Stochastic convergence

Patrick Breheny

September 16, 2024

Intro

- In our analysis review, we went over some results pertaining to the convergence of deterministic sequences of numbers and functions
- For statistical theory, we must extend these ideas to the convergence of random variables
- In doing so, there are actually several different ways we can characterize convergence
- You've seen these ideas in previous classes; our goal for today is to review these ideas and discuss their relationships in some ways that maybe you haven't seen before

Convergence in distribution

- Probably the most useful type of convergence (and the most straightforward extension from regular analysis) is convergence in distribution (also called convergence in law, or weak convergence)
- **Definition:** A sequence of random variables $\mathbf{x}_n$ *converges in distribution* to the random variable $\mathbf{x}$, denoted $\mathbf{x}_n \xrightarrow{\text{d}} \mathbf{x}$, if for all points $\mathbf{a}$ at which F is continuous, we have $F_n(\mathbf{a}) \to F(\mathbf{a})$.
- In other words, convergence in distribution is just regular pointwise convergence, applied to CDFs

Convergence in distribution and continuity

- Why the continuity requirement?
- Consider the following example:
  - $X_n = 1/n$ with probability 1
  - $X = 0$ with probability 1
- Seems obvious that the distribution of X_n converges to the distribution of X, and yet
  - $F_n(0) \to 0$
  - $F(0) = 1$
- In other words, $F_n \to F$ everywhere except the discontinuity point at 0

Generalized inequalities

- Now is a good time to mention something about generalized inequalities
- For a univariate distribution, $F_X(a)$ means $\mathbb{P}\{X \leq a\}$; however, the notation $\mathbf{x} \leq \mathbf{a}$ is potentially ambiguous for vectors (what if some $x_j > a_j$ and some $x_j < a_j$?)
- To be explicit, the notation $\preceq$ is sometimes used instead, where $\mathbf{x} \preceq \mathbf{a}$ is defined to mean $\mathbf{a} - \mathbf{x} \in \mathbb{R}^d_+$
- In other words, $\mathbf{x} \preceq \mathbf{a}$ means that $x_j \leq a_j$ for all j
- Returning to CDFs, $F_{\mathbf{x}}(\mathbf{a})$ means $\mathbb{P}\{\mathbf{x} \preceq \mathbf{a}\}$ or equivalently, $\mathbb{P}\{\cap_{j=1}^d x_j \leq a_j\}$
- It does **not** mean $\mathbb{P}\{\|\mathbf{x}\| \preceq a\}$,

Generalized inequalities (cont'd)

- The strict inequality is defined similarly, with $\mathbf{x} \prec \mathbf{a}$ defined to mean $\mathbf{a} - \mathbf{x} \in \mathbb{R}^d_{++}$, or $x_j < a_j$ for all j
- This definition probably seems unnecessarily abstract; however, it is useful in defining inequalities for more complex objects
- In particular, the generalized inequality $\mathbf{A} \preceq \mathbf{B}$ means that the matrix $\mathbf{B} - \mathbf{A}$ belongs to the set of positive semidefinite matrices (and $\mathbf{A} \prec \mathbf{B}$ meaning that $\mathbf{B} - \mathbf{A}$ is positive definite)

Convergence in probability

- **Definition:** A sequence of random vectors $\mathbf{x}_n$ *converges in probability* to the random vector $\mathbf{x}$, denoted $\mathbf{x}_n \xrightarrow{\text{P}} \mathbf{x}$, if for all $\delta > 0$,

$$\mathbb{P}\{\|\mathbf{x}_n - \mathbf{x}\| > \delta\} \to 0$$

- Recalling the definition of convergence, note what convergence in probability requires: for all $\epsilon > 0$ and all $\delta > 0$, there exists N such that

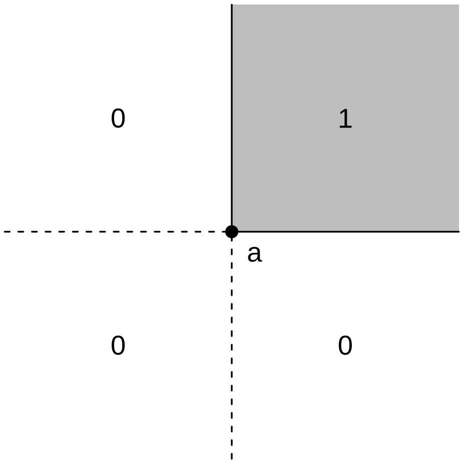
$$\mathbb{P}\{\|\mathbf{x}_n - \mathbf{x}\| > \delta\} < \epsilon$$

for all $n > N$

Convergence in probability vs convergence in distribution

- Often, we are interested in convergence in probability to a constant; in this case convergence in probability and convergence in distribution are equivalent
- **Theorem:** Let $\mathbf{a} \in \mathbb{R}^d$. Then $\mathbf{x}_n \xrightarrow{\mathrm{P}} \mathbf{a}$ if and only if $\mathbf{x}_n \xrightarrow{\mathrm{d}} \mathbf{a}$.
- Note, of course, that if $\mathbf{x}_n \xrightarrow{\mathrm{d}} \mathbf{x}$, where $\mathbf{x}$ is a random vector, then certainly convergence in distribution does not imply convergence in probability (two random variables with the same distribution can be very far apart)

Illustration: CDF of a point (relevant to proof)



Weak law of large numbers

- In statistics, convergence in probability to a constant is often connected to the idea of consistency; an estimator $\hat{\boldsymbol{\theta}}$ is said to be a *consistent* estimator of $\boldsymbol{\theta}$ if $\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{P}} \boldsymbol{\theta}$
- For example, later in the course we will take up the question: under what conditions are maximum likelihood estimators consistent?
- The most important and well-known consistency result is the law of large numbers
- **Theorem (Weak law of large numbers):** Let $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, \ldots$ be independently and identically distributed random vectors such that $\mathbb{E}\|\mathbf{x}\| < \infty$. Then $\bar{\mathbf{x}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\mu}$, where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{x})$.

Note on existence of moments

- A brief aside on the existence of means
- In the univariate case, the mean exists if and only if $\mathbb{E}|X| < \infty$
- The equivalent idea in multiple dimensions is that $\mathbb{E}(\mathbf{x})$ exists if and only if $\mathbb{E}\|\mathbf{x}\| < \infty$ (not as obvious as in the one-dimensional case, but easiest to see if we consider $\|\cdot\|_1$)
- The same goes for higher moments as well: $\mathbb{E}\|\mathbf{x}\|^2 < \infty$ is equivalent to saying that the variance exists
- Authors (including me) often use the norm version in stating proof conditions just because it's more compact, but keep in mind that it has more to do with existence of mean/variance than the norm itself

o_p notation

- A few weeks ago, we introduced O, o notation and mentioned that they had analogs for probabilistic convergence; let's return to that idea now
- **Definition:** A sequence of random vectors $\mathbf{x}_n$ is said to be $o_p(1)$ if it converges to $\mathbf{0}$ in probability. Furthermore, $\mathbf{x}_n$ is said to be $o_p(r_n)$ if

$$\frac{\mathbf{x}_n}{r_n} \xrightarrow{\mathrm{P}} \mathbf{0}.$$

- Note that this is exactly the same definition as $o(\cdot)$, only with $\rightarrow$ replaced by $\xrightarrow{\mathrm{P}}$

Bounded in probability

- $O_p(\cdot)$ is conceptually the same as $O(\cdot)$, but somewhat more complicated to define
- **Definition:** A sequence of random variables $\mathbf{x}_n$ is *bounded in probability* if for any $\epsilon > 0$, there exist M and N such that $\mathbb{P}\{\|\mathbf{x}_n\| > M\} < \epsilon$ for all $n > N$.
- Often, this is easiest to show through convergence in distribution: if there exists a random variable $\mathbf{x}$ such that $\mathbf{x}_n \xrightarrow{\text{d}} \mathbf{x}$, then $\mathbf{x}_n$ is bounded in probability
- **Definition:** A sequence of random variables $\mathbf{x}_n$ is said to be $O_p(1)$ if it is bounded in probability. Furthermore, $\mathbf{x}_n$ is said to be $O_p(r_n)$ if $\mathbf{x}_n/r_n$ is bounded in probability.

Algebra of O_p, o_p notation

The rules of working with O_p and o_p terms are exactly the same as in the deterministic case:

Theorem: For $a \le b$:

$$O_p(1) + O_p(1) = O_p(1) \qquad O_p\{O_p(1)\} = O_p(1)$$

$$o_p(1) + o_p(1) = o_p(1) \qquad o_p\{O_p(1)\} = o_p(1)$$

$$o_p(1) + O_p(1) = O_p(1) \qquad o_p(r_n) = r_n o_p(1)$$

$$O_p(1)O_p(1) = O_p(1) \qquad O_p(r_n) = r_n O_p(1)$$

$$O_p(1)o_p(1) = o_p(1) \qquad O_p(n^a) + O_p(n^b) = O_p(n^b)$$

$$\{1 + o_p(1)\}^{-1} = O_p(1) \qquad o_p(n^a) + o_p(n^b) = o_p(n^b)$$

Examples

Suppose $x \overset{\text{iid}}{\sim} (\mu, \sigma^2)$, with $\sigma^2 > 0$ [1]

- $x_i = O_p(1)$
- $\sum_i x_i = O_p(\sqrt{n})$ if $\mu = 0$
- $\sum_i x_i = O_p(n)$ if $\mu \neq 0$
- $\sum_i x_i^2 = O_p(n)$

[1] the notation (μ, σ^2) means that the mean is μ and variance is σ^2 but the distribution is otherwise left unspecified

$\sqrt{n}$-consistency

- The idea of consistency is often too weak to be interesting: any reasonable estimator is consistent given an infinite amount of data – often, many estimators satisfy this requirement
- We can reveal more about their relative accuracy by noting the rate of convergence
- For example, $\hat{\boldsymbol{\theta}}$ is said to be a $\sqrt{n}$-*consistent* estimator of $\boldsymbol{\theta}$ if $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| = O_p(1/\sqrt{n})$
- This means that not only does $\hat{\boldsymbol{\theta}}$ get close to $\boldsymbol{\theta}$ as $n \to \infty$, but it converges to $\boldsymbol{\theta}$ so fast that it stays within an ever-shrinking neighborhood $N_{M/\sqrt{n}}(\boldsymbol{\theta})$ with high probability
- As we will see, the MLE $\hat{\boldsymbol{\theta}}$ tends to be $\sqrt{n}$-consistent

Convergence in rth mean

- An alternative form of convergence is to consider the convergence of moments
- **Definition:** For any real number $r > 0$, $\mathbf{x}_n$ *converges in* r*th mean* to $\mathbf{x}$, denoted $\mathbf{x}_n \xrightarrow{r} \mathbf{x}$, if

$$\mathbb{E}\|\mathbf{x}_n - \mathbf{x}\|^r \to 0.$$

- This is most useful in the case when $r = 2$, where it is called *convergence in quadratic mean* and denoted $\mathbf{x}_n \xrightarrow{\text{qm}} \mathbf{x}$

Convergence in mean vs convergence in probability

- Convergence in quadratic mean is particularly useful as it often provides an easy way to prove consistency due to the following two facts
- **Theorem:** If $\mathbf{x}_n \xrightarrow{r} \mathbf{x}$ for some $r > 0$, then $\mathbf{x}_n \xrightarrow{\text{P}} \mathbf{x}$.
- **Theorem:** If $\mathbf{a} \in \mathbb{R}^d$, then $\mathbf{x}_n \xrightarrow{\text{qm}} \mathbf{a}$ if and only if $\mathbb{E}\mathbf{x}_n \to \mathbf{a}$ and $\mathbb{V}\mathbf{x}_n \to \mathbf{0}$.

Another law of large numbers

- It is also worth pointing out and proving another law of large numbers, this time involving convergence in quadratic mean
- **Theorem (Law of large numbers):** Let $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, \ldots$ be independently and identically distributed random vectors such that $\mathbb{E}\|\mathbf{x}\|^2 < \infty$. Then $\bar{\mathbf{x}}_n \xrightarrow{\text{qm}} \boldsymbol{\mu}$, where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{x})$.
- Note that this proof did not actually require $\{\mathbf{x}_i\}$ to be independent or identically distributed, only that they are uncorrelated and have the same mean and variance

Convergence in distribution vs convergence of means

- Now, suppose $\mathbf{x}_n \xrightarrow{\text{d}} \mathbf{x}$; can we conclude that $\mathbb{E}\mathbf{x}_n \to \mathbb{E}\mathbf{x}$?
- As it turns out, no, not necessarily
- As a counterexample,

$$X_n = \begin{cases} n & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 1/n \end{cases}$$

- We have $X_n \xrightarrow{\text{P}} 0$ (and thus, $X_n \xrightarrow{\text{d}} 0$), but $\mathbb{E}X_n \to 1$

Dominated convergence theorem

- The problem here is that the sequence $\{X_n\}$ is not bounded (or more accurately, not uniformly bounded)
- If the sequence $\{X_n\}$ is able to be bounded, however, then the moments do converge; this result is known as the *dominated convergence theorem*, which we previously encountered in pure integral form
- **Theorem (Dominated convergence):** If there exists a random variable Z such that $\|\mathbf{x}_n\| \leq Z$ for all n and $\mathbb{E}Z < \infty$, then $\mathbf{x}_n \xrightarrow{\mathrm{d}} \mathbf{x}$ implies that $\mathbb{E}\mathbf{x}_n \to \mathbb{E}\mathbf{x}$.
- Note that if $\{X_n\}$ is uniformly bounded by a constant, then the DCT clearly applies

Almost sure convergence

- The final type of convergence we will discuss is called almost sure convergence, also known as strong convergence or convergence with probability 1 (sometimes abbreviated wp1)
- **Definition:** A sequence of random variables $\mathbf{x}_n$ *converges almost surely* to the random variable $\mathbf{x}$, denoted $\mathbf{x}_n \xrightarrow{\text{as}} \mathbf{x}$, if

$$\mathbb{P}\left\{\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}\right\} = 1.$$

- This type of convergence is a bit more abstract than convergence in distribution, probability, or mean, but on the other hand is sometimes easier to work with since the limit operation takes place inside the probability expression, and thus involves only deterministic considerations

Strong law of large numbers

- As with convergence in probability, we are often concerned with almost sure convergence to a constant, and in particular with convergence of an estimator $\hat{\boldsymbol{\theta}}$ to the true value $\boldsymbol{\theta}$
- If $\hat{\boldsymbol{\theta}} \xrightarrow{\text{as}} \boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}$ is said to be a *strongly consistent* estimator of $\boldsymbol{\theta}$
- The sample mean, for example, is a strongly consistent estimator of $\mathbb{E}(X)$
- **Theorem (Strong law of large numbers):** Let $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, \ldots$ be independently and identically distributed random vectors such that $\mathbb{E}\|\mathbf{x}\| < \infty$. Then $\bar{\mathbf{x}}_n \xrightarrow{\text{as}} \boldsymbol{\mu}$, where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{x})$.

Alternate definition

- The following equivalent definition of almost sure convergence helps to highlight the difference between almost sure convergence and convergence in probability
- **Definition:** A sequence of random variables $\mathbf{x}_n$ *converges almost surely* to the random variable $\mathbf{x}$, denoted $\mathbf{x}_n \xrightarrow{\text{as}} \mathbf{x}$, if for every $\epsilon > 0$,

$$\mathbb{P}\{\|\mathbf{x}_k - \mathbf{x}\| < \delta \text{ for all } k \geq n\} \to 1$$

as $n \to \infty$

Alternate definition (cont'd)

- In other words, for every $\epsilon > 0$,
  - Convergence in probability requires that $\mathbb{P}\{\|\mathbf{x}_n - \mathbf{x}\| < \delta\} \to 1$
  - Convergence almost surely requires that $\|\mathbf{x}_k - \mathbf{x}\| < \delta$ for all $k > n$
- Since the second event is a subset of the first, we can immediately see that convergence almost surely implies convergence in probability
- **Theorem:** $\mathbf{x}_n \xrightarrow{\text{as}} \mathbf{x} \implies \mathbf{x}_n \xrightarrow{\text{P}} \mathbf{x}$.

Converse?

- The converse, however, is not true: it is possible for $\mathbf{x}_n$ to converge to $\mathbf{x}$ in probability, but not almost surely
- That said, finding a counterexample is not trivial and most of them that you find in textbooks tend to be rather contrived, making it hard to see how this idea is relevant to statistics
- However, let us consider an interesting and rather surprising result known as the Law of the Iterated Logarithm

Law of the iterated logarithm

- **Theorem (Law of the iterated logarithm):** Let $Z_1, Z_2, \ldots \overset{\text{iid}}{\sim} (0,1)$. Then

  $$\limsup_{n\to\infty} \frac{\sum_{i=1}^n Z_i}{\sqrt{n \log \log n}} = \sqrt{2}$$

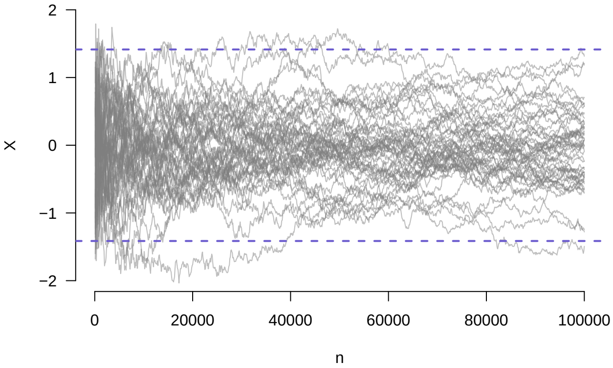
  almost surely.
- From the central limit theorem, we know that $\sqrt{n}\bar{Z}$ is spread out over the entire real line: $\sqrt{n}\bar{Z} \xrightarrow{\text{d}} \text{N}(0,1)$
- On the other hand, the LIL is saying that if we divide $\sqrt{n}\bar{Z}$ by $\sqrt{\log \log n}$, then this is no longer the case – it stays within the region $[-\sqrt{2}, \sqrt{2}]$ almost surely

Convergence of Z_n

- Let $X_n = \sqrt{n}\bar{Z}/\sqrt{\log \log n}$ denote the quantity on the previous slide
- We know that $X_n \xrightarrow{\mathrm{P}} 0$ (why?)
- However, the LIL indicates that X_n does not converge almost surely to 0; indeed, it doesn't converge to anything and instead wanders around the interval $[-\sqrt{2}, \sqrt{2}]$ forever

Illustration

A picture helps to illustrate the situation:



Eventually, each of these lines will reach the dotted lines at $\pm\sqrt{2}$ infinitely often (i.e., get within ϵ of them)

Implications for inference

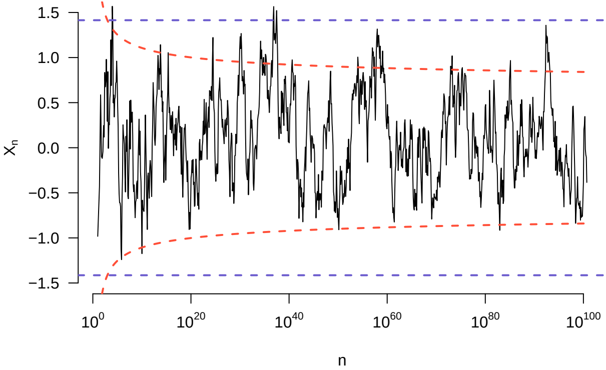
- So, what are the implications for statistical inference?
- Suppose we construct standard 95% confidence intervals for μ based on the known $\sigma^2 = 1$: $\bar{Z} \pm 1.96/\sqrt{n}$
- Those confidence intervals will contain the true value of $\mu = 0$ if and only if

$$|X_n| \leq \frac{1.96}{\sqrt{\log \log n}}$$

- However, when n is large enough $(n > 921)$, this quantity is less than $\sqrt{2}$, so the law of the iterated logarithm says that with probability 1, our confidence interval will eventually exclude the true value of μ

Another illustration

The red lines are $\pm 1.96/\sqrt{\log\log n}$, any time X_n passes outside them, the CI excludes the truth



Weak vs strong convergence

- So on the one hand, by convergence in distribution (weak), we have that for any ϵ, we can limit our Type I error below ϵ at any specific value of n (provided n is large enough)
- On the other hand, by convergence almost surely (strong), we know that no matter how low our Type I error is (ϵ), we will eventually make Type I errors for some n (in fact, infinitely often as $n \to \infty$)

Is this bad?

- Is this a problem?
- On one hand, we can be certain that no matter what, our confidence intervals will eventually be wrong – sounds bad
- On the other hand, imagine that we have 100 statisticians collecting another data point every minute and updating their 95% confidence intervals
- We know that 5 of them will be wrong; the LIL tells us that everyone eventually has to take a turn being wrong, which seems only fair
- Finally, note that we have to go out to pretty ridiculous sample sizes ($n = 10^{200}$) in order to start seeing the difference between strong and weak convergence (i.e., not a huge issue in practice)