Optimization methods
Lecture 5: Introduction to numerical optimization methods. Gradient descent

Alexandr Katrutsa

Modern State of Artificial Intelligence Masters Program
Moscow Institute of Physics and Technology

Brief reminder of the previous lectures

- How convexity of the problem helps to solve it
- Disciplined convex programming
- CVXPy
- Optimality conditions

Problem statement

$$
\begin{aligned}
&\min_{\mathbf{x} \in S} f_0(\mathbf{x}) \\
\text{s.t. } & f_j(\mathbf{x}) = 0,\ j = 1, \dots, m \\
& g_k(\mathbf{x}) \leq 0,\ k = 1, \dots, p
\end{aligned}
$$

where $S \subseteq \mathbb{R}^n,\ f_j : S \to \mathbb{R},\ j = 0, \dots, m,$
$g_k : S \to \mathbb{R},\ k = 1, \dots, p$

- All functions here are at least continuous
- **Nonlinear** optimization problems are numerically intractable in general case

Reminder of analytical results

First order necessary condition

If $\mathbf{x}^*$ is a local minimizer of differentiable function $f(\mathbf{x})$, then

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

Reminder of analytical results

First order necessary condition

If $\mathbf{x}^*$ is a local minimizer of differentiable function $f(\mathbf{x})$, then

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

Second order necessary condition

If $\mathbf{x}^*$ is a local minimizer of twice differentiable function $f(\mathbf{x})$, then

$$f'(\mathbf{x}^*) = 0 \quad \text{и} \quad f''(\mathbf{x}^*) \succeq 0$$

Reminder of analytical results

First order necessary condition

If $\mathbf{x}^*$ is a local minimizer of differentiable function $f(\mathbf{x})$, then

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

Second order necessary condition

If $\mathbf{x}^*$ is a local minimizer of twice differentiable function $f(\mathbf{x})$, then

$$f'(\mathbf{x}^*) = 0 \quad \text{и} \quad f''(\mathbf{x}^*) \succeq 0$$

Sufficient condition

Let $f(\mathbf{x})$ be twice differentiable functio, and $\mathbf{x}^*$ satisfies the following conditions

$$f'(\mathbf{x}^*) = 0 \quad f''(\mathbf{x}^*) \succ 0,$$

then $\mathbf{x}^*$ is a local minimizer of $f(\mathbf{x})$

Features of numerical solving

- Exact solution can not be obtained due to machine precision

Features of numerical solving

- Exact solution can not be obtained due to machine precision
- The stopping criterion is needed

Features of numerical solving

- Exact solution can not be obtained due to machine precision
- The stopping criterion is needed
- Information about the problem

General scheme

- Initial guess $\mathbf{x}_0$
- Desired tolerance ε
- Update solution approximation

Questions

1. What stopping criteria are possible?

Questions

1. What stopping criteria are possible?
2. How the next point is computed?

Stopping criteria

1. Convergence by argument:

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2 < \varepsilon$$

Stopping criteria

1. Convergence by argument:

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2 < \varepsilon$$

2. Convergence by objective:

$$\|f_k - f^*\|_2 < \varepsilon$$

Stopping criteria

1. Convergence by argument:

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2 < \varepsilon$$

2. Convergence by objective:

$$\|f_k - f^*\|_2 < \varepsilon$$

3. Necessary condition

$$\|f'(\mathbf{x}_k)\|_2 < \varepsilon$$

How to update point

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{h}_k$$

How to update point

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{h}_k$$

Line search

1. Find a direction $\mathbf{h}_k$
2. After that compute «optimal» value α_k

How to compare optimization methods?

For fixed class of problems the following ways of comparison are possible

1. Complexity
2. Convergence speed
3. Experiments

Convergence speed

1. Sublinear
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Ck^{\alpha},$$
where $\alpha < 0$ and $0 < C < \infty$

Convergence speed

1. Sublinear
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Ck^{\alpha},$$
where $\alpha < 0$ and $0 < C < \infty$
2. Linear (geometric)
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Cq^{k},$$
where $q \in (0, 1)$ и $0 < C < \infty$

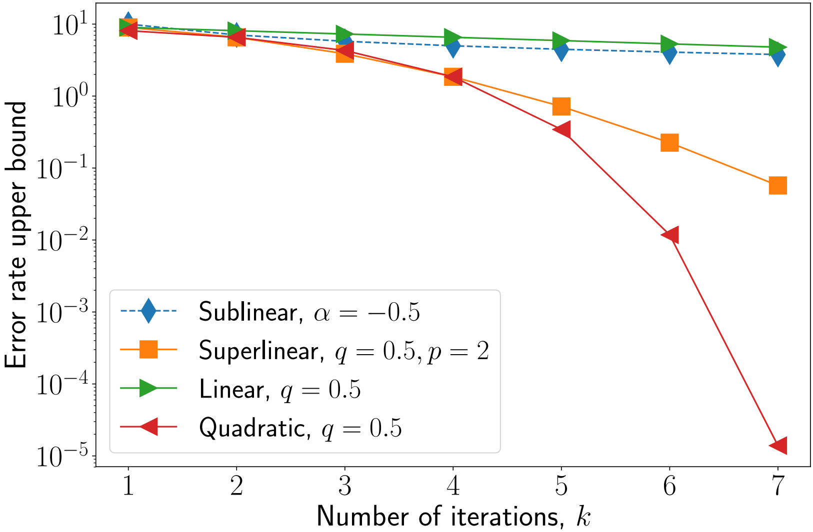
Convergence speed

1. Sublinear
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Ck^{\alpha},$$
where $\alpha < 0$ and $0 < C < \infty$
2. Linear (geometric)
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Cq^k,$$
where $q \in (0, 1)$ и $0 < C < \infty$
3. Super-linear
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Cq^{k^p},$$
where $q \in (0, 1)$, $0 < C < \infty$ and $p > 1$

Convergence speed

1. Sublinear
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Ck^{\alpha},$$
where $\alpha < 0$ and $0 < C < \infty$
2. Linear (geometric)
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Cq^{k},$$
where $q \in (0, 1)$ и $0 < C < \infty$
3. Super-linear
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Cq^{k^p},$$
where $q \in (0, 1)$, $0 < C < \infty$ and $p > 1$
4. Quadratic
$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq C\|\mathbf{x}_k - \mathbf{x}^*\|_2^2, \qquad \text{or} \qquad \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq Cq^{2^k}$$
where $q \in (0, 1)$ and $0 < C < \infty$

Comparison of convergence speed



Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable

Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable
  - convexity

Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable
  - convexity
  - smoothness

Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable
  - convexity
  - smoothness
- qualitative behavior of method

Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable
  - convexity
  - smoothness
- qualitative behavior of method
  - is initial guess significant for convergence or not

Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable
  - convexity
  - smoothness
- qualitative behavior of method
  - is initial guess significant for convergence or not
  - what type of convergence can we expect

Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable
  - convexity
  - smoothness
- qualitative behavior of method
  - is initial guess significant for convergence or not
  - what type of convergence can we expect
- convergence speed estimate

Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable
  - convexity
  - smoothness
- qualitative behavior of method
  - is initial guess significant for convergence or not
  - what type of convergence can we expect
- convergence speed estimate
  - theoretical estimate without any experiments

Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable
  - convexity
  - smoothness
- qualitative behavior of method
  - is initial guess significant for convergence or not
  - what type of convergence can we expect
- convergence speed estimate
  - theoretical estimate without any experiments
  - identification of factors that affect convergence

Convergence theorems interpretation

What profits can give convergence theorems

- class of problems for which the method is applicable
  - convexity
  - smoothness
- qualitative behavior of method
  - is initial guess significant for convergence or not
  - what type of convergence can we expect
- convergence speed estimate
  - theoretical estimate without any experiments
  - identification of factors that affect convergence
  - sometimes it is possible to set number of iterations in advance to achieve pre-defined tolerance

Convergence theorems interpretation

What facts convergence theorems do not establish

- if method converges, this fact does not signal of any practical gain of using this method

Convergence theorems interpretation

What facts convergence theorems do not establish

- if method converges, this fact does not signal of any practical gain of using this method
- convergence estimate depends on the unknown constants

Convergence theorems interpretation

What facts convergence theorems do not establish

- if method converges, this fact does not signal of any practical gain of using this method
- convergence estimate depends on the unknown constants
- rounding errors are out of scope

What classes of methods do we consider

What classes of methods do we consider

Method order

- Zero-order methods: oracle provides only the value of $f(\mathbf{x})$

What classes of methods do we consider

Method order

- Zero-order methods: oracle provides only the value of $f(\mathbf{x})$
- First-order methods: oracle provides objective function value $f(\mathbf{x})$ and gradient $f'(\mathbf{x})$

What classes of methods do we consider

Method order

- Zero-order methods: oracle provides only the value of $f(\mathbf{x})$
- First-order methods: oracle provides objective function value $f(\mathbf{x})$ and gradient $f'(\mathbf{x})$
- Second-order method: oracle provides objective function $f(\mathbf{x})$, gradient $f'(\mathbf{x})$ and hessian $f''(\mathbf{x})$.

What classes of methods do we consider

Method order

- Zero-order methods: oracle provides only the value of $f(\mathbf{x})$
- First-order methods: oracle provides objective function value $f(\mathbf{x})$ and gradient $f'(\mathbf{x})$
- Second-order method: oracle provides objective function $f(\mathbf{x})$, gradient $f'(\mathbf{x})$ and hessian $f''(\mathbf{x})$.

Q: do higher-order methods exist?

What classes of methods do we consider

Method order

- Zero-order methods: oracle provides only the value of $f(\mathbf{x})$
- First-order methods: oracle provides objective function value $f(\mathbf{x})$ and gradient $f'(\mathbf{x})$
- Second-order method: oracle provides objective function $f(\mathbf{x})$, gradient $f'(\mathbf{x})$ and hessian $f''(\mathbf{x})$.

Q: do higher-order methods exist?
A: yes, but they are still mostly under development to make them a technology not art. Original paper[1]

[1] Nesterov Y. Implementable tensor methods in unconstrained convex optimization //Mathematical Programming. – 2019. – P. 1-27.

How can we use previous points?

1. One-step methods

$$\mathbf{x}_{k+1} = \Phi(\mathbf{x}_k)$$

How can we use previous points?

1. One-step methods

$$\mathbf{x}_{k+1} = \Phi(\mathbf{x}_k)$$

2. Multi-step methods

$$\mathbf{x}_{k+1} = \Phi(\mathbf{x}_k, \mathbf{x}_{k-1}, \ldots)$$

Main facts from intro

- Why do we need numerical methods
- General scheme
- How can we compare optimization methods
- Zoo of problems and methods

Descent methods

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{h}_k$$

such that

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$$

Descent methods

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{h}_k$$

such that

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$$

Definition

Direction $\mathbf{h}_k$ is called *descent direction*

Descent methods

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{h}_k$$

such that

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$$

Definition

Direction $\mathbf{h}_k$ is called *descent direction*

Remark

There exist methods that do not require monotonic decreasing of the objective function but converge faster

L-smooth function: reminder

Definition

Let $L > 0$. A function f is called L-smooth if it is differentiable and satisfies

$$\|f'(\mathbf{x}) - f'(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2.$$

L-smooth function: reminder

Definition

Let $L > 0$. A function f is called L-smooth if it is differentiable and satisfies

$$\|f'(\mathbf{x}) - f'(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2.$$

- L is a Lipschitz constant of gradient

L-smooth function: reminder

Definition

Let $L > 0$. A function f is called L-smooth if it is differentiable and satisfies

$$\|f'(\mathbf{x}) - f'(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2.$$

- L is a Lipschitz constant of gradient
- The minimal possible $L > 0$ is especially important

L-smooth function: reminder

Definition

Let $L > 0$. A function f is called L-smooth if it is differentiable and satisfies

$$\|f'(\mathbf{x}) - f'(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2.$$

- L is a Lipschitz constant of gradient
- The minimal possible $L > 0$ is especially important

Descent lemma

Let f be an L-smooth function. Then for any $\mathbf{x}, \mathbf{y}$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$$

Gradient descent

Global upper bound for function f at point $\mathbf{x}_k$:

$$f(\mathbf{y}) \leq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{y} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}_k\|_2^2 \equiv g(\mathbf{y}),$$

where $\lambda_{\max}(f''(\mathbf{x})) \leq L$ for all feasible $\mathbf{x}$.

Gradient descent

Global upper bound for function f at point $\mathbf{x}_k$:

$$f(\mathbf{y}) \leq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{y} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}_k\|_2^2 \equiv g(\mathbf{y}),$$

where $\lambda_{\max}(f''(\mathbf{x})) \leq L$ for all feasible $\mathbf{x}$.
The right-hand side is a quadratic form such that its minimizer has analytical expression:

$$\begin{aligned}
& g'(\mathbf{y}^*) = 0 \\
& f'(\mathbf{x}_k) + L(\mathbf{y}^* - \mathbf{x}_k) = 0 \\
& \mathbf{y}^* = \mathbf{x}_k - \frac{1}{L} f'(\mathbf{x}_k) \equiv \mathbf{x}_{k+1}
\end{aligned}$$

This approach estimates step size as $\frac{1}{L}$.

Step size selection

- Constant step size $\alpha_k \equiv \text{const} < \frac{2}{L}$
- Decreasing sequence such that $\sum\limits_{k=1}^{\infty} \alpha_k = \infty$, e.g. $\frac{1}{k}, \frac{1}{\sqrt{k}}$
- Adaptive step size
- The steepest descent rule: the search of the best α_k

Important note

The best step size does not provide better theoretical convergence rate

Convergence to a stationary point

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 =$$

$$f(\mathbf{x}_k) - \alpha_k \|f'(\mathbf{x}_k)\|_2^2 + \frac{L\alpha_k^2}{2}\|f'(\mathbf{x}_k)\|_2^2 =$$

$$f(\mathbf{x}_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right)\|f'(\mathbf{x}_k)\|_2^2$$

Convergence to a stationary point

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 =$$

$$f(\mathbf{x}_k) - \alpha_k \|f'(\mathbf{x}_k)\|_2^2 + \frac{L\alpha_k^2}{2}\|f'(\mathbf{x}_k)\|_2^2 =$$

$$f(\mathbf{x}_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right)\|f'(\mathbf{x}_k)\|_2^2$$

- Descent condition: $\alpha_k - \frac{L\alpha_k^2}{2} > 0 \Rightarrow \alpha_k < \frac{2}{L}$

Convergence to a stationary point

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 =$$
$$f(\mathbf{x}_k) - \alpha_k \|f'(\mathbf{x}_k)\|_2^2 + \frac{L\alpha_k^2}{2}\|f'(\mathbf{x}_k)\|_2^2 =$$
$$f(\mathbf{x}_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right)\|f'(\mathbf{x}_k)\|_2^2$$

- Descent condition: $\alpha_k - \frac{L\alpha_k^2}{2} > 0 \Rightarrow \alpha_k < \frac{2}{L}$
- $\alpha_k^* = \arg\max\limits_{\alpha_k} \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) = \frac{1}{L}$

Convergence to a stationary point

$$
\begin{aligned}
f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 = \\
f(\mathbf{x}_k) - \alpha_k \|f'(\mathbf{x}_k)\|_2^2 + \frac{L\alpha_k^2}{2} \|f'(\mathbf{x}_k)\|_2^2 = \\
f(\mathbf{x}_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) \|f'(\mathbf{x}_k)\|_2^2
\end{aligned}
$$

- Descent condition: $\alpha_k - \frac{L\alpha_k^2}{2} > 0 \Rightarrow \alpha_k < \frac{2}{L}$
- $\alpha_k^* = \arg\max\limits_{\alpha_k} \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) = \frac{1}{L}$
- $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2L} \|f'(\mathbf{x}_k)\|_2^2$

Convergence to a stationary point

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 =$$

$$f(\mathbf{x}_k) - \alpha_k \|f'(\mathbf{x}_k)\|_2^2 + \frac{L\alpha_k^2}{2}\|f'(\mathbf{x}_k)\|_2^2 =$$

$$f(\mathbf{x}_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right)\|f'(\mathbf{x}_k)\|_2^2$$

- Descent condition: $\alpha_k - \frac{L\alpha_k^2}{2} > 0 \Rightarrow \alpha_k < \frac{2}{L}$
- $\alpha_k^* = \arg\max_{\alpha_k} \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) = \frac{1}{L}$
- $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2L}\|f'(\mathbf{x}_k)\|_2^2$
- $\frac{1}{2L}\sum_{k=0}^{T} \|f'(\mathbf{x}_k)\|_2^2 \leq f(\mathbf{x}_0) - f(\mathbf{x}_{T+1}) \leq f(\mathbf{x}_0) - f^*$

Convergence to a stationary point

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 =$$

$$f(\mathbf{x}_k) - \alpha_k \|f'(\mathbf{x}_k)\|_2^2 + \frac{L\alpha_k^2}{2}\|f'(\mathbf{x}_k)\|_2^2 =$$

$$f(\mathbf{x}_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right)\|f'(\mathbf{x}_k)\|_2^2$$

- Descent condition: $\alpha_k - \frac{L\alpha_k^2}{2} > 0 \Rightarrow \alpha_k < \frac{2}{L}$
- $\alpha_k^* = \arg\max_{\alpha_k} \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) = \frac{1}{L}$
- $f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2L}\|f'(\mathbf{x}_k)\|_2^2$
- $\frac{1}{2L}\sum_{k=0}^{T} \|f'(\mathbf{x}_k)\|_2^2 \leq f(\mathbf{x}_0) - f(\mathbf{x}_{T+1}) \leq f(\mathbf{x}_0) - f^*$
- f is bounded below, $\|f'(\mathbf{x}_k)\|_2 \to 0, \; k \to \infty$

Convergence: convex function

Theorem

Let f be convex function with Lipschitz gradient and $\alpha = \frac{1}{L}$, then gradient descent converges like

$$f(\mathbf{x}_{k+1}) - f^* \leq \frac{2L\|\mathbf{x} - \mathbf{x}_0\|_2^2}{k+4} = \mathcal{O}(1/k)$$

Convergence: strongly convex case

- As a consequence of strongly convexity

$$f(\mathbf{z}) \geq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{z} - \mathbf{x}_k \rangle + \frac{\mu}{2}\|\mathbf{z} - \mathbf{x}_k\|_2^2$$

Convergence: strongly convex case

- As a consequence of strongly convexity

$$f(\mathbf{z}) \geq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{z} - \mathbf{x}_k \rangle + \frac{\mu}{2} \|\mathbf{z} - \mathbf{x}_k\|_2^2$$

- Minimize both sides by $\mathbf{z}$

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) - \frac{1}{2\mu} \|f'(\mathbf{x}_k)\|_2^2, \quad \|f'(\mathbf{x}_k)\|_2^2 \geq 2\mu(f(\mathbf{x}_k) - f^*)$$

Convergence: strongly convex case

- As a consequence of strongly convexity

$$f(\mathbf{z}) \geq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{z} - \mathbf{x}_k \rangle + \frac{\mu}{2} \|\mathbf{z} - \mathbf{x}_k\|_2^2$$

- Minimize both sides by $\mathbf{z}$

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) - \frac{1}{2\mu} \|f'(\mathbf{x}_k)\|_2^2, \quad \|f'(\mathbf{x}_k)\|_2^2 \geq 2\mu(f(\mathbf{x}_k) - f^*)$$

- Remember that $\alpha_k \equiv \frac{1}{L}$

$$f^* \leq f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2L} \|f'(\mathbf{x}_k)\|_2^2$$

Convergence: strongly convex case

- As a consequence of strongly convexity

$$f(\mathbf{z}) \geq f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{z} - \mathbf{x}_k \rangle + \frac{\mu}{2} \|\mathbf{z} - \mathbf{x}_k\|_2^2$$

- Minimize both sides by $\mathbf{z}$

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) - \frac{1}{2\mu} \|f'(\mathbf{x}_k)\|_2^2, \quad \|f'(\mathbf{x}_k)\|_2^2 \geq 2\mu(f(\mathbf{x}_k) - f^*)$$

- Remember that $\alpha_k \equiv \frac{1}{L}$

$$f^* \leq f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2L} \|f'(\mathbf{x}_k)\|_2^2$$

- And finally get linear convergence

$$f(\mathbf{x}_{k+1}) - f^* \leq \left(1 - \frac{1}{\kappa}\right) (f(\mathbf{x}_k) - f^*)$$

Theorem about strongly convex functions

Theorem

Let f be L-smooth and μ-strongly convex, $\alpha_k = \frac{2}{\mu+L}$, then gradient descent converges like

$$f(\mathbf{x}_k) - f^* \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

What affect the linear convergence?

$$q^* = \frac{L-\mu}{L+\mu} = \frac{L/\mu - 1}{L/\mu + 1} = \frac{\kappa - 1}{\kappa + 1},$$

where κ is an estimate of condition number of $f''(\mathbf{x})$.

What affect the linear convergence?

$$q^* = \frac{L - \mu}{L + \mu} = \frac{L/\mu - 1}{L/\mu + 1} = \frac{\kappa - 1}{\kappa + 1},$$

where κ is an estimate of condition number of $f''(\mathbf{x})$.

- If $\kappa \gg 1$, $q^* \to 1 \Rightarrow$, we have very slow convergence. For example, if $\kappa = 100$, then $q^* \approx 0.98$

What affect the linear convergence?

$$q^* = \frac{L-\mu}{L+\mu} = \frac{L/\mu - 1}{L/\mu + 1} = \frac{\kappa - 1}{\kappa + 1},$$

where κ is an estimate of condition number of $f''(\mathbf{x})$.

- If $\kappa \gg 1$, $q^* \to 1 \Rightarrow$, we have very slow convergence. For example, if $\kappa = 100$, then $q^* \approx 0.98$
- If $\kappa \simeq 1$, $q^* \to 0 \Rightarrow$ implies acceleration of convergence. For example, if $\kappa = 4$, then $q^* = 0.6$

Can we do better?

What do we know by now?

- Gradient descent converges like $\mathcal{O}(1/k)$ for convex L-smooth functions
- Gradient descent converges linearly with factor $q = \frac{\kappa-1}{\kappa+1}$ for strongly convex L-smooth functions

Q: do methods with faster convergence exist and how can we derive them?

Take home message

- General scheme of numerical optimization methods
- Convergence speed
- Gradient descent
- Properties and convergence theorems