PROBABILITY
AND
MATHEMATICAL STATISTICS

Vol. 44, Fasc. 1 (2024), pp. 1–13
Published online 24.5.2024
doi:10.37190/0208-4147.00162

A NOTE ON A BERNSTEIN-TYPE INEQUALITY FOR THE LOG-LIKELIHOOD FUNCTION OF CATEGORICAL VARIABLES WITH INFINITELY MANY LEVELS

BY

YUNPENG ZHAO* (FORT COLLINS, CO)

Abstract. We prove a Bernstein-type bound for the difference between the average of the negative log-likelihoods of independent categorical variables with infinitely many levels – that is, a countably infinite number of categories, and its expectation – namely, the Shannon entropy. The result holds for the class of discrete random variables with tails lighter than or of the same order as a discrete power-law distribution. Most commonly used discrete distributions, such as the Poisson distribution, the negative binomial distribution, and the power-law distribution itself, belong to this class. The bound is effective in the sense that we provide a method to compute the constants within it. The new technique we develop allows us to obtain a uniform concentration inequality for categorical variables with a finite number of levels with the same optimal rate as in the literature, but with a much simpler proof.

2020 Mathematics Subject Classification: Primary 60E15.

Key words and phrases: Bernstein-type bound, categorical variables with infinitely many levels, concentration inequality, Shannon entropy.

1. INTRODUCTION

Concentration inequalities provide powerful tools for various subjects, including information theory [9], algorithm analysis [7], and statistics [14, 13]. The objective of this paper is to establish an exponential decay bound, with computable constants, for the difference between the negative log-likelihood of categorical variables with infinitely many levels and its expectation, i.e., the Shannon entropy.

Let X be a discrete random variable that takes an infinite set of possible values on $\mathcal{X} = \{x_1, \ldots, x_k, \ldots\}$. Let $p_k = \mathbb{P}(X = x_k)$ be the probability mass at x_k. Assume, without loss of generality, that $p_k > 0$ for each k; otherwise, simply remove x_k with $p_k = 0$ from $\mathcal{X}$. Let $P(X)$ be a random variable with $P(X) = p_k$ if $X = x_k$, $k \geqslant 1$. Then $\mathbb{E}[-\log P(X)] = -\sum_{k=1}^{\infty} p_k \log p_k$ is the Shannon

* This research was supported by the U.S. National Science Foundation grant DMS-1840203.

entropy,[1] which is a key concept in information theory [12, 5] Note that neither $P(X)$ nor the entropy depends on the elements in $\mathcal{X}$. In fact, $\mathcal{X}$ is not necessarily a set of numbers; the set can contain generic symbols such as letters and is therefore named the alphabet. Consequently, we can equivalently define $P(X)$ and entropy for a categorical variable with infinitely many levels. Let $\mathbf{z} = (z_1, \dots, z_k, \dots)$ be a dummy coding of a categorical variable with a countably infinite number of categories, in which one and only one entry is 1, and the others are 0.

Let $\mathbf{z}_1, \dots, \mathbf{z}_n$ be independently and identically distributed (i.i.d.) copies of $\mathbf{z}$. Then $\sum_{i=1}^n \sum_{k=1}^\infty z_{ik} \log p_k$ is the joint log-likelihood of $\mathbf{z}_1, \dots, \mathbf{z}_n$, where z_{ik} is the kth entry of $\mathbf{z}_i$. A natural question is to study the concentration of the log-likelihood and its expectation – namely, the negative entropy. By the weak law of large numbers,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{\infty} z_{ik}\log p_k - \sum_{k=1}^{\infty} p_k \log p_k\right| \geqslant \epsilon\right) \to 0,$$

provided that the entropy is finite. This result, particularly for the case of $\mathbf{z}$ with finite categories, is called the asymptotic equipartition property in the information theory literature. It serves as the foundation for many important results in this field [5, 6].

Exponential decay concentration bounds for log-likelihoods of categorical variables have recently attracted attention. Originally motivated by theoretical research in the statistical analysis of network data [4], Zhao [15] proved a Bernstein-type inequality for log-likelihoods of categorical variables:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log p_k - \sum_{k=1}^{K} p_k \log p_k\right| \geqslant \epsilon\right) \leqslant 2K\exp\left\{-\frac{n\epsilon^2}{2K(K+\epsilon)}\right\},$$

where n is the number of variables and K is the number of categories. The bound is uniform over p_k and shrinks to zero if $(K^2 \log K)/n = o(1)$. Ren [10] improved the inequality in [15] by obtaining the optimal constant for the case when $K = 2$. Zhao [16] proved another uniform concentration bound that improves the rate to $(\log K)^2/n = o(1)$ and demonstrated that the new rate is optimal.

All of the aforementioned works studied inequalities for categorical variables with a finite number of levels, while our focus in this work is on variables with infinitely many levels. Zhao [16] pointed out that a uniform concentration bound does not exist over the class of $\{p_k\}_{k\geqslant 1}$ if no additional conditions are imposed beyond the requirement that the distributions have finite entropies. In this paper, we prove a Bernstein-type inequality for categorical variables with infinitely many levels, assuming that $\sum_{k=1}^\infty p_k^{1-r}$ has a finite upper bound for certain r. The concentration bound depends solely on the value of r and on the upper bound of $\sum_{k=1}^\infty p_k^{1-r}$. The theme of the present paper is not directly focused on entropy estimation (see [1, 3]

[1] Throughout the paper, "log" denotes the natural logarithm.

for examples) because $\sum_{k=1}^{\infty} z_{ik} \log p_k$ contains the parameters of the distribution. However, this type of concentration inequalities has recently been applied to the concentration of empirical relative entropy [8].

In Section 2, we prove the main result. In Section 3, we show that the assumption of $\sum_{k=1}^{\infty} p_k^{1-r}$ being finite holds if the tail of $\{p_k\}_{k\geqslant 1}$ drops faster or on the same order as a discrete power-law distribution; conversely, the assumption cannot be satisfied if the tail drops slower than all power-law distributions. Most commonly used discrete distributions such as the Poisson distribution, the negative binomial distribution, and the power-law distribution itself, satisfy this assumption. Furthermore, we propose a method to compute the constants in the concentration bound. In Section 4, we apply the same proof technique to categorical variables with a finite number of levels and obtain a uniform concentration inequality with the same optimal rate as in [16], albeit with a better constant.

2. MAIN RESULT

Our result requires only one assumption on $\{p_k\}_{k\geqslant 1}$:

ASSUMPTION 1. There exists $0 < r < 1$ such that

$$\sum_{k=1}^{\infty} p_k^{1-r} < \infty.$$

In the following, we denote by C_r an upper bound for $\sum_{k=1}^{\infty} p_k^{1-r}$, a quantity that will appear in the concentration bound. An estimate of C_r will be provided in Section 3.

Assumption 1 implies that the tail of $\{p_k\}_{k\geqslant 1}$ cannot be too heavy. In Section 3, we will elaborate on this assumption by showing that the assumption holds if the tail of $\{p_k\}_{k\geqslant 1}$ is lighter than or on the same order as a discrete power-law distribution; conversely, it cannot be satisfied if the tail is heavier than all power-law distributions.

First, note that Assumption 1 ensures the finiteness of the entropy.

PROPOSITION 2.1. *Under Assumption* 1, $-\sum_{k=1}^{\infty} p_k \log p_k < \infty$.

Proof. We have

$$-\sum_{k=1}^{\infty} p_k \log p_k = \sum_{k=1}^{\infty} p_k^{1-r}(-p_k^r \log p_k) \leqslant \frac{1}{er} \sum_{k=1}^{\infty} p_k^{1-r}.$$

The last inequality holds because $-p_k^r \log p_k$ on $[0,1]$ is maximized at $p_k = e^{-1/r}$. This result can be easily verified by comparing the function value at the stationary point in $(0,1)$, which is unique for this function, with the values at the boundaries. Here, we use the convention $q^r \log q = 0$ at $q = 0$, which ensures the continuity of the function on $[0,1]$, as $\lim_{q\to 0+} q^r \log q = 0$. ■

Readers are referred to [2] for a more thorough study of the conditions for the finiteness of entropy on categorical variables with infinitely many levels.

Let $Y_i = \sum_{k=1}^{\infty} z_{ik} \log p_k - \sum_{k=1}^{\infty} p_k \log p_k$. The key ingredient of the proof of the main result is to bound the moment generating function (MGF) of Y_i, which is defined as

$$\mathbb{E}[e^{\lambda Y_i}] = \Big(\sum_{k=1}^{\infty} p_k^{\lambda+1}\Big) \exp\Big(-\lambda \sum_{k=1}^{\infty} p_k \log p_k\Big).$$

Let the MGF of Y_i be denoted by $M_{Y_i}(\lambda)$. Under Assumption 1, $M_{Y_i}(\lambda)$ is finite for $|\lambda| < r$ because

$$\sum_{k=1}^{\infty} p_k^{\lambda+1} \leqslant \sum_{k=1}^{\infty} p_k^{1-r} < \infty.$$

Conversely, if Assumption 1 does not hold then $\sum_{k=1}^{\infty} p_k^{\lambda+1}$ diverges for all $\lambda < 0$, because if $\sum_{k=1}^{\infty} p_k^{\lambda+1}$ converges for a certain negative λ then it must be within the interval $(-1, 0)$ and one can take $r = -\lambda$.

We now give the main result.

THEOREM 2.1 (Main result). *Under Assumption* 1*, specifically, if there exists* $0 < r < 1$ *such that*

$$\sum_{k=1}^{\infty} p_k^{1-r} \leqslant C_r < \infty,$$

then for $|\lambda| < r$*,*

$$M_{Y_i}(\lambda) \leqslant \exp\Big(\frac{C_r \lambda^2}{r^2} \frac{1}{1 - |\lambda|/r} \frac{1}{2\sqrt{\pi}}\Big).$$

Furthermore, for all $\epsilon > 0$*,*

$$\text{(2.1)} \quad \mathbb{P}\Big(\Big|\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{\infty} z_{ik} \log p_k - \sum_{k=1}^{\infty} p_k \log p_k\Big| \geqslant \epsilon\Big) \leqslant 2 \exp\Big(-\frac{n\epsilon^2}{2C_r/(\sqrt{\pi} r^2) + 2\epsilon/r}\Big).$$

Proof. For $|\lambda| < r$,

$$\begin{aligned} \text{(2.2)} \quad \log M_{Y_i}(\lambda) &= \log\Big(\sum_{k=1}^{\infty} p_k^{\lambda+1}\Big) - \lambda \sum_{k=1}^{\infty} p_k \log p_k \\ &\leqslant \sum_{k=1}^{\infty} p_k^{\lambda+1} - 1 - \lambda \sum_{k=1}^{\infty} p_k \log p_k \\ &= \sum_{k=1}^{\infty} p_k \exp(\lambda \log p_k) - 1 - \lambda \sum_{k=1}^{\infty} p_k \log p_k \\ &= \sum_{k=1}^{\infty} \Big(p_k + \lambda p_k \log p_k + \sum_{m=2}^{\infty} \frac{1}{m!} \lambda^m p_k (\log p_k)^m\Big) - 1 - \lambda \sum_{k=1}^{\infty} p_k \log p_k, \end{aligned}$$

where the inequality follows from $\log x \leqslant x - 1$ for $x > 0$.

For $m \geqslant 2$, it is easy to check that the minimum of $p_k^r(\log p_k)^m$ on $[0,1]$ when m is an odd number, and the maximum when m is an even number, are both achieved at $e^{-m/r}$. This can be verified by comparing the function value at the unique stationary point within (0, 1) with the values at the boundaries. Here we use the convention $q^r(\log q)^m = 0$ at $q = 0$ as before, which ensures the continuity of the function on $[0,1]$, as $\lim_{q\to 0+} q^r(\log q)^m = 0$.

Therefore, for $m \geqslant 2$,

$$\begin{aligned} \left|\frac{1}{m!}\lambda^m p_k(\log p_k)^m\right| &= p_k^{1-r}\frac{1}{m!}|\lambda|^m|p_k^r(\log p_k)^m| \\ &\leqslant p_k^{1-r}\frac{1}{m!}|\lambda|^m e^{-m}\left(\frac{m}{r}\right)^m \\ &\leqslant p_k^{1-r}\frac{1}{m!}(|\lambda|/r)^m\frac{m!}{\sqrt{2\pi m}} \\ &\leqslant p_k^{1-r}\left(\frac{|\lambda|}{r}\right)^m\frac{1}{2\sqrt{\pi}}, \end{aligned} \tag{2.3}$$

where the first inequality is obtained by replacing $|p_k^r(\log p_k)^m|$ with its maximum and the second inequality follows from Stirling's formula (see [11] for example):

$$m! \geqslant \sqrt{2\pi m}\left(\frac{m}{e}\right)^m \qquad \text{for } m \geqslant 1.$$

It follows that for $|\lambda| < r$,

$$\begin{aligned} \left|\sum_{m=2}^{\infty}\frac{1}{m!}\lambda^m p_k(\log p_k)^m\right| &\leqslant \sum_{m=2}^{\infty}\left|\frac{1}{m!}\lambda^m p_k(\log p_k)^m\right| \\ &\leqslant p_k^{1-r}\sum_{m=2}^{\infty}\left(\frac{|\lambda|}{r}\right)^m\frac{1}{2\sqrt{\pi}} = p_k^{1-r}\frac{\lambda^2}{r^2}\frac{1}{1-|\lambda|/r}\frac{1}{2\sqrt{\pi}}, \end{aligned}$$

and

$$\sum_{k=1}^{\infty}\left|\sum_{m=2}^{\infty}\frac{1}{m!}\lambda^m p_k(\log p_k)^m\right| \leqslant C_r\frac{\lambda^2}{r^2}\frac{1}{1-|\lambda|/r}\frac{1}{2\sqrt{\pi}}.$$

Since the three terms under the first sum in the last line of (2.2) all converge absolutely for $|\lambda| < r$, one can take the sum term by term. Therefore, for $|\lambda| < r$,

$$\log M_{Y_i}(\lambda) \leqslant \sum_{k=1}^{\infty}\left|\sum_{m=2}^{\infty}\frac{1}{m!}\lambda^m p_k(\log p_k)^m\right| \leqslant C_r\frac{\lambda^2}{r^2}\frac{1}{1-|\lambda|/r}\frac{1}{2\sqrt{\pi}},$$

and

$$M_{Y_i}(\lambda) \leqslant \exp\left(\frac{C_r\lambda^2}{r^2}\frac{1}{1-|\lambda|/r}\frac{1}{2\sqrt{\pi}}\right). \tag{2.4}$$

The second part follows from a standard argument using the Chernoff bound, which can be found in [14, Chapter 2]. We give the details for completeness. For $t > 0$ and $0 < \lambda < r$,

$$\mathbb{P}\Big(\sum_{i=1}^{n} Y_i \geqslant t\Big) = \mathbb{P}(e^{\lambda \sum_{i=1}^{n} Y_i} \geqslant e^{\lambda t}) \leqslant \frac{\prod_{i=1}^{n} M_{Y_i}(\lambda)}{e^{\lambda t}}$$
$$\leqslant \exp\Big\{\frac{nC_r\lambda^2}{r^2}\,\frac{1}{1-|\lambda|/r}\,\frac{1}{2\sqrt{\pi}} - \lambda t\Big\},$$

where the first inequality is Markov's inequality and the second inequality follows from (2.4). By setting

$$\lambda = \frac{t}{nC_r/(\sqrt{\pi}r^2) + t/r} \in (0, r),$$

we obtain

$$\mathbb{P}\Big(\sum_{i=1}^{n} Y_i \geqslant t\Big) \leqslant \exp\Big(-\frac{t^2}{2nC_r/(\sqrt{\pi}r^2) + 2t/r}\Big).$$

The left tail bound can be derived similarly by setting $\lambda = -\frac{t}{nC_r/(\sqrt{\pi}r^2)+t/r}$. Therefore,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} Y_i\Big| \geqslant t\Big) \leqslant 2\exp\Big(-\frac{t^2}{2nC_r/(\sqrt{\pi}r^2) + 2t/r}\Big).$$

Finally, letting $t = n\epsilon$, we get

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} Y_i\Big| \geqslant \epsilon\Big) \leqslant 2\exp\Big(-\frac{n\epsilon^2}{2C_r/(\sqrt{\pi}r^2) + 2\epsilon/r}\Big). \quad \blacksquare$$

Theorem 2.1 can be generalized to $\{\mathbf{z}_i\}_{i=1}^n$ with independent but non-identical distributions. Let $p_{ik} = \mathbb{P}(z_{ik} = 1)$ be the probability that the ith observation belongs to category k, and$-\sum_{k=1}^{\infty} p_{ik} \log p_{ik}$ be the entropy of $\mathbf{z}_i$. In addition, redefine Y_i and $M_{Y_i}(\lambda)$ accordingly. We have the following result for non-identical distributions:

COROLLARY 2.1. *If there exists* $0 < r < 1$ *such that*

$$\sum_{k=1}^{\infty} p_{ik}^{1-r} \leqslant C_{r,i} < \infty, \quad i = 1, \ldots, n,$$

then for $|\lambda| < r$*,*

$$M_{Y_i}(\lambda) \leqslant \exp\Big(\frac{C_{r,i}\lambda^2}{r^2}\,\frac{1}{1-|\lambda|/r}\,\frac{1}{2\sqrt{\pi}}\Big).$$

Furthermore, for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{\infty}(z_{ik}-p_{ik})\log p_{ik}\right| \geqslant \epsilon\right)$$
$$\leqslant 2\exp\left(-\frac{n\epsilon^2}{2\sum_{i=1}^{n} C_{r,i}/(n\sqrt{\pi}r^2)+2\epsilon/r}\right).$$

The proof is the same as that of Theorem 2.1.

3. DETERMINING THE CONSTANTS IN THE BOUND

The radius of convergence r in (2.3) and the upper bound C_r for $\sum_{k=1}^{\infty} p_k^{1-r}$ are the constants to be determined if one wants to use (2.1) as an effective upper bound for a given distribution $\{p_k\}_{k\geqslant 1}$.

We first determine the types of distributions and the range of r that can make $\sum_{k=1}^{\infty} p_k^{1-r}$ converge. Intuitively speaking, for distributions that satisfy Assumption 1, the tail of $\{p_k\}_{k\geqslant 1}$ cannot be too heavy. We make the above statement precise in the following proposition.

PROPOSITION 3.1. *The distribution* $\{p_k\}_{k\geqslant 1}$ *satisfies Assumption* 1 *if the tail of* $\{p_k\}_{k\geqslant 1}$ *is lighter than or on the same order as a discrete power-law distribution; conversely, Assumption* 1 *cannot be satisfied if the tail is heavier than all power-law distributions. Specifically:*

(i) *If*

$$\lim_{k\to\infty}\frac{p_k}{k^{-\alpha}} = 0 \quad \textit{for all } \alpha > 1,$$

then

$$\sum_{k=1}^{\infty} p_k^{1-r} < \infty \quad \textit{for all } 0 < r < 1.$$

(ii) *If*

$$0 < \liminf_{k\to\infty}\frac{p_k}{k^{-\alpha}} \leqslant \limsup_{k\to\infty}\frac{p_k}{k^{-\alpha}} < \infty \quad \textit{for some } \alpha > 1,$$

then

$$\sum_{k=1}^{\infty} p_k^{1-r} < \infty \quad \textit{if and only if} \quad 0 < r < \frac{\alpha-1}{\alpha}.$$

(iii) *If*

$$\lim_{k\to\infty}\frac{p_k}{k^{-\alpha}} = \infty \quad \textit{for all } \alpha > 1,$$

then

$$\sum_{k=1}^{\infty} p_k^{1-r} = \infty \quad \textit{for all } 0 < r < 1.$$

Proof. Recall that $\sum_{k=1}^{\infty} k^{-\beta}$ converges for $\beta > 1$, and diverges for $\beta \leqslant 1$. Statement (i) is obvious by taking $\alpha > 1/(1-r)$. Statement (ii) is also obvious by noticing that the assumption implies that there exist positive constants a_1, a_2 such that $a_1 k^{-\alpha} \leqslant p_k \leqslant a_2 k^{-\alpha}$ for sufficiently large k. We prove (iii) by contradiction. If there exists $0 < r < 1$ such that $\sum_{k=1}^{\infty} p_k^{1-r} < \infty$, then

$$\liminf_{k\to\infty} \frac{p_k^{1-r}}{k^{-1}} = 0.$$

This implies

$$\liminf_{k\to\infty} \frac{p_k}{k^{-1/(1-r)}} = 0,$$

which contradicts the assumption since $1/(1-r) > 1$. ∎

Proposition 3.1 implies that there is a wide class of discrete distributions satisfying Assumption 1, including the most commonly used ones such as the Poisson distribution, the negative binomial distribution, and the power-law distribution itself. The class even contains certain discrete random variables that do not have finite expectations. In fact, if X follows a discrete power-law distribution with $1 < \alpha \leqslant 2$ then $\mathbb{E}[X] = \infty$ since $\sum_{k=1}^{\infty} k^{-(\alpha-1)}$ diverges. But such distributions satisfy Assumption 1 by Proposition 3.1(ii).

REMARK 3.1. It may be surprising, at first glance, to get an exponential decay bound for a power-law distribution, which itself is heavy-tailed. But note that (2.1) is a concentration bound for $\log P(X)$, not for X. The log-likelihood $\log P(X)$ is typically better-behaved than X that takes values on non-negative integers and follows a power-law distribution. For example, the MGF of X is infinite if X follows a power-law distribution while the MGF of $\log P(X)$ can be finite. This phenomenon can be explained by noticing that $-\log(k^{-\alpha})$ grows much slower than k.

Finally, we discuss how to compute C_r after r is determined by Proposition 3.1. In practice, one can compute the partial sum of $\sum_{k=1}^{\infty} p_k^{1-r}$ until the increment is negligible. The value obtained in this way, however, is a lower bound for $\sum_{k=1}^{\infty} p_k^{1-r}$ as in principle, the tail behavior cannot be predicted by a finite number of terms[2].

If the tail of $\{p_k\}_{k\geqslant 1}$ is dominated by a power-law distribution, we propose a method that can compute an upper bound for $\sum_{k=1}^{\infty} p_k^{1-r}$ at any tolerance level. Specifically, the next proposition shows how to compute an upper bound C_r for $\sum_{k=1}^{\infty} p_k^{1-r}$ with $|\sum_{k=1}^{\infty} p_k^{1-r} - C_r|$ smaller than a pre-specified tolerance level if we find k_0 such that $p_k \leqslant c_0 k^{-\alpha}$ for $k > k_0$. Note that such a k_0 exists if $\{p_k\}_{k\geqslant 1}$ satisfies the condition in (i) or (ii) in Proposition 3.1.

[2]This issue is minor in practice especially when p_k drops exponentially. The series $\sum_{k=1}^{\infty} p_k^{1-r}$ converges fast in this case. There is nothing wrong with taking the partial sum until the increment is negligible. The method in Proposition 3.2 is useful to someone who needs a rigorous upper bound.

PROPOSITION 3.2. *Suppose k_0 is a positive integer such that $p_k \leqslant c_0 k^{-\alpha}$ for a certain $\alpha > 1$ and all $k > k_0$, where $c_0 > 0$. Pick r such that $0 < r < (\alpha - 1)/\alpha$. For all $\epsilon > 0$, let*

$$k_1 = \max\left\{k_0, \left\lceil \left(\frac{\epsilon(\alpha(1-r)-1)}{c_0^{1-r}}\right)^{-1/[\alpha(1-r)-1]} \right\rceil \right\},$$

where $\lceil \cdot \rceil$ indicates rounding up to the next integer. Then

$$C_r = \sum_{k=1}^{k_1} p_k^{1-r} + \epsilon$$

satisfies

$$0 \leqslant C_r - \sum_{k=1}^{\infty} p_k^{1-r} \leqslant \epsilon.$$

Proof. We only need to bound the tail probability for $k > k_1$:

$$\begin{aligned} \sum_{k=k_1+1}^{\infty} p_k^{1-r} &\leqslant c_0^{1-r} \sum_{k=k_1+1}^{\infty} k^{-\alpha(1-r)} \\ &= c_0^{1-r} \sum_{k=k_1}^{\infty} \int_k^{k+1} (k+1)^{-\alpha(1-r)}\, dx \\ &\leqslant c_0^{1-r} \int_{k_1}^{\infty} x^{-\alpha(1-r)}\, dx \\ &= \frac{c_0^{1-r}}{\alpha(1-r)-1} k_1^{-(\alpha(1-r)-1)} \leqslant \epsilon, \end{aligned}$$

where the first inequality holds because $p_k \leqslant c_0 k^{-\alpha}$ for all $k > k_0$ and the last inequality holds because

$$k_1 \geqslant \left\lceil \left(\frac{\epsilon(\alpha(1-r)-1)}{c_0^{1-r}}\right)^{-1/[\alpha(1-r)-1]} \right\rceil.$$

Therefore,

$$\sum_{k=1}^{\infty} p_k^{1-r} = \sum_{k=1}^{k_1} p_k^{1-r} + \sum_{k=k_1+1}^{\infty} p_k^{1-r} \leqslant \sum_{k=1}^{k_1} p_k^{1-r} + \epsilon. \quad \blacksquare$$

Proposition 3.2 provides a general method for estimating the upper bound of $\sum_k p_k^{1-r}$. For power-law, Poisson, and negative binomial distributions, we offer more explicit estimates of the upper bound of $\sum_k p_k^{1-r}$ below.

PROPOSITION 3.3. *For* $p_k = k^{-\alpha}/\zeta(\alpha)$ $(\alpha > 1,\ k = 1, 2, \dots)$*, and all* r *such that* $0 < r < (\alpha-1)/\alpha$*,*

$$\sum_{k=1}^{\infty} p_k^{1-r} = \frac{1}{[\zeta(\alpha)]^{1-r}}\zeta(\alpha(1-r)),$$

where $\zeta(\alpha)$ *is the Riemann zeta function.*

The proof is straightforward.

PROPOSITION 3.4. *For* $p_k = e^{-\mu}\mu^k/k!$ $(\mu > 0,\ k = 0, 1, 2, \dots)$*, all* r *such that* $0 < r < 1$*, and all integers* k_0 *such that* $k_0 > e\mu$*,*

$$\begin{aligned}&\sum_{k=0}^{\infty} p_k^{1-r}\\ &\quad\leqslant e^{-\mu(1-r)}\left[\sum_{k=0}^{k_0-1}\left(\frac{\mu^k}{k!}\right)^{1-r} + (2\pi k_0)^{-\frac{1}{2}(1-r)}\left(\frac{e\mu}{k_0}\right)^{k_0(1-r)}\frac{1}{1-(e\mu/k_0)^{1-r}}\right].\end{aligned}$$

Proof. We have

$$\begin{aligned}&\sum_{k=0}^{\infty} p_k^{1-r}\\ &\leqslant e^{-\mu(1-r)}\left[\sum_{k=0}^{k_0-1}\left(\frac{\mu^k}{k!}\right)^{1-r} + \sum_{k=k_0}^{\infty}\mu^{k(1-r)}(2\pi k)^{-\frac{1}{2}(1-r)}\left(\frac{e}{k}\right)^{k(1-r)}\right]\\ &\leqslant e^{-\mu(1-r)}\left[\sum_{k=0}^{k_0-1}\left(\frac{\mu^k}{k!}\right)^{1-r} + (2\pi k_0)^{-\frac{1}{2}(1-r)}\sum_{k=k_0}^{\infty}\left(\frac{e\mu}{k_0}\right)^{k(1-r)}\right]\\ &= e^{-\mu(1-r)}\left[\sum_{k=0}^{k_0-1}\left(\frac{\mu^k}{k!}\right)^{1-r} + (2\pi k_0)^{-\frac{1}{2}(1-r)}\left(\frac{e\mu}{k_0}\right)^{k_0(1-r)}\frac{1}{1-(e\mu/k_0)^{1-r}}\right]. \qquad \blacksquare\end{aligned}$$

PROPOSITION 3.5. *Let* X *follow a negative binomial distribution, i.e.,*

$$p_k = \binom{k+s-1}{k}(1-p)^k p^s, \qquad k = 0, 1, 2, \dots,$$

where $0 < p < 1$ *and* s *is a positive integer. Then for all* r *such that* $0 < r < 1$*, we have*

$$\sum_{k=0}^{\infty} p_k^{1-r} \leqslant \left(\frac{p}{1-\sqrt{1-p}}\right)^{s(1-r)}\frac{1}{1-(1-p)^{(1-r)/2}}.$$

Proof. The MGF of X is

$$\mathbb{E}[e^{\lambda X}] = \left(\frac{p}{1-(1-p)e^{\lambda}}\right)^{s} \quad \text{for } \lambda < -\log(1-p).$$

By Markov's inequality, for $0 < \lambda < -\log(1-p)$,

$$p_k \leqslant \mathbb{P}(X \geqslant k) = \mathbb{P}(e^{\lambda X} \geqslant e^{\lambda k}) \leqslant \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda k}} = \left(\frac{p}{1-(1-p)e^{\lambda}}\right)^s e^{-\lambda k}.$$

Letting $\lambda = -\frac{1}{2}\log(1-p)$, we obtain

$$p_k \leqslant \left(\frac{p}{1-\sqrt{1-p}}\right)^s (1-p)^{k/2}.$$

Therefore, for $0 < r < 1$,

$$\begin{aligned}\sum_{k=0}^{\infty} p_k^{1-r} &\leqslant \left(\frac{p}{1-\sqrt{1-p}}\right)^{s(1-r)} \sum_{k=0}^{\infty} (1-p)^{k(1-r)/2} \\ &= \left(\frac{p}{1-\sqrt{1-p}}\right)^{s(1-r)} \frac{1}{1-(1-p)^{(1-r)/2}}. \quad \blacksquare\end{aligned}$$

4. UNIFORM CONCENTRATION INEQUALITY FOR CATEGORICAL VARIABLES WITH A FINITE NUMBER OF LEVELS

The same technique used in the proof of Theorem 2.1 can be applied to the case of categorical variables with a finite number of levels to obtain a uniform concentration inequality with the same optimal rate as in [16], but with a much simpler proof. Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be independent categorical variables with K categories and $p_{ik} = P(z_{ik} = 1)$ for $i = 1, \ldots, n$, $k = 1, \ldots, K$, and $\mathbf{p}_i = (p_{i1}, \ldots, p_{iK})$ for $i = 1, \ldots, n$. The entropy of $\mathbf{z}_i$ is defined as $-\sum_{k=1}^K p_{ik} \log p_{ik}$. Finally, let $\mathcal{C} = \{\mathbf{q} = (q_1, \ldots, q_K) : 0 < q_k < 1, k = 1, \ldots, K, \sum_{k=1}^K q_k = 1\}$ be the constraint on $\mathbf{p}_1, \ldots, \mathbf{p}_n$. We have the following uniform concentration inequalities:

THEOREM 4.1. *For $2 \leqslant K \leqslant 7$ and all $\epsilon > 0$,*

$$\sup_{\mathbf{p}_1,\ldots,\mathbf{p}_n \in \mathcal{C}} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K (z_{ik} - p_{ik}) \log p_{ik}\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{n\epsilon^2}{2K/\sqrt{\pi} + 2\epsilon}\right).$$

For $K \geqslant 8$ and all $\epsilon > 0$,

$$\begin{aligned}&\sup_{\mathbf{p}_1,\ldots,\mathbf{p}_n \in \mathcal{C}} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K (z_{ik} - p_{ik}) \log p_{ik}\right| \geqslant \epsilon\right) \\ &\qquad \leqslant 2\exp\left(-\frac{n\epsilon^2}{e^2(\log K)^2/(2\sqrt{\pi}) + \epsilon \log K}\right).\end{aligned}$$

Proof. Let $Y_i = \sum_{k=1}^{K}(z_{ik} - p_{ik})\log p_{ik}$. Similar to the proof of Theorem 2.1, for $0 < r \leqslant 1$ and $|\lambda| < r$,

$$\begin{aligned}\log M_{Y_i}(\lambda) &= \log\Big(\sum_{k=1}^{K} p_{ik}^{\lambda+1}\Big) - \lambda\sum_{k=1}^{K} p_{ik}\log p_{ik}\\ &\leqslant \sum_{k=1}^{K}\sum_{m=2}^{\infty} p_{ik}^{1-r}\left(\frac{|\lambda|}{r}\right)^m \frac{1}{2\sqrt{\pi}} = \Big(\sum_{k=1}^{K} p_{ik}^{1-r}\Big)\frac{\lambda^2}{r^2}\,\frac{1}{1-|\lambda|/r}\,\frac{1}{2\sqrt{\pi}}.\end{aligned}$$

Since p_{ik}^{1-r} is a concave function of p_{ik} for $0 < r \leqslant 1$, by Jensen's inequality,

$$\sum_{k=1}^{K} p_{ik}^{1-r} = K\frac{\sum_{k=1}^{K} p_{ik}^{1-r}}{K} \leqslant K\left(\frac{\sum_{k=1}^{K} p_{ik}}{K}\right)^{1-r} = K^r.$$

Therefore, for $0 < r \leqslant 1$ and $|\lambda| < r$,

$$M_{Y_i}(\lambda) \leqslant \exp\left(K^r\frac{\lambda^2}{r^2}\,\frac{1}{1-|\lambda|/r}\,\frac{1}{2\sqrt{\pi}}\right).$$

Similar to the proof of Theorem 2.1,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{n\epsilon^2}{2K^r/(\sqrt{\pi}r^2) + 2\epsilon/r}\right) \quad \text{for } 0 < r \leqslant 1.$$

Finally, we pick r that minimizes K^r/r^2 over $r \in (0, 1]$. For $2 \leqslant K \leqslant 7$, we take $r = 1$, which gives

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{n\epsilon^2}{2K/\sqrt{\pi} + 2\epsilon}\right).$$

For $K \geqslant 8$, we take $r = 2/\log K < 1$, which gives

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{n\epsilon^2}{e^2(\log K)^2/(2\sqrt{\pi}) + \epsilon\log K}\right). \quad ■$$

REMARK 4.1. In [16] we proved that for sufficiently small positive ϵ and $K \geqslant 5$,

$$\sup_{\mathbf{p}_1,\ldots,\mathbf{p}_n \in \mathcal{C}} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(z_{ik} - p_{ik})\log p_{ik}\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{n\epsilon^2}{4(\log K)^2}\right),$$

and the rate $(\log K)^2/n = o(1)$ is optimal. Theorem 2.1 achieves the same optimal rate with a better constant.

Acknowledgments. We thank the editor, the associate editor, and the anonymous reviewer for their valuable feedback and constructive comments.

REFERENCES

[1] A. Antos and I. Kontoyiannis, *Convergence properties of functional estimates for discrete distributions*, Random Structures Algorithms 19 (2001), 163–193.

[2] V. Baccetti and M. Visser, *Infinite Shannon entropy*, J. Statist. Mech. Theory Exp. 4 (2013), art. P04010, 12 pp.

[3] J. Beirlant, E. J. Dudewicz, L. Györfi, E. C. Van der Meulen, *Nonparametric entropy estimation: An overview*, Int. J. Math. Math. Sci. 6 (1997), 17–39.

[4] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, *Stochastic blockmodels with a growing number of classes*, Biometrika 99 (2012), 273–284.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley-Interscience, Hoboken, NJ, 2006.

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Cambridge Univ. Press, Cambridge, 2011.

[7] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*, Cambridge Univ. Press, Cambridge, 2009.

[8] Y. Li and B. Tian, *Optimal non-asymptotic concentration of centered empirical relative entropy in the high-dimensional regime*, Statist. Probab. Lett. 197 (2023), art. 109803, 5 pp.

[9] M. Raginsky and I. Sason, *Concentration of measure inequalities in information theory, communications and coding*, arXiv:1212.4663 (2012).

[10] Z. Ren, *Optimal distribution-free concentration for the log-likelihood function of Bernoulli variables*, J. Inequal. Appl. 2023, art. 81, 11 pp.

[11] H. Robbins, *A remark on Stirling's formula*, Amer. Math. Monthly 62 (1955), 26–29.

[12] C. E. Shannon, *A mathematical theory of communication*, Bell Labs Tech. J. 27 (1948), 379–423.

[13] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Univ. Press, Cambridge, 2018.

[14] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Univ. Press, Cambridge, 2019.

[15] Y. Zhao, *A note on new Bernstein-type inequalities for the log-likelihood function of Bernoulli variables*, Statist. Probab. Lett. 163 (2020), art. 108779, 5 pp.

[16] Y. Zhao, *An optimal uniform concentration inequality for discrete entropies on finite alphabets in the high-dimensional setting*, Bernoulli 28 (2022), 1892–1911.

Yunpeng Zhao
Department of Statistics
Colorado State University
Fort Collins, CO 80521, USA
E-mail: Yunpeng.Zhao@colostate.edu

Received 1.2.2024;
accepted 16.4.2024