
DEPARTMENT OF STATISTICS

University of Wisconsin
Madison, Wisconsin

TECHNICAL REPORT NO. 620

October 1980

A NOTE: ON AN UPPER BOUND OF SANOV INEQUALITY FOR THE PROBABILITY OF LARGE DEVIATION

by

James C. Fu

A Note:
ON AN UPPER BOUND OF SANOV INEQUALITY FOR THE PROBABILITY OF LARGE DEVIATION

James C. Fu

Department of Statistics, University of Wisconsin
and
Department of Statistics, University of Manitoba

ABSTRACT

In this paper, a sharp upper bound of Sanov inequality for the probability of large deviations under the multinomial distribution is obtained. An example of computing the upper bound for the probability of misclassification in discriminant analysis is also given.

1. Introduction

There are many probability and statistics problems involving computation (or approximation) of the tail probability or its error bound; for instance, to find the probability of misclassification, to find the probability of type I error for testing hypothesis with fixed critical region, and to find the rate of convergence for the probability of a consistent estimator in an ϵ-neighbourhood of the parameter θ, etc. These were usually studied under the name-large deviation probabilities. There are two fundamental tools for finding the upper bounds for these probabilities. One is the so-called Bernstain-Chernoff-Bahadur inequality (Chernoff [4] and Bahadur [1]) which gives an upper bound for the tail probability of a sequences of sums of independent identically distributed (i.i.d.) random variables. It can be stated as follows; for a sequence of i.i.d. non-degenerate random variables with mean zero and moment generating function $\phi_X(t)$ that exists in a neighbourhood of $t = 0$, then for any $\epsilon > 0$

$$P(X_1 + \dots + X_n \geq n\epsilon) \leq (\rho(\epsilon))^n, \quad \text{for all } n, \tag{1.1}$$

where $\rho(\epsilon) = \inf_t\{e^{-\epsilon t}\phi_X(t): t \geq 0\}$. Another one is the Sanov [6]) which gives an upper bound for the probability of multinomial distribution. The Sanov inequality can be stated as follows; let $S = \{x = (x_1,\dots,x_k): 0 \leq x_i \leq 1, i = 1,\dots,k, \text{ and } \sum_{i=1}^k x_i \equiv 1\}$ be a subset of a k-dimensional vector space, and for each n let $p_n = (n_1/n,\dots,n_k/n)$ be the vector of relative frequencies of n observations from a multinomial distribution with the parameper $p = (p_1,\dots,p_k)$ belonging to S if A is a compact subset in S and $p \notin A$ then

$$P(p_n \epsilon A|p) \leq (n+1)^k e^{-nK(A,p)}, \quad \text{for all } n, \tag{1.2}$$

where $K(A,p) = \inf_q\{K(q,p): q\epsilon A\}$ and $K(q,p)$ is the Kullback-Liebler information of q with respect to p defined as

$$K(q,p) = \sum_{i=1}^k q_i \log \frac{q_i}{p_i}. \tag{1.3}$$

Recently, Bartfai [3] and Bahadur and Zabell [2] have generalized the Bernstain-Chernoff-Bahadur inequality for the sum of i.i.d. random vectors. Based on the fact that p_n can be represented as a sum of i.i.d. random vectors and applying the above result, the Sanov inequality (1.2) can be reduced to

$$P(p_n \epsilon A|p) \leq e^{-nK(A,p)}, \quad \text{for all } n. \tag{1.4}$$

The upper bound given by the above inequality is better than the upper bound given by (1.2). For a special case $k = 2$, Glick [5, page 246] gave a very interesting result which shows that the probability of misclassification in discriminant analysis satisfies the following inequality

$$(1.5)\qquad P(p_n \in A|p) \le \frac{1}{2}\, e^{-nK(A,p)}, \qquad \text{for all } n,$$

where $K(A,p) = -\log[1-|\sqrt{p_1} - \sqrt{p_2}|^2]$. Clearly, the upper bound given by (1.5) is better than the upper bounds given by (1.2) and (1.4). For the case $k > 2$, the author points out that the direct extension of the method used in his paper seems difficult.

After observing the above inequalities (1.2), (1.4), and (1.5) one would naturally expect a more general result ($k \ge 2$) that there exists $0 \le L(A,n) \le 1$ ($L(A,n)$ may depend on the subset A and the sample size n) such that the following inequality holds;

$$(1.6)\qquad P(p_n \in A\, p) \le L(A,n)\, e^{-nK(A,p)}, \qquad \text{for all } n.$$

In section 2, we prove this. In section 3, an application is given. It shows that Glick's result is a very special case of the general result (1.6).

2. The Main Theorem

Let A be a compact subset in S and $p \notin A$. The subset A is defined to be K-regular with respect to (w.r.t.) p if there exists a point $p^* \in A$ such that

$$(2.1)\qquad \sum_{i=1}^{k} (x_i - p_i^*) \log \frac{p_i^*}{p_i} \ge 0,$$

for all $x \in A$. Let A(K) be the subset in A which contains all the points p^* and satisfies the inequality (2.1); i.e. $A(K) = \{p^*: p^* \in A$ and $\sum_{i=1}^{k} (x_i - p_i^*) \log(p_i^*|p_i) \ge 0\ \ \forall x \in A\}$. To prove our main result, we first prove the following lemmas.

Lemma 1. For given $y \in A$,

$$(2.2)\qquad \sup_{x \in A} \sum_{i=1}^{k} y_i \log x_i = \sum_{i=1}^{k} y_i \log y_i .$$

Proof: For any x any y, the Kullback-Liebler information of y with respect to x is

$$(2.3)\qquad K(y,x) = \sum_{i=1}^{k} y_i \log y_i - \sum_{i=1}^{k} y_i \log x_i \ge 0.$$

Since $y \in A$ and the above inequality holds for all $x \in A$, the result follows directly.

Lemma 2. For any x and y,

$$(2.4)\qquad \frac{1}{n} \log \frac{P(p_n|x)}{P(p_n|y)} = \sum_{i=1}^{k} \frac{n_i}{n} \log \frac{x_i}{y_i}, \quad \text{for every } n.$$

Proof: If n observations are taken from a multinomial distribution having parameter $x \in A$, then the probability of observing p_n is

$$(2.5)\qquad P(p_n|x) = n! \prod_{i=1}^{k} (x_i^{n_i}/(n_i!)).$$

The result follows immediately from above equality.

Lemma 3. If A is K-regular with respect to p then for every p^* belonging to A(K)

$$(2.6)\qquad K(p^*,p) = K(A,p).$$

<u>Proof:</u> Since A is K-regular with respect to p, it follows for every given $p^* \epsilon A(K)$ we have

$$(2.7) \qquad \sum_{i=1}^{k} p_i^* \log \frac{p_i^*}{p_i} \leq \sum_{i=1}^{k} x_i \log \frac{p_i^*}{p_i}, \quad \text{for all } x \epsilon A.$$

By using Lemma 1, it follows

$$(2.8) \qquad \sum_{i=1}^{k} x_i \log \frac{p_i^*}{p_i} \leq \sum_{i=1}^{k} x_i \log \frac{x_i}{p_i}, \quad \text{for all } x \epsilon A(K).$$

Again, it follows from the inequalities (2.7) and (2.8) that

$$(2.9) \qquad \sum_{i=1}^{k} p_i^* \log \frac{p_i^*}{p_i} \leq \sum_{i=1}^{k} x_i \log \frac{p_i^*}{p_i} \leq \sum_{i=1}^{k} x_i \log \frac{x_i}{p_i}, \quad \text{for all } x \epsilon A(K).$$

Since A is compact, the result follows from the definition of $K(A,p)$ and the above inequality (2.9) directly.

<u>Lemma 4.</u> If A is K-regular with respect to p then there exists a unique p^* which satisfies the inequality (2.1) (i.e. $A(K)$ contains only one point) and $K(p^*,p) = K(A,p)$.

<u>Proof:</u> Since A is K-regular with respect to p the second part of the Lemma follows immediately from Lemma 3, hence we need only to prove the uniqueness. Suppose there are two points, say p^* and p^{**} ($p^* \neq p^{**}$), both belonging to $A(K)$. Since both p^* and p^{**} belong to $A(K)$ it follows from Lemma 3 that

$$(2.10) \qquad K(p^*,p) = K(p^{**},p) = K(A,p).$$

It follows from (2.1) that

$$(2.11) \qquad K(p^{**},p) \geq K(p^{**},p^*) + K(p^*,p).$$

The facts $K(p^*,p) = K(p^{**},p)$ and $K(p^{**},p^*) > 0$ together shows that the inequality (2.11) is impossible. This completes the proof of uniqueness.

<u>Theorem:</u> If the subset A is K-regular with respect to p, and $p \not\epsilon A$ then

$$(2.12) \qquad P(p_n \epsilon A|p) \leq L(A,n)\, e^{-nK(A,p)}, \quad \text{for all } n,$$

where $0 \leq L(A,n) = \sum_{p_n \epsilon A} P(p_n\ p^*) \leq 1$.

<u>Proof:</u> Let p^* be the point satisfying the condition (2.1). By Lemmas 3 and 4 p^* is unique and $K(p^*,p) = K(A,p)$. For given n and $p(p \not\epsilon A)$, we have

$$\alpha_n = P(p_n\ A|p) = \sum_{p_n \epsilon A} P(p_n|p),$$

$$= \sum_{p_n \epsilon A} \frac{P(p_n|p)}{P(p_n|p^*)} P(p_n|p^*)$$

$$= \sum_{p_n \epsilon A} e^{n \cdot \frac{1}{n} \log \frac{P(p_n|p)}{P(p_n|p^*)}} P(p_n|p^*)$$

$$\leq e^{n \sup_{p_n \epsilon A} \frac{1}{n} \log \frac{P(p_n|p)}{P(p_n|p^*)}} \sum_{p_n \epsilon A} P(p_n|p^*).$$

It follows from Lemma 2, that

$$\alpha_n \le e^{-n \inf_{p_n \epsilon A} \sum_{i=1}^{k} \frac{n_i}{n} \log \frac{p_i^*}{p_i}} \sum_{p_n \epsilon A} P(p_n|p^*) .$$

Since A is K-regular with respect to p, it follows from (2.7) that

$$\alpha_n \le e^{-n \sum_{i=1}^{k} p_i^* \log \frac{p_i^*}{p_i}} \sum_{p_n \epsilon A} P(p_n|p^*).$$

By Lemma 3, it follows

$$\alpha_n \le e^{-nK(A,p)} \sum_{p_n \epsilon A} P(p_n|p^*).$$

This completes the proof.

The condition (2.1) is a simple condition. In a typical statistical problem it is satisfied. Lemmas 3 and 4 suggest a direct method for checking the K regularity of the subset A with respect to p. The method can be stated as follows: (a) Use the Lagrange multiplier technique to find the unique point p* such that $K(p^*,p) = K(A,p)$, (b) Construct a hyperplane

$$\text{(2.13)} \qquad L: \sum_{i=1}^{k} a_i x_i = a_o,$$

where $a_o = K(p^*,p)$, $a_i = \log \frac{p_i^*}{p_i}$, $i=1,2,\dots,k$, the hyperplane L will cut the hyperplane S into two parts, say

$$\text{(2.14)} \qquad S(L^+) = \{x: x \epsilon S \text{ and } \sum_{i=1}^{k} a_i x_i \ge a_0\}$$

$$\text{(2.15)} \qquad \text{and} \quad S(L^-) = \{x: x \epsilon S \text{ and } \sum_{i=1}^{k} a_i x_i < a_0\},$$

(c) if A belongs to the subset $S(L^+)$ then A is K-regular with respect to p.

It is worthwhile to point out that the hyperplane L constructs above is unique and passes through the point p*. The subset A does not have to be convex. The results holds as long as the set A contains p* and is a subset of $S(L^+)$.

Two subsets A_1 and A_2 are defined to be K-equivalent with respect to p if (a) both A_1 and A_2 are K-regular with respect to p, and (b) $K(A_1,p) = K(A_2,p)$. Let A_j: $j=1,2,\dots,m$ be a family of finite subsets which are K-equivalent with respect to p, and $A = \bigcup_{j=1}^{m} A_j$. The main Theorem given above can be extended to the following general case: if $p \not\epsilon A$, then

$$\text{(2.16)} \qquad P(p_n \epsilon A|p) \le mL(A,n)e^{-nK(A,p)},$$

where

$$\text{(2.17)} \qquad L(A,n) = \max_{1 \le j \le m} \sum_{p_n \epsilon A_j} P(p_n|p^*(j))$$

and $K(p^*(j),p) = K(A_j,p) = K(A,p)$, $j=1,\dots,m$.

Under the condition (2.1), the upper bound given by (2.12) is stronger than the upper bound given by Sonov [6], Bartfai [3], and Bahadur and Zabell [2]. Compared to the methods used in [6] our proof is elementary and simple.

3. An Example

To illustrate the theorem, we give an example for the case of binomial distribution (k=2). We take k=2 for two reasons. First we can compare our result with Glick's result. Secondly the computation is simpler and provided a clearer illustration of the method.

Let $S = \{x = (x_1,x_2): \ 0 \le x_1,x_2 \le 1 \text{ and } x_1 + x_2 = 1\}$,

$p = (p_1,p_2)$ belongs to S, $p_1 < \frac{1}{2}$, and $A = \{x: x_1 \ge \frac{1}{2}\}$. Clearly the subset A is compact and $p \notin A$. The Kullback-Liebler information of p' with respect to p is

$$K(p',p) = p_1' \log \frac{p_1'}{p_1} + p_2' \log \frac{p_2'}{p_2} .$$

By using the Lagrange multiplier method to minimize the quantity $K(p',p)$ subject to the condition $p' \in A$, we have unique $p^* = (\frac{1}{2},\frac{1}{2})$ and

$$(3.1) \qquad K(p^*,p) = K(A,p) = \frac{1}{2}\log\frac{1}{2p_1} + \frac{1}{2}\log\frac{1}{2p_2}$$

$$= -\log[1-|\sqrt{p_1} - \sqrt{p_2}|^2] .$$

The hyperplane L induced by p* and p is given as

$$(3.2) \qquad L: \quad x_1 \log\frac{1}{2p_1} + x_2\log\frac{1}{2p_2} = \frac{1}{2}\log\frac{1}{2p_1} + \frac{1}{2}\log\frac{1}{2p_1} .$$

Thy hyperplane L cuts the hyperplane S into two parts

$$(3.3) \qquad S(L^+) = \{x = (x_1,x_2): \ x \in S, \ x_1 \ge \tfrac{1}{2}\} , \text{ and}$$

$$(3.4) \qquad S(L^-) = \{x = (x_1,x_2): \ x \in S, \ x_1 < \tfrac{1}{2}\} .$$

The subset A belongs to $S(L^+)$ hence the A is K-regular with respect to p. It follows that

$$(3.5) \qquad L(A,n) = \sum_{p_n \in A} P(p_n|(\tfrac{1}{2},\tfrac{1}{2}))$$

$$= \sum_{n_1 \ge n/2} \binom{n}{n_1}(\tfrac{1}{2})^n$$

$$= \begin{cases} \frac{1}{2} & \text{if n is odd} \\ \sum_{n_1 \ge \frac{n}{2}} \binom{n}{n_1}(\frac{1}{2})^n & \text{if n is even,} \end{cases}$$

and

$$(3.6) \qquad P(p_n \in A|p) = \begin{cases} \frac{1}{2}(1-|\sqrt{p_1} - \sqrt{p_2}|^2)^n & \text{if n is odd,} \\ \sum_{n_1 \ge \frac{n}{2}} \binom{n}{n_1}(\frac{1}{2})^n (1-|\sqrt{p_1} - \sqrt{p_2}|^2)^n & \text{if n is even.} \end{cases}$$

The subset considered by Glick [5] is slightly different from the subset A considered here. Glick [5, page 245] defines

$$(3.7) \qquad A(n) = \begin{cases} \{x: x_1 \ge \frac{1}{2}\} & \text{if n is odd,} \\ \{x: x_1 > \frac{1}{2} & \text{plus with probability } \frac{1}{2} \text{ of including the point } (\frac{1}{2},\frac{1}{2})\} \text{ if n is even.} \end{cases}$$

The sets A and B differ only for even n where A contains the point $(\frac{1}{2},\frac{1}{2})$ with A(n) containing the point $(\frac{1}{2},\frac{1}{2})$ only with probability $\frac{1}{2}$. Since for even n

$$\frac{1}{2}\binom{n}{\frac{n}{2}}\left(\frac{1}{2}\right)^n + \sum_{n_1 > \frac{n}{2}} \binom{n}{n_1}\left(\frac{1}{2}\right)^n \equiv \frac{1}{2}, \tag{3.8}$$

we have $L(A(n),n) \equiv \frac{1}{2}$ for both n odd or even. Therefore

$$P(p_n \epsilon A(n)|p) \le \frac{1}{2}(1-|\sqrt{p_1} - \sqrt{p_2}|^2)^n . \tag{3.9}$$

It is clear that Glick's result is a very special case of our general Theorem. Our result also applies in the case $k > 2$ (see Steineback [7]).

4. Acknowledgment

The author wishes to thank Professor J. Steineback for introducing this problem and for his stimulating and fruitful discussions on this subject. The author would also like to thank Professor J. Klotz for his valuable comments.

5. References

[1] Bahadur, R.R. (1971). Some limit theorems in statistics. Regional Conferences Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia.

[2] Bahadur, R.R. and Zabell, S.L. (1979). Large deviations of the sample mean in general vector spaces. Ann. Prob. Vol. 7, No. 4, 587-621.

[3] Bartfai, P. (1978). Large deviations of the sample mean in Euclidean Spaces. Mimeograph Series No. 78-13, Statist. Depart., Purdue Univ.

[4] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on sum of observations. Ann. Math. Statist. 23, 493-502.

[5] Glick, N. (1973). Sample-based Multinomial Classification. Biometrics 29, 241-256.

[6] Sanov, I.N. (1957). On the probability of large deviations of random variables. Sel. Trans. Math. Statist. Prob. 1, 213-244.

[7] Steinebach, J. (1980). (Personal Communication).