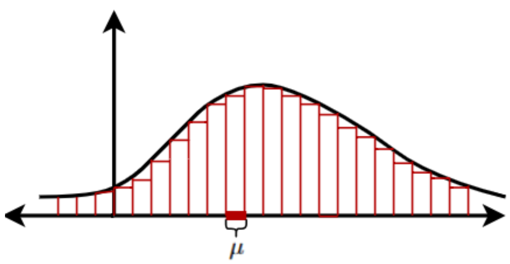
CSE 840: Computational Foundations of Artificial Intelligence — March 17, 2025

Differentiation on $\mathbb{R}^n$: partial, total, and directional derivatives

Instructor: Vishnu Boddeti — *Scribe: Muhammed Salih Kayhan*

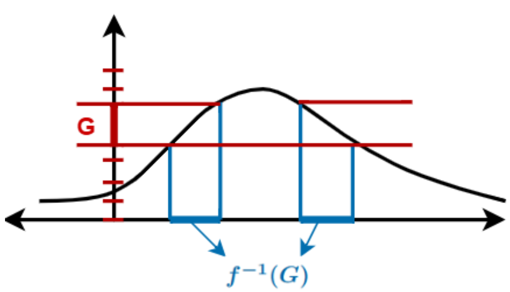
The Lebesgue Integral on $\mathbb{R}^n$

Reimann



- bounded
- continuous
- finite set of rectangles

Lebesgue

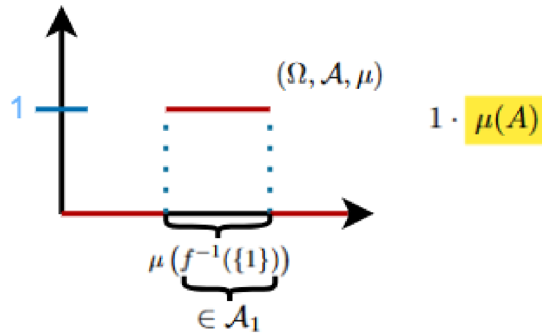


- not bounded
- need not be continuous
- countable sets

Definition 1 *: A function $f : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ between two measurable spaces is called measurable if pre-images of measurable sets are measurable:*

$$\forall A_2 \in \mathcal{A}_2 : f^{-1}(A_2) \in \mathcal{A}_1$$

where $f^{-1}(A_2) =: \{x \in \Omega_1 \mid f(x) \in A_2\}$



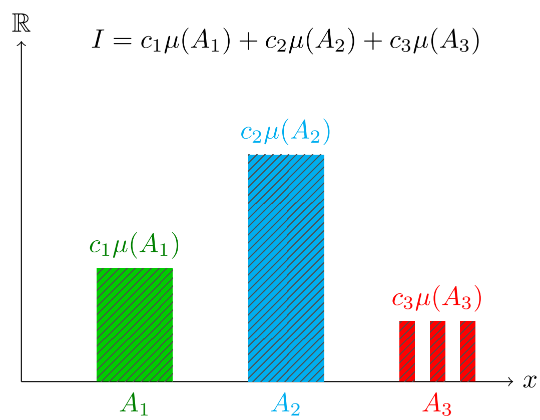
$(\Omega, \mathcal{A})$, $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

Characteristic function (also *indicator function*)

$$\chi_A : \Omega \to \mathbb{R}, \quad \chi_A(\omega) := \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

Definition 2 *: A function $\phi : \mathbb{R}^n \to \mathbb{R}$ is called a simple function if there exist measurable sets $A_i \subset \mathbb{R}^n$, $c_i \in \mathbb{R}$ such that:*

$$\phi(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(x)$$



$$\phi(x) = c_1\chi_{A_1}(x) + c_2\chi_{A_2}(x) + c_3\chi_{A_3}(x)$$

The Lebesgue integral for a simple function is defined as:

$$I(\phi) = \int \phi \, d\mu = \sum_{i=1}^{n} c_i \mu(A_i)$$

For a function $f^+ : \mathbb{R}^n \to [0, \infty)$, we define its Lebesgue integral as:

$$\int f^+ \, d\mu = \sup \left\{ \int \phi \, d\mu \mid \phi \leq f, \phi \text{ simple} \right\}$$

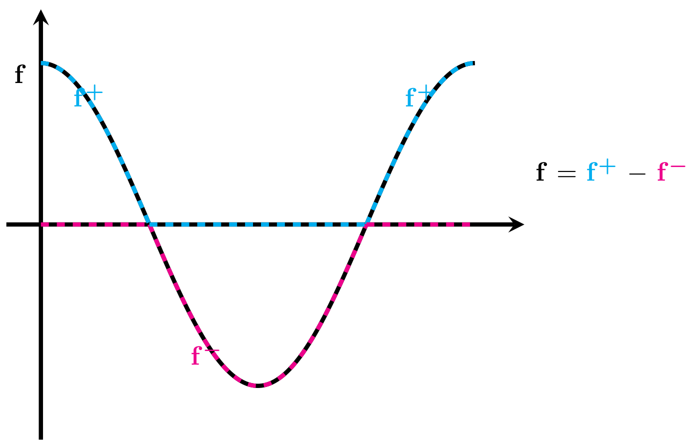
Note that this integral might be infinite.

For a general function $f : \mathbb{R}^n \to \mathbb{R}$, we split the function into positive and negative parts:

$$f = f^+ - f^-, \quad f^+ \geq 0, \quad f^- \geq 0$$

where:

$$f^+(x) = \begin{cases} f(x), & \text{if } f(x) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$



Note: f^+ and f^- are measurable if f is measurable.

If both f^+ and f^- satisfy $\int f^+ \, d\mu < \infty$ and $\int f^- \, d\mu < \infty$, then we call f integrable and define:

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu$$

This is a much more powerful notion than the Riemann integral.

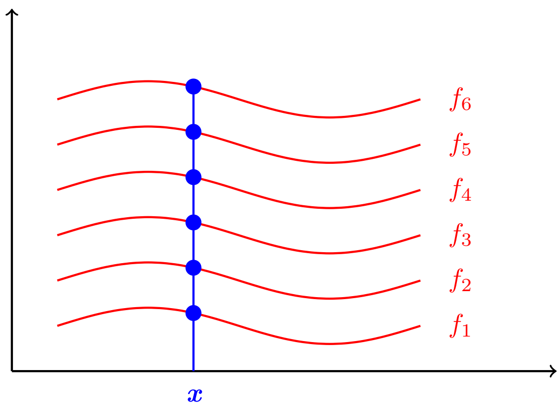
Example

$$\int \chi_{\mathbb{Q}} \, d\mu = 1 \cdot \mu(\mathbb{Q}) = 0$$

Two Important Theorems

Theorem (Monotone Convergence) *: Consider a sequence of functions* $f_n : \mathbb{R}^n \to [0, \infty)$ *that is pointwise non-decreasing:* $\forall x \in \mathbb{R}^n, f_{k+1}(x) \geq f_k(x)$*. Assume that all* f_k *are measurable, and that the pointwise limit exists:*

$$\forall x, \lim f_k(x) =: f(x).$$



Then:

$$\int \lim_{k\to\infty} f_k(x)\, dx = \lim_{k\to\infty} \int f_k(x)\, dx$$

That is,

$$\int f(x)\, dx = \lim_{k\to\infty} \int f_k(x)\, dx$$

Theorem (Dominated Convergence) *:* $f_k : B \to \mathbb{R}$ *be a sequence of functions such that* $|f_k(x)| \leq g(x)$ *on* B*, where* $g(x)$ *is integrable. Assume that the pointwise limit exists:*

$$\forall x \in B, \quad f(x) := \lim_{k\to\infty} f_k(x).$$

Then:

$$\int \lim_{k\to\infty} f_k(x)\, dx = \lim_{k\to\infty} \int f_k(x)\, dx.$$

That is,

$$\int f(x)\, dx = \lim_{k\to\infty} \int f_k(x)\, dx.$$

Partial Derivatives on $\mathbb{R}^n$

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$.

Definition *: f is called partially differentiable with respect to variable x_j at point $\xi \in \mathbb{R}^n$ if the function*

$$x_j \mapsto g(x_j) := f(\xi_1, \xi_2, \ldots, \xi_{j-1}, x_j, \xi_{j+1}, \ldots, \xi_n)$$

$g : \mathbb{R} \to \mathbb{R}$ is differentiable at $\xi_j \in \mathbb{R}$.

Notation:

$$\frac{\partial f(\xi)}{\partial x_j} = \lim_{h \to 0} \frac{f(\xi + e_j \cdot h) - f(\xi)}{h}$$

where h is a scalar, and e_j is the j-th unit vector, which has a 1 at the j-th index and zeros everywhere else.

For example, if $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ and $f(x) = x_1^2 + x_2^2 \cdot x_1$, then $f : \mathbb{R}^n \to \mathbb{R}$.

If all partial derivatives exist, then the vector of all partial derivatives is called the gradient:

$$\text{grad}(f)(\xi) = \nabla f(\xi) = \begin{pmatrix} \frac{\partial f(\xi)}{\partial x_1} \\ \vdots \\ \frac{\partial f(\xi)}{\partial x_n} \end{pmatrix} \in \mathbb{R}^n$$

If $f : \mathbb{R}^n \to \mathbb{R}^m$, we decompose f into its m component functions $f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix}$. We define the Jacobian matrix:

$$Df(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} (\nabla f_1(x)) \\ \vdots \\ (\nabla f_m(x)) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

The i-th row of the Jacobian matrix is the gradient of f_i.

Caution: Even if all partial derivatives exist at ξ, we do not know if f is continuous at ξ.

Example: Consider $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x, y) = \begin{cases} \frac{x \cdot y}{x^2 + y^2}, & \text{if } (x, y) \neq (0, 0) \\ 0, & \text{if } x = y = 0 \end{cases}$$

For $(x,y) \neq (0,0)$,

$$\nabla f(x,y) = \left(y \cdot \frac{y^2 - x^2}{(x^2+y^2)^2}, x \cdot \frac{x^2 - y^2}{(x^2+y^2)^2} \right)$$

$\nabla f(0,0) = 0$ since $f(x,0) = 0\ \forall x$ and $f(0,y) = 0\ \forall y$, but f is not continuous at 0.

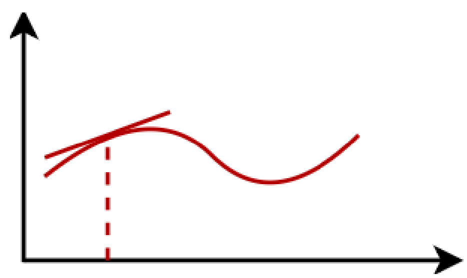
Total Derivative

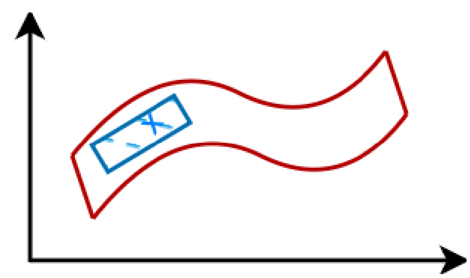
$f : \mathbb{R}^n \to \mathbb{R}^m$ and $\xi \in U$. f is differentiable at ξ if there exists a linear mapping $L : \mathbb{R}^n \to \mathbb{R}^m$ such that for $h \in \mathbb{R}^n$,

$$f(\xi + h) - f(\xi) = L(h) + r(h)$$

with

$$\lim_{h \to 0} \frac{r(h)}{|h|} \to 0.$$





Intuition: f is "locally linear"

Theorem : *$f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at ξ:*

- *Then f is continuous at ξ*
- *The linear functional L coincides with the gradient:*

$$f(\xi + h) - f(\xi) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\xi) \cdot h_j + r(h) = \langle \nabla f(\xi), h \rangle + r(h)$$

If $f : \mathbb{R}^n \to \mathbb{R}^m$, it is differentiable if all coordinate functions $f_1, f_2, \ldots, f_m$ are differentiable. Then all partial derivatives exist and $L(h) = (\text{Jacobian matrix}) \cdot h$.

Theorem : *If all partial derivatives exist and are all continuous, then f is differentiable.*

Cautioun: If partial derivatives exist but are not continuous, then f doesn't need to be differentiable.

Directional Derivatives

Definition *: Assume $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $v \in \mathbb{R}^n$ with $\|v\| = 1$. The directional derivative of f at ξ in the direction of v is defined as:*

$$D_v f(\xi) = \lim_{t \to 0} \frac{f(\xi + t \cdot v) - f(\xi)}{t}$$

In this equation, $t \in \mathbb{R}$ is a scalar and $\mathbf{v} \in \mathbb{R}^n$ is a unit vector corresponding to a direction.

Theorem *: $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at ξ, then all the directional derivatives exist, and we can compute them as:*

$$D_v f(\xi) = (\nabla f(\xi))^T \cdot v = \sum_{i=1}^{n} v_i \frac{\partial f(\xi)}{\partial x_i}$$

In this equation, $v_i \in \mathbb{R}$ is a scalar, and $\mathbf{v}$ is a vector.

The largest value of all directional derivatives is attained in the direction:

$$v = \frac{\nabla f(\xi)}{\|\nabla f(\xi)\|}$$

Explored Supplementary Concepts for CSE 840

Vertical or Horizontal Slices? Riemann and Lebesgue Integration

Riemann Integration: Vertical Slices

The Riemann integral partitions the domain into subintervals and approximates the area under a curve using vertical slices.

Definition: Let $f : [a, b] \to \mathbb{R}$ be bounded. A partition π of $[a, b]$ is defined as:

$$\pi := \{a = t_0, t_1, \ldots, t_N = b\}$$

Define:

$$m_j = \inf_{t \in [t_{j-1}, t_j]} f(t), \quad M_j = \sup_{t \in [t_{j-1}, t_j]} f(t)$$

Then, the lower and upper Darboux sums are:

$$S_\pi[f] := \sum_{j=1}^{N} m_j \Delta t_j, \qquad S^\pi[f] := \sum_{j=1}^{N} M_j \Delta t_j$$

f is Riemann integrable if:

$$\int_* f := \sup_\pi S_\pi[f] = \inf_\pi S^\pi[f] =: \int^* f$$

Limitation Example

Let f be the indicator function of rational numbers in $[0,1]$. Since rationals are dense, both $m_j = 0$ and $M_j = 1$ in every subinterval, and the upper and lower sums do not converge. Hence, f is not Riemann integrable.

Lebesgue Integration: Horizontal Slices

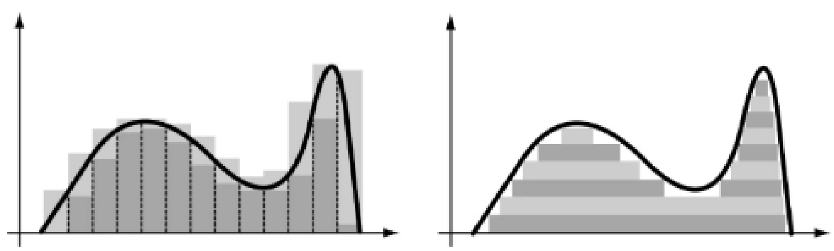
Lebesgue's approach partitions the *range* of the function, grouping points in the domain that map to the same function value. This leads to **horizontal slicing**.

Core Idea: Measure how much of the domain maps to a given function value and weight that value accordingly.

Advantages

- Can integrate more "wild" functions, such as the characteristic function of $\mathbb{Q} \cap [0,1]$
- Enables powerful theorems like dominated and monotone convergence
- Lebesgue measure assigns measure 0 to $\mathbb{Q} \cap [0,1]$, making its integral 0

Visual Comparison



Left: Riemann—vertical slices based on domain subdivision. *Right:* Lebesgue—horizontal slices based on function values.

Conclusion

The Riemann integral is convenient for calculating the primitive, or anti-derivative, of the integrand of 'reasonably behaved' functions. However, it fails to provide a meaningful results for more exotic functions. The Lebesgue theory comes to the rescue, and it provides very powerful theorems that justify the interchange of limits and integrals.

Further Reading

- Schilling (2005), *Measures, Integrals and Martingales. Cambridge Univ. Press, 381pp. ISBN 978-0-5216-1525-9.*

Partial Derivatives in Machine Learning

Partial derivatives play a vital role in machine learning, particularly in optimization algorithms like gradient descent. They help us understand how a function changes with respect to its input variables, allowing us to optimize model parameters effectively, even in high-dimensional spaces.

Definition *Let $f(x_1, x_2, \ldots, x_n)$ be a multivariable function. The partial derivative with respect to x_i is:*

$$\frac{\partial f}{\partial x_i}$$

It represents the rate of change of the function f with respect to x_i, keeping all other variables constant.

Gradient and Gradient Descent

The **gradient** of a function is a vector of all partial derivatives:

$$\nabla f = \left\langle \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right\rangle$$

It points in the direction of the function's steepest ascent.

Gradient Descent is an iterative optimization method that updates parameters in the direction of steepest descent (i.e., the negative gradient) in order to minimize a cost function.

Application: Linear Regression Model

Consider the basic linear regression model:

$$f(x) = wx + b$$

where:

- x is the input feature,
- w is the weight (slope),
- b is the bias (intercept).

We aim to minimize the Mean Squared Error cost function:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^{m} (wx_i + b - y_i)^2$$

To minimize this, we compute the partial derivatives:

$$\frac{\partial J}{\partial w} = \frac{1}{m} \sum_{i=1}^{m} (wx_i + b - y_i)x_i, \qquad \frac{\partial J}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} (wx_i + b - y_i)$$

Gradient Descent Update Rules:

$$w := w - \alpha \frac{\partial J}{\partial w}, \quad b := b - \alpha \frac{\partial J}{\partial b}$$

where α is the learning rate, controlling the step size during optimization.

Example: Implementation in Python

```
import numpy as np

# Sample dataset
X = np.array([1, 2, 3, 4, 5])  # House sizes
y = np.array([100, 200, 300, 400, 500])  # House prices

# Initialize parameters
w = 0
b = 0
learning_rate = 0.01
epochs = 100

# Gradient Descent Loop
for epoch in range(epochs):
    predictions = w * X + b
    dw = (1 / len(X)) * np.sum((predictions - y) * X)
    db = (1 / len(X)) * np.sum(predictions - y)
    w -= learning_rate * dw
    b -= learning_rate * db

print("Optimal parameters: w =", w, "b =", b)
```

Output:
Optimal parameters: `w = 93.98`, `b = 21.72`

This simple example demonstrates how partial derivatives are used to compute gradients, which are then used to iteratively optimize model parameters using gradient descent.

Conclusion

Partial derivatives are fundamental in training machine learning models. They allow the computation of gradients, which are used in gradient-based optimization algorithms like gradient descent. Understanding and applying partial derivatives enables effective model training and performance improvements.

FAQs

What is a partial derivative?

A partial derivative measures how a multivariable function changes with respect to one of its input variables, keeping the others fixed.

How does gradient descent work?
Gradient descent is an optimization algorithm that iteratively updates model parameters in the opposite direction of the gradient to minimize the cost function.

Why are partial derivatives important in machine learning?
They help compute gradients needed for optimization, making them crucial for training models efficiently.

What is the role of the learning rate in gradient descent?
The learning rate controls the step size during parameter updates. If it's too small, convergence is slow; if too large, the algorithm may overshoot or diverge.

The Gradient and Directional Derivative

The gradient of a function $w = f(x, y, z)$ is the vector function:

$$\nabla f = \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right\rangle$$

For a function of two variables $z = f(x, y)$, the gradient is the two-dimensional vector:

$$\nabla f = \langle f_x(x, y), f_y(x, y) \rangle$$

This definition generalizes in a natural way to functions of more than three variables.

Examples

For the function $z = f(x, y) = 4x^2 + y^2$, the gradient is:

$$\nabla f = \langle 8x, 2y \rangle$$

For the function $w = g(x, y, z) = \exp(xyz) + \sin(xy)$, the gradient is

$$\text{grad } g = \langle yz\, e^{xyz} + y \cos(xy),\ xz\, e^{xyz} + x \cos(xy),\ xy\, e^{xyz} \rangle$$

Geometric Description of the Gradient Vector

There is a nice way to describe the gradient geometrically. Consider $z = f(x, y) = 4x^2 + y^2$. The surface defined by this function is an **elliptical paraboloid** — a bowl-shaped surface. The bottom of the bowl lies at the origin.

The level curves are defined by $f(x, y) = c$, i.e., the ellipses:

$$4x^2 + y^2 = c$$

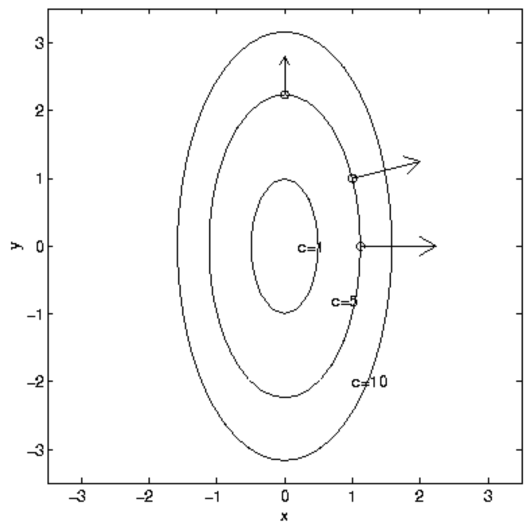


Figure: Level curves of the function $f(x,y) = 4x^2 + y^2$ ***for*** $c = 1, 5, 10$***. The arrows represent the gradient vectors*** $\nabla f = \langle 8x, 2y \rangle$ ***at selected points.***

The gradient vector $\langle 8x, 2y \rangle$ is plotted at the 3 points:

$$(\sqrt{1.25}, 0), \quad (1,1), \quad (0, \sqrt{5})$$

As the plot shows, the gradient vector at (x, y) is normal (perpendicular) to the level curve through (x, y). As we will see below, the gradient vector points in the direction of greatest rate of increase of $f(x, y)$.

In three dimensions, the level curves become **level surfaces**. Again, the gradient vector at (x, y, z) is normal to the level surface through that point.

Directional Derivatives

For a function $z = f(x, y)$:

- The partial derivative with respect to x gives the rate of change of f in the x direction.
- The partial derivative with respect to y gives the rate of change of f in the y direction.

How do we compute the rate of change of f in an arbitrary direction?

The rate of change of a function of several variables in the direction $\mathbf{u}$ is called the **directional derivative** in the direction $\mathbf{u}$. Here, $\mathbf{u}$ is assumed to be a unit vector. Assuming $w = f(x, y, z)$ and $\mathbf{u} = \langle u_1, u_2, u_3 \rangle$, we have:

$$D_{\mathbf{u}} f = \nabla f \cdot \mathbf{u} = \frac{\partial f}{\partial x} u_1 + \frac{\partial f}{\partial y} u_2 + \frac{\partial f}{\partial z} u_3$$

Hence, the directional derivative is the dot product of the gradient and the vector $\mathbf{u}$. Note that if $\mathbf{u}$ is a unit vector in the x direction, $\mathbf{u} = \langle 1, 0, 0 \rangle$, then the directional derivative is simply the partial derivative with respect to x. For a general direction, the directional derivative is a combination of all three partial derivatives.

Example

What is the directional derivative in the direction $\langle 2, 1\rangle$ of the function $z = f(x, y) = 4x^2 + y^2$ at the point $x = 1$, $y = 1$?

- The gradient is $\nabla f = \langle 8x, 2y\rangle = \langle 8, 2\rangle$ at $(1, 1)$.
- The direction vector is $\langle 2, 1\rangle$.
- Convert this to a unit vector:

$$\mathbf{u} = \frac{1}{\sqrt{5}}\langle 2, 1\rangle$$

- Compute the directional derivative:

$$D_{\mathbf{u}} f = \nabla f \cdot \mathbf{u} = \langle 8, 2\rangle \cdot \left\langle \frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \right\rangle = \frac{16 + 2}{\sqrt{5}} = \frac{18}{\sqrt{5}}$$

Directions of Greatest Increase and Decrease

The directional derivative can also be written as:

$$D_{\mathbf{u}} f = \nabla f \cdot \mathbf{u} = |\nabla f||\mathbf{u}| \cos\theta$$

where θ is the angle between the gradient vector and $\mathbf{u}$.

The directional derivative takes on its greatest positive value if $\theta = 0$. Hence, the direction of greatest increase of f is the same as the direction of the gradient vector.

The directional derivative takes on its greatest negative value if $\theta = \pi$ (or 180 degrees). Hence, the direction of greatest decrease of f is the direction opposite to the gradient vector.

References

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.

[Gee24] GeeksforGeeks, *Partial derivatives in machine learning*, April 2024, Accessed: March 23, 2025.

[Sch17] René L Schilling, *Measures, integrals and martingales*, Cambridge University Press, 2017.

[Ste12] James Stewart, *Calculus: early transcendentals*, Cengage learning, 2012.

[Tha23] Published by Thatsmaths, *Vertical or horizontal slices? riemann and lebesgue integration*, June 2023.