Lecture 2 Part 1: Derivatives in Higher Dimensions: Jacobians and Matrix Functions

MIT 18.S096 Matrix Calculus For Machine Learning And Beyond

March 20, 2024

All definitions and theorems follow from Chapter 5 of Magnus and Neudecker (1988).

1 Vector-to-vector functions

Differentiability. Let f be a function defined on an open subset[1] $S \subseteq \mathbb{R}^n$ mapping to $\mathbb{R}^m$. Let $\boldsymbol{x}$ be an interior point of S and $B(\boldsymbol{x}, r)$ be an n-ball inside S $(r > 0)$. Let $d\boldsymbol{x} \in \mathbb{R}^n$ be such that $\boldsymbol{x} + d\boldsymbol{x} \in B(\boldsymbol{x}, r)$. If there exists a real $m \times n$ matrix A, depending on $\boldsymbol{x}$ but not $d\boldsymbol{x}$, such that

$$f(\boldsymbol{x} + d\boldsymbol{x}) = f(\boldsymbol{x}) + A(\boldsymbol{x})(d\boldsymbol{x}) + r_{\boldsymbol{x}}(d\boldsymbol{x})$$

for all $d\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{x} + d\boldsymbol{x} \in B(\boldsymbol{x}, r)$ and

$$\lim_{d\boldsymbol{x} \to 0} \frac{r_{\boldsymbol{x}}(d\boldsymbol{x})}{\|d\boldsymbol{x}\|} = 0,$$

then f is said to be *differentiable* at $\boldsymbol{x}$.

Derivative. The matrix $\boldsymbol{A}(\boldsymbol{x}) \in \mathbb{R}^{m \times n}$ is called the (first) *derivative* of f at $\boldsymbol{x}$. The *gradient* is simply the transpose of the derivative.

Magnus and Neudecker (1988) goes on to prove several important facts about differentiable functions:

- The derivative $A(\boldsymbol{x})$ is unique.
- If f is differentiable at $\boldsymbol{x}$, then f is continuous at $\boldsymbol{x}$.
- If f is differentiable at $\boldsymbol{x}$, then all partial derivatives

  $$D_j f_i(\boldsymbol{x}) \triangleq \lim_{t \to 0} \frac{f_i(\boldsymbol{x} + t e_j) - f_i(\boldsymbol{x})}{t}$$

  exist at $\boldsymbol{x}$[2]. Let $Df(\boldsymbol{x})$ be a matrix where the ij-entry is $D_j f_i(\boldsymbol{x})$, i.e., the *Jacobian*. One can show that $A(\boldsymbol{x}) = Df(\boldsymbol{x})$. This "reveals that the elements $a_{ij}(\boldsymbol{x})$ of the matrix $A(\boldsymbol{x})$ are, in fact, precisely the partial derivatives $D_j f_i(\boldsymbol{x})$".

2 A very simple example from the lecture

Too simple to be included.

1. While a lot of functions are defined on the entire $\mathbb{R}^n$, a lot of functions aren't (e.g., $f(\boldsymbol{x}) = 1 \oslash \boldsymbol{x}$).

2. The converse is not true. For the converse to be true, we must add the assumption that the partial derivatives are continuous at $\boldsymbol{x}$. This result can be found in college-level calculus textbooks. See, for example, Chapter 14 Partial Derivatives in *Calculus – Early Transcendentals* (8th edition) by James Stewart. I must admit that I didn't pay attention to this in college.

3 My first example

Function:

$$f(\boldsymbol{x}) \;=\; \boldsymbol{x} \odot \boldsymbol{x}$$

Difference:

$$\begin{aligned} f(\boldsymbol{x}+d\boldsymbol{x}) - f(\boldsymbol{x}) &= (\boldsymbol{x}+d\boldsymbol{x}) \odot (\boldsymbol{x}+d\boldsymbol{x}) - \boldsymbol{x} \odot \boldsymbol{x} \\ &= \boldsymbol{x} \odot \boldsymbol{x} + 2\boldsymbol{x} \odot (d\boldsymbol{x}) + (d\boldsymbol{x}) \odot (d\boldsymbol{x}) - \boldsymbol{x} \odot \boldsymbol{x} \\ &= 2\boldsymbol{x} \odot (d\boldsymbol{x}) + (d\boldsymbol{x}) \odot (d\boldsymbol{x}) \\ &= \mathrm{diag}(2\boldsymbol{x})(d\boldsymbol{x}) + (d\boldsymbol{x}) \odot (d\boldsymbol{x}) \end{aligned}$$

Differential:

$$df(\boldsymbol{x}; d\boldsymbol{x}) \;=\; \mathrm{diag}(2\boldsymbol{x})(d\boldsymbol{x})$$

Gradient:

$$\nabla f(\boldsymbol{x}) \;=\; \mathrm{diag}(2\boldsymbol{x})$$

Verification:

derivative of `x .* x` w.r.t. `x`

$$\frac{\partial}{\partial x}(x \odot x) = \mathbf{2} \cdot \mathbf{diag}(x)$$

4 My second example

Function:

$$f(\boldsymbol{x}) \;=\; \boldsymbol{x}\boldsymbol{x}^T\boldsymbol{x}$$

Difference:

$$\begin{aligned} & f(\boldsymbol{x}+d\boldsymbol{x}) - f(\boldsymbol{x}) \\ = \; & (\boldsymbol{x}+d\boldsymbol{x})(\boldsymbol{x}+d\boldsymbol{x})^T(\boldsymbol{x}+d\boldsymbol{x}) - \boldsymbol{x}\boldsymbol{x}^T\boldsymbol{x} \\ = \; & (\boldsymbol{x}+d\boldsymbol{x})(\boldsymbol{x}^T+(d\boldsymbol{x})^T)(\boldsymbol{x}+d\boldsymbol{x}) - \boldsymbol{x}\boldsymbol{x}^T\boldsymbol{x} \\ = \; & (\boldsymbol{x}\boldsymbol{x}^T + \boldsymbol{x}(d\boldsymbol{x})^T + (d\boldsymbol{x})\boldsymbol{x}^T + (d\boldsymbol{x})(d\boldsymbol{x})^T)(\boldsymbol{x}+d\boldsymbol{x}) - \boldsymbol{x}\boldsymbol{x}^T\boldsymbol{x} \\ = \; & \boldsymbol{x}(d\boldsymbol{x})^T\boldsymbol{x} + (d\boldsymbol{x})\boldsymbol{x}^T\boldsymbol{x} + (d\boldsymbol{x})(d\boldsymbol{x})^T\boldsymbol{x} + \boldsymbol{x}\boldsymbol{x}^T(d\boldsymbol{x}) + \boldsymbol{x}(d\boldsymbol{x})^T(d\boldsymbol{x}) + (d\boldsymbol{x})\boldsymbol{x}^T(d\boldsymbol{x}) + (d\boldsymbol{x})(d\boldsymbol{x})^T(d\boldsymbol{x}) \end{aligned}$$

Differential (the first line is actually what the product rule says):

$$\begin{aligned} df &= \boldsymbol{x}(d\boldsymbol{x})^T\boldsymbol{x} + (d\boldsymbol{x})\boldsymbol{x}^T\boldsymbol{x} + \boldsymbol{x}\boldsymbol{x}^T(d\boldsymbol{x}) \\ &= \boldsymbol{x}\boldsymbol{x}^T(d\boldsymbol{x}) + \mathbb{I}(d\boldsymbol{x})\boldsymbol{x}^T\boldsymbol{x} + \boldsymbol{x}\boldsymbol{x}^T(d\boldsymbol{x}) \\ &= (2\boldsymbol{x}\boldsymbol{x}^T + (\boldsymbol{x}^T\boldsymbol{x})\mathbb{I})(d\boldsymbol{x}) \end{aligned}$$

Gradient:

$$\nabla f \;=\; 2\boldsymbol{x}\boldsymbol{x}^T + (\boldsymbol{x}^T\boldsymbol{x})\mathbb{I}$$

Verification:

derivative of `x * x' * x` w.r.t. `x`

$$\frac{\partial}{\partial x}\left(\boldsymbol{x}\cdot\boldsymbol{x}^\top\cdot\boldsymbol{x}\right) = \boldsymbol{x}^\top\cdot\boldsymbol{x}\cdot\mathbb{I} + 2\cdot\boldsymbol{x}\cdot\boldsymbol{x}^\top$$

5 Chain rule: derivative of a composition of vector-to-vector functions

Theorem (chain rule). Let $f\colon\mathbb{R}^m\to\mathbb{R}^n$ and $g\colon\mathbb{R}^n\to\mathbb{R}^p$ be differentiable. Then the composition function $h\colon\mathbb{R}^m\to\mathbb{R}^n$ defined by $h(x)=g(f(x))$ is differentiable. Further, the Jacobian of h is

$$Dh(\boldsymbol{x}) = (Dg(\boldsymbol{y}))(Df(\boldsymbol{x})),$$

where $\boldsymbol{x}\in\mathbb{R}^m$ and $\boldsymbol{y}=f(\boldsymbol{x})\in\mathbb{R}^n$.

Practical note. Suppose we need to multiply two Jacobians together. Their shapes are $(m\times q)$ and $(q\times p)$. There would be mp dot products of length q, which involves about mpq scalar operations in total. Now consider multiplying three Jacobians together. Their shapes are $(1\times n)$, $(n\times n)$ and $(n\times n)$. This can be done in two opposite orders, each offering a different cost:

$$\underbrace{\underbrace{(1\times n),(n\times n)}_{(1,n)\text{ with cost }n^2},(n\times n)}_{(1,p)\text{ with cost }n^2} \quad\text{vs.}\quad \underbrace{(1\times n),\underbrace{(n\times n),(n\times n)}_{(n,n)\text{ with cost }n^3}}_{(1,p)\text{ with cost }n^2}.$$

Clearly, the first way is cheaper, so order does matter here.

Theorem (Cauchy's rule of invariance). (Using the setup from the theorem above)

$$dh(\boldsymbol{x};d\boldsymbol{x}) \;=\; dg(\boldsymbol{y};df(\boldsymbol{x};d\boldsymbol{x})).$$

Proof. From the theorem above, it follows that

$$\begin{aligned} dh(\boldsymbol{x};d\boldsymbol{x}) &= Dh(\boldsymbol{x})\,d\boldsymbol{x} \\ &= (Dg(\boldsymbol{y}))(Df(\boldsymbol{x}))d\boldsymbol{x} \\ &= (Dg(\boldsymbol{y}))\,\underbrace{df(\boldsymbol{x};d\boldsymbol{x})}_{d\boldsymbol{y}} \\ &= dg(\boldsymbol{y};df(\boldsymbol{x};d\boldsymbol{x})). \end{aligned}$$