Mathematics/Statistics Bootcamp

Part II: Calculus

Steven Winter   Christine Shen

Department of Statistical Science
Duke University

MSS Orientation, August 2022

Overview

Sequences and Series

Limits of Sequences

A **sequence** is an ordered list of numbers. We write $\{a_1, a_2, a_3, \dots\}$, $\{a_n\}$, $\{a_n\}_{n=1}^{\infty}$, (a_n), etc.

A sequence $\{a_n\}$ has **limit** L ($\lim_{n\to\infty} a_n = L$, or $a_n \to L$ as $n \to \infty$) if

$$\forall \varepsilon > 0,\ \exists N_\varepsilon \in \mathbb{N} \text{ such that } n \geq N_\varepsilon \implies |a_n - L| < \varepsilon.$$

If all limits exist, then

1. $\lim_{n\to\infty}(ca_n) = c(\lim_{n\to\infty} a_n)$.
2. $\lim_{n\to\infty}(a_n + b_n) = (\lim_{n\to\infty} a_n) + (\lim_{n\to\infty} b_n)$.
3. $\lim_{n\to\infty}(a_n b_n) = (\lim_{n\to\infty} a_n)(\lim_{n\to\infty} b_n)$.
4. $a_n \leq b_n \leq c_n \implies \lim_{n\to\infty} a_n \leq \lim_{n\to\infty} b_n \leq \lim_{n\to\infty} c_n$ (Squeeze theorem).

Sequence Theorems (Optional)

If for every $n \in \mathbb{N}$, $a_n \leq a_{n+1}$ (increasing) or $a_n \geq a_{n+1}$ (decreasing), then the sequence $\{a_n\}$ is **monotonic**.

If there exists a number $M > 0$ such that $|a_n| \leq M$ for every n then the sequence $\{a_n\}$ is **bounded**.

Monotonic Sequence theorem: Every bounded, monotonic sequence is convergent (has a limit).

Bolzano–Weierstrass theorem: Every bounded sequence has a convergent subsequence.

Series Basics

Fix a sequence (a_k) and let

$$s_n = \sum_{k=1}^{n} a_k$$

be the sequence of partial sums. A **series** is the limit of s_n (written $\sum_{n=1}^{\infty} a_n$).

The series converges if (s_n) has a limit. Same properties as sequence limits (except for products).

Key example is the **geometric series**:

$$\sum_{n=1}^{\infty} rx^{n-1} = r + rx + rx^2 + \cdots$$

which converges to $r/(1-x)$ if $|x| < 1$ and diverges otherwise.

Series Theorems

The comparison test: Suppose that $\sum a_n$ and $\sum b_n$ are series with positive terms.

(i) If $\sum b_n$ is convergent and $a_n \leq b_n$ for all n, then $\sum a_n$ is also convergent;

(ii) If $\sum b_n$ is divergent and $a_n \geq b_n$ for all n, then $\sum a_n$ is also divergent.

Many other tests:
`https://en.wikipedia.org/wiki/Series_(mathematics)`

Discussion

1. Fix $x, p \in \mathbb{R}$. Discuss convergence of the sequences

$$a_n = \frac{1}{n^p}, \quad b_n = \left(1 + \frac{x}{n}\right)^n, \quad c_n = \cos(nx).$$

2. Fix $p \in \mathbb{R}$. Discuss convergence of the series

$$A = \sum_{n=1}^{\infty} \frac{1}{n^p}, \quad B = \sum_{n=1}^{\infty} \frac{\log(n)}{n}.$$

3. Assume

$$\left| \sum_{n=1}^{\infty} a_n \right| < \infty.$$

   What, if anything, can we say about the limit of a_n?

Limits and Derivatives

Pointwise Limits

A function $f : \mathbb{R} \to \mathbb{R}$ has **limit** L at a (written $\lim_{x \to a} f(x) = L$) if

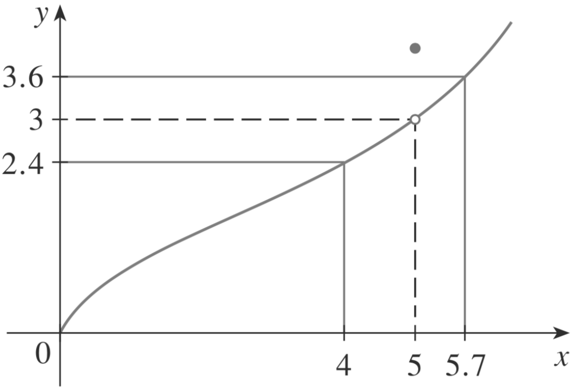
$$\forall \varepsilon > 0, \ \exists \delta_\varepsilon > 0 \text{ such that } |x - a| < \delta_\varepsilon \implies |f(x) - L| < \varepsilon.$$

Equivalently, if the sequence $(f(x_1), f(x_2), \ldots)$ converges to L for *any* sequence x_n converging to a. Same properties as sequences.

Left-hand limit: $\lim_{x \to a^-} f(x) = L$ if for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $a - \delta < x < a$.

Right-hand limit: $\lim_{x \to a^+} f(x) = L$ if for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $a < x < a + \delta$.

Example



- What is $\lim_{x \to 5^-} f(x)$?
- What is $\lim_{x \to 5^+} f(x)$?
- What is $\lim_{x \to 5} f(x)$?
- What is $\lim_{x \to 4} f(x)$?

Continuity Basics

A function f is continuous at a if

$$\lim_{x \to a} f(x) = f(a).$$

Right continuous at a if $\lim_{x \to a^-} f(x) = f(a)$; **left continuous** at a if $\lim_{x \to a^+} f(x) = f(a)$.

You should know:

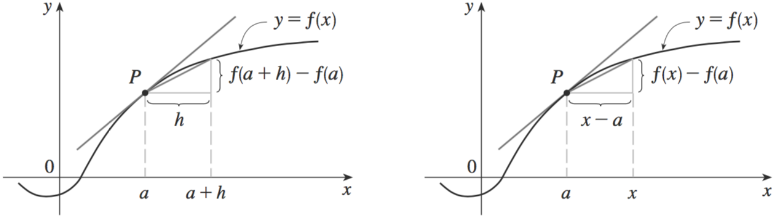
- Continuous functions form a vector space!
- Composition of continuous functions is continuous.
- Results like the intermediate value theorem.

Derivative Basics

The derivative of function f at $a \in X$, denoted by $f'(a)$, is

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

if this limit exists. Differentiable implies continuous.



(a) $f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$
$=$ slope of tangent at P
$=$ slope of curve at P

(b) $f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$
$=$ slope of tangent at P
$=$ slope of curve at P

Derivative Rules

Derivatives of some common functions:

- $f(x) = const$, then $f'(x) = 0$;
- $f(x) = x^\alpha, \alpha \neq 0$, then $f'(x) = \alpha x^{\alpha-1}$;
- $(e^x)' = e^x$, $(\ln x)' = 1/x$ $(x > 0)$;
- $(\sin x)' = \cos x$, $(\cos x)' = -\sin x$, $(\tan x)' = 1/\cos^2 x$;

If both $f(x)$ and $g(x)$ are differentiable:

- $(cf(x))' = cf'(x)$, $(f(x) + g(x))' = f'(x) + g'(x)$;
- $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$;
- $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}$ (assume $g(x) > 0$);
- The **chain rule**: if $F = f \circ g$, then $F'(x) = f'(g(x))g'(x)$.

Taylor Series

If f is infinitely differentiable at a, then it can be expressed as a power series:

$$\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \end{aligned}$$

This is called the **Taylor series** of f at a.

Intuition: we can approximate nice functions arbitrarily well with polynomials.

Very useful in statistics: CLT, delta method, optimization algorithms, etc.

Exercises

1. Differentiate

$$f(x) = xe^{-x}, \quad g(x) = 1 - cos^2(x), \quad h(x) = \frac{\log(x)}{x}.$$

2. Fix $\mu, x \in \mathbb{R}$ and $\gamma > 0$. Let

$$f(x) = \frac{1}{\sqrt{\gamma}} \exp\left(-\frac{(x-\mu)^2}{\gamma}\right)$$

   Find $x_0 \in \mathbb{R}$ such that the tangent line of $f(x)$ at x_0 is horizontal.

3. Find the Taylor series of $f(x) = e^x$ around 0.
4. Find $\lim_{x\to 0}(1+x)^{1/x}$.

Integrals

The Fundamental Theorem of Calculus

If f is continuous on $[a, b]$, then the function

$$g(x) = \int_a^x f(s)ds$$

is continuous on $[a, b]$, differentiable on (a, b), and $g'(x) = f(x)$.

If F is any anti-derivative of f $(F' = f)$, then

$$\int_a^b f(x)dx = F(b) - F(a).$$

Definite Integral Rules

Let $a \leq d \leq b \in \mathbb{R}$:

- If $c \in \mathbb{R}$ is a constant, then $\int_a^b c dx = c(b-a)$;
- $\int_a^b cf(x)dx = c\int_a^b f(x)dx$;
- $\int_a^b [f(x)+g(x)]dx = \int_a^b f(x)dx + \int_a^b g(x)dx$;
- $\int_a^d f(x)dx + \int_d^b f(x)dx = \int_a^b f(x)dx$;
- If $f(x) \geq g(x)$ for $a \leq x \leq b$, then $\int_a^b f(x)dx \geq \int_a^b g(x)dx$;
- If $m \leq f(x) \leq M$ for $a \leq x \leq b$, then $m(b-a) \leq \int_a^b f(x)dx \leq M(b-a)$.

Useful Formulas for Integration

- **Substitution**: If $u = g(x)$ is continuously differentiable on $[a, b]$ and f is continuous on the range of u, then

$$\int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(u)du.$$

  Proof: chain rule. Example: $\int_a^b (x-1)^{50}dx$.

- **Integration by parts**: If functions u and v are both continuously differentiable on $[a, b]$, then

$$\int_a^b u(x)v'(x)dx = [u(x)v(x)]|_a^b - \int_a^b v(x)u'(x)dx.$$

  Proof: product rule. Example: $\int_a^b xe^x dx$.

Improper Integrals

Assume $\int_a^b f(x)dx$ exists for every $b \geq a$ and define

$$\int_a^\infty f(x)dx = \lim_{b\to\infty} \int_a^b f(x)dx.$$

if the limit exists. Likewise for $\int_{-\infty}^b f(x)dx$.

Extend these to define

$$\int_{-\infty}^\infty f(x)dx = \lim_{b\to-\infty} \lim_{a\to\infty} \int_a^b f(x)dx.$$

if all limits exist.

Intuition: solve integral on a “safe“ domain, take limits.

Discontinuous Integrand

If f is continuous on $[a, b)$ and is discontinuous at b, then

$$\int_a^b f(x)dx = \lim_{t \to b^-} \int_a^t f(x)dx$$

if this limit exists. Likewise if f is continuous on $(a, b]$.

Same intuition: solve integral on a “safe“ domain, take limits.

Exercises

1. Calculate $\int_1^e \frac{\ln(x)}{x} dx$.
2. Calculate $\int_0^\pi x\cos(x) dx$.
3. Fix $p \in \mathbb{R}$. Calculate $I_1 = \int_0^1 \frac{1}{x^p} dx$ and $I_2 = \int_1^\infty \frac{1}{x^p} dx$.

Multivariate Calculus

Partial Derivatives

Suppose $f : \mathbb{R}^n \to \mathbb{R}$. The **partial derivative** with respect to the ith variable x_i is

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h}.$$

Strategy: treat other variables as constants.

If $\frac{\partial^2 f}{\partial x_i \partial x_j}$ and $\frac{\partial^2 f}{\partial x_j \partial x_i}$ are both continuous on $\mathbb{R}^n$, then $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$.

Gradients and Hessians

Suppose $f : \mathbb{R}^n \to \mathbb{R}$. If all first-order partial derivatives exist, then the **gradient**[1] is

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)^T .$$

Intuition: points uphill.

If all second-order partial derivatives exist, then the **Hessian** is

$$H = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix} .$$

Intuition: local curvature. When is this symmetric?

[1] Technically $\nabla f(\mathbf{x})$ is a map $T_{\mathbf{x}}\mathbb{R}^n \to T_{f(\mathbf{x})}\mathbb{R}$.

Jacobians

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$. If all first-order partial derivatives exist, then the **Jacobian** is

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

.

Intuition: locally approximates f as a linear transformation (stretching, rotating, permuting, etc).

When $n = m$, $|J|$ describes how f locally distorts volume.

Change of Variables

Consider integrating a function $f : \mathbb{R}^n \to \mathbb{R}$ over a set $U \subseteq \mathbb{R}^n$. Let $\varphi : U \to \mathbb{R}^n$ denote a change of coordinates with Jacobian J. Then under some conditions,

$$\int_U f(\mathbf{x}) d\mathbf{x} = \int_{\varphi(U)} f(\varphi(\mathbf{u})) |J(\mathbf{u})| d\mathbf{u}.$$

Extremely useful for finding probability density functions.

Note: need Fubini's theorem to freely change order of integrals.

Matrix Calculus

We sometimes need to differentiate with respect to a matrix.

E.g., if we model data $\mathbf{x}_1, ..., \mathbf{x}_n \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and want the MLE of $\boldsymbol{\Sigma}$, then we differentiate

$$\ell(\boldsymbol{\Sigma}) = -\frac{n}{2}|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mu)$$

with respect to $\boldsymbol{\Sigma}$.

Rely on

- The Matrix Cookbook
- `https://en.wikipedia.org/wiki/Matrix_calculus`
- Matrix Algebra From a Statistician's Perspective by Harville.

Optimization

Extrema

Derivative condition: If f has a local minimum or maximum at c and $f'(c)$ exists, then $f'(c) = 0$. Converse is false.

Second derivative test: If $f'(c) = 0$ and f has second derivative on $(c - \epsilon_0, c + \epsilon_0)$ for some $\epsilon_0 > 0$, then

- $f''(c) > 0$ implies c is a local minimum,
- $f''(c) < 0$ implies c is a local maximum.

Why? Draw pictures.

Multivariate analogue: local optima satisfy $\nabla f(\mathbf{c}) = 0$. Minimum if Hessian is positive definite; maximum if negative definite.

Useful for minimizing error, finding MLEs, etc.

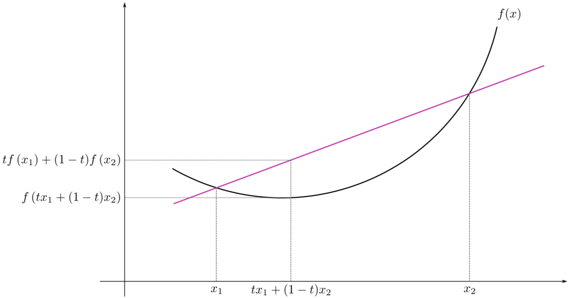
Convexity Basics

A subset X of $\mathbb{R}^n$ is **convex** if for any $x, y \in X$ and $t \in [0, 1]$.

$$tx + (1 - t)y \in X.$$

A function $f : X \to \mathbb{R}$ is **convex** if for any $x, y \in X$ and $t \in [0, 1]$,

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y).$$



Convexity Theorems

Suppose a function $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable on an open set. The following are equivalent

- f is convex.
- $f(x) > f(y) + f'(y)(x - y)$; the graph is above all tangents.
- $f''(x) \geq 0$.

Similar tests for multivariate functions.

Any local minimum of a convex function is also a global minimum.

Lagrange Multipliers

Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be differentiable. **Lagrange multipliers** are a method of optimizing f subject to $g = 0$.

Procedure:

1. Solve the following system of equations:

$$\begin{aligned} \nabla f(\mathbf{x}) &= \lambda \nabla g(\mathbf{x}) \\ g(\mathbf{x}) &= 0 \end{aligned}$$

2. Plug all solutions into f to find the global optima.

Often introduce the Lagrangian,

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x}).$$

Example

Consider rolling a k-sided die n times. Let p_j denote the true probability of face j and X_j count the number of times we see face j. Mathematically, $(X_1 ... X_k) \sim Multinomial(n, p_1 ... p_k)$.

We want to infer $\mathbf{p} = (p_1, ..., p_k)$ from the data, e.g. with the MLE. Requires maximizing the log-likelihood subject to $\sum_{i=1}^k p_i = 1$.

The Lagrangian is

$$\begin{aligned}
\mathcal{L}(\mathbf{p}, \lambda) &= \ell(\mathbf{p}; X_1, ..., X_k) + \lambda\left(1 - \sum_{i=1}^k p_i\right) \\
&= log(n!) - \sum_{i=1}^k log(x_i!) + \sum_{i=1}^k x_i log(p_i) + \lambda\left(1 - \sum_{i=1}^k p_i\right)
\end{aligned}$$

Example

First solve $\nabla\mathcal{L}(\mathbf{p}, \lambda) = 0$. The partial derivatives are

$$\frac{\partial\mathcal{L}}{\partial p_j} = \frac{x_j}{p_j} - \lambda.$$

Setting to zero gives $\hat{p}_j = x_j/\lambda$.

Now plug this into $g(\mathbf{p}) = 0$ to find λ.

$$1 = \sum_{i=1}^{k} \hat{p}_i = \sum_{i=1}^{k} \frac{x_i}{\lambda} = \frac{1}{\lambda}\sum_{i=1}^{k} x_i.$$

Solving for λ gives $\lambda = \sum_{i=1}^{k} x_j$, hence the MLE is $\hat{p}_j = x_j / \sum_{i=1}^{k} x_i$

Exercises

1. Let $f : \mathbb{R}_+ \times [0, 2\pi] \to \mathbb{R}^2$ transform polar coordinates to Cartesian coordinates:

   $$(r, \theta) \mapsto (r \cos(\theta), r \sin(\theta)).$$

   Find the Jacobian and its determinant.
2. Prove/disprove convexity for the following functions:

   $$f(x) = |x|, \quad g(x) = \log(x^2 + 1), \quad h(x) = e^{-x}.$$

3. Fix $\alpha, \beta > 0$. Find the global maximum of

   $$f(x) = x^{\alpha-1} e^{-\beta x} \mathbf{1}(x > 0).$$

   Justify all claims.

Acknowledgements

Past contributors:

- Jordan Bryan, PhD student
- Brian Cozzi, MSS alumni
- Michael Valancius, MSS alumni
- Graham Tierney, PhD student
- Becky Tang, PhD student