Machine Learning: Multivariate Calculus

Hao Tang

January 28, 2026

- What is the derivative of $f(x) = x^2 + 2x + 3$?
- What is the gradient of $f(x, y, z) = x + 2y + 3z$?

- What is the derivative of $f(x) = x^2 + 2x + 3$?
- What is the gradient of $f(x, y, z) = x + 2y + 3z$?
- I assume you have taken CAP (MATH08058).
- In this session, we will study
  - why we need calculus in this course
  - what 1D derivative means
  - how to generalize 1D derivative to multiple dimensions

Maximum likelihood of i.i.d. Gaussian

- Suppose $x_1, x_2, \ldots, x_n$ are i.i.d. Gaussian samples. What is the maximum likelihood estimate of the Gaussian mean?

Maximum likelihood of i.i.d. Gaussian

- Suppose $x_1, x_2, \ldots, x_n$ are i.i.d. Gaussian samples. What is the maximum likelihood estimate of the Gaussian mean?
- The likelihood is defined as the distribution value given the data points, i.e.,

$$p(x_1, x_2, \ldots, x_n) \tag{1}$$

- Due to the i.i.d. assumption,

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)^\top \Sigma^{-1} (x_i - \mu)\right) \tag{2}$$

- The likelihood is a function of the parameters, in this case, μ and Σ.

Maximum likelihood of i.i.d. Gaussian

- The maximum likelihood estimation of the mean is the μ that maximizes $p(x_1, \ldots, x_n)$.

- In other words, let $L(\mu) = p(x_1, \ldots, x_n)$, the maximum likelihood estimation of the mean can be written as

$$\operatorname*{argmax}_{\mu} L(\mu). \tag{3}$$

Maximum likelihood of i.i.d. Gaussian

- The first thing to notice is that

$$\operatorname*{argmax}_{\mu} L(\mu) = \operatorname*{argmax}_{\mu} \log L(\mu). \tag{4}$$

- The log likelihood can be written as

$$\log p(x_1, \ldots, x_n) = \log \prod_{i=1}^{n} p(x_i) = \sum_{i=1}^{n} \log p(x_i) \tag{5}$$

$$= \sum_{i=1}^{n} \left[-\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) \right]. \tag{6}$$

Maximum likelihood of i.i.d. Gaussian

- We find the maximum by taking the derivative and setting it to zero.

$$\frac{\partial}{\partial \mu} \log L(\mu) = 0 \tag{7}$$

- With some calculus, we have

$$\frac{\partial}{\partial \mu} \log L(\mu) = -\Sigma^{-1} \left(\sum_{i=1}^{n} x_i - n\mu \right). \tag{8}$$

- We can conclude that the maximum likelihood estimate of the mean is

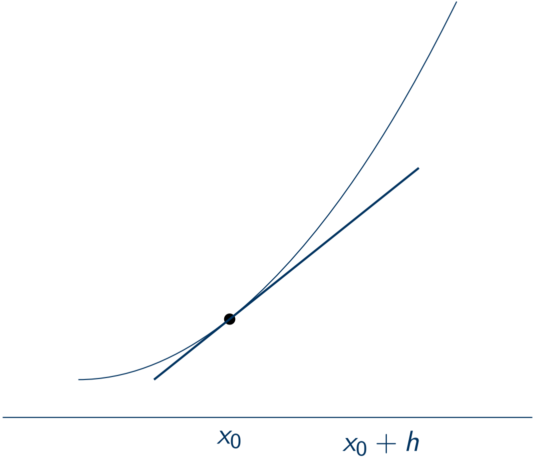
$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{9}$$

Derivative in 1D

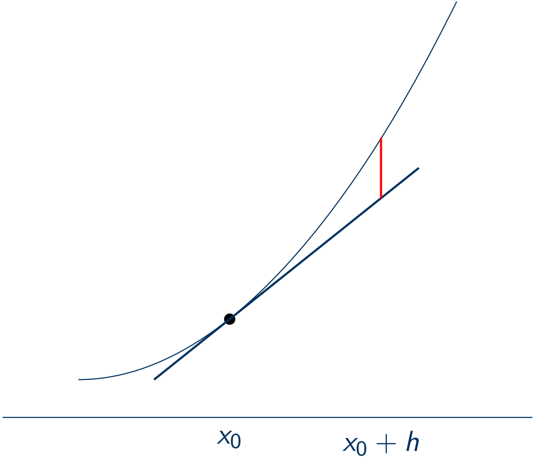
The derivative of a function $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$(D_x f)(x_0) = \left(\frac{d}{dx} f \right)(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}. \tag{10}$$

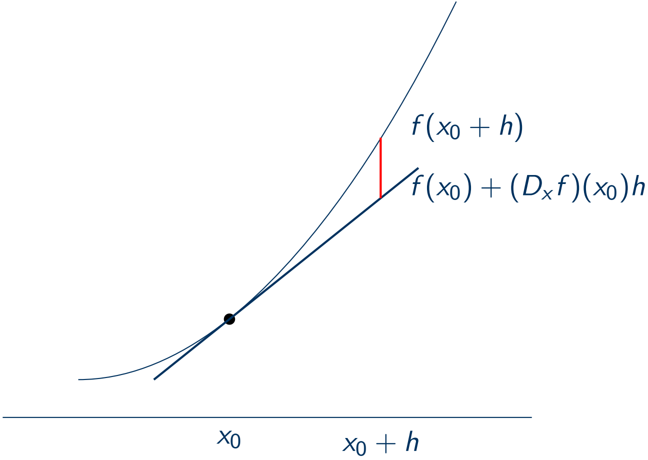
Derivative as linear approximation



Derivative as linear approximation



Derivative as linear approximation



Derivative as linear approximation

- Consider the line $g_{x_0}(x) = f(x_0) + (D_x f)(x_0)(x - x_0)$.
- The term $E(x) = |f(x) - g_{x_0}(x)|$ defines the vertical distance between the line and the function.
- Think of approximating the function with the line, and $E(x)$ tells us how bad this approximation is.
- The error has to become small as we get close to x_0, i.e.,

$$\lim_{x \to x_0} \frac{E(x)}{x - x_0} = 0. \tag{11}$$

Alternative definition of derivative

- Suppose we have a function T that is linear. If

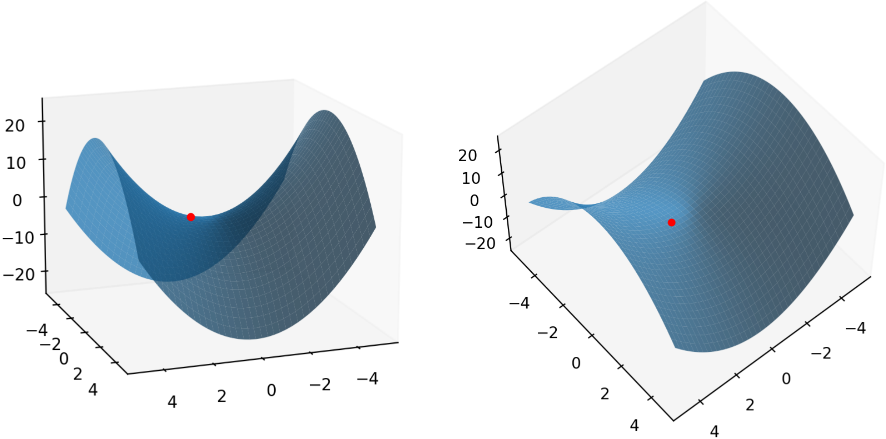
$$\lim_{h \to 0} \frac{f(x+h) - [f(x) + T(x)h]}{h} = 0 \tag{12}$$

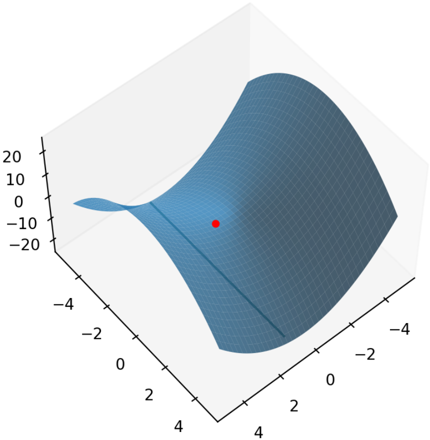
for all x, we have

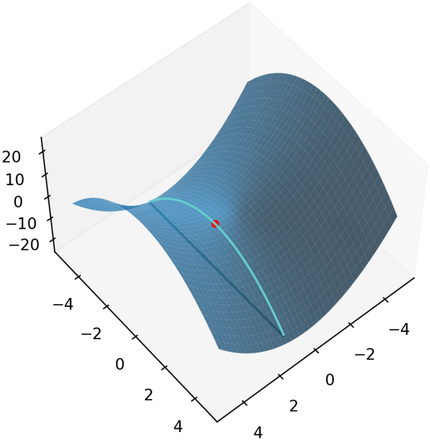
$$T(x) = (D_x f)(x) \tag{13}$$

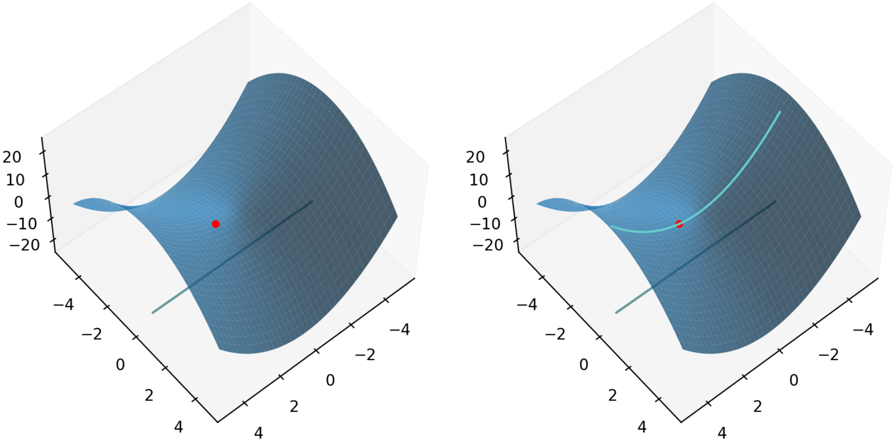
for all x.

- In words, the derivative locally gives the best linear approximation of a function.


20
10
0
−10
−20
−4
−2
0
2
4
4
2
0
−2
−4
20
10
0
−10
−20
−4
−2
0
2
4
4
2
0
−2
−4


20
10
0
−10
−20
−4
−2
0
2
4
−4
−2
0
2
4


20
10
0
−10
−20
−4
−2
0
2
4
−4
−2
0
2
4


20
10
0
−10
−20
−4
−2
0
2
4
−4
−2
0
2
4
20
10
0
−10
−20
−4
−2
0
2
4
−4
−2
0
2
4

Directional derivative

- The directional derivative of $f : \mathbb{R}^d \to \mathbb{R}$ along the direction v at $x_0 \in \mathbb{R}^d$ is defined as

$$(D_v f)(x_0) = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t}. \tag{14}$$

Directional derivative

- The directional derivative of $f : \mathbb{R}^d \to \mathbb{R}$ along the direction v at $x_0 \in \mathbb{R}^d$ is defined as

$$(D_v f)(x_0) = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t}. \tag{14}$$

- If we let $g_{x_0}(t) = f(x_0 + tv)$, then

$$(D_t g)(0) = \lim_{t \to 0} \frac{g(0 + t) - g(0)}{t} = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} = (D_v f)(x_0) \tag{15}$$

Example

- Consider the function $f(x, y) = x^2 - y^2$.

- If we are at $(2, 0)$, the directional derivative along $(1, 0)$ is 4.

Example

- Consider the function $f(x, y) = x^2 - y^2$.
- If we are at $(2, 0)$, the directional derivative along $(1, 0)$ is 4.
- If we take a line at $\{(x, y) : (x, y) = (2, 0) + t(1, 0) = (2 + t, 0) \text{ for } t \in \mathbb{R}\}$, we have $g(t) = f(2 + t, 0) = (2 + t)^2$. The derivative $(D_t g)(t) = 2(2 + t)$, and $(D_t g)(0) = 2 \cdot (2 + 0) = 4$.

Partial derivatives

- A partial derivative is a directional derivative along the direction of coordinate axes.

Partial derivatives

- A partial derivative is a directional derivative along the direction of coordinate axes.

- In a three-dimensional space, the direction of the axes are

$$(1,0,0) \quad (0,1,0) \quad (0,0,1). \tag{16}$$

For a function $f : \mathbb{R}^3 \to \mathbb{R}$, the partial derivatives along the axes are

$$\frac{\partial}{\partial x} f \quad \frac{\partial}{\partial y} f \quad \frac{\partial}{\partial z} f. \tag{17}$$

Example

- Given a function $f(x, y) = x^2 - y^2$,

$$\left(\frac{\partial}{\partial x} f\right)(x, y) \qquad \left(\frac{\partial}{\partial y} f\right)(x, y) \tag{18}$$

Example

- Given a function $f(x, y) = x^2 - y^2$,

$$\left(\frac{\partial}{\partial x} f\right)(x, y) = 2x \qquad \left(\frac{\partial}{\partial y} f\right)(x, y) = -2y. \tag{18}$$

Example

- Given a function $f(x, y) = x^2 - y^2$,

$$\left(\frac{\partial}{\partial x} f\right)(x, y) = 2x \qquad \left(\frac{\partial}{\partial y} f\right)(x, y) = -2y. \tag{18}$$

- The x-axis is the direction $(1, 0)$. At any point (x, y), the line along that direction is $(x + t, y)$. The function value along that line is $g(t) = f(x + t, y) = (x + t)^2 - y^2$. We then have $(D_t g)(t) = 2(x + t)$, and

$$\left(\frac{\partial}{\partial x} f\right)(x, y) = (D_t g)(0) = 2x. \tag{19}$$

Example

- Given a function $f(x, y) = x^2 - y^2$,

$$\left(\frac{\partial}{\partial x} f\right)(x, y) = 2x \qquad \left(\frac{\partial}{\partial y} f\right)(x, y) = -2y. \tag{18}$$

- The x-axis is the direction $(1, 0)$. At any point (x, y), the line along that direction is $(x + t, y)$. The function value along that line is $g(t) = f(x + t, y) = (x + t)^2 - y^2$. We then have $(D_t g)(t) = 2(x + t)$, and

$$\left(\frac{\partial}{\partial x} f\right)(x, y) = (D_t g)(0) = 2x. \tag{19}$$

- Treat other variables as constants and take 1D derivatives.

Example

- Given a function $f(x, y, z) = (x + 2y - 3z)^2$,

$$\left(\frac{\partial}{\partial x} f\right)(x, y, z) \tag{20}$$

$$\left(\frac{\partial}{\partial y} f\right)(x, y, z) \tag{21}$$

$$\left(\frac{\partial}{\partial z} f\right)(x, y, z) \tag{22}$$

Example

- Given a function $f(x, y, z) = (x + 2y - 3z)^2$,

$$\left(\frac{\partial}{\partial x} f\right)(x, y, z) = 2(x + 2y - 3z) \tag{20}$$

$$\left(\frac{\partial}{\partial y} f\right)(x, y, z) = 2(x + 2y - 3z) \cdot 2 \tag{21}$$

$$\left(\frac{\partial}{\partial z} f\right)(x, y, z) = 2(x + 2y - 3z) \cdot (-3) \tag{22}$$

Example

- Given a function

$$f(w, b) = \frac{1}{1 + \exp(-(w^\top x + b))}, \tag{23}$$

show that

$$\left(\frac{\partial}{\partial b} f\right)(w, b) = f(w, b)(1 - f(w, b)). \tag{24}$$

Gradients

- The gradient of a function is the vector consisting of all partial derivatives.

Gradients

- The gradient of a function is the vector consisting of all partial derivatives.
- For a function $f : \mathbb{R}^3 \to \mathbb{R}$, its gradient is

$$(\nabla f)(x, y, z) = \begin{bmatrix} \left(\frac{\partial}{\partial x} f\right)(x, y, z) \\ \left(\frac{\partial}{\partial y} f\right)(x, y, z) \\ \left(\frac{\partial}{\partial z} f\right)(x, y, z) \end{bmatrix}. \tag{25}$$

Example

- Given a function $f(x, y, z) = (x + 2y - 3z)^2$, show that its gradient is

$$(\nabla f)(x, y, z) = \begin{bmatrix} 2(x + 2y - 3z) \\ 2(x + 2y - 3z) \cdot 2 \\ 2(x + 2y - 3z) \cdot (-3) \end{bmatrix}. \tag{26}$$

Example

- Given a function $f(a) = b^\top a$, show that its gradient is

$$(\nabla f)(a) = b. \tag{27}$$

Example

- Given a function $f(a) = b^\top a$, show that its gradient is

$$(\nabla f)(a) = b. \tag{27}$$

- Given a function $f(a) = b^\top A a$, show that its gradient is

$$(\nabla f)(a) = A^\top b. \tag{28}$$

Example

- Given a function $f(a) = \|a\|_2^2$, show that its gradient is

$$(\nabla f)(a) = 2a. \tag{29}$$

Example

- Given a function $f(w) = (w^\top x + b - y)^2$, show that

$$(\nabla f)(w) = 2(w^\top x + b - y)x. \tag{30}$$

Example

- Given a function

$$f(w) = \frac{1}{1 + \exp(-(w^\top x + b))}, \tag{31}$$

show that its gradient is

$$(\nabla f)(w) = f(w)(1 - f(w))x. \tag{32}$$

Theorem

- For a function $f : \mathbb{R}^d \to \mathbb{R}$ and any direction v at any point x, show that

$$(D_v f)(x) = (\nabla f)(x)^\top v. \tag{33}$$

- Once we know the gradient, we know all directional derivatives.

Second-order derivative

For a function $f : \mathbb{R} \to \mathbb{R}$, its second-order derivative is defined and written as

$$\frac{\partial^2}{\partial x^2} f = \frac{\partial}{\partial x} \left(\frac{\partial}{\partial x} f \right). \tag{34}$$

Example

- Given a function $f(x) = x^2$, it's second-order derivative is 2.

Example

- Given a function $f(x) = x^2$, it's second-order derivative is 2.

- The second-order derivative tells us whether the function looks like a cup or an upside-down cup.

Hessian

- The Hessian of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\begin{bmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f & \frac{\partial^2}{\partial x_1 \partial x_2} f & \dots & \frac{\partial^2}{\partial x_1 \partial x_d} f \\ \frac{\partial^2}{\partial x_2 \partial x_1} f & \frac{\partial^2}{\partial x_2 \partial x_2} f & \dots & \frac{\partial^2}{\partial x_2 \partial x_d} f \\ \vdots f & & & \vdots \\ \frac{\partial^2}{\partial x_d \partial x_1} f & \frac{\partial^2}{\partial x_d \partial x_2} f & \dots & \frac{\partial^2}{\partial x_d \partial x_d} f \end{bmatrix}. \tag{35}$$

- Because

$$\frac{\partial^2}{\partial x_j \partial x_i} f = \frac{\partial^2}{\partial x_i \partial x_j} f, \tag{36}$$

the Hessian matrix is always symmetric.

Example

- Given a function $f(x, y) = x^2 - y^2$, its Hessian is

Example

- Given a function $f(x,y) = x^2 - y^2$, its Hessian is $\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$.

Taylor approximation

- The function

$$f(x_0) + \nabla f(x_0)^{(}x - x_0) \tag{37}$$

is locally the best first-order approximation.

- The function

$$f(x_0) + \nabla f(x_0)^\top (x - x_0) + \frac{1}{2}(x - x_0)^\top H(x - x_0) \tag{38}$$

is locally the best linear approximation.

Further reading

- Apostol, “Calculus,” Wiley, 1975
- Spivak, “Calculus on manifolds: A modern approach to classical theorems of advanced calculus,” Westview Press, 1971
- Colley and Cañez, “Vector calculus,” Pearsonn, 2023
- Marsden and Tromba, “Vector calculus,” W. H. Freeman, 2011