Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

Optimal distribution-free concentration for the log-likelihood function of Bernoulli variables

Zhonggui Ren[1*]

*Correspondence:
renly0125@163.com
[1]College of Foundation Science, Harbin University of Commerce, Harbin, China

Abstract

This paper aims to establish distribution-free concentration inequalities for the log-likelihood function of Bernoulli variables, which means that the tail bounds are independent of the parameters. Moreover, Bernstein's and Bennett's inequalities with optimal constants are obtained. The simulation study shows significant improvements over the previous results.

Keywords: Bernstein's inequality; Bennett's inequality; Log-likelihood function; Bernoulli distribution

1 Introduction

Concentration inequality has been applied in a variety of scenarios, including statistical inference, information theory, machine learning, etc., see [1, 6, 8, 9]. Let $X_1, X_2, \dots, X_n$ be independent Bernoulli random variables with parameters p_i, respectively. For simplicity, denote $X_i \sim \mathrm{Ber}(p_i)$. Inspired by theoretical studies of likelihood-based methods for binary data, in particular for community detection in networks (see [3, 5, 12] for details), it is of significance to investigate the concentration behavior of the joint likelihood function of $X_1, X_2, \dots, X_n$, say, $L_n = \prod_{i=1}^n p_i^{X_i}(1-p_i)^{1-X_i}$.

Consider a simple case that $p_1 = p_2 = \cdots = p_n = p \in [0,1]$. The asymptotic equipartition property (AEP), one of the most classical results in information theory [7], asserts that

$$n^{-1}\log L_n = \frac{1}{n}\sum_{i=1}^{n}\big(X_i \log p + (1-X_i)\log(1-p)\big) \xrightarrow{\mathbb{P}} p\log p + (1-p)\log(1-p),$$

which can be obtained by the law of large numbers. Indeed, this relation implies that the sample averaged Shannon entropy (scaled log-likelihood function $n^{-1}\log L_n$) converges to the population Shannon entropy in probability. To get a clearer perspective of the AEP, this paper aims to derive a nonasymptotic concentration inequality for the tail probability. Zhao [10] demonstrated a novel Bernstein-type inequality for $\sum_{i=1}^n X_i \log p_i$ (a part of

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Springer

$\log L_n$), which asserts that, for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i \log p_i - \sum_{i=1}^{n} p_i \log p_i\right| \geq n\epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2(1+\epsilon)}\right) \tag{1}$$

by providing an upper bound, independent of the parameter p_i, for the moment generating function (MGF) of $(X_i - p_i)\log p_i$.

Invoke the definition of sub-gamma variable (see [1, 8] for more details), which states that a random variable X is sub-gamma with the variance factor $\nu > 0$ and the scale parameter $b > 0$ (denoted as $\text{sub}\,\Gamma(\nu, b)$) if its MGF satisfies, for all $|\lambda| < b^{-1}$,

$$\mathbb{E}\left(e^{\lambda(X-\mathbb{E}X)}\right) \leq \exp\left(\frac{\nu\lambda^2}{2(1-b|\lambda|)}\right).$$

Indeed, Theorem 2 of [10] shows that the random variable $X_i \log p_i$ is $\text{sub}\,\Gamma(1,1)$. As pointed out by Remark 2 of [10], the scale factor $b = 1$ on the denominator is optimal, while the variance factor $\nu = 1$ is not sharp. A natural question is whether we can improve the variance factor as well as the Bernstein-type inequality (1). Moreover, moving back to the joint log-likelihood function, can we derive a sharp Bernstein-type inequality for $\log L_n$?

In this paper, we are interested in studying optimal distribution-free (independent of parameters p_i) concentration bounds for $\sum_{i=1}^{n}(X_i - p_i)\log p_i$ and $\log L_n = \prod_{i=1}^{n}(X_i \log p_i + (1-X_i)\log(1-p_i))$ with independent $X_i \sim \text{Ber}(p_i)$ for $p_i \in [0,1]$. Those results, which cannot be derived by classical Hoeffding's inequality (see [4, 10]), will be particularly useful in cases where the assumptions for p_i are not convenient to be made. In addition, the improvements by optimal uniform constants are nonnegligible in the sense of nonasymptotic, especially in the case of small sample datasets.

The rest of this paper is organized as follows. Section 2 establishes the Bernstein-type inequalities for $\sum_{i=1}^{n}(X_i - p_i)\log p_i$ and the log-likelihood function of the Bernoulli variable, which both enjoy the optimal variance factors and scale factors. Inspired by Bennett's inequality for the bounded random variable, we use Bennett's inequality to improve the tail bounds on the right tail in Sect. 3. Some extensions are illustrated in Sect. 4. The improvements by the optimal constants are demonstrated by various simulation studies in Sect. 5. Finally, Sect. 6 concludes the article with a discussion.

Notation: Throughout this paper, set $0\log 0 = 0$ for convention. All logarithms and exponentials are in the natural base.

2 Bernstein's inequality

Based on the classical Chernoff method (see [1, 8, 10]), this section illustrates the optimal Bernstein-type inequalities for $\sum_{i=1}^{n}(X_i - p_i)\log p_i$ and $\log L_n$.

Theorem 1 *Let X_i be independent* $\text{Ber}(p_i)$ *for $i = 1, \ldots, n$, where $p_i \in [0,1]$. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - p_i)\log p_i\right| \geq n\epsilon\right) \leq 2\exp\left(-n\gamma h\left(\frac{\epsilon}{\gamma}\right)\right), \tag{2}$$

where $h(u) = 1 + u - \sqrt{1+2u}$ for $u > 0$ and $\gamma := \max_{0\leq x\leq 1} x(1-x)(\log x)^2$.

Proof Denote $Y_i := (X_i - p_i)\log p_i$, where $X_i \sim \mathrm{Ber}(p_i)$. It follows that $\mathbb{E}Y_i = 0$ and $\mathrm{Var}(Y_i) = p_i(1-p_i)(\log p_i)^2$. Note $\gamma = \max_{0\le x\le 1} x(1-x)(\log x)^2$, which is approximately 0.477365 at $x_0 \approx 0.104058$. Hence, $\mathrm{Var}(Y_i) \le \gamma$ with the equality if and only if $p_i = x_0$. Now, we claim that the moments of Y_i satisfy the Bernstein condition (i.e. see Theorem 1 of [10] or Theorem 2.10 of [1]) as follows:

$$\left|\mathbb{E}(Y_i)^k\right| \le \frac{1}{2}k!\gamma \quad \text{for all } k \ge 2. \tag{3}$$

By the power series expansion of the exponential function and Fubini's theorem, it follows that, for all $|\lambda| < 1$,

$$\begin{aligned}\mathbb{E}(\exp\lambda Y_i) &= 1 + \frac{\lambda^2}{2}\mathrm{Var}(Y_i) + \sum_{k=3}^{\infty}\lambda^k\frac{\mathbb{E}(Y_i)^k}{k!}\\ &\le 1 + \frac{\lambda^2\gamma}{2} + \sum_{k=3}^{\infty}\frac{|\lambda|^k\gamma}{2}\\ &= 1 + \frac{\gamma\lambda^2}{2}\frac{1}{1-|\lambda|}\\ &\le \exp\left(\frac{\gamma\lambda^2}{2(1-|\lambda|)}\right),\end{aligned}$$

where γ is defined as above, and the last inequality uses $1 + x \le \exp(x)$ for $x \in \mathbb{R}$. Therefore, the random variable Y_i is sub-gamma $(\gamma, 1)$, which is independent of p_i. To this end, applying the standard Chernoff method (cf. Theorem 2.10 of [1]) gives, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_i\right| \ge \epsilon\right) \le 2\exp\left(-n\gamma h\left(\frac{\epsilon}{\gamma}\right)\right),$$

where $h(u) = 1 + u - \sqrt{1+2u}$ for $u > 0$ and γ is defined as above.

It remains to show claim (3). The case of $k = 2$ follows trivially. Similar to Theorem 1 of [10], for $k \ge 3$,

$$\begin{aligned}\left|\mathbb{E}(Y_i)^k\right| &= \left|p_i(1-p_i)^k(\log p_i)^k + (1-p_i)(-p_i\log p_i)^k\right|\\ &\le \left|p_i(\log p_i)^k\right| + \left|(-p_i\log p_i)^k\right|\\ &\le \left(\frac{k}{e}\right)^k + \exp(-k),\end{aligned}$$

where the first inequality follows from $p_i \in [0,1]$ and the last inequality is implied by the fact that the function $f(x) := x^r(\log x)^k$ for $x \in (0,1)$ achieves its optimum at $x = e^{-k/r}$ for any $r > 0$ and integer $k \ge 1$. Now we prove the claim by induction. The case of $k = 3$ follows from simple computations, that is, $28\exp(-3) \approx 1.394038 < 3\gamma$. Suppose for some $k \ge 3$,

$(k^k + 1)\exp(-k) \le k!\gamma/2$. For the case of $k+1$, it follows that

$$\begin{aligned}\frac{(k+1)^{k+1}+1}{\exp(k+1)} &\le \frac{k!\gamma}{2}\frac{(k+1)^{k+1}+1}{e(k^k+1)}\\ &< \frac{k!\gamma}{2e}\frac{(k+1)^{k+1}}{k^k}\\ &= \frac{k!\gamma}{2e}(k+1)\left(1+\frac{1}{k}\right)^k\\ &\le \frac{(k+1)!\gamma}{2},\end{aligned}$$

which completes the induction and finishes the proof of this proposition. □

Remark 1 The constant γ is optimal because Y_i, being sub-gamma (v, b), satisfies $\mathrm{Var}(Y_i) \le v$ by the property of sub-gamma variables. Indeed, one can choose $p_i = x_0$, which achieves $\mathrm{Var}(Y_i) = \gamma$, thus $\gamma \le v$, implying its optimality. As pointed out by Remarks 1 and 2 of [10], the constant $b = 1$ is optimal since $\mathbb{E}(Y_i)^k$ can achieve $c(k/e)^k$ for some constant $0 < c < 1$ at $p_i = \exp(-k)$ and the Stirling approximation

$$c(k/e)^k > c\frac{k!}{2\sqrt{2\pi k}} > \frac{k!\gamma}{2}b^{k-2}$$

for large enough k and any $0 < b < 1$. Hence, both the variance factor and the scale factor in Theorem 1 are optimal.

Remark 2 To give a nice form of (2), by the elementary inequality (Exercise 2.8 of [1])

$$h(u) \ge \frac{u^2}{2(1+u)} \quad \text{for } u > 0,$$

we have, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - p_i)\log p_i\right| \ge n\epsilon\right) \le 2\exp\left(-\frac{n\epsilon^2}{2(\gamma+\epsilon)}\right), \tag{4}$$

where γ is defined as Theorem 1. Compared to equation (3) of [10], we improved the constant $\sigma^2 = 1$ to $\gamma \approx 0.477$, which is optimal. Furthermore, one can generalize 1 to the multinoulli variables and the grouped observation, analogously to Corollary 1 and Theorem 2 of [10].

Analogously, Theorem 2 provides the optimal Bernstein inequality of $\log L_n$.

Theorem 2 *Let X_i be independent* $\mathrm{Ber}(p_i)$ *for $i = 1, \dots, n$, where $p_i \in [0, 1]$. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\big(\big|\log L_n - \mathbb{E}(\log L_n)\big| \ge n\epsilon\big) \le 2\exp\left(-n\gamma_0 h\left(\frac{\epsilon}{\gamma_0}\right)\right), \tag{5}$$

where $h(u) = 1 + u - \sqrt{1+2u}$ for $u > 0$ and $\gamma_0 := \max_{0\le x\le 1} x(1-x)(\log\frac{x}{1-x})^2$.

Proof Similar to the proof of 1, denote $Z_i := X_i \log p_i + (1 - X_i)\log(1 - p_i) - p_i \log p_i - (1 - p_i)\log(1 - p_i)$. It follows that $\log L_n - \mathbb{E}(\log L_n) = \sum_{i=1}^n Z_i$. One can easily verify that $\mathbb{P}(Z_i = (1 - p_i)\log(p_i/(1 - p_i))) = p_i$ and $\mathbb{P}(Z_i = -p_i \log(p_i/(1 - p_i))) = 1 - p_i$ for $p_i \in (0, 1)$. It follows that $\mathbb{E}(Z_i) = 0$ and

$$\operatorname{Var}(Z_i) = p_i(1 - p_i)\big[\log\big(p_i/(1 - p_i)\big)\big]^2 \le \gamma_0$$

by the definition of $\gamma_0 = \max_{0\le x\le 1} x(1 - x)(\log \frac{x}{1-x})^2$, which is approximately 0.439229 at $x_0 \approx 0.083222$ and $1 - x_0 \approx 0.916778$. In what follows, we shall show that the moments of Z_i are well bounded, satisfying the Bernstein condition as

$$\big|\mathbb{E}(Z_i)^k\big| \le \frac{1}{2}k!\gamma_0 \quad \text{for all } k \ge 2, \tag{6}$$

which implies the desired results by the same arguments of Theorem 1. To this end, the case of $k = 2$ follows trivially by $\mathbb{E}(Z_i)^2 = \operatorname{Var}(Z_i) \le \gamma_0$. For $k \ge 3$ and all $p_i \in [0, 1]$, we have

$$\big|\mathbb{E}(Z_i)^k\big| = \left|p_i(1 - p_i)^k\left(\log\frac{p_i}{1 - p_i}\right)^k + (1 - p_i)(p_i)^k\left(-\log\frac{p_i}{1 - p_i}\right)^k\right|,$$

which is symmetric about $p_i = 1/2$. It suffices to consider $p_i \in [0, 1/2]$. Indeed, it follows that, for $0 \le p_i \le 1/2$,

$$\begin{aligned}
\big|\mathbb{E}(Z_i)^k\big| &\le \left|p_i(1 - p_i)^k\left(\log\frac{p_i}{1 - p_i}\right)^k\right| + \left|(1 - p_i)(p_i)^k\left(-\log\frac{p_i}{1 - p_i}\right)^k\right| \\
&\le \big|p_i\big(\log p_i - \log(1 - p_i)\big)^k\big| + \left|(1 - p_i)^{k+1}\left(-\frac{p_i}{1 - p_i}\log\frac{p_i}{1 - p_i}\right)^k\right| \\
&\le p_i|\log p_i|^k + \max_{0\le t\le 1}(-t\log t)^k \\
&\le \big(k^k + 1\big)\exp(-k),
\end{aligned}$$

where the first inequality uses $|a + b| \le |a| + |b|$, the second inequality follows from $(1 - p_i)^k \le 1$ for $p_i \in [0, 1]$, the third inequality uses the fact that $|(\log p_i - \log(1 - p_i))^k| \le |\log p_i|^k$ and $0 \le p_i/(1 - p_i) \le 1$ for $0 \le p_i \le 1/2$, and the last inequality is implied by that the function $f(x) := x^r(\log x)^k$ for $x \in (0, 1)$ achieves its optimum at $x = e^{-k/r}$ for any $r > 0$ and integer $k \ge 1$. Note that the case of $k = 3$ follows from simple computations, that is,

$$\begin{aligned}
\big|\mathbb{E}(Z_i)^3\big| &= \left|p_i(1 - p_i)(1 - 2p_i)\left(\log\frac{p_i}{1 - p_i}\right)^3\right| \\
&\le \max_{0\le t\le 1/2} t(1 - t)(2t - 1)\left(\log\frac{t}{1 - t}\right)^3 \\
&\approx 1.14929 < 3\gamma_0
\end{aligned}$$

by $\gamma_0 > 2/5$. While for $k \ge 4$, we shall prove (6) by induction with the bound $|\mathbb{E}(Z_i)^k| \le (k^k + 1)\exp(-k)$. The case of $k = 4$ can be verified by

$$\big|\mathbb{E}(Z_i)^4\big| \le \big(4^4 + 1\big)\exp(-4) \approx 4.707119 < \frac{1}{2}4!\gamma_0,$$

where $\gamma_0 > 2/5$. Suppose for some $k \geq 4$, $(k^k + 1)\exp(-k) \leq k!\gamma_0/2$. For the case of $k+1$, it follows that

$$\begin{aligned}\frac{(k+1)^{k+1}+1}{\exp(k+1)} &\leq \frac{k!\gamma_0}{2}\frac{(k+1)^{k+1}+1}{e(k^k+1)}\\ &< \frac{k!\gamma_0}{2e}\frac{(k+1)^{k+1}}{k^k}\\ &= \frac{k!\gamma_0}{2e}(k+1)\left(1+\frac{1}{k}\right)^k\\ &\leq \frac{(k+1)!\gamma_0}{2}\end{aligned}$$

by $(1+1/k)^k \leq e$ for all $k > 4$, which completes the induction and finishes the proof of this proposition. □

Remark 3 Similar to Remark 1, the constants γ_0 and $b = 1$ are optimal for Bernstein-type inequality (5). One can also obtain a friendly form of (5) as

$$\mathbb{P}\big(\big|\log L_n - \mathbb{E}(\log L_n)\big| \geq n\epsilon\big) \leq 2\exp\left(-\frac{n\epsilon^2}{2(\gamma_0+\epsilon)}\right), \tag{7}$$

in which γ_0 is defined as Theorem 2.

3 Bennett's inequality

The Bernstein-type inequalities in Sect. 2 are useful on both left and right tails, while one may be more interested in the right tail in practice. A natural question to ask at this point is whether we can derive a tighter bound on right tails for the log-likelihood function of binary data. Thanks to the definition of the Bernoulli variables, the components of the log-likelihood function are well upper-bounded, which inspires us to use the Bennett inequality to derive a more informative upper bound on the right tail, see [1, 8] for more details on Bennett's inequality.

Theorem 3 *Under the condition of Theorem 1, for all $\epsilon > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - p_i)\log p_i \geq n\epsilon\right) \leq \exp\left(-e^2 n\gamma g\left(\frac{\epsilon}{e\gamma}\right)\right), \tag{8}$$

where $g(u) = (1+u)\log(1+u) - u$ for $u > 0$ and $\gamma := \max_{0\leq x\leq 1} x(1-x)(\log x)^2$.

Proof Consider $Y_i := (X_i - p_i)\log p_i$ and $S = \sum_{i=1}^n Y_i$. It is not hard to verify that $Y_i \leq -p_i \log p_i \leq 1/e$ for all $p_i \in [0,1]$. Following the classical method of Bennett's inequality, let $\phi(u) := e^u - u - 1$ for all $u \in \mathbb{R}$. By $u^{-2}\phi(u)$ is a nondecreasing function of $u \in \mathbb{R}$ (where at 0 we continuously extend the function). Hence, for all $\lambda > 0$, $e^{\lambda Y_i} - \lambda Y_i - 1 \leq e^2 Y_i^2 \phi(\lambda/e)$, implying that $\mathbb{E}(e^{\lambda Y_i}) \leq 1 + e^2\mathbb{E}(Y_i^2)\phi(\lambda/e)$ by $\mathbb{E}(Y_i) = 0$. Since Y_i's are independent, it follows that

$$\mathbb{E}\big(e^{\lambda S}\big) = \prod_{i=1}^n \mathbb{E}\big(e^{\lambda Y_i}\big) \leq \prod_{i=1}^n \big(1 + e^2\mathbb{E}\big(Y_i^2\big)\phi(\lambda/e)\big) \leq \exp\left(e^2\phi(\lambda/e)\sum_{i=1}^n \mathbb{E}\big(Y_i^2\big)\right) \tag{9}$$

by $\log(1+u) \le u$ for all $u \ge 0$. Note $\mathbb{E}(Y_i^2) = p_i(1-p_i)(\log p_i)^2 \le \gamma$, where γ is defined as in Theorem 3, which implies $\log \mathbb{E}(e^{\lambda S}) \le e^2 n\gamma\phi(\lambda/e)$ for all $\lambda > 0$. Then the Cramér transform of S is bounded by that of a corresponding Poisson random variable (see Chap. 2 of [1]), that is,

$$\mathbb{P}(S \ge n\epsilon) \le \exp\left(-e^2 n\gamma g\left(\frac{\epsilon}{e\gamma}\right)\right),$$

in which γ is defined as above and $g(u) = (1+u)\log(1+u) - u$ for $u > 0$, completing the proof of (8). □

Analogously, we can obtain Bennett's inequality for the joint log-likelihood function $\log L_n$.

Theorem 4 *Under the condition of Theorem 2, for all $\epsilon > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - p_i)\log\left(\frac{p_i}{1-p_i}\right) \ge n\epsilon\right) \le \exp\left(-\frac{n\gamma_0}{\beta^2} g\left(\frac{\beta\epsilon}{\gamma_0}\right)\right), \tag{10}$$

where $g(u) = (1+u)\log(1+u) - u$ for $u > 0$, $\beta := \max_{0<x<1}(1-x)\log(x/(1-x))$ and $\gamma_0 := \max_{0\le x\le 1} x(1-x)(\log\frac{x}{1-x})^2$.

Proof The proof follows similar arguments of Theorem 3. Indeed, set $Z_i := X_i \log p_i + (1 - X_i)\log(1-p_i) - p_i \log p_i - (1-p_i)\log(1-p_i)$, which satisfies $\mathbb{E}(Z_i) = 0$ and $\mathbb{E}(Z_i^2) \le \gamma_0$ by Theorem 2. To this end, we have $Z_i \le \max_{0\le t\le 1}(1-t)\log\frac{t}{1-t} = \beta$ for all $1 \le i \le n$. Hence, by similar arguments as those of Theorem 3, for all $\epsilon > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - p_i)\log\left(\frac{p_i}{1-p_i}\right) \ge n\epsilon\right) \le \exp\left(-\frac{n\gamma_0}{\beta^2} g\left(\frac{\beta\epsilon}{\gamma_0}\right)\right),$$

where $\beta, g(u)$ are defined in Theorem 4 and γ_0 is defined as in Theorem 2. □

Remark 4 To get a nice form of (8), one can verify that

$$g(u) \ge \frac{u^2}{2(1+u/3)},$$

which delivers that, under the condition of Theorem 3,

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - p_i)\log p_i \ge n\epsilon\right) \le \exp\left(-\frac{n\epsilon^2}{2(\gamma + \epsilon/(3e))}\right), \tag{11}$$

improving upon the Bernstein-type inequality by a factor of $3e$ on the right tail (compared with (4) in Remark 2). One can also obtain that

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - p_i)\log\left(\frac{p_i}{1-p_i}\right) \ge n\epsilon\right) \le \exp\left(-\frac{n\epsilon^2}{2(\gamma_0 + \beta\epsilon/3)}\right), \tag{12}$$

improving upon the Bernstein-type inequality by a factor of $3/\beta$ on the right tail (compared with (7) in Remark 3).

4 Extensions

The results in Sects. 2 and 3 are distribution-free, it is of interest to investigate some extensions of the concentration inequalities. In this section, we point out a possible direction that generalizes the results above.

As illustrated in Sect. 3, the improved upper tail relies on (9), which uses an elemental inequality $1 + x \leq \exp(x)$ for all $x \in \mathbb{R}$. In addition to the distribution-free concentration, it is also valuable to obtain some concentration inequalities dependent on parameters, which might be better than that when faced with specific problems. Inspired by the refined Bennett inequality provided by [13], we can sharpen Theorems 3 and 4 by equipping the arithmetic-geometric (AG) mean inequality. Under the condition of Theorem 3, we have

$$\prod_{i=1}^{n} \mathbb{E}\big(e^{\lambda Y_i}\big) \leq \prod_{i=1}^{n} \big(1 + \mathbb{E}\big(Y_i^2\big)\phi(\lambda/e)\big) \leq \big(1 + \bar{s}\phi(\lambda/e)\big)^n,$$

where $\bar{s} := \sum_{i=1}^{n} \mathbb{E}(Y_i^2)/n$ by $\prod_{k=1}^{n} x_k \leq (\bar{x})^n$ for $x_k \geq 0$ with $\bar{x} = \sum_{i=1}^{n} x_k/n$. Then we can deduce a similar Bennett inequality by the classical Chernoff method. Moreover, following the arguments in Sect. 3 of [13], we can obtain refined Bennett inequalities by applying a refined AG mean inequality, introduced by [2], which improves the upper bound of the MGF of S by extracting the difference between the parameters p_is.

5 Simulation study

In this section, we are going to show some simulation studies to express the improvements in our results. Because of the distribution-free property of the tail bounds, we shall ignore the parameters p_is and compare the logarithmic tail probabilities for various sample sizes n and the error rates ϵ. Both the two-sides tail and the right-side tail are illustrated with different n and ϵ.

At first, consider the improvement of the concentration bound for $\sum_{i=1}^{n}(X_i - p_i)\log p_i$. The sample size takes four values with $n = 100, 200, 500$, and 1000. For each n, the error rate varies from 0.1 to 1. For each fixed (n, ϵ), the two-sides Bernstein-type tail probability bounds (2) and (4) with optimal constants and the result due to [10] (see (1)) are compared, which can be found in Fig. 1. The sharp Bernstein-type inequalities (Theorem 1) improve the previous result (1) significantly for various cases.

For the joint log-likelihood function, the result provided by [11] asserts that (see Theorem 1 of [11] with $K = 2$), for small $\epsilon > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - p_i)\log\left(\frac{p_i}{1-p_i}\right)\right| \geq n\epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{4(\max\{\log K, \log 5\})^2}\right). \tag{13}$$

Similar to the first example above, the sample size $n = 100, 200, 500$, and 1000. For each n, the error rate varies from 0.1 to 1. The tail probabilities (5), (7), and (13) are demonstrated. From Fig. 2, our results perform better than (13) over different n and small ϵ, which are interesting cases in practice. Moreover, the right-tail concentration results (Theorems 3 and 4) are also compared. To this end, let the sample size take the values from $\{100, 300, 1000, 2000\}$, and the error rate varies from 0.1 to 1. The one-side tail bounds (1), (2), and (8) can be found in Fig. 3, where the factor 2 in (1) and (2) is removed for the

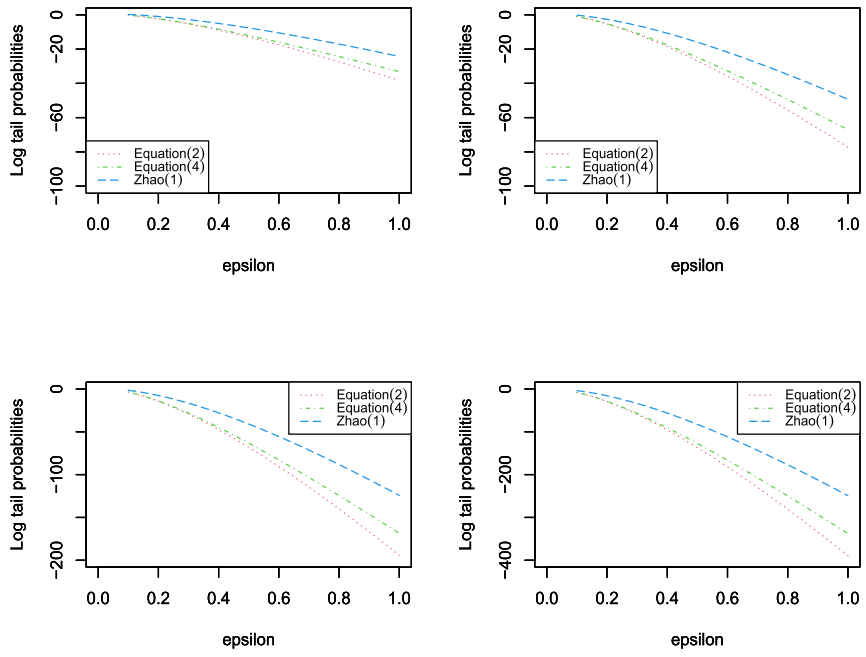

Figure 1 The logarithmic tail probabilities for (1), (2), and (4) with different n (upper left $n = 100$, upper right $n = 200$, bottom left $n = 500$, and bottom right $n = 1000$) and ϵ

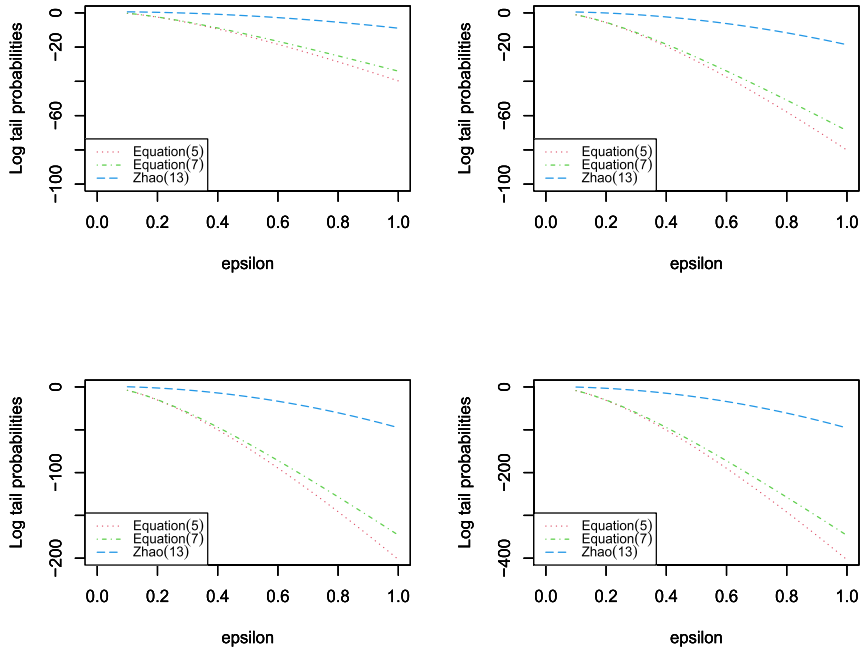

Figure 2 The logarithmic tail probabilities for (5), (7), and (13) with different n (upper left $n = 100$, upper right $n = 200$, bottom left $n = 500$, and bottom right $n = 1000$) and ϵ

right-tail case. Similarly, the right-tail bounds (5), (10), and (13) can be found in Fig. 4 in which the factor 2 in (5) and (13) is removed for the right-tail case. The sharp Bennett inequalities improve the right-tail bounds significantly when the error rate increases.

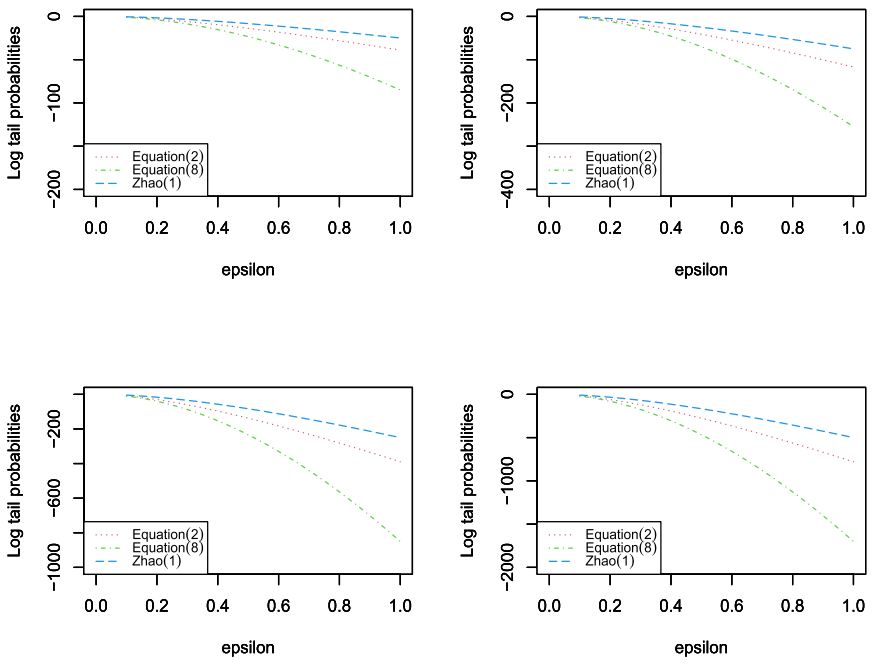

Figure 3 The logarithmic right-tail probabilities for (1), (2), and (8) with different n (upper left $n = 100$, upper right $n = 300$, bottom left $n = 1000$, and bottom right $n = 2000$) and ϵ

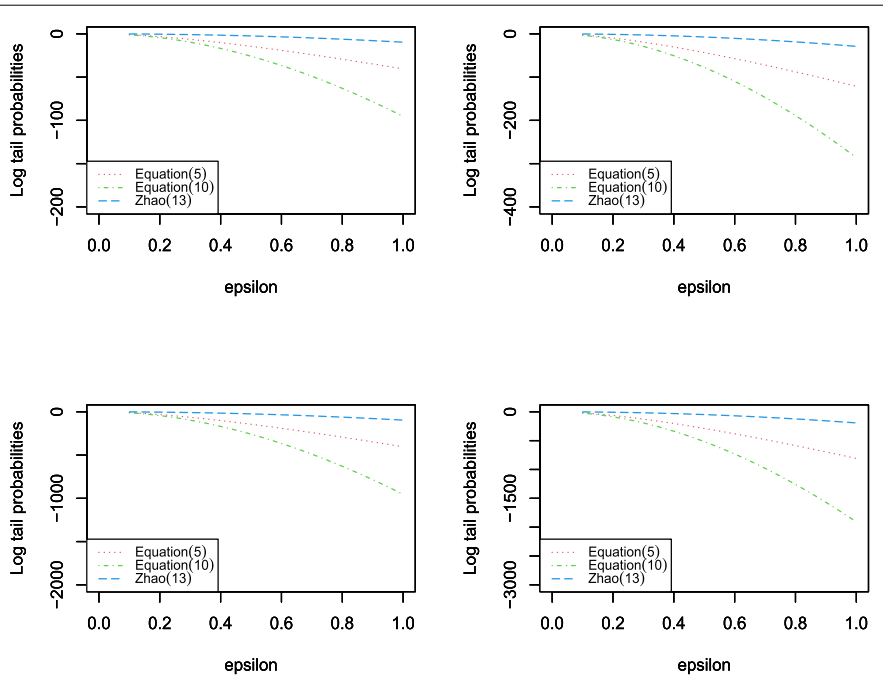

Figure 4 The logarithmic right-tail probabilities for (5), (10), and (13) with different n (upper left $n = 100$, upper right $n = 300$, bottom left $n = 1000$, and bottom right $n = 2000$) and ϵ

6 Conclusion

We study the distribution-free concentration inequalities for the log-likelihood function of Bernoulli variables. Indeed, we established the optimal Bernstein-type inequalities with the best constants of variance factor and scale factor in the sense of sub-gamma random variable. Moreover, Bennett's inequalities with sharp constants are also illustrated, which improves the scale factors of Bernstein-type inequalities on the right tails.

There are some limitations of this study. First, it is more interesting to consider the concentration with multiple discrete distributions, see [11] for more details. Secondly, we focus on the distribution-free concentration, while it is also valuable to obtain some concentration inequalities dependent on parameters, which might be better than that when faced with specific problems. Furthermore, one can consider the concentration of the likelihood ratio statistics, which is an interesting direction for further study.

Funding
No funding was received for this research.

Availability of data and materials
No data were used to support this study.

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
Zhonggui Ren wrote the main manuscript text, prepared Figs. 1-4, and reviewed the manuscript

Received: 2 January 2023 Accepted: 30 May 2023 Published online: 06 June 2023

References

1. Boucheron, S., Lugosi, G., Massart, P.: Concentration Inequalities: A Non-asymptotic Theory of Independence. Oxford University Press, Oxford (2013)
2. Cartwright, D.I., Field, M.J.: A refinement of the arithmetic mean-geometric mean inequality. Proc. Am. Math. Soc. **71**, 36–38 (1978)
3. Choi, D.S., Wolfe, P.J., Airoldi, E.M.: Stochastic blockmodels with a growing number of classes. Biometrika **99**, 273–284 (2012)
4. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963)
5. Paul, S., Chen, Y.: Consistent community detection in multi-relational data through restricted multi-layer stochastic block model. Electron. J. Stat. **10**, 3807–3870 (2016)
6. Raginsky, M., Sason, I.: Concentration of measure inequalities in information theory, communications, and coding. Found. Trends Commun. Inf. Theory **10**, 1–246 (2013)
7. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
8. Zhang, H., Chen, S.: Concentration inequalities for statistical inference. Commun. Math. Sci. **37**, 1–85 (2021)
9. Zhang, H., Wei, H.: Sharper sub-Weibull concentrations. Mathematics **10**, 2252 (2022)
10. Zhao, Y.: A note on new Bernstein-type inequalities for the log-likelihood function of Bernoulli variables. Stat. Probab. Lett. **163**, 108779 (2020)
11. Zhao, Y.: An optimal uniform concentration inequalities for discrete entropy in the high-dimensional setting. Bernoulli **28**, 1892–1911 (2022)
12. Zhao, Y., Weko, C.: Network inference from grouped observations using hub models. Stat. Sin. **29**, 225–244 (2019)
13. Zheng, S.: An improved Bennett's inequality. Commun. Stat., Theory Methods **47**, 4152–4159 (2017)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.