Analysis review: O notation, Taylor series, and linear algebra

Patrick Breheny

September 1

Introduction

One final lecture of analysis review, in which we go over three indispensable tools that we will use constantly in the remainder of the course:

- O, o notation
- Taylor series expansions
- Linear algebra

o-notation: Motivation

- When investigating the asymptotic behavior of functions, 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

- There is where something called o-notation comes in: 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 $n \to \infty$.
- For example,

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

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_0 such that

  $$|X_n| < M$$

  for all $n > n_0$. Likewise, X_n is said to be $O(r_n)$ if there exist M and n_0 such that for all $n > n_0$,

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

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

Algebra of O, o notation

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

Theorem: For $a \leq b$:

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

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

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

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

- 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

Taylor series: Introduction

- It is difficult to overstate the importance of Taylor series expansions to statistical theory, and for that reason we are now going to cover them fairly extensively
- In particular, Taylor's theorem comes in a number of versions, and it is worth knowing several of them, since they come up in statistics quite often
- Furthermore, students often have not seen the multivariate versions of these expansions

Taylor's theorem

- **Theorem (Taylor):** Suppose n is a positive integer and $f : \mathbb{R} \to \mathbb{R}$ is n times differentiable at a point x_0. Then

  $$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + R_{n+1}(x, x_0),$$

  where the remainder R_{n+1} satisfies

  $$R_{n+1}(x, x_0) = o(|x - x_0|^n) \text{ as } x \to x_0$$

- If $f^{(n+1)}(x_0)$ exists, you could also say that R_{n+1} is $O(|x - x_0|^{n+1})$
- This form of the remainder is sometimes called the *Peano* form

Taylor's theorem: Lagrange form

- **Theorem (Taylor):** Suppose $f : \mathbb{R} \to \mathbb{R}$ is n times differentiable on an open interval containing x_0. Then for any point x in that interval, there exists $\bar{x} \in (x, x_0)$:

$$R_n(x, x_0) = \frac{f^{(n)}(\bar{x})}{(n)!}(x - x_0)^n.$$

- This is also known as the *mean-value form*, as the mean value theorem is the central idea in proving the result
- Note that we have a stronger result, but at the cost of stronger assumptions: $f^{(n)}$ must exist along the entire interval from x to x_0, not just at the point x_0

Example: Absolute value

- For example, consider approximating the function $f(x) = |x|$ at $x_0 = -0.1$
- Note that f' exists at x_0, but not at 0
- The basic form of Taylor's theorem says that if we get close enough to x_0, the approximation $f(-0.1) + f'(-0.1)(x + 0.1)$ becomes very accurate – indeed, the remainder is exactly zero for any x within 0.1 of x_0
- However, suppose $x = 0.2$; since f is not differentiable at zero, we are not guaranteed the existence of a point $\bar{x}$ such that

$$f(0.2) = f(-0.1) + 0.3f'(\bar{x});$$

and indeed in this case no such point exists

Lagrange bound

- One reason why the Lagrange form is more powerful is that it allows us to establish error bounds – to know exactly how close x must be to x_0 in order to ensure that the approximation error is less than ϵ
- In particular, if there exists an M such that $\left|f^{(n+1)}(x)\right| \leq M$ over the interval (x, x_0), then

$$|R_{n+1}(x)| \leq \frac{M}{(n+1)!} \left|x - x_0\right|^{n+1}$$

Multivariable forms of Taylor's theorem

- We now turn our attention to the multivariate case
- For the sake of clarity, I'll present the first- and second-order expansions for each of the previous forms, rather than abstract formulae involving $f^{(n)}$
- Lastly, I'll provide a form that goes out to third order, although higher orders are less convenient as they can't be represented compactly using vectors and matrices

Taylor's theorem

- **Theorem (Taylor):** Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at a point $\mathbf{x}_0$. Then

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|)$$

- **Theorem (Taylor):** Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable at a point $\mathbf{x}_0$. Then

$$\begin{aligned} f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \\ \tfrac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|^2) \end{aligned}$$

Taylor's theorem: Lagrange form

- **Theorem (Taylor):** Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable on $N_r(\mathbf{x}_0)$. Then for any $\mathbf{x} \in N_r(\mathbf{x}_0)$, there exists $\bar{\mathbf{x}}$ on the line segment connecting $\mathbf{x}$ and $\mathbf{x}_0$ such that

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \mathbf{x}_0)$$

- **Theorem (Taylor):** Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable on $N_r(\mathbf{x}_0)$. Then for any $\mathbf{x} \in N_r(\mathbf{x}_0)$, there exists $\bar{\mathbf{x}}$ on the line segment connecting $\mathbf{x}$ and $\mathbf{x}_0$ such that

$$\begin{aligned} f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \\ \tfrac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\bar{\mathbf{x}})(\mathbf{x} - \mathbf{x}_0) \end{aligned}$$

- "$\bar{\mathbf{x}}$ on the line segment connecting $\mathbf{x}$ and $\mathbf{x}_0$" means that there exists $w \in [0, 1]$ such that $\bar{\mathbf{x}} = w\mathbf{x} + (1 - w)(\mathbf{x}_0)$

Taylor's theorem: Third order

Theorem (Taylor): Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is three times differentiable on $N_r(\mathbf{x}_0)$. Then for any $\mathbf{x} \in N_r(\mathbf{x}_0)$, there exists $\bar{\mathbf{x}}$ on the line segment connecting $\mathbf{x}$ and $\mathbf{x}_0$ such that

$$\begin{aligned}
f(\mathbf{x}) = f(\mathbf{x}_0) &+ \sum_{j=1}^{d} \frac{\partial f(\mathbf{x}_0)}{\partial x_j}(x_j - x_{0j}) \\
&+ \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d} \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_j \partial x_k}(x_j - x_{0j})(x_k - x_{0k}) \\
&+ \frac{1}{6}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{\ell=1}^{d} \frac{\partial^3 f(\bar{\mathbf{x}})}{\partial x_j \partial x_k \partial x_\ell}(x_j - x_{0j})(x_k - x_{0k})(x_\ell - x_{0\ell}),
\end{aligned}$$

where $\partial f(\mathbf{x}_0)/\partial x_j$ is shorthand for $\partial f(\mathbf{x})/\partial x_j$ evaluated at $\mathbf{x}_0$

Vector-valued functions

- The preceding slides represent the most common uses of Taylor series approximations in statistics, although in this course we will also occasionally need to take approximations of vector-valued functions
- This can be represented in a variety of ways, but the following is simple and suffices for our purposes
- **Theorem:** Suppose $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^k$ is twice differentiable on $N_r(\mathbf{x}_0)$, and that $\nabla^2 f$ is bounded on $N_r(\mathbf{x}_0)$. Then for any $\mathbf{x} \in N_r(\mathbf{x}_0)$,

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \left[\nabla \mathbf{f}(\mathbf{x}_0) + O(\|\mathbf{x} - \mathbf{x}_0\|)\mathbf{1}_{d\times k}\right]^\top (\mathbf{x} - \mathbf{x}_0),$$

  where $\mathbf{1}$ is a matrix of ones (i.e., every element equals one)

Remark

- The reason we need a separate theorem along these lines is that unfortunately, there is not a Lagrange-type result for vector-valued functions
- In other words, it is **not** true that there exists an $\bar{\mathbf{x}}$ such that

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \nabla \mathbf{f}(\bar{\mathbf{x}})^\top (\mathbf{x} - \mathbf{x}_0);$$

  such a point exists for each element of $\mathbf{f}$ separately, but usually the same point will not work for both f_1 and f_2 (and so on)
- Thus, instead of $\nabla \mathbf{f}(\bar{\mathbf{x}})$, we would have a matrix with columns $\nabla f_1(\bar{\mathbf{x}}_1), \nabla f_2(\bar{\mathbf{x}}_2)$, and so on

Linear algebra: Linear and quadratic forms

- Our last mathematical topic to review/reference is linear algebra, which we will use right away in our next lecture on the multivariate normal distribution
- I'll start by just listing some useful identities/relationships involving matrix products and their scalar representations:

$$\mathbf{a}^\top \mathbf{x} = \sum_i a_i x_i; \quad \mathbf{1}^\top \mathbf{x} = \sum_i x_i$$

$$\mathbf{A}^\top \mathbf{x} = (\sum_i a_{i1} x_i \quad \cdots \quad \sum_i a_{ik} x_i)^\top$$

$$\mathbf{a}^\top \mathbf{W} \mathbf{x} = \sum_i \sum_j a_i w_{ij} x_j; \quad \mathbf{a}^\top \mathbf{1} \mathbf{x} = \sum_i \sum_j a_i x_j$$

Inverses

- **Definition:** The *inverse* of an $n \times n$ matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix satisfying $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.
- Note: We're sort of getting ahead of ourselves by saying that $\mathbf{A}^{-1}$ is "the" matrix satisfying $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$, but it is indeed the case that if a matrix has an inverse, the inverse is unique

Singular matrices

- However, not all matrices have inverses; for example

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

- There does not exist a matrix such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_2$
- Such matrices are said to be *singular*
- Remark: Only square matrices have inverses; an $n \times m$ matrix $\mathbf{A}$ might, however, have a *left inverse* (satisfying $\mathbf{B}\mathbf{A} = \mathbf{I}_m$) or *right inverse* (satisfying $\mathbf{A}\mathbf{B} = \mathbf{I}_n$)

Positive definite

- A related notion is that of a "positive definite" matrix, which applies to symmetric matrices
- **Definition:** A symmetric $n \times n$ matrix $\mathbf{A}$ is said to be *positive definite* if for all $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 \qquad \text{if } \mathbf{x} \neq 0$$

- The two notions are related in the sense that if $\mathbf{A}$ is positive definite, then (a) $\mathbf{A}$ is not singular and (b) $\mathbf{A}^{-1}$ is also positive definite
- If $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$, then $\mathbf{A}$ is said to be *positive semidefinite*
- In statistics, these classifications are particularly important for variance-covariance matrices, which are always positive semidefinite (and positive definite, if they aren't singular)

Square root of a matrix

- These concepts are important with respect to knowing whether a matrix has a "square root"
- **Definition:** An $n \times n$ matrix $\mathbf{A}$ is said to have a *square root* if there exists a matrix $\mathbf{B}$ such that $\mathbf{BB} = \mathbf{A}$.
- **Theorem:** Let $\mathbf{A}$ be a positive definite matrix. Then there exists a unique matrix $\mathbf{A}^{1/2}$ such that $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$.
- Positive semidefinite matrices have square roots as well, although they aren't necessarily unique

Rank

- We also need to be familiar with the concept of matrix rank (there are many ways of defining rank; all are equivalent)
- **Definition:** The *rank* of a matrix is the dimension of its largest nonsingular submatrix.
- For example, the following 3×3 matrix is singular, but contains a nonsingular 2×2 submatrix, so its rank is 2:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & \cancel{3} \\ \cancel{2} & \cancel{4} & \cancel{6} \\ 1 & 0 & \cancel{1} \end{bmatrix}$$

- Note that a nonsingular $n \times n$ matrix has rank n, and is said to be *full rank*

Rank and multiplication

- There are many results and theorems involving rank; we're not going to cover them all, but it is important to know that rank cannot be increased through the process of multiplication
- **Theorem:** For any matrices $\mathbf{A}$ and $\mathbf{B}$ with appropriate dimensions, rank$(\mathbf{AB}) \leq$ rank$(\mathbf{A})$ and rank$(\mathbf{AB}) \leq$ rank$(\mathbf{B})$.
- In particular, rank$(\mathbf{A}^\top\mathbf{A}) =$ rank$(\mathbf{A}\mathbf{A}^\top) =$ rank$(\mathbf{A})$.

Expectation and variance

- In addition, we need some results on expected values of vectors and functions of vectors
- First of all, we need to define expectation and variance as they pertain to random vectors
- **Definition:** Let $\mathbf{x} = (X_1 \;\, X_2 \;\, \cdots X_d)^\top$ denote a vector of random variables, then $\mathbb{E}(\mathbf{x}) = (\mathbb{E}X_1 \;\, \mathbb{E}X_2 \;\, \cdots \mathbb{E}X_d)^\top$. Meanwhile, $\mathbb{V}\mathbf{x}$ is a $d \times d$ matrix:

$$\mathbb{V}\mathbf{x} = \mathbb{E}\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\} \text{ with elements}$$
$$(\mathbb{V}\mathbf{x})_{ij} = \mathbb{E}\left\{(X_i-\mu_i)(X_j-\mu_j)\right\},$$

where $\mu_i = \mathbb{E}X_i$. The matrix $\mathbb{V}\mathbf{x}$ is referred to as the *variance-covariance matrix* of $\mathbf{x}$.

Linear and quadratic forms

- Letting $\mathbf{A}$ denote a matrix of constants and $\mathbf{x}$ a random vector with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$,

$$\begin{aligned} \mathbb{E}(\mathbf{A}^\top \mathbf{x}) &= \mathbf{A}^\top \boldsymbol{\mu} \\ \mathbb{V}(\mathbf{A}^\top \mathbf{x}) &= \mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A} \\ \mathbb{E}(\mathbf{x}^\top \mathbf{A} \mathbf{x}) &= \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu} + \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}), \end{aligned}$$

  where $\operatorname{tr}(\mathbf{A}) = \sum_i A_{ii}$ is the trace of $\mathbf{A}$
- Some useful facts about traces:

$$\begin{aligned} \operatorname{tr}(\mathbf{AB}) &= \operatorname{tr}(\mathbf{BA}) \\ \operatorname{tr}(\mathbf{A}+\mathbf{B}) &= \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B}) \\ \operatorname{tr}(c\mathbf{A}) &= c \operatorname{tr}(\mathbf{A}) \\ \operatorname{tr}(\mathbf{A}) &= \operatorname{rank}(\mathbf{A}) \quad \text{if } \mathbf{AA} = \mathbf{A} \end{aligned}$$

Eigendecompositions

- Finally, we'll also take a moment to introduce some facts about eigenvalues
- The most important thing about eigenvalues is that they allow us to "diagonalize" a matrix: if $\mathbf{A}$ is a symmetric $d \times d$ matrix, then it can be factored into:

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top,$$

  where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$ of $\mathbf{A}$ and the columns of $\mathbf{Q}$ are its eigenvectors
- Furthermore, eigenvectors are orthonormal, so we have $\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$

Eigenvalues and "size"

- This is very helpful from a conceptual standpoint, as it allows us to separate the "size" of a matrix ($\boldsymbol{\Lambda}$) from its "direction(s)" ($\mathbf{Q}$)
- For example, we have already seen that one measure of the size of a matrix is based on $\lambda_{\max}$ (for a symmetric matrix, its spectral norm is its largest eigenvalue)
- In addition, the trace and determinant, two other ways of quantifying the "size" of a matrix, are simple functions of the eigenvalues:
    - $\text{tr}(\mathbf{A}) = \sum_i \lambda_i$
    - $|\mathbf{A}| = \prod_i \lambda_i$

Eigenvalues and inverses

- Once one has obtained the eigendecomposition of $\mathbf{A}$, calculating its inverse is straightforward
- If $\mathbf{A}$ is not singular, then $\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^\top$; note that since $\mathbf{\Lambda}$ is diagonal, its inverse is trivial to calculate
- Even if $\mathbf{A}$ is singular, we can obtain something called a "generalized inverse": $\mathbf{A}^- = \mathbf{Q}\mathbf{\Lambda}^-\mathbf{Q}^\top$, where $(\mathbf{\Lambda}^-)_{ii} = \lambda_i^{-1}$ if $\lambda_i \neq 0$ and $(\mathbf{\Lambda}^-)_{ii} = 0$ otherwise
- Many other important properties of matrices can be deduced entirely from their eigenvalues:
    - $\mathbf{A}$ is positive definite if and only if $\lambda_i > 0$ for all i
    - $\mathbf{A}$ is positive semidefinite if and only if $\lambda_i \geq 0$ for all i
    - If $\mathbf{A}$ has rank r, then $\mathbf{A}$ has r nonzero eigenvalues and the remaining $d - r$ eigenvalues are zero

Extreme values

- Lastly, there is a connection between a matrix's eigenvalues and the extreme values of its quadratic form
- Let the eigenvalues $\lambda_1, \ldots, \lambda_d$ of $\mathbf{A}$ be ordered from largest to smallest. Over the set of all vectors $\mathbf{x}$ such that $\|\mathbf{x}\|_2 = 1$,

$$\max \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_1$$

and

$$\min \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_d$$