McDiarmid's inequalities of Bernstein and Bennett forms

Yiming Ying
Department of Mathematics, City University of Hong Kong
Kowloon, Hong Kong, CHINA.
ymying@cityu.edu.hk

Abstract

In this note, we will reprove the McDiarmid's inequality with more elementary analysis. Moreover we derive its variant Bennett and Berstein's inequalities.

McDiarmid's inequlity was first proved in paper [1] using Martingale theory. This method has been widely used in combinatorial applications [1] and in learning theory [3, 4]. However if we assume the variables are independent, the proof will be very elementary. In this note, we will reprove it and give its variants of Bennett and Berstein's types.

Before we list the propositions, let's give some notations. Given a family of independent random variables $\mathbf{z} = (z_1, z_2, \cdots, z_n)$ with z_k in a set Ω_k according to a distribution P_k for each k . Suppose that the real valued function $f : \prod_{k=1}^{n} \Omega_k \to \mathbb{R}$. In order to prove our results, it is useful to introduce new functions

$$\begin{cases} g_n(z_1, \cdots, z_n) = f(z_1, z_2, \cdots, z_n) - \mathbb{E}_{z_n}\Big(f(z_1, z_2, \cdots, z_n)\Big) \\ g_{n-1}(z_1, z_2, \cdots, z_{n-1}) = \mathbb{E}_{z_n}\Big(f(z_1, \cdots, z_{n-1}, z_n)\Big) - \mathbb{E}_{z_{n-1},z_n}\Big(f(z_1, z_2, \cdots, z_{n-1}, z_n)\Big) \\ \vdots \\ g_k(z_1, z_2, \cdots, z_k) = \mathbb{E}_{z_{k+1},\cdots,z_n}\Big(f(\mathbf{z})\Big) - \mathbb{E}_{z_k,\cdots,z_n}\Big(f(\mathbf{z})\Big) \\ \vdots \\ g_1(z_1) = \mathbb{E}_{z_2,\cdots,z_n}\Big(f(z_1, \cdots, z_n)\Big) - \mathbb{E}_{z_1,\cdots,z_n}\Big(f(z_1, z_2, \cdots, z_n)\Big) \end{cases}$$

then we have

$$\begin{cases} \sum_{k=1}^{n} g_k(z_1, z_2, \cdots, z_k) = f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \\ \mathbb{E}_{z_k}(g(z_1, z_2, \cdots, z_k)) = 0 \qquad \text{for all} \quad 1 \le k \le n \end{cases} \tag{1}$$

Definition 1. *We say the function* $f : \prod_{k=1}^{n} \Omega_k \to \mathbb{R}$ *with bounded differences* $\{c_k\}_{k=1}^{n}$ *if, for all* $1 \le k \le n$*,*

$$\sup_{z_1,\cdots,z_{k-1},z_k,z_k',\cdots,z_n} |f(z_1, \cdots, z_{k-1}, z_k, z_{k+1}, \cdots, z_n) - f(z_1, \cdots, z_{k-1}, z_k', z_{k+1}, \cdots, z_n)| \le c_k$$

Proposition 1. *(McDiarmid's inequality) Suppose* $f : \prod_{k=1}^{n} \Omega_k \to \mathbb{R}$ *with bounded differences* $\{c_k\}_{k=1}^{n}$ *then , for all* $\epsilon > 0$*, there holds*

$$\mathbf{Pr}_{\mathbf{z}} \Big\{ f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \ge \epsilon \Big\} \le e^{-\frac{2\epsilon^2}{\sum_{k=1}^{n} c_k^2}}.$$

Lemma 1. *If a random variable* X *satisfies* $\mathbb{E}X = 0$ *and* $a \le X \le b$*, then* $\mathbb{E}(e^{hX}) \le e^{\frac{1}{8}h^2(b-a)^2}$ *for all* $h > 0$

Proof of Lemma 1 . One can find the the proof in [2]. For the completeness we list the proof here. Indeed, by the convexity of e^{hX}, we have

$$e^{hX} \le (\frac{X-a}{b-a})e^{hb} + (\frac{b-X}{b-a})e^{ha}.$$

Therefore

$$\mathbb{E}(e^{hX}) \le \frac{b}{b-a}e^{ha} + \frac{-a}{b-a}e^{hb} = (1-p)e^{-py} + pe^{(1-p)y} = e^{f(y)}$$

where $p = (\frac{-a}{b-a}), y = (b-a)h, f(y) = -py + \log(1-p+pe^y)$

The fact $f(0) = f'(0) = 0$ and $f''(y) = \frac{p(1-p)e^{-y}}{(p+(1-p)e^{-y})^2} \le \frac{1}{4}$ gives the claim. □

Proof of Proposition 1 . Set

$$b_k := \sup_{z_k} g_k(z_1, z_2, \cdots, z_k) \qquad a_k := \inf_{z_k} g_k(z_1, z_2, \cdots, z_k).$$

hence we get

$$a_k \le g_k(z_1, \cdots, z_k) \le b_k \quad \text{and } 0 \le b_k - a_k \le c_k.$$

Moreover, utilizing property (1) and the Markov inequality we obtain for any $h > 0$,

$$\mathbf{Pr_z}\left\{f(\mathbf{z}) - \mathbb{E}_\mathbf{z} f(\mathbf{z}) \geq \epsilon\right\} \leq e^{-h\epsilon}\mathbb{E}_\mathbf{z}\left(e^{h(f(\mathbf{z})-\mathbb{E}_\mathbf{z} f(\mathbf{z}))}\right) = e^{-h\epsilon}\mathbb{E}_\mathbf{z}\left(e^{h\sum_{k=1}^n g_k}\right).$$

Rewrite the expected value as an integral, we have

$$\int_{z_1,z_2,\cdot\cdot,z_n} e^{h\sum_{k=1}^n g_k} dP_1(z_1)\cdots dP_n(z_n)$$

$$:= \int_{z_1,\cdots,z_{n-1}} \prod_{k=1}^{n-1} e^{hg_k}\left(\int_{z_n} e^{hg_n} dP_n(z_n)\right) dP_1(z_1)\cdots dP_{n-1}(z_{n-1})$$

Now we estimate the term in the brace. Since $\mathbb{E}_{z_n} g_n(z_1,\cdots,z_{n-1},z_n) = 0$ in seen in (1), we get by the Lemma 1,

$$\int_{z_n} e^{hg_n(z_1,z_2,\cdots,z_{n-1},z_n)} dP_n(z_n) \leq e^{\frac{1}{8}h^2(b_n-a_n)^2} \leq e^{\frac{1}{8}h^2c_n^2}.$$

Therefore we have

$$\mathbb{E}_\mathbf{z}\left(e^{h\sum_{k=1}^n g_k}\right) \leq \exp\left\{\frac{h^2c_n^2}{8}\right\}\int_{z_1,\cdots,z_{n-1}} \prod_{k=1}^{n-1} e^{hg_k} dP_1(z_1)\cdots dP_{n-1}(z_{n-1}).$$

Using the property $\mathbb{E}_{z_{n-1}} g_{n-1}(z_1,z_2,\cdots,z_{n-1}) = 0$ again, repeat the above procedure we have

$$\mathbb{E}_\mathbf{z}\left\{e^{h\sum_{k=1}^n g_k}\right\} \leq \exp\left(\frac{h^2(c_n^2+c_{n-1}^2)}{8}\right)\int_{z_1,\cdots,z_{n-2}} \prod_{k=1}^{n-2} e^{hg_k} dP_1(z_1)\cdots dP_{n-2}(z_{n-2}).$$

Repeat the above procedure n times, we finally obtain

$$\mathbb{E}_\mathbf{z}\left(e^{h\sum_{k=1}^n g_k}\right) \leq \exp\left\{\frac{h^2\sum_{k=1}^n c_k^2}{8}\right\}$$

which yields

$$\mathbf{Pr_z}\left\{f(\mathbf{z}) - \mathbb{E}_\mathbf{z} f(\mathbf{z}) \geq \epsilon\right\} \leq \exp\left\{-h\epsilon + \frac{h^2}{8}\sum_{k=1}^n c_k^2\right\}.$$

Set $h = \frac{4\epsilon}{\sum_{k=1}^n c_k^2}$, we get the McDiarmid inequality

$$\mathbf{Pr_z}\left\{f(\mathbf{z}) - \mathbb{E}_\mathbf{z} f(\mathbf{z}) \geq \epsilon\right\} \leq \exp\left\{-\frac{2\epsilon^2}{\sum_{k=1}^n c_k^2}\right\}.$$

$\square$

In the following, we will derive inequalities of "Bennett" and "Bernstein" forms . It is necessary to introduce some notations. Denote

$$\begin{cases} V_k := \sup_{\mathbf{z}\setminus z_k} \mathbb{E}_{z_k}\Big(f(\mathbf{z}) - \mathbb{E}_{z_k} f(\mathbf{z})\Big)^2 \\ \tilde{\sigma}^2 := \sum_{k=1}^{n} V_k \qquad , \\ B := \max_{1\le k\le n} \sup_{\mathbf{z}} |f(\mathbf{z}) - \mathbb{E}_{z_k} f(\mathbf{z})| \end{cases} \tag{2}$$

Proposition 2. *With the notations above and $\epsilon > 0$, we have*

$$\mathbf{Pr}_{\mathbf{z}}\Big\{f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \ge \epsilon\Big\} \le \exp\Big\{-\frac{\epsilon}{2B}\log\Big(1 + \frac{B\epsilon}{\tilde{\sigma}^2}\Big)\Big\}$$

and

$$\mathbf{Pr}_{\mathbf{z}}\Big\{f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \ge \epsilon\Big\} \le \exp\Big\{-\frac{\epsilon^2}{2(\tilde{\sigma}^2 + \frac{B\epsilon}{3})}\Big\}.$$

Proof. For any $h > 0$, we have

$$\mathbf{Pr}_{\mathbf{z}}\Big\{f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \ge \epsilon\Big\} \le e^{-h\epsilon}\mathbb{E}_{\mathbf{z}}\Big(e^{h\sum_{k=1}^{n} g_k}\Big).$$

Now we write the expected value as integral

$$\int_{z_1,z_2,\cdots,z_n} e^{h\sum_{k=1}^{n} g_k} dP_1(z_1)\cdots dP_n(z_n)$$

$$:= \int_{z_1,\cdots,z_{n-1}} \prod_{k=1}^{n-1} e^{hg_k}\Big(\int_{z_n} e^{hg_n} dP_n(z_n)\Big) dP_1(z_1)\cdots dP_{n-1}(z_{n-1}). \tag{3}$$

Applying Taylor expansion to e^{hg_n} combined with $\mathbb{E}_{z_n} g_n(z_1,\cdots,z_{n-1},z_n) = 0$ as shown in (1), we get

$$\begin{aligned} \int_{z_n} e^{hg_n(z_1,z_2,\cdots,z_n)} dP_n(z_n) &= \int_{z_n}\Big(1 + hg_n + \frac{h^2 g_n^2}{2} + \cdots\Big) dP_n(z_n) \\ : &= \int_{z_n}\Big[1 + h^2 g_n^2 G(hg_n)\Big] dP_n(z_n) \end{aligned}$$

where $G(x) := \frac{e^x - 1 - x}{x^2}$.

Observe $G(x)$ is increasing. This can be seen by the following

$$G'(x) = x^{-3}\Big((x-2)e^x + 2 + x\Big)$$

Set $h(x) = (x-2)e^x + 2 + x$, then $h'(x) = (x-1)e^x + 1, h''(x) = xe^x$ and $h'(0) = 0, h(0) = 0$. Then one can see 0 is the only minimum point of $h'(x)$. Hence $h'(x) \geq 0$ for all x. That is, $h(x)$ is nondecreasing. Note that $h(0) = 0$, which means $h(x) \leq 0$ for $x < 0$ and $h(x) \geq 0$ for $x > 0$. Therefore we have $G'(x) \geq 0$.

Hence we get, by the definition of B, V_n as shown in (2),

$$\int_{z_n} e^{hg_n(z_1,z_2,\cdots,z_n)} dP_n(z_n) \leq 1 + h^2 V_n G(hB) \leq \exp\Big\{h^2 V_n G(hB)\Big\}.$$

Repeating the procedure above n times, we finally have

$$\begin{aligned}
(3) \leq\ & \exp\Big\{h^2 V_n G(hB)\Big\} \int_{z_1,z_2,\cdots,z_{n-1}} \prod_{k=1}^{n-1} e^{hg_k} dP_1(z_1)\cdots dP_{n-1}(z_{n-1}) \\
\leq\ & \exp\Big\{h^2 \Big[V_n + V_{n-1}\Big] G(hB)\Big\} \int_{z_1,\cdots,z_{n-2}} \prod_{k=1}^{n-2} e^{hg_k} dP_1(z_1)\cdots dP_{n-2}(z_{n-2}) \\
\leq\ & \exp\Big\{h^2 \tilde{\sigma}^2 G(hB)\Big\}.
\end{aligned}$$

Therefore we obtain

$$\begin{aligned}
\mathbf{Pr}_{\mathbf{z}}\Big\{f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \geq \epsilon\Big\} &\leq \exp\Big\{-h\epsilon + h^2\tilde{\sigma}^2 G(hB)\Big\} \\
&= \exp\Big\{-h\epsilon + \tfrac{\tilde{\sigma}^2}{B^2}\Big[e^{hB} - 1 - hB\Big]\Big\}.
\end{aligned}$$

Set $h = \frac{1}{B}\log\Big(1 + \frac{B\epsilon}{\tilde{\sigma}^2}\Big)$, then

$$\mathbf{Pr}_{\mathbf{z}}\Big\{f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \geq \epsilon\Big\} \leq \exp\Big\{-\frac{\tilde{\sigma}^2}{B^2}\Big[(1 + \frac{B\epsilon}{\tilde{\sigma}^2})\log(1 + \frac{B\epsilon}{\tilde{\sigma}^2}) - \frac{B\epsilon}{\tilde{\sigma}^2}\Big]\Big\}.$$

If we apply the inequality

$$(1+x)\log(1+x) - x \geq \frac{x}{2}\log(1+x) \qquad \text{for all} \quad x \geq 0$$

we get the "Bennett" inequality,

$$\mathbf{Pr}_{\mathbf{z}}\Big\{f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \geq \epsilon\Big\} \leq \exp\Big\{-\frac{\epsilon}{2B}\log\Big(1 + \frac{B\epsilon}{\tilde{\sigma}^2}\Big)\Big\}.$$

If we use the inequality

$$(1+x)\log(1+x) - x \geq \frac{3x^2}{6+2x} \qquad \text{for all} \quad x \geq 0$$

then we get the "Bernstein" inequality,

$$\mathbf{Pr}_{\mathbf{z}}\left\{ f(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} f(\mathbf{z}) \geq \epsilon \right\} \leq \exp\left\{ -\frac{\epsilon^2}{2(\tilde{\sigma}^2 + \frac{B\epsilon}{3})} \right\}.$$

□

Acknowledgement: I would like to thank Professor D. X. Zhou for his constructive suggestions and insights.

References

[1] C.McDiarmid. Surveys in Combinatorics, Chapter On the methods of bounded differences, page 148-188. Cambridge University Press,Cambridge(UK),1989.

[2] D.Pollard, Convergence of Stochastic Processes, Springer-Verlag, New York, 1984.

[3] Olivier Bousquet and Andre Elisseeff, Stabilty and Generalization, J. of Machine Learning Research 2(2002) 499-526.

[4] P. L. Bartlett and S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research, 3:463-482, 2002.