Large deviations for random walks with finite M.G.F.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d random variables such that

$$E[X_1] = \mu < \infty$$
and
$$Var(X_1) = \sigma^2 \in (0,\infty)$$

Define $S_n := X_1 + X_2 + \cdots + X_n$, $n \in \mathbb{N}$.

We recall two fundamental results from basic probability theory that tells us about the path properties of S_n as n grows very large.

The first one tells the 'speed' or the drift of the walk.

Theorem (i) Strong law of large numbers (SLLN):

$$S_n/n \xrightarrow{a.s} \mu \quad \text{as} \quad n \to \infty$$

i.e. $$P\left(\lim_{n\to\infty} S_n/n = \mu\right) = 1$$

The second is about second order fluctuation from the 'drift'.

Theorem (ii) Central limit theorem (CLT):

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{dist.} N(0,1) \quad \text{as} \quad n \to \infty$$

$N(0,1)$ is the standard normal random variable.

CLT specifies that typically S_n differs from $n\mu$ by an amount of order $\sqrt{n}$. Deviations of this order from $n\mu$ are called normal deviations.

Now consider the following event:

$$\{S_n \geq (\mu + a)n\}, \quad a \geq 0.$$

We shall show that SLLN implies the probability of this event decays to 0 as $n \to \infty$.

By theorem (i) we have,

$$P\left(\lim_{n\to\infty} S_n/n = \mu\right) = 1$$

$$\Rightarrow P\left(\bigcap_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \left\{\left|S_n/n - \mu\right| < 1/k\right\}\right) = 1$$

Given $a \geq 0$ let k be such that $0 < 1/k < a$

say $A_n^k := \left\{ \left| S_n/n - \mu \right| < 1/k \right\}$

and $B_m^k = \bigcap_{n=m}^{\infty} A_n^k$.

Now B_m^k is a sequence of increasing events and it increases to $\bigcup_{m=1}^{\infty} B_m^k$.

Now $P\left(\bigcup_{m=1}^{\infty} B_m^k\right) = 1$ [$\because$ Probability of a smaller event is 1.]

So given $\delta > 0$ $\exists$ $M_0 \in \mathbb{N}$ such that $P(B_m^k) \geq 1-\delta$ if $m \geq M_0$

and as $A_m^k \supseteq B_m^k$ we've

$$P(A_n^k) \geq 1-\delta \quad \forall \; n \geq M_0$$

$$\Rightarrow P\left(\left|S_n/n - \mu\right| < 1/k\right) \geq 1-\delta \quad \forall \; n \geq M_0$$

$$\Rightarrow P\left(\left|S_n/n - \mu\right| \geq 1/k\right) < \delta \quad \forall \; n \geq M_0$$

$$\Rightarrow P\left(S_n/n - \mu \geq 1/k\right) < \delta \quad \forall \; n \geq M_0$$

$$\Rightarrow P\left(S_n/n - \mu \geq a\right) < \delta \quad \forall \; n \geq M_0.$$

$$\Rightarrow \lim_{n\to\infty} P\left(S_n/n \geq \mu + a\right) = 0$$

S_n is the main process we'd deal with and try to understand the precise decay rate of rare events such as the above deviation of S_n from μn of order n.

We'd show the rate at which $P\left(S_n/n \geq \mu + a\right)$ decays is exponential in n.

An intuition behind this is as follows:

$$P(S_n \geq (\mu + a)n)$$

$\approx P(X_i \geq \mu + a ;\ 1 \leq i \leq n)$ [$\because$ n is large the fraction of X_i's deviating from $\mu + a$ is very small.]

$= \{P(X_1 \geq \mu + a)\}^n$ [$\because$ The X_i's are i.i.d.]

$= (f(a))^n$ [$f(a) = P(X_1 \geq \mu + a) \leq 1$]

$= e^{n \log (f(a))}$

Deviations of S_n from $n\mu$ of the order n are called Large deviations.

We are now ready to state and prove our main result:

Theorem (iii)

Let $\{X_i\}_{i \geq 1}$ be i.i.d. random variables with $P(X_1 = 0) = P(X_1 = 1) = 1/2$

Then, for $a > 1/2$

$$\lim_{n \to \infty} \frac{1}{n} \log(P\{S_n \geq an\}) = -I(a)$$

where $I(z) = \begin{cases} \log 2 + z\log z + (1-z)\log(1-z) & \text{if } z \in [0,1] \\ \infty & \text{otherwise.} \end{cases}$

Proof: If $a > 1$ then $P(S_n \geq an) = 0$ as $S_n \leq n$ and hence $\log(P\{S_n \geq an\}) = -\infty$ if $a > 1$. So consider $1/2 \leq a \leq 1$.

$$P(S_n \geq an) = \sum_{an \leq k \leq n} 2^{-n} \cdot {}^nC_k$$

Put $Q_n(a) = \max_{an \leq k \leq n} {}^nC_k$

Then $2^{-n} Q_n(a) \leq P(S_n \geq an) \leq 2^{-n}(n+1) Q_n(a)$

[∵ at least one term equal to $Q_n(a)$ is present in the sum defining $P(S_n \geq an)$ and at most there are $(n+1)$ terms in the sum all equal to $Q_n(a)$.]

Claim: $\lim_{n\to\infty} \frac{1}{n} \log Q_n(a) = -a\log(a) - (1-a)\log(1-a)$

We'd prove the present theorem assuming the above claim:

$$\frac{1}{n}\log 2^{-n} Q_n(a) \leq \frac{1}{n}\log(P\{S_n \geq an\}) \leq \frac{1}{n}\log 2^{-n}(n+1)Q_n(a)$$

[∵ log is an increasing function]

$$\Rightarrow -\log 2 + \frac{1}{n}\log Q_n(a) \leq \frac{1}{n}\log(P\{S_n \geq an\})$$
$$\leq -\log 2 + \frac{\log(n+1)}{n} + \frac{\log Q_n(a)}{n}$$

Now, $\lim_{n\to\infty} \frac{\log(n+1)}{n} = 0$ and hence using the Sandwich theorem we get,

$$\lim_{n\to\infty} \frac{1}{n}\log(P\{S_n \geq an\}) = -\log 2 + \lim_{n\to\infty}\frac{1}{n}\log Q_n(a)$$
$$= -\log 2 - a\log a - (1-a)\log(1-a).$$

This completes the proof.

Proof of the claim:

$$Q_n(a) = \max_{an \leq k \leq n} {}^nC_k = {}^nC_{\lceil an \rceil} \text{ as } a > 1/2$$

as ${}^nC_k$ is greatest around $n/2$.

[where $\lceil an \rceil$ is the smallest integer greater than or equal to an.]

$$\Rightarrow Q_n(a) = \frac{n!}{\lceil an \rceil ! \, (n - \lceil an \rceil)!}$$

By Stirling's formula for $n \to \infty$, $n! \sim (n/e)^n \sqrt{2\pi n}$

[we'll say $a_n \sim b_n$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$]

Thus we have that,

$$\frac{n!}{\lceil an \rceil ! \, (n - \lceil an \rceil)!} \sim \frac{(n/e)^n \sqrt{2\pi n}}{\left(\frac{\lceil an \rceil}{e}\right)^{\lceil an \rceil} \sqrt{2\pi \lceil an \rceil} \cdot \left(\frac{n - \lceil an \rceil}{e}\right)^{(n - \lceil an \rceil)} \sqrt{2\pi (n - \lceil an \rceil)}}$$

$$\left[\because\ n \to \infty \Rightarrow \lceil an\rceil \to \infty \text{ and } (n - \lceil an\rceil) \to \infty\right]$$

$$= \frac{n^{n+\frac{1}{2}}}{(\lceil an\rceil)^{\lceil an\rceil + \frac{1}{2}}\ (n - \lceil an\rceil)^{(n - \lceil an\rceil + \frac{1}{2})}\ \sqrt{2\pi}}$$

So for large enough n,

$$\frac{1}{n}\log(Q_n(a)) \sim \frac{1}{n}\Big\{(n+\tfrac{1}{2})\log n - (\lceil an\rceil + \tfrac{1}{2})\log\lceil an\rceil - (n - \lceil an\rceil + \tfrac{1}{2})\log(n - \lceil an\rceil) - \log\sqrt{2\pi}\Big\}$$

$$= \frac{1}{2}\left(\frac{\log n}{n} - \frac{\log\lceil an\rceil}{n} - \frac{\log(n - \lceil an\rceil)}{n}\right) - \frac{1}{n}\log\sqrt{2\pi} - \frac{\lceil an\rceil}{n}\log\frac{\lceil an\rceil}{n} - \frac{(n - \lceil an\rceil)}{n}\log\frac{(n - \lceil an\rceil)}{n}$$

Now for $n \to \infty$, $\frac{\lceil an\rceil}{n} = a$ and $\frac{n - \lceil an\rceil}{n} = 1 - a$.

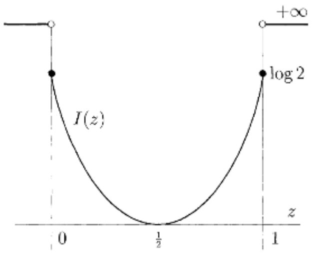
So, $$\lim_{n\to\infty} \frac{1}{n}\log(Q_n(a)) = -a\log a - (1-a)\log(1-a)$$

$$\left[\because \lim_{n\to\infty}\frac{\log n}{n} = 0 \text{ and } \lceil an\rceil \leq n \Rightarrow \log\lceil an\rceil \leq \log n \Rightarrow \frac{\log\lceil an\rceil}{n} \leq \frac{\log(n)}{n} \to 0\right]$$

This proves the claim.

Observations:

(i) The graph of the function $I(z)$ is as follows:



Observe that $I(z)$ has a minima at $\frac{1}{2}$ and increases in both directions away from $\frac{1}{2}$. This is to be expected since as a increases from $\frac{1}{2}$ to 1 then we'd intuitively have a faster exponential decay for the probability $P(S_n \geq an)$

(ii) The symmetry of the graph about the line $z = \frac{1}{2}$ suggests that a similar result holds for a large deviation on the right side of the mean as well. Stated precisely we have

$$\lim_{n\to\infty} \frac{1}{n} \log \{ P(S_n \le an) \} = -I(a) \quad \text{with} \quad a \le \frac{1}{2}$$

But a rigorous proof is required.

(iii) We provide a proof of SLLN for Ber($\frac{1}{2}$) as a corollary of the above theorem but for that we'd need the following lemma from basic probability.

Borel – Cantelli Lemma: Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{A_n\}_{n\ge 1}$ is a sequence of events. Say

$$\sum_{k=1}^{\infty} P(A_k) < \infty$$

then the probability that infinitely many of the events A_n occur is 0. i.e.,

$$P\Big(\bigcap_{n=1}^{\infty} \big(\bigcup_{k\ge n} A_k\big)\Big) = 0$$

Proof of SLLN for Ber($\frac{1}{2}$) random variables:

We know that $\lim_{n\to\infty} \frac{1}{n} \log (P\{S_n \ge an\}) = -I(a)$

Given $\frac{1}{2} > \delta > 0$, let $\varepsilon > 0$ be such that $\varepsilon < I(\frac{1}{2}+\delta) < \infty$

say $n_0 \in \mathbb{N}$ be such that $n \ge n_0$ implies

$$\left| \frac{1}{n} \log (P\{S_n \ge (\tfrac{1}{2}+\delta)n\}) + I(\tfrac{1}{2}+\delta) \right| < \varepsilon$$

and $\left| \frac{1}{n} \log (P\{S_n \le (\frac{1}{2}-\delta)n\}) + I(\frac{1}{2}-\delta) \right| < \varepsilon$.

$\Rightarrow$ for $n \ge n_0$,

$P\{S_n \ge (\frac{1}{2}+\delta)n\} \le e^{n\varepsilon} e^{-nc}$

and $P\{S_n \le (\frac{1}{2}-\delta)n\} \le e^{n\varepsilon} e^{-nc}$

[where $c = I(\frac{1}{2}+\delta) = I(\frac{1}{2}-\delta)$]

So, $\displaystyle\sum_{n=1}^{k} P\{|S_n/n - \tfrac{1}{2}| \ge \delta\}$ (for $k > n_0$)

$$= \sum_{n=1}^{n_0} P\{|S_n/n - \tfrac{1}{2}| \ge \delta\} + \sum_{n=n_0+1}^{k} P\{|S_n/n - \tfrac{1}{2}| \ge \delta\}$$

$$\leq \sum_{n=1}^{n_0} P\left\{ \left|\frac{S_n}{n} - \frac{1}{2}\right| \geq \delta \right\} + \sum_{n=n_0+1}^{K} 2e^{-n(c-\varepsilon)}$$

$$< \infty \qquad [\because \; c - \varepsilon > 0.]$$

Thus $\displaystyle\sum_{n=1}^{\infty} P\left\{ \left|\frac{S_n}{n} - \frac{1}{2}\right| \geq \delta \right\}$

Now, if we denote the event $\left\{ \left|\frac{S_n}{n} - \frac{1}{2}\right| \geq \delta \right\}$ by A_n^{δ} then using the Borel - Cantelli lemma we get,

$$P\left(\bigcap_{n=1}^{\infty} \left(\bigcup_{k \geq n} A_k^{\delta} \right) \right) = 0$$

So, $\displaystyle P\left(\bigcup_{m=1}^{\infty} \left(\bigcap_{n=1}^{\infty} \left(\bigcup_{k \geq n} A_k^{1/m} \right) \right) \right) \leq \sum_{m=1}^{\infty} P\left(\bigcap_{n=1}^{\infty} \left(\bigcup_{k \geq n} A_k^{1/m} \right) \right) = 0$

Now, $\displaystyle\left\{ \lim_{n \to \infty} \frac{S_n}{n} = n \right\} = \left(\bigcup_{m=1}^{\infty} \left(\bigcap_{n=1}^{\infty} \left(\bigcup_{k \geq n} A_k^{1/m} \right) \right) \right)^c$

$\Rightarrow \displaystyle P\left\{ \lim_{n \to \infty} \frac{S_n}{n} = n \right\}$

$$= 1 - P\left\{ \bigcup_{m=1}^{\infty} \left(\bigcap_{n=1}^{\infty} \left(\bigcup_{k \geq n} A_k^{1/m} \right) \right) \right\}$$

$$= 1$$

Thus $\displaystyle P\left(\lim_{n \to \infty} \frac{S_n}{n} = \kappa \right) = 1$

A generalization to random walks with finite M.G.F.

Theorem (iv): Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. random variables with $\phi(t) = E(e^{tX_1}) < \infty \quad \forall t \in \mathbb{R}$

Define: $S_n = X_1 + X_2 + \cdots + X_n, \quad n \in \mathbb{N}$

Then for $a > E[X_1]$

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \geq an) = -I(a)$$

where $I(z) = \sup_{t\in\mathbb{R}} (zt - \log(\phi(t)))$

Proof: We can assume $a=0$ and $E(X_1) < 0$ by considering the following translation of X_i's: $\tilde{X}_i := X_i - a$

Then $\{\tilde{X}_i\}_{i\geq 1}$ are i.i.d. as well and

$$\tilde{\phi}(t) = E(e^{t\tilde{X}_1}) = E(e^{tX_1 - ta}) = e^{-ta}\phi(t)$$

So, $\tilde{I}(0) = \sup_{t\in\mathbb{R}} (-\log(\tilde{\phi}(t)))$

$$= \sup_{t\in\mathbb{R}} (at - \log(\phi(t))) = I(a)$$

So WLOG assume $E[X_1] < 0$ and $a = 0$

Suppose X_i's are degenerate as say $P(X_1 = c) = 1$, then $I(c) = \sup_{t\in\mathbb{R}} (ct - \log(e^{ct})) = 0$

and if $z \neq a$ $I(z) = \sup_{t\in\mathbb{R}} (zt - \log(e^{ct})) = \infty$

and in that case $P(S_n \geq an) = \begin{cases} 1 & \text{if } a = c \\ 0 & \text{if } a > c \end{cases}$

[Note that $E[X_1] = c$ and so we didn't consider $a < c$ above.]

So consider X_i's to be non-degenerate.

Now say $\rho = \inf_{t\in\mathbb{R}} \phi(t)$ and note that

$I(0) = -\log \rho$ [take $I(0) = \infty$ if $\rho = 0$.]

We'd have to prove that $\lim_{n\to\infty} \frac{1}{n} \log\{P(S_n \geq 0)\} = \log(\rho)$

Say $F(x) = P(X_1 \leq x)$ be the cdf of X_1

then $\phi'(t) = \int_{-\infty}^{\infty} x e^{tx}\, dF(x)$

and $\phi''(t) = \int_{-\infty}^{\infty} x^2 e^{tx}\, dF(x)$ [$dF(x)$ is such that we'd get appropriate expressions for discrete and continuous random variables]

Note that $\phi''(t) > 0$ [$\because X_1$ is non-degenerate] and so $\phi(t)$ is strictly convex.

Let us now consider 3 separate cases:

(i) $P(X_1 < 0) = 1$, then $\phi'(t) < 0 \quad \forall t$ and hence $\phi(t)$ is strictly decreasing and $\lim_{t\to\infty} \phi(t) = \rho = 0$ and in that case $P(S_n \geq 0) = 0$

$$\Rightarrow \frac{1}{n} \log\{P(S_n \geq 0)\} = -\infty \quad \forall n$$

$$\Rightarrow \lim_{n\to\infty} \frac{1}{n} \log\{P(S_n \geq 0)\} = \log \rho = -\infty.$$

(ii) If $P(X_i \le 0) = 1$ and $P(X_1 = 0) > 0$, then again $\phi'(t) < 0 \;\; \forall t$ as X_1 is non-degenerate and $\lim_{t \to \infty} \phi(t) = \rho = P(X_1 = 0)$ and then,

$$P(S_n \ge 0) = P(X_1 = X_2 = \cdots = X_n = 0) = \rho^n$$

$$\Rightarrow \lim_{n \to \infty} \frac{1}{n} \log \{P(S_n \ge 0)\} = \log \rho$$

(iii) If $P(X_1 < 0) > 0$ and $P(X_1 > 0) > 0$ then, $\lim_{t \to \pm\infty} \phi(t) = \infty$

and as $\phi(t)$ is strictly convex, $\exists$ a unique $\tau > 0$ with $\phi'(\tau) = 0$ and $\phi(\tau) = \rho$

$[\tau > 0$ as $\phi'(t) < 0 \;\; \forall t < 0]$

Hence, $P(S_n \ge 0) = P(e^{\tau S_n} \ge 1) \overset{(*)}{\le} \dfrac{E(e^{\tau S_n})}{1}$

$$= E\left(e^{\tau \left(\sum_{i=1}^{n} X_i\right)}\right) = (\phi(\tau))^n = \rho^n$$

[(*) By chebyshev's inequality.]

$$\Rightarrow \limsup_{n \to \infty} \frac{1}{n} \log \{P(S_n \ge 0)\} \le \log \rho$$

To prove the other inequality, consider the i.i.d. sequence $\{\hat{X}_i\}_{i \ge 1}$ of random variables with c.d.f. given by:

$$\hat{F}(x) = \frac{1}{\rho} \int_{(-\infty, x]} e^{\tau y} \, dF(y)$$

$\hat{F}$ is called the Cramer-transform of F.

Define $\hat{S}_n := \hat{X}_1 + \hat{X}_2 + \cdots + \hat{X}_n$.

$$P(S_n \ge 0) = \int_{\{x_1 + x_2 + \cdots + x_n \ge 0\}} dF(x_1)\, dF(x_2) \cdots dF(x_n)$$

[$\because X_i$'s are independent]

$$= \int_{\{x_1 + x_2 + \cdots + x_n \ge 0\}} \left(\rho e^{-\tau x_1} d\tilde{F}(x_1)\right)\left(\rho e^{-\tau x_2} d\tilde{F}(x_2)\right) \cdots \left(\rho e^{-\tau x_n} d\tilde{F}(x_n)\right)$$

$$= \rho^n E\left(e^{-\tau \tilde{S}_n} 1_{\{\tilde{S}_n \ge 0\}}\right) \quad \cdots\cdots \quad (i)$$

Now, $\tilde{\phi}(t) = E(e^{t\hat{X}_1}) = \frac{1}{\rho} \phi(t + \tau) < \infty$

and hence $E[\hat{X}_1] = \tilde{\phi}'(0) = \frac{1}{\rho} \phi'(0 + \tau) = 0$

and $\sigma = Var[\tilde{X}_1] = \tilde{\phi}''(0) = \frac{1}{\rho} \phi''(0 + \tau) \in (0, \infty)$

So we can apply CLT to $\tilde{S}_n$ i.e.,

$$\tilde{S}_n / \sigma\sqrt{n} \xrightarrow{dist.} N(0, 1)$$

Say $c > 0$ be such that $\frac{1}{\sqrt{2\pi}} \int_0^c e^{-x^2/2} \, dx > 1/4$

Thus, $E\left[e^{-\tau \tilde{S}_n} 1_{\{\tilde{S}_n \geq 0\}}\right] =$

$$= \int_{\{x_1+x_2+\cdots+x_n \geq 0\}} \left(e^{-\tau x_1}\, d\tilde{F}(x_1)\right)\left(e^{-\tau x_2}\, d\tilde{F}(x_2)\right)\cdots\left(e^{-\tau x_n}\, d\tilde{F}(x_n)\right)$$

$$\geq \int_{\{c\sigma\sqrt{n} \geq x_1+x_1+\cdots+x_n \geq 0\}} \left(e^{-\tau x_1}\, d\tilde{F}(x_1)\right)\left(e^{-\tau x_2}\, d\tilde{F}(x_2)\right)\cdots\left(e^{-\tau x_n}\, d\tilde{F}(x_n)\right)$$

$$\geq e^{-\tau c\sigma\sqrt{n}}\, P\left(\frac{S_n}{\sigma\sqrt{n}} \in [0, c)\right)$$

and the last probability is at least $\frac{1}{4}$ for large enough n. Hence we have,

$$\liminf_{n\to\infty} \frac{1}{n} \log E\left(e^{-\tau S_n} 1_{\{S_n \geq 0\}}\right) \geq \lim_{n\to\infty} \left(-\frac{\tau c\sigma}{\sqrt{n}} - \frac{\log 4}{n}\right) = 0$$

So from (i) $\liminf_{n\to\infty} \frac{1}{n} \log\{P(S_n \geq 0)\} \geq \log \rho$

Hence, $\lim_{n\to\infty} \frac{1}{n} \log\{P(S_n \geq 0)\} = \log \rho$

Note that $P(X_1 \geq 0)$ isn't possible as $E[X_1] < 0$

We now conclude with an application:

Consider an insurance company with n customers. Let X_i be the random variable denoting the amount the i^{th} customer claims. Assume that X_i's are i.i.d random variables with finite mean μ and finite variance. Say, the company charges each customer a premium $a > \mu$. Then the total capital of the company $= an$. Then the event that the company gets bankrupt is:

$$\{X_1 + X_2 + \cdots + X_n > an\}$$

which is a large deviation event. The probability of the above event decays exponentially. So perhaps it'd be safer to buy an insurance policy from a company with a larger customer base.

References:

Frank den Hollander – 'Large deviations' American Mathematical Society, RI, 2000.