Sergey Lototsky, USC

A summary of large deviations

Motivating example: when CLT is not a good approximation. Consider a particular location where, on average, there is one earthquake every four days. What is the probability to have more that 100 earthquakes in one year (365 days)? The Poisson process model implies that the number of earthquakes in one year is $\mathcal{P}(91.25)$, that is, the Poisson random variable with mean $91.25 = 365/4$. Then the answer, $\mathbb{P}(\mathcal{P}(91.25) > 100) \approx 0.17$, can be computed using a statistical package. Without such a package, one can use the normal approximation of the Poisson distribution, $\mathcal{P}(\theta) \approx \mathcal{N}(\theta, \theta),\ \theta \to \infty$, leading to the approximate answer

$$\mathbb{P}(\mathcal{P}(91.25) > 100) \approx \mathbb{P}\left(\mathcal{N}(0,1) > \frac{100 - 91.25}{\sqrt{91.25}}\right) \approx \mathbb{P}(\mathcal{N}(0,1) > 0.97) \approx 0.17.$$

Now let us change the numbers, and assume that, on average, there is one earthquake every two days. What is the probability to have more than 128 earthquakes in 132 days? The same statistical package produces the answer $\mathbb{P}(\mathcal{P}(66) > 128) \approx 4.6 \cdot 10^{-12}$, which is very small. An application of the normal approximation leads to $\mathbb{P}(\mathcal{P}(66) > 128) \approx \mathbb{P}(\mathcal{N}(0,1) > 7.6) \approx 1.5 \cdot 10^{-14}$ which is also very small, but is off by a factor of more than 300.

Deviations from the law of large numbers.

The setting. Let $X, X_1, X_2, \dots$ be iid with $\mathbb{E}X = 0,\ \mathbb{E}X^2 = 1$, and define $S_n = X_1 + \ldots + X_n$. If $x > 0$, then, by the LLN, $\lim_{n\to\infty} \mathbb{P}(S_n/n > x) = 0$, and, by CLT, $\lim_{n\to\infty} \mathbb{P}(S_n/\sqrt{n} > x) = 1 - \Phi(x)$, where Φ is the standard normal cdf. We now consider a sequence $\{x_n,\ n \geq 1\}$ of positive numbers and look at the asymptotic of the probability $\mathbb{P}(S_n/\sqrt{n} > x_n) = \mathbb{P}(S_n/n > x_n/\sqrt{n})$, $n \to \infty$, either by itself or by comparing it with $\mathbb{P}(\mathcal{N}(0,1) > x_n)$. We get the following regimes depending on the behavior of x_n as $n \to \infty$,

- `Normal deviations` (CLT) if $x_n = O(1),\ n \to \infty$;
- `Moderate deviation` if $x_n \to +\infty$ but $x_n/\sqrt{n} \to 0$;
- `Large deviations` if $x_n = O(\sqrt{n})$;
- `Super-large deviations` if $x_n/\sqrt{n} \to +\infty$.

Logarithmic asymptotic: for two sequences $a_n > 0, b_n > 0$ we write $a_n \asymp b_n$ if $\lim_{n\to\infty} \frac{\ln a_n}{\ln b_n} = 1$.

The result. Assume that the moment generating function $M_X(\lambda) = \mathbb{E}e^{\lambda X}$ of X is defined in some neighborhood of $\lambda = 0$. Then, for every measurable set $A \subset \mathbb{R}$ with $0 \notin A$ and every $\beta > 1/2$,

$$\mathbb{P}\left(\frac{S_n}{n^\beta} \in A\right) \asymp \exp\Big(-n^{\gamma(\beta)} \inf_{x\in A} I_\beta(x)\Big), \tag{1.1}$$

where

- If $1/2 < \beta < 1$ (moderate deviations), then $\gamma(\beta) = 2\beta - 1$ and $I_\beta(x) = x^2/2$;
- If $\beta = 1$ (large deviations), then $\gamma = 1$ and $I_\beta(x) = \sup_\lambda \Big(\lambda x - \ln M_X(\lambda)\Big)$ is the *Legendre transform* of M_X;
- If $\beta > 1$ (super-large deviations), then more detailed information about M_X is necessary to determine γ and I_β. [For example, if $|X| < 1$ and $x > 0$, then $\mathbb{P}(|S_n|/n^2 > x) = 0$ for all $n > 1/x$.]

We see that the moderate deviations are, in a sense, trivial (the right-hand side of (1.1) does not depend on X), whereas the super-large deviations depend on X too much; large deviations are *accurate enough to be useful and loose enough to be correct*[1].

General theory.

Definition. A family $\{P_\varepsilon, \varepsilon > 0\}$ of probability measures on a complete separable metric space $\mathbb{V}$ with Borel sigma-algebra $\mathcal{B}(\mathbb{V})$ obeys/satisifies the `large deviations principle` (LDP) with rate function $\mathbf{I} : \mathbb{V} \to [0, +\infty]$ if

- the function $\mathbf{I}$ is lower semi-continuous;
- for every closed set $C \subset \mathbb{V}$,

$$\limsup_{\varepsilon\to 0} \varepsilon \ln P_\varepsilon(C) \leq -\inf_{x\in C} \mathbf{I}(x); \tag{1.2}$$

- for every open set $G \subset \mathbb{V}$,

$$\liminf_{\varepsilon\to 0} \varepsilon \ln P_\varepsilon(G) \geq -\inf_{x\in G} \mathbf{I}(x). \tag{1.3}$$

[1] A. Dembo and O. Zeitouni, *Large Deviations: Techniques and Applications*, Springer.

Immediate extensions.

(1) Lower semi-continuity of the function $\mathbf{I}$ means that the set $\mathbf{I}_a = \{x \in \mathbb{V} : \mathbf{I}(x) \le a\}$ is closed in $\mathbb{V}$ for every $a \in \mathbb{R}$. If instead all sets $\mathbf{I}_a$ are *compact*, then the rate function $\mathbf{I}$ is called `good`.

(2) If $A \in \mathcal{B}(\mathbb{V})$ is a measurable set with interior A^{int} and closure A^{cl}, then (1.2) and (1.3) can be combined into

$$-\inf_{x\in A^{int}} \mathbf{I}(x) \le \liminf_{\varepsilon\to 0} \varepsilon \ln P_\varepsilon(A) \le \limsup_{\varepsilon\to 0} \varepsilon \ln P_\varepsilon(A) \le -\inf_{x\in A^{cl}} \mathbf{I}(x). \tag{1.4}$$

The set A is called `regular` if

$$P_\varepsilon(A) \asymp e^{-\varepsilon^{-1}\inf_{x\in A}\mathbf{I}(x)},$$

that is, $\lim_{\varepsilon\to 0}\varepsilon \ln P_\varepsilon(A) = -\inf_{x\in A}\mathbf{I}(x)$

(3) The family $\{P_\varepsilon, \varepsilon > 0\}$ is called `exponentially tight` if, for every $c > 0$, there exists a compact set $K_c \subset \mathbb{V}$ such that $\limsup_{\varepsilon\to 0}\varepsilon \ln P_\varepsilon(\mathbb{V}\setminus K_c) \le -c$.

(4) The family $\{P_\varepsilon, \varepsilon > 0\}$ satisfies `local/weak/vague` LDP if there is a lower semi-continuous function $\mathbf{I} : \mathbb{V} \to [0, +\infty]$ such that, for every $x \in \mathbb{V}$,

$$\lim_{\delta\to 0}\lim_{\varepsilon\to 0} \varepsilon \ln P_\varepsilon\big(B_\delta(x)\big) = -\mathbf{I}(x), \tag{1.5}$$

where $B_\delta(x)$ is the ball in $\mathbb{V}$ with center at x and radius δ.

(5) **Theorem.**
- (1.3) + (1.2) for *compact* sets C $\Rightarrow$ (1.5) $\Rightarrow$ uniqueness of $\mathbf{I}$.
- LDP + Expo. tightness $\Rightarrow$ Good rate function; LDP + Good rate function $\Rightarrow$ Expo. tightness
- (1.5) $\Rightarrow$ (1.3); (1.5) $\Rightarrow$ (1.2) for *compact* sets C.
- (1.5)+ Exponential tightness $\Rightarrow$ LDP (with good rate function).

(6) The original setting can be further extended to a regular Hausdorff topological space $\mathbb{V}$ with a sigma-algebra other than $\mathcal{B}(\mathbb{V})$, but then some of the above results (e.g. equivalence of (1.4) and (1.2)+(1.3)) might no longer hold. Still, such an extension can be useful, for example, if we want to do LDP on a non-separable space like $L_\infty((0,T))$.

Contraction principle. If $\{P_\varepsilon, \varepsilon > 0\}$ obeys LDP on $\mathbb{V}$ and $f : \mathbb{V} \to \mathbb{U}$ is a continuous mapping, then $\{P_\varepsilon \circ f^{-1}, \varepsilon > 0\}$ obeys LDP on $\mathbb{U}$.

Main examples.

(1) `Level 1 large deviations: Cramér's theorem.` If $X, X_1, X_2, \dots$ are iid with $\mathbb{E}X = 0$, $M_X(\lambda) = \mathbb{E}e^{\lambda X} < \infty$, $|\lambda| < \delta$, $\delta > 0$, and $P_\varepsilon(A) = \mathbb{P}(S_n/n \in A)$, $\varepsilon = 1/n$, then the family $\{P_\varepsilon,\ \varepsilon > 0\}$ satisfies LDP on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with the good rate function $\mathbf{I}(x) = \sup_\lambda \Big(\lambda x - \ln M_X(\lambda)\Big)$.

(2) `Level 2 large deviations: Sanov's theorem.` Let $X, X_1, X_2, \dots$ be iid with distribution $\mu_X(A) = \mathbb{P}(X \in A)$, and define the `empirical measure` $\bar\mu_n(A) = n^{-1}\sum_{k=1}^n 1(X_k \in A)$. Let $\mathbb{V}$ be the collection of probability measures on $\mathbb{R}$ equipped with the Lèvy-Prokhorov metric. With $\varepsilon = 1/n$ and $A \in \mathcal{B}(\mathbb{V})$, define $P_\varepsilon(A) = \mathbb{P}(\bar\mu_n \in A)$. Then the family $\{P_\varepsilon,\ \varepsilon > 0\}$ satisfies LDP on $(\mathbb{V}, \mathcal{B}(\mathbb{V}))$ with the good rate function $\mathbf{I}(\nu) = \int_{\mathbb{R}} \ln(d\nu/d\mu_X)\, d\nu$ if $\nu \ll \mu_X$ and $\mathbf{I}(\nu) = +\infty$ otherwise. Note that the rate function in this case is the *relative entropy/Kullback-Liebler divergence*, and an extension to random elements with values in a complete separable metric space is immediate.

(3) `Level 3 large deviations: Donsker-Varadhan theory.` The main object is the `empirical process` that encodes not only the usual empirical distribution, but also all possible *joint* empirical distributions (for pairs, triplets, etc.)

(4) `Functional/Pathwise LDP: Mogulskii's theorem.` Let $X, X_1, X_2, \dots$ be iid, $\mathbb{E}X = 0$, $\mathbb{E}X^2 = 1$, $M_X(\lambda) = \mathbb{E}e^{\lambda X} < \infty$, $\lambda \in \mathbb{R}$, and let $S_n(t)$, $t \in [0,1]$, be the continuous interpolation of $n^{-1}\sum_{k=1}^{[nt]} X_k$. With $\varepsilon = 1/n$, define P_ε as the distribution of S_n in the space of continuous functions on $[0,1]$. Then the family $\{P_\varepsilon,\ \varepsilon > 0\}$ satisfies LDP with a good rate function $\mathbf{J}(\varphi) = \int_0^1 \mathbf{I}(\dot\varphi(t))\, dt$ for absolutely continuous φ with $\varphi(0) = 0$; $\mathbf{I}$ is the rate function from Cramér' theorem.

(5) `Functional/Pathwise LDP: Wentzell-Freidlin theory.` If P_ε is the distribution, in the space of continuous functions on $[0,T]$, of the solution of $dX = b(X)dt + \sqrt{\varepsilon}\sigma(X)dW(t)$, $t \in [0,T]$, then, under some natural conditions on b and σ, we get LDP with a good rate function.

(6) `The Gaussian case.` If X is a $\mathbb{V}$-valued, zero-mean Gaussian random element with *Cameron-Martin space* H_X and $P_\varepsilon(A) = \mathbb{P}(\sqrt{\varepsilon}X \in A)$, then the LDP holds with the good rate function $\mathbf{I}(x) = \|x\|^2_{H_X}/2$, $x \in H_X$, $\mathbf{I}(x) = +\infty$, $x \notin H_X$. The lower bound (1.3) follows immediately from Borell's inequality for shifted measures.