# Analysis review, Part 1

Patrick Breheny

August 26

# Introduction

- 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

# Norms: Introduction

- Central to this pursuit is the idea of measuring 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
- However, in order to be a meaningful measure of size, 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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,
  - $\|\mathbf{x}\| \geq 0$, with $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = \mathbf{0}$ (positivity)
  - $\|a\mathbf{x}\| = |a| \, \|\mathbf{x}\|$ for any $a \in \mathbb{R}$ (homogeneity)
  - $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality)
- The triangle inequality is also sometimes expressed as

$$\|\mathbf{x} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|,$$

or

$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}),$$

where $d(\mathbf{x}, \mathbf{y})$ quantifies the distance between $\mathbf{x}$ and $\mathbf{y}$

# Reverse triangle inequality

- A related inequality:
- **Theorem (reverse triangle inequality):** For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\|\mathbf{x}\| - \|\mathbf{y}\| \leq \|\mathbf{x} - \mathbf{y}\|$$

- **Corollary:** For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\|\mathbf{x}\| - \|\mathbf{y}\| \leq \|\mathbf{x} + \mathbf{y}\|$$
$$\|\mathbf{y}\| - \|\mathbf{x}\| \leq \|\mathbf{x} + \mathbf{y}\|$$
$$\|\mathbf{y}\| - \|\mathbf{x}\| \leq \|\mathbf{x} - \mathbf{y}\|$$

# Examples of norms

- By far the most common norm is the Euclidean ($L_2$) norm:

$$\|\mathbf{x}\|_2 = \sqrt{\sum\nolimits_i x_i^2}$$

- However, there are many other norms; for example, the Manhattan ($L_1$) norm:

$$\|\mathbf{x}\|_1 = \sum\nolimits_i |x_i|$$

- Both Euclidean and Manhattan norms are members of the $L_p$ family of norms: for $p \geq 1$,

$$\|\mathbf{x}\|_p = \left(\sum\nolimits_i |x_i|^p\right)^{1/p}$$

# Examples of norms (cont'd)

- Another norm worth knowing about is the $L_\infty$ norm:

$$\|\mathbf{x}\|_\infty = \max_i |x_i|,$$

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:

$$\|\mathbf{x}\|_0 = \sum\nolimits_i 1\{x_i \neq 0\};$$

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*:

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\| \, ;$$

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

$$\|\mathbf{A}\|_F = \sqrt{\sum\nolimits_{i,j} a_{ij}^2}$$

# Spectral norm

- Another common matrix norm is the *spectral norm*:

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}},$$

  where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{A}^\top\mathbf{A}$
- 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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\mathbf{x}^\top \mathbf{y} \leq \|\mathbf{x}\|_2 \, \|\mathbf{y}\|_2 \, ,$$

  where equality holds only if $\mathbf{x} = a\mathbf{y}$ for some scalar $a$
- Note: the above is *the* Cauchy-Schwarz inequality, but in statistics, its probabilistic version goes by the same name:

$$\mathbb{E}\,|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$$

  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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\mathbf{x}^\top \mathbf{y} \leq \|\mathbf{x}\|_p \, \|\mathbf{y}\|_q \, ,$$

  again with exact equality iff $\mathbf{x} = a\mathbf{y}$ for some scalar $a$ (unless $p$ or $q$ is exactly 1)
- Probabilistic analogue:

$$\mathbb{E}\,|XY| \leq \sqrt[p]{\mathbb{E}\,|X|^p}\,\sqrt[q]{\mathbb{E}\,|Y|^q}$$

# Jensen's inequality

- 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

# Relationships between norms

- Getting back to the different norms, there are many important relationships between norms that are often useful to know
- **Theorem:** For all $\mathbf{x} \in \mathbb{R}^d$,

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

- Obvious, but useful:

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq d\,\|\mathbf{x}\|_\infty$$
$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{d}\,\|\mathbf{x}\|_\infty$$

# 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

$$\|\mathbf{x}\|_a \leq c_1 \|\mathbf{x}\|_b \leq c_2 \|\mathbf{x}\|_a$$

- 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:

$$A = B + \|\mathbf{r}\|$$

and show that $\|\mathbf{r}\| \to 0$, so $A \approx B$

# Equivalence of norms (cont'd)

- 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 $\|\mathbf{x}\|$ 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

# Equivalence of matrix norms

- Like vector norms, matrix norms are also equivalent
- For example,

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{r}\,\|\mathbf{A}\|_2\,,$$

where $r$ is the rank of $\mathbf{A}$

# Continuity

- One essential use of norms is to define what it means for elements of a vector space to be "local"
- Specifically, the *neighborhood* of a point $\mathbf{p} \in \mathbb{R}^d$ is the set $\{\mathbf{x} : \|\mathbf{x} - \mathbf{p}\| < \delta\}$, abbreviated $N_\delta(\mathbf{p})$
- Needed, for example, in the definition of a continuity for a vector-valued function:
- **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 $\|\cdot\|_1$

# 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
- **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 $\mathbf{x}$.
- **Theorem:** $\mathbf{x}_n \to \mathbf{x}$ if and only if $\|\mathbf{x}_n - \mathbf{x}\| \to 0$.

# Real-valued functions: Derivative and gradient

- This brings us to the important topic of vector calculus, which we will use frequently in this course
- **Definition:** For a function $f : \mathbb{R}^d \to \mathbb{R}$, its *derivative* is the $1 \times d$ row vector

$$\dot{f}(\mathbf{x}) = \left[ \frac{\partial f}{\partial x_1} \cdots \frac{\partial f}{\partial x_d} \right]$$

- In statistics, it is generally more common (but not always the case) to use the gradient (also called "denominator layout" or the "Hessian formulation")

$$\nabla f(\mathbf{x}) = \dot{f}(\mathbf{x})^\top;$$

i.e., $\nabla f(\mathbf{x})$ is a $d \times 1$ column vector

# Vector-valued functions

- **Definition:** For a function $f : \mathbb{R}^d \to \mathbb{R}^k$, its *derivative* is the $k \times d$ matrix with $ij$th element

$$\dot{\mathbf{f}}(\mathbf{x})_{ij} = \frac{\partial f_i(\mathbf{x})}{\partial x_j}$$

- Correspondingly, the gradient is a $d \times k$ matrix:

$$\nabla \mathbf{f}(\mathbf{x}) = \dot{\mathbf{f}}(\mathbf{x})^\top$$

- In our course, this will usually come up in the context of taking second derivatives; however, by the symmetry of second derivatives, we have

$$\nabla^2 f(\mathbf{x}) = \ddot{f}(\mathbf{x})$$

# Vector calculus identities

| | |
|---|---|
| Inner product: | $\nabla_{\mathbf{x}}(\mathbf{A}^\top \mathbf{x}) = \mathbf{A}$ |
| Quadratic form: | $\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}^\top \mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$ |
| Chain rule: | $\nabla_{\mathbf{x}}\mathbf{f}(\mathbf{y}) = \nabla_{\mathbf{x}}\mathbf{y}\nabla_{\mathbf{y}}\mathbf{f}$ |
| Product rule: | $\nabla(\mathbf{f}^\top \mathbf{g}) = (\nabla \mathbf{f})\mathbf{g} + (\nabla \mathbf{g})\mathbf{f}$ |
| Inverse function theorem: | $\nabla_{\mathbf{x}}\mathbf{y} = (\nabla_{\mathbf{y}}\mathbf{x})^{-1}$ |

Note that for the inverse function theorem to apply, the gradient must be invertible

# Vector calculus identities (row-vector layout)

| | |
|---|---|
| Inner product: | $D_{\mathbf{x}}(\mathbf{A}\mathbf{x}) = \mathbf{A}$ |
| Quadratic form: | $D_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}^\top \mathbf{x}) = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top)$ |
| Chain rule: | $D_{\mathbf{x}}\mathbf{f}(\mathbf{y}) = D_{\mathbf{y}}\mathbf{f}\, D_{\mathbf{x}}\mathbf{y}$ |
| Product rule: | $D(\mathbf{f}^\top \mathbf{g}) = \mathbf{g}^\top \dot{\mathbf{f}} + \mathbf{f}^\top \dot{\mathbf{g}}$ |
| Inverse function theorem: | $D_{\mathbf{x}}\mathbf{y} = (D_{\mathbf{y}}\mathbf{x})^{-1}$ |

I don't expect to use these, but for your future reference, here they are

# Practice

**Exercise:** In linear regression, the ridge regression estimator is obtained by minimizing the function

$$\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2 + \lambda \left\|\boldsymbol{\beta}\right\|^2,$$

where $\lambda$ is a prespecified tuning parameter. Show that

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}.$$

# Integration and measure: Introduction

- Our final topic for today is a brief treatment of measure 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
- In particular, we will go over
  - The Riemann-Stieltjes integral
  - The Lebesgue decomposition theorem

# 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 $\mu$ be a bounded, nondecreasing function on $[a, b]$, and let

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

  note that $\mu_i \geq 0$

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

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

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

# 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 (measurable) with respect to* $\mu$ over $[a, b]$, and we denote their common value

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

or sometimes

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

# Implications for probability

- The application to probability is clear: any CDF can play the role of $\mu$ (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 (measurable),

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

- 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

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

where
  - $F_{\mathsf{D}}$ is the discrete component (i.e., probability is given by a sum of point masses)
  - $F_{\mathsf{AC}}$ is the absolutely continuous component (i.e., probability is given by an integral with respect to a density function)
  - $F_{\mathsf{SC}}$ 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)