Overview of Stochastic Convergence

Motivation: Functions of Random Variables

1. Let $X_1, ..., X_n$ be i.i.d. with $\mathbb{E}X_i = \mu$ and $\text{var}[X_i] = \sigma^2$. Consider:

   $$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

   - If $X_i \sim \mathcal{N}(\mu, \sigma^2)$ or $X_i \sim \exp(1/\mu)$ then know $\text{dist}[\bar{X}_n]$.
   - But $X_i$ may be from some more general distribution
   - Joint distribution of $X_i$ may not even be completely understood

   Would like to be able to say something about $\bar{X}_n$ even in those cases!

   Perhaps this is not easy for fixed $n$, but what about letting $n \to \infty$?

   \(\Rightarrow\) (a very common approach in mathematics)

2. Stochastic Convergence
   - How does a R.V. “Converge”?
   - Convergence in Probability and in Distribution

3. Useful Theorems
   - Weak Convergence of Random Vectors

4. Stronger Notions of Convergence

5. The Two “Big” Theorems

Statistical Theory (Week 2) Stochastic Convergence

Once we assume that $n \to \infty$ we start understanding $\text{dist}[\bar{X}_n]$ more:

- At a crude level $\bar{X}_n$ becomes concentrated around $\mu$

\[ \mathbb{P}[|\bar{X}_n - \mu| < \epsilon] \approx 1, \quad \forall \epsilon > 0, \text{ as } n \to \infty \]

- Perhaps more informative is to look at the “magnified difference”

\[ \mathbb{P}[\sqrt{n}(\bar{X}_n - \mu) \leq x] \overset{n \to \infty}{\approx} ? \quad \text{could yield } \mathbb{P}[\bar{X}_n \leq x] \]

More generally \(\Rightarrow\) Want to understand distribution of $Y = g(X_1, ..., X_n)$ for some general $g$:

- Often intractable
- Resort to asymptotic approximations to understand behaviour of $Y$

Warning: While lots known about asymptotics, often they are misused ($n$ small!)
Convergence of Random Variables

Need to make precise what we mean by:
- $Y_n$ is “concentrated” around $\mu$ as $n \to \infty$
- More generally what “$Y_n$ behaves like $Y$” for large $n$ means
- $\text{dist}[g(X_1, \ldots, X_n)] \overset{n \to \infty}{\approx} ?$

↔ Need appropriate notions of convergence for random variables

Recall: random variables are functions between measurable spaces

$\implies$ Convergence of random variables can be defined in various ways:
- Convergence in probability (convergence in measure)
- Convergence in distribution (weak convergence)
- Convergence with probability 1 (almost sure convergence)
- Convergence in $L^p$ (convergence in the $p$-th moment)

Each of these is qualitatively different - Some notions stronger than others

Convergence in Probability

Definition (Convergence in Probability)

Let $\{X_n\}_{n \geq 1}$ and $X$ be random variables defined on the same probability space. We say that $X_n$ converges in probability to $X$ as $n \to \infty$ (and write $X_n \overset{p}{\to} X$) if for any $\epsilon > 0$,

$$\mathbb{P}[|X_n - X| > \epsilon] \overset{n \to \infty}{\to} 0.$$ Intuitively, if $X_n \overset{p}{\to} X$, then with high probability $X_n \approx X$ for large $n$.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}[0, 1]$, and define $M_n = \max\{X_1, \ldots, X_n\}$. Then,

$$F_{M_n}(x) = x^n \implies \mathbb{P}[|M_n - 1| > \epsilon] = \mathbb{P}[M_n < 1 - \epsilon] = (1 - \epsilon)^n \overset{n \to \infty}{\to} 0$$

for any $0 < \epsilon < 1$. Hence $M_n \overset{p}{\to} 1$.

Some Comments on “$\overset{p}{\to}$” and “$\overset{d}{\to}$”

- Convergence in probability implies convergence in distribution.
- Convergence in distribution does NOT imply convergence in probability
  ⇐ Consider $X \overset{d}{\sim} \mathcal{N}(0, 1), -X + \frac{1}{n} \overset{d}{\to} X$ but $-X + \frac{1}{n} \overset{p}{\to} -X$.
- “$\overset{d}{\to}$” relates distribution functions
  ⇐ Can use to approximate distributions (approximation error?).
- Both notions of convergence are metrizable
  ⇐ i.e. there exist metrics on the space of random variables and distribution functions that are compatible with the notion of convergence.
  ⇐ Hence can use things such as the triangle inequality etc.
- “$\overset{d}{\to}$” is also known as “weak convergence” (will see why).

Equivalent Def: $X \overset{d}{\to} X \iff \mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all measurable $f$. 

Convergence in Distribution

Definition (Convergence in Distribution)

Let $\{X_n\}_{n \geq 1}$ and $X$ be random variables (not necessarily defined on the same probability space). We say that $X_n$ converges in distribution to $X$ as $n \to \infty$ (and write $X_n \overset{d}{\to} X$) if for any continuous and bounded $f$

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X) \quad \forall \text{cts and bounded } f$$

Statistical Theory (Week 2) Stochastic Convergence 5 / 21

Statistical Theory (Week 2) Stochastic Convergence 6 / 21

Statistical Theory (Week 2) Stochastic Convergence 7 / 21

Statistical Theory (Week 2) Stochastic Convergence 8 / 21
**Theorem**

(a) \( X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X \)

(b) \( X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c, c \in \mathbb{R} \).

**Proof**

Let \( x \) be a continuity point of \( F_X \) and \( \epsilon > 0 \). Then,

\[
P[X_n \leq x] = P[X_n \leq x, |X_n - X| \leq \epsilon] + P[X_n \leq x, |X_n - X| > \epsilon]
\]

\[
\leq P[X \leq x + \epsilon] + P[|X_n - X| > \epsilon]
\]

since \( \{X \leq x + \epsilon\} \) contains \( \{X_n \leq x, |X_n - X| \leq \epsilon\} \). Similarly,

\[
P[X \leq x - \epsilon] = P[X \leq x - \epsilon, |X_n - X| \leq \epsilon] + P[X \leq x - \epsilon, |X_n - X| > \epsilon]
\]

\[
\leq P[X_n \leq x] + P[|X_n - X| > \epsilon]
\]

which yields \( P[|X_n - X| > \epsilon] \)

\[
\leq P[X \leq x - \epsilon, |X_n - X| \leq \epsilon] + P[|X_n - X| > \epsilon]
\]

Combining the two inequalities and “sandwiching” yields the result.

(b) Let \( F \) be the distribution function of a constant \( c \). Then

\[
F(x) = P[c \leq x] = \begin{cases} 1 & \text{if } x \geq c, \\ 0 & \text{if } x < c. \end{cases}
\]

**Exercise**

Prove part (a). You may assume without proof the Subsequence Lemma: \( X_n \xrightarrow{p} X \) if and only if every subsequence \( X_{nm} \) of \( X_n \) has a further subsequence \( X_{nm(k)} \) such that \( P[|X_{nm(k)} - X| \to 0] = 1 \).

**Theorem (Slutsky’s Theorem)**

Let \( X_n \xrightarrow{d} X \) and \( Y_n \xrightarrow{d} c \in \mathbb{R} \). Then

(a) \( X_n + Y_n \xrightarrow{d} X + c \)

(b) \( X_n Y_n \xrightarrow{d} cX \)

**Proof of Slutsky’s Theorem.**

(a) We may assume \( c = 0 \). Let \( x \) be a continuity point of \( F_X \). We have

\[
P[X_n + Y_n \leq x] = P[X_n + Y_n \leq x, |Y_n| \leq \epsilon] + P[X_n + Y_n \leq x, |Y_n| > \epsilon]
\]

\[
\leq P[X_n \leq x + \epsilon] + P[|Y_n| > \epsilon]
\]

Similarly,

\[
P[X_n \leq x - \epsilon] \leq P[X_n + Y_n \leq x] + P[|Y_n| > \epsilon]
\]

Therefore,

\[
P[X_n \leq x - \epsilon] - P[|Y_n| > \epsilon] \leq P[X_n + Y_n \leq x] \leq P[X_n \leq x + \epsilon] + P[|Y_n| > \epsilon]
\]

Taking \( n \to \infty \), and then \( \epsilon \to 0 \) proves (a).

(b) By (a) we may assume that \( c = 0 \) (check). Let \( \epsilon, M > 0 \):

\[
P[|X_n Y_n| > \epsilon] \leq P[|X_n Y_n| > \epsilon, |Y_n| \leq 1/M] + P[|Y_n| \geq 1/M]
\]

\[
\leq P[|X_n| > \epsilon M] + P[|Y_n| \geq 1/M]
\]

\[
\to P[|X| > \epsilon M] + 0
\]

The first term can be made arbitrarily small by letting \( M \to \infty \).
Theorem (General Version of Slutsky’s Theorem)

Let \( g : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \) be continuous and suppose that \( X_n \xrightarrow{d} X \) and \( Y_n \xrightarrow{d} c \in \mathbb{R} \). Then, \( g(X_n, Y_n) \to g(X, c) \) as \( n \to \infty \).

\[ \text{\footnotesize Notice that the general version of Slutsky’s theorem does not follow immediately from the continuous mapping theorem.} \]

- The continuous mapping theorem would be applicable if \( (X_n, Y_n) \) weakly converged jointly (i.e. their joint distribution) to \( (X, c) \).
- But here we assume only marginal convergence (i.e. \( X_n \xrightarrow{d} X \) and \( Y_n \xrightarrow{d} c \) separately, but their joint behaviour is unspecified).
- The key of the proof is that in the special case where \( Y_n \xrightarrow{d} c \) where \( c \) is a constant, then marginal convergence \( \iff \) joint convergence.
- However if \( X_n \xrightarrow{d} X \) where \( X \) is non-degenerate, and \( Y_n \xrightarrow{d} Y \) where \( Y \) is non-degenerate, then the theorem fails.
- Notice that even the special cases (addition and multiplication) of Slutsky’s theorem fail of both \( X \) and \( Y \) are non-degenerate.

Exercise: Give a counterexample to show that neither of \( X_n \xrightarrow{p} X \) or \( X_n \xrightarrow{d} X \) ensures that \( \mathbb{E}X_n \to \mathbb{E}X \) as \( n \to \infty \).

Theorem (Convergence of Expectations)

If \( |X_n| < M < \infty \) and \( X_n \xrightarrow{d} X \), then \( \mathbb{E}X \) exists and \( \mathbb{E}X_n \xrightarrow{a.s.} \mathbb{E}X \).

Proof.

Assume first that \( X_n \) are non-negative \( \forall n \). Then,

\[
|\mathbb{E}X_n - \mathbb{E}X| = \left| \int_0^M \mathbb{P}[X_n > x] - \mathbb{P}[X > x] \, dx \right| \\
\leq \int_0^M |\mathbb{P}[X_n > x] - \mathbb{P}[X > x]| \, dx \xrightarrow{n \to \infty} 0.
\]

since \( M < \infty \) and the integration domain is bounded.

Exercise: Generalise the proof to arbitrary random variables.

Remarks on Weak Convergence

- Often difficult to establish weak convergence directly (from definition)
- Indeed, if \( F_n \) known, establishing weak convergence is “useless”
- Need other more “handy” sufficient conditions

Scheffé’s Theorem

Let \( X_n \) have density functions (or probability functions) \( f_n \), and let \( X \) have density function (or probability function) \( f \). Then

\[ f_n \xrightarrow{n \to \infty} f \text{ (a.e.)} \iff X_n \xrightarrow{d} X \]

- The converse to Scheffé’s theorem is NOT true (why?).

Continuity Theorem

Let \( X_n \) and \( X \) have characteristic functions \( \varphi_n(t) = \mathbb{E}[e^{itX_n}] \), and \( \varphi(t) = \mathbb{E}[e^{itX}] \), respectively. Then,

(a) \( X_n \xrightarrow{d} X \iff \varphi_n \to \varphi \text{ pointwise} \)
(b) If \( \varphi_n(t) \) converges pointwise to some limit function \( \psi(t) \) that is continuous at zero, then:

(i) \( \exists \) a measure \( \nu \) with c.f. \( \psi \)
(ii) \( F_{X_n} \xrightarrow{w} \nu \).

Proof.

Taylor expanding around \( \theta \) gives:

\[ g(X_n) = g(\theta_n^*) + g'(\theta_n^*)(X_n - \theta_n^*) \]

Thus \( \theta_n^* \to \theta \). By the continuous mapping theorem \( g'(\theta_n^*) \xrightarrow{p} g'(\theta) \).

Thus \( a_n(g(X_n) - g(\theta)) = a_n(g(\theta_n^*) + g'(\theta_n^*)(X_n - \theta_n^*) - g(\theta)) = g'(\theta_n^*)a_n(X - \theta) \xrightarrow{d} g'(\theta)Z \).

The delta method actually applies even when \( g'(\theta) \) is not continuous (proof uses Skorokhod representation).
Stochastic Convergence

Almost Sure Convergence and Convergence in $L^p$

There are also two stronger convergence concepts (that do not compare)

**Definition (Almost Sure Convergence)**

Let $(X_n)_{n \geq 1}$ and $X$ be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $A := \{\omega \in \Omega : X_n(\omega) \xrightarrow{n \to \infty} X(\omega)\}$. We say that $X_n$ converges almost surely to $X$ as $n \to \infty$ (and write $X_n \overset{a.s.}{\to} X$ if $\mathbb{P}[A] = 1$.

More plainly, we say $X_n \overset{a.s.}{\to} X$ if $\mathbb{P}[X_n \to X] = 1$.

**Definition (Convergence in $L^p$)**

Let $(X_n)_{n \geq 1}$ and $X$ be random variables defined on the same probability space. We say that $X_n$ converges to $X$ in $L^p$ as $n \to \infty$ (and write $X_n \xrightarrow{L^p} X$) if

$$\mathbb{E}|X_n - X|^p \xrightarrow{n \to \infty} 0.$$  

Note that $||X||_{L^p} := (\mathbb{E}|X|^p)^{1/p}$ defines a complete norm (when finite).

**Recalling two basic Theorems**

**Multivariate Random Variables $\to$ “$d$“ defined coordinatewise**

**Theorem (Strong Law of Large Numbers)**

Let $(X_n)_{n \geq 1}$ be pairwise iid random variables with $\mathbb{E}X_k = \mu$ and $\mathbb{E}|X_k| < \infty$, for all $k \geq 1$. Then,

$$\frac{1}{n} \sum_{k=1}^{n} X_k \overset{a.s.}{\to} \mu.$$ 

- “Strong” is as opposed to the “weak” law which requires $\mathbb{E}X_k^2 < \infty$ instead of $\mathbb{E}|X_k| < \infty$ and gives “$p$” instead of “$a.s.$”

**Theorem (Central Limit Theorem)**

Let $(X_n)$ be an iid sequence of random vectors in $\mathbb{R}^d$ with mean $\mu$ and covariance $\Sigma$ and define $\bar{X}_n := \sum_{m=1}^{n} X_m/n$. Then,

$$\sqrt{n}\Sigma^{-\frac{1}{2}}(\bar{X} - \mu) \overset{d}{\to} Z \sim \mathcal{N}_d(0, I_d).$$

**Relationship Between Different Types of Convergence**

- $X_n \overset{a.s.}{\to} X \implies X_n \overset{p}{\to} X \implies X_n \overset{d}{\to} X$
- $X_n \overset{L^p}{\to} X$, for $p > 0 \implies X_n \overset{p}{\to} X \implies X_n \overset{d}{\to} X$
- for $p \geq q$, $X_n \overset{L^p}{\to} X \implies X_n \overset{L^q}{\to} X$
- There is no implicative relationship between “$a.s.$” and “$L^p$”

**Definition**

Let $(X_n)$ be a sequence of random vectors of $\mathbb{R}^d$, and $X$ a random vector of $\mathbb{R}^d$ with $X_n = (X_n^{(1)}, \ldots, X_n^{(d)})^T$ and $X = (X^{(1)}, \ldots, X^{(d)})^T$. Define the distribution functions $F_{X_n}(x) = \mathbb{P}[X_n^{(1)} \leq x^{(1)}, \ldots, X_n^{(d)} \leq x^{(d)}]$ and $F_X(x) = \mathbb{P}[X^{(1)} \leq x^{(1)}, \ldots, X^{(d)} \leq x^{(d)}]$, for $x = (x^{(1)}, \ldots, x^{(d)})^T \in \mathbb{R}^d$. We say that $X_n$ converges in distribution to $X$ as $n \to \infty$ (and write $X_n \xrightarrow{d} X$) if for every continuity point of $F_X$ we have

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

There is a link between univariate and multivariate weak convergence:

**Theorem (Cramér-Wold Device)**

Let $(X_n)$ be a sequence of random vectors of $\mathbb{R}^d$, and $X$ a random vector of $\mathbb{R}^d$. Then, $X_n \overset{d}{\to} X \iff \theta^T X_n \overset{d}{\to} \theta^T X$, $\forall \theta \in \mathbb{R}^d$.

**Exercise**

Prove part (b) of the continuous mapping theorem.

Weak Convergence of Random Vectors
Convergence Rates

Often convergence not enough → How fast?
↪→ [quality of approximation]

- Law of Large Numbers: assuming finite variance, $L^2$ rate of $n^{-1/2}$
- What about Central Limit Theorem?

**Theorem (Berry-Essen)**

Let $X_1, ..., X_n$ be iid random vectors taking values in $\mathbb{R}^d$ and such that $\mathbb{E}[X_i] = 0$, $\text{cov}[X_i] = I_d$. Define,

$$S_n = \frac{1}{\sqrt{n}}(X_1 + ... + X_n).$$

If $A$ denotes the class of convex subsets of $\mathbb{R}^d$, then for $Z \sim \mathcal{N}_d(0, I_d)$,

$$\sup_{A \in A} |\mathbb{P}[S_n \in A] - \mathbb{P}[Z \in A]| \leq C d^{1/4} \frac{\mathbb{E}\|X_i\|^3}{\sqrt{n}}.$$