Large Deviation Theory

Jan M. Swart (Czech Academy of Sciences)

Wednesday, March 3rd, 2021

The weak law of large numbers

Let $(X_k)_{k\geq 1}$ be i.i.d. real random variables.
Assume $\mathbb{E}[|X_1|] < \infty$ and set $\rho := \mathbb{E}[X_1]$.

The *weak law of large numbers* states that the *empirical averages*

$$T_n := \frac{1}{n}\sum_{k=1}^{n} X_k \qquad (n \geq 1).$$

satisfy

$$\mathbb{P}\big[|T_n - \rho| \geq \varepsilon\big] \underset{n\to\infty}{\longrightarrow} 0 \qquad (\varepsilon > 0).$$

Question How fast is this convergence?

Cramér's theorem

Theorem (Harald Cramér, 1938) Assume that

$$Z(\lambda) := \mathbb{E}[e^{\lambda X_1}] < \infty \qquad (\lambda \in \mathbb{R}).$$

Then

$$\text{(i)} \quad \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[T_n \geq y] = -I(y) \qquad (y > \rho),$$

$$\text{(ii)} \quad \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[T_n \leq y] = -I(y) \qquad (y < \rho),$$

where I is defined by

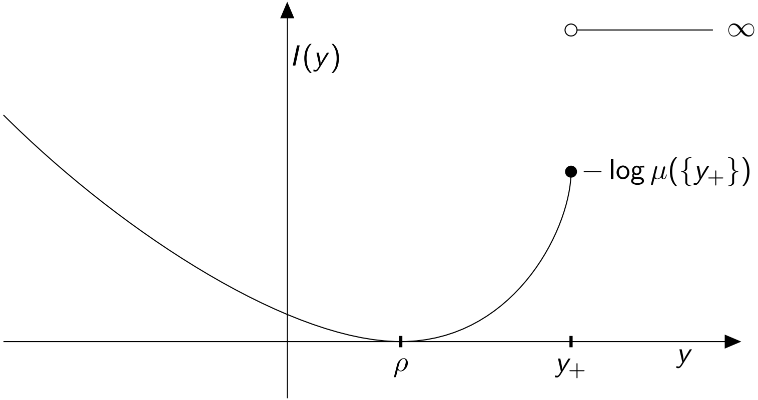
$$I(y) := \sup_{\lambda\in\mathbb{R}} \big[\lambda y - \log Z(\lambda)\big] \qquad (y \in \mathbb{R}).$$

Cramér's theorem

$Z(\lambda)$	*moment generating function,*
$\log Z(\lambda)$	*logarithmic moment generating function,*
$I(y)$	*rate function.*

$$
\begin{aligned}
\mu &:= \text{the law of } X_1,\\
\rho &:= \text{mean of } \mu,\\
\sigma &:= \text{variance of } \mu,\\
y_- &:= \inf(\mathrm{support}(\mu)),\\
y_+ &:= \sup(\mathrm{support}(\mu)),\\
\mathcal{D}_I &:= \{y \in \mathbb{R} : I(y) < \infty\},\\
\mathcal{U}_I &:= \mathrm{int}(\mathcal{D}_I).
\end{aligned}
$$

The rate function



The rate function

Assume $\sigma > 0$. Then:

(i) I is convex.

(ii) I is lower semi-continuous.

(iii) $0 \leq I(y) \leq \infty$ for all $y \in \mathbb{R}$.

(iv) $I(y) = 0$ if and only if $y = \rho$.

(v) $\mathcal{U}_I = (y_-, y_+)$.

(vi) I is infinitely differentiable on $\mathcal{U}_I$.

(vii) $\lim_{y \downarrow y_-} I'(y) = -\infty$ and $\lim_{y \uparrow y_+} I'(y) = \infty$.

(viii) $I'' > 0$ on $\mathcal{U}_I$ and $I''(\rho) = 1/\sigma^2$.

(ix) If $-\infty < y_-$, then $I(y_-) = -\log \mu(\{y_-\})$, and if $y_+ < \infty$, then $I(y_+) = -\log \mu(\{y_+\})$.

Cramér's theorem

Recall $T_n := \frac{1}{n}\sum_{k=1}^{n} X_k \qquad (n \geq 1)$.

$$\mathbb{P}[T_n \geq y] = e^{-nI(y) + o(n)} \qquad \text{as} \quad n \to \infty \qquad (y > \rho),$$

Also

$$\mathbb{P}[T_n \leq y_- \text{ or } T_n \geq y_+] = e^{-nI(y_-) + o(n)} + e^{-nI(y_+) + o(n)}$$

$$= e^{-n\big(I(y_-) \wedge I(y_+)\big) + o(n)} \qquad \text{as} \quad n \to \infty.$$

The slowest exponential decay wins.

Moderate deviations

Let $S_n := \sum_{k=1}^{n} X_k \qquad (n \geq 1)$. Then Cramér says:

$$\mathbb{P}\big[S_n - \rho n \geq yn\big] = e^{-nI(\rho + y) + o(n)} \quad \text{as } n \to \infty \qquad (y > 0).$$

The central limit theorem says

$$\mathbb{P}\big[S_n - \rho n \geq y\sqrt{n}\big] \underset{n\to\infty}{\longrightarrow} \Phi(y/\sigma) \qquad (y \in \mathbb{R}).$$

Moderate Deviations Theorem For $\frac{1}{2} < \alpha < 1$ and $y > 0$

$$\mathbb{P}[S_n - \rho n \geq yn^{\alpha}] = e^{-n^{2\alpha-1}\frac{1}{2\sigma^2}y^2 + o(n^{2\alpha-1})} \quad \text{as } n \to \infty.$$

Relative entropy

$N_n(x) :=$ number of times x thrown in n dice rolls $(x = 1, \ldots, 6)$.
$M_n(x) := N_n(x)/n$ relative frequency of x.
$\Delta_n := \max_{1 \le x \le 6} M_n(x) - \min_{1 \le x \le 6} M_n(x)$.

The strong law of large numbers gives $M_n(x) \underset{n\to\infty}{\longrightarrow} 1/6 \ \forall x$.

As a consequence $\Delta_n \underset{n\to\infty}{\longrightarrow} 0$.

Proposition There exists a continuous, strictly increasing function $I : [0,1] \to \mathbb{R}$ with $I(0) = 0$ and $I(1) = \log 6$, such that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[\Delta_n \ge \varepsilon\big] = -I(\varepsilon) \qquad (0 \le \varepsilon \le 1).$$

Relative entropy

Let $\mathcal{M}_1(S) :=$ the space of probability measures on a finite set S.

For $\mu, \nu \in \mathcal{M}_1(S)$, the *relative entropy* of ν with respect to μ is defined as

$$H(\nu|\mu) := \sum_{x \in S} \nu(x) \log \frac{\nu(x)}{\mu(x)} = \sum_{x \in S} \mu(x) \frac{\nu(x)}{\mu(x)} \log \frac{\nu(x)}{\mu(x)},$$

with $\log(0) := -\infty$ and $0 \cdot \infty := 0$.

$H(\nu|\mu)$ is also known as the *Kullback-Leibler distance* or *divergence*.

Relative entropy

Assume that $\mu(x) > 0$ for all $x \in S$. Then

(i) $0 \leq H(\nu|\mu) < \infty$ for all $\nu \in \mathcal{M}_1(S)$.

(ii) $H(\mu|\mu) = 0$.

(iii) $H(\nu|\mu) > 0$ for all $\nu \neq \mu$.

(iv) $\nu \mapsto H(\nu|\mu)$ is convex and continuous on $\mathcal{M}_1(S)$.

(v) $\nu \mapsto H(\nu|\mu)$ is infinitely differentiable on the interior of $\mathcal{M}_1(S)$.

Relative entropy

Boltzmann-Sanov Theorem Let C be a closed subset of $\mathcal{M}_1(S)$ such that C is the closure of its interior. Then

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[M_n \in C] = -\min_{\nu\in C} H(\nu|\mu).$$

Application For each $0 \leq \varepsilon < 1$, the set

$$C_\varepsilon := \big\{\nu \in \mathcal{M}_1(S) : \max_{x\in S} \nu(x) - \min_{x\in S} \nu(x) \geq \varepsilon\big\}$$

is the closure of its interior and hence

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[\Delta_n \geq \varepsilon\big] &= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[M_n \in C_\varepsilon\big] \\ &= -\min_{\nu\in C_\varepsilon} H(\nu|\mu) =: -I(\varepsilon).\end{aligned}$$

Relative entropy

Remark It is quite tricky to calculate the function I explicitly.

For ε sufficiently small, it seems that the minimizers of the entropy $H(\,\cdot\,|\mu)$ on the set C_ε are (up to permutations of the coordinates) of the form

$$\nu(1) = \tfrac{1}{6} - \tfrac{1}{2}\varepsilon, \quad \nu(2) = \tfrac{1}{6} + \tfrac{1}{2}\varepsilon, \quad \nu(3), \ldots, \nu(6) = \tfrac{1}{6}.$$

For $\varepsilon > \frac{1}{3}$, this solution is of course no longer well-defined and the minimizer must look differently.

Markov chains

$(X_t)_{t\geq 0}$ continuous-time Markov chain with finite state space S, initial law μ, and *transition probabilities* $P_t(x,y)$.
For $0 = t_0 < \cdots < t_n$,

$$\begin{aligned}&\mathbb{P}\big[X_{t_0} = x_0, \ldots, X_{t_n} = x_n\big]\\ &\quad = \mu(x_0) P_{t_1 - t_0}(x_0, x_1) P_{t_2 - t_1}(x_1, x_2) \cdots P_{t_n - t_{n-1}}(x_n, x_n).\end{aligned}$$

Semigroup $(P_t)_{t\geq 0}$ defined by *generator* G through

$$P_t = e^{Gt} = \sum_{n=0}^{\infty} \frac{1}{n!} G^n t^n.$$

Here

$$\begin{aligned}&G(x,y) \geq 0 \quad (x \neq y) \quad \text{rate of jumps } x \mapsto y,\\ &G(x,x) = -\sum_{y:\, y\neq x} G(x,y).\end{aligned}$$

Non-exit probability

Let $U \subset S$ and assume that the transition rates are irreducible on U. Then by the Perron-Frobenius theorem, there exists a function f, unique up to a multiplicative constant, and a constant $\lambda \geq 0$, such that

$$\begin{aligned} &\text{(i)} \quad f > 0 \text{ on } U, \\ &\text{(ii)} \quad f = 0 \text{ on } S \backslash U, \\ &\text{(iii)} \quad Gf(x) = -\lambda f(x) \qquad (x \in U). \end{aligned}$$

Theorem The process X started in any initial law such that $\mathbb{P}[X_0 \in U] > 0$ satisfies

$$\lim_{t \to \infty} \frac{1}{t} \log \mathbb{P}\big[X_s \in U \ \forall 0 \leq s \leq t\big] = -\lambda.$$

Large deviation theory

Large deviation theory provides a *unified framework* for proving the results we have seen:

- Rare events happen in the *least unlikely way*.
- Cramér's theorem is closely linked to relative entropy.
- Even the Perron-Frobenius theorem is closely linked to a certain rate function, which can be expressed in terms of relative entropy.

Large deviation theory

Chapter 1 Abstract theory. Similarity between a *Large Deviation Principle* and *weak convergence* of probability measures.

Chapter 2 Large deviations of i.i.d. random variables. Cramér's theorem, moderate deviations, Sanov's theorem. Use of *convex analysis* and the *Legendre transform*.

Chapter 3 Large deviations for Markov chains.

No time for: Large deviations of random fields, connection to *Gibbs measures* and *phase transitions*. Large deviations of random graphs. And more...

History of the subject

- 1870ies: the physicist Boltzmann works on the probabilistic interpretation of entropy.
- 1938: the Swedish mathematician Cramér proves his theorem.
- 1966 Varadhan defines a *large deviation principle* and shows that it implies Varadhan's lemma.
- 1977 and 1984 Gärtner and Ellis establish the Gärtner-Ellis theorem.
- 1985 Ellis' influential book on large deviation theory and statistical mechanics.
- 1989 The book by Deuschel and Stroock.
- 1990 Bryc proves that Varadhan's lemma conversely implies a large deviation principle.
- 1991 O'Brian en Verwaat, and Puhalskii prove that exponential tightness implies a subsequential LDP.
- 1993 The book by Dembo and Zeitouni.

References

- F. Rassoul-Agha and Timo Seppäläinen (2015) *A Course on Large Deviations with an Introduction to Gibbs Measures.*
- F. den Hollander (2000) *Large Deviations.*
- A. Dembo and O. Zeitouni (1998) *Large deviations techniques and applications 2nd edition.*
- J.-D. Deuschel and D.W. Stroock (1989) *Large Deviations.*
- R.S. Ellis (1985) *Entropy, Large Deviations, and Statistical Mechanics.*
- P. Dupuis and R.S. Ellis (1997) *A Weak Convergence Approach to the Theory of Large Deviations.*
- A. Puhalskii (2001) *Large Deviations and Idempotent Probability.*