Some useful results from probability theory
STA355H1S

1 Modes of Convergence

Note: In this section (and subsequently), all random variables are real-valued unless otherwise specified.

Convergence in probability. $\{X_n\}$ converges in probability to X ($X_n \xrightarrow{p} X$) if, for each $\epsilon > 0$,

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

Convergence in r-th mean. $\{X_n\}$ converges to X in r-th mean ($X_n \xrightarrow{L_r} X$) if

$$\lim_{n\to\infty} E[|X_n - X|^r] = 0.$$

(This type of convergence is also known as L_r convergence.)

Convergence in distribution. $\{X_n\}$ converges in distribution to X ($X_n \xrightarrow{d} X$) if

$$\lim_{n\to\infty} P(X_n \le x) = P(X \le x)$$

for all x where $P(X = x) = 0$.

Notes:

1. Convergence in probability of $\{X_n\}$ to X means that when n is sufficiently large, the random variable X_n is well-approximated by the random variable X. The limiting random variable X is often a constant and so if $X_n \xrightarrow{p} X$ with $P(X = a) = 1$ then $X_n \approx a$ (with probability close to 1) for sufficiently large n.

2. A somewhat stronger type of convergence is **almost sure convergence**: $\{X_n\}$ converges almost surely to X ($X_n \xrightarrow{a.s.} X$) if $X_n(\omega) \to X(\omega)$ as $n \to \infty$ for all $\omega \in A$ where $P(A) = 1$. (In other words, $P(X_n \to X) = 1$.) $X_n \xrightarrow{a.s.} X$ if, and only if, for each $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\bigcup_{k=n}^{\infty} [|X_k - X| > \epsilon]\right) = 0.$$

Almost sure convergence is important in probability theory and more advanced statistical theory, but we will not use it in this course. For completeness, we will include here a number of results that are related to almost sure convergence; some of these results turn out to be quite useful in practice.

3. It is important to note that convergence in distribution refers to convergence of probability measures (distributions) rather than the random variables themselves. For this reason, we often $F_n \xrightarrow{d} F$ where F_n is the distribution function of X_n and F is the distribution function of X. Convergence in distribution is also sometimes referred to as *weak* convergence.

4. Convergence in probability, almost sure convergence, and convergence in distribution can be defined if $\{X_n\}$, X take values in a metric space $(\mathcal{X}, d)$:

   (a) $X_n \xrightarrow{p} X$ if $d(X_n, X) \xrightarrow{p} 0$;

   (b) $X_n \xrightarrow{a.s.} X$ if $d(X_n, X) \xrightarrow{a.s.} 0$;

   (c) $X_n \xrightarrow{d} X$ if $E[g(X_n)] \to E[g(X)]$ for all bounded, continuous real-valued functions g defined on $\mathcal{X}$.

Proving convergence in distribution

For sequences of random variables, convergence in distribution refers to convergence of distribution functions and not of random variables *per se*. If $\{X_n\}$ is a sequence of random variables then $X_n \xrightarrow{d} X$ if $F_n(x) = P(X_n \leq x) \to P(X \leq x) = F(x)$ (as $n \to \infty$) for each x at which F is continous. This condition is not always easy to check; however, there are several sufficient (and in some cases, necessary and sufficient) conditions that can be use to establish $X_n \xrightarrow{d} X$.

Characteristic functions. Define $\exp(\iota x) = \cos(x) + \iota \sin(x)$ where $\iota = \sqrt{-1}$. $X_n \to X$ if, and only if, $\phi_n(t) = E[\exp(\iota t X_n)] \to E[\exp(\iota t X)] = \phi(t)$ for all t. The functions $\{\phi_n(t)\}$ and $\phi(t)$ are called the characteristic functions of $\{X_n\}$ and X.

Moment generating functions. Define moment generating functions $m_n(t) = E[\exp(tX)]$ and $m(t) = E[\exp(tX)]$. If $m_n(t) \to m(t) < \infty$ for all $t \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$ then $X_n \xrightarrow{d} X$. Similarly, if $\{X_n\}$ and X are non-negative random variables then $m_n(t) \to m(t)$ for all $t \leq 0$ implies that $X_n \xrightarrow{d} X$.

Density and probability functions. Suppose that $\{X_n\}$ and X are continuous random variables with density functions $\{f_n(x)\}$ and $f(x)$. If $f_n(x) \to f(x)$ as $n \to \infty$ for all x then $X_n \xrightarrow{d} X$. Likewise, if $\{X_n\}$ and X are discrete random variables and $P(X_n = x) \to P(X = x)$ for all x then $X_n \xrightarrow{d} X$.

As mentioned above, for $\{X_n\}$ and X taking values in more general metric spaces, we define convergence in distribution (weak convergence) in terms of expected values of bounded continuous real-valued functions: $X_n \xrightarrow{d} X$ if (and only if) $E[g(X_n)] \to E[g(X)]$ for all

bounded, continuous real-valued function g. For random variables $\{X_n\}$ and X, it turns out to be sufficient to verify convergence of $E[g(X_n)]$ to $E[g(X)]$ for all functions g belonging to a somewhat more manageable class of functions.

Relationships between modes of convergence

Almost sure convergence requires that $\{X_n\}$ and X be defined on a common probability space. The same is true in general for convergence in probability and convergence in r-th mean except when the limit X is a constant; in this case, the definitions of $\xrightarrow{p}$ and $\xrightarrow{L_r}$ do not require the X_n's to be defined on the same probability space.

The following relationships between modes of convergence hold.

(a) $X_n \xrightarrow{a.s.} X$ implies that $X_n \xrightarrow{p} X$.

(b) $X_n \xrightarrow{p} X$ implies that $X_n \xrightarrow{d} X$.

(c) $X_n \xrightarrow{L_r} X$ implies that $X_n \xrightarrow{p} X$.

(d) $X_n \xrightarrow{d} X$ implies that $X_n \xrightarrow{p} X$ if X is a constant with probability 1.

Note that neither $X_n \xrightarrow{a.s.} X$ nor $X_n \xrightarrow{p} X$ implies that $X_n \xrightarrow{L_r} X$; this seems reasonable since the expectation $E[|X_n - X|^r]$ need not be finite for any n. (See section 3 for more details.) It is easy to see that almost sure convergence is much stronger than either convergence in probability or convergence in distribution. If $X_n \xrightarrow{p} X$, then it is always possible to find a subsequence $\{X_{n_k}\}$ such that $X_{n_k} \xrightarrow{a.s.} X$ as $n_k \to \infty$. For example, pick n_k so that $P[|X_n - X| > \epsilon] \leq 2^{-k}$ for $n \geq n_k$; then

$$\begin{aligned} P\left(\lim_{k\to\infty} X_{n_k} = X\right) &= \lim_{k\to\infty} P\left(\bigcup_{j=k}^{\infty} [|X_{n_j} - X| > \epsilon]\right) \\ &\leq \lim_{k\to\infty} \sum_{j=k}^{\infty} P[|X_{n_j} - X| > \epsilon] \\ &\leq \lim_{k\to\infty} 2^{-k+1} = 0 \end{aligned}$$

and so $X_{n_k} \xrightarrow{a.s.} X$. While no similar relationship exists for convergence in distribution, the following result gives an interesting connection between convergence in distribution and almost sure convergence.

Skorokhod representation theorem. Suppose that $X_n \xrightarrow{d} X$. Then there exist random variables $\{X_n^*\}$ and X^* defined on a common probability space such that

(a) $X_n^* \stackrel{d}{=} X_n$ for all n and $X^* \stackrel{d}{=} X$ (where $\stackrel{d}{=}$ denotes equality in distribution).

(b) $X_n^* \xrightarrow{a.s.} X^*$.

The Skorokhod representation theorem also applies if $\{X_n\}$ and X are random vectors and can be extended to elements of more general spaces (for example, spaces of random functions). The proof of the result as stated here for sequences of random variables is very simple (at least conceptually). If F_n and F are the distribution functions of X_n and X, we define F_n^{-1} and F^{-1} to be their (left-continuous) inverses where, for example, $F^{-1}(t) = \inf\{x : F(x) \geq t\}$ for $0 < t < 1$. Then given a random variable U which has a uniform distribution on $(0, 1)$, we define $X_n^* = F_n^{-1}(U)$ and $X^* = F^{-1}(U)$. It is then easy to verify that $X_n^* \stackrel{d}{=} X_n$ and $X^* \stackrel{d}{=} X$ and $X_n^* \xrightarrow{a.s.} X^*$. This simple proof cannot be generalized (even to sequences of random vectors) as it takes advantage of the natural ordering of the real line.

The Skorokhod representation theorem is used almost exclusively as an analytical tool for proving convergence in distribution results. It allows one to replace sequences of random variables by sequences of numbers and thereby makes many convergence in distribution proofs completely transparent. For example, this result allows for virtually trivial proofs of both the continuous mapping theorem and the delta method stated in the section 2.

O_p and o_p notation

At this point, we will introduce some very useful notation. Suppose that $\{X_n\}$ is a sequence of random variables. We say that $\{X_n\}$ is *bounded in probability* (or *tight*) and write $X_n = O_p(1)$ if for each $\epsilon > 0$, there exists $M_\epsilon < \infty$ such that $P(|X_n| > M_\epsilon) < \epsilon$ for all n. If $\{Y_n\}$ is another sequence of random variables (or constants) then $X_n = O_p(Y_n)$ is equivalent to $X_n/Y_n = O_p(1)$.

A complement to the "O_p" notation is the "o_p" notation. We say that $X_n = o_p(1)$ (as $n \to \infty$) if $X_n \xrightarrow{p} 0$ as $n \to \infty$. Likewise $X_n = o_p(Y_n)$ means that $X_n/Y_n = o_p(1)$. This "o_p" notation is often used to denote that two sequences $\{X_n\}$ and $\{Y_n\}$ are approximately equal for sufficiently large n; for example, we can write $X_n = Y_n + o_p(1)$, which is equivalent to saying that $X_n - Y_n \xrightarrow{p} 0$.

2 Limit theorems

Continuous Mapping Theorem. Let g be a continuous function. Then

1. $X_n \xrightarrow{a.s.} X$ implies $g(X_n) \xrightarrow{a.s.} g(X)$.
2. $X_n \xrightarrow{p} X$ implies $g(X_n) \xrightarrow{p} g(X)$.
3. $X_n \xrightarrow{d} X$ implies $g(X_n) \xrightarrow{d} g(X)$.

In fact, g need not be everywhere continuous provided that the set of points A at which g is discontinuous has $P(X \in A) = 0$. Moreover, the Continuous Mapping Theorem applies for random vectors provided that g is a continuous mapping from space to the other.

The Delta Method. Let $\{a_n\}$ be a sequence of constants with $a_n \to \infty$ and suppose that $a_n(X_n - \theta) \xrightarrow{d} X$. Then if g is a function which is differentiable at θ

$$a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)X.$$

The Delta Method can be extended to, for example, real-valued functions of random vectors: If $a_n(\boldsymbol{X}_n - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{X}$ and g is a real-valued function with gradient ∇g then

$$a_n(g(\boldsymbol{X}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} [\nabla g(\boldsymbol{\theta})]^T \boldsymbol{X}.$$

Slutsky's Theorem. Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} \theta$. Then

1. $X_n + Y_n \xrightarrow{d} X + \theta$.
2. $Y_n X_n \xrightarrow{d} \theta X$.

A variation of part 2 of Slutsky's Theorem can be given using the "O_p"/"o_p" notation of the previous section. Specifically, if $X_n = O_p(1)$ and $Y_n = o_p(1)$ then $X_n Y_n = o_p(1)$ (or, equivalently, $X_n Y_n \xrightarrow{p} 0$).

Cramér-Wold device. Suppose that $\{\boldsymbol{X}_n\}$, $\boldsymbol{X}$ are k-dimensional random vectors. Then $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ if, and only if, $\boldsymbol{a}^T \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{a}^T \boldsymbol{X}$ for all k-dimensional vectors $\boldsymbol{a}$.
The Cramér-Wold device can be used to obtain a more general version of Slutsky's Theorem. Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} \theta$; then $a_1 X_n + a_2 Y_n \xrightarrow{d} a_1 X + a_2 \theta$ by the "vanilla" version of Slutsky's Theorem for any a_1, a_2 and so $(X_n, Y_n) \xrightarrow{d} (X, \theta)$ by the Cramér-Wold device. Thus if $g : R^2 \to R^p$ is a continuous function, we have $g(X_n, Y_n) \xrightarrow{d} g(X, \theta)$.

Laws of Large Numbers. Laws of large numbers give conditions under which a sample mean of random variables will converge to some appropriate population mean. The classical law of large numbers (strong and weak) is given in (a) below.

(a) Suppose that $X_1, X_2, \cdots$ be independent, identically distributed random variables with $E[|X_i|] < \infty$ and define $\mu = E(X_i)$. Then

$$\frac{1}{n}(X_1 + \cdots + X_n) \left\{ \begin{array}{c} \xrightarrow{a.s.} \\ \xrightarrow{p} \end{array} \right\} \mu$$

as $n \to \infty$.

(b) Suppose that $X_1, X_2, \cdots$ are independent random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma_i^2$. If

$$\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \to 0$$

then

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{p} \mu$$

as $n \to \infty$.

Central Limit Theorems. Under fairly broad conditions, sums of random variables will be approximately normally distributed. We give three cases where this holds for sums of independent random variables.

(a) If $X_1, X_2, \cdots$ are independent, identically distributed random variables with $E(X_i) = 0$ and $\text{Var}(X_i) = \sigma^2$ for $i = 1, \cdots, n$ then

$$Z_n = \frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n) \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

(b) If $X_1, X_2, \cdots$ are independent random variables with $E(X_i) = 0$, $\text{Var}(X_i) = \sigma_i^2$ and $E[|X_i|^3] = \gamma_i < \infty$. If

$$\frac{\gamma_1 + \cdots + \gamma_n}{(\sigma_1^2 + \cdots + \sigma_n^2)^{3/2}} \to 0$$

as $n \to \infty$ then

$$Z_n = \frac{X_1 + \cdots + X_n}{(\sigma_1^2 + \cdots + \sigma_n^2)^{1/2}} \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

(c) If $X_1, X_2, \cdots$ are independent, identically distributed random variables with $E(X_i) = 0$, $\text{Var}(X_i) = \sigma^2$, and $a_1, a_2, \cdots$ are constants satisfying the condition

$$\max_{1 \le i \le n} \frac{a_i^2}{a_1^2 + \cdots + a_n^2} \to 0 \quad \text{as } n \to \infty$$

then

$$Z_n = \frac{a_1 X_1 + \cdots + a_n X_n}{\sigma(a_1^2 + \cdots + a_n^2)^{1/2}} \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

Cases (b) and (c) give conditions under which sums of independent random variables are approximately normal when the summands do not have the same distribution. These conditions essentially imply that the amount of variability that each summand contributions to the overall variability is small so that Z_n is not dominated by a small number of summands.

Poisson convergence theorem. Suppose that $X_{n1}, \cdots, X_{nn}$ are independent random variables with $P(X_{nk} = 1) = p_{nk} = 1 - P(X_{nk} = 0)$. If

$$\sum_{k=1}^{n} p_{nk} \to \lambda > 0 \quad \text{and} \quad \max_{1 \le k \le n} p_{nk} \to 0$$

as $n \to \infty$ then

$$X_{n1} + \cdots + X_{nn} \xrightarrow{d} Y \sim \text{Poisson}(\lambda).$$

Note that the conditions on the probabilities $\{p_{nk}\}$ imply that only a small number of the $\{X_{nk}\}$ can equal 1 for large n.

3 Convergence of moments

It is often tempting to say that convergence of X_n to X (in some sense) implies convergence of moments. Unfortunately, this is not true in general; a small amount of probability in the tail of the distribution of X_n will destroy the convergence of $E(X_n)$ to $E(X)$. However, if certain bounding conditions are put on the sequence $\{X_n\}$ then convergence of moments is possible.

Fatou's Lemma. Let $\{X_n\}$ be a sequence of random variables defined on a common probability space and for each ω in the sample space, define $Y(\omega) = \liminf_{n\to\infty} |X_n(\omega)|$. Then

$$E[Y] \le \liminf_{n\to\infty} E[|X_n|].$$

(Note that if $X_n \xrightarrow{a.s.}$ some X then $Y = |X|$; thus if $X_n \xrightarrow{d} X$, it follows from the Skorokhod representation theorem that $E(|X|) \le \liminf_{n\to\infty} E(|X_n|)$.)

Dominated convergence theorem. Suppose that $X_n \xrightarrow{a.s.} X$ and $|X_n| \le Y$ (for all $n \ge$ some n_0) where $E(Y) < \infty$. Then $E(X_n) \to E(X)$.
(A similar result holds if $\xrightarrow{a.s.}$ is replaced by $\xrightarrow{d}$; see below.)

Uniform integrability. A sequence $\{X_n\}$ is uniformly integrable if

$$\lim_{x\to\infty} \limsup_{n\to\infty} E[|X_n| I(|X_n| > x)] = 0.$$

The following results involve uniform integrability:

1. If $\{X_n\}$ is uniformly integrable then $\sup_n E[|X_n|] < \infty$.
2. If $\sup_n E[|X_n|^{1+\delta}] < \infty$ for some $\delta > 0$ then $\{X_n\}$ is uniformly integrable.

3. If there exists a random variable Y with $E[|Y|] < \infty$ and $P(|X_n| > x) \leq P(|Y| > x)$ for all $n \geq$ some n_0 and $x > 0$ then $\{X_n\}$ is uniformly integrable.

4. Suppose that $X_n \xrightarrow{d} X$. If $\{X_n\}$ is uniformly integrable then $E[|X|] < \infty$ and $E(X_n) \to E(X)$. (It also follows that $E[|X_n|] \to E[|X|]$.)

4 Inequalities

Hölder's inequality. $E[|XY|] \leq E^{1/r}[|X|^r]E^{1/s}[|Y|^s]$ where $r > 1$ and $r^{-1} + s^{-1} = 1$. When $r = 2$, this inequality is called the *Cauchy-Schwarz inequality*.

Minkowski's inequality. $E^{1/r}[|X + Y|^r] \leq E^{1/r}[|X|^r] + E^{1/r}[|Y|^r]$ for $r \geq 1$.

Chebyshev's inequality. Let $g(\cdot)$ be a positive, even function which is increasing on $[0, \infty)$ (for example, $g(x) = x^2$ or $|x|$). Then for any $\epsilon > 0$

$$P(|X| > \epsilon) \leq \frac{E(g(X))}{g(\epsilon)}.$$

(Chebyshev's inequality usually refers to the special case where $g(x) = x^2$; the more general result is sometimes called Markov's inequality.)

Jensen's inequality. If $g(\cdot)$ is a convex function and $E(X)$ exists and is finite then $g(E(X)) \leq E(g(X))$.