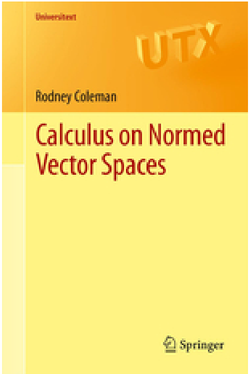
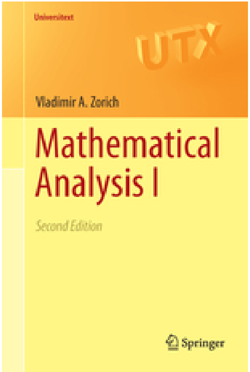
Review of High-Dimensional Calculus

Ju Sun
Computer Science & Engineering
University of Minnesota, Twin Cities

September 13, 2022

Recommended references





[Munkres, 1997, Coleman, 2012, Zorich, 2015]

Our notation

- scalars: x, vectors: $\boldsymbol{x}$, matrices: $\boldsymbol{X}$, tensors: $\mathcal{X}$, sets: S
- 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
- $[n] \doteq \{1, \dots, n\}$

Differentiability — first order

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

- Definition: **First-order differentiable** at a point $\boldsymbol{x}$ if there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ such that

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

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

- $\boldsymbol{B}$ is called the (Fréchet) derivative. When $m = 1$, $\boldsymbol{b}^\intercal$ (i.e., $\boldsymbol{B}^\intercal$) called **gradient**, denoted as $\nabla f(\boldsymbol{x})$. For general m, also called **Jacobian** matrix, denoted as $\boldsymbol{J}_f(\boldsymbol{x})$.
- Calculation: $b_{ij} = \frac{\partial f_i}{\partial x_j}(\boldsymbol{x})$
- **Sufficient condition**: if all partial derivatives exist and are **continuous** at $\boldsymbol{x}$, then $f(\boldsymbol{x})$ is differentiable at $\boldsymbol{x}$.

Calculus rules

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 $\nabla \left[\lambda_1 f + \lambda_2 g\right](\boldsymbol{x}) = \lambda_1 \nabla f(\boldsymbol{x}) + \lambda_2 \nabla g(\boldsymbol{x})$
- **product**: assume $m = 1$, fg is differentiable at $\boldsymbol{x}$ and $\nabla \left[fg\right](\boldsymbol{x}) = f(\boldsymbol{x}) \nabla g(\boldsymbol{x}) + g(\boldsymbol{x}) \nabla f(\boldsymbol{x})$
- **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

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

  When $k = 1$,

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

Put the definition in good use!

First-order differentiable at a point $\boldsymbol{x}$ if there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$, called Jacobian, such that

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

- prove the chain rule for $h \circ f\left(\boldsymbol{x}\right)$ (whiteboard)
- derive Jacobian (white board)
  - $f\left(\boldsymbol{x}\right) = \boldsymbol{A}\boldsymbol{x}$
  - $g\left(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3\right) = \boldsymbol{y} - \boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{W}_3\boldsymbol{x}$

Differentiability — second order

Consider $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ and assume f is 1st-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 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)**:

$$\nabla^2 f(\boldsymbol{x}) \doteq \left[\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})\right]_{j,i}, \tag{1}$$

  where $\left[\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})\right]_{j,i} \in \mathbb{R}^{n \times n}$ has its (j,i)-th element as $\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})$.
- $\nabla^2 f$ is symmetric.
- **Sufficient condition**: if all $\frac{\partial f^2}{\partial x_j \partial x_i}(\boldsymbol{x})$ exist and are **continuous**, f is 2nd-order differentiable at $\boldsymbol{x}$ (not converse; we omit the definition due to its technicality).

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

$$f(\boldsymbol{x} + \boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle + o(\|\boldsymbol{\delta}\|_2) \text{ as } \boldsymbol{\delta} \to \mathbf{0}.$$

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

$$f(\boldsymbol{x} + \boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle + \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{\delta} \right\rangle + o(\|\boldsymbol{\delta}\|_2^2) \text{ as } \boldsymbol{\delta} \to \mathbf{0}.$$

Matrix version: consider $f(\boldsymbol{X}) : \mathbb{R}^{m \times n} \to \mathbb{R}$

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

$$f(\boldsymbol{X} + \boldsymbol{\Delta}) = f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta} \rangle + o(\|\boldsymbol{\Delta}\|_F) \text{ as } \boldsymbol{\Delta} \to \mathbf{0}.$$

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

$$f(\boldsymbol{X} + \boldsymbol{\Delta}) = f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{\Delta} \rangle + \frac{1}{2} \left\langle \boldsymbol{\Delta}, \nabla^2 f(\boldsymbol{X}) \boldsymbol{\Delta} \right\rangle + o(\|\boldsymbol{\Delta}\|_F^2)$$

$$\text{as } \boldsymbol{\Delta} \to \mathbf{0}.$$

Put Taylor in good use!

- derive gradient and Hessian for $f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ (whiteboard)

- derive gradient (and Hessian) for

$$g(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3) = \|\boldsymbol{y} - \boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{W}_3\boldsymbol{x}\|_F^2$$

  (whiteboard)

before: **gradient, Hessian** $\Longrightarrow$ **Taylor expansion**
now: **Taylor expansion** $\Longrightarrow$ **gradient, Hessian**

But why?

Taylor approximation — asymptotic uniqueness

Let $f : \mathbb{R} \to \mathbb{R}$ be k ($k \geq 1$ integer) times differentiable at a point x. If $P(\delta)$ is a k-th order polynomial satisfying $f\left(x+\delta\right) - P\left(\delta\right) = o(\delta^k)$ as $\delta \to 0$, then $P\left(\delta\right) = P_k(\delta) \doteq f(x) + \sum_{i=1}^{k} \frac{1}{k!} f^{(k)}\left(x\right)\delta^k$.

Generalization to the vector version

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

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

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

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

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

then $P\left(\boldsymbol{\delta}\right) = f\left(\boldsymbol{x}\right) + \langle \nabla f\left(\boldsymbol{x}\right), \boldsymbol{\delta} \rangle + \frac{1}{2}\left\langle \boldsymbol{\delta}, \nabla^2 f\left(\boldsymbol{x}\right)\boldsymbol{\delta} \right\rangle$, i.e., the 2nd-order Taylor expansion. We can read off ∇f and $\nabla^2 f$ if we know the expansion!

Similarly for the matrix version. See Chap 5 of [Coleman, 2012] for other forms of Taylor theorems and proofs of the asymptotic uniqueness.

Directional derivatives and curvatures

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

- **directional derivative**: $D_{\boldsymbol{v}} f(\boldsymbol{x}) \doteq \frac{d}{dt} f(\boldsymbol{x} + t\boldsymbol{v})$
- 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 $D_{\boldsymbol{v}} f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$, what is $D_{\boldsymbol{u}} (D_{\boldsymbol{v}} f)(\boldsymbol{x})$?

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

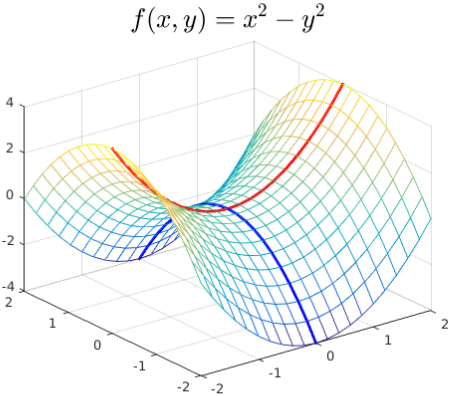
- When $\boldsymbol{u} = \boldsymbol{v}$,

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

- $\frac{\left\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{u} \right\rangle}{\|\boldsymbol{u}\|_2^2}$ is the **directional curvature** along $\boldsymbol{u}$ independent of the norm of $\boldsymbol{u}$

Directional curvature

$\frac{\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{u} \rangle}{\|\boldsymbol{u}\|_2^2}$ is the **directional curvature** along $\boldsymbol{u}$ independent of the norm of $\boldsymbol{u}$



Blue: negative curvature (bending down)

Red: positive curvature (bending up)

References i

[Coleman, 2012] Coleman, R. (2012). **Calculus on Normed Vector Spaces.** Springer New York.

[Munkres, 1997] Munkres, J. R. (1997). **Analysis On Manifolds.** Taylor & Francis Inc.

[Zorich, 2015] Zorich, V. A. (2015). **Mathematical Analysis I.** Springer Berlin Heidelberg.