高级工程数学
2021-2022 (1)

沈超敏
计算机科学与技术学院
cmshen@cs.ecnu.edu.cn

教书院 219

Elements of Calculus

5.1 Sequences and Limits

A *sequence of real numbers* is a function whose domain is the set of natural numbers $1, 2, \ldots, k, \ldots$ and whose range is contained in $\mathbb{R}$. Thus, a sequence of real numbers can be viewed as a set of numbers $\{x_1, x_2, \ldots, x_k, \ldots\}$, which is often also denoted as $\{x_k\}$ (or sometimes as $\{x_k\}_{k=1}^{\infty}$, to indicate explicitly the range of values that k can take).

A sequence $\{x_k\}$ is *increasing* if $x_1 < x_2 < \cdots < x_k \cdots$; that is, $x_k < x_{k+1}$ for all k. If $x_k \leq x_{k+1}$, then we say that the sequence is *nondecreasing*. Similarly, we can define *decreasing* and *nonincreasing sequences*. Nonincreasing or nondecreasing sequences are called *monotone sequences*.

A number $x^* \in \mathbb{R}$ is called the *limit* of the sequence $\{x_k\}$ if for any positive ε there is a number K (which may depend on ε) such that for all $k > K$, $|x_k - x^*| < \varepsilon$; that is, x_k lies between $x^* - \varepsilon$ and $x^* + \varepsilon$ for all $k > K$. In this case we write

$$x^* = \lim_{k \to \infty} x_k$$

or

5.1 Sequences and Limits

The notion of a sequence can be extended to sequences with elements in $\mathbb{R}^n$. Specifically, a sequence in $\mathbb{R}^n$ is a function whose domain is the set of natural numbers $1, 2, \ldots, k, \ldots$ and whose range is contained in $\mathbb{R}^n$. We use the notation $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots\}$ or $\{\boldsymbol{x}^{(k)}\}$ for sequences in $\mathbb{R}^n$. For limits of sequences in $\mathbb{R}^n$, we need to replace absolute values with vector norms. In other words, $\boldsymbol{x}^*$ is the limit of $\{\boldsymbol{x}^{(k)}\}$ if for any positive ε there is a number K (which may depend on ε) such that for all $k > K$, $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| < \varepsilon$. As before, if a sequence $\{\boldsymbol{x}^{(k)}\}$ is convergent, we write $\boldsymbol{x}^* = \lim_{k\to\infty} \boldsymbol{x}^{(k)}$ or $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^*$.

5.1 Sequences and Limits

Theorem 5.1 *A convergent sequence has only one limit.* □

Proof. We prove this result by contradiction. Suppose that a sequence $\{\boldsymbol{x}^{(k)}\}$ has two different limits, say $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Then, we have $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| > 0$. Let

$$\varepsilon = \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.$$

From the definition of a limit, there exist K_1 and K_2 such that for $k > K_1$ we have $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_1\| < \varepsilon$, and for $k > K_2$ we have $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_2\| < \varepsilon$. Let $K = \max\{K_1, K_2\}$. Then, if $k > K$, we have $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_1\| < \varepsilon$ and $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_2\| < \varepsilon$. Adding $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_1\| < \varepsilon$ and $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_2\| < \varepsilon$ yields

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}_1\| + \|\boldsymbol{x}^{(k)} - \boldsymbol{x}_2\| < 2\varepsilon.$$

Applying the triangle inequality gives

$$\begin{aligned}
\|-\boldsymbol{x}_1 + \boldsymbol{x}_2\| &= \|\boldsymbol{x}^{(k)} - \boldsymbol{x}_1 - \boldsymbol{x}^{(k)} + \boldsymbol{x}_2\| \\
&= \|(\boldsymbol{x}^{(k)} - \boldsymbol{x}_1) - (\boldsymbol{x}^{(k)} - \boldsymbol{x}_2)\| \\
&\leq \|\boldsymbol{x}^{(k)} - \boldsymbol{x}_1\| + \|\boldsymbol{x}^{(k)} - \boldsymbol{x}_2\|.
\end{aligned}$$

Therefore,

$$\|-\boldsymbol{x}_1 + \boldsymbol{x}_2\| = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| < 2\varepsilon.$$

However, this contradicts the assumption that $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| = 2\varepsilon$, which completes the proof. ∎

Video 22 结束

5.1 Sequences and Limits

A sequence $\{\boldsymbol{x}^{(k)}\}$ in $\mathbb{R}^n$ is *bounded* if there exists a number $B \geq 0$ such that $\|\boldsymbol{x}^{(k)}\| \leq B$ for all $k = 1, 2, \ldots$.

Theorem 5.2 *Every convergent sequence is bounded.* □

Proof. Let $\{\boldsymbol{x}^{(k)}\}$ be a convergent sequence with limit $\boldsymbol{x}^*$. Choose $\varepsilon = 1$. Then, by definition of the limit, there exists a natural number K such that for all $k > K$,

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| < 1.$$

By the result of Exercise 2.9, we get

$$\|\boldsymbol{x}^{(k)}\| - \|\boldsymbol{x}^*\| \leq \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| < 1 \quad \text{for all } k > K.$$

Therefore,

$$\|\boldsymbol{x}^{(k)}\| < \|\boldsymbol{x}^*\| + 1 \quad \text{for all } k > K.$$

Letting

$$B = \max\left\{\|\boldsymbol{x}^{(1)}\|, \|\boldsymbol{x}^{(2)}\|, \ldots, \|\boldsymbol{x}^{(K)}\|, \|\boldsymbol{x}^*\| + 1\right\},$$

we have

$$B \geq \|\boldsymbol{x}^{(k)}\| \quad \text{for all } k,$$

which means that the sequence $\{\boldsymbol{x}^{(k)}\}$ is bounded. ∎

5.1 Sequences and Limits

For a sequence $\{x_k\}$ in $\mathbb{R}$, a number B is called an *upper bound* if $x_k \leq B$ for all $k = 1, 2, \ldots$. In this case, we say that $\{x_k\}$ is *bounded above*. Similarly, B is called a *lower bound* if $x_k \geq B$ for all $k = 1, 2, \ldots$. In this case, we say that $\{x_k\}$ is *bounded below*. Clearly, a sequence is bounded if it is both bounded above and bounded below.

Any sequence $\{x_k\}$ in $\mathbb{R}$ that has an upper bound has a *least upper bound* (also called the *supremum*), which is the smallest number B that is an upper bound of $\{x_k\}$. Similarly, any sequence $\{x_k\}$ in $\mathbb{R}$ that has a lower bound has a *greatest lower bound* (also called the *infimum*). If B is the least upper bound of the sequence $\{x_k\}$, then $x_k \leq B$ for all k, and for any $\varepsilon > 0$, there exists a number K such that $x_K > B - \varepsilon$. An analogous statement applies to the greatest lower bound: If B is the greatest lower bound of $\{x_k\}$, then $x_k \geq B$ for all k, and for any $\varepsilon > 0$, there exists a number K such that $x_K < B + \varepsilon$.

Video 23 结束

5.1 Sequences and Limits

Theorem 5.3 *Every monotone bounded sequence in $\mathbb{R}$ is convergent.* □

Proof. We prove the theorem for nondecreasing sequences. The proof for nonincreasing sequences is analogous.

Let $\{x_k\}$ be a bounded nondecreasing sequence in $\mathbb{R}$ and x^* the least upper bound. Fix a number $\varepsilon > 0$. Then, there exists a number K such that $x_K > x^* - \varepsilon$. Because $\{x_k\}$ is nondecreasing, for any $k \geq K$,

$$x_k \geq x_K > x^* - \varepsilon.$$

Also, because x^* is an upper bound of $\{x_k\}$, we have

$$x_k \leq x^* < x^* + \varepsilon.$$

Therefore, for any $k \geq K$,

$$|x_k - x^*| < \varepsilon,$$

which means that $x_k \to x^*$. ■

Video 24 结束

5.1 Sequences and Limits

Theorem 5.4 *Consider a convergent sequence* $\{\boldsymbol{x}^{(k)}\}$ *with limit* $\boldsymbol{x}^*$. *Then, any subsequence of* $\{\boldsymbol{x}^{(k)}\}$ *also converges to* $\boldsymbol{x}^*$. □

Proof. Let $\{\boldsymbol{x}^{(m_k)}\}$ be a subsequence of $\{\boldsymbol{x}^{(k)}\}$, where $\{m_k\}$ is an increasing sequence of natural numbers. Observe that $m_k \geq k$ for all $k = 1, 2, \ldots$. To show this, first note that $m_1 \geq 1$ because m_1 is a natural number. Next, we proceed by induction by assuming that $m_k \geq k$. Then, we have $m_{k+1} > m_k \geq k$, which implies that $m_{k+1} \geq k + 1$. Therefore, we have shown that $m_k \geq k$ for all $k = 1, 2, \ldots$.

Let $\varepsilon > 0$ be given. Then, by definition of the limit, there exists K such that $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\| < \varepsilon$ for any $k > K$. Because $m_k \geq k$, we also have $\|\boldsymbol{x}^{(m_k)} - \boldsymbol{x}^*\| < \varepsilon$ for any $k > K$. This means that

$$\lim_{k \to \infty} \boldsymbol{x}^{(m_k)} = \boldsymbol{x}^*.$$

■

It turns out that any bounded sequence contains a convergent subsequence. This result is called the *Bolzano-Weierstrass theorem* (see [2, p. 70]).

5.1 Sequences and Limits

Consider a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and a point $\boldsymbol{x}_0 \in \mathbb{R}^n$. Suppose that there exists $\boldsymbol{f}^*$ such that for any convergent sequence $\{\boldsymbol{x}^{(k)}\}$ with limit $\boldsymbol{x}_0$, we have

$$\lim_{k\to\infty} \boldsymbol{f}(\boldsymbol{x}^{(k)}) = \boldsymbol{f}^*.$$

Then, we use the notation

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \boldsymbol{f}(\boldsymbol{x})$$

to represent the limit $\boldsymbol{f}^*$.

It turns out that $\boldsymbol{f}$ is continuous at $\boldsymbol{x}_0$ if and only if for any convergent sequence $\{\boldsymbol{x}^{(k)}\}$ with limit $\boldsymbol{x}_0$, we have

$$\lim_{k\to\infty} \boldsymbol{f}(\boldsymbol{x}^{(k)}) = \boldsymbol{f}\left(\lim_{k\to\infty} \boldsymbol{x}^{(k)}\right) = \boldsymbol{f}(\boldsymbol{x}_0)$$

(see [2, p. 137]). Therefore, using the notation introduced above, the function $\boldsymbol{f}$ is continuous at $\boldsymbol{x}_0$ if and only if

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0).$$

5.1 Sequences and Limits

矩阵的极限

We say that a sequence $\{\boldsymbol{A}_k\}$ of $m \times n$ matrices converges to the $m \times n$ matrix $\boldsymbol{A}$ if

$$\lim_{k\to\infty} \|\boldsymbol{A} - \boldsymbol{A}_k\| = 0.$$

Lemma 5.1 *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Then, $\lim_{k\to\infty} \boldsymbol{A}^k = \boldsymbol{O}$ if and only if the eigenvalues of $\boldsymbol{A}$ satisfy $|\lambda_i(\boldsymbol{A})| < 1$, $i = 1, \ldots, n$.* □

Proof. To prove this theorem, we use the *Jordan form* (see, e.g., [47]). Specifically, it is well known that any square matrix is similar to the Jordan form: There exists a nonsingular $\boldsymbol{T}$ such that

$$\boldsymbol{T}\boldsymbol{A}\boldsymbol{T}^{-1} = \operatorname{diag}\left[\boldsymbol{J}_{m_1}(\lambda_1), \ldots, \boldsymbol{J}_{m_s}(\lambda_1), \boldsymbol{J}_{n_1}(\lambda_2), \ldots, \boldsymbol{J}_{t_v}(\lambda_q)\right] \triangleq \boldsymbol{J},$$

Video 25 结束

5.2 Differentiability

Differential calculus is based on the idea of approximating an arbitrary function by an *affine function*. A function $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ is *affine* if there exists a *linear function* $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $\boldsymbol{y} \in \mathbb{R}^m$ such that

$$\mathcal{A}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}) + \boldsymbol{y}$$

for every $\boldsymbol{x} \in \mathbb{R}^n$. Consider a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and a point $\boldsymbol{x}_0 \in \mathbb{R}^n$. We wish to find an affine function $\mathcal{A}$ that approximates $\boldsymbol{f}$ near the point $\boldsymbol{x}_0$. First, it is natural to impose the condition

$$\mathcal{A}(\boldsymbol{x}_0) = \boldsymbol{f}(\boldsymbol{x}_0).$$

Because $\mathcal{A}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}) + \boldsymbol{y}$, we obtain $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}_0) - \mathcal{L}(\boldsymbol{x}_0)$. By the linearity of $\mathcal{L}$,

$$\mathcal{L}(\boldsymbol{x}) + \boldsymbol{y} = \mathcal{L}(\boldsymbol{x}) - \mathcal{L}(\boldsymbol{x}_0) + \boldsymbol{f}(\boldsymbol{x}_0) = \mathcal{L}(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{f}(\boldsymbol{x}_0).$$

Hence, we may write

$$\mathcal{A}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{f}(\boldsymbol{x}_0).$$

Next, we require that $\mathcal{A}(\boldsymbol{x})$ approaches $\boldsymbol{f}(\boldsymbol{x})$ faster than $\boldsymbol{x}$ approaches $\boldsymbol{x}_0$; that is,

$$\lim_{\boldsymbol{x} \to \boldsymbol{x}_0, \boldsymbol{x} \in \Omega} \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \mathcal{A}(\boldsymbol{x})\|}{\|\boldsymbol{x} - \boldsymbol{x}_0\|} = 0.$$

The conditions above on $\mathcal{A}$ ensure that $\mathcal{A}$ approximates $\boldsymbol{f}$ near $\boldsymbol{x}_0$ in the sense that the error in the approximation at a given point is "small" compared with the distance of the point from $\boldsymbol{x}_0$.

5.3 The Derivative Matrix

Any linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, and in particular the derivative $\mathcal{L}$ of $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$, can be represented by an $m \times n$ matrix. To find the matrix representation $\boldsymbol{L}$ of the derivative $\mathcal{L}$ of a differentiable function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$, we use the natural basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ for $\mathbb{R}^n$. Consider the vectors

$$\boldsymbol{x}_j = \boldsymbol{x}_0 + t\boldsymbol{e}_j, \quad j = 1, \ldots, n.$$

By the definition of the derivative, we have

$$\lim_{t \to 0} \frac{\boldsymbol{f}(\boldsymbol{x}_j) - (t\boldsymbol{L}\boldsymbol{e}_j + \boldsymbol{f}(\boldsymbol{x}_0))}{t} = \boldsymbol{0}$$

for $j = 1, \ldots, n$. This means that

$$\lim_{t \to 0} \frac{\boldsymbol{f}(\boldsymbol{x}_j) - \boldsymbol{f}(\boldsymbol{x}_0)}{t} = \boldsymbol{L}\boldsymbol{e}_j$$

for $j = 1, \ldots, n$. But $\boldsymbol{L}\boldsymbol{e}_j$ is the jth column of the matrix $\boldsymbol{L}$. On the other hand, the vector $\boldsymbol{x}_j$ differs from $\boldsymbol{x}_0$ only in the jth coordinate, and in that coordinate the difference is just the number t. Therefore, the left side of the preceding equation is the partial derivative

$$\frac{\partial \boldsymbol{f}}{\partial x_j}(\boldsymbol{x}_0).$$

Because vector limits are computed by taking the limit of each coordinate function, it follows that if

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix},$$

then

$$\frac{\partial \boldsymbol{f}}{\partial x_j}(\boldsymbol{x}_0) = \begin{bmatrix} \frac{\partial f_1}{\partial x_j}(\boldsymbol{x}_0) \\ \vdots \\ \frac{\partial f_m}{\partial x_j}(\boldsymbol{x}_0) \end{bmatrix},$$

5.3 The Derivative Matrix

Theorem 5.5 *If a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\boldsymbol{x}_0$, then the derivative of $\boldsymbol{f}$ at $\boldsymbol{x}_0$ is determined uniquely and is represented by the $m \times n$ derivative matrix $D\boldsymbol{f}(\boldsymbol{x}_0)$. The best affine approximation to $\boldsymbol{f}$ near $\boldsymbol{x}_0$ is then given by*

$$\mathcal{A}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0) + D\boldsymbol{f}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0),$$

in the sense that

$$\boldsymbol{f}(\boldsymbol{x}) = \mathcal{A}(\boldsymbol{x}) + \boldsymbol{r}(\boldsymbol{x})$$

and $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \|\boldsymbol{r}(\boldsymbol{x})\|/\|\boldsymbol{x} - \boldsymbol{x}_0\| = 0$. The columns of the derivative matrix $D\boldsymbol{f}(\boldsymbol{x}_0)$ are vector partial derivatives. The vector

$$\frac{\partial \boldsymbol{f}}{\partial x_j}(\boldsymbol{x}_0)$$

is a tangent vector at $\boldsymbol{x}_0$ to the curve $\boldsymbol{f}$ obtained by varying only the jth coordinate of $\boldsymbol{x}$. □

5.3 The Derivative Matrix

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then the function ∇f defined by

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{bmatrix} = Df(\boldsymbol{x})^\top$$

is called the *gradient* of f. The gradient is a function from $\mathbb{R}^n$ to $\mathbb{R}^n$, and can be pictured as a *vector field*, by drawing the arrow representing $\nabla f(\boldsymbol{x})$ so that its tail starts at $\boldsymbol{x}$.

Given $f : \mathbb{R}^n \to \mathbb{R}$, if ∇f is differentiable, we say that f is *twice differentiable*, and we write the derivative of ∇f as

$$D^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

(The notation $\frac{\partial^2 f}{\partial x_i \partial x_j}$ represents taking the partial derivative of f with respect to x_j first, then with respect to x_i.) The matrix $D^2 f(\boldsymbol{x})$ is called the *Hessian matrix* of f at $\boldsymbol{x}$, and is often also denoted $\boldsymbol{F}(\boldsymbol{x})$.

5.3 The Derivative Matrix

Note that the Hessian matrix of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x}$ is symmetric if f is twice continuously differentiable at $\boldsymbol{x}$. This is a well-known result from calculus called *Clairaut's theorem* or *Schwarz's theorem*. However, if the second partial derivatives of f are not continuous, then there is no guarantee that the Hessian is symmetric, as shown in the following well-known example.

Example 5.1 Consider the function

$$f(\boldsymbol{x}) = \begin{cases} x_1 x_2 (x_1^2 - x_2^2)/(x_1^2 + x_2^2) & \text{if } \boldsymbol{x} \neq \boldsymbol{0} \\ 0 & \text{if } \boldsymbol{x} = \boldsymbol{0}. \end{cases}$$

Video 26 结束

5.4 Differentiation Rules

We now introduce the *chain rule* for differentiating the composition $g(\boldsymbol{f}(t))$, of a function $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^n$ and a function $g : \mathbb{R}^n \to \mathbb{R}$.

Theorem 5.6 *Let $g : \mathcal{D} \to \mathbb{R}$ be differentiable on an open set $\mathcal{D} \subset \mathbb{R}^n$, and let $\boldsymbol{f} : (a, b) \to \mathcal{D}$ be differentiable on (a, b). Then, the composite function $h : (a, b) \to \mathbb{R}$ given by $h(t) = g(\boldsymbol{f}(t))$ is differentiable on (a, b), and*

$$h'(t) = Dg(\boldsymbol{f}(t))D\boldsymbol{f}(t) = \nabla g(\boldsymbol{f}(t))^\top \begin{bmatrix} f_1'(t) \\ \vdots \\ f_n'(t) \end{bmatrix}.$$

□

Proof. By definition,

$$h'(t) = \lim_{s \to t} \frac{h(s) - h(t)}{s - t} = \lim_{s \to t} \frac{g(\boldsymbol{f}(s)) - g(\boldsymbol{f}(t))}{s - t}$$

if the limit exists. By Theorem 5.5 we write

$$g(\boldsymbol{f}(s)) - g(\boldsymbol{f}(t)) = Dg(\boldsymbol{f}(t))(\boldsymbol{f}(s) - \boldsymbol{f}(t)) + r(s),$$

where $\lim_{s \to t} r(s)/(s - t) = 0$. Therefore,

$$\frac{h(s) - h(t)}{s - t} = Dg(\boldsymbol{f}(t))\frac{\boldsymbol{f}(s) - \boldsymbol{f}(t)}{s - t} + \frac{r(s)}{s - t}.$$

Letting $s \to t$ yields

$$h'(t) = \lim_{s \to t} Dg(\boldsymbol{f}(t))\frac{\boldsymbol{f}(s) - \boldsymbol{f}(t)}{s - t} + \frac{r(s)}{s - t} = Dg(\boldsymbol{f}(t))D\boldsymbol{f}(t).$$

∎

5.4 Differentiation Rules

Next, we present the *product rule*. Let $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ be two differentiable functions. Define the function $h : \mathbb{R}^n \to \mathbb{R}$ by $h(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{g}(\boldsymbol{x})$. Then, h is also differentiable and

$$Dh(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^\top D\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})^\top D\boldsymbol{f}(\boldsymbol{x}).$$

We end this section with a list of some useful formulas from multivariable calculus. In each case, we compute the derivative with respect to $\boldsymbol{x}$. Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ be a given matrix and $\boldsymbol{y} \in \mathbb{R}^m$ a given vector. Then,

$$\begin{aligned} D(\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x}) &= \boldsymbol{y}^\top \boldsymbol{A} \\ D(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}) &= \boldsymbol{x}^\top (\boldsymbol{A} + \boldsymbol{A}^\top) \quad \text{if } m = n. \end{aligned}$$

It follows from the first formula above that if $\boldsymbol{y} \in \mathbb{R}^n$, then

$$D(\boldsymbol{y}^\top \boldsymbol{x}) = \boldsymbol{y}^\top.$$

It follows from the second formula above that if $\boldsymbol{Q}$ is a symmetric matrix, then

$$D(\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}) = 2\boldsymbol{x}^\top \boldsymbol{Q}.$$

In particular,

$$D(\boldsymbol{x}^\top \boldsymbol{x}) = 2\boldsymbol{x}^\top.$$

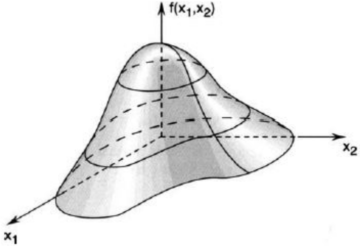
Video 27 结束

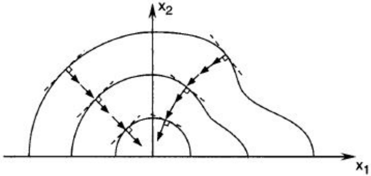
5.5 Level Sets and Gradients

The *level set* of a function $f : \mathbb{R}^n \to \mathbb{R}$ at level c is the set of points

$$S = \{\boldsymbol{x} : f(\boldsymbol{x}) = c\}.$$

For $f : \mathbb{R}^2 \to \mathbb{R}$, we are usually interested in S when it is a curve. For $f : \mathbb{R}^3 \to \mathbb{R}$, the sets S most often considered are surfaces.



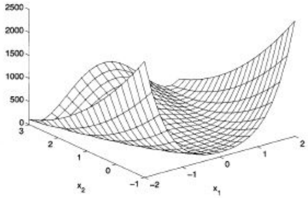


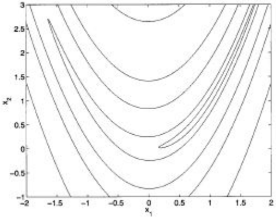
5.5 Level Sets and Gradients

Example 5.2 Consider the following real-valued function on $\mathbb{R}^2$:

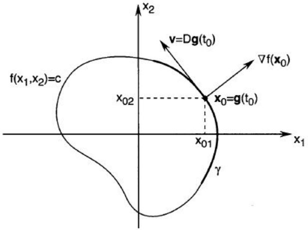
$$f(\boldsymbol{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2, \qquad \boldsymbol{x} = [x_1, x_2]^\top.$$

The function above is called *Rosenbrock's function*. A plot of the function f is shown in Figure 5.2. The level sets of f at levels 0.7, 7, 70, 200, and 700 are depicted in Figure 5.3. These level sets have a particular shape resembling





5.5 Level Sets and Gradients



To say that a point $\boldsymbol{x}_0$ is on the level set S at level c means that $f(\boldsymbol{x}_0) = c$. Now suppose that there is a curve γ lying in S and parameterized by a continuously differentiable function $\boldsymbol{g} : \mathbb{R} \to \mathbb{R}^n$. Suppose also that $\boldsymbol{g}(t_0) = \boldsymbol{x}_0$ and $D\boldsymbol{g}(t_0) = \boldsymbol{v} \neq \mathbf{0}$, so that $\boldsymbol{v}$ is a tangent vector to γ at $\boldsymbol{x}_0$ (see Figure 5.4). Applying the chain rule to the function $h(t) = f(\boldsymbol{g}(t))$ at t_0 gives

$$h'(t_0) = Df(\boldsymbol{g}(t_0))D\boldsymbol{g}(t_0) = Df(\boldsymbol{x}_0)\boldsymbol{v}.$$

But since γ lies on S, we have

$$h(t) = f(\boldsymbol{g}(t)) = c;$$

that is, h is constant. Thus, $h'(t_0) = 0$ and

$$Df(\boldsymbol{x}_0)\boldsymbol{v} = \nabla f(\boldsymbol{x}_0)^\top \boldsymbol{v} = 0.$$

5.6 Taylor Series

Theorem 5.8 ***Taylor's Theorem.*** *Assume that a function $f : \mathbb{R} \to \mathbb{R}$ is m times continuously differentiable (i.e., $f \in C^m$) on an interval $[a, b]$. Denote $h = b - a$. Then,*

$$f(b) = f(a) + \frac{h}{1!} f^{(1)}(a) + \frac{h^2}{2!} f^{(2)}(a) + \cdots + \frac{h^{m-1}}{(m-1)!} f^{(m-1)}(a) + R_m,$$

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

$$R_m = \frac{h^m (1-\theta)^{m-1}}{(m-1)!} f^{(m)}(a + \theta h) = \frac{h^m}{m!} f^{(m)}(a + \theta' h),$$

with $\theta, \theta' \in (0, 1)$. □

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \frac{1}{1!} D f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \\ + \frac{1}{2!} (\boldsymbol{x} - \boldsymbol{x}_0)^\top D^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + o(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2).$$

5.6 Taylor Series

Theorem 5.9 *If a function* $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ *is differentiable on an open set* $\Omega \subset \mathbb{R}^n$, *then for any pair of points* $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, *there exists a matrix* $\boldsymbol{M}$ *such that*

$$\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{M}(\boldsymbol{x} - \boldsymbol{y}).$$

□

The mean value theorem follows from Taylor's theorem (for the case where $m = 1$) applied to each component of $\boldsymbol{f}$. It is easy to see that $\boldsymbol{M}$ is a matrix whose rows are the rows of $D\boldsymbol{f}$ evaluated at points that lie on the line segment joining $\boldsymbol{x}$ and $\boldsymbol{y}$ (these points may differ from row to row).

Video 28 结束

Topic 1: Introduction

★ Problem

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

- f: objective function
- $\Omega \subset \mathbb{R}^n$: feasible set / constraint set

$\max f(\boldsymbol{x})$ can be changed to $\min -f(\boldsymbol{x})$

★ Optimization

1. **Modeling:** Practical problems $\rightarrow$ Optimization problems
2. **Algorithms:** Methods to solve
3. **Software:** Implement

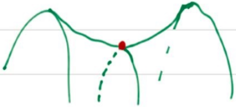
We focus on 2 and 3

Topic 1: Introduction

★ Types of Optimization Problems

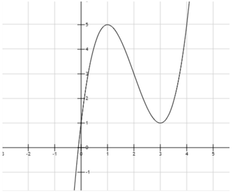
- Constrained vs. Non-Constrained
  $\boldsymbol{x} \in \Omega$ $\boldsymbol{x} \in \mathbb{R}^n$, i.e., $\boldsymbol{x}$ is free
- Convex vs. Non-convex
  both f and Ω are convex
- Smooth vs. Non-Smooth
  ∇f ∂f
- 1-d $\mathbb{R}^1$ vs. n-d $\mathbb{R}^n$

$\left\{\begin{array}{l}\text{illustration}\\ \text{motivation}\\ \text{understanding}\end{array}\right.$ Harder due to infinite directions



★ **Examples**

1. $\min_{x\in[0,5]} f(x), \quad f(x) = x^3 - 6x^2 + 9x + 1$



This is a 1-d smooth, non-convex, and constrained optimization

sol. (only outline, not detailed solution)

critical pts: $f'(x) = 0, \quad x = 1$ or $x = 3$

$\min$ or $\max$: $f''(1) < 0$ local max, $\quad f''(3) > 0$ local min

Video 29 结束

Topic 1: Introduction

★ Examples

Question: how about the case $f''(x) = 0$?

Can we conclude the answer is $x = 3$?

No! This is a constrained opt. problem. Check $x = 3 \in \Omega = [0, 5]$
And compare with values on boundary: $f(0), f(5)$

The final answer is $\min_{x\in[0,5]} f(x) = 1$ and is achieved at $x = 0, 3$

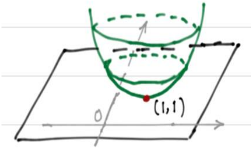
Question: Change the constraint $\min f(x)$ s.t. $x \in [-1, 1]$

Topic 1: Introduction

★ **Examples**

2.

$$f(x, y) = (x-1)^2 + (y-1)^2$$
$$\min_{x,y \in \mathbb{R}^2} f(x, y)$$



Critical pts:

$$\begin{cases} \frac{\partial f}{\partial x}(x, y) = 0, & 2(x-1) = 0 \quad \Rightarrow \quad x = 1 \\ \frac{\partial f}{\partial y}(x, y) = 0, & 2(y-1) = 0 \quad \Rightarrow \quad y = 1 \end{cases}$$

Topic 1: Introduction

★ **Examples**

Second order conditions.

- 1-d: $f''(x) > 0$
- 2-d: Hessian matrix

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y \partial x} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}, \text{ formally } \nabla^2 f \succ 0$$

For a symmetric matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$,

$$A \succ 0 \text{ if } \quad (u, v) A \begin{pmatrix} u \\ v \end{pmatrix} > 0 \quad \forall (u, v) \in \mathbb{R}^2,$$

i.e., $\alpha_{11} u^2 + 2a_{12} uv + a_{22} v^2 > 0, \quad \forall u, v$

特殊值: Set $v = 0, u = 1 \quad \rightarrow \quad a_{11} > 0$

Set $u = 0, v = 1 \quad \rightarrow \quad a_{22} > 0$

$\det(A) = a_{11} a_{22} - a_{12}^2 > 0$

Topic 1: Introduction

2×2 symmetric $A \succ 0$ if a_{11} and $\det(A) > 0$

Important. For a number, either $a > 0$ or $a \leq 0$

For a symmetric matrix, $A \succ 0$ or $A \nsucc 0$ including $\begin{cases} A \preccurlyeq 0 \\ \text{Saddle point} \end{cases}$

Better to view as $A = Q^T \bigwedge Q$, where $\bigwedge = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ and Q consists of corresponding eigenvectors, and $Q^T Q = I$

$A \succ 0 \Leftrightarrow \lambda_1 > 0, \lambda_2 > 0$

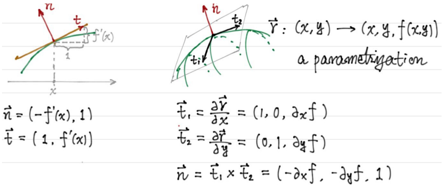
$A \nsucc 0 \begin{cases} \lambda_1 \leq 0, \lambda_2 \leq 0 \\ \lambda_1 \lambda_2 \leq 0 \quad (+, -) \quad \text{saddle pt} \end{cases}$

Topic 1: Introduction

Notation. $f : \mathbb{R}^n \to \mathbb{R} \quad \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x} = (x_1, x_2, ..., x_n)$

$$Df \triangleq (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}), \quad \nabla f = (Df)^T = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \cdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

特例: 1-d, $f'(x)$ is the slope of the tangent line $\frac{\partial f}{\partial x_n}$ at x

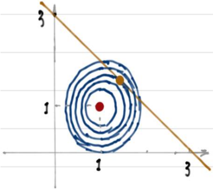


$H = \nabla^2 f = (\frac{\partial^2 f}{\partial x_i \partial x_j})$. As $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$, ∇^2 is symmetric

Topic 1: Introduction

★ **Examples**

3. $\min f(x,y)$ s.t. $x+y=3$
$f(x,y)=(x-1)^2+(y-1)^2, \quad \Omega=\{(x,y)\in\mathbb{R}^2 | x+y=3\}$
This is a 2-d smooth, convex and constrained optimization.



Level set $\mathrm{Sc}=\{\boldsymbol{x}: \quad f(\boldsymbol{x})=c\}$
Level set Sc is a curve in $\mathbb{R}^2$
Graph of $f \quad (\boldsymbol{x}, f(\boldsymbol{x}))\in\mathbb{R}^3$ is a surface in $\mathbb{R}^3$
Non-constrained minimum is at $(1,1)$ red pt $\nabla f(1,1)=0$
With an equality constraint, the minimum is changed to brown where $\nabla f \neq 0$! More complicat

★ **Examples**

For this problem, we can eliminate y to get a 1-d non-constraint smooth and convex opt problem.

$$y = 3 - x \quad \tilde{f}(x) \triangleq f(x, 3 - x) = (x-1)^2 + (x-2)^2$$

$$\tilde{f}'(x) = 0, \text{ so } x = \frac{3}{2}, \quad y = \frac{3}{2} \quad \tilde{f}''(x) = 4 > 0 \forall x$$

$x = \frac{3}{2}$ is a local minimum and as $\tilde{f}$ is convex, it is a global min. So $\min f(x, y)$ s.t. $x + y = 3$ is $\frac{1}{2}$ and the minimum pt $(\frac{3}{2}, \frac{3}{2})$

Fact: for a convex function, a local minimum is also a global one (to be proved soon)

Video 30 结束

Topic 1: Introduction

★ level sets and gradient

$\mathrm{Sc} = \{\boldsymbol{x} | f(\boldsymbol{x}) = c\}$. This is a smooth curve for most c.
How to represent/describe a curve? Parametrization.
$g: \quad \mathbb{R} \quad \to \quad \mathbb{R}^n \quad g(\boldsymbol{t}) = (x(\boldsymbol{t}), y(\boldsymbol{t}))$ in $\mathbb{R}^2$
$h: \quad \mathbb{R} \quad \to \quad \mathbb{R} \quad h(\boldsymbol{t}) \triangleq f(g(\boldsymbol{t}))$
$h(\boldsymbol{t}) = c$ by definition. So $h'(\boldsymbol{t}) = 0$
By chain rule, $h'(\boldsymbol{t}) = \nabla f(g(\boldsymbol{t})) \cdot g'(\boldsymbol{t})$. So for a pt $\boldsymbol{x_0} \in \mathrm{Sc}$, we have $\nabla f(\boldsymbol{x_0}) \cdot \boldsymbol{v} = 0$
where $\boldsymbol{v}$ is a tangent vector of Sc at $\boldsymbol{x}_0$.
Theorem. $\nabla f(\boldsymbol{x_0}) \perp \boldsymbol{v}, \quad \forall \boldsymbol{v}$ tangent vector at $\boldsymbol{x_0}$ of the level set Sc for $c = f(\boldsymbol{x_0})$
$\nabla f(\boldsymbol{x})$ is the direction of maximum rate of increase of f at $\boldsymbol{x}$
$-\nabla f(\boldsymbol{x})$ is the direction of maximum rate of decrease of f at $\boldsymbol{x}$
$-\nabla f(x)$: steepest descent direction

Topic 1: Introduction

★ Examples

4. Rosenbrock function $f(x_1, x_2) = 100(x_2 - {x_1}^2)^2 + (1 - x_1)^2$

$$\min_{\boldsymbol{x} \in \mathbb{R}^2} f$$

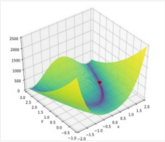
non-constrained, smooth, but non-convex

$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} -400x_1(x_2 - {x_1}^2) - 2(1 - x_1) \\ 200(x_2 - {x_1}^2) \end{pmatrix}, \text{ critical point } (1, 1)$$

$$H(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} 1200{x_1}^2 - 400x_2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}$$

$$H(1, 1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}, \text{ so } (1, 1) \text{ is a local minimum.}$$

$(1, 1)$ is also a global minimum. It is inside a long, narrow, parabolic shaped flab valley.



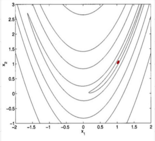


Figure 5.3 Level sets of Rosenbrock's (banana) function.