15.093 Optimization Methods

Lecture 18: Optimality Conditions and
Gradient Methods
for Unconstrained Optimization

1 Outline

SLIDE 1

1. Necessary and sufficient optimality conditions
2. Gradient methods
3. The steepest descent algorithm
4. Rate of convergence
5. Line search algorithms

2 Optimality Conditions

SLIDE 2

Necessary Conds for Local Optima
"If $\bar{x}$ is local optimum then $\bar{x}$ must satisfy ..."

Identifies all candidates for local optima.

Sufficient Conds for Local Optima
"If $\bar{x}$ satisfies ...,then $\bar{x}$ must be a local optimum "

3 Optimality Conditions

3.1 Necessary conditions

SLIDE 3

Consider

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

Zero first order variation along all directions

Theorem
Let $f(\boldsymbol{x})$ be continuously differentiable.
If $\boldsymbol{x}^* \in \Re^n$ is a local minimum of $f(\boldsymbol{x})$, then

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0} \text{ and } \nabla^2 f(x^*) \text{ PSD}$$

3.2 Proof

SLIDE 4

Zero slope at local min x^*

- $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}^* + \lambda \boldsymbol{d})$ for all $\boldsymbol{d} \in \Re^n$, $\lambda \in \Re$

- Pick $\lambda > 0$

$$0 \leq \frac{f(\boldsymbol{x}^* + \lambda \boldsymbol{d}) - f(\boldsymbol{x}^*)}{\lambda}$$

- Take limits as $\lambda \to 0$

$$0 \leq \nabla f(\boldsymbol{x}^*)' \boldsymbol{d}, \qquad \forall \boldsymbol{d} \in \Re^n$$

- Since $\boldsymbol{d}$ arbitrary, replace with $-d \Rightarrow \nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

SLIDE 5

Nonnegative curvature at a local min x^*

- $f(x^* + \lambda d) - f(x^*) = \nabla f(x^*)'(\lambda d) + \frac{1}{2}(\lambda d)' \nabla^2 f(x^*)(\lambda d) + ||\lambda d||^2 R(x^*; \lambda d)$ where $R(x^*; y) \to 0$ as $y \to 0$. Since $\nabla f(x^*) = 0$,

$$= \frac{1}{2}\lambda^2 d' \nabla^2 f(x^*) d + \lambda^2 ||d||^2 R(x^*; \lambda d) \Rightarrow$$

$$\frac{f(\boldsymbol{x}^* + \lambda \boldsymbol{d}) - f(\boldsymbol{x}^*)}{\lambda^2} = \frac{1}{2} d' \nabla^2 f(x^*) d + ||d||^2 R(x^*; \lambda d)$$

If $\nabla^2 f(x^*)$ is not PSD, $\exists \bar{d}$: $\bar{d}\,' \nabla^2 f(x^*) \bar{d} < 0 \Rightarrow f(x^* + \lambda \bar{d}) < f(\bar{x})$, $\forall \lambda$ suff. small QED.

3.3 Example

SLIDE 6

$f(x) = \frac{1}{2}x_1^2 + x_1.x_2 + 2x_2^2 - 4x_1 - 4x_2 - x_2^3$
$\nabla f(x) = (x_1 + x_2 - 4,\ x_1 + 4x_2 - 4 - 3x_2^2)$ Candidates $\boldsymbol{x}^* = (4, 0)$ and $\bar{\boldsymbol{x}} = (3, 1)$

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} 1 & 1 \\ 1 & 4 - 6x_2 \end{bmatrix}$$

$$\nabla^2 f(\boldsymbol{x}^*) = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

PSD

SLIDE 7

$\bar{\boldsymbol{x}} = (3, 1)$

$$\nabla^2 f(\bar{\boldsymbol{x}}) = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}$$

Indefinite matrix
$\boldsymbol{x}^*$ is the only candidate for local min

3.4 Sufficient conditions

SLIDE 8

Theorem f twice continuously differentiable. If $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x})$ PSD in $B(\boldsymbol{x}^*, \epsilon)$, then $\boldsymbol{x}^*$ is a local minimum.
Proof: Taylor series expansion: For all $\boldsymbol{x} \in B(\boldsymbol{x}^*, \epsilon)$

$$f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)'(\boldsymbol{x} - \boldsymbol{x}^*)$$
$$+ \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)' \nabla^2 f(\boldsymbol{x}^* + \lambda(\boldsymbol{x} - \boldsymbol{x}^*))(\boldsymbol{x} - \boldsymbol{x}^*)$$

for some $\lambda \in [0, 1]$

$$\Rightarrow f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$$

3.5 Example Continued...

SLIDE 9

At $x^* = (4, 0)$, $\nabla f(x^*) = 0$ and
$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} 1 & 1 \\ 1 & 4 - 6x_2 \end{bmatrix}$
is PSD for $\boldsymbol{x} \in B(\boldsymbol{x}^*, \epsilon)$

SLIDE 10

$f(x) = x_1^3 + x_2^2$ and $\nabla f(x) = (3x_1^2, 2x_2)$ $x^* = (0, 0)$
$\nabla^2 f(x) = \begin{bmatrix} 6x_1 & 0 \\ 0 & 2 \end{bmatrix}$ is not PSD in $B(\mathbf{0}, \epsilon)$
$f(-\epsilon, 0) = -\epsilon^3 < 0 = f(\boldsymbol{x}^*)$

3.6 Characterization of convex functions

SLIDE 11

Theorem Let $f(\boldsymbol{x})$ be continuously differentiable.
Then $f(\boldsymbol{x})$ is convex if and only if

$$\nabla f(\boldsymbol{x})'(\overline{\boldsymbol{x}} - \boldsymbol{x}) \leq f(\overline{\boldsymbol{x}}) - f(\boldsymbol{x})$$

3.7 Proof

SLIDE 12

By convexity

$$f(\lambda\overline{\boldsymbol{x}} + (1-\lambda)\boldsymbol{x}) \leq \lambda f(\overline{\boldsymbol{x}}) + (1-\lambda) f(\boldsymbol{x})$$

$$\frac{f(\boldsymbol{x} + \lambda(\overline{\boldsymbol{x}} - \boldsymbol{x})) - f(\boldsymbol{x})}{\lambda} \leq f(\overline{\boldsymbol{x}}) - f(\boldsymbol{x})$$

As $\lambda \to 0$,

$$\nabla f(\boldsymbol{x})'(\overline{\boldsymbol{x}} - \boldsymbol{x}) \leq f(\overline{\boldsymbol{x}}) - f(\boldsymbol{x})$$

3.8 Convex functions

SLIDE 13

Theorem Let $f(\boldsymbol{x})$ be a continuously differentiable convex function. Then $\boldsymbol{x}^*$ is a minimum of f if and only if

$$\nabla f(\boldsymbol{x}^*) = \mathbf{0}$$

Proof: If f convex and $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \nabla f(\boldsymbol{x}^*)'(\boldsymbol{x} - \boldsymbol{x}^*) = 0$$

3.9 Descent Directions

Slide 14

Interesting Observation

f diff/ble at $\bar{x}$
$\exists d$: $\nabla f(\bar{x})'d < 0 \Rightarrow \forall \lambda > 0$, suff. small, $f(\bar{x} + \lambda d) < f(\bar{x})$
(d: descent direction)

3.10 Proof

Slide 15

$$f(\bar{x} + \lambda d) = f(\bar{x}) + \lambda \nabla f(\bar{x})^t d + \lambda ||d|| R(\bar{x}, \lambda d)$$

where $R(\bar{x}, \lambda d) \longrightarrow_{\lambda \to 0} 0$

$$\frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda} = \nabla f(\bar{x})^t d + ||d|| R(\bar{x}, \lambda d)$$

$\nabla f(\bar{x})^t d < 0$, $R(\bar{x}, \lambda d) \longrightarrow_{\lambda \to 0} 0 \Rightarrow$
$\forall \lambda > 0$ suff. small $f(\bar{x} + \lambda d) < f(\bar{x})$. QED

4 Algorithms for unconstrained optimization

4.1 Gradient Methods-Motivation

Slide 16

- Decrease $f(\boldsymbol{x})$ until $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$
- $$f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) \approx f(\bar{\boldsymbol{x}}) + \lambda \nabla f(\bar{\boldsymbol{x}})' \boldsymbol{d}$$
- If $\nabla f(\bar{\boldsymbol{x}})' \boldsymbol{d} < 0$, then for small $\lambda > 0$,
$$f(\bar{\boldsymbol{x}} + \lambda \boldsymbol{d}) < f(\bar{\boldsymbol{x}})$$

5 Gradient Methods

5.1 A generic algorithm

Slide 17

- $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \lambda^k \boldsymbol{d}^k$
- If $\nabla f(\boldsymbol{x}^k) \neq \mathbf{0}$, direction $\boldsymbol{d}^k$ satisfies:
$$\nabla f(\boldsymbol{x}^k)' \boldsymbol{d}^k < 0$$
- Step-length $\lambda^k > 0$
- Principal example:
$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \lambda^k \boldsymbol{D}^k \nabla f(\boldsymbol{x}^k)$$
$\boldsymbol{D}^k$ positive definite symmetric matrix

5.2 Principal directions

Slide 18

- Steepest descent:
$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \lambda^k \nabla f(\boldsymbol{x}^k)$$
- Newton's method:
$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \lambda^k (\nabla^2 f(\boldsymbol{x}^k))^{-1} \nabla f(\boldsymbol{x}^k)$$

5.3 Other directions

Slide 19

- Diagonally scaled steepest descent
$$\boldsymbol{D}^k = \text{Diagonal approximation to } (\nabla^2 f(\boldsymbol{x}^k))^{-1}$$
- Modified Newton's method
$$\boldsymbol{D}^k = \text{Diagonal approximation to } (\nabla^2 f(\boldsymbol{x}^0))^{-1}$$
- Gauss-Newton method for least squares problems $f(\boldsymbol{x}) = ||g(\boldsymbol{x})||^2$ $\boldsymbol{D}^k = (\nabla g(\boldsymbol{x}^k) \nabla g(\boldsymbol{x}^k)')^{-1}$

6 Steepest descent

6.1 The algorithm

Slide 20

Step 0 Given $\boldsymbol{x}^0$, set $k := 0$.

Step 1 $\boldsymbol{d}^k := -\nabla f(\boldsymbol{x}^k)$. If $||\boldsymbol{d}^k|| \leq \epsilon$, then stop.

Step 2 Solve $\min_\lambda h(\lambda) := f(\boldsymbol{x}^k + \lambda \boldsymbol{d}^k)$ for the step-length λ^k, perhaps chosen by an exact or inexact line-search.

Step 3 Set $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{x}^k + \lambda^k \boldsymbol{d}^k$, $k \leftarrow k+1$. Go to **Step 1**.

6.2 An example

Slide 21

minimize $f(x_1, x_2) = 5x_1^2 + x_2^2 + 4x_1x_2 - 14x_1 - 6x_2 + 20$

$\boldsymbol{x}^* = (x_1^*, x_2^*)' = (1,1)'$

$f(\boldsymbol{x}^*) = 10$

Slide 22

Given $\boldsymbol{x}^k$

$$\boldsymbol{d}^k = -\nabla f(x_1^k, x_2^k) = \begin{pmatrix} -10x_1^k - 4x_2^k + 14 \\ -2x_2^k - 4x_1^k + 6 \end{pmatrix} = \begin{pmatrix} d_1^k \\ d_2^k \end{pmatrix}$$

$$\begin{aligned} h(\lambda) &= f(x^k + \lambda d^k) \\ &= 5(x_1^k + \lambda d_1^k)^2 + (x_2^k + \lambda d_2^k)^2 + 4(x_1^k + \lambda d_1^k)(x_2^k + \lambda d_2^k) - \\ &\quad -14(x_1^k + \lambda d_1^k) - 6(x_2^k + \lambda d_2^k) + 20 \end{aligned}$$

$$\lambda^k = \frac{(d_1^k)^2 + (d_2^k)^2}{2(5(d_1^k)^2 + (d_2^k)^2 + 4d_1^k d_2^k)}$$

SLIDE 23

Start at $\boldsymbol{x} = (0, 10)'$

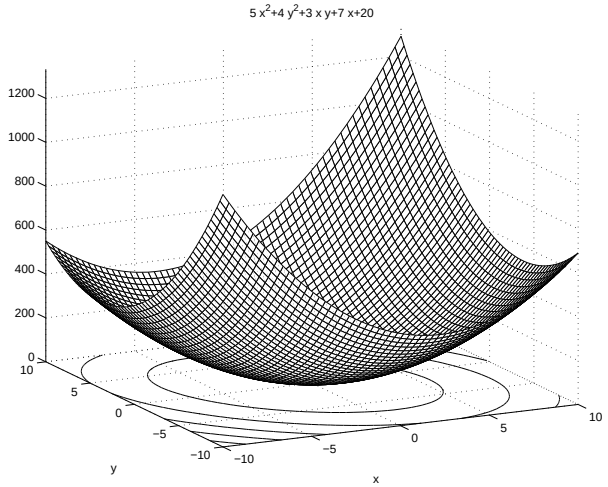
$\varepsilon = 10^{-6}$

k	x_1^k	x_2^k	d_1^k	d_2^k	$\|\|d^k\|\|_2$	λ^k	$f(x^k)$
1	0.000000	10.000000	−26.000000	−14.000000	29.52964612	0.0866	60.000000
2	−2.252782	8.786963	1.379968	−2.562798	2.91071234	2.1800	22.222576
3	0.755548	3.200064	−6.355739	−3.422321	7.21856659	0.0866	12.987827
4	0.204852	2.903535	0.337335	−0.626480	0.71152803	2.1800	10.730379
5	0.940243	1.537809	−1.553670	−0.836592	1.76458951	0.0866	10.178542
6	0.805625	1.465322	0.082462	−0.153144	0.17393410	2.1800	10.043645
7	0.985392	1.131468	−0.379797	−0.204506	0.43135657	0.0866	10.010669
8	0.952485	1.113749	0.020158	−0.037436	0.04251845	2.1800	10.002608
9	0.996429	1.032138	−0.092842	−0.049992	0.10544577	0.0866	10.000638
10	0.988385	1.027806	0.004928	−0.009151	0.01039370	2.1800	10.000156

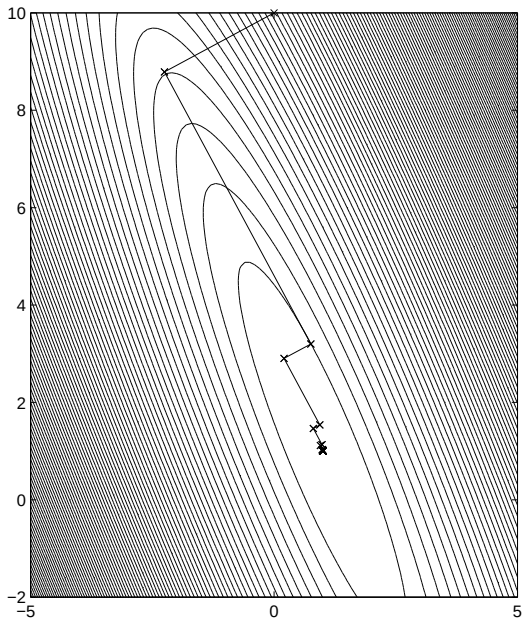
SLIDE 24

k	x_1^k	x_2^k	d_1^k	d_2^k	$\|\|d^k\|\|_2$	λ^k	$f(x^k)$
11	0.999127	1.007856	−0.022695	−0.012221	0.02577638	0.0866	10.000038
12	0.997161	1.006797	0.001205	−0.002237	0.00254076	2.1800	10.000009
13	0.999787	1.001920	−0.005548	−0.002987	0.00630107	0.0866	10.000002
14	0.999306	1.001662	0.000294	−0.000547	0.00062109	2.1800	10.000001
15	0.999948	1.000469	−0.001356	−0.000730	0.00154031	0.0866	10.000000
16	0.999830	1.000406	0.000072	−0.000134	0.00015183	2.1800	10.000000
17	0.999987	1.000115	−0.000332	−0.000179	0.00037653	0.0866	10.000000
18	0.999959	1.000099	0.000018	−0.000033	0.00003711	2.1800	10.000000
19	0.999997	1.000028	−0.000081	−0.000044	0.00009204	0.0866	10.000000
20	0.999990	1.000024	0.000004	−0.000008	0.00000907	2.1803	10.000000
21	0.999999	1.000007	−0.000020	−0.000011	0.00002250	0.0866	10.000000
22	0.999998	1.000006	0.000001	−0.000002	0.00000222	2.1817	10.000000
23	1.000000	1.000002	−0.000005	−0.000003	0.00000550	0.0866	10.000000
24	0.999999	1.000001	0.000000	−0.000000	0.00000054	0.0000	10.000000

SLIDE 25



SLIDE 26



6.3 Important Properties

Slide 27

- $f(\boldsymbol{x}^{k+1}) < f(\boldsymbol{x}^k) < \cdots < f(\boldsymbol{x}^0)$ (because $\boldsymbol{d}^k$ are descent directions)
- Under reasonable assumptions of $f(\boldsymbol{x})$, the sequence $\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots,$ will have at least one cluster point $\bar{\boldsymbol{x}}$
- Every cluster point $\bar{\boldsymbol{x}}$ will satisfy $\nabla f(\bar{\boldsymbol{x}}) = \mathbf{0}$
- *Implication*: If $f(\boldsymbol{x})$ is a convex function, $\bar{\boldsymbol{x}}$ will be an optimal solution

7 Global Convergence Result

Slide 28

Theorem:
$f : R^n \to R$ is continuously diff/ble on $\mathcal{F} = \{x \in R^n : \ f(x) \leq f(x^0)\}$ closed, bounded set
Every cluster point $\bar{x}$ of $\{x_k\}$ satisfies $\nabla f(\bar{x}) = 0$.

7.1 Work Per Iteration

Slide 29

Two computation tasks at each iteration of steepest descent:

- Compute $\nabla f(\boldsymbol{x}^k)$ (for quadratic objective functions, it takes $O(n^2)$ steps) to determine $\boldsymbol{d}^k = -\nabla f(\boldsymbol{x}^k)$

- Perform line-search of $h(\lambda) = f(\boldsymbol{x}^k + \lambda \boldsymbol{d}^k)$
  to determine $\lambda^k = \arg\min_\lambda h(\lambda) = \arg\min_\lambda f(\boldsymbol{x}^k + \lambda \boldsymbol{d}^k)$

8 Rate of convergence of algorithms

SLIDE 30

Let $z_1, \ldots, z_n, \ldots \to z$ be a convergent sequence. We say that the order of convergence of this sequence is p^* if

$$p^* = \sup \left\{ p : \lim \sup_{k \to \infty} \frac{|z_{k+1} - z|}{|z_k - z|^p} < \infty \right\}$$

Let

$$\beta = \lim \sup_{k \to \infty} \frac{|z_{k+1} - z|}{|z_k - z|^{p^*}}$$

The larger p^*, the faster the convergence

8.1 Types of convergence

SLIDE 31

1. $p^* = 1$, $0 < \beta < 1$, then linear (or geometric) rate of convergence
2. $p^* = 1$, $\beta = 0$, super-linear convergence
3. $p^* = 1$, $\beta = 1$, sub-linear convergence
4. $p^* = 2$, quadratic convergence

8.2 Examples

SLIDE 32

- $z_k = a^k$, $0 < a < 1$ converges linearly to zero, $\beta = a$
- $z_k = a^{2^k}$, $0 < a < 1$ converges quadratically to zero
- $z_k = \frac{1}{k}$ converges sub-linearly to zero
- $z_k = \left(\frac{1}{k}\right)^k$ converges super-linearly to zero

8.3 Steepest descent

SLIDE 33

- $z_k = f(\boldsymbol{x}^k)$, $z = f(\boldsymbol{x}^*)$, where $\boldsymbol{x}^* = \arg\min f(\boldsymbol{x})$

- Then an algorithm exhibits *linear convergence* if there is a constant $\delta < 1$ such that

$$\frac{f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)} \leq \delta \ ,$$

for all k sufficiently large, where $\boldsymbol{x}^*$ is an optimal solution.

8.3.1 Discussion

Slide 34

$$\frac{f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)} \leq \delta < 1$$

- If $\delta = 0.1$, every iteration adds another digit of accuracy to the optimal objective value.

- If $\delta = 0.9$, every 22 iterations add another digit of accuracy to the optimal objective value, because $(0.9)^{22} \approx 0.1$.

9 Rate of convergence of steepest descent

9.1 Quadratic Case

9.1.1 Theorem

Slide 35

Suppose $f(\boldsymbol{x}) = \frac{1}{2}\,\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{c}'\boldsymbol{x}$
Q is psd
$\lambda_{\max}$ = largest eigenvalue of $\boldsymbol{Q}$
$\lambda_{\min}$ = smallest eigenvalues of $\boldsymbol{Q}$

Linear Convergence Theorem: If $f(\boldsymbol{x})$ is a quadratic function and $\boldsymbol{Q}$ is psd, then

$$\frac{f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)} \leq \left(\frac{\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) - 1}{\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) + 1} \right)^2$$

9.1.2 Discussion

Slide 36

$$\frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} \leq \left(\frac{\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) - 1}{\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) + 1} \right)^2$$

- $\kappa(\boldsymbol{Q}) := \frac{\lambda_{\max}}{\lambda_{\min}}$ is the *condition number* of $\boldsymbol{Q}$

- $\kappa(\boldsymbol{Q}) \geq 1$
- $\kappa(\boldsymbol{Q})$ plays an extremely important role in analyzing computation involving $\boldsymbol{Q}$

SLIDE 37

$$\frac{f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)} \leq \left(\frac{\kappa(\boldsymbol{Q}) - 1}{\kappa(\boldsymbol{Q}) + 1}\right)^2$$

$\kappa(Q) = \frac{\lambda_{\max}}{\lambda_{\min}}$	Upper Bound on Convergence Constant δ	Number of Iterations to Reduce the Optimality Gap by 0.10
1.1	0.0023	1
3.0	0.25	2
10.0	0.67	6
100.0	0.96	58
200.0	0.98	116
400.0	0.99	231

SLIDE 38

For $\kappa(Q) \sim O(1)$ converges fast.
For large $\kappa(Q)$

$$\left(\frac{\kappa(\boldsymbol{Q}) - 1}{\kappa(\boldsymbol{Q}) + 1}\right)^2 \sim (1 - \frac{1}{\kappa(\boldsymbol{Q})})^2 \sim 1 - \frac{2}{\kappa(\boldsymbol{Q})}$$

Therefore

$$(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)) \leq (1 - \frac{2}{\kappa(\boldsymbol{Q})})^k (f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*))$$

In $k \sim \frac{1}{2}\kappa(\boldsymbol{Q})(-ln\epsilon)$ iterations, finds $\boldsymbol{x^k}$:

$$(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)) \leq \epsilon (f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*))$$

9.2 Example 2

SLIDE 39

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x'Qx} - \boldsymbol{c'x} + 10$$

$$\boldsymbol{Q} = \begin{bmatrix} 20 & 5 \\ 5 & 1 \end{bmatrix} \qquad c = \begin{pmatrix} 14 \\ 6 \end{pmatrix}$$

$\kappa(\boldsymbol{Q}) = 30.234$
$\delta = \left(\frac{\kappa(\boldsymbol{Q})-1}{\kappa(\boldsymbol{Q})+1}\right)^2 = 0.8760$

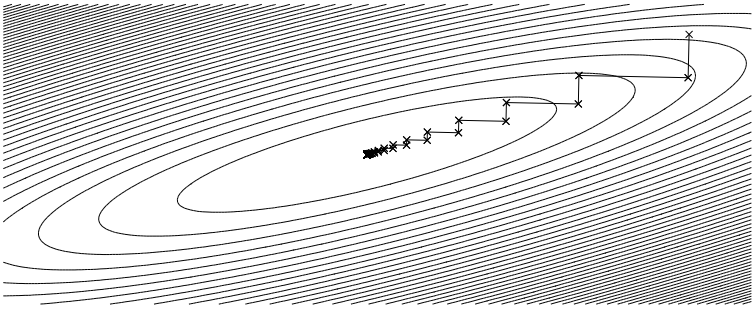
SLIDE 40

k	x_1^k	x_2^k	$\|\|\boldsymbol{d}^k\|\|_2$	λ^k	$f(\boldsymbol{x}^k)$	$\frac{f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^{k-1}) - f(\boldsymbol{x}^*)}$
1	40.000000	−100.000000	286.06293014	0.0506	6050.000000	
2	25.542693	−99.696700	77.69702948	0.4509	3981.695128	0.658079
3	26.277558	−64.668130	188.25191488	0.0506	2620.587793	0.658079
4	16.763512	−64.468535	51.13075844	0.4509	1724.872077	0.658079
5	17.247111	−41.416980	123.88457127	0.0506	1135.420663	0.658079
6	10.986120	−41.285630	33.64806192	0.4509	747.515255	0.658079
7	11.304366	−26.115894	81.52579489	0.0506	492.242977	0.658079
8	7.184142	−26.029455	22.14307211	0.4509	324.253734	0.658079
9	7.393573	−16.046575	53.65038732	0.0506	213.703595	0.658079
10	4.682141	−15.989692	14.57188362	0.4509	140.952906	0.658079

SLIDE 41

k	x_1^k	x_2^k	$\|\|\boldsymbol{d}^k\|\|_2$	λ^k	$f(\boldsymbol{x}^k)$	$\frac{f(\boldsymbol{x}^k)-f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^{k-1})-f(\boldsymbol{x}^*)}$
20	0.460997	0.948466	1.79847660	0.4509	3.066216	0.658079
30	−0.059980	3.038991	0.22196980	0.4509	0.965823	0.658079
40	−0.124280	3.297005	0.02739574	0.4509	0.933828	0.658079
50	−0.132216	3.328850	0.00338121	0.4509	0.933341	0.658079
60	−0.133195	3.332780	0.00041731	0.4509	0.933333	0.658078
70	−0.133316	3.333265	0.00005151	0.4509	0.933333	0.658025
80	−0.133331	3.333325	0.00000636	0.4509	0.933333	0.654656
90	−0.133333	3.333332	0.00000078	0.0000	0.933333	0.000000

SLIDE 42



9.3 Example 3

SLIDE 43

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x'Qx} - \boldsymbol{c'x} + 10$$

$$\boldsymbol{Q} = \begin{bmatrix} 20 & 5 \\ 5 & 16 \end{bmatrix} \qquad \boldsymbol{c} = \begin{pmatrix} 14 \\ 6 \end{pmatrix}$$

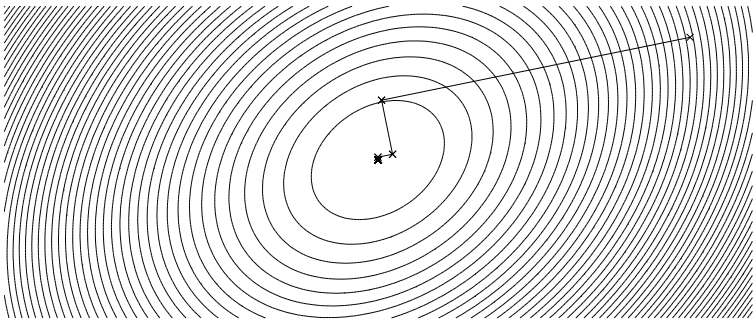
$\kappa(\boldsymbol{Q}) = 1.8541$

$$\delta = \left(\frac{\kappa(\boldsymbol{Q})-1}{\kappa(\boldsymbol{Q})+1}\right)^2 = 0.0896$$

SLIDE 44

k	x_1^k	x_2^k	$\|\|\boldsymbol{d}^k\|\|_2$	λ^k	$f(\boldsymbol{x}^k)$	$\frac{f(\boldsymbol{x}^k)-f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^{k-1})-f(\boldsymbol{x}^*)}$
1	40.000000	−100.000000	1434.79336491	0.0704	76050.000000	
2	19.867118	−1.025060	385.96252652	0.0459	3591.615327	0.047166
3	2.513241	−4.555081	67.67315150	0.0704	174.058930	0.047166
4	1.563658	0.113150	18.20422450	0.0459	12.867208	0.047166
5	0.745149	−0.053347	3.19185713	0.0704	5.264475	0.047166
6	0.700361	0.166834	0.85861649	0.0459	4.905886	0.047166
7	0.661755	0.158981	0.15054644	0.0704	4.888973	0.047166
8	0.659643	0.169366	0.04049732	0.0459	4.888175	0.047166
9	0.657822	0.168996	0.00710064	0.0704	4.888137	0.047166
10	0.657722	0.169486	0.00191009	0.0459	4.888136	0.047166
11	0.657636	0.169468	0.00033491	0.0704	4.888136	0.047166
12	0.657632	0.169491	0.00009009	0.0459	4.888136	0.047161
13	0.657628	0.169490	0.00001580	0.0704	4.888136	0.047068
14	0.657627	0.169492	0.00000425	0.0459	4.888136	0.045002
15	0.657627	0.169491	0.00000075	0.0000	4.888136	0.000000

SLIDE 45



9.4 Empirical behavior

Slide 46

- The convergence constant bound is not just theoretical. It is typically experienced in practice.

- Analysis is due to Leonid Kantorovich, who won the Nobel Memorial Prize in Economic Science in 1975 for his contributions to optimization and economic planning.

Slide 47

- What about non-quadratic functions?
  - Suppose $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$
  - $\nabla^2 f(\boldsymbol{x}^*)$ is the Hessian of $f(\boldsymbol{x})$ at $\boldsymbol{x} = \boldsymbol{x}^*$
  - Rate of convergence will depend on $\kappa(\nabla^2 f(\boldsymbol{x}^*))$

10 Summary

Slide 48

1. Optimality Conditions
2. The steepest descent algorithm - Convergence
3. Rate of convergence of Steepest Descent

15.093J / 6.255J Optimization Methods

Fall 2009