**MASSACHUSETTS INSTITUTE OF TECHNOLOGY**

6.265/15.070J Fall 2013
Lecture 2 9/9/2013

**Large Deviations for i.i.d. Random Variables**

**Content.** Chernoff bound using exponential moment generating functions. Properties of a moment generating functions. Legendre transforms.

## 1 Preliminary notes

The Weak Law of Large Numbers tells us that if $X_1, X_2, \ldots,$ is an i.i.d. sequence of random variables with mean $\mu \triangleq \mathbb{E}[X_1] < \infty$ then for every $\epsilon > 0$

$$\mathbb{P}(|\frac{X_1 + \ldots + X_n}{n} - \mu| > \epsilon) \to 0,$$

as $n \to \infty$.

But how quickly does this convergence to zero occur? We can try to use Chebyshev inequality which says

$$\mathbb{P}(|\frac{X_1 + \ldots + X_n}{n} - \mu| > \epsilon) \leq \frac{\text{Var}(X_1)}{n\epsilon^2}.$$

This suggest a "decay rate" of order $\frac{1}{n}$ if we treat Var$(X_1)$ and $\epsilon$ as a constant. Is this an accurate rate? Far from so ...

In fact if the higher moment of $X_1$ was finite, for example, $\mathbb{E}[X_1^{2m}] < \infty$, then using a similar bound, we could show that the decay rate is at least $\frac{1}{n^m}$ (exercise).

The goal of the large deviation theory is to show that in many interesting cases the decay rate is in fact *exponential*: $e^{-cn}$. The exponent $c > 0$ is called the *large deviations rate*, and in many cases it can be computed explicitly or numerically.

## 2 Large deviations upper bound (Chernoff bound)

Consider an i.i.d. sequence with a common probability distribution function $F(x) = \mathbb{P}(X \le x), x \in \mathbb{R}$. Fix a value $a > \mu$, where $\mu$ is again an expectation corresponding to the distribution $F$. We consider probability that the average of $X_1, \ldots, X_n$ exceeds $a$. The WLLN tells us that this happens with probability converging to zero as $n$ increases, and now we obtain an estimate on this probability. Fix a positive parameter $\theta > 0$. We have

$$
\begin{aligned}
\mathbb{P}(\frac{\sum_{1\le i\le n} X_i}{n} > a) &= \mathbb{P}(\sum_{1\le i\le n} X_i > na) \\
&= \mathbb{P}(e^{\theta \sum_{1\le i\le n} X_i} > e^{\theta na}) \\
&\le \frac{\mathbb{E}[e^{\theta \sum_{1\le i\le n} X_i}]}{e^{\theta na}} \qquad \text{Markov inequality} \\
&= \frac{\mathbb{E}[\prod_i e^{\theta X_i}]}{(e^{\theta a})^n},
\end{aligned}
$$

But recall that $X_i$'s are i.i.d. Therefore $\mathbb{E}[\prod_i e^{\theta X_i}] = (\mathbb{E}[e^{\theta X_1}])^n$. Thus we obtain an upper bound

$$
\mathbb{P}(\frac{\sum_{1\le i\le n} X_i}{n} > a) \le \left(\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}\right)^n. \tag{1}
$$

Of course this bound is meaningful only if the ratio $\mathbb{E}[e^{\theta X_1}]/e^{\theta a}$ is less than unity. We recognize $\mathbb{E}[e^{\theta X_1}]$ as the moment generating function of $X_1$ and denote it by $M(\theta)$. For the bound to be useful, we need $\mathbb{E}[e^{\theta X_1}]$ to be at least finite. If we could show that this ratio is less than unity, we would be done – exponentially fast decay of the probability would be established.

Similarly, suppose we want to estimate

$$
\mathbb{P}(\frac{\sum_{1\le i\le n} X_i}{n} < a),
$$

for some $a < \mu$. Fixing now a negative $\theta < 0$, we obtain

$$
\begin{aligned}
\mathbb{P}(\frac{\sum_{1\le i\le n} X_i}{n} < a) &= \mathbb{P}(e^{\theta \sum_{1\le i\le n} X_i} > e^{\theta na}) \\
&\le \left(\frac{M(\theta)}{e^{\theta a}}\right)^n,
\end{aligned}
$$

and now we need to find a negative $\theta$ such that $M(\theta) < e^{\theta a}$. In particular, we need to focus on $\theta$ for which the moment generating function is finite. For this purpose let $\mathcal{D}(M) \triangleq \{\theta : M(\theta) < \infty\}$. Namely $\mathcal{D}(M)$ is the set of values $\theta$ for which the moment generating function is finite. Thus we call $\mathcal{D}$ the domain of $M$.

## 3 Moment generating function. Examples and properties

Let us consider some examples of computing the moment generating functions.

- **Exponential distribution.** Consider an exponentially distributed random variable $X$ with parameter $\lambda$. Then

$$\begin{aligned} M(\theta) &= \int_0^\infty e^{\theta x} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^\infty e^{-(\lambda-\theta)x} dx. \end{aligned}$$

When $\theta < \lambda$ this integral is equal to $\frac{-1}{\lambda-\theta} e^{-(\lambda-\theta)x} \Big|_0^\infty = 1/(\lambda-\theta)$. But when $\theta \geq \lambda$, the integral is infinite. Thus the exp. moment generating function is finite iff $\theta < \lambda$ and is $M(\theta) = \lambda/(\lambda-\theta)$. In this case the domain of the moment generating function is $\mathcal{D}(M) = (-\infty, \lambda)$.

**Standard Normal distribution.** When $X$ has standard Normal distribution, we obtain

$$\begin{aligned} M(\theta) = \mathbb{E}[e^{\theta X}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{\theta x} e^{-\frac{x^2}{2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{x^2 - 2\theta x + \theta^2 - \theta^2}{2}} dx \\ &= e^{\frac{\theta^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{(x-\theta)^2}{2}} dx \end{aligned}$$

Introducing change of variables $y = x - \theta$ we obtain that the integral is equal to $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{y^2}{2}} dy = 1$ (integral of the density of the standard Normal distribution). Therefore $M(\theta) = e^{\frac{\theta^2}{2}}$. We see that it is always finite and $\mathcal{D}(M) = \mathbb{R}$.

In a retrospect it is not surprising that in this case $M(\theta)$ is finite for all $\theta$. The density of the standard Normal distribution "decays like" $\approx e^{-x^2}$ and this is faster than just exponential growth $\approx e^{\theta x}$. So no matter how large is $\theta$ the overall product is finite.

- **Poisson distribution.** Suppose $X$ has a Poisson distribution with parameter $\lambda$. Then

$$\begin{aligned} M(\theta) = \mathbb{E}[e^{\theta X}] &= \sum_{m=0}^{\infty} e^{\theta m} \frac{\lambda^m}{m!} e^{-\lambda} \\ &= \sum_{m=0}^{\infty} \frac{(e^{\theta}\lambda)^m}{m!} e^{-\lambda} \\ &= e^{e^{\theta}\lambda - \lambda}, \end{aligned}$$

(where we use the formula $\sum_{m\geq 0} \frac{t^m}{m!} = e^t$). Thus again $\mathcal{D}(M) = \mathbb{R}$. This again has to do with the fact that $\lambda^m/m!$ decays at the rate similar to $1/m!$ which is faster then any exponential growth rate $e^{\theta m}$.

We now establish several properties of the moment generating functions.

**Proposition 1.** *The moment generating function $M(\theta)$ of a random variable $X$ satisfies the following properties:*

*(a)* $M(0) = 1$. *If $M(\theta) < \infty$ for some $\theta > 0$ then $M(\theta') < \infty$ for all $\theta' \in [0, \theta]$. Similarly, if $M(\theta) < \infty$ for some $\theta < 0$ then $M(\theta') < \infty$ for all $\theta' \in [\theta, 0]$. In particular, the domain $\mathcal{D}(M)$ is an interval containing zero.*

*(b) Suppose $(\theta_1, \theta_2) \subset \mathcal{D}(M)$. Then $M(\theta)$ as a function of $\theta$ is differentiable in $\theta$ for every $\theta_0 \in (\theta_1, \theta_2)$, and furthermore,*

$$\frac{d}{d\theta} M(\theta)\Big|_{\theta=\theta_0} = \mathbb{E}[X e^{\theta_0 X}] < \infty.$$

*Namely, the order of differentiation and expectation operators can be changed.*

*Proof.* Part (a) is left as an exercise. We now establish part (b). Fix any $\theta_0 \in (\theta_1, \theta_2)$ and consider a $\theta$-indexed sequence of random variables

$$Y_\theta \triangleq \frac{\exp(\theta X) - \exp(\theta_0 X)}{\theta - \theta_0}.$$

Since $\frac{d}{d\theta}\exp(\theta x) = x\exp(\theta x)$, then almost surely $Y_\theta \to X\exp(\theta_0 X)$, as $\theta \to \theta_0$. Thus to establish the claim it suffices to show that convergence of expectations holds as well, namely $\lim_{\theta\to\theta_0}\mathbb{E}[Y_\theta] = \mathbb{E}[X\exp(\theta_0 X)]$, and $\mathbb{E}[X\exp(\theta_0 X)] < \infty$. For this purpose we will use the Dominated Convergence Theorem. Namely, we will identify a random variable $Z$ such that $|Y_\theta| \le Z$ almost surely in some interval $(\theta_0 - \epsilon, \theta_0 + \epsilon)$, and $\mathbb{E}[Z] < \infty$.

Fix $\epsilon > 0$ small enough so that $(\theta_0 - \epsilon, \theta_0 + \epsilon) \subset (\theta_1, \theta_2)$. Let $Z = \epsilon^{-1}\exp(\theta_0 X + \epsilon|X|)$. Using the Taylor expansion of $\exp(\cdot)$ function, for every $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$, we have

$$Y_\theta = \exp(\theta_0 X)\left(X + \frac{1}{2!}(\theta-\theta_0)X^2 + \frac{1}{3!}(\theta-\theta_0)^2X^3 + \cdots + \frac{1}{n!}(\theta-\theta_0)^{n-1}X^n + \cdots\right),$$

which gives

$$\begin{aligned}
|Y_\theta| &\le \exp(\theta_0 X)\left(|X| + \frac{1}{2!}(\theta-\theta_0)|X|^2 + \cdots + \frac{1}{n!}(\theta-\theta_0)^{n-1}|X|^n + \cdots\right)\\
&\le \exp(\theta_0 X)\left(|X| + \frac{1}{2!}\epsilon|X|^2 + \cdots + \frac{1}{n!}\epsilon^{n-1}|X|^n + \cdots\right)\\
&= \exp(\theta_0 X)\epsilon^{-1}\left(\exp(\epsilon|X|) - 1\right)\\
&\le \exp(\theta_0 X)\epsilon^{-1}\exp(\epsilon|X|)\\
&= Z.
\end{aligned}$$

It remains to show that $\mathbb{E}[Z] < \infty$. We have

$$\begin{aligned}
\mathbb{E}[Z] &= \epsilon^{-1}\mathbb{E}[\exp(\theta_0 X + \epsilon X)\mathbf{1}\{X \ge 0\}] + \epsilon^{-1}\mathbb{E}[\exp(\theta_0 X - \epsilon X)\mathbf{1}\{X < 0\}]\\
&\le \epsilon^{-1}\mathbb{E}[\exp(\theta_0 X + \epsilon X)] + \epsilon^{-1}\mathbb{E}[\exp(\theta_0 X - \epsilon X)]\\
&= \epsilon^{-1}M(\theta_0 + \epsilon) + \epsilon^{-1}M(\theta_0 - \epsilon)\\
&< \infty,
\end{aligned}$$

since $\epsilon$ was chosen so that $(\theta_0 - \epsilon, \theta_0 + \epsilon) \subset (\theta_1, \theta_2) \subset \mathcal{D}(M)$. This completes the proof of the proposition. □

**Problem 1.**

*(a) Establish part (a) of Proposition 1.*

*(b) Construct an example of a random variable for which the corresponding interval is trivial $\{0\}$. Namely, $M(\theta) = \infty$ for every $\theta > 0$.*

(c) *Construct an example of a random variable $X$ such that $\mathcal{D}(M) = [\theta_1, \theta_2]$ for some $\theta_1 < 0 < \theta_2$. Namely, the the domain $\mathcal{D}$ is a non-zero length closed interval containing zero.*

Now suppose the i.i.d. sequence $X_i, i \geq 1$ is such that $0 \in (\theta_1, \theta_2) \subset \mathcal{D}(M)$, where $M$ is the moment generating function of $X_1$. Namely, $M$ is finite in a neighborhood of 0. Let $a > \mu = \mathbb{E}[X_1]$. Applying Proposition 1, let us differentiate this ratio with respect to $\theta$ at $\theta = 0$:

$$\frac{d}{d\theta}\frac{M(\theta)}{e^{\theta a}} = \frac{\mathbb{E}[X_1 e^{\theta X_1}]e^{\theta a} - a e^{\theta a}\mathbb{E}[e^{\theta X_1}]}{e^{2\theta a}} = \mu - a < 0.$$

Note that $M(\theta)/e^{\theta a} = 1$ when $\theta = 0$. Therefore, for sufficiently small positive $\theta$, the ratio $M(\theta)/e^{\theta a}$ is smaller than unity, and (1) provides an exponential bound on the tail probability for the average of $X_1, \ldots, X_n$.

Similarly, if $a < \mu$, the ratio $M(\theta)/e^{\theta a} < 1$ for sufficiently small negative $\theta$.

We now summarize our findings.

**Theorem 1 (Chernoff bound).** *Given an i.i.d. sequence $X_1, \ldots, X_n$ suppose the moment generating function $M(\theta)$ is finite in some interval $(\theta_1, \theta_2) \ni 0$. Let $a > \mu = \mathbb{E}[X_1]$. Then there exists $\theta > 0$, such that $M(\theta)/e^{\theta a} < 1$ and*

$$\mathbb{P}(\frac{\sum_{1\leq i\leq n} X_i}{n} > a) \leq \left(\frac{M(\theta)}{e^{\theta a}}\right)^n.$$

*Similarly, if $a < \mu$, then there exists $\theta < 0$, such that $M(\theta)/e^{\theta a} < 1$ and*

$$\mathbb{P}(\frac{\sum_{1\leq i\leq n} X_i}{n} < a) \leq \left(\frac{M(\theta)}{e^{\theta a}}\right)^n.$$

How small can we make the ratio $M(\theta)/\exp(\theta a)$? We have some freedom in choosing $\theta$ as long as $\mathbb{E}[e^{\theta X_1}]$ is finite. So we could try to find $\theta$ which minimizes the ratio $M(\theta)/e^{\theta a}$. This is what we will do in the rest of the lecture. The surprising conclusion of the large deviations theory is very often that such a minimizing value $\theta^*$ exists and is tight. Namely it provides *the correct decay rate*! In this case we will be able to say

$$\mathbb{P}(\frac{\sum_{1\leq i\leq n} X_i}{n} > a) \approx \exp(-I(a, \theta^*)n),$$

where $I(a, \theta^*) = -\log\left(M(\theta^*)/e^{\theta^* a}\right)$.

## 4 Legendre transforms

Theorem 1 gave us a large deviations bound $(M(\theta)/e^{\theta a})^n$ which we rewrite as $e^{-n(\theta a - \log M(\theta))}$. We now study in more detail the exponent $\theta a - \log M(\theta)$.

**Definition 1.** *A Legendre transform of a random variable $X$ is the function $I(a) \triangleq \sup_{\theta\in\mathbb{R}}(\theta a - \log M(\theta))$.*

Let us go over the examples of some distributions and compute their corresponding Legendre transforms.

- **Exponential distribution with parameter $\lambda$.** Recall that $M(\theta) = \lambda/(\lambda-\theta)$ when $\theta < \lambda$ and $M(\theta) = \infty$ otherwise. Therefore when $\theta < \lambda$

  $$\begin{aligned} I(a) &= \sup_\theta (a\theta - \log \frac{\lambda}{\lambda-\theta}) \\ &= \sup_\theta (a\theta - \log\lambda + \log(\lambda-\theta)), \end{aligned}$$

  and $I(a) = -\infty$ otherwise. Setting the derivative of $g(\theta) = a\theta - \log\lambda + \log(\lambda-\theta)$ equal to zero we obtain the equation $a - 1/(\lambda-\theta) = 0$ which has the unique solution $\theta^* = \lambda - 1/a$. For the boundary cases, we have $a\theta - \log\lambda + \log(\lambda-\theta)) \to -\infty$ when either $\theta \uparrow \lambda$ or $\theta \to -\infty$ (check). Therefore

  $$\begin{aligned} I(a) &= a(\lambda - 1/a) - \log\lambda + \log(\lambda - \lambda + 1/a) \\ &= a\lambda - 1 - \log\lambda + \log(1/a) \\ &= a\lambda - 1 - \log\lambda - \log a. \end{aligned}$$

  The large deviations bound then tells us that when $a > 1/\lambda$

  $$\mathbb{P}(\frac{\sum_{1\leq i\leq n} X_i}{n} > a) \approx e^{-(a\lambda - 1 - \log\lambda - \log a)n}.$$

  Say $\lambda = 1$ and $a = 1.2$. Then the approximation gives us $\approx e^{-(.2-\log 1.2)n}$.

  Note that we can obtain an exact expression for this tail probability. Indeed, $X_1, X_1 + X_2, \ldots, X_1 + X_2 \cdots + X_n, \ldots$ are the events of a Poisson process with parameter $\lambda = 1$. Therefore we can compute the probability $\mathbb{P}(\quad_{1\leq i\leq n} X_i > 1.2n)$ exactly: it is the probability that the Poisson process has at most $n-1$ events before time $1.2n$. Thus

$$\mathbb{P}(\frac{\sum_{1\le i\le n} X_i}{n} > 1.2) = \mathbb{P}(\sum_{1\le i\le n} X_i > 1.2n)$$
$$= \sum_{0\le k\le n-1} \frac{(1.2n)^k}{k!} e^{-1.2n}.$$

It is not at all clear how revealing this expression is. In hindsight, we know that it is approximately $e^{-(.2-\log 1.2)n}$, obtained via large deviations theory.

- **Standard Normal distribution.** Recall that $M(\theta) = e^{\frac{\theta^2}{2}}$ when $X_1$ has the standard Normal distribution. The expected value $\mu = 0$. Thus we fix $a > 0$ and obtain

$$I(a) = \sup_\theta (a\theta - \frac{\theta^2}{2})$$
$$= \frac{a^2}{2},$$

achieved at $\theta^* = a$. Thus for $a > 0$, the large deviations theory predicts that

$$\mathbb{P}(\frac{\sum_{1\le i\le n} X_i}{n} > a) \approx e^{-\frac{a^2}{2}n}.$$

Again we could compute this probability directly. We know that $\frac{\sum_{1\le i\le n} X_i}{n}$ is distributed as a Normal random variable with mean zero and variance $1/n$. Thus

$$\mathbb{P}(\frac{\sum_{1\le i\le n} X_i}{n} > a) = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_a^\infty e^{-\frac{t^2 n}{2}} dt.$$

After a little bit of technical work one could show that this integral is "dominated" by its part around $a$, namely, $\int_a^{a+\epsilon} \cdot$, which is further approximated by the value of the function itself at $a$, namely $\frac{\sqrt{n}}{\sqrt{2\pi}} e^{-\frac{a^2}{2}n}$. This is consistent with the value given by the large deviations theory. Simply the lower order magnitude term $\frac{\sqrt{n}}{\sqrt{2\pi}}$ disappears in the approximation on the log scale.

- **Poisson distribution.** Suppose $X$ has a Poisson distribution with parameter $\lambda$. Recall that in this case $M(\theta) = e^{e^{\theta}\lambda - \lambda}$. Then

  $$I(a) = \sup_{\theta}(a\theta - (e^{\theta}\lambda - \lambda)).$$

  Setting derivative to zero we obtain $\theta^* = \log(a/\lambda)$ and $I(a) = a\log(a/\lambda) - (a - \lambda)$. Thus for $a > \lambda$, the large deviations theory predicts that

  $$\mathbb{P}(\frac{\sum_{1\le i\le n} X_i}{n} > a) \approx e^{-(a\log(a/\lambda) - a + \lambda)n}.$$

  In this case as well we can compute the large deviations probability explicitly. The sum $X_1 + \cdots + X_n$ of Poisson random variables is also a Poisson random variable with parameter $\lambda n$. Therefore

  $$\mathbb{P}(\sum_{1\le i\le n} X_i > an) = \sum_{m > an} \frac{(\lambda n)^m}{m!} e^{-\lambda n}.$$

  But again it is hard to infer a more explicit rate of decay using this expression

## 5 Additional reading materials

- Chapter 0 of [2]. This is non-technical introduction to the field which describes motivation and various applications of the large deviations theory. Soft reading.
- Chapter 2.2 of [1].

## References


[1] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Springer, 1998.

[2] A. Shwartz and A. Weiss, *Large deviations for performance analysis*, Chapman and Hall, 1995.

15.070J / 6.265J Advanced Stochastic Processes
Fall 2013