INFR10086 Machine Learning (MLG) Semester 2, 2023/4

Multivariate Calculus

Lecturer: Hao Tang

Definition 1. The derivative of a function $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined and written as

$$(D_x f)(x_0) = \left(\frac{d}{dx} f\right)(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}. \tag{1}$$

- The x in $\frac{d}{dx} f$ is just the name of the variable, not an actual variable that we can plug in.
- The object $\frac{d}{dx} f$ is a function $\mathbb{R} \to \mathbb{R}$. and $\left(\frac{d}{dx} f\right)(x_0)$ means that we plug in x_0 to the function.
- The more common notation is $\frac{df(x)}{dx}$. This way of writing can be confusing. The problem is that $f(x)$ can be the function itself of f evaluated at x. We will avoid this notation when possible.

Example 1. Consider the line $g_{x_0}(x) = f(x_0) + (D_x f)(x_0)(x - x_0)$. How well this line approximates f can be described by the error $E(x) = |f(x) - g_{x_0}(x)|$. Show that the appoximation becomes arbitrarily good as we get close to x_0.

- The error term looks a lot like a derivative if we devide it by $x - x_0$.

$$E(x) = \frac{|f(x) - g_{x_0}(x)|}{x - x_0}(x - x_0) = \frac{|f(x) - f(x_0) - (D_x f)(x_0)(x - x_0)|}{x - x_0}(x - x_0) \tag{2}$$

$$= \left| \frac{f(x) - f(x_0)}{x - x_0} - (D_x f)(x_0) \right| (x - x_0) \tag{3}$$

In other words,

$$\lim_{x \to x_0} \frac{E(x)}{x - x_0} = 0. \tag{4}$$

- Instread of looking at a different point x, we can also look at how far off we are from x_0. If we let $x = x_0 + h$,

$$E(x_0 + h) = |f(x_0 + h) - f(x_0) - (D_x f)(x_0) h|. \tag{5}$$

We have

$$\lim_{h \to 0} \frac{E(x_0 + h)}{h} = 0. \tag{6}$$

- In words, the derivative of a function offers a good linear approximation locally for any point.

Example 2. Suppose we have a function T that is linear. Show that if

$$\lim_{h \to 0} \frac{f(x+h) - [f(x) + T(x)h]}{h} = 0 \tag{7}$$

for all x, we have

$$T(x) = (D_x f)(x) \tag{8}$$

for all x.

- Because

  $$\lim_{h \to 0} \frac{f(x+h) - [f(x) + T(x)h]}{h} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} - T(x) = (D_x f)(x) - T(x) = 0, \tag{9}$$

  we have the desired property for all x.

- The result seems almost trivial, but we double down on the property of linear approximation. It offers an alternative definition of differentiation.

Definition 2. 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

$$(D_v f)(x_0) = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t}. \tag{10}$$

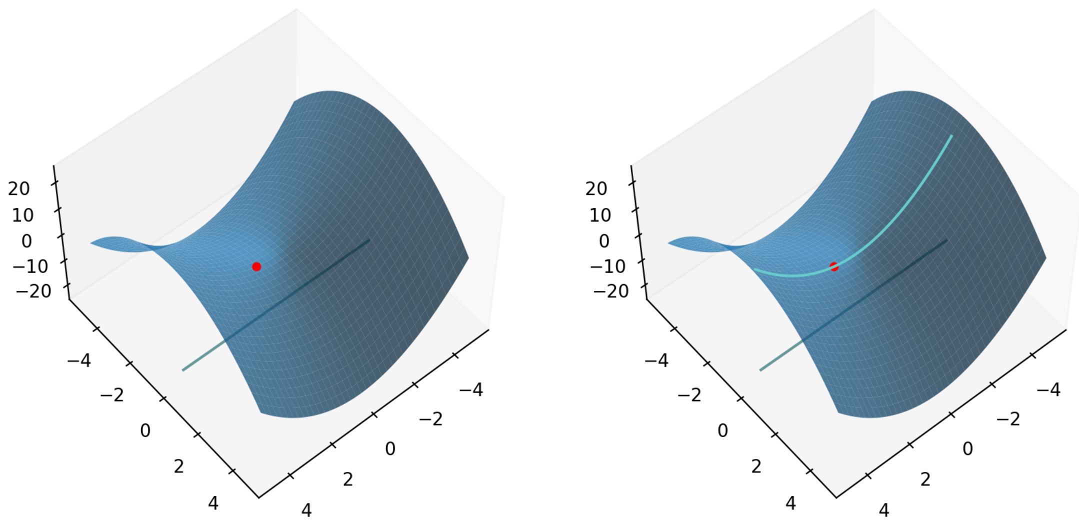
This is a simple way of extending single-variate functions to multivariate functions. To see this, we can write $g(t) = f(x_0 + tv)$. Note that $g : \mathbb{R} \to \mathbb{R}$ is a single-variate function. If we take the derivative of g at 0, we have

$$(D_t g)(0) = \lim_{t \to 0} \frac{g(0+t) - g(0)}{t} = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} = (D_v f)(x_0) \tag{11}$$

Example 3. Consider the function $f(x, y) = x^2 - y^2$. Show that 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 $(D_t g)(0) = 2 \cdot (2 + 0) = 4$.

- If we plot the line $(2, 0) + t(1, 0)$ and walk along the line on the function, we should be able to see the parabola $(t + 2)^2$. Taking the derivative of a parabola is something we know how to do.



Definition 3. A partial derivative is a directional derivative along the direction of coordinate axes.

Example 4. In a three-dimensional space, the direction of the axes are

$$(1, 0, 0) \quad (0, 1, 0) \quad (0, 0, 1). \tag{12}$$

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

$$\frac{\partial}{\partial x} f \quad \frac{\partial}{\partial y} f \quad \frac{\partial}{\partial z} f. \tag{13}$$

Example 5. Given a function $f(x, y) = x^2 - y^2$, show that

$$\left(\frac{\partial}{\partial x} f\right)(x, y) = 2x \qquad \left(\frac{\partial}{\partial y} f\right)(x, y) = -2y. \tag{14}$$

- 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

$$\left(\frac{\partial}{\partial x} f\right)(x, y) = (D_t g)(0) = 2x. \tag{15}$$

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

$$\left(\frac{\partial}{\partial y} f\right)(x, y) = (D_t g)(0) = -2y. \tag{16}$$

- We can always compute partial derivatives from first principles. However, it is much simpler to treat other variables as constants, pretending that everything is single variate and taking the derivative.

Example 6. Given a function $f(x, y, z) = (x + 2y - 3z)^2$, show that

$$\left(\frac{\partial}{\partial x} f\right)(x, y, z) = 2(x + 2y - 3z) \tag{17}$$

$$\left(\frac{\partial}{\partial y} f\right)(x, y, z) = 2(x + 2y - 3z) \cdot 2 \tag{18}$$

$$\left(\frac{\partial}{\partial z} f\right)(x, y, z) = 2(x + 2y - 3z) \cdot (-3) \tag{19}$$

Example 7. Given a function

$$f(w, b) = \frac{1}{1 + \exp(-(w^\top x + b))}, \tag{20}$$

show that

$$\left(\frac{\partial}{\partial b} f\right)(w, b) = f(w, b)(1 - f(w, b)). \tag{21}$$

$$\left(\frac{\partial}{\partial b} f\right)(w, b) = \frac{1}{1 + \exp(-(w^\top x + b))} \frac{\exp(-(w^\top x + b))}{1 + \exp(-(w^\top x + b))} = f(w, b)(1 - f(w, b)) \tag{22}$$

Definition 4. The gradient of a function is the vector consisting of all partial derivatives.

Example 8. For a function $f : \mathbb{R}^3 \to \mathbb{R}$, its gradient is

$$(\nabla f)(x, y, z) = \begin{bmatrix} \left(\frac{\partial}{\partial x} f\right)(x, y, z) \\ \left(\frac{\partial}{\partial y} f\right)(x, y, z) \\ \left(\frac{\partial}{\partial z} f\right)(x, y, z) \end{bmatrix}. \tag{23}$$

Example 9. Given a function $f(x, y, z) = (x + 2y - 3z)^2$, show that its gradient is

$$(\nabla f)(x, y, z) = \begin{bmatrix} 2(x + 2y - 3z) \\ 2(x + 2y - 3z) \cdot 2 \\ 2(x + 2y - 3z)(-3) \end{bmatrix}. \tag{24}$$

Example 10. Given a function $f(a) = b^\top a$, show that its gradient is

$$(\nabla f)(a) = b. \tag{25}$$

Based on the definition of gradient,

$$\frac{\partial}{\partial a_i} f = \frac{\partial}{\partial a_i} \sum_{i=1}^{d} a_i b_i = b_i. \tag{26}$$

Putting the partial derivatives into a vector, we have

$$(\nabla f)(a) = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{bmatrix} = b. \tag{27}$$

Note that the result is a constant vector and does not depend on a.

Example 11. Given a function $f(a) = b^\top A a$, show that its gradient is

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

Example 12. Given a function $f(a) = \|a\|_2^2$, show that its gradient is

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

Following the definition of gradient, we have

$$\frac{\partial}{\partial a_i} f = \frac{\partial}{\partial a_i} \sum_{i=1}^{d} a_i^2 = 2a_i \qquad (\nabla f)(a) = \begin{bmatrix} 2a_1 \\ 2a_2 \\ \vdots \\ 2a_d \end{bmatrix} = 2a. \tag{30}$$

Example 13. Given a function $f(w) = (w^\top x + b - y)^2$, show that

$$(\nabla f)(w) = 2(w^\top x + b - y)x. \tag{31}$$

Example 14. Given a function

$$f(w) = \frac{1}{1 + \exp(-(w^\top x + b))}, \tag{32}$$

show that its gradient is

$$(\nabla f)(w) = f(w)(1 - f(w))x. \tag{33}$$

$$(\nabla f)(w) = \frac{1}{1+\exp(-(w^\top + b))} \frac{\exp(-(w^\top x + b))}{1+\exp(-(w^\top x + b))} x = f(w)(1-f(w))x. \tag{34}$$

Example 15. For a function $f : \mathbb{R}^d \to \mathbb{R}$ and any direction v at any point x, show that

$$(D_v f)(x) = (\nabla f)(x)^\top v. \tag{35}$$

- Let $g(t) = f(u(t))$, where $u = x + tv$. By chain rule,

$$\begin{aligned}
(D_t g)(t) &= \sum_{i=1}^{d} \Big[\left(D_{u_i} f\right)(u(t)) \Big] \Big[\left(D_t u_i\right)(t) \Big] && (36) \\
&= \begin{bmatrix} \left(\frac{\partial}{\partial u_1} f\right)(u(t)) & \dots & \left(\frac{\partial}{\partial u_d} f\right)(u(t)) \end{bmatrix}^\top \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix} && (37) \\
&= (\nabla f)(u(t))^\top v. && (38)
\end{aligned}$$

  Finally,

$$(D_v f)(x) = (D_t g)(0) = (\nabla f)(x)^\top v. \tag{39}$$

Definition 5. For a function $f : \mathbb{R} \to \mathbb{R}$, its second-order derivative is defined and written as

$$\frac{\partial^2}{\partial x^2} f = \frac{\partial}{\partial x} \left(\frac{\partial}{\partial x} f \right). \tag{40}$$

Example 16. 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.
- For this particular example, the second-order derivative is possitive, so the paraballa looks like a cup.

Definition 6. The Hessian of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\begin{bmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f & \frac{\partial^2}{\partial x_1 \partial x_2} f & \dots & \frac{\partial^2}{\partial x_1 \partial x_d} f \\ \frac{\partial^2}{\partial x_2 \partial x_1} f & \frac{\partial^2}{\partial x_2 \partial x_2} f & \dots & \frac{\partial^2}{\partial x_2 \partial x_d} f \\ \vdots f & & & \vdots \\ \frac{\partial^2}{\partial x_d \partial x_1} f & \frac{\partial^2}{\partial x_d \partial x_2} f & \dots & \frac{\partial^2}{\partial x_d \partial x_d} f \end{bmatrix}. \tag{41}$$

- Because

$$\frac{\partial^2}{\partial x_i \partial x_j} f = \frac{\partial^2}{\partial x_j \partial x_i} f, \tag{42}$$

  the Hessian matrix is always symmetric.

Example 17. Given a function $f(x, y) = x^2 - y^2$, show that its Hessian is $\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$.

The 2 indicates that it looks like a cup along the x-axis, while the -2 indicates that it looks like an upside-down cup along the y-axis.