Large Deviations

We first introduce some general definitions. Suppose S is a a complete, separable metric space and ν_n is a sequence of probability measures on $(S, \mathcal{S})$.

Recall that a function $f : S \mapsto [-\infty, \infty]$ is a lower semicontinuous function if

$$f(x) \leq \liminf_{y \to x} f(y), \quad \forall x \in S.$$

f is lower semicontinuous if and only if for any real number c, $\{x : f(x) > c\}$ is an open subset of S. From this one can easily show that if $\{f_a(x) : a \in A\}$ is a family of continuous functions on S, then the function

$$f(x) = \sup_{a \in A} f_a(x), \quad x \in S$$

is a lower semicontinuous function on S.

Definition 1 *A rate function I is a lower semicontinuous function $I : S \mapsto [0, \infty]$. A good rate function is a rate function I for which all level sets $\Psi_I(\alpha) = \{x : I(x) \leq \alpha\}$, $\alpha \geq 0$, are compact subsets of S.*

Definition 2 *We say that $\{\nu_n\}$ satisfies the large deviation principle with a rate function I if*

(a) for any closed subset F of S,

$$\limsup \frac{1}{n} \log \nu_n(F) \leq - \inf_{x \in F} I(x);$$

(b) for any open subset G of S,

$$\liminf \frac{1}{n} \log \nu_n(G) \geq - \inf_{x \in G} I(x).$$

Suppose that $\{\nu_n\}$ satisfies the large deviation principle with a rate function I. If $\Gamma \in \mathcal{S}$ is a continuity set of I, i.e.,

$$\inf_{x \in \Gamma^o} I(x) = \inf_{x \in \overline{\Gamma}} I(x) := I_\Gamma,$$

then

$$\lim_{n\to\infty}\frac{1}{n}\log\nu_n(\Gamma) = -I_\Gamma$$

which can be written heuristically as

$$\nu_n(\Gamma) \asymp exp(-nI_\Gamma).$$

Now suppose that $X_1, X_2, \cdots$ are iid random variables. Let $\mu = EX_1$ and let ν be the distribution of X_1. Let ν_n be the distribution of $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n}$. The logarithmic moment generating function associated with ν is

$$\Lambda(\lambda) := \log M(\lambda) := \log E(e^{\lambda X_1}), \qquad \lambda \in \mathbf{R}.$$

Obviously, $\Lambda(0) = 0$, $\Lambda(\lambda) > -\infty$ for any λ. It is possible to have $\Lambda(\lambda) = \infty$.

Definition 3 *The Legendre-Fenchel transform of* $\Lambda(\lambda)$ *is*

$$\Lambda^*(x) = \sup_{\lambda\in\mathbf{R}}\{\lambda x - \Lambda(\lambda)\}.$$

From the definition one can easily see that Λ^* is a nonnegative lower semicontinuous function on R, so it is a rate function.

Theorem 1 *(Cramer's Theorem) The sequence of probability measures* $\{\nu_n\}$ *satisfies the large deviation principle with the convex rate function* $\Lambda^*(\cdot)$*, namely*

(a) for any closed subset F *of* $\mathbf{R}$*,*

$$\limsup\frac{1}{n}\log\nu_n(F) \le -\inf_{x\in F} I(x); \tag{1}$$

(b) for any open subset G *of* $\mathbf{R}$*,*

$$\liminf\frac{1}{n}\log\nu_n(G) \ge -\inf_{x\in G} I(x). \tag{2}$$

First we state the properties of $\Lambda^*(\cdot)$ and Λ that are needed for proving Cramer's Theorem.

Lemma 2 *(a)* Λ *and* Λ^* *are both convex functions on* R*.*

(b) If $\mathcal{D}_\Lambda := \{\lambda : \Lambda(\lambda) < \infty\} = \{0\}$*, then* Λ^* *is identically zero. If* $\Lambda(\lambda) < \infty$ *for some* $\lambda > 0$*, then* $\mu < \infty$ *and for all* $x \ge \mu$*,*

$$\Lambda^*(x) = \sup_{\lambda\ge 0}\{\lambda x - \Lambda(\lambda)\}, \tag{3}$$

is a nondeacreasing function on (μ, ∞). *Similarly, if* $\Lambda(\lambda) < \infty$ *for some* $\lambda < 0$, *then* $\mu > -\infty$ *and for all* $x \le \mu$,

$$\Lambda^*(x) = \sup_{\lambda \le 0}\{\lambda x - \Lambda(\lambda)\}, \tag{4}$$

is a nonincreasing function on $(-\infty, \mu)$. *When* μ *is finite,* $\Lambda^*(\mu) = 0$, *and we always have*

$$\inf_{x \in \mathbf{R}} \Lambda^*(x) = 0. \tag{5}$$

(c) $\Lambda(\cdot)$ *is differentiable in* $\mathcal{D}_\Lambda^o$ *with*

$$\Lambda'(\eta) = \frac{1}{M(\eta)} E(X_1 e^{\eta X_1}) \tag{6}$$

and

$$\Lambda'(\eta) = y \Longrightarrow \Lambda^*(y) = \eta y - \Lambda(\eta). \tag{7}$$

Proof. (a) The convexity of Λ follows from Hölder's inequality

$$\begin{aligned} \Lambda(\theta\lambda_1 + (1-\theta)\lambda_2) &= \log E[(e^{\lambda_1 X_1})^\theta (e^{\lambda_2 X_1})^{1-\theta}] \\ &\le \log\{E(e^{\lambda_1 X_1})^\theta E(e^{\lambda_2 X_1})^{1-\theta}\} \\ &= \theta\Lambda(\lambda_1) + (1-\theta)\Lambda(\lambda_2). \end{aligned}$$

The convexity of Λ^* follows from its definition

$$\begin{aligned} \theta\Lambda^*(x_1) + (1-\theta)\Lambda^*(x_2) &= \sup_{\lambda \in \mathbf{R}}\{\theta\lambda x_1 - \theta\Lambda(\lambda)\} + \sup_{\lambda \in \mathbf{R}}\{(1-\theta)\lambda x_2 - (1-\theta)\Lambda(\lambda)\} \\ &\ge \sup_{\lambda \in \mathbf{R}}\{(\theta x_1 + (1-\theta)x_2)\lambda - \Lambda(\lambda)\} \\ &= \Lambda^*(\theta x_1 + (1-\theta)x_2). \end{aligned}$$

(b) If $\mathcal{D}_\Lambda = \{0\}$, then $\Lambda^*(x) = \Lambda(0) = 0$ for all $x \in \mathbf{R}$. If $\Lambda(\lambda) = \log M(\lambda) < \infty$ for some $\lambda > 0$, then $\int_0^\infty x d\nu \le M(\lambda)/\lambda < \infty$, implying $\mu = \int x d\nu < \infty$. Now, for all $\lambda \in \mathbf{R}$, by Jensen's inequality

$$\Lambda(\lambda) = \log E(e^{\lambda X_1}) \ge E(\log e^{\lambda X_1}) = \lambda\mu.$$

If $\mu = -\infty$, then $\Lambda(\lambda) = \infty$ for λ negative and (3) holds trivially. When μ is finite, it follows from the proceeding inequality that $\Lambda^*(\mu) = 0$. In this case, for any $x \ge \mu$ and every $\lambda < 0$,

$$\lambda x - \Lambda(\lambda) \le \lambda\mu - \Lambda(\lambda) \le \Lambda^*(\mu) = 0,$$

and (3) follows. Observe that (3) implies the monotonicity of Λ^* on (μ, ∞), since for any $\lambda \ge 0$, $\lambda x - \Lambda(\lambda)$ is a nondecreasing function of x.

When $\Lambda(\lambda) < \infty$ for some $\lambda < 0$, both (4) and the monotonicity of Λ^* on $(-\infty, \mu)$ follows by considering the logarithmic moment generating function of $-X_1$, for which the proceeding proof applies.

It remains to prove that $\inf_{x\in\mathbf{R}} \Lambda^*(x) = 0$. This is already established for $\mathcal{D}_\Lambda = \{0\}$, in which case $\Lambda^* \equiv 0$, and when μ is finite, in which case, $\Lambda^*(\mu) = 0$. Now consider the case $\mu = -\infty$ while $\Lambda(\lambda) < \infty$ for some $\lambda > 0$. Then by (3) and Chebyshev

$$\begin{aligned}\log \nu([x,\infty)) &\le \inf_{\lambda\ge 0} \log E(e^{\lambda(X_1-x)}) \\ &= -\sup_{\lambda\ge 0}\{\lambda x - \Lambda(\lambda)\} = -\Lambda^*(x).\end{aligned}$$

Hence

$$\lim_{x\to-\infty} \Lambda^*(x) \le \lim_{x\to-\infty}\{-\log \nu([x,\infty))\} = 0,$$

and (5) follows.

The last remaining case, that of $\mu = \infty$ while $\Lambda(\lambda) < \infty$ for some $\lambda < 0$, is settled by considering the logarithmic moment generating function of $-X_1$.

(c) The identity (6) follows by interchanging the order of differentiation and integration. This is justified by the dominated convergence theorem since $f_\epsilon(x) := (e^{(\eta+\epsilon)x} - e^{\eta x})/\epsilon$ converges pointwise to $xe^{\eta x}$ as $\epsilon \to 0$, and $|f_\epsilon(x)| \le e^{\eta x}(e^{\delta|x|} - 1)/\delta := h(x)$ for every $\epsilon \in (-\delta, \delta)$, while $E|h(X_1)| < \infty$ for δ small enough.

Let $\Lambda'(\eta) = y$ and consider the function $g(\lambda) := \lambda y - \Lambda(\lambda)$. Since $g(\cdot)$ is a concave function and $g'(\eta) = 0$, it follows that $g(\eta) = \sup_{\lambda\in\mathbf{R}} g(\lambda)$ and (7) is established.

Proof of Cramer's Theorem (a) Let F be a nonempty closed subset of $\mathbf{R}$. (1) holds trivially when $I_F := \inf_{x\in F} \Lambda^*(x) = 0$. Assume $I_F > 0$. It follows from part (b) of the lemma that μ exists. For all x and every $\lambda > 0$, Chebyshev yields

$$\begin{aligned}\nu_n([x,\infty)) &= E(1_{\{\frac{S_n}{n}-x\ge 0\}}) \\ &\le E(e^{n\lambda(\frac{S_n}{n}-x)}) \\ &= e^{-n\lambda x}\prod_{i=1}^n E(e^{\lambda X_i}) \\ &= e^{-n(\lambda x - \Lambda(\lambda))} \qquad (8)\end{aligned}$$

Therefore, if $\mu < \infty$, then by (3) for every $x > \mu$,

$$\nu_n([x,\infty)) \le e^{-n\Lambda^*(x)}. \qquad (9)$$

By a similar argument, if $\mu > -\infty$ and $x < \mu$, then

$$\nu_n((-\infty, x]) \le e^{-n\Lambda^*(x)}. \qquad (10)$$

First, consider the case of μ finite. Then $\Lambda^*(\mu) = 0$, and because by assumption $I_F > 0$, μ must be in F^c. Let (x_-, x_+) be the component of F^c containing μ. Note that $x_- < x_+$ and either x_- or x_+ must be finite since F is nonempty. If x_- is finite, then $x_- \in F$, and consequently $\Lambda^*(x_-) \geq I_F$. Likewise, $\Lambda^*(x_+) \geq I_F$ whenever x_+ is finite. Applying (9) for $x = x_+$ and (10) for $x = x_-$ we get

$$\nu_n(F) \leq \nu_n(-\infty, x_-]) + \nu_n([x_+, \infty)) \leq 2e^{-nI_F}$$

and the upper bounds follows.

Suppose now that $\mu = -\infty$. Then, since Λ^* is nondecreasing, it follows from (5) that $\lim_{x\to-\infty} \Lambda^*(x) = 0$, and hence $x_+ = \inf\{x : x \in F\}$ is finite for otherwise $I_F = 0$. Since F is closed, $x_+ \in F$ and consequently $\Lambda^*(x_+) \geq I_F$. Moreover, $F \subset [x_+, \infty)$ and, therefore, the large deviation upper bounds follows by applying (9) for $x = x_+$.

The case of $\mu = \infty$ is handled analogously.

(b) We prove next that for every $\delta > 0$

$$\liminf \frac{1}{n} \log \nu_n((-\delta, \delta)) \geq \inf_{\lambda \in \mathbf{R}} \Lambda(\lambda) = -\Lambda^*(0). \tag{11}$$

Since the transform $Y = X - x$ results in $\Lambda_Y(\lambda) = \Lambda(\lambda) - \lambda x$ and hence $\Lambda^*_Y(\cdot) = \Lambda^*(\cdot + x)$, it follows from the proceeding inequality that for every x and every $\delta > 0$,

$$\liminf \frac{1}{n} \log \nu_n((x - \delta, x + \delta)) \geq -\Lambda^*(x). \tag{12}$$

For any open subset G of $\mathbf{R}$ and any $x \in G$, and any $\delta > 0$ small enough, $(x-\delta, x+\delta) \subset G$. Thus the large deviation lower bounds follows from (12).

Turning to the proof of the key inequality (11), first suppose that $\nu((-\infty, 0)) > 0$, $\nu((0, \infty)) > 0$ and ν is supported by a bounded subset of $\mathbf{R}$. By the former assumption, $\Lambda(\lambda) \to \infty$ as $|\lambda| \to \infty$, and by the later assumption, $\Lambda(\cdot)$ is finite everywhere. Accordingly, $\Lambda(\cdot)$ is a continuous, differentiable function, and hence there exists a finite η such that $\Lambda(\eta) = \inf_{\lambda\in\mathbf{R}} \Lambda(\lambda)$ and $\Lambda'(\eta) = 0$. Define a new probability measure $\tilde{\nu}$ in terms of ν by

$$\frac{d\tilde{\nu}}{d\nu}(x) = e^{\eta x - \Lambda(\eta)}$$

and observe that $\tilde{\nu}$ is a probability measure because

$$\int_{\mathbf{R}} d\tilde{\nu} = \frac{1}{M(\eta)} \int e^{\eta x} d\nu = 1.$$

Let $\tilde{\nu}_n$ be the distribution of $\frac{S_n}{n}$ when $X_1, X_2, \cdots$ are iid with distribution $\tilde{\nu}$. Note that for every $\epsilon > 0$,

$$\begin{aligned} \nu_n((-\epsilon, \epsilon)) &= \int_{|\sum_{i=1}^n x_i| < n\epsilon} \nu(dx_1) \cdots \nu(dx_n) \\ &\geq e^{-n\epsilon|\eta|} \int_{|\sum_{i=1}^n x_i| < n\epsilon} \exp(\eta \sum_{i=1}^n x_i) \nu(dx_1) \cdots \nu(dx_n) \\ &= e^{-n\epsilon|\eta|} e^{n\Lambda(\eta)} \tilde{\nu}_n((-\epsilon, \epsilon)) \end{aligned} \tag{13}$$

By (6) and the choice of η

$$E_{\tilde{\nu}}(X_1) = \frac{1}{M(\eta)} \int x e^{\eta x} d\nu = \Lambda'(\eta) = 0.$$

Hence, by the weak law of large numbers,

$$\lim_{n\to\infty} \tilde{\nu}_n((-\epsilon, \epsilon)) = 1. \tag{14}$$

It now follows from (13) that for every $0 < \epsilon < \delta$,

$$\begin{aligned} \liminf \frac{1}{n} \log \nu_n((-\delta, \delta)) &\geq \liminf \frac{1}{n} \log \nu_n((-\epsilon, \epsilon)) \\ &\geq \Lambda(\eta) - \epsilon|\eta| \end{aligned}$$

and (11) follows by letting $\epsilon \to 0$.

Suppose that ν is of unbounded support, while both $\nu((-\infty, 0)) > 0$ and $\nu((0, \infty)) > 0$. Fix M large enough so that $\nu([-M, 0)) > 0$ and $\nu((0, M]) > 0$, and let

$$\Lambda_M(\lambda) = \log \int_{-M}^{M} e^{\lambda x} d\nu.$$

Let $\overline{\nu}$ be the probability measure defined by

$$\overline{\nu}(A) = \frac{\nu(A \cap [-M, M])}{\nu([-M, M])},$$

and let $\overline{\nu}_n$ be the distribution of $\frac{S_n}{n}$ when $X_1, X_2, \cdots$ are iid with distribution $\overline{\nu}$. The for all n and every $\delta > 0$,

$$\nu_n((-\delta, \delta)) \geq \overline{\nu}_n((-\delta, \delta))\nu(([-M, M])^n.$$

Observe that by the proceeding paragraph, (11) holds for $\overline{\nu}_n$. Therefore, with the logarithmic moment generating function associated with $\overline{\nu}$ being $\Lambda_M(\lambda) - \log \nu([-M, M])$,

$$\begin{aligned} \liminf \frac{1}{n} \log \nu_n((-\delta, \delta)) &\geq \log \nu([-M, M]) + \liminf \frac{1}{n} \log \tilde{\nu}_n((-\delta, \delta)) \\ &\geq \inf_{\lambda \in \mathbf{R}} \Lambda_M(\lambda). \end{aligned}$$

With $I_M = -\inf_{\lambda\in\mathbf{R}} \Lambda_M(\lambda)$ and $I^* = \limsup_{M\to\infty} I_M$, it follows that

$$\liminf \frac{1}{n} \log \nu_n((-\delta, \delta)) \geq -I^*. \tag{15}$$

Note that $\Lambda_M(\cdot)$ is nondecreasing in M, and so is $-I_M$. Moreover, $-I_M \leq \Lambda_M(0) \leq \Lambda(0) = 0$, and hence $-I^* \leq 0$. Now, since $-I_M$ is finite for all M large enough, $-I^* > -\infty$. Therefore the level sets $\{\lambda : \Lambda_M(\lambda) \leq -I^*\}$ are nonempty, compact sets that are decreasing with respect to M, and hence there exists at least one point, denoted λ_0, in their intersection.

By the monotone convergence theorem, $\Lambda(\lambda_0) = \lim \Lambda_M(\lambda_0) \leq -I^*$, and consequently the bounds (15) yields (11), now for ν of unbounded support.

The proof of (11) for an arbitrary μ is completed by observing that if either $\mu((-\infty,0)) = 0$ or $\mu((0,\infty)) = 0$, then Λ is a monotone function with $\inf_{\lambda\in\mathbf{R}} \Lambda(\lambda) = \log \nu(\{0\})$. Hence, in this case, (11) follows from

$$\nu_n((-\delta,\delta)) \geq \nu_n(\{0\}) = \nu(\{0\})^n.$$

Lemma 3 *If $0 \in \mathcal{D}_\Lambda^o$ then Λ^* is a good rate function. Moreover, if $\mathcal{D}_\Lambda = \mathbf{R}$, then*

$$\lim_{|x|\to\infty} \frac{\Lambda^*(x)}{|x|} = \infty. \tag{16}$$

Proof. As $0 \in \mathcal{D}_\Lambda^o$, there exist $\lambda_- < 0$ and $\lambda_+ > 0$ that are both in $\mathcal{D}_\Lambda$. Since for any $\lambda \in \mathbf{R}$,

$$\frac{\Lambda^*(x)}{|x|} \geq \lambda \text{sign}(x) - \frac{\Lambda(\lambda)}{|x|},$$

it follows that

$$\liminf_{|x|\to\infty} \frac{\Lambda^*(x)}{|x|} \geq \min\{\lambda_+, -\lambda_-\} > 0.$$

In particular, $\Lambda^*(x) \to \infty$ as $|x| \to \infty$, and its level sets are closed and bounded, hence compact. Thus Λ^* is a good rate function. Note that (16) follows for $\mathcal{D}_\Lambda = \mathbf{R}$ by considering $-\lambda_- = \lambda_+ \to \infty$.

Examples

(a) If the iid sequence $X_1, X_2, \cdots$ has the normal ditribution with mean μ and variance σ^2 as their common distribution, then $M(\lambda) = \exp(\mu\lambda + \frac{\sigma^2\lambda^2}{2})$, $\Lambda(\lambda) = \mu\lambda + \frac{\sigma^2\lambda^2}{2}$, and $\Lambda^*(x) = \frac{(x-\mu)^2}{2\sigma^2}$.

(b) If the iid sequence $X_1, X_2, \cdots$ has the exponetial ditribution with parameter $\theta > 0$ as their common distribution, then $M(\lambda) = \frac{\theta}{\theta-\lambda}$ for $\lambda < \theta$, and ∞ otherwise; $\Lambda(\lambda) = \log\frac{\theta}{\theta-\lambda}$ for $\lambda < \theta$, and ∞ otherwise; $\Lambda^*(x) = \theta x - 1 - \log(\theta x)$ for $x > 0$, and ∞ otherwise.

(c) If the iid sequence $X_1, X_2, \cdots$ has the Bernoulli ditribution with parameter $p \in (0,1)$ as their common distribution, then $M(\lambda) = pe^\lambda + 1 - p$, $\Lambda^*(x) = x\log(\frac{x}{p}) + (1-p)\log(\frac{1-x}{1-p})$ for $x \in [0,1]$ and ∞ otherwise.

(d) If the iid sequence $X_1, X_2, \cdots$ has the Poisson ditribution with parameter $\theta > 0$ as their common distribution, then $M(\lambda) = \exp(\theta(e^\lambda - 1))$ and $\Lambda^*(x) = \theta - x + x\log(\frac{x}{\theta})$ for $x \geq 0$ and ∞ otherwise.