Large Deviations and Statistical Mechanics

Stefan Adams
October 2012

(1) Introduction and Cramér's Theorem

1.1 Introduction

In the lecture we will give an elementary and rigorous introduction to large deviations techniques and methods and we will use later this knowledge to study large systems of interacting subsystems (many-particle systems) which have among others spatial dependency structures. The overall aim is to introduce the free energy in equilibrium statistical mechanics solely from mathematical principles. That is, we will show that the free energy is nothing else than the rate function of certain large deviations principles. Once having established this link we study different ensembles (probability measures) in statistical mechanics and show that different large deviations principles (level-1 to level-3) provide macroscopic descriptions of varying information depth. This demonstrates that large

deviations principles provide a natural bridge between microscopic description and certain macroscopic description given as variational principles or PDEs providing effective modelling of the interacting system.
Another objet of study will be large deviations for stochastic processes and hydrodynamic limits.
We will study a couple of applications, e.g., as gradient systems (modelling of random interfaces and materials); interacting Brownian motions/bridges as effective modelling for interacting particles; variational principle (maximum entropy principle).

The student might wonder what exactly is meant by large deviations.
Precise definitions and statements are postponed.
Let us elaborate a bit on a simple example.

Let $X_1, X_2, \dots$ be a sequence of independent, Standard Normal, real-valued random variables, and consider the empirical mean

$$\hat{S}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Since $\hat{S}_n$ is again a Normal random variable with zero mean and variance $1/n$, it follows that for any $\delta > 0$,

(1) $\mathbb{P}(|\hat{S}_n| \geq \delta) \xrightarrow[n\to\infty]{} 0$, and, for any interval $A \subset \mathbb{R}$

(2) $\mathbb{P}(\sqrt{n}\,\hat{S}_n \in A) \xrightarrow[n\to\infty]{} \frac{1}{\sqrt{2\pi}} \int_A e^{-\frac{1}{2}x^2} dx$.

Note that

$$\mathbb{P}(|\hat{S}_n| \geq \delta) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} e^{-1/2 x^2} dx,$$

therefore,

(3) $\frac{1}{n} \log \mathbb{P}(|\hat{S}_n| \geq \delta) \xrightarrow[n\to\infty]{} -\frac{\delta^2}{2}$ (exercise)

Equation (3) is an example of a large deviations statement.

"The typical value of $\hat{S}_n$ is, by (2), of the order $\frac{1}{\sqrt{n}}$, but with small probability (of the order $e^{-n\delta^2/2}$), $|\hat{S}_n|$ takes relatively large values.

Since both, (1) and (2), remain valid as long as (X_i)

are i.i.d. r.v.s with mean zero and unit variance, it could be asked whether (3) also holds for non-Normal r.v.s X_i. The answer is that while

$$\lim_{n\to\infty} 1/n \log \mathbb{P}(|\hat{S}_n| \geq \delta) \quad \text{always exists},$$

its value depends on the distribution of the X_i.

This is the content of Cramér's Theorem in Sation 1.2.

Before we briefly outline the curriculum of the lecture.

(1) Cramér's Theorem and Sanov's Theorem
(real-valued i.i.d. r.v.s)

(2) General Principles
(definition, uniqueness, exp. tightness, properties, Varadhan's Lemma ; Bryc's Lemma)

(3) LDP for abstract empirical measures
(sub-additivity ; Spatial dependency structures

④ LDP for statistical mechanics
(ensembles; free energy; variational principle;
Gibbs measure; phase transitions)

⑤ Examples
(gradient models; surface large deviations;
Wulff shape LDP; hydrodynamic limit)

References (see slide)

1.2 Cramér's Theorem for the empirical mean

Let $X_1, X_2, \ldots$ be i.i.d. random variables on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ (i.e., X_i are real-valued).
Write $\mathbb{E}$ to denote the expectation under $\mathbb{P}$, let

$$\mathbb{E}[X_1] = \mu \in \mathbb{R}$$
$$\operatorname{var}(X_1) = \sigma^2 \in (0, \infty),$$

and let $S_n := X_1 + \cdots + X_n$ be the partial sum.

We recall the fundamental theorems dealing with such sequences.

WLLN (Weak Law of Large Numbers)

$$\forall \varepsilon > 0 \qquad \lim_{n \to \infty} \mathbb{P}\left(\left| \frac{1}{n} S_n - \mu \right| > \varepsilon \right) = 0$$

SLLN (Strong Law of Large Numbers)

$$\frac{1}{n} S_n \xrightarrow[\text{as } n \to \infty]{} \mu \qquad \mathbb{P}\text{-a.s.}$$

CLT (Central Limit Theorem)

$$\frac{1}{\sigma \sqrt{n}} (S_n - \mu n) \xrightarrow[\text{as } n \to \infty]{} Z \quad \text{in law w.r.t. } \mathbb{P},$$

where Z is standard normal random variable.

SLLN asserts that the empirical average $\frac{1}{n} S_n$ converges to μ as $n \to \infty$; the CLT quantifies the probability that S_n differs from μn by an amount of order $\sqrt{n}$. Deviations of this size are called "normal".

In this lecture we are studying deviations of the order n, so well beyond what is described by the CLT. Deviations of this size are called "large". A large deviation event $\{S_n \geq (\mu + a)n\}$, $a > 0$, has a probability which goes to zero as $n \to \infty$.
Under certain conditions on the tail of the distribution of X_1, the decay is exponential in n:

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq (\mu + a)n) = -I(a) < \infty, \quad a > 0.$$

Notation: Given two sequences of positive numbers (α_n) and (β_n), we write

$$\alpha_n \simeq \beta_n \iff \lim_{n \to \infty} \frac{1}{n} (\log \alpha_n - \log \beta_n) = 0$$

$$\alpha_n + \beta_n \simeq \alpha_n \vee \beta_n \quad (\text{"largest-exponent-wins"-principle})$$

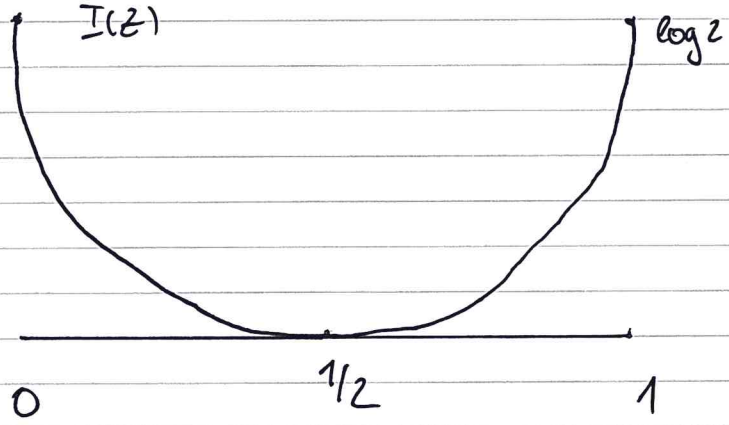
Example: coin-tossing

Let (X_i) be i.i.d. with $\mathbb{P}(X_1=0) = \mathbb{P}(X_1=1) = 1/2$, $S_n = \sum_{i=1}^{n} X_i$. Then, for all $a > 1/2$,

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

where

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



Proof (sketch):

- $a > 1$ ✓
- $a \in (1/2, 1]$: $\mathbb{P}(S_n \geq an) = 2^{-n} \sum_{k \geq an} \binom{n}{k}$

$$2^{-n} \underbrace{Q_n(a)}_{= \max_{k \geq an} \binom{n}{k}} \leq \mathbb{P}(S_n \geq an) \leq (n+1)\, 2^{-n}\, Q_n(a)$$

- $I(1-z) = I(z)$ □

Obervation:

$z \longmapsto I(z)$ is infinite outside $[0,1]$

finite and strictly convex inside $[0,1]$

unique zero at $z = 1/2$

This zero corresponds to the SLLN as it implies that

$$\sum_{n \in \mathbb{N}} \mathbb{P}(|1/n\, S_n - 1/2| > \delta) < \infty \qquad \forall \ \delta > 0.$$

Recall that an application of Chebyshev's inequality gives an estimate $\mathbb{P}(|1/n\, S_n - 1/2| > \delta) \leq \frac{1}{\delta^2 n} \to 0$ as $n \to \infty$ but this estimate is of order $1/n$ and therefore not summable.

It is therefore desirable to find out exactly how fast the large deviation probabilities

$$\mathbb{P}(|1/n\, S_n - 1/2| > \delta) \quad \text{decay.}$$

This depends on finer features of the random variable X than merely the finiteness of its variance.

We begin our program by introducing the logarithmic moment generating function:

Let $\mu \in M_1(\mathbb{R})$

$$\Lambda_\mu(\lambda) := \log\left(\int_{\mathbb{R}} \exp\{\lambda x\}\, \mu(dx)\right),$$

$\lambda \in \mathbb{R}$.

If μ is the law of a family $X_1, X_2, \ldots$ of real-valued i.i.d. random variables we write

$$\Lambda_\mu(\lambda) = \Lambda(\lambda) = \log \mathbb{E}\left[e^{\lambda X_1}\right], \lambda \in \mathbb{R}.$$

Note that $\lambda \in \mathbb{R} \longmapsto \Lambda_\mu(\lambda) \in (-\infty, \infty]$ is a lower semicontinuous convex function.
Indeed, by truncation, it is easy to write Λ_μ as the non-decreasing limit of smooth functions, and the convexity of Λ_μ follows from Hölder's inequality.
Next, let Λ_μ^* be the Legendre transform of Λ_μ:

$$\Lambda_\mu^*(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda_\mu(\lambda)\}, x \in \mathbb{R}.$$

Note that, by its definition as the point-wise supremum of linear functions, Λ_μ^* is necessarily lower semi-continuous and convex.

In order to develop some feeling for the relationship between Λ_μ, Λ_μ^*, and μ, we present the following elementary lemma.
Define $\mathcal{D}_{\Lambda_\mu} := \{\lambda : \Lambda_\mu(\lambda) < \infty\}$ and $\mathcal{D}_{\Lambda_\mu^*} := \{x : \Lambda_\mu^*(x) < \infty\}$

Lemma 1: Let $\mu \in \mathcal{M}_1(\mathbb{R})$. Then $\Lambda_\mu^* \geq 0$.

Moreover:

(a) If $\int_{\mathbb{R}} |x| \mu(dx) < \infty$ and $p = (\mathbb{E}(X))$

$= \int_{\mathbb{R}} x \mu(dx)$, then $\Lambda_\mu^*(p) = 0$,

Λ_μ^* is non-decreasing on $[p, \infty)$ and non-increasing on $(-\infty, p]$.

In addition, for $q \geq p$,

$\Lambda_\mu^*(q) = \sup_{\lambda \geq 0} \{\lambda q - \Lambda_\mu(\lambda)\}$

and $\mu([q, \infty)) \leq \exp\{-\Lambda_\mu^*(q)\}$; and

for $q \leq p$, $\Lambda_\mu^*(q) = \sup_{\lambda \leq 0} \{\lambda q - \Lambda_\mu(\lambda)\}$ and

$\mu((-\infty, q]) \leq \exp\{-\Lambda_\mu^*(q)\}$

(b) If $\mathcal{D}_{\Lambda_\mu} = \{0\}$, then $\Lambda_\mu^* \equiv 0$

If $\Lambda_\mu(\lambda) < \infty$ for some $\lambda > 0 \Rightarrow p = \mathbb{E}[X_1] < \infty$ (possibly $p = -\infty$). Similarly, if $\Lambda_\mu(\lambda) < \infty$ for some $\lambda < 0$, then $p > -\infty$ (possibly $p = \infty$).

(c) If $\Lambda_\mu(\lambda) < \infty$ for λ's in a neighborhood of 0, then $\Lambda_\mu^*(x) \to \infty$ as $|x| \to \infty$.

If $\mathcal{D}_{\Lambda_\mu} = \mathbb{R}$, then $\Lambda_\mu \in C^\infty(\mathbb{R})$

and $\Lambda_\mu^*(x)/|x| \longrightarrow \infty$ as $|x| \longrightarrow \infty$

(d) $\Lambda_\mu(\cdot)$ is differentiable in $\mathcal{D}_\Lambda^\circ$ with

$$\Lambda_\mu'(\eta) = \frac{1}{M(\eta)} \mathbb{E}[X_1 e^{\eta X_1}]$$

with $M(\eta) = \mathbb{E}[X_1]$,

and $\Lambda_\mu'(\eta) = y \Rightarrow \Lambda^*(y) = \eta y - \Lambda_\mu(\eta)$.

Proof:

Since $\lambda x - \Lambda_\mu(\lambda) = 0$ for $\lambda = 0$ and $\forall x \in \mathbb{R}$,

$\Lambda_\mu^*(x) \geq 0$.

ad (a): To see that $\Lambda_\mu^*(p) = 0$, we use Jensen's inequality to obtain $\Lambda_\mu(\lambda) \geq \lambda p \quad \forall \lambda \in \mathbb{R}$.

This shows that $\lambda p - \Lambda_\mu(\lambda) \leq 0$ for all $\lambda \in \mathbb{R}$ and so $\Lambda_\mu^*(p) \leq 0$. Since it is non-negative and convex, this proves that $\Lambda_\mu^*(p) = 0$, Λ_μ^* is non-decreasing in $[p, \infty)$, and Λ_p^* is non-increasing on $(-\infty, p]$.

To complete (a), we first note that, if $q \geq p$ then

$$\Lambda_\mu^*(q) = \sup_{\lambda \geq 0} \{ \lambda q - \Lambda_\mu(\lambda) \} \quad \text{and if} \quad q \leq p$$

then $\Lambda_\mu^*(q) = \sup_{\lambda \leq 0} \{ \lambda q - \Lambda_\mu(\lambda) \}$

(If $p = \mathbb{E}[X_1] = -\infty$, then $\Lambda_\mu(\lambda) = \infty$ for λ negative,

$X \geq p$; $\lambda X - \Lambda_\mu(\lambda) \leq \lambda p - \Lambda_\mu(\lambda) \leq \Lambda_\mu^*(p) = 0$

and the assertion follows. Hence, since (Chebychev's inequality)
$\mu([q,\infty)) \leq \exp\{-(\lambda q - \Lambda_\mu(\lambda))\}$, $\lambda \geq 0$, $\mu([q,\infty)) \leq \exp\{-\Lambda_\mu^*(q)\}$.

(b) If $\mathcal{D}_{\Lambda_\mu} = \{0\}$ $\Rightarrow$ $\Lambda_\mu^*(x) = \Lambda_\mu(0) = 0$ $\forall x \in \mathbb{R}$

If $\Lambda_\mu(\lambda) < \infty$ for some $\lambda > 0$, then

$\int_0^\infty x \, \mu(dx) < \frac{M(\lambda)}{\lambda} < \infty$, implying

$p = \mathbb{E}[X_1] < \infty$ (possibly $p = -\infty$)

(c) If $\lambda > 0$ ($\lambda < 0$) and $\Lambda_\mu(\lambda) < \infty$,
then $\varliminf_{x \to \infty} \Lambda_\mu^*(x)/x \geq \lambda \left(\varliminf_{x \to -\infty} \Lambda_\mu^*(x)/x \leq \lambda \right)$

But, by Taylor's Theorem and the Lebesgue dominated Convergence Theorem, it is easy to check that $\lambda \in (-\delta, \delta) \longmapsto \Lambda_\mu(\lambda)$ is, in fact, real-analytic as long as $\Lambda_\mu(\pm\delta) < \infty$. $\square$

Theorem 1: Let $X_1, X_2, \dots$ be i.i.d. on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ with mean μ, satisfying

$$\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X_1}] < \infty \quad \forall \lambda \in \mathbb{R},$$

and S_n their partial sums. Then, for any $x > \mu$ we have

$$\lim \frac{1}{n} \log \mathbb{P}(1/n S_n \geq x) = -\Lambda^*(x).$$

Proof: Upper bound: We use Chebyshev again, but in an optimised form. For any nonnegative, increasing function ψ we get

$$\mathbb{P}(1/n S_n \geq x) \leq \mathbb{P}(\psi(S_n) \geq \psi(nx)) \leq \frac{1}{\psi(nx)} \mathbb{E}[\psi(S_n)].$$

We now choose $\psi(x) = e^{\lambda x}$ and optimize over $\lambda \geq 0$ later. This yields

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}(1/n S_n \geq x) \leq -\lambda x + \overline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{E}[\exp(\lambda S_n)]$$

$$= -\lambda x + \Lambda(\lambda)$$

$$\Rightarrow \overline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}(1/2 S_n \geq x) \leq -\sup_{\lambda \geq 0} \{\lambda x - \Lambda(\lambda)\}$$

Λ is a convex function with $\Lambda'(0) = \mu < x$.

Henceforth the expression in the curly bracket on the right is negative for $\lambda < 0$, and vanishes for $\lambda = 0$.

Hence the supremum may be taken over all $\lambda \in \mathbb{R}$, which completes the proof of the upper bound.

(λ neg. implies $\lambda x - \Lambda(\lambda) \leq \lambda\mu - \Lambda(\lambda) \leq \Lambda^*(\mu) = 0$).

Lower bound: change of measure or tilting argument.

Replace the law P of X_1 by the law Q

$$dQ(x) = e^{-\Lambda(\lambda) + \lambda x} \, dP(x)$$

Assumption (to be justified later): $\forall \varepsilon > 0 \;\; \exists \;\; \lambda > 0$

such that

$$Q\left(x + \varepsilon > \tfrac{1}{n} S_n \geq x\right) \to 1 \quad \text{as } n \to \infty,$$

where $\mathbb{Q} = Q^{\otimes n}$

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left(\tfrac{1}{n} S_n \geq x\right) \geq \liminf_{n\to\infty} \frac{1}{n} \log \underbrace{P\left(x + \varepsilon > \tfrac{1}{n} S_n \geq x\right)}_{E_Q\left[e^{n\Lambda(\lambda) - \lambda S_n} \, 1\{x + \varepsilon > \frac{1}{n} S_n \geq x\}\right]}$$

$$\geq \Lambda(\lambda) - \lambda(x + \varepsilon) + \liminf_{n\to\infty} \frac{1}{n} \log Q\left(x + \varepsilon > \tfrac{1}{n} S_n \geq x\right)$$

$$= \Lambda(\lambda) - \lambda(x + \varepsilon) \geq -\Lambda^*(x + \varepsilon) \quad \text{, then } \varepsilon \downarrow 0$$

to conclude with the lower bound.
We are left to prove the above assumption.
It suffices to show that $\lambda > 0$ can be chosen such that $\Lambda'(\lambda) = e^{-\Lambda(\lambda)} \mathbb{E}[X e^{\lambda X}]$
$= \mathbb{E}_Q[X]$.

That is, by WLLN, we shall show that $\mathbb{E}_Q[X] = x + \frac{\varepsilon}{2}$.

$\Lambda'(0) = \mathbb{E}[X] = \mu$ and $\Lambda'(\infty) = \operatorname{ess\,sup} X = M$

If $\mu < x < M$, by the intermediate value theorem, we can find $\forall \varepsilon > 0$, some $\lambda > 0$ with $\Lambda'(\lambda) = x + \varepsilon/2$ as required.
To complete the argument note that, in the case $M < \infty$, for $x > M$ both sides of the statement in the theorem are equal to $-\infty$, and if $x = M$ they are both equal to $\log \mathbb{P}(X = M)$. □
(Hint: $\lambda x - \log \mathbb{E}[e^{\lambda X_1}] \geq \lambda(x - \mathbb{E}[X_1])$).

In conjunction with Lemma 1 one can prove the following statements in the setting of Theorem 1.

$I''(\mu) = \frac{1}{\sigma^2}$, where $I(z) = \Lambda^*(z)$.

Since Λ is smooth, we have

$I(z) = z\tau(z) - \Lambda(\tau(z))$ with $\tau(z)$ the unique solution of $z = (\log M)'(\tau(z))$

recall that $M(\lambda) = \mathbb{E}[e^{\lambda X_1}]$.

$$I'(z) = \tau(z) \; ; \; I''(z) = \tau'(z) = \frac{1}{[\log M]''(\tau(z))}$$

Since $\Lambda(0) = 1$, $\Lambda'(0) = \mu$ and $\Lambda''(0) = \mu^2 + \sigma^2$ it follows with $\tau(\mu) = 0$ that $I(\mu) = 0$, $I'(\mu) = 0$, $I''(\mu) = 1/\sigma^2$.

From the properties of Λ^* we get, if $a > \mu$,

$I(z) \geq I(a)$ for all $z \geq a$

Hence, Cramér's theorem can be written as

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(\tfrac{1}{n} S_n \in A) = - \inf_{z \in A} I(z)$$

with $A = [a, \infty)$

Hence the large deviation $\{\frac{1}{n} S_n \in A\}$ is essentially carried by the event where $\frac{1}{n} S_n$ is close to the minimiser $\bar{z}$ of $I(z)$ on A (which happens to be $\bar{z} = a$ if $a > \mu$).
The event cost $\exp(-n I(\bar{z}) + o(n))$ and is therefore the cheapest realisation of A

> Any large deviation is done in the least unlikely of all the unlikely ways!

We are going to generalise the above theorem to a statement about the limiting frequencies at which the random variables $X_1, X_2, \ldots$ take their values in some finite set E.

Let E be finite set and $X_1, X_2, \ldots$ i.i.d. random variables having values in E with law $\mu \in M_1(E)$ with the additional assumption that $\mu(s) > 0 \quad \forall s \in E$. (*)

The empirical measure of the sequence (X_i) is defined as

$$L_n = \frac{1}{n} \sum_{i=1}^{n} X_i ,$$

with δ_x denoting the point-mass at $x \in \mathbb{R}$.
Note that L_n is a random probability measure on E.

$$\mathcal{M}_1(E) = \left\{ \nu = (\nu_1, \dots, \nu_{|E|}) \in [0,1]^{|E|} : \sum_{s=1}^{|E|} \nu_s = 1 \right\}$$

is the probability simplex in $\mathbb{R}^{|E|}$, which may be identified with the set of prob. measures on E.
On $\mathcal{M}_1(E)$ we define the total variation distance

$$d(\mu, \nu) = \frac{1}{2} \sum_{s \in E} |\mu_s - \nu_s|$$

which turns $\mathcal{M}_1(E)$ into a Polish space.

$$\text{SLLN} \Rightarrow \quad d(L_n, \mu) \xrightarrow[\text{as } n \to \infty]{} 0 \quad \mathbb{P}\text{-a.s.}$$

The following theorem is a statement about the large deviations of L_n away from μ.

<u>Theorem 2:</u> Let (X_i) be i.i.d. rvs satisfying (*) above. Then, for all $a > 0$,

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(L_n \in B_a^c(\mu)) = - \inf_{\nu \in B_a^c(\mu)} I_\mu(\nu),$$

where $B_a(\mu) = \{\nu \in \mathcal{M}_1(E) : d(\nu,\mu) \le a\}$

$B_a^c(\mu) = \mathcal{M}_1(E) \setminus B_a(\mu)$ and

$$I_\mu(\nu) = \sum_{s\in E} \nu_s \log \frac{\nu_s}{\mu_s}$$

(convention $\inf \phi\, I_\mu = \infty$ and $0 \log 0 = 0$).

Proof: We only give a very brief sketch. Consult the book by Dembo/Zeitouni for further details.

$$K_n = \{k = (k_1, \dots, k_{|E|}) \in \mathbb{N}_0^{|E|} : \sum_{\substack{s\in E \\ s=1,\dots,|E|}} k_s = n\}$$

(we may give labels/numbers to the elements in E, i.e. $E \hat{=} \{1, \dots, |E|\}$)

$\frac{1}{n} K_n \subset \mathcal{M}_1(E)$ $\forall n \in \mathbb{N}$. L_n has multinomial distribution

$$\mathbb{P}\left(L_n(s) = \frac{k_s}{n} \ \forall s\right) = n! \prod_{s=1}^{|E|} \frac{\mu_s^{k_s}}{k_s!}, \quad k \in K_n$$

For $k \in K_n$, let $\nu_n(k) = \frac{1}{n} k \in M_1(E)$,

and put $Q_n(a) = \max_{k \in K_n : \nu_n(k) \in B_a^c(\mu)} \left(n! \prod_{s=1}^{|E|} \frac{\mu_s^{k_s}}{k_s!} \right)$.

Then $Q_n(a) \le \mathbb{P}(L_n \in B_a^c(\mu)) \le |K_n| \, Q_n(a)$

Stirling's formula gives

$$\frac{1}{n} \log \left(n! \prod_{s=1}^{|E|} \frac{\mu_s^{k_s}}{k_s!} \right) = \sum_{s=1}^{|E|} \frac{k_s}{n} \left(\log \mu_s - \log \frac{k_s}{n} \right)$$

$+ O\left(\frac{\log n}{n}\right)$ uniformly on K_n (note that $\sum_{s \in E} k_s = n$)

The sum equals $- I_\mu(\nu_n(k))$ and

using $|K_n| = \binom{n+|E|}{|E|-1} = O(n^{|E|-1})$, we find that

$$\frac{1}{n} \log \mathbb{P}(L_n \in B_a^c(\mu)) = O\left(\frac{\log n}{n}\right) + \frac{1}{n} \log Q_n(a)$$

$$= O\left(\frac{\log n}{n}\right) - \min_{k \in K_n : \nu_n(k) \in B_a^c(\mu)} \{ I_\mu(\nu_n(k)) \}$$

To complete the proof it remains to show that

- $\bigcup_{n \in \mathbb{N}} \{ \nu_n(k) : k \in K_n \}$ is dense in K_1

and

- $\nu \longmapsto I_\mu(\nu)$ is continuous on $\mathcal{M}_1(E)$

$\forall\, \nu \in \mathcal{M}_1(E)$ $\exists$ a sequence $(k_n)_{n\in\mathbb{N}}$ s.t.

$$\left.\begin{aligned} d(\nu_n(k_n), \nu) &\to 0 \\ I_\mu(\nu_n(k_n)) &\to I_\mu(\nu) \end{aligned}\right\} \text{ as } n \to \infty$$

As $B_a^c(\mu)$ is open, this implies

$$\limsup_{n\to\infty} \min_{k\in k_n : \nu_n(k) \in B_a^c(\mu)} I_\mu(\nu_n(k)) \leq I_\mu(\nu)$$

$\forall\, \nu \in B_a^c(\mu)$. Optimizing over ν, we get

$$\limsup_{n\to\infty} \min_{k\in k_n : \nu_n(k) \in B_a^c(\mu)} I_\mu(\nu_n(k)) \leq \inf_{\nu \in B_a^c(\mu)} I_\mu(\nu)$$

□

The function $I_\mu : \mathcal{M}_1(E) \to [0, \infty]$ is nothing else than the relative entropy, i.e.,

$I_\mu(\nu) = H(\nu|\mu)$.

<u>Definition:</u> Let $\mu, \nu \in \mathcal{M}_1(E)$, E finite

(a) The <u>entropy</u> of μ is defined as

$$H(\mu) = -\sum_{x\in E} \mu(x) \log \mu(x). \quad \text{(note the convention } 0\log 0 \hat{=} 0;\ 0\log\tfrac{0}{0} \hat{=} 0\text{)}$$

(b) The relative entropy of ν with respect to μ is defined as

$$H(\nu|\mu) = \sum_{x \in E} \nu(x) \log \frac{\nu(x)}{\mu(x)} .$$

Remark: (1) $H(\cdot|\mu) \geq 0$, and finite and continuous on $\{\nu \in M_1(E) : E_\nu \subset E_\mu\}$, where $E_\nu = \{x \in E : \nu(x) > 0\}$.
$H(\cdot|\mu) = \infty$ outside of that set.

(2) Relative entropy is a key notion in ergodic theory, information theory and statistical physics and stochastic processes.

(3) The choice of the distance in the Theorem is obviously flexible. All that we really need is the following lemma
(i.e. we may replace $B_a^c(\mu)$ by an arbitrary open subset of $M_1(E)$)

Lemma 2: (i) I_μ is finite, continuous and strictly convex on $M_1(E)$
(ii) $I_\mu(\nu) \geq 0$ with equality $\Leftrightarrow \nu = \mu$.