Optimization methods
Lecture 7: Conjugate gradient method, heavy-ball method and accelerated gradient method

Alexandr Katrutsa

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

Brief reminder of the previous lecture

- Introduction to numerical optimization methods
- Convergence speed
- Gradient descent
- Convergence and condition number

Conjugate gradient method

- Consider the problem

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

where $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbf{b}^\top\mathbf{x}$ and $\mathbf{A} \in \mathbb{S}^n_{++}$

Conjugate gradient method

- Consider the problem

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

  where $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$ and $\mathbf{A} \in \mathbb{S}^n_{++}$
- From the FOOC follows that

$$\mathbf{A}\mathbf{x}^* = \mathbf{b}$$

Conjugate gradient method

- Consider the problem

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

  where $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$ and $\mathbf{A} \in \mathbb{S}^n_{++}$
- From the FOOC follows that

$$\mathbf{A}\mathbf{x}^* = \mathbf{b}$$

- Denote $f'(\mathbf{x}_k) = \mathbf{A}\mathbf{x}_k - \mathbf{b} = \mathbf{r}_k$

Conjugate gradient method

- Consider the problem

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

  where $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$ and $\mathbf{A} \in \mathbb{S}^n_{++}$
- From the FOOC follows that

$$\mathbf{A}\mathbf{x}^* = \mathbf{b}$$

- Denote $f'(\mathbf{x}_k) = \mathbf{A}\mathbf{x}_k - \mathbf{b} = \mathbf{r}_k$
- We reduce optimization problem to the problem of solving linear system

Motivation

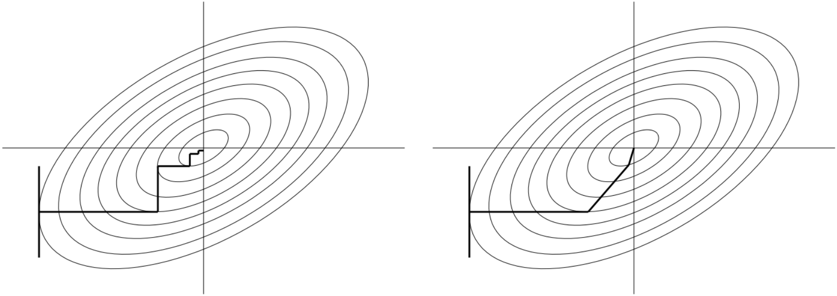
- Gradient descent convergence depends on the condition number

Motivation

- Gradient descent convergence depends on the condition number
- How to develop the method that converges at most after n iterations independently on the condition number?

Motivation

- Gradient descent convergence depends on the condition number
- How to develop the method that converges at most after n iterations independently on the condition number?



Plot is from this page

Conjugate directions

Definition

Non-zero vectors $\{\mathbf{p}_0, \dots, \mathbf{p}_l\}$ are called conjugate with respect to matrix $\mathbf{A} \in \mathbb{S}^n_{++}$, if

$$\mathbf{p}_i^\top \mathbf{A} \mathbf{p}_j = 0, \qquad i \neq j$$

Conjugate directions

Definition

Non-zero vectors $\{\mathbf{p}_0, \ldots, \mathbf{p}_l\}$ are called conjugate with respect to matrix $\mathbf{A} \in \mathbb{S}^n_{++}$, if

$$\mathbf{p}_i^\top \mathbf{A} \mathbf{p}_j = 0, \qquad i \neq j$$

Properties

Conjugate directions

Definition

Non-zero vectors $\{\mathbf{p}_0, \ldots, \mathbf{p}_l\}$ are called conjugate with respect to matrix $\mathbf{A} \in \mathbb{S}^n_{++}$, if

$$\mathbf{p}_i^\top \mathbf{A} \mathbf{p}_j = 0, \qquad i \neq j$$

Properties

- linear independence

Conjugate directions

Definition

Non-zero vectors $\{\mathbf{p}_0, \ldots, \mathbf{p}_l\}$ are called conjugate with respect to matrix $\mathbf{A} \in \mathbb{S}^n_{++}$, if

$$\mathbf{p}_i^\top \mathbf{A} \mathbf{p}_j = 0, \qquad i \neq j$$

Properties

- linear independence
- conjugate directions + the steepest descent step size selection = method that converges after n iterations

Conjugate directions

Definition

Non-zero vectors $\{\mathbf{p}_0, \ldots, \mathbf{p}_l\}$ are called conjugate with respect to matrix $\mathbf{A} \in \mathbb{S}^n_{++}$, if

$$\mathbf{p}_i^\top \mathbf{A} \mathbf{p}_j = 0, \qquad i \neq j$$

Properties

- linear independence
- conjugate directions + the steepest descent step size selection = method that converges after n iterations
- $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \to \mathbf{r}_{k+1} = \mathbf{r}_k + \alpha_k \mathbf{A} \mathbf{p}_k$

Conjugate directions

Definition

Non-zero vectors $\{\mathbf{p}_0, \ldots, \mathbf{p}_l\}$ are called conjugate with respect to matrix $\mathbf{A} \in \mathbb{S}^n_{++}$, if

$$\mathbf{p}_i^\top \mathbf{A}\mathbf{p}_j = 0, \qquad i \neq j$$

Properties

- linear independence
- conjugate directions + the steepest descent step size selection = method that converges after n iterations
- $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \to \mathbf{r}_{k+1} = \mathbf{r}_k + \alpha_k \mathbf{A}\mathbf{p}_k$

Q: how to derive conjugate directions from the set of linear independent vectors?

Convergence

Theorem

Assume $\mathbf{x}_k$ is generated by conjugate direction method. Then

1. $\langle \mathbf{r}_k, \mathbf{p}_i \rangle = 0,\ i = 1, \ldots, k-1$
2. $\mathbf{x}_k = \arg\min_{\mathbf{x} \in P} f(\mathbf{x})$, where $P = \mathbf{x}_0 + \mathtt{span}(\mathbf{p}_0, \ldots, \mathbf{p}_{k-1})$

Proof

Convergence

Theorem

Assume $\mathbf{x}_k$ is generated by conjugate direction method. Then

1. $\langle \mathbf{r}_k, \mathbf{p}_i \rangle = 0, \; i = 1, \ldots, k-1$
2. $\mathbf{x}_k = \arg\min_{\mathbf{x} \in P} f(\mathbf{x})$, where $P = \mathbf{x}_0 + \mathtt{span}(\mathbf{p}_0, \ldots, \mathbf{p}_{k-1})$

Proof

1. $\phi(\gamma) = f(\mathbf{x}_0 + \gamma_0 \mathbf{p}_0 + \ldots + \gamma_{k-1} \mathbf{p}_{k-1})$

Convergence

Theorem

Assume $\mathbf{x}_k$ is generated by conjugate direction method. Then

1. $\langle \mathbf{r}_k, \mathbf{p}_i \rangle = 0, \ i = 1, \ldots, k-1$
2. $\mathbf{x}_k = \arg\min_{\mathbf{x} \in P} f(\mathbf{x})$, where $P = \mathbf{x}_0 + \mathtt{span}(\mathbf{p}_0, \ldots, \mathbf{p}_{k-1})$

Proof

1. $\phi(\gamma) = f(\mathbf{x}_0 + \gamma_0 \mathbf{p}_0 + \ldots + \gamma_{k-1} \mathbf{p}_{k-1})$
2. $\phi(\gamma)$is strictly convex $\rightarrow$ there exists γ^*

Convergence

Theorem

Assume $\mathbf{x}_k$ is generated by conjugate direction method. Then

1. $\langle \mathbf{r}_k, \mathbf{p}_i \rangle = 0, \; i = 1, \dots, k-1$
2. $\mathbf{x}_k = \arg\min_{\mathbf{x} \in P} f(\mathbf{x})$, where $P = \mathbf{x}_0 + \mathtt{span}(\mathbf{p}_0, \dots, \mathbf{p}_{k-1})$

Proof

1. $\phi(\gamma) = f(\mathbf{x}_0 + \gamma_0 \mathbf{p}_0 + \ldots + \gamma_{k-1} \mathbf{p}_{k-1})$
2. $\phi(\gamma)$is strictly convex $\to$ there exists γ^*
3. According to the FOOC $\phi'(\gamma^*) = \langle f'(\mathbf{x}_0 + \gamma_0^* \mathbf{p}_0 + \ldots + \gamma_{k-1}^* \mathbf{p}_{k-1}), \mathbf{p}_i \rangle = 0, \; i = 0, \dots, k-1$

Convergence

Theorem

Assume $\mathbf{x}_k$ is generated by conjugate direction method. Then

1. $\langle \mathbf{r}_k, \mathbf{p}_i \rangle = 0, \; i = 1, \ldots, k-1$
2. $\mathbf{x}_k = \underset{\mathbf{x} \in P}{\arg\min} f(\mathbf{x})$, where $P = \mathbf{x}_0 + \mathtt{span}(\mathbf{p}_0, \ldots, \mathbf{p}_{k-1})$

Proof

1. $\phi(\gamma) = f(\mathbf{x}_0 + \gamma_0 \mathbf{p}_0 + \ldots + \gamma_{k-1} \mathbf{p}_{k-1})$
2. $\phi(\gamma)$is strictly convex $\to$ there exists γ^*
3. According to the FOOC
   $\phi'(\gamma^*) = \langle f'(\mathbf{x}_0 + \gamma_0^* \mathbf{p}_0 + \ldots + \gamma_{k-1}^* \mathbf{p}_{k-1}), \mathbf{p}_i \rangle = 0, \; i = 0, \ldots, k-1$
4. From the definition of $\mathbf{r}_k$ follows
   $\langle \mathbf{r}_k, \mathbf{p}_i \rangle = 0, \; i = 0, \ldots, k-1$

Convergence

Theorem

Assume $\mathbf{x}_k$ is generated by conjugate direction method. Then

1. $\langle \mathbf{r}_k, \mathbf{p}_i \rangle = 0, \; i = 1, \ldots, k-1$
2. $\mathbf{x}_k = \arg\min\limits_{\mathbf{x} \in P} f(\mathbf{x})$, where $P = \mathbf{x}_0 + \mathtt{span}(\mathbf{p}_0, \ldots, \mathbf{p}_{k-1})$

Proof

1. $\phi(\gamma) = f(\mathbf{x}_0 + \gamma_0 \mathbf{p}_0 + \ldots + \gamma_{k-1} \mathbf{p}_{k-1})$
2. $\phi(\gamma)$is strictly convex $\to$ there exists γ^*
3. According to the FOOC
   $\phi'(\gamma^*) = \langle f'(\mathbf{x}_0 + \gamma_0^* \mathbf{p}_0 + \ldots + \gamma_{k-1}^* \mathbf{p}_{k-1}), \mathbf{p}_i \rangle = 0, \; i = 0, \ldots, k-1$
4. From the definition of $\mathbf{r}_k$ follows
   $\langle \mathbf{r}_k, \mathbf{p}_i \rangle = 0, \; i = 0, \ldots, k-1$
5. Thus, $(1) \Leftrightarrow (2)$

6. Proof (1) by induction:

6. Proof (1) by induction:
   - base: $\langle \mathbf{r}_1, \mathbf{p}_0 \rangle = 0$ by construction

6. Proof (1) by induction:
   - base: $\langle \mathbf{r}_1, \mathbf{p}_0 \rangle = 0$ by construction
   - hypothesis: $\langle \mathbf{r}_{k-1}, \mathbf{p}_i \rangle = 0, \ i = 1, \ldots, k-2$

6. Proof (1) by induction:
   - base: $\langle \mathbf{r}_1, \mathbf{p}_0 \rangle = 0$ by construction
   - hypothesis: $\langle \mathbf{r}_{k-1}, \mathbf{p}_i \rangle = 0, \; i = 1, \ldots, k-2$
7. $\mathbf{r}_k = \mathbf{r}_{k-1} + \alpha_{k-1} \mathbf{A} \mathbf{p}_{k-1}$

6. Proof (1) by induction:
   - base: $\langle \mathbf{r}_1, \mathbf{p}_0 \rangle = 0$ by construction
   - hypothesis: $\langle \mathbf{r}_{k-1}, \mathbf{p}_i \rangle = 0, \ i = 1, \ldots, k-2$
7. $\mathbf{r}_k = \mathbf{r}_{k-1} + \alpha_{k-1} \mathbf{A} \mathbf{p}_{k-1}$
8. $\langle \mathbf{p}_{k-1}, \mathbf{r}_k \rangle = \langle \mathbf{p}_{k-1}, \mathbf{r}_{k-1} \rangle + \alpha_{k-1} \langle \mathbf{p}_{k-1}, \mathbf{A} \mathbf{p}_{k-1} \rangle = 0$ by construction α_{k-1}

6. Proof (1) by induction:
   - base: $\langle \mathbf{r}_1, \mathbf{p}_0 \rangle = 0$ by construction
   - hypothesis: $\langle \mathbf{r}_{k-1}, \mathbf{p}_i \rangle = 0, \ i = 1, \ldots, k-2$
7. $\mathbf{r}_k = \mathbf{r}_{k-1} + \alpha_{k-1}\mathbf{A}\mathbf{p}_{k-1}$
8. $\langle \mathbf{p}_{k-1}, \mathbf{r}_k \rangle = \langle \mathbf{p}_{k-1}, \mathbf{r}_{k-1} \rangle + \alpha_{k-1}\langle \mathbf{p}_{k-1}, \mathbf{A}\mathbf{p}_{k-1} \rangle = 0$ by construction α_{k-1}
9. $\langle \mathbf{p}_i, \mathbf{r}_k \rangle = \langle \mathbf{p}_i, \mathbf{r}_{k-1} \rangle + \alpha_{k-1}\langle \mathbf{p}_i, \mathbf{A}\mathbf{p}_{k-1} \rangle, \ i = 1, \ldots, k-2$

6. Proof (1) by induction:
   - base: $\langle \mathbf{r}_1, \mathbf{p}_0 \rangle = 0$ by construction
   - hypothesis: $\langle \mathbf{r}_{k-1}, \mathbf{p}_i \rangle = 0,\ i = 1, \ldots, k-2$
7. $\mathbf{r}_k = \mathbf{r}_{k-1} + \alpha_{k-1} \mathbf{A} \mathbf{p}_{k-1}$
8. $\langle \mathbf{p}_{k-1}, \mathbf{r}_k \rangle = \langle \mathbf{p}_{k-1}, \mathbf{r}_{k-1} \rangle + \alpha_{k-1} \langle \mathbf{p}_{k-1}, \mathbf{A} \mathbf{p}_{k-1} \rangle = 0$ by construction α_{k-1}
9. $\langle \mathbf{p}_i, \mathbf{r}_k \rangle = \langle \mathbf{p}_i, \mathbf{r}_{k-1} \rangle + \alpha_{k-1} \langle \mathbf{p}_i, \mathbf{A} \mathbf{p}_{k-1} \rangle,\ i = 1, \ldots, k-2$
10. $\langle \mathbf{p}_i, \mathbf{r}_{k-1} \rangle = 0$ according to hypothesis

6. Proof (1) by induction:
   - base: $\langle \mathbf{r}_1, \mathbf{p}_0 \rangle = 0$ by construction
   - hypothesis: $\langle \mathbf{r}_{k-1}, \mathbf{p}_i \rangle = 0,\ i = 1, \ldots, k-2$
7. $\mathbf{r}_k = \mathbf{r}_{k-1} + \alpha_{k-1} \mathbf{A} \mathbf{p}_{k-1}$
8. $\langle \mathbf{p}_{k-1}, \mathbf{r}_k \rangle = \langle \mathbf{p}_{k-1}, \mathbf{r}_{k-1} \rangle + \alpha_{k-1} \langle \mathbf{p}_{k-1}, \mathbf{A} \mathbf{p}_{k-1} \rangle = 0$ by construction α_{k-1}
9. $\langle \mathbf{p}_i, \mathbf{r}_k \rangle = \langle \mathbf{p}_i, \mathbf{r}_{k-1} \rangle + \alpha_{k-1} \langle \mathbf{p}_i, \mathbf{A} \mathbf{p}_{k-1} \rangle,\ i = 1, \ldots, k-2$
10. $\langle \mathbf{p}_i, \mathbf{r}_{k-1} \rangle = 0$ according to hypothesis
11. $\langle \mathbf{p}_i, \mathbf{A} \mathbf{p}_{k-1} \rangle = 0$ by the conjugacy of $\{\mathbf{p}_i\}$

Conjugate gradients

- $\mathbf{p}_0 = -\mathbf{r}_0$

Conjugate gradients

- $\mathbf{p}_0 = -\mathbf{r}_0$
- $\mathbf{p}_{k+1} = -\mathbf{r}_{k+1} + \beta_{k+1}\mathbf{p}_k$, where β_{k+1} guarantees that $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$ are conjugate:

$$\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_{k+1} = \mathbf{p}_k^\top \mathbf{A}(-\mathbf{r}_{k+1} + \beta_{k+1}\mathbf{p}_k) = 0$$

$$\beta_{k+1} = \frac{\mathbf{p}_k^\top \mathbf{A}\mathbf{r}_{k+1}}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k}$$

Pseudocode: basic version

```
def ConjugateGradientQuadratic(x0, A, b, eps):
    r = A.dot(x0) - b
    p = -r
    while np.linalg.norm(r) > eps:
        alpha = -r.dot(p) / p.dot(A.dot(p))
        x = x + alpha * p
        r = A.dot(x) - b
        beta = r.dot(A.dot(p)) / p.dot(A.dot(p))
        p = -r + beta * p
    return x
```

Modifications of basic version

- How to compute α_k:

$$\alpha_k = -\frac{\mathbf{r}_k^\top \mathbf{p}_k}{\mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k} = -\frac{\mathbf{r}_k^\top (-\mathbf{r}_k + \beta_k \mathbf{p}_{k-1})}{\mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k} = \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{p}_k^\top \mathbf{A} \mathbf{p}_k}$$

Modifications of basic version

- How to compute α_k:

$$\alpha_k = -\frac{\mathbf{r}_k^\top \mathbf{p}_k}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k} = -\frac{\mathbf{r}_k^\top(-\mathbf{r}_k + \beta_k \mathbf{p}_{k-1})}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k} = \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k}$$

- How to compute β_k:

$$\beta_{k+1} = \frac{\mathbf{r}_{k+1}^\top \mathbf{A}\mathbf{p}_k}{\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_k} = \frac{\mathbf{r}_{k+1}^\top(\mathbf{r}_{k+1} - \mathbf{r}_k)}{(-\mathbf{r}_k + \beta_k \mathbf{p}_{k-1})^\top(\mathbf{r}_{k+1} - \mathbf{r}_k)} = \frac{\|\mathbf{r}_{k+1}\|_2^2}{\|\mathbf{r}_k\|_2^2}$$

Pseudocode: faster version

```
def ConjugateGradientQuadratic(x0, A, b, eps)
    r = A.dot(x0) - b
    p = -r
    while np.linalg.norm(r) > eps:
        alpha = r.dot(r) / p.dot(A.dot(p))
        x = x + alpha * p
        r_next = r + alpha * A.dot(p)
        beta = r_next.dot(r_next) / r.dot(r)
        p = -r_next + beta * p
        r = r_next
    return x
```

Why conjugate gradients are conjugate?

Theorem

Assume that after k iterations $\mathbf{x}_k \neq \mathbf{x}^*$. Then

1. $\langle \mathbf{r}_k, \mathbf{r}_i \rangle = 0, \; i = 1, \dots k-1$
2. $\mathtt{span}(\mathbf{r}_0, \dots, \mathbf{r}_k) = \mathtt{span}(\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \dots, \mathbf{A}^k\mathbf{r}_0)$
3. $\mathtt{span}(\mathbf{p}_0, \dots, \mathbf{p}_k) = \mathtt{span}(\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \dots, \mathbf{A}^k\mathbf{r}_0)$
4. $\mathbf{p}_k^\top \mathbf{A}\mathbf{p}_i = 0$, $i = 1, \dots, k-1$

Krylov subspace

Definition

A subspace $\mathcal{K}_k(\mathbf{A}) = \mathtt{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b})$ is called Krylov subspace.

Krylov subspace

Definition

A subspace $\mathcal{K}_k(\mathbf{A}) = \mathtt{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b})$ is called Krylov subspace.

The main property

$$\mathbf{A}^{-1}\mathbf{b} \in \mathcal{K}_n(\mathbf{A})$$

Krylov subspace

Definition

A subspace $\mathcal{K}_k(\mathbf{A}) = \mathtt{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b})$ is called Krylov subspace.

The main property

$$\mathbf{A}^{-1}\mathbf{b} \in \mathcal{K}_n(\mathbf{A})$$

Proof

Krylov subspace

Definition

A subspace $\mathcal{K}_k(\mathbf{A}) = \mathtt{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b})$ is called Krylov subspace.

The main property

$$\mathbf{A}^{-1}\mathbf{b} \in \mathcal{K}_n(\mathbf{A})$$

Proof

- According to the Cayley–Hamilton theorem: $p(\mathbf{A}) = 0$, where $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$

Krylov subspace

Definition

A subspace $\mathcal{K}_k(\mathbf{A}) = \mathtt{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b})$ is called Krylov subspace.

The main property

$$\mathbf{A}^{-1}\mathbf{b} \in \mathcal{K}_n(\mathbf{A})$$

Proof

- According to the Cayley–Hamilton theorem: $p(\mathbf{A}) = 0$, where $p(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I})$
- $p(\mathbf{A})\mathbf{b} = \mathbf{A}^n\mathbf{b} + a_1\mathbf{A}^{n-1}\mathbf{b} + \ldots + a_{n-1}\mathbf{A}\mathbf{b} + a_n\mathbf{b} = 0$

Krylov subspace

Definition

A subspace $\mathcal{K}_k(\mathbf{A}) = \mathtt{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b})$ is called Krylov subspace.

The main property

$$\mathbf{A}^{-1}\mathbf{b} \in \mathcal{K}_n(\mathbf{A})$$

Proof

- According to the Cayley–Hamilton theorem: $p(\mathbf{A}) = 0$, where $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$
- $p(\mathbf{A})\mathbf{b} = \mathbf{A}^n\mathbf{b} + a_1\mathbf{A}^{n-1}\mathbf{b} + \ldots + a_{n-1}\mathbf{A}\mathbf{b} + a_n\mathbf{b} = 0$
- $\mathbf{A}^{-1}p(\mathbf{A})\mathbf{b} =$
  $\mathbf{A}^{n-1}\mathbf{b} + a_1\mathbf{A}^{n-2}\mathbf{b} + \ldots + a_{n-1}\mathbf{b} + a_n\mathbf{A}^{-1}\mathbf{b} = 0$

Krylov subspace

Definition

A subspace $\mathcal{K}_k(\mathbf{A}) = \mathtt{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b})$ is called Krylov subspace.

The main property

$$\mathbf{A}^{-1}\mathbf{b} \in \mathcal{K}_n(\mathbf{A})$$

Proof

- According to the Cayley–Hamilton theorem: $p(\mathbf{A}) = 0$, where $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$
- $p(\mathbf{A})\mathbf{b} = \mathbf{A}^n\mathbf{b} + a_1\mathbf{A}^{n-1}\mathbf{b} + \ldots + a_{n-1}\mathbf{A}\mathbf{b} + a_n\mathbf{b} = 0$
- $\mathbf{A}^{-1}p(\mathbf{A})\mathbf{b} =$ $\mathbf{A}^{n-1}\mathbf{b} + a_1\mathbf{A}^{n-2}\mathbf{b} + \ldots + a_{n-1}\mathbf{b} + a_n\mathbf{A}^{-1}\mathbf{b} = 0$
- $\mathbf{A}^{-1}\mathbf{b} = -\frac{1}{a_n}(\mathbf{A}^{n-1}\mathbf{b} + a_1\mathbf{A}^{n-2}\mathbf{b} + \ldots + a_{n-1}\mathbf{b})$

Interpretation

- Search of the best approximation in the k-th Krylov subspace

$$\mathbf{x}_k = \arg\min_{\mathbf{x} \in \mathcal{K}_k} f(\mathbf{x})$$

Interpretation

- Search of the best approximation in the k-th Krylov subspace
$$\mathbf{x}_k = \arg\min_{\mathbf{x} \in \mathcal{K}_k} f(\mathbf{x})$$
- Directions $\{\mathbf{p}_i\} \neq \{\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b}\}$. Why?

Interpretation

- Search of the best approximation in the k-th Krylov subspace

$$\mathbf{x}_k = \arg\min_{\mathbf{x} \in \mathcal{K}_k} f(\mathbf{x})$$

- Directions $\{\mathbf{p}_i\} \neq \{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}\}$. Why?

Brief description of the method

Search of the solution in the orthonormal Krylov basis

Convergence by f and by $\mathbf{x}$

- Solution: $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$

Convergence by f and by $\mathbf{x}$

- Solution: $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$
- The minimal objective:

$$f^* = \frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-\top}\mathbf{A}\mathbf{A}^{-1}\mathbf{b} - \mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b} = -\frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b} = -\frac{1}{2}\|\mathbf{x}^*\|_{\mathbf{A}}^2$$

Convergence by f and by $\mathbf{x}$

- Solution: $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$
- The minimal objective:

$$f^* = \frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-\top}\mathbf{A}\mathbf{A}^{-1}\mathbf{b} - \mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b} = -\frac{1}{2}\mathbf{b}^\top \mathbf{A}^{-1}\mathbf{b} = -\frac{1}{2}\|\mathbf{x}^*\|_{\mathbf{A}}^2$$

- Convergence by objective:

$$\begin{aligned} f(\mathbf{x}) - f^* &= \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} + \frac{1}{2}\|\mathbf{x}^*\|_{\mathbf{A}}^2 \\ &= \frac{1}{2}\|\mathbf{x}\|_{\mathbf{A}}^2 - \mathbf{x}^\top \mathbf{A}\mathbf{x}^* + \frac{1}{2}\|\mathbf{x}^*\|_{\mathbf{A}}^2 \\ &= \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{A}}^2 \end{aligned}$$

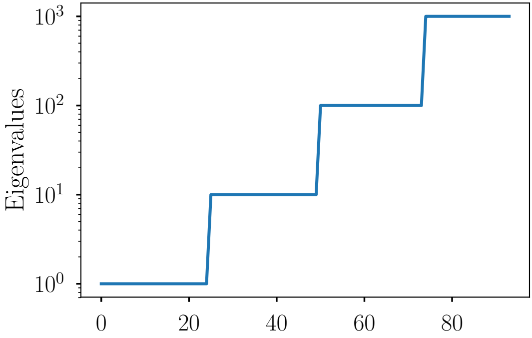
Convergence

Theorem

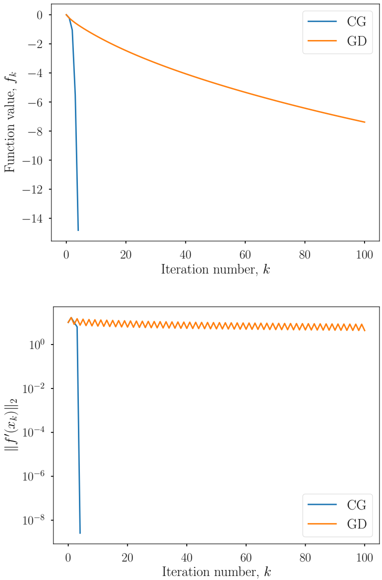
If matrix $\mathbf{A}$ has only m different eigenvalues, then conjugate gradient method converges in m iterations.

Example

- $n = 100$
- Spectrum of $\mathbf{A}$: $\{1, 10, 100, 1000\}$
- $\kappa = 1000$



Convergence plot



Other estimates

- If no information about spectrum, then

$$f_k - f^* \leq C \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} - 1} \right)^k$$

Non-linear conjugate gradient method

1. Adaptive step size selection α_k

Non-linear conjugate gradient method

1. Adaptive step size selection α_k
2. Coefficient β_k is found with gradients $f'(\mathbf{x}_{k-1}), f'(\mathbf{x}_{k-2})$

Non-linear conjugate gradient method

1. Adaptive step size selection α_k
2. Coefficient β_k is found with gradients $f'(\mathbf{x}_{k-1}), f'(\mathbf{x}_{k-2})$

Examples

Non-linear conjugate gradient method

1. Adaptive step size selection α_k
2. Coefficient β_k is found with gradients $f'(\mathbf{x}_{k-1}), f'(\mathbf{x}_{k-2})$

Examples

- Fletcher-Reeves method

$$\beta_k = \frac{\|f'(\mathbf{x}_{k-1})\|_2^2}{\|f'(\mathbf{x}_{k-2})\|_2^2}$$

Non-linear conjugate gradient method

1. Adaptive step size selection α_k
2. Coefficient β_k is found with gradients $f'(\mathbf{x}_{k-1}), f'(\mathbf{x}_{k-2})$

Examples

- Fletcher-Reeves method

$$\beta_k = \frac{\|f'(\mathbf{x}_{k-1})\|_2^2}{\|f'(\mathbf{x}_{k-2})\|_2^2}$$

- Polak-Ribière method

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

Non-linear conjugate gradient method

1. Adaptive step size selection α_k
2. Coefficient β_k is found with gradients $f'(\mathbf{x}_{k-1}), f'(\mathbf{x}_{k-2})$

Examples

- Fletcher-Reeves method

$$\beta_k = \frac{\|f'(\mathbf{x}_{k-1})\|_2^2}{\|f'(\mathbf{x}_{k-2})\|_2^2}$$

- Polak-Ribière method

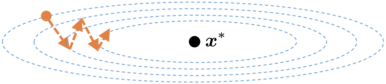
$$\beta_k = \frac{\langle f'(\mathbf{x}_{k-1}), f'(\mathbf{x}_{k-1}) - f'(\mathbf{x}_{k-2})\rangle}{\|f'(\mathbf{x}_{k-2})\|_2^2}$$

- Hestenes-Stiefel method

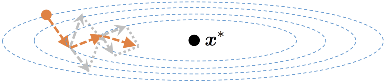
$$\beta_k = \frac{\langle f'(\mathbf{x}_{k-1}), f'(\mathbf{x}_{k-1}) - f'(\mathbf{x}_{k-2})\rangle}{\langle \mathbf{p}_{k-1}, f'(\mathbf{x}_{k-1}) - f'(\mathbf{x}_{k-2})\rangle}$$

Heavy-ball method (B.T. Polyak, 1964)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k f'(\mathbf{x}_k) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})$$



gradient descent



heavy-ball method

This plot is from this slides

- Two-step non-monotone method
- CG is a particular case

Convergence

Theorem

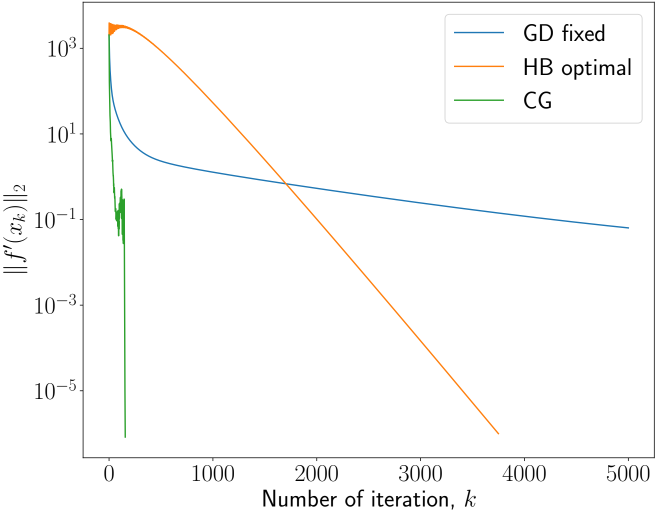
If f is L-smooth strongly convex function, then $\alpha_k = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta_k = \max(|1-\sqrt{\alpha_k L}|, |1-\sqrt{\alpha_k \mu}|)^2$ gives

$$\left\| \begin{bmatrix} \mathbf{x}_{k+1} - \mathbf{x}^* \\ \mathbf{x}_k - \mathbf{x}^* \end{bmatrix} \right\|_2 \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k \left\| \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}^* \\ \mathbf{x}_0 - \mathbf{x}^* \end{bmatrix} \right\|_2$$

- Parameters depend on L and μ
- Faster than gradient descent
- Analogue of CG for non-quadratic but strongly convex function

Example

- $n = 100$
- Random strongly convex quadratic problem



Accelerated gradient method (Nesterov, 1983)

One of the form

$$
\begin{aligned}
\mathbf{y}_0 &= \mathbf{x}_0 \\
\mathbf{x}_{k+1} &= \mathbf{y}_k - \alpha_k f'(\mathbf{y}_k) \\
\mathbf{y}_{k+1} &= \mathbf{x}_{k+1} + \frac{k}{k+3}(\mathbf{x}_{k+1} - \mathbf{x}_k)
\end{aligned}
$$

Accelerated gradient method (Nesterov, 1983)

One of the form

$$\begin{aligned}
\mathbf{y}_0 &= \mathbf{x}_0 \\
\mathbf{x}_{k+1} &= \mathbf{y}_k - \alpha_k f'(\mathbf{y}_k) \\
\mathbf{y}_{k+1} &= \mathbf{x}_{k+1} + \frac{k}{k+3}(\mathbf{x}_{k+1} - \mathbf{x}_k)
\end{aligned}$$

- Comparison with heavy-ball method

Accelerated gradient method (Nesterov, 1983)

One of the form

$$\mathbf{y}_0 = \mathbf{x}_0$$
$$\mathbf{x}_{k+1} = \mathbf{y}_k - \alpha_k f'(\mathbf{y}_k)$$
$$\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \frac{k}{k+3}(\mathbf{x}_{k+1} - \mathbf{x}_k)$$

- Comparison with heavy-ball method
- Non-monotone

Convergence

Theorem

If f is convex and L-smooth and step size $\alpha_k = \frac{1}{L}$, then accelerated gradient method converges as

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

Convergence

Theorem

If f is convex and L-smooth and step size $\alpha_k = \frac{1}{L}$, then accelerated gradient method converges as

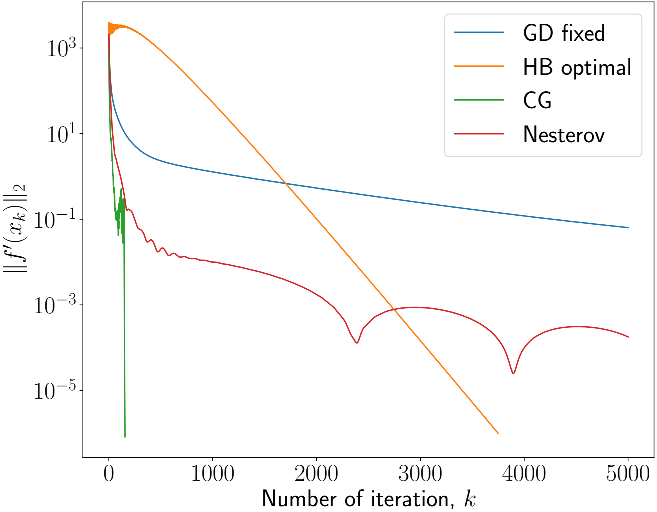
$$f(\mathbf{x}_k) - f^* \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{(k+1)^2} = \mathcal{O}(1/k^2)$$

Theorem

Accelerated gradient method used for minimizing strongly convex functions with step size $\alpha_k = \frac{1}{L}$ converges as

$$f(\mathbf{x}_k) - f^* \leq L\|\mathbf{x}_k - \mathbf{x}_0\|_2^2 \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k$$

Example


GD fixed
HB optimal
CG
Nesterov
10^{3}
10^{1}
10^{-1}
10^{-3}
10^{-5}
$\|f'(x_k)\|_2$
0
1000
2000
3000
4000
5000
Number of iteration, k

What do we know?

- Convergence of gradient descent can be improved

What do we know?

- Convergence of gradient descent can be improved
- Conjugate gradient method has to be used for minimizing quadratic strongly convex functions

What do we know?

- Convergence of gradient descent can be improved
- Conjugate gradient method has to be used for minimizing quadratic strongly convex functions
- Accelerated gradient method provides the fastest convergence in theory

What do we know?

- Convergence of gradient descent can be improved
- Conjugate gradient method has to be used for minimizing quadratic strongly convex functions
- Accelerated gradient method provides the fastest convergence in theory

Questions

What do we know?

- Convergence of gradient descent can be improved
- Conjugate gradient method has to be used for minimizing quadratic strongly convex functions
- Accelerated gradient method provides the fastest convergence in theory

Questions

- How to process inexact gradients?

What do we know?

- Convergence of gradient descent can be improved
- Conjugate gradient method has to be used for minimizing quadratic strongly convex functions
- Accelerated gradient method provides the fastest convergence in theory

Questions

- How to process inexact gradients?
- How to select step sizes since in all methods they depend on unknown constants?

What do we know?

- Convergence of gradient descent can be improved
- Conjugate gradient method has to be used for minimizing quadratic strongly convex functions
- Accelerated gradient method provides the fastest convergence in theory

Questions

- How to process inexact gradients?
- How to select step sizes since in all methods they depend on unknown constants?
- How convergence speed is changed?

Summary

- Conjugate gradient method

Summary

- Conjugate gradient method
- Heavy-ball method

Summary

- Conjugate gradient method
- Heavy-ball method
- Accelerated gradient method