

# Analysis review: Vector calculus and measure

Patrick Breheny

August 28

# Introduction

- Next up, we'll be reviewing the central tools of calculus: derivatives and integrals
- I assume that you're already quite familiar with ordinary scalar derivatives, but not necessarily with vector derivatives
- Likewise, I assume that you know how to take integrals, but perhaps not with its 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 theory and allows for a nice unification of continuous and discrete probability theory

# Real-valued functions: Derivative and gradient

- Vector calculus is extremely important in statistics, and we will use it frequently in this course
- **Definition:** For a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , 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  $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ , 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:

$$\nabla_{\mathbf{x}}(\mathbf{x}^{\top} \mathbf{A} \mathbf{x}) = (\mathbf{A} + \mathbf{A}^{\top}) \mathbf{x}$$

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:

$$\nabla_{\mathbf{x}} \mathbf{y} = (\nabla_{\mathbf{y}} \mathbf{x})^{-1}$$

## 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

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

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

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

# Integration and measure: Introduction

- Our other 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  $x_0, x_1, \dots, x_n$  such that

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

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

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

note that  $\Delta\mu_i \geq 0$

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

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

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

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

or sometimes

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

# 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 measurable functions such that  $f_n \rightarrow f$ . If there exists a measurable function  $g$  such that  $|f_n(x)| \leq g(x)$  for all  $n$  and all  $x$ , then

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

- 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  $\mu$  (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 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

$$F = F_D + F_{AC} + F_{SC},$$

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)