Chapter 3

Bounds, Convergence, and Limit Theorems

Bounds on Probability

Markov's Inequality If X is a random variable that takes only nonnegative values, then for any $z > 0$,

$$P(X \geq z) \leq \frac{\mathrm{E}[X]}{z}.$$

Indeed, if X is a discrete random variable,

$$\begin{aligned} P(X \geq z) &= \sum_{x \geq z} p_X(x) \leq \sum_{x \geq z} \frac{x}{z} p_X(x) \\ &\leq \frac{1}{z} \sum_{x} x p_X(x) = \frac{\mathrm{E}[X]}{z}. \end{aligned}$$

If X is a continuous random variable,

$$\begin{aligned} P(X \geq z) &= \int_z^\infty f_X(x)dx \leq \int_z^\infty \frac{x}{z} f_X(x)dx \\ &\leq \frac{1}{z}\int_{-\infty}^\infty x f_X(x)dx = \frac{\mathrm{E}[X]}{z}. \end{aligned}$$

Chebyshev's Inequality If X is a random variable with mean m_X and variance K_X, then for any $z > 0$,

$$P(|X - m_X| \geq z) \leq \frac{K_X}{z^2}.$$

Indeed, since $(X - m_X)^2$ takes only nonnegative values, then by Markov's inequality with z replaced by z^2, we have

$$P((X - m_X)^2 \geq z^2) \leq \frac{\mathrm{E}[(X - m_X)^2]}{z^2} = \frac{K_X}{z^2}.$$

This is equivalent to

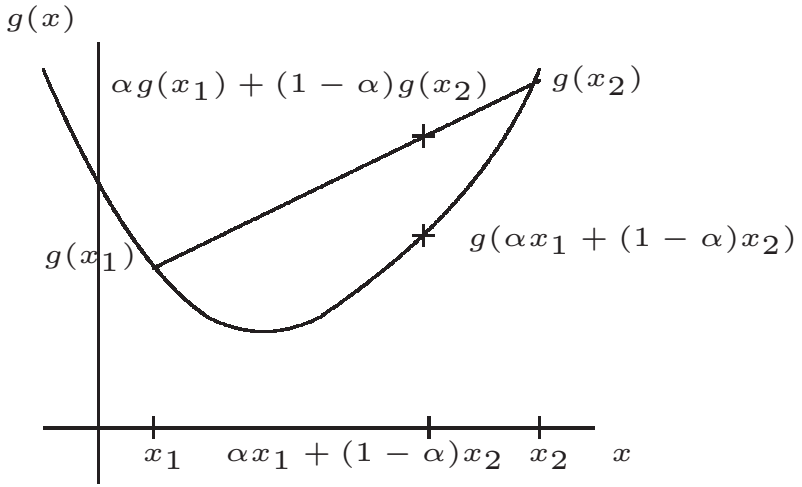
$$P(|X - m_X| \geq z) \leq \frac{K_X}{z^2}.$$

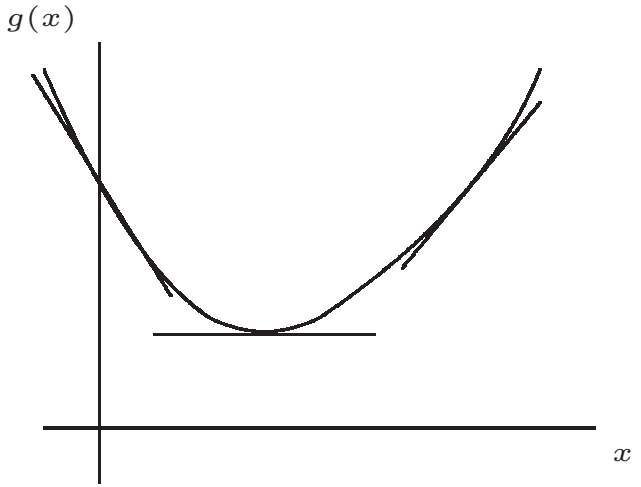
Bounds on Means

Recall that a function $g : \mathbb{R} \to \mathbb{R}$ is convex if for all α, $0 \leq \alpha \leq 1$, and all real x_1 and x_2,

$$g(\alpha x_1 + (1-\alpha)x_2) \leq \alpha g(x_1) + (1-\alpha)g(x_2).$$

Geometrically, this means that g lies above any of its tangents.



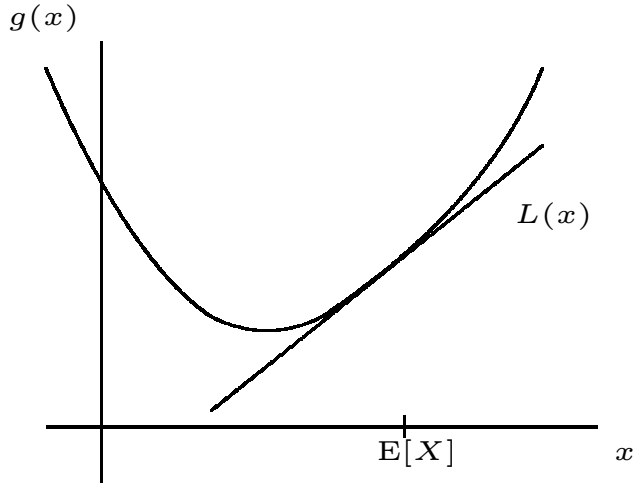


Jensen's Inequality Let $g : \mathbb{R} \to \mathbb{R}$ be a convex function and X be a random variable, then $g(\mathrm{E}[X]) \leq \mathrm{E}[g(X)]$.

Indeed, let $L(x) = ax + b$ be a line tangent to g at $\mathrm{E}[X]$. Since g is convex, it lies above L, i.e., $g(x) \geq L(x)$ for all x, which implies that $\mathrm{E}[g(X)] \geq \mathrm{E}[L(X)]$. We have

$$\mathrm{E}[L(X)] = \mathrm{E}[aX + b] = a\mathrm{E}[X] + b = L(\mathrm{E}[X]) = g(\mathrm{E}[X]),$$

where the last equality follows as L touches g at $\mathrm{E}[X]$.



Cauchy-Schwartz Inequality Let X and Y be random variables with finite variances. Then,

$$|\mathrm{E}[XY]| \leq \sqrt{\mathrm{E}[X^2]\mathrm{E}[Y^2]}.$$

The proof is the same as that used to prove that $|K_{X,Y}| \leq \sqrt{K_X K_Y}$ in Chapter 2 by considering the nonnegativity of $\mathrm{E}[(aX - Y)^2] = a^2\mathrm{E}[X^2] - 2a\mathrm{E}[XY] + \mathrm{E}[Y^2]$ for all real a which implies that the discriminant $4(\mathrm{E}[XY])^2 - 4\mathrm{E}[X^2]\mathrm{E}[Y^2]$ is nonpositive.

Convergence of Random Variables

Given a sequence $x_1, x_2, \ldots$ of real numbers, we say that it converges to some real number x if for every $\epsilon > 0$, there exists a positive integer n_ϵ such that $|x_n - x| < \epsilon$ for all $n \geq n_\epsilon$.

Let $X, X_1, X_2, \ldots$ be random variables defined on the probability space $(\Omega, \mathcal{F}, P)$.

The sequence $X_1, X_2, \ldots$ converges to X *surely*, denoted by $X_n \xrightarrow{s} X$, if for every $\omega \in \Omega$,

$$X_n(\omega) \to X(\omega).$$

The sequence $X_1, X_2, \ldots$ converges to X *almost surely* or *with probability* 1, denoted by $X_n \xrightarrow{a.s.} X$, if

$$P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1.$$

The sequence $X_1, X_2, \ldots$ converges to X *in the mean square sense*, denoted by $X_n \xrightarrow{m.s.} X$, if

$$\lim_{n\to\infty} \mathrm{E}[(X_n - X)^2] = 0.$$

The sequence $X_1, X_2, \ldots$ converges to X *in probability*, denoted by $X_n \xrightarrow{p} X$, if for every $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

The sequence $X_1, X_2, \ldots$ converges to X in *distribution* or *weakly*, denoted by $X_n \xrightarrow{d} X$, if for every x for which $F_X(x)$ is continuous,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Example 1 Consider a probability space in which the sample space is $[0, 1]$ and the probability of the event that the outcome ω belongs to the interval $[0, a]$ is a, where $0 \leq a \leq 1$. Let $X_1, X_2, \ldots$ be a sequence of random variables defined by

$$X_n(\omega) = \begin{cases} n^\alpha & \omega \in [0, 1/n] \\ 0 & \text{otherwise,} \end{cases}$$

where α is a real number. Thus,

$$X_n = \begin{cases} n^\alpha & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 1/n. \end{cases}$$

If $\alpha < 0$, then for all ω, $X_n(\omega) \to 0$ as $n \to \infty$ and the sequence converges surely and almost surely to the constant random variable 0. However, if $\alpha \geq 0$, then $X_n(\omega) \to 0$ for all $\omega \neq 0$ but not for $\omega = 0$. Therefore, if $\alpha \geq 0$, the sequence does not converge surely but it converges almost surely to 0.

We have

$$\mathrm{E}[(X_n - 0)^2] = n^{2\alpha} \times 1/n + 0^2 \times (1 - 1/n) = n^{2\alpha - 1}.$$

Hence, the sequence converges in mean square sense if and only if $\alpha < 1/2$ and in this case it converges to 0.

We have for any $\epsilon > 0$,

$$P(|X_n - 0| > \epsilon) \leq P(X_n \neq 0) = 1/n.$$

Hence, the sequence converges to 0 in probability.

We have

$$F_{X_n}(x) = \begin{cases} 0 & x < 0 \\ 1 - \frac{1}{n} & 0 \leq x < n^\alpha \\ 1 & x \geq n^\alpha, \end{cases}$$

which converges, as $n \to \infty$, to

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0. \end{cases}$$

Hence, the sequence converges to 0 in distribution.

It can be shown that[a]:

- Sure convergence implies all other types of convergence.
- Almost sure convergence implies convergence in probability.
- Convergence in mean square sense implies convergence in probability.
- Convergence in probability implies convergence in distribution.

No other general implications can be added to this list.

[a] See Section 3.3 in textbook.

Laws of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each with mean m. Let $\hat{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Then, $\hat{X}_n$ is a random variable which is called the *sampled mean*.

Weak Law of Large Numbers $\hat{X}_n \xrightarrow{p} m$.

Strong Law of Large Numbers $\hat{X}_n \xrightarrow{a.s.} m$ provided that $\mathrm{E}[|X_1|] < \infty$.

We give a proof of the weak law in the special case in which the variance of the random variables, K, is finite. Then, $\mathrm{E}[\hat{X}_n] = m$ and $K_{\hat{X}_n} = K/n$. The result follows from Chebyshev's inequality which implies that for any $\epsilon > 0$,

$$P(|\hat{X}_n - m| \geq \epsilon) \leq \frac{K_{\hat{X}_n}}{\epsilon^2} = \frac{K}{\epsilon^2 n} \to 0 \text{ as } n \to \infty.$$

We give a proof of the strong law in the special case in which the fourth moment of the random variables, $\mathrm{E}[X_n^4]$, is finite and we denote it by m_4. Let $m_2 = \mathrm{E}[X_n^2]$. Then, $m_4 \geq m_2^2$ since

$$m_4 - m_2^2 = \mathrm{E}[X_n^4] - (\mathrm{E}[X_n^2])^2 = K_{X_n^2} \geq 0.$$

It suffices to give a proof in case $m = 0$ otherwise we replace X_n by $X_n - m$ in the proof.

Notice that $\hat{X}_n^4$ is $1/n^4$ times the sum of terms of the form X_i^4, $X_i^2 X_j^2$, $X_i^3 X_j$, $X_i^2 X_j X_k$, and $X_i X_j X_k X_l$, where i, j, k, l are distinct. Since the random variables have zero means and are independent, the last three terms have zero means. For each i, the term X_i^4 appears once and for each pair of i and j, $i \neq j$, the term $X_i^2 X_j^2$ appears six times.

$$
\begin{aligned}
\mathrm{E}[\hat{X}_n^4] &= \frac{1}{n^4}\left(\sum_{i=1}^{n}\mathrm{E}[X_i^4] + \sum_{1\le i<j\le n} 6\ \mathrm{E}[X_i^2 X_j^2]\right) \\
&= \frac{1}{n^4}\left((nm_4 + 6\binom{n}{2}m_2^2\right) \quad \text{(because of independence)} \\
&= \frac{1}{n^3}\left(m_4 + 3(n-1)m_2^2\right) < \frac{3m_4}{n^2} \quad (\text{as } m_2^2 \le m_4).
\end{aligned}
$$

Hence, $\sum_{n=1}^{\infty}\mathrm{E}[\hat{X}_n^4] < 3m_4\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty$.

By Markov's inequality, $P(\hat{X}_n^4 \ge \epsilon) \le \mathrm{E}[\hat{X}_n^4]/\epsilon$ for any $\epsilon > 0$. Hence, $\sum_{n=1}^{\infty} P(\hat{X}_n^4 \ge \epsilon) < \infty$ and, by the Borel-Cantelli Lemma, the probability that $\hat{X}_n^4 \ge \epsilon$ infinitely often is zero. Thus, $\hat{X}_n^4$ and, consequently, $\hat{X}_n$ converge almost surely to 0.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables each having mean m and finite variance K. For $n = 1, 2, \ldots$, let

$$Y_n = \frac{\sum_{k=1}^{n}(X_k - m)}{\sqrt{nK}} = \frac{S_n - nm}{\sqrt{nK}} = \sqrt{\frac{n}{K}}(\hat{X}_n - m),$$

where $S_n = X_1 + X_2 + \cdots + X_n$ and
$\hat{X}_n = (X_1 + X_2 + \cdots + X_n)/n$ are the sum and the sampled mean of the random variables $X_1, X_2, \ldots, X_n$, respectively.

The Central Limit Theorem states that $Y_n \xrightarrow{d} Y$ where Y is Gaussian$(0, 1)$, i.e., for any y,

$$P(Y_n \leq y) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{x^2}{2}}\, dx = \varphi(y)$$

as $n \to \infty$.

It suffices to show this in case $m = 0$ and $K = 1$ otherwise we can replace X_k by $(X_k - m)/\sqrt{K}$ in the proof. Let $\Phi(\omega)$ be the characteristic function of X_k. Then, $\Phi(\omega/\sqrt{n})$ is the characteristic function of $X_k/\sqrt{n}$. The characteristic function of $Y_n = (X_1 + X_2 + \cdots + X_n)/\sqrt{n}$ is then $(\Phi(\omega/\sqrt{n}))^n$ while that of Y which is Gaussian$(0, 1)$ is $e^{-\omega^2/2}$. We will show that

$$(\Phi(\omega/\sqrt{n}))^n \to e^{-\omega^2/2} \qquad \text{as } n \to \infty$$

or, equivalently,

$$nL(\omega/\sqrt{n}) \to -\omega^2/2 \qquad \text{as } n \to \infty,$$

where $L(\omega) = \ln \Phi(\omega)$.

From the definition of $\Phi(\omega) = \mathrm{E}[e^{\jmath\omega X_k}]$, we have $\Phi(0) = 1$, $\Phi'(0) = \mathrm{E}[\jmath X_k] = 0$, and $\Phi''(0) = \mathrm{E}[(\jmath X_k)^2] = -1$. Hence, $L(0) = 0$, $L'(0) = \Phi'(0)/\Phi(0) = 0$, and

$$L''(0) = \frac{\Phi(0)\Phi''(0) - (\Phi'(0))^2}{(\Phi(0))^2} = -1.$$

The Maclaurin series for $L(\omega)$ is

$$\begin{aligned} L(\omega) &= L(0) + L'(0)\omega + L''(0)\omega^2/2! + O(\omega^3) \\ &= -\omega^2/2 + O(\omega^3). \end{aligned}$$

Hence,

$$nL(\omega/\sqrt{n}) = -n \times (\omega/\sqrt{n})^2/2 + O(n(\omega/\sqrt{n})^3).$$

As $n \to \infty$, we have $nL(\omega/\sqrt{n}) \to -\omega^2/2$.

Example 2 A fair coin is tossed 100 times. Estimate the probability that the number of heads, S_{100}, is at least 40 and at most 60.

The result of each toss is Bernoulli(0.5) with mean 0.5 and variance 0.25 and S_{100} is the sum of these independent random variables.

$$\begin{aligned} P(40 \le S_{100} \le 60) &= P(-10 \le S_{100} - 100 \times 0.5 \le 10) \\ &= P(-2 \le \frac{S_{100} - 50}{\sqrt{100 \times 0.25}} \le 2) \\ &\approx \varphi(2) - \varphi(-2) \\ &\approx 0.9772 - 0.0227 = 0.9545. \end{aligned}$$

The exact value is

$$P(40 \le S_{100} \le 60) = \sum_{i=40}^{60} \binom{100}{i} (0.5)^i (1-0.5)^{100-i} = 0.9648$$

up to four decimal digits.

Large Deviations and Rate Function

The Central Limit Theorem estimates the probability that $P(S_n > nz)$, where S_n is the sum of n independent and identically distributed random variables $X_1, X_2, \ldots, X_n$, by

$$1 - \varphi\left(\sqrt{\frac{n}{K_{X_1}}}(z - m_{X_1})\right) = Q\left(\sqrt{\frac{n}{K_{X_1}}}(z - m_{X_1})\right).$$

Using the bounds on the Q-function, we get an estimate of $\frac{1}{n}\ln P(S_n > nz)$ that equals $-(z - m_{X_1})^2/2K_{X_1}$.

This estimate is poor if the deviation of z from m_{X_1} is large and may not be asymptotically correct as $n \to \infty$.

Large deviations are dealt with using different techniques as shown next.

Chernoff Bound If X is a random variable with moment generating function $M_X(s)$, then for every real $s \geq 0$

$$P(X \geq z) \leq \frac{M_X(s)}{e^{sz}}. \tag{1}$$

Indeed, the result holds trivially for $s = 0$. Assume $s > 0$, by replacing X by e^{sX} and z by e^{sz} in Markov's inequality, we obtain

$$P(X \geq z) = P(e^{sX} \geq e^{sz}) \leq \frac{\mathrm{E}[e^{sX}]}{e^{sz}} = \frac{M_X(s)}{e^{sz}}.$$

Let

$$I_X(z) = \sup_{s \geq 0}(sz - \ln M_X(s)).$$

Then,

$$P(X > z) \leq e^{-I_X(z)}$$

or equivelently

$$\ln P(X > z) \leq -I_X(z).$$

The function $I_X(z)$ is called the *rate function* of X or the *Legendre transform* of $\ln M_X(s)$.

Properties of the Rate Function:

(1) $I_X(z)$ is convex. Indeed, for $0 \leq \alpha \leq 1$, we have

$$\begin{aligned} I_X(\alpha z_1 + (1-\alpha)z_2) &= \sup_{s \geq 0}(s(\alpha z_1 + (1-\alpha)z_2) - \ln M_X(s)) \\ &\leq \alpha \sup_{s \geq 0}(sz_1 - \ln M_X(s)) \\ &\qquad +(1-\alpha)\sup_{s \geq 0}(sz_2 - \ln M_X(s)) \\ &= \alpha I_X(z_1) + (1-\alpha) I_X(z_2). \end{aligned}$$

(2) $I_X(z) \geq 0$. Indeed, since $M_X(0) = 1$, then

$$I_X(z) = \sup_{s \geq 0}(sz - \ln M_X(s)) \geq 0 \times z - \ln M_X(0) = 0.$$

(3) $I_X(m_X) = 0$. Indeed, by Jensen's inequality,

$$M_X(s) = \mathrm{E}[e^{sX}] \geq e^{\mathrm{E}[sX]} = e^{sm_X}.$$

In particular,

$$I_X(m_X) = \sup_{s \geq 0}(sm_X - \ln M_X(s)) \leq 0$$

and the result follows from the above property.

Example 3 Let X be Bernoulli$(1/2)$. The rate function of X is

$$I_X(z) = \sup_{s \geq 0}(sz - \ln M_X(s)).$$

where $M_X(s) = \mathrm{E}[e^{sX}] = \frac{1}{2} + \frac{1}{2}e^s$. Let

$$g_z(s) = sz - \ln M_X(s) = sz - \ln \frac{1}{2}(1 + e^s).$$

Then, $I_X(z)$ is the supremum of $g_z(s)$ over all $s \geq 0$.

We have

$$g'_z(s) = z - \frac{e^s}{1+e^s} = z - \frac{1}{e^{-s}+1}.$$

Hence, $g'_z(s) = 0$ if and only if $s = \ln \frac{z}{1-z}$. Therefore, for $1/2 < z < 1$, this value of s is positive and

$$\begin{aligned} I_X(z) &= sz - \ln \frac{1}{2}(1+e^s)\Big|_{s=\ln \frac{z}{1-z}} \\ &= z \ln \frac{z}{1-z} - \ln \frac{1}{2(1-z)} \\ &= z \ln z + (1-z)\ln(1-z) + \ln 2. \end{aligned}$$

The rate function of the sampled mean Let $X_1, X_2, \ldots X_n$ be independent and identically distributed random variables and $\hat{X}_n = (X_1 + X_2 + \cdots + X_n)/n$. Then,

$$\begin{aligned} M_{\hat{X}_n}(s) &= \mathrm{E}[e^{s\hat{X}_n}] = \mathrm{E}[e^{s(X_1+X_2+\cdots+X_n)/n}] \\ &= \mathrm{E}[e^{\frac{s}{n}X_1}]\ \mathrm{E}[e^{\frac{s}{n}X_2}]\ \cdots\ \mathrm{E}[e^{\frac{s}{n}X_n}] \\ &= M_{X_1}\left(\frac{s}{n}\right) M_{X_2}\left(\frac{s}{n}\right) \cdots M_{X_n}\left(\frac{s}{n}\right) \\ &= \left(M_{X_1}\left(\frac{s}{n}\right)\right)^n. \end{aligned}$$

The rate function of $\hat{X}_n$ is then

$$\begin{aligned} I_{\hat{X}_n}(z) &= \sup_{s\geq 0}(sz - \ln M_{\hat{X}_n}(s)) \\ &= \sup_{s\geq 0}\left(sz - n\ln M_{X_1}\left(\frac{s}{n}\right)\right) \\ &= n\sup_{s\geq 0}(sz - \ln M_{X_1}(s)) = nI_{X_1}(z) \end{aligned}$$

as we replaced s by ns in the last supremum. Chernoff bound then implies that

$$\frac{1}{n}\ln P(\hat{X}_n > z) \leq -I_{X_1}(z)$$

or, equivalently,

$$\frac{1}{n}\ln P(S_n > nz) \leq -I_{X_1}(z),$$

where $S_n = X_1 + X_2 + \cdots + X_n$.

Example 4 Let $X_1, X_2, \ldots$ be independent Bernoulli$(1/2)$ random variables. Then, for $1/2 < z < 1$,

$$P(\hat{X}_n > z) \leq e^{-nI_{X_1}(z)},$$

where from Example 3,

$$I_{X_1}(z) = z \ln z + (1-z) \ln(1-z) + \ln 2.$$

Since $\hat{X}_n = S_n/n$, where $S_n = X_1 + X_2 + \cdots + X_n$, we have

$$P(S_n > nz) \leq e^{-nI_{X_1}(z)}.$$

We would like to see how tight is this upper bound on $P(S_n > nz)$ for $1/2 < z < 1$. Notice that S_n is Binomial$(n, 1/2)$. We have

$$P(S_n > nz) = \frac{1}{2^n} \sum_{k=a_n}^{n} \binom{n}{k},$$

where $a_n = \lfloor nz \rfloor + 1$. Since $\binom{n}{k}$ is decreasing for $n/2 \leq k \leq n$, we have

$$\frac{1}{2^n} \binom{n}{a_n} \leq P(S_n > nz) \leq \frac{1}{2^n} (n+1) \binom{n}{a_n},$$

i.e.,

$$\ln \binom{n}{a_n} - n \ln 2 \leq \ln P(S_n > nz) \leq \ln(n+1) + \ln \binom{n}{a_n} - n \ln 2.$$

Hence,

$$\lim_{n \to \infty} \frac{1}{n} \ln P(S_n > nz) = \lim_{n \to \infty} \frac{1}{n} \ln \binom{n}{a_n} - \ln 2.$$

By Stirling's formula,

$$\lim_{n\to\infty} \frac{\sqrt{2\pi n} n^n e^{-n}}{n!} = 1.$$

Hence,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \ln \binom{n}{a_n} &= \lim_{n\to\infty} \frac{1}{n} \ln \frac{n!}{a_n!(n-a_n)!} \\
&= \lim_{n\to\infty} \frac{1}{n} \ln \frac{n^n}{a_n^{a_n}(n-a_n)^{n-a_n}} \\
&= \lim_{n\to\infty} \left(\ln n - \frac{a_n}{n} \ln a_n - \frac{n-a_n}{n} \ln(n-a_n) \right).
\end{aligned}$$

As $a_n = \lfloor nz \rfloor + 1$, we get

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \ln \binom{n}{a_n} \\
&\quad = \lim_{n\to\infty} (\ln n - z(\ln z + \ln n) - (1-z)(\ln(1-z) + \ln n)) \\
&\quad = -\left(z \ln z + (1-z) \ln(1-z)\right).
\end{aligned}$$

Hence, for $1/2 < z < 1$,

$$\lim_{n\to\infty} \frac{1}{n} \ln P(S_n > nz) = -\left(z \ln z + (1-z) \ln(1-z)\right) - \ln 2.$$

i.e., $P(S_n > nz) \approx e^{-nI_{X_1}(z)}$ for large n and the bound is tight. Notice that the Central Limit Theorem gives the estimate $-2(z - \frac{1}{2})^2$ of the limit, which is not correct.

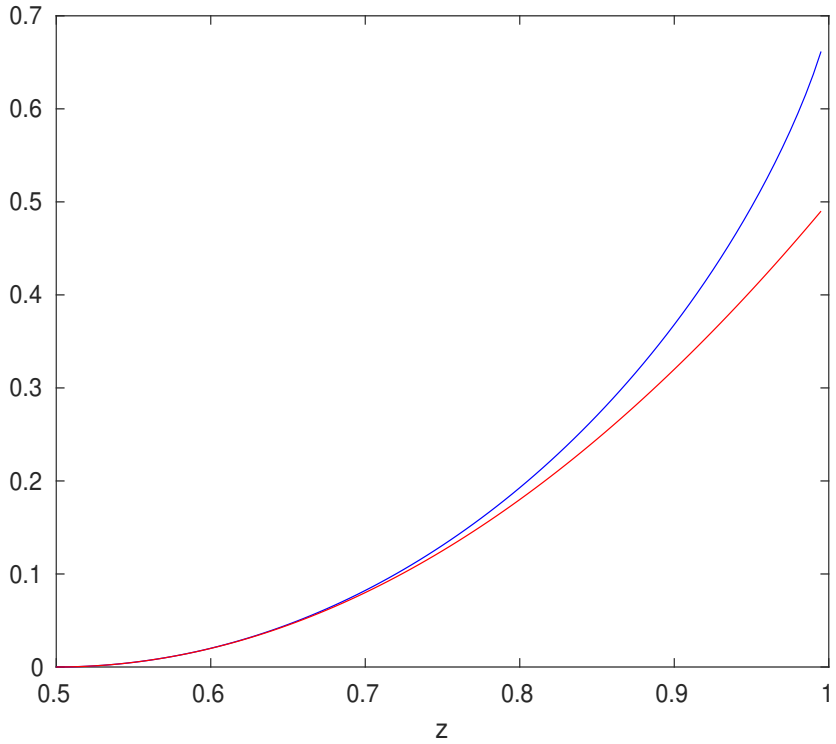


Figure 1: $(z \ln z + (1-z)\ln(1-z)) + \ln 2$ (blue) and $2(z-\frac{1}{2})^2$ (red) as a function of z.

Actually, Cramér's Theorem shows that the asymptotic tightness of the bound based on the rate function is not specific to this example.

Cramér's Theorem Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then, for $z > \mathrm{E}[X_1]$,

$$\lim_{n\to\infty} \frac{1}{n} \ln P(S_n > nz) = -I_{X_1}(z).$$

The proof is rather technical and will be skipped.

Hoeffding's Inequalities

Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ with probability 1 for each $i = 1, 2, \ldots, n$. Let $S_n = \sum_{i=1}^{n} X_i$. Then, for any $z > 0$,

$$P(S_n - \mathrm{E}[S_n] \geq nz) \leq e^{-2n^2z^2/\sum_{i=1}^{n}(a_i-b_i)^2}, \quad (2)$$

$$P(S_n - \mathrm{E}[S_n] \leq -nz) \leq e^{-2n^2z^2/\sum_{i=1}^{n}(a_i-b_i)^2}, \quad (3)$$

$$P(|S_n - \mathrm{E}[S_n]| \geq nz) \leq 2e^{-2n^2z^2/\sum_{i=1}^{n}(a_i-b_i)^2}, \quad (4)$$

$$P(|\hat{X}_n - \mathrm{E}[\hat{X}_n]| \geq z) \leq 2e^{-2n^2z^2/\sum_{i=1}^{n}(a_i-b_i)^2}. \quad (5)$$

The proof of Hoeffding's inequalities is based on the following: Let Z be a random variable such that $\mathrm{E}[Z] = 0$ and $a \leq Z \leq b$ with probability 1. Then for all $s \geq 0$,

$$M_Z(s) = \mathrm{E}[e^{sZ}] \leq e^{(a-b)^2 s^2/8}. \tag{6}$$

To prove (6), notice that since the function $g(x) = e^{sx}$ is convex, it follows from the definition of convex functions that for all $0 \leq \alpha \leq 1$,

$$g(\alpha x_1 + (1-\alpha)x_2) \leq \alpha g(x_1) + (1-\alpha)g(x_2).$$

Let $a \leq x \leq b$. With $x_1 = a$, $x_2 = b$, $\alpha = (b-x)/(b-a)$, which implies that $1 - \alpha = (x-a)/(b-a)$ and $\alpha x_1 + (1-\alpha)x_2 = x$, we get

$$e^{sx} \leq \frac{b-x}{b-a} e^{sa} + \frac{x-a}{b-a} e^{sb}.$$

Replacing x by Z and taking expectation, we get

$$\begin{aligned} \mathrm{E}[e^{sZ}] &\leq \mathrm{E}\Big[\frac{b-Z}{b-a}e^{sa} + \frac{Z-a}{b-a}e^{sb}\Big] \\ &\overset{\text{(a)}}{=} \frac{b-\mathrm{E}[Z]}{b-a}e^{sa} + \frac{\mathrm{E}[Z]-a}{b-a}e^{sb} \\ &\overset{\text{(b)}}{=} \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}, \end{aligned} \tag{7}$$

where in (a) we used the fact that expectation is linear and in (b) the assumption that $\mathrm{E}[Z] = 0$.

Define

$$f(s) = \ln\left(\frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}\right). \tag{8}$$

We have

$$\begin{aligned} f'(s) &= \frac{\frac{ab}{b-a}e^{sa} - \frac{ab}{b-a}e^{sb}}{\frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}} = ab\,\frac{e^{sa} - e^{sb}}{be^{sa} - ae^{sb}} && (9) \\ f''(s) &= ab\,\frac{(be^{sa} - ae^{sb})(ae^{sa} - be^{sb}) - ab(e^{sa} - e^{sb})^2}{(be^{sa} - ae^{sb})^2} \\ &= -ab(a-b)^2\,\frac{e^{s(a+b)}}{(be^{sa} - ae^{sb})^2} \leq \frac{1}{4}(a-b)^2, && (10) \end{aligned}$$

where the last inequality follows from the fact that

$$(be^{sa} - ae^{sb})^2 + 4abe^{s(a+b)} = (be^{sa} + ae^{sb})^2 \geq 0,$$

which implies that

$$-ab\,\frac{e^{s(a+b)}}{(be^{sa} - a^{sb})^2} \leq \frac{1}{4}.$$

By Taylor's Theorem, for every real number s, there exists a real number s_0 such that

$$f(s) = f(0) + f'(0)s + \frac{1}{2}f''(s_0)s^2.$$

From (8), (9), and (10), it follows that $f(0) = 0$, $f'(0) = 0$ and, therefore,

$$f(s) \leq \frac{1}{8}(a-b)^2 s^2.$$

Combining this result with (7) and (8), we get (6).

To prove Hoeffding's inequalities, we apply Chernoff bound (1) to the random variable $S_n - \mathrm{E}[S_n]$ to get

$$P(S_n - \mathrm{E}[S_n] \geq nz) \leq e^{-snz}\mathrm{E}[e^{s(S_n - \mathrm{E}[S_n])}] \tag{11}$$

for all $s \geq 0$. Since $S_n = \sum_{i=1}^{n} X_i$, we have

$$\begin{aligned} \mathrm{E}\Big[e^{s(S_n - \mathrm{E}[S_n])}\Big] &= \mathrm{E}\Big[e^{s\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathrm{E}[X_i]\right)}\Big] \\ &= \mathrm{E}\Big[e^{s\left(\sum_{i=1}^{n} (X_i - \mathrm{E}[X_i])\right)}\Big]. \end{aligned}$$

As $X_1, X_2, \ldots, X_n$ are independent,

$$\mathrm{E}[e^{s\sum_{i=1}^{n}(X_i - \mathrm{E}[X_i])}] = \prod_{i=1}^{n} \mathrm{E}[e^{s(X_i - \mathrm{E}[X_i])}].$$

Applying (6) to the random variable $X_i - \mathrm{E}[X_i]$ which has zero mean and for which $a_i - \mathrm{E}[X_i] \leq X - \mathrm{E}[X_i] \leq b_i - \mathrm{E}[X_i]$, we get

$$\mathrm{E}[e^{s(S_n - \mathrm{E}[S_n])}] \leq \prod_{i=1}^{n} e^{(a_i - b_i)^2 s^2/8} = e^{\sum_{i=1}^{n}(a_i - b_i)^2 s^2/8}.$$

From (11),

$$P(S_n - \mathrm{E}[S_n] \geq nz) \leq e^{-snz + \sum_{i=1}^{n}(a_i - b_i)^2 s^2/8}.$$

This holds for every $s \geq 0$. The tightest upperbound on $P(S_n - \mathrm{E}[S_n]) \geq nz)$ is obtained by selecting $s \geq 0$ to minimize $-snz + \sum_{i=1}^{n}(a_i - b_i)^2 s^2/8$, which gives $s = 4nz/\sum_{i=1}^{n}(a_i - b_i)^2$. With such s, we get

$$P(S_n - \mathrm{E}[S_n] \geq nz) \leq e^{-2n^2 z^2 / \sum_{i=1}^{n}(a_i - b_i)^2}.$$

This proves (2). Replacing X_i by $-X_i$ in (2) gives (3). From (2) and (3), we get (4) from which (5) follows.

Example 5 Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli(p) random variables. Then, Hoeffding's inequality (5) gives

$$P(|\hat{X}_n - p| \geq z) \leq 2e^{-2nz^2}.$$