Review of Multivariate Calculus

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

September 21, 2020
Recommended references

Our notation

- scalars: $x$, vectors: $x$, matrices: $X$, tensors: $\mathcal{X}$, sets: $S$
- vectors are always **column vectors**, unless stated otherwise
- $x_i$: $i$-th element of $x$, $x_{ij}$: $(i, j)$-th element of $X$, $x^i$: $i$-th row of $X$ as a row vector, $x_j$: $j$-th column of $X$ 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
- $[n] \doteq \{1, \ldots, n\}$
Consider \( f(x) : \mathbb{R}^n \to \mathbb{R}^m \)

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

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

i.e.,

\[
f(x + \delta) = f(x) + B \delta + o(\|\delta\|_2) \quad \text{as} \quad \delta \to 0.
\]

- \( B \) is called the (Fréchet) derivative. When \( m = 1 \), \( b^\top \) (i.e., \( B^\top \)) called **gradient**, denoted as \( \nabla f(x) \). For general \( m \), also called **Jacobian** matrix, denoted as \( J_f(x) \).

- **Calculation:** \( b_{ij} = \frac{\partial f_i}{\partial x_j}(x) \)

- **Sufficient condition:** if all partial derivatives exist and are **continuous** at \( x \), then \( f(x) \) is differentiable at \( x \).
Calculus rules

Assume $f, g : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable at a point $x \in \mathbb{R}^n$.

- **linearity**: $\lambda_1 f + \lambda_2 g$ is differentiable at $x$ and
  \[ \nabla [\lambda_1 f + \lambda_2 g] (x) = \lambda_1 \nabla f (x) + \lambda_2 \nabla g (x) \]

- **product**: assume $m = 1$, $fg$ is differentiable at $x$ and
  \[ \nabla [fg] (x) = f (x) \nabla g (x) + g (x) \nabla f (x) \]

- **quotient**: assume $m = 1$ and $g(x) \neq 0$, $\frac{f}{g}$ is differentiable at $x$ and
  \[ \nabla \left[ \frac{f}{g} \right] (x) = \frac{g(x) \nabla f(x) - f(x) \nabla g(x)}{g^2(x)} \]

- **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 $x$ and $y = f(x)$ and $h$ is differentiable at $y$. Then, $h \circ f : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable at $x$, and
  \[ J_{[h \circ f]} (x) = J_h (f (x)) J_f (x) . \]

When $k = 1$,

\[ \nabla [h \circ f] (x) = J_f^{\top} (x) \nabla h (f (x)) . \]
Consider $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ and assume $f$ is 1st-order differentiable in a small ball around $\mathbf{x}$

- Write $\frac{\partial f^2}{\partial x_j \partial x_i}(\mathbf{x}) = \left[ \frac{\partial}{\partial x_j} \left( \frac{\partial f}{\partial x_i} \right) \right](\mathbf{x})$ provided the right side well defined

- **Symmetry**: If both $\frac{\partial f^2}{\partial x_j \partial x_i}(\mathbf{x})$ and $\frac{\partial f^2}{\partial x_i \partial x_j}(\mathbf{x})$ exist and both are continuous at $\mathbf{x}$, then they are equal.

- **Hessian (matrix)**:

\[
\nabla^2 f(\mathbf{x}) = \left[ \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}) \right]_{j,i},
\]

where $\left[ \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}) \right]_{j,i} \in \mathbb{R}^{n \times n}$ has its $(j, i)$-th element as $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x})$.

- $\nabla^2 f$ is symmetric.

- **Sufficient condition**: if all $\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x})$ exist and are **continuous**, $f$ is 2nd-order differentiable at $\mathbf{x}$ (not converse; we omit the definition due to its technicality).
Taylor’s theorem

**Vector version:** consider $f(x) : \mathbb{R}^n \to \mathbb{R}$

- If $f$ is 1st-order differentiable at $x$, then
  \[
  f(x + \delta) = f(x) + \langle \nabla f(x), \delta \rangle + o(\|\delta\|_2) \quad \text{as} \quad \delta \to 0.
  \]
- If $f$ is 2nd-order differentiable at $x$, then
  \[
  f(x + \delta) = f(x) + \langle \nabla f(x), \delta \rangle + \frac{1}{2} \langle \delta, \nabla^2 f(x) \delta \rangle + o(\|\delta\|_2^2) \quad \text{as} \quad \delta \to 0.
  \]

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

- If $f$ is 1st-order differentiable at $X$, then
  \[
  f(X + \Delta) = f(X) + \langle \nabla f(X), \Delta \rangle + o(\|\Delta\|_F) \quad \text{as} \quad \Delta \to 0.
  \]
- If $f$ is 2nd-order differentiable at $X$, then
  \[
  f(X + \Delta) = f(X) + \langle \nabla f(X), \Delta \rangle + \frac{1}{2} \langle \Delta, \nabla^2 f(X) \Delta \rangle + o(\|\Delta\|_F^2) \quad \text{as} \quad \Delta \to 0.
  \]
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 $f(x + \delta) - P(\delta) = o(\delta^k)$ as $\delta \to 0$, then $P(\delta) = P_k(\delta) = f(x) + \sum_{i=1}^{k} \frac{1}{i!} f^{(i)}(x) \delta^i$.

Generalization to the vector version

- Assume $f(x) : \mathbb{R}^n \to \mathbb{R}$ is 1-order differentiable at $x$. If $P(\delta) = f(x) + \langle \nu, \delta \rangle$ satisfies that $f(x + \delta) - P(\delta) = o(\|\delta\|^2_2)$ as $\delta \to 0$, then $P(\delta) = f(x) + \langle \nabla f(x), \delta \rangle$, i.e., the 1st-order Taylor expansion.

- Assume $f(x) : \mathbb{R}^n \to \mathbb{R}$ is 2-order differentiable at $x$. If $P(\delta) = f(x) + \langle \nu, \delta \rangle + \frac{1}{2} \langle \delta, H \delta \rangle$ with $H$ symmetric satisfies that $f(x + \delta) - P(\delta) = o(\|\delta\|^2_2)$ as $\delta \to 0$, then $P(\delta) = f(x) + \langle \nabla f(x), \delta \rangle + \frac{1}{2} \langle \delta, \nabla^2 f(x) \delta \rangle$, i.e., the 2nd-order Taylor expansion. We can read off $\nabla f$ 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 \( f(x) = \|y - Ax\|_2^2 \).
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
\( g(W) = \|y - \sigma(Wx)\|_2^2 \), where \( \sigma \) is applied to the vector \( Wx \) elementwise. This is regression based on a simplified one-neuron network. Please include your calculation details.

(a) Consider the least-squares objective \( f(x) = \|y - Ax\|_2^2 \) again. Recall that for any two vectors \( a, b, \|a - b\|_2^2 = \|a\|_2^2 - 2a^Tb + \|b\|_2^2 \). Now \( f(x + \delta) = \|(y - Ax) - A\delta\|_2^2 \). 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 \( u, v \in \mathbb{R}^n \) and any matrix \( M \in \mathbb{R}^{n \times n} \), \( \langle u, Mv \rangle = \langle M^Tu, v \rangle \). This can be derived from the trace cyclic property above.)

(b) Consider the one-neuron network regression again: \( g(W) = \|y - \sigma(Wx)\|_2^2 \) 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 \( \sigma((W + \Delta)x) = \sigma(Wx) + \sigma'(Wx) \odot (\Delta x) + o(\|\Delta\|_F) \) when \( \Delta \to 0 \). Here, both \( \sigma \) and \( \sigma' \) are applied elementwise, and \( \odot \) denotes the elementwise (Hadamard) product.
- So \( y - \sigma((W + \Delta)x) = (y - \sigma(Wx)) - \sigma'(Wx) \odot (\Delta x) - o(\|\Delta\|_F) \) when \( \Delta \to 0 \). Substitute this back into the square and use the identity \( \|a + b + c\|_2^2 = \|a\|_2^2 + \|b\|_2^2 + \|c\|_2^2 + 2a^Tb + 2a^Tc + 2b^Tc \) to obtain the first-order approximation to \( g(W + \Delta) \).
Remember that any terms lower order than \( \|\Delta\|_F \) are not interesting and we can always assume \( \Delta \) 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

$$f(W) = \sum_i \|y_i - W_k W_{k-1} \ldots W_2 W_1 x_i\|_F^2$$

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]
Consider \( f(x) : \mathbb{R}^n \rightarrow \mathbb{R} \)

- **directional derivative**: \( D_v f(x) = \frac{d}{dt} f(x + tv) \)

- When \( f \) is 1-st order differentiable at \( x \),

\[
D_v f(x) = \langle \nabla f(x), v \rangle.
\]

- Now \( D_v f(x) : \mathbb{R}^n \rightarrow \mathbb{R} \), what is \( D_u(D_v f)(x) \)?

\[
D_u(D_v f)(x) = \langle u, \nabla^2 f(x) v \rangle.
\]

- When \( u = v \),

\[
D_u(D_u f)(x) = \langle u, \nabla^2 f(x) u \rangle = \frac{d^2}{dt^2} f(x + tu).
\]

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

\[ \frac{\langle u, \nabla^2 f(x) u \rangle}{\|u\|^2} \]

is the directional curvature along \( u \) independent of the norm of \( u \).

\[ f(x, y) = x^2 - y^2 \]

Blue: negative curvature (bending down)
Red: positive curvature (bending up)


