Statistics and Probability Letters 131 (2017) 87–92

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A refined Hoeffding's upper tail probability bound for sum of independent random variables

Songfeng Zheng

Department of Mathematics, Missouri State University, 901 S. National Ave., Springfield, MO 65897, USA

ARTICLE INFO

Article history:
Received 23 March 2017
Received in revised form 16 August 2017
Accepted 21 August 2017
Available online 30 August 2017

Keywords:
Hoeffding's inequality
Arithmetic-geometric mean inequality
Variance

ABSTRACT

Using a refined arithmetic–geometric mean inequality, this paper gives an improved version of Hoeffding's inequality which has a closed form and is easy to evaluate. Numerical simulation comparing the performance of the proposed inequality to the original version of Hoeffding's inequality is also presented.

1. Introduction

In probability theory, Hoeffding's inequality (Hoeffding, 1963) provides an upper bound on the probability that the sum of independent random variables deviates from its expected value. It has wide applications in random algorithm analysis (Dubhashi and Panconesi, 2012), statistical learning theory (Bousquet et al., 2004; Mendelson, 2003), information theory (Raginsky and Sason, 2013), and random matrix theory (Bai, 1999), to name a few. For convenience of reference, we give the full statement of Hoeffding's inequality and the complete proof below.

Theorem 1 (*Hoeffding's Inequality*). *Let $X_1, X_2, \ldots, X_n$ be independent (not necessarily identically distributed) random variables with $P(0 \le X_i \le 1) = 1$ for $i = 1, 2, \ldots, n$. Let $p_i = \mathrm{E}(X_i)$ for $i = 1, 2, \ldots, n$, $p = \frac{1}{n}\sum_{i=1}^{n} p_i$, and $q = 1 - p$. Then for any $0 < t < q$*

$$P(\bar{X} > p + t) \le \exp\left(-n\left[(p+t)\log\frac{p+t}{p} + (q-t)\log\frac{q-t}{q}\right]\right) \equiv U_H, \tag{1}$$

where $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

Proof. Let $X = \sum_{i=1}^{n} X_i$. For any $\lambda > 0$, we have

$$P(\bar{X} > p + t) = P(X > n(p+t)) = P(e^{\lambda X} > e^{n\lambda(p+t)}) \le \frac{\mathrm{E}[e^{\lambda X}]}{e^{n\lambda(p+t)}} \tag{2}$$

$$= \frac{\mathrm{E}[e^{\lambda \sum_{i=1}^{n} X_i}]}{e^{n\lambda(p+t)}} = \frac{\mathrm{E}[\prod_{i=1}^{n} e^{\lambda X_i}]}{e^{n\lambda(p+t)}} = \frac{\prod_{i=1}^{n} \mathrm{E}[e^{\lambda X_i}]}{e^{n\lambda(p+t)}}, \tag{3}$$

where the inequality in Eq. (2) follows from Markov's inequality, and Eq. (3) uses the independence assumption.

E-mail address: SongfengZheng@MissouriState.edu.

http://dx.doi.org/10.1016/j.spl.2017.08.012
0167-7152/

Using the convexity of function $e^{\lambda x}$ on the interval $[0, 1]$, there is

$$e^{\lambda x} \leq 1 - x + xe^{\lambda}. \tag{4}$$

Applying Eq. (4) to Eq. (3), yields

$$\begin{aligned} P(\bar{X} > p+t) &\leq e^{-n\lambda(p+t)} \prod_{i=1}^{n} \mathrm{E}[1 - X_i + X_i e^{\lambda}] \\ &= \exp\left(-n\lambda(p+t)\right) \prod_{i=1}^{n} (1 - p_i + p_i e^{\lambda}) \\ &\leq \exp\left(-n\lambda(p+t)\right) \left(\frac{\sum_{i=1}^{n}(1 - p_i + p_i e^{\lambda})}{n} \right)^{n} && (5) \\ &= \exp\left(-n\lambda(p+t)\right) (q + pe^{\lambda})^{n} && (6) \\ &= \exp\left(-n\left[\lambda(p+t) - \log(q + pe^{\lambda})\right]\right), && (7) \end{aligned}$$

where in Eq. (5), we used the arithmetic–geometric mean inequality which states that for positive $x_1, x_2, \ldots, x_n$,

$$\prod_{i=1}^{n} x_i \leq \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^{n}. \tag{8}$$

Choosing λ to minimize Eq. (7), we obtain

$$\lambda^* = \log\left[\frac{(p+t)q}{p(q-t)} \right]. \tag{9}$$

Substituting λ^* into Eq. (7), we can obtain the desired result in Eq. (1). □

2. The refined Hoeffding's inequality

In the proof of Theorem 1, the arithmetic–geometric mean inequality in Eq. (8) was used as a major ingredient. There is a refined version of the arithmetic–geometric mean inequality which makes it possible to improve the Hoeffding's inequality in Eq. (1).

Lemma 2 (*Refined Arithmetic-Geometric Mean Inequality*)**.** *Suppose that $x_k \in [a, b]$ with $a > 0$, $p_k \geq 0$ for $k = 1, 2, \ldots, n$, and further assume that $\sum_{k=1}^{n} p_k = 1$. Let $\bar{x} = \sum_{k=1}^{n} p_k x_k$. We have*

$$\frac{1}{2b} \sum_{k=1}^{n} p_k (x_k - \bar{x})^2 \leq \bar{x} - \prod_{k=1}^{n} x_k^{p_k} \leq \frac{1}{2a} \sum_{k=1}^{n} p_k (x_k - \bar{x})^2. \tag{10}$$

The proof of Lemma 2 is given in Cartwright and Field (1978). We are interested in a special case in which each $p_k = 1/n$, then

$$\sum_{k=1}^{n} p_k (x_k - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2 = \mathrm{var}(\mathbf{x}),$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and $\mathrm{var}(\mathbf{x})$ denotes the variance[1] of the array $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ with equiprobable outcomes. Thus, for this special case, there is

$$\frac{1}{2b} \mathrm{var}(\mathbf{x}) \leq \frac{x_1 + x_2 + \cdots + x_n}{n} - \left(\prod_{k=1}^{n} x_k \right)^{1/n} \leq \frac{1}{2a} \mathrm{var}(\mathbf{x}),$$

and consequently

$$\prod_{k=1}^{n} x_k \leq \left(\frac{x_1 + x_2 + \cdots + x_n}{n} - \frac{1}{2b} \mathrm{var}(\mathbf{x}) \right)^{n}. \tag{11}$$

By comparing Eq. (8) to Eq. (11), clearly, Eq. (11) provides a tighter bound for the product of the given numbers by incorporating the variance information.

[1] Although it is not exactly the variance, calling it as variance simplifies our notation greatly.

Theorem 3 (*Improved Hoeffding's Inequality*). *With the same notations and settings as in Theorem 1, let*

$$A = (q-t)\left(p - \frac{\sigma^2}{2}\right), \quad B = -(p+t)(q+\sigma^2), \quad C = \frac{\sigma^2}{2}(1+p+t),$$

where $\sigma^2 = \sum_{i=1}^n (p_i - p)^2/n$, *and denote*

$$w = \frac{-B + \sqrt{B^2 - 4AC}}{2A},$$

where $B^2 - 4AC$ *is always positive. Then,*

$$P(\bar{X} > p+t) \le w^{-n(p+t)}\left(q + pw - \frac{\sigma^2(w-1)^2}{2w}\right)^n \equiv U_N. \tag{12}$$

Before we give the proof of Theorem 3, we summarize some useful properties of the numbers A, B, and C as follows:

Lemma 4. *With the same notations as Theorem 3, we have* $A > 0$, $A + B + C < 0$, *and* $B^2 - 4AC > 0$.

Proof. Since

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n (p_i - p)^2 = \frac{1}{n}\sum_{i=1}^n p_i^2 - p^2,$$

there is

$$p - \sigma^2 = p - \frac{1}{n}\sum_{i=1}^n p_i^2 + p^2 = \frac{1}{n}\sum_{i=1}^n (p_i - p_i^2) + p^2 > 0,$$

since every $p_i \in [0, 1]$. Consequently,

$$A = (q-t)\left(p - \frac{\sigma^2}{2}\right) > 0, \tag{13}$$

because $q > t$. Moreover,

$$\begin{aligned} A + B + C &= (q-t)\left(p - \frac{\sigma^2}{2}\right) - (p+t)(q+\sigma^2) + \frac{\sigma^2}{2}(1+p+t) \\ &= (pq - pt - q\sigma^2/2 + t\sigma^2/2) - (pq + tq + p\sigma^2 + t\sigma^2) + (\sigma^2/2 + p\sigma^2/2 + t\sigma^2/2) \\ &= -pt - qt - q\sigma^2/2 - p\sigma^2 + \sigma^2/2 + p\sigma^2/2 = -t < 0, \end{aligned} \tag{14}$$

where the last equality follows since $p + q = 1$.

Finally, let us calculate

$$\begin{aligned} B^2 - 4AC &= (p+t)^2(q+\sigma^2)^2 - 4(q-t)\left(p - \frac{\sigma^2}{2}\right)\frac{\sigma^2}{2}(1+p+t) \\ &= (p^2q^2 + p^2\sigma^4 - 2pq\sigma^2 + q\sigma^4 + pq\sigma^4) + (2pt\sigma^4 - t\sigma^4 - pt\sigma^4 + qt\sigma^4) \\ &\quad + (4pqt\sigma^2 - 2pqt\sigma^2) + (q^2t^2 + 2pq^2t + 2q\sigma^2t^2 + 2p\sigma^2t + 2p^2\sigma^2t + 2p\sigma^2t^2). \end{aligned} \tag{15}$$

In Eq. (15),

$$2pt\sigma^4 - t\sigma^4 - pt\sigma^4 + qt\sigma^4 = (p+q)t\sigma^4 - t\sigma^4 = 0,$$

since $p + q = 1$; and

$$\begin{aligned} p^2q^2 + p^2\sigma^4 - 2pq\sigma^2 + q\sigma^4 + pq\sigma^4 &= (pq)^2 - 2pq\sigma^2 + \sigma^4(p^2 + pq + q) \\ &= (pq)^2 - 2pq\sigma^2 + \sigma^4 = (pq - \sigma^2)^2 \ge 0; \end{aligned}$$

and

$$\begin{aligned} r(t) &= (4pqt\sigma^2 - 2pqt\sigma^2) + (q^2t^2 + 2pq^2t + 2q\sigma^2t^2 + 2p\sigma^2t + 2p^2\sigma^2t + 2p\sigma^2t^2) \\ &= q^2t^2 + 2pq^2t + 2(p+q)t^2\sigma^2 + 2t\sigma^2(pq + p + p^2) \\ &= q^2t^2 + 2pq^2t + 2t^2\sigma^2 + 4pt\sigma^2 > 0, \end{aligned}$$

since every term in $r(t)$ is positive. Hence,

$$B^2 - 4AC = (pq - \sigma^2)^2 + r(t) > 0. \quad \square \tag{16}$$

We now proceed to

Proof of Theorem 3. For each i, since $p_i \in [0, 1]$, we have $1 + p_i(e^\lambda - 1) \in [1, e^\lambda]$. Thus, according to Eq. (11),

$$\prod_{i=1}^{n}(1 - p_i + p_i e^\lambda) \le \left(\frac{\sum_{i=1}^{n}(1 - p_i + p_i e^\lambda)}{n} - \frac{\sigma^2(e^\lambda - 1)^2}{2e^\lambda}\right)^n = \left(q + pe^\lambda - \frac{\sigma^2(e^\lambda - 1)^2}{2e^\lambda}\right)^n. \tag{17}$$

Similar to the proof of Theorem 1, it follows that

$$\begin{aligned} P(\bar{X} > p + t) &\le \exp\left(-n\lambda(p + t)\right)\prod_{i=1}^{n}(1 - p_i + p_i e^\lambda) \\ &\le \exp\left(-n\left[\lambda(p + t) - \log\left(q + pe^\lambda - \frac{\sigma^2}{2}e^\lambda + \sigma^2 - \frac{\sigma^2}{2}e^{-\lambda}\right)\right]\right), \end{aligned} \tag{18}$$

where Eq. (18) follows by applying Eq. (17).

To minimize Eq. (18), let

$$g(\lambda) = \lambda(p + t) - \log\left(q + pe^\lambda - \frac{\sigma^2}{2}e^\lambda + \sigma^2 - \frac{\sigma^2}{2}e^{-\lambda}\right),$$

and we maximize $g(\lambda)$ with respect to $\lambda > 0$. For this purpose, set

$$g'(\lambda) = p + t - \frac{\left(p - \frac{\sigma^2}{2}\right)e^\lambda + \frac{\sigma^2}{2}e^{-\lambda}}{q + pe^\lambda - \frac{\sigma^2}{2}e^\lambda + \sigma^2 - \frac{\sigma^2}{2}e^{-\lambda}} = 0,$$

which is equivalent to

$$p + t = \frac{\left(p - \frac{\sigma^2}{2}\right)e^{2\lambda} + \frac{\sigma^2}{2}}{\left(p - \frac{\sigma^2}{2}\right)e^{2\lambda} + (q + \sigma^2)e^\lambda - \frac{\sigma^2}{2}}$$

or

$$\left(p - \frac{\sigma^2}{2}\right)e^{2\lambda} + \frac{\sigma^2}{2} = (p + t)\left(p - \frac{\sigma^2}{2}\right)e^{2\lambda} + (p + t)(q + \sigma^2)e^\lambda - \frac{\sigma^2}{2}(p + t),$$

which is

$$(q - t)\left(p - \frac{\sigma^2}{2}\right)e^{2\lambda} - (p + t)(q + \sigma^2)e^\lambda + \frac{\sigma^2}{2}(1 + p + t) = 0. \tag{19}$$

Denote

$$A = (q - t)\left(p - \frac{\sigma^2}{2}\right), \quad B = -(p + t)(q + \sigma^2), \quad C = \frac{\sigma^2}{2}(1 + p + t), \quad \text{and} \quad W = e^\lambda,$$

then Eq. (19) is

$$AW^2 + BW + C = 0. \tag{20}$$

According to Lemma 4, $B^2 - 4AC > 0$. Thus, Eq. (20) has two real roots:

$$W = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.$$

Since $\lambda > 0$, $W = e^\lambda > 1$, therefore we should choose the root which is greater than 1. Let us check the two roots.

1. Since $A > 0$, $\frac{-B+\sqrt{B^2-4AC}}{2A} > 1$ is equivalent to

$$\sqrt{B^2 - 4AC} > 2A + B. \tag{21}$$

If $2A + B < 0$, Eq. (21) is true automatically. If $2A + B > 0$, Eq. (21) is equivalent to

$$B^2 - 4AC > 4A^2 + 4AB + B^2,$$

which is further equivalent to

$$4A(A + B + C) < 0,$$

which is true because $A > 0$ and $A + B + C < 0$ according to Eqs. (13) and (14).

2. Similarly, $\frac{-B-\sqrt{B^2-4AC}}{2A} > 1$ is equivalent to

$$\sqrt{B^2 - 4AC} < -(2A + B). \tag{22}$$

If $2A + B > 0$, Eq. (22) is false automatically. If $2A + B < 0$, Eq. (22) is equivalent to

$$B^2 - 4AC < 4A^2 + 4AB + B^2,$$

which is further equivalent to

$$4A(A + B + C) > 0,$$

which is false because $A > 0$ and $A + B + C < 0$ according to Eqs. (13) and (14).

Hence, we should take

$$w = e^{\lambda^*} = \frac{-B + \sqrt{B^2 - 4AC}}{2A}$$

as the solution to Eq. (20). According to Eq. (18), we finally have

$$P(\bar{X} > p + t) \le \exp\left(-n\lambda^*(p + t)\right)\left(q + pe^{\lambda^*} - \frac{\sigma^2(e^{\lambda^*} - 1)^2}{2e^{\lambda^*}}\right)^n = w^{-n(p+t)}\left(q + pw - \frac{\sigma^2(w - 1)^2}{2w}\right)^n,$$

which is what to be proved. □

Remark 5. It is straightforward to verify that the original Hoeffding's inequality in Eq. (1) can be derived from Eq. (18) by setting $\sigma^2 = 0$, that is, ignoring the variance of p_i's. In this sense, we see that the original Hoeffding's inequality wastes part of the information contained in p_i's. We also observe from Eq. (18) that larger σ^2 could lead to tighter bound.

Remark 6. Let $p_M = \max(p_1, p_2, \ldots, p_n) \le 1$, there is $1 + p_i(e^\lambda - 1) \le 1 + p_M(e^\lambda - 1) \le e^\lambda$. Hence, Eq. (17) can be improved as

$$\prod_{i=1}^{n}(1 - p_i + p_i e^\lambda) \le \left(q + pe^\lambda - \frac{\sigma^2(e^\lambda - 1)^2}{2(1 + p_M(e^\lambda - 1))}\right)^n. \tag{23}$$

Consequently, Eq. (18) becomes

$$P(\bar{X} > p + t) \le \exp\left(-n\left[\lambda(p + t) - \log\left(q + pe^\lambda - \frac{\sigma^2(e^\lambda - 1)^2}{2(1 + p_M(e^\lambda - 1))}\right)\right]\right). \tag{24}$$

However, to minimize the right hand side of Eq. (24), we need to solve a high order equation, which is not easy. Since Eq. (24) holds for any $\lambda > 0$, we can just choose the λ which minimizes Eq. (18), that is,

$$w = e^{\lambda^*} = \frac{-B + \sqrt{B^2 - 4AC}}{2A}.$$

With this choice, the further improved bound is

$$\begin{aligned} P(\bar{X} > p + t) &\le \exp\left(-n\lambda^*(p + t)\right)\left(q + pe^{\lambda^*} - \frac{\sigma^2(e^{\lambda^*} - 1)^2}{2(1 + p_M(e^{\lambda^*} - 1))}\right)^n \\ &= w^{-n(p+t)}\left(q + pw - \frac{\sigma^2(w - 1)^2}{2(1 + p_M(w - 1))}\right)^n. \end{aligned} \tag{25}$$

Clearly, compared to Eq. (12), Eq. (25) provides a tighter bound because $p_M \le 1$.

3. Numerical comparison

In this section, we compare the performance of the original Hoeffding's inequality given in Eq. (1) to that of the improved version in Eq. (12). In Theorems 1 and 3, we assume that every $X_i \in [0, 1]$ almost surely, thus $p_i = E(X_i) \in [0, 1]$ almost surely.

We consider that p_i's are uniformly distributed in $(0, 1)$. Table 1 displays the values of U_H and U_N for various combinations of n and λ. To get an intuitive idea about the improvement of the proposed inequality to the original inequality, we also list the ratio between these two values, i.e., $R = U_H/U_N$. From Table 1, it is observed that our method provides a modest improvement when λ is small, but for large λ values such that $p + \lambda$ is close to 1, the improvement is significant. Table 1 also demonstrates that for large n, the improvement is more prominent.

Table 1
Comparison of the original Hoeffding's inequality to the proposed inequality when p_i's are uniformly distributed in (0, 1).

n	p	σ^2	λ	U_H	U_N	$R = U_H/U_N$
10	0.3762	0.0485	0.2	0.4412	0.3884	1.1359
			0.4	0.0359	0.0250	1.4357
			0.45	0.0139	0.0090	1.5326
50	0.4887	0.0868	0.2	0.0167	0.0052	3.2055
			0.4	1.3950×10^{-8}	5.1546×10^{-10}	27.0629
			0.45	3.2958×10^{-11}	6.7021×10^{-13}	49.1751
150	0.5259	0.0970	0.2	3.4907×10^{-6}	6.0417×10^{-8}	57.7763
			0.4	6.9625×10^{-26}	7.5374×10^{-31}	9.2373×10^{4}
			0.45	2.3836×10^{-35}	3.2747×10^{-41}	7.2790×10^{5}

4. Conclusions

By applying a refined arithmetic–geometric inequality which incorporates variance information, this paper introduces an improved version to Hoeffding's inequality. Comparing to the original version of Hoeffding's inequality, the proposed improvement incorporates the variance information of the means of the random variables. The proposed version has closed form and it can be evaluated easily. We also make better use of the given information of the mean values to further improve the upper bound. Numerical results show that the improvement is moderate for small λ, but it is significant in far tail area, that is, for $p + \lambda$ close to 1. Furthermore, the improvement becomes more significant as the sample size gets larger.

Acknowledgments

The author would like to extend his sincere gratitude to the anonymous reviewer for the suggestions, which helped to improve the quality of this paper. This work was partially supported by a Summer Faculty Research Fellowship from Missouri State University.

References

Bai, Z.D., 1999. Methodologies in spectral analysis of large dimensional random matrices, a review. Statist. Sinica 9, 611–677.
Bousquet, O., Boucheron, S., Lugosi, G., 2004. Introduction to statistical learning theory. In: Advanced Lectures on Machine Learning. Springer-Verlag, New York, NY, pp. 169–207.
Cartwright, D.I., Field, M.J., 1978. A refinement of the arithmetic mean-geometric mean inequality. Proc. Amer. Math. Soc. 71, 36–38.
Dubhashi, D.P., Panconesi, A., 2012. Concentration of Measure for the Analysis of Randomized Algorithms. Cambridge University Press.
Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58 (301), 13–30.
Mendelson, S., 2003. A few notes on statistical learning theory. In: Advanced Lectures on Machine Learning. Springer-Verlag, New York, NY, pp. 1–40.
Raginsky, M., Sason, I., 2013. Concentration of measure inequalities in information theory, communications, and coding. Found. Trends Commun. Inf. Theory 10 (1–2), 1–246.