Advanced Stochastic Processes.

David Gamarnik

LECTURE 3
Large Deviations for i.i.d. Random Variables

Outline of Lecture

- Chernoff bound using exponential moment generating functions. Examples. Legendre transforms.

3.1. 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 $\epsilon > 0$

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

as $n \rightarrow \infty$.

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

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

This suggest a "decay rate" of order $\frac{1}{n}$ if we treat $\mathrm{Var}(X_1)$ and ϵ 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 $\frac{1}{n^m}$ (we skip the proof).

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 *large deviations rate*.

3.2. 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 $a > \mu$. 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

$$\begin{aligned}
\mathbb{P}(\frac{\sum_{1\le i\le n} X_i}{n} > a) &= \mathbb{P}(\sum_{1\le i\le n} X_i > na) \\
&= \mathbb{P}(e^{\theta \sum_{1\le i\le n} X_i} > e^{\theta na}) \\
&\le \frac{\mathbb{E}[e^{\theta \sum_{1\le i\le n} X_i}]}{e^{\theta na}} \qquad \text{Markov inequality} \\
&= \frac{\mathbb{E}[e^{\theta X_1} \cdots e^{\theta X_n}]}{(e^{\theta a})^n},
\end{aligned}$$

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

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

Of course this bound is meaningful only if the ratio $\mathbb{E}[e^{\theta X_1}]/e^{\theta a}$ is less than unity. At least we need $\mathbb{E}[e^{\theta X_1}]$ to be 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.

Can we expect to find θ for which this is indeed the case? First assume for a moment that $\mathbb{E}[e^{\theta X_1}]$ is finite for all θ in some interval $[0, \theta_0)$. Note that when $\theta = 0$ the ratio $\mathbb{E}[e^{\theta X_1}]/e^{\theta a}$ is equal to unity. Now let us differentiate this ratio with respect to θ at $\theta = 0$:

$$\frac{d}{d\theta}\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}} = \frac{\mathbb{E}[X_1 e^{\theta X_1}]e^{\theta a} - a e^{\theta a}\mathbb{E}[e^{\theta X_1}]}{e^{2\theta a}}$$

When we set $\theta = 0$ we obtain that the derivative is $\mathbb{E}[X_1] - a = \mu - a < 0$. Therefore, for sufficiently small θ the ratio $\mathbb{E}[e^{\theta X_1}]/e^{\theta a}$ is indeed smaller than unity!

We have established an upper bound part of the large deviations theory:

Theorem 3.1 (**Chernoff bound**)**.** *Given an i.i.d. sequence $X_1, \ldots, X_n$ suppose $\mathbb{E}[e^{\theta X_1}]$ is finite for all θ in some interval $[0, \theta_0)$. Let $a > \mu = \mathbb{E}[X_1]$. Then for some sufficiently small $\theta > 0$ there holds $\mathbb{E}[e^{\theta X_1}]/e^{\theta a} < 1$ and, moreover*

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

In words the large deviations probability is exponentially (geometrically) small.

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

3.3. Exponential moment generating function

Definition 3.2. An exponential moment generating function of a random variable X with parameter θ is defined to be $M(\theta) \triangleq \mathbb{E}[e^{\theta X}] = \int_{-\infty}^{\infty} e^{\theta x} dF(x)$.

We will be primarily interested in the case when $\theta \geq 0$. Note that when $\theta = 0$, we have $M(0) = 1$. Also when $\theta \geq 0$ the part of the integral corresponding to negative values $\int_{-\infty}^{0} e^{\theta x} dF(x)$ is at most unity. Indeed since $e^{\theta x} \leq 1$ when $x < 0$ then

$$\begin{aligned}\int_{-\infty}^{0} e^{\theta x} dF(x) &\leq \int_{-\infty}^{0} dF(x)\\ &\leq \int_{-\infty}^{\infty} dF(x)\\ &= 1\end{aligned}$$

However, the positive part $\int_{0}^{\infty} e^{\theta x} dF(x)$ can be infinite (we will have examples later).

The important case for us is when it is finite at least for *some* $\theta > 0$.

Proposition 1. Suppose $M(\theta) < \infty$ for some $\theta > 0$. Then for every $0 \leq \theta' < \theta$, also $M(\theta') < \infty$. Thus the set of all θ for which $M(\theta)$ is finite is some interval. This interval can be open $[0, \theta_0)$ or closed $[0, \theta_0]$.

Proof. We showed already that $\int_{-\infty}^{0} e^{\theta' x} dF(x) < \infty$. Now consider an integral over the positive part:

$$\begin{aligned}\int_{0}^{\infty} e^{\theta' x} dF(x) &\leq \int_{0}^{\infty} e^{\theta x} dF(x)\\ &\leq \int_{-\infty}^{\infty} e^{\theta x} dF(x)\\ &= M(\theta)\\ &< \infty.\end{aligned}$$

Therefore the entire integral $M(\theta') = \int_{-\infty}^{\infty} e^{\theta' x} dF(x)$ is finite. Thus the set of θ for which $M(\theta)$ is finite is an interval. We will show later that this interval is open from the right when the underlying distribution is exponential. Exercise 1 below asks you to find an example of a distribution for which the corresponding interval is closed. □

Exercise 1.
- (a) Construct example of a random variable for which the corresponding interval is trivial $\{0\}$. Namely, $M(\theta) = \infty$ for every $\theta > 0$.
- (b) (*) Construct an example of a random variable X for which the corresponding interval $[0, \theta_0]$ is closed. That is $M(\theta_0) < \infty$, but $M(\theta') = \infty$ for all $\theta' > \theta_0$.

3.4. Examples

Let us do some examples of computing exponential moment generating functions.

- **Exponential distribution.** Consider an exponentially distributed random variable X with parameter λ. Then

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

When $\theta < \lambda$ this integral is equal to $\frac{-1}{\lambda-\theta}e^{-(\lambda-\theta)x}\Big|_0^\infty = 1/(\lambda-\theta)$. But when $\theta \ge \lambda$, 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 $\theta_0 = \lambda$ and the corresponding interval $[0, \theta_0)$ is open.

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

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

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. That is $\theta_0 = \infty$.

In a retrospect it is not surprising that in this case $M(\theta)$ is finite for all θ. 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 θ the overall product is finite.

- **Poisson distribution.** Suppose X has a Poisson distribution with parameter λ. Then

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

(where we use the formula $\sum_{m\ge 0}\frac{t^m}{m!} = e^t$). Thus again the exp. moment generating function $M(\theta)$ is finite always. 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 $e^{\theta m}$.

Exercise 2. (a) Let X have a uniform distribution on an interval $[0, 1]$. Find $M(\theta)$ for all $\theta \ge 0$.

(b) Suppose X is a bounded random variable. Namely $a \leq X \leq b$ for some finite constants $a < b$. Prove that $M(\theta)$ is finite for *all* θ.

3.5. Legendre transforms

Theorem 3.4 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 3.3. A Legendre transform of a random variable X is the function $l(a) \triangleq \sup_\theta(\theta a - \log M(\theta))$.

Proposition 2. (a) $l(a)$ is a convex function of a.

(b) Suppose $a > \mu$. Then $l(a) = \sup_{\theta \geq 0}(\theta a - \log M(\theta))$ (no need to consider negative θ).

(c) Consider θ_0 – the largest θ for which $M(\theta) < \infty$. There exists $0 < \theta^* < \theta_0$ at which value $l(a)$ is achieved: $l(a) = a\theta^* - \log M(\theta^*)$.

Proof. (a) Note that, for every *fixed* θ, the function $\theta a - \log M(\theta)$, as a function of a is linear. Therefore $l(a)$ is a supremum of linear functions. As such it is convex.

(b) Since $e^{\theta z}$ is a convex function then using Jensen's inequality $M(\theta) = \mathbb{E}[e^{\theta X_1}] \geq e^{\theta E[X_1]} = e^{\theta\mu}$. When $\theta < 0$ we have $e^{\theta\mu} \geq e^{\theta a}$, implying $\log M(\theta) \geq \theta a$ or $a\theta - \log M(\theta) < 0$. But when $\theta = 0$ this difference is $0 - M(0) = 0$. So, indeed, there is no reason to consider negative θ when computing $l(a)$.

(c) This part is more difficult and technical. We skip the proof.

□

Armed with the results of this proposition, we can restate Theorem 3.4 in terms of θ^*:

Theorem 3.4 (Chernoff bound). *Given an i.i.d. sequence $X_1, \ldots, X_n$ suppose $\mathbb{E}[e^{\theta X_1}]$ is finite for all θ in some interval $[0, \theta_0)$. Let $a > \mu = \mathbb{E}[X_1]$. Then*

$$\mathbb{P}(\frac{\sum_{1 \leq i \leq n} X_i}{n} > a) \leq (\frac{\mathbb{E}[e^{\theta^* X_1}]}{e^{\theta^* a}})^n = e^{-l(a)n},$$

where θ^* solves $a\theta^* - \log M(\theta^*) = \sup_\theta(a\theta - \log M(\theta))$.

3.6. Additional reading materials

- Course packet. Section 1.1 and Section 1.2 up to including page 15.
- Chapter 0 of [**1**]. This is non-technical introduction to the field which describes motivation and various applications of the large deviations theory. Soft reading.

BIBLIOGRAPHY

1. A. Shwartz and A. Weiss, *Large deviations for performance analysis*, Chapman and Hall, 1995.