Large deviations of the sample mean in Euclidean spaces

by

P. Bártfai
Mathematical Institute of the
Hungarian Academy of Sciences

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series No. 78-13

Large deviations of the sample mean in Euclidean spaces

by P. Bartfai

Summary.

Set $\frac{S_n}{n}$ be the sample mean of i.i.d. random k-dimensional vectors and A an open k-dimensional set. Among the questions concerning whether the expression $(P(\frac{S_n}{n} \in A))^{\frac{1}{n}}$ has a limit and what is this limit, the lower estimation is completely solved, while the upper one is partly open. Our aim is to give a view on the solved cases of the upper estimation and to show the methods used here.

1. Introduction. One-dimensional case.

1.1. In connection with any limit theorem stating that $G_n(x) \to G(x)$, where G_n and G are distribution functions, one can formulate a large deviation problem as well. If $G_n(x) \to G(x)$ and $x_n \to +\infty$ then obviously $G_n(x_n) \to 1$ and $G_n(-x_n) \to 0$, but the next question in this direction is what is the rate of this convergence. The large deviation problem is to find the asymptotic behavior of the expressions $1-G_n(x_n)$ and $G_n(-x_n)$ if $n \to \infty$. These expressions, in most cases, are symmetrical and therefore, usually, only the first one is investigated.

The large deviation problem arising connected with the central limit theorem is the one most frequently discussed (for other types see e.g. [4], [14]). Here we deal only with the i.i.d. case, when the r.v.'s $X_1, X_2, \dots,$ are independent and identically distributed r.v.'s $S_n = X_1+X_2+\dots X_n$, $F_n(x) = P(S_n < x)$ and $G_n(x) = F_n(\sqrt{n}x)$ ($E(X_1) = 0$ is assumed for the time being). General results for the non-i.i.d. case are in Sievers [17], Steinebach [19].

1.2. There exists a natural classification of the large deviation problems according to the order of magnitude of x_n.

The case $x_n = o(\sqrt{n})$ was solved by Cramèr [8]. If the moment generating function $R(t) = E(e^{tX_1}) < \infty$ for some $t > 0$, $E(X_1) = 0$, $D(X_1) = 1$, $x_n > 0$ then

$$(1.1) \qquad 1 - F_n(\sqrt{n}x_n) = (1 - \Phi(x_n))e^{\frac{x_n^2}{2}} \rho^n\left(\frac{x_n}{\sqrt{n}}\right)\left(1 + O\left(\frac{x_n+1}{\sqrt{n}}\right)\right) \quad (n \to \infty),$$

where

$$(1.2) \qquad \rho(x) = \inf_t e^{-tx}R(t).$$

The behavior of the other tail is similar assuming that $R(t) < \infty$ for some $t < 0$. As a refinement of Cramèr's theorem there are results when $R(t)$ is replaced by some moment condition (Nagaev [11]). In this case $\log \rho$ is replaced by a partial sum of its power series with a remainder term.

If the deviation is not large, to be exact if its order of magnitude is less than $c\sqrt{\log n}$, then the existence of the kth moment guarantees the validity of (1.1) which has the form now $1 - F_n(\sqrt{n}x_n) \sim 1 - \phi(x_n)$, where k depends on c, (Rubin-Sethuraman [16]). Deviations of this type are called moderate deviations. Osipov [12] gave the exact condition of the validity (1.1), if the order of magnitude of x_n is known and less than $o(\sqrt{n})$.

The cases mentioned above have the important property that the statement (1.1) uses only the values of ρ in a neighborhood of the expectation (which is now the origin). If the deviation is somewhat larger, $x_n \sim c\sqrt{n}$ or $\frac{x_n}{\sqrt{n}} \to +\infty$; then this no longer remains true. Therefore if we deal with large deviations of the sample mean, when $x_n = x\sqrt{n}$, and x takes different values, we have to use the complete function $\rho(x)$. This argument explains why the methods of this so-called Chernoff case differ from those of the earlier Cramèr case.

For "very large" deviations ($\frac{x_n}{\sqrt{n}} \to \infty$) there are only a few and very special results (see Linnik [10]).

1.2. The large deviations of the sample mean represent a boundary case in the above sense, and it is not necessary to emphasize their importance. The investigations are mostly done differently from the Cramèr case.

Let us assume that $E(X_1) = m$ (not necessarily $m = 0$), then the Chernoff theorem [7] states that

$$(1.3) \qquad (P(\frac{S_n}{n} \geq x))^{\frac{1}{n}} \to \rho(x) \qquad (x \geq m),$$

and if $R(t) < \infty$ for some $t > 0$, then the convergence uniform in $[m_1, +\infty)$. Complete this result by a very often used inequality which takes the role of the classical Bernstein inequality and therefore it is called by the same name:

$$(1.4) \qquad P(\frac{S_n}{n} \geq x) \leq \rho^n(x) \qquad (x \geq m).$$

There exists a sharpened version of the Bernstein inequality, namely

$$P(\sup_{1 \leq k \leq n} S_k \geq x) \leq \rho^n(x)$$

(see Steiger [18]) and the statement (1.3) can be similarly modified.

The function $\rho(x)$ defined by (1.2) is called the Chernoff function. Obviously $0 \leq \rho(x) \leq 1$. The Chernoff function has always the following shape (assuming now that $R(t) < \infty$ for $|t| < \delta$): there exists an open (finite or infinite) interval around the expectation point in which $\rho(x)$ is an analytic function, it will be called the domain of analyticity, and outside of its closure $\rho(x) = 0$ or $\rho(x)$ is an exponential function. In the first case the endpoint of the interval is not necessarily a point of continuity of $\rho(x)$. The proof of the Chernoff theorem in the domain of analyticity can be carried out in a manner similar to Cramèr's while the extension always uses an approximation (or truncation). As to the complete proof we refer to Bahadur [1], for the approximations the general theorem of Bartfai [4] (Theorem 10-11) is very useful.

In the domain of analyticity Bahadur and Rao [2] gave a very often used sharper asymptotic formula.

$$P\left(\frac{S_n}{n} \geq x\right) \sim (2\pi n)^{-\frac{1}{2}} b_n(x)\rho^n(x)$$

where $|\log b_n(x)|$ is a bounded expression for a fixed x and in the case when the distribution is not of lattice type $b_n(x)$ is independent of n (see also Petrov [13]).

1.3. The only method of the theory of large deviations is (excepting only one case: Sievers [17]) the method of conjugate distributions (or exponential families). To any distribution function F(x) and to any t , for which $R(t) < \infty$, we can assign another distribution function

$$\bar{F}(x) = R^{-1}(t) \int_{-\infty}^{x} e^{ty} dF(y). \tag{1.5}$$

This transformation has a very important property, namely $(\overline{F_n}) = (\bar{F})_n$, where the lower index n denotes the nth convolution power. From the from of the moment generating function of $\bar{F}$, which is equal to $\frac{R(t+s)}{R(t)}$ (t is fixed now), we can see that this transformation shifts the coordinate system of the function

R(t) and then renorms it to be again a moment generating function.

The transformation (1.5) has a very simple inversion formula:

$$(1.6) \qquad 1 - F(x) = R(t) \int_x^\infty e^{-ty} d\bar{F}(y)$$

and the t value here must be the same as in (1.5), in this case (1.6) is an (almost trivial) identity.

Apply (1.6) to F_n, then

$$(1.7) \quad 1 - F_n(nx) = R^n(t) \int_{nx}^\infty e^{-ty} d\bar{F}_n(y) = (e^{-tx}R(t))^n \int_x^\infty e^{-nt(y-x)} d\bar{F}_n(ny)$$

and by choosing t in an appropriate manner the factor before the integration is equal (or nearly equal) to $\rho^n(x)$. In evaluating the integral the fact is often used that $\bar{F}_n(\sqrt{n}y)$ is nearly normal. This is the basic formula of the theory of large deviations.

The Bernstein inequality can be read directly from (1.7) because for $t > 0$ the integrand is less than 1, and $x \geq m$ implies that

$$(1.8) \qquad \sup_{t \geq 0} e^{-tx}R(t) = \sup_t e^{-tx}R(t) = \rho(x),$$

namely $e^{-tm} R(t) \geq 1$ for all t and for $t < 0$ $e^{-tx}R(t) = e^{-tm}R(t)e^{t(m-x)} \geq 1$.

2. Introduction. Multi-dimensional case.

2.1. The techniques used in the one-dimensional case works in this case, too. We can define the basic notions in a similar way.

Let $X_1, X_2, \ldots$ be i.i.d. random vectors from the k-dimensional Euclidean space R^k. Let the distribution function and the moment generating function of X_1 be F(x) and $R(t) = E(e^{<t,x>})$, resp. ($x \in R^k$, $t \in R^k$ and $<t,x>$ denotes their inner product). The function $R(t) = +\infty$ in the case when the expectation does not exist.

Define the Chernoff function of X_1 (or F) by

$$\rho(x) = \inf_{t \in R_k} e^{-<t,x>} R(t) \qquad (x \in R^k). \tag{2.1}$$

The conjugate distributions can be defined in a similar way and the inversion formula as well as the basic formula also hold in multi-dimensional spaces.

2.2. In order to investigate the properties of the Chernoff function there is a very useful aid, the theory of convex analysis. For this we refer to the book of Rockafellar [15].

A function $f(x)$ $(x \in R^k)$ is said to be convex if the set $\{(x,y): y \geq f(x),\ x \in R^k,\ y \in R^1\}$ is convex and $f(x)$ is closed if this set is closed. There is a duality (Fenchel-duality) among the closed convex functions:

$$\bar{f}(x) = \sup_t \,(<t,x> - f(t))$$

and

$$\bar{\bar{f}}(x) = f(x)$$

in words $\bar{f}$ is the convex conjugate of f.

Using this notion we can establish that the functions log R(t) and -log $\rho(x)$ are convex conjugate each to other. In fact, it is well-known that log R(t) is a convex function and it is easy to prove that R(t) is closed, because - by truncation - it can be approximated from below by continuous functions. Consequently - log $\rho(x)$ is a closed convex function, too and from the duality we can deduce the inversion formula:

$$R(t) = \sup_{x \in R^k} e^{<t,x>} \rho(x). \tag{2.2}$$

2.3. Define the effective domain of $\rho(x)$ by $\mathrm{dom}\,\rho = \{x: \rho(x) > 0\}$.

Theorem 1. The closure of dom ρ is equal to the convex hull of the support of the distribution of X_1.

(The support of the distribution of X_1 is the smallest closed set S such that $P(X_1 \in S) = 1$.)

Proof. Define the recession function of a convex function f by

$$f^{0+}(y) = \lim_{\lambda\to\infty} \frac{1}{\lambda} f(x + \lambda y)$$

where x is an arbitrary point for which $|f(x)| < \infty$. The limit is independent of the choice of x.

Let X be a random variable with moment generating function $R_1(t)$, then by a simple calculation

$$\lim_{t\to\infty} \frac{1}{t} \log R_1(t) = \text{ess sup } X.$$

Consider for any fixed $y \in R^k$ the function $R_1(\lambda) = R(\lambda y)$, then $R_1(\lambda)$ is the moment generating function of $<X_1, y>$. By denoting $f(t) = \log R(t)$

$$f^{0+}(y) = \lim_{\lambda\to\infty} \frac{1}{\lambda} \log R(\lambda y) = \text{ess sup} <X_1, y> = \sup_{x\in S} <x,y>.$$

For any closed set S we can define a support function

$$\bar{\delta}_S(t) = \sup_x\{<x,t>: \ x \in S\}.$$

The support function of S is equal to that of conv S = C (conv S is the convex hull of S). Therefore $f^{0+}(y) = \bar{\delta}_C(y)$.

According to Theorem 13.3 of [15] $f^{0+}(y) = \bar{\delta}_{\text{dom } \bar{f}} = \bar{\delta}_{\text{cl dom } \rho}$ and, since $\bar{\delta}_C$ and the indicator funciton of C,

$$\delta_C(x) = \begin{cases} 1 & x \in C, \\ +\infty & x \notin C, \end{cases}$$

are convex conjugate (Theorem 13.2 of [15]), the equality $\bar{\delta}_C = \bar{\delta}_{\text{cl dom } \rho}$ implies that $C = \text{cl dom } \rho$.

2.4. It is easy to verify that applying a non-singular linear transformation to X, say $X' = Ax + b$ ($\det A \neq 0$) the transformed Chernoff function is

$$\rho_{X'}(x) = \rho_X(A^{-1}(x-b)).$$

In particular, if we represent X and x in the same space, the Chernoff function $\rho(x)$ of X is independent of the choice of the coordinate system.

The situation is not so simple if $\det A = 0$. Let us consider for example a projection to the subspace determined by the first k' ($k' < k$) coordinates. The moment generating function of the projected distribution is $R(t_1,\ldots,t_{k'},0,\ldots,0)$. According to (2.2),

$$R(t_1,\ldots,t_k) = \sup_{x_1,\ldots,x_k} e^{t_1x_1+\ldots+t_kx_k}\rho(x_1,\ldots,x_k),$$

hence

$$R(t_1,\ldots,t_{k'},0,\ldots,0) = \sup_{x_1,\ldots,x_{k'}} e^{t_1x_1+\ldots+t_{k'}x_{k'}} \sup_{x_{k'+1},\ldots,x_k} \rho(x_1,\ldots,x_k).$$

The function $-\log \sup_{x_{k'+1},\ldots,x_k} \rho(x_1,\ldots,x_k)$ is a convex function again, but not necessarily closed, therefore we can state only that the Chernoff function of the projection is

$$\hat{\rho}(x_1,\ldots,x_{k'}) = \text{cl} \sup_{x_{k'+1},\ldots,x_k} \rho(x_1,\ldots,x_k). \tag{2.3}$$

(If f is a closed convex function cl f denotes the greatest closed convex function which is not greater than f. Here the closure operation is defined as $\exp\{-\mathrm{cl}(-\log \rho)\}$). We note that the closure operation can modify the function only on the boundary of its effective domain.

Example 5.1 of Bahadur - Zabell [3] shows that the closure operation, in general, cannot be omitted in (2.3). If we suppose that $R(t) < \infty$ in a neighborhood of the origin, then sup ρ is closed, therefore - in this case - the Chernoff function of the projection is $\sup\limits_{x_{k'+1},\ldots,x_k} \rho(x_1,\ldots,x_k)$.

2.5. Let H be a supporting hyperplane of dom ρ. By some heuristic arguments we can hope that $\rho(x)$ for $x \in H$ is equal to the Chernoff function of the conditional distribution of X_1 under the condition that $X_1 \in H$. But similar difficulties arise in this case, too. This statement is valid in the one-dimensional case, but if $k \geq 2$ we have to assume that $R(t) < \infty$ everywhere.

2.6. Introduce the notation

$$\rho(A) = \sup_{x \in A} \rho(x)$$

where A is an arbitrary set in R^k. The main, partly open problem concerning large deviations of the sample mean is whether the relation

$$(P(\frac{S_n}{n} \in A))^{\frac{1}{n}} \to \rho(A) \tag{2.4}$$

holds. If A is a closed set, even if the interior of A is not empty, then this relation is not valid: Example of [5] shows that we can find a closed set A for which $P(\frac{S_n}{n} \in A) = 0$ $(n = 1,2,\ldots)$ and $\rho(A) > 0$.

Statement (2.4) can be divided into two parts, a lower and an upper part. The lower part, namely that

$$\liminf \left(P\left(\frac{S_n}{n} \in A\right)\right)^{\frac{1}{n}} \geq \rho(A)$$

holds for every open set without any restrictions. Many authors have proved this statement. Lanford [9], Bahadur-Zabell [3], Bartfai [6].

The upper part seems to be more difficult, and , in its complete generality it is an open problem. There are three easily solvable cases, the first one has special importance.

Theorem 2. (Bernstein inequality). *If* A *is a convex open set*

$$P\left(\frac{S_n}{n} \in A\right) \leq \rho^n(A). \tag{2.5}$$

This statement is valid for closed sets, too, but in this case the inequality not necessarily sharp.

Proof. First we prove (2.5) for a closed convex set B instead of A. If $\rho(B) = 0$, (2.5) follows from Theorem 1, so we can assume that $0 < \rho(B) = c < 1$. The disjoint closed convex sets, B and $C = \{x: \rho(x) > c+\varepsilon\}$ can be separated by a hyperplane H and according to Sec. 2.4. we can assume that H is orthogonal to the first coordinate axis: $H = \{(x_1,x_2,\ldots,x_k): x_1 = \alpha\}$, and $\rho(x) \leq c+\varepsilon$ for $x_1 \geq \alpha$ $(x = (x_1,\ldots,x_n))$. Applying the proof of the one-dimensional Bernstein inequality (Sec. 1.3) we obtain

$$P\left(\frac{S_n}{n} \in B\right) \leq P\left(\frac{1}{n}(S_n)_1 \geq \alpha\right) \leq \left(\inf_{t \geq 0} e^{-t\alpha}R_1(t)\right)^n,$$

where $(S_n)_1$ denotes the first coordinate of S_n and R_1 is the moment generating function of the first coordinate of X_1.

Let $\rho_1(\alpha) = \inf_t e^{-t\alpha}R_1(t)$. If $\rho_1(\alpha) = \inf_{t \leq 0} e^{-t\alpha}R_1(t)$, then we have for $\alpha' < \alpha$ and $t \leq 0$

$$e^{-t\alpha}R(t) \geq \inf_{t \leq 0} e^{-t\alpha'}R(t) \geq \rho_1(\alpha'),$$

i.e. $\rho_1(\alpha) \geq \rho_1(\alpha')$, but because of the construction of H and of Sec. 2.4 the last inequality holds in the case $\alpha' \geq \alpha$ too, therefore $\rho_1(\alpha) = \sup \rho_1(\alpha') = 1$. This is impossible, consequently

$$P\left(\frac{S_n}{n} \in B\right) \leq \rho_1^n(\alpha) \leq (c+\varepsilon)^n$$

which proves the statement for closed convex sets.

Now construct a sequence $B_1 \subset B_2 \subset \ldots$ of closed convex sets such that $\bigcup_{i=1}^{\infty} B_i = A$, then for fixed n

$$P\left(\frac{S_n}{n} \in B_k\right) \to P\left(\frac{S_n}{n} \in A\right) \qquad (k \to +\infty),$$

and for every k

$$P\left(\frac{S_n}{n} \in B_k\right) \leq \rho^n(A) \qquad (n = 1,2,\ldots).$$

These two relations prove Theorem 2.

Other cases when the upper estimation does not cause any trouble are

(i) $\rho(A) = 0$ (follows from Theorem 1),

(ii) $\rho(A) = 1$ (upper estimation is trivial).

3. Elimination of the singularities.

3.1. If we fix the value $\rho(A) = c$ $(0 < c < 1)$, then-from the point of view of the upper estimation-we can choose the set A as large as possible, and further on we always assume that $A = \{x: \ \rho(x) < c\}$. This set is always open and its complement A^c is convex.

If A^c is bounded - which is true iff $R(t) < \infty$ in a neighborhood of the origin - and the condition

$$\rho(A) = \rho(\text{cl } A) \tag{3.1}$$

holds then the proof of (2.4) is very easy.

Let $A_\varepsilon = \{x: \rho(x) < c+\varepsilon\}$, then A_ε^c and cl A are disjoint closed sets, A_ε^c is bounded, therefore they can be separated by a polyhedron P such that $A \subset P^c \subset A_\varepsilon$, and the statement (2.4) holds for P^c because it can be divided into finite many convex sets. Consequently

$$\limsup \left(P\left(\frac{S_n}{n} \in A\right)\right)^{\frac{1}{n}} \leq \limsup \left(P\left(\frac{S_n}{n} \in P^c\right)\right)^{\frac{1}{n}} \leq c+\varepsilon,$$

which proves (2.4) provided that A^c is bounded and (3.1) holds.

3.2. By an example we can show that (3.1) is not always true. Let P_1 be the uniform distribution on the two-dimensional unit disc, and P_2 be the probability measure concentrated at (1,0), and form their mixture: $P = 1/2(P_1 + P_2)$. The Chernoff function of P vanishes in the outside of the disc and on the periphery but $\rho(1,0) = 1/2$, i.e. $\rho(x)$ is not continuous at this point. Let $A = \{x: \rho(x) < 1/3\}$, then A and A_ε^c cannot be separated by any polygon with finitely many sides.

This example gives the idea of the next proofs: the distribution must be decomposable into two parts of this type.

3.3. If (3.1) does not hold then there exists a point $x_0 \in$ cl A such that $\rho(x_0) > \rho(A) > 0$. This point will be called an outstanding singularities of ρ with respect to A. Obviously $x_0 \in$ bound dom ρ. The next lemma states that the outstanding singularities can be covered by finite many supporting hyperplanes of dom ρ.

Lemma 1. If the condition (3.1) is not fulfilled by an open set A, but $P(X_1 \in \text{int}\,(A_n \text{ dom } \rho)) > 0$, then there exist finite many supporting hyperplanes of dom ρ, such that the condition (3.1) is fulfilled by the Chernoff function of the conditional distribution of X_1 under the condition that X_1 does not belong to these hyperplanes.

Proof. Let us fix an open convex set $C \subset A$ with positive measure: $P(X_1 \in C) = p_0 > 0$. Then $\rho(A) \geq p_0$.

If (3.1) does not hold then there exists a point $x_1 \in$ bound A such that $\rho(x_1) > \rho(A) \geq p_0$. Since $\rho(x)$ is not continuous at x_1, $x_1 \in$ bound dom ρ. Consider a supporting hyperplane H_1 at the point x_1 and let $p_1 = P(X_1 \in H_1)$. According to Sec. 2.4 $p_1 > c$. Introduce a new probability measure P_1 by

$$P_1(B) = P(X_1 \in B | X_1 \notin H_1).$$

If the Chernoff function of P_1 does not satisfy (3.1), then there exists another supporting hyperplane H_2 for which $p_2 = P_1(H_2) > P_1(C) = \frac{p_0}{1-p_1}$, similarly to the first step.

If the procedure could be continued without limit then

$$p_n > p_0 \left[\prod_{i=1}^{n-1} (1-p_i) \right]^{-1} \qquad (n = 1,2,\ldots).$$

Hence

$$1 - p_n < \frac{1-np_0}{1-(n-1)p_0}$$

and thus the procedure must stop at step $\frac{1}{p_0} \leq n < \frac{1}{p_0} + 1$ at the latest.

Lemma 2. Let two probability measures P_1 and P_2 be given with moment generating functions R_1 and R_2 and with Chernoff functions ρ_1 and ρ_2, resp. Then for the Chernoff function ρ of the mixed distribution $pP_1 + qP_2$ $(p \geq 0,\ q \geq 0,\ p + q = 1)$ the inequality

$$\rho(x) \geq \sup_{\substack{0 \leq \lambda \leq 1 \\ \lambda y + (1-\lambda) z = x}} c_\lambda (p\rho_1(y))^\lambda (q\rho_2(y))^{1-\lambda} \tag{3.2}$$

holds where $c_\lambda = \lambda^{-\lambda}(1-\lambda)^{-1+\lambda}$.

Proof. By definition

$$\rho^n(x) = \inf_t e^{-<t,x>n}(pR_1(t)+qR_2(t))^n =$$

$$\inf_t \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (R_1(t)e^{-<tx_1>})^k (R_2(t)e^{-<t,x_2>})^{n-k},$$

hence

$$\rho^n(x) \geq \binom{n}{k} p^k q^{n-k} \rho_1^k(x_1)\rho_2^{n-k}(x_2) \tag{3.3}$$

provided that $kx_1 + (n-k)x_2 = nx$.

If

$$\sup_{\substack{\lambda \\ \lambda y+(1-\lambda)z=x}} c_\lambda (p\rho_1(y))^\lambda (q\rho_2(y))^{1-\lambda} = \hat{\rho}(x)$$

then there exists λ_0, y_0 and z_0 such that z_0 is a continuity point of ρ_2, $\lambda_0 y_0 + (1-\lambda)z_0 = x$ and

$$c_{\lambda_0}(p\rho_1(y_0))^{\lambda_0}(q\rho_2(z_0))^{1-\lambda_0} \geq \hat{\rho}(x) - \varepsilon \tag{3.4}$$

Let $x_1 = y_0$ and k_n be a sequence such that $\frac{k_n}{n} \to \lambda_0$. If x_2 is equal to the solution of $k_n x_1 + (n-k_n)x_2 = nx$, then $x_2 \to z_0$. If $n \to \infty$ and then $\varepsilon \to 0$ (3.3) and (3.4) yield $\rho(x) \geq \hat{\rho}(x)$.

3.4. Now we are able to eliminate the condition (3.1). Our result is the following

Theorem 3. If $R(t) < \infty$ in a neighborhood of the origin then (2.4) is valid.

Proof. The elimination procedure of (3.1) is based upon an induction by dimension. Since in the one-dimensional case there is no problem (there are no outstanding singularities at all) we may assume that (2.4) is valid in the k-1 dimensional case and we prove it in R^k.

Let us suppose that

$$(3.5) \qquad P(X_1 \in \operatorname{int}(A \cap \operatorname{dom} \rho)) > 0$$

then the outstanding singularities of ρ with respect to A can be covered by finite many, say ν, hyperplanes (Lemma 1). We apply another induction by ν. If $\nu = 0$ we have no singularities. Let us suppose, therefore, that (2.4) holds if the outstanding singularities can be covered by $\nu - 1$ hyperplanes and we have to prove it for ν hyperplanes.

Consider one of the supporting hyperplanes H constructed in Lemma 1. Let the set $H \cap \operatorname{dom} \rho$ be partitioned into sets B_j $(j = 1,2,\ldots,m)$ such that $\rho(B_i) - \inf_{x \in B_j} \rho(x) < \varepsilon_1$ $(j = 1,2,\ldots,m)$ and each B_i is in $H \cap A$ or in $H \cap A^c$.

Now we estimate $P(\frac{S_n}{n} \in A)$. Evidently

$$(3.6) \qquad P(\frac{S_n}{n} \in A) \leq \sum_{j=1}^{m} \sum_{\ell=0}^{n} \binom{n}{\ell} p^{\ell} q^{n-\ell} P(\frac{1}{\ell} \sum_{i=1}^{\ell} X_i' \in B_j) P(\frac{1}{n-\ell} \sum_{i=1}^{n-\ell} X_i'' \in B_{j,n,\ell}),$$

Here $X_1', X_2', \ldots$ and $X_1'', X_2'', \ldots$ are i.i.d sequences of r.v.'s, the distributions of which are equal to the conditional distributions of X_1 under the conditions $X_1 \in H$ and $X_1 \notin H$, resp., further

$$B_{j,n,\ell} = \{x: \frac{\ell}{n} y + (1 - \frac{\ell}{n}) x \in A \quad \text{for some } y \in B_j\}$$

and $p = P(X_1 \in H)$, $q = 1-p$. For $0 \leq \lambda \leq 1$ define the set

$$B_{j,\lambda} = \{x: \lambda y + (1-\lambda) x \in A \text{ for some } y \in B_j\}, \quad (j = 1,2,\ldots,m).$$

Let the Chernoff function X' and X'' be denoted by ρ_1 and ρ_2, resp. We denote $\Gamma_1 = \{j: B_j \subset H \cap A^c\}$. The functions $g_j(\lambda) = \rho_2(B_{j,\lambda})$ $(j \in \Gamma_1)$ are monotone functions therefore we can choose a division $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_N = 1$ of [0,1] such that λ_ℓ's are continuity points of each $g_i(\lambda)$ $(j \in \Gamma_1)$. By using the argument of Sec. 3.1

$$(3.7)\quad P\left(\frac{1}{n-\ell}\sum_{i=1}^{n-\ell} X_i'' \in B_{j,n,k}\right) \le P\left(\frac{1}{n-\ell}\sum_{i=1}^{n-\ell} X_i'' \in B_{j,\lambda_{r-1}}\right) \le$$

$$\le (\rho_2(B_{j,\lambda_{r-1}}) + \varepsilon)^{n(1-\lambda_r)}$$

provided that $\lambda_{r-1} \le \frac{\ell}{n} \le \lambda_r$ and n is large enough $(j \in \Gamma_1,\ r = 2,3,\ldots,N)$. If $r = 1$, then $B_{j,0} = A$ then (3.7) is also valid because of the condition of induction on the number of supporting hyperplanes.

If $j \in \Gamma_2 = \{j: B_j \subset H \cap A\}$, then (3.7) is also true, because $B_{j,\lambda_{r-1}}$ can be enlarged to $\tilde{B}_{j,\lambda_{r-1}} = \{x: \rho_2(x) < \rho_2(\tilde{B}_{j,\lambda_{r-1}})\} \subset A^c$ and for $\hat{B}_{j,\lambda_{r-1}}$ the condition of induction (on the number of hyperplanes covering the singularities) is satisfied. The case $r = 1$ is as same as earlier.

According to the condition of induction by dimension

$$P\left(\frac{1}{\ell}\sum_{i=1}^{\ell} X_i' \in B_j\right) \le (\rho_1(B_j) + \varepsilon)^{\ell} \le (\rho_1(B_j) + \varepsilon)^{n\lambda_{r-1}} \quad \left(\lambda_{r-1} \le \frac{\ell}{n} \le \ell_r\right).$$

By substituting this and (3.7) into (3.6)

$$P\left(\frac{S_n}{n} \in A\right) \le mn \sup_j \sum_{r=1}^{N} \binom{n}{n\lambda_r^*} p^{n\lambda_{r-1}} q^{n(1-\lambda_r)} [\rho_1(B_j) + \varepsilon]^{n\lambda_{r-1}} [\rho_2(B_{j,\lambda_{r-1}}) + \ ]^{n(1-\lambda_r)}$$

where $\lambda_{r-1} \le \lambda_r^* \le \lambda_r$. It follows from this by letting n tend to infinity and then ε to zero that

$$\limsup_{n\to\infty} \left(P\left(\frac{S_n}{n} \in A\right)\right)^{\frac{1}{n}} \le \sup_{j,r} c_{\lambda_r^*} p^{\lambda_{r-1}} q^{1-\lambda_r} \rho_1^{\lambda_{r-1}}(B_j)\rho_2^{1-\lambda_r}(B_j,\lambda_{r-1}) \le$$

$$\le \sup_r c_{\lambda_r^*} p^{\lambda_{r-1}} q^{1-\lambda_r} \sup_{\substack{\lambda_{r-1}y(1-\lambda_{r-1})x\in A \\ y\in \bigcup_{j=1}^m B_j}} (\rho_1(y)+\varepsilon_1)^{\lambda_{r-1}} \rho_2^{1-\lambda_r}(x)$$

Now let $\varepsilon_1 \to 0$ together with $\max_r (\lambda_r - \lambda_{r-1}) \to 0$, then by Lemma 2

$$\limsup_{n\to\infty} \left(P\left(\frac{S_n}{n} \in A\right)\right)^{\frac{1}{n}} \le \sup_{\lambda} c_{\lambda} p^{\lambda} q^{1-\lambda} \sup_{\substack{\lambda y+(1-\lambda)x\in A \\ y\in H}} \rho_1^{\lambda}(y)\rho_2^{1-\lambda}(x) \le \rho(A).$$

If the condition (3.5) does not hold then we modify the distribution of X_1. Let the distribution of the i.i.d. r.v.'s $\tilde{X}_1, \tilde{X}_2, \ldots$ be the mixture of the distribution of X_1 with weight $1-\varepsilon$ and of the probability measure concentrated on the point $x_0 \in \text{int}\,(A \cap \text{dom}\,\rho)$ with weight ε. Then

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \in A\right) \ge P\left(\frac{S_n}{n} \in A\right)(1-\varepsilon)^n$$

and

$$\limsup_{n\to\infty} \left(P\left(\frac{S_n}{n} \in A\right)\right)^{\frac{1}{n}} \le \frac{1}{1-\varepsilon}\, \rho_\varepsilon(A)$$

where ρ_ε is the Chernoff function of the modified distribution. Obviously

$$\rho_\varepsilon(x) = \inf_t (\varepsilon e^{-\langle t, x-x_0\rangle} + (1-\varepsilon)e^{-\langle t,x\rangle}R(t)), \tag{3.8}$$

but from the inequality $0 < \rho(x_0) \le e^{-\langle t,x_0\rangle} R(t)$ we can express $e^{\langle t,x_0\rangle}$, by substituting it into (3.8) we get

$$\rho_\varepsilon(x) \leq \left(\frac{\varepsilon}{\rho(x_0)} + 1-\varepsilon\right)\rho(x).$$

Therefore we have the upper estimate

$$\limsup_{n\to\infty} \left(P\left(\frac{S_n}{n} \in A\right)\right)^{\frac{1}{n}} \leq \left(1 + \frac{\varepsilon}{(1-\varepsilon)\rho(x_0)}\right) \rho(A)$$

and after tending ε to zero Theorem 3 is proved.

3.5. We note that under special circumstances it is possible to give an exact upper bound for $P(\frac{S_n}{n} \in A)$. We refer to [6], Theorem 5.

4. Non-bounded level sets.

4.1. In the last Section the case where $A^c = \{x: \rho(x) \geq c\}$ is bounded or, what is the same, $R(t) < \infty$ in a neighborhood of the origin, has been completely solved. Now we give a result for the unbounded case.

Theorem 4. If $E(\log^k(1+|X_1|)) < \infty$ for every $k > 0$ and the second derivative of the level curve of ρ is monotone for large values of x (see more exactly in Sec. 4.4) then (2.4) holds in R^2.

The second condition seems to be only a technical one and probably it can be eliminated by a refinement of the calculation, but in order to eliminate the first condition new ideas are needed. We have not yet tried to extend the calculations to higher dimensions.

4.2. Before proving Theorem 4 we sketch the main idea of the proof. At a large distance Y_n from the origin we locate the set A^c such that $P(|\frac{S_n}{n}| > Y_n) < c^n$. The remaining part of A^c can be approximated by a polygon with number of sides N_n such that the values of ρ on this polygon are between c and $c+\varepsilon$. On estimating N_n we find that $N_n < (1+\delta)^n$. Then by the Bernstein inequality

$$P(\frac{S_n}{n} \in A) \leq c^n + (1+\delta)^n(c+\varepsilon)^n$$

which proves the statement (2.4).

4.3. Proof. Without restriction of generality we can suppose the following (i) There exist no outstanding singularities. (If this is not so after this proof we can apply the method of Sec. 3), (ii) Let the explicit form of the curve $\rho(x) = c$ be $\eta = \eta(\xi)$ $(x = (\xi,\eta))$, then by an appropriate choice of the coordinate system we can have $\eta'(\xi) \to +\infty$ if $\xi \to +\infty$. (iii) $\eta(\xi) \to +\infty$ if $\xi \to +\infty$. (If this is not so then the complete curve can be approximated by a polygon similarly to Sec. 3.1.)

4.4. To every point of A^c we can assign a point $t(x) = (\tau(x),\sigma(x))$ for which the expression $e^{-tx}R(t)$ takes its minimum ($t(x)$ can be infinite too).

Lemma 3. $\eta'(\xi) = -\frac{\tau(x)}{\sigma(x)}$.

Proof. Taking the derivative of the identity

$$\log c = -\tau(\xi,\eta(\xi))\xi - \sigma(\xi,\eta(\xi))\eta(\xi) + \log R(\tau,\sigma)$$

and using the equality grad log $R(t) = x$ we obtain Lemma 3.

Lemma 4. $\limsup_{\substack{\xi\to\infty \\ \rho(x) = c}} \langle t(x), x\rangle \leq K$

Proof of Lemma 4. Condition (iii) implies that the Chernoff function of the first coordinate of X_1 (see Sec. 2.4) tends to 1 if $\xi \to +\infty$, consequently $R(t) = +\infty$ for $\sigma > 0$ and for $\sigma = 0$, $\tau > 0$. Consider a level curve $R(t) = T$ ($T > 1$, fixed), this is a convex curve going through the origin. First we prove that $\langle t, \text{grad} \log R(t)\rangle \to +\infty$ if $t \to 0$ on this level curve. By a simple calculation

$$\langle t, \text{grad} \log R(t)\rangle = \frac{1}{T} \int_{-\infty}^{\infty} \langle t,x\rangle e^{\langle t,x\rangle} dF(x).$$

According to the bounded convergence criterion

$$\int_{<t,x><0} <t,x>e^{<t,x>}dF(x) \to 0 \qquad t \to 0,$$

therefore this part is negligible. Let us choose a $c > 0$ value arbitrarily then

$$<t,\text{grad } \log R(t)> \geq \frac{1}{T} \int_{<t,x>\geq c} <t,x>e^{<t,x>} dF(x) + o(1) \geq$$

$$\geq \frac{C}{T}(T- \int_{<t,x><C} e^{<t,x>}dF(x)) + o(1).$$

If we choose now a function $\omega = \omega(t)$ such that $\omega \to \infty$ but $|t|\omega \to 0$ $(t \to 0)$ then

$$\int_{<t,x><C} e^{<t,x>}dF(x) = \int_{<t,x>\leq|t|\omega} e^{<t,x>}dF(x) + \int_{|t|\omega<<t,x><C} e^{<t,x>}dF(x) \; \_$$

$$\leq e^{|t|\omega} + e^{C}P(\frac{<t,X_1>}{|t|} >\omega) = 1 + o(1).$$

These inequalities yield

$$<t, \text{ grad } \log R(t)> \geq C(1 - \frac{1}{T}) + o(1),$$

(4.1) $$<t, \text{ grad } \log R(t)> \to +\infty$$

if $t \to 0$ such that $R(t) = T$.

The set $B = \{t(x): \; x \in A^c\}$ can not contain the points where $R(t) = T$ for small t values, since for the points of B

(4.2) $$-<t, \text{ grad } \log R(t)> + \log R(t) \geq - \log c$$

and for the points $R(t) = T$, if they are in B as well,

$$\langle t, \text{grad} \log R(t) \leq \log c + \log T$$

and this contradicts (4.1).

On every half line starting from the origin the function $-\langle t, \text{grad} \log R(t)\rangle +$ $\log R(t)$ is monotone decreasing, therefore the section of the half line with B is always an interval (or empty). From this fact and from the statement of the last paragraph it follows that the condition (ii) - using also Lemma 1 - can hold only if $t(x) \to 0$ where $x = (\xi, \eta(\xi))$ and $\xi \to +\infty$. But for small t values $R(t) \leq T$, therefore

$$\langle t, \text{grad} \log R(t)\rangle = \log R(t) - \log \rho(x) \leq \log T - \log c.$$

With this Lemma 4 is proved.

4.4. Proof of Theorem 4 (continued). From our point of view it is more suitable to consider the inverse function $\xi(\eta)$ of $\eta(\xi)$. The distance between two level curve $\rho(x) = c$ and $\rho(x) = c + \varepsilon$ can be estimated by

$$d \leq \varepsilon(|t(x')|)^{-1}$$

where $\rho(x') = c + \varepsilon$. For the length h of a chord of $\xi(\eta)$ we get by an easy calculation that

$$h^2 \geq C\varepsilon[\xi''(\eta^*)|t(\eta')|]^{-1}$$

where η^* is a point of the interval determined by the second coordinates of the endpoints of the chord.

Now let us make a suitable polygon - approximation of the curve on the interval $A \leq \eta \leq Y$ and denote the second coordinates of its vertices by η_i

taking $\eta_0 = A$, $\eta_N = Y$. In the following estimation we use the second condition of Theorem 4: We assume that $\xi''(\eta)$ is a monotone function of η for $\eta \geq A$ (A is large enough). The monotonicity of $|t(\eta)|$ follows from the statements of the last section.

Having these properties we can estimate the following integral:

$$(C\varepsilon)^{-\frac{1}{2}} \int_A^Y \sqrt{|\xi''(\eta)|\,|t(\eta)|}\, dy \geq (C\varepsilon)^{-\frac{1}{2}} \sum_{i=1}^{N} (\eta_i - \eta_{i-1}) \sqrt{|\xi''(\eta_i)|\,|t(\eta_i)|} \geq$$

$$\geq (C\varepsilon)^{-\frac{1}{2}} \sum_{i=2}^{N} (\eta_{i-1} - \eta_{i-2}) \sqrt{|\xi''(\eta^*_{i-1})|\,|t(\eta^!_{i-1})|} \geq \sum_{i=2}^{N} \frac{\eta_{i-1} - \eta_{i-2}}{h_{i-1}} \geq$$

$$\geq (N-1)(1-\delta)$$

provided that A large enough and $Y > A$. Consequently the number of chords

$$N_n \leq C_1 \int_A^Y \sqrt{|\xi''(\eta)|\,|t(\eta)|}\, d\eta. \tag{4.3}$$

Let the intersection of a tangent drawn to a point $(\xi(\eta),\eta)$ and the ξ-axis be $(0,\zeta(\eta)) = z(\eta)$. Then for $x = (\xi(\eta),\eta)$ the inner product

$$\langle X,t(x)\rangle = \langle z(\eta),t(\eta)\rangle \sim \zeta(\eta)\,|t(\eta)| \qquad (\eta \to +\infty)$$

and because of Lemma 4

$$|t(\eta)| \leq \frac{K}{\zeta(\eta)}.$$

Another simple calculation shows that

$$\zeta'(\eta) = \eta\,|\xi''(\eta)|.$$

By substituting these relations into (4.3) we have

$$N_n \leq C_1 \int_A^{Y_n} \sqrt{\frac{\zeta'(\eta)}{\zeta(\eta)} \frac{1}{\eta}} \, d\eta \leq C_1 \left[\int_A^{Y_n} \frac{\zeta'(\eta)}{\zeta(\eta)} \, d\eta \int_A^{Y_n} \frac{1}{\eta} d\eta \right]^{\frac{1}{2}} \leq$$

$$\leq C_2 \sqrt{\log \zeta(Y_n) \log Y_n}.$$

If Y_n is large enough $\zeta(Y_n) \leq \xi(Y_n) \leq Y_n$, therefore

$$N_n \leq C_2 \log Y_n,$$

and, according to the requirement of Sec. 4.2, $N_n \leq (1+\delta)^n$ if

$$\log Y_n \leq \frac{1}{C_2} (1+\delta)^n.$$

Denote the second coordinate of S_n by $(S_n)_2$, then

$$P(\tfrac{1}{n}(S_n)_2 > Y_n) \leq n\, P(|X_1| > Y_n) = n\, P(\log^k(1+|X_1|) \leq \log^k(1+Y_n)) \leq$$

$$\leq n\, E(\log^k(1+|X_1|)) \log^{-k}(1+Y_n) \leq$$

$$\leq \frac{n\, C(k)}{(1+\delta)^{nk}} \leq nC(k) c^n$$

for large enough n provided that $(\frac{1}{1+\delta})^k \leq c$.

With this, Theorem 4 is proved.

Acknowledgement: The author wishes to thank the Department of Statistics and Mathematics of Purdue University for financial support for research during his visit.

REFERENCES

[1] Bahadur, R. R.: Some limit theorems in statistics, SIAM, Philadelphia, 1971.

[2] Bahadur, R. R., Ranga Rao, R.: On deviation of the sample mean; Ann. Math. Statist, 31 (1960), 43-54.

[3] Bahadur, R. R., Zabell, S. L.: Large deviations of the sample mean in general vector spaces (to appear).

[4] Bártfai, P.: Large deviations in the queuing theory, Periodica Math. Hungar., 2 (1972), 165-172.

[5] Bártfai, P.: Connections between convex analysis and theory of large deviations, Preprint of the Math. Inst. of Hung. Acad. Sci., 1977.

[6] Bártfai, P.: On the multivariate Chernoff theorem, Preprint of the Math. Inst. of Hung. Acad. Sci., 1977.

[7] Chernoff, H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann. Math. Statist., 23 (1952), 493-507.

[8] Cramér, H.: Sur un nouvean theoreme-limite de la theorie des probabilites, Act. Sci. et Ind., 736 (1938), 5-23.

[9] Lanford, O. E.: Entropy and equilibrium states in classical statistical mechanics. Statistical Mechanics and Mathematical Problems. Lecture Notes in Physics 20 (1971), 1-113.

[10] Linnik, Ju. V.: In the probability of large deviations for sums of independent variables, Proc. 4th Berkeley Symp. Math. Stat. and Prob., Vol. 2, 289-306.

[11] Nagaev, A. V.: Limit theorems that take into account large deviations when Cramer's condition is violated, lzv. Ak. Nauk UzSSR, Ser. Math. Fiz., 13 (1969), 17-22.

[12] Osipov, L. V.: On probability of large deviations for sum of independent random variables, Theory of Prob. Appl. 17(1972), 309-331.

[13] Petrov, V. V.: On the probabilities of large deviations for sums of independent random variables. Theory of Prob. Appl., 10 (1965), 287-298.

[14] Plachky, D., Steineback, J.: A theorem about probabilities of large deviations with an application to queuing theory, Periodic Math. Hungar., 6 (1975), 343-345.

[15] Rockafellar, R. T.: Convex Analysis, Princeton Univ. Press, 1970.

[16] Rubin, H. Sethuraman, J.: Probabilities of moderate derivations, Sankhyā Ser. A. 27 (1965), 325-346.

[17] Sievers, G. L.: On the probability of large deviations and exact slopes, Ann. Math. Statist., 40 (1969), 1908-1921.

[18] Steiger, W. L.: Bernstein's inequality for martingales, Zeitchinft fur Wahrscheinlichkeitstheorie, 16 (1970), 107-106.

[19] Steinebach, J." Convergence rates of large deviation probabilities in the multi-dimensional case, (to appear).