

Sanov's theorem

Restriction and conditioning

Consider a sequence (\mathbb{P}_n) satisfying the LDP with rate function I and an open set $G \subset K$ such that $\inf_G I = \inf_{\mathbb{E}} I$.

Proposition 1. (Restriction) For every $f \in C(K)$

$$\lim_n \int_G f d\mathbb{P}_n = \lim_n \int_{\mathbb{E}} f d\mathbb{P}_n = \sup_G \left(\int f / e^{I} \right) = \max_{\mathbb{E}} \left(\int f / e^{I} \right).$$

Proof. We can restrict ourselves to $f \geq 0$ and by density assume that $f > 0$. We thus write $f = e^{h}$ with $h \in C(K)$ and define new probability measures by $\mathbb{P}_n = c_n e^{h} \mathbb{P}_n$ for suitable constants c_n . The sequence (\mathbb{P}_n) satisfy the LDP with rate function $J = I + h$ $\min_K (I + h)$, since h is continuous $\inf_G J = \inf_{\mathbb{E}} J$ so

$$\int_G f d\mathbb{P}_n = \int f d\mathbb{P}_n (\mathbb{P}_n(G))^{1/n} \exp(\inf_G J - \min_K (I + h)) = \exp(\inf_G (I + h)).$$

If $I(x) < +\infty$ for some $x \in \mathbb{E}$ then $\mathbb{P}_n(\mathbb{E}) > 0$ for n large enough and we can introduce conditional measures \mathbb{P}_n such that $\mathbb{P}_n(f) = \mathbb{P}_n(f 1_{\mathbb{E}}) / \mathbb{P}_n(\mathbb{E})$ for all f bounded Borel on K . The set \mathbb{E} is another compact metrizable space.

Corollary 2. Assume $\min_{\mathbb{E}} I < +\infty$. The sequence of conditional measures (\mathbb{P}_n) obey the LDP with rate function $J: \mathbb{E} \rightarrow [0, +\infty]$ given by $J(x) = I(x) - \min_{\mathbb{E}} I$ for all $x \in \mathbb{E}$.

Proof. Take $f \in C(\mathbb{E})$ and let $\hat{f} \in C(K)$ any continuous extension of f (which exists for example due to separability of $C(K)$, think about it). Then $\hat{f} 1_{\mathbb{E}} d\mathbb{P}_n = \max_{\mathbb{E}} (\hat{f} / e^{I})$ so $\int f d\mathbb{P}_n = \hat{f} 1_{\mathbb{E}} d\mathbb{P}_n / 1_{\mathbb{E}} d\mathbb{P}_n = \max_{\mathbb{E}} (\hat{f} / e^{I}) / \max_{\mathbb{E}} (e^{I})$.

Tensorization and projections

Theorem 3. Consider two compact Polish spaces K_1 and K_2 . Let (\mathbb{P}_n^1) and (\mathbb{P}_n^2) be sequences resp in (K_1) and (K_2) which obey the LDP with rate functions I_1 and I_2 . Then the sequence $(\mathbb{P}_n = \mathbb{P}_n^1 \otimes \mathbb{P}_n^2)$ in $(K_1 \times K_2)$ obey the LDP with rate function $I(x_1, x_2) = I_1(x_1) + I_2(x_2)$.

Proof. Take $f(x_1, x_2) = (f_1 - f_2)(x_1, x_2) f_1(x_1) f_2(x_2)$. Then for any LD-converging sub-sequence of (\mathbb{P}_n) we have, for some rate function I ,

$$\lim_k \int f_1 - f_2 d\mathbb{P}_k = \sup_{x \in K_1 \times K_2} (f_1(x_1) f_2(x_2) e^{I(x_1, x_2)}).$$

On the other hand $\int f_1 - f_2 d\mathbb{P}_k = \int f_1 d\mathbb{P}_k^1 \int f_2 d\mathbb{P}_k^2$ and then

$$\sup_{x \in K_1 \times K_2} (f_1(x_1) f_2(x_2) e^{I(x_1, x_2)}) = \sup_{x_1 \in K_1} (f_1(x_1) e^{I_1(x_1)}) \sup_{x_2 \in K_2} (f_2(x_2) e^{I_2(x_2)})$$

For some fixed $z \in K_1 \cup \dots \cup K_2$ choose $f_i(x_i) = \exp(-N d_i(x_i, z))$ for $i = 1, 2$. Letting $N \rightarrow \infty$ we get

$$I(z_1, z_2) = I_1(z_1) + I_2(z_2) = I(z)$$

(prove it!) thus all possible accumulation points of $(\mu_n)_n$ have the same rate functions so the whole sequence satisfy the LDP with rate function I .

Exercise 1. Prove that $(\mu_n^1 \otimes \mu_n^2)_n$ is LD-convergent if and only if $(\mu_n^1)_n$ and $(\mu_n^2)_n$ are LD-convergent.

Theorem 4. (Dawson-Gartner) Consider a sequence of measures $(\mu_n)_n$. Let $\{g_k\}_{k=1}^\infty \subset C(K)$ be a family of continuous functions which separates the points of K . Define $G_k: K \rightarrow \mathbb{R}^k$ as $G_k(x) = (g_1(x), \dots, g_k(x))$. Assume that for all $k \geq 1$ the laws $\mu_n^k = (G_k)_\# \mu_n$ of the vector (g_1, \dots, g_k) obey the LDP with rate function I_k on the compact set $G_k(K)$. Then $(\mu_n)_n$ satisfy the LDP with rate function

$$I(x) = \sup_k I_k(G_k(x)).$$

Proof. By the Stone-Weierstrass theorem the functions of the form $f = g(G_k)$ are dense in $C(K_0)$ and the limit $\lim_n \int f d\mu_n = \int f d\mu$ exists. Convergence of $\int f d\mu_n$ for a dense set of f imply LD-convergence. Let us call I the rate function, then

$$I(x) = \log \sup \{f(x) : \lim_n \int f d\mu_n = 1\} = \log \sup \{g(G_k(x)) : \lim_n \int g(G_k) d\mu_n = 1\} = I_k(G_k(x))$$

so $I(x) = \sup_k I_k(G_k(x)) = I(x)$. Now for every $f \in C(K)$ such that $\lim_n \int f d\mu_n = 1$ choose k and g such that $f \approx g \circ G_k$. Then

$$f(x) = g(G_k(x)) + \epsilon \quad \lim_n \int g(G_k) d\mu_n = e^{I_k(G_k(x))} + \epsilon \quad (1 + \epsilon) e^{I(x)} + \epsilon$$

so $I(x) = \log[(1 + \epsilon) e^{I(x)} + \epsilon]$ for arbitrary $\epsilon > 0$. Then $I = I$.

Large deviations for coin tossing and Boltzmann discovery

Let $(X_n)_{n=1}^\infty$ be an iid sequence with law Bernoulli(p) for some $p \in [0, 1]$. Consider the r.v. $N_n = \sum_{k=1}^n X_k = \#\{X_k = 1 : 1 \leq k \leq n\}$ which counts the number of ones in the sequence. Of course $N_n \sim B(n, p)$ and if μ_n is the law of N_n/n we have

$$\mu_n = \sum_{k=0}^n \frac{1}{n!} \left(\frac{f(k/n)}{p^k (1-p)^{n-k}} \right)^{1/n}$$

Recall that given μ, ν ($\{0, \dots, N\}$) the relative entropy of μ wrt ν is given by

$$H(\mu/\nu) = \sum_{i=0}^N \mu(i) \log \frac{\mu(i)}{\nu(i)}.$$

Exercise 2. Prove that $(\mu_n)_n$ satisfy the LDP with rate function

$$I(x) = x \log(x/p) + (1-x) \log((1-x)/(1-p)) = H(\text{Ber}(x)/\text{Ber}(p)).$$

Hints:

- a) Prove that $f_n = \max_{0 \leq k \leq n} (|f(k/n)|)^{1/n} p^{k/n} (1-p)^{1-k/n}$ using the fact that the cardinality of the summation in the definition of f_n is of order n .
- b) Prove that, uniformly in $0 \leq k \leq n$,

$$\frac{k}{n}^{1/n} \leq \frac{k}{n}^{k/n} \leq 1 \leq \frac{k}{n}^{(1-k)/n}$$

by observing that the bound

$$\int_{1/k}^1 \log x dx \leq \frac{1}{k} \sum_{m=1}^k \log(m/k) \leq \int_0^1 \log x dx$$

imply $(k!)^{1/k} \sim (k/e)$ as $k \rightarrow \infty$ and then conclude that $(k!)^{1/n} \sim (k/e)^{k/n}$ uniformly in k by using different arguments for small and large k .

For sequences $(X_n)_{n \geq 1}$ of iid variables on the finite set $K = \{1, \dots, N\}$ with common law (μ) we can define the empirical vector L_n with values in the compact metrizable space $\mathcal{K} = \{p \in [0, 1]^N : p_1 + \dots + p_N = 1\}$ as

$$L_n(i) = \frac{1}{n} \sum_{k=1}^n 1_{X_k=i} = \frac{\#\{1 \leq k \leq n : X_k=i\}}{n}$$

and let μ_n to be the law on \mathcal{K} (thus $\mu_n = (\mu_n(K))$).

Theorem 5. (Boltzmann, 1877) The sequence $(\mu_n)_n$ satisfy the LDP on \mathcal{K} with (convex) rate function $I(\cdot) = H(\cdot | \mu)$.

The key point of a direct proof of this theorem is that the set of all possible empirical vectors of a sample of size n is of cardinality not larger than $(n+1)^N$ (each of the N components can take at most $n+1$ values). This magnitude disappear in the LD limit since it is sub-exponential in n . Only the asymptotic size of the set of the microscopic configurations compatible with a given empirical vector will contribute to the rate function, as in the coin tossing ($N=2$) case.

Another possible proof of this theorem goes via Cramér's theorem on \mathbb{R}^N . Replace each X_n by the vector of Bernoulli variables (Y_n^1, \dots, Y_n^N) : $\{0, 1\}^N$ where $Y_n^i = 1_{X_n=i}$ and observe that $L_n(i) = n^{-1} \sum_{j=1}^n Y_n^i$ so that empirical measure becomes an empirical mean. Then Cramér's theorem gives that the rate function on \mathcal{K} is given by the Fenchel-Legendre transform $I : \mathcal{K} \rightarrow \mathbb{R}$ of the log mgf of the vector Y_1 , but

$$I(x_1, \dots, x_N) = \log E(e^{x_1 Y_1^1 + \dots + x_N Y_1^N}) = \log \sum_{i=1}^N e^{x_i} \mu(i)$$

so, for every $x_1, \dots, x_N \in [0, 1]$ with $x_1 + \dots + x_N = 1$ we have

$$\begin{aligned} I(x_1, \dots, x_N) &= \sup_{\lambda_1, \dots, \lambda_N} [\lambda_1 x_1 + \dots + \lambda_N x_N - \log \sum_{i=1}^N e^{x_i} \mu(i)] \\ &= H((x_i)_{i=1}^N | (\mu(i))_{i=1}^N) = \sum_{i=1}^N x_i \log \frac{x_i}{\mu(i)}. \end{aligned}$$

Remark 6. Boltzmann discovered this asymptotic probability in 1877 during his attempt to ground thermodynamics of the perfect gas on a microscopic statistical theory of a system of free particles. His proof, using Stirling asymptotic formula for $n!$, was not completely rigorous but nonetheless right. His work showed that the physical entropy of thermodynamics is linked with the mathematical entropy of a probability distribution as a measure of its unevenness.

Now we are ready to generalize Sanov theorem to compactly supported measures on \mathbb{R}^d , not necessarily discrete. Let K a compact subset of \mathbb{R}^d and $(X_n)_{n \geq 1}$ an iid sequence with values in K with law $\mu = (X_1) \sim P$ on (K) and \mathbb{P}_μ on (K) the law of the empirical measure L_n defined via $L_n(f) = \frac{1}{n} \sum_{i=1}^n f(X_i)$ for all bounded measurable $f: K \rightarrow \mathbb{R}$ (alternatively $L_n(dx) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(dx)$).

Theorem 7. (Sanov) The sequence $(\mathbb{P}_\mu)_n$ obey the LDP on (K) with rate function $I(\cdot) = H(\cdot / \mu)$.

Proof. Let $\{x_k\}_{k=1}^\infty$ be a countable dense set in $C(K)$ and let $F_k = (x_1, \dots, x_k)$ the associated filtration of the Borel σ -algebra $B(K)$. Let $F_k: (K) \rightarrow [0, 1]^k$ be the continuous function $F_k(\cdot) = (x_1(\cdot)/x_1, \dots, x_k(\cdot)/x_k)$. By Cramer's theorem the image law $\mathbb{P}_\mu^k = (F_k)_* \mathbb{P}_\mu$ satisfy the LDP on $[0, 1]^k$ with rate function $I_k(x_1, \dots, x_k)$ given by

$$I_k(x_1, \dots, x_k) = \sup_{\nu_1, \dots, \nu_k} \left[\nu_1 x_1 + \dots + \nu_k x_k - \log E[e^{i_1 x_1 + \dots + i_k x_k}] \right].$$

Then by the Dawson-Gartner's theorem

$$\begin{aligned} I(\cdot) &= \sup_k I_k(F_k(\cdot)) = \sup_k \sup_{\nu_1, \dots, \nu_k} \left[\nu_1 x_1 + \dots + \nu_k x_k - \log E[e^{i_1 x_1 + \dots + i_k x_k}] \right] \\ &= \sup \left[\langle \cdot, \mu \rangle - \log E[e^{\langle \cdot, X_1 \rangle}] \right] = H(\cdot / \mu). \end{aligned}$$

Sanov's theorem can be used to prove Cramer's theorem. Indeed the empirical mean \hat{m}_n of a vector (X_1, \dots, X_n) is a function of the empirical distribution L_n : $\hat{m}_n = \int x L_n(dx) = m(L_n)$. If the variables take values on a compact subset of \mathbb{R}^d then $m: (K) \rightarrow \mathbb{R}$ is a continuous function and its image is compact. By the contraction principle the laws \mathbb{P}_n of \hat{m}_n obey the LDP with rate function $J(x) = \inf \{H(\cdot / \mu): (K), \int x d\mu = x\}$. Introduce the probability measures $\nu = e^{-x \log \mu(\cdot)}$ and observe that for any $y \in \text{supp } \mu$ we can find ν such that $y = m(\nu)$ and that

$$H(\nu / \mu) = \log \frac{d\nu}{d\mu} = H(\nu / \mu) + \log \mu(\nu)$$

so $J(y) = \inf \{H(\nu / \mu): (K), \int x d\nu = y\} + \log \mu(\nu) = \sup \left[\langle y, \mu \rangle - \log \mu(\nu) \right]$ (think why). Conclusion: J is the Fenchel-Legendre transform of $\log \mu(\cdot)$.

Gibbsian conditioning

Sanov's theorem allow us to discuss another physical phenomenon related to Gibbsian distributions. Let $(X_n)_{n \geq 1}$ be an iid sequence with values in K and law μ on (K) . Fix some integer $k \geq 1$ and consider the law \mathbb{P}_μ^k on (K^k) of (X_1, \dots, X_k) conditional of an event involving $L_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$, the empirical measure of the vector (X_1, \dots, X_n) :

$$\mathbb{P}_\mu^k(f) = \int_{K^k} f(x) \mathbb{P}_\mu^k(dx) = E[f(X_1, \dots, X_k) / L_n \in B]$$

where $A \subset B(K^k)$ and $B \subset B(K)$. We will work with $k = 1$ generalization to higher k being easy.

Exercise 3. Let $k = 1$, show that

$$\mathbb{P}_{q_n}(f) = \mathbb{E}[f(X_1)/L_n \mid B] = \mathbb{E}[L_n(f)/L_n \mid B] = \int_{(K)} (f) \mu_n(d)$$

where μ_n is the law of L_n conditional to the event B .

Assume that B is closed and that $\inf_{B^c} H = \min_B H = H(\hat{\cdot})$ for a unique minimum point $\hat{\cdot} \in B$. Then the LDP for $(L_n)_{n \geq 1}$ and the restriction theorem imply that the sequence $(\mu_n)_{n \geq 1}$ obey the LDP on $B \subset (K)$ with rate function $J(\cdot) = H(\cdot) - H(\hat{\cdot})$ and that $\mu_n \rightarrow \hat{\mu}$ as $n \rightarrow \infty$. Then, for every continuous $f \in C(K)$, $\mathbb{P}_{q_n}(f) = \int_{(K)} (f) \mu_n(d) \rightarrow \int_{(K)} (f) \hat{\mu}(d)$ and we have proved that the sequence $(\mathbb{P}_{q_n})_{n \geq 1}$ converge weakly to the probability measure $\hat{\mu}$ which is the solution of the minimization of H over B .

Interesting case is where the conditioning set B is of the form $B = \{ \mu \in (K) : \int [e, e + \epsilon] \}$ for some small $\epsilon > 0$ and $e \in \mathbb{R}$ is such that $\mathbb{E}[f(X_1)] < e < \sup_K f$. This last condition is to render the event $\{L_n \in B\} = \{L_n(\cdot) \in [e, e + \epsilon]\}$ atypical : by the LLN we have $L_n(\cdot) \rightarrow \mathbb{E}[f(X_1)]$ a.s. In this case $\hat{\mu}$ can be described explicitly as an exponential perturbation of μ . Let μ and introduce the tilted measures $\mu^\epsilon = e^{\epsilon f} / Z(\epsilon)$ with $Z(\epsilon) = \int e^{\epsilon f} d\mu$ and observe that $H(\mu^\epsilon) = H(\mu) + \epsilon \int f d\mu - \log Z(\epsilon)$.

Exercise 4. Show that there exists $\epsilon > 0$ for which $\mu^\epsilon(f) = e$ and that $H(\mu^\epsilon) = \min_B H$.

Now

$$H(\mu^\epsilon) = e + \log Z(\epsilon) = \min_{(f) \in [e, e + \epsilon]} [H(\mu) + \epsilon(f) + \log Z(\epsilon)] = \min_B H$$

so $\hat{\mu} = \mu^\epsilon$.

Let us extend the result to $k > 1$. It is enough to consider new block variables $\tilde{X}_i = (X_{1+k(i-1)}, \dots, X_{k+k(i-1)})$ and observe if $\tilde{L}_n \in (K^k)$ is the empirical law of $(\tilde{X}_1, \dots, \tilde{X}_n)$ then $L_{nk}(f) = \tilde{L}_n(\tilde{f})$ where $\tilde{f}(x_1, \dots, x_k) = (f(x_1) + \dots + f(x_k))/k$.

Exercise 5. Let $\mu \in (K^2)$, show that $H(\mu) = H(\mu_1, \mu_2) = H(\mu_1) + H(\mu_2)$ where μ_1, μ_2 are the marginals of μ . Hint: in the variational formula for $H(\mu)$ take test functions ϕ of the form $\phi_1 \phi_2$.

Exercise 6. For B of the form $B = \{L_n(f) \leq I\}$ derive the LDP for $(\tilde{L}_i)_{i \geq 1}$ on (K^k) and obtain that the law of Y_1 conditional on B is given by the minimum $\tilde{\mu} \in (K^k)$ of the functional $H_{\tilde{\mu}}$ over $\{ \mu \in (K^k) : \int \tilde{f} d\mu \leq I \}$ where $\tilde{\mu} = \mu^{\otimes k}$ is the k -fold product measure with marginals μ . Conclude that if $\hat{\mu}$ is the unique minimum of H over $\{ \mu \in (K) : \int f d\mu \leq I \}$ then $\hat{\mu}^{\otimes k}$ is the unique minimum of the variational problem on (K^k) and observe that this imply the independence of (X_1, \dots, X_k) in the limit law.

The physical interpretation of this phenomenon goes as follows: consider an assembly of n independent particles each of them characterized by some quantity X_i , $i = 1, \dots, n$ taking values in K (e.g. energy, momentum, position, etc...) and assume that the allowed configurations of the whole system are those compatible with a given mean value of some function $f: K \rightarrow \mathbb{R}$: $\int f(X_i)/n \approx e$ (e.g. energy per particle, density, etc..). This constraint is macroscopic in the sense that involves only an average over all the particles. Then in the limit of a infinite system ($n \rightarrow \infty$, in reality $n \approx 10^{23}$) the configurations of a very small subsystem of size k (in our model k is fixed as $n \rightarrow \infty$) are described by iid configurations, each particle distributes as $\hat{\mu}$, the Gibbs distribution compatible with the macroscopic constraint.

Large deviations for processes

Let $(X_n)_{n \geq 1}$ be an iid sequence of Bernoulli(p) r.v. For every n the vectors $X_n = (X_1, \dots, X_n)$ are random elements in $\{0, 1\}^n$ which we will embed in $L^1([0, 1])$ as follows : for each n let

$$F_n(x_1, \dots, x_n)(\cdot) = \sum_{i=1}^n x_i 1_{[(i-1)/n, i/n)}(\cdot)$$

so that $F_n(X_n)$ is a random element in $K = \{f \in L^1([0, 1]) : f \in \{0, 1\} \text{ a.e.}\}$ and we denote by \mathcal{P}_K its law. On K we consider the weak-topology, i.e. the smallest topology which renders all the linear maps $f \mapsto \int_0^1 f(\cdot) g(\cdot) d\lambda$ continuous for every $g \in L^1([0, 1])$. With this topology K is compact and metrizable. A possible metric is obtained by taking a countable dense subset $\{g_k\}_{k \geq 1}$ of the unit ball of L^1 and letting

$$d(f, g) = \sum_{k=1}^{\infty} \frac{|\int_0^1 f(\cdot) g_k(\cdot) d\lambda - \int_0^1 g(\cdot) g_k(\cdot) d\lambda|}{2^k}$$

Another possible metric is given by $d(f, g) = \sup_{0 \leq t \leq 1} |\int_0^t (f(\cdot) - g(\cdot)) d\lambda|$. Let $J_p(x) = H(\text{Ber}(x)/\text{Ber}(p))$. Then we have the following result

Theorem 8. (Mogulskii) *The sequence $(\mathcal{P}_K)_n$ obey the LDP on K with rate function*

$$I(f) = \int_0^1 J_p(f(\cdot)) d\lambda.$$

Proof. We only need to uniquely identify the rate function I of possible accumulation points. For each k define $Q_{k,I} = ((I \wedge 1)/k, (I \vee 1)/k]$ and $G_k: K \rightarrow [0, 1]^k$ as $G_k(f) = (f_{k,1}, \dots, f_{k,k})$ where $f_{k,i} = \int_{Q_{k,I}} f(\cdot) d\lambda$ is the mean of f over $Q_{k,I}$ so that $f_k(f) = F_k(G_k(f))$ f in K (why?). By Cramér's theorem the laws \mathcal{P}_K^k of $G_k(F_k(X_{kn}))$ on $[0, 1]^k$ satisfy the LDP with speed n/k and rate function $I_k(x_1, \dots, x_k) = \sum_{i=1}^k J_p(x_i)$ taking into account the change of speed we have that, for every $g \in K$ and for every k ,

$$\min \{I(f) : G_k(f) = G_k(g)\} = \frac{1}{k} \sum_{i=1}^k J_p(G_k(g)_i) = \int_0^1 J_p(g(\cdot)) d\lambda = I_k(g)$$

Now using Fatou lemma it is easy to compute the \liminf -limit of the functional $\int_0^1 J_p(g(\cdot)) d\lambda$ as

$$\lim_k \int_0^1 J_p(g(\cdot)) d\lambda = \int_0^1 J_p(g(\cdot)) d\lambda = I(g)$$

(exercise) while another easy argument gives $\lim_k \min \{I(f) : G_k(f) = G_k(g)\} = I(g)$ since I is lsc. Then we can conclude that $I(g) = I(g)$.