Large deviations

Francis Comets
Université Paris Diderot, France
http://www.proba.jussieu.fr/∼comets/

Large deviations is concerned with the study of rare events and of small probabilities. Let $X_i, 1 \le i \le n$, be independent identically distributed (i.i.d.) real random variables with expectation m, and $\bar{X}_n = (X_1 + \dots X_n)/n$ their empirical mean. The law of large numbers shows that, for any Borel $A \subset \mathbb{R}$ not containing m in its closure, $P(\bar{X}_n \in A) \to 0$ as $n \to \infty$, but does not tell us how fast the probability vanishes. Large deviations state it is exponential in n, and give us the rate of decay. **Cramér's theorem** states that,

$$P(\bar{X}_n \in A) = \exp -n\big(\inf\{I(x); x \in A\} + o(1)\big)$$

as $n \to \infty$, for all interval A. The rate function I can be computed as the Legendre conjugate of the logarithmic moment generating function of X,

$$I(x) = \sup\{\lambda x - \ln E \exp(\lambda X_1); \lambda \in \mathbb{R}\},$$

and is called the Cramér transform of the common law of the X_i's. The natural assumption is the finiteness of the moment generating function in a neighborhood of the origin, i.e., the property of exponential tails. The function $I : \mathbb{R} \to [0, +\infty]$ is convex with $I(m) = 0$.
• In the Gaussian case $X_i \sim \mathcal{N}(m, \sigma^2)$, we find $I(x) = (x - m)^2/(2\sigma^2)$;
• In the Bernoulli case $P(X_i = 1) = p = 1 - P(X_i = 0)$, we find the entropy function $I(x) = x \ln(x/p) + (1-x) \ln(1-x)/(1-p)$ for $x \in [0, 1]$, and $I(x) = +\infty$ otherwise.

To emphasize the importance of rare events, let us mention a consequence, the **Erdös-Rényi law**: consider an infinite sequence $X_i, i \ge 1$, of Bernoulli i.i.d. variables with parameter p, and define R_n the length of the longest consecutive run, contained within the first n tosses, in which the fraction of 1's is at least a ($a > p$). Erdös and Rényi proved that, almost surely as $n \to \infty$,

$$R_n / \ln n \longrightarrow I(a)^{-1},$$

with the function I from the Bernoulli case above. Though it may look paradoxical, large deviations are at the core of this event of full probability. This result is the basis of bioinformatics applications like sequence matching, and of statistical tests for sequence randomness.

The theory does not only apply to independent variables, but allows for many variations, including weakly dependent variables in a general state space, Markov or Gaussian processes, large deviations from ergodic theorems, non-asymptotic bounds, asymptotic expansions (Edgeworth expansions), ...

Here is the **formal definition**. Given a Polish space (i.e., a separable complete metric space) $\mathcal{X}$, let $\{\mathbb{P}_n\}$ be a sequence of Borel probability measures on $\mathcal{X}$, let a_n be a positive sequence tending to infinity, and finally let $I : \mathcal{X} \to [0, +\infty]$ be a lower continuous functional on X which level sets$\{x : I(x) \leq a\}$ are compact for all $a < \infty$. We say that the sequence $\{\mathbb{P}_n\}$ satisfies a large deviation principle with speed a_n and rate I, if for each measurable set $E \subset X$

$$-\inf_{x\in E^\circ} I(x) \leq \underline{\lim}_n a_n^{-1} \ln \mathbb{P}_n(E) \leq \overline{\lim}_n a_n^{-1} \ln \mathbb{P}_n(E) \leq -\inf_{x\in \bar{E}} I(x)$$

where $\bar{E}$ and E° denote respectively the closure and interior of E. The rate function can be obtained as

$$I(x) = -\lim_{\delta\searrow 0} \lim_{n\to\infty} a_n^{-1} \ln \mathbb{P}_n(B(x, \delta)),$$

with $B(x, \delta)$ the ball of center x and radius δ. Large deviation theory allows for an abstract version of Laplace method for estimating integrals: Varadhan's lemma states that, for any continuous function $F : X \to \mathbb{R}$ with

$$\lim_{M\to\infty} \limsup_{n\to\infty} a_n^{-1} \ln \int_{F(x)\geq M} e^{a_n F(x)} dP_n(x) = -\infty$$

(a bounded F is fine), we have

$$\lim_{n\to\infty} a_n^{-1} \ln \int_X e^{a_n F(x)} dP_n(x) = \sup_x \{F(x) - I(x)\},$$

and the sequence of probability measures $e^{a_n F} dP_n / \int e^{a_n F} dP_n$ concentrates on the set of maximizers.

Sanov's theorem and sampling with replacement: let μ be a probability measure on a set Σ that we assume finite for simplicity, with $\mu(y) > 0$ for all $y \in \Sigma$. Let $Y_i, i \geq 1$, an i.i.d. sequence with law μ, and N_n the score vector of the n-sample,

$$N_n(y) = \sum_{i=1}^n \mathbf{1}_y(Y_i).$$

By the law of large numbers, $N_n/n \to \mu$ a.s. From the multinomial distribution, one can check that, for all ν such that $n\nu$ is a possible score vector for the n-sample,

$$(n+1)^{-|\Sigma|} e^{-nH(\nu|\mu)} \leq P(n^{-1}N_n = \nu) \leq e^{-nH(\nu|\mu)},$$

where $H(\nu|\mu) = \sum_{y\in\Sigma} \nu(y) \ln \frac{\nu(y)}{\mu(y)}$ is the relative entropy of ν with respect to μ. The large deviations theorem holds for the empirical distribution of a general n-sample, with speed n and rate $I(\nu) = H(\nu|\mu)$ given by the natural generalization of the above formula. This result, due to Sanov, has many consequences in information theory and statistical mechanics [2, 5], and for exponential families in statistics. Applications in statistics also include point estimation (by giving

the exponential rate of convergence of M-estimators) and for hypothesis testing (Bahadur efficiency) [7], and concentration inequalities [2].

Consider now a **Markov chain** $(Y_n, n \geq 0)$. For simplicity we assume that it is irreducible with a finite state space Σ. We denote by $Q = (Q(i,j); i,j \in \Sigma)$ the transition matrix, and for any $V : \Sigma \to \mathbb{R}$, $Q_V(i,j) = Q(i,j)e^{V(j)}$. By Perron-Frobenius theorem, $n^{-1} \ln \sum_j Q_V^n(i,j) \to \ln \lambda_V(Q)$ as $n \to \infty$, with $\lambda_V(Q)$ the principal eigenvalue of the positive matrix Q_V. By the ergodic theorem, N_n/n converges to the (unique) invariant law for Q. The law of the empirical distribution N_n/n satisfes a large deviation principle with speed n and rate I_Q given by

$$I_Q(\nu) = \sup_V \{\sum_j V(j)\nu(j) - \ln \lambda_V(Q)\}$$

for any law ν on Σ.

Consider next a **Markov process** $(Y_t, t \in \mathbb{R}^+)$. We assume it is irreducible on the finite state space Σ, and denote by $a(i,j)$ the transition rate from i to j ($a(i,j) \geq 0$ for $i \neq j$, $\sum_j a(i,j) = 0$). Then, similarly to the time-discrete case, the law of the empirical distribution $(t^{-1} \int_0^t \mathbf{1}_j(Y_s)ds; j \in \Sigma)$ satisfes a large deviation principle with speed t and rate I_a given by

$$I_a(\nu) = \sup_V \{\sum_j V(j)\nu(j) - \lambda_V(a)\},$$

with $\lambda_V(a)$ the principal eigenvalue of the matrix $(a(i,j) + \delta(i,j)V(j))_{i,j}$. Now, assume in addition that the process is reversible with respect to a probability measure π, i.e. $\pi(i)a(i,j) = \pi(j)a(j,i)$ for all i,j. Then, π is the invariant measure for the process and $\pi(i) > 0$ for all $i \in \Sigma$. Using the variational formula for eigenvalues of symmetric operators, Donsker and Varadhan found that the rate function takes a simple form in the reversible case,

$$I_a(\nu) = \frac{1}{2} \sum_{i,j} \pi(i)a(i,j) \left(\sqrt{\frac{\nu(i)}{\pi(i)}} - \sqrt{\frac{\nu(j)}{\pi(j)}} \right)^2 ,$$

that is the value of the Dirichlet form of the reversible process on the square root of the density of ν with respect to the invariant measure.

The **Freidlin-Wentzell theory** deals with diffusion processes with small noise,

$$dX_t^\epsilon = b(X_t^\epsilon)dt + \sqrt{\epsilon}\, \sigma(X_t^\epsilon)dB_t \,, \qquad X_0^\epsilon = y.$$

The coefficients b, σ are uniformly lipshitz functions, and B is a standard Brownian motion. The sequence X^ϵ can be viewed as $\epsilon \searrow 0$ as a small random perturbation of the ordinary differential equation (ODE)

$$dx_t = b(x_t)dt \,, \qquad x_0 = y.$$

Indeed, $X^\epsilon \to x$ in the supremum norm on bounded time-intervals. Freidlin and Wentzell have shown that, on a finite time interval $[0,T]$, the sequence X^ϵ with values in the path space obeys the LDP with speed ϵ^{-1} and rate function

$$I_{0,T}(\phi) = \frac{1}{2}\int_0^T \sigma(\phi(t))^{-2}\Big(\dot{\phi(t)} - b(\phi(t))\Big)^2 dt$$

if ϕ is absolutely continuous with square-integrable derivative and $\phi(0) = y$; $I(\phi) = \infty$ otherwise. (To fit in the above formal definition, take a sequence $\epsilon = \epsilon_n \searrow 0$, and for $\mathbb{P}_n$ the law of X^{ϵ_n}.) Note that $I(\phi) = 0$ if and only if ϕ is a solution of the above ODE. To follow closely any other path ϕ during a finite time T is an event of probability $\exp\{-\epsilon^{-1}I_{0,T}(\phi)\}$ at leading order, and therefore is very rare for small ϵ.

A simple case is $\sigma = 1$ and $b(x) = -V'(x)$ with a smooth V; we view $V(x)$ as the height of x, and a key role is played by the local minima of V. With an overwhelming probability as $\epsilon \searrow 0$, the picture will be as follows in a generic situation. The process X^ϵ will stay close to the solution of the ODE starting from y, and will eventually come near the local minimum (say z_0) which attracts y, and stay around for times of order $\exp o(\epsilon^{-1})$. But, by ergodicity, it will leave the neighborhood of z_0 at some time, and, even more, it will visit all points: it is then important how these large deviations occur. Let D be domain of attraction of z_0, h be its depth (i.e., the height difference between z_0 and the lowest point on the boundary of D, that we call the lowest pass). Up to times of order $\exp(\epsilon^{-1}h)$, the process remains in the part of D of relative height smaller than h, and will occupy this region with density proportional to $\exp -(\epsilon^{-1}V(\cdot))$. At some random time τ of order $\exp(\epsilon^{-1}h)$, it will leave D through the lowest path, and fall down towards a new local minimum, following roughly a path of the ODE. The piece of path just before leaving D is the time-reversed of an ODE path. The ratio of τ to its expected value converges to an exponential law.

The Freidlin-Wentzell theory has applications in physics (metastability phenomena, analysis of rare events) and engineering (tracking loops, statistical analysis of signals, stabilization of systems and algorithms) [6, 1, 2, 9].

Acknowledgement: This article is based on an article from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science+Business Media, LLC

References

[1] Azencott, R. *Grandes déviations et applications.* (French) Eighth Saint Flour Probability Summer School—1978, 1–176, Lecture Notes in Math. 774, Springer, Berlin, 1980

[2] Dembo, Amir; Zeitouni, Ofer: *Large deviations techniques and applications.* Springer, New York, 1998.

[3] Deuschel, J.-D., Stroock, D. *Large deviations.* Academic Press, Inc., Boston, MA, 1989

[4] Feng, J., Kurtz, T. *Large deviations for stochastic processes.* American Mathematical Society, Providence, RI, 2006

[5] den Hollander, Frank: *Large deviations.* American Mathematical Society, Providence, RI, 2000.

[6] Freidlin, M. I.; Wentzell, A. D.: *Random perturbations of dynamical systems.* Springer-Verlag, New York, 1998.

[7] Kester, A.: *Some large deviation results in statistics.* CWI Tract, 18. Centrum voor Wiskunde en Informatica, Amsterdam, 1985.

[8] Kipnis, C., Landim, C.: *Scaling limits of interacting particle systems.* Springer-Verlag, Berlin, 1999

[9] Olivieri, Enzo; Vares, Maria Eulália: *Large deviations and metastability.* Cambridge University Press, 2005.

[10] Varadhan, S. R. S.: Large deviations. *Ann. Probab.* **36** (2008), 397–419