Overview of Stochastic Convergence

Statistical Theory

Guillaume Dehaene
Ecole Polytechnique Fédérale de Lausanne

Motivation: Functions of Random Variables

Functions of Random Variables

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 $X_i \sim \exp(\lambda = 1/\mu)$ 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 $n \to \infty$?
$\hookrightarrow$(a very common approach in mathematics)

Functions of Random Variables

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 μ

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

- 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 $\longrightarrow$ Want to understand distribution of $Y = g(X_1, \ldots, X_n)$ for some general g:

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

Warning: asymptotics are often abused (n small!)

Stochastic Convergence

Convergence of Random Variables

Need to make precise what we mean by:

- Y_n is "concentrated" around μ as $n \to \infty$
- More generally what "Y_n behaves like Y" for large n means
- $\text{dist}[g(X_1, ..., X_n)] \overset{n\to\infty}{\approx}$?

$\hookrightarrow$ Need appropriate notions of convergence for random variables

Recall: random variables are *functions* between *measurable spaces*

$\Longrightarrow$ 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 $X_n \xrightarrow{p} X$) if for any $\epsilon > 0$,

$$\mathbb{P}[|X_n - X| > \epsilon] \xrightarrow{n\to\infty} 0.$$

Intuitively, if $X_n \xrightarrow{p} X$, 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 $M_n = \max\{X_1, ..., X_n\}$. Then,

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

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

Convergence in Probability

Lemma (Ky-Fan definition of convergence in probability)

$X_n \xrightarrow{p} X$ if and only if there exists some $\alpha_n \downarrow 0$ such that

$$\mathbb{P}[|X_n - X| > \alpha_n] \leq \alpha_n, \qquad \forall\, n \geq 1.$$

Proof.

Suppose that there exists such an α_n. Then given any $\epsilon > 0$, there exists n sufficiently large so that $\alpha_n < \epsilon$. It follows that

$$\mathbb{P}[|X_n - X| > \epsilon] \leq \mathbb{P}[|X_n - X| > \alpha_n] \leq \alpha_n \overset{n\to\infty}{\longrightarrow} 0.$$

and thus $X_n \xrightarrow{p} X$. For the converse, suppose that $X_n \xrightarrow{p} X$. Then, $\exists\ \{n_k\}_{k\geq 1}$ s.t.

$$n_k < n_{k+1}, \quad \& \quad \mathbb{P}[|X_n - X| > 1/k] \leq \frac{1}{k}, \forall\, n \geq n_k.$$

Define $\alpha_n = \sum_{k=1}^{\infty} \frac{1}{k} \mathbf{1}\{n_k \leq n < n_{k+1}\}$, and observe that $\mathbb{P}[|X_n - X| > \alpha_n] \leq \alpha_n$, for all $n \geq 1$, and $\alpha_n \downarrow 0$, thus completing the proof. $\square$

Convergence in Distribution

Definition (Convergence in Distribution)

Let $\{X_n\}$ 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 $X_n \xrightarrow{d} X$) if

$$\mathbb{P}[X_n \leq x] \overset{n\to\infty}{\longrightarrow} \mathbb{P}[X \leq x],$$

at every continuity point of $F_X(x) = \mathbb{P}[X \leq x]$.

Example

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{U}[0,1]$, $M_n = \max\{X_1, ..., X_n\}$, and $Q_n = n(1 - M_n)$.

$$\mathbb{P}[Q_n \leq x] = \mathbb{P}[M_n \geq 1 - x/n] = 1 - \left(1 - \frac{x}{n}\right)^n \overset{n\to\infty}{\longrightarrow} 1 - e^{-x}$$

for all $x \geq 0$. Hence $Q_n \xrightarrow{d} Q$, with $Q \sim exp(1)$.

Some Comments on "$\xrightarrow{p}$" and "$\xrightarrow{d}$"

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

Equivalent Def: $X \xrightarrow{d} X \iff \mathbb{E}f(X_n) \to \mathbb{E}f(X)\ \forall$ cts and bounded f

Useful Theorems

Some Basic Results

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

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

$$\begin{aligned} \mathbb{P}[X_n \leq x] &= \mathbb{P}[X_n \leq x, |X_n - X| \leq \epsilon] + \mathbb{P}[X_n \leq x, |X_n - X| > \epsilon] \\ &\leq \mathbb{P}[X \leq x + \epsilon] + \mathbb{P}[|X_n - X| > \epsilon] \end{aligned}$$

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

$$\begin{aligned} \mathbb{P}[X \leq x - \epsilon] &= \mathbb{P}[X \leq x - \epsilon, |X_n - X| \leq \epsilon] + \mathbb{P}[X \leq x - \epsilon, |X_n - X| > \epsilon] \\ &\leq \mathbb{P}[X_n \leq x] + \mathbb{P}[|X_n - X| > \epsilon] \end{aligned}$$

(proof cont'd).

which yields $\mathbb{P}[X \leq x - \epsilon] - \mathbb{P}[|X_n - X| > \epsilon] \leq \mathbb{P}[X_n \leq x]$.

Combining the two inequalities and "sandwitching" yields the result.

(b) Let F be the distribution function of a constant r.v. c,

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

$$\begin{aligned}
\mathbb{P}[|X_n - c| > \epsilon] &= \mathbb{P}[\{X_n - c > \epsilon\} \cup \{c - X_n > \epsilon\}] \\
&= \mathbb{P}[X_n > c + \epsilon] + \mathbb{P}[X_n < c - \epsilon] \\
&\leq 1 - \mathbb{P}[X_n \leq c + \epsilon] + \mathbb{P}[X_n \leq c - \epsilon] \\
&\xrightarrow{n\to\infty} 1 - F(\underbrace{c + \epsilon}_{\geq c}) + F(\underbrace{c - \epsilon}_{< c}) = 0
\end{aligned}$$

Since $X_n \xrightarrow{d} c$. $\square$

Theorem (Continuous Mapping Theorem)

Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then,

(a) $X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X)$

(b) $Y_n \xrightarrow{d} Y \implies g(Y_n) \xrightarrow{d} g(Y)$

Exercise

Prove part (a). You may assume without proof the *Subsequence Lemma*: $X_n \xrightarrow{p} X$ if and only if every subsequence X_{n_m} of X_n, has a further subsequence $X_{n_{m(k)}}$ such that $\mathbb{P}[X_{n_{m(k)}} \xrightarrow{k\to\infty} X] = 1$.

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

$$\begin{aligned}\mathbb{P}[X_n + Y_n \leq x] &= \mathbb{P}[X_n + Y_n \leq x, |Y_n| \leq \epsilon] + \mathbb{P}[X_n + Y_n \leq x, |Y_n| > \epsilon] \\ &\leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]\end{aligned}$$

Similarly, $\quad \mathbb{P}[X_n \leq x - \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x] + \mathbb{P}[|Y_n| > \epsilon], \quad$ therefore,

$$\mathbb{P}[X_n \leq x - \epsilon] - \mathbb{P}[|Y_n| > \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x] \leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]$$

Since ϵ is arbitrary, this proves (a) by taking $n \to \infty$.

$\square$

Proof.

(b) It suffices to assume $c = 0$ (since $(Y_n + c)X_n = X_n Y_n + X_n c$, so if we can show $X_n Y_n \stackrel{d}{\to} 0$, then (a) gives conclusion). Let $\epsilon, M > 0$:

$$\begin{aligned} \mathbb{P}[|X_n Y_n| > \epsilon] &\leq \mathbb{P}[|X_n Y_n| > \epsilon, |Y_n| \leq 1/M] + \mathbb{P}[|Y_n| \geq 1/M] \\ &\leq \mathbb{P}[|X_n| > \epsilon M] + \mathbb{P}[|Y_n| \geq 1/M] \\ &\xrightarrow{n\to\infty} \mathbb{P}[|X| > \epsilon M] + 0 \end{aligned}$$

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 $n \to \infty$.

$\hookrightarrow$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.

Theorem (The Delta Method)

Let $Z_n := a_n(X_n - \theta) \xrightarrow{d} Z$ *where* $a_n, \theta \in \mathbb{R}$ *for all* n *and* $a_n \uparrow \infty$. *Let* $g(\cdot)$ *be continuously differentiable at* θ. *Then,* $a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z$.

Proof

Taylor expanding around θ gives:

$$g(X_n) = g(\theta) + g'(\theta_n^*)(X_n - \theta), \quad \theta_n^* \text{ between } X_n, \theta.$$

Thus $|\theta_n^* - \theta| < |X_n - \theta| = a_n^{-1} \cdot |a_n(X_n - \theta)| = a_n^{-1} Z_n \xrightarrow{p} 0$ [by Slutsky]
Therefore, $\theta_n^* \xrightarrow{p} \theta$. By the continuous mapping theorem $g'(\theta_n^*) \xrightarrow{p} g'(\theta)$.

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

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

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 Expecations)

If $|X_n| < M < \infty$ and $X_n \xrightarrow{d} X$, then $\mathbb{E}X$ exists and $\mathbb{E}X_n \xrightarrow{n\to\infty} \mathbb{E}X$.

Proof.

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

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

since $M < \infty$ and the integration domain is bounded. $\square$

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

$$f_n \stackrel{n\to\infty}{\longrightarrow} f \text{ (a.e.)} \implies X_n \stackrel{d}{\to} X$$

- 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) $X_n \stackrel{d}{\to} X \Leftrightarrow \phi_n \to \phi$ pointwise

(b) If $\phi_n(t)$ converges pointwise to some limit function $\psi(t)$ that is continuous at zero, then:

(i) $\exists$ a measure ν with c.f. ψ

(ii) $F_{X_n} \stackrel{w}{\to} \nu$.

Weak Convergence of Random Vectors

Definition

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

$$F_{\mathbf{X}_n}(\mathbf{X}) \xrightarrow{n\to\infty} F_{\mathbf{X}}(\mathbf{x}).$$

There is a link between univariate and multivariate weak convergence:

Theorem (Cramér-Wold Device)

Let $\{\mathbf{X}_n\}$ be a sequence of random vectors of $\mathbb{R}^d$, and $\mathbf{X}$ a random vector of $\mathbb{R}^d$. Then,

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \Leftrightarrow \boldsymbol{\theta}^\mathsf{T}\mathbf{X}_n \xrightarrow{d} \boldsymbol{\theta}^\mathsf{T}\mathbf{X}, \ \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

Stronger Notions of 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 $A := \{\omega \in \Omega : X_n(\omega) \overset{n\to\infty}{\to} X(\omega)\}$. We say that X_n converges almost surely to X as $n \to \infty$ (and write $X_n \xrightarrow{a.s.} X$) if $\mathbb{P}[A] = 1$.

More plainly, we say $X_n \xrightarrow{a.s.} X$ 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 $X_n \overset{L^p}{\to} X$) 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)

Relationship Between Different Types of Convergence

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

Theorem (Skorokhod's Representation Theorem)

Let $\{X_n\}_{n\geq 1}, X$ be random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $X_n \xrightarrow{d} X$. Then, there exist random variables $\{Y_n\}_{n\geq 1}, Y$ defined on some probability space $(\Omega', \mathcal{G}, \mathbb{Q})$ such that:

(i) $Y \stackrel{d}{=} X$ & $Y_n \stackrel{d}{=} X_n$, $\forall n \geq 1$,

(ii) $Y_n \xrightarrow{a.s.} Y$.

Exercise

Prove part (b) of the continuous mapping theorem.

The Two "Big" Theorems

Recalling two basic Theorems

Multivariate Random Variables $\rightarrow$ " $\xrightarrow{d}$ " defined coordinatewise

Theorem (Strong Law of Large Numbers)

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

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

- "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 "$\xrightarrow{p}$" instead of "$\xrightarrow{a.s.}$"
- This is insanely strong: $(E)|X_k| < \infty$ is the weakest condition for it to have an expected value. The theorem reads: if there is an expected value, we can find it with the empirical mean.
- The strong law says **nothing useful** about the **size** of the error.

Recalling two basic theorems

Theorem (Central Limit Theorem)

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

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

- Insanely strong theorem: as soon as the covariance exist, we are in business.
- Once more, no control on the size of error.
- There are many variants of this basic CLT.

Convergence Rates

The mathematician rarely cares about convergence speed. The statistician does (should?) because **data is money**.

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

The Berry-Esseen theorem

Theorem (Berry-Essen {Bentkus, 2005, Theory Prob Appl})

Let $\mathbf{X}_1, ..., \mathbf{X}_n$ *be iid random vectors taking values in* $\mathbb{R}^d$ *and such that* $\mathbb{E}[\mathbf{X}_i] = 0$, $cov[\mathbf{X}_i] = I_d$. *Define,*

$$\mathbf{S}_n = \frac{1}{\sqrt{n}}(\mathbf{X}_1 + \ldots + \mathbf{X}_n).$$

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

$$\sup_{A \in \mathcal{A}} |\mathbb{P}[\mathbf{S}_n \in A] - \mathbb{P}[\mathbf{Z} \in A]| \leq C \frac{d^{1/4}\mathbb{E}\|\mathbf{X}_i\|^3}{\sqrt{n}}.$$

The constant C *is universal.* $C \leq 4$.

Using this theorem, we can construct confidence regions with guaranteed coverage.