Derivatives and optimization

AI503
Shan Shan

Lecture 6

Definition of Partial Derivatives

Partial derivatives of f at (a, b)

For all points at which the limit exist, we define the partial derivatives at the point (a, b) by

$$f_x(a, b) = \lim_{h \to 0} \frac{f(a + h, b) - f(a, b)}{h}$$
$$f_y(a, b) = \lim_{h \to 0} \frac{f(a, b + h) - f(a, b)}{h}$$

- Functions of (x, y): $f_x(x, y)$, $f_y(x, y)$.
- Alternative notation: $\frac{\partial z}{\partial x}$, $\frac{\partial z}{\partial y}$.
- Measures the rate of change in x or y direction.

Partial Derivatives in n Variables

Partial derivatives of f at $(x_1, \dots, x_n)$

For all points at which the limit exist, we define the partial derivatives at the point $(x_1, \dots, x_n)$ by

$$\frac{\partial f}{\partial x_i}(x_1, \dots, x_n) = \lim_{h \to 0} \frac{f(x_1, \dots, x_i + h, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)}{h}$$

- Compute by treating all other variables as constants.
- Example:

$$f(x, y, z) = x^2 yz$$

$$\frac{\partial f}{\partial x} = 2xyz, \quad \frac{\partial f}{\partial y} = x^2 z, \quad \frac{\partial f}{\partial z} = x^2 y$$

Local linearity

Tangent hyperplane approximation/ First order approximation

For $\mathbf{x} = (x_1, \ldots, x_n)$ near $\mathbf{a} = (a_1, \ldots, a_n)$:

$$f(\mathbf{x}) \approx L(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a})(x_i - a_i)$$

Error behavior: If f has continuous second derivatives, then

$$f(\mathbf{x}) - L(\mathbf{x}) = O\left(\|\mathbf{x} - \mathbf{a}\|^2\right).$$

Definition: Directional Derivative

Directional Derivative

The **directional derivative** of f at (a, b) in the direction of a *unit vector* $\mathbf{u} = u_1\mathbf{i} + u_2\mathbf{j}$ is

$$f_{\mathbf{u}}(a, b) = \lim_{h \to 0} \frac{f(a + hu_1, b + hu_2) - f(a, b)}{h},$$

provided the limit exists.

Questions

Calculate the directional derivative of $f(x, y) = x^2 + y^2$ at $(1, 0)$ in the direction of $\mathbf{i} + \mathbf{j}$.

Definition: Directional Derivative in n Variables

Directional Derivative

Let $f : \mathbb{R}^n \to \mathbb{R}$ and let $\mathbf{a} = (a_1, \ldots, a_n)$. The **directional derivative** of f at $\mathbf{a}$ in the direction of a *unit vector* $\mathbf{u} = (u_1, \ldots, u_n)$ is

$$f_{\mathbf{u}}(\mathbf{a}) = \lim_{h \to 0} \frac{f(a_1 + hu_1, \ldots, a_n + hu_n) - f(a_1, \ldots, a_n)}{h},$$

provided the limit exists.

Computing Directional Derivatives from Partial Derivatives

- Approximate $f(a_1 + hu_1, \ldots, a_n + hu_n)$ with the first-order approximation at $(hu_1, \ldots, hu_n)$
- What happens when you simplify the limit in the previous definition?

Gradient Vector in n Variables

Gradient

Let $f : U \subset \mathbb{R}^n \to \mathbb{R}$ be differentiable at $\mathbf{a} = (a_1, \dots, a_n)$. The **gradient** of f at $\mathbf{a}$ is

$$\nabla f(\mathbf{a}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{a}), \frac{\partial f}{\partial x_2}(\mathbf{a}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{a}) \right).$$

Directional Derivative via Gradient

If $\mathbf{u} = (u_1, \dots, u_n)$ is a unit vector, the directional derivative of f at $\mathbf{a}$ in the direction of $\mathbf{u}$ is

$$f_{\mathbf{u}}(\mathbf{a}) = \nabla f(\mathbf{a}) \cdot \mathbf{u} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a}) u_i.$$

Alternative notation

$$\nabla f = \text{grad} f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right).$$

Example

Find the gradient of $f(x, y) = x + e^y$ at $(1, 1)$. Use it to compute the directional derivative in the direction of $\mathbf{i} + \mathbf{j}$.

Question

We computed the directional derivative by the inner product with the gradient vector

$$f_{\mathbf{u}}(\mathbf{a}) = \nabla f(\mathbf{a}) \cdot \mathbf{u}$$

Recall that the inner product measures the alignment of two vectors:

$$\nabla f(\mathbf{a}) \cdot \mathbf{u} = \|\nabla f(\mathbf{a})\| \, \|\mathbf{u}\| \cos\theta = \|\nabla f(\mathbf{a})\| \cos\theta$$

Think, what does this tell us about the gradient vector?

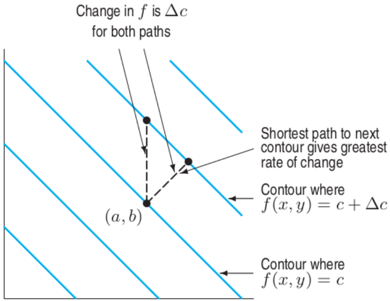
Properties of the gradient vector I

Direction of Fastest Increase

Assume $\nabla f(\mathbf{x}) \neq 0$. Then $\nabla f(\mathbf{x})$ points in the direction along which f is increasing the fastest.

Question

In the following picture, mark the gradient vector of f at (a, b).



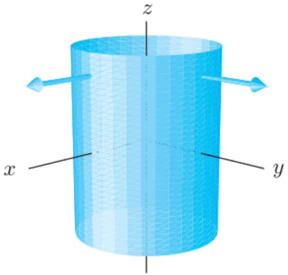
Properties of the gradient vector II

Normal to the level set

If f is reasonably well behaved, the gradient and the level set will be perpendicular.

Example

Let $f(x, y, z) = x^2 + y^2$. The following picture draws the level surface of $f(x, y, z) = 1$. Note how the gradient of f at the point (0,1,1) and (1,0,1) are perpendicular to the level surface.



Application to optimization problems

Definition

Given a real-valued function $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$, the general problem of finding the value that minimizes f is formulated as follows.

$$\min_{x \in \Omega} f(x).$$

In this context, f is the objective function (sometimes referred to as loss function or cost function).

Gradient descent

Given an initial point $\mathbf{x}_0$, find iterates $\mathbf{x}_{n+1}$ recursively using

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma \nabla f(\mathbf{x}_n)$$

for some $\gamma > 0$. The parameter γ is called the step length or the learning rate.

Demo

https://fa.bianp.net/teaching/2018/eecs227at/gradient_descent.html

Example

Let's consider a simple classification example where we need to classify whether a student passes an exam based on two features:

1. x_1 = Hours studied
2. x_2 = Hours slept

The goal is to make a binary classification such that

1. $y = 1$ if the student passes (positive class).
2. $y = 0$ if the student fails (negative class)

Example (continued)

1. Suppose a data set $\{x_1^{(i)}, x_2^{(i)}, y^{(i)}\}$ with $i = 1, \ldots, m$ is given.
2. The logistic regression model is defined by

$$f(x_1, x_2) = \frac{1}{1 + e^{-(\theta_1 x_1 + \theta_2 x_2 + b)}}$$

   where θ_1 and θ_2 are parameters (weights) for the features. and b is some bias term.
3. The predicted outcome using the logistic model is

$$y^{(i)} = f(x_1^{(i)}, x_2^{(i)}).$$

Think: What is the range of f?

Example (continued)

The amount of error we made is characterized by the following function. Let us define

$$L(\theta_1, \theta_2, b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right]$$

What is the gradient vector? HW: Check

$$\frac{\partial L}{\partial \theta_1} = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})x_1^{(i)}$$

$$\frac{\partial L}{\partial \theta_2} = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})x_2^{(i)}$$

$$\frac{\partial L}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})$$

Second-order Partial Derivatives (n variables)

Since the partial derivatives of a function are themselves functions, we can differentiate them, giving second-order partial derivatives.

A function $f(x_1, \ldots, x_n)$ has n first-order partial derivatives

$$\frac{\partial f}{\partial x_i}, \quad i = 1, \ldots, n$$

How many second-order partial derivatives does it have?

Second-order Partial Derivatives (n variables)

Second-Order Partial Derivatives of $f(x_1, \ldots, x_n)$

$$\frac{\partial^2 f}{\partial x_i^2}, \quad \frac{\partial^2 f}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n, \; i \neq j$$

Total: $n^2 = n$ pure second derivatives $+\ n(n-1)$ mixed derivatives.

Example

Compute the second-order partial derivatives of

$$f(x_1, x_2) = x_1 x_2^2 + 3x_1^2 e^{x_2}.$$

Equality of mixed partial derivatives

Observe in the previous example $\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1}$. This is not an accident! In general

Equality of Mixed Partial Derivatives

If all mixed partial derivatives are continuous, then

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

Local Extrema

Definition (Local Extrema)

- f has a local maximum at P_0 if $f(P_0) \geq f(P)$ for all P near P_0.
- f has a local minimum at P_0 if $f(P_0) \leq f(P)$ for all P near P_0.

Local extrema can only occur at *critical points* or at the boundary of the function domain.

Definition (Critical Points)

Points where $\nabla f = \mathbf{0}$ or undefined.

Second Derivative Test for n variables

At a critical point P_0, let H be the Hessian matrix:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

Compute eigenvalues of H.

- All positive eigenvalues $\implies$ local minimum
- All negative eigenvalues $\implies$ local maximum
- Mixed signs $\implies$ saddle point
- Zero eigenvalue(s) $\implies$ test inconclusive

Example

Find and analyze the critical points of $f(x, y) = x^2 - 2x + y^2 - 4y + 5$.

Example

Find and analyze any critical points of $f(x, y) = -\sqrt{x^2 + y^2}$.

Example

Find and analyze any critical points of $f(x, y) = x^2 - y^2$.

Example

Classify the critical points of $f(x, y) = x^4 + y^4$, and $g(x, y) = -x^4 - y^4$ and $h(x, y) = x^4 - y^4$.

Global Extrema

Definition (Global Extrema)

For f defined on $R \subset \mathbb{R}^n$:

- f has a global maximum at P_0 if $f(P_0) \geq f(P)$ for all $P \in R$.
- f has a global minimum at P_0 if $f(P_0) \leq f(P)$ for all $P \in R$.

Global extrema occur either at critical points or on the boundary of R.

Extreme Value Theorem (Multivariable)

Theorem

If f is continuous on a closed and bounded region R, then f attains both a global maximum and minimum somewhere in R.

- Closed region: contains its boundary
- Bounded region: does not stretch to infinity