Introduction to Inequalities, Law of Large Number, and Large Deviation Theory

John C.S. Lui
Department of Computer Science & Engineering
The Chinese University of Hong Kong

Outline

Introduction to Inequality

Convex Function

- Def: A function $h(x)$,where $x \in R^n$, is said to be convex if

$$h(\alpha x_1 + (1-\alpha)x_2) \leq \alpha h(x_1) + (1-\alpha)h(x_2)$$

  h is *concave* if $-h$ is convex.

- For $x \in R$ and h has a second derivative, then it is convex if

$$h^{(2)}(x) \geq 0 \qquad \forall x$$

- If h is defined on the integers, $x \in N$,

$$h(x+1) + h(x-1) - 2h(x) \geq 0 \qquad \text{for} \quad x \in N$$

Introduction to inequality

Jensen's Inequality

- Suppose that h is a differentiable convex function defined on R, then

$$E[h(X)] \geq h(E[X])$$

- Useful way to remember Jensen's Inequality

$$E[X^2] \geq (E[X])^2$$

- Why convex? Because variance ≥ 0, $E[X^2] - (E[X])^2 \geq \phi$.
- Generally, X^{2n} is a convex function, therefore

$$E[X^{2n}] \geq (E[X])^{2n}$$

Introduction to inequality

Momemt Generatin Function

- We learn of transform before, for example, Laplace transform and $Z-$transform.
- Assume X is a continuous random variable, the Laplace transform of X is $E[e^{-sX}] = \int e^{-sx} f_X(x)dx$.
- For moment generating function:

$$M_X(\theta) = E[e^{\theta X}]$$

Since exponential is a convex function, we have:

$$E[e^{\theta X}] \geq e^{\theta E[X]}$$

Introduction to inequality

Jensen's Inequality for concave function

- If h is convex, $g(x) = -h(x)$ is concave, we have

$$E[h(X)] \geq h(E[X]) \Rightarrow E[-h(X)] \leq -h(E[X]).$$

- Therefore,

$$E[g(X)] \leq g(E[X]).$$

- Example:

$$E[min\{X_1, X_2, \cdots X_n\}] \leq \min\{E[X_1], E[X_2], \cdots E[X_n]\}$$

Introduction to inequality

Simple Markov Inequality:

- If X is a non-negative random variable, we have:

$$E[X] = \int x f_X(x) dx.$$

- If N is a discrete non-negative random varaible, we have:

$$E[N] = \sum n \text{Prob}[N = n]$$

- Another way to express $E[X]$, where X is a non-negative R.V. is:

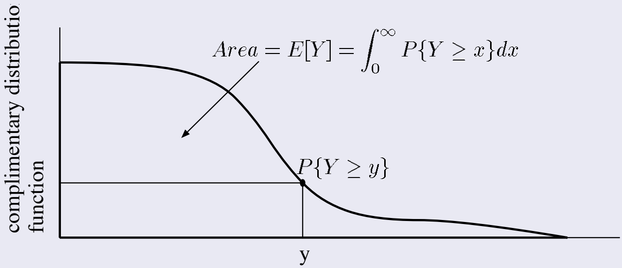
$$E[X] = \int (1 - F_X(x))\, dx.$$

Introduction to inequality

Simple Markov Inequality: continue

Assume Y is a non-negative random variable, "*Simple Markov Inequality*" states that

$$\text{Prob}[Y \geq y] \leq \frac{E[Y]}{y}.$$



Generalized Markov Inequality

Let h be a *nonnegative, nondecreasing* function and let X be a random variable.

$$
\begin{aligned}
E[h(X)] &= \int_{z=-\infty}^{\infty} h(z) f_X(z) dz \\
E[h(X)] &= \int_{z=-\infty}^{\infty} h(z) f_X(z) dz \geq \int_{t}^{\infty} h(z) f_X(z) dz \\
&\geq h(t) \underbrace{\int_{t}^{\infty} f_X(z) dz}_{P[X \geq t]}
\end{aligned}
$$

$$P[X \geq t] \leq \frac{E[h(X)]}{h(t)}$$

- Example : Let $h(x) = (x)^+$
- By Markov inequality ,

$$P[X > t] \leq \frac{E[X^+]}{t}$$

- We can use this result to estimate tail distribution! If expected response time of a job is $E[X] = 1$ sec

$$\text{Prob[response time } \geq 10 \text{ sec]} = P[X \geq 10] \leq \frac{E[X]}{10} \leq \frac{1}{10} = 0.1$$

$\Rightarrow$ at most 10% of the response time is greater than 10 sec.

Chebyshev's inequality

(2^{nd} order inequality assuming δ_X^2 is known)

$$\begin{aligned}
Y &= (X - E[X])^2 \ \text{ and } \ h(x) = x \\
P[Y \geq t^2] &\leq \frac{E[Y]}{t^2} \ \text{(simple Markov's Inequality)} \\
P[Y \geq t^2] &= P[(X - E[X])^2 \geq t^2] = P[\mid X - E[X] \mid \geq t] \\
\text{also} \quad E[Y] &= E[(X - E[X])^2] = \sigma_X^2 \\
P[\mid X - E[X] \mid \geq t] &\leq \frac{\sigma_X^2}{t^2}
\end{aligned}$$

(It provides intuition about the meaning of the variance of a r.v. since it shows that wide dispersions from the mean ($E[X]$) are unlikely if σ_X^2 is small.)
ex: $t = c\sigma_X$ where σ_X is the standard deviation
$P[\mid X - E[X] \mid] \geq c\sigma_X] \leq \frac{1}{c^2}$

Chernoff's Bound

- Assume we know the moment generating functions
- Let $h(x) = e^{\theta X}$ for $\theta \geq 0$

$$
\begin{aligned}
P[X \geq t] &\leq \frac{E[h(X)]}{h(t)} = M_X(\theta)e^{-\theta t} \\
P[X \geq t] &\leq \inf_{\theta \geq \phi} e^{-\theta t} M_X(\theta)
\end{aligned}
$$

- intuitively, this provides tighter bound than Markov and Chebyshev because we need higher moments.

Application

- Application : Let $Y_i, i = 1, 2, \cdots$ be independent Bernoulli r.v. with parameter $\frac{1}{2}$

$$X_n = Y_1 + Y_2 + \cdots + Y_n$$

be the total no. of heads obtained in n tosses.

$$\begin{aligned} E[Y_i] &= \phi(\frac{1}{2}) + 1(\frac{1}{2}) = \frac{1}{2} \\ Var(Y_i) &= E[Y^2] - E^2[Y] = \frac{1}{2} - \frac{1}{4} = \frac{1}{4} \end{aligned}$$

Application: continue

- Moment generating function of Y

$$
\begin{aligned}
E[e^{\theta Y}] &= e^{\theta \phi}(\frac{1}{2}) + e^{\theta(1)}(\frac{1}{2}) = \frac{1+e^{\theta}}{2} \\
E[X_n] &= E[Y_1 + \ldots Y_n] = E[Y_1] + \ldots E[Y_n] = \frac{n}{2} \\
Var[X_n] &= Var[Y_1 + \ldots Y_n] = Var[Y_1] + \ldots + Var[Y_n] \\
&= \frac{n}{4} \quad \text{(due to independence of } Y_i\text{)}
\end{aligned}
$$

- Moment generating function: $E[e^{\theta X_n}] = (E[e^{\theta Y_i}])^n = (\frac{1+e^{\theta}}{2})^n$

Application: continue

- $\alpha > \frac{1}{2}$, consider $P[X_n \geq \alpha n]$
- by Markov's Inequality , $E[X \geq t] \leq \frac{E[X]}{t}$

$$P[X_n \geq \alpha n] \leq \frac{\frac{n}{2}}{\alpha n} = \frac{1}{2\alpha}$$

- Chebyshev's inequality: Let $\alpha n = \frac{n}{2} + (\alpha - \frac{1}{2})n$. Observe

$$P[X_n \geq \alpha n] = P[X_n - \frac{n}{2} \geq (\alpha - \frac{1}{2})n]$$

$$P[X_n - \frac{n}{2} \geq (\alpha - \frac{1}{2})n] \leq P[| X_n - \frac{n}{2} | \geq (\alpha - \frac{1}{2})n] \leq \frac{\frac{n}{4}}{[(\alpha - \frac{1}{2})n]^2}$$

$$P[X_n - \frac{n}{2} \geq (\alpha - \frac{1}{2})n] \leq \frac{1}{4n(\alpha - \frac{1}{2})^2}$$

- Note: this is also equal to $P[X_n \geq \alpha n] \leq 1/\left(4n(\alpha - \frac{1}{2})^2\right)$

Application: continue:

- for Chernoff's bound

$$P[X_n \geq \alpha n] \leq \inf_{\theta \geq 0} e^{-\theta \alpha n} [\frac{1+e^{\theta}}{2}]^n \quad (*)$$

- To find the optimal θ^*, we perform

$$\begin{aligned} \frac{d}{d\theta}[e^{-\theta\alpha n}(\frac{1+e^{\theta}}{2})^n] &= \phi \\ \theta^* &= \ln[\frac{\alpha}{1-\alpha}] \end{aligned}$$

- Substitute θ^* into the expression $(*)$

$$P[X_n \geq \alpha n] \leq \frac{[\frac{1}{(2(1-\alpha))}]^n}{[\frac{\alpha}{(1-\alpha)}]^{\alpha n}}$$

- Note: $P[X_n \geq \alpha n]$ is exponentially decreasing in n and it's a tighter bound

Application: continue

- $n = 100$

α	0.55	0.60	0.80
Markov's Inequality	0.90	0.83	0.62
Chebyshev's Inequality	0.1	0.025	0.002
Chernoff's Bound	0.006	1.8×10^{-9}	1.9×10^{-84}

Weak Law of Large Numbers

- Let $X_i, i = 1, 2, \cdots$ be i.i.d. r.v. with finite mean $E[X]$ and variance σ_X^2.

$$S_n = X_1 + X_2 + \ldots + X_n$$

- The statistical average of the first n experiments is $\frac{S_n}{n}$
- Intuition tells us that as $n \to \infty$, $E[\frac{S_n}{n}] \to E[X]$

$$\begin{aligned} E[\frac{S_n}{n}] &= E[\frac{X_1}{n} + \frac{X_2}{n} + \ldots + \frac{X_n}{n}] = \frac{E[X]}{n} + \ldots + \frac{E[X]}{n} \\ &= E[X] \\ Var[\frac{S_n}{n}] &= Var[\frac{X_1}{n} + \frac{X_2}{n} + \ldots + \frac{X_n}{n}] \\ &= \frac{1}{n^2} Var[X_1] + \ldots + \frac{1}{n^2} Var[X_n] = \frac{\sigma_X^2}{n} \end{aligned}$$

Using Cherbyshev's inequality

$$P\left[\left|\frac{S_n}{n} - E[X]\right| \geq \varepsilon\right] = \frac{Var[\frac{S_n}{n}]}{\varepsilon^2} = \frac{\sigma_X^2}{n\varepsilon^2}$$

- This says than as $n \to \infty$(or no. of experiment increases), it becomes less likely the "statistical average" differs from the mean $E[X]$

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - E[X]\right| \geq \varepsilon\right] = 0$$

Theory

We studied Chernoff's bound from Markov inequality. For a random variable X, Chernoff's bound implies

$$P[X \geq t] \leq \inf_{\theta \geq 0} e^{-\theta t} M_X(\theta) \tag{1}$$

where $M_X(\theta)$ is the *moment generating function* of X. Taking the log on both sides, we have

$$\begin{aligned} \ln P[X \geq t] &\leq \inf_{\theta \geq 0} \left(-\theta t + \ln M_X(\theta)\right) \\ &= -\sup_{\theta \geq 0} \left(\theta t - \ln M_X(\theta)\right). \end{aligned} \tag{2}$$

Define $I(t)$ as *large deviation rate function*:

$$I(t) = \sup_{\theta \geq 0} \left(\theta t - \ln M_X(\theta)\right). \tag{3}$$

Consider an example statistical average

Assume X_i are i.i.d, we have:

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{4}$$

The strong law of large number says that:

$$S_n \rightarrow E[X], \qquad \text{as } n \rightarrow \infty,$$

but provides no information about the *rate of convergence*. We are interested in the probability that S_n is larger than some value t, where $t \geq E[X]$. For large n, large deviation theory shows that:

$$P[S_n \geq t] = e^{-nI(t)+o(n)}, \qquad t \geq E[X], \tag{5}$$

or deviations away from the mean decrease *exponentially fast* with n at the rate of $-I(t)$.

Proof: the upper bound

We first observe that $M_{S_n}(\theta) = M_X^n(\theta/n)$, using the result of Chernoff's bound, we have

$$\begin{aligned} \ln P[S_n \geq t] &\leq -\sup_{\theta \geq 0}(\theta t - n \ln M_X(\theta/n)) \\ &= -n \sup_{\theta \geq 0}((\theta/n)t - \ln M_X(\theta/n)). \end{aligned} \quad (6)$$

In (6), we replace the "dummy" variable θ with $n\theta$. Doing this, dividing by n, we can rewrite (6) as:

$$\frac{1}{n} \ln P[S_n \geq t] \leq -I(t).$$

Since it holds for *all* n, it also holds for the limit supremum:

$$\lim_{n \to \infty} \sup \frac{1}{n} \ln P[S_n \geq t] \leq -I(t). \quad (7)$$

Proof: the lower bound

Suppose θ^* is the value obtained in the supremum of the rate function:

$$I(t) = \theta^* t - \ln M_X(\theta^*). \tag{8}$$

Define a new random variable (or the *twisted distribution*) Y with density function given by:

$$f_Y(z) = \frac{e^{\theta^* z} f_X(z)}{M_X(\theta^*)}. \tag{9}$$

One key feature of $f_Y(z)$ is that:

$$E[Y] = t.$$

Proof: the lower bound (cont)

To see this,

$$
\begin{aligned}
E[Y] &= \int_{z=-\infty}^{\infty} \frac{z e^{\theta^* z} f_X(z) dz}{M_X(\theta^*)} \\
&= \frac{1}{M_X(\theta^*)} \frac{d}{d\theta} \int_{z=-\infty}^{\infty} e^{\theta z} f_X(z) dz \bigg|_{\theta=\theta^*} \\
&= \frac{M_X'(\theta^*)}{M_X(\theta^*)} = \frac{d}{d\theta} \ln M_X(\theta) \bigg|_{\theta=\theta^*}
\end{aligned}
$$

From Equation (8), implies that

$$
\frac{d}{d\theta} \ln M_X(\theta) \bigg|_{\theta=\theta^*} = t.
$$

So $E[Y] = t$ as claimed.

Proof: the lower bound (cont)

To obtain a lower bound on $(1/n)\ln P[S_n \geq t]$, we can write

$$P[S_n \geq t] = \int_{nt \leq z_1 + \cdots + z_n} f_X(z_1) \cdots f_X(z_n) dz_1 \cdots dz_n.$$

Rewriting in terms of the density of Y (Eq. (9)) yields

$$P[S_n \geq t] = M_X^n(\theta^*) \int_{nt \leq z_1 + \cdots + z_n} e^{-\theta^*(z_1 + \cdots + z_n)} f_Y(z_1) \cdots f_Y(z_n) dz_1 \cdots dz_n.$$

Let ϵ be a positive constant that is used to restrct the range of the integral above, we have:

$$\begin{aligned} P[S_n \geq t] &\geq M_X^n(\theta^*) \int_{nt \leq z_1 + \cdots + z_n \leq n(t+\epsilon)} e^{-\theta^*(z_1 + \cdots + z_n)} f_Y(z_1) \cdots f_Y(z_n) dz_1 \cdots dz_n \\ &\geq M_X^n(\theta^*) e^{-\theta^* nt} \int_{nt \leq z_1 + \cdots + z_n \leq n(t+\epsilon)} f_Y(z_1) \cdots f_Y(z_n) dz_1 \cdots dz_n. \quad (10) \end{aligned}$$

Proof: the lower bound (cont)

Since $E[Y] = t$, the strong law of large number implies that the Equation (10) converges to 1 as $n \to \infty$. This is easy to show

$$\lim_{n\to\infty} \int_{t \leq \frac{z_1+\cdots+z_n}{n} \leq t+\epsilon} f_Y(z_1)\cdots f_Y(z_n) dz_1 \cdots dz_n = 1.$$

Taking the log of both side on (10) and dividing n implies

$$\lim_{n\to\infty} \inf \frac{1}{n} \ln P[S_n \geq t] \geq -I(t).$$

Combining with (7), we see that as $n \to \infty$, the upper and lower bounds converge, yields

$$\lim_{n\to\infty} \inf \frac{1}{n} \ln P[S_n \geq t] = -I(t) = -\theta^* t + M_X(\theta^*).$$

Large Deviation Bound for Exp. Random Variables

Let $X_i, i = 1, 2, \ldots,$ be independent, identically distributed exponential random variables with $E[X_i] = 1$.
The moment generating function of X is

$$M_X(\theta) = \int_{z=0}^{\infty} e^{\theta z} e^{z} dz = \frac{1}{1-\theta}. \tag{11}$$

To find θ^* the rate function (3), we use calculus and yield:

$$\frac{d}{d\theta}(\theta t - \ln M_X(\theta)) = t - \frac{M'_X(\theta)}{M_X(\theta)} = 0.$$

Since $M_X(\theta) = \frac{1}{1-\theta}$, substitute it to the above equation yields

$$\theta^* = \frac{t-1}{t}.$$

Large Deviation Bound for Exp. V.R (cont)

The large deviation rate function is $l(t) = \theta t - \ln M_X(\theta)$, substituting θ^*, we have:

$$l(t) = (t-1) - \ln t.$$

We can now find the tail distribution of S_n (with respect to the rate of convergence), or $P[S_n \geq t]$ for $t \geq E[X] = 1$:

$$\begin{aligned} P[S_n \geq t] &= e^{-nl(t)} \\ &= e^{-n(t-1)+n\ln t} \\ &= t^n e^{-n(t-1)} \qquad \text{for } t \geq E[X] = 1. \end{aligned}$$