Review of High-Dimensional Calculus

Ju Sun*

February 8, 2021

High-dimensional calculus is typically not covered in basic calculus course series, but it is the language of modern machine learning—we almost always express quantities of interest as vectors, matrices, or even tensors. In this set of notes, we quickly go over some basics of high-dimensional calculus that are most useful for machine learning, and especially highlight certain computational techniques that are not often taught elsewhere. Two recommended sources for picking up high-dimensional calculus are [Mun97, Col12]. [Zor15, Zor16] is another set of useful references.

1 Our notations

- scalars: x (small letters) vectors: $\boldsymbol{x}$ (bold small) matrices: $\boldsymbol{X}$ (bold capital) tensors: $\mathcal{X}$ (script capital) sets: S (capital)
- vectors are always **column vectors**, unless stated otherwise
- x_i: i-th element of $\boldsymbol{x}$ x_{ij}: (i,j)-th element of $\boldsymbol{X}$ $\boldsymbol{x}^i$: i-th row of $\boldsymbol{X}$ as a **row vector** $\boldsymbol{x}_j$: j-th column of $\boldsymbol{X}$ as a **column vector**
- $\mathbb{R}$: real numbers $\mathbb{R}_+$: positive reals $\mathbb{R}^n$: space of n-dimensional vectors $\mathbb{R}^{m\times n}$: space of $m\times n$ matrices $\mathbb{R}^{m\times n\times k}$: space of $m\times n\times k$ tensors, etc
- $[n] \doteq \{1,\dots,n\}$ (a notation often used by theoretical computer scientists)

2 Differentiability

2.1 First-order differentiability

Definition 2.1 (First-order derivative or Jacobian)**.** *Consider $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m$. f is (Fréchet) differentiable at a point $\boldsymbol{x}$ if there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{m\times n}$ such that*

$$\frac{\|f(\boldsymbol{x}+\boldsymbol{\delta}) - f(\boldsymbol{x}) - \boldsymbol{B}\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}\|_2} \to \mathbf{0} \quad \text{as} \quad \boldsymbol{\delta} \to \mathbf{0}, \tag{2.1}$$

or equivalently,

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = f(\boldsymbol{x}) + \boldsymbol{B}\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2) \quad \text{as} \quad \boldsymbol{\delta} \to \mathbf{0}. \tag{2.2}$$

Here $\boldsymbol{B}$ is called the (Fréchet) derivative, or the Jacobian of f at $\boldsymbol{x}$, denoted as $\boldsymbol{J}_f(\boldsymbol{x})$.

*Department of Computer Science and Engineering and Department of Neurosurgery, University of Minnesota at Twin Cities. Email: jusun@umn.edu.

Here, a vector-valued function $h(\boldsymbol{\delta})$ is $o(\|\boldsymbol{\delta}\|_2)$ if $\frac{\|h(\boldsymbol{\delta})\|_2}{\|\boldsymbol{\delta}\|_2} \to 0$ as $\boldsymbol{\delta} \to \mathbf{0}$.

Definition 2.2 (Gradient). *For function $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$. The gradient $\nabla f(\boldsymbol{x}) \in \mathbb{R}^{n\times 1}$, which is a column vector, is the transpose of the Jacobian $\boldsymbol{J}_f(\boldsymbol{x}) \in \mathbb{R}^{1\times n}$.*

According to this convention, for $f(\boldsymbol{X}) : \mathbb{R}^{m\times n} \to \mathbb{R}$, the gradient should be $\mathbb{R}^{(m\times n)\times 1}$, which is a length-$(m \times n)$ vector. This is *inconvenient* for many purposes. So in practice, the vector is shaped into an $m \times n$ matrix, i.e., having the same shape as $\boldsymbol{X}$. We will use this convention, i.e., $\nabla f(\boldsymbol{X}) = \left[\frac{\partial f}{\partial x_{i,j}}(\boldsymbol{X})\right]_{i,j} \in \mathbb{R}^{m\times n}$.[1]

For computation of Jacobian $\boldsymbol{J}_f(\boldsymbol{x})$ for $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m$, we have

$$\boldsymbol{J}_f(\boldsymbol{x}) = \left[\frac{\partial f_i}{\partial x_j}(\boldsymbol{x})\right]_{i,j}, \tag{2.3}$$

i.e., $\boldsymbol{J}_f(\boldsymbol{x})$ is the collection of all the first-order partial derivatives $\frac{\partial f_i}{\partial x_j}(\boldsymbol{x})$ for all i, j.

Theorem 2.3 (Sufficient condition for first-order differentiability). *If all first-order partial derivatives exist and are continuous at $\boldsymbol{x}$, then $f(\boldsymbol{x})$ is first-order differentiable at $\boldsymbol{x}$.*

But the condition is not necessary for first-order differentiability.

Theorem 2.4 (Calculus rules of Jacobian). *Assume $f, g : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable at a point $\boldsymbol{x} \in \mathbb{R}^n$.*

- ***linearity**: $\lambda_1 f + \lambda_2 g$ is differentiable at $\boldsymbol{x}$ and $\boldsymbol{J}_{\lambda_1 f+\lambda_2 g}(\boldsymbol{x}) = \lambda_1 \boldsymbol{J}_f(\boldsymbol{x}) + \lambda_2 \boldsymbol{J}_g(\boldsymbol{x})$.*
- ***product**: assume $m = 1$, fg is differentiable at $\boldsymbol{x}$ and $\nabla[fg](\boldsymbol{x}) = f(\boldsymbol{x})\nabla g(\boldsymbol{x}) + g(\boldsymbol{x})\nabla f(\boldsymbol{x})$.*
- ***quotient**: assume $m = 1$ and $g(\boldsymbol{x}) \neq 0$, $\frac{f}{g}$ is differentiable at $\boldsymbol{x}$ and $\nabla\left[\frac{f}{g}\right](\boldsymbol{x}) = \frac{g(\boldsymbol{x})\nabla f(\boldsymbol{x}) - f(\boldsymbol{x})\nabla g(\boldsymbol{x})}{g^2(\boldsymbol{x})}$.*
- ***Chain rule**: Let $f : \mathbb{R}^m \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^k$, and f is differentiable at $\boldsymbol{x}$ and $\boldsymbol{y} = f(\boldsymbol{x})$ and h is differentiable at $\boldsymbol{y}$. Then, $h \circ f : \mathbb{R}^m \to \mathbb{R}^k$ is differentiable at $\boldsymbol{x}$, and*

  $$\boldsymbol{J}_{[h\circ f]}(\boldsymbol{x}) = \boldsymbol{J}_h(f(\boldsymbol{x}))\boldsymbol{J}_f(\boldsymbol{x}). \tag{2.4}$$

  When $k = 1$,

  $$\nabla[h \circ f](\boldsymbol{x}) = \boldsymbol{J}_f^\top(\boldsymbol{x})\nabla h(f(\boldsymbol{x})). \tag{2.5}$$

2.2 Derive the Jacobian

There are two major methods. One is by invoking Eq. (2.3) and the calculus rules in Theorem 2.4. The other is by perturbation-expansion method based on Definition 2.1, as explained below.

Let's take an example $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$ for $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{b} \in \mathbb{R}^n$ and try to derive the Jacobian.

- **Elementwise calculation and calculus rules**. Obviously $\boldsymbol{J}_b(\boldsymbol{x}) = \mathbf{0}$ so we can focus on the $\boldsymbol{A}\boldsymbol{x}$ term. Now

  $$\frac{\partial(\boldsymbol{A}\boldsymbol{x})_i}{\partial x_j} = \frac{\partial(\boldsymbol{a}^i\boldsymbol{x})}{\partial x_j} = \frac{\partial \sum_k a_{ik}x_k}{\partial x_j} = a_{ij}. \tag{2.6}$$

  So $\boldsymbol{J}_{\boldsymbol{A}\boldsymbol{x}}(\boldsymbol{x}) = \boldsymbol{A}$, and by linearity $\boldsymbol{J}_f(\boldsymbol{x}) = \boldsymbol{A}$.

[1] We write $\boldsymbol{M} = [m_{ij}]_{i,j}$ to mean $\boldsymbol{M}$ is a matrix in which the (i, j)-th entry takes the form m_{ij}.

- **Perturbation-expansion method**. We make a sufficiently small perturbation $\boldsymbol{\delta}$ to $\boldsymbol{x}$, so

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = \boldsymbol{A}(\boldsymbol{x}+\boldsymbol{\delta}) - \boldsymbol{b} = (\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) + \boldsymbol{A}\boldsymbol{\delta} = f(\boldsymbol{x}) + \boldsymbol{A}\boldsymbol{\delta}. \tag{2.7}$$

  Comparing this to Eq. (2.2) in Definition 2.1, we easily obtain that $\boldsymbol{J}_f(\boldsymbol{x}) = \boldsymbol{A}$.

In the perturbation-expansion method, after the infinitesimal perturbation, we rearrange the terms to match the form of Eq. (2.2), i.e.,

$$f(\boldsymbol{x}) + \text{linear term in } \boldsymbol{\delta} + \text{lower-order term in } \|\boldsymbol{\delta}\|_2 \tag{2.8}$$

so that we can read off the Jacobian from the linear term.

Example 2.5. *Consider the least-squares objective $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2$. We will derive the Jacobian, which is the transpose of the gradient.*

- ***Chain rule***. *We can view f as composition of $g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}$ and $h(\boldsymbol{y}) = \|\boldsymbol{y}\|_2^2$ so that $f = h \circ g(\boldsymbol{x})$. From our last example, $\boldsymbol{J}_g(\boldsymbol{x}) = \boldsymbol{A}$. For h, it is easy to check that $\boldsymbol{J}_h(\boldsymbol{y}) = 2\boldsymbol{y}^\intercal$. Applying the chain rule, we obtain that*

$$\boldsymbol{J}_f(\boldsymbol{x}) = \boldsymbol{J}_h(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\boldsymbol{J}_g(\boldsymbol{x}) = 2(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^\intercal \boldsymbol{A}. \tag{2.9}$$

- ***Perturbation-expansion method***. *Making an infinitesimal perturbation $\boldsymbol{\delta}$ to $\boldsymbol{x}$, we obtain*

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{\delta}) &= \|\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{\delta}) - \boldsymbol{b}\|_2^2 && (2.10)\\ &= \|(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) + \boldsymbol{A}\boldsymbol{\delta}\|_2^2 && (2.11)\\ &= \|(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\|_2^2 + \|\boldsymbol{A}\boldsymbol{\delta}\|_2^2 + 2\left\langle \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}, \boldsymbol{A}\boldsymbol{\delta}\right\rangle. && (2.12) \end{aligned}$$

  Let us make some clarification before proceeding. We use $\langle \cdot, \cdot \rangle$ to mean inner product for vectors, i.e., for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, $\langle \boldsymbol{u}, \boldsymbol{v} \rangle \doteq \sum_i u_i v_i$. For all $p \ge 1$, the ℓ_p norm of the vector $\boldsymbol{u} \in \mathbb{R}^n$ is defined as $\|\boldsymbol{u}\|_p \doteq \left(\sum_i |u_i|^p\right)^{1/p}$. For $p = 2$, the norm is also called the Euclidean norm and it can be easily verified that $\|\boldsymbol{u}\|_2 = \sqrt{\langle \boldsymbol{u}, \boldsymbol{u} \rangle}$. So for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$,

$$\|\boldsymbol{u}+\boldsymbol{v}\|_2^2 = \langle \boldsymbol{u}+\boldsymbol{v}, \boldsymbol{u}+\boldsymbol{v} \rangle = \langle \boldsymbol{u}, \boldsymbol{u} \rangle + \langle \boldsymbol{v}, \boldsymbol{v} \rangle + 2\left\langle \boldsymbol{u}, \boldsymbol{v} \right\rangle = \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2 + 2\left\langle \boldsymbol{u}, \boldsymbol{v} \right\rangle. \tag{2.13}$$

  We have used this identity to arrive at Eq. (2.12). In Eq. (2.12), $\|(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\|_2^2 = f(\boldsymbol{x})$, and $\|\boldsymbol{A}\boldsymbol{\delta}\|_2^2 \in O\left(\|\boldsymbol{\delta}\|_2^2\right) \Longrightarrow \|\boldsymbol{A}\boldsymbol{\delta}\|_2^2 \in o(\|\boldsymbol{\delta}\|_2)$ which we do not care. The linear term is $2\left\langle \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}, \boldsymbol{A}\boldsymbol{\delta}\right\rangle$. We now invoke another identity $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\intercal \boldsymbol{v}$ to obtain that

$$2\left\langle \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}, \boldsymbol{A}\boldsymbol{\delta}\right\rangle = 2(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^\intercal \boldsymbol{A}\boldsymbol{\delta}. \tag{2.14}$$

  Comparing this with Eq. (2.2), we conclude that

$$\boldsymbol{J}_f(\boldsymbol{x}) = 2(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^\intercal \boldsymbol{A}. \tag{2.15}$$

The chain rule is fine for simple compositions. But it quickly leads to fatigue when there are many compositions. On other other hand, when the intermediate variables involve matrices, often tensors will be involved. An example is when deriving gradients for functions involving deep neural networks, e.g.,

$$f(\boldsymbol{W}) = \sum_i \|\boldsymbol{y}_i - \boldsymbol{W}_k \sigma(\boldsymbol{W}_{k-1} \dots \sigma(\boldsymbol{W}_2 \sigma(\boldsymbol{W}_1 \boldsymbol{x}_i)))\|_2^2. \tag{2.16}$$

Example 2.6 (Chain rule follows from perturbation-expansion)**.** *The chain rule in Theorem 2.4 can be easily derived from the perturbation-expansion method. Consider an infinitesimal perturbation $\boldsymbol{\delta}$ to $\boldsymbol{x}$ in $h \circ f$:*

$$h \circ f(\boldsymbol{x}+\boldsymbol{\delta}) = h(f(\boldsymbol{x}+\boldsymbol{\delta})) = h(f(\boldsymbol{x}) + \boldsymbol{J}_f(\boldsymbol{x})\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2)), \tag{2.17}$$

where we expanded $f(\boldsymbol{x}+\boldsymbol{\delta})$ as is because f is differentiable at $\boldsymbol{x}$. Now $h(f(\boldsymbol{x}) + \boldsymbol{J}_f(\boldsymbol{x})\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2))$ is h at the point $f(\boldsymbol{x})$ perturbed by the infinitesimal quantity $\boldsymbol{J}_f(\boldsymbol{x})\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2)$. Since h is differentiable at the point $f(\boldsymbol{x})$, we can invoke Eq. (2.2) again and obtain that

$$\begin{aligned} & h(f(\boldsymbol{x}) + \boldsymbol{J}_f(\boldsymbol{x})\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2)) \\ &= h(f(\boldsymbol{x})) + \boldsymbol{J}_h(f(\boldsymbol{x}))(\boldsymbol{J}_f(\boldsymbol{x})\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2)) + \underbrace{o(\|\boldsymbol{J}_f(\boldsymbol{x})\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2)\|_2)}_{o(\|\boldsymbol{\delta}\|_2)} \qquad (2.18) \\ &= h(f(\boldsymbol{x})) + \boldsymbol{J}_h(f(\boldsymbol{x}))\boldsymbol{J}_f(\boldsymbol{x})\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|_2). \qquad (2.19) \end{aligned}$$

So $h \circ f$ is differentiable at $\boldsymbol{x}$, and $\boldsymbol{J}_{h\circ f}(\boldsymbol{x}) = \boldsymbol{J}_h(f(\boldsymbol{x}))\boldsymbol{J}_f(\boldsymbol{x})$.

2.3 Second-order differentiability

It is possible to define second-order or even higher-order differentiability for general $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m$. For our purposes, it is sufficient to consider real-valued functions $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$, which we focus exclusively on here. Assume f is first-order differentiable in a small ball around $\boldsymbol{x}$.

- Write $\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x}) \doteq \left[\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)\right](\boldsymbol{x})$ provided the right side is well defined.
- **Symmetry**: If both $\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})$ and $\frac{\partial f^2}{\partial x_i \partial x_j}(\boldsymbol{x})$ exist and both are continuous at $\boldsymbol{x}$, then they are equal.
- **Hessian (matrix)**:

$$\nabla^2 f(\boldsymbol{x}) \doteq \left[\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})\right]_{j,i}.$$

  $\nabla^2 f$ is symmetric due to the symmetry property above.
- **Sufficient condition**: if all $\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})$ exist and are **continuous** near $\boldsymbol{x}$, f is 2nd-order differentiable at $\boldsymbol{x}$ (the converse is not true; we omit the precise definition of 2nd-order differentiability due to its technicality).

3 Taylor's theorems

Taylor's theorems take several forms. Here we focus on the form useful for gradient and Hessian derivation.

Theorem 3.1 (Taylor's theorem—scalar version)**.** *Consider $f(x) : \mathbb{R} \to \mathbb{R}$.*

- *If f is 1st-order differentiable at x, then*

$$f(x+\delta) = \underbrace{f(x) + \delta \nabla f(x)}_{\text{first-order Taylor expansion}} + o(|\delta|) \quad \text{as } \delta \to 0. \tag{3.1}$$

- *If f is 2nd-order differentiable at x, then*

$$f(x+\delta) = \underbrace{f(x) + \delta \nabla f(x) + \frac{1}{2}\delta^2 \nabla^2 f(x)}_{\text{second-order Taylor expansion}} + o(|\delta|^2) \quad \text{as } \delta \to 0. \tag{3.2}$$

The result can be easily generalized to real-valued vector- and matrix-variable functions.

Theorem 3.2 (Taylor's theorem—vector version)**.** *Consider $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$.*

- *If f is 1st-order differentiable at $\boldsymbol{x}$, then*

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = \underbrace{f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle}_{\text{first-order Taylor expansion}} + o(\|\boldsymbol{\delta}\|_2) \quad \text{as } \boldsymbol{\delta} \to \mathbf{0}. \tag{3.3}$$

- *If f is 2nd-order differentiable at $\boldsymbol{x}$, then*

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = \underbrace{f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle + \frac{1}{2}\left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x})\boldsymbol{\delta} \right\rangle}_{\text{second-order Taylor expansion}} + o(\|\boldsymbol{\delta}\|_2^2) \quad \text{as } \boldsymbol{\delta} \to \mathbf{0}. \tag{3.4}$$

To present the matrix version, we need to clarify the definitions of inner product and Euclidean norm for matrices, both natural generalization of those for vectors. For $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{m \times n}$,

$$\langle \boldsymbol{U}, \boldsymbol{V} \rangle = \sum_{i,j} u_{ij} v_{ij} \quad \text{and} \quad \|\boldsymbol{U}\|_F = \sqrt{\sum_{i,j} u_{ij}^2} = \sqrt{\langle \boldsymbol{U}, \boldsymbol{U} \rangle}. \tag{3.5}$$

In other words, let $\mathrm{vec}\,(\boldsymbol{U})$ be the vectorized version of $\boldsymbol{U}$ by sequentially stacking its columns into a long vector. We have

$$\langle \boldsymbol{U}, \boldsymbol{V} \rangle = \langle \mathrm{vec}\,(\boldsymbol{U}), \mathrm{vec}\,(\boldsymbol{V}) \rangle \quad \text{and} \quad \|\boldsymbol{U}\|_F = \|\mathrm{vec}\,(\boldsymbol{U})\|_2. \tag{3.6}$$

Theorem 3.3 (Taylor's theorem—matrix version)**.** *Consider $f(\boldsymbol{X}) : \mathbb{R}^{m \times n} \to \mathbb{R}$.*

– *If f is 1st-order differentiable at $\boldsymbol{X}$, then*

$$f(\boldsymbol{X}+\boldsymbol{\Delta}) = \underbrace{f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta} \rangle}_{\text{first-order Taylor expansion}} + o(\|\boldsymbol{\Delta}\|_F) \quad \text{as } \boldsymbol{\Delta} \to \mathbf{0}. \tag{3.7}$$

– *If f is 2nd-order differentiable at $\boldsymbol{X}$, then*

$$f(\boldsymbol{X}+\boldsymbol{\Delta}) = \underbrace{f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta} \rangle + \frac{1}{2}\left\langle \boldsymbol{\Delta}, \nabla^2 f(\boldsymbol{X})\boldsymbol{\Delta} \right\rangle}_{\text{second-order Taylor expansion}} + o(\|\boldsymbol{\Delta}\|_F^2) \quad \text{as } \boldsymbol{\Delta} \to \mathbf{0}. \tag{3.8}$$

Now we want to put Taylor's theorems into good use. But before that, we need another important property of Taylor expansion. In short, *Taylor expansion is unique*.

Theorem 3.4 (Asymptotic uniqueness of Taylor expansion—scalar version)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be k ($k \ge 1$ integer) times differentiable at a point x. If $P(\delta)$ is a k-th order polynomial satisfying $f(x+\delta) - P(\delta) = o(\delta^k)$ as $\delta \to 0$, then $P(\delta) = f(x) + \sum_{i=1}^{k} \frac{1}{k!} f^{(k)}(x)\delta^k$, i.e., k-th order Taylor expansion.*

Why is this useful? Typically, we calculate derivatives to obtain the Taylor expansion. This theorem enables the converse path. Suppose we somehow obtain a k-th order polynomial $P(\delta)$ satisfying $f(x+\delta) = P(\delta) + o(|\delta|)$—e.g., by the perturbation-expansion technique described above, we can compare it to the standard Taylor expansion form and read off the derivatives.

Example 3.5. *Consider $f(x) = x^3$ and let us calculate $\nabla f(x)$ and $\nabla^2 f(x)$. For any infinitesimal perturbation δ,*

$$f(x+\delta) = (x+\delta)^3 = x^3 + 3x^2\delta + 3x\delta^2 + \delta^3. \tag{3.9}$$

First-order term in δ is $3x^2\delta$ and so $\nabla f(x) = 3x^2$. Second-order term in δ is $3x\delta^2$ and so $\nabla^2 f(x) = 6x$.

The uniqueness property also holds for the vector and matrix versions.

Theorem 3.6 (Asymptotic uniqueness of Taylor expansion—vector version)**.** *Consider $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$.*

- *Assume $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is 1st-order differentiable at $\boldsymbol{x}$. If $P(\boldsymbol{\delta}) \doteq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{\delta} \rangle$ satisties that*

$$f(\boldsymbol{x}+\boldsymbol{\delta}) - P(\boldsymbol{\delta}) = o(\|\boldsymbol{\delta}\|_2) \quad \text{as } \boldsymbol{\delta} \to \boldsymbol{0},$$

  then $P(\boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle$, i.e., the 1st-order Taylor expansion, and $\nabla f(\boldsymbol{x}) = \boldsymbol{v}$.

- *Assume $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is 2nd-order differentiable at $\boldsymbol{x}$. If $P(\boldsymbol{\delta}) \doteq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{\delta} \rangle + \frac{1}{2} \langle \boldsymbol{\delta}, \boldsymbol{H}\boldsymbol{\delta} \rangle$ with $\boldsymbol{H}$ symmetric satisfies that*

$$f(\boldsymbol{x}+\boldsymbol{\delta}) - P(\boldsymbol{\delta}) = o(\|\boldsymbol{\delta}\|_2^2) \quad \text{as } \boldsymbol{\delta} \to \boldsymbol{0},$$

  then $P(\boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle + \frac{1}{2} \langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{\delta} \rangle$, i.e., the 2nd-order Taylor expansion, and $\nabla f(\boldsymbol{x}) = \boldsymbol{v}$, $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{H}$.

The matrix version, as well as proofs of the asymptotic uniqueness properties and other forms of Taylor's theorems can be found in Chapter 5 of [Col12].

Now we provide a couple of examples on how the perturbation-expansion technique can help us to move from Taylor expansion to derivatives.

Example 3.7. *Let's consider $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$ again and try to derive $\nabla f(\boldsymbol{x})$ and $\nabla^2 f(\boldsymbol{x})$. From Example 2.5, we know that*

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \underbrace{2\langle \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\delta} \rangle}_{\text{1st-order in } \boldsymbol{\delta}} + \underbrace{\|\boldsymbol{A}\boldsymbol{\delta}\|_2^2}_{\text{2nd-order in } \boldsymbol{\delta}}. \tag{3.10}$$

To read off the gradient, we need to rearrange the 1-st order term into the form $\langle \clubsuit, \boldsymbol{\delta} \rangle$ for some $\clubsuit$. Now we need a useful rule for manipulating vector/matrix inner products.

> *Any leading matrix can be transposed and moved to leading position of the other side of the inner product; similarly, any trailing matrix can be transposed and moved to the trailing position of the other side of the inner product. For example, consider matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ with compatible dimensions so that $\langle \boldsymbol{A}\boldsymbol{B}, \boldsymbol{C}\boldsymbol{D} \rangle$ is well defined. Then*
>
> $$\langle \boldsymbol{A}\boldsymbol{B}, \boldsymbol{C}\boldsymbol{D} \rangle = \langle \boldsymbol{B}, \boldsymbol{A}^\intercal \boldsymbol{C}\boldsymbol{D} \rangle = \langle \boldsymbol{C}^\intercal \boldsymbol{A}\boldsymbol{B}, \boldsymbol{D} \rangle = \langle \boldsymbol{A}, \boldsymbol{C}\boldsymbol{D}\boldsymbol{B}^\intercal \rangle = \langle \boldsymbol{A}\boldsymbol{B}\boldsymbol{D}^\intercal, \boldsymbol{C} \rangle. \tag{3.11}$$
>
> *This property can be derived from the cyclic property of matrix traces, as we illustrate in Homework Set 1.*

So we can rearrange the 1st *order term as*

$$2\langle \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}, \boldsymbol{A}\boldsymbol{\delta}\rangle = \langle 2\boldsymbol{A}^\intercal(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}), \boldsymbol{\delta}\rangle, \tag{3.12}$$

implying that $\nabla f(\boldsymbol{x}) = 2\boldsymbol{A}^\intercal(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})$. *For the* 2nd *order term,*

$$\|\boldsymbol{A}\boldsymbol{\delta}\|_2^2 = \langle \boldsymbol{A}\boldsymbol{\delta}, \boldsymbol{A}\boldsymbol{\delta}\rangle = \langle \boldsymbol{\delta}, \boldsymbol{A}^\intercal\boldsymbol{A}\boldsymbol{\delta}\rangle, \tag{3.13}$$

which is to be compared to $\frac{1}{2}\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x})\boldsymbol{\delta}\rangle$, *implying that* $\nabla^2 f(\boldsymbol{x}) = 2\boldsymbol{A}^\intercal\boldsymbol{A}$.

Example 3.8. *We now consider a matrix-variable problem with two blocks of variables*

$$f(\boldsymbol{W}_1, \boldsymbol{W}_2) = \sum_i \|\boldsymbol{y}_i - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{x}_i\|_2^2 = \|\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}\|_F^2 \tag{3.14}$$

and try to derive the gradient. This is an objective corresponding to a two-layer linear neural network. Making infinitesimal perturbation to $\boldsymbol{W}_1, \boldsymbol{W}_2$, *we obtain*

$$f(\boldsymbol{W}_1+\boldsymbol{\Delta}_1, \boldsymbol{W}_2+\boldsymbol{\Delta}_2) = \|\boldsymbol{Y} - (\boldsymbol{W}_2+\boldsymbol{\Delta}_2)(\boldsymbol{W}_1+\boldsymbol{\Delta}_1)\boldsymbol{X}\|_F^2 \tag{3.15}$$

$$= \|(\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}) - \boldsymbol{W}_2\boldsymbol{\Delta}_1\boldsymbol{X} - \boldsymbol{\Delta}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{\Delta}_1\boldsymbol{\Delta}_2\boldsymbol{X}\|_F^2. \tag{3.16}$$

Now we need the identity: for matrices $\boldsymbol{U}, \boldsymbol{V}$ *of the same size,* $\|\boldsymbol{U}+\boldsymbol{V}\|_F^2 = \|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2 + 2\langle \boldsymbol{U}, \boldsymbol{V}\rangle$, *which can be derived similarly to the vector version. First,*

$$\|(\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}) - \boldsymbol{W}_2\boldsymbol{\Delta}_1\boldsymbol{X} - \boldsymbol{\Delta}_2\boldsymbol{W}_1\boldsymbol{X} - \boldsymbol{\Delta}_1\boldsymbol{\Delta}_2\boldsymbol{X}\|_F^2 \tag{3.17}$$

$$= \|(\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}) - \boldsymbol{W}_2\boldsymbol{\Delta}_1\boldsymbol{X} - \boldsymbol{\Delta}_2\boldsymbol{W}_1\boldsymbol{X}\|_F^2 + \underbrace{\|\boldsymbol{\Delta}_1\boldsymbol{\Delta}_2\boldsymbol{X}\|_F^2}_{o(\|\boldsymbol{\Delta}\|_F)}$$

$$\underbrace{-2\langle(\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}) - \boldsymbol{W}_2\boldsymbol{\Delta}_1\boldsymbol{X} - \boldsymbol{\Delta}_2\boldsymbol{W}_1\boldsymbol{X}, \boldsymbol{\Delta}_1\boldsymbol{\Delta}_2\boldsymbol{X}\rangle}_{o(\|\boldsymbol{\Delta}\|_F)}, \tag{3.18}$$

where we use $o(\|\boldsymbol{\Delta}\|_F)$ *to mean* $\min\left(o(\|\boldsymbol{\Delta}_1\|_F), o(\|\boldsymbol{\Delta}_2\|_F)\right)$. *So we only need to focus on*

$$\|(\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}) - \boldsymbol{W}_2\boldsymbol{\Delta}_1\boldsymbol{X} - \boldsymbol{\Delta}_2\boldsymbol{W}_1\boldsymbol{X}\|_F^2 = \underbrace{\|\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}\|_F^2}_{f(\boldsymbol{W}_1,\boldsymbol{W}_2)} + \underbrace{\|\boldsymbol{W}_2\boldsymbol{\Delta}_1\boldsymbol{X} + \boldsymbol{\Delta}_2\boldsymbol{W}_1\boldsymbol{X}\|_F^2}_{o(\|\boldsymbol{\Delta}\|_F)}$$

$$- 2\langle \boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}, \boldsymbol{W}_2\boldsymbol{\Delta}_1\boldsymbol{X} + \boldsymbol{\Delta}_2\boldsymbol{W}_1\boldsymbol{X}\rangle. \tag{3.19}$$

We now only need to compare the linear term $-2\langle \boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}, \boldsymbol{W}_2\boldsymbol{\Delta}_1\boldsymbol{X} + \boldsymbol{\Delta}_2\boldsymbol{W}_1\boldsymbol{X}\rangle$ *with*

$$\left\langle \begin{bmatrix} \nabla_{\boldsymbol{W}_1} f \\ \nabla_{\boldsymbol{W}_2} f \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Delta}_1 \\ \boldsymbol{\Delta}_2 \end{bmatrix} \right\rangle = \langle \nabla_{\boldsymbol{W}_1} f, \boldsymbol{\Delta}_1\rangle + \langle \nabla_{\boldsymbol{W}_2} f, \boldsymbol{\Delta}_2\rangle. \tag{3.20}$$

We have that

$$-2\langle \boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}, \boldsymbol{W}_2\boldsymbol{\Delta}_1\boldsymbol{X}\rangle = -2\langle \boldsymbol{W}_2^\intercal(\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X})\boldsymbol{X}^\intercal, \boldsymbol{\Delta}_1\rangle, \tag{3.21}$$

$$-2\langle \boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X}, \boldsymbol{\Delta}_2\boldsymbol{W}_1\boldsymbol{X}\rangle = -2\langle (\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X})\boldsymbol{X}^\intercal\boldsymbol{W}_1^\intercal, \boldsymbol{\Delta}_2\rangle, \tag{3.22}$$

implying that

$$\nabla_{\boldsymbol{W}_1} f = -2\boldsymbol{W}_2^\intercal(\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X})\boldsymbol{X}^\intercal, \tag{3.23}$$

$$\nabla_{\boldsymbol{W}_2} f = -2(\boldsymbol{Y} - \boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{X})\boldsymbol{X}^\intercal\boldsymbol{W}_1^\intercal. \tag{3.24}$$

Note that in this example, due to Eq. (3.20), one can also perturb $\boldsymbol{W}_1$ only (i.e., $\boldsymbol{\Delta}_2 = \mathbf{0}$) to obtain $\nabla_{\boldsymbol{W}_1} f$ and similarly perturb $\boldsymbol{W}_2$ only to obtain $\nabla_{\boldsymbol{W}_2} f$. This tends to make the process less messy. Similarly, for functions with multiple blocks, one can take turns to perturb one block each time to derive block-wise gradients (this does NOT work for higher-order derivatives!).

Final words on this: we have discussed two or three techniques for deriving derivatives. For practical problems, it is often that a mixture of these techniques works the best. So stay flexible!

4 Directional derivatives and curvatures

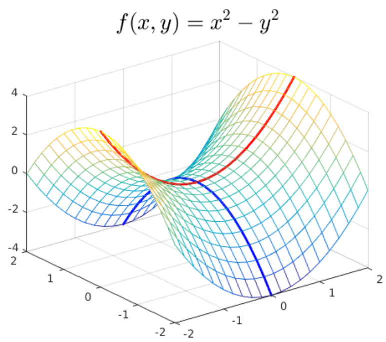

Figure 1: Blue: negative curvature (bending down); Red: positive curvature (bending up)

Consider $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$.

- **directional derivative**: $D_{\boldsymbol{v}} f(\boldsymbol{x}) \doteq \frac{d}{dt} f(\boldsymbol{x} + t\boldsymbol{v})\Big|_{t=0}$, i.e., rate of change at $\boldsymbol{x}$ along $\boldsymbol{v}$
- When f is 1-st order differentiable at $\boldsymbol{x}$,

$$D_{\boldsymbol{v}} f(\boldsymbol{x}) = \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle .$$

- Now $D_{\boldsymbol{v}} f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is another function. What is $D_{\boldsymbol{u}}(D_{\boldsymbol{v}} f)(\boldsymbol{x})$? If f is 2nd-order differentiable at $\boldsymbol{x}$,

$$D_{\boldsymbol{u}}(D_{\boldsymbol{v}} f)(\boldsymbol{x}) = \left\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{v} \right\rangle .$$

When $\boldsymbol{u} = \boldsymbol{v}$,

$$D_{\boldsymbol{u}}(D_{\boldsymbol{u}} f)(\boldsymbol{x}) = \left\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{u} \right\rangle = \frac{d^2}{dt^2} f(\boldsymbol{x} + t\boldsymbol{u})\Big|_{t=0} ,$$

which is the **directional curvature** along $\boldsymbol{u}$ and grows quadratically with respect to $\|\boldsymbol{u}\|_2$. To make it independent of the norm $\|\boldsymbol{u}\|_2$, one can consider $\frac{\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{u} \rangle}{\|\boldsymbol{u}\|_2^2}$.

Obviously, the spectral property (i.e., distribution of eigenvalues and eigenvectors) of $\nabla^2 f(\boldsymbol{x})$ determines directional curvatures. Particularly, eigenvector directions corresponding to negative (positive) eigenvalues of $\nabla^2 f(\boldsymbol{x})$ have negative (positive) curvatures.

Further reading

Chapters 3 & 5 of [DFO20] are particularly relevant and you are encouraged to go over the materials there.

Disclaimer

This set of notes is preliminary and has not been thoroughly proofread. Typos and factual errors are well expected and hence use it with caution. Bug reports are very welcome and should be sent to Prof. Ju Sun via `jusun@umn.edu`.

References

[Col12] Rodney Coleman, *Calculus on normed vector spaces*, Springer New York, 2012.

[DFO20] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong, *Mathematics for machine learning*, Cambridge University Press, 2020.

[Mun97] James R. Munkres, *Analysis on manifolds*, Taylor & Francis Inc, 1997.

[Zor15] Vladimir A. Zorich, *Mathematical analysis I*, Springer Berlin Heidelberg, 2015.

[Zor16] ______, *Mathematical analysis II*, Springer Berlin Heidelberg, 2016.