Why study large deviations?

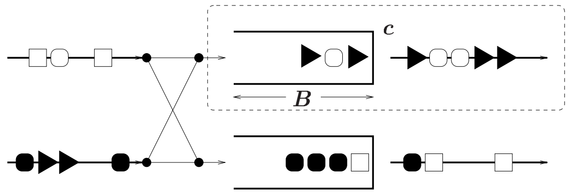
- The performance of many systems is limited by events which have a small probability of occurring, but which have severe consequences when they occur.
- The theory deals with rare events, and is asymptotic in nature.
- It can be viewed as a refinement of the law of large numbers.
- It is useful when simulation or numerical techniques become increasingly difficult as a parameter tends to its limit.
- It has many applications:
  queueing and communications models,
  information theory,
  simulation techniques,
  parameter estimation,
  hypothesis testing, $\ldots$

These slides were written to support an informal discussion of Chapters 1 and 2 of "Large Deviations for Performance Analysis", by Shwartz and Weiss.

Richard Weber, 12 October 1995

The problem of estimating buffer overflow frequency

The figure below shows a 2×2 switch, where output links are served at rate c.

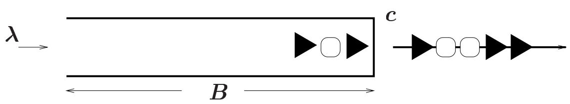


In order to know how many virtual circuits may be allowed to use this output link, for a given Quality of Service constraint, we need to estimate the probability that the content of the queue, Q_t, exceeds the buffer of size B.

$P(Q_t \geq B)$ should be small.

The overflow probability in a $M/M/1/B$ queue

Simply to illustrate ideas consider a single server $M/M/1/B$ queue, with finite buffer, here being shared by two traffic sources, with combined Poisson arrivals at rate λ



We know

$$P(Q_t = B) = \left[\frac{1-(\lambda/c)}{1-(\lambda/c)^{B+1}}\right](\lambda/c)^B.$$

Hence

$$P(Q_t = B) \sim e^{-B\log(c/\lambda)} \quad \text{for large } B,$$

where $\sim$ means

$$\lim_{B\to\infty}\frac{1}{B}\log P(Q_t = B) = -\log(c/\lambda).$$

This is typical.

Elements of large deviation theory

Here is another result of large deviation theory.

Suppose $x_1, x_2, \ldots$ are i.i.d. r.v.s then

$$P\left(\frac{1}{n}\sum_{i=1}^{n} x_i \in [a,b]\right) \sim e^{-n[\inf_{x\in[a,b]}\ell(x)+o(n)]}$$

We had for the queue:

$$P(Q_t = B) \sim e^{-B\log(c/\lambda)} \quad \text{for large } B.$$

These are typical. The general conclusions are:

- The asymptotic frequency of occurence of rare events depends in an exponential manner on some parameters of the problem. E.g., n, B.
- If a rare events occurs then it occurs in the most likely way. E.g., $\inf_{x\in[a,b]}\ell(x)$.
- Rare events occur as a Poisson process.

Chernoff's theorem (upper bound)

Suppose $x_1, x_2, \ldots$ is a sequence of i.i.d. random variables and $a \geq Ex_1$. Let $S_n = x_1 + \cdots + x_n$. Then for all $\theta > 0$,

$$\begin{aligned} P\left(S_n \geq na\right) &= E\,1[x_1 + \cdots + x_n - na \geq 0] \\ &\leq E\left(e^{\theta[x_1+\cdots+x_n-na]}\right) \\ &= e^{-n\left[a\theta - \log Ee^{\theta x_1}\right]} \end{aligned}$$

Hence

$$P\left(S_n \geq na\right) \leq e^{-n\sup_{\theta\geq 0}\left[\theta a - \log Ee^{\theta x_1}\right]}$$

Note that by Jensen's inequality that for all θ,

$$Ee^{\theta x_1} \geq e^{\theta E x_1}$$

and hence $\theta a - \log Ee^{\theta x_1} \leq \theta(a - Ex_1)$.

Thus

$$\begin{aligned} \ell(a) &\stackrel{\text{def}}{=} \sup_\theta \left[\theta a - \log Ee^{\theta x_1}\right] \\ &= \sup_{\theta \geq 0} \left[\theta a - \log Ee^{\theta x_1}\right] \end{aligned}$$

and we conclude

$$\boxed{P\left(S_n \geq na\right) \leq e^{-n\ell(a)}}$$

Note $\ell(Ex_1) = 0$.

Observations

- Note the key role of *moment generating function*, $M(\theta) = Ee^{\theta x_1}$ and *logarithmic moment generating function*, $\log M(\theta)$ (also called the cumulant generating function.)
- $\log M(\theta)$ is a convex function of θ.
- $\ell(a) := \sup_\theta\left[\theta a - \log M(\theta)\right]$ is called the *Legendre transform* of $\log M(\theta)$.
- $\ell(a)$ is a convex function of a.
- $\ell(a)$ and $\log M(\theta)$ are Legendre transform duals, i.e.,

$$\begin{aligned} \sup_a[a\theta - \ell(a)] &= \sup_a[a\theta - \sup_\phi(\phi a - \log M(\phi))] \\ &= \sup_a \inf_\phi[\log M(\phi) - a(\theta - \phi)] \\ &= \inf_{\phi:\phi=\theta} \log M(\phi) \\ &= \log M(\theta) \end{aligned}$$

- The optimizing θ, say θ^*, satisfies

$$a = M'(\theta^*)/M(\theta^*).$$

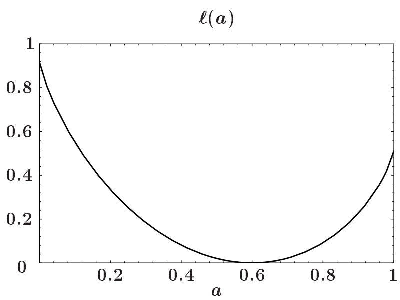
A typical rate function

Suppose $x_i = 0, 1$ with probabilities q, p. Then

$$\log M(\theta) = \log(q + pe^\theta),$$

and

$$\ell(a) = \begin{cases} a\log\left(\frac{a}{p}\right) + (1-a)\log\left(\frac{1-a}{1-p}\right), & 0 \leq a \leq 1 \\ \infty, & \text{otherwise.} \end{cases}$$



Here $Ex_1 = p = 0.6$.

- $\ell(a)$ is convex.
- $|\ell'(a)| \longrightarrow \infty$ as $a \longrightarrow$ boundary of the set where $\ell(a)$ is finite.
- $\ell(Ex_1) = 0$.

Chernoff's theorem (lower bound)

Suppose F is the distribution of x_1 and define

$$G(y) = M(\theta^*)^{-1}\int_{-\infty}^{y} e^{\theta^* x}dF(x)$$

where θ^* is as above. Then G is a distribution. It is called a *tilted distribution*. Note that if $\tilde{x} \sim G$,

$$E(\tilde{x}) = M(\theta^*)^{-1}\int_{-\infty}^{y} xe^{\theta^* x}dF(x) = \frac{M'(\theta^*)}{M(\theta^*)} = a.$$

Now $dG(y) = M(\theta^*)^{-1}e^{\theta^* y}dF(y)$, so

$$\begin{aligned} P\left(S_n \geq na\right) &= \int\cdots\int_{y_1+\cdots+y_n\geq na} dF(y_1)\ldots dF(y_n) \\ &= M(\theta^*)^n \int\cdots\int_{y_1+\cdots+y_n\geq na} e^{-\theta^*(y_1+\cdots+y_n)}dG(y_1)\ldots dG(y_n) \\ &\geq M(\theta^*)^n \int\cdots\int_{na+n\epsilon\geq y_1+\cdots+y_n\geq na} e^{-\theta^*(y_1+\cdots+y_n)}dG(y_1)\ldots dG(y_n) \\ &\geq e^{-n[\theta^*(a+\epsilon)-\log M(\theta^*)]}\int\cdots\int_{na+n\epsilon\geq y_1+\cdots+y_n\geq na} dG(y_1)\ldots dG(y_n) \\ &= e^{-n\ell(a)-n\epsilon}P(na+n\epsilon \geq \tilde{x}_1+\cdots+\tilde{x}_n \geq na) \\ &= e^{-n\ell(a)-n\epsilon}P\left(\sqrt{n}\epsilon \geq \frac{\tilde{x}_1+\cdots+\tilde{x}_n - na}{\sqrt{n}} \geq 0\right) \end{aligned}$$

Chernoff's theorem (lower bound), continued

$$P\left(S_n \geq na\right)$$
$$\geq e^{-n\ell(a)-n\epsilon} P\left(\sqrt{n}\epsilon \geq \frac{\tilde{x}_1 + \cdots + \tilde{x}_n - na}{\sqrt{n}} \geq 0\right)$$

Now

$$P\left(\sqrt{n}\epsilon \geq \frac{\tilde{x}_1 + \cdots + \tilde{x}_n - na}{\sqrt{n}} \geq 0\right) \to \frac{1}{2}$$

So since ϵ is arbitrarily small,

$$\boxed{\liminf_{n\to\infty} \frac{1}{n} \log P\left(S_n \geq na\right) \geq -\ell(a)}$$

- The upper and lower bounds together imply
$$P\left(S_n \geq na\right) = e^{-n[\ell(a)+o(n)]}$$
- We need conditions to ensure that $\theta a - \log M(\theta)$ is differentiable at θ^* and that its derivative is 0. It is enough to assume that $M(\theta)$ is finite in some neighborhood of 0 and that there is a θ^* in the interior of this neighborhood such that $\ell(a) = \theta^* a - \log M(\theta^*)$.

Illustration with the normal distribution

In the simple case that $x_i \sim N(\mu, \sigma^2)$,

$$\log M(\theta) = \theta\mu + \tfrac{1}{2}\theta^2\sigma^2$$

and

$$\ell(a) = (\mu - a)^2/2\sigma^2$$

A more refined estimate can be obtained from

$$\frac{1}{y+y^{-1}} e^{-\frac{1}{2}y^2} \leq \int_y^\infty e^{-\frac{1}{2}t^2} dt \leq \frac{1}{y} e^{-\frac{1}{2}y^2} \implies$$

$$P(S_n \geq na) \approx \frac{1}{\sqrt{2\pi n}(a-\mu)/\sigma} e^{-n(a-\mu)^2/2\sigma^2}$$
$$= \frac{1}{\sqrt{2\pi n}(a-\mu)/\sigma} e^{-n\ell(a)}$$

- The appearance of $1/\sqrt{n}$ $(= e^{-\frac{1}{2}\log n})$ is typical.
- An application of the theory would be to approximate $P(S_n \geq na)$ by $e^{-n(a-\mu)^2/2\sigma^2}$.
- Sometimes one can get refined approximations, e.g., as above, or the Bahadur-Rao approximation for the binomial distribution.

Generalization to i.i.d. vectors

Theorem 1.22. *Suppose $x_1, x_2, \ldots \in \mathbb{R}^d$ is a sequence of random vectors and*

$$M(\theta) = Ee^{<\theta, x_1>}.$$

Define the rate function

$$\ell(a) = \sup_\theta [< \theta, a > - \log M(\theta)].$$

Then for any set $C \subset \mathbb{R}^d$

$$\underline{\lim}_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{t=1}^n x_t \in C\right) \geq -\inf_{a\in C^o} \ell(a),$$

$$\overline{\lim}_{n\to\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{t=1}^n x_t \in C\right) \leq -\inf_{a\in \bar{C}} \ell(a),$$

where C^o and $\bar{C}$ are respectively the interior and closure of C.

Note: If you go back to the proof of Chernoff's theorem, you will see that you can easily extend the proof to statements about $P(S_n/n \in C)$. You can take C a closed set when doing the upper bound, but will need to take C to be an open set for the lower bound. (You'll want to let a^* be the minimizer of $\ell(a)$ and bound the probability of being in C by the probability of being in a neighbouhood of a^*; so you'll need that if $a^* \in C$ then a neighborhood of a^* is also in C.)

General statement of a large deviation principle

Suppose $z_1, z_2, \ldots$ is a sequence of random vectors in a probability space $(\mathcal{X}, \Omega, \mathcal{F})$. Here $\mathcal{X}$ might be $\mathbb{R}^d$, or perhaps $C[0,T]$, the space of continuous functions.
E.g., think of $z_n = (x_1 + \cdots + x_n)/n$.

Definition 2.1. *A real valued function I on $\mathcal{X}$ is called a "rate function" if*

(i) $I(x) \geq 0$,

(ii) I is lower semi-continuous; i.e., if $y_1, y_2, \ldots$ is a sequence such that $y_n \to y$ in $\mathcal{X}$ then $\liminf_{n\to\infty} I(y_n) \geq I(y)$.

Definition 2.2. *We say $z_1, z_2, \ldots$* satisfy a large deviation principle with rate function *I if for every set $C \subset \mathcal{X}$*

$$\underline{\lim}_{n\to\infty} \frac{1}{n} \log P\left(z_n \in C\right) \geq -\inf_{x\in C^o} I(x),$$

$$\overline{\lim}_{n\to\infty} \frac{1}{n} \log P\left(z_n \in C\right) \leq -\inf_{x\in \bar{C}} I(x),$$

where C^o and $\bar{C}$ are respectively the interior and closure of C.

If $\inf_{x\in C^o} I(x) = \inf_{x\in\bar{C}}$ then the two bounds coincide and C is said to be an *I-continuity* set for I.

Varadhan's lemma

Theorem 2.12. *Suppose that $z_1, z_2, \ldots$ satisfy a large deviation principle with rate function I. Then for any bounded continuous function g on $\mathcal{X}$,*

$$\lim_{n\to\infty} \frac{1}{n} \log E\left(e^{ng(z_n)}\right) = \sup_x [g(x) - I(x)].$$

The intuitive idea is that

$$\begin{aligned} E\left(e^{ng(z_n)}\right) &= \int_x e^{ng(x)} P(z_n \approx x) dx \\ &\approx \int_x e^{ng(x)} e^{-nI(x)} dx \\ &\approx e^{n \sup_x [g(x) - I(x)]} \end{aligned}$$

where the last line follows from Laplace's argument, that the rate of growth of an integral (or sum) is obtained by approximating it by its largest term.
E.g.,

$$4e^{-2n} + 6e^{-3n} + e^{-100n} \approx 4e^{-2n}$$

for large n.

The contraction principle

Definition 2.1. *A rate function is said to be a* good rate *function if*

(iii) The set $\{x : I(x) \le a\}$ is compact for every a.

Suppose that $z_1, z_2, \ldots$ satisfy a large deviation principle with rate function I. Let f be a continuous function and let $y_i = f(z_i)$. Define

$$I'(y) = \begin{cases} \inf\{I(x) : x \in \mathcal{X}, f(x) = y\} \\ \infty, \text{ if } y = f(x) \text{ for no } x \in \mathcal{X} \end{cases}$$

Theorem 2.13.

(i) If I is a good rate function then I' is a good rate function.

(ii) If $z_1, z_2, \ldots$ satisfy a large deviation principle with good rate function I then $y_1, y_2, \ldots$ satisfy a large deviation principle with good rate function I'.

Again, Laplace's argument gives the right intuition why this is true.

Sanov's theorem

Definition A.82. *Let $x_1, x_2, \ldots$ be a sequence random variables with distribution F and values in some metric space $\mathcal{X}$. The empirical distribution μ_n of a measurable set A is*

$$\mu_n(A) := \frac{1}{n} \sum_{i=1}^n 1[x_i \in A].$$

Suppose $x_1, x_2, \ldots$ are i.i.d. with distribution μ, i.e., $P(x_1 \le y) = \mu((-\infty, y])$. Define

$$I(\nu) = \int \log\left(\frac{d\nu}{d\mu}(y)\right) d\nu(y).$$

In the case that of a discrete distribution $(p_1, \ldots, p_d)$, over a discrete set of d points this would be

$$I(q) = \sum_{j=1}^d q_j \log\left(\frac{q_j}{p_j}\right) = H(q \mid p).$$

Theorem 1.22. *Consider the sequence $\mu_1, \mu_2, \ldots$. For every set C contained in the space of probability distributions,*

$$\underline{\lim}_{n\to\infty} \frac{1}{n} \log P(\mu_n \in C) \ge -\inf_{\nu \in C^o} I(\nu),$$

$$\overline{\lim}_{n\to\infty} \frac{1}{n} \log P(\mu_n \in C) \le -\inf_{\nu \in \bar{C}} I(\nu),$$

where C^o and $\bar{C}$ are respectively the interior and closure of C.

Sanov's theorem for a discrete distribution

Suppose $\mathcal{X} = \{1, \ldots, d\}$. Given $x_i = j$, let $y_i \in \{0,1\}^d$ be a vector whose jth component is equal to 1 and all others are equal to 0. For any $z \in \mathbb{R}^d$, let $|z| = \max_{1 \le j \le d} |z_j|$. Then

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n y_i - \bar{q}\right| \ge \epsilon\right) \sim \exp\left(-n \inf_{q : |q - \bar{q}| \ge \epsilon} \sum_{j=1}^d q_j \log\left(\frac{q_j}{p_j}\right)\right).$$

Example. Suppose we roll a die n times and the total is $\ge 4n$. The expected value is $3.5n$. So we have seen a rare event. How did this happen? For $\bar{q} = (.103, .123, .146, .174, .207, .247)$, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n y_i - \bar{q}\right| \ge \epsilon \;\middle|\; \frac{1}{n}\sum_{j=1}^n j x_j \ge 4\right) \lesssim$$

$$\frac{\exp\left(-n \inf_{q : |q - \bar{q}| \ge \epsilon, \sum_j j q_j \ge 4} \sum_{j=1}^d q_j \log\left(\frac{q_j}{1/6}\right)\right)}{\exp\left(-n \inf_{q : \sum_j j q_j > 4} \sum_{j=1}^d q_j \log\left(\frac{q_j}{1/6}\right)\right)} \to 0$$

where $\bar{q}$ is chosen as infimizer of the denominator.

Information theory and large deviations

Suppose a source generates letters from an alphabet of d symbols. Letters are i.i.d. choices amongst the d symbols, with probabilities $q_1, \ldots, q_d$. The empirical distribution of the symbols in a string of n symbols will be close to q, so without losing much information, we could ignore strings for which the empirical distribution is far from q.

There are d^n possible strings of length n, but we would be using only a fraction of these. The number we would be using, say M_n, is given by

$$\frac{m_n}{d^n} \approx \exp\left(-n \sum_{j=1}^{d} q_j \log\left(\frac{q_j}{1/d}\right)\right) = \frac{2^{nh(q)}}{d^n},$$

where $h(q) = -\sum_j q_j \log_2 q_j$. Hence

$$m_n = 2^{nh(q)} \leq 2^{n \log_2 d}.$$

This shows that the source has *information rate* $h(q) \leq \log_2 d$.