

Introduction to Large Deviations Theory

I. Introduction

Let $\{Y_n\}$ be a sequence of random variables and suppose that $Y_n \xrightarrow{P} m$, for some constant m . If $A \subset (m - \epsilon, m + \epsilon)^c$, where $\epsilon > 0$, the event $\{Y_n \in A\}$ is represents a large deviation Y_n from where it ‘wants’ to be with high probability when n is large, namely, near m . Large deviation theory addresses the rate at which the probability of such events tends to 0 as $n \rightarrow \infty$. This question is interesting in and of itself and is important in applications, especially when the limit $Y_n \xrightarrow{P} m$ arises from a law of large numbers. In general, Y_n may not represent simply a real-valued random variable, but a random variable taking values in an infinite dimensional vector space, or even a random measure. Here, we study the simplest case: $Y_n = (1/n) \sum_1^n X_i$, where X_1, X_2, \dots are i.i.d. random variables.

The starting point of the theory is a simple upper bound that is a consequence of Markov’s inequality. If Y is a random variable, $M_Y(t) = E[e^{tY}]$, $t \in \mathbb{R}$, denotes its moment generating function (m.g.f.). This may be infinite for some t , but we think of $M_Y(t)$ as taking values in the extended reals, so we consider it to be defined for all t . The log moment generating function is $\Lambda_Y(t) = \ln M_Y(t)$, where, by convention, $\ln(\infty) = \infty$ and $e^\infty = \infty$. For any number a and for any $t \geq 0$, $\{Y \geq a\} = \{e^{tY} \geq e^{ta}\}$. Therefore, by Markov’s inequality,

$$\mathbb{P}(Y \geq a) \leq \frac{E[e^{tY}]}{e^{ta}} = e^{-(at - \Lambda_Y(t))}.$$

To get the best bound from this inequality, we should optimize over t . Thus,

$$\mathbb{P}(Y \geq a) \leq \exp \left\{ - \sup_{t \geq 0} [at - \Lambda_Y(t)] \right\} \quad (1)$$

Suppose now that X_1, X_2, \dots are i.i.d., with common m.g.f. $M(t)$ and log m.g.f. $\Lambda(t) = \ln M(t)$. Let $S_n := \sum_{i=0}^n X_i$. Then, using the independence,

$$\begin{aligned} M_{S_n/n}(t) &= E \left[\exp \left\{ \frac{1}{n} \sum_1^n X_i \right\} \right] = M^n \left(\frac{t}{n} \right), \quad \text{and hence} \\ \Lambda_{S_n/n}(t) &= n\Lambda(t/n). \end{aligned}$$

Since $\sup_{t \geq 0} at - n\Lambda(t/n) = n \sup_{t \geq 0} a(t/n) - \Lambda(t/n) = n \sup_{t \geq 0} (at - \Lambda(t))$, it follows from applying (1) to $Y = S_n$ that

$$\mathbb{P}(S_n/n \geq a) \leq \exp\{-n \sup_{t \geq 0} (at - \Lambda(t))\}. \quad (2)$$

Replacing X by $-X$ and a by $-a$ in this inequality leads to

$$\mathbb{P}(S_n/n \leq a) \leq \exp\{-n \sup_{t \leq 0} (at - \Lambda(t))\}. \quad (3)$$

When $\sup_{t \geq 0} at - \Lambda(t) > 0$ in (2), $\mathbb{P}(S_n \geq na)$ decays at least at an exponential rate. Both for this case and in the general setting, $\{Y_n \in A\}$, we can ask if large deviations probabilities decay at exact exponential rates. By this we mean,

$$\mathbb{P}(Y_n \in A) = e^{-n(I_A + o(1))},$$

where $o(1)$ denotes a correction term that tends to 0 as $n \rightarrow \infty$. Equivalently,

$$\text{does } \lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}(Y_n \in A) \text{ exist?}$$

The theorems of large deviation theory are usually expressed in terms of this limit. It is not always possible to identify an exact rate. A large deviation upper bound is any quantity U satisfying

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}(Y_n \in A) \leq U,$$

and a large deviation lower bound is any quantity L satisfying

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}(Y_n \in A) \geq L.$$

Thus (2) says that $-\sup_{t \geq 0} (at - \Lambda(t))$ is a large deviation upper bound for $\mathbb{P}(S_n \geq na)$. Of course, if we can prove $U = L$, then we have an exact exponential rate of decay.

It is interesting to note the following for empirical means of i.i.d. sequences.

Theorem 1 *If X_1, X_2, \dots are i.i.d. and $S_n = \sum_1^n X_i$, then $\frac{1}{n} \ln \lim_{n \rightarrow \infty} \mathbb{P}(S_n \geq na)$ exists.*

The proof uses a lemma about sequences that is useful in other contexts. A sequence $\{a_n\}$ is *subadditive* if $a_{n+m} \geq a_n + a_m$ for all n and m . For example, $\{\ln \mathbb{P}(S_n \geq na)\}$ defines a subadditive sequence. If $\mathbb{P}(X_n < a) = 1$, then $\mathbb{P}(S_n \geq na) = 0$ for all n , and so $\ln \mathbb{P}(S_n \geq na) = -\infty$ for all n , a sequence which is subadditive in a trivial sense. If $\mathbb{P}(X_n \geq a) > 0$, then $\mathbb{P}(S_n \geq na) > 0$ for all n , and hence $\{\ln \mathbb{P}(S_n \geq na)\}$ is a sequence of finite numbers. Since $\{S_n \geq na\} \cap \{S_{n+m} - S_n \geq ma\} \subset \{S_{m+n} \geq (m+n)a\}$, since the two events $\{S_n \geq na\}$ and $\{S_{n+m} - S_n \geq ma\}$ are independent, and since S_m and $S_{n+m} - S_n$ are identically distributed,

$$\begin{aligned} \mathbb{P}(S_n \geq na) \mathbb{P}(S_m \geq n) &= \mathbb{P}(S_n \geq na) \mathbb{P}(S_{n+m} - S_n \geq ma) \\ &\leq \mathbb{P}(S_{n+m} \geq (n+m)a). \end{aligned}$$

It follows by taking logs that $\{\ln \mathbb{P}(S_n \geq na)\}$ is subadditive.

Theorem 1 is a direct consequence of the following result.

Lemma 1 *If $\{a_n\}$ is a subadditive sequence, $\lim_{n \rightarrow \infty} a_n/n = \sup_n (a_n/n)$.*

Proof: Let $a_0 = 0$. Let k be any positive integer. For each m and $0 \leq j < k$, $a_{mk+j} \geq ma_k + a_j$. Thus

$$\frac{a_{mk+j}}{mk+j} \geq \frac{a_k}{k} \frac{mk}{mk+j} + \frac{a_j}{mk+j} = \frac{a_k}{k} + \frac{1}{m} \left[-\frac{a_k}{k} \frac{mj}{mk+j} + \frac{a_j m}{mk+j} \right].$$

As the second term is bounded by $(1/m)[|a_k/k| + \sup\{|a_j|; 0 < j < k\}/k]$ in absolute value, it follows that

$$\sup_n \frac{a_n}{n} \geq \liminf \frac{a_n}{n} \geq \frac{a_k}{k} \quad \text{for all } k.$$

The lemma follows by taking a supremum over k . ◇

II. Example: Chernoff bounds for binomial random variables.

Let X_1, X_2, \dots be i.i.d. Bernoulli random variables with $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1 - p$, where $0 < p < 1$. Then $E[X_i] = p$ and $\Lambda(t) = \ln(pe^t + (1-p))$. The application of the upper bounds in (2) and (3) to $S_n = \sum_1^n X_i$ is developed in the following exercise.

Exercise. (a) With the convention that $0 \cdot \ln 0 = 0$,

$$\sup_t (at - \Lambda(t)) = \begin{cases} a \ln \left(\frac{a}{p} \right) + (1-a) \left(\frac{1-a}{1-p} \right), & \text{if } 0 \leq a \leq 1; \\ \infty, & \text{otherwise.} \end{cases} \quad (4)$$

(b) If $a \geq p$, then $\sup_{t \geq 0}(at - \Lambda(t)) = \sup_t(at - \Lambda(t))$ and if $a \leq p$, then $\sup_{t \leq 0}(at - \Lambda(t)) = \sup_t(at - \Lambda(t))$.

(c) By direct application of (2) and (a) and (b), it follows that for $a > p$

$$\mathbb{P}(S_n \geq na) \leq \exp \left\{ -n \left[a \ln \left(\frac{a}{p} \right) + \left(\frac{1-a}{1-p} \right) \right] \right\}.$$

Derive from this that for $a > p$

$$\mathbb{P}(S_n/n \geq p + \epsilon) \leq e^{-2n(a-p)^2}.$$

Show that

$$\mathbb{P}(|S_n/n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Derive from this the strong law of large numbers for $\{X_1, X_2, \dots\}$.

(d) Show that for $a = 0$ or $a = 1$, we in fact have $\mathbb{P}(S_n = na) = e^{-n \sup_t(at - \Lambda(t))}$.

III. Facts about moment generating functions; the Legendre-Fenchel transform of the log m.g.f.

To proceed with the general theory, it is clearly important to study the log moment generating function. We present the important results in this section. Throughout, X is a non-degenerate random variable (meaning that $\mathbb{P}(X = x) < 1$ for every x), $M_X(t) = E[e^{tX}]$, and $\Lambda_X(t) = \ln M_X(t)$. We use the conventions that $\ln \infty = \infty$, $\ln 0 = -\infty$, $e^\infty = \infty$, $e^{-\infty} = 0$. We shall use μ to denote the mean, $E[X]$ when it exists, although we allow it to be infinite. (The mean μ exists as long as $E[X] = E[X^+] - E[X^-]$ makes sense as an extended real number, where X^+ and X^- are the positive and negative parts of X ; μ is not defined only when $E[X^+] = E[X^-] = \infty$.) Any statement made involving μ is made under the assumption it exists.

Finally, define

$$I(a) := \sup_t at - \Lambda_X(t).$$

Notice that the supremum is taken over all $t \in \mathbb{R}$ in this definition. As a function of a , $I(a)$ is called the Legendre-Fenchel transform of Λ . Since $I(a) \geq a \cdot 0 - \Lambda_X(0) = 0$, I is non-negative. The Legendre-Fenchel transform for Bernoulli random variables was computed in equation (4) above.

It is useful to note that

$$at - \Lambda_X(t) = -\ln E[e^{t(X-a)}] = \Lambda_{X-a}(t) \tag{5}$$

Some basic facts are presented and proved in the following items. Convexity is important in these considerations. A function $f : \mathbb{R} \rightarrow (-\infty, \infty]$ is strictly convex if $f((1-\theta)s + \theta t) \leq (1-\theta)f(s) + \theta f(t)$ for all s and t , and the inequality is strict if $f(s)$ and $f(t)$ are finite. Strict concavity is defined in a similar way.

Lemma 2 (a) *Let $\mathcal{D} = \{t; M_X(t) < \infty\}$. Then \mathcal{D} is an interval which includes $t = 0$, $M_X(t)$ is continuous on the closure of \mathcal{D} , it is infinitely differentiable on $\text{int}(\mathcal{D})$, and its derivative of order n is $M_X^{(n)}(t) = E[X^n e^{tX}]$.*

(Saying that $M_X(t)$ is continuous on the closure of \mathcal{D} entails that $\lim_{t \rightarrow \alpha, t \in \mathcal{D}} M_X(t) = M_X(\alpha)$, if α is an endpoint of \mathcal{D} , whether or not α belongs to \mathcal{D} .)

(b) *If $\mathbb{P}(X < 0) > 0$ and $\mathbb{P}(X > 0) > 0$, then $\lim_{t \rightarrow \infty} M_X(t) = \lim_{t \rightarrow -\infty} M_X(t) = \infty$.*

(c) *$M_X(t) \geq e^{t\mu}$, or, equivalently, $\Lambda_X(t) \geq \mu t$.*

It is a consequence of (a) of this Lemma that on \mathcal{D} ,

$$\Lambda'(t) = \frac{M_X'(t)}{M_X(t)} \quad \text{and} \quad \Lambda''(t) = \frac{E[X^2 e^{tX}]E[e^{tX}] - (E[X e^{tX}])^2}{M_X^2(t)} \quad (6)$$

These formulae will be useful later.

Proofs: Observe that $M_X(0) = 1$ whatever X is and so $0 \in \mathcal{D}$. Write $M_X(t) = E[e^{tX} \mathbf{1}_{\{X < 0\}}] + E[e^{tX} \mathbf{1}_{\{X \geq 0\}}]$. (Claim (b) of the theorem is a direct consequence of this representation.) The first term is finite whenever $t \geq 0$, no matter what the distribution of X is. Suppose that $t_1 > 0$ and $M_X(t_1) < \infty$. Then $E[e^{t_1 X} \mathbf{1}_{\{X \geq 0\}}] < \infty$ and it follows by monotonicity of e^{tx} that $E[e^{tX} \mathbf{1}_{\{X \geq 0\}}] < \infty$ for all $0 \leq t \leq t_1$. Hence $[0, t_1] \subset \mathcal{D}$. Similarly, if $t_0 \in \mathcal{D}$, then $[t_0, 0] \subset \mathcal{D}$. If $a = \inf\{t; t \in \mathcal{D}\}$ and $b = \sup\{t; t \in \mathcal{D}\}$, it follows that $(a, b) \subset \mathcal{D}$. The endpoints may or may not be in \mathcal{D} . But if α is an endpoint of \mathcal{D} , the monotone and dominated convergence theorems applied to $E[e^{tX} \mathbf{1}_{\{X < 0\}}] + E[e^{tX} \mathbf{1}_{\{X \geq 0\}}]$ imply $\lim_{t \rightarrow \alpha, t \in \mathcal{D}} M_X(t) = M_X(\alpha)$.

For any $\epsilon > 0$, $|x|^n \leq K_\epsilon(e^{\epsilon x} + e^{-\epsilon x})$ for some finite constant K_ϵ . Thus, $E[|X|^n e^{tX}] < \infty$ if t is in the interior of \mathcal{D} . The existence of a derivative of any order is proved inductively using the bound $|(e^{(t+h)x} - e^{tx})/h| < |x|(e^{(t+\epsilon)x} + e^{(t-\epsilon)x}) \leq K(e^{(t+2\epsilon)x} + e^{(t-2\epsilon)x})$, which holds for $|h| < \epsilon$ and an appropriate finite K , and dominated convergence.

To prove part (c), consider first $|\mu| < \infty$. By the convexity of e^{tx} , $e^{tx} > e^{t\mu} + e^{t\mu}(x - \mu)$ for any x . Hence $E[e^{tX}] \geq e^{t\mu} + te^{t\mu}E[X - \mu] = e^{t\mu}$. (The inequality and this proof are a special case of Jensen's inequality.)

Suppose $\mu = -\infty$. Then $E[X^-] = \infty$ and hence $M_X(t) \geq E[e^{tX} \mathbf{1}_{\{X < 0\}}] = \infty$ if $t < 0$. But if $t < 0$, then $t\mu = \infty$ also, and so $e^{t\mu} = e^\infty = \infty \leq M_X(t)$. For $t > 0$, $e^{t\mu} = e^{-\infty} = 0 < M_X(t)$. The case $\mu = \infty$ is similar. \diamond

Lemma 3 Λ_X is strictly convex.

This fact uses the standing assumption that X is non-degenerate. By the Cauchy-Schwarz inequality, $E^2[Xe^{tX}] \leq E[X^2e^{tX}]E[e^{tX}]$. Equality can only hold if $e^{(t/2)X} = cXe^{(t/2)X}$ a.s. for some constant c ; if X is non-degenerate, as we are assuming, this cannot happen and hence the inequality is strict. From (6), it follows that $\Lambda''(t) > 0$ in the interior of \mathcal{D} . This and the fact that Λ is continuous on the closure of \mathcal{D} imply strict convexity. \diamond

We turn now to a study of the function I , which is called the *rate function*, since it will be used to characterize the exponential rate of decay of large deviation probabilities.

As a first case, consider a random variable X such that $E[e^{tX}] = \infty$ for all $t \neq 0$; this will occur for example whenever the mean μ is not defined, but also when the tail probabilities $\mathbb{P}(X > x)$ and $\mathbb{P}(X < -x)$ are both bounded below by c/x^n as $x \rightarrow \infty$. In this case $at - \lambda_X(t) = -\infty$ whenever $t \neq 0$. It follows that $I \equiv 0$.

If X is degenerate, say $\mathbb{P}(X = \mu) = 1$, then $I(a) = \sup(at - \mu t)$ and so $I(\mu) = 0$, while $I(a) = \infty$ whenever $a \neq \mu$.

The next lemma describes how I behaves when \mathcal{D} contains a non-trivial open interval, (which implies that the mean μ of X exists, although it may be infinite).

Lemma 4 Assume $\mu = E[X]$ exists.

- (i) If $a \geq \mu$, $I(a) = \sup_{t \geq 0}(at - \Lambda(t))$; if $a \leq \mu$, $I(a) = \sup_{t \leq 0}(at - \Lambda(t))$.
- (ii) I is decreasing on $(-\infty, \mu)$ and increasing on (μ, ∞) . If μ is finite, $I(\mu) = 0$. If $|\mu|$ is infinite, $\lim_{a \rightarrow \mu} I(a) = 0$.
- (iii) I is convex and lower semi-continuous.
- (iv) If there exists $t_a \in \mathcal{D}$ such that $\Lambda'(t_a) = M'(t_a)/M(t_a) = a$, then $I(a) = at_a - \Lambda(t_a)$.
- (v) If a is finite and $\mathbb{P}(X \leq a) = 0$ (or $\mathbb{P}(X \geq a) = 0$), then

$$I_X(a) = -\ln \mathbb{P}(X = a). \quad (7)$$

Proof: (i). By Lemma 2(c), $at - \Lambda_X(t) = -\Lambda_{X-a}(t) \leq -(\mu - a)t = (a - \mu)t$. Thus if $a \geq \mu$, $at - \Lambda_X(t) \leq 0$ if $t < 0$, implying that $\sup_{t \leq 0}(at - \Lambda_X(t)) = \sup_{t \geq 0}(at - \Lambda_X(t))$.

Claim (ii) is a direct consequence of (i). Only the claim that $\lim_{a \rightarrow \mu} I(a) = 0$ if μ is infinite, requires argument. Without loss of generality, assume $\mu = -\infty$ so that $I(a) = \sup_{t \geq 0} (at - \Lambda_X(t))$. Since $\Lambda(t)$ is a convex function, there exists a linear function $ct + b$ such that $\Lambda(t) \geq ct + b$ for all t . Then $at - \Lambda(t) \leq (a - c)t - b$. It follows that $\sup_{t \geq \epsilon} (at - \Lambda(t)) \leq (a - c)\epsilon - b$ if $a < c$. Thus $\lim_{a \rightarrow -\infty} I(a) = \lim_{a \rightarrow -\infty} \sup_{0 \leq t < \epsilon} (at - \Lambda_X(t)) \leq -\inf_{0 \leq t \leq \epsilon} \Lambda_X(t)$ for all $\epsilon < 0$. However, either $\Lambda_X(t) = \infty$ for all $t > 0$ or $\lim_{t \downarrow 0} \Lambda_X(t) = 0$. Thus, letting $\epsilon \downarrow$ implies $\lim_{a \rightarrow -\infty} I(a) = 0$.

The supremum of a family of convex functions is convex. Since I is a supremum of linear functions, it is convex. Let $\lim a_k = a$; since $\lim_{k \rightarrow \infty} (a_k t - \Lambda(t)) = (at - \Lambda(t))$ for every t , $\liminf I(a_k) \geq at - \Lambda(t)$ for all t , and hence $\liminf I(a_k) \geq I(a)$, proving lower semi-continuity.

For (iv), observe that $at - \Lambda_X(t) = -\Lambda_{X-a}(t)$ is concave on \mathcal{D} and hence it has a unique maximum at a critical point.

To prove (7) observe that if $\mathbb{P}(X \geq a) = 0$, $at - \Lambda_X(t) = -\ln E[e^{t(X-a)}]$ is an increasing function of t . By the dominated convergence theorem, its limit as $t \rightarrow \infty$ is $-\ln \mathbb{P}(X = a)$. \diamond

The nicest case to deal with mathematically is that for which $M_X(t) \leq \infty$ for all t . This will certainly be true when X is a bounded random variable. It is also true that

$$M_{X,K}(t) := E \left[e^{tX} \mathbf{1}_{\{|X| \leq K\}} \right] < \infty \quad \text{for all } t,$$

when K is finite, which suggests truncation will be a useful technique. However, $M_{X,K}(0) = \mathbb{P}(|X| \leq K)$ and thus it is not the moment generating function of a random variable if $\mathbb{P}(|X| \leq K) < 1$. To get around this define the measure

$$\mathbb{F}(U) := \mathbb{P}(X \in U \mid |X| \leq K) := \frac{\mathbb{P}(X \in U, |X| \leq K)}{\mathbb{P}(|X| \leq K)}.$$

on the Borel subsets of \mathbb{R} . This is clearly a probability measure. Let Y be a random variable with \mathbb{F} as its distribution: $\mathbb{P}(Y \in U) = \mathbb{F}(U)$. (For convenience of notation, we assume Y is defined on the same probability space as X ; this can always be done by augmenting the original probability space so that it supports such a Y independent of X .) Then,

$$M_Y(t) = \frac{M_{X,K}(t)}{\mathbb{P}(|X| \leq K)}.$$

To see this, it is only necessary to observe that for positive, Borel measurable functions h ,

$$E[h(Y)] = \frac{E[h(X)\mathbf{1}_{|X| \leq K}]}{\mathbb{P}(|X| \leq K)}.$$

This is proved first for simple functions h , for which it follows directly from the definition of the distribution of Y , and then for general Borel measurable functions by approximation and passing to the limit.

If Y_1, Y_2, \dots , are i.i.d. with the distribution of Y and if X_1, \dots are i.i.d. with the distribution of X , then for bounded Borel $h : \mathbb{R}^n \rightarrow \mathbb{R}$,

$$E[h(Y_1, \dots, Y_n)] = \frac{E \left[h(X_1, \dots, X_n) \prod_1^n \mathbf{1}_{\{|X_i| \leq K\}} \right]}{\mathbb{P}^n(|X| \leq K)} \quad (8)$$

This formula is proved in a similar way. The family of sets $U \subset \mathbb{R}^n$ such that (8) is true for $h = \mathbf{1}_U$ is a monotone class. But (8) is certainly true of $h = \mathbf{1}_U$ when U is in the algebra of finite disjoint unions of rectangles of the form $A_1 \times \dots \times A_n$, where the A_i are Borel subsets of \mathbb{R} . Hence by the monotone class theorem, (8) holds for $\mathbf{1}_U$ whenever U is any Borel set. Therefore it holds true for simple Borel measurable functions, and by approximations and limit arguments for any bounded Borel function.

We shall use (8) in proving a large deviation lower bound. Indeed, it implies that

$$\begin{aligned} \mathbb{P}\left(\frac{\sum_1^n X_i}{n} \in A\right) &\geq \mathbb{P}\left(\left\{\frac{\sum_1^n X_i}{n} \in A\right\} \cap \bigcap_1^n \{|X_i| \leq K\}\right) \\ &= \mathbb{P}^n(|X| \leq K) \mathbb{P}\left(\frac{\sum_1^n Y_i}{n} \in A\right) \end{aligned} \quad (9)$$

With regard to this construction, the following will be helpful. Let $I_K(a) = \sup_t(at - \Lambda_{X,K}(t))$, where $\Lambda_{X,K}(t) = \ln M_{X,K}(t)$.

Lemma 5 *If $\lim_{|t| \rightarrow \infty} M_{X,K}(t) = \infty$ for some $K > 0$, then $\lim_{K \rightarrow \infty} I_K(a) = I(a)$.*

Proof: Since $M_{X,K} \leq M_{X,K'}$ if $K' > K$, $I_K(a)$ is decreasing as K increases and $I_K(a) \geq I(a)$ for all K . Let $\alpha > I(a)$ and consider the sets

$$V_K = \{t; at - \Lambda_{X,K}(t) \geq \alpha\}.$$

$at - \Lambda_{X,K}(t)$ is concave and continuous on the closure of the domain where it is finite; this is a consequence of Lemmas (2)–(4), since $\Lambda_{X,K}(t)$ is a constant

multiple of a log moment generating function. Hence V_K is closed for every K . Since $M_{X,K} \leq M_{X,K'}$ if $K' > K$, the sets V_K decrease as K increases, and by the assumption of the Lemma, they eventually become compact. If they were all non-empty, there would be a value t such that $at - \Lambda_{X,K}(t) \geq \alpha$ for all K , and hence, since $\lim_{K \rightarrow \infty} \Lambda_{X,K}(t) = \Lambda_X(t)$, $I(a) \geq \alpha$, which is a contradiction. Therefore there must be a K such that $V_{K'}$ is empty for $K' \geq K$. This implies that $\lim_{K \rightarrow \infty} I_K(a) \leq \alpha$. Letting $\alpha \downarrow I(a)$, it follows that $\lim_{K \rightarrow \infty} I_K(a) = I(a)$. \diamond

IV. Cramér's theorem.

Let X be a random variable. Let X_1, X_2, \dots be i.i.d. random variables with the same distribution as X ; hence $M_{X_i}(t) = M_X(t)$ and $\Lambda_{X_i} = \Lambda_X$ for all i . Define $I(a) = \sup_t (at - \Lambda_X(t))$, as in section III, and $S_n = \sum_1^n X_i$. We deduce from equation (2) and Lemma 4 (i) and (ii), that for $a \geq \mu$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P} \left(\frac{1}{n} S_n \geq a \right) \leq -I(a) = - \inf_{x \geq a} I(x). \quad (10)$$

Similarly, from equation (3),

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P} \left(\frac{1}{n} S_n \geq a \right) \leq -I(a) = - \inf_{x \geq a} I(x). \quad (11)$$

These equations reveal the general form that the upper and lower large deviation bounds take. In what follows \bar{A} denotes the closure of a subset A of \mathbb{R} .

Theorem 2 *Let X_1, X_2, \dots be independent, identically distributed random variables. For any Borel subset A of \mathbb{R} ,*

$$- \inf_{\text{int}(A)} I(x) \leq \liminf \frac{1}{n} \mathbb{P} \left(\frac{S_n}{n} \in A \right) \leq \limsup \frac{1}{n} \mathbb{P} \left(\frac{S_n}{n} \in A \right) \leq - \inf_{\bar{A}} I(x) \quad (12)$$

If the mean μ of the X_i exists and $a > \mu$,

$$\lim \frac{1}{n} \mathbb{P} \left(\frac{S_n}{n} \geq a \right) = - \inf_{x \geq a} I(x) \quad (13)$$

Proof of the upper bound. We have more or less proved this by the previous calculations. Let us first dispose of the uninteresting cases. If the mean does not exist, $M_x(t) = \infty$ whenever $t \neq 0$, and so $I(a) \equiv 0$. But $\ln \mathbb{P}(S_n/n \in A) \leq \ln(1) = 0$ always, and so the upper bound is trivially true.

Suppose that μ exists, is finite, and is in \bar{A} , then $I(\mu) = 0$, and again the upper bound is trivially true. If μ is ∞ or $-\infty$ and there is a subsequence $\{a_n\}$ converging to μ then, by Lemma 4 (ii) of the previous section, it follows that $\inf_A I(x) = 0$ and so the bound is again trivially.

It remains to deal with the case in which μ exists and is not in the closure of A . Consider first the case in which μ is finite. Then $A \subset (-\infty, a_1] \cup [a_2, \infty)$ where $a_1 = \sup\{x \in A; x < \mu\}$ and $a_2 = \inf\{x \in A; x > \mu\}$. Because I is decreasing on $(-\infty, \mu)$ and increasing on (μ, ∞) . $\min\{I(a_1), I(a_2)\} = \inf_A I(x)$. From (10) and (11)

$$\mathbb{P}\left(\frac{1}{n}S_n \in A\right) \leq e^{-nI(a_1)} + e^{-nI(a_2)},$$

which implies that

$$\limsup \frac{1}{n} \ln \mathbb{P}\left(\frac{1}{n}S_n \in A\right) = -\min\{I(a_1), I(a_2)\} = -\inf_A I(x).$$

If, say, $\mu = -\infty$, and $a = \inf\{x; s \in A\}$, then, since I is decreasing everywhere as a function of A , $I(a) = \inf_A I(x)$ and the upper bound in (12) is a direct consequence of (10).

Proof of the lower bound. It is enough to show that for any a and any $\delta > 0$,

$$\liminf \frac{1}{n} \ln \mathbb{P}\left(\frac{S_n}{n} \in (a-\delta, a+\delta)\right) \geq -I(a) \quad (14)$$

Since the left hand side is decreasing in δ , it is actually enough to show that

$$\liminf \frac{1}{n} \ln \mathbb{P}\left(\frac{S_n}{n} \in (a-\delta, a+\delta)\right) \geq -I(a) - r(\delta) \quad \text{where } r(\delta) \rightarrow 0, \text{ as } \delta \downarrow 0. \quad (15)$$

We first dispose of a relatively uninteresting case. Assume that $\mathbb{P}(X > a) = 0$. Then, using Lemma 4 (v),

$$\begin{aligned} \liminf \frac{1}{n} \ln \mathbb{P}\left(\frac{S_n}{n} \in (a-\epsilon, a+\epsilon)\right) &\geq \liminf \frac{1}{n} \ln \mathbb{P}\left(\frac{S_n}{n} = a\right) \\ &= \ln \mathbb{P}(X_i = a) = -I(a), \end{aligned}$$

Applying this result to $-X_i$, proves the lower bound also when $\mathbb{P}(X_i < a) = 0$.

The remaining cases will employ the important technique of changing probability measures. We pause to discuss this idea in general. Suppose that Z

is a non-negative random vector, on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $E[Z] = 1$. Then

$$\mathbb{P}_Z(U) = E[\mathbf{1}_U Z], \quad U \in \mathcal{F},$$

defines a new probability measure on (Ω, \mathcal{F}) . This is easy to check; \mathbb{P}_Z is countable additive by the monotone convergence theorem and $\mathbb{P}_Z(\Omega) = 1$. Let E_Z denote expectation with respect to \mathbb{P}_Z . The definition of \mathbb{P}_Z implies that for any simple random variable Y ,

$$E_Z[Y] = E[ZY]. \quad (16)$$

By approximating random variables from below by simple functions, one proves this formula is true for all Y such that the right-hand side makes sense. If, in addition, $\mathbb{P}(Z > 0) = 1$, the same formulas in reverse give

$$E[Y] = E_Z[Z^{-1}Y]. \quad (17)$$

To prove the large deviation lower bound, we will use the change of probability measure defined by letting $Z = e^{tS_n}/M_X^n(t)$ for a value of t in the interior of $\mathcal{D} = \{t; M_X(t) < \infty\}$. To check this is valid, we need to show $E[Z] = 1$, but this is an easy consequence of the fact that $E[e^{tS_n}] = M_X^n(t)$. Thus, for a given n , let $\mathbb{P}_{n,t}$ be the probability defined by

$$\mathbb{P}_{n,t}(U) = \frac{E[\mathbf{1}_U e^{tS_n}]}{M_X^n(t)}.$$

Lemma 6 *Let $t \in \text{int}(\mathcal{D})$. With respect to the probability measure $\mathbb{P}_{n,t}(U)$, X_1, \dots, X_n are i.i.d. random variables with the probability distribution measure*

$$\mathbb{P}(B) = \frac{E[\mathbf{1}_B(X)e^{tX}]}{M_X(t)} \quad \text{and mean} \quad \frac{E[Xe^{tX}]}{M_X(t)} \quad (18)$$

Also

$$\mathbb{P}\left(\frac{S_n}{n} \in A\right) = e^{n\Lambda_X(t)} E_{n,t} \left[\mathbf{1}_A(S_n/n) e^{-tS_n} \right] \quad (19)$$

The important point of this lemma is that the new probability measure shifts the mean of the random variables. Also the new distribution and mean are *independent* of n .

Proof: For any Borel measurable subsets B_1, \dots, B_n of \mathbb{R} , using the fact that X_1, \dots, X_n are i.i.d. with the distribution of X under the original probability

measure,

$$\mathbb{P}_{n,t}(X_1 \in B_1, \dots, X_n \in B_n) = \frac{E\left[\prod_1^n \mathbf{1}_{B_i}(X_i) e^{t \sum_1^n X_i}\right]}{M_X^n(t)} = \prod_1^n \frac{E[\mathbf{1}_{B_i}(X) e^{tX}]}{M_X(t)},$$

which proves that X_1, \dots, X_n are also i.i.d. under $\mathbb{P}_{n,t}$ with probability distribution measure \mathbb{F} . Moreover, applying (16),

$$E_{n,t}[X_1] = \frac{E[X_1 e^{tX_1} e^{t(\sum_2^n X_i)}]}{M_X^n(t)} = \frac{E[X e^{tX}] E[e^{t \sum_2^n X_i}]}{M_X^n(t)} = \frac{E[X e^{tX}]}{M_X(t)}.$$

Equation (19) is an immediate consequence of (18) and formula (17). \diamond

We now turn to the case in which $\mathbb{P}(X_i > a)$ and $\mathbb{P}(X_i < a)$ are strictly positive and $M_X(t) < \infty$ for all t . By writing,

$$at - \Lambda_X(t) = -\Lambda_{X-a}(t) = -\ln \left[E[e^{t(X-a)} \mathbf{1}_{\{X>a\}}] + E[e^{t(X-a)} \mathbf{1}_{\{X \leq a\}}] \right].$$

we see that in this case $\lim_{|t| \rightarrow \infty} -\Lambda_{X-a}(t) = -\infty$. Also $at - \Lambda_X(t)$ is an everywhere differentiable, strictly concave function, by Lemma 2 (a) and Lemma 3. Therefore, it achieves its supremum $I(a)$ at a unique t_a satisfying

$$a = \Lambda'(t_a) = \frac{E[e^{t_a X}]}{M_X(t_a)}.$$

Therefore Lemma 6 implies that under the new measure

$$\mathbb{P}_{n,t_a}(U) = \frac{E[\mathbf{1}_U e^{t_a S_n}]}{M_X^n(t_a)},$$

X_1, \dots, X_n are i.i.d. with mean a and a distribution that does not depend on n . It follows from the weak law of large numbers that

$$\lim_{n \rightarrow \infty} \mathbb{P}_{n,t_a} \left(\frac{S_n}{n} \in (a - \delta, a + \delta) \right) = 1 \quad (20)$$

This is basically what is needed to complete the proof. By first using (19) and then bounding $e^{t_a S_n}$ from below on the event $\{S_n \in (a - \delta, a + \delta)\}$

$$\begin{aligned} \frac{1}{n} \ln \left[\mathbb{P} \left(\frac{S_n}{n} \in (a - \delta, a + \delta) \right) \right] &= \Lambda_X(a) + \frac{1}{n} \ln \left[E_{n,t_a} \left[\mathbf{1}_{|S_n/n - a| < \delta} e^{-t_a S_n} \right] \right] \\ &\geq \Lambda_X(a) + \ln \left[e^{-n a t_a - n |t_a| \delta} \mathbb{P}_{n,t_a} \left(\frac{S_n}{n} \in (a - \delta, a + \delta) \right) \right] \\ &= -I(a) - |t_a| \delta + \frac{1}{n} \ln \left[\mathbb{P}_{n,t_a} \left(\frac{S_n}{n} \in (a - \delta, a + \delta) \right) \right] \end{aligned}$$

Thus, because of (20),

$$\frac{1}{n} \ln [\mathbb{P}\left(\frac{S_n}{n} \in (a-\delta, a+\delta)\right)] \geq -I(a) - |t_a|\delta.$$

This proves (15) and so completes the proof when $M_X(t) < \infty$ for all t .

The only case that remains is that for which $\mathbb{P}(X > a)$ and $\mathbb{P}(X < a)$ but for which $M_X(t)$ is not finite for all t . To handle this, we will use the truncation outlined in Section III. Recall the notation, $M_{X,K}(t) = E[e^{tX} \mathbf{1}_{|X| \leq K}]$, $\Lambda_{X,K}(t) = \ln M_{X,K}(t)$, and $I_K(a) = \sup_t (at - \Lambda_{X,K}(t))$. Let Y_1, Y_2, \dots be i.i.d. with distribution measure $\mathbb{F}_Y(B) = \mathbb{P}(X \in B \mid |X| \leq K)$, as defined in section III. Then $M_Y(t) = M_{X,K}(t)/\mathbb{P}(|X| \leq K)$ and so $\Lambda_Y(t) = \Lambda_{X,K}(t) - \ln \mathbb{P}(|X| \leq K)$. It follows that $I^Y(a) := \sup_t (at - \Lambda_Y(t)) = \sup_t (at - \Lambda_{X,K}(t)) + \ln \mathbb{P}(|X| \leq K) = I_K(a) + \ln \mathbb{P}(|X| \leq K)$. When K is large enough, $\mathbb{P}(Y > a)$ and $\mathbb{P}(Y < a)$, and, since $M_Y(t)$ is finite for all t , we then know the large deviation lower bound holds for Y_1, Y_2, \dots . Thus

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}\left(\frac{\sum_1^n Y_i}{n} \in (a-\delta, a+\delta)\right) \geq -I_K(a) - \ln \mathbb{P}(|X| \leq K) \quad (21)$$

Now we shall use the inequality (9). Taking logarithms of both sides,

$$\frac{1}{n} \ln \mathbb{P}\left(\frac{S_n}{n}\right) \geq \frac{1}{n} \ln \mathbb{P}\left(\frac{\sum_1^n Y_i}{n} \in (a-\delta, a+\delta)\right) + \ln \mathbb{P}(|X| \leq K).$$

and it follows that for every finite K

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}\left(\frac{S_n}{n}\right) \geq -I_K(a).$$

But Lemma 5 implies that $I_K(a) \rightarrow I(a)$ as $K \rightarrow \infty$. This completes the proof of the lower bound.

Proof of (11): We want to prove that if $a \geq \mu$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}\left(\frac{1}{n} S_n \geq a\right) \leq -I(a) = -\inf_{x \geq a} I(x).$$

Because of the form of the upper and lower bounds, this will certainly be true if $\inf_{x > a} I(x) = \inf_{x \geq a} I(x)$. Suppose on the other hand that $\inf_{x > a} I(x) > \inf_{x \geq a} I(x)$. This occurs only if I is discontinuous at a . Since I is convex and lower semicontinuous, it is continuous on $\{y; I(y) < \infty\}$. Therefore, since I is increasing on (μ, ∞) , a discontinuity occurs at a only if $I(x) = \infty$ for all $x > a$.

But this can be true only if $\mathbb{P}(X > a) = 0$. In this case we know (see Lemma 5 (v)) that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P} \left(\frac{1}{n} S_n \geq a \right) = \ln \mathbb{P}(X = a) = -I(a) = - \inf_{x \geq a} I(x). \quad \diamond$$

Remarks. 1. In the case when $M_X(t) = \infty$ for all $t \neq 0$, $I(a) \equiv 0$. Therefore the large deviation upper and lower bounds coincide for any open set A and, when A is open,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbb{P}(S_n/n \in A) = 0.$$

Thus, if $\mathbb{P}(S_n/n \in A)$ does decay to 0, it does so more slowly than $e^{-n\gamma}$ for any positive γ .

2. Consider a random variable with the following properties: (i) $b := \sup\{t; M_X(t) < \infty\}$ satisfies $0 < b < \infty$; (ii) μ exists; and (iii) $\alpha := \sup_t \Lambda'_X(t) < \infty$.

An example with $b = 1$ is the random variable with density,

$$f_X(x) = c \frac{e^{-x}}{1+x^3} \mathbf{1}_{\{x>0\}},$$

where c is an appropriate normalizing constant.

Since $\Lambda_X(t)$ is convex, $\Lambda'_X(t)$ is increasing on its domain of definition. Since it is bounded by α , it follows that $\Lambda_X(b) < \infty$ and, from monotone convergence that

$$\alpha = \lim_{t \uparrow b} \Lambda'_X(t) = \frac{E[Xe^{bX}]}{E[e^{bX}]}.$$

Therefore, for $a \geq \alpha$, $I(a) = ab - \Lambda_X(b)$.

These notes are based on material in Chapter 1 of Durrett, *Probability: Theory and Examples*, first edition, and in Chapter 2 of Dembo and Zeitouni, *Large Deviations; Techniques and Applications*.