Open Journal of Mathematical Sciences

PSR Press

*Article*

# New-type Hoeffding's inequalities and application in tail bounds

**Pingyi Fan**

Beijing National Research Center for Information Science and Technology and the Department of Electronic Engineering, Tsinghua University, Beijing 10084, China; fpy@tsinghua.edu.cn

Academic Editor: Mobeen Munir
Received: 26 March 2021; Accepted: 24 May 2021; Published: 9 June 2021.

**Abstract:** It is well known that Hoeffding's inequality has a lot of applications in the signal and information processing fields. How to improve Hoeffding's inequality and find the refinements of its applications have always attracted much attentions. An improvement of Hoeffding inequality was recently given by Hertz [1]. Eventhough such an improvement is not so big, it still can be used to update many known results with original Hoeffding's inequality, especially for Hoeffding-Azuma inequality for martingales. However, the results in original Hoeffding's inequality and its refined version by Hertz only considered the first order moment of random variables. In this paper, we present a new type of Hoeffding's inequalities, where the high order moments of random variables are taken into account. It can get some considerable improvements in the tail bounds evaluation compared with the known results. It is expected that the developed new type Hoeffding's inequalities could get more interesting applications in some related fields that use Hoeffding's results.

**Keywords:** Hoeffding's lemma; Hoeffding's tail bounds; Azuma inequality; Chernoff's bound.

**MSC:** 60E15; 60A05.

## 1. Introduction

It is well known that Hoeffding's inequality has been applied in many scenarios in the signal and information processing fields. Since Hoeffding's inequality was first found in 1963 [2], it has been attracting much attentions in the academic research, i.e., ([3–9]) and industry.

In [3], it employed the Markov inequality similar to that in the deriving of Chernoff-Hoeffding inequality and considered the tail probability bound of the sums of bounded random variables with limited independence. In [4], it presented a probability bound for a reversible Markov chain where the occupation measure of a set exceeds the stationary probability of a set by a positive quantity. In [5], it discussed the irreducible finite state Markov chains and developed bounds on the distribution function of the empirical mean, especially, employed Gillman's approach to estimate the rate of convergence through bounding the largest eigenvalue of a perturbation of the transition matrix for the Markov chain. In [6], it considered the finite reversible Markov chain and presented some optimal exponential bounds for the probabilities of large deviations about the sums of an arbitrary bounded function of random variables. In [7], it presented Hoeffding-type inequalities for geometrically ergodic Markov chains on general state space, where these bounds depend only on the stationary mean spectral gap and the end-points of support of the bounded function of random variables. In [8], it presented a refined version of the arithmetic geometric mean inequality to improve the Hoeffding's inequality. In [9], it also presented a refinement of Hoeffding's inequality and showed some numerical results to demonstrate its effectiveness.

Especially, in the last decade, it has been used to evaluate the channel code design [10,11] and achievable rate over nonlinear channels [12] as well as delay performance in CSMA with linear virtual channels under a general topology [13] in information theory [14]. As one key tool, it also found the applications in machine learning and big data processing, i.e., PAC-Bayesian method analysis and Markov model analysis in machine learning [15,16], statistical mode bias analysis [17], concept drift in online learning for big data mining [18] and compressed sensing of high dimensional sparse functions [19], etc. It also has been employed in biomedical

fields, i.e., developing the computational molecular modelling tools [20] and analyzing the level set estimation in medical image and pattern recognition [21], etc.

Due to its widely applications, the refined results and improvements of Hoeffding's inequality and Hoeffding-Azuma inequality in martingales usually resulted in more new insights on the developments of related fields. Recently, Hertz [1] presented an improvement result on the original Hoeffding's inequality by utilizing the asymmetric feature of finite distribution interval of random variables. It can reduce the related exponential coefficient from its arithmetic means to the geometric means of $|a|$ and $b$, where $[a,b]$ $(a<0,b>0)$ is the distributed interval of random variable $X$. This improvement motivates us to improve the Hoeffding's inequality. For simplicity, let us first review the result of Hoeffding's inequality [2] and its improvement obtained by Hertz [1].

### 1.1. Hoeffding's Inequality and An Improvement

Assume that $X$ is a zero mean real valued random variable and $X\in[a,b]$ with $a<0,b>0$. Hoeffding's lemma state that for all $s\in\mathbf{R},s>0$,

$$E[e^{sX}]\leq\exp\left\{\frac{s^2(b-a)^2}{8}\right\}. \tag{1}$$

Recently, D. Hertz presented an improved result with the following form

$$E[e^{sX}]\leq\exp\left\{\frac{s^2\Phi^2(a,b)}{2}\right\}, \tag{2}$$

where

$$\Phi(a,b)=\begin{cases}\frac{|a|+b}{2} & b>|\ a\ |,\\ \sqrt{|a|b}, & b\leq|\ a\ |.\end{cases}$$

Since $\sqrt{|a|b}\leq\frac{|a|+b}{2}$, it gives a tighter upper bound for $-a>b$, compared with the original Hoeffding's inequality.

Motivated by this result, an interesting question raises. Can we further improve the Hoeffding's inequality? If so, how to do it.

In this paper, we derive a new type of Hoeffding's inequalities, where higher order moments of random variable $X$ are taken into account, except $E(X)=0$, i.e., $E(X^k)=m_k(k=2,3,...)$.

### 1.2. Main theorem

To give a clear picture of this paper, the new type of Hoeffding's inequality is given as follows;

**Theorem 1.** *Assume that $X$ is a real valued random variable with $E(X)=0$, $X\in[a,b]$ with $a<0,b>0$. For all $s\in\boldsymbol{R},s>0$ and an integer $k$ $(k\geq1)$, we have*

$$E[e^{sX}]\leq\Upsilon_k(a,b)\exp\left\{\frac{s^2}{2k}\Phi^2(a,b)\right\} \tag{3}$$

*where* $\Upsilon_k(a,b)=\left[1+\frac{\max\{|a|,b\}}{|a|}\right]^k-k\frac{\max\{|a|,b\}}{|a|}$ *and* $\Phi(a,b)=\begin{cases}\frac{|a|+b}{2} & b>|\ a\ |,\\ \sqrt{|a|b}, & b\leq|\ a\ |.\end{cases}$

**Remark 1.** When $k=1$, it is easy to check that $\Upsilon_1(a,b)=1$. This indicates that the new type Hoeffding's inequality will be reduced to the improved Hoeffding's inequality (2), still better than the original Hoeffding's inequality. When $k=2$, $\Upsilon_1(a,b)=1+\{\frac{\max\{|a|,b\}}{|a|}\}^2$ and the exponential coefficient has been decreased by 2 times compared to the improved Hoeffding's inequality (2). In fact, such a result can be refined, which is given by the following Corollary.

**Corollary 1.** *Under the same assumption of Theorem 1 for $k=2$, we have*

$$E[e^{sX}]\leq\left[1+\frac{m_2}{a^2}\right]\exp\left\{\frac{s^2}{4}\Phi^2(a,b)\right\}$$

*where $m_2=E(X^2)$.*

*If $E(X^2)$ is unknown, the inequalities can be relaxed as*

$$E[e^{sX}] \leq \left[1+\frac{b}{|a|}\right]\exp\left\{\frac{s^2}{4}\Phi^2(a,b)\right\} \quad if \quad |a|<b, \tag{4}$$

*and*

$$E[e^{sX}] \leq 2\exp\left\{\frac{s^2}{4}\Phi^2(a,b)\right\} \quad if \quad |a|\geq b.$$

Comparing the result in Equation (4) with that presented in Theorem 1, it is easy to check that

$$\left[1+\frac{b}{|a|}\right] \leq 1+\left\{\frac{\max\{|a|,b\}}{|a|}\right\}^2$$

holds. This indicates that Corollary 1 really improves the result presented in Theorem 1 for $k=2$. Comparing to that in Equation (2), the exponential coefficient has be reduced by 2 times. That is to say, when parameter $s$ is relatively large, the new type of Hoeffding's inequalities will give much tighter results than original Hoeffding's inequality and its improvement obtained by Hertz.

The remaining part of this paper is organized as follows: In Section 2, we first present the proof of Corollary 1 and show the insight by taking higher order moments of real valued random variables into account and then present the proof of main theorem in this paper. In Section 3, we present the new type Hoeffding's inequalities applications in the one sided and two sided tail bounds. We also discuss how to select the integer parameter $k$ to give a tighter bound in Section 4. Finally, in Section 5, we give the conclusion.

## 2. The Proof of Main Theoretical Results

Let us first introduce some Lemmas.

### 2.1. Some Useful Lemmas

**Lemma 1.** *Supposed $f(x)$ is a convex function of $x$, $f(x)>0$ with $x\in[a,b]$, then we have the following results:*

1. $$f(x) \leq \frac{b-x}{b-a}f(a)+\frac{x-a}{b-a}f(b),$$

2. *$f^2(x)$ is also a convex function of $x$ and*

$$f^2(x) \leq \left[\frac{b-x}{b-a}f(a)+\frac{x-a}{b-a}f(b)\right]^2$$

*and*

$$f^2(x) \leq \frac{b-x}{b-a}f^2(a)+\frac{x-a}{b-a}f^2(b).$$

The proof of Lemma 1 can be directly derived by using the definition of Convex function and $(f^2(x))' = 2f(x)f'(x)$ and $(f^2(X))'' = 2(f'(x))^2+2f(x)f''(x)>0$.

**Lemma 2.** *Assume that $X$ is a real valued random variable with $E(X)=0$, $P(X\in[a,b])=1$ with $a<0, b>0$, then we have*

1. $$E(X^2) \leq |a|b. \tag{5}$$

2. $$E(X^4) \leq |a|b(a^2+ab+b^2) \leq |a|b(a^2+b^2). \tag{6}$$

**Proof.** .

1. Since $f(x)=x^2$ is a convex function of $x$ in $[a,b]$, we have

$$x^2 \leq \frac{b-x}{b-a}a^2+\frac{x-a}{b-a}b^2$$

$$E(X^2) \leq \frac{b}{b-a}a^2 + \frac{-a}{b-a}b^2 = |a|b.$$

2. Since $f(x) = x^2$ is a convex function and $f(x) \geq 0$. We know that $f^2(x) = x^4$ is also a convex function of $x$ according to Lemma 1. Then we have

$$x^4 \leq \frac{b-x}{b-a}a^4 + \frac{x-a}{b-a}b^4 \tag{7}$$

$$E(X^4) \leq \frac{b}{b-a}a^4 + \frac{-a}{b-a}b^4 = |a|b(a^2+ab+b^2) \leq |a|b(a^2+b^2).$$

□

**Lemma 3.** *For $0 < \lambda < 1$ and $u > 0$, let $\psi(u) = -\lambda u + \ln(1-\lambda+\lambda e^u)$. Then we have*

$$\psi(u) = 0.5\tau(1-\tau)u^2$$

*where $\tau = \frac{\lambda}{(1-\lambda)e^{-\xi}+\lambda}$, $\xi \in [0,u]$. In addition, we have*

$$\psi(u) \leq \begin{cases} \frac{u^2}{8} & \lambda \leq 0.5 \\ \lambda(1-\lambda)\frac{u^2}{2} & \lambda > 0.5 \end{cases}$$

This lemma was derived in [1]. For completeness, we reorganize it as follows:

**Proof.** Since

$$\psi(u) = -\lambda u + \ln(1-\lambda+\lambda e^u).$$

For $u > 0$, one can use Taylor's expansion and obtain

$$\psi(u) = \psi(0) + \psi^{'}(0)u + 0.5\psi^{''}(\xi)u^2.$$

It is easy to check that $\psi(0) = 0$ and

$$\psi^{'}(u) = -\lambda + \frac{\lambda e^u}{1-\lambda+\lambda e^u},$$
$$\psi^{''}(u) = \frac{\lambda e^u}{1-\lambda+\lambda e^u}(1 - \frac{\lambda e^u}{1-\lambda+\lambda e^u}).$$

That means $\psi^{'}(0) = 0$ and $\psi^{''}(\xi) = 0.5\tau(1-\tau)$ where $\tau = \frac{\lambda}{(1-\lambda)e^{-\xi}+\lambda}$, $\xi \in [0,u]$. That is,

$$\psi(u) = 0.5\tau(1-\tau)u^2.$$

Now let us divide it into two cases to discuss:

(i) If $\lambda > 0.5$, then

$$\tau = \frac{\lambda}{(1-\lambda)e^{-\xi}+\lambda} \geq \lambda > 0.5.$$

That means, $\tau(1-\tau)$ reaches its maximum at $\tau = \lambda$. In other word, $\tau(1-\tau) \leq \lambda(1-\lambda)$.

(ii) If $\lambda \leq 0.5$, then we have $\tau(1-\tau) \leq \frac{1}{4}$.

By combining cases (i) and (ii), we get

$$\psi(u) \leq \begin{cases} \frac{u^2}{8} & \lambda \leq 0.5, \\ \lambda(1-\lambda)\frac{u^2}{2} & \lambda > 0.5. \end{cases}$$

The proof is completed.

□

### 2.2. Observation from Corollary 1

Now we first review the Corollary 1. It claimed that under the same assumption of Theorem 1, for $k = 2$, we have

$$E[e^{sX}] \leq \left[1 + \frac{m_2}{a^2}\right] \exp\left\{\frac{s^2}{4}\Phi^2(a,b)\right\}$$

where $m_2 = E(X^2)$.

Before we present the proof of Corollary 1, let us analyze why such a new type of Hoeffding's inequality can decrease its exponential factor by 2 times in philosophy.

Since

$$f(x) = \exp(\alpha x)$$

is a convex function for any $\alpha > 0$.

Let $\alpha = 2\tilde{s}$, then

$$\begin{aligned} E(\exp(2\tilde{s}X)) &\leq \frac{b^2 + m_2}{(b-a)^2}\exp(2\tilde{s}a) + \frac{m_2 + a^2}{(b-a)^2}\exp(2\tilde{s}b) + \frac{-2ab - 2m_2}{(b-a)^2}\exp(\tilde{s}a)\exp(\tilde{s}b) \\ &= \frac{b^2 + m_2}{(b-a)^2}\exp(2\tilde{s}a) + \frac{m_2 + a^2}{(b-a)^2}\exp(2\tilde{s}b) + \frac{-2ab - 2m_2}{(b-a)^2}\exp\left(2\tilde{s}\frac{a+b}{2}\right). \end{aligned}$$

The equation above can be rewritten as

$$E(\exp(\alpha X)) \leq \frac{b^2 + m_2}{(b-a)^2}\exp(\alpha a) + \frac{m_2 + a^2}{(b-a)^2}\exp(\alpha b) + \frac{-2ab - 2m_2}{(b-a)^2}\exp\left(\alpha\frac{a+b}{2}\right). \tag{8}$$

Using Lemma 2 above, it is easy to see that all of the weighting coefficients of $\exp(\alpha a), \exp(\alpha b)$ and $\exp(\alpha\frac{a+b}{2})$ are non-negative and

$$\frac{b^2 + m_2}{(b-a)^2} + \frac{m_2 + a^2}{(b-a)^2} + \frac{-2ab - 2m_2}{(b-a)^2} = 1.$$

Now, by using $s = \alpha$ in the inequality (8), we have

$$E(\exp(sX)) \leq \frac{b^2 + m_2}{(b-a)^2}\exp(sa) + \frac{m_2 + a^2}{(b-a)^2}\exp(sb) + \frac{-2ab - 2m_2}{(b-a)^2}\exp\left(s\frac{a+b}{2}\right) \tag{9}$$

It is easy to see that the right hand side of equation is equal to the linear weighting sum of $\exp(sa), \exp(sb)$ and $\exp(s\frac{a+b}{2})$. That is to say, one can use the information provided by three points to estimate the upper bound of $E(\exp(sX))$. It exactly provides more information than that only using two point linear weighting sum of $\exp(sa)$ and $\exp(sb)$ to estimate the upper bound of $E(\exp(sX))$. Similarly, if one can use the information of function $\exp(sx)$ at multiple points, the upper bound of estimation $E(\exp(sX))$ may be improved further, this is why we consider the high order moments of random variables to discuss Hoeffding's inequality improvement.

### 2.3. Proof of Corollary 1

Now let us present the proof of Corollary 1.

**Proof.** Following the inequality (9), we have that

$$\begin{aligned} E(\exp(sX)) &\leq \frac{b^2 + m_2}{(b-a)^2}\exp(sa) + \frac{m_2 + a^2}{(b-a)^2}\exp(sb) + \frac{-2ab - 2m_2}{(b-a)^2}\exp\left(s\frac{a+b}{2}\right) \\ &= \frac{b^2}{(b-a)^2}\exp(sa) + \frac{a^2}{(b-a)^2}\exp(sb) + \frac{-2ab}{(b-a)^2}\exp\left(s\frac{a+b}{2}\right) \\ &\quad + \frac{m_2}{(b-a)^2}\left\{\exp\left(s\frac{b}{2}\right) - \exp\left(s\frac{a}{2}\right)\right\}. \end{aligned} \tag{10}$$

Let $u = \frac{s(b-a)}{2}$,$\lambda = \frac{-a}{b-a}$, and $\beta^2 = \frac{m_2}{(b-a)^2}$, then we have $s = \frac{2u}{b-a}$, $\frac{b}{b-a} = 1-\lambda$. The inequality (10) can be rewritten as

$$\begin{aligned} E(\exp(sX)) &\leq \left[(1-\lambda)e^{-\lambda u} + \lambda e^{(1-\lambda)u}\right]^2 + \beta^2(e^{(1-\lambda)u} - e^{-\lambda u})^2 \\ &\leq \left[(1-\lambda)e^{-\lambda u} + \lambda e^{(1-\lambda)u}\right]^2 \left(1 + \frac{\beta^2}{\lambda^2}\right) \\ &= \exp(2\psi(u)) \left(1 + \frac{\beta^2}{\lambda^2}\right). \end{aligned}$$

By using Lemma 3, we have

$$E(\exp(sX)) \leq \begin{cases} \exp\left(\frac{u^2}{4}\right)\left(1 + \frac{\beta^2}{\lambda^2}\right) & |a| < b, \\ \exp\left(\lambda(1-\lambda)u^2\right)\left(1 + \frac{\beta^2}{\lambda^2}\right) & |a| \geq b. \end{cases}$$

Now we shall discuss the exponential coefficient and the multiply factor $\left(1 + \frac{\beta^2}{\lambda^2}\right)$ in two different cases:

(a) If $|a| \geq b$, then by using $u = \frac{s(b-a)}{2}$, $\lambda = \frac{-a}{b-a}$, and $\beta^2 = \frac{m_2}{(b-a)^2}$, we have

$$\lambda(1-\lambda)u^2 \leq \frac{-ab}{(b-a)^2}\frac{s^2(b-a)^2}{4} = \frac{s^2|a|b}{4}$$

as well as

$$1 + \frac{\beta^2}{\lambda^2} = 1 + \frac{m_2}{a^2} \leq 1 + \frac{b}{|a|} \leq 2.$$

(b) If $|a| < b$, then we have

$$\frac{u^2}{4} = \frac{1}{4}\frac{s^2(b-a)^2}{4} = \frac{s^2(b-a)^2}{16}$$

and

$$1 + \frac{\beta^2}{\lambda^2} = 1 + \frac{m_2}{a^2} \leq 1 + \frac{b}{|a|}.$$

Combining the two difference cases and using

$$\Phi(a,b) = \begin{cases} \frac{|a|+b}{2} & b > |\ a\ |, \\ \sqrt{|a|b}, & b \leq |\ a\ |. \end{cases}$$

The proof is completed. □

### 2.4. Proof of Theorem 1

**Proof.** If $k = 1$, it is the improved Hoeffding's inequality (2). Now we mainly focus on the case of $k \geq 2$. Since $f(x) = e^{\alpha x}$ is a convex function of $x$ for all $\alpha > 0$ and $f(X) > 0$, we have

$$e^{\alpha x} \leq \frac{b-x}{b-a}e^{\alpha a} + \frac{x-a}{b-a}e^{\alpha b}.$$

For an positive integer $k$ ($k \geq 2$), we have

$$\begin{aligned} e^{k\alpha x} &\leq \left[\frac{b-x}{b-a}e^{\alpha a} + \frac{x-a}{b-a}e^{\alpha b}\right]^k \\ &= \left\{\left[\frac{b}{b-a}e^{\alpha a} + \frac{-a}{b-a}e^{\alpha b}\right] + x\left[\frac{e^{\alpha b} - e^{\alpha a}}{b-a}\right]\right\}^k \end{aligned}$$

and

$$E\left(e^{k\alpha X}\right) \leq E\left\{\left[\frac{b}{b-a}e^{\alpha a} + \frac{-a}{b-a}e^{\alpha b}\right] + X\left[\frac{e^{\alpha b} - e^{\alpha a}}{b-a}\right]\right\}^k.$$

By using $s = k\alpha$ and $\lambda = \frac{-a}{b-a}$, $u = \frac{s}{k}(b-a)$, then we have

$$E\left(e^{sX}\right) \leq E\left\{[(1-\lambda)e^{-\lambda u} + \lambda e^{(1-\lambda)u}] + \frac{X}{|a|}[\lambda e^{(1-\lambda)u} - \lambda e^{-\lambda u}]\right\}^k.$$

Let $e^{\psi(u)} = (1-\lambda)e^{-\lambda u} + \lambda e^{(1-\lambda)u}$ and $\varphi(u) = \lambda e^{(1-\lambda)u} - \lambda e^{-\lambda u}$ then

$$\begin{aligned}
E\left(e^{sX}\right) &\leq E\left[e^{\psi(u)} + \frac{X}{|a|}\varphi(u)\right]^k \\
&= e^{k\psi(u)} + ke^{(k-1)\psi(u)}E\left(\frac{X}{|a|}\right)\varphi(u) + \sum_{i=2}^{k} \mathrm{C}_k^i e^{(k-i)\psi(u)}E\left(\frac{X}{|a|}\right)^i \varphi^i(u) \\
&= e^{k\psi(u)} + \sum_{i=2}^{k} \mathrm{C}_k^i e^{(k-i)\psi(u)}E\left(\frac{X}{|a|}\right)^i \varphi^i(u) \\
&\leq e^{k\psi(u)} + \sum_{i=2}^{k} \mathrm{C}_k^i e^{(k-i)\psi(u)}E\left(\frac{|X|}{|a|}\right)^i \varphi^i(u) \\
&\leq [(1-\lambda)e^{-\lambda u} + \lambda e^{(1-\lambda)u}]^k \left\{\left[1 + \frac{\max\{-a,b\}}{|a|}\right]^k - k\frac{\max\{-a,b\}}{|a|}\right\}
\end{aligned}$$

where $\mathrm{C}_k^i = \frac{k!}{i!(k-i)!}$.

By using $(b-a)\lambda = -a$, and $\psi(u) = 0.5\tau(1-\tau)u^2$, $u = \frac{s}{k}(b-a)$, we have

$$\begin{aligned}
E(e^{sX}) &\leq e^{\frac{k}{2}\tau(1-\tau)u^2}\left\{[1 + \frac{\max\{-a,b\}}{|a|}]^k - k\frac{\max\{-a,b\}}{|a|}\right\} \\
&\leq \left\{[1 + \frac{\max\{|a|,b\}}{|a|}]^k - k\frac{\max\{-a,b\}}{|a|}\right\}\exp\left(\frac{s^2}{2k}\Phi^2\right) \\
&= \mathrm{Y}_k(a,b)\exp\left(\frac{s^2}{2k}\Phi^2\right)
\end{aligned}$$

where $\mathrm{Y}_k(a,b) = \left[1 + \frac{\max\{|a|,b\}}{|a|}\right]^k - k\frac{\max\{|a|,b\}}{|a|}$, and $\Phi = \begin{cases} \frac{(b-a)}{2} & -a < b, \\ \sqrt{|a|b} & -a \geq b. \end{cases}$

The proof is completed. □

**Remark 2.** The proof of Theorem 1 create a new routine on how to use multipoint values of $\exp(sx)$ to get tighter approximation of $E(\exp(sX))$ for any random distribution in a finite interval with $P(X \in [a,b]) = 1$. Comparing with the original Hoeffding's inequality and its improvement obtained by Hertz, the advantages is that it can exactly reduce the exponential coefficients by $k$ times when all the moments of less than $k$ order statistics are taken into account, but the cost is that it will almost enlarge the multiply factor with $C_1^k$ times, as shown by $\mathrm{Y}_k(a,b)$, where $C_1$ is a constant with $C_1 > 1$. That means there exists a trade off between the exponential coefficient reduction and the multiply factor increment. It needs to be considered in specific applications.

In some scenarios, one may interested in the case of $k = 4$. The following Corollary shows one refinement of Theorem 1.

**Corollary 2.** *Assume that $X$ is a real valued random variable, $P(X \in [a,b]) = 1$ with $a < 0, b > 0$ and $E(X) = 0$, $E(X^2) = m_2$, $E(X^3) = 0$ and $E(X^4) = m_4$. For all $s \in \boldsymbol{R}, s > 0$, we have*

$$E\left[e^{sX}\right] \leq \left[1 + \frac{6m_2}{a^2} + \frac{m_4}{a^4}\right]\exp\left(\frac{s^2}{8}\Phi^2(a,b)\right) \tag{11}$$

*where* $\Phi(a,b) = \begin{cases} \frac{(b-a)}{2} & |a| < b, \\ \sqrt{|a|b} & |a| \geq b. \end{cases}$

**Proof.** Since $f(x) = e^{\alpha x}$ is a convex function of $x$ for all $\alpha > 0$ and $f(X) > 0$, we have

$$e^{\alpha x} \leq \frac{b-x}{b-a}e^{\alpha a} + \frac{x-a}{b-a}e^{\alpha b}$$

and

$$\begin{aligned} e^{4\alpha x} &\leq \left[\frac{b-x}{b-a}e^{\alpha a} + \frac{x-a}{b-a}e^{\alpha b}\right]^4 \\ &= \left\{\left[\frac{b}{b-a}e^{\alpha a} + \frac{-a}{b-a}e^{\alpha b}\right] + x\left[\frac{e^{\alpha b} - e^{\alpha a}}{b-a}\right]\right\}^4. \end{aligned}$$

Let $s = 4\alpha$, and using $E(X) = 0, E(X^2) = m_2$ $E(X^3) = 0$ and $E(X^4) = m_4$, we have

$$E(e^{sX}) \leq \left(\frac{b}{b-a}e^{\frac{s}{4}a} + \frac{-a}{b-a}e^{\frac{s}{4}b}\right)^4 + 6m_2\left(\frac{b}{b-a}e^{\frac{s}{4}a} + \frac{-a}{b-a}e^{\frac{s}{4}b}\right)^2\left(\frac{e^{\frac{s}{4}b} - e^{\frac{s}{4}a}}{b-a}\right)^2 + m_4\left(\frac{e^{\frac{s}{4}b} - e^{\frac{s}{4}a}}{b-a}\right)^4.$$

Let $\lambda = \frac{-a}{b}$, $u = \frac{s}{4}(b-a)$, then we have $\frac{b}{b-a} = 1-\lambda$, $\frac{s}{4}a = -\lambda u$, $\frac{s}{4}b = (1-\lambda)u$. Then the inequality above can be rewritten as

$$\begin{aligned} E\left(e^{sX}\right) &\leq \left[1 - \lambda e^{-\lambda u} + \lambda e^{(1-\lambda)u}\right]^4 + \frac{6m_2}{(b-a)^2\lambda^2}\left[1 - \lambda e^{-\lambda u} + \lambda e^{(1-\lambda)u}\right]^2\left[\lambda e^{(1-\lambda)u} - \lambda e^{-\lambda u}\right]^2 \\ &\quad + \frac{m_4}{(b-a)^4\lambda^4}\left[\lambda e^{(1-\lambda)u} - \lambda e^{-\lambda u}\right]^4 \\ &\leq \left[1 - \lambda e^{-\lambda u} + \lambda e^{(1-\lambda)u}\right]^4\left[1 + \frac{6m_2}{(b-a)^2\lambda^2} + \frac{m_4}{(b-a)^4\lambda^4}\right] \\ &= e^{4\psi(u)}\left[1 + \frac{6m_2}{(b-a)^2\lambda^2} + \frac{m_4}{(b-a)^4\lambda^4}\right]. \end{aligned}$$

By using $(b-a)\lambda = -a$, and Lemma 3, we have

$$E\left(e^{sX}\right) \leq \left[1 + \frac{6m_2}{a^2} + \frac{m_4}{a^4}\right]e^{\frac{s^2\Phi^2}{8}}$$

where $\Phi = \begin{cases} \frac{(b-a)}{2} & -a < b, \\ \sqrt{|a|b} & -a \geq b. \end{cases}$
The proof is completed. □

If the $E(X^2)$ and $E(X^4)$ are not exactly known and $|a| = b$, we have the following result:

**Corollary 3.** *Assume that $X$ is a real valued random variable, $P(X \in [-a, a]) = 1$ with $a > 0$ and $E(X) = 0$ and $E(X^3) = 0$. For all $s \in \boldsymbol{R}, s > 0$, we have*

$$E\left[e^{sX}\right] \leq 8\exp\left(\frac{a^2s^2}{8}\right).$$

**Proof.** By using $m_2 \leq a^2$, $m_4 \leq a^4$ and the inequality in Corollary 2, we can get the result directly. □

## 3. Applications in Tail Bound Evaluation

Let us consider the scenario, where $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in [a_i, b_i], a_i < 0, b_i > 0$ and $EX_i = 0$ for $i = 1, 2, \ldots, n$.

Define $S_n = \sum_{i=1}^n x_i$.

It is easy to check that $ES_n = 0$. For all $s > 0$, we have

$$\begin{aligned} P(S_n \geq t) &= P\left(e^{sS_n} \geq e^{st}\right) \quad \text{Chernoff} \\ &\leq e^{-st} E e^{sS_n} \quad \text{Markov} \\ &= e^{-st} \prod_{i=1}^{n} E e^{sX_i}. \end{aligned} \tag{12}$$

Using the results of Theorem 1 and its Corollaries, one can obtain that

$$E e^{sX_i} \leq A_{k_i} \exp\left(\frac{s^2}{2k_i}\Phi_i\right)$$

where $A_{k_i}$ and $k_i$ are based on which one inequality of $X_i$ being selected in Theorem 1 with

$$A_{k_i} = \begin{cases} 1 & k_i = 1, \\ 1 + \frac{\max\{|a|,b\}}{|a|} & k_i = 2, \\ \left[1 + \frac{\max\{|a|,b\}}{|a|}\right]^{k_i} - k_i \frac{\max\{|a|,b\}}{|a|} & k_i \geq 3, \end{cases}$$

or others presented in Corollaries, and $\Phi_i = \Phi(a_i, b_i)$.

In this case, we get

$$P\left(S_n \geq t\right) \leq \left(\prod_{i=1}^{n} A_{k_i}\right) \exp\left\{-st + s^2\left(\sum_{i=1}^{n} \frac{\Phi_i^2}{2k_i}\right)\right\}.$$

Now selecting $s = \frac{t}{2\left(\sum_{i=1}^{n} \frac{\Phi_i^2}{2k_i}\right)}$ to minimize the exponent in inequality (12), we obtain

$$P\left(S_n \geq t\right) \leq \left(\prod_{i=1}^{n} A_{k_i}\right) \exp\left\{-t^2\left(2\sum_{i=1}^{n} \frac{\Phi_i^2}{k_i}\right)^{-1}\right\}. \tag{13}$$

In particular, if all the $k_i$, $(i = 1, 2, \cdots, n)$ are selected as 1, then $A_{k_i} = 1$, it reduces to the improved Hoeffding's one side tail bound.

If all the $k_i$, $(i = 1, 2, \cdots, n)$ are selected as 2 and $|a_i| = b_i$, then $A_{k_i} = 2$, and the inequality can be rewritten as

$$P(S_n \geq t) \leq 2^n \exp\left\{-\frac{t^2}{\sum_{i=1}^{n} a_i^2}\right\}.$$

Furthermore,

$$P\left(\frac{S_n}{n} \geq l\right) \leq \left(\prod_{i=1}^{n} A_{k_i}\right) \exp\left\{\frac{-nl^2}{2\tilde{\Phi}_i^2}\right\}$$

where $l$ is a positive number and $\tilde{\Phi}_i^2 = \frac{1}{n}\left(\sum_{i=1}^{n} \frac{\Phi_i^2}{2k_i}\right)$.

The two sided tail bound can be given by

$$P\left(|S_n| \geq t\right) \leq \left(\prod_{i=1}^{n} A_{k_i}\right) \exp\left\{-t^2\left(2\sum_{i=1}^{n} \frac{\Phi_i^2}{k_i}\right)^{-1}\right\} + \left(\prod_{j=1}^{n} B_{k_j}\right) \exp\left\{-t^2\left(2\sum_{j=1}^{n} \frac{\Phi_j^2}{k_j}\right)^{-1}\right\}$$

where $\{B_{k_j}, j = 1, 2, \cdots, n\}$ is a sort of $\{A_{k_i}, i = 1, 2, \cdots, n\}$ complement. That is to say, the calculation of $B_{k_j}$ is just changing the positions of $a_j$ and $b_j$ in such a way $-a_j \to b_i$ and $-b_j \to a_i$ in the calculation of $A_{k_i}$ if the integer index $k_j$ of $B_{k_j}$ is equal to the integer index $k_i$ of $A_{k_i}$. In other word, for the same $X_i$, it may select two different integer parameter values of $k_i$ to estimate both sided tail bounds for the positive and the negative directions.

## 4. Selection of Integer Parameter $k$

On the selection of integer parameter $k_i$, we shall discuss it firstly from one sided tail bound. For simplicity, let us consider $n = 1$. The first question is when selecting a larger $k$ will get a tighter bound. The question can be solved by

$$A_{k+1} \exp\left\{-t^2\left(2\frac{\Phi^2}{k+1}\right)^{-1}\right\} < A_k \exp\left\{-t^2\left(2\frac{\Phi^2}{k}\right)^{-1}\right\}. \tag{14}$$

Using logarithm on both sides of inequality (14) and after some manipulations, we get

$$\frac{t^2}{2\Phi^2} > \ln A_{k+1} - \ln A_k.$$

That is

$$t > \Phi\sqrt{2\ln\frac{A_{k+1}}{A_k}}.$$

To clear illustrate the effect of $k$ selection, we give following three examples:

**Example 1.** For $a = -1$ and $b = 1$, the selection rule of $k$ ($k = 1, 2, 3$) is given by

$$k = \begin{cases} 1, & 0 < t < \sqrt{2\ln 2} \approx 1.177, \\ 2, & \sqrt{2\ln 2} < t < \sqrt{2\ln(2.5)} \approx 1.3537, \\ 3, & t > \sqrt{2\ln(2.5)}. \end{cases}$$

**Example 2.** For $a = -1$ and $b = 5$, the selection rule of $k$ ($k = 1, 2, 3$) is given by

$$k = \begin{cases} 1, & 0 < t < 3\sqrt{2\ln 6} \approx 5.679, \\ 2, & 3\sqrt{2\ln 6} < t < 3\sqrt{2\ln(191/6)} \approx 7.892, \\ 3,, & t > 3\sqrt{2\ln(191/6)}. \end{cases}$$

**Example 3.** For $a = -5$ and $b = 1$, the selection rule of $k$ ($k = 1, 2, 3$) is given by

$$k = \begin{cases} 1, & 0 < t < \frac{1}{2}\sqrt{10\ln(6/5)} \approx 0.6751, \\ 2, & \frac{1}{2}\sqrt{10\ln(6/5)} < t < \sqrt{10\ln(25/6)} \approx 3.778, \\ 3, & t > \sqrt{10\ln(25/6)}. \end{cases}$$

**Remark 3.** All the three examples show that when $t$ is relatively small, i.e., close to zero, selecting parameter $k = 1$ is the best one. The results in Example 3 show that when $t = 0.8$, selecting $k = 2$ will give a tighter tail bound. The results in Example 2 and Example 3 also indicates when random variable $X$ with $P(X \in [-1, 5]) = 1$, where $a = -1, b = 5$, one need to estimate $P(|X| > 0.8)$, the right hand sided bound should select $k = 1$ as its estimation while the left hand sided bound should select $k = 2$ as its estimation. This result shows that one may not consistently select the same parameter $k$ to deal with both sided tail bounds when $|a| \neq b$.

Now let consider the general case.

The goal of parameters $k_i$ selection is basically to minimize the right hand of inequality (13). Thus, one can set up an optimization problem as follows:

**Problem 1.** For a given $t > 0$,

$$\min_{k_i}\left(\prod_{i=1}^{n} A_{k_i}\right)\exp\left\{-t^2\left(2\sum_{i=1}^{n}\frac{\Phi_i^2}{k_i}\right)^{-1}\right\}$$

where $A_{k_i}$ are calculated by using the theoretical results in Theorem 1 and its Corollaries for a given $k_i$. It is equivalent to

$$\min_{k_i} \left( \sum_{i=1}^{n} \ln(A_{k_i}) \right) - t^2 \left( 2 \sum_{i=1}^{n} \frac{\Phi_i^2}{k_i} \right)^{-1}$$

and

$$\max_{k_i} \frac{1}{\left( 2 \sum_{i=1}^{n} \frac{\Phi_i^2}{k_i} \right)} - \frac{\left( \sum_{i=1}^{n} \ln(A_{k_i}) \right)}{t^2}.$$

In fact, such an optimization problem can be solved by using computer search. Here, in order to provide a tractable mode, we relax $A_{k_i}$ with the form $\left[ 1 + \frac{\max\{|a|,b\}}{|a|} \right]^k$ given in Theorem 1. In this case, the optimization problem can be transformed into the following problem.

**Problem 2.** Take

$$\max_{k_i} \frac{1}{\left( 2 \sum_{i=1}^{n} \frac{\Phi_i^2}{k_i} \right)} - \frac{\left( \sum_{i=1}^{n} k_i \ln \left( 1 + \frac{max\{|a_i|,b_i\}}{|a_i|} \right) \right)}{t^2}.$$

Let us define

$$g\left(k_1, k_2, \ldots, k_n\right) = \frac{1}{\left( 2 \sum_{i=1}^{n} \frac{\Phi_i^2}{k_i} \right)} - \frac{\left( \sum_{i=1}^{n} k_i \ln \left( 1 + \frac{max\{|a_i|,b_i\}}{|a_i|} \right) \right)}{t^2}.$$

In order to get some insights, let us consider $k_j$ to be a real number rather than an integer, then the partial derivative of function $g(.)$ to $k_j$ is given by

$$\frac{\partial g}{\partial k_j} = \frac{2\Phi_i^2 k_j^{-2}}{\left( 2 \sum_{i=1}^{n} \frac{\Phi_i^2}{k_i} \right)^2} - \frac{\ln \left( 1 + \frac{max\{|a_j|,b_j\}}{|a_j|} \right)}{t^2}.$$

Let $\frac{\partial g}{\partial k_j} = 0$, after some manipulations, we obtain

$$k_j = \frac{\Phi_j}{\sqrt{2 \ln \left( 1 + \frac{max\{|a_j|,b_j\}}{|a_j|} \right)}} \frac{t}{\sum_{i=1}^{n} \frac{\Phi_i^2}{k_i}} = \frac{\Phi(a_j, b_j)}{\sqrt{2 \ln \left( 1 + \frac{max\{|a_j|,b_j\}}{|a_j|} \right)}} \frac{t}{\sum_{i=1}^{n} \frac{\Phi_i^2}{k_i}}.$$

Since $\frac{t}{\sum_{i=1}^{n} \frac{\Phi_i^2}{k_i}}$ is a common factor for all the $k_j$,$(j = 1, 2, \ldots, n)$. This means

$$k_j \propto \frac{\Phi(a_j, b_j)}{\sqrt{2 \ln \left( 1 + \frac{max\{|a_j|,b_j\}}{|a_j|} \right)}}.$$

That is to say, the near optimal value of $k_j$ is mainly determined by $a_j$ and $b_j$ except a common factor, the parameters of distribution interval of $X_j$. This is an interesting result, which can provide more insight. In most of applications, all the $X_i$ $(i = 1, 2, \ldots, n)$ are distributed with the same interval. In this case, one can select the same $k_i$ value for all of them, so that it can approximate the near optimal tighter tail bound. Such a discussion can be extended to the scenarios of two sided tail bound.

**Remark 4.** Consider the distribution interval is symmetric, where $|a_i| = b_i$. In this case, we have

$$\frac{\Phi(a_j, b_j)}{\sqrt{2 \ln \left( 1 + \frac{max\{|a_j|,b_j\}}{|a_j|} \right)}} = \frac{|a_j|}{\sqrt{2 \ln 2}}.$$

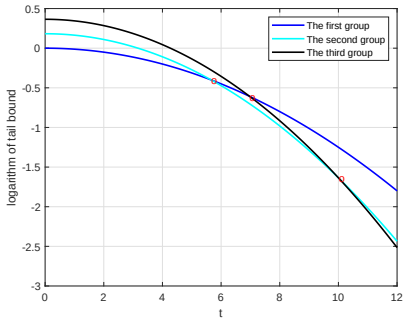


**Figure 1.** The logarithm of the one sided tail bound of three different group selection of parameters $k$ with $n = 4$ in Example 5.

This means

$$k_j \propto |a_j|$$

It indicates the integer parameter $k_j$ selection is proportional to the distribution interval length. When $|a_j|$ is relatively small, i.e., $|a_j|$ is close to zero, the linear interpolation of two points with $x_1 = a_j$ and $x_2 = |a_j|$ is good enough to approximate the random curve of $e^{sX}$. That is to say, select $k_j = 1$ is good enough.

When $|a_j|$ is relatively large, the linear interpolation of two points with $x_1 = a_j$ and $x_2 = |a_j|$ may not be good enough to approximate the curve of $e^{sX}$. It needs more points in the curve of $e^{sx}$ to do the interpolation so that it could have a good approximation to the random curve of $e^{sX}$. That is to say, selecting a larger $k_j$ is necessary. Such an observation is consistent with our "intuitive feeling" on the function approximation in philosophy. We shall illustrate such phenomenon in detail with some examples below.

**Example 4.** Let $a = -5$, $b = 5$ and $m_2 = 5$. The selection rule of $k$ ($k = 1, 2, 3$) is given by

$$k = \begin{cases} 1, & 0 < t < 5\sqrt{2\ln 2(6/5)} \approx 3.019, \\ 2, & 5\sqrt{2\ln 2(6/5)} < t < 5\sqrt{2\ln(25/6)} \approx 8.447, \\ 3, & t > 5\sqrt{2\ln(25/6)}. \end{cases}$$

**Remark 5.** The results in Example 1 show, selecting $k = 1$ is always the best since the best working region for $t$ of $k \geq 2$ is out of the $X$ distributed interval, which can not occur in practice. Example 4 shows that when $m_2$ is given, it is possible to select $k \geq 2$ to get a tighter tail bound, i.e., $t = 4$, the best selection of $k$ is $k = 2$, which also show that when the distribution interval is relatively larger, it is possible to select the larger integer value of $k$ for the tail bound estimation.

**Example 5.** Let us consider $n = 4$, where $X_1 \in [-1, 1]$, $X_2 \in [-5, 5]$, $X_3 \in [-1, 5]$ and $X_4 \in [-5, 1]$ with $E(X_1) = E(X_2) = E(X_3) = E(X_4) = 0$, $E(X_2^2) = 5$ and $S_4 = X_1 + X_2 + X_3 + X_4$. It is easy to check that $S_4 \in [-12, 12]$.

Figure 1 shows different curves of one sided tail bounds, in which Group one: $k_1 = k_2 = k_3 = k_4 = 1$. Group two: $k_1 = k_3 = k_4 = 1, k_2 = 2$ and Group three: $k_1 = k_3 = 1, k_2 = k_4 = 2$, where the y-label is the logarithm of the one sided tail bound, $\left(\sum_{i=1}^{n} \ln(A_{k_i})\right) - t^2 \left(2\sum_{i=1}^{n} \frac{\Phi_i^2}{k_i}\right)^{-1}$, the x-label is $t$. It is observed that among the three groups of parameter $k$ selection, when $0 < t < 5.6647$, the curve of Group one provides the tightest bound. When $5.6647 < t < 10.0138$, the curve of Group two provides the tightest bound and when $10.0138 < t < 12$, the curve of Group three provides the tightest bound.

The results in Example 5 exactly demonstrate that the new type Hoeffding's inequalities are useful in the tail bound estimation.

**Remark 6.** In real applications, one would not like to pay more attention on the selection of parameter $k_i$ in order to make the system analysis simplified. It recommends to select $k_i$ to be 1 or 2.

## 5. Conclusion

In this paper, we presented new type of Hoeffding's inequalities by using higher order moments of random variables. Some applications in one and two sided tail bound improvements can also be obtained by using the exponential function positiveness and Chernoff inequality. Perhaps, future research may focus on trying to improve the related inequalities that use Hoeffding's Lemma.

**Conflicts of Interest:** The author declares no conflict of interest.

**Data Availability:** No data is required for this research.

**Funding Information:** This work was partially supported by Beijing Natural Science Foundation under Grant 4202030.

**Acknowledgments:** The author is thankful to the two anonymous referees for their constructive comments.

### References

[1] Hertz, D. (2020). Improved Hoeffding's lemma and Hoeffding's tail bounds. *arXiv preprint* arXiv:2012.03535.

[2] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association, 58*, 13–30.

[3] Schmidt, J. P., Siegel, A., & Srinivasan, A. (1995). Chernoff-Hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics, 8*(2), 223–250.

[4] Dinwoodie, I. H. (1995). A probability inequality for the occupation measure of a reversible Markov chain. *The Annals of Applied Probability, 5*(1), 37–43.

[5] Lezaud, P. (1998). Chernoff-type bound for finite Markov chains. *Annals of Applied Probability, 8*(3), 849–867.

[6] León, C. A., & Perron, F. (2004). Optimal Hoeffding bounds for discrete reversible Markov chains. *The Annals of Applied Probability, 14*(2), 958–970.

[7] Miasojedow, B. (2014). Hoeffding's inequalities for geometrically ergodic Markov chains on general state space. *Statistics and Probability Letters, 87*, 115–120.

[8] Zheng, S. (2017). A refined Hoeffding's upper tail probability bound for sum of independent random variables. *Statistics and Probability Letters, 131*, 87–92.

[9] From, S. G., & Swift, A. W. (2013). A refinement of Hoeffding's inequality. *Journal of Statistical Computation and Simulation, 83*(5), 977-983.

[10] Scarlett, J., Martinez, A., & i Fàbregas, A. G. (2014). Second-order rate region of constant-composition codes for the multiple-access channel. *IEEE Transactions on Information Theory, 61*(1), 157–172.

[11] Sason, I., & Eshel, R. (2011, July). On concentration of measures for LDPC code ensembles. *In 2011 IEEE International Symposium on Information Theory Proceedings*, (pp. 1268–1272). IEEE.

[12] Xenoulis, K., Kalouptsidis, N., & Sason, I. (2012, July). New achievable rates for nonlinear Volterra channels via martingale inequalities. *In 2012 IEEE International Symposium on Information Theory Proceedings*, (pp. 1425-1429), IEEE.

[13] Yun, D., Lee, D., Yun, S. Y., Shin, J., & Yi, Y. (2015). Delay optimal CSMA with linear virtual channels under a general topology. *IEEE/ACM Transactions on Networking, 24*(5), 2847–2857.

[14] Raginsky, M., & Sason, I. (2018). *Concentration of Measure Inequalities in Information Theory, Communications, and Coding:* Third Edition. Now Foundations and Trends.

[15] Seldin Y., Laviolette F., Cesa-Bianchi N., Shawe-Taylor J., & Auer P. (2012) Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory, 58*(12), 7086–7093.

[16] Fan J., Jiang B., & Sun Q. (2018) Hoeffding's lemma for markov chains and its applications to statistical learning. *arXiv preprint* arXiv:1802.00211.

[17] Gourgoulias, K., Katsoulakis, M. A., Rey-Bellet, L., & Wang, J. (2020). How biased is your model? Concentration inequalities, information and model bias. *IEEE Transactions on Information Theory, 66*(5), 3079–3097.

[18] Frias-Blanco, I., del Campo-Ăvila, J., Ramos-Jimenez, G., Morales-Bueno, R., Ortiz-Diaz, A., & Caballero-Mota, Y. (2014). Online and non-parametric drift detection methods based on Hoeffding's bounds. *IEEE Transactions on Knowledge and Data Engineering, 27*(3), 810–823.

[19] Schnass, K., & Vybiral, J. (2011, May). Compressed learning of high-dimensional sparse functions. *In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 3924–3927). IEEE.

[20] Rasheed M., Clement N., Bhowmick A., & Bajaj C. L. (2019). Statistical framework for uncertainty quantification in computational molecular modeling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 16*(4), 1154–1167.

[21] Willett R. M., & Nowak R. D. (2007). Minimax optimal level-set estimation. *IEEE Transactions on Image Processing, 16*(12), 2965–2979.