Analysis review: Norms, convergence, and continuity

Patrick Breheny

August 25
Before we get to likelihood theory, we are going to spend the first part of this course reviewing/extending/deepening our knowledge of mathematical and statistical tools.

In particular, lower-level analysis and mathematical statistics courses often focus on single-variable results.

In practice, however, statistics is almost always a multivariate pursuit.

Thus, one of the things we will focus on in this review is covering results you may have seen for single variables in terms of vectors.
Asymptotic theory

- A large amount (but not all) of statistical theory is based on asymptotic, or large sample, arguments.
- Exact theoretical results are often very complicated and difficult to obtain, but we can typically simplify the problem greatly by considering what happens as $n \to \infty$.
- A core idea here from analysis is that of a convergent sequence: $x_n$ converges to $x$ if, as $n$ gets larger, $x_n$ gets closer and closer to $x$.
- We’ll provide a formal definition later (and of course, discuss probabilistic versions), but first, we need to take a step back and define what it means for $x_n$ to be “close” to $x$. 
Norms: Introduction

- Throughout this course, we need to be able to measure the distance between two vectors, or equivalently, the size of a vector; such a measurement is called a norm.
- This is straightforward for scalars – you can simply take the absolute value.
- Vectors are more complicated; as we will see, there are many ways of measuring the size of a vector.
- In order to be a meaningful measure of size, however, there are certain conditions any norm must satisfy.
**Norm: Definition**

- **Definition:** A *norm* is a function \( \| \cdot \| : \mathbb{R}^d \to \mathbb{R} \) such that for all \( x, y \in \mathbb{R}^d \),
  1. \( \| x \| \geq 0 \), with \( \| x \| = 0 \) iff \( x = 0 \) (positivity)
  2. \( \| ax \| = |a| \| x \| \) for any \( a \in \mathbb{R} \) (homogeneity)
  3. \( \| x + y \| \leq \| x \| + \| y \| \) (triangle inequality)

- The triangle inequality is also sometimes expressed as

\[
\| x - z \| \leq \| x - y \| + \| y - z \|,
\]

or

\[
d(x, z) \leq d(x, y) + d(y, z),
\]

where \( d(x, y) \) quantifies the distance between \( x \) and \( y \).
Reverse triangle inequality

- A related inequality:
- **Theorem (reverse triangle inequality):** For any \( x, y \in \mathbb{R}^d \),
  \[
  \|x\| - \|y\| \leq \|x - y\|
  \]
- **Corollary:** For any \( x, y \in \mathbb{R}^d \),
  \[
  \|x\| - \|y\| \leq \|x + y\|
  \]
  \[
  \|y\| - \|x\| \leq \|x + y\|
  \]
  \[
  \|y\| - \|x\| \leq \|x - y\|
  \]
Examples of norms

- By far the most common norm is the Euclidean ($L_2$) norm:
  \[ \|x\|_2 = \sqrt{\sum_i x_i^2} \]

- However, there are many other norms; for example, the Manhattan ($L_1$) norm:
  \[ \|x\|_1 = \sum_i |x_i| \]

- Both Euclidean and Manhattan norms are members of the $L_p$ family of norms: for $p \geq 1$,
  \[ \|x\|_p = \left( \sum_i |x_i|^p \right)^{1/p} \]
Another norm worth knowing about is the $L_\infty$ norm:

$$\|x\|_\infty = \max_i |x_i|,$$

which is the limit of the family of $L_p$ norms as $p \to \infty$.

One last “norm” worth mentioning is the $L_0$ norm:

$$\|x\|_0 = \sum_i 1\{x_i \neq 0\};$$

be careful, however: this is not a proper norm! (why not?)
Matrix norms

- There are also matrix norms, although we will not work with these as often.
- In addition to the three requirements listed earlier, matrix norms must also satisfy a requirement of submultiplicativity:

  \[ \|AB\| \leq \|A\| \|B\|; \]

  unlike the other requirements, this only applies to \(n \times n\) matrices.
- The simplest matrix norm is the Frobenius norm

  \[ \|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2} \]
Another common matrix norm is the *spectral norm*:

\[ \|A\|_2 = \sqrt{\lambda_{\text{max}}}, \]

where \( \lambda_{\text{max}} \) is the largest eigenvalue of \( A^\top A \).

There are many other matrix norms.
Cauchy-Schwarz

• There are several important inequalities involving norms that you should be aware of; the most important is the Cauchy-Schwarz inequality, arguably the most useful inequality in all of mathematics

• **Theorem (Cauchy-Schwarz):** For $x, y \in \mathbb{R}^d$,

\[ x^\top y \leq \|x\|_2 \|y\|_2, \]

where equality holds only if $x = ay$ for some scalar $a$

• Note: the above is *the* Cauchy-Schwarz inequality, but in statistics, its probabilistic version goes by the same name:

\[ \mathbb{E}\left|XY\right| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)} \]

for random variables $X$ and $Y$, with equality iff $X = aY$
Hölder’s inequality

- The Cauchy-Schwarz inequality is actually a special case of Hölder’s inequality:

- **Theorem (Hölder):** For $1/p + 1/q = 1$ and $x, y \in \mathbb{R}^d$,

  $$x^\top y \leq \|x\|_p \|y\|_q,$$

  again with exact equality iff $x = ay$ for some scalar $a$ (unless $p$ or $q$ is exactly 1)

- Probabilistic analogue:

  $$\mathbb{E}|XY| \leq \sqrt[p]{\mathbb{E}|X|^p} \sqrt[q]{\mathbb{E}|Y|^q}$$
Another extremely important inequality is Jensen’s inequality; surely you’ve seen it before, but perhaps not in vector form.

**Theorem (Jensen):** For \( \mathbf{a}, \mathbf{x} \in \mathbb{R}^d \) with \( a_i > 0 \) for all \( i \), if \( g \) is a convex function, then

\[
 g \left( \frac{\sum_i a_i x_i}{\sum_i a_i} \right) \leq \frac{\sum_i a_i g(x_i)}{\sum_i a_i}
\]

**Probabilistic analog:**

\[
 g(\mathbb{E}X) \leq \mathbb{E}g(X)
\]

The inequalities are reversed if \( g \) is concave.
• Getting back to the different norms, there are many important relationships between norms that are often useful to know

• **Theorem:** For all $x \in \mathbb{R}^d$,

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$$

• Obvious, but useful:

$$\|x\|_\infty \leq \|x\|_1 \leq d\|x\|_\infty$$

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\|x\|_\infty$$
Equivalence of norms

• The relationships on the previous slide suggest the following statement, which is in fact always true: for any two norms $a$ and $b$, there exist constants $c_1$ and $c_2$ such that

$$\|x\|_a \leq c_1 \|x\|_b \leq c_2 \|x\|_a$$

• This result is known as the equivalence of norms and means that we can often generalize results for any one norm to all norms

• For example, we will often encounter results that look like:

$$A = B + \|r\|$$

and show that $\|r\| \to 0$, so $A \to B$
By the equivalence of norms, if, say, $\|r\|_1 \to 0$, then $\|r\|_2 \to 0$ and so on for all norms (except not the $L_0$ “norm”!)

In this course, we will almost always be working with the Euclidean norm, so much so that I will typically write $\|x\|$ to mean the Euclidean norm and not even bother with the subscript.

That said, it is important to note that with these relationships, we can always derive corollaries that extend results to other norms.
Like vector norms, matrix norms are also equivalent.

For example,

\[ \|A\|_2 \leq \|A\|_F \leq \sqrt{r} \|A\|_2, \]

where \( r \) is the rank of \( A \).
One essential use of norms is to define what it means for elements of a vector space to be “close”

**Definition:** The neighborhood of a point $p \in \mathbb{R}^d$, denoted $N_\delta(p)$, is the set $\{x : \|x - p\| < \delta\}$.

This will come up quite often in this course

- For example, we will often need to make assumptions about the likelihood function $L(\theta)$
- However, we don’t necessarily need these assumptions to hold everywhere – it’s enough that they hold in a neighborhood of $\theta^*$, the true value of the parameter
Convergence (scalar)

- Let’s now go back and provide a formal definition of convergence, starting with the scalar case.
- A sequence of scalar values $x_n$ is said to converge to $x$, which we denote $x_n \to x$, if for every $\epsilon > 0$, there is a number $N$ such that $n > N$ implies that $|x_n - x| < \epsilon$.
- If you’ve never taken a course in real analysis, pay very close attention to the wording here:
  - We are not saying that there is a single $N$ that always works.
  - Instead, we are saying that if you (1) pick an $\epsilon$, then (2) you can always find an $N$ that works, where $N$ is allowed to depend on $\epsilon$ (and typically, must).
There are two potential ways we could extend this idea to the multivariate case

**Definition:** We say that the vector $x_n$ converges to $x$, denoted $x_n \rightarrow x$, if each element of $x_n$ converges to the corresponding element of $x$.

Alternatively, we can use norms to construct a more direct definition

**Definition:** A sequence $x_n$ is said to converge to $x$, which we denote $x_n \rightarrow x$, if for every $\epsilon > 0$, there is a number $N$ such that $n > N$ implies that $\|x_n - x\| < \epsilon$.

We’ll establish in a moment that these two definitions are equivalent
Continuity

- It’s fairly obvious that, say, \( x_n + y_n \rightarrow x + y \), but what about more complicated functions? Does \( \sqrt{x_n} \rightarrow \sqrt{x} \)? Does \( f(x_n) \rightarrow f(x) \) 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, so functions that possess it are given a specific name: continuous functions.
- **Definition:** A function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is said to be continuous at a point \( p \) if for all \( \epsilon > 0 \), there exists \( \delta > 0 \):

\[
\| x - p \| < \delta \implies | f(x) - f(p) | < \epsilon
\]

- 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 \( \| \cdot \|_1 \).
The norm itself is a continuous function:

**Theorem:** Let $f(x) = \|x\|$, where $\|\cdot\|$ is any norm. Then $f(x)$ is continuous.

One consequence of this result is that element-wise convergence is equivalent to convergence in norm.

**Theorem:** $x_n \to x$ element-wise if and only if $\|x_n - x\| \to 0$. 
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 $x$, the sequence $f_n(x)$ converges. We can then define the *limit function* $f$ by

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

Sequences of functions come up constantly in statistics, the most relevant example being the likelihood function $L(\theta|x_n) = L_n(\theta)$
• Furthermore, we are often interested in combining convergence of the function with convergence of the argument.
• For example, does \( f_n(\hat{\theta}) \to f(\theta) \) as \( \hat{\theta} \to \theta \)?
• 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 \( f_n(x) \to f(x_0) \) as \( x \to x_0 \)?
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

\[ \lim_{x \to 1^-} \lim_{n \to \infty} f_n(x) = 0 \neq f(1) \]
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 $f_1, f_2, \ldots : \mathbb{R}^d \rightarrow \mathbb{R}$ converges uniformly on a set $E$ to a function $f$ if for every $\epsilon > 0$ there exists $N$ such that $n > N$ implies

  $$|f_n(x) - f(x)| < \epsilon$$

  for all $x \in E$.

- **Corollary:** $f_n \rightarrow f$ uniformly on $E$ if and only if

  $$\sup_{x \in E} |f_n(x) - f(x)| \rightarrow 0.$$
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, $\alpha$ is a least upper bound of the set $E$ if (i) $\alpha$ is an upper bound of $E$ and (ii) if $\gamma < \alpha$, then $\gamma$ 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
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 \( f_n(x) \to f(x_0) \) as \( x \to x_0 \).
- Note that this argument does not work without uniform convergence.
The theorem on the previous page can actually be made somewhat stronger:

**Theorem:** Suppose $f_n \to f$ uniformly on $E$ and that $\lim_{x \to x_0} f_n(x)$ exists for all $n$. Then for any limit point $x_0$ of $E$,

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

**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 $E$. 

Patrick Breheny University of Iowa
Likelihood Theory (BIOS 7110)
• There are number of related concepts similar to uniform convergence

• **Definition:** A function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is called *uniformly continuous* if for all \( \epsilon > 0 \), there exists \( \delta > 0 \) such that for all \( x, y \in \mathbb{R}^d : \|x - y\| < \delta \), we have \( |f(x) - f(y)| < \epsilon \).

• 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 \( X_n \).