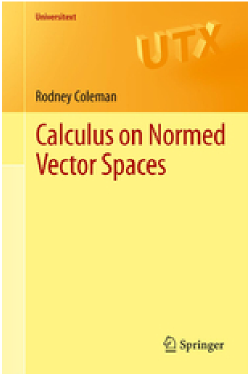
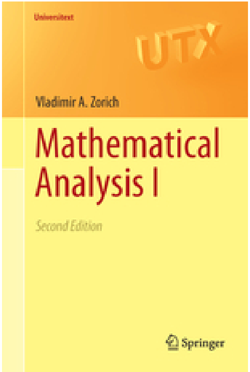
Review of Multivariate Calculus

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

September 21, 2020

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, $h \circ f : \mathbb{R}^n \to \mathbb{R}^k$ 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 $k = 1$**,**

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

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

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

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

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

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

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

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

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

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

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

Taylor approximation — asymptotic uniqueness

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 $P(\delta) = P_k(\delta) \doteq f(x) + \sum_{i=1}^{k} \frac{1}{k!} f^{(k)}(x)\, \delta^k$.

Generalization to the vector version

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

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

  then $P(\boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle$, i.e., the 1st-order Taylor expansion.
- Assume $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is 2-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 satisties that

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

  then $P(\boldsymbol{\delta}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\delta} \rangle + \frac{1}{2} \left\langle \boldsymbol{\delta}, \nabla^2 f(\boldsymbol{x})\, \boldsymbol{\delta} \right\rangle$, 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.

Asymptotic uniqueness — why interesting?

Two ways of deriving gradients and Hessians (Recall HW0!)

(a) Derive the gradient and Hessian of the linear least-squares function $f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$. Please include your calculation details.

(b) Let $\sigma = \frac{1}{1+e^{-x}}$, i.e., the *logistic function*. Derive the gradient of the matrix-variable function $g(\boldsymbol{W}) = \|\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})\|_2^2$, where σ is applied to the vector $\boldsymbol{W}\boldsymbol{x}$ elementwise. This is regression based on a simplified one-neuron network. Please include your calculation details.

(a) Consider the least-squares objective $f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ again. Recall that for any two vectors $\boldsymbol{a}, \boldsymbol{b}$, $\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 - 2\boldsymbol{a}^\intercal \boldsymbol{b} + \|\boldsymbol{b}\|_2^2$. Now $f(\boldsymbol{x} + \boldsymbol{\delta}) = \|(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) - \boldsymbol{A}\boldsymbol{\delta}\|_2^2$. Expand this square by the previous formula, and compare it to the 2nd order Taylor expansion by plugging your results from **Problem 1(a)**. Are they equal or not? Why? (Hint: You may find this fact useful: for any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ and any matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, $\langle \boldsymbol{u}, \boldsymbol{M}\boldsymbol{v}\rangle = \langle \boldsymbol{M}^\intercal \boldsymbol{u}, \boldsymbol{v}\rangle$. This can be derived from the trace cyclic property above.)

(b) Consider the one-neuron network regression again: $g(\boldsymbol{W}) = \|\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})\|_2^2$ with $\sigma = \frac{1}{1+e^{-x}}$, i.e., the *logistic function*. Let's try to work out its 1st order Taylor expansion by direct expansion as follows.

- Show that $\sigma((\boldsymbol{W} + \boldsymbol{\Delta})\boldsymbol{x}) = \sigma(\boldsymbol{W}\boldsymbol{x}) + \sigma'(\boldsymbol{W}\boldsymbol{x}) \odot (\boldsymbol{\Delta}\boldsymbol{x}) + o(\|\boldsymbol{\Delta}\|_F)$ when $\boldsymbol{\Delta} \to \boldsymbol{0}$. Here, both σ and σ' are applied elementwise, and $\odot$ denotes the elementwise (Hadamard) product.
- So $\boldsymbol{y} - \sigma((\boldsymbol{W} + \boldsymbol{\Delta})\boldsymbol{x}) = (\boldsymbol{y} - \sigma(\boldsymbol{W}\boldsymbol{x})) - \sigma'(\boldsymbol{W}\boldsymbol{x}) \odot (\boldsymbol{\Delta}\boldsymbol{x}) - o(\|\boldsymbol{\Delta}\|_F)$ when $\boldsymbol{\Delta} \to \boldsymbol{0}$. Substitute this back into the square and use the identity $\|\boldsymbol{a} + \boldsymbol{b} + \boldsymbol{c}\|_2^2 = \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2 + \|\boldsymbol{c}\|_2^2 + 2\boldsymbol{a}^\intercal \boldsymbol{b} + 2\boldsymbol{a}^\intercal \boldsymbol{c} + 2\boldsymbol{b}^\intercal \boldsymbol{c}$ to obtain the first-order approximation to $g(\boldsymbol{W} + \boldsymbol{\Delta})$. Remember that any terms lower order than $\|\boldsymbol{\Delta}\|_F$ are not interesting and we can always assume $\boldsymbol{\Delta}$ as small as needed.
- Substitute the result from **Problem 1(b)** into the 1st order Taylor expansion formula above and compare it to the result obtained here. Are they equal or not?

Asymptotic uniqueness — why interesting?

Think of neural networks with identity activation functions

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

How to derive the gradient?

- Scalar chain rule?
- Vector chain rule?
- First-order Taylor expansion

Why interesting? See e.g.,
[Kawaguchi, 2016, Lampinen and Ganguli, 2018]

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

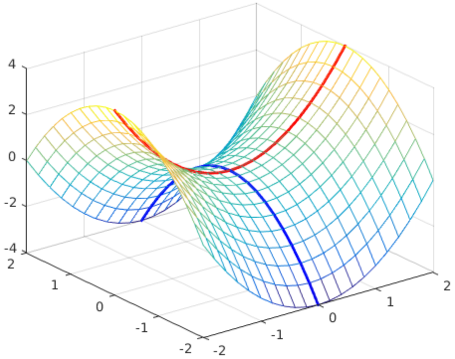
$$D_{\boldsymbol{u}} (D_{\boldsymbol{u}} f)(\boldsymbol{x}) = \left\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{u} \right\rangle = \frac{d^2}{dt^2} f(\boldsymbol{x} + t\boldsymbol{u}) .$$

- $\frac{\left\langle \boldsymbol{u}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{u} \right\rangle}{\|\boldsymbol{u}\|_2^2}$ 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 $\boldsymbol{u}$

$$f(x, y) = x^2 - y^2$$



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.

[Kawaguchi, 2016] Kawaguchi, K. (2016). **Deep learning without poor local minima.** *arXiv:1605.07110*.

[Lampinen and Ganguli, 2018] Lampinen, A. K. and Ganguli, S. (2018). **An analytic theory of generalization dynamics and transfer learning in deep linear networks.** *arXiv:1809.10374*.

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

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