Statistical Theory — Week 1
Overview of Stochastic Convergence

Course contents

- Likelihood theory
- Decision theory
- Limitation of the likelihood approach
- Some nonparametric approaches
- Optimal transport

Today: stochastic convergence

Functions of random variables

Let $X_1, \ldots, X_n$ be identically distributed with $E[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$, and consider

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

- If the X_i are independent and $X_i \sim N(\mu, \sigma^2)$ or $X_i \sim \text{Exp}(\lambda = 1/\mu)$ or $X_i \sim \text{Poisson}(\mu)$ then we know $\text{dist}[\bar{X}_n]$.
- But the X_i may be from some more general distribution.
- The joint distribution of X_i may not even be completely understood/known.

We 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 $n \to \infty$?

- A very common approach in mathematics.

Functions of random variables

- Once we assume $n \to \infty$ we start to better understand dist$[\overline{X}_n]$:
    - It concentrates around μ, by Chebishev: for $\epsilon > 0$

$$P[|\overline{X}_n - \mu| < \epsilon] \geq 1 - \frac{\sigma^2}{n\epsilon} \to 1, \qquad n \to \infty.$$

    - A rate of convergence can be understood as a sequence $r_n \to \infty$ such that

$$P[r_n(\overline{X}_n - \mu) \leq x] \to ?$$

  which could provide statements about $P(\overline{X}_n \leq t)$.

- More generally, we want to understand distribution of $Y_n = g(X_1, \ldots, X_n)$ for some function g
    - Often infeasible.
    - Thus, we resort to asymptotic approximations to understand it
- Such approximations are appropriate if n is large (perhaps not if n is small!)

Convergence of random variables

- Need to make precise what we mean by "Y_n is close to Y" for n large.
- Recall that random variables are functions between measurable spaces. $\Rightarrow$ Convergence of random variables can be defined in various ways:

- Convergence in probability (convergence in measure).
- Convergence in distribution (weak / narrow convergence).
- Convergence with probability 1 (almost sure convergence).
- Convergence in L_p (convergence in the p-th moment).

All these notions are qualitatively different. Some modes of convergence are stronger than others.

Convergence in probability

Definition (Convergence in probability)

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

$$P[\|X_n - X\| > \varepsilon] \to 0 \quad \textit{as } n \to \infty.$$

Intuitively, if $X_n \xrightarrow{p} X$, then for large n, $X_n \approx X$ with probability close to 1.

Example

Let $X_1, \ldots, X_n$ be iid Unif$(0,1)$, and define $M_n = \max\{X_1, ..., X_n\}$. Then $F_{M_n}(x) = x^n$ for $x \in [0,1]$ and

$$P[|M_n - 1| > \varepsilon] = P[M_n < 1 - \varepsilon] = (1-\varepsilon)^n \to 0, \quad \text{for any } 0 < \varepsilon < 1.$$

Hence, $M_n \xrightarrow{p} 1$.

Convergence in distribution

Definition (Convergence in distribution)

Let $\{X_n\}$ and X be random vectors (not necessarily defined on the same probability space). We say that X_n converges in distribution to X as $n \to \infty$ (denoted $X_n \xrightarrow{d} X$) if

$$P[X_n \leq x] \to P[X \leq x], \quad \textit{at every continuity point of } F_X(x) = P[X \leq x].$$

Restriction to continuity points is important.

Example

Let $X_1, \ldots, X_n$ be iid Unif$(0,1)$, $M_n = \max\{X_1, \ldots, X_n\}$, and define $Q_n = n(1 - M_n)$. Then,

$$P[Q_n \leq x] = P[M_n \geq 1 - x/n] = 1 - (1 - x/n)^n \to 1 - e^{-x}, \quad \text{for all } x \geq 0.$$

Hence, $Q_n \xrightarrow{d} Q$, where $Q \sim$ Exp(1).

Some comments on convergence in distribution and in probability

- Convergence in probability involves the random vectors themselves.
- Convergence in distributions pertains only to the distribution functions.
  $\hookrightarrow$ Can be used to approximate distributions (approximation error?)
- Both notions of convergence are metrisable : There exist metrics on the space of random vectors and on the space of distribution functions that are compatible with these notions of convergence.
- Convergence in probability implies convergence in distribution.
- Convergence in distribution does NOT imply convergence in probability: let $X \sim N(0,1)$, $X_n = -X + 1/n \xrightarrow{d} X$ but $X_n \xrightarrow{p} -X \neq X$.

Portmanteau lemma

$X_n \xrightarrow{d} X \iff \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all real-valued continuous and bounded functions f.
(More on this later)

Theorem

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

(b) *For any constant* $c \in \mathbb{R}^k$, $X_n \xrightarrow{d} c \Rightarrow X_n \xrightarrow{p} c$.

Proof. **(a)** Let x be a continuity point of F_X. Then, for any $\varepsilon > 0$,

$$\begin{aligned} P[X_n \le x] &= P[X_n \le x, \|X_n - X\| \le \varepsilon] + P[X_n \le x, \|X_n - X\| > \varepsilon] \\ &\le P[X \le x + \varepsilon] + P[\|X_n - X\| > \varepsilon]. \end{aligned}$$

Similarly,

$$\begin{aligned} P[X \le x - \varepsilon] &\le P[X_n \le x, \|X_n - X\| \le \varepsilon] + P[\|X_n - X\| > \varepsilon] \\ &\le P[X_n \le x] + P[\|X_n - X\| > \varepsilon]. \end{aligned}$$

Conclude that

$$P[X \le x-\varepsilon] - P[\|X_n - X\| > \varepsilon] \le P(X_n \le x) \le P[X \le x+\varepsilon] + P[\|X_n - X\| > \varepsilon]$$

Taking limits as $n \to \infty$ and $\varepsilon \to 0$ gives the result (exercise).

Proof (cont'd)

(b) Assume first X_n are one-dimensional, so $c \in \mathbb{R}$.
Let F be the CDF of a degenerate random variable at c. Then

$$F(x) = \begin{cases} 1, & x \geq c \\ 0, & x < c \end{cases}.$$

We have

$$\begin{aligned} P[|X_n - c| > \varepsilon] &= P[X_n > c + \varepsilon] + P[X_n < c - \varepsilon] \\ &= 1 - P[X_n \leq c + \varepsilon] + P[X_n \leq c - \varepsilon] \\ &\to 1 - 1 + 0, \qquad n \to \infty \end{aligned}$$

since $c \pm \varepsilon$ are continuity points of F.
If $X_n, c \in \mathbb{R}^k$, work on each coordinate separately and use

$$\{\|X_n - c\| > \varepsilon\} \subset \bigcup_{i=1}^{k} \{|X_n^i - c^i| > \sqrt{\varepsilon/k}\}.$$

Almost Sure Convergence and Convergence in L_p

Let $\{X_n\}$ and X be random vectors on the same probability space (Ω, F, P).

Definition (almost sure convergence)

We say X_n converges almost surely to X (denoted $X_n \xrightarrow{a.s.} X$) if

$$P[X_n \to X] := P(\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}) = 1.$$

Almost sure convergence is NOT metrisable.

Definition (convergence in L_p)

For $p > 0$, X_n converges to X in L_p (denoted $X_n \xrightarrow{L_p} X$) if

$$E[\|X_n - X\|^p] \to 0 \quad \textit{as } n \to \infty.$$

Convergence in L_p is metrisable with a norm if $p \geq 1$. It is useful because if in addition $E[|X|^p] < \infty$ and either p is an integer or $X \geq 0$, then $EX_n^p \to EX^p$.

Both $X_n \xrightarrow{L_p} X$ and $X_n \xrightarrow{a.s.} X$ imply $X_n \xrightarrow{p,d} X$ but are not comparable.

Continuous Mapping Theorem

Theorem (Continuous mapping theorem)

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

- *If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.*
- *If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.*
- *If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$ (trivial!).*

Proof. The almost sure case is obvious. The other cases can be reduced to it using the subsequence lemma or Skorokhod representation theorem. But we can give elementary and more instructive proofs.
If $X_n \xrightarrow{p} X$, then for any $R > 0$,

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

The first term is bounded by $P(\|X\| > R)$ and vanishes uniformly in n as $R \to \infty$.

Proof continued

For the second term $P(\|g(X_n) - g(X)\| > \epsilon, \|X\| \leq R)$, note that $B_{R+1} := \{x : \|x\| \leq R + 1\}$ is compact so g is uniformly continuous there. Let δ be such that if $x, y \in B_{R+1}$ are such that $\|x - y\| \leq \delta$, then $\|g(x) - g(y)\| \leq \epsilon$. We may assume that $\delta \leq 1$. Thus

$$\begin{aligned}
&P(\|g(X_n) - g(X)\| > \epsilon, \|X\| \leq R) \\
&\leq P(\|X_n - X\| > \delta, \|X\| \leq R) \\
&\leq P(\|X_n - X\| > \delta) \to 0, \qquad n \to \infty.
\end{aligned}$$

Next, if $X_n \xrightarrow{d} X$, then by the portmanteau lemma it suffices to show that for a

continuous bounded function $f : \mathbb{R}^l \to \mathbb{R}$,

$$E[f(g(X_n))] \to E[f(g(X))].$$

But this follows from the portmanteau lemma, since $X_n \xrightarrow{d} X$ and $f \circ g$ is continuous.

Slutsky's theorem

Theorem (Slutsky, general version)

Let $X_n \xrightarrow{d} X$ on $\mathbb{R}^k$ and $Y_n \xrightarrow{d} c$ on $\mathbb{R}^l$, where $c \in \mathbb{R}^l$ is a constant. Then $g(X_n, Y_n) \xrightarrow{d} g(X, c)$.

Proof. By the continuous mapping theorem, it suffices to show that $(X_n, Y_n) \xrightarrow{d} (X, c)$. While in general it is not true that convergence of marginals yields convergence of the joint distributions, it does hold in the particular case when one of the limits is a constant. Indeed, suppose $F_{(X,Y)}$ is continuous at (x, y).
Case 1: $F_Y(y) = 0$, i.e., $y \geq c$ does NOT hold. In this case $P(X_n \leq x, Y_n \leq y) \leq P(Y_n \leq y) \to 0$ since $y_i < c_i$ for some i.
Case 2: $\exists i : y_i = c_i$. Then it must be that $F_X(x) = 0$ (exercise), and therefore F_X is continuous at x and $P(X_n \leq x, Y_n \leq y) \leq P(X_n \leq x) \to F_X(x) = 0$.
Case 3: $y > c$. Then locally around (x, y), $F_{X,Y}(x', y') = F_X(x')$ so it must be that F_X is continuous at x. Thus
$P(X_n \leq x, Y_n \leq y) \leq P(X_n \leq x) \to F_X(x) = F_{X,Y}(x, y)$ and
$P(X_n \leq x, Y_n > y) \leq P(Y_n > y) \to 0$. (Assumed for simplicity $l = 1$.)

Theorem (Delta method)

Let $Z_n := r_n(X_n - \theta) \xrightarrow{d} Z$, where $\theta \in \mathbb{R}^k$ is a constant and $r_n \to \infty$ are also constants. If $g : \mathbb{R}^k \to \mathbb{R}^l$ is differentiable at θ, then $r_n(g(X_n) - g(\theta)) \xrightarrow{d} \nabla g(\theta)^\top Z$.

Proof. Let

$$f(x) = \frac{g(x) - g(\theta) - \nabla g(\theta)^\top (x - \theta)}{\|x - \theta\|}$$

with $f(\theta) = 0$. Then applying Slutsky three times

$$\begin{aligned} r_n(g(X_n) - g(\theta)) &= r_n\|X_n - \theta\| f(X_n) + r_n \nabla g(\theta)^\top (X_n - \theta) \\ &= \|Z_n\| f(X_n) + \nabla g(\theta)^\top Z_n \xrightarrow{d} \|Z\| 0 + \nabla g(\theta)^\top Z, \end{aligned}$$

if we can show that $f(X_n) \xrightarrow{d} 0$. Now, f is continuous around θ, so for $\epsilon > 0$ there exists $\delta > 0$ such that

$$P(\|f(X_n)\| > \epsilon) \leq P(\|X_n - \theta\| > \delta) = P(r_n^{-1}\|Z_n\| > \delta) \to 0$$

by Slutsky (again!) as $r_n^{-1} \to 0$ and $\|Z_n\| \xrightarrow{d} \|Z\|$ (continuous mapping theorem).

Dominated convergence, convergence of expectations

$X_n \xrightarrow{p} X$ does not give $EX_n \to EX$ unless

Theorem (dominated convergence)

If $\|X_n\| \leq Y$ for all n, $X_n \xrightarrow{p} X$, and $EY < \infty$, then $EX_n \to EX \in \mathbb{R}^k$.

Proof. We also have $\|X\| \leq Y$ (exercise). It suffices to show $E\|X_n - X\| \to 0$. Let $R, \epsilon > 0$ and write

$$\begin{aligned} E\|X_n - X\| = & E[\|X_n - X\| 1(\|X_n - X\| \leq \epsilon, Y \leq R)] \\ & + E[\|X_n - X\| 1(\|X_n - X\| > \epsilon, Y \leq R)] \\ & + E[\|X_n - X\| 1(Y > R)] = E_1 + E_2 + E_3, \end{aligned}$$

where $E_1 \leq \epsilon$, $E_2 \leq 2R\ P(\|X_n - X\| > \epsilon) \overset{n\to\infty}{\to} 0$ and

$$E_3 \leq 2E[Y1(Y > R)] = 2\int_0^\infty P([Y1(Y > R)] > t)dt = 2\int_R^\infty P(Y > t)dt.$$

Since $EY < \infty$, this is a tail of a convergent integral, so $E_3 \to 0$ as $R \to \infty$. Moreover, $E_1 \to 0$ as $\epsilon \to 0$.

Aside: sufficient conditions for convergence in distribution

- Often difficult to establish weak convergence directly (from definition).
- Indeed, in most interesting cases F_n inconvenient to work with.
- Class of continuous functions in large.

Scheffé theorem

If X_n, X have densities f_n, f (with respect to the same measure μ) and $f_n \to f$ (μ-almost surely), then $X_n \xrightarrow{d} X$.

For example, law of small numbers

$$\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \to \frac{\lambda^x}{x!}e^{-\lambda}, n \to \infty,$$

for all $x \in \{0\} \cup \mathbb{N}$.
The convergence in Scheffé theorem is called convergence in total variation.

Aside: sufficient conditions for convergence in distribution

Class of continuous functions in large, but we do not need to check $Ef(X_n) \to Ef(X)$ for all continuous bounded functions.

- For light-tailed distributions, it suffices to look at (unbounded) functions $x \mapsto e^{t^\top x}$ for $t \in \mathbb{R}^k$ (moment generating functions).
- In general, it suffices to look as sines and cosines of arbitrary frequencies (characteristic functions).

Can reduce to one dimension by Cramér-Wold device: $X_n \xrightarrow{d} X$ on $\mathbb{R}^k$ if and only if for all $\theta \in \mathbb{R}^k$, $\theta^\top X_n \xrightarrow{d} \theta^\top X$.

Two important nontrivial theorems

Theorem (strong law of large numbers)

Let (X_n) be independent and identically distributed random vectors with $E\|X_1\| < \infty$. Then EX_1 is finite and

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{a.s.} EX_1$$

- "Strong" is as opposed to the "weak" law which gives $\xrightarrow{p}$ instead of $\xrightarrow{a.s.}$
- This is insanely strong: $E\|X_1\| < \infty$ is the weakest condition for the expectation to be well defined. The theorem reads: if there is an expected value, we can find it with the empirical mean.
- The strong law says nothing about the size of the error.

Two important nontrivial theorems

Theorem (central limit theorem)

Let (X_n) be independent and identically distributed random vectors with expectation $\mu \in \mathbb{R}^k$ and invertible covariance matrix Σ. Then

$$\sqrt{n}\Sigma^{-1/2}(\overline{X}_n - \mu) \xrightarrow{d} Z,$$

where $\overline{X}_n = \sum_{i=1}^n X_i/n$ and Z has k independent coordinates following $N(0,1)$.

- Insanely strong theorem: as soon as the covariance exists, we are in business (there are versions if it is not invertible).
- Once more, no control about the size of the error.
- There are many variants of this theorem.

Berry-Esseen Theorem

Theorem (Berry–Esseen, Bentkus (2005) version)

Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\mathbb{R}^k$ with $E[X_i] = 0$ and Cov$(X_i) = I_d$ (identity matrix). Define

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i.$$

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

$$\sup_{A \in \mathcal{A}} \left| P(Z_n \in A) - P(Z \in A) \right| \leq C \frac{d^{1/4} E\|X_i\|^3}{\sqrt{n}},$$

where C is a universal constant satisfying $C \leq 4$.

- Quantifies the approximation error in the Central Limit Theorem (CLT).
- Provides explicit error bounds for normal approximation in high dimensions.
- Useful for constructing confidence regions with guaranteed coverage probabilities.