LMS/EPSRC Short Course Stability, Coupling Methods and Rare Events
Heriot-Watt University, Edinburgh, 4-9 September 2006
Lecture Notes on:
Large deviations principle
Takis Konstantopoulos

1 History

The story begins with Khinchin[1] (1929) who proves the CLT for a general random walk, followed by Cramér[2] (1938) who obtained the first LD result under additional assumptions. These were removed later by Chernoff[3] (1952) and hence the standard LD theorem is called Chernoff's theorem. This is what we are going to talk about here, albeit cast in more modern terms.

2 Random walks with exponential tails; upper bound

We consider $S_n = \sum_{j=1}^n X_j$, $n \geq 1$, $S_0 = 0$, where the X_j are i.i.d. real-valued random variables with common distribution F that possesses some exponential moments, which means that $Ee^{\theta X_1} < \infty$ for some $\theta \neq 0$ (otherwise the theorem below says nothing). Let $\mu = EX_1$.

We know that $S_n/n \to 1$, a.s. Chernoff's theorem tells us how fast the probability $P(S_n/n > x)$ decays to 0 for some $x > \mu$. It is not a surprise that the rate is exponential, but Chernoff's theorem gives the exact rate, i.e. the exact rate at which the logarithm $\log P(S_n/n > x)$ goes to zero, that's why it's called a logarithmic asymptotic. What is surprising is that there is a *beautiful* theory, both mathematically and physically about this rate of convergence. It appears in Physics (Hamiltonian vs Lagrangian formulation of Newtonian Mechanics), in Statistics, in Laplace's method for asymptotics of integrals, in Optimisation (Duality transforms), and, in Information Theory (Shannon's theorems) and, of course, in Probability.

To discover the theorem, let us start with some "trivialities". Markov's inequality says that, for any nondecreasing function $\varphi : \mathbb{R} \to \mathbb{R}_+$,

$$P(S_n/n > x) \leq P(\varphi(S_n) \geq \varphi(nx)) \leq \frac{E\varphi(S_n)}{\varphi(nx)},$$

so we get the BEST upper bound by finding the least value of the RHS over all such functions. This is (a) a hard job and (b) probably futile for many functions will give $E\varphi(S_n) = \infty$. If we restrict to the exponential class of functions $\varphi_\theta(x) = \exp(\theta x)$, $\theta > 0$ then we know that at least one of them won't give ∞ (this we assumed). So

$$P(S_n/n > x) \leq \frac{Ee^{\theta S_n}}{e^{\theta n x}} = \big(Ee^{\theta X_1}\big)^n e^{-\theta n x} = \exp\big\{ - n\big(\theta x - \log Ee^{\theta X_1}\big)\big\}$$

and so if we *maximise* $\theta x - \log Ee^{\theta X_1}$ over all θ we have the best upper bound within the class of functions considered. So then let $\Lambda(\theta) := \log Ee^{\theta X_1}$ (this is called cumulative-generating

[1] A Russian who wrote in German
[2] A Swede who wrote in French.
[3] An American who wrote in English.

function of X, or, simply, cumulant) and denote the best exponent by

$$\Lambda^*(x) := \sup_{\theta>0}(\theta x - \Lambda(\theta))$$

So we have a Theorem:

Theorem 1.

$$P(S_n/n > x) \le e^{-n\Lambda^*(x)}, \quad x > \mu, \quad n \in \mathbb{N}.$$

Incidentally, this implies that (Borel-Cantelli at work here)

$$P(S_n/n > x \text{ i.o. }) = 0, \quad \text{for all } x > \mu,$$

which is $\frac{1}{2}$SLLN, and the other half can be obtained by doing the same with $-S_n$.

3 The Legendre-Fenchel transform

Let us take a closer look at Λ and Λ^*. (Rockafellar (1970) is a classic on these things.)

A word of "convention": Let $\overline{\mathbb{R}}$ be $\mathbb{R}$ with $+\infty$ attached to it. When $f : \mathbb{R} \to \overline{\mathbb{R}}$ is a convex function, i.e. the set $\{(x, y) \subseteq \mathbb{R} \times \mathbb{R} : y \ge f(x)\}$ is a convex set, we let $\mathcal{D}_f$ be the set $\{f < \infty\}$. Recall that any convex function is the sup of all affine functions below it:

$$f(x) = \sup_{\substack{a,b\in\mathbb{R} \\ ax+b\le f(y)\ \forall y}} (ax + b).$$

Side note: an affine function g is called *supporting* to f at x if $f \ge g$ and $f(x) = g(x)$. Incidentally, this immediately proves *Jensen's inequality*

$$Ef(X) \ge f(EX),$$

for, clearly, $E \sup_j Z_j \ge \sup_j EZ_j$ (for any collection $\{Z_j\}$ of r.v.'s) and, trivially, $E(aY + b) = aEY + b$, for any r.v. Y. Putting these together, along with DCT[4], gives a few nice properties. (Whenever I talk about $\Lambda'(\theta)$, I do assume that Λ is defined on an open neighbourhood of θ.)

1. Λ is convex and C^1 on $\mathcal{D}_\Lambda$.
2. $\Lambda(0) = 0$, $\Lambda'(0) = \mu$.
3. $\Lambda^*(x) = \sup_{\theta\in\mathbb{R}}(\theta x - \Lambda(\theta))$ is a convex function : $\mathbb{R} \to \overline{\mathbb{R}}$, that agrees with the previously defined $\Lambda^*(x)$ for $x > \mu$.
4. $\Lambda(\theta) = \sup_{x\in\mathbb{R}}(\theta x - \Lambda^*(x))$.
5. $\theta x \le \Lambda(\theta) + \Lambda^*(x)$ (*Fenchel-Young inequality*)
6. $\boxed{\Lambda'(\theta) = \dfrac{EXe^{\theta X}}{Ee^{\theta X}} = E^1_\theta X}$, where E^1_θ is expectation with respect to a probability measure.

[4]Dominated Convergence Theorem

7. If $\Lambda'(\theta) = y$ then $\Lambda^*(y) = \theta y - \Lambda(\theta)$.

The mapping $\Lambda \mapsto \Lambda^*$ is called *Legendre-Fenchel transform* and is invertible on the space of convex functions. The geometric interpretation is important: Fix an x and find the supporting affine function to Λ that has *slope* x. This is the function $\theta \mapsto x\theta - \Lambda^*(x)$. Since the LF transform is invertible, all properties have their dual analogues.

Note that $\Lambda \mapsto \Lambda^*$ does not require convexity of Λ, yet it always yields a convex Λ^*. However, convexity of Λ ensures invertibility of the FL transform.

Note also that, in general, a convex function may have many supporting affine functions at a point. However, here, differentiability ensures uniqueness.

Some examples:

1. Gaussian: Let X be standard normal. Then $\Lambda(\theta) = \frac{1}{2}\theta^2$. The sup is $\sup_x(\theta x - \frac{1}{2}\theta^2)$ is achieved at $x = \theta$ and so $\Lambda^*(x) = \frac{1}{2}x^2$. The function $\Lambda(\theta) = \frac{1}{2}\theta^2$ is the only one that satisfies $\Lambda = \Lambda^*$.

2. Bernoulli: $P(X = 1) = p$, $P(X = 0) = 1 - p$. Then $\Lambda(\theta) = \log(pe^\theta + (1 - p))$. Since $\Lambda'(\theta) \in (0, 1)$, we know that the interior of the domain of Λ^* is $(0, 1)$ and we easily find
$$\Lambda^*(x) = x \log(x/p) + (1 - x) \log((1 - x)/(1 - p)), \quad x \in [0, 1],$$
where $\Lambda^*(1) = \Lambda^*(0) = 0$. Note: The *Kullback-Leibler* distance of the probability measure Q from the probability measure P is defined by $D(Q \parallel P) := \int dQ \log \frac{dQ}{dP}$. So $\Lambda^*(x) = D(\mathrm{BER}_x \parallel \mathrm{BER}_p)$. This is not fortuitous.

Physics The LF transform is the one that links the Lagrangian and Hamiltonian formulations of Mechanics. The equations of motion of a particle from point A at time 0 to point B at time 1 in space (here space=$\mathbb{R}^1$ but it could be $\mathbb{R}^d$) are found by looking as the extrema of the action functional
$$\int_0^1 L(x, \dot{x})dt$$
The function $L(x, v)$ is the Lagrangian (which in a motion in a field generated by potential $U(x)$ equals the kinetic energy minus the potential). Euler's equations (necessary conditions for optimality in the Calculus of Variations) result into the differential equation of motion (mass times acceleration equals force). Hamilton's formulation is cast in terms of another functional known as Hamiltonian $H(x, p)$. The relation between $v \mapsto L(x, v)$ and $p \mapsto H(x, p)$ is precisely that of a LF transform:
$$H(x, p) = \sup_v \big(pv - L(x, v)\big).$$
So, the v that achieves the supremum satisfies $p = \frac{\partial}{\partial v}L(x, v)$. All this is generalisable to $\mathbb{R}^d$ and, indeed, to a d-dimensional manifold.

Optimisation Every convex optimisation problem (minimising a convex function on a convex set) has an associated dual. This duality is expressed by an LF pair. See Luenberger (1969) for a readable account with geometric intuition.

4 Lower bound; Chernoff's theorem

So far we have an upper bound. We now deal with a lower bound. To make things concrete, we work, without loss of generality, on the canonical probability space $\Omega = \mathbb{R}^{\mathbb{N}}$ with the product σ-field, P the product measure $F^{\otimes\mathbb{N}}$ and let $X_n(\omega) = \omega_n$, $n \in \mathbb{N}$. We would like to derive a lower bound for $P(S_n/n > x)$ when $x > \mu$. Fix $a > \mu$. We will estimate $P(|S_n/n - a| < \varepsilon)$ for some small $\varepsilon > 0$. The idea is to change probability measure and make the event $\{|S_n/n - a| < \varepsilon\}$ be as likely as possible. We have $a \in \mathcal{D}^o_{\Lambda^*}$ and so there is $\theta > 0 \equiv \theta(a) > 0$ such that

$$\Lambda'(\theta) = a.$$

Recall that $\Lambda'(\theta)$ is the expectation of X under a different probability measure:

$$F_\theta(B) := \frac{E\big(e^{\theta X_1}; X_1 \in B\big)}{e^{\Lambda(\theta)}}.$$

Now consider a different probability measure on Ω, namely the one that is defined by products of F_θ, i.e. $P_\theta = F_\theta^{\otimes\mathbb{N}}$. We then have

$$E_\theta X_1 = a, \quad \lim_{n\to\infty} P_\theta(|S_n/n - a| < \varepsilon) = 1 \text{ for all } \varepsilon > 0.$$

This allows us to perform the following magic:

$$P(|S_n/n - a| < \varepsilon) = P_\theta(|S_n/n - a| < \varepsilon; \frac{dP}{dP_\theta}) = P_\theta(|S_n/n - a| < \varepsilon; e^{n\Lambda(\theta)} e^{-\theta S_n})$$
$$\geq e^{n\Lambda(\theta)} e^{-\theta n(a+\varepsilon)} P_\theta(|S_n/n - a| < \varepsilon) = e^{-n[\theta(a+\varepsilon) - \Lambda(\theta)] + o(1)} \geq e^{-n\Lambda^*(a+\varepsilon) + o(1)}, \quad \text{as } n \to \infty,$$

and so

$$\underline{\lim}_{n\to\infty} \frac{1}{n} \log P(|S_n/n - a| < \varepsilon) \geq -\Lambda^*(a + \varepsilon),$$

But $P(S_n/n > x) \geq P(|S_n/n - (x + \varepsilon)| < \varepsilon)$ and so

$$\underline{\lim}_{n\to\infty} \frac{1}{n} \log P(|S_n/n - x| > \varepsilon) \geq -\Lambda^*(x + 2\varepsilon),$$

and, because Λ^* is continuous, we have proved

Theorem 2.

$$\underline{\lim}_{n\to\infty} \frac{1}{n} \log P(S_n/n > x) \geq -\Lambda^*(x), \quad x > \mu, \quad x \in \mathcal{D}^o_{\Lambda^*}.$$

Theorem 3 (Chernoff)**.** *Let $S_n = X_1 + \cdots + X_n$ be a random walk in $\mathbb{R}^1$ with mean $\mu = EX_1$, cumulant $\Lambda(\theta) = Ee^{\theta X}$ whose LF transform is $\Lambda^*(x)$. Then, if $x > \mu$ and for some $\varepsilon > 0$, $\Lambda^*(x + \varepsilon) < \infty$,*

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n/n > x) = -\Lambda^*(x).$$

Proof. Theorem 1 + Theorem 2. □

Terminology: $\Lambda^*(x)$ is called RATE FUNCTION because it quantifies the rate of convergence of $P(S_n/n > x)$ to zero.

5 Large deviations

We pass on to estimating $P(S_n/n \in B)$ for arbitrary Borel sets B, not necessarily semi-infinite intervals.

We assume that $\Lambda(\theta) < \infty$, to avoid technicalities. (Modifications, otherwise, are easy.)

We first assume that $B = [a, b]$ is compact interval and that $\mu < a < b$. Then $\Lambda^*(a) \leq \Lambda^*(b)$, so the rate of convergence of $P(S_n/n \geq a)$ dominates that of $P(S_n/n \geq b)$ and that is why

$$\frac{1}{n} \log P(S_n/n \in [a, b]) \to -\Lambda^*(a),$$

when $\mu < a < b$.

We then consider an arbitrary $B = [a, b]$ and, arguing likewise, we find

$$\frac{1}{n} \log P(S_n/n \in [a, b]) \to -\min_{x \in [a,b]} \Lambda^*(x).$$

This motivates the definition that, for an arbitrary Borel set B, the quantity

$$\Lambda^*(B) := \inf_{x \in B} \Lambda^*(x)$$

represents the "slowest rate in B".

With a bit extra work of topological nature (cover compact sets by finite ε-covers and take an appropriate ε-interval within an open set–one over which the rate is close to the infimum), we can prove:

Theorem 4 (large deviations theorem)**.** *For any Borel set B,*

$$-\Lambda^*(B^o) \leq \underline{\lim} \frac{1}{n} P(S_n/n \in B) \leq \overline{\lim} \frac{1}{n} P(S_n/n \in B) \leq -\Lambda^*(\overline{B}),$$

where B^o is the interior and $\overline{B}$ the closure of B.

Equivalent to the last statement is the conjunction of the following two statements:

$$\text{For all closed } F, \quad \overline{\lim} \frac{1}{n} P(S_n/n \in F) \leq -\Lambda^*(\overline{F}),$$
$$\text{For all open } G, \quad \underline{\lim} \frac{1}{n} P(S_n/n \in G) \geq -\Lambda^*(\overline{G}).$$

Weak convergence Recall that if $P, (P_n)_{n \in \mathbb{N}}$ are probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ then P_n *converges weakly*[5] to P if and only if

$$\text{For all closed } F, \quad \overline{\lim} P_n(F) \leq P(F),$$
$$\text{For all open } G, \quad \underline{\lim} P_n(G) \geq P(G).$$

You will see the connection between the two in the lectures by Tolya Puhalskii and in more general framework. For the time being, just accept the consequence of the last theorem as a definition in more general situations.

Let $\mathbb{U}$ be a Hausdorff topological space. A *rate* function $I : \mathbb{U} \to \mathbb{R}_+$ is a lower-semicontinuous.[6] function with closed sub-level sets[7] The Borel sets $\mathscr{B}(\mathbb{U})$ is the class

[5] This means that $\int f dP_n \to \int f dP$ for all bounded continuous $f : \mathbb{R} \to \mathbb{R}$ (something that in Analysis is known as weak* convergence).

[6] Lower semicontinuity means that the *epigraph* $\{(x, y) \in \mathbb{U} \times \mathbb{R} : \ y \geq I(x)$ is a closed set in the product topology of $\mathbb{U} \times \mathbb{R}$.

[7] A sub-level set is a set of the form $\{x \in \mathbb{U} : \ I(x) \leq y\}$ for some $y \in \mathbb{R}$.

of all sets that are obtained from countable set operations on the open sets of $\mathbb{U}$. For any $B \in \mathscr{B}(\mathbb{U})$, we shall let

$$I(B) := \inf_{x \in B} I(x).$$

Definition 1. Let $\mathbb{U}$ be Hausdorff. Let I be a rate function on $\mathbb{U}$. Let $(P_n)_{n\in\mathbb{N}}$ be probability measures on $(\mathbb{U}, \mathscr{B}(\mathbb{U}))$. We say that (P_n) satisfies LDP with rate function I on $\mathbb{U}$ if

$$\text{For all closed } F, \quad \overline{\lim} \frac{1}{n} \log P_n(F) \le -I(F),$$
$$\text{For all open } G, \quad \underline{\lim} \frac{1}{n} \log P_n(G) \ge -I(G).$$

An I-continuity set B is such that $I(B^o) = I(\overline{B})$. In such a case, $\frac{1}{n} \log P_n(B) \to -I(B)$.

The rate function is called *good* if all sub-level sets are not just closed, but *compact*.

Just as weak convergence is transferable from space to space via continuous mappings[8], so is LDP:

Theorem 5 ("contraction principle")**.** *If (P_n) satisfies LDP on $\mathbb{U}$ with good rate function I and if $f : \mathbb{U} \to \mathbb{U}'$ is continuous then $(P_n \circ f^{-1})$ satisfies LDP on $\mathbb{U}'$ with good rate function $I'(y) := \inf\{I(x) : \ x \in f^{-1}\{y\}\}$.*

The big picture Warning: The blurb below is pure heuristics. Large deviations, loosely speaking, means the study of deviations of a random phenomenon around its "mean behaviour", provided that the "randomness" that kicks the system away from its mean is not too bad (e.g. exponential tails). A big area of applications of Large Deviations is in studying the deviation of the trajectory of a stochastic dynamical system from its "mean" trajectory. The mean trajectory is typically obtained via so-called fluid limits, which is tantamount to a functional law of large numbers. A concrete example of this is: Consider the deterministic system

$$\dot{x} = -\mu x, \quad x(0) = a > 0.$$

Nothing could be simpler than this. The physical phenomenon could be, for example, discharging a capacitor, or watching the way that the evolution of the number of mathematical schools that maintain some standards. We can see this system as the limit of a Markov chain $(X_t, t \ge 0)$ in $\mathbb{Z}_+$ with rates

$$q(k+1, k) = \mu k, \quad n \in \mathbb{Z}_+.$$

Indeed, if we let $X_0 = [na]$ then the sequence of random functions $(\frac{1}{n} X_{nt}, t \ge 0)$, $n = 1, 2, \ldots$, has as a limit, as $n \to \infty$ the function $x(t)$ that solves the ODE above, in the sense that

$$\lim_{n\to\infty} \sup_{0 \le t \le T} \left| \frac{1}{n} X(nt) - x(t) \right| = 0,$$

for all $T > 0$. Large deviations is the study of the probability that $\frac{1}{n} X(n\cdot)$ differs from $x(\cdot)$ by some amount. Precisely, if we let P_n be the law on $C[0, T]$ of a continuous interpolation to $\frac{1}{n} X(n\cdot)$ and then the sequence (P_n) does obey an LDP with a rate function that can be computed explicitly. Now imagine this situation in a more complicated dynamical system, and you get a feel of what Large Deviations does in relation to dynamical systems.

[8] If P_n converges weakly to P on some space $\mathbb{U}$ and if $f : \mathbb{U} \mapsto \mathbb{U}'$ then $P_n \circ f^{-1}$ converges weakly to $P \circ f^{-1}$ on $\mathbb{U}'$. This you can call CONTRACTION PRINCIPLE for you may think of $\mathbb{U}$ as a BIG SPACE–like a function space–and of $\mathbb{U}'$ as a small space–like $\mathbb{R}^2$.

6 Large deviations in $\mathbb{R}^d$

We can easily extend Chernoff's theorem to random walks in $\mathbb{R}^d$. We let $S_n = X_1 + \cdots + X_n$, where the X_i are i.i.d. with common distribution F for which we assume $\Lambda(\theta) = \log Ee^{\langle\theta, X_1\rangle} < \infty$ for all $\theta \in \mathbb{R}^d$. This is a restriction but it can be removed. We let

$$\Lambda^*(x) := \sup_{\theta\in\mathbb{R}^d} \big(\langle\theta, x\rangle - \Lambda(\theta)\big).$$

Theorem 6. *The sequence of probability measures $P(S_n/n \in \cdot)$, $n \in \mathbb{N}$, satisfies LDP with good rate function Λ^*.*

7 Sanov's theorem

Consider a finite set $\mathcal{A} = \{1, \ldots, d\}$ called *alphabet*, not just for want of a better name, but because it has applications in Information Theory. The elements of $\mathcal{A}$ are called *letters*. A probability measure μ on $\mathcal{A}$ is, obviously, an assignment $\mathcal{A} \ni \alpha \mapsto \mu(\alpha) \in \mathbb{R}_+$ of numbers with $\sum_{\alpha\in\mathcal{A}} \mu(\alpha) = 1$. Given i.i.d. random variables $X_1, X_2, \ldots$ with values in $\mathcal{A}$ and common law μ, we let

$$\mu_n(\alpha) := \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(X_j = \alpha), \quad \alpha \in \mathcal{A},$$

be the *empirical distribution* or *type*[9] of $(X_1, \ldots, X_n)$. By the SLLN,

$$P(\mu_n \to \mu) = 1.$$

Here $\mu_n \to \mu$ means $\mu_n(\alpha) \to \mu(\alpha)$ for all $\alpha \in \mathcal{A}$. The random vectors μ_n live in $\mathbb{R}^d$ and, in particular, are concentrated in the unit simplex

$$\Sigma^d := \big\{x \in \mathbb{R}^d_+ : \sum_{\alpha=1}^{d} x_\alpha = 1\big\}.$$

Sanov's theorem provides the rate of convergence in the the way that $P(|\mu_n - \mu| > \varepsilon)$ approaches 0 as $n \to \infty$, for any $\varepsilon > 0$.

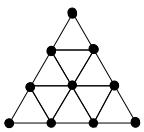
Now let us look at the assignment of the empirical distribution (type) to a specific sequence of length n:

$$A^n \ni (x_1, \ldots, x_n) \longmapsto \mu_n^x \in \Sigma^d,$$

where $\mu_n^x = (\mu_n^x(1), \ldots, \mu_n^x(d))$, $\mu_n^x(\alpha) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(x_j = \alpha)$. The range of this map is denoted by F_n:

$$F_n := \left\{ \left(\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(x_j = \alpha),\ \alpha \in \mathcal{A} \right) \subseteq \Sigma^d :\ x_1, x_2, \ldots, x_n \in \mathcal{A} \right\}.$$

For example, with $\mathcal{A} := \{\alpha, \beta\}$ and $n = 3$, the possible sequences are $(\alpha,\alpha,\alpha), (\alpha,\alpha,\beta)$,..., (β,β,β), and so $F_3(\mathcal{A}) = \{(1,0,0), (0,1,0), (0,0,1),\ (0,1/3,2/3), (0,2/3,1/3), (1/3,0,2/3), (2/3,0,1/3), (1/3,2/3,0), (2/3,1/3,0), (1/3,1/3,1/3)\}$. Since the set F_n is isomorphic to the



F_3 if $d = 2$

[9] Empirical distribution is a terminology used in Statistics: If μ is an unknown distribution then the empirical distribution is estimates μ. Type is a terminology used in Information Theory.

set of ways to put d indistinguishable balls in n boxes, we have that

$$|F_n| = \binom{n+d-1}{d} \leq (n+1)^d.$$

The inverse image of a $\nu \in \Sigma^d$ under the map $x \mapsto \mu_n^x$ is denoted by $T_n(\nu)$: it is the set of sequences of length n that have type ν. So $A^n = \cup_{\nu \in \Sigma^d} T_n(\nu)$. The cardinality of $T_n(\nu)$ is found as follows: Amongst the n letters in a sequence we require that there are $n\nu(\alpha)$ which have value α, $\alpha = 1, \ldots, d$. Think of letters as colours. In how many ways can we arrange n objects, when $n\nu(1)$ of them are of colour 1, etc., $n\nu(d)$ of them are of colour d? The answer is, of course, the multinomial coefficient

$$|T_n(\nu)| = \binom{n}{n\nu(1), \ldots, n\nu(d)} = \frac{n!}{(n\nu(1))! \cdots (n\nu(d))!}.$$

Now let μ_n be the empirical distribution of the random sequence $(X_1, \ldots, X_n)$, which are i.i.d. with common law μ. Then for any $\nu \in F_n$, we have

$$\begin{aligned} P(\mu_n = \nu) &= P\Big(\sum_{k=1}^n \mathbf{1}(X_k = \alpha) = n\nu(\alpha),\ 1 \leq \alpha \leq d \Big) \\ &= \binom{n}{n\nu(1), \ldots, n\nu(d)} \mu(1)^{n\nu(1)} \cdots \mu(d)^{n\nu(d)}. \end{aligned}$$

By using Stirling's approximation, and by doing some algebra we find

$$(n+1)^{-d} e^{-nD(\nu\|\mu)} \leq P(\mu_n = \nu) \leq e^{-nD(\nu\|\mu)}.$$

So if ν is away from the true distribution μ, this probability decays exponentially to zero. Using this we GUESS what the LDP should look like. Define

$$I(\nu) := D(\nu \parallel \mu), \quad \nu \in \Sigma^d.$$

A bit more work and we obtain:

Theorem 7 (Sanov's theorem)**.** *If $\mathcal{A}$ is a finite set and $X_1, X_2, \ldots$ i.i.d. random variables with values in $\mathcal{A}$ and common law μ then, letting μ_n be the empirical distribution of $(X_1, \ldots, X_n)$, the sequence of probability measures $P(\mu_n \in \cdot)$ satisfy an LDP with good rate function I.*

8 Asymptotics for supremum of a random walk (stationary regime of a queue)

Consider a random walk S_n as before, with cumulant Λ, but now suppose that mean of its increment is negative:

$$EX_1 = -\mu < 0.$$

If so, then, clearly, the random variable

$$M := \sup_{n \geq 0} S_n$$

is a.s. finite. This is intimately related to the stationary regime of a queue, something we will see in the Stability Lectures. For now, we shall prove how the distribution of M decays to zero:

Theorem 8 (effective bandwidth theorem[10]). *If* $EX_1 = -\mu < 0$, *then*

$$\lim_{x\to\infty} \frac{1}{x} \log P(M \geq x) = -c,$$

where

$$c := \inf_{t>0} t\Lambda^*(1/t) = \sup\{\theta > 0 : \ \Lambda(\theta) \leq 0\}$$

Note: The exact and logarithmic asymptotics for the probability $P(M > x)$ have been already studied in Lecture 1. Here we propose a different way to obtain the logarithmic asymptotics in the light-tail case.

Proof. First we deal with the lower bound. It rests on what-appears-to-be a trivial observation. Since, obviously,

$$P(M \geq x) \geq P(S_n \geq x) \quad \text{for all } n \in \mathbb{N},\ x > 0,$$

we have

$$P(M \geq x) \geq P(S_{[tx]} \geq x) = P\left(\frac{S_{[tx]}}{tx} \geq \frac{1}{t}\right), \quad \text{for all } t, x > 0.$$

Using Chernoff's theorem, we have

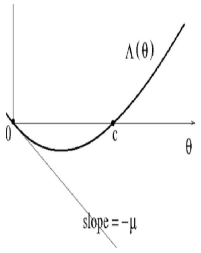
$$\lim_{x\to\infty} \ \frac{1}{tx} \ P\left(\frac{S_{[tx]}}{tx} \geq \frac{1}{t}\right) = -\Lambda^*(1/t),$$

and so

$$\underline{\lim}_{x\to\infty} \ \frac{1}{x} \ P(M \geq x) \geq -\inf_{t>0} t\Lambda^*(1/t) =: -c.$$

The upper bound uses Boole's inequality[11], so it is even more "trivial":

$$P(M \geq x) \leq \sum_{n=1}^{\infty} P(S_n \geq x) = \sum_{n=1}^{\infty} P(e^{\theta S_n} \geq e^{\theta x}) \leq \sum_{n=1}^{\infty} e^{n\Lambda(\theta)-\theta x}, \quad \text{for any } \theta > 0.$$



Pick $\theta > 0$ so that $\Lambda(\theta) < 0$, and then sum the geometric series to get

$$P(M \geq x) \leq e^{-\theta x} \frac{e^{\Lambda(\theta)}}{1 - e^{\Lambda(\theta)}}.$$

The best upper bound is

$$\inf_{\theta>0, \Lambda(\theta)<0} e^{-\theta x} \frac{e^{\Lambda(\theta)}}{1 - e^{\Lambda(\theta)}}$$

which turns out to be e^{-cx}. □

Actuarial Mathematics Take what we said and interpret everything in terms of ruin of a risk process. Then we have the so-called Cramér-Lundberg estimate for a ruin probability. You see, different people may be working in seemingly different things, but, mathematically, they are the same.

[10] This interpretation will be discussed in terms of queues as models of buffers for things like streaming video on the Internet

[11] $P(\cup_n A_n) \leq \sum_n P(A_n)$

9 Non-i.i.d. sequences

Frequently, we need to deal with situations where the random walk does not have independent increments.

H-Fractional Gaussian Walk A Fractional Brownian motion $B_H = (B_H(t), t \geq 0)$ is a zero-mean Gaussian process with correlation function

$$EB_H(t)B_H(s) = \frac{1}{2}\big(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\big),$$

where $H \in (0, 1]$. This specifies the finite-dimensional distributions of the process. (It turns out that there is a version of the process with continuous paths.) If $H = 1$, then $B_1(t) \equiv 0$, and this is not an interesting case.

If $H = 1/2$ then $EB_H(t)B_H(s) = \frac{1}{2}\big(|t| + |s| - |t-s|\big) = \min(t, s)$, and so $B_{1/2}$ is a standard Brownian motion. In this case, the sequence $(B_{1/2}(n), n \in \mathbb{Z}_+)$ is a random walk with i.i.d. increments which have $\mathscr{N}(0, 1)$ distribution.

If $H \in (0, 1)$ then the increments of B_H are correlated (positively if $H > 1/2$, negatively if $H < 1/2$). We consider the sequence

$$S_n^H := B_H(n), \quad n \in \mathbb{Z}_+,$$

has identically distributed but not independent increments. The correct normalisation for the obtention of a SLLN is n^{2H}, namely,

$$\frac{S_n^H}{n^{2H}} \to 0, \quad a.s.$$

Twisting things around, we let

$$Z_n = \frac{S^H_{[n^{1/2H}]}}{n},$$

which also $\to 0$, a.s. If P_n is the law of Z_n then it turns out that (P_n) does obey an LDP with (good) rate function $x^2/2$. Note that, here, the cumulant of Z_n (i.e. of P_n) is

$$\Lambda_{Z_n}(\theta) = \log Ee^{\theta Z_n} = \frac{1}{2}\theta^2 \operatorname{var}(Z_n) = \frac{1}{2}\theta^2 \frac{[n^{1/2H}]^{2H}}{n^2}$$

and so

$$\frac{1}{n}\Lambda_{Z_n}(n\theta) = \frac{1}{2}\theta^2 \frac{[n^{1/2H}]^{2H}}{n} \xrightarrow[n\to\infty]{} \frac{1}{2}\theta^2.$$

This is the situation that is taken care of by:

Theorem 9 (Gärtner-Ellis)**.** *Let (P_n) be a sequence of probability measures on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$,*

$$\Lambda_n(\theta) := \log \int_{\mathbb{R}^d} e^{\langle \theta, x\rangle}\ P_n(dx),$$

and assume that

$$\frac{1}{n}\Lambda_n(n\theta) \xrightarrow[n\to\infty]{} \Lambda(\theta) \in \overline{\mathbb{R}}, \quad \text{for all } \theta \in \mathbb{R}^d,$$

$$\Lambda \text{ is smooth and lower semi-continuous},$$

$$0 \in \Lambda^{-1}(\mathbb{R}).$$

Then (P_n) obeys an LDP with good rate function Λ^*.

Proof: See Dembo and Zeitouni (1993).

10 Discussion; problems

1. Say that x_n is logarithmically equivalent to y_n if $n^{-1}\log x_n - n^{-1}\log y_n \to 0$, as $n \to \infty$. Show that the sum of two sequences is logarithmically equivalent to their maximum. (Does this remind you of the definition of subexponentiality?)

2. Let S_n be Binomial$(n, 1/2)$. Prove, only using Stirling's approximation, that

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

   for all $a \in (1/2, 1)$.

3. Derive all listed properties of the LF transform.

4. Compute the rate function for Poisson and exponential laws.

5. Use Hölder's inequality and Fatou's lemma to show that the cumulant is convex and lower-semicontinuous.

6. Suppose X takes values 1 and 1.5 with equal probability. Let S_n be a random walk with increments distributed according to X. Compute

$$\lim \frac{1}{n} \log E(S_n/n)^n.$$

7. Compare the rough asymptotics for $P(M > x)$ in the light-tailed case with the precise asymptotics derived in yesterday's lecture.

8. Starting from Sanov's theorem and the contraction principle rederive Chernoff's theorem for the special case of simple random variables.

9. I toss a fair coin 1000 times and, amazingly enough, I get strictly more than 990 heads. What is, approximately, the distribution of heads that I got? (Is it, e.g., uniform on $[991, 1000]$?)

10. Consider $I(x) = \int_0^1 L(\dot{x})dt$, where $x : [0,1] \to \mathbb{R}$ is absolutely continuous (and, therefore, $\dot{x}$ denotes a version of its Radon-Nikodým derivative with respect to the Lebesgue measure), and L is a convex function. Guess what $\inf_x I(x)$ is and use Jensen's inequality to prove it. (This is an important rate function associated with processes with independent increments.)

References

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations. *Ann. Math. Stat.* **23**, 493-507.

Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités. *Act. Sci. Indust.* **736**, 5-23.

Dembo, A. and Zeitouni, O. (1993). *Large Deviations Techniques and Applications.* Springer

KHINCHIN, A.Y. (1923). Über einen neuen Grenzwertsatz der Wahrsheinlichkeitsrechnung. *Math. Annalen* **101**, 745-752.

LUENBERGER, D.G. (1969). *Optimization by Vector Space Methods.* Wiley.

ROCKAFELLAR, T.R. (1970). *Convex Analysis.* Princeton.