BEHIND EVERY LIMIT THEOREM THERE IS AN INEQUALITY

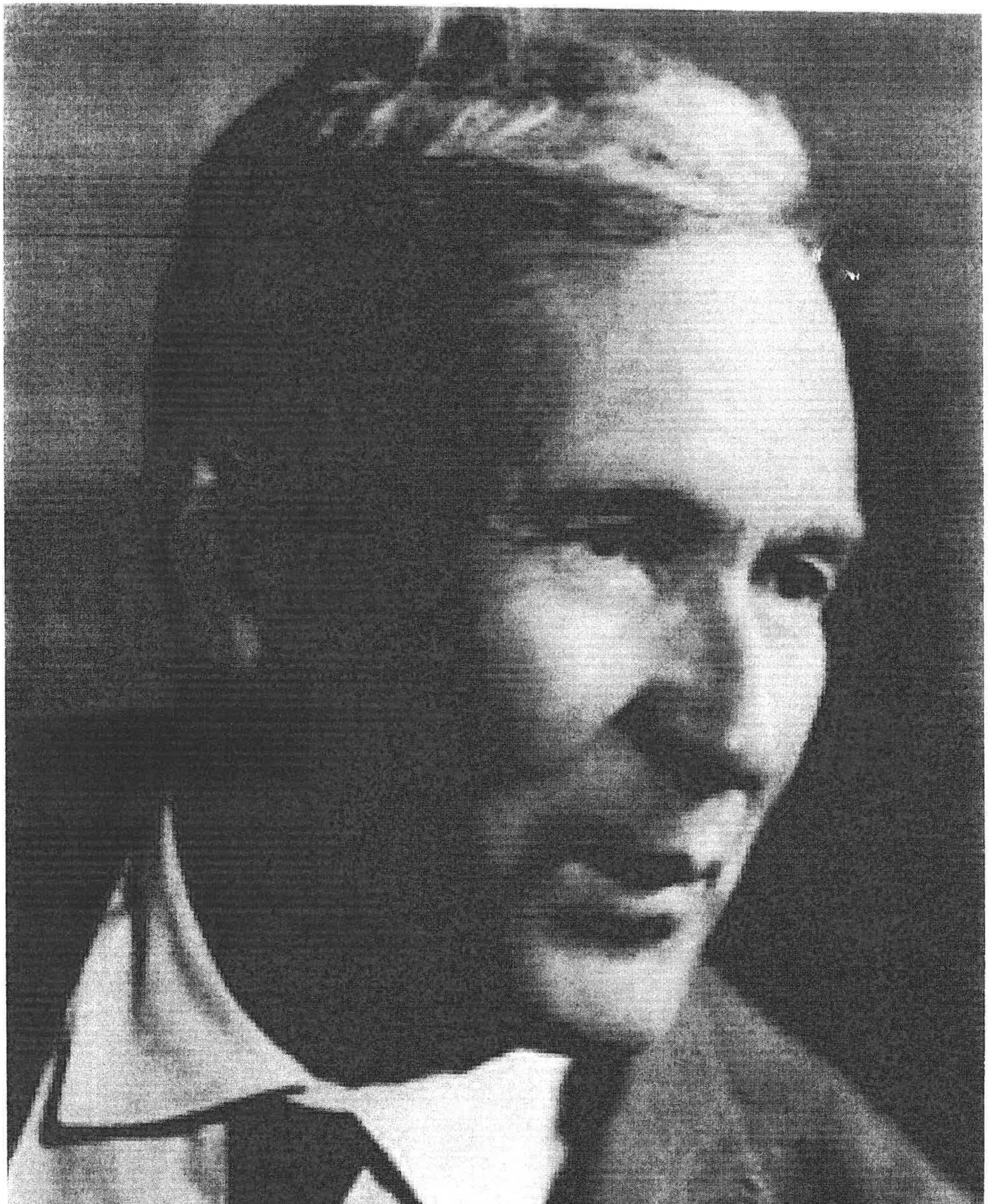
FOUNDATIONS
OF THE
THEORY OF PROBABILITY

BY

A. N. KOLMOGOROV

TRANSLATION EDITED BY
NATHAN MORRISON

CHELSEA PUBLISHING COMPANY
NEW YORK
1950



Example. Let $\frac{S_n}{a_n}$ be a sequence of real random variables. Then, to show that $\frac{S_n}{a_n} \to \mu$ in probability, Markov s inequality is often used:

$$P(|\frac{S_n}{a_n} - \mu| > \epsilon) \leq E\frac{|\frac{S_n}{a_n} - \mu|^p}{\epsilon^p},$$

Let $\{X_i\}$ be a sequence of i.i.d. random variables and

$$S_n = X_1 + ... + X_n.$$

The weak law of large numbers for sums of i.i.d. random variables with finite variance uses the case $p = 2$ and the fact that the variance of the sum is the sum of the variances. What happens when the variance is infinite, and when a_n depends on $(X_1, X_2, ..., X_n)$?

Self-Normalized Processes in Dependent Variables

by

Victor H. de la Peña
Department of Statistics
Columbia University

Self-normalized processes are frequently found in statistical applications. They have the property of (in standard form) being unit free and frequently eliminate or weaken moment assumptions. The prototypical example of a self-normalized random process is Student's t-statistic which replaces the population standard deviation in the standard form $\sqrt{n}(\bar{X} - \mu)/\sigma$ by the sample standard deviation. In a more general context, a (standard) self-normalized process can take on the form A_n/B_n or A_t/B_t in continuous time, where B_t is a random variable that frequently is used to estimate a dispersion measure of the process A_t.

OUTLINE

A SELF-NORMALIZATION: EXAMPLES

a1) Introduction and Motivation

a2) Martingales, MLE's, Auto Regressive Processes

a3) Examples (Martingales and Super-Martingales)

B PSEUDO-MAXIMIZATION

b1) Psedudo-Maximization by density integration

b2) Exponential inequalities

b3) Boundary Crossing

b4) LIL for self-normalized normal and continuous martingales

A. Self-normalization: Examples

Self-normalized processes are frequently found in statistical applications. They have the property of (in standard form) being unit free and frequently eliminate or weaken moment assumptions. The prototypical example of a self-normalized random process is Student's t-statistic which replaces the population standard deviation in the standard form $\sqrt{n}(\bar{X}-\mu)/\sigma$ by the sample standard deviation. In a more general context, a (standard) self-normalized process can take on the form A_n/B_n or A_t/B_t in continuous time, where B_t is a random variable that frequently is used to estimate a dispersion measure of the process A_t.

Let $\{X_i\}$ be i.i.d normal $N(\mu, \sigma^2)$,

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \qquad s_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1},$$

$T_n = \frac{\bar{X}_n - \mu}{s_n/\sqrt{n}}$ is t-distributed t_{n-1}. Let $Y_i = X_i - \mu$, $A_n = \sum_{i=1}^n Y_i$, $B_n^2 = \sum_{i=1}^n Y_i^2$.

Then

$$T_n = \frac{\frac{A_n}{B_n}}{\sqrt{(n-(A_n/B_n)^2)/(n-1)}}$$

Information on the properties of T_n can be derived from the self-normalized process:

$$\frac{A_n}{B_n} = \frac{\sum_{i=1}^n Y_i}{\sqrt{\sum_{i=1}^n Y_i^2}}$$

In recent years, there has been increasing interest in limit theorems and moment bounds for self-normalized sums of i.i.d. zero-mean random variables X_i. In particular, Bentkus and Götze (1996), where a Berry-Esseen bound for the Student t-statistic is obtained, and Giné, Götze and Mason (1997), where it is proved that the Student t-statistic has a limiting standard normal distribution if and only if X_1 is in the domain of attraction of a normal law, by making use of *exponential and L_p bounds* for the self-normalized sums $U_n = A_n/B_n$, where $A_n = \sum_{i=1}^{n} X_i$ and $B_n^2 = \sum_{i=1}^{n} X_i^2$.

Shao (1997) provides large deviation results for U_n without moment conditions and moderate deviation results when X_1 is the domain of attraction of a normal or stable law. Egorov (1998) gives exponential inequalities for a centered variant of U_n. de la Peña, Lai and Shao (2009) contains an extensive treatment of Self-normalized processes and their applications.

a2. Martingales, MLE's, Auto-Regressive Processes

The first example of Self-normalized processes in dependent variables that we treat in this survey arises in the context of Maximum Likelihood Estimators (MLE)'s for the parameter in a linear regression model.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. $\mathcal{F}_n$ and increasing sequence of σ-fields in $\mathcal{F}$. Let $M_n = d_1 + d_2 + ... + d_n$, $n = 1, ...$ be a martingale adapted to $\mathcal{F}_n$, with martingale difference sequence $\{d_i\}$ Then, $E(d_n|\mathcal{F}_{n-1}) = 0$, $E|d_n| < \infty$ for $i = 2, ...$

Example. Consider the linear regression model, $Y_0 = 0$,

$$Y_i = \alpha Y_{i-1} + \epsilon_i,$$

where $\alpha \neq 0$ is a fixed (unknown) parameter and ϵ_i are (for simplicity) independent standard normal random variables $N(0,1)$. We will show that the maximum likelihood estimator (MLE) for α is a self-normalized process.

The MLE for α can be obtained by taking the derivative of the log likelihood function $L_{Y_1,...,Y_n}(\alpha)$.

$$L_{Y_1,...,Y_n}(\alpha) = \log f(Y_1, ..., Y_n) = \log f(Y_1, ..., Y_{n-1}) f_{Y_n|Y_1,...,Y_{n-1}}$$

$$= \log[f(Y_1, ..., Y_{n-1}) \frac{\exp\{(Y_n - \alpha Y_{n-1})^2/2\}}{\sqrt{2\pi}}]$$

$$\cdots$$

$$= \sum_{j=1}^{n} (Y_j - \alpha Y_{j-1})^2/2 - n \log(\sqrt{2\pi}).$$

Taking derivative with respect to α, equating to zero and solving for α, we obtain the MLE for α

$$\hat{\alpha}_n = \frac{\sum_{j=1}^{n} Y_{j-1} Y_j}{\sum_{j=1}^{n} Y_{j-1}^2}.$$

Using the definition of Y_j we obtain

$$\hat{\alpha}_n = \frac{\sum_{j=1}^{n} Y_{j-1}(\alpha Y_{j-1} + \epsilon_j)}{\sum_{j=1}^{n} Y_{j-1}^2} =$$

$$\alpha + \frac{\sum_{j=1}^{n} Y_{j-1} \epsilon_j}{\sum_{j=1}^{n} Y_{j-1}^2}.$$

Therefore, in testing a hypothesis about α, we can use the self-normalized random variable

$$\hat{\alpha}_n - \alpha = \frac{\sum_{j=1}^{n} Y_{j-1} \epsilon_j}{\sum_{j=1}^{n} Y_{j-1}^2},$$

as a measure of the distance between our estimator $\hat{\alpha}_n$ and the parameter of interest α.

It is interesting to note that the the random variable

$$A_n = \sum_{j=1}^{n} Y_{j-1}\epsilon_j$$

is a martingale with respect to the sequence

$$\mathcal{F}_n = \sigma(Y_1, ..., Y_n; \quad \epsilon_1, ..., \epsilon_n).$$

Setting $d_j = Y_{j-1}\epsilon_j$ we see that the conditional variance of A_n is

$$B_n^2 = \sum_{j=1}^{n} E[(Y_{j-1}\epsilon_j)^2 | \mathcal{F}_{n-1}] = \sum_{j=1}^{n} Y_{j-1}^2.$$

Therefore, in this case $\hat{\alpha} - \alpha = \frac{A_n}{B_n^2}$ is a self-normalized process where the normalization is done by using the conditional variance.

One should note the following fact that holds for all $n \geq 1$ and λ with $-\infty < \lambda < \infty$

$$E\exp\{\lambda \sum_{j=1}^{n} Y_{j-1}\epsilon_j - \frac{\lambda^2 \sum_{j=1}^{n} Y_{j-1}^2}{2}\} \leq 1.$$

The random variable

$$T_n = \exp\{\lambda \sum_{j=1}^{n} Y_{j-1}\epsilon_j - \frac{\lambda^2 \sum_{j=1}^{n} Y_{j-1}^2}{2}\},$$

is an example of an exponential super-martingale. It provides a good example of a random variable satisfying the canonical condition that we will use to develop a large body of theory for self-normalized variables.

Canonical Assumption.

For arbitrary random variables A, B, with $B > 0$

$$E \exp\{\lambda A - \lambda^2 B^2/2\} \leq 1,$$

for all λ, $\quad -\infty < \lambda < \infty$.

The canonical assumption in particular implies several moment, and exponential bounds including the following bound connected to the law of large numbers which derives from de la Peña (1999).

$$P(A/B^2 > x, 1/B^2 \leq y) \leq \exp\{-x^2/2y\},$$

for all x, $\; y > 0$.

Applying this bound to our example to A_n and $-A_n$ and letting $z = 1/y$ we get for all $x, z > 0$,

$$P(|\hat{\alpha}_n - \alpha| > x, \sum_{j=1}^{n} Y_{j-1}^2 \geq z) \leq 2 \exp\{-x^2 z^2/2\},$$

giving us a sense as to how far our estimator is from the true parameter when the data exceeds a certain level.

The following estimates provide related control on the tail, using different means of attaining the control. In analogy to the previous result, they are connected to the central limit theorem.

Assume that the canonical assumption holds, then for all $x > 0$, from de la Peña, Klass and Lai (2004),

$$P(A/\sqrt{B^2 + (EB)^2} > x) \leq \sqrt{2} \exp\{-x^2/4\}.$$

From de la Peña and Pang (2009), for all $p,\ q \geq 1$ with $\frac{1}{p}+\frac{1}{q}=1$.

$$P(\frac{|A|}{\sqrt{\frac{2q-1}{q}(B^2+(E|A|^p)^{1/p})}} \geq x) \leq$$

$$(\frac{q}{2q-1})^{\frac{q}{2q-1}} x^{\frac{-q}{2q-1}} \exp\{-x^2/2\}.$$

Compare these with traditional bound for a normal rv with mean 0 and variance 1.

$$P(|Z| \geq x) \leq \frac{2}{\sqrt{\Pi}} x^{-1} \exp\{-x^2/2\}.$$

In the next section we provide several examples of random variables that satisfy the canonical assumption.

a3. Useful Exponential Martingales

Lemmas 1 and 2 are considered classical results in martingale theory. Lemma 3 comes from Bercu and Touati (2008). Lemma 4 deals with the case of conditionally symmetric random variables without any moment or dependence condition.

Lemma 1. *Let W_t be a standard Brownian motion. Assume that T is a stopping time such that $T < \infty$ a.s.. Then*

$$E\exp\{\lambda W_T - \lambda^2 T/2\} \leq 1,$$

for all $\lambda \in \mathbf{R}$.

Lemma 2. *Let M_t be a continuous, square-integrable martingale, with $M_0 = 0$. Then*

$$E\exp\{\lambda M_t - \lambda^2 \langle M\rangle_t/2\} \leq 1,$$

for all $t \geq 0$. Moreover, $\exp\{\lambda M_t - \lambda^2\langle M\rangle_t/2\}$ is a supermartingale for all $\lambda \in \mathbf{R}$.

Lemma 3. *Let $\{d_i\}$ be a martingale difference sequence with respect to the filtration $\{\mathcal{F}_n\} : n \geq 1$. Suppose that $Ed_i^2 < \infty$ for all $i \geq 1$. Then, for all $\lambda \in \mathbf{R}$*

$$E[\exp(\lambda \sum_{i=1}^n d_i - \frac{\lambda^2}{2}(\sum_{i=1}^n d_i^2 + \sum_{i=1}^n E(d_{i^2}|\mathcal{F}_{n-1})) \leq 1.$$

The following lemma holds without any integrability conditions on the variables involved.

It is a generalization of the fact that if X is any symmetric random variable, then $A = X$ and $B = X^2$ satisfy the canonical condition.

Lemma 4. *Let $\{d_i\}$ be a sequence of variables adapted to an increasing sequence of σ-fields $\{\mathcal{F}_i\}$. Assume that the d_i's are conditionally* *symmetric (that is* $\mathcal{L}(d_i|\mathcal{F}_{i-1}) = \mathcal{L}(-d_i|\mathcal{F}_{i-1})$*). Then* $\exp\{\lambda\Sigma_{i=1}^n d_i - \lambda^2\Sigma_{i=1}^n d_i^2/2\}$, $n \geq 1$, *is a supermartingale with mean* ≤ 1, *for all* $\lambda \in \mathbf{R}$.

Note that any sequence of real-valued random variables X_i can be "symmetrized" to produce an exponential supermartingale by introducing random variables X_i' such that

$$\mathcal{L}(X_n'|X_1, X_1', \ldots, X_{n-1}, X_{n-1}', X_n) = \mathcal{L}(X_n|X_1, \ldots, X_{n-1})$$

and setting $d_n = X_n - X_n'$; see Section 6.1 of de la Peña and Giné (1999).

B. Pseudo-Maximization (Method of Mixtures)

b1. Pseudo-Maximization by density integration

Recall that in our previous section, we presented several examples of pairs of random variables A, B with $B > 0$ satisfying the canonical inequality

$$E\exp\{\lambda A - \frac{\lambda^2 B^2}{2}\} \leq 1, \qquad (*)$$

for all $-\infty < \lambda < \infty$.

We are interested in developping moment and exponential inequalities as well as LIL's on the sole basis of (*).

Note that if one was to maximize over λ inside the expectation (as can be done in the case when A, B are non-random), taking the value $\lambda = A/B^2$ one would get $E\exp\{\frac{A}{2B^2}\} \leq 1$. This in turn would give the "ideal" inequality $P(\frac{A}{B} > x) \leq \exp\{\frac{-x^2}{2}\}$. Since in general A, B can not be taken to be non-random, we need to find an alternative method for dealing with this maximization. One approach for attaining a similar effect involves the integrating over a probability measure F, and using Fubini's theorem to interchange the order of integration with respect to P and F. The adequate F would need to grow as slow as possible so as to "pick" the maximum value of $T(\lambda) = \exp\{\lambda A - \lambda^2/2B^2\}$.

This approach will be used in what follows to provide exponential and moment inequalities for

$$\frac{A}{B}, \quad \frac{A}{\sqrt{B^2 + (EB)^2}}, \quad \frac{A}{B\sqrt{\log\log(B \vee e^2)}},$$

as well as LIL's. We begin with the second case where the proof is more transparent. The approach used was pioneered by Robbins and Siegmund (1970) and is commonly known as the method of mixtures.

b2. An exponential inequality

Theorem 2. (de la Peña, Klass and Lai (AOP)) *Let A, B with $B > 0$ be random variables satisfying the canonical assumption for all $\lambda \in \mathbf{R}$. Then*

$$P(\frac{|A|}{\sqrt{B^2 + (EB)^2}} > x) \le \sqrt{2} \exp(-x^2/4)$$

for all $x > 0$.

The proof of this result is based on the following lemma.

Lemma *Let A, B with $B > 0$ be two random variables satisfying the canonical condition for all $\lambda \in \mathbf{R}$. Then for all $y > 0$,*

$$E \frac{y}{\sqrt{B^2 + y^2}} \exp\{\frac{A^2}{2(B^2 + y^2)}\} \le 1.$$

Proof: Multiplying both sides of the canonical condition by $(2\pi)^{-1/2} y \exp(-\lambda^2 y^2/2)$ (with $y > 0$) and integrating over λ, we obtain by using Fubini's theorem that

$$1 \ge \int_{-\infty}^{\infty} E \frac{y}{\sqrt{2\pi}} \exp\left(\lambda A - \frac{\lambda^2}{2} B^2\right) \exp\left(- \frac{\lambda^2 y^2}{2}\right) d\lambda$$

$$= E\Big[\frac{y}{\sqrt{B^2 + y^2}} \exp\left\{\frac{A^2}{2(B^2 + y^2)}\right\} \times$$

$$\int_{-\infty}^{\infty} \frac{\sqrt{B^2 + y^2}}{\sqrt{2\pi}} \exp\{- \frac{B^2 + y^2}{2}\left(\lambda^2 - 2\frac{A}{B^2 + y^2}\lambda + \frac{A^2}{(B^2 + y^2)^2}\right)\} d\lambda\Big]$$

$$= E\Big[\frac{y}{\sqrt{B^2 + y^2}} \exp\left(\frac{A^2}{2(B^2 + y^2)}\right)\Big]. \qquad \square$$

By Schwarz's inequality

$$E\exp\left\{\frac{A^2}{4(B^2+y^2)}\right\} \le \left\{\left(E\frac{y\exp\{\frac{A^2}{2(B^2+y^2)}\}}{B^2+y^2}\right)\left(E\sqrt{\frac{B^2+y^2}{y^2}}\right)\right\}^{1/2}$$
$$\le \left(E\sqrt{\frac{B^2}{y^2}+1}\right)^{1/2}.$$

Since $E\sqrt{\frac{B^2}{y^2}+1} \le E(\frac{B}{y}+1)$, the special case $y = EB$ above gives

$$E\exp(A^2/[4(B^2+(EB)^2]) \le \sqrt{2}.$$

Then, using Markov's inequality and this we get

$$P\{\frac{|A|}{\sqrt{B^2+(EB)^2}} \ge x\} = P\{\frac{A^2}{4(B^2+(EB)^2)} \ge \frac{x^2}{4}\} \le \sqrt{2}\,\exp(-x^2/4).$$

□

Heuristic Interpretation

When $EB \to \infty$ we are roughly locating point mass around zero and since for all $\epsilon > 0$, $P(A/B^2 > \epsilon, B^2 > z) \le \exp\{-\epsilon^2 z^2/2\} \to 0$ as $z \to \infty$, the maximum value of $T(\lambda, A, B)$ is also located around $\lambda = 0$ with high probability. A similar argument gives a heuristic in other situations.

b3) Boundary Crossing

In what follows we study boundary crossing by using the method of mixtures first introduced in Robbins and Siegmund (1970), and extended in de la Peña, Klass and Lai (2004) to treat the general case of self-normalization. As in Section b1, the method relies on the construction of appropriate super-martingales through mixing the random variables in our cannonical assumption by integrating over appropriate functions.

We begin by introducing the "Robbins-Siegmund" (R-S) boundaries which are extensively discussed in Lai (1976).

Let F be a finite measure on $(0, \lambda_0)$ and assume that $F(0, \lambda_0) > 0$.

Let $\Psi(u, v^2) = \int \exp(\lambda u - \lambda^2 v^2/2\} dF(\lambda)$. Given $c > 0$ and $v^2 > 0$, the equation $\Psi(u, v^2) = c$ has a unique solution $u = \beta_F(v^2, c)$. In addition, $\beta_F(v^2, c)$ is a concave function of v^2 and $\lim_{v^2 \to \infty} \frac{\beta_F(v^2, c)}{v^2} = b/2$ where $b = \{\sup y > 0 : \int_0^b F(d\lambda) = 0\}$, with sup over the empty set equal to cero.

In order to use the R-S boundaries $\beta_F(v^2, c)$ in boundary crossing we proceed as follows.

Consider the problem of estimating the probability

$$P(A_t \geq g(B_t) \ \text{ for some} t \geq 0 \ ).$$

If g is an R-S boundary, then $g(B_t) = \beta_F(B_t^2, c)$ for some F and $c > 0$. This probability equals

$$P(A_t \geq \beta_F(B_t^2, c) \ \text{ for some } 0 \leq t \leq t_0 \ ) =$$

$$P(\Psi(A_t, B_t^2) \geq c \ \text{ for some } 0 \leq t \leq t_0 \ ) \leq$$

$$E\Psi(A_{t_0}, B_{t_0}^2)/c \leq F(0, \lambda_0)/c,$$

by using Doob's inequality on the super-martingale $\Psi(A_t, B_t), 0 \leq t \leq t_0$.

In what follows we present an application of this methodology in which the upper LIL for continuous martingales is obtained. We begin by taking

$$dF_\delta(\lambda) = \frac{1}{\lambda(\log(1/\lambda))(\log\log(1/\lambda)^{1+\delta}},$$

for all $\delta > 0$ and $0 < \lambda < e^{-e}$.

Let $\log_2(x) = \log\log(x)$ and $\log_3(x) = \log\log_2(x)$. As shown in Example 4 of Robbins and Siegmund (1970), in this case

$$\beta_{F_\delta}(v^2, c) = \sqrt{2v^2[\log_2 v^2 + (3/2+\delta)\log_3 v^2 + \log(c/2\sqrt{\pi}) + o(1)]},$$

as $v^2 \to \infty$ Using the above we have that the probability of interest is bounded by with $F_\delta(0, e^{-e})/c$ for all $c > 0$. For $\epsilon > 0$ given, take δ small enough and and $c = c(\delta)$ large enough so that $F_\delta(0, e^{-e})/c(\delta) < \epsilon$. Since ϵ can be arbitrarily small and for fixed $c = c(\delta)$, $\beta_{F_\delta}(v^2, c) \to \sqrt{2v^2 \log\log v}$ as $v^2 \to \infty$,

$$\limsup_{t\to\infty} \frac{A_t}{\sqrt{2B_t \log\log B_t}} \leq 1,$$

on the set $B_t \to \infty$.

The heuristic here is that as we are integrating over $dF_\delta(\lambda)$, which is chosen to decrease very slowly. In fact, for $\delta = 0$, this function is not integrable. Integrating over this density and then as $\delta \to 0$ allows one to retreive the largest almost sure growth of A_t, as measured by B_t.

To prove the reverse inequality one uses Lemma 1 of Robbins and Siegmund (1970).