}_k\} - E\{g(X_1, \ldots, X_n) | \mathcal{A}_{k-1}\} \\
&\equiv H_k(X_1, \ldots, X_k) - \int H_k(X_1, \ldots, X_{k-1}, u) dF_k(u)
\end{aligned}
$$

where F_k is the distribution of X_k. Let

$$
W_k \equiv \sup_u \left(H_k(X_1, \ldots, X_{k-1}, u) - \int H_k(X_1, \ldots, X_{k-1}, u') dF_k(u') \right)
$$

and

$$
Z_k \equiv \inf_v \left(H_k(X_1, \ldots, X_{k-1}, v) - \int H_k(X_1, \ldots, X_{k-1}, v') dF_k(v') \right).
$$

Thus $Z_k \le V_k \le W_k$ almost surely, and

$$W_k - Z_k \le \sup_u \sup_v \left(H_k(X_1,\ldots,u) - H_k(X_1,\ldots,v)\right) \le c_k$$

for $k = 1,\ldots,n$ by the hypothesis (1.1). Hence the claimed bounds hold by the Hoeffding - Azuma theorem. $\square$

Now the goal is to use the Hoeffding-Azuma-McDiarmid inequalities to prove the following two lemmas due to Koltchinskii (1999). The basic idea is closely related to the methods of Yurinskii (1974), (1976). Suppose that $X_1,\ldots,X_n$ are i.i.d. P on $(\mathcal{X},\mathcal{A})$. Let $\mathbb{P}_n \equiv n^{-1}\sum_{i=1}^n \delta_{X_i}$, and for any class of functions $\mathcal{F}$ from $\mathcal{X}$ to R, set

$$\Delta_n(\mathcal{F}) \equiv \|\mathbb{P}_n - P\|_{\mathcal{F}}\,.$$

If $r_1,\ldots,r_n$ are i.i.d. Rademacher random variables with $P(r_i = \pm 1) = 1/2$, then we define

$$\mathbb{P}_n^s \equiv n^{-1}\sum_{i=1}^n r_i \delta_{X_i}\,, \qquad \text{and} \qquad R_n(\mathcal{F}) \equiv \|\mathbb{P}_n^s\|_{\mathcal{F}}\,.$$

Lemma 2.2. For any countable class of functions $\mathcal{F}$ with $f : \mathcal{X} \to [0,1]$ for each $f \in \mathcal{F}$, we have

$$\begin{aligned}
&Pr(\Delta_n(\mathcal{F}) \ge E\Delta_n(\mathcal{F}) + \epsilon) \le \exp(-2n\epsilon^2)\,,\\
&Pr(E\Delta_n(\mathcal{F}) \ge \Delta_n(\mathcal{F}) + \epsilon) \le \exp(-2n\epsilon^2)\,,\\
&Pr(|\Delta_n(\mathcal{F}) - E\Delta_n(\mathcal{F})| \ge \epsilon) \le 2\exp(-2n\epsilon^2)\,.
\end{aligned}$$

Lemma 2.3. For any countable class of functions $\mathcal{F}$ with $f : \mathcal{X} \to [0,1]$ for each $f \in \mathcal{F}$, we have

$$\begin{aligned}
&Pr(R_n(\mathcal{F}) \ge ER_n(\mathcal{F}) + \epsilon) \le \exp(-n\epsilon^2/2)\,,\\
&Pr(ER_n(\mathcal{F}) \ge R_n(\mathcal{F}) + \epsilon) \le \exp(-n\epsilon^2/2)\,,\\
&Pr(|R_n(\mathcal{F}) - ER_n(\mathcal{F})| \ge \epsilon) \le 2\exp(-n\epsilon^2/2)\,.
\end{aligned}$$

Proof of Lemma 2.2. Let $Z_i \equiv \delta_{X_i} - P$. Then take

$$V \equiv n\|\mathbb{P}_n - P\|_{\mathcal{F}} = \|\sum_{i=1}^n Z_i\|_{\mathcal{F}} \equiv n\Delta_n(\mathcal{F})\,,$$

and with $\mathcal{A}_i = \sigma\{X_1,\ldots,X_i\}$, $i = 1,\ldots,n$, define

$$V_i \equiv \left(E^{\mathcal{A}_i} - E^{\mathcal{A}_{i-1}}\right)(V) = E(V|\mathcal{A}_i) - E(V|\mathcal{A}_{i-1})\,.$$

Note that

$$V = \| \sum_{j=1}^{n} Z_j \|_{\mathcal{F}} \equiv g(X_1, \ldots, X_n)$$

satisfies

$$\begin{aligned}
&|g(X_1, \ldots, X_n) - g(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_n)| \\
&= ||\| \sum_{j \neq i} Z_j + Z_i \|_{\mathcal{F}} - \| \sum_{j \neq i} Z_j + Z_i' \|_{\mathcal{F}}| \\
&\leq \| Z_i - Z_i' \|_{\mathcal{F}} = \| f(X_i) - Pf - (f(X_i') - Pf) \|_{\mathcal{F}} \\
&= \| f(X_i) - f(X_i') \|_{\mathcal{F}} \leq 1 \,.
\end{aligned}$$

Thus the hypotheses of McDiarmid's theorem hold with $c_i = 1, \ldots, n$, and $\sum_{i=1}^{n} c_i^2 = n$. Hence it follows that

$$Pr(V - EV > t) \quad \leq \quad \exp\left(-\frac{2t^2}{n} \right)$$

and this yields

$$Pr(\Delta_n(\mathcal{F}) - E\Delta_n(\mathcal{F}) > \epsilon) \leq \exp(-2n\epsilon^2) \,.$$

The proof of the second inequality is the same, and the third statement follows from the first two. $\square$

Proof of Lemma 2.3. Let $Z_i \equiv r_i \delta_{X_i}$. Then take

$$V \equiv n \| \mathbb{P}_n^s \|_{\mathcal{F}} = \| \sum_{i=1}^{n} Z_i \|_{\mathcal{F}} \equiv n R_n(\mathcal{F}) \,,$$

and with $\mathcal{A}_i = \sigma\{\tilde{X}_1, \ldots, \tilde{X}_i\}$, $\tilde{X}_i \equiv (r_i, X_i)$, $i = 1, \ldots, n$, define

$$V_i \equiv \left(E^{\mathcal{A}_i} - E^{\mathcal{A}_{i-1}} \right) (V) = E(V|\mathcal{A}_i) - E(V|\mathcal{A}_{i-1}) \,.$$

Note that

$$V = \| \sum_{i=1}^{n} Z_j \|_{\mathcal{F}} \equiv g(\tilde{X}_1, \ldots, \tilde{X}_n)$$

satisfies

$$\begin{aligned}
&|g(\tilde{X}_1, \ldots, \tilde{X}_n) - g(\tilde{X}_1, \ldots, \tilde{X}_{i-1}, \tilde{X}_i', \tilde{X}_{i+1}, \ldots, \tilde{X}_n)| \\
&= ||\| \sum_{j \neq i} Z_j + Z_i \|_{\mathcal{F}} - \| \sum_{j \neq i} Z_j + Z_i' \|_{\mathcal{F}}| \\
&\leq \| Z_i - Z_i' \|_{\mathcal{F}} = \| r_i f(X_i) - r_i' f(X_i') \|_{\mathcal{F}} \\
&= \| r_i f(X_i) - r_i' f(X_i') \|_{\mathcal{F}} \leq 2 \,.
\end{aligned}$$

Thus the hypotheses of McDiarmid's theorem hold with $c_i = 2$ for $i = 1, \ldots, n$, and $\sum_{i=1}^n c_i^2 = 4n$. Hence it follows that

$$Pr(V - EV > t) \quad \leq \quad \exp\left(-\frac{t^2}{2n}\right)$$

and this yields

$$Pr(R_n(\mathcal{F}) - ER_n(\mathcal{F}) > \epsilon) \leq \exp(-n\epsilon^2/2)\,.$$

The proof of the second inequality is the same, and the third statement follows from the first two. $\square$

To see what Lemma 2.2 means in terms of exponential bounds for $\Delta_n(\mathcal{F})$, we have the following Lemma.

Lemma 2.3. For all $t \geq E[\sqrt{n}\Delta_n(\mathcal{F})]$,

$$Pr(\sqrt{n}\Delta_n(\mathcal{F}) \geq t) \quad \leq \quad \exp(4tE(\sqrt{n}\Delta_n(\mathcal{F})))\exp(-2t^2)\,.$$

Proof. By using the first inequality of Lemma 2.2 we find that, for $t \geq E[\sqrt{n}\Delta_n(\mathcal{F})]$,

$$\begin{aligned} Pr(\sqrt{n}\Delta_n(\mathcal{F}) \geq t) &= Pr\left(\Delta_n(\mathcal{F}) - E[\Delta_n(\mathcal{F})] \geq (t - E[\sqrt{n}\Delta_n(\mathcal{F})])/\sqrt{n}\right) \\ &\leq \exp\left(-2(t - E[\sqrt{n}\Delta_n(\mathcal{F})])^2\right) \\ &\leq \exp\left(4tE[\sqrt{n}\Delta_n(\mathcal{F})]\right)\exp(-2t^2)\,. \end{aligned}$$

$\square$

Note that this is completely consistent with the bounds resulting from Kiefer (1961), Alexander (1983), and Talagrand (1994): for any Donsker class $\mathcal{F}$, and in particular for $VC-$classes $\mathcal{F}$, we can bound $E[\sqrt{n}\Delta_n(\mathcal{F})]$ in terms of covering numbers and the envelope of the class $\mathcal{F}$; see e.g. Van der Vaart and Wellner (1996), Theorems 2.14.1 and 2.14.2, pages 239 and 240. Thus if we assume that

$$E[\sqrt{n}\Delta_n(\mathcal{F})] \leq K$$

for a constant K independent of n, the bound of Lemma 2.3 can be further bounded by

$$\exp(4Kt)\exp(-2t^2)\,,$$

and the term $\exp(4Kt)$ is larger than the polynomial functions of t appearing in the bounds of Talagrand (1994).

REFERENCES

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tokuku Math J.* **19**, 357 - 367.

Koltchinskii, V. (1999). Rademacher penalties and structural risk minization. *Technical Report*, University of New Mexico.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics, 1989*, pages 148 - 188. Cambridge University Press, Cambridge.

Yurinskii, Y.V. (1974). Exponential bounds for large deviations. *Theor. Probab. Appl.* **19**, 154 - 155.

Yurinskii, Y.V. (1976). Exponential inequalities for sums of random vectors. *J. Mult. Anal.* **6**, 473 - 499.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probability* **22**, 28 -76.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Department of Statistics
University of Washington
P.O. Box 354322
Seattle, Washington 98195-4322
U.S.A.
e-mail: jaw@stat.washington.edu