Convergence, continuity, and measure

Patrick Breheny

September 4, 2024

Introduction

- In the previous lecture, we introduced (a) the idea of convergence and (b) the concept of a norm to measure the distance between two vectors
- Today, we will combine these two ideas to discuss the convergence of vectors as well as the related concepts of continuity and uniform convergence
- In addition, we will go over the basics of measure theory – you don't need to be an expert in this topic as a statistician, but a little goes a long way

Neighborhoods

- The set of vectors that is "close" to a vector $\mathbf{x}$ is known as its "neighborhood"
- **Definition:** The *neighborhood* of a point $\mathbf{p} \in \mathbb{R}^d$, denoted $N_\delta(\mathbf{p})$, is the set $\{\mathbf{x} : \|\mathbf{x} - \mathbf{p}\| < \delta\}$.
- This will come up quite often in this course
  - For example, we will often need to make assumptions about the likelihood function $L(\boldsymbol{\theta})$
  - However, we don't necessarily need these assumptions to hold everywhere – it's enough that they hold in a neighborhood of $\boldsymbol{\theta}^*$, the true value of the parameter

Convergence

- There are two potential ways we could extend this idea to the multivariate case
- **Definition:** We say that the vector $\mathbf{x}_n$ *converges* to $\mathbf{x}$, denoted $\mathbf{x}_n \to \mathbf{x}$, if each element of $\mathbf{x}_n$ converges to the corresponding element of $\mathbf{x}$.
- Alternatively, we can use norms to construct a more direct definition
- **Definition:** A sequence $\mathbf{x}_n$ is said to *converge* to $\mathbf{x}$, which we denote $\mathbf{x}_n \to \mathbf{x}$, if for every $\epsilon > 0$, there is a number N such that $n > N$ implies that $\|\mathbf{x}_n - \mathbf{x}\| < \epsilon$.
- We'll establish in a moment that these two definitions are equivalent

Continuity

- It's fairly obvious that, say, $\mathbf{x}_n + \mathbf{y}_n \to \mathbf{x} + \mathbf{y}$, but what about more complicated functions? Does $\sqrt{x_n} \to \sqrt{x}$? Does $f(\mathbf{x}_n) \to f(\mathbf{x})$ for all functions?
- The answer to the second question is no: not all functions possess this property at all points
- This is obviously a very useful property though, so functions that possess it are given a specific name: continuous functions

Continuity (cont'd)

- **Definition:** A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *continuous* at a point $\mathbf{p}$ if for all $\epsilon > 0$, there exists $\delta > 0$:

$$\|\mathbf{x} - \mathbf{p}\| < \delta \implies |f(\mathbf{x}) - f(\mathbf{p})| < \epsilon$$

- Note that by the equivalence of norms, we can just say that a function is continuous – it can't be, say, continuous with respect to $\|\cdot\|_2$ and not continuous with respect to $\|\cdot\|_1$
- **Theorem:** Suppose $\mathbf{x}_n \to \mathbf{x}_0$ and $f : \mathbb{R}^d \to \mathbb{R}$ is continuous at $\mathbf{x}_0$. Then $f(\mathbf{x}_n) \to f(\mathbf{x}_0)$.

Continuity and convergence

- The norm itself is a continuous function
- **Theorem:** Let $f(\mathbf{x}) = \|\mathbf{x}\|$, where $\|\cdot\|$ is any norm. Then $f(\mathbf{x})$ is continuous.
- One consequence of this result is that element-wise convergence is equivalent to convergence in norm
- **Theorem:** $\mathbf{x}_n \to \mathbf{x}$ element-wise if and only if $\|\mathbf{x}_n - \mathbf{x}\| \to 0$.

Convergence of functions

- One final important concept with respect to convergence is the convergence of functions
- **Definition:** Suppose $f_1, f_2, \ldots$ is a sequence of functions and that for all $\mathbf{x}$, the sequence $f_n(\mathbf{x})$ converges. We can then define the *limit function* f by

$$f(\mathbf{x}) = \lim_{n \to \infty} f_n(\mathbf{x})$$

- Sequences of functions come up constantly in statistics, the most relevant example being the likelihood function $L(\boldsymbol{\theta}|\mathbf{x}_n) = L_n(\boldsymbol{\theta})$

Combining the two types of convergence

- Furthermore, we are often interested in combining convergence of the function with convergence of the argument
- For example, does $f_n(\hat{\theta}_n) \to f(\theta)$ as $\hat{\theta}_n \to \theta$?
- This raises a number of additional issues we have not encountered before
- We'll return to the probabilistic question later in the course; for now, let's discuss the problem in deterministic terms: does $f_n(x_n) \to f(x_0)$ as $x_n \to x_0$?

Counterexample

- Unfortunately, the answer is no – in general, this is not true
- For example:

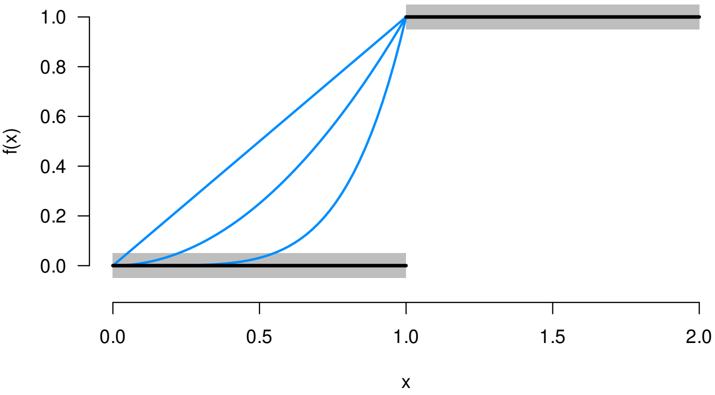
$$f_n(x) = \begin{cases} x^n & x \in [0, 1] \\ 1 & x \in (1, \infty) \end{cases}$$

- We have:

$$\lim_{x \to 1^-} \lim_{n \to \infty} f_n(x) = 0 \neq f(1)$$

Illustration

The underlying issue is that f_n doesn't really converge to f in the sense of always lying within $\pm\epsilon$ of it:



Uniform convergence

- The relationship between f_n and f is one of *pointwise convergence*; we need something stronger
- **Definition:** A sequence of functions $f_1, f_2, \ldots : \mathbb{R}^d \to \mathbb{R}$ *converges uniformly* on a set E to a function f if for every $\epsilon > 0$ there exists N such that $n > N$ implies

$$|f_n(\mathbf{x}) - f(\mathbf{x})| < \epsilon$$

  for all $x \in E$
- **Corollary:** $f_n \to f$ uniformly on E if and only if

$$\sup_{\mathbf{x} \in E} |f_n(\mathbf{x}) - f(\mathbf{x})| \to 0.$$

Supremum and infimum

- In case you haven't seen it before, the $\sup$ notation on the previous slide stands for *supremum*, or *least upper bound*
- As the name implies, α is a least upper bound of the set E if (i) α is an upper bound of E and (ii) if $\gamma < \alpha$, then γ is not an upper bound of E
- Similarly, the *greatest lower bound* of a set is known as the *infimum*, denoted $\alpha = \inf E$
- The concept is similar to the maximum/minimum of E, but if E is an infinite set, it doesn't necessarily have a largest/smallest element, which is why we need sup/inf

Supremum and infimum: Example

- For example, consider the set $\{x^2 : x \in (0,1)\}$
- Its least upper bound (sup) is 1, but 1 is not an element of the set
- To prove that 1 is the least upper bound, note that (a) 1 is an upper bound and (b) if I choose any number $b < 1$, then b is not an upper bound; this is standard technique
- Similarly, the greatest lower bound (inf) of the set is 0, but 0 is not an element of the set

Why uniform convergence is useful

- Uniform convergence is useful because it allows us to reach the kind of conclusion we originally sought
- **Theorem:** Suppose $f_n \to f$ uniformly, with f_n continuous for all n. Then $f_n(\mathbf{x}) \to f(\mathbf{x}_0)$ as $\mathbf{x} \to \mathbf{x}_0$.
- Note that this argument does not work without uniform convergence

Preview

- Later on in the course, this idea will be quite relevant to likelihood theory: we will often require that $\mathcal{I}_n(\hat{\theta}_n)$ is close to $\mathscr{I}(\theta^*)$
- A common way of ensuring uniform convergence is by bounding the derivative; here, this would mean requiring that

$$\left|\frac{\partial}{\partial\theta}\mathcal{I}_n(\theta)\right| \leq M$$

for all n and for all θ
- Note that this must be a *uniform* bound in the sense that the bound M does not depend on θ or n

Extensions

- The theorem we just proved can actually be made somewhat stronger:
- **Theorem:** Suppose $f_n \to f$ uniformly on E and that $\lim_{x \to x_0} f_n(\mathbf{x})$ exists for all n. Then for any limit point x_0 of E,

$$\lim_{x \to x_0} \lim_{n \to \infty} f_n(\mathbf{x}) = \lim_{n \to \infty} \lim_{x \to x_0} f_n(\mathbf{x}).$$

- **Corollary:** If $\{f_n\}$ is a sequence of continuous functions on E and if $f_n \to f$ uniformly on E, then f is continuous on E.

Related concepts

- There are number of related concepts similar to uniform convergence
- **Definition:** A function $f : \mathbb{R}^d \to \mathbb{R}$ is called *uniformly continuous* if for all $\epsilon > 0$, there exists $\delta > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{y}\| < \delta$, we have $|f(\mathbf{x}) - f(\mathbf{y})| < \epsilon$.
- For example, $f(x) = x^2$ is uniformly continuous over $[0, 1]$ but not over $[0, \infty)$
- **Definition:** A sequence $X_1, X_2, \ldots$ of random variables is said to be *uniformly bounded* if there exists M such that $|X_n| < M$ for all X_n.

o-notation: Motivation

- When dealing with convergence, it is often convenient to replace unwieldy expressions with compact notation
- For example, if we encountered the mathematical expression

$$x^2 + a - a,$$

we would obviously want to replace it with x^2 since $a - a = 0$

- However, what if we encounter something like

$$x^2 + \frac{5\theta}{\sqrt{n}} - \frac{3\theta}{n+5}?$$

- We can no longer just replace this with x^2

o-notation: Motivation (cont'd)

- However, as n gets larger, the expression gets closer and closer to x^2
- It would be convenient to have a shorthand notation for this, something like $x^2 + o_n$, where o_n represents some quantity that becomes negligible as n becomes large
- This is the basic idea behind o-notation, and its simplifying powers become more apparent as the mathematical expression we are dealing with becomes more complicated:

$$\frac{x^2 + \frac{5\theta}{\sqrt{n}} - \frac{3\theta}{n+5}}{(n^2+5n-2)/(n^2-3n+1)} + \frac{\exp\{-\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}\|^2\}}{2\sqrt{n}\theta\int_0^\infty g(s)ds}$$

o-notation

- This is where o-notation comes in: it provides a formal way of handling terms that effectively "cancel out" as we take limits
- **Definition:** A sequence of numbers x_n is said to be $o(1)$ if it converges to zero. Likewise, x_n is said to be $o(r_n)$ if

$$\frac{x_n}{r_n} \to 0$$

  as $n \to \infty$.
- When the rate is constant, o notation is pretty straightforward:

$$x^2 + \frac{5\theta}{\sqrt{n}} - \frac{3\theta}{n+5} = x^2 + o(1)$$

o-notation remarks

- When the rate is not constant, expressions are a bit harder to think about – it helps to go over some cases:
- For example:
  - $x_n \to \infty$, but $r_n \to \infty$ even faster:

$$n = o(n^2)$$

  - $r_n \to 0$, but $x_n \to 0$ even faster:

$$\frac{1}{n^2} = o(1/n)$$

O-notation

- A very useful companion of o-notation is O-notation, which denotes whether or not a term remains bounded as $n \to \infty$
- **Definition:** A sequence of numbers x_n is said to be $O(1)$ if there exist M and N such that

$$|x_n| < M$$

for all $n > N$. Likewise, x_n is said to be $O(r_n)$ if there exist M and N such that for all $n > N$,

$$\left| \frac{x_n}{r_n} \right| < M.$$

O-notation remarks

- For example,

$$\frac{\exp\{-\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}\|^2\}}{2\sqrt{n}\theta\int_0^\infty g(s)ds} = O(n^{-1/2})$$

- Note that $x_n = O(1)$ does not necessarily mean that x_n is bounded, just that it is eventually bounded
- Note also that just because a term is $O(1)$, this does not necessarily mean that it has a limit; for example,

$$\sin\left(\frac{n\pi}{2}\right) = O(1),$$

even though the sequence does not converge

O-notation remarks (cont'd)

- You may encounter the ambiguous phrase "x_n is of order r_n"
- The author may mean that $x_n = O(r_n)$
- However, it might also mean something stronger: that there exist positive constants m and M such that

$$m \leq \left| \frac{x_n}{r_n} \right| \leq M$$

for large enough n; i.e., the ratio is bounded above but also bounded below

- In other words, $x_n = O(r_n)$ but in addition $x_n \neq o(r_n)$; some authors use the notation $x_n \asymp r_n$ to denote this situation

Informative-ness of o and O notation

- There are typically many ways of writing an expression using O notation, although not all of them will be equally informative
- For example, if $x_n = \frac{1}{n}$, then all of the following are true:

$$
\begin{aligned}
x_n &= o(1) & & \\
x_n &= O(1) & & \text{(least informative)} \\
x_n &= O(\tfrac{1}{n}) & & \text{(more informative)} \\
x_n &\asymp \tfrac{1}{n} & & \text{(most informative)}
\end{aligned}
$$

Algebra of O, o notation

O, o-notation are useful in combination because simple rules govern how they interact with each other

Theorem: For $a \leq b$:

$$
\begin{aligned}
O(1) + O(1) &= O(1) & O\{O(1)\} &= O(1) \\
o(1) + o(1) &= o(1) & o\{O(1)\} &= o(1) \\
o(1) + O(1) &= O(1) & o(r_n) &= r_n o(1) \\
O(1)O(1) &= O(1) & O(r_n) &= r_n O(1) \\
O(1)o(1) &= o(1) & O(n^a) + O(n^b) &= O(n^b) \\
\{1 + o(1)\}^{-1} &= O(1) & o(n^a) + o(n^b) &= o(n^b)
\end{aligned}
$$

Remarks

- O, o "equations" are meant to be read left-to-right; for example, $O(\sqrt{n}) = O(n)$ is a valid statement, but $O(n) = O(\sqrt{n})$ is not
- **Exercise:** Determine the order of

$$n^{-2}\left\{(-1)^n \sqrt[n]{2} + n(1 + \tfrac{1}{n})^n\right\}.$$

- As we will see in a week or two, there are stochastic equivalents of these concepts, involving convergence in probability and being bounded in probability
- As such, we won't do a great deal with O, o-notation right now, but will use the stochastic equivalents extensively

Introduction

- We now turn our attention to integration – I assume that you know how to take integrals, but perhaps not their underlying theoretical development, and not with the Riemann-Stieltjes form of integrals
- This form is useful to be aware of, as it has a deep connection with probability and measure theory and allows for a nice unification of continuous and discrete probability theory
- This is not a measure theory-based course, but it is worth knowing some basic results that will help you read papers that use measure theoretical language

Introduction to Riemann-Stieltjes integration

- Probability and expectation are intimately connected with integration
- The basic forms of integration that you learn as an undergraduate are known as Riemann integrals; a more rigorous form is the Lebesgue integral, but that rests on quite a bit of measure theory
- The Riemann-Stieltjes integral is a useful bridge between the two, and particularly useful in statistics

Partitions and lower/upper sums

- **Definition:** A *partition* P of the interval $[a, b]$ is a finite set of points $x_0, x_1, \ldots, x_n$ such that

$$a = x_0 < x_1 < \cdots < x_n = b.$$

- Let μ be a bounded, nondecreasing function on $[a, b]$, and let

$$\Delta\mu_i = \mu(x_i) - \mu(x_{i-1});$$

  note that $\Delta\mu_i \geq 0$

- Finally, for any function g define the lower and upper sums

$$L(P, g, \mu) = \sum_{i=1}^{n} m_i \Delta\mu_i \qquad m_i = \inf_{[x_{i-1}, x_i]} g$$

$$U(P, g, \mu) = \sum_{i=1}^{n} M_i \Delta\mu_i \qquad M_i = \sup_{[x_{i-1}, x_i]} g$$

Refinements

- **Definition:** A partition P^* is a *refinement* of P if $P^* \supset P$ (every point of P is a point of P^*). Given partitions P_1 and P_2, we say that P^* is their *common refinement* if $P^* = P_1 \cup P_2$.
- **Theorem:** If P^* is a refinement of P, then

$$L(P, g, \mu) \leq L(P^*, g, \mu)$$

and

$$U(P^*, g, \mu) \leq U(P, g, \mu)$$

- **Theorem:** $L(P_1, g, \mu) \leq U(P_2, g, \mu)$

The Riemann-Stieltjes integral

Definition: If the following two quantities are equal:

$$\inf_P U(P, g, \mu)$$

$$\sup_P L(P, g, \mu),$$

then g is said to be *integrable with respect to* μ over $[a, b]$, and we denote their common value

$$\int_a^b g d\mu$$

or sometimes

$$\int_a^b g(x) d\mu(x)$$

Dominated convergence theorem

- One of the most important results in measure theory is the dominated convergence theorem
- **Theorem (Dominated convergence):** Let f_n be a sequence of integrable functions such that $f_n \to f$. If there exists an integrable function g such that $|f_n(x)| \leq g(x)$ for all n and all x, then

$$\lim_{n \to \infty} \int f_n \, d\mu = \int f \, d\mu.$$

- The theorem can be restated in terms of expected values, which we will go over (and use) in a later lecture

Implications for probability

- The application to probability is clear: any CDF can play the role of μ (CDFs are bounded and nondecreasing), so expected values can be written

$$\mathbb{E}g(X) = \int g(x)\,dF(x)$$

- Why is this more appealing than the usual Riemann integral?
- The main reason is that the above statement is valid regardless of whether X has a continuous or discrete distribution (or some combination of the two) – we require only that F is nondecreasing, not that it is continuous

Continuous and discrete measures

- Suppose F is the CDF of a discrete random variable that places point mass p_i on support point s_i; then

$$\int g\,dF = \sum_{i=1}^{\infty} g(s_i)p_i$$

- Suppose F is the CDF of a continuous random variable with corresponding density $f(x)$; then assuming $g(X)$ is integrable with respect to F,

$$\int g\,dF = \int g(x)f(x)\,dx$$

- In other words, the Riemann-Stieltjes integral reduces to familiar forms in both continuous and discrete cases

Example

- However, the Riemann-Stieltjes integral also works in mixed cases
- **Exercise:** Suppose X has a distribution such that $P(X=0)=1/3$, but if $X \neq 0$, then it follows an exponential distribution with $\lambda = 2$. Suppose $g(x) = x^2$; what is $\int g \, dF$?

Decomposing random variables

- Now, you might be wondering: can we always do this?
- Can we always just separate out any random variable into its continuous and discrete components and handle them separately like this?
- The answer, unfortunately, is no

Lebesgue decomposition theorem

- **Theorem (Lebesgue decomposition):** Any probability distribution F can uniquely be decomposed as

$$F = F_{\text{D}} + F_{\text{AC}} + F_{\text{SC}},$$

  where
  - F_{D} is the discrete component (i.e., probability is given by a sum of point masses)
  - F_{AC} is the absolutely continuous component (i.e., probability is given by an integral with respect to a density function)
  - F_{SC} is the singular continuous component (i.e., it's weird)
- The theorem is typically stated in terms of measures, but I'm using (sub)distribution functions here for the sake of familiarity

Important takeaways

- Obviously, we're skipping the technical details of measure theory as well as the proof of this theorem, but you don't need a technical understanding to see why it's important
- It's not the case that all distributions can be decomposed into discrete and "continuous" components – there is a third possibility: singular
- However, if we add the restriction that we are dealing with *non-singular* (or *regular*) distributions, then yes, all distributions can be decomposed into the familiar continuous and discrete cases
- To be technically accurate, one might wish to clarify "absolutely continuous" instead of continuous when you're referring to a distribution with a density (in non-technical contexts, this is implicit)