高级工程数学

2122(1)

沈超敏

计算机科学与技术学院

cmshen@cs.ecnu.edu.cn

Topic 2: Condition

Three acronyms:

- FONC (First Order Necessary Condition)
- SONC (Second Order Necessary Condition)
- SOSC (Second Order Sufficient Condition)

$\boxed{x^* \text{ is a local minimizer of } f}$

Necessary

Sufficient

interior pt $\nabla f(\boldsymbol{x}^*) = 0, \nabla^2 f(\boldsymbol{x}^*) \geqslant 0$ $\nabla f(\boldsymbol{x}^*)$ $\nabla f(\boldsymbol{x}^*) = 0, \nabla^2 f(\boldsymbol{x}^*) > 0$

general pt $\nabla f(\boldsymbol{x}^*) \cdot \boldsymbol{d} \geqslant 0 \quad \forall d$

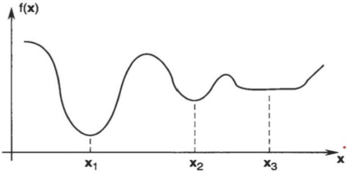
If $\nabla f(\boldsymbol{x}^*) \cdot \boldsymbol{d} = 0$ for some $\boldsymbol{d}, \boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{d} \geqslant 0$

Topic 2: Condition

★ Conditions for Local Minimizers

- Global minimizer $\boldsymbol{x}^*$: $f(\boldsymbol{x}) \geqslant f(\boldsymbol{x}^*) \quad \forall \boldsymbol{x} \in \Omega \setminus \{\boldsymbol{x}^*\}$
- Strict global minimizer $\boldsymbol{x}^*$: $f(\boldsymbol{x}) > f(\boldsymbol{x}^*) \quad \forall \boldsymbol{x} \in \Omega \setminus \{\boldsymbol{x}^*\}$
- Mathematically, “near” can be characterized as $\|\boldsymbol{x} - \boldsymbol{x}^*\| < \varepsilon$
- $\boldsymbol{x}^*$ is a local minimizer if $\exists \varepsilon > 0$, s.t.

$$f(\boldsymbol{x}) \geqslant f(\boldsymbol{x}^*) \quad \forall \boldsymbol{x} \in \Omega \setminus \{\boldsymbol{x}^*\} \,\&\, \|\boldsymbol{x} - \boldsymbol{x}^*\| < \varepsilon$$



Topic 2: Condition

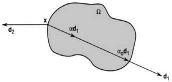
★ **First Order Necessary Condition**

Theorem. $\boldsymbol{x}^*$ is a local minimizer of f over Ω. Then for any feasible direction $\boldsymbol{d}$ at $\boldsymbol{x}^*$, we have

$$\nabla f\left(\boldsymbol{x}^*\right) \cdot \boldsymbol{d} \geqslant 0$$

Explanation. ① feasible direction $\boldsymbol{d}$ at a point $\boldsymbol{x} \in \Omega$ is a direction so that: starting from $\boldsymbol{x}$ and moving towards $\boldsymbol{d}$ remains in Ω.

Math language: $\exists \alpha_0 > 0$ s.t. $\boldsymbol{x} + \alpha \boldsymbol{d} \in \Omega, \forall \alpha \in [0, \alpha_0]$



② $\nabla f\left(\boldsymbol{x}^*\right) \cdot \boldsymbol{d}$: inner product of two vectors.

Also write as $\boldsymbol{d}^\top \nabla f\left(\boldsymbol{x}^*\right)$ or $\left(\nabla f\left(\boldsymbol{x}^*\right), \boldsymbol{d}\right)$, $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{d}\rangle$

$\frac{\partial f}{\partial \boldsymbol{d}} \triangleq \nabla f \cdot \boldsymbol{d}$ is the directional derivative when $\|\boldsymbol{d}\| = 1$

- ③ Define $\phi(\alpha) = f\left(\boldsymbol{x}^* + \alpha \boldsymbol{d}\right)$ for $\alpha \in [0, \alpha_0]$, then

$$\phi'(0) = \begin{cases} \lim_{\alpha \to 0^+} \frac{\phi(\alpha) - \phi(0)}{\alpha} = \lim_{\alpha \to 0^+} \frac{f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) - f(\boldsymbol{x}^*)}{\alpha} & \text{Def} \\ \nabla f\left(\boldsymbol{x}^*\right) \cdot \boldsymbol{d} & \text{Chain rule} \end{cases}$$

Topic 2: Condition

- Proof. Let $\boldsymbol{d}$ be any feasible direction at $\boldsymbol{x}^*$. Define $\phi(\alpha) = f(\boldsymbol{x}^* + \alpha \boldsymbol{d})$
  Then $f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) - f(\boldsymbol{x}^*) = \phi(\alpha) - \phi(0) = \phi'(0)\alpha + o(\alpha)$
  $$= \langle \nabla f(\boldsymbol{x}^*) \cdot \boldsymbol{d} \rangle \alpha + o(\alpha)$$
  If $\boldsymbol{x}^*$ is a local minimizer,
  (i.e., $\exists \varepsilon$, s.t. $f(\boldsymbol{x}) \geqslant f(\boldsymbol{x}^*)$, $\forall \boldsymbol{x} \in \Omega \setminus \{\boldsymbol{x}^*\} \,\&\, \|\boldsymbol{x} - \boldsymbol{x}^*\| < \varepsilon$)
  for sufficiently small α (e.g. $\|\alpha \boldsymbol{d}\| < \varepsilon$), $f(\boldsymbol{x}^* + \alpha d) - f(\boldsymbol{x}^*) \geqslant 0$
  then $\phi'(0) = \nabla f(\boldsymbol{x}^*) \cdot \boldsymbol{d} \geqslant 0$
- FONC Two possibilities for a given feasible direction $\boldsymbol{d}$.
  $$\begin{cases} \nabla f(\boldsymbol{x}^*) \cdot \boldsymbol{d} > 0 \text{ then } f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) > f(\boldsymbol{x}^*) \text{ for all sufficienty small } \alpha > 0 \\ \nabla f(\boldsymbol{x}^*) \cdot \boldsymbol{d} = 0 \text{ . check second-order derivative} \end{cases}$$

Topic 2: Condition

★ Second Order Necessary Condition

- Theorem If $\boldsymbol{x}^*$ is a local minimizer of f over Ω, and there exists a feasible direction $\boldsymbol{d}$ at $\boldsymbol{x}^*$ s.t. $\nabla f\left(\boldsymbol{x}^*\right) \cdot \boldsymbol{d}=0$, then

$$\boldsymbol{d}^{\top} \nabla^2 f\left(\boldsymbol{x}^*\right) \boldsymbol{d} \geqslant 0$$

- Proof Consider $\phi(\alpha)=f\left(\boldsymbol{x}^*+\alpha \boldsymbol{d}\right)$ and its Taylor series at $\alpha=0$

$$\phi(\alpha)=\phi(0)+\phi^{\prime}(0) \alpha+\frac{1}{2} \phi^{\prime \prime}(0) \alpha^2+o\left(\alpha^2\right).$$

$$\text{as } \phi^{\prime}(0)=\nabla f\left(\boldsymbol{x}^*\right) \cdot \boldsymbol{d}=0$$

So we have $\phi(\alpha)-\phi(0)=\frac{1}{2} \phi^{\prime \prime}(0) \alpha^2+o\left(\alpha^2\right)$. Written in terms f is $f\left(\boldsymbol{x}^*+\alpha \boldsymbol{d}\right)-f\left(\boldsymbol{x}^*\right)=\frac{\alpha^2}{2} \boldsymbol{d}^{\top} \nabla^2 f\left(x^*\right) \boldsymbol{d}+o\left(\alpha^2\right)$.

If $\boldsymbol{d}^{\top} \nabla^2 f\left(\boldsymbol{x}^*\right) \boldsymbol{d}<0$, then for sufficiently small α (how small?)

which contradicts that x^* is a local minimizer.

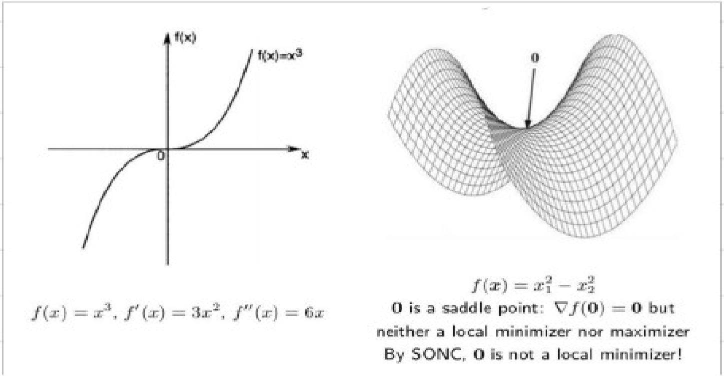
so, $\boldsymbol{d}^{\top} \nabla^2 f\left(\boldsymbol{x}^*\right) \boldsymbol{d} \geq 0$.

Topic 2: Condition

- Corollary $\boldsymbol{x}^*$ is an interior local minimizer of f. Then
- FONC $\nabla f(\boldsymbol{x}^*) = 0$
- SONC $\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}^*)\, \boldsymbol{d} \geqslant 0, \quad \forall \boldsymbol{d} \in \mathbb{R}^n$

  Examples. 6.3 (p.86), 6.5 (p.89)

  Necessary conditions are not sufficient



Video 32 结束

Topic 2: Condition

★ Second Order Sufficient Condition

- Th 6.3 (SOSC) $f \in C^2(\Omega), \boldsymbol{x}^* \in \Omega$ is an interior point.

  Suppose that (1) $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$; (2) $\nabla^2 f(\boldsymbol{x}^*) \succ 0$.

  Then $\boldsymbol{x}^*$ is a strict local minimizer of f.

  Pf. $\nabla^2 f(x^*) \succ 0 \Leftrightarrow \lambda_{\min}\left(\nabla^2 f(x^*)\right) > 0$

  (Prove by diagonaligation of $\nabla^2 f(\boldsymbol{x}^*) = Q^\top \wedge Q\Big)$

  For a feasible direction $\boldsymbol{d} \neq 0$, define $\phi(\alpha) = f(\boldsymbol{x}^* + \alpha \boldsymbol{d})$

$$
\begin{aligned}
f(\boldsymbol{x}^* + \alpha \boldsymbol{d}) - f(\boldsymbol{x}^*) = \phi(\alpha) - \phi(0) &= \frac{1}{2}\phi''(0)\alpha^2 + o\left(\alpha^2\right) \\
&= \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}^*)\, \boldsymbol{d}\alpha^2 + o\left(\alpha^2\right) \\
&\geqslant \frac{1}{2}\lambda_{\min}\|d\|^2\alpha^2 + o\left(\alpha^2\right) > 0
\end{aligned}
$$

  if α is sufficiently small.

Topic 2: Condition

x^* is a local minimizer of f

Necessary

Sufficient

interior pt $\nabla f\left(\boldsymbol{x}^*\right)=0, \nabla^2 f\left(\boldsymbol{x}^*\right) \geqslant 0 \quad \nabla f\left(\boldsymbol{x}^*\right)$ $\qquad \nabla f\left(\boldsymbol{x}^*\right)=0, \nabla^2 f\left(\boldsymbol{x}^*\right)>0$

general pt $\nabla f\left(\boldsymbol{x}^*\right) \cdot \boldsymbol{d} \geqslant 0 \quad \forall d$

If $\nabla f\left(\boldsymbol{x}^*\right) \cdot \boldsymbol{d}=0$ for some $\boldsymbol{d}, \boldsymbol{d}^{\top} \nabla^2 f\left(\boldsymbol{x}^*\right) \boldsymbol{d} \geqslant 0$

Video 33 结束

Ch 7. One-Dimensional Search Methods

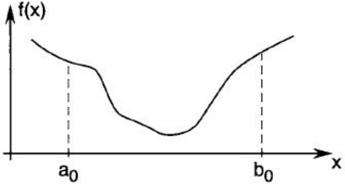


Figure 7.1 Unimodal function.

7.1 Introduction

7.2 Golden Section Search

7.3 Fibonacci Method

7.4 Bisection Method

7.5 Newton's Method

7.6 Secant Method

7.7 Bracketing

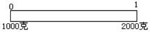
7.8 Line Search in Multidimensional Optimization

Ch 7. One-Dimensional Search Methods

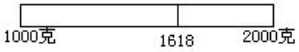
§2 单 因 素

我们知道，钢要用某种化学元素来加强其强度，太少不好，太多也不好. 例如，碳太多了成为生铁，碳太少了成为熟铁，都不成钢材，每吨要加多少碳才能达到强度最高？假定已从理论上算出)每吨在1000克到2000克之间. 普通的方法是加1001克，1002克，……，做下去，做了一千次以后，才能发现最好的选择，这种方法称为均分法. 做一千次实验既浪费时间又浪费原材料. 为了迅速找出最优方案，我们建议以下的“折迭纸条法”.

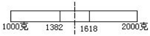
请牢记一个数0.618.



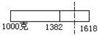
用一个有刻度的纸条表达1000～2000克，在这纸条长度的0.618的地方划一条线，在这条线所指示的刻度做一次实验，也就是按1618克做一次实验.



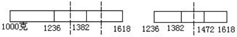
然后把纸条对中迭起，前一线落在另一层上的地方，再划一条线，这条线在1382克处，再按1382克做一次实验.



两次实验进行比较，如果1382克的好一些，我们在1618处把纸条的右边一段剪掉，得：



(如果1618克比较好，则在1382克处剪掉左边一段). 再依中对折起来，又可划出一条线在1236克处：



依1236克做实验，再和1382克的结果比较. 如果，仍然是1382克好，则在1236处剪掉左边：

再依中对折，找出一个试点是1472，按1472克做实验，做出后再剪掉一段，等等. 注意每次留下的纸条的长度是上次长度的0.618(留下的纸条长=0.618×上次长).

就这样，实验、分析、再实验、再分析，矛盾的解决和又出现的过程中，一次比一次地更加接近所需要的加入量，直到所能达到的精度.

Ch 7. One-Dimensional Search Methods

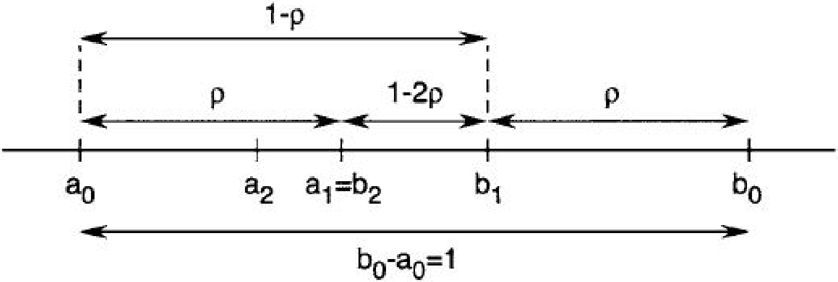


Figure 7.4 Finding value of ρ resulting in only one new evaluation of

Video 34 结束

Ch 7. One-Dimensional Search Methods

7.3 Fibonacci Method

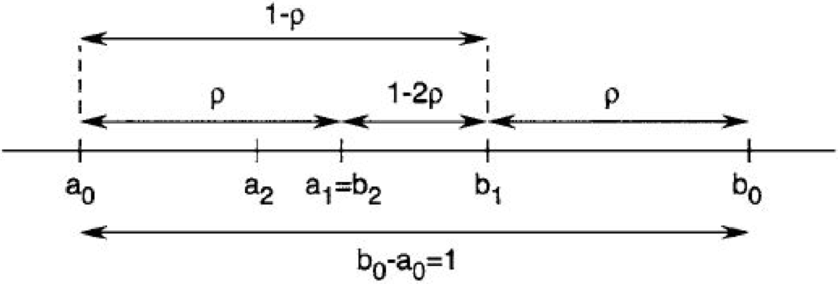
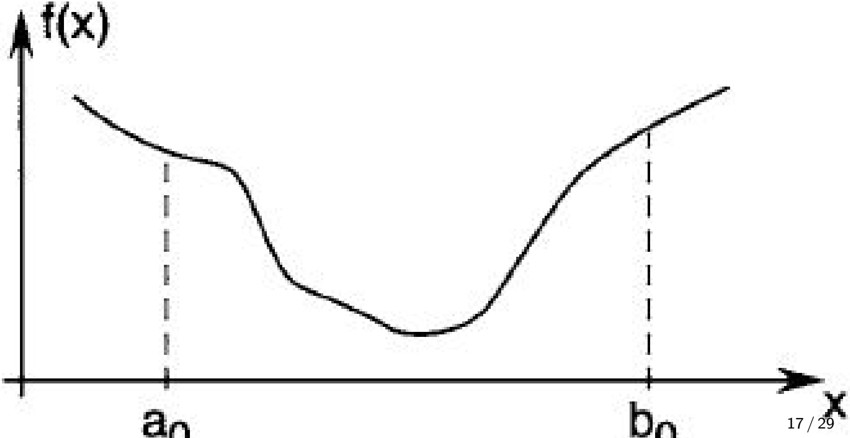


Figure 7.4 Finding value of ρ resulting in only one new evaluation of

7.4 Bisection Method



Video 35 结束

Topic 3: Ch 8. Gradient Methods

$\nabla f(x)$ is the direction of maximum rate of increase of f at x.

$-\nabla f(x)$ is the direction of maximum rate of decrease of f at x.

Lemma: (Cauchy-Schwarz inequality) For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$

$$\boldsymbol{a} \cdot \boldsymbol{b} = (\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} a_i b_i \leqslant \left(\sum_{i=1}^{n} a_i^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} b_i^2 \right)^{\frac{1}{2}} = \|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|$$

Apply to $(\nabla f, \boldsymbol{d}) \leqslant \|\nabla f\| \cdot \|\boldsymbol{d}\| = \|\nabla f\|$ if $\|\boldsymbol{d}\| = 1$

Equality holds $\Leftrightarrow \boldsymbol{d} = \frac{\nabla f}{\|\nabla f\|}$

Therefore $-\nabla f(\boldsymbol{x})$ is the max-rate descending direction. When $\nabla f(\boldsymbol{x}) \neq \mathbf{0}$, for α sufficiently small, $f(\boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})) < f(\boldsymbol{x})$.

Topic 3: Gradient Methods

Gradient Descent Algorithm:

Start from $\boldsymbol{x}^0$ for $k = 0, 1, 2, \cdots$ till converge

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha_k \nabla f\left(\boldsymbol{x}^k\right)$$

Ideal condition: $\left\|\nabla f\left(\boldsymbol{x}^k\right)\right\| = 0$

Practical conditions:

gradient condition $\left\|\nabla f\left(\boldsymbol{x}^k\right)\right\| < \varepsilon$

success objective condition $\frac{\left|f\left(\boldsymbol{x}^{k+1}\right) - f\left(\boldsymbol{x}^k\right)\right|}{\left|f\left(\boldsymbol{x}^k\right)\right|} < \varepsilon$

successive point difference $\frac{\left\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\right\|}{\left\|\boldsymbol{x}^k\right\|} < \varepsilon$

Replace denominator by $\max\left\{1, \left|f\left(\boldsymbol{x}^k\right)\right|\right\}$ or $\max\{1, \|\boldsymbol{x}^k\|\}$ to avoid division by tiny numbers.

Topic 3: Gradient Methods

Step size:

1. Exact line search. Expensive and not worth

2. Fixed estimate value.

3. Line search (Ch. 7).

具体地

1. Exact or "best" for $\phi_k(\alpha) = f\left(x_k - \alpha \nabla f\left(x_k\right)\right)$

$$\alpha_k = \underset{\alpha \geqslant 0}{\arg\min}\, \phi_k(\alpha)$$

This is called Steepest Descent method.

Topic 3: Gradient Methods

2. Based on properties of f, choose a fixed value

small: converges slow

large: may diverge faster

convergence efficiency

3. Several practical line search algorithms

• Golden section • Newton's method

• Fibonacci • Secant method

• Bisection • Bracketing

Topic 3: Gradient Methods

Quadratic Programming

For a symmetric and positive definite (SPD) matrix Q, i.e., $Q > 0$,

we can define a new norm $\|\boldsymbol{x}\|_Q = \left(\boldsymbol{x}^\top Q \boldsymbol{x}\right)^{\frac{1}{2}}$ and a new inner product

$$(\boldsymbol{x}, \boldsymbol{y})_Q \triangleq (Q\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, Q\boldsymbol{y}) = \boldsymbol{y}^T Q \boldsymbol{x}$$

Let $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_Q^2 - (\boldsymbol{b}, \boldsymbol{x})$. Consider non-constrained, convex, and smooth optimization problem:

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

$$\nabla f(\boldsymbol{x}) = Q\boldsymbol{x} - \boldsymbol{b}, \quad \nabla^2 f(\boldsymbol{x}) = Q > 0$$

Topic 3: Quadratic Programming

As f is strictly convex, the global minimum pt is $\nabla f(x) = \mathbf{0}$

Solve $Q\boldsymbol{x} - \boldsymbol{b} = \mathbf{0}$ to get the solution $\boldsymbol{x} = Q^{-1}\boldsymbol{b}$.

Why not computing $Q^{-1}\boldsymbol{b}$ directly?

1. Q^{-1} is expensive, $O(n^3)$ complexity.

2. Want a method for more general problems.

Steepest descent for quadratic programming

$$\phi_k(\alpha) = f\left(\boldsymbol{x}_k - \alpha \nabla f\left(\boldsymbol{x}_k\right)\right), \quad \alpha_k = \arg\min_{\alpha} \phi_k(\alpha) \quad \phi_k'\left(\alpha_k\right) = 0$$

$$\begin{aligned}
\phi_k'(\alpha) &= \langle \nabla f(\boldsymbol{x}_k - \alpha \nabla f(\boldsymbol{x}_k)), -\nabla f\left(\boldsymbol{x}_k\right) \rangle \\
&= \langle Q\left(\boldsymbol{x}_k - \alpha \nabla f\left(\boldsymbol{x}_k\right)\right) - \boldsymbol{b}, -\nabla f\left(\boldsymbol{x}_k\right) \rangle \\
&= \alpha \|\nabla f(\boldsymbol{x}_k)\|_Q^2 - \langle Q\boldsymbol{x}_k - \boldsymbol{b}, \nabla f\left(\boldsymbol{x}_k\right) \rangle = \alpha \left\|\nabla f\left(\boldsymbol{x}_k\right)\right\|_Q^2 - \left\|\nabla f\left(\boldsymbol{x}_k\right)\right\|^2
\end{aligned}$$

Video 36 结束

Topic 3: Quadratic Programming

If we denote by $\boldsymbol{g}_k = \nabla f\left(\boldsymbol{x}_k\right)$, then we can write as

$$\alpha_k = \left\|\boldsymbol{g}_k\right\|^2 / \left\|\boldsymbol{g}_k\right\|_Q^2$$

As $\lambda_{\min}\|\boldsymbol{v}\|^2 \leqslant \boldsymbol{v}^\top Q \boldsymbol{v} \leqslant \lambda_{\max}\|\boldsymbol{v}\|^2$

(which can be proved first for diagonal matrix and then $Q = U^T \Lambda U$)

so $\frac{1}{\lambda_{\max}} \leqslant \alpha_k \leqslant \frac{1}{\lambda_{\min}}$

Video 37 结束

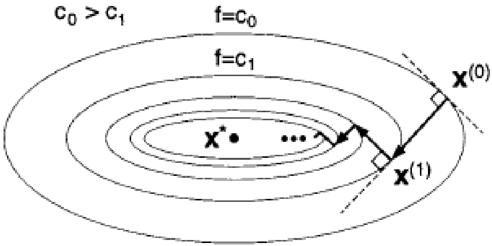


Figure 8.7 Steepest descent method in search for minimizer in a narrow valley.

Topic 4: Newton's method

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

- $n = 1$ $\quad x_{k+1} = x_k - \left(f''(x_k)\right)^{-1} f'(x_k)$
- $n > 1$ $\quad x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$

Another form ① solve $\nabla^2 f(x_k)\, d_k = -\nabla f(x_k)$

② update $x_{k+1} = x_k + d_k$

Remark. Do not require $\nabla^2 f(x_k) \succ 0$ only needs non-singular (invertible).

Namely, Newton's method also works for non-convex optimization problems.

but may not find local min.