Derivatives and optimization

AI503
Shan Shan

Lecture 7

Gradient Vector in n Variables

Gradient

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

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

Directional Derivative via Gradient

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

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

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.

Normal to the level set

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

Gradient descent

The minimization problem of an arbitrary function f whose gradient exists everywhere on the domain can be solved by the following method.

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.

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

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

The Hessian Matrix

- **Definition:** For $f : \mathbb{R}^n \to \mathbb{R}$, the Hessian is the matrix of second-order partial derivatives:

$$H(f)(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- **Properties:**
  - Symmetric (think why?)
  - Describes local curvature of f.
- **Applications:**
  - Optimization: minima, maxima, saddle points.
  - Machine learning: Newton's method, second-order methods.
  - Physics: stability of equilibria.

Why Does the Hessian Describe Curvature?

- **1D case:** Second derivative measures curvature:

$$
\begin{aligned}
f''(x) > 0 &\Rightarrow \text{ curves upward (convex)}, \\
f''(x) < 0 &\Rightarrow \text{ curves downward (concave)}.
\end{aligned}
$$

- **Multivariate case:** The Hessian generalizes this idea, and it gives the second derivative of f along direction v.

$$
\begin{aligned}
D_v^2 f(x) = D_v(D_v f(x)) = D_v\left(\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} v_i\right) &= \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j} v_i v_j \\
&= v^\top H(f)(x) v
\end{aligned}
$$

- Interpretation:
  - Positive in all directions $\Rightarrow$ local minimum.
  - Negative in all directions $\Rightarrow$ local maximum.
  - Mixed signs $\Rightarrow$ saddle point.

Application to optimization

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 $H = H(f)(P_0)$ be the Hessian matrix:
Use the quadratic form $v^\top Hv$ to test curvature along all directions v:

- $v^\top Hv > 0$ for all $v \neq 0$ (H is positive definite) $\implies$ local minimum
- $v^\top Hv < 0$ for all $v \neq 0$ (H is negative definite) $\implies$ local maximum
- $v^\top Hv$ can be positive in some directions and negative in others (H is indefinite) $\implies$ saddle point
- $v^\top Hv = 0$ in some nonzero direction (H is semidefinite) $\implies$ test inconclusive

Positive Definiteness and Eigenvalues

Definition

Let $A \in \mathbb{R}^{n \times n}$. A scalar $\lambda \in \mathbb{C}$ is called an **eigenvalue** of A if there exists a nonzero vector $v \in \mathbb{C}^n$ such that

$$Av = \lambda v.$$

The vector v is called an **eigenvector** corresponding to λ.

- Equivalently, λ is an eigenvalue if it satisfies the **characteristic equation**:

$$\det(A - \lambda I) = 0.$$

Positive Definiteness and Eigenvalues

Theorem (Characterization via Eigenvalues)

Let $H \in \mathbb{R}^{n \times n}$ be a symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then:

- *H is* **positive definite** $\iff \lambda_i > 0$ *for all i*
- *H is* **negative definite** $\iff \lambda_i < 0$ *for all i*
- *H is* **indefinite** $\iff$ *it has both positive and negative eigenvalues*
- *H is* **semidefinite** $\iff$ *all eigenvalues are nonnegative (positive semidefinite) or nonpositive (negative semidefinite)*

Second Derivative Test for n variables

At a critical point P_0, let $H = H(f)(P_0)$ be the Hessian matrix: 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.

Question

Do all functions have global extrema? Can you think of a counterexample?

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

Question

Does the function $f = 1/(x^2 + y^2)$ have global maxima and minima on the region R given by $0 < x^2 + y^2 \leq 1$?

Question

Does the function $f = x^2y^2$ have global maxima and minima in the xy-plane?

Chain rule

Theorem

If f, g, and h are differentiable and if $z = f(x, y)$, and $x = g(t)$, and $y = h(t)$, then

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}.$$

Chain rule in a diagram

The following diagram keeps track of how a change in t propagates through the chain of composed functions.

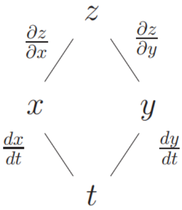

Figure: There are two paths from t to z, one through x and one through y. For each path, we multiply together the derivatives along the path. Then, to calculate dz/dt, we add the contributions from the two paths.

Questions

Suppose that $z = f(x, y) = x \sin y$, where $x = t^2$ and $y = 2t + 1$. Let $z = g(t)$. Compute $g'(t)$ directly and using the chain rule.

Questions

If f, g, h are differentiable and if $z = f(x, y)$, with $x = g(u, v)$ and $y = h(u, v)$. What is $\frac{\partial z}{\partial u}$ and $\frac{\partial z}{\partial v}$?

Chain rule in multivariable

Theorem (Multivariable Chain Rule)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, and let each x_i be a differentiable function of $t_1, \ldots, t_m$:

$$x_i = x_i(t_1, \ldots, t_m), \quad i = 1, \ldots, n.$$

Define

$$z = f(x_1, \ldots, x_n).$$

Then z is differentiable with respect to t_j, and

$$\frac{\partial z}{\partial t_j} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial t_j}, \quad j = 1, \ldots, m.$$