Some useful results from probability theory
ECO2402F

1 Modes of Convergence

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

Almost sure convergence. $\{X_n\}$ converges almost surely to X ($X_n \to_{a.s.} X$) if $X_n(\omega) \to X(\omega)$ as $n \to \infty$ for all $\omega \in A$ where $P(A) = 1$. (In other words, $P(X_n \to X) = 1$.) $X_n \to_{a.s.} X$ if, and only if, for each $\epsilon > 0$,

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

Convergence in probability. $\{X_n\}$ converges in probability to X ($X_n \to_p X$) 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 ($X_n \to_{L^r} X$) 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 ($X_n \to_d X$) if

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

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

Notes:

1. An equivalent (and more general) definition of convergence in distribution says that $X_n \to_d X$ if $E[f(X_n)] \to E[f(X)]$ for all bounded, continuous functions f; this definition applies to sequences of random elements of both Euclidean and non-Euclidean metric spaces and is often useful for proving results involving convergence in distribution. We may also be able to restrict the class of bounded, continuous functions to a more manageable class; for example, for random variables $\{X_n\}$ and X, it is sufficient to prove convergence of $E[f(X_n)]$ to $E[f(X)]$ for all sinusoidal functions f.

2. 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 write $P_n \to_d P$ or $F_n \to_d F$ where P_n (F_n) is the probability measure (distribution function) of X_n and P (F) is the probability measure (distribution function) of X. Convergence in distribution is also referred to as *weak* convergence.

3. Convergence in probability and almost sure convergence can be defined if $\{X_n\}$, X take values in a metric space (with metric d): $X_n \to_p X$ ($X_n \to_{a.s.} X$) if $d(X_n, X) \to_p 0$ ($d(X_n, X) \to_{a.s.} 0$).

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 $\to_p$ and $\to_{L^r}$ do not require the X_n's to be defined on the same probability space.

The following relationships between the four modes of convergence hold.

(a) $X_n \to_{a.s.} X$ implies that $X_n \to_p X$.

(b) $X_n \to_p X$ implies that $X_n \to_d X$.

(c) $X_n \to_{L^r} X$ implies that $X_n \to_p X$.

(d) $X_n \to_d X$ implies that $X_n \to_p X$ if X is a constant with probability 1.

Note that neither $X_n \to_{a.s.} X$ nor $X_n \to_p X$ implies that $X_n \to_{L^r} X$; this seems reasonable since the expectation $E[|X_n - X|^r]$ need not be finite for any n. (See section III 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 $X_n \to_p X$, then it is always possible to find a subsequence $\{X_{n_k}\}$ such that $X_{n_k} \to_{a.s.} X$ 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 $X_{n_k} \to_{a.s.} X$. 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 $X_n \to_d X$. Then there exist random variables $\{X_n^*\}$ and X^* defined on a common probability space such that

(a) $X_n^* =_d X_n$ for all n and $X^* =_d X$ (where $=_d$ denotes equality in distribution).

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

The Skorokhod representation theorem also applies if $\{X_n\}$ and X are random vectors or, in fact, random elements of any separable and complete metric space. The proof of the result as stated here is very simple. 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 $X_n^* =_d X_n$ and $X^* =_d X$ and $X_n^* \to_{a.s.} X^*$. 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 II.

Tightness of probability measures

Suppose that $\{P_n\}$ is a sequence of probability measures defined on a Euclidean space. (We can think of P_n as the probability distribution of a random variable X_n.) We say that the sequence is *tight* if for each $\epsilon > 0$, there exists a compact (i.e. closed and bounded) set K_ϵ such that $P_n(K_\epsilon) > 1 - \epsilon$ for all n. This definition is quite general as it applies to probability measures on general metric spaces (not just the real line). In the case of probability measures on the real line, we can take the compact set K_ϵ to be a closed interval $[-M_\epsilon, M_\epsilon]$ without loss of generality.

Prohorov's Theorem. Suppose that $\{P_n\}$ is a tight sequence of probability measures on a Euclidean space (or any separable and complete metric space). Then there exists a subsequence $\{n_k\}$ such that $P_{n_k} \to_d$ some probability measure P.

There is also a slight modification of tightness called asymptotic (or uniform) tightness. We say that a sequence of probability measures $\{P_n\}$ is *asymptotically tight* if for each $\epsilon > 0$, there exists a compact set K_ϵ such that $\liminf_{n\to\infty} P_n(K_\epsilon) > 1 - \epsilon$. Prohorov's Theorem still holds if we replace "tight" by "asymptotically tight" in the statement of the theorem. (It's easy to see that tightness implies asymptotic tightness.)

Prohorov's Theorem is important as it provides us with a tool to prove the existence of a limiting probability measure. This is not particularly critical for probability measures on Euclidean spaces (such as the real line) but is quite important for non-Euclidean spaces. In many problems, we are able to identify certain characteristics (for example, moments or finite dimensional distributions) of a possible limiting measure P but it is not always clear that a measure satisfying these characteristics actually exists. However, if $\{P_n\}$ is asymptotically tight, then Prohorov's Theorem may be used to imply existence of the limiting measure P.

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* (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. Note that $X_n = O_p(1)$ is equivalent to saying that the corresponding sequence of probability measures is tight. If $\{Y_n\}$ is another sequence of random variables 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 $X_n \to_p 0$ as $n \to \infty$. Likewise $X_n = o_p(Y_n)$ means that $X_n/Y_n = o_p(1)$.

2 Limit theorems

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

1. $X_n \to_{a.s.} X$ implies $g(X_n) \to_{a.s.} g(X)$.
2. $X_n \to_p X$ implies $g(X_n) \to_p g(X)$.
3. $X_n \to_d X$ implies $g(X_n) \to_d g(X)$.

In fact, g need not be everywhere continuous provided that the set of discontinuities has P_X-probability 0. Moreover, the Continuous Mapping Theorem applies for random elements of metric spaces (for example, random vectors) provided that g is a continuous mapping from one metric space to another.

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

$$a_n(g(X_n) - g(\theta)) \to_d g'(\theta)X.$$

Slutsky's Theorem. Suppose that $X_n \to_d X$ and $Y_n \to_p \theta$. Then

1. $X_n + Y_n \to_d X + \theta$.
2. $Y_n X_n \to_d \theta X$.

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 $X_nY_n = o_p(1)$ (or, equivalently, $X_nY_n \to_p 0$).

Cramér-Wold device. Let $\{X_n\}$, X be k-dimensional random vectors. Then $X_n \to_d X$ if, and only if, $a^TX_n \to_d a^TX$ for all k-dimensional vectors a.

The Cramér-Wold device can be used to obtain a more general version of Slutsky's Theorem. Suppose that $X_n \to_d X$ and $Y_n \to_p \theta$; then $a_1X_n + a_2Y_n \to_d a_1X + a_2\theta$ by the "vanilla" version of Slutsky's Theorem for any a_1, a_2 and so $(X_n, Y_n) \to_d (X, \theta)$ by the Cramér-Wold device. Thus if $g : R^2 \to R^p$ is a continuous function, we have $g(X_n, Y_n) \to_d g(X, \theta)$.

Strong Law of Large Numbers. Let $X_1, X_2, \cdots$ be independent, identically distributed random variables with $E[|X_i|] < \infty$ and let $\mu = E(X_i)$. Then

$$\frac{1}{n}(X_1 + \cdots + X_n) \to_{a.s.} \mu$$

as $n \to \infty$.

Central Limit Theorems. Let $\{X_{ni} : n \geq 1, i = 1, \cdots, k_n\}$ be a triangular array of random variables such that for each n, $X_{n1}, \cdots, X_{nk_n}$ are independent random variables with $E(X_{ni}) = 0$ and $\mathrm{Var}(X_{ni}) = \sigma_{ni}^2$. Define

$$\sigma_n^2 = \sum_{i=1}^{k_n} \sigma_{ni}^2 \quad \text{and} \quad Z_n = \frac{1}{\sigma_n^2}\left(X_{n1} + \cdots + X_{nk_n}\right).$$

(a) Suppose that the *Lindeberg condition*

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \sum_{i=1}^{k_n} E\left[|X_{ni}|^2 I(|X_{ni}| > \epsilon\sigma_n^2)\right] = 0 \quad \text{(for each } \epsilon > 0)$$

is satisfied. Then $Z_n \to_d Z$ where Z has a standard normal distribution.

(b) Suppose that

$$\max_{1\leq i\leq k_n} \frac{\sigma_{ni}^2}{\sigma_n^2} \to 0 \quad \text{as} \quad n \to \infty.$$

Then $\{Z_n\}$ converges in distribution to a standard normal random variable if, and only if, the Lindeberg condition (given in (a)) holds.

(c) If $X_1, X_2, \cdots$ are independent, identically distributed random variables with $E(X_i) = 0$ and $\mathrm{Var}(X_i) = \sigma^2$, and $X_{ni} = X_i$ for $i = 1, \cdots, n$ then the Lindeberg condition holds and so

$$Z_n = \frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n) \to_d Z.$$

(d) Suppose that $X_1, X_2, \cdots$ are independent random variables with $E(X_i) = 0$, $\mathrm{Var}(X_i) = \sigma_i^2$ 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 the Lindeberg condition holds (again setting $X_{ni} = X_i$) and so

$$Z_n = \frac{X_1 + \cdots + X_n}{(\sigma_1^2 + \cdots + \sigma_n^2)^{1/2}} \to_d Z.$$

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

Dominated convergence theorem. Suppose that $X_n \to_{a.s.} X$ and $|X_n| \le Y$ (for all $n \ge$ some n_0) where $E(Y) < \infty$. Then $E(X_n) \to E(X)$.
(This is the Lebesgue dominated convergence theorem applied to expectations; however, we require $\{X_n\}$ and X to be defined on the same probability space. A similar result holds if $\to_{a.s.}$ is replaced by $\to_d$; 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 $P(|X_n| > x) \le P(|Y| > x)$ for all $n \ge$ some n_0 and $x > 0$ then $\{X_n\}$ is uniformly integrable.

4. Suppose that $X_n \to_d X$. If $\{X_n\}$ is uniformly integrable then $E[|X|] < \infty$ and $E(X_n) \to E(X)$. (It also follows that $E[|X_n|] \to E[|X|]$. This result is a refinement of the dominated convergence theorem; this refinement is possible since probability measures are finite measures.)

Vitali's Theorem. Suppose that $X_n \to_p X$ and $E[|X_n|^r] < \infty$ for all n and some $0 < r < \infty$. Then the following are equivalent:

(a) $\{|X_n|^r\}$ is uniformly integrable;

(b) X_n converges in r-th mean to X (that is, $E[|X_n - X|^r] \to 0$);

(c) $E[|X_n|^r] \to E[|X|^r]$.

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

Kolmogorov's inequality. Suppose that $X_1, \cdots, X_n$ are independent r.v.'s with mean 0 and finite variances and let $S_n = X_1 + \cdots + X_n$. Then for $\epsilon > 0$

$$P\left[\max_{1 \leq k \leq n} |S_k| > \epsilon\right] \leq \frac{\mathrm{Var}(S_n)}{\epsilon^2}.$$

5 Conditional probability and expectation

In order to develop a formal mathematically consistent definition of conditional probability and conditional expectation, we first need to backtrack a bit and define formally what we mean by a probability space. A probability space is formally defined to be a triple $(\Omega, \mathcal{F}, P)$

where Ω is the set of all possible outcomes (the sample space), $\mathcal{F}$ is a σ-field[1] consisting of subsets of Ω and P is a probability measure defined on the sets in $\mathcal{F}$. In measure theoretic terms, a random variable X defined on the probability space $(\Omega, \mathcal{F}, P)$ is simply a $\mathcal{F}$-measurable real-valued function.

Given a random variable X on $(\Omega, \mathcal{F}, P)$, we would like to define the conditional expectation $E(X|\mathcal{A})$ where $\mathcal{A}$ is a σ-field contained in $\mathcal{F}$. Typically, we think of $\mathcal{A}$ as representing "partial information" about the experiment. For example, if the value of a random variable Y is known then $\mathcal{A} = \sigma(Y)$, which is simply the smallest σ-field with respect to which Y is measurable; on the other hand, if no information is known then $\mathcal{A}$ consists of only two sets $\{\emptyset, \Omega\}$.

Conditional expectation. Suppose that X is a random variable with $E[|X|] < \infty$. Then $E(X|\mathcal{A})$ is defined to be an $\mathcal{A}$-measurable random variable with
(a) $E[|E(X|\mathcal{A})|] < \infty$, and (b) $E[XI(A)] = E[E(X|\mathcal{A})I(A)]$ for any $A \in \mathcal{A}$.

With the definition of conditional expectation in hand, we can, for any set $B \subset \Omega$, define $P(B|\mathcal{A}) = E[I(B)|\mathcal{A}]$. Below, we enumerate some of the properties of conditional expectation; we will assume that X is $\mathcal{F}$-measurable with $E[|X|] < \infty$.

1. If $\mathcal{A} = \mathcal{F}$ then $E(X|\mathcal{A}) = X$.
2. If $\mathcal{A} = \{\emptyset, \Omega\}$ then $E(X|\mathcal{A}) = E(X)$.
3. $E(X) = E[E(X|\mathcal{A})]$ for any $\mathcal{A}$.
4. If Y is $\mathcal{A}$-measurable (e.g. $\mathcal{A} = \sigma(Y)$) then $E(XY|\mathcal{A}) = YE(X|\mathcal{A})$.
5. If $E(X^2) < \infty$ then $\mathrm{Var}(X) \geq \mathrm{Var}[E(X|\mathcal{A})]$.

[1] A σ-field (or σ-algebra) is a class of subsets $\mathcal{F}$ of Ω satisfying the following three conditions:

1. $\Omega \in \mathcal{F}$;
2. $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$;
3. $A_1, A_2, \cdots \in \mathcal{F}$ implies $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.