Large deviations: introductory notes

S. Ramasubramanian
Statistics and Mathematics Unit
Indian Statistical Institute
Bangalore - 560 059

1 Introduction

Large deviations is a part of probability theory. It provides asymptotic estimates for probabilities of certain types of rare events; these probabilities can basically be expressed on an exponential scale. It may be pointed out that the law of large numbers and the central limit theorem, the classical limit theorems of probability theory, concern typical events.

2 Heuristics

Let $X_1, X_2, \ldots$ be real valued independent identically distributed random variables (i.i.d.r.v.'s, for short) with common distribution function F. So $F(x) = P(X_i \leq x),\ x \in \mathbb{R}$; let μ_F denote the probability measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ associated with F; note that $\mu_F = PX_i^{-1}$, $\mu_F(B) = P(X_i \in B)$, $B \in \mathcal{B}_{\mathbb{R}}$ for each i. Assume that the mean (or expectation) exists; so $\int |x|\ d\mu_F(x) < \infty$. Denote the mean by m; note that $m = E(X_i) = \int\limits_{\Omega} X_i(\omega)\ dP(\omega) = \int\limits_{\mathbb{R}} x\ d\mu_F(x) = \int\limits_{\mathbb{R}} x\ dF(x)$.

We write $S_n = X_1 + X_2 + \ldots + X_n$ for $n \geq 1$; so $\{S_n\}$ is the sequence of partial sums. We know by the strong law of large numbers that $\frac{1}{n}\ S_n \to m$ with probability one (that is, almost surely) as $n \to \infty$. So for large n, the random variable $\frac{1}{n}\ S_n$ is very likely to take value in a small neighbourhood of m. In other words, $\{|\frac{1}{n}\ S_n - m| < \epsilon\}$, where $\epsilon > 0$, is an example of a typical event. In addition, if $\sigma^2 = \mathrm{Var}(X_i)$ exists, then using the central

limit theorem it can be shown that for any $\epsilon > 0$

$$\lim_{n\to\infty} \left| P\left\{ \left| \frac{1}{n}\, S_n - m \right| < \epsilon \right\} - \int_{-a_n}^{a_n} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2} dx \right| = 0 \qquad (2.1)$$

where $a_n = \sqrt{n}\ \epsilon/\sigma$. So the probability of above mentioned typical event can be estimated. (To prove (2.1) one needs to use the fact: If $G_n,\ n \geq 1$, G are distribution functions on $\mathbb{R}$, $G_n \to G$ in distribution as $n \to \infty$, and $G(\cdot)$ is continuous, then $\sup_x\ |G_n(x) - G(x)| \to 0$ as $n \to \infty$.)

Let $a > m$. Then by the strong law of large numbers

$$\lim_{n\to\infty}\ P\left(\frac{1}{n}\, S_n \geq a \right) = 0. \qquad (2.2)$$

A question of interest is the rate of convergence in the above. In fact, the question arose in insurance, and was investigated first by Esscher and later by Cramer. Note that $\{\frac{1}{n}\ S_n \geq a\}$ is not a typical event as its probability of occurrence becomes very small as n increases. Can the central limit theorem answer the question?

For simplicity assume that variance exists and $\mathrm{Var}(X_i) = 1$. Denote by Φ the distribution function of the $N(0,1)$ distribution, that is,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}y^2}\ dy,\ x \in \mathbb{R}.$$

By the central limit theorem we get

$$\begin{aligned} P\left(\frac{1}{n}\, S_n \geq a \right) &= P\left(\frac{S_n - nm}{\sqrt{n}} \geq \sqrt{n}(a-m) \right) \\ &\approx 1 - \Phi(\sqrt{n}(a-m)) \to 0,\ \text{as } n \to \infty. \end{aligned}$$

So the (Lindeberg-Levy form of) central limit theorem is not sensitive enough to give information about the rate.

We now consider two examples where we can say something.

Example 2.1 Let $\{X_i : i \geq 1\}$ be a sequence of i.i.d. Bernoulli random variables with parameter $\frac{1}{2}$. So $P(X_i = 0) = \frac{1}{2} = P(X_i = 1)$ for all i, and

$m = \frac{1}{2}$. We know that $S_n = X_1 + \ldots + X_n$ has a binomial distribution with parameters n and $\frac{1}{2}$.

Let $a \in (\frac{1}{2}, 1)$ and $Q_n(a) = \max\left\{ \begin{pmatrix} n \\ k \end{pmatrix} : k \geq an \right\}$. Note that maximum is attained at $k_0 = [an]$ = smallest integer $\geq an$. (To prove this, consider when the ratio $\begin{pmatrix} n \\ k \end{pmatrix} \Big/ \begin{pmatrix} n \\ k+1 \end{pmatrix}$ can exceed 1 as k varies). It is clear that

$$\frac{1}{2^n}\ Q_n(a) \leq P(S_n \geq an) \leq \frac{(n+1)}{2^n}\ Q_n(a). \qquad (\ 2.3\ )$$

Because of the two sided estimate (2.3) involving $Q_n(a)$ it may be fruitful to analyse the asymptotic behaviour of $Q_n(a)$. As factorials are involved, Stirling's formula comes to mind, which says that

$$(n!) \sim \sqrt{2\pi}\ e^{-n}\ n^{(n+\frac{1}{2})},\ \text{as}\ n \uparrow \infty. \qquad (\ 2.4\ )$$

Recall that the above means that the ratio of the l.h.s. and r.h.s. converges to 1 as $n \to \infty$. As $\frac{1}{2} < a < 1$, note that $k_0 \to \infty$, $(n - k_0) \to \infty$ as $n \to \infty$. So one can use Stirling approximation for $(n!)$, $(k_0!)$, $((n-k_0)!)$ as $n \to \infty$. Also the r.h.s. of (2.4) hints that it may be more convenient to look at $\log\ Q_n(a)$. Using (2.4) it is easily seen that

$$\lim_{n\to\infty}\ \frac{1}{n}\ \log Q_n(a) = \lim_{n\to\infty}\ \frac{1}{n}\ \log\left[\frac{n^{n+\frac{1}{2}}}{k_0^{(k_0+\frac{1}{2})}(n-k_0)^{(n-k_0+\frac{1}{2})}}\right]. \qquad (\ 2.5\ )$$

Put $\epsilon_n = an - k_0 = an - [an]$; clearly $0 \leq \epsilon_n \leq 1$. It is not difficult to check that

$$k_0^{k_0+\frac{1}{2}} = [an]^{[an]+\frac{1}{2}} = a^{(an+\frac{1}{2}-\epsilon_n)}\ n^{(an+\frac{1}{2}-\epsilon_n)}\left(1 - \frac{\epsilon_n}{an}\right)^{an+\frac{1}{2}-\epsilon_n}.$$

In a similar manner one can get analogous expression for $(n-k_0)^{(n-k_0+\frac{1}{2})}$ by replacing $a, -\epsilon_n$ respectively by $(1-a), +\epsilon_n$ on the right side. Since an and $(1-a)n$ go to $+\infty$ as $n \to \infty$, and $\{\epsilon_n\}$ is a bounded sequence, (2.5) now becomes

$$\begin{aligned} \lim_{n\to\infty}\ \frac{1}{n}\log\ Q_n(a) &= -a\log a - (1-a)\log(1-a) + \lim_{n\to\infty}\ O\left(\frac{\log n}{n}\right) \\ &= -a\log a - (1-a)\ \log(1-a). \qquad (\ 2.6\ ) \end{aligned}$$

From (2.3), (2.6) we get for $a \in (\frac{1}{2}, 1)$

$$\begin{aligned} & \lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \geq a\right) \\ = & -[\log 2 + a \log a + (1-a) \log(1-a)]. \end{aligned} \quad (2.7)$$

Now suppose $0 < a < \frac{1}{2}$. Using symmetry about $\frac{1}{2}$, (that is, reversing the roles of success and failure), and (2.7)

$$\begin{aligned} & \lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \leq a\right) \\ = & \lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \geq (1-a)\right) \\ = & -[\log 2 + a \log a + (1-a)\log(1-a)]. \end{aligned} \quad (2.8)$$

It is easy to see that (2.7) (resp. (2.8)) holds when $a = 1$ (resp. $a = 0$).

Note: What happens if $a = \frac{1}{2}$ in the above?

Example 2.2 Let $\{X_i\}$ be a sequence of i.i.d. $N(0,1)$ random variables. We know that the empirical mean $\frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ has $N(0, \frac{1}{n})$ distribution. Denote the distribution function and the probability density function of $N(0,1)$ distribution by Φ and φ respectively. For $a > 0$ note that

$$P\left(\left|\frac{1}{n} S_n\right| \geq a\right) = 2[1 - \Phi(a\sqrt{n})]. \quad (2.9)$$

For $y > 0$ note that

$$\left(1 - \frac{3}{y^4}\right) \varphi(y) < \varphi(y) < \left(1 + \frac{1}{y^2}\right) \varphi(y).$$

Let $z > 0$. Integrate the above over $[z, \infty)$ to get

$$\left(\frac{1}{z} - \frac{1}{z^3}\right) \varphi(z) < [1 - \Phi(z)] < \frac{1}{z} \varphi(z). \quad (2.10)$$

Putting $z = a\sqrt{n}$ with $a > 0$ it follows that

$$\lim_{n\to\infty} \frac{1}{n} \log\left[\left(\frac{1}{a\sqrt{n}} - \frac{1}{(a\sqrt{n})^3}\right) \varphi(a\sqrt{n})\right] = -\frac{1}{2} a^2, \quad (2.11)$$

$$\lim_{n\to\infty} \frac{1}{n} \log\left[\frac{1}{a\sqrt{n}} \varphi(a\sqrt{n})\right] = -\frac{1}{2} a^2. \quad (2.12)$$

Clearly (2.9) - (2.12) imply that for $a > 0$

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\left|\frac{1}{n} S_n\right| \geq a\right) = -\frac{1}{2} a^2. \qquad (2.13)$$

In view of the two preceding examples we want to see if $P(\frac{1}{n} S_n \geq a)$, where $a > m = E(X_i)$, decays at an exponential rate. And if so, we would like to know the rate.

3 Functions $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$

We continue with the notation of the preceding section. As we are investigating exponential decay, it is reasonable to assume that the moment generating function (or Laplace transform) exists. Assume that

$$\begin{aligned} M(t) &= E[e^{tX_1}] = \int_{\mathbb{R}} e^{tx}\, dF(x) \\ &= \int_{\mathbb{R}} e^{tx}\, d\mu_F(x) < \infty, \ \forall\ t \in \mathbb{R}. \end{aligned} \qquad (3.1)$$

Then we know that $E(X_1)$ exists. Let $a > E(X_1)$. Using Chebyshev inequality we have

$$\begin{aligned} P\left(\frac{1}{n} S_n \geq a\right) &= P\left(\exp\left(\frac{\theta}{n} S_n\right) \geq e^{\theta a}\right) \\ &\leq e^{-\theta a}\, E\left[\exp\left(\frac{\theta}{n} S_n\right)\right] = e^{-\theta a} \left[M\left(\frac{\theta}{n}\right)\right]^n \end{aligned}$$

for any $\theta > 0$. Put $\theta = nt$, $\theta > 0$ being arbitrary, to get

$$\frac{1}{n} \log P\left(\frac{1}{n} S_n \geq a\right) \leq \{-ta + \log M(t)\}$$

for any $t \geq 0$. Consequently

$$\begin{aligned} & \limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \geq a\right) \\ \leq\ & \inf_{t\geq 0} \{-ta + \log M(t)\} \\ =\ & -\sup \{ta - \log M(t) : t \geq 0\} \\ =\ & -\sup \{ta - \log M(t) : t \in \mathbb{R}\}. \end{aligned} \qquad (3.2)$$

The justification for replacing supremum over $[0, \infty)$ by supremum over $\mathbb{R}$ in the last step will be given later in Proposition 3.4; (this is due to $a > E(X_1)$).

Now, in Example 2.2, where X_i's are i.i.d. $N(0,1)$ random variables, $M(t) = \exp\ (\frac{1}{2}\ t^2)$, $t \in \mathbb{R}$. Hence

$$\sup\ \{ta - \log M(t) : t \in \mathbb{R}\} = \frac{1}{2}\ a^2.$$

So in this case, the upper bound got in (3.2) is the same as the rate obtained in (2.13).

Thus we now have a candidate for the "rate". Before we take up our problem, we shall try to understand two objects encountered above.

With $M(\cdot)$ as in (3.1), define

$$\Lambda(\theta) = \log M(\theta) = \log \int e^{\theta x}\ d\mu_F(x),\ \theta \in \mathbb{R}. \qquad (\ 3.3\ )$$

Note that $\Lambda(\cdot)$ is well defined and $-\infty < \Lambda(\theta) < \infty$ for all $\theta \in \mathbb{R}$. Denote $m = E(X) = \int\limits_{\mathbb{R}} x\ d\mu_F(x)$. The function $\Lambda(\cdot)$ is called the *logarithmic moment generating function* of the distribution μ_F.

Proposition 3.1 *$\theta \mapsto \Lambda(\theta)$ is convex.*

Proof: Let $\theta_1, \theta_2 \in \mathbb{R}$, $0 < c < 1$. By Holder's inequality applied to $\frac{1}{c}$ we get

$$\begin{aligned}
& \int \exp[(c\theta_1 + (1-c)\theta_2)x]\ d\mu_F(x) \\
= & \int e^{c\theta_1 x}\ e^{(1-c)\theta_2 x}\ d\mu_F(x) \\
\leq & \left[\int (e^{c\theta_1 x})^{\frac{1}{c}}\ d\mu_F(x)\right]^c\ \left[\int \left(e^{(1-c)\theta_2 x}\right)^{\frac{1}{1-c}}\ d\mu_F(x)\right]^{(1-c)} \\
= & \left[\int e^{\theta_1 x}\ d\mu_F(x)\right]^c\ \left[\int e^{\theta_2 x}\ d\mu_F(x)\right]^{1-c}.
\end{aligned}$$

Taking log on both sides we get the result. □

For a convex function, a useful (or rather dual) notion is the *Fenchel-Legendre transform*, which has its roots in the calculus of variations. Define the Fenchel-Legendre transform of $\Lambda(\cdot)$ by

$$\begin{aligned} \Lambda^*(x) &= \sup \{\theta x - \Lambda(\theta) : \theta \in \mathbb{R}\} \\ &= \sup \{\theta x - \log M(\theta) : \theta \in \mathbb{R}\}. \end{aligned} \qquad (3.4)$$

Before investigating the properties of $\Lambda^*(\cdot)$, we need a preliminary result; we state it in a general context to be useful later as well.

Proposition 3.2 *Let (S, d) be a complete separable metric space. Let $f : S \to [-\infty, \infty]$ be a function; (so f can possibly take values $+\infty$, $-\infty$ also). Then the following are equivalent:*
(i) $\liminf_{n\to\infty} f(x_n) \geq f(x)$ *whenever* $x_n \to x$.
(ii) $\lim_{\epsilon \downarrow 0} \inf_{y \in B_\epsilon(x)} f(y) = f(x)$, *for any* x.
(iii) $f^{-1}([-\infty, c]) = \{x \in S : f(x) \leq c\}$ *is a closed set for any* $c \in \mathbb{R}$; *that is, f has closed level sets.*

A function f satisfying the above equivalent properties is said to be lower-semicontinuous.

Proof: (i) $\Rightarrow$ (ii): Suppose (ii) does not hold; that is, $\lim_{\epsilon \downarrow 0} \inf_{y \in B_\epsilon(x)} f(y) < f(x)$; (note that strict inequality in the other direction cannot happen.) Then there is a sequence $\{x_n\}$ and $\delta > 0$ with $x_n \to x$ such that $f(x_n) < f(x) - \delta$ for all n. This contradicts (i).

(ii) $\Rightarrow$ (iii): Fix $c \in \mathbb{R}$. Note that the set $\{x \in S : f(x) > c\}$ is open, because (ii) and $f(x) > c$ imply $f(y) > c$ for all $y \in B_\epsilon(x)$ with $\epsilon > 0$ sufficiently small.

(iii) $\Rightarrow$ (i): Suppose (i) fails; that is, $\exists$ a sequence x_n such that $x_n \to x$, $f(x_n) \to c < f(x)$. Let $0 < \delta < f(x) - c$. Then $x_n \in \{y \in S : f(y) \leq c + \delta\}$ which is a closed set (by (iii)). So $f(x) \leq c + \delta$. This contradicts the choice of δ. $\square$

Proposition 3.3 *(i) $x \mapsto \Lambda^*(x)$ is a nonnegative convex function. (It can possibly take the value $+\infty$.) (ii) $x \mapsto \Lambda^*(x)$ is a lower semicontinuous function.*

Proof: (i) $\Lambda(0) = 0$ and hence $\Lambda^*(\cdot) \geq 0$ by definition (3.4). As $\sup\{a(\theta) + b(\theta)\} \leq \sup\{a(\theta)\} + \sup\{b(\theta)\}$, convexity of $\Lambda^*(\cdot)$ can be directly verified.

(ii) For fixed $\theta \in \mathbb{R}$, as $-\infty < \Lambda(\theta) < \infty$, note that $x \mapsto \theta x - \Lambda(\theta)$ is continuous. Note that supremum h of a family $\{h_\alpha\}$ of continuous functions is lower semicontinuous, because $\{h(\cdot) > a\} = \bigcup_\alpha \{h_\alpha > a\}$. □

Proposition 3.4 *(i) If $x \geq m$, then*

$$\Lambda^*(x) = \sup\{\theta x - \Lambda(\theta) : \theta \geq 0\}. \qquad (3.5)$$

(Thus the last step in (3.2) is justified.) Also $\Lambda^*(\cdot)$ is nondecreasing on $[m, \infty)$.

(ii) If $x \leq m$, then $\Lambda^*(x) = \sup\{\theta x - \Lambda(\theta) : \theta \leq 0\}$. Also $\Lambda^*(\cdot)$ is nonincreasing on $(-\infty, m]$.

(iii) $\inf\{\Lambda^*(x) : x \in \mathbb{R}\} = 0 = \Lambda^*(m)$.

Proof: (i) As log is a concave function, by Jensen's inequality for any $\theta \in \mathbb{R}$

$$\Lambda(\theta) = \log \int e^{\theta x}\, d\mu_F(x) \geq \int \log(e^{\theta x})\, d\mu_F(x) = \theta m.$$

So far any $\theta \in \mathbb{R}$ we have $\theta m - \Lambda(\theta) \leq 0$, with equality holding if $\theta = 0$. This shows that $\Lambda^*(m) = 0$.

Now if $x \geq m$, $\theta < 0$ then $\theta x - \Lambda(\theta) \leq \theta m - \Lambda(\theta) \leq 0$. As $\Lambda^*(\cdot) \geq 0$, it follows now that (3.5) holds.

By the above, on $[m, \infty)$ we need to consider only $\theta \geq 0$ in the definition of $\Lambda^*(\cdot)$. Let $y \geq x \geq m$, $\theta \geq 0$. Then $\theta y - \Lambda(\theta) \geq \theta x - \Lambda(\theta)$, and hence we get that $\Lambda^*(\cdot)$ is nondecreasing on $[m, \infty)$.

(ii) Consider the logarithmic moment generating function of $(-X)$, and apply (i)

(iii) is now clear from the above. □

Proposition 3.5 $\Lambda^*(\cdot)$ *has compact level sets; that is,* $\{x \in \mathbb{R} : \Lambda^*(x) \leq a\}$ *is compact for every $a \in [0, \infty)$.*

Proof: By lower semicontinuity we know that $\{\Lambda^* \leq a\}$ is closed; so we need to show that it is bounded. Suppose it is not bounded. Then $\exists$ a sequence $\{z_n\}$ with $\Lambda^*(z_n) \leq a$ for all n, and $z_n \to +\infty$; (the case $z_n \to -\infty$ can be handled similarly). Hence, by definition of Λ^*, we have $z_n t - \log M(t) \leq a$ for all $t \in \mathbb{R}$, $n \geq 1$. Put $t = 2a/z_n$. Then $2a - \log M(\frac{2a}{z_n}) \leq a$ for all n. Let $n \uparrow \infty$ to get $2a - \log M(0) = 2a$. So $2a \leq a$ which is a contradiction for any $a > 0$. Thus $\{\Lambda^*(\cdot) \leq a\}$ is compact for any $a > 0$. Hence $\{\Lambda^* \leq 0\}$ is also compact. $\square$

Proposition 3.6 *Let $\mu_F \neq \delta_0$. Then $\Lambda(\cdot)$ is strictly convex on $\mathbb{R}$.*

Proof: Note that $M'(t) = \int x e^{tx}\, d\mu_F(x)$, $M''(t) = \int x^2 e^{tx}\, d\mu_F(x)$. Since $\mu_F \neq \delta_0$, note that $M''(t) > 0$ for all t, and hence $M(\cdot)$ is strictly convex on $\mathbb{R}$. Clearly $\theta \mapsto \Lambda(\theta)$ is a smooth function, $\Lambda'(t) = \frac{1}{M(t)} M'(t)$ and

$$\Lambda''(t) = \frac{1}{(M(t))^2}\ \{M(t)M''(t) - (M'(t))^2\}.$$

We need to show that $\Lambda''(t) > 0$. For fixed $t \in \mathbb{R}$, note that $dN_t(x) = \frac{1}{M(t)}\, e^{tx}\, d\mu_F(x)$ is a probability measure on $\mathbb{R}$. It is easy to check that $\Lambda''(t)$ is the variance for this probability measure. Since $dN_t(x)$ is nondegenerate it follows that $\Lambda''(t) > 0$, completing the proof. (Thanks are due to M.G. Nadkarni for the suggestion leading to a shorter proof.) $\square$

Proposition 3.7 *Denote*

$$D_{\Lambda^*} = \{z \in \mathbb{R} : \Lambda^*(z) < \infty\}. \qquad (3.6)$$

Then $D_{\Lambda^*} = Range(\Lambda')$.

Proof: From the preceding proof, $\Lambda'(\cdot)$ is a well defined differentiable function taking values in $\mathbb{R}$. Suppose $\mu_F = \delta_a$ for some $a \in \mathbb{R}$. Then $\Lambda^*(a) = 0$, $\Lambda^*(x) = +\infty$, $x \neq a$. So $D_{\Lambda^*} = \{a\}$. It is easy to check that Range $(\Lambda') = \{a\}$.

Let μ_F be not degenerate. For fixed $z \in \mathbb{R}$, define

$$g_z(\theta) = z\theta - \Lambda(\theta), \quad \theta \in \mathbb{R}. \qquad (3.7)$$

By the preceding proposition $g_z(\cdot)$ is a strictly concave function on $\mathbb{R}$.

Now let $x \in D_{\Lambda^*}$. Then $\Lambda^*(x) = \sup\{g_x(\theta) : \theta \in \mathbb{R}\} < \infty$. Note that $g_x'(\theta) = 0 \Leftrightarrow \Lambda'(\theta) = x$. Since $g_x(\cdot)$ is strictly concave, it now follows that g_x has a unique maximum point, which is the point θ such that $\Lambda'(\theta) = x$. Thus $x \in$ Range(Λ').

Conversely let $x \in$ Range(Λ'). Then $\exists\ \tilde{\theta} \in \mathbb{R}$ with $\Lambda'(\tilde{\theta}) = x$. Note that $g_x'(\tilde{\theta}) = 0$ and hence $\tilde{\theta}$ is a local optimum for g_x. By strict concavity of g_x we see that $\tilde{\theta}$ is the unique global maximum for $g_x(\cdot)$. Hence $\Lambda^*(x) = g_x(\tilde{\theta}) = x\tilde{\theta} - \Lambda(\tilde{\theta})$ which is finite. So $x \in D_{\Lambda^*}$. □

Proposition 3.8 *With notation as in the preceding proposition, $\Lambda^*(\cdot)$ is a continuous, strictly convex and twice differentiable function on Int(D_{Λ^*}), which is the interior of D_{Λ^*}.*

Proof: Enough to consider the case when μ_F is nondegenerate. By the proof of the preceding proposition, for $z \in \text{Int}(D_{\Lambda^*})$ there is a unique point $\tau(z)$ such that $\Lambda'(\tau(z)) = z$. Moreover $\Lambda^*(z) = z\tau(z) - \Lambda(\tau(z))$. By Proposition 3.6, $\Lambda'(\cdot)$ is differentiable, $\Lambda^{''}(\theta) > 0$ for all θ. Consequently $z \mapsto \tau(z)$ is differentiable on $\text{Int}(D_{\Lambda^*})$ and

$$\tau'(z) = \frac{1}{\Lambda''(\tau(z))} > 0, \quad z \in \text{Int}(D_{\Lambda^*}). \tag{3.8}$$

So $\Lambda^*(\cdot)$ is differentiable on $\text{Int}(D_{\Lambda^*})$ and

$$\begin{aligned} (\Lambda^*)'(z) &= \tau(z) + \tau'(z)[z - \Lambda'(\tau(z))] \\ &= \tau(z), \quad z \in \text{Int}(D_{\Lambda^*}). \end{aligned} \tag{3.9}$$

As τ is differentiable, by (3.8), (3.9) it now follows that Λ^* is twice differentiable and $(\Lambda^*)^{''}(\cdot) > 0$ on $\text{Int}(D_{\Lambda^*})$. All the required conclusions now follow. □

4 Cramer's theorem

We now proceed to Cramer's theorem.

Theorem 4.1 *(Cramer): Let $X_1, X_2, \ldots$ be real valued i.i.d. random variables with common distribution μ_F. Assume that (3.1) holds. Denote $S_n = X_1 + \ldots + X_n$, $n \geq 1$. Then for $a > E(X_1)$*

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \geq a\right) = -\Lambda^*(a) \qquad (4.1)$$

where Λ^* is given by (3.4). Similarly for $a < E(X_1)$

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \leq a\right) = -\Lambda^*(a). \qquad (4.2)$$

Proof: It is enough to consider the case $a > E(X_1)$. (Because: Put $Y_i = -X_i$; note that $\Lambda_Y^*(-a) = \Lambda_X^*(a)$.) We may assume that μ_F is nondegenerate. (Why?) Also we may take $a = 0$, $E(X_1) < 0$.

As $M(t) < \infty$ for all $t \in \mathbb{R}$, by Proposition 3.4(i), all the steps in the derivation of (3.2) are rigorous. So we already have the upper bound. Hence it is enough to prove the lower bound (as $a = 0$)

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \geq 0\right) \geq -\Lambda^*(0). \qquad (4.3)$$

Put $\rho = \inf\{M(t) : t \in \mathbb{R}\}$; note that $\Lambda^*(0) = -\log \rho$; (if $\rho = 0$ then $\Lambda^*(0) = +\infty$). So we need to prove

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \geq 0\right) \geq \log \rho. \qquad (4.4)$$

As μ_F is nondegenerate we know that $M(\cdot)$ is strictly convex; also $M'(0) = E(X_1) < 0$.

We now have 3 possible cases. (Why?)

Case (i): $P(X_i < 0) = 1$. In this case $M(\cdot)$ is strictly decreasing, $\lim_{t\to\infty} M(t) = 0$; hence $\log \rho = -\infty$. As $P(\frac{1}{n} S_n \geq 0) = 0$, (4.4) clearly holds.

Case (ii): $P(X_i \leq 0) = 1$, $P(X_i = 0) > 0$. Since X_i is nondegenerate, again $M(\cdot)$ is strictly decreasing. Note that $\lim_{t\to\infty} M(t) = \inf_t M(t) = \rho = P(X_i = 0) > 0$. Therefore for any $n = 1, 2, \ldots$ we have $P(\frac{1}{n} S_n \geq 0) = P(\frac{1}{n} S_n = 0) = P(X_1 = 0, \ldots, X_n = 0) = \rho^n$. Clearly (4.4) follows.

Case (iii): $P(X_1 < 0) > 0,\ P(X_1 > 0) > 0$. In this case note that $\lim_{t\to\infty} M(t) = \infty,\ \lim_{t\to-\infty} M(t) = +\infty$. Since $M(\cdot)$ is strictly convex there exists a unique minimum point $\tau \in \mathbb{R}$ so that $M(\tau) = \rho,\ M'(\tau) = 0$. It remains to prove (4.4) in case (iii). For this we are going to make a change of measure. We have encountered a similar change in the proof of Proposition 3.6. Put

$$\hat{F}(x) = \frac{1}{\rho} \int\limits_{(-\infty,x]} e^{\tau y}\, dF(y), \quad x \in \mathbb{R}; \tag{4.5}$$

that is,

$$\mu_{\hat{F}}(A) = \frac{1}{M(\tau)} \int\limits_{A} e^{\tau y}\, d\mu_F(y), \quad A \in \mathcal{B}(\mathbb{R}). \tag{4.6}$$

It is easily verified that $\hat{F}$ is a distribution function on $\mathbb{R}$; equivalently $\mu_{\hat{F}}$ is a probability measure. The distribution $\mu_{\hat{F}}$ (or equivalently the distribution function F) is called an *Esscher transform* (or a *Cramer transform* or an *exponential tilting*) of the distribution μ_F (or of $F(\cdot)$).

Let $\hat{X}_1, \hat{X}_2, \ldots$ be an i.i.d. sequence with common distribution $\mu_{\hat{F}}$. We need a few lemmas.

Lemma 4.2 *$E(\hat{X}_i) = \hat{m} = 0$, $Var(\hat{X}_i) = \hat{\sigma}^2 \in (0, \infty)$.*

Proof: Easy.

Lemma 4.3 *Put $\hat{S}_n = \hat{X}_1 + \ldots + \hat{X}_n,\ n \geq 1$. Then*

$$\begin{aligned} P\left(\frac{1}{n}S_n \geq 0\right) &= P\left(\sum_{i=1}^{n} X_i \geq 0\right) \\ &= \rho^n E\left[I_{[0,\infty)}(\hat{S}_n) \cdot \exp(-\tau \hat{S}_n)\right]. \end{aligned} \tag{4.7}$$

Proof: Denote $H_n = \{(x_1, \ldots, x_n) : x_1 + \ldots + x_n \geq 0\}$. Then

$$P\left(\frac{1}{n}S_n\right) = P\left(\sum_{i=1}^{n} X_i \geq 0\right) = \int\limits_{H_n} dF(x_1) \ldots dF(x_n)$$

$$
\begin{aligned}
&= \int_{H_n} (\rho e^{-\tau x_1} d\hat{F}(x_1)) \ldots (\rho e^{-\tau x_n} d\hat{F}(x_n)) \\
&= \rho^n \int_{H_n} e^{-\tau(x_1+\ldots+x_n)}\ d\hat{F}(x_1) \ldots d\hat{F}(x_n) \\
&= \text{r.h.s. of (4.7).}
\end{aligned}
$$

□

Lemma 4.4 *We have*

$$\liminf_{n\to\infty} \frac{1}{n} \log E\left[I_{[0,\infty)}(\hat{S}_n) e^{-\tau \hat{S}_n}\right] \geq 0. \qquad (4.8)$$

Proof: By Lemma 4.2, Lindeberg-Levy form of the central limit theorem is applicable to $\hat{X}_1, \hat{X}_2, \ldots$. Choose $C > 0$ so that $\Phi(C) - \Phi(0) > \frac{1}{4}$. Therefore

$$
\begin{aligned}
& E\left[I_{[0,\infty)}(\hat{S}_n)\ \exp(-\tau \hat{S}_n)\right] \\
= \ & E\left[I_{[0,\infty)}\left(\frac{1}{\sqrt{n}\ \hat{\sigma}} \sum_{i=1}^{n} \hat{X}_i\right) \cdot \exp\left(-\tau\sqrt{n}\ \hat{\sigma} \frac{1}{\sqrt{n}\ \hat{\sigma}} \sum_{i=1}^{n} \hat{X}_i\right)\right]. \\
\geq \ & E\left[I_{[0,C]}\left(\frac{1}{\sqrt{n}\ \hat{\sigma}} \sum_{i=1}^{n}\ \hat{X}_i\right) \cdot \exp\left(-\tau\sqrt{n}\ \hat{\sigma} \frac{1}{\sqrt{n}\ \hat{\sigma}} \sum_{i=1}^{n} \hat{X}_i\right)\right] \\
\geq \ & \exp(-\tau\sqrt{n}\ \hat{\sigma}\ C) \cdot P\left(0 \leq \frac{1}{\sqrt{n}\ \hat{\sigma}} \sum_{i=1}^{n} \hat{X}_i \leq C\right) \\
> \ & \frac{1}{2}[\Phi(C) - \Phi(0)] \cdot \exp(-\tau\sqrt{n}\ \hat{\sigma}\ C) \\
> \ & \frac{1}{8}\ \exp(-\tau\sqrt{n}\ \hat{\sigma}\ C) \qquad (4.9)
\end{aligned}
$$

for all large n. The required conclusion (4.8) now follows from (4.9). □

The two preceding lemmas now imply (4.4) in case (iii), completing the proof of Theorem 4.1. □

Remark 4.5 Due to Esscher transform (or exponential tilting) the rare event $S_n \geq 0$ becomes a typical even under the tilted distribution. While the original rare event was too small for Lindeberg-Levy form of the CLT to discern the smallness, the CLT could be effectively applied for the tilted event. □

While Theorem 4.1 gives a fairly sharp estimate, it would be too optimistic to expect such a result in a general context. In the remainder of this section we work towards a version of Cramer's theorem that could be amenable to generalization in more complicated set up.

Proposition 4.6 *Notation and hypotheses are as in Theorem 4.1. For* $A \in \mathcal{B}(\mathbb{R})$*, write*

$$\Lambda^*(A) = \inf\{\Lambda^*(x) : x \in A\}. \qquad (4.10)$$

Then for any $a \in \mathbb{R}$

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in [a,\infty)\right) = -\Lambda^*([a,\infty)), \qquad (4.11)$$

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in (-\infty,a]\right) = -\Lambda^*((-\infty,a]). \qquad (4.12)$$

Proof: By Proposition 3.4, Λ^* is nonincreasing on $(-\infty, m]$ and is nondecreasing on $[m,\infty)$. So (4.11) (resp. (4.12)) follows from Theorem 4.1 if $a > m$ (resp. $a < m$).

Next we shall prove (4.11) where $a \le m$. Using the CLT we get

$$P\left(\frac{1}{n} S_n \ge a\right) \ge P\left(\frac{1}{n} S_n \ge m\right) = P\left(\frac{S_n - mn}{\sqrt{n}\,\sqrt{\mathrm{Var}(X_1)}} \ge 0\right)$$

is bounded away from 0. Therefore l.h.s. of (4.11) is 0. Also r.h.s. of (4.11) is 0 as $\Lambda^*(m) = 0$ and $m \in [a,\infty)$.

In an analogous manner (4.12) can be established when $a \ge m$. □

The next result gives another class of sets for which we have similar expressions.

Proposition 4.7 *Let* D_{Λ^*} *be given by (3.6). Then for any* $a \in \mathbb{R}\backslash(\partial D_{\Lambda^*})$

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in (a,\infty)\right) = -\Lambda^*((a,\infty)) \qquad (4.13)$$

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in (-\infty,a)\right) = -\Lambda^*((-\infty,a)). \qquad (4.14)$$

Proof: By Proposition 3.4 note that D_{Λ^*} is a singleton or an interval. So ∂D_{Λ^*} has at most two points.

First let $a \in \text{Int}(D_{\Lambda^*})$. Then $\exists\ \delta_0 > 0$ such that $a + \delta \in \text{Int}(D_{\Lambda^*})$ for all $0 < \delta < \delta_0$. Applying (4.11) twice we get

$$\begin{aligned} -\Lambda^*([a+\delta,\infty)) &= \lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\ S_n \in [a+\delta,\infty)\right) \\ &\leq \liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\ S_n \in (a,\infty)\right) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\ S_n \in (a,\infty)\right) \\ &\leq \limsup_{n\to\infty} \frac{1}{n}\ \log P\left(\frac{1}{n}\ S_n \in [a,\infty)\right) \\ &= -\Lambda^*([a,\infty)). \end{aligned} \qquad (4.15)$$

By Proposition 3.8, Λ^* is continuous at a; so $\Lambda^*([a,\infty)) = \Lambda^*((a,\infty))$. Also letting $\delta \downarrow 0$ in (4.15) we see that limit on l.h.s. of (4.13) exists and that (4.13) holds.

We know that $m \in D_{\Lambda^*}$. If m is an interior point it is covered above. (Note that the proposition does not make any claim if $m \in \partial D_{\Lambda^*}$.)

Now suppose $a \notin \overline{D}_{\Lambda^*} = \text{Int}(D_{\Lambda^*}) \cup \partial D_{\Lambda^*}$, and $a > m$. Then $\Lambda^*(x) = +\infty$ for $x > a$. An argument as in (4.15) yields that both sides of (4.13) are equal to $(-\infty)$.

Next let $a < m$. As in the proof of the preceding proposition, irrespective of $a \in \partial D_{\Lambda^*}$ or not, note that $P(\frac{1}{n}\ S_n \in (a,\infty))$ is bounded away from 0. So both sides of (4.13) are equal to 0. Thus (4.13) has been established.

Proof of (4.14) is entirely similar. □

Lemma 4.8 *Let $\{a_n\}, \{b_n\}$ be two sequences of positive numbers. Then*

$$\limsup_{n\to\infty}\ \frac{1}{n} \log(a_n + b_n) = \max\left\{\limsup \frac{1}{n}\ \log a_n,\ \ \limsup \frac{1}{n}\ \log b_n\right\}.$$

Proof: As $\max\{a_n, b_n\} \leq a_n + b_n \leq 2\max\{a_n, b_n\}$ we see that

$$\begin{aligned} \log(\max\{a_n, b_n\}) &\leq \log(a_n + b_n) \\ &\leq \log 2 + \log(\max\{a_n, b_n\}) \end{aligned}$$

and hence

$$\begin{aligned}\max\{\log a_n, \log b_n\} &\leq \log(a_n + b_n) \\ &\leq \log 2 + \max\{\log a_n, \log b_n\}.\end{aligned}$$

Therefore we get

$$\begin{aligned}\max\left\{\frac{1}{n}\log a_n, \frac{1}{n}\log b_n\right\} &\leq \frac{1}{n}\log(a_n + b_n) \\ &\leq \frac{1}{n}\log 2 + \max\left\{\frac{1}{n}\log a_n, \frac{1}{n}\log b_n\right\}.\end{aligned}$$

Taking lim sup as $n \to \infty$ we are done. □

The next result is a forerunner to the way one can deal with more general situations.

Proposition 4.9 *Let the hypotheses be as in Theorem 4.1. Then for any closed set $F \subseteq \mathbb{R}$*

$$\limsup_{n\to\infty} \frac{1}{n}\log P\left(\frac{1}{n}\ S_n \in F\right) \leq -\Lambda^*(F) \qquad (4.16)$$

where we have used the notation (4.10).

Proof: Let $a = \inf\{x \in F : x > m\}$, $b = \sup\{x \in f : x < m\}$. As F is closed note that $a \in F, b \in F$; moreover $a = b$ implies $a = b = m$. Clearly $F \subseteq (-\infty, b] \cup [a, \infty)$, and hence

$$\frac{1}{n}\ \log P\left(\frac{1}{n}\ S_n \in F\right) \leq \frac{1}{n}\ \log\left[P\left(\frac{1}{n}\ S_n \geq a\right) + P\left(\frac{1}{n}\ S_n \leq b\right)\right].$$

(In the above note that limit as $n \to \infty$ may not exist on the l.h.s.) Taking lim sup on both sides and using the preceding lemma and Proposition 4.6 we get

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{n}\ \log P\left(\frac{1}{n}\ S_n \in F\right) \\ \leq\ &\max\left\{\left[\overline{\lim}\ \frac{1}{n}\log P\left(\frac{1}{n}\ S_n \geq a\right)\right], \left[\overline{\lim}\ \frac{1}{n}\ \log P\left(\frac{1}{n}\ S_n \leq b\right)\right]\right\} \\ =\ &\max\{[-\Lambda^*([a,\infty))], [-\Lambda^*((-\infty, b]))\} \\ =\ &-\min\{\Lambda^*([a,\infty)), \Lambda^*((-\infty, b])\} \\ =\ &-\Lambda^*((-\infty, b] \cup [a,\infty)) = -\Lambda^*(F)\end{aligned}$$

as $a, b \in F$. $\square$

If the closed set F is a finite interval we can say more.

Proposition 4.10 *Hypotheses as in Theorem 4.1. Then for any $a, b \in \mathbb{R}$ with $a < b$*

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in [a,b]\right) = -\Lambda^*([a,b]). \qquad (4.17)$$

Proof: We shall only sketch the proof, and that too only when $b > a > E(X_1)$. In view of the upper bound (4.16) in the preceding proposition, it is enough to prove that

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in [a,b]\right) \geq -\Lambda^*([a,b]). \qquad (4.18)$$

For this we can follow the proof of the lower bound (4.3) in the proof of Theorem 4.1. We adopt the notation used in that proof, and indicate only the modifications needed. Apply the Esscher transform given by (4.6). In the place of (4.7) we need to prove

$$P\left(0 \leq \frac{1}{n} S_n \leq b\right) = \rho^n E\left[I_{[0,b]}\left(\frac{1}{n} \hat{S}_n\right) e^{-\tau \hat{S}_n}\right]. \qquad (4.19)$$

Take $H_n = \{(x_1, \ldots, x_n) : 0 \leq x_1 + \ldots + x_n \leq nb\}$ and proceed as in the proof of Lemma 4.3 to obtain (4.19). Next, since $I_{[0,b]}(\frac{1}{n}\hat{S}_n) = I_{[0,nb]}(\hat{S}_n)$, one can mimic the proof of Lemma 4.4 (using the CLT) to get

$$\liminf_{n\to\infty} \frac{1}{n} \log E\left[I_{[0,b]}\left(\frac{1}{n}\hat{S}_n\right) \exp(\tau \hat{S}_n)\right] \geq 0 \qquad (4.20)$$

(4.19), (4.20) now imply (4.18). $\square$

Similar to the derivation of Proposition 4.7 using Proposition 4.6, the next result can be obtained using Proposition 4.10.

Proposition 4.11 *Let $a < b$ with both $a, b \in \mathbb{R}\backslash(\partial D_{\Lambda^*})$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in (a,b)\right) = -\Lambda^*((a,b)). \qquad (4.21)$$

Proposition 4.12 *Hypotheses as in Theorem 4.1. Then for any* $-\infty \leq a < b \leq \infty$,

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in (a,b)\right) \geq -\Lambda^*((a,b)). \qquad (4.22)$$

Proof: Let $m \leq a$ and $a \in \partial D_{\Lambda^*}$. Note that $\Lambda^*(x) = +\infty,\ x > a$. Let $a < b < \infty$. Note that $b \notin \partial D_{\Lambda^*}$. For sufficiently small $\delta > 0$, clearly $a + \delta < b$, $a + \delta \notin \partial D_{\Lambda^*}$. Therefore by the preceding proposition

$$\begin{aligned} \text{l.h.s. of } 4.22 &\geq \liminf \frac{1}{n} \log P\left(\frac{1}{n} S_n \in (a+\delta, b)\right) \\ &= \lim \frac{1}{n} \log P\left(\frac{1}{n} S_n \in (a+\delta, b)\right) \\ &= -\Lambda^*((a+\delta, b)) = -\infty = \text{r.h.s. of (4.22)}. \end{aligned}$$

If $b = \infty$ then an analogous argument using Proposition 4.7 in the place of Proposition 4.12 can be adopted.

In a similar fashion if $b \leq m$, $b \in \partial D_{\Lambda^*}$, (4.22) can be proved for any a. The case when $m \in (a,b)$ is left as an exercise. The cases when a, b are not boundary points have been already dealt with in previous propositions. □

The preceding result now leads to the following

Proposition 4.13 *Hypotheses as in Theorem 4.1. Then for any open set* $U \subseteq \mathbb{R}$,

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in U\right) \geq -\Lambda^*(U). \qquad (4.23)$$

Proof: We can write $U = \bigcup_{i=1}^{\infty} U_i$ where each U_i is an open interval, $U_i \cap U_j = \phi,\ i \neq j$. By the preceding proposition, for any $k \geq 1$

$$\begin{aligned} \text{l.h.s. of (4.23)} &\geq \liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in U_k\right) \\ &\geq -\Lambda^*(U_k). \end{aligned}$$

Consequently

$$\begin{aligned} \text{l.h.s. of (4.23)} &\geq \sup_k [-\Lambda^*(U_k)] \\ &= -\inf_k [\inf\{\Lambda^*(x) : x \in U_k\}] = -\Lambda^*(U), \end{aligned}$$

completing the proof. □

Combining Propositions 4.9 and 4.13 we have the following version of Cramer's theorem; this will be the formulation to be adopted in general situations.

Theorem 4.14 *Hypotheses as in Theorem 4.1. Then for any closed set* $F \subseteq \mathbb{R}$*,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in F\right) \leq -\inf\{\Lambda^*(x) : x \in F\}$$

and for any open set $U \subseteq \mathbb{R}$

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in U\right) \geq -\inf\{\Lambda^*(x) : x \in U\}.$$

Remark 4.15 Let $A \in \mathcal{B}(\mathbb{R})$ be such that $\Lambda^*(\text{Int}(A)) = \Lambda^*(A) = \Lambda^*(Cl(A))$. In some of the preceding results we have seen examples of such sets. By the theorem above it then follows that

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n} S_n \in A\right) = -\Lambda^*(A).$$

Therefore it can be seen that

$$P\left(\frac{1}{n} S_n \in A\right) = \exp[-n\Lambda^*(A) + \alpha(n)]$$

with $\lim_{n\to\infty} \frac{\alpha(n)}{n} = 0$; in other words

$$P\left(\frac{1}{n} S_n \in A\right) = \exp[-n\Lambda^*(A) + o(n)],$$

which is an intuitively appealing form of the large deviations statement. □

We end this section by identifying $\Lambda^*(\cdot)$ in a few examples. We shall be using the notation $g_z(\cdot)$ as given in (3.7).

Example 4.16 Let $\{X_i\}$ be a sequence of i.i.d. Bernoulli (p) random variables, where $0 < p < 1$. In this case $M(t) = 1 - p + pe^t$, $t \in \mathbb{R}$. Hence for $z \in \mathbb{R}$,

$$g_z(t) = \log \frac{e^{zt}}{1 - p + pe^t} = \log \left[\frac{e^{zt}}{pe^t} \frac{pe^t}{(1 - p + pe^t)}\right].$$

It can be seen that $g_z(\cdot)$ is bounded above if and only if $z \in [0,1]$. So $D_{\Lambda^*} = [0,1]$. Let $z \in [0,1]$. Note that $g_z'(t) = 0 \Leftrightarrow t = \log\left[\frac{z(1-p)}{p(1-z)}\right]$. By strict concavity of $g_z(\cdot)$ (see the proof of Proposition 3.7), it follows that $g_z(\cdot)$ has a unique maximum at the above point. It is easy now to obtain

$$\Lambda^*(x) = \begin{cases} x \log\left(\frac{x}{p}\right) + (1-x) \log \left(\frac{1-x}{1-p}\right), & 0 \le x \le 1 \\ +\infty, \quad \text{otherwise.} \end{cases} \qquad (4.24)$$

For $p = \frac{1}{2}$, this is in agreement with the earlier expression

Example 4.17 Let $\{X_i\}$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Here $M(t) = \exp\ (\mu t + \frac{1}{2}\sigma^2 t^2),\ t \in \mathbb{R}$ and hence it follows that

$$\Lambda^*(x) = \frac{1}{2}\ \frac{(x-\mu)^2}{\sigma^2}, \quad x \in \mathbb{R}.$$

Example 4.18 Let $\{X_i\}$ be i.i.d. Poisson (λ) random variables where $\lambda > 0$. We know that $M(t) = \exp[\lambda(e^t - 1)],\ t \in \mathbb{R}$, and hence $g_z(t) = zt + \lambda(e^t - 1),\ t \in \mathbb{R}$ for any $z \in \mathbb{R}$. If $z < 0$, note that $g_z(t) \to +\infty$ as $t \downarrow (-\infty)$. Let $z \ge 0$. In this case, though $g_z(\cdot)$ is not bounded below, it is bounded above. It can be checked that $g_z(\cdot)$ has a unique maximum point $t = \log\left(\frac{z}{\lambda}\right)$. Therefore

$$\Lambda^*(x) = \begin{cases} x \log\left(\frac{x}{\lambda}\right) - x + \lambda, & \text{if } x \ge 0 \\ +\infty \quad , \quad \text{if} \quad x < 0. \end{cases}$$

5 Sanov's theorem for finite sets

The function $\Lambda^*(\cdot)$ given by (4.24) in Example 4.16 arises also in information theory / statistics. This turns out to be a feature of large deviations at another level. We begin by describing the set up.

Let $A = \{1, 2, \ldots, k\}$ where k is a fixed integer. Let M be the set of all probability measures on A; so

$$M = \left\{ \mu = (\mu_1, \ldots, \mu_k) : 0 \le \mu_i \le 1, \sum_{i=1}^{k}\ \mu_i = 1 \right\}. \qquad (5.1)$$

Note that M is a compact subset of $\mathbb{R}^k$. M is given the relative topology, which is nothing but coordinate wise convergence.

For $\nu \in M$ define the *entropy of* ν by

$$H(\nu) = -\sum_{i=1}^{k} \nu_i \log \nu_i. \qquad (5.2)$$

For $\mu, \nu \in M$ define the *relative entropy of* ν *w.r.t.* μ by

$$H(\nu \mid \mu) = \sum_{i=1}^{k} \nu_i \log \left(\frac{\nu_i}{\mu_i}\right); \qquad (5.3)$$

it is also called *Kullback-Leibler information* in statistics. We use the convention: if $\nu_i = 0$ then $\nu_i \log\left(\frac{\nu_i}{\mu_i}\right) = 0$; if $\mu_i = 0$, $\nu_i > 0$ then $\nu_i \log\left(\frac{\nu_i}{\mu_i}\right) = \infty$.

Note: Let $\Lambda^*(\cdot)$ be as in (4.24), with $0 < p < 1$. For $x \in [0,1]$ it is easily seen that $\Lambda^*(x) = H(\nu \mid \mu)$, where $\mu =$ Bernoulli $(p), \nu =$ Bernoulli (x) distributions.

Proposition 5.1 *Fix $\mu \in M$. Assume $\mu_i > 0$ for all i. Then $\nu \mapsto H(\nu \mid \mu)$ is a finite, continuous, and strictly convex function on M. Moreover $H(\nu \mid \mu) \geq 0$, with equality holding if and only if $\nu = \mu$.*

Proof: Let $h(x) = x \log x$, $x \in [0,1]$. Note that $h(\cdot)$ is a finite, continuous, strictly convex function on $[0,1]$. Next, as $\mu_i > 0$ for all i, we can write

$$H(\nu \mid \mu) = \sum_i \mu_i \left(\frac{\nu_i}{\mu_i}\right) \log\left(\frac{\nu_i}{\mu_i}\right) = \sum_i \mu_i h\left(\frac{\nu_i}{\mu_i}\right).$$

So the first assertion follows.

Moreover by Jensen's inequality

$$\begin{aligned} H(\nu \mid \mu) &= \sum_i \mu_i h\left(\frac{\nu_i}{\mu_i}\right) \\ &\geq h\left(\sum_i \mu_i \frac{\nu_i}{\mu_i}\right) = h(1) = 0; \end{aligned}$$

we could apply Jensen's inequality because h is convex and $\sum_i$ above is a convex combination. Since $h(\cdot)$ is strictly convex, equality in the above occurs if and only if $\frac{\nu_i}{\mu_i} = C \ \forall\ i$. But $C = 1$ as $\sum \nu_i = \sum \mu_i = 1$. $\square$

Remark 5.2 Let $\hat{\mu}$ be the discrete uniform distribution over $A = \{1, 2, \ldots, k\}$. So $\hat{\mu}(\{i\}) = \frac{1}{k}$ for any $1 \leq i \leq k$. Then for any probability measure ν on A

$$\begin{aligned} H(\nu \mid \hat{\mu}) &= \sum_{i=1}^{k} \nu_i \log \nu_i + (\log k) \sum_{i=1}^{k} \nu_i \\ &= -H(\nu) + \log k. \end{aligned} \qquad (5.4)$$

Now the preceding proposition applied to $\mu = \hat{\mu}$, together with (5.4), implies that $H(\nu) \leq \log k$ for any $\nu \in M$. Moreover we also get that the equality $H(\nu) = \log k$ holds if and only if $\nu = \hat{\mu}$. Thus the entropy $H(\nu)$ attains the maximum value $(\log k)$ if and only if ν is the discrete uniform distribution. Thus the discrete uniform distribution is in a sense the most random probability distribution on a finite set. □

Fix $\mu \in M$ with $\mu_i > 0$ for all i. Let $X_1, X_2, \ldots$ be an i.i.d. sequence taking values in A and having common distribution μ. Note that X_i's are defined on some probability space $(\Omega, \mathcal{F}, P)$. Define the *empirical distribution* by

$$L_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} \delta(X_i(\omega)), \quad \omega \in \Omega \qquad (5.5)$$

for $n = 1, 2, \ldots$ where $\delta(y)$ denotes the degenerate probability measure at y. For each $\omega \in \Omega$, clearly $L_n(\omega)$ is a probability measure on A; that is, $L_n(\omega) \in M \ \forall\ \omega \in \Omega, \ \forall\ n$. For fixed $n \geq 1$, $\omega \in \Omega$ note that

$$\begin{aligned} L_n(\omega)(\{j\}) &= \frac{1}{n} \#\{\ell : X_\ell(\omega) = j\} \\ &= \text{proportion among the sample points} \\ &\quad X_1(\omega), \ldots, X_n(\omega) \text{ taking the value } j, \end{aligned}$$

for $j \in A$; hence the reason for its name. For each $n \geq 1$, it can be seen that $\omega \mapsto L_n(\omega)$ is a measurable function; so each L_n is an M-valued random variable.

We can now state Sanov's theorem for finite sets.

Theorem 5.3 *Fix $\mu \in M$ with $\mu_i > 0$ for all i. Then for any open set $U \subseteq M$*

$$\liminf_{n \to \infty} \frac{1}{n} \log P(L_n \in U) \geq -\inf\{H(\nu \mid \mu) : \nu \in U\}, \qquad (5.6)$$

and for any closed set $F \subseteq M$

$$\limsup_{n\to\infty} \frac{1}{n} \log P(L_n \in F) \leq -\inf\{H(\nu \mid \mu) : \nu \in F\}. \qquad (5.7)$$

Proof: For a set Γ we shall write $\Gamma^\circ = \mathrm{Int}(\Gamma)$ = interior of Γ, $\bar{\Gamma} = \mathrm{Cl}(\Gamma)$ = closure of Γ. We shall prove the result in the following form: For any Borel set $\Gamma \subseteq M$,

$$\liminf_{n\to\infty} \frac{1}{n} \log P(L_n \in \Gamma) \geq -\inf\{H(\nu \mid \mu) : \nu \in \Gamma^\circ\} \qquad (5.8)$$

$$\limsup_{x\to\infty} \frac{1}{n} \log P(L_n \in \Gamma) \leq -\inf\{H(\nu \mid \mu) : \nu \in \bar{\Gamma}\}. \qquad (5.9)$$

Obviously when Γ is open or closed we get the corresponding statements of the theorem.

We shall denote

$$S_n = \left\{ \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_k}{n}\right) : n_i \geq 0 \text{ integers}, \sum_{i=1}^{k} n_i = n \right\}. \qquad (5.10)$$

Clearly $S_n \subset M$. Note that $S_n = \mathrm{Range}(L_n)$ is the set of all possible n-sample empiricals.

Temporarily assume: For $\nu \in S_n$

$$\begin{aligned} \frac{1}{(n+1)^k} \exp[-nH(\nu \mid \mu)] &\leq P(L_n = \nu) \\ &\leq \exp[-nH(\nu \mid \mu)]. \end{aligned} \qquad (5.11)$$

We shall indicate how the proof can be completed assuming (5.11).

Let $C = \inf\{H(\nu \mid \mu) : \nu \in \Gamma\}$. Note that $C \geq 0$. By (5.11), clearly $P(L_n = \nu) \leq e^{-nC}$, $\forall\ \nu \in \Gamma$. Note also that $|S_n| \leq (n+1)^k$, as $n_i \in \{0, 1, \ldots, n\}$ for $1 \leq i \leq k$. Therefore it follows that

$$\begin{aligned} P(L_n \in \Gamma) &= \sum_{\nu \in S_n \cap \Gamma} P(L_n = \nu) \leq |S_n \cap \Gamma|\ e^{-nC} \\ &\leq (n+1)^k\ e^{-nC}, \end{aligned}$$

from which (5.9) is clear.

To derive (5.8), fix $\nu \in \Gamma^\circ$. We will show that

$$\liminf_{n\to\infty} \frac{1}{n} \log P(L_n \in \Gamma) \geq -H(\nu \mid \mu); \qquad (5.12)$$

then taking supremum over $\nu \in \Gamma^\circ$ we get the desired result.

Note that $\nu = (\nu_1, \ldots, \nu_k)$ with $\nu_i \geq 0 \ \forall\ i$ and $\sum_{i=1}^{k} \nu_i = 1$. It is not difficult to see that $\exists\ \nu^{(n)} \in S_n$, $n \geq 1$ such that $\nu^{(n)} \to \nu$. As Γ° is open and $\nu \in \Gamma^\circ$ it now follows that $\nu^{(n)} \in \Gamma^\circ \cap S_n \subseteq \Gamma \cap S_n$ for all large n and $\nu^{(n)} \to \nu$. Now for large n

$$P(L_n \in \Gamma) \geq P(L_n = \nu^{(n)}) \geq \frac{1}{(n+1)^k} \exp[-nH(\nu^{(n)} \mid \mu)]$$

by the left inequality in (5.11). By Proposition 5.1 $H(\nu^{(n)} \mid \mu) \to H(\nu \mid \mu) < \infty$. From the above now (5.12) follows.

We need to prove (5.11). Fix $\nu \in S_n$; let $\nu = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_k}{n}\right)$ with $n_i \geq 0$ being integers and $\sum_{i=1}^{k} n_i = n$. Now let $S_n(\nu)$ denote the set of samples of size n whose empirical is ν. So $S_n(\nu)$ is a subset of A^n. In fact

$$S_n(\nu) = \{(x_1, \ldots, x_n) \in A^n : \text{exactly } n_i \text{ among } x_\ell\text{'s are } i, 1 \leq i \leq k\}.$$

Now temporarily assume that

$$\frac{1}{(n+1)^k} \exp[nH(\nu)] \leq |S_n(\nu)| \leq \exp[nH(\nu)]. \qquad (5.13)$$

We will show that (5.13) implies (5.11). Note that

$$\begin{aligned}
\prod_{i=1}^{k} (\mu_i^{n_i}) &= \exp\left(\sum_{i=1}^{k} n_i \log \mu_i\right) \\
&= \exp\left(n \sum_{i=1}^{k} \nu_i \log \mu_i\right) \\
&= \exp\left[n \sum_{i=1}^{k} (\nu_i \log \nu_i - \nu_i \log \nu_i + \nu_i \log \mu_i)\right] \\
&= \exp\ [n\{-H(\nu) - H(\nu \mid \mu)\}]
\end{aligned}$$

where we have used the notation $\nu_i = \frac{n_i}{n}\ \forall\ i$. Consequently by the description of $S_n(\nu)$ given above we get

$$\begin{aligned} P(L_n = \nu) &= |S_n(\nu)| \prod_{i=1}^{k} (\mu_i^{n_i}) \\ &= |S_n(\nu)|\ \exp\ [n\{-H(\nu) - H(\nu \mid \mu)\}]. \end{aligned}$$

It is now clear that (5.13) implies (5.11). Observe that there is no reference to μ in (5.13).

To establish (5.13), fix $\nu \in S_n$ where $\nu = (\nu_1, \ldots, \nu_k) = \left(\frac{n_1}{n}, \ldots, \frac{n_k}{n}\right)$ with usual meaning. Let $Y_1, Y_2, \ldots$ be an i.i.d. sequence taking values in A and having common distribution ν. Define

$$M_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} \delta(Y_i(\omega)),\ \omega \in \Omega,\ n \geq 1. \qquad (\ 5.14\ )$$

Clearly $M_n(\omega) \in M\ \ \forall\, \omega, n$. Note that M_n is the empirical distribution when i.i.d. random variables have distribution ν. Again we make a temporary assumption: Assume

$$P(M_n = \nu) \geq P(M_n = \hat{\nu}),\quad \forall\ \ \hat{\nu} \in S_n. \qquad (\ 5.15\ )$$

We will now derive (5.13) using (5.15). Using (5.15), by an argument similar to the one using the description of $S_n(\nu)$ we get

$$\begin{aligned} 1 &= \sum_{\hat{\nu} \in S_n} P(M_n = \hat{\nu}) \leq |S_n|\ P(M_n = \nu) \\ &= |S_n|\ |S_n(\nu)|\ \prod_{i=1}^{k} (\nu_i^{n_i}) \\ &\leq (n+1)^k |S_n(\nu)|\ \exp(-nH(\nu)), \end{aligned}$$

whence the left inequality in (5.13) follows. To obtain the other inequality in (5.13) note that

$$\begin{aligned} 1 &\geq P(M_n = \nu) = |S_n(\nu)|\ \prod_{i=1}^{k} (\nu_i^{n_i}) \\ &= |S_n(\nu)|\ \exp(-nH(\nu)), \end{aligned}$$

from which the desired inequality follows.

Finally it remains to prove (5.15).

Let $\hat{\nu} = (\hat{\nu}_1, \ldots, \hat{\nu}_k) = \left(\frac{\hat{n}_1}{n}, \ldots, \frac{\hat{n}_k}{n}\right)$. Note that computing probabilities involving M_n is the same as computing appropriate multinomial probabilities with cell probabilities $\nu_1, \ldots, \nu_k$. Clearly $M_n(\omega) = \hat{\nu}$ is the same as $\frac{1}{n} \sum_{j=1}^{n} \delta(Y_j(\omega)) = \frac{1}{n} \sum_{i=1}^{k} \hat{n}_i \delta(i)$, that is exactly $\hat{n}_i$ among $\{Y_j(\omega), 1 \le j \le n\}$ being i for each $i \in A = \{1, \ldots, k\}$. Consequently, under ν,

$$\begin{aligned} P(M_n = \hat{\nu}) &= \frac{n!}{(\hat{n}_1!)\ldots(\hat{n}_k!)} (\nu_1)^{\hat{n}_1} \ldots (\nu_k)^{\hat{n}_k} \\ &= |S_n(\hat{\nu})| \, (\nu_1)^{n\hat{\nu}_1} \ldots (\nu_k)^{n\hat{\nu}_k}. \end{aligned} \qquad (5.16)$$

If $\hat{\nu}_i > 0$ but $\nu_i = 0$ for some i, then it is clear from (5.16) that $P(M_n = \hat{\nu}) = 0$, and hence (5.15) holds in this case. (So we may assume that $\nu_i = 0$ implies $\hat{\nu}_i = 0$; in such a case the analysis may have to be done with S_{n-1} or S_ℓ for some $\ell \le n-1$.) Hence, in the following, we may take $\nu_i > 0,\ \hat{\nu}_i > 0$ for all i. For positive integers α, β note that $\frac{\alpha!}{\beta!} \ge \beta^{(\alpha-\beta)}$.

Consequently

$$\begin{aligned} \frac{|S_n(\nu)|}{|S_n(\hat{\nu})|} &= \frac{(n\hat{\nu}_1)!\ldots(n\hat{\nu}_k)!}{(n\nu_1)!\ldots(n\nu_k)!} \\ &\ge \prod_{i=1}^{k} (n\nu_i)^{n\hat{\nu}_i - n\nu_i} \\ &= \left[\prod_{i=1}^{k} n^{(n\hat{\nu}_i - n\nu_i)}\right] \left[\prod_{i=1}^{k} \nu_i^{(n\hat{\nu}_i - n\nu_i)}\right] \\ &= \prod_{i=1}^{k} \nu_i^{(n\hat{\nu}_i - n\nu_i)} \end{aligned}$$

because $\sum \hat{\nu}_i = \sum \nu_i = 1$. Therefore by (5.16)

$$\frac{P(M_n = \nu)}{P(M_n = \hat{\nu})} = \frac{|S_n(\nu)| \prod_{i=1}^{k} (\nu_i)^{n\nu_i}}{|S_n(\hat{\nu})| \prod_{i=1}^{k} (\nu_i)^{n\hat{\nu}_i}} \ge 1,$$

establishing (5.15).

This completes the proof of the theorem. □

Now using Sanov's theorem we want to derive Cramer's theorem for i.i.d. random variables taking values in a fixed finite set of $\mathbb{R}$. Let $A = \{a_1, a_2, \ldots, a_k\} \subset \mathbb{R}$ be a fixed finite set. Without loss of generality let $a_1 < a_2 < \ldots < a_k$. Let μ be a fixed probability measure on A with $\mu(\{a_i\}) > 0$ for all i. Let $X_1, X_2, \ldots$ be i.i.d. random variables having μ as the common distribution. Let $S_n = X_1 + \ldots + X_n$, $n \geq 1$. Note that $\frac{1}{n} S_n$, $n \geq 1$ take value in the compact interval $[a_1, a_k]$. Note that the moment generating function is

$$0 < M_\mu(t) = \sum_{i=1}^{k} e^{ta_i} \mu(\{a_i\}) < \infty, \quad \forall\ t \in \mathbb{R},$$

and $\Lambda(\theta) = \log M_\mu(\theta)$ is finite for all $\theta \in \mathbb{R}$.

We shall denote by M the set of all probability measures on $A = \{a_1, \ldots, a_k\}$. For any $\nu \in M$ note that the relative entropy is

$$H(\nu \mid \mu) = \sum_{i=1}^{k} \nu(\{a_i\}) \log \left(\frac{\nu(\{a_i\})}{\mu(\{a_i\})} \right). \qquad (5.17)$$

Proposition 5.4 *Let the notation and hypotheses be as above. Set* $\Lambda^*(x) = \sup\{\theta x - \Lambda(\theta) : \theta \in \mathbb{R}\}$. *Then*

$$\begin{aligned} \Lambda^*(x) &= \inf\{H(\nu \mid \mu) : \nu \in M \text{ with } \sum_i a_i \nu(\{a_i\}) = x\} \quad (5.18) \\ &= +\infty, \text{ if there is no such } \nu. \end{aligned}$$

Proof: For notational simplicity write $\mu_i = \mu(\{a_i\})$, $\nu_i = \nu(\{a_i\})$.

If $x < a_1$, there is no $\nu \in M$ such that $\sum_i a_i \nu_i = x$. So r.h.s. of (5.18) is $+\infty$. For such an x note that $(a_i - x) > 0$ for all i; so $e^{\theta(a_i - x)} \to 0$ as $\theta \to (-\infty)$ for all i. Therefore

$$\theta x - \Lambda(\theta) = -\log \left[\sum_{i=1}^{k} \mu_i\ e^{\theta(a_i - x)} \right] \to (+\infty)$$

as $\theta \to (-\infty)$. Thus $\Lambda^*(x) = +\infty$ if $x < a_1$. In a similar fashion $\Lambda^*(x) = +\infty$ if $x > a_k$.

Next let $x = a_1$. Let $\nu^* \in M$ be such that $\sum_i a_i \nu_i^* = a_1$. It is not difficult to see that $\nu_i^* = 0 \quad \forall\ i \geq 2$. Hence $\nu^* = \delta(a_1)$. So r.h.s. of (5.18) is $H(\nu^* \mid \mu) = -\log \mu_1$. Also

$$\theta a_1 - \Lambda(\theta) = -\log \left[\sum_{i=1}^{k} \mu_i e^{\theta(a_i - a_1)} \right]$$

decreases as θ increases, and as $\theta \to (-\infty)$ it converges to $-\log \mu_1$. Thus $\Lambda^*(a_1) = -\log \mu_1$. Thus (5.18) holds when $x = a_1$.

By an analogous argument $\Lambda^*(a_k) = -\log \mu_k =$ r.h.s. of (5.18).

Finally let $a_1 < x < a_k$. Let $\nu \in M$ be such that $\sum_{i=1}^{k} a_i \nu_i = x$. Note that for any $\theta \in \mathbb{R}$, by Jensen's inequality

$$\begin{aligned}
\Lambda(\theta) &= \log \sum e^{\theta a_i} \mu_i = \log \sum \left(\frac{\mu_i}{\nu_i}\ e^{\theta a_i} \right) \cdot \nu_i \\
&\geq \sum \nu_i \log \left(\frac{\mu_i}{\nu_i}\ e^{\theta a_i} \right) \\
&= \theta \sum a_i \nu_i - \sum \nu_i \log \left(\frac{\nu_i}{\mu_i} \right) \\
&= \theta x - H(\nu \mid \mu). \qquad (5.19)
\end{aligned}$$

So $\theta x - \Lambda(\theta) \leq H(\nu \mid \mu)$ for any $\theta \in \mathbb{R}$ and any $\nu \in M$ with $\sum_i a_i \nu_i = x$. Hence

$$\begin{aligned}
\Lambda^*(x) &= \sup_\theta \ [\theta x - \Lambda(\theta)] \\
&\leq \inf\{H(\nu \mid \mu) : \nu \in M \text{ with } \sum a_i \nu_i = x\} \\
&= \text{r.h.s. of (5.18).} \qquad (5.20)
\end{aligned}$$

We will now exhibit $\nu^x \in M$ for which $\sum_i a_i \nu_i^x = x$ and equality is attained in (5.19).

Note that $\Lambda(\cdot)$ is twice differentiable and $a_1 \leq \Lambda'(\theta) \leq a_k$ for all θ; also Range$(\Lambda') = (a_1, a_k)$ as μ is nondegenerate. By the proof of Proposition 3.7, as $x \in$ Range(Λ'), there is a unique $\hat{\theta} \in \mathbb{R}$ such that $\Lambda'(\hat{\theta}) = x$ and

$\Lambda^*(x) = \hat{\theta}x - \Lambda(\hat{\theta})$. Define the probability measure $\nu^x \in M$ by

$$\begin{aligned} \nu^x(\{a_i\}) &= \nu_i^x = \frac{1}{\sum_\ell \mu_\ell \exp(\hat{\theta} a_\ell)} \ \mu_i \exp(\hat{\theta} a_i) \\ &= \frac{1}{M_\mu(\hat{\theta})} \ \mu_i \exp(\hat{\theta} a_i). \end{aligned}$$

Then $\sum_i a_i \nu_i^x = \Lambda'(\hat{\theta}) = x$. Observe that by definition of ν^x,

$$\begin{aligned} H(\nu^x \mid \mu) &= \sum_i \nu_i^x \ \log\left(\frac{\nu_i^x}{\mu_i}\right) \\ &= \sum_i \ \frac{1}{M_\mu(\hat{\theta})} \ \mu_i e^{\hat{\theta} a_i} \ \log\left(\frac{e^{\hat{\theta} a_i}}{M_\mu(\hat{\theta})}\right) \\ &= \frac{1}{M_\mu(\hat{\theta})} \ \sum_i \ \mu_i \ e^{\hat{\theta} a_i} [\hat{\theta} a_i - \Lambda(\hat{\theta})] \\ &= \frac{\hat{\theta}}{M_\mu(\hat{\theta})} \ \sum_i \ a_i \mu_i e^{\hat{\theta} a_i} - \frac{\Lambda(\hat{\theta})}{M_\mu(\hat{\theta})} \ \sum_i \ \mu_i e^{\hat{\theta} a_i} \\ &= \hat{\theta} \ \sum_i \ a_i \nu_i^x - \Lambda(\hat{\theta}) = \hat{\theta} x - \Lambda(\hat{\theta}). \end{aligned}$$

Thus equality holds in (5.19). Also $\Lambda^*(x) = H(\nu^* \mid \mu)$. The proof is now complete. □

Proposition 5.5 *Let $A = \{a_1, \ldots, a_k\} \subset \mathbb{R}$, and μ a probability measure on A with $\mu_i = \mu(\{a_i\}) > 0$ for all i. Let $X_1, X_2, \ldots$ be i.i.d. random variables taking value in A with common distribution μ. Let $S_n = X_1 + \ldots + X_n$, $n \geq 1$. If $U \subseteq \mathbb{R}$ is open, then*

$$\liminf_{n\to\infty} \ \frac{1}{n} \ \log P\left(\frac{1}{n} \ S_n \in U\right) \geq -\inf\{\Lambda^*(x) : x \in U\}.$$

If $F \subseteq \mathbb{R}$ is closed, then

$$\limsup_{n\to\infty} \ \frac{1}{n} \ \log P\left(\frac{1}{n} \ S_n \in F\right) \leq -\inf\{\Lambda^*(x) : x \in F\}.$$

Proof: In Sanov's theorem take $A = \{a_1, \ldots, a_k\}$ (in the place of $\{1, 2, \ldots, k\}$). Note that the theorem is valid in this set up as well, with relative entropy given by (5.17). For $\nu \in M$ we shall write $\nu = (\nu_1, \ldots, \nu_k)$ with $\nu_i = \nu(\{a_i\})$, $1 \leq i \leq k$.

Now, for any Borel set $B \subseteq \mathbb{R}$ let

$$\begin{aligned}\tilde{B} &= \left\{\nu \in M : \sum_{i=1}^{k} a_i\nu_i \in B\right\} \\ &= \bigcup_{x\in B} \left\{\nu \in M : \sum_{i=1}^{k} a_i\nu_i = x\right\}.\end{aligned}$$

Note that $\sum\limits_i a_i\nu_i =$ expectation of the probability measure ν. Observe that $\nu \mapsto \sum\limits_{i=1}^{k} a_i\nu_i$ is a continuous map from M into $\mathbb{R}$. So if B is closed or open or Borel then so is $\tilde{B}$.

Let $\omega \in \Omega$ be such that $L_n(\omega) = \nu \in M$, where $L_n(\cdot)$ is given by (5.5).

Note that

$$L_n(\omega) = \frac{1}{n}\sum_{j=1}^{n} \delta(X_j(\omega)) = \sum_{i=1}^{k} \frac{n_i}{n}\,\delta(a_i)$$

where exactly n_i among $\{X_j(\omega) : 1 \le j \le n\}$ are a_i, for $1 \le i \le k$. So $L_n(\omega) = \nu \Leftrightarrow \nu = \left(\frac{n_1}{n}, \ldots, \frac{n_k}{n}\right)$ with $n_i \ge 0$ integers, $\sum\limits_{i=1}^{k} n_i = n$; (that is, $\nu_i = \frac{n_i}{n}$ for all i) $\Leftrightarrow \frac{1}{n}\, S_n(\omega) = \frac{1}{n}\sum\limits_{j=1}^{n} X_j(\omega) = \frac{1}{n}\sum\limits_{i=1}^{k} n_i a_i = \sum\limits_{i=1}^{k} a_i\,\frac{n_i}{n} = \sum\limits_{i=1}^{k} a_i\nu_i$. Therefore $L_n = \nu \Leftrightarrow \frac{1}{n}\, S_n = \sum\limits_{i=1}^{k} a_i\nu_i$, and hence $\frac{1}{n}\, S_n \in B \Leftrightarrow L_n \in \tilde{B}$.

Consequently using Theorem 5.3 and Proposition 5.4, we get for any open set $U \subseteq \mathbb{R}$,

$$\begin{aligned}&\liminf_{n\to\infty} \frac{1}{n}\,\log P\left(\frac{1}{n}\, S_n \in U\right) \\ =\ &\liminf_{n\to\infty} \frac{1}{n}\,\log P(L_n \in \tilde{U}) \\ \ge\ &-\inf\{H(\nu \mid \mu) : \nu \in \tilde{U}\} \\ =\ &-\inf_{x\in U}\ \inf\{H(\nu \mid \mu) : \sum_i a_i\nu_i = x\} \\ =\ &-\inf\{\Lambda^*(x) : x \in U\}.\end{aligned}$$

The assertion concerning any closed set $F \subseteq \mathbb{R}$ can be proved in a similar fashion. $\square$

We end this section with a few remarks.

Remark 5.6 (i) By Proposition 5.1 we know that $\nu \mapsto H(\nu \mid \mu)$ is continuous. If U is an open set, it now follows that r.h.s. of (5.8) and (5.9) are equal. Consequently for an open set U we have

$$\lim_{n\to\infty} \frac{1}{n} \log P(L_n \in U) = -\inf\{H(\nu \mid \mu) : \nu \in U\}.$$

(ii) Note that proving Sanov's theorem involves the following: (a) $\nu \mapsto H(\nu \mid \mu)$ is continuous; (b) $\bigcup_n S_n$ is dense in M; (c) Proof when $\mu \in S_n$.

Remark 5.7 Let $A = \{a_1, \ldots, a_k\} \subset \mathbb{R}$ and μ a probability measure on A. Let X denote a random variable with distribution μ. Note that we can write $\mu = \sum_{i=1}^{k} \mu_i \ \delta(a_i)$ where $\mu_i = \mu(\{a_i\})$. Now define $\delta_X(\omega) \stackrel{\Delta}{=} \delta(X(\omega)) = \delta(a_i)$, if $X(\omega) = a_i,\ 1 \le i \le k$. Clearly $\omega \mapsto \delta_X(\omega)$ is an M-valued random variable; also $P(\delta_X = \delta(a_i)) = P(X = a_i) = \mu_i$. Consequently

$$E(\delta_X) = \sum_{i=1}^{k} \delta(a_i)\ P(\delta_X = \delta(a_i)) = \sum \mu_i \delta(a_i) = \mu.$$

If $X_1, X_2, \ldots$ are i.i.d. A-valued random variables, with distribution μ, then $\delta_{X_1}, \delta_{X_2}, \ldots$ are i.i.d. M-valued random variables and $E(\delta_{X_i}) = \mu$. Therefore by the law of large numbers

$$L_n = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j} \to \mu \text{ a.s.} \quad \text{as } n \to \infty.$$

So $\{L_n \notin B_\epsilon(\mu)\}, \{L_n \notin \overline{B_\epsilon(\mu)}\}$ are examples of rare events, where $B_\epsilon(\mu)$ denotes a neighbourhood of μ. Clearly Theorem 5.3 gives information about asymptotic behaviour of $\frac{1}{n} \log P(L_n \notin B_\epsilon(\mu))$, $\frac{1}{n} \log P(L_n \notin \overline{B_\epsilon(\mu)})$. Thus Sanov's theorem is a result on large deviations (from the mean μ). $\square$

Remark 5.8 Let U be an open set in M with $\mu \notin U$. As noted in the preceding remark, $\{L_n \in U\}$ is a rare event whose probability can be very

small for large n. Let $\nu \in U \cap S_n$ for some n. Note that proof of Sanov's theorem involves a transformation from PL_n^{-1} to PM_n^{-1}. Both PL_n^{-1} and PM_n^{-1} are probability measures on M (which itself is a space of probability measures!); this is because L_n, M_n are M-valued random variables and PL_n^{-1}, PM_n^{-1} are their respective distributions. While the former is centered around μ, the latter is centered around ν. However, note that $\{M_n \in U\}$ is a typical event, as indicated clearly by (5.15). Thus the above transformation is another instance of tilting (similar to Esscher transform seen earlier).

6 General framework of Varadhan

Expressions like $\sup_{\xi} \{g(\xi) - I(\xi)\}$ are encountered as solutions to certain differential equations, where ξ varies over subsets of function spaces. These are not generally easy to handle/evaluate. In some cases the function $g(\cdot)$ is a nice function, while $I(\cdot)$ is a function having properties similar to $\Lambda^*(\cdot)$ of Cramer's theorem, though defined on a complicated space. To see how this apparent similarity could be used to advantage, we look at a situation on the real line.

Let λ be a σ-finite measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. For $f \in L^p(\mathbb{R}, \lambda)$ we shall write $\|f\|_{p,\lambda}$ its L^p-norm; if $d\lambda(x) = dx$ is the Lebesgue measure we shall simply write $\|f\|_p$. Also $\|f\|_\infty$ denotes the L^∞-norm.

Proposition 6.1 *Let $f \in L^r(\mathbb{R}, dx)$ for some $r \geq 1$. Then $\|f\|_p \to \|f\|_\infty$ as $p \to \infty$.* □

To prove this result we need some lemmas.

Lemma 6.2 *Let λ be a finite measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Let $h \in L^\infty(\mathbb{R})$. Then*

$$\lim_{p \to \infty} \|h\|_{p,\lambda} = \|h\|_\infty.$$

Proof: Note that $h \in L^p(\mathbb{R}, \lambda)$ for any $p \geq 1$. If $\|h\|_\infty = 0$ then the assertion of the lemma holds trivially. So assume $\|h\|_\infty > 0$.

Let $0 < \epsilon < \|h\|_\infty$ be arbitrary. Let $H_\epsilon = \{x \in \mathbb{R} : |h(x)| > \|h\|_\infty - \epsilon\}$. Note that $\lambda(H_\epsilon) > 0$. Then for any $p \geq 1$

$$(\|h\|_\infty - \epsilon)(\lambda(H_\epsilon))^{1/p} \leq \|h\|_{p,\lambda} \leq \|h\|_\infty (\lambda(\mathbb{R}))^{1/p}.$$

For any real number $a > 0$, observe that $a^{1/p} \to 1$ as $p \to \infty$. Hence letting $p \to \infty$ in the above we get

$$\|h\|_\infty - \epsilon \leq \liminf_{p\to\infty} \|h\|_{p,\lambda} \leq \limsup_{p\to\infty} \|h\|_{p,\lambda} \leq \|h\|_\infty.$$

As ϵ is arbitrary the required result follows. □

Lemma 6.3 *Let $f \in L^r(\mathbb{R}, dx)$, for some $r \geq 1$ and $0 \leq \|f\|_\infty < \infty$. Then $f \in L^p(\mathbb{R}, dx)$ for all $r \leq p \leq \infty$.*

Proof: We may assume $\|f\|_\infty > 0$. Let $r < p < \infty$. Without loss of generality $\|f\|_\infty = 1$; (otherwise one may consider $\frac{1}{\|f\|_\infty} f(\cdot)$). So $|f(\cdot)| \leq 1$ a.s. Hence $\int |f(x)|^p dx \leq \int |f(x)|^r \ dx < \infty$, proving the lemma. □

Proof of Proposition 6.1: Case (i): $\|f\|_\infty < \infty$. We may assume $\|f\|_\infty > 0$. By the Lemma 6.3, $f \in L^p(dx) \ \ \forall\ p \geq r$. Now define the measure λ by $d\lambda(x) = |f(x)|^r dx$. Note that λ is a finite measure on $\mathbb{R}$. Now observe that

$$\begin{aligned} \|f\|_p &= \left(\int |f(x)|^p \, dx\right)^{1/p} = \left(\int |f(x)|^{p-r} \, d\lambda(x)\right)^{1/p} \\ &= (\|f\|_{p-r,\lambda})^{(p-r)/p} = \|f\|_{p-r,\lambda} \cdot (\|f\|_{p-r,\lambda})^{-r/p} \end{aligned}$$

for any $p \geq r$. By Lemma 6.2 note that r.h.s. $\to \|f\|_\infty$ as $p \to \infty$.

Case (ii): $\|f\|_\infty = +\infty$. In this case we need to show that $\liminf_{p\to\infty} \|f\|_p = +\infty$. Suppose not. Then there is a sequence $p_n \to \infty$ such that $\|f\|_{p_n} \leq K$ for all n, for some constant K. As $\|f\|_\infty = +\infty$, note that Leb. meas. $(|f| \geq 4K) = C_K > 0$. For every n, it follows that

$$\|f\|_{p_n} \geq \left(\int\limits_{\{|f|\geq 4K\}} |f(x)|^{p_n} dx\right)^{1/p_n} \geq 4K \ (C_K)^{1/p_n}.$$

We know that $(C_K)^{1/p_n} \to 1$ as $n \to \infty$. Choose n_0 such that $(C_K)^{1/p_n} > \frac{1}{2}$ for all $n \geq n_0$. Then $\|f\|_{p_n} \geq 2K$ for all $n \geq n_0$, which is a contradiction. $\square$

Note: Suppose $f(\cdot) \equiv 1$. The $\|f\|_\infty = 1$, but $f \notin L^p(dx)$ for any $p \geq 1$. So the hypothesis $f \in L^r(dx)$ for some $r \geq 1$ cannot be dropped. $\square$

Remark 6.4 *(Heuristics)* Let $\gamma(\cdot)$ be a nice bounded function on $\mathbb{R}$, in the sense that all integrability requirements below hold. Using Proposition 6.1 we have

$$\begin{aligned}
\sup\{\gamma(x) : x \in \mathbb{R}\} &= \log \| \exp(\gamma(\cdot))\|_\infty \\
&= \lim_{n\to\infty} \log \|e^{\gamma(\cdot)}\|_n \\
&= \lim_{n\to\infty} \log \left(\int_{\mathbb{R}} \exp(n\gamma(x))\ dx \right)^{1/n} \\
&= \lim_{n\to\infty} \frac{1}{n} \log \left(\int_{\mathbb{R}} e^{n\gamma(x)} dx \right). \qquad (6.1)
\end{aligned}$$

Now take $\gamma(x) = g(x) - I(x)$, $x \in \mathbb{R}$. Then by (6.1)

$$\sup\{g(x) - I(x) : x \in \mathbb{R}\} = \lim_{n\to\infty} \frac{1}{n} \log \left(\int_{\mathbb{R}} e^{ng(x)}\ e^{-nI(x)}\ dx \right). \quad (6.2)$$

(6.1) or (6.2) is called *Laplace's method* in classical asymptotic analysis.

We had indicated earlier that the function $I(\cdot)$ could be similar to $\Lambda^*(\cdot)$ of Cramer's theorem. So define $P_n(dx) = e^{-nI(x)}\ dx$, and *assume* that P_n is a probability measure for each $n \geq 1$. Now let $a \in \mathbb{R}$ be a sufficiently large number. By an analysis similar to the one leading to (6.1), (6.2), but on the set $[a, \infty)$ with $g(\cdot) \equiv 0$ we get

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log P_n([a,\infty)) \\
= &\lim_{n\to\infty} \frac{1}{n} \log \int_{[a,\infty)} e^{-nI(x)} dx \\
= &\sup\{-I(x) : x \geq a\} = -\inf\{I(x) : x \geq a\}. \qquad (6.3)
\end{aligned}$$

Clearly (6.3) reminds us of Cramer's theorem. Moreover, if $\{P_n\}$ are as above then by (6.2)

$$\lim_{n\to\infty} \frac{1}{n} \log \int e^{ng(x)}\, dP_n(x) = \sup\{g(x) - I(x) : x \in \mathbb{R}\} \qquad (6.4)$$

for any reasonable function $g(\cdot)$.

The above discussion suggests the following: If $\{P_n\}$ is a family of suitable probability measures (as encountered in Cramer's theorem) then expressions like (6.2) can be obtained, serving as tractable approximations to solutions to certain differential equations. What is needed now is a general framework as suggested by Varadhan. □

Let (S, d) be a complete separable metric space. Let $\{P_\epsilon : \epsilon > 0\}$ be a family of probability measures on $(S, \mathcal{B}(S))$. The idea is to characterize the limiting behaviour of $\{P_\epsilon\}$, as $\epsilon \downarrow 0$, in terms of a function. The required abstraction is contained in two key definitions.

Definition 6.5 *(i) A function $I : S \to [0, \infty]$ is called a* rate function *if $I \not\equiv \infty$ and if the level set $\{x \in S : I(x) \leq c\}$ is compact in S for each $c < \infty$. (In particular, a rate function is lower semicontinuous.)*

(ii) Let $\{P_\epsilon : \epsilon > 0\}$ be a family of probability measures on $(S, \mathcal{B}(S))$. The family $\{P_\epsilon\}$ is said to satisfy the large deviation principle with rate function *I if*
(a) $I(\cdot)$ is a rate function
(b) for every closed set $C \subseteq S$

$$\limsup_{\epsilon\to 0} \epsilon \log P_\epsilon(C) \leq -\inf\{I(y) : y \in C\}. \qquad (6.5)$$

(c) for every open set $U \subset S$

$$\liminf_{\epsilon\to 0} \epsilon \log P_\epsilon(U) \geq -\inf\{I(y) : y \in U\}. \qquad (6.6)$$

Note: In the above definition we have taken a family $\{P_\epsilon : \epsilon > 0\}$ of probability measures, instead of sequences as in previous sections; quite a few applications to stochastic processes involve asymptotic behaviour as parameter $\epsilon \downarrow 0$. To apply above definition and results on LDP to sequence of probability measures / random variables one can take $\epsilon = \frac{1}{n}$.

Example 6.6 Let $X_1, X_2, \ldots$ be i.i.d. real valued random variables with common distribution μ. Take $\epsilon = \frac{1}{n}$. Let P_n denote the distribution of $\frac{1}{n}\ S_n = \frac{1}{n}\ (X_1 + \ldots + X_n),\ n \geq 1$. That is, $P_n(A) = P(\frac{1}{n} S_n \in A) = \mu^{*(n)}(nA), A \in \mathcal{B}(\mathbb{R})$, where $\mu^{*(n)}$ denotes the n-fold convolution of μ. So Cramer's theorem says that $\{P_n\}$ satisfies the large deviation principle with rate function Λ^*, the Fenchel-Legendre transform of the logarithmic moment generating function of μ.

Example 6.7 Let $A = \{a_1, a_2, \ldots, a_k\} \subset \mathbb{R}$, and $S = M = \{$all probability measures on $A\}$. Note that S can be identified with a compact subset of $\mathbb{R}^k$, and hence is a complete separable metric space. Take $\epsilon = \frac{1}{n}$. Let P_n denote the distribution of the empirical distribution L_n, $n \geq 1$. Note that $P_n(B) = P(\{\omega : L_n(\omega) \in B\})$, $B \in \mathcal{B}(S)$; so P_n is a probability measure on $(S, \mathcal{B}(S))$. Hence Sanov's theorem says that $\{P_n\}$ satisfies the large deviation principle with rate function $H(\cdot \mid \mu)$, the relative entropy function w.r.t. μ. □

Proposition 6.8 *Let $\{P_\epsilon\}$ satisfy the large deviation principle. Then the associated rate function is unique.*

Proof: Suppose $I(\cdot), J(\cdot)$ are two rate functions for $\{P_\epsilon\}$. We write $I(B) = \inf\{I(z) : z \in B\}$, $J(B) = \inf\{J(z) : z \in B\}$. Fix $x \in S$. Let $\delta > 0$. By definition of large deviation principle we have

$$\begin{aligned} -I(x) &\leq -I(B(x:\delta')) \\ &\leq \liminf_{\epsilon \to 0}\ \epsilon \log\ P_\epsilon(B(x:\delta')) \\ &\leq \limsup_{\epsilon \to 0}\ \epsilon \log\ P_\epsilon(\overline{B(x:\delta')}) \\ &\leq -J\ (\overline{B(x:\delta')}) \leq -J(B(x:\delta)) \end{aligned}$$

for all $\delta' < \delta$. Let $\delta \downarrow 0$. As J is lower semicontinuous, by equivalent condition (ii) in Proposition 3.2, we get $\lim_{\delta \downarrow 0}\ J(B(x:\delta)) = J(x)$. Thus $I(x) \geq J(x)$. Reversing the roles of $I(\cdot)$ and $J(\cdot)$ we get $I(x) \leq J(x)$. So $I(\cdot) = J(\cdot)$. □

Note: In the course of the above proof we have proved

$$\lim_{\delta \downarrow 0}\ \liminf_{\epsilon \downarrow 0}\ \epsilon \log P_\epsilon(B(x:\delta)) \quad = \quad -I(x)$$

$$\lim_{\delta\downarrow 0} \limsup_{\epsilon\downarrow 0} \epsilon \log P_\epsilon(\overline{B(x:\delta)}) \;=\; -I(x).$$

□

The next result fulfills the promise made in Remark 6.4. It is an extension of Laplace's method to an abstract setting, and is known as Varadhan's lemma.

Theorem 6.9 *Let $\{P_\epsilon\}$ satisfy the large deviation principle with rate function $I(\cdot)$. Then for any bounded continuous function $g(\cdot)$ on S*

$$\lim_{\epsilon\downarrow 0} \epsilon \log \left[\int_S \exp\left(\frac{1}{\epsilon} g(x)\right) \, dP_\epsilon(x) \right] = \sup\{g(x) - I(x) : x \in S\}. \quad (6.7)$$

Proof: Upper bound. Let $\delta > 0$ be arbitrary but fixed. As g is bounded, Range$(g) \subseteq$ compact interval. Taking a partition of that interval of mesh $< \delta$, and using continuity of $g(\cdot)$, one can find an integer $N = N(\delta)$, closed sets $C_1, C_2, \ldots, C_N$ such that $S = \bigcup_{i=1}^{N} C_i$, $g(x) \le g_j + \delta$ for all $x \in C_j$ where $g_j = \inf\{g(y) : y \in C_j\}$, $1 \le j \le N$. Then

$$\begin{aligned} \int_S \exp\left(\frac{1}{\epsilon}\, g(x)\right) \, dP_\epsilon(x) \;&\le\; \sum_{j=1}^{N} \int_{C_j} \exp\left(\frac{1}{\epsilon}\, g(x)\right) \, dP_\epsilon(x) \\ &\le\; \sum_{j=1}^{N} P_\epsilon(C_j) \, \exp\left[\frac{1}{\epsilon}(g_j + \delta)\right]. \end{aligned}$$

By Lemma 4.8, and (6.5) in Definition 6.5 (ii) we now get

$$\begin{aligned} & \limsup_{\epsilon\downarrow 0} \epsilon \log \left[\int_S \exp\left(\frac{1}{\epsilon}\, g(x)\right) \, dP_\epsilon(x) \right] \\ \le\; & \max_{1\le j\le N} \limsup_{\epsilon\downarrow 0} \epsilon \, \log \left[P_\epsilon(C_j) \, \exp\left(\frac{1}{\epsilon}(g_j + \delta)\right) \right] \\ =\; & \max_{1\le j\le N} \limsup_{\epsilon\downarrow 0} \left[(g_j + \delta) + \epsilon \, \log \, P_\epsilon(C_j) \right] \\ \le\; & \max_{1\le j\le N} \left[(g_j + \delta) + (-I(C_j)) \right] \end{aligned}$$

$$
\begin{aligned}
&= \max_{1\le j\le N} \sup_{x\in C_j} [g_j - I(x)] + \delta \\
&\le \max_{1\le j\le N} \sup_{x\in C_j} [g(x) - I(x)] + \delta \\
&= \sup_{x\in S} (g(x) - I(x)) + \delta.
\end{aligned}
$$

As $\delta > 0$ is arbitrary we get the required upper bound.

Lower bound. Let $\delta > 0$ be arbitrary. Choose $y \in S$ such that $g(y) - I(y) \ge \sup_x [g(x) - I(x)] - \frac{1}{2}\delta$. One can find an open neighbourhood U of y such that $g(x) \ge g(y) - \frac{1}{2}\delta$ for all $x \in U$ by continuity of $g(\cdot)$. Then using (6.6) in Definition 6.5(ii) we have

$$
\begin{aligned}
&\liminf_{\epsilon\downarrow 0} \ \epsilon \ \log \left(\int_S \ \exp\left(\frac{1}{\epsilon} g(x)\right) \ dP_\epsilon(x) \right) \\
\ge\ &\liminf_{\epsilon\downarrow 0} \ \epsilon \ \log \left[\int_U \ \exp\left(\frac{1}{\epsilon} \ g(x)\right) \ dP_\epsilon(x) \right] \\
\ge\ &\liminf_{\epsilon\downarrow 0} \ \epsilon \ \log \left[\left\{ \exp \ \frac{1}{\epsilon} \ \left(g(y) - \frac{1}{2}\delta \right) \right\} \ P_\epsilon(U) \right] \\
=\ &\liminf_{\epsilon\downarrow 0} \ \left[\left(g(y) - \frac{1}{2} \ \delta \right) + \epsilon \ \log \ P_\epsilon(U) \right] \\
\ge\ &\left[g(y) - \frac{1}{2} \ \delta \right] - I(U) \ge [g(y) - I(y)] - \frac{1}{2} \ \delta \\
\ge\ &\sup_z \ [g(z) - I(z)] - \delta
\end{aligned}
$$

by our choice of y. As $\delta > 0$ is arbitrary we get the desired lower bound.

This completes the proof. □

The preceding result is the key to many deep applications of large deviations. However, in this introductory notes we will be content with the following toy example.

Example 6.10 Let $X_1, X_2, \ldots$ be i.i.d.'s such that $P(X_i = \frac{1}{2}) = \frac{1}{2} = P(X_i = \frac{3}{2})$. So $m = E(X_i) = 1$. Let P_n be the distribution of $\frac{1}{n} \ S_n = \frac{1}{n} \ (X_1 + X_2 + \ldots + X_n)$, $n \ge 1$. By Cramer's theorem $\{P_n\}$ satisfies large

deviation principle with rate function

$$I(x) = \begin{cases} \log 2 + (x - \frac{1}{2}) \ \log(x - \frac{1}{2}) + (\frac{3}{2} - x) \ \log(\frac{3}{2} - x), \ \frac{1}{2} \le x \le \frac{3}{2} \\ +\infty, \quad \text{otherwise.} \end{cases}$$

Now, as P_n is supported on $[\frac{1}{2}, \frac{3}{2}]$,

$$\begin{aligned} E\left[\left(\frac{1}{n} \ S_n\right)^n\right] &= \int\limits_{\mathbb{R}} x^n \ dP_n(x) \\ &= \int\limits_{[\frac{1}{2}, \frac{3}{2}]} \exp(n \log x) \ dP_n(x) \\ &= \int\limits_{\mathbb{R}} \exp(n \ g(x)) \ dP_n(x) \end{aligned}$$

where $g(x) = 1_{[\frac{1}{2}, \frac{3}{2}]}(x) \log x$. Now by Varadhan's lemma

$$\begin{aligned} & \lim_{n \to \infty} \ \frac{1}{n} \ \log E\left[\left(\frac{1}{n} \ S_n\right)^n\right] \\ = \ & \lim_{n \to \infty} \ \frac{1}{n} \ \log \left[\int \exp(n \ g(x)) \ dP_n(x)\right] \\ = \ & \sup \ \{g(x) - I(x) : x \in \mathbb{R}\} \\ = \ & \sup \ \left\{\log x - I(x) : \frac{1}{2} \le x \le \frac{3}{2}\right\} \stackrel{\Delta}{=} b, \text{ say.} \end{aligned} \qquad (6.8)$$

Put $h(x) = \log x - I(x)$; note that $h(1) = 0$, $h'(1) > 0$. So $h(\cdot)$ does not attain its maximum at $x = 1$. Hence $b > 0$. Though this example is somewhat artificial, it is instructive. By the law of large numbers we know that $\frac{1}{n} \ S_n \to 1$ a.s. Hence one might naively expect l.h.s. of (6.8) to be zero. But as seen above it is not so. Thus $E[(\frac{1}{n} \ S_n)^n]$ is basically determined (for large n) not by the almost sure behaviour but by the rare event when $\frac{1}{n} \ S_n$ takes values near b. (It turns out $b \approx 0.1$.) $\square$

We end this section with the following result, which is called contraction principle.

Theorem 6.11 *Let $\{P_\epsilon\}$ be a family of probability measures on a complete separable metric space (S, d) satisfying the large deviation principle with*

rate function $I(\cdot)$. Let $(\hat{S}, \hat{d})$ be another complete separable metric space; let $\pi : S \to \hat{S}$ be a continuous function. Put $\hat{P}_\epsilon = P_\epsilon \pi^{-1}$, $\epsilon > 0$. Then $\{\hat{P}_\epsilon\}$ also satisfies the large deviation principle with rate function $\hat{I}(\cdot)$ given by

$$\hat{I}(y) = \begin{cases} \inf\{I(x) : \pi(x) = y\}, & y \in \textit{ Range } (\pi) \\ +\infty, & y \notin \textit{ Range } (\pi). \end{cases} \qquad (\ 6.9\ )$$

Proof: Let $0 \leq c < \infty$. As π is continuous and I is rate function note that $\{y : \hat{I}(y) \leq c\} = \pi(\{x : I(x) \leq c\})$. Hence $\{y : \hat{I}(y) \leq c\}$ is compact. Clearly $\hat{I}(\cdot) \not\equiv \infty$. So $\hat{I}(\cdot)$ is a rate function. For any $\hat{A} \subseteq \hat{S}$ note that

$$\inf\{\hat{I}(y) : y \in \hat{A}\} = \inf\{I(x) : x \in \pi^{-1}(\hat{A})\}.$$

(This works even if $\pi^{-1}(\hat{A}) = \phi$, as infimum over an empty set is $+\infty$.) Moreover by continuity of $\pi(\cdot)$, if $\hat{A}$ is open (resp. closed) then $\pi^{-1}(\hat{A})$ is also open (resp. closed). Consequently the defining conditions (6.5), (6.6) of Definition 6.5 (ii) can easily be checked for $\{\hat{P}_\epsilon\}$. □

We have already come across an instance of contraction principle in Proposition 5.5, where we derived a special case of Cramer's theorem from Sanov's theorem for finite sets.

We have barely scratched the surface of the subject. A curious reader can dig into the references given below for more information.

Acknowledgement: This write up is an enlarged version of a series of lectures given at a Summer School on Probability and Applications held at Kerala School of Mathematics, Kozhikode during May - June 2010.

References

[1] J.A. Bucklew : *Large Deviations Techniques in Decision, Simulation and Estimation*. Wiley, New York, 1990.

[2] A. Dembo and O. Zeitouni : *Large Deviations Techniques and Applications*. Second edition (corrected printing), Springer, New York, 2010.

[3] J.D. Deuschel and D.W. Stroock : *Large Deviations*. Academic Press, Boston, 1989.

[4] P. Dupuis and R.S. Ellis : *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, New York, 1997.

[5] R.S. Ellis : *Entropy, Large Deviations and Statistical Mechanics*. Springer, New York, 2006.

[6] M.I. Freidlin and A.D. Wentzell : *Random Perturbations of Dynamical Systems*. Springer, New York, 1998.

[7] F. den Hollander : *Large Deviations*. Amer. Math. Soc., Providence, 2000.

[8] B.V. Rao : *Large Deviations*. (Informal Notes). Unpublished, 2009.

[9] A. Schwartz and A. Weiss : *Large Deviations for Performance Analysis, Queues, Communications and Computing*. Chapman and Hall, London, 1995.

[10] S.R.S. Varadhan : *Large Deviations and Applications*. SIAM, Philadelphia, 1984.

[11] S.R.S. Varadhan : Workshop on Large Deviations : Lecture Notes. *Bulletin of Kerala Mathematics Association*, October 2009 (Special Issue), pp. 1-14.