The Markov and Chebychev Inequalities

Some of the simplest results which address the atypical behavior of a random process are for large deviations from the mean or the average.

The Markov Inequality: Let $S$ be a nonnegative valued random variable with mean $\mu$. Then

$$\text{Prob} \ (S \geq a) \leq \frac{\mu}{a}, \quad a > 0.$$ 

Proof: Let $p(\cdot)$ denote the probability density function with respect to the Lebesgue measure of the random variable $S$. (An analogous proof can be written with the counting measure if $S$ is a discrete valued random variable, etc.) Observe that since $S$ has a finite mean,

$$\mu = \int_{s=0}^{\infty} sp(s)ds$$ 

$$= \int_{s=0}^{\infty} \int_{t=0}^{s} dt \ p(s)ds$$ 

$$= \int_{t=0}^{\infty} \int_{s=t}^{\infty} p(s)ds \ dt$$ 

$$\geq \int_{t=0}^{a} \int_{s=a}^{\infty} p(s)ds \ dt$$ 

$$= a \ \text{Prob} \ (S \geq a),$$

for every $a > 0$ and the assertion follows. \[\square\]

The Chebychev Inequality: Let $S$ be a real-valued random variable with finite mean $\mu$ and variance $\sigma^2$. Then

$$\text{Prob} \ (|S - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}, \quad \epsilon > 0.$$
**Proof:** Define the nonnegative valued random variable $(S - \mu)^2$, and observe that

$$\text{Prob} \left( |S - \mu| \geq \epsilon \right) = \text{Prob} \left( (S - \mu)^2 \geq \epsilon^2 \right).$$

A simple application of the Markov inequality then yields the Chebychev inequality. ■

If the random variable $S$ represents the sample mean $S_n$ of an ergodic process based on $n$ observations, the event considered by the Markov and Chebychev inequalities can be interpreted as a rare event: that the sample mean exceeds the statistical one. In this situation, however, the bounds provided by these inequalities are very crude. The variance of the sample mean of an i.i.d. process, for instance, decays as $O \left( \frac{1}{n} \right)$, and all we can conclude from the Chebychev inequality then is that $S_n$ converges in probability to the mean value of the process. To prove stronger results, such as its almost sure convergence, we need to sharpen these bounds. In general, if $A_n$ is a sequence of events such that $\text{Prob} \left( A_n \right) \rightarrow 0$, we may wish to determine the rate with respect to $n$ at which this probability decays to zero. We shall see shortly that in several interesting cases, this probability decays to zero exponentially fast. In such cases, we may wish to determine the exponent in the decay-rate. Large deviations theory is concerned with problems of this nature. Typical large deviations results provide upper and lower bounds for the exponent of the rate of decay.

**Loose definition:** We shall say that a random process $S = \{S_n\}_{n=0}^{\infty}$ satisfies a large deviations principle (LDP) if

$$\text{Prob} \left( S_n \in A \right) \approx K(n, A) \exp\{-nI(A)\},$$

where $A$ is the event of interest, $K(n, \cdot)$ is a *slowly varying* function in $n$, and $I(\cdot)$ is an $\mathbb{R}^+\text{-valued function which is called the rate function of the LDP.}

One of the first results in this spirit of large deviations theory is perhaps the Chernoff bound, which is frequently used in the analysis of communication systems. It provides, much like the Markov or Chebychev inequalities, an upper bound on the probability of a rare deviation from the expected behavior of a random variable.

**The Chernoff Bound:** Let $S$ be a real-valued random variable whose moment generating function exists in a neighborhood of zero\(^1\). Then

$$\text{Prob} \left( S \geq a \right) \leq \exp \left\{ - \sup_{\theta \geq 0} \left( a \theta - \log M(\theta) \right) \right\}, \quad a, \theta \in \mathbb{R},$$

where $M(\theta) = E \left[ e^{\theta S} \right]$ is the moment generating function of the random variable $S$.

\(^1\)This is slightly stronger than assuming that all moments of $S$ exist.
**Proof:** Observe $e^{\theta S}$ is a nonnegative valued and monotone function of $s$. For $\theta > 0$, it is increasing in $s$. Therefore,

\[
\text{Prob } (S \geq a) = \text{Prob } (\theta S \geq \theta a) = \text{Prob } (e^{\theta S} \geq e^{\theta a}) \leq \frac{E[e^{\theta S}]}{e^{\theta a}} = \exp \{-\theta a + \log M(\theta)\}.
\]

Since this inequality holds trivially for $\theta = 0$, we get the tightest bound by minimizing the expression on the right side of the inequality, *i.e.*, by maximizing

\[I_\theta(a) = \theta a - \log M(\theta)\]

over $\theta \geq 0$. \(\blacksquare\)

If $S_n = \frac{1}{n} \sum_{t=1}^{n} X_t$ is the sample mean of an i.i.d process $\{X_t, \ t = 1, 2, \ldots\}$ with common mean $\mu$ and moment generating function $M_X(\cdot)$, then $S_n$ has mean $\mu$ and moment generating function

\[M_{S_n}(\theta) = \left[M_X\left(\frac{\theta}{n}\right)\right]^n.
\]

Therefore, by the Chernoff bound,

\[
\text{Prob } (S_n \geq \mu + \epsilon) \leq \exp \left\{ - \sup_{\theta \geq 0} \left( (\mu + \epsilon)\theta - n \log M_{S_n}\left(\frac{\theta}{n}\right) \right) \right\}
\]

\[
= \exp \left\{ - \sup_{\theta' \geq 0} \left( ((\mu + \epsilon)n\theta' - n \log M_{S_n}(\theta')) \right) \right\}
\]

\[
= \exp \left\{ -n \sup_{\theta' \geq 0} ((\mu + \epsilon)\theta' - \log M_{S_n}(\theta')) \right\}
\]

We show later that the supremum is strictly positive for $\epsilon > 0$, making this a stronger result than the inequalities discussed previously.

A natural question to ask is whether the bound of Chernoff is tight, *i.e.*, whether the exact probability (which may be difficult to compute) is significantly smaller than the bound determined above. We shall see shortly that empirical means of a large class of processes satisfy a LDP, and that the Chernoff bound, in particular, is tight up to the first order in the exponent.
A Frequently Encountered Rate Function\(^2\)

The rate function in several large deviations results for \(\mathbb{R}\)-valued processes is related to the extended real-valued function

\[
I(x) = \sup_{\theta \in \mathbb{R}} \theta x - \log M(\theta), \quad x \in \mathbb{R},
\]

where \(M(\theta)\) is a suitable moment generating function. We now state some properties of this function which are useful for proving other results later.

- The function \(I(\cdot)\) is convex on \(\mathbb{R}\).
  - For any \(\lambda \in [0, 1]\),
    \[
    I(\lambda x_1 + (1 - \lambda) x_2) = \sup_{\theta \in \mathbb{R}} \theta (\lambda x_1 + (1 - \lambda) x_2) - \log M(\theta)
    \leq \lambda \sup_{\theta \in \mathbb{R}} (\theta x_1 - \log M(\theta)) + (1 - \lambda) \sup_{\theta \in \mathbb{R}} (\theta x_2 - \log M(\theta))
    \]
    and the assertion follows.

- The function \(I(\cdot)\) is nonnegative, and \(I(\mu) = 0\), where \(\mu\) is the mean of the random variable whose moment generating function is \(M(\cdot)\).
  - Note that \(I_0(x) = 0x - \log M(0) = 0\) for every \(x \in \mathbb{R}\). Therefore \(I(x)\), which is the supremum of \(I_\theta(x)\) over \(\theta\), is nonnegative. Further, by Jensen’s inequality
    \[
    \log M(\theta) = \log E [e^{\theta X}] \geq \log e^{\theta E[X]} = \theta \mu, \quad \theta \in \mathbb{R}.
    \]
    Thus \(I(\mu) = \sup_\theta \theta \mu - \log M(\theta) \leq 0\), and \(I(x)\) achieves its minimum at \(x = \mu\).

Recall that in the Chernoff bound for \(P(S_n \geq \mu + \epsilon)\), the exponent is

\[
\sup_{\theta \geq 0} (\mu + \epsilon) \theta - \log M_{S_n}(\theta).
\]

Now, we just showed that \(\theta \mu - \log M_{S_n}(\theta) \leq 0\) for all \(\theta\), and thus if \(\epsilon > 0\) then

\[
\theta < 0 \implies (\mu + \epsilon) \theta - \log M_{S_n}(\theta) = \epsilon \theta + (\mu \theta - \log M_{S_n}(\theta)) \leq \theta \epsilon < 0.
\]

But the supremum over \(\theta \in \mathbb{R}\) is nonnegative. Thus the supremum over all \(\theta \in \mathbb{R}\) must be the supremum over \(\theta \geq 0\). Finally, the convexity of \(I(\cdot)\) provides that the exponent is strictly positive if \(\epsilon > 0\).

\(^2\)This section and the following two sections dealing with Cramer’s Theorem and the Gartner-Ellis Theorem are adapted from J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*, John Wiley and Sons, 1990.
• If $M(\theta) < \infty$ for every $\theta \in \mathbb{R}$, and the smallest interval whose probability is one is all of $\mathbb{R}$, then for any $x \in \mathbb{R}$

$$I(x) = \theta_0 x - \log M(\theta_0), \quad \text{for some } \theta_0(x) = \theta_0 \in \mathbb{R},$$

i.e. the supremum in the definition of $I(x)$ is attained for every $x$.

- Fix any $M > 0$, and observe that

$$\liminf_{\theta \to \infty} \frac{\log M(\theta)}{\theta} = \liminf_{\theta \to \infty} \frac{\log \int_\mathbb{R} \exp \{\theta s\} \, dP(s)}{\theta}$$

$$\geq \liminf_{\theta \to \infty} \frac{\log \int_{-\infty}^{M} \exp \{\theta s\} \, dP(s)}{-\theta}$$

$$\geq \liminf_{\theta \to \infty} \frac{\log \left(\exp \{\theta M\} \int_{-\infty}^{M} \, dP(s)\right)}{-\theta}$$

$$= \liminf_{\theta \to \infty} \frac{\theta M + \log P([M, \infty))}{-\theta}$$

$$= M.$$  

Since the choice of $M$ is arbitrary, it must hold that

$$\liminf_{\theta \to \infty} \frac{\log M(\theta)}{\theta} = \infty.$$  

Similarly, for any $M < 0$,

$$\liminf_{\theta \to -\infty} \frac{\log M(\theta)}{-\theta} = \liminf_{\theta \to -\infty} \frac{\log \int_\mathbb{R} \exp \{\theta s\} \, dP(s)}{-\theta}$$

$$\geq \liminf_{\theta \to -\infty} \frac{\log \int_{-\infty}^{M} \exp \{\theta s\} \, dP(s)}{-\theta}$$

$$\geq \liminf_{\theta \to -\infty} \frac{\log \left(\exp \{\theta M\} \int_{-\infty}^{M} \, dP(s)\right)}{-\theta}$$

$$= \liminf_{\theta \to -\infty} \frac{\theta M + \log P((-\infty, M])}{-\theta}$$

$$= -M,$$

and therefore

$$\limsup_{\theta \to -\infty} \frac{\log M(\theta)}{-\theta} = -\infty.$$  

For every $x$, therefore, $\theta x - \log M(\theta) = \theta \left( x - \frac{\log M(\theta)}{\theta} \right) \to -\infty$ as $|\theta| \to \infty$, and thus the supremum is attained for some $\theta_0 \in (-\infty, \infty)$.

The finiteness of $M(\theta)$ for all $\theta$ also implies that it is differentiable at the maximum, and hence the maximizer of $I_\theta(x)$ may be obtained by solving

$$\frac{d I_\theta(x)}{d\theta} = 0 \quad \implies \quad x = \frac{M'(\theta)}{M(\theta)}.$$  

---

This is the same as saying that the random variable is essentially unbounded, both above and below.
Homework:

1. Compute the moment generating function and the “rate function” $I(x)$ when
   
   (a) $S$ is Gaussian, with mean $\mu$ and variance $\sigma^2$.
   
   (b) $S = S_n$ is the sample mean of an i.i.d. sequence of Bernoulli random variables, with $\text{Prob}(1) = p$ for some $0 < p < 1$.

   In either case, is there a $\theta_0$ which achieves $I(x)$? If so, when is $\theta_0 \geq 0$?

2. Use the Chernoff bound for the case 1b, together with the Borel-Cantelli lemma, to show that $S_n \to p$ a.s.