Introduction

Oscar Dalmau
dalmau@cimat.mx

Centro de Investigación en Matemáticas
CIMAT A.C. Mexico

January 2018

Outline

1. Introduction
2. Level sets and gradient
3. Taylor Series

Optimization Problem

- **Optimization Problem**:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \qquad f(\boldsymbol{x})$$

where $f(\cdot)$ is class $\mathcal{C}^1$ or continuously differentiable that consists of all differentiable functions whose derivative is continuous or $f()$ could be class $\mathcal{C}^2$

Optimization Problem

- **Optimization Problem**:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x})$$

- **Examples**: $\mathbf{x} = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$,

$$\min_{\boldsymbol{x}} \frac{\boldsymbol{x}^T \mathbf{A} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}, \ \mathbf{A} \text{ is simmetric}$$

$$\min_{\boldsymbol{x}} \sum_{i=1}^{n} (x_i - y_i)^2 + \lambda \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2$$

$$\min_{\boldsymbol{x}} \sum_{i=1}^{N-1} 100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2$$

Notation

$$
\begin{aligned}
\mathbf{1} &= [1, 1, \cdots, 1]^T \\
\mathbf{0} &= [0, 0, \cdots, 0]^T \\
\mathbf{I} &= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \\
f(\boldsymbol{x}) &: \mathbb{R}^n \rightarrow \mathbb{R}^m \\
f(\boldsymbol{x}) &= [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x})]^T
\end{aligned}
$$

Definitions

- A real $n \times n$ matrix $\mathbf{A}$ is said to be positive definite if the scalar $\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}$ is positive for every non-zero column vector $\boldsymbol{x}$
- A real $n \times n$ matrix $\mathbf{A}$ is said to be the **negative definite**, **positive semi-definite**, and **negative semi-definite** matrices are defined in the same way, except that the expression $\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}$ is required to be always negative, non-negative, and non-positive, respectively.
- We can determine that a **symmetric matrix** is **positive definite** by computing its eigenvalues and verifying that they are all positive, or by performing a Cholesky factorization (Cholesky factorization gives error for non-positive-definite matrices)

Some theorems

Lema: (Sign Preserving Property) Let f be continuous at a and $f(a) \neq 0$. Then there is an interval $(a-\delta, a+\delta)$ about a in which f has the same sign as $f(a)$.
Proof: Without loss of generality, assume that $f(a) > 0$. Using the continuity of f. For all $\epsilon > 0$ there exist $\delta > 0$ such that

$$|x-a| < \delta \;\Rightarrow\; |f(x) - f(a)| < \epsilon$$

Then $f(x) > f(a) - \epsilon$ and taking $0 < \epsilon < f(a)$ (for example, $\epsilon = \frac{f(a)}{2}$) one obtains

$$f(x) > f(a) - \epsilon > 0$$

for $|x-a| < \delta$, ie, $x \in (a-\delta, a+\delta)$

Some theorems

Theorem: (Mean value theorem) Let $f : [a, b] \to \mathbb{R}$ be a continuous function on the closed interval $[a, b]$, and differentiable on the open interval (a, b), where $a < b$. Then there exists some $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

The *mean value theorem* is a generalization of *Rolle's theorem*, which assumes $f(a) = f(b)$, so that the right-hand side above is zero.

Norms

For a vector $\boldsymbol{x} \in \mathbb{R}^n$, we have the following norms:

$$\|\boldsymbol{x}\|_1 \stackrel{def}{=} \sum_{i=1}^{n} |x_i| \tag{1}$$

$$\|\boldsymbol{x}\|_2 \stackrel{def}{=} \sqrt{\sum_{i=1}^{n} x_i^2} \tag{2}$$

$$\|\boldsymbol{x}\|_\infty \stackrel{def}{=} \max_{i=1,2,\ldots n} |x_i| \tag{3}$$

The norm $\|\cdot\|_2$ is often called the *Euclidean norm*, $\|\cdot\|_1$ is called the ℓ_1-norm and $\|\cdot\|_\infty$ is called the ℓ_∞-norm.

Norms

In general, a norm is any mapping $\|\cdot\|$ from $\mathbb{R}^n$ to the nonnegative real numbers that satisfies the following properties:

$$\begin{aligned} \|\boldsymbol{x}\| &= 0 \Leftrightarrow \boldsymbol{x} = 0;\ \boldsymbol{x} \in \mathbb{R}^n \\ \|\alpha\boldsymbol{x}\| &= |\alpha|\|\boldsymbol{x}\|;\ \boldsymbol{x} \in \mathbb{R}^n;\ \alpha \in \mathbb{R} \\ \|\boldsymbol{x} + \boldsymbol{y}\| &\leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|;\ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n \end{aligned}$$

Cauchy-Schwarz inequality

A property that holds for the Euclidean norm is the Cauchy-Schwarz inequality, which states that:

$$|\boldsymbol{x}^T \boldsymbol{y}| \quad \leq \quad \|\boldsymbol{x}\| \|\boldsymbol{y}\|$$

and the equality holds if and only if one of these vectors is a multiple of the other, ie, the vectors are parallels.

Matrix norms

Let $\|\cdot\|$ be generic notation for the three norms listed above, we define the corresponding matrix norm as:

$$\|\mathbf{A}\| \stackrel{def}{=} sup_{\boldsymbol{x}\neq 0}\frac{\|\mathbf{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$$

Matrix norms

Explicit formulas

$$\begin{aligned}
\|\mathbf{A}\|_1 &= \max_{j=1,2,\ldots,m} \sum_{i=1}^{n} |A_{ij}| \\
\|\mathbf{A}\|_\infty &= \max_{i=1,2,\ldots,n} \sum_{j=1}^{m} |A_{ij}| \\
\|\mathbf{A}\|_2 &= \sqrt{\lambda_{\mathsf{max}}(A^T A)} = \sigma_{\mathsf{max}}(A)
\end{aligned}$$

Matrix norms

- The Frobenius norm is defined as follows

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j}^{n,m} A_{ij}^2}$$

- Inner product $\langle A, B\rangle = \mathrm{tr}(A^T B)$

$$\begin{aligned}\|\mathbf{A}\|_F^2 &= \langle A, A\rangle = \mathrm{tr}(A^T A) = \sum_j (A^T A)_{jj} \\ &= \sum_j \sum_i A_{ji}^T A_{ij} = \sum_i \sum_j A_{ij} A_{ij} = \sum_i \sum_j A_{ij}^2\end{aligned}$$

Condition number of a Matrix

The condition number of a nonsingular matrix is defined as

$$\kappa(\mathbf{A}) \quad = \quad \|\mathbf{A}\|\|\mathbf{A}^{-1}\|,$$

where any matrix norm can be used in the definition. We can use a subscript for the different norms, ie, $\kappa_1(\cdot)$, $\kappa_2(\cdot)$, and $\kappa_\infty(\cdot)$ respectively. We use κ to denote $\kappa_2(\cdot)$.

Definitions

In many optimization methods the information of the first and/or second derivative of $f(\cdot)$ is required

- **Gradient**: If $f : \mathbb{R}^n \to \mathbb{R}$ such that $f \in \mathcal{C}^1$, i.e., f has continuous partial derivatives of first order, then the *gradient* of $f(\cdot)$ is defined as

$$\begin{aligned} \nabla f(\boldsymbol{x}) &= [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n}]^T \stackrel{def}{=} \boldsymbol{g}(\boldsymbol{x}) \\ Df(\boldsymbol{x}) &= \nabla f(\boldsymbol{x})^T \end{aligned}$$

Definitions

In many optimization methods the information of the first and/or second derivative of $f(\cdot)$ is required

- **Hessian**: If $f \in \mathcal{C}^2$, i.e., f has continuous partial derivatives of second order, then the *Hessian* of $f(\cdot)$ is defined as

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

$\mathbf{H}(\boldsymbol{x})$ is a symmetric square matrix, ie, $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$. For the *equality of mixed partial derivatives Theorem*, due to the mixed partial derivatives exist and are continuous.

Definitions

In many optimization methods the information of the first and/or second derivative of $f(\cdot)$ is required

- The function $f(\cdot)$, the gradient $\boldsymbol{g}(\cdot)$ and the Hessian $\mathbf{H}(\cdot)$ at $\boldsymbol{x} = \boldsymbol{x}_k$ are denoted as f_k, $\boldsymbol{g}_k$ and $\mathbf{H}_k$ respectively in order to simplify notation, i.e.,

$$\begin{aligned} f_k &\overset{def}{=} f(\boldsymbol{x}_k) \\ \boldsymbol{g}_k &\overset{def}{=} \boldsymbol{g}(\boldsymbol{x}_k) \\ \mathbf{H}_k &\overset{def}{=} \mathbf{H}(\boldsymbol{x}_k) \end{aligned}$$

Definitions

Directional derivative: Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is defined on an open ball about a point $\boldsymbol{x}_0$. Given a unit vector $\boldsymbol{v}$, we call

$$D_{\boldsymbol{v}} f(\boldsymbol{x}_0) = \lim_{h \to 0} \frac{f(\boldsymbol{x}_0 + h\boldsymbol{v}) - f(\boldsymbol{x}_0)}{h},$$

provided the limit exists, the **directional derivative** of f in the direction of $\boldsymbol{v}$ at $\boldsymbol{x}_0$. For example

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}_0) = D_{e_i} f(\boldsymbol{x}_0)$$

where $e_i = [0, 0, ..., 1, ..., 0]^T$.

Differentiation Rules

- **Chain rule**: Let $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^k$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable functions at $\boldsymbol{x} \in \mathbb{R}^n$ then

$$D_{\boldsymbol{x}}(\boldsymbol{f} \circ \boldsymbol{g}) \quad = \quad D\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x}))D\boldsymbol{g}(\boldsymbol{x})$$

  where $D\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x})) \in \mathbb{R}^{k\times m}$, $D\boldsymbol{g}(\boldsymbol{x})\mathbb{R}^{m\times n}$ and $D_{\boldsymbol{x}}(\boldsymbol{f} \circ \boldsymbol{g}) \in \mathbb{R}^{k\times n}$ are matrices and $(\boldsymbol{f} \circ \boldsymbol{g}) : \mathbb{R}^n \to \mathbb{R}^k$.

Differentiation Rules

- **Example**: $\boldsymbol{f}(x,y) = [x^2 + y, x + y^2]^T$, $g(t) = [t^2, t]^T$.
  Compute $D_t(\boldsymbol{f} \circ \boldsymbol{g})$.
  Solution 1: Using the chain rule

$$\begin{aligned} D\boldsymbol{f}(x,y) &= \begin{bmatrix} 2x & 1 \\ 1 & 2y \end{bmatrix} \\ Dg(t) &= [2t, 1]^T \\ D\boldsymbol{f}(g(t)) &= \begin{bmatrix} 2t^2 & 1 \\ 1 & 2t \end{bmatrix} \\ D\boldsymbol{f}(g(t))Dg(t) &= \begin{bmatrix} 4t^3 + 1 \\ 4t \end{bmatrix} \end{aligned}$$

Differentiation Rules

- **Example**: $\boldsymbol{f}(x,y) = [x^2 + y, x + y^2]^T$, $g(t) = [t^2, t]^T$.
  Compute $D_t(\boldsymbol{f} \circ \boldsymbol{g})$.
  Solution 2: Computing $(\boldsymbol{f} \circ \boldsymbol{g})(t)$

$$h(t) = (\boldsymbol{f} \circ \boldsymbol{g})(t) = \begin{bmatrix} t^4 + t \\ 2t^2 \end{bmatrix}$$

$$h'(t) = \begin{bmatrix} 4t^3 + 1 \\ 4t \end{bmatrix}$$

- **Chain rule** (case 1): Let $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{z} : \mathbb{R} \to \mathbb{R}^n$ be differentiable functions at $t \in \mathbb{R}$ then

$$D_t(f \circ \boldsymbol{z}) \quad = \quad Df(\boldsymbol{z}(t))D\boldsymbol{z}(t)$$

Note that:

$$\begin{aligned} f \circ \boldsymbol{z} : \mathbb{R} &\to \mathbb{R} \\ Df(\boldsymbol{z}(t)) &= \nabla f(\boldsymbol{z}(t))^T \\ D\boldsymbol{z}(t) &= [z_1'(t), z_2'(t), \cdots, z_n'(t)]^T \\ D_t(f \circ \boldsymbol{z}) &= \nabla f(\boldsymbol{z}(t))^T D\boldsymbol{z}(t) = \langle \nabla f(\boldsymbol{z}(t)), D\boldsymbol{z}(t) \rangle \in \mathbb{R} \end{aligned}$$

- **Chain rule** (case 1 Example): $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$, $\boldsymbol{z}(t) = \boldsymbol{x}_0 + t\boldsymbol{v}$.
  Compute $D_t(f \circ \boldsymbol{z})$.
  Solution:

$$\begin{aligned} Df(\boldsymbol{x}) &= \nabla f(\boldsymbol{x})^T = 2\boldsymbol{x}^T \\ D\boldsymbol{z}(t) &= [z_1'(t), z_2'(t), \cdots, z_n'(t)]^T = \boldsymbol{v} \end{aligned}$$

  Then $D_t(f \circ \boldsymbol{z}) = 2\boldsymbol{z}(t)^T\boldsymbol{v} = 2(\boldsymbol{x}_0 + t\boldsymbol{v})^T\boldsymbol{v}$

- **Product rule**: Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be two differentiable functions. Define the function $h : \mathbb{R}^n \to \mathbb{R}$ by $h(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^T \boldsymbol{g}(\boldsymbol{x})$ then

$$D_{\boldsymbol{x}} h(\boldsymbol{x}) \quad = \quad \boldsymbol{f}(\boldsymbol{x})^T D\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^T D\boldsymbol{f}(\boldsymbol{x})$$

  where $D\boldsymbol{g}(\boldsymbol{x})$ and $D\boldsymbol{f}(\boldsymbol{x})$ are matrices (the Jacobian matrix).

- **Product rule** (Example): $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$, $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$. Compute $D_{\boldsymbol{x}} h(\boldsymbol{x})$ with $h(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^T \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}^T \mathbf{A}\boldsymbol{x}$.
  Solution:

$$\begin{aligned} D\boldsymbol{f}(\boldsymbol{x}) &= \mathbf{I} \\ D\boldsymbol{g}(\boldsymbol{x}) &= \mathbf{A} \end{aligned}$$

  Then

$$\begin{aligned} D_{\boldsymbol{x}} h(\boldsymbol{x}) &= \boldsymbol{f}(\boldsymbol{x})^T D\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^T D\boldsymbol{f}(\boldsymbol{x}) \\ &= \boldsymbol{x}^T \mathbf{A} + \boldsymbol{x}^T \mathbf{A}^T \mathbf{I} = \boldsymbol{x}^T (\mathbf{A} + \mathbf{A}^T) \end{aligned}$$

  and $\nabla h(\boldsymbol{x}) = (\mathbf{A} + \mathbf{A}^T)\boldsymbol{x}$.

Note: If $\mathbf{A}^T = \mathbf{A}$ then $\nabla h(\boldsymbol{x}) = 2\mathbf{A}\boldsymbol{x}$ and $\nabla^2 h(\boldsymbol{x}) = 2\mathbf{A}$.

Level sets

The level set of a function $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ at level l is the set of points

$$S \quad = \quad \{\boldsymbol{x} : \ f(\boldsymbol{x}) = l\}$$

- If $f : \ \mathbb{R}^2 \to \mathbb{R}$ then S is a curve
- If $f : \ \mathbb{R}^3 \to \mathbb{R}$ then S is a surface

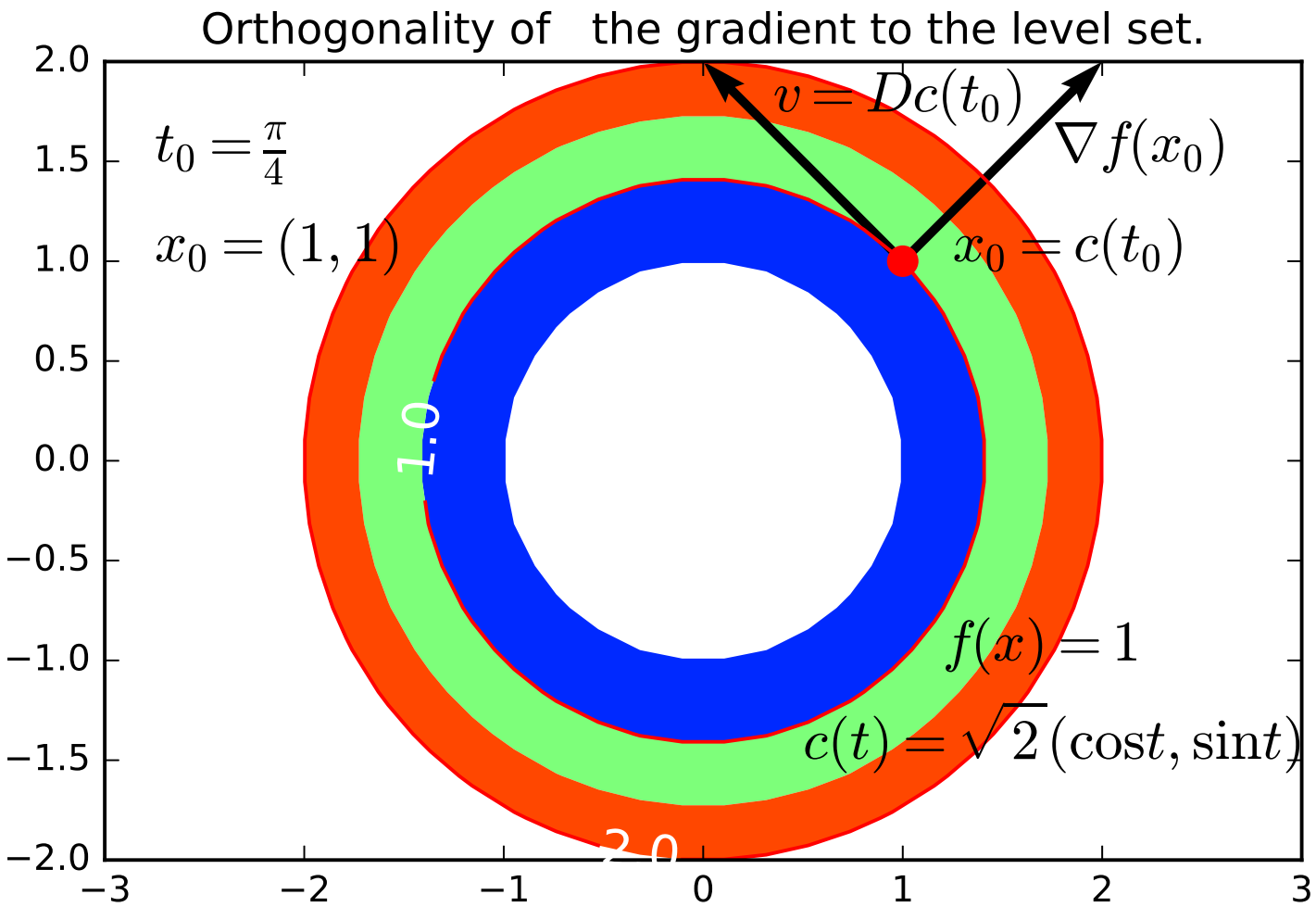
Level sets

- Let $\boldsymbol{x}_0$ a point in the level set S, i.e., $f(\boldsymbol{x}_0) = l$
- Let $\boldsymbol{c}(t)$ be a parametrization of a curve γ that lies on S, such that

$$\begin{aligned} \boldsymbol{c}(t) &: \mathbb{R} \to \mathbb{R}^n \\ \boldsymbol{c}(t_0) &= \boldsymbol{x}_0 \\ D\boldsymbol{c}(t_0) &= \boldsymbol{v} \neq \mathbf{0} \end{aligned}$$

  where $\boldsymbol{v}$ is the tangent vector to γ at $\boldsymbol{x}_0$
- Example: $f(\boldsymbol{x}) = \frac{1}{2}(x_1^2 + x_2^2)$ and the level set $f(\boldsymbol{x}) = 1$, ie, $x_1^2 + x_2^2 = 2$. A parametrization could be $c(t) = [x_1(t), x_2(t)]^T = \sqrt{2}[\cos t, \sin t]^T$, see next slide.

Level sets and gradient



Theorem: The vector $\nabla f(\boldsymbol{x}_0)$ is orthogonal to the tangent vector to an arbitrary smooth curve passing through $\boldsymbol{x}_0$ on the level set determined by $f(\boldsymbol{x}) = f(\boldsymbol{x}_0)$.
Proof: (It is straightforward)
Applying the chain rule to the function $h(t) = f(\boldsymbol{c}(t))$

$$\begin{aligned} h'(t_0) &= Df(\boldsymbol{c}(t_0))D\boldsymbol{c}(t_0) = Df(\boldsymbol{x}_0)\boldsymbol{v} = \nabla f(\boldsymbol{x}_0)^T\boldsymbol{v} \\ h(t) &= f(\boldsymbol{c}(t)) = \text{constant} \\ h'(t) &= 0 \end{aligned}$$

then $\nabla f(\boldsymbol{x}_0)^T\boldsymbol{v} = 0$, ie, $\nabla f(\boldsymbol{x}_0) \perp \boldsymbol{v}$

Comments

- It is said that $\nabla f(\boldsymbol{x}_0)$ is orthogonal or normal to the level set S corresponding to $\boldsymbol{x}_0$
- The tangent plane to S at $\boldsymbol{x}_0$ the set of all points $\boldsymbol{x}$ satisfying

$$\nabla f(\boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{x}_0) = 0$$

if $\nabla f(\boldsymbol{x}_0) \neq \mathbf{0}$.

Comments

Theorem: Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^1$ on an open ball containing the point $\boldsymbol{x}_0$. Then for any unit vector $\boldsymbol{v}$, $D_{\boldsymbol{v}} f(\boldsymbol{x}_0)$ exists and $D_{\boldsymbol{v}} f(\boldsymbol{x}_0) = Df(\boldsymbol{x}_0)\boldsymbol{v} = \nabla f(\boldsymbol{x}_0)^T \boldsymbol{v}$.
Proof: Note that

$$D_{\boldsymbol{v}} f(\mathbf{x}_0) = \left[\frac{d}{d\alpha} f(z(\alpha))\right]_{\alpha=0} = \left[\frac{d}{d\alpha} f(\mathbf{x}_0 + \alpha\ \boldsymbol{v})\right]_{\alpha=0}$$

with $z(\alpha) = \mathbf{x}_0 + \alpha\ \boldsymbol{v}$. Using the Chain rule

$$D_{\boldsymbol{v}} f(\mathbf{x}_0) = \left[Df(z(\alpha))z'(\alpha)\right]_{\alpha=0} = Df(z(0))\boldsymbol{v} = Df(\mathbf{x}_0)\boldsymbol{v}$$

Comments

Theorem: Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^1$ on an open ball containing the point $\boldsymbol{x}_0$. Then for any unit vector $\boldsymbol{v}$, $D_{\boldsymbol{v}} f(\boldsymbol{x}_0)$ exists and $D_{\boldsymbol{v}} f(\boldsymbol{x}_0) = Df(\boldsymbol{x}_0)\boldsymbol{v} = \nabla f(\boldsymbol{x}_0)^T \boldsymbol{v}$.
Note: Using the Cauchy-Schwarz inequality

$$\begin{aligned} |D_{\boldsymbol{v}} f(\boldsymbol{x}_0)| &= |\nabla f(\boldsymbol{x}_0)^T \boldsymbol{v}| \leq \|\nabla f(\boldsymbol{x}_0)\| \|\boldsymbol{v}\| \\ |D_{\boldsymbol{v}} f(\boldsymbol{x}_0)| &\leq \|\nabla f(\boldsymbol{x}_0)\| \end{aligned}$$

if $\nabla f(\boldsymbol{x}_0) \neq \mathbf{0}$ and taking $\boldsymbol{v} = \frac{\nabla f(\boldsymbol{x}_0)}{\|\nabla f(\boldsymbol{x}_0)\|}$

$$\begin{aligned} D_{\boldsymbol{v}} f(\boldsymbol{x}_0) &= \|\nabla f(\boldsymbol{x}_0)\| \\ D_{-\boldsymbol{v}} f(\boldsymbol{x}_0) &= -\|\nabla f(\boldsymbol{x}_0)\| \end{aligned}$$

Comments

Proposition: Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^1$ on an open ball containing the point $\boldsymbol{x}_0$. Then $D_{\boldsymbol{v}} f(\boldsymbol{x}_0)$ has a maximum value of $\|\nabla f(\boldsymbol{x}_0)\|$ when $\boldsymbol{v}$ is the direction of $\nabla f(\boldsymbol{x}_0)$ and a minimum value of $-\|\nabla f(\boldsymbol{x}_0)\|$ when $\boldsymbol{v}$ is the direction of $-\nabla f(\boldsymbol{x}_0)$.
Proof:

$$\begin{aligned} D_{\boldsymbol{v}} f(\boldsymbol{x}_0) &= \nabla f(\boldsymbol{x}_0)^T \boldsymbol{v} \\ &= \|\nabla f(\boldsymbol{x}_0)\| \|\boldsymbol{v}\| \cos \angle(\nabla f(\boldsymbol{x}_0), \boldsymbol{v}) \\ &= \|\nabla f(\boldsymbol{x}_0)\| \cos \angle(\nabla f(\boldsymbol{x}_0), \boldsymbol{v}) \end{aligned}$$

The maximum and minimum of $D_{\boldsymbol{v}} f(\boldsymbol{x}_0)$ is obtained when $\cos \angle(\nabla f(\boldsymbol{x}_0), \boldsymbol{v}) = 1$ and $\cos \angle(\nabla f(\boldsymbol{x}_0), \boldsymbol{v}) = -1$ respectively, i.e., when $\angle(\nabla f(\boldsymbol{x}_0), \boldsymbol{v}) = 0$ and $\angle(\nabla f(\boldsymbol{x}_0), \boldsymbol{v}) = \pi$

Comments

1. The gradient , $\nabla f(\cdot)$, of a real-valued differentiable function at a point is orthogonal to the level set of the function at that point.
2. The gradient vector points in the direction of the maximum rate of increase of the function and the negative of the gradient vector points in the direction of the maximum rate of decrease of the function.
3. The length of the gradient vector tells us the rate of increase in the direction of maximum increase and its negative tells us the rate of decrease in the direction of maximum decrease.

Theorem: Taylor's Theorem. Assume that a function $f : \mathbb{R} \to \mathbb{R}$ is $m+1$ times continuously differentiable (i.e., $f \in \mathcal{C}^{m+1}$) at $a \in \mathbb{R}$. Denote $h = x - a$. Then,

$$f(x) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots \frac{h^m}{m!} f^{(m)}(a) + R_{m+1},$$

(called Taylor's formula) where $f^{(i)}$ is the i-th derivative of f, and

$$R_{m+1} = \frac{h^{m+1}}{(m+1)!} f^{(m+1)}(a + \theta h), \quad \theta \in (0, 1)$$

Theorem: Taylor's Theorem. Assume that a function $f : \mathbb{R} \to \mathbb{R}$ is $m+1$ times continuously differentiable (i.e., $f \in \mathcal{C}^{m+1}$) at $a \in \mathbb{R}$. Denote $h = x - a$. Then,

$$f(x) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots \frac{h^m}{m!} f^{(m)}(a) + R_{m+1},$$

(called Taylor's formula) where $f^{(i)}$ is the i-th derivative of f, and

$$R_{m+1} = \frac{h^{m+1}}{(m+1)!} f^{(m+1)}(a + \theta h), \quad \theta \in (0, 1)$$

The remainder R_{m+1} using **Big O** and **little o** notation?

Big O and Little o Notation

- **Definition**. We say $f(x) = O(g(x))$ as $x \to a$ if there exists a constant C such that

$$|f(x)| \leq C|g(x)|$$

  in some neighborhood of a, that is, for $x \in (a - \delta, a + \delta) \setminus \{a\}$ for some $\delta > 0$.
- We say $f(x) = O(g(x))$ as $x \to \infty$ if there exist positive constants x_0 and C such that $|f(x)| \leq C|g(x)|$ for all $x > x_0$.

Big O and Little o Notation

Example: $x^3 - 2x + 1$ is $O(x^3)$ as $x \to \infty$

$$|x^3 - 2x + 1| \quad \leq \quad |x^3| + |2x| + |1|$$

for $x \geq 1$ we have

$$\begin{aligned} |x^3| &\leq 1|x^3| \\ |2x| &\leq 2|x^3| \\ |1| &\leq |x^3| \\ |x^3 - 2x + 1| &\leq 4|x^3| \end{aligned}$$

then $f(x) = O(g(x))$ as $x \to \infty$.

Big O and Little o Notation

- **Theorem 1**: If $\lim_{x \to a} \frac{f(x)}{g(x)} = L$ then $f(x) = O(g(x))$ as $x \to a$
- **Example**: $x^3 - 2x + 1$ is $O(x^3)$ as $x \to \infty$
- **By the theorem**:

$$\lim_{x \to \infty} \frac{x^3 - 2x + 1}{x^3} = 1$$

then $f(x) = O(g(x))$ as $x \to \infty$

Big O and Little o Notation

- **Definition**. We say $f(x) = o(g(x))$ as $x \to a$ if

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 0$$

- Generally, some results assume that $a = 0$ since changing $x - a$ to x is just a change of coordinates.

Big O and Little o Notation

- The previous notation is standard, but it is not good. The statement $f(x) = O(g(x))$ has a meaning, but $O(g(x)) = f(x)$ is meaningless, i.e., the equal sign $(=)$ is not symmetric, then this is an abuse of notation.
- A better option would be $f(x) \in O(g(x))$ or $f(x)$ is $O(g(x))$
- The limit a is important, for example, $\frac{1}{x} = o(1)$ as $x \to \infty$, but $\frac{1}{x} \neq o(1)$ as $x \to 0$

Big O and Little o Notation. Examples

- Big O examples

$$\begin{aligned} x &= O(x), \text{ as } x \to \infty \\ x &= O(x^2), \text{ as } x \to \infty \\ ax^n &= O(x^m),\ m \geq n, \text{ as } x \to \infty \\ ax^n &\neq O(x^m),\ m < n, \text{ as } x \to \infty \end{aligned}$$

- Little o examples

$$\begin{aligned} x^2 &= o(x), \text{ as } x \to 0 \\ x &\neq o(x^2), \text{ as } x \to 0 \\ x - \sin x &= o(x), \text{ as } x \to 0 \\ x - \sin x &= o(x^2), \text{ as } x \to 0 \end{aligned}$$

Big O and Little o Notation. Properties

1. $f(x) = O(f(x))$
2. If $f(x) = O(g(x))$, then $cf(x) = O(g(x))$ for any constant c.
3. If $f_1(x)$ and $f_2(x)$ are both $O(g(x))$, then so is $f_1(x) + f_2(x)$.
4. If $f(x) = o(g(x))$, then $f(x) = O(g(x))$.
5. If $f(x) = O(g(x))$, then $O(f(x)) + O(g(x)) = O(g(x))$.
6. If $f(x) = O(g(x))$, then $o(f(x)) + o(g(x)) = o(g(x))$.
7. If $f_1(x) = O(g(x))$ but $f_2(x) = o(g(x))$,then $f_1(x) + f_2(x) = O(g(x))$
8. If $f(x) = O(g(x))$, and $g(x) = o(h(x))$, then $f(x) = o(h(x))$.
9. Let $c \neq 0$, then $cO(g(x)) = O(g(x))$ and $co(g(x)) = o(g(x))$.

Big O and Little o Notation. Properties

1. $O(f(x))O(g(x)) = O(f(x)g(x))$
2. $o(f(x))O(g(x)) = o(f(x)g(x))$
3. if $\lim_{x\to 0} \frac{h(x)}{g(x)} = L$ then $h(x) = O(g(x))$. (We can use the definition of limit). then for any $\epsilon > 0$ there exists $\delta > 0$, such that for $0 < |x| < \delta$ it holds $|\frac{f(x)}{g(x)} - L| < \epsilon$. Using the inequality $|a| \leq |a - b| + |b|$

$$\begin{aligned} |\frac{f(x)}{g(x)}| &\leq |\frac{f(x)}{g(x)} - L| + |L| < \epsilon + |L| \stackrel{def}{=} M \\ |f(x)| &< M|g(x)|, \text{ for } 0 < |x| < \delta \end{aligned}$$

therefore $f(x) \in O(g(x))$ as $x \to 0$!

Taylor's formula with Big o and little o notation

$$R_{m+1} = \frac{h^{m+1}}{(m+1)!} f^{(m+1)}(a+\theta h), \quad \theta \in (0,1)$$

Therefore,

$$\begin{aligned} R_{m+1} &= o(h^m), \text{ as } h \to 0 \\ R_{m+1} &= O(h^{m+1}), \text{ as } h \to 0 \end{aligned}$$

Then,if $f(x) \in \mathcal{C}^{m+1}$, we may write Taylor's formula as

$$\begin{aligned} f(x) &= f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^m}{m!} f^{(m)}(a) + o(h^m), \\ f(x) &= f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^m}{m!} f^{(m)}(a) + O(h^{m+1}) \end{aligned}$$

If we assume that $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R} \in \mathcal{C}^3$, we have the formula for the remainder term R_3

$$\begin{aligned} f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + \frac{1}{1!} Df(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2!}(\boldsymbol{x} - \boldsymbol{x}_0)^T D^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \\ &\quad + o(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2) \\ f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + \frac{1}{1!} Df(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2!}(\boldsymbol{x} - \boldsymbol{x}_0)^T D^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \\ &\quad + O(\|\boldsymbol{x} - \boldsymbol{x}_0\|^3) \end{aligned}$$

Taylor Theorem

Suppose that $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R} \in \mathcal{C}^2$. Then we have that

$$\begin{aligned} f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + Df(\boldsymbol{x}_0 + \theta\boldsymbol{h})\boldsymbol{h} \\ f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + Df(\boldsymbol{x}_0)\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T D^2 f(\boldsymbol{x}_0 + \theta\boldsymbol{h})\boldsymbol{h} \end{aligned}$$

for some $\theta \in (0, 1)$ and $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}_0$.

Taylor Theorem

Let $g : \mathbb{R} \to \mathbb{R} \in \mathcal{C}^1$. By Taylor's theorem in 1D (or mean value theorem)

$$g(1) = g(0) + g'(\theta)$$

with $\theta \in (0, 1)$. Let define $g(t) = f(\boldsymbol{z}(t))$ with $\boldsymbol{z}(t) = \boldsymbol{x}_0 + t\boldsymbol{h}$ and $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}_0$. Therefore,

$$\begin{aligned} g(1) &= f(\boldsymbol{z}(1)) = f(\boldsymbol{x}) \\ g(0) &= f(\boldsymbol{z}(0)) = f(\boldsymbol{x}_0) \\ g'(\theta) &= Df(\boldsymbol{z}(\theta))\boldsymbol{h} = \nabla f(\boldsymbol{x}_0 + \theta\boldsymbol{h})^T\boldsymbol{h} \\ f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0 + \theta\boldsymbol{h})^T\boldsymbol{h} \end{aligned}$$

Taylor Theorem

The proof of the second part is similar. Let $g : \mathbb{R} \to \mathbb{R} \in \mathcal{C}^1$. By Taylor's theorem in 1D

$$g(1) = g(0) + g'(0) + \frac{1}{2} g''(\theta)$$

with $\theta \in (0, 1)$. Let define $g(t) = f(\boldsymbol{z}(t))$ with $\boldsymbol{z}(t) = \boldsymbol{x}_0 + t\boldsymbol{h}$ and $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}_0$. Therefore, for $t = \|\boldsymbol{h}\|$

$$\begin{aligned} g(1) &= f(\boldsymbol{z}(1)) = f(\boldsymbol{x}) \\ g(0) &= f(\boldsymbol{z}(0)) = f(\boldsymbol{x}_0) \\ g'(0) &= Df(\boldsymbol{z}(0))\boldsymbol{h} = \nabla f(\boldsymbol{x}_0)^T \boldsymbol{h} \end{aligned}$$

Taylor Theorem

On the other hand

$$\begin{aligned} g'(\theta) &= Df(\boldsymbol{z}(\theta))\boldsymbol{h} = \nabla f(z(\theta))^T \boldsymbol{h} \\ g''(\theta) &= \boldsymbol{h}^T D^2 f(\boldsymbol{z}(t)) D\boldsymbol{z}(t) = \boldsymbol{h}^T D^2 f(\boldsymbol{z}(\theta))\boldsymbol{h} \\ &= \boldsymbol{h}^T D^2 f(\boldsymbol{x}_0 + \theta \boldsymbol{h})\boldsymbol{h} \end{aligned}$$

now using $g(1) = g(0) + g'(0) + \frac{1}{2}g''(\theta)$ and substituting one obtains the result.

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^T D^2 f(\boldsymbol{x}_0 + \theta \boldsymbol{h})\boldsymbol{h}$$