CIS6930/4930

Inequalities, Convergence of Random Variables

Prof. Ye Xia

Review: Convergence and Limits

Let $\{s_n\}_{n\geq 1} = s_1, s_2, \ldots$ be a sequence of real numbers. We say the sequence **converges to** s, written as $\lim_{n\to\infty} s_n = s$ (or $s_n \to s$ as $n \to \infty$), if for every $\epsilon > 0$, there exists N such that for **all** $n \geq N$, $|s_n - s| < \epsilon$.

Example: For $s_n = \frac{1}{n}$, $\lim_{n\to\infty} \frac{1}{n} = 0$.

If $\{s_n\}_{n\geq 1}$ is a non-decreasing sequence, then either $s_n \to \infty$ or $s_n \to s$ for some finite s. If, in addition, the sequence is bounded from above, then it converges to a finite limit.

Similar statements can be said about a non-increasing sequence.

If a sequence is bounded, it may not converge, but it must have a limit point.

A **limit point** of $\{s_n\}_{n\geq 1}$ is a number s such that for any $\epsilon > 0$, there exist infinitely many s_k such that $|s_k - s| < \epsilon$. In other words, however small ϵ is, there are always infinitely many points from the sequence $\{s_n\}$ that are within ϵ distance from s.

A fact or an alternative definition: s is a limit point of $\{s_n\}_{n\geq 1}$ if and only if there is a subsequence $\{s_{n_i}\}$ such that $s_{n_i} \to s$ as $i \to \infty$.

Example: The sequence $\{(-1)^n\} = -1, 1, -1, 1, -1, \ldots$ has two limit points: -1 and 1.

If $\{s_n\}_{n\geq 1}$ has a single limit point, then the sequence converges to that limit point.

We cannot write $\lim_{n\to\infty} s_n$ without affirming that the sequence converges. However, we can always write $\limsup_{n\to\infty} s_n$ and $\liminf_{n\to\infty} s_n$, provided we allow $+\infty$ and $-\infty$.

$\limsup s_n$ is the supremum of the set of limit points for $\{s_n\}$. $\liminf s_n$ is the infimum of the set of limit points for $\{s_n\}$. They are usually define as follows:

$$\limsup_{n\to\infty} s_n = \lim_{n\to\infty} \sup\{s_m : m \geq n\}.$$

$$\liminf_{n\to\infty} s_n = \lim_{n\to\infty} \inf\{s_m : m \geq n\}.$$

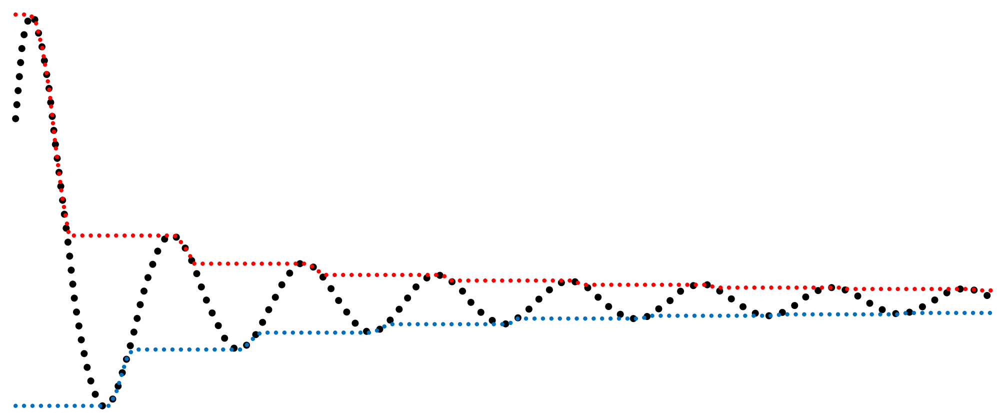

Figure 1: Sequences s_n (thicker black dots), $\sup\{s_m : m \geq n\}$ (red dots) and $\inf\{s_m : m \geq n\}$ (blue dots). If the gap between the latter two eventually goes to 0, then, the $\{s_n\}$ sequence has a single limit point and s_n converges to it. If the gap never closes, then $\{s_n\}$ has more than one limit points.

Note that $\sup\{s_m : m \geq n\}$ forms a non-increasing sequence in n. It is possible that the sequence is $\{\infty, \infty, \ldots\}$, in which case $\limsup s_n = \infty$. In other situations, it either goes to $-\infty$ or converges to a finite limit.

Similarly, $\inf\{s_m : m \geq n\}$ forms a non-decreasing sequence in n. Unless it is $\{-\infty, -\infty, \ldots\}$, it either goes to $+\infty$ or converges to a finite limit.

Clearly, $\liminf s_n \leq \limsup s_n$.

Fact: $\{s_n\}_{n\geq 1}$ converges if $\liminf s_n = \limsup s_n < \infty$ and in that case, $\lim s_n = \liminf s_n = \limsup s_n$.

There is a possibility that $\liminf s_n = \limsup s_n = \infty$ or $-\infty$. In those cases, we still can write $\lim s_n = \liminf s_n = \limsup s_n$, with the understanding that the notations $\lim s_n = \infty$ and $\lim s_n = -\infty$ are allowed. However, by convergence, we usually mean convergence to a finite limit.

Consider the sequence given by $s_n = n$ for $n \geq 1$. We have $\sup\{s_m : m \geq n\} = \{\infty, \infty, \ldots\}$, and therefore, $\limsup s_n = \infty$. Also, $\inf\{s_m : m \geq n\} = \{1, 2, \ldots\}$, and therefore, $\liminf s_n = \infty$. It is still true that $\lim s_n = \liminf s_n = \limsup s_n$.

Let $f : \mathbb{R} \to \mathbb{R}$. By $\lim_{x\to x_o} f(x) = a$, we mean for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - a| < \epsilon$ for all x with $|x - x_o| < \delta$.

There is a sequential version of the definition. $\lim_{x\to x_o} f(x) = a$ if for every sequence $\{x_n\}$ converging to x_o, $\lim_{n\to\infty} f(x_n) = a$. In other words,

$f(x_n) \to a$ as $x_n \to x_o$.

The function f is **continuous** at x_o if $\lim_{x \to x_o} f(x) = f(x_o)$. The sequential version of this: f is continuous at x_o if $f(x_n) \to f(x_o)$ for every sequence $\{x_n\}$ with $x_n \to x_o$.

$\limsup_{x \to x_o} f(x)$ and $\liminf_{x \to x_o} f(x)$ can also be defined.

Most of the above can be extended to general metric spaces. A metric is a generalized notion of distance. A **metric space** is a pair (M, d) where M is a set and d is the metric on M, which is a function $d : M \times M \to \mathbb{R}$. The metric must satisfy the following:

(i) $d(x, y) = 0$ if and only if $x = y$.

(ii) $d(x, y) = d(y, x)$.

(iii) $d(x, z) \leq d(x, y) + d(y, z)$. (triangle inequality)

Properties (i)-(iii) imply nonnegativity $d(x, y) \geq 0$.

Example: An Euclidean space is a metric space where $M = \mathbb{R}^n$ and the metric is the Euclidean distance.

For defining sequential convergence on a metric space, we only need to replace the absolute value in $|s_n - s|$ by $d(s_n, s)$. We can still talk about limit points of a sequence. However, $\limsup$ and $\liminf$ are undefined.

For a function $f : (M_1, d_1) \to (M_2, d_2)$ where (M_1, d_1) and (M_2, d_2) are two metric spaces, by $\lim_{x \to x_o} f(x) = a$, we mean for any $\epsilon > 0$, there exists $\delta > 0$ such that $d_2(f(x), a) < \epsilon$ for all x with $d_1(x, x_o) < \delta$.

Note: $\{x : d_1(x, x_o) < \delta\}$ is an open ball of radius δ around x_o.

Important Inequalities

Markov Inequality: If X is any random variable, then

$$P(|X| \geq a) \leq \frac{E|X|}{a}, \quad \text{for any } a > 0.$$

Proof: Let $A = \{|X| \geq a\}$. Let I_A be the indicator function of A. Then, $|X| \geq aI_A$. Why? On A, $|X| \geq a = aI_A$. On A^c, $|X| \geq 0 = aI_A$.

Taking the expectations on both sides, we have

$$E|X| \geq E[aI_A] = aP(A) = aP(|X| \geq a).$$

Generalized Markov Inequality: Let $h : \mathbb{R} \to [0, \infty)$ be a non-negative function. Then,

$$P(h(X) \geq a) \leq \frac{E[h(X)]}{a}, \quad \text{for any } a > 0.$$

Chebyshev's Inequality:

$$P(|X| \geq a) \leq \frac{E[X^2]}{a^2}, \quad \text{for any } a > 0.$$

Proof: Let $h(x) = x^2$ and notice $P(|X| \geq a) = P(X^2 \geq a^2)$.

Chernoff Bound/Inequality

For any a,

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{E[e^{tX}]}{e^{ta}} = e^{-ta} M_X(t), \quad \text{for every } t > 0.$$

Note: The proof is by applying the generalized Markov inequality with $h(x) = e^{tx}$.

Recall $E[e^{tX}]$ is the moment generating function $M_X(t)$.

The inequality is only useful for t such that $M_X(t) < \infty$.

The inequality holds for all a, but it is more useful for $a > 0$. This is in part because of the applications such as large deviations (see later), and in part because the bound is often trivial for $a < 0$. For instance, for $a < 0$, the best possible bound on the right hand is $M_X(t)$.

For a fixed $t > 0$ where $M(t) < \infty$, the probability decays at least exponentially fast in a.

Side Note: Recall that $M_X(t)$ is the most useful when it is finite in some neighborhood containing 0. For such X, the tail of the distribution function must decay at least exponentially fast.

For a given a, the bound can be tightened by minimizing $\frac{E[e^{tX}]}{e^{ta}}$ over $t > 0$. Sometimes, the bound is tight asymptotically (as in the results of large deviations).

The Chernoff bound is often applied to the case where X is the sum of independent random variables. For that case, let us write X by S_n. Let $S_n = X_1 + \cdots + X_n$. Then,

$$M_{S_n}(t) = \prod_{i=1}^{n} E[e^{tX_i}] = \prod_{i=1}^{n} M_{X_i}(t).$$

$$P(S_n \geq a) \leq \frac{E[e^{tS_n}]}{e^{ta}} = e^{-ta} \prod_{i=1}^{n} M_{X_i}(t), \quad \text{for every } t > 0.$$

$$P(S_n \geq a) \leq \inf_{t>0} e^{-ta} \prod_{i=1}^{n} M_{X_i}(t).$$

In the case where $X_1, X_2, \ldots$ are IID with mean μ, $S_n - n\mu$ is about the size of $n^{1/2}$ (probabilistically). Using the Central Limit Theorem, we can estimate

$P(S_n - n\mu > \sqrt{n}a)$ for large n.

But, sometimes we are interested in the probability of large deviations, i.e., $P(S_n - n\mu > na)$ for some $a > 0$.

Without loss of generality, let us assume $\mu = 0$. We have the following theorem.

Theorem (Large Deviations): Let $X_1, X_2, \ldots$ be a sequence of IID random variables with mean 0. Suppose their generating function $M(t)$ is finite in some interval containing 0. If $a > 0$ and $P(X_i > a) > 0$, then

$$P(S_n > na)^{1/n} \to e^{-\psi(a)}, \quad \text{as } n \to \infty,$$

where

$$\psi(a) = -\log\left(\inf_{t>0} e^{-at} M(t)\right).$$

- The issue is only interesting when $P(X_i > a) > 0$. Otherwise, $P(S_n > na) = 0$.
- The result can also be written as:

$$\frac{1}{n} \log P(S_n > na) \to -\psi(a), \quad \text{as } n \to \infty.$$

- Roughly, $P(S_n > na) \approx e^{-n\psi(a)}$ for large n. Therefore, $P(S_n > na)$ decays exponentially in n.

  More precisely, $P(S_n > na) \approx e^{-n\psi(a)+o(n)}$ for large n, where the $o(n)$ term satisfies $\frac{o(n)}{n} \to 0$ as $n \to \infty$. Sometimes, one writes $P(S_n > na) \approx A(n)e^{-n\psi(a)}$, with $A(n)$ representing the $e^{o(n)}$ term. In practice, $A(n)$ may be substantial.

Proof for the upper bound: The upper bound is just due to the Chernoff bound.

$$P(S_n > na) \leq P(S_n \geq na) \leq \inf_{t>0} e^{-nta}(M(t))^n,$$

which leads to

$$P(S_n > na)^{1/n} \leq \inf_{t>0} e^{-ta} M(t).$$

Letting $n \to \infty$, we get

$$\limsup_{n\to\infty} P(S_n > na)^{1/n} \leq \inf_{t>0} e^{-ta} M(t).$$

One also needs to show the lower bound

$$\liminf_{n\to\infty} P(S_n > na)^{1/n} \geq \inf_{t>0} e^{-ta} M(t).$$

This is the more difficult step (see Grimmett and Stirzaker, p. 184). We will omit it.

There is a second part of the Chernoff bound: For any a,

$$P(X \leq a) = P(e^{-tX} \geq e^{-ta}) \leq \frac{E[e^{-tX}]}{e^{-ta}} = e^{ta} M_X(-t), \quad \text{for every } t > 0.$$

It is more useful for $a < 0$.

For the theorem of large deviations quoted earlier, since $a > 0$, $P(S_n > na)$ is the large deviation above the mean (which is 0 by the theorem assumption). The theorem also has a part for large deviations below the mean: $P(S_n < na)$ for $a < 0$. It is related to the Chernoff bound in the above form.

The following special case of the Chernoff bound is often encountered in computer science.

Chernoff-Hoeffding Theorem: Suppose $X_1, X_2, \ldots, X_n$ are independent, $\{0, 1\}$-valued random variables with $E[X_i] = p_i$. Let $S_n = X_1 + \cdots + X_n$, and let $s_n = E[S_n]$ and let $p = s_n/n$. For any $\epsilon > 0$,

$$P(S_n \geq s_n + n\epsilon) = P(S_n \geq n(p + \epsilon)) \leq e^{-nD_p(p+\epsilon)}$$

$$P(S_n \leq s_n - n\epsilon) = P(S_n \leq n(p - \epsilon)) \leq e^{-nD_p(p-\epsilon)},$$

where $D_y(x) = x \ln \frac{x}{y} + (1 - x) \ln \frac{1-x}{1-y}$ is the relative entropy or Kullback-Leibler divergence of two Bernoulli distributions with success probability equal to x or y, respectively.

By loosening the bounds, the theorem can take other forms, e.g., for $0 < \delta < 1$,

$$P(S_n \geq (1 + \delta)s_n) \leq \left(\frac{e^{\delta}}{(1 + \delta)^{1+\delta}}\right)^{np} \leq e^{\frac{-\delta^2 pn}{2+\delta}},$$

$$P(S_n \leq (1 - \delta)s_n) \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}}\right)^{np} \leq e^{\frac{-\delta^2 pn}{2}}.$$

More Inequalities

Hölder's Inequality: If $p, q > 1$ and $p^{-1} + q^{-1} = 1$, then

$$E|XY| \leq E[|X|^p]^{1/p} E[|Y|^q]^{1/q}.$$

For $r \geq 1$, $E[|X|^r]^{1/r}$ defines a r-norm. Hölder's inequality says:

$$\|XY\|_1 \leq \|X\|_p \ \|Y\|_q \, .$$

When $p = q = 2$, Hölder's inequality yields the Cauchy-Schwarz inequality:
$E[XY]^2 \leq E[X^2]E[Y^2]$ because $E[XY]^2 \leq E[|XY|]^2 \leq E[X^2]E[Y^2]$.

Hölder's inequality also holds when $p = 1$ and $q = \infty$ or vice versa. More definitions are needed to make sense of those cases, which we omit.

Fact: If $E[|X|^p] < \infty$ and $E[|Y|^q] < \infty$, then $E|XY| < \infty$.

Fact: Suppose $0 < r < s$. If $E[|X|^s] < \infty$, then $E[|X|^r] < \infty$. Define $p = s/r > 1$ and $q = p/(p-1)$. Apply Hölder's inequality to two random variables $|X|^r$ and 1_Ω. Note that $E[(1_\Omega)^q] = E[1_\Omega] = 1$.

$$E[|X|^r] = E[|X|^r 1_\Omega] \leq E[(|X|^r)^p]^{1/p} \cdot 1 = E[|X|^s]^{r/s}.$$

Therefore, if $E[|X|^s] < \infty$, then $E[|X|^r] < \infty$.

Minkowski's Inequality: If $p \geq 1$, then

$$E[|X+Y|^p]^{1/p} \leq E[|X|^p]^{1/p} + E[|Y|^p]^{1/p}.$$

Minkowski's inequality says:

$$\|X+Y\|_p \leq \|X\|_p + \|Y\|_p .$$

This is the triangle inequality.

Minkowski's inequality can be proved by using Hölder's inequality (omitted).

Definition: A function $f : \mathbb{R} \to \mathbb{R}$ is convex if for all $y, z \in \mathbb{R}$ and all $0 \leq \lambda \leq 1$,

$$f(\lambda y + (1-\lambda)z) \leq \lambda f(y) + (1-\lambda) f(z).$$

The definition implies that, for constants $\lambda_1, \lambda_2, \ldots, \lambda_n$ with $\lambda_i \in [0,1]$ for each i and $\sum_{i=1}^n \lambda_i = 1$,

$$f(\sum_{i=1}^n \lambda_i x_i) \leq \sum_{i=1}^n \lambda_i f(x_i),$$

for every tuple $(x_1, \ldots, x_n)$.

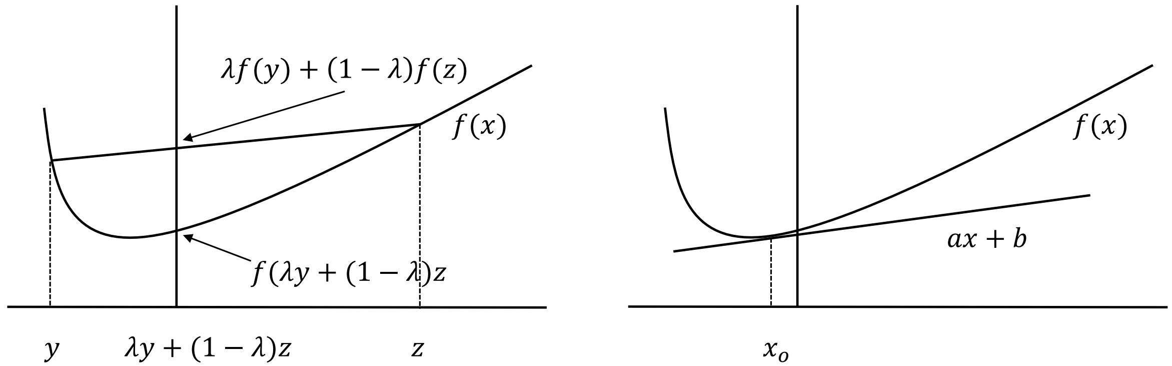

Figure 2: Left: definition. Right: useful fact

A useful fact is that if f is convex, it has a linear global under-estimator at every point x_o. In other words, there exist constants a and b, which depend on x_o, such that $ax_o + b = f(x_o)$, and for all x,

$$f(x) \geq ax + b.$$

When f is differentiable at x_o, then $a = f'(x_o)$ and $b = f(x_o) - f'(x_o)x_o$. In other words, we have

$$f(x) \geq f'(x_o)(x - x_o) + f(x_o).$$

Jensen's Inequality: If $f : \mathbb{R} \to \mathbb{R}$ is a convex function and X is a random variable,

$$f(E[X]) \leq E[f(X)],$$

provided $E|X| < \infty$ and $E|f(X)| < \infty$.

See the figure to remember the direction of the inequality.

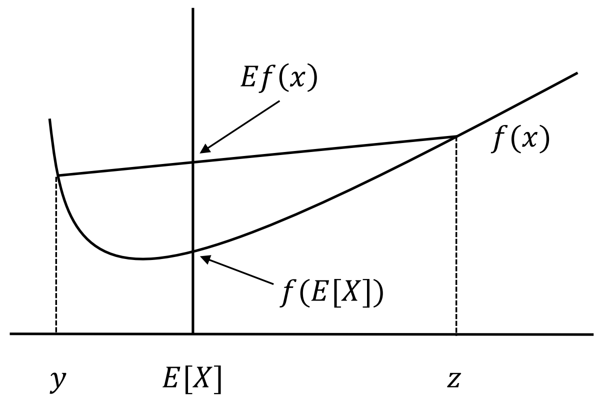

Figure 3: Special Case: $P(X = y) = \lambda$; $P(X = z) = 1 - \lambda$

Convergence of Random Variables

Definition: Let $X_1, X_2, \ldots$ and X be random variables on some probability space $(\Omega, \mathcal{F}, P)$. We say that

(a) $X_n \to X$ **almost surely**, written $X_n \stackrel{\text{a.s.}}{\to} X$, if $\{\omega \in \Omega : X_n(\omega) \to X(\omega),\ \text{as } n \to \infty\}$ is an event with probability 1.

(b) $X_n \to X$ **in** r**th mean or in** L^r, where $r \geq 1$, written $X_n \stackrel{L^r}{\to} X$, if $E|X_n^r| < \infty$ for all n and

$$E[|X_n - X|^r] \to 0,\ \text{as } n \to \infty.$$

(c) $X_n \to X$ **in probability**, written $X_n \stackrel{\text{p}}{\to} X$, if for any $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \to 0,\ \text{as } n \to \infty.$$

(d) $X_n \to X$ **in distribution**, written $X_n \stackrel{\text{d}}{\to} X$, if

$$P(X_n \leq x) \to P(X \leq x),\ \text{as } n \to \infty.$$

for all points x at which $F_X(x) = P(X \leq x)$ is continuous.

Comments:

- $X_n \stackrel{\text{a.s.}}{\to} X$ is also called (i) $X_n \to X$ almost everywhere; or (ii) $X_n \to X$ with probability 1.
- Convergence in distribution is really about distribution functions $F_1, F_2, \ldots$ converging to another distribution function F at all points of continuity of F, i.e., $F_n \to F$.
- In general,

$$X_n \stackrel{\text{a.s.}}{\to} X \implies X_n \stackrel{\text{p}}{\to} X \implies X_n \stackrel{\text{d}}{\to} X.$$

For any $r \geq 1$,

$$X_n \stackrel{L^r}{\to} X \implies X_n \stackrel{\text{p}}{\to} X \implies X_n \stackrel{\text{d}}{\to} X.$$

If $r > s \geq 1$,

$$X_n \stackrel{L^r}{\to} X \implies X_n \stackrel{L^s}{\to} X.$$

- In general, $X_n \stackrel{\text{p}}{\to} X \not\Longrightarrow X_n \stackrel{\text{a.s.}}{\to} X$.

$X_n \stackrel{\text{p}}{\to} X$ can also be understood as: for every $\epsilon > 0$, $P(|X_n - X| \leq \epsilon) \to 1$. That is, for any $\delta > 0$, there exists $N(\epsilon, \delta)$ such that for all $n > N(\epsilon, \delta)$, $P(|X_n - X| \leq \epsilon) > 1 - \delta$.

Let $B(n, \epsilon) = \{\omega : |X_n(\omega) - X(\omega)| \leq \epsilon\}$. It can happen that different ω takes turn goes into and out of $B(n, \epsilon)$ for different n indefinitely, while still having $P(B(n, \epsilon)) > 1 - \delta$ for every $n > N(\epsilon, \delta)$.

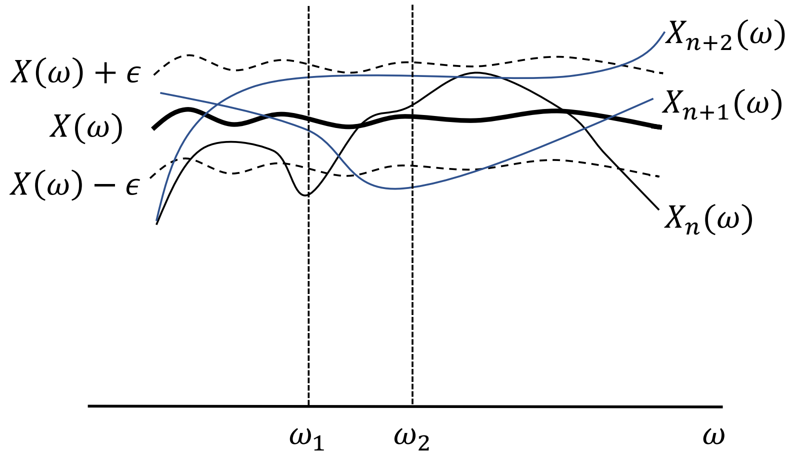

Figure 4: For each n, most of ω are in $B(n,\epsilon)$. At ω_1, $|X_{n+1}(\omega_1) - X(\omega_1)| \le \epsilon$ and $|X_{n+2}(\omega_1) - X(\omega_1)| \le \epsilon$. At ω_2, $|X_n(\omega_2) - X(\omega_2)| \le \epsilon$ and $|X_{n+2}(\omega_2) - X(\omega_2)| \le \epsilon$.

It can happen, for some $\epsilon > 0$, $|X_n(\omega) - X(\omega)| > \epsilon$ infinitely often (that is, for infinitely many n) for a substantial 'portion' of ω. In such a case, $P\{\omega : X_n(\omega) \to X(\omega)\} < 1$. That is, the almost sure convergence does not hold.

Example: Let $\{X_n\}_{n\ge 1}$ be a sequence of **independent** random variables, where

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

Then, for any $0 < \epsilon < 1$,

$$P(|X_n - 0| > \epsilon) = P(|X_n| > \epsilon) = \frac{1}{n} \to 0.$$

Hence, $X_n \stackrel{\text{p}}{\to} 0$.

It can be shown that, for any $0 < \epsilon < 1$, $|X_n(\omega) - X(\omega)| > \epsilon$ infinitely often for almost all ω. That is, $P\{\omega : X_n(\omega) \to X(\omega)\} = 0$. X_n converges to 0 at almost nowhere (see Grimmett and Stirzaker, p. 279).

- $X_n \stackrel{L^1}{\to} X \;\not\Longrightarrow\; X_n \stackrel{\text{a.s.}}{\to} X$.

  For the same example above, $E|X_n| = 1/n \to 0$. Hence, $X_n \stackrel{L^1}{\to} 0$. But, X_n converges to 0 at almost nowhere.

- $X_n \stackrel{\text{p}}{\to} X \;\not\Longrightarrow\; X_n \stackrel{L^r}{\to} X$.

  Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables, where

$$X_n = \begin{cases} n^3 & \text{with probability } \frac{1}{n^2} \\ 0 & \text{with probability } 1 - \frac{1}{n^2}. \end{cases}$$

  Then, for any $0 < \epsilon < 1$,

$$P(|X_n - 0| > \epsilon) = P(|X_n| > \epsilon) = \frac{1}{n^2} \to 0.$$

Hence, $X_n \xrightarrow{\text{p}} 0$. But, for $r \geq 1$, $E|X_n - 0|^r = n^{3r} \times \frac{1}{n^2} = n^{3r-2} \to \infty$. Therefore, X_n does not converge to 0 in L^r.

- $X_n \xrightarrow{\text{a.s.}} X \;\not\Longrightarrow\; X_n \xrightarrow{L^r} X$.

  In the previous example, for any $0 < \epsilon < 1$, $\sum_{n=1}^{\infty} P(|X_n - 0| > \epsilon) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$. By a later theorem, that is a sufficient condition for $X_n \xrightarrow{\text{a.s.}} 0$. But, the convergence in the r-th mean does not hold for any $r \geq 1$.

Bounding Probabilities

Computing precise probabilities are often either difficult or impossible due to incomplete information. Much of the probabilistic analysis is about bounding probabilities.

If events $A \subseteq B$, we say A **implies** B. What that means is that if A occurs, then B occurs.

More precisely, suppose the experimental outcome is ω and suppose $\omega \in A$. In that case, we say event A occurred. Since $A \subseteq B$, $\omega \in B$. Therefore, B also occurred.

If A implies B, clearly, $P(A) \leq P(B)$.

The trick in upper-bounding the probability of an event of interest A is to come up with another event B, which is implied by A and for which we know how to calculate the probability.

Similarly, for lower-bounding the probability of an event A, try to come up with an event C that implies A, i.e., $C \subseteq A$.

Example: Show that if $X_n \xrightarrow{\text{p}} X$ then $X_n \xrightarrow{\text{d}} X$.

Proof: Suppose $X_n \xrightarrow{\text{p}} X$ and write

$$F_n(x) = P(X_n \leq x), \quad F(x) = P(X \leq x),$$

for the distribution functions of X_n and X, respectively. Fix an $\epsilon > 0$. We have

$$\begin{aligned} F_n(x) =& P(X_n \leq x) = P(X_n \leq x, X \leq x + \epsilon) + P(X_n \leq x, X > x + \epsilon) \\ \leq& P(X \leq x + \epsilon) + P(|X_n - X| > \epsilon) \\ =& F(x + \epsilon) + P(|X_n - X| > \epsilon). \end{aligned}$$

For the inequality above is due to two upper bounds. First, clearly $\{X_n \leq x, X \leq x + \epsilon\} \subseteq \{X \leq x + \epsilon\}$.

Second, if $X_n \leq x$ and $X > x + \epsilon$, then $|X_n - X| > \epsilon$. This shows if the event $\{X_n \leq x, X > x + \epsilon\}$ occurs, then the event $\{|X_n - X| > \epsilon\}$ also occurs. In other words, the event $\{X_n \leq x, X > x + \epsilon\}$ implies the event $\{|X_n - X| > \epsilon\}$. Therefore,

$$P(X_n \leq x, X > x + \epsilon) \leq P(|X_n - X| > \epsilon).$$

We now continue with the proof. Similarly,

$$
\begin{aligned}
F(x-\epsilon) =& P(X \le x-\epsilon) = P(X \le x-\epsilon, X_n \le x) + P(X \le x-\epsilon, X_n > x) \\
\le& P(X_n \le x) + P(|X_n - X| > \epsilon) \\
=& F_n(x) + P(|X_n - X| > \epsilon).
\end{aligned}
$$

We have

$$
F(x-\epsilon) - P(|X_n - X| > \epsilon) \le F_n(x) \le F(x+\epsilon) + P(|X_n - X| > \epsilon)
$$

$X_n \xrightarrow{\text{p}} X$ means that $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$. Letting $n \to \infty$, we have

$$
\begin{aligned}
\limsup F_n(x) \le& F(x+\epsilon) + \limsup P(|X_n - X| > \epsilon) \\
=& F(x+\epsilon) + \lim_{n\to\infty} P(|X_n - X| > \epsilon) \\
=& F(x+\epsilon).
\end{aligned}
$$

$$\begin{aligned}\liminf F_n(x) \geq & F(x-\epsilon) - \limsup P(|X_n - X| > \epsilon) \\ = & F(x-\epsilon) - \lim_{n\to\infty} P(|X_n - X| > \epsilon) \\ = & F(x-\epsilon).\end{aligned}$$

In the first line, we used the fact that, for a sequence $\{a_n\}$, $\liminf(-a_n) = -\limsup a_n$.

In summary, we have

$$F(x-\epsilon) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(x+\epsilon),$$

for all $\epsilon > 0$.

If F is continuous at x, then

$$F(x-\epsilon) \uparrow F(x) \quad \text{and} \quad F(x+\epsilon) \downarrow F(x), \text{ as } \epsilon \downarrow 0.$$

Letting $\epsilon \downarrow 0$, we get

$$F(x) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(x).$$

Then, it must be that

$$\liminf F_n(x) = \limsup F_n(x) = F(x),$$

which shows $\lim_{n\to\infty} F_n(x) = F(x)$ at any continuity point of F.

More on Convergence of Random Variables

Theorem:

(a) If $X_n \stackrel{\text{d}}{\to} c$, where c is a constant, then $X_n \stackrel{\text{p}}{\to} c$.

(b) If $P_n(\epsilon) = P(|X_n - X| > \epsilon)$ satisfies $\sum_n P_n(\epsilon) < \infty$ for any $\epsilon > 0$, then $X_n \stackrel{\text{a.s.}}{\to} X$.

Proof of (a): Let F be the distribution function of the constant random variable $X = c$. That is, $F(x) = 0$ for $x < c$ and $F(x) = 1$ for $x \geq c$. For any $\epsilon > 0$, $c - \epsilon$ and $c + \epsilon$ are both continuity points of F.

$X_n \stackrel{\text{d}}{\to} c$ implies that $P(X_n \leq c - \epsilon) \to 0$ and $P(X_n \leq c + \epsilon) \to 1$ as $n \to \infty$. The latter implies $P(X_n > c + \epsilon) \to 0$. Then,

$$P(|X_n - c| > \epsilon) = P(X_n > c + \epsilon) + P(X_n < c - \epsilon) \to 0.$$

Proof of (b): Let

$$A_n(\epsilon) = \{|X_n - X| > \epsilon\}, \qquad B_m(\epsilon) = \bigcup_{n \geq m} A_n(\epsilon).$$

Interpretation: If $\omega \in B_m(\epsilon)$, there exists $n \geq m$ such that $\omega \in A_n(\epsilon)$, which means $|X_n(\omega) - X(\omega)| > \epsilon$.

$\{B_m(\epsilon) : m \geq 1\}$ is a non-increasing sequence of events, that is, $B_1(\epsilon) \supseteq B_2(\epsilon) \supseteq \cdots$. Recall $\lim_{m\to\infty} B_m(\epsilon)$ is defined as $\bigcap_{m=1}^{\infty} B_m(\epsilon)$. Let $A(\epsilon) \triangleq \bigcap_{m=1}^{\infty} B_m(\epsilon)$.

$A(\epsilon) = \bigcap_{m\geq 1} \bigcup_{n\geq m} A_n$ is known as the set of ω for which the events in the $\{A_n(\epsilon)\}$ sequence occur infinitely often. $A(\epsilon)$ is a measurable set and therefore an event. It is known as the event that **infinitely many** $\{A_n(\epsilon)\}$ occur or $\{A_n(\epsilon)\}$ occur **infinitely often**.

Interpretation: Suppose $\omega \in A(\epsilon)$. Then, there are **infinitely many** n for which $A_n(\epsilon)$ occurs. That is, $|X_n(\omega) - X(\omega)| > \epsilon$ for infinitely many n. Why? We need to show for any $N \geq 1$, there exists $n > N$ such that $\omega \in A_n(\epsilon)$.

$\omega \in A(\epsilon)$ implies that $\omega \in B_m(\epsilon)$ for all $m \geq 1$. Therefore, for the given N, $\omega \in B_{N+1}(\epsilon)$. Then, $\omega \in \bigcup_{n\geq N+1} A_n(\epsilon)$, which implies there is some $n \geq N+1$, such that $\omega \in A_n(\epsilon)$.

Continue with the proof. If we can show $P(A(\epsilon)) = 0$, that would mean for almost all ω, $|X_n(\omega) - X(\omega)| \leq \epsilon$ eventually. That is, there exists $N \geq 1$, which may depend on ω, such that for all $n \geq N$, $|X_n(\omega) - X(\omega)| \leq \epsilon$. Then, since $\epsilon > 0$ is arbitrarily chosen, we can conclude that for almost all $\omega \in \Omega$, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$, and hence, $X_n \stackrel{\text{a.s.}}{\to} X$.

By the continuity of probability measure (a homework 1 question),

$$P(A(\epsilon)) = P(\lim_{m\to\infty} B_m(\epsilon)) = \lim_{m\to\infty} P(B_m(\epsilon)).$$

Since $B_m(\epsilon) = \bigcup_{n\geq m} A_n(\epsilon)$, $P(B_m(\epsilon)) \leq \sum_{n\geq m} P(A_n(\epsilon))$. This is know as the **union bound**.

The condition of the theorem says $\sum_{n\geq 1} P_n(\epsilon) < \infty$ for any $\epsilon > 0$. That is, $\sum_{n\geq 1} P(A_n(\epsilon)) < \infty$ for any $\epsilon > 0$. That implies that the tail part $\sum_{n\geq m} P(A_n(\epsilon)) \to 0$ as $m \to \infty$ (since the partial sum of $P(A_n(\epsilon))$ is a Cauchy sequence). Therefore, $P(B_m(\epsilon)) \to 0$ as $m \to \infty$, and hence, $P(A(\epsilon)) = 0$.

Theorem: If $X_n \xrightarrow{\text{p}} X$, then there exists a non-random increasing sequence of integers $n_1, n_2, \ldots$ such that $X_{n_i} \xrightarrow{\text{a.s.}} X$ as $i \to \infty$.

Proof: Since $X_n \xrightarrow{\text{p}} X$, we have, for any $\epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$. Pick an increasing sequence $n_1, n_2, \ldots$ of positive integers such that $P(|X_{n_i} - X| > 1/i) \leq 1/i^2$.

How to do this? For each fixed i, let $\delta = 1/i^2$. Since $P(|X_n - X| > 1/i) \to 0$ as $n \to \infty$, there exists N such that for **all** $n > N$, $P(|X_n - X| > 1/i) \leq \delta = 1/i^2$. Choose $n_i > \max(N, n_{i-1})$. We thus ensure n_i is an increasing sequence and $P(|X_{n_i} - X| > 1/i) \leq 1/i^2$.

Then, for any $\epsilon > 0$,

$$\sum_{i \geq \epsilon^{-1}} P(|X_{n_i} - X| > \epsilon) \leq \sum_{i \geq \epsilon^{-1}} P(|X_{n_i} - X| > 1/i) \leq \sum_{i \geq \epsilon^{-1}} 1/i^2 < \infty,$$

since for each $i \geq \epsilon^{-1}$, the event $\{|X_{n_i} - X| > \epsilon\}$ implies the event $\{|X_{n_i} - X| > 1/i\}$.

Therefore, for any $\epsilon > 0$, $\sum_{i \geq 1} P(|X_{n_i} - X| > \epsilon) < \infty$.

We can then apply part (b) of the previous theorem.

Borel-Cantelli Lemma

The Borel-Cantelli lemma is often used to show almost sure convergence.

Let $\{A_n\}$ be a sequence of events, i.e., measurable subsets of the probability space $(\Omega, \mathcal{F}, P)$.

Let $A = \cap_{m\geq 1} \cup_{n\geq m} A_n$, the event (a measurable set) that infinitely many A_n occur, or A_n occur infinity often.

Side Note: Let $B = \cup_{m\geq 1} \cap_{n\geq m} A_n$. B is the event (set) that A_n occur **eventually**, i.e., for all $n > N$ for some N (which may depend on ω). Why? For any $\omega \in B$, there exists $N(\omega)$ such that $\omega \in \cap_{n\geq N(\omega)} A_n$. Then, $\omega \in A_n$ for all $n \geq N(\omega)$.

Borel-Cantelli Lemma: (a) $P(A) = 0$ if $\sum_{n=1}^{\infty} P(A_n) < \infty$.

(b) $P(A) = 1$ if $\sum_{n=1}^{\infty} P(A_n) = \infty$ and $A_1, A_2, \ldots$ are independent events.

Proof: We have already shown (a) in the proof of the previous theorem.

For (b), consider $A^c = \cup_{m\geq 1} \cap_{n\geq m} A_n^c$. We need to show $P(A^c) = 0$.

$B_m \triangleq \cap_{n=m}^{\infty} A_n^c$ is a non-decreasing sequence of sets. Therefore, we can write $\lim_{m\to\infty} B_m$, which is equal to $A^c = \cup_{m\geq 1} B_m$, and by the continuity of probability (homework 1 question),

$$P(A^c) = P(\lim_{m\to\infty} B_m) = \lim_{m\to\infty} P(B_m) = \lim_{m\to\infty} P(\cap_{n=m}^{\infty} A_n^c).$$

Consider $P(B_m) = P(\cap_{n=m}^{\infty} A_n^c)$. Note that $E_r \triangleq \cap_{n=m}^{r} A_n^c$ is a non-increasing sequence of events and $\lim_{r\to\infty} E_r = \cap_{n=m}^{\infty} A_n^c$. Then, by the continuity of probability,
$P(\cap_{n=m}^{\infty} A_n^c) = P(\lim_{r\to\infty} E_r) = \lim_{r\to\infty} P(E_r) = \lim_{r\to\infty} P(\cap_{n=m}^{r} A_n^c)$.

Then, for any $m \geq 1$,

$$
\begin{aligned}
P(B_m) &= \lim_{r\to\infty} P(\cap_{n=m}^{r} A_n^c) \\
&= \lim_{r\to\infty} \prod_{n=m}^{r} (1 - P(A_n)) \qquad \text{by independence} \\
&= \prod_{n=m}^{\infty} (1 - P(A_n)) \\
&\leq \prod_{n=m}^{\infty} \exp(-P(A_n)) \qquad \text{because } 1 - x \leq e^{-x} \text{ for } x \geq 0 \\
&= \exp\left(-\sum_{n=m}^{\infty} P(A_n)\right) = 0.
\end{aligned}
$$

Therefore, $P(A^c) = \lim_{m\to\infty} P(B_m) = 0$.

Note: Independence of an infinity collection of events $\{A_n\}_{n\in I}$ is defined as: any **finite** subset of the events are independent. This is why we didn't directly write $P(\cap_{n=m}^{\infty} A_n^c) = \prod_{n=m}^{\infty}(1 - P(A_n))$. When I is a countable set, which is the case here, it turns out the expression is true by the continuity of probability.

More on Convergence of Random Variables

Theorem (Skorokhod's Representation Theorem: If $\{X_n\}_{n\geq 1}$ and X, with distribution functions $\{F_n\}_{n\geq 1}$ and F, are such that

$$X_n \stackrel{\mathrm{d}}{\to} X(\text{ or, equivalently, } F_n \to F) \text{ as } n \to \infty,$$

then there exists a probability space $(\Omega', \mathcal{F}', P')$ and random variables $\{Y_n\}_{n\geq 1}$ and Y, which maps Ω' to $\mathbb{R}$, such that

(a) $\{Y_n\}$ and Y have distribution functions $\{F_n\}$ and F,

(b) $Y_n \stackrel{\text{a.s.}}{\to} Y$ as $n \to \infty$.

Theorem: Suppose $g : \mathbb{R} \to \mathbb{R}$ is continuous function. Then,

(a) $X_n \stackrel{\mathrm{d}}{\to} X \Rightarrow g(X_n) \stackrel{\mathrm{d}}{\to} g(X)$;

(b) $X_n \stackrel{\mathrm{p}}{\to} X \Rightarrow g(X_n) \stackrel{\mathrm{p}}{\to} g(X)$;

(c) $X_n \stackrel{\text{a.s.}}{\to} X \Rightarrow g(X_n) \stackrel{\text{a.s.}}{\to} g(X)$;

Proof: (a) Let Y_n and Y as given in the Skorokhod's Representation Theorem. By the continuity of g,

$$\{\omega : Y_n(\omega) \to Y(\omega)\} \subseteq \{\omega : g(Y_n(\omega)) \to g(Y(\omega))\}.$$

Since $P(\{\omega : Y_n(\omega) \to Y(\omega)\}) = 1$, we have $P(\{\omega : g(Y_n(\omega)) \to g(Y(\omega))\}) = 1$. That is, $g(Y_n) \overset{\text{a.s.}}{\to} g(Y)$, which implies $g(Y_n) \overset{\text{d}}{\to} g(Y)$.

Since $g(Y_n)$ has the same distribution as $g(X_n)$ and $g(Y)$ has the same distribution as $g(X)$, we have $g(X_n) \overset{\text{d}}{\to} g(X)$.

(b) Fix $\epsilon > 0$. For any $\delta > 0$ define

$$B_\delta = \{x \in \mathbb{R} : \text{ there exists } y \in \mathbb{R} \text{ such that } |x - y| < \delta \text{ and } |g(x) - g(y)| > \epsilon\}.$$

$$\begin{aligned} P(|g(X_n) - g(X)| > \epsilon) =& P(|g(X_n) - g(X)| > \epsilon, X \in B_\delta^c) \\ &+ P(|g(X_n) - g(X)| > \epsilon, X \in B_\delta) \end{aligned}$$

The event $\{|g(X_n) - g(X)| > \epsilon, X \in B_\delta^c\}$ implies the event $\{|X_n - X| \geq \delta\}$, which means $\{|g(X_n) - g(X)| > \epsilon, X \in B_\delta^c\} \subseteq \{|X_n - X| \geq \delta\}$. Therefore,

$$P(|g(X_n) - g(X)| > \epsilon, X \in B_\delta^c) \leq P(|X_n - X| \geq \delta).$$

Also,

$$P(|g(X_n) - g(X)| > \epsilon, X \in B_\delta) \leq P(X \in B_\delta).$$

Therefore,

$$P(|g(X_n) - g(X)| > \epsilon) \leq P(|X_n - X| \geq \delta) + P(X \in B_\delta).$$

Letting $n \to \infty$, $\lim_{n\to\infty} P(|X_n - X| \geq \delta) = 0$. We have

$$\limsup_{n\to\infty} P(|g(X_n) - g(X)| > \epsilon) \leq P(X \in B_\delta).$$

The above is true for all $\delta > 0$. Letting $\delta \downarrow 0$, $B_\delta \to \emptyset$ since g is continuous, and therefore, $P(X \in B_\delta) \to 0$. Therefore,

$$\limsup_{n\to\infty} P(|g(X_n) - g(X)| > \epsilon) \leq 0,$$

which implies that, as $n \to \infty$, the limit of $P(|g(X_n) - g(X)| > \epsilon)$ exists and it is equal to 0, i.e., $\lim_{n\to\infty} P(|g(X_n) - g(X)| > \epsilon) = 0$. Hence,
$g(X_n) \stackrel{\text{p}}{\to} g(X)$.

(c) Same as the proof for (a) that shows $Y_n \stackrel{\text{a.s.}}{\to} Y$.

Theorem: The following statements are equivalent.

(a) $X_n \xrightarrow{\text{d}} X$.

(b) $E[g(X_n)] \to E[g(X)]$ for all bounded continuous functions g.

Theorem: Let a and b be any real numbers.

(a) If $X_n \xrightarrow{\text{a.s.}} X$ and $Y_n \xrightarrow{\text{a.s.}} Y$, then $aX_n + bY_n \xrightarrow{\text{a.s.}} aX + bY$, and $X_nY_n \xrightarrow{\text{a.s.}} XY$.

(b) If $X_n \xrightarrow{L^r} X$ and $Y_n \xrightarrow{L^r} Y$, then $aX_n + bY_n \xrightarrow{L^r} aX + bY$.

(c) If $X_n \xrightarrow{\text{p}} X$ and $Y_n \xrightarrow{\text{p}} Y$, then $aX_n + bY_n \xrightarrow{\text{p}} aX + bY$, and $X_nY_n \xrightarrow{\text{p}} XY$.

Weak Law of Large Numbers

Theorem: Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite mean μ. Let S_n be the partial sum:

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

Then, we have

$$\frac{1}{n} S_n \xrightarrow{\mathrm{p}} \mu, \quad \text{as } n \to \infty.$$

Earlier when we discuss characteristic functions, we showed that under the same condition,

$$\frac{1}{n} S_n \xrightarrow{\mathrm{d}} \mu, \quad \text{as } n \to \infty.$$

One of the previous theorems says: If $X_n \xrightarrow{\mathrm{d}} c$, where c is a constant, then $X_n \xrightarrow{\mathrm{p}} c$. Hence, we have

$$\frac{1}{n} S_n \xrightarrow{\mathrm{p}} \mu.$$

Strong Law of Large Numbers

Theorem: Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite mean μ. Let S_n be the partial sum:

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

Then, we have

$$\frac{1}{n} S_n \overset{\text{a.s.}}{\to} \mu, \ \text{ as } n \to \infty.$$

Dominated Convergence Theorem

Theorem: Suppose $|X_n| \leq Z$ for all n, where $E[Z] < \infty$. If $X_n \xrightarrow{\text{p}} X$, then $X_n \xrightarrow{L^1} X$.

Proof: $X_n \xrightarrow{\text{p}} X$ implies that that is a subsequence $n_1, n_2, \ldots$ such that $X_{n_i} \xrightarrow{\text{a.s.}} X$ as $i \to \infty$. Then, $|X_{n_i}| \xrightarrow{\text{a.s.}} |X|$ as $i \to \infty$. Since each $|X_{n_i}| \leq Z$ for each i, $|X| \leq Z$ almost surely.

Let $Z_n = |X_n - X|$. We have $Z_n \leq |X_n| + |X| \leq 2Z$ almost surely. In the following, let I_A be an indicator random variable for event A.

$$\begin{aligned}
E[Z_n] =& E[Z_n I_{\{Z_n \leq \epsilon\}}] + E[Z_n I_{\{Z_n > \epsilon\}}] \\
\leq& \epsilon + E[Z_n I_{\{Z_n > \epsilon\}}] \\
\leq& \epsilon + E[2Z I_{\{Z_n > \epsilon\}}]
\end{aligned}$$

It can be shown that, if $E[Z] < \infty$ and $P(A_n) \to 0$ for a sequence of events

$\{A_n\}_{n\geq 1}$, then $E[ZI_{A_n}] \to 0$. Here, $A_n = \{Z_n > \epsilon\}$. Then, we have

$$\limsup_{n\to\infty} E[Z_n] \leq \epsilon.$$

Letting $\epsilon \downarrow 0$, we have $\limsup_{n\to\infty} E[Z_n] = 0$. Since $E[Z_n] \geq 0$ for each n, we have $\lim_{n\to\infty} E[Z_n] = 0$.

The Dominated Convergence Theorem usually takes the following form:

Corollary (Dominated Convergence Theorem): Suppose $|X_n| \leq Z$ for all n, where $E[Z] < \infty$. If $X_n \stackrel{\text{a.s.}}{\to} X$, then $E[X_n] \to E[X]$.

Proof: $X_n \stackrel{\text{a.s.}}{\to} X$ implies $X_n \stackrel{\text{p}}{\to} X$. Also, $|E[X_n] - E[X]| \leq E|X_n - X|$. Then, apply the above theorem.

Corollary (Bounded Convergence Theorem): Suppose $|X_n| \leq M$ for all n, where M is a constant. If $X_n \stackrel{\text{a.s.}}{\to} X$, then $E[X_n] \to E[X]$.

Monotone Convergence Theorem

Monotone Convergence Theorem: Suppose $0 \le X_n \uparrow X$ almost surely. Then, $E[X_n] \uparrow E[X]$.

Here, it is allowed that $E[X] = \infty$.

The theorem is a result of the properties of integral.

For random variables, the non-negativity of the functions can be relaxed in various ways. For instance, suppose C is a constant, possibly negative.

Suppose $C \le X_n \uparrow X$ almost surely. Then, $E[X_n] \uparrow E[X]$.

Since $0 \le X_n - C \uparrow X - C$, we can apply the regular version of the theorem and get

$$E[X_n] - C = E[X_n - C] \uparrow E[X - C] = E[X] - C.$$

Then, we have $E[X_n] \uparrow E[X]$.

Fatou's Lemma

Fatou's Lemma is about interchanging limit with expectation (integral).

Fatou's Lemma: If $X_n \geq 0$ for all n, then

$$\liminf_{n\to\infty} E[X_n] \geq E[\liminf_{n\to\infty} X_n].$$

Proof: Let $Y_n = \inf_{k\geq n} X_k$. Note that $0 \leq Y_n \uparrow \liminf_{n\to\infty} X_n$. By the Monotone Convergence Theorem,

$$E[Y_n] \uparrow E[\liminf_{n\to\infty} X_n].$$

Since $X_n \geq Y_n$, we have $E[X_n] \geq E[Y_n]$ for each n. Then,

$$\liminf_{n\to\infty} E[X_n] \geq \liminf_{n\to\infty} E[Y_n] = \lim_{n\to\infty} E[Y_n] = E[\liminf_{n\to\infty} X_n].$$

Example: Let $\Omega = (0,1)$ with the Borel σ-field. Let the probability measure P be derived from the Lesbegue measure (i.e., the measure of an interval is equal to its length). Let $X_n = n1_{(0,1/n)}$.

For any $\omega \in (0,1)$, $\lim_{n\to\infty} X_n(\omega) = 0$ and hence

$$E[\liminf_{n\to\infty} X_n] = E[\lim_{n\to\infty} X_n] = 0.$$

But, for each n, $E[X_n] = n \times \frac{1}{n} = 1$, and hence,

$$\liminf_{n\to\infty} E[X_n] = \lim_{n\to\infty} E[X_n] = 1.$$

We see $\liminf_{n\to\infty} E[X_n] \geq E[\liminf_{n\to\infty} X_n]$.

This simple example helps to remember the direction of the inequality.