Conjugate gradients method

Optimization methods in machine learning

Aleksandr Beznosikov

Innopolis University

2 October 2023

Back to Cauchy again.

- Again we solve the system of linear equations:

$$Ax = b.$$

Try to find $x \in \mathbb{R}^d$

- $A \in \mathbb{R}^{d \times d}$ positive definite and $b \in \mathbb{R}^d$.

Conjugate directions

Definition of conjugate directions

A set of non-zero vectors $\{p_i\}_{i=0}^{n-1}$ is called conjugate with respect to a positive definite matrix A if for any $i \neq j \in \{0, \dots n-1\}$ follows

$$p_i^T A p_j = 0.$$

Conjugate directions: linear independence

Linear independence of conjugate directions

The conjugate vectors $\{p_i\}_{i=0}^{n-1}$ are linearly independent.

Conjugate directions: linear independence

Proof

- By contradiction:

Conjugate directions: linear independence

Proof

- By contradiction: let there exist p_i such that there are:

$$p_i = \sum_{i \neq j} \lambda_j p_j \text{ for some } \lambda_j \in \mathbb{R}.$$

Conjugate directions: linear independence

Proof

- By contradiction: let there exist p_i such that there are:

$$p_i = \sum_{i \neq j} \lambda_j p_j \text{ for some } \lambda_j \in \mathbb{R}.$$

- Let us use a definition: let's take the scalar product with Ap_m, where $m \neq i$

Conjugate directions: linear independence

Proof

- By contradiction: let there exist p_i such that there are:

$$p_i = \sum_{i \neq j} \lambda_j p_j \text{ for some } \lambda_j \in \mathbb{R}.$$

- Let us use a definition: let's take the scalar product with Ap_m, where $m \neq i$

$$0 = p_m^T A p_i = \sum_{i \neq j} \lambda_j p_m^T A p_j = \lambda_m p_m^T A p_m.$$

Question: why are the first and last transitions performed?

Conjugate directions: linear independence

Proof

- By contradiction: let there exist p_i such that there are:

$$p_i = \sum_{i \neq j} \lambda_j p_j \text{ for some } \lambda_j \in \mathbb{R}.$$

- Let us use a definition: let's take the scalar product with Ap_m, where $m \neq i$

$$0 = p_m^T A p_i = \sum_{i \neq j} \lambda_j p_m^T A p_j = \lambda_m p_m^T A p_m.$$

  Question: why are the first and last transitions performed? Because of the definition of contiguity.

- **Question:** what did we get?

Conjugate directions: linear independence

Proof

- By contradiction: let there exist p_i such that there are:

$$p_i = \sum_{i \neq j} \lambda_j p_j \text{ for some } \lambda_j \in \mathbb{R}.$$

- Let us use a definition: let's take the scalar product with Ap_m, where $m \neq i$

$$0 = p_m^T A p_i = \sum_{i \neq j} \lambda_j p_m^T A p_j = \lambda_m p_m^T A p_m.$$

  Question: why are the first and last transitions performed? Because of the definition of contiguity.

- **Question:** what did we get? $\lambda_m = 0$.

Conjugate directions: linear independence

Proof

- By contradiction: let there exist p_i such that there are:

$$p_i = \sum_{i \neq j} \lambda_j p_j \text{ for some } \lambda_j \in \mathbb{R}.$$

- Let us use a definition: let's take the scalar product with Ap_m, where $m \neq i$

$$0 = p_m^T A p_i = \sum_{i \neq j} \lambda_j p_m^T A p_j = \lambda_m p_m^T A p_m.$$

  Question: why are the first and last transitions performed? Because of the definition of contiguity.

- **Question:** what did we get? $\lambda_m = 0$.
- **Question:** what does that mean?

Conjugate directions: linear independence

Proof

- By contradiction: let there exist p_i such that there are:

$$p_i = \sum_{i \neq j} \lambda_j p_j \text{ for some } \lambda_j \in \mathbb{R}.$$

- Let us use a definition: let's take the scalar product with Ap_m, where $m \neq i$

$$0 = p_m^T A p_i = \sum_{i \neq j} \lambda_j p_m^T A p_j = \lambda_m p_m^T A p_m.$$

  Question: why are the first and last transitions performed? Because of the definition of contiguity.

- **Question:** what did we get? $\lambda_m = 0$.
- **Question:** what does that mean? We can run through all of them $m \neq i$ and get $\lambda_m = 0$, and then $p_i = 0$. Contradiction.

Conjugate directions: how to use

- We have some kind of basis. **Question:** how can it be used?

Conjugate directions: how to use

- We have some kind of basis. **Question:** how can it be used? If we have d conjugate vectors, they form a basis. Decompose the solution:

$$x^* = \sum_{i=0}^{d-1} \lambda_i p_i.$$

Conjugate directions: how to use

- We have some kind of basis. **Question:** how can it be used? If we have d conjugate vectors, they form a basis. Decompose the solution:

$$x^* = \sum_{i=0}^{d-1} \lambda_i p_i.$$

- **Question:** how to find λ_i?

Conjugate directions: how to use

- We have some kind of basis. **Question:** how can it be used? If we have d conjugate vectors, they form a basis. Decompose the solution:
$$x^* = \sum_{i=0}^{d-1} \lambda_i p_i.$$
- **Question:** how to find λ_i? Take the scalar product with Ap_j:
$$p_j^T A x^* = \sum_{i=0}^{d-1} \lambda_i p_j^T A p_i = \lambda_j p_j^T A p_j.$$
Here again we used the definition of contiguity.

Conjugate directions: how to use

- We have some kind of basis. **Question:** how can it be used? If we have d conjugate vectors, they form a basis. Decompose the solution:
$$x^* = \sum_{i=0}^{d-1} \lambda_i p_i.$$
- **Question:** how to find λ_i? Take the scalar product with Ap_j:
$$p_j^T A x^* = \sum_{i=0}^{d-1} \lambda_i p_j^T A p_i = \lambda_j p_j^T A p_j.$$
Here again we used the definition of contiguity.
- Take into account that $Ax^* = b$, then $p_j^T b = \lambda_j p_j^T A p_j$.

Conjugate directions: how to use

- We have some kind of basis. **Question:** how can it be used? If we have d conjugate vectors, they form a basis. Decompose the solution:

$$x^* = \sum_{i=0}^{d-1} \lambda_i p_i.$$

- **Question:** how to find λ_i? Take the scalar product with Ap_j:

$$p_j^T A x^* = \sum_{i=0}^{d-1} \lambda_i p_j^T A p_i = \lambda_j p_j^T A p_j.$$

  Here again we used the definition of contiguity.
- Take into account that $Ax^* = b$, then $p_j^T b = \lambda_j p_j^T A p_j$.
- Hence,

$$\lambda_j = \frac{p_j^T b}{p_j^T A p_j}.$$

Conjugate directions: how to use

- **Question:** and can you see any problems?

Conjugate directions: how to use

- **Question:** and can you see any problems? Everything is good except that we ourselves invented conjugate directions, we ourselves said that they exist, but how to get them in reality is still unclear.

Conjugate directions: how to use

- **Question:** and can you see any problems? Everything is good except that we ourselves invented conjugate directions, we ourselves said that they exist, but how to get them in reality is still unclear.
- Let us start turning the reasoning into some iterative method:

$$\boxed{x^{k+1} = x^k + \alpha_k p_k.}$$

  That is, we are supposed to look for a new p_k at each iteration and find α_k for it.

Conjugate gradients method: α

We more or less figured out how to look for α when we try to find λ.
Question: $\lambda = \alpha$?

Conjugate gradients method: α

We more or less figured out how to look for α when we try to find λ.
Question: $\lambda = \alpha$? Not always. Iterative scheme with α:

$$x^{k+1} = x^0 + \sum_{i=0}^{k} \alpha_i p_i.$$

Conjugate gradients method: α

We more or less figured out how to look for α when we try to find λ.
Question: $\lambda = \alpha$? Not always. Iterative scheme with α:

$$x^{k+1} = x^0 + \sum_{i=0}^{k} \alpha_i p_i.$$

Iterative scheme with λ:

$$x^{k+1} = \sum_{i=0}^{k} \lambda_i p_i.$$

Conjugate gradients method: α

We more or less figured out how to look for α when we try to find λ.
Question: $\lambda = \alpha$? Not always. Iterative scheme with α:

$$x^{k+1} = x^0 + \sum_{i=0}^{k} \alpha_i p_i.$$

Iterative scheme with λ:

$$x^{k+1} = \sum_{i=0}^{k} \lambda_i p_i.$$

It turns out that $\alpha_i = \lambda_i$ if $x^0 = 0$.

Conjugate gradients method: α

We more or less figured out how to look for α when we try to find λ.
Question: $\lambda = \alpha$? Not always. Iterative scheme with α:

$$x^{k+1} = x^0 + \sum_{i=0}^{k} \alpha_i p_i.$$

Iterative scheme with λ:

$$x^{k+1} = \sum_{i=0}^{k} \lambda_i p_i.$$

It turns out that $\alpha_i = \lambda_i$ if $x^0 = 0$. We need a formula to find α, since starting from 0 is good, but we may have a closer candidate as a starting point.

Conjugate gradients method: α

- We can decompose x^0 into a basis and find $\tilde{\lambda}_i$ for it:

Conjugate gradients method: α

- We can decompose x^0 into a basis and find $\tilde{\lambda}_i$ for it:

$$x^0 = \sum_{i=0}^{d-1} \tilde{\lambda}_i p_i, \text{ where } \tilde{\lambda}_i = \frac{p_i^T A x^0}{p_i^T A p_i}.$$

Conjugate gradients method: α

- We can decompose x^0 into a basis and find $\tilde{\lambda}_i$ for it:

$$x^0 = \sum_{i=0}^{d-1} \tilde{\lambda}_i p_i, \text{ where } \tilde{\lambda}_i = \frac{p_i^T A x^0}{p_i^T A p_i}.$$

- Then the following statement is true:

$$x^0 + \sum_{i=0}^{d-1} \alpha_i p_i = \sum_{i=0}^{d-1} \left(\frac{p_i^T A x^0}{p_i^T A p_i} + \alpha_i \right) p_i = \sum_{i=0}^{d-1} \lambda_i p_i = \sum_{i=0}^{d-1} \frac{p_i^T b}{p_i^T A p_i} p_i.$$

Conjugate gradients method: α

- We can decompose x^0 into a basis and find $\tilde{\lambda}_i$ for it:

$$x^0 = \sum_{i=0}^{d-1} \tilde{\lambda}_i p_i, \text{ where } \tilde{\lambda}_i = \frac{p_i^T A x^0}{p_i^T A p_i}.$$

- Then the following statement is true:

$$x^0 + \sum_{i=0}^{d-1} \alpha_i p_i = \sum_{i=0}^{d-1} \left(\frac{p_i^T A x^0}{p_i^T A p_i} + \alpha_i \right) p_i = \sum_{i=0}^{d-1} \lambda_i p_i = \sum_{i=0}^{d-1} \frac{p_i^T b}{p_i^T A p_i} p_i.$$

- We get

$$\alpha_k = \frac{p_k^T (b - A x^0)}{p_k^T A p_k}.$$

Conjugate gradients method: α

- The result is already normal, a little more can be done:

$$p_k^T A(x^k - x^0) = 0.$$

Question: why? $(x^k - x^0) = \sum_{i=0}^{k-1} \alpha_i p_i$, a p_i и p_k conjugate with respect to A.

Conjugate gradients method: α

- The result is already normal, a little more can be done:

$$p_k^T A(x^k - x^0) = 0.$$

Question: why? $(x^k - x^0) = \sum_{i=0}^{k-1} \alpha_i p_i$, a p_i и p_k conjugate with respect to A.

- Then we can do like this:

$$\alpha_k = \frac{p_k^T(b - Ax^k)}{p_k^T A p_k} = -\frac{p_k^T r_k}{p_k^T A p_k}.$$

Here we add notation: $r_k = Ax^k - b$.

Conjugate gradients method: physical meaning α

- Consider the step of the method $x^{k+1} = x^k + \alpha_k p_k$, as well as the function

$$f(x) = \frac{1}{2}x^T Ax - bx.$$

Conjugate gradients method: physical meaning α

- Consider the step of the method $x^{k+1} = x^k + \alpha_k p_k$, as well as the function
$$f(x) = \frac{1}{2}x^T Ax - bx.$$
- **Question:** what is this function?

Conjugate gradients method: physical meaning α

- Consider the step of the method $x^{k+1} = x^k + \alpha_k p_k$, as well as the function
$$f(x) = \frac{1}{2}x^T Ax - bx.$$
- **Question:** what is this function? Its minimization is equivalent to finding a solution to a system of equations: $\nabla f(x^*) = Ax^* - b = 0$.

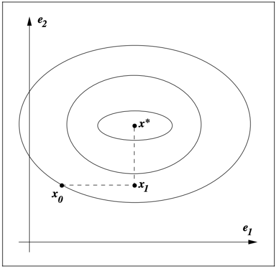
Conjugate gradients method: physical meaning α

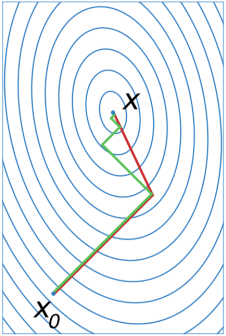
- Consider the step of the method $x^{k+1} = x^k + \alpha_k p_k$, as well as the function
$$f(x) = \frac{1}{2}x^T Ax - bx.$$
- **Question:** what is this function? Its minimization is equivalent to finding a solution to a system of equations: $\nabla f(x^*) = Ax^* - b = 0$.
- Consider:
$$g(\alpha) = f(x^k + \alpha p_k).$$
Where this function has a minimum on α^*?

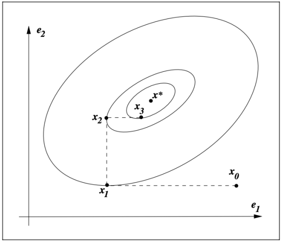
Conjugate gradients method: physical meaning α

- Consider the step of the method $x^{k+1} = x^k + \alpha_k p_k$, as well as the function
$$f(x) = \frac{1}{2}x^T A x - bx.$$
- **Question:** what is this function? Its minimization is equivalent to finding a solution to a system of equations: $\nabla f(x^*) = Ax^* - b = 0$.
- Consider:
$$g(\alpha) = f(x^k + \alpha p_k).$$
Where this function has a minimum on α^*? $\alpha^* = \frac{p_k^T(b - Ax^k)}{p_k^T A p_k} = \alpha_k$. That's physics — minimizing along p_k.

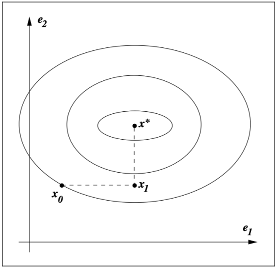
Conjugate gradients method: physical meaning α

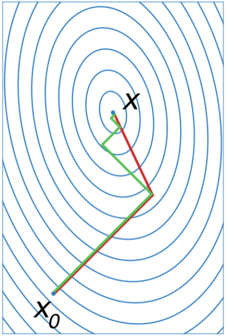


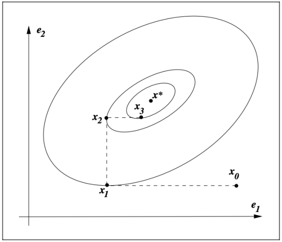




Conjugate gradients method: physical meaning α

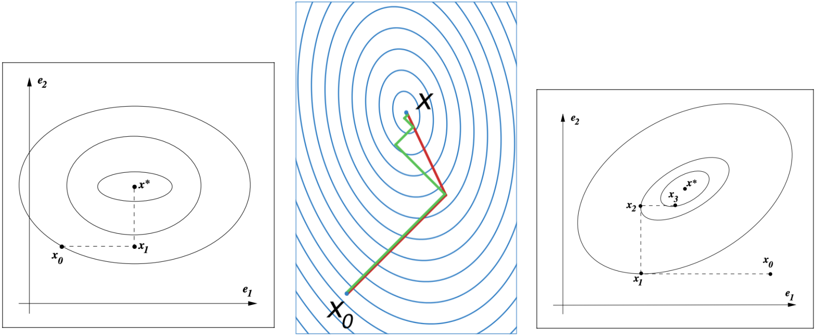






- The second picture shows that the conjugate directions are not orthogonal (in the usual sense of the word).

Conjugate gradients method: physical meaning α



- The second picture shows that the conjugate directions are not orthogonal (in the usual sense of the word).
- The third picture shows that the directions are not conjugate with respect to A, which causes problems.

Conjugate gradients method: physical meaning p

Physical meaning of p

If $\{p_i\}_{i=0}^k$ conjugate directions, then for any $k \geq 0$ and $i \leq k$ it holds:

$$r_{k+1}^T p_i = 0 \text{ same as } \langle \nabla f(x^{k+1}), p_i \rangle.$$

Conjugate gradients method: physical meaning p

Proof

- By induction. **Base:**

Conjugate gradients method: physical meaning p

Proof

- By induction. **Base:** $r_1 = Ax^1 - b = Ax^0 - b + \alpha_0 A p_0 = r_0 + \alpha_0 A p_0$, by virtue of choice $\alpha_0 = 0$:

$$p_0^T r_1 = p_0^T r_0 + \alpha_0 p_0^T A p_0 = 0.$$

Conjugate gradients method: physical meaning p

Proof

- By induction. **Base:** $r_1 = Ax^1 - b = Ax^0 - b + \alpha_0 A p_0 = r_0 + \alpha_0 A p_0$, by virtue of choice $\alpha_0 = 0$:
$$p_0^T r_1 = p_0^T r_0 + \alpha_0 p_0^T A p_0 = 0.$$
- **Assumption:** let the assumption be true for all $i \leq k$.

Conjugate gradients method: physical meaning p

Proof

- By induction. **Base:** $r_1 = Ax^1 - b = Ax^0 - b + \alpha_0 A p_0 = r_0 + \alpha_0 A p_0$, by virtue of choice $\alpha_0 = 0$:
$$p_0^T r_1 = p_0^T r_0 + \alpha_0 p_0^T A p_0 = 0.$$
- **Assumption:** let the assumption be true for all $i \leq k$.
- **Step:** let us prove for $k+1$.

Conjugate gradients method: physical meaning p

Proof

- By induction. **Base:** $r_1 = Ax^1 - b = Ax^0 - b + \alpha_0 A p_0 = r_0 + \alpha_0 A p_0$, by virtue of choice $\alpha_0 = 0$:
$$p_0^T r_1 = p_0^T r_0 + \alpha_0 p_0^T A p_0 = 0.$$
- **Assumption:** let the assumption be true for all $i \leq k$.
- **Step:** let us prove for $k+1$. Consider:
$$r_{k+1} = Ax^{k+1} - b = Ax^k - b + \alpha_k A p_k = r_k + \alpha_k A p_k.$$

Conjugate gradients method: physical meaning p

Proof

- By induction. **Base:** $r_1 = Ax^1 - b = Ax^0 - b + \alpha_0 Ap_0 = r_0 + \alpha_0 Ap_0$, by virtue of choice $\alpha_0 = 0$:
$$p_0^T r_1 = p_0^T r_0 + \alpha_0 p_0^T Ap_0 = 0.$$
- **Assumption:** let the assumption be true for all $i \leq k$.
- **Step:** let us prove for $k+1$. Consider:
$$r_{k+1} = Ax^{k+1} - b = Ax^k - b + \alpha_k Ap_k = r_k + \alpha_k Ap_k.$$

Whence, by virtue of choice α_k
$$p_k^T r_{k+1} = p_k^T r_k + \alpha_k p_k^T Ap_k = 0.$$

For $i < k$:
$$p_i^T r_{k+1} = p_i^T r_k + \alpha_k p_i^T Ap_k = 0.$$

Question: why?

Conjugate gradients method: physical meaning p

Proof

- By induction. **Base:** $r_1 = Ax^1 - b = Ax^0 - b + \alpha_0 A p_0 = r_0 + \alpha_0 A p_0$, by virtue of choice $\alpha_0 = 0$:
$$p_0^T r_1 = p_0^T r_0 + \alpha_0 p_0^T A p_0 = 0.$$
- **Assumption:** let the assumption be true for all $i \leq k$.
- **Step:** let us prove for $k+1$. Consider:
$$r_{k+1} = Ax^{k+1} - b = Ax^k - b + \alpha_k A p_k = r_k + \alpha_k A p_k.$$

Whence, by virtue of choice α_k
$$p_k^T r_{k+1} = p_k^T r_k + \alpha_k p_k^T A p_k = 0.$$

For $i < k$:
$$p_i^T r_{k+1} = p_i^T r_k + \alpha_k p_i^T A p_k = 0.$$

Question: why? By virtue of induction and conjugation.

Conjugate gradients method: p

- It's time to look for the p already.
- **Question:** what do we want to demand from the p search technique (remember, for example, the Gram-Schmidt procedure)?

Conjugate gradients method: p

- It's time to look for the p already.
- **Question:** what do we want to demand from the p search technique (remember, for example, the Gram-Schmidt procedure)? « Cheapness» of counting p_k:

$$p_k = -r_k + \beta_k p_{k-1},$$

where β_k is some coefficient. To find p_k you only need to know p_{k-1} and r_k, and you can already forget the old r_i and p_i (they are accounted for in x^k).

Conjugate gradients method: p

- It's time to look for the p already.
- **Question:** what do we want to demand from the p search technique (remember, for example, the Gram-Schmidt procedure)? « Cheapness» of counting p_k:

$$p_k = -r_k + \beta_k p_{k-1},$$

  where β_k is some coefficient. To find p_k you only need to know p_{k-1} and r_k, and you can already forget the old r_i and p_i (they are accounted for in x^k).
- **Question:** how to find β_k?

Conjugate gradients method: p

- It's time to look for the p already.
- **Question:** what do we want to demand from the p search technique (remember, for example, the Gram-Schmidt procedure)? « Cheapness» of counting p_k:

$$p_k = -r_k + \beta_k p_{k-1},$$

  where β_k is some coefficient. To find p_k you only need to know p_{k-1} and r_k, and you can already forget the old r_i and p_i (they are accounted for in x^k).
- **Question:** how to find β_k? The conjugacy of p_k and p_{k-1}:

$$0 = p_{k-1}^T A p_k = -p_{k-1}^T A r_k + \beta_k p_{k-1}^T A p_{k-1},$$

Conjugate gradients method: p

- It's time to look for the p already.
- **Question:** what do we want to demand from the p search technique (remember, for example, the Gram-Schmidt procedure)? « Cheapness» of counting p_k:

$$p_k = -r_k + \beta_k p_{k-1},$$

  where β_k is some coefficient. To find p_k you only need to know p_{k-1} and r_k, and you can already forget the old r_i and p_i (they are accounted for in x^k).
- **Question:** how to find β_k? The conjugacy of p_k and p_{k-1}:

$$0 = p_{k-1}^T A p_k = -p_{k-1}^T A r_k + \beta_k p_{k-1}^T A p_{k-1},$$

  from where

$$\beta_k = \frac{p_{k-1}^T A r_k}{p_{k-1}^T A p_{k-1}}.$$

Conjugate gradients method

Алгоритм 1 Conjugate gradients method

Input: starting point $x^0 \in \mathbb{R}^d$, $r_0 = Ax_0 - b$, $p_0 = -r_0$ number of iterations K

for $k = 0, 1, \ldots, K-1$ **do**

$\quad \alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}$

$\quad x^{k+1} = x^k + \alpha_k p_k$

$\quad r_{k+1} = Ax^{k+1} - b$

$\quad \beta_{k+1} = \frac{r_{k+1}^T A p_k}{p_k^T A p_k}$

$\quad p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$

end for

Output: x^K

Conjugate gradients method

Алгоритм 2 Conjugate gradients method

Input: starting point $x^0 \in \mathbb{R}^d$, $r_0 = Ax_0 - b$, $p_0 = -r_0$ number of iterations K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: $\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}$

3: $x^{k+1} = x^k + \alpha_k p_k$

4: $r_{k+1} = Ax^{k+1} - b$

5: $\beta_{k+1} = \frac{r_{k+1}^T A p_k}{p_k^T A p_k}$

6: $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$

7: **end for**

Output: x^K

Question: why the gradients?

Conjugate gradients method

Алгоритм 3 Conjugate gradients method

Input: starting point $x^0 \in \mathbb{R}^d$, $r_0 = Ax_0 - b$, $p_0 = -r_0$ number of iterations K

for $k = 0, 1, \ldots, K-1$ **do**

$\quad \alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}$

$\quad x^{k+1} = x^k + \alpha_k p_k$

$\quad r_{k+1} = Ax^{k+1} - b$

$\quad \beta_{k+1} = \frac{r_{k+1}^T A p_k}{p_k^T A p_k}$

$\quad p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$

end for

Output: x^K

Question: why the gradients? $r_k = Ax^k - b = \nabla f(x^k)$. That's worth remembering.

Conjugate gradients method: proof

- **Question:** maybe we've already proved the convergence estimate?

Conjugate gradients method: proof

- **Question:** maybe we've already proved the convergence estimate? Close to this, we know that if all $\{p_i\}$ are conjugate directions, then we have enough d steps to recover the coefficients for x^* in the basis from $\{p_i\}$.

Conjugate gradients method: proof

- **Question:** maybe we've already proved the convergence estimate? Close to this, we know that if all $\{p_i\}$ are conjugate directions, then we have enough d steps to recover the coefficients for x^* in the basis from $\{p_i\}$.
- **Question:** Do we know that all $\{p_i\}$ are conjugate?

Conjugate gradients method: proof

- **Question:** maybe we've already proved the convergence estimate? Close to this, we know that if all $\{p_i\}$ are conjugate directions, then we have enough d steps to recover the coefficients for x^* in the basis from $\{p_i\}$.
- **Question:** Do we know that all $\{p_i\}$ are conjugate? No, we only know that p_k and p_{k-1} are conjugate by virtue of the selection of β_k. We need to show a broader statement:

$$\text{For all } k \geq 1 \text{ for all } i < k \text{ it holds } p_k^T A p_i = 0.$$

Conjugate gradients method: proof

- By induction

Conjugate gradients method: proof

- By induction
- **Base:** p_0 and p_1 are conjugated by virtue of the selection β_1.

Conjugate gradients method: proof

- By induction
- **Base:** p_0 and p_1 are conjugated by virtue of the selection β_1.
- **Assumption:** let all $\{p_i\}_{i=0}^k$ conjugate for $k \geq 1$.

Conjugate gradients method: proof

- By induction
- **Base:** p_0 and p_1 are conjugated by virtue of the selection β_1.
- **Assumption:** let all $\{p_i\}_{i=0}^k$ conjugate for $k \geq 1$.
- **Step:** Let us prove for $k+1$.

Conjugate gradients method: proof

- By induction
- **Base:** p_0 and p_1 are conjugated by virtue of the selection β_1.
- **Assumption:** let all $\{p_i\}_{i=0}^k$ conjugate for $k \geq 1$.
- **Step:** Let us prove for $k+1$. p_{k+1} и p_k are conjugate by virtue of the selection of β_{k+1}.

Conjugate gradients method: proof

- By induction
- **Base:** p_0 and p_1 are conjugated by virtue of the selection β_1.
- **Assumption:** let all $\{p_i\}_{i=0}^k$ conjugate for $k \geq 1$.
- **Step:** Let us prove for $k+1$. p_{k+1} и p_k are conjugate by virtue of the selection of β_{k+1}. Consider $i < k$:

$$p_{k+1}^T A p_i = -r_{k+1}^T A p_i + \beta_{k+1} p_k^T A p_i = -r_{k+1}^T A p_i.$$

Conjugate gradients method: proof

- By induction
- **Base:** p_0 and p_1 are conjugated by virtue of the selection β_1.
- **Assumption:** let all $\{p_i\}_{i=0}^k$ conjugate for $k \geq 1$.
- **Step:** Let us prove for $k+1$. p_{k+1} и p_k are conjugate by virtue of the selection of β_{k+1}. Consider $i < k$:

$$p_{k+1}^T A p_i = -r_{k+1}^T A p_i + \beta_{k+1} p_k^T A p_i = -r_{k+1}^T A p_i.$$

Question: why is the second transition valid?

Conjugate gradients method: proof

- By induction
- **Base:** p_0 and p_1 are conjugated by virtue of the selection β_1.
- **Assumption:** let all $\{p_i\}_{i=0}^k$ conjugate for $k \geq 1$.
- **Step:** Let us prove for $k+1$. p_{k+1} и p_k are conjugate by virtue of the selection of β_{k+1}. Consider $i < k$:

$$p_{k+1}^T A p_i = -r_{k+1}^T A p_i + \beta_{k+1} p_k^T A p_i = -r_{k+1}^T A p_i.$$

Question: why is the second transition valid? Because of the induction assumption and the fact that $i < k$.

Conjugate gradients method: proof

- By induction
- **Base:** p_0 and p_1 are conjugated by virtue of the selection β_1.
- **Assumption:** let all $\{p_i\}_{i=0}^k$ conjugate for $k \geq 1$.
- **Step:** Let us prove for $k+1$. p_{k+1} и p_k are conjugate by virtue of the selection of β_{k+1}. Consider $i < k$:

$$p_{k+1}^T A p_i = -r_{k+1}^T A p_i + \beta_{k+1} p_k^T A p_i = -r_{k+1}^T A p_i.$$

  Question: why is the second transition valid? Because of the induction assumption and the fact that $i < k$.
- It remains to show that $r_{k+1}^T A p_i = 0$. Let us remember that.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds
  $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и
  $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.
- By induction. **Base:** follows from the initialization.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.
- By induction. **Base:** follows from the initialization.
- **Assumption:** we assume that it is true for all $i \leq k$.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds
  $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и
  $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.
- By induction. **Base:** follows from the initialization.
- **Assumption:** we assume that it is true for all $i \leq k$.
- **Step:** let us prove for $k+1$. Using assumption:
  $r_k \in \text{span}\{r_0, \ldots A^k r_0\}$ and $p_k \in \text{span}\{r_0, \ldots A^k r_0\}$.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.
- By induction. **Base:** follows from the initialization.
- **Assumption:** we assume that it is true for all $i \leq k$.
- **Step:** let us prove for $k + 1$. Using assumption: $r_k \in \text{span}\{r_0, \ldots A^k r_0\}$ and $p_k \in \text{span}\{r_0, \ldots A^k r_0\}$. Then $Ap_k \in \text{span}\{Ar_0, \ldots A^{k+1} r_0\}$.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.
- By induction. **Base:** follows from the initialization.
- **Assumption:** we assume that it is true for all $i \leq k$.
- **Step:** let us prove for $k+1$. Using assumption: $r_k \in \text{span}\{r_0, \ldots A^k r_0\}$ and $p_k \in \text{span}\{r_0, \ldots A^k r_0\}$. Then $Ap_k \in \text{span}\{Ar_0, \ldots A^{k+1} r_0\}$.Knowing that $r_{k+1} = r_k + \alpha_k A p_k$, we get $r_{k+1} \in \{r_0, \ldots A^{k+1} r_0\}$.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds
  $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и
  $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.
- By induction. **Base:** follows from the initialization.
- **Assumption:** we assume that it is true for all $i \leq k$.
- **Step:** let us prove for $k + 1$. Using assumption:
  $r_k \in \text{span}\{r_0, \ldots A^k r_0\}$ and $p_k \in \text{span}\{r_0, \ldots A^k r_0\}$. Then
  $Ap_k \in \text{span}\{Ar_0, \ldots A^{k+1} r_0\}$.Knowing that $r_{k+1} = r_k + \alpha_k A p_k$, we get $r_{k+1} \in \{r_0, \ldots A^{k+1} r_0\}$. From where
  $\text{span}\{r_0, \ldots r_{k+1}\} \subseteq \text{span}\{r_0, \ldots A^{k+1} r_0\}$, but we need equality.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds
  $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и
  $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.
- By induction. **Base:** follows from the initialization.
- **Assumption:** we assume that it is true for all $i \leq k$.
- **Step:** let us prove for $k+1$. Using assumption:
  $r_k \in \text{span}\{r_0, \ldots A^k r_0\}$ and $p_k \in \text{span}\{r_0, \ldots A^k r_0\}$. Then
  $Ap_k \in \text{span}\{Ar_0, \ldots A^{k+1} r_0\}$.Knowing that $r_{k+1} = r_k + \alpha_k A p_k$, we get $r_{k+1} \in \{r_0, \ldots A^{k+1} r_0\}$. From where
  $\text{span}\{r_0, \ldots r_{k+1}\} \subseteq \text{span}\{r_0, \ldots A^{k+1} r_0\}$, but we need equality.
  Note that from the second assumption:
  $A^{k+1} r_0 = A(A^k r_0) \in \text{span}\{Ap_0, \ldots Ap_k\}$.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds
  $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и
  $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.
- By induction. **Base:** follows from the initialization.
- **Assumption:** we assume that it is true for all $i \leq k$.
- **Step:** let us prove for $k+1$. Using assumption:
  $r_k \in \text{span}\{r_0, \ldots A^k r_0\}$ and $p_k \in \text{span}\{r_0, \ldots A^k r_0\}$. Then
  $Ap_k \in \text{span}\{Ar_0, \ldots A^{k+1} r_0\}$.Knowing that $r_{k+1} = r_k + \alpha_k A p_k$, we get $r_{k+1} \in \{r_0, \ldots A^{k+1} r_0\}$. From where
  $\text{span}\{r_0, \ldots r_{k+1}\} \subseteq \text{span}\{r_0, \ldots A^{k+1} r_0\}$, but we need equality.
  Note that from the second assumption:
  $A^{k+1} r_0 = A(A^k r_0) \in \text{span}\{Ap_0, \ldots Ap_k\}$. Since $(r_{i+1} - r_i)/\alpha_i = Ap_i$ we get that $A^{k+1} r_0 \in \text{span}\{r_0, \ldots r_{k+1}\}$.

Conjugate gradients method: proof

- We prove that for $k \geq 0$ the following holds $\text{span}\{r_0, \ldots r_k\} = \text{span}\{r_0, \ldots A^k r_0\}$ и $\text{span}\{p_0, \ldots p_k\} = \text{span}\{r_0, \ldots A^k r_0\}$.
- By induction. **Base:** follows from the initialization.
- **Assumption:** we assume that it is true for all $i \leq k$.
- **Step:** let us prove for $k+1$. Using assumption: $r_k \in \text{span}\{r_0, \ldots A^k r_0\}$ and $p_k \in \text{span}\{r_0, \ldots A^k r_0\}$. Then $Ap_k \in \text{span}\{Ar_0, \ldots A^{k+1} r_0\}$.Knowing that $r_{k+1} = r_k + \alpha_k A p_k$, we get $r_{k+1} \in \{r_0, \ldots A^{k+1} r_0\}$. From where $\text{span}\{r_0, \ldots r_{k+1}\} \subseteq \text{span}\{r_0, \ldots A^{k+1} r_0\}$, but we need equality. Note that from the second assumption: $A^{k+1} r_0 = A(A^k r_0) \in \text{span}\{Ap_0, \ldots Ap_k\}$. Since $(r_{i+1} - r_i)/\alpha_i = Ap_i$ we get that $A^{k+1} r_0 \in \text{span}\{r_0, \ldots r_{k+1}\}$. From where $\text{span}\{r_0, \ldots A^{k+1} r_0\} \subseteq \text{span}\{r_0, \ldots r_{k+1}\}$.
  The inclusion of both sides is proven.

Conjugate gradients method: proof

- There's still a transition left for the second part.

Conjugate gradients method: proof

- There's still a transition left for the second part.
- According to update of p_{k+1}:

$$\text{span}\{p_0, \ldots p_{k+1}\} = \text{span}\{p_0, \ldots p_k, r_{k+1}\}.$$

Conjugate gradients method: proof

- There's still a transition left for the second part.
- According to update of p_{k+1}:

$$\text{span}\{p_0, \ldots p_{k+1}\} = \text{span}\{p_0, \ldots p_k, r_{k+1}\}.$$

- By the second assumption of induction:

$$\text{span}\{p_0, \ldots p_{k+1}\} = \text{span}\{r_0, \ldots A^k r_0, r_{k+1}\}.$$

Conjugate gradients method: proof

- There's still a transition left for the second part.
- According to update of p_{k+1}:

$$\text{span}\{p_0, \dots p_{k+1}\} = \text{span}\{p_0, \dots p_k, r_{k+1}\}.$$

- By the second assumption of induction:

$$\text{span}\{p_0, \dots p_{k+1}\} = \text{span}\{r_0, \dots A^k r_0, r_{k+1}\}.$$

- By the first assumption:

$$\text{span}\{p_0, \dots p_{k+1}\} = \text{span}\{r_0, \dots r_k, r_{k+1}\}.$$

Conjugate gradients method: proof

- There's still a transition left for the second part.
- According to update of p_{k+1}:

$$\text{span}\{p_0, \dots p_{k+1}\} = \text{span}\{p_0, \dots p_k, r_{k+1}\}.$$

- By the second assumption of induction:

$$\text{span}\{p_0, \dots p_{k+1}\} = \text{span}\{r_0, \dots A^k r_0, r_{k+1}\}.$$

- By the first assumption:

$$\text{span}\{p_0, \dots p_{k+1}\} = \text{span}\{r_0, \dots r_k, r_{k+1}\}.$$

- According to what has just been proven:

$$\text{span}\{p_0, \dots p_{k+1}\} = \text{span}\{r_0, \dots, A^{k+1} r_0\}.$$

Conjugate gradients method: proof

- Let's go back to $-r_{k+1}^T A p_i = 0$ для $i < k$.

Conjugate gradients method: proof

- Let's go back to $-r_{k+1}^T A p_i = 0$ для $i < k$.
- Now we know that
$$p_i \in \text{span}\{r_0, \ldots, A^i r_0\}.$$

Conjugate gradients method: proof

- Let's go back to $-r_{k+1}^T A p_i = 0$ для $i < k$.
- Now we know that

$$p_i \in \text{span}\{r_0, \ldots, A^i r_0\}.$$

- From where $$Ap_i \in \text{span}\{Ar_0, \ldots, A^{i+1} r_0\}.$$

Conjugate gradients method: proof

- Let's go back to $-r_{k+1}^T A p_i = 0$ для $i < k$.
- Now we know that
$$p_i \in \text{span}\{r_0, \ldots, A^i r_0\}.$$
- From where $$Ap_i \in \text{span}\{Ar_0, \ldots, A^{i+1} r_0\}.$$
- From what's just been proven.
$$Ap_i \in \text{span}\{Ar_0, \ldots, A^{i+1} r_0\} \subseteq \text{span}\{p_0, \ldots, p_{i+1}\}.$$

Conjugate gradients method: proof

- Let's go back to $-r_{k+1}^T A p_i = 0$ для $i < k$.
- Now we know that
$$p_i \in \mathrm{span}\{r_0, \ldots, A^i r_0\}.$$
- From where $$Ap_i \in \mathrm{span}\{Ar_0, \ldots, A^{i+1} r_0\}.$$
- From what's just been proven.
$$Ap_i \in \mathrm{span}\{Ar_0, \ldots, A^{i+1} r_0\} \subseteq \mathrm{span}\{p_0, \ldots, p_{i+1}\}.$$
- But all p_j for j from 0 to i are orthogonal to r^{k+1} by virtue of the fact that $\{p_j\}$ are conjugate by virtue of the induction assumption. So we have what we need.

Conjugate gradients method: convergence

Theorem on convergence of conjugate gradients method

The conjugate gradient method for solving a system of linear equations with a square positive definite matrix of size d finds an exact solution in at most d iterations.

Conjugate gradients method: convergence

Theorem on convergence of conjugate gradients method

The conjugate gradient method for solving a system of linear equations with a square positive definite matrix of size d finds an exact solution in at most d iterations.

Equivalent to minimizing a strong convex quadratic problem.

Conjugate gradients method: convergence

- **Question:** what we got is bad or good?

Conjugate gradients method: convergence

- **Question:** what we got is bad or good? Not really. And the method faced this problem at the moment of its appearance 70 years ago. I.e. for *exact* solution of the system of equations it is competitive, but not very popular.

Conjugate gradients method: convergence

- **Question:** what we got is bad or good? Not really. And the method faced this problem at the moment of its appearance 70 years ago. I.e. for *exact* solution of the system of equations it is competitive, but not very popular.
- The key word in the previous paragraph is «*exact*». The method of conjugate gradients can be stopped earlier, it is iterative. And this is already more interesting.

Conjugate gradients method: convergence

- **Question:** what we got is bad or good? Not really. And the method faced this problem at the moment of its appearance 70 years ago. I.e. for *exact* solution of the system of equations it is competitive, but not very popular.
- The key word in the previous paragraph is «*exact*». The method of conjugate gradients can be stopped earlier, it is iterative. And this is already more interesting.
- There are convergence features that make the method even faster.

Conjugate gradients method: convergence

Theorem on convergence of conjugate gradients method

The conjugate gradient method for solving a system of linear equations with a square positive definite matrix of size d finds an exact solution in at most r iterations, where r is the number of unique eigenvalues of the matrix.

Conjugate gradients method: convergence

Theorem on convergence of conjugate gradients method

The conjugate gradient method for solving a system of linear equations with a square positive definite matrix of size d finds an exact solution in at most r iterations, where r is the number of unique eigenvalues of the matrix.

Theorem on convergence of conjugate gradients method

The method of conjugate gradients for solving a system of linear equations with a square positive definite matrix of size d has the following convergence estimate:

$$\|x^k - x^*\|_A^2 \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k \|x^0 - x^*\|_A^*.$$

Here $\|x\|_A^2 = x^T A x$ and $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$.

Conjugate gradients method: convergence

Theorem on convergence of conjugate gradients method

The conjugate gradient method for solving a system of linear equations with a square positive definite matrix of size d finds an exact solution in at most r iterations, where r is the number of unique eigenvalues of the matrix.

Theorem on convergence of conjugate gradients method

The method of conjugate gradients for solving a system of linear equations with a square positive definite matrix of size d has the following convergence estimate:

$$\|x^k - x^*\|_A^2 \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^k \|x^0 - x^*\|_A^*.$$

Here $\|x\|_A^2 = x^T A x$ and $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$.

Question: and for which method is a similar estimate valid?

Conjugate gradients method: convergence

Theorem on convergence of conjugate gradients method

The conjugate gradient method for solving a system of linear equations with a square positive definite matrix of size d finds an exact solution in at most r iterations, where r is the number of unique eigenvalues of the matrix.

Theorem on convergence of conjugate gradients method

The method of conjugate gradients for solving a system of linear equations with a square positive definite matrix of size d has the following convergence estimate:

$$\|x^k - x^*\|_A^2 \leq 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^k \|x^0 - x^*\|_A^*.$$

Here $\|x\|_A^2 = x^T Ax$ and $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$.

Question: and for which method is a similar estimate valid? Accelerated

Conjugate gradients method

Алгоритм 4 Conjugate gradients method (classical version)

Input: starting point $x^0 \in \mathbb{R}^d$, $r_0 = Ax_0 - b$, $p_0 = -r_0$ number of iterations K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: $\alpha_k = \frac{r_k^T r_k}{p_k^T A p_k}$

3: $x^{k+1} = x^k + \alpha_k p_k$

4: $r_{k+1} = r_k + \alpha_k A p_k$

5: $\beta_{k+1} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$

6: $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$

7: **end for**

Output: x^K

Conjugate gradients method

Алгоритм 5 Conjugate gradients method (classical version)

Input: starting point $x^0 \in \mathbb{R}^d$, $r_0 = Ax_0 - b$, $p_0 = -r_0$ number of iterations K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: $\alpha_k = \frac{r_k^T r_k}{p_k^T A p_k}$

3: $x^{k+1} = x^k + \alpha_k p_k$

4: $r_{k+1} = r_k + \alpha_k A p_k$

5: $\beta_{k+1} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$

6: $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$

7: **end for**

Output: x^K

Recall that the gradient is $r_k = Ax^k - b = \nabla f(x^k)$.

Conjugate gradients method for general problems

Алгоритм 6 Conjugate gradients method (Fletcher - Reeves)

Input: staring point $x^0 \in \mathbb{R}^d$, $p_0 = -\nabla f(x_0)$ number of iterations K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\alpha_k = ?$
3: $x^{k+1} = x^k + \alpha_k p_k$
4: $\beta_{k+1} = \frac{\langle \nabla f(x^{k+1}), \nabla f(x^{k+1}) \rangle}{\langle \nabla f(x^k), \nabla f(x^k) \rangle}$
5: $p_{k+1} = -\nabla f(x^{k+1}) + \beta_{k+1} p_k$
6: **end for**

Output: x^K

Conjugate gradients method for general problems

Алгоритм 7 Conjugate gradients method (Fletcher - Reeves)

Input: staring point $x^0 \in \mathbb{R}^d$, $p_0 = -\nabla f(x_0)$ number of iterations K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\alpha_k = ?$
3: $x^{k+1} = x^k + \alpha_k p_k$
4: $\beta_{k+1} = \frac{\langle \nabla f(x^{k+1}), \nabla f(x^{k+1}) \rangle}{\langle \nabla f(x^k), \nabla f(x^k) \rangle}$
5: $p_{k+1} = -\nabla f(x^{k+1}) + \beta_{k+1} p_k$
6: **end for**

Output: x^K

Question: how to find the step α_k?

Conjugate gradients method for general problems

Алгоритм 8 Conjugate gradients method (Fletcher - Reeves)

Input: staring point $x^0 \in \mathbb{R}^d$, $p_0 = -\nabla f(x_0)$ number of iterations K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: $\quad \alpha_k = ?$

3: $\quad x^{k+1} = x^k + \alpha_k p_k$

4: $\quad \beta_{k+1} = \frac{\langle \nabla f(x^{k+1}), \nabla f(x^{k+1}) \rangle}{\langle \nabla f(x^k), \nabla f(x^k) \rangle}$

5: $\quad p_{k+1} = -\nabla f(x^{k+1}) + \beta_{k+1} p_k$

6: **end for**

Output: x^K

Question: how to find the step α_k? We want to minimize along the direction p_k, we get a one-dimensional function depending on α. Let's remember about dichotomy and the golden ratio.

Conjugate gradients method for general problems

Алгоритм 9 Conjugate gradients method (Polak - Ribiere)

Input: starting point $x^0 \in \mathbb{R}^d$, $p_0 = -\nabla f(x_0)$ number of iterations K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: $\alpha_k =$ Linesearch

3: $x^{k+1} = x^k + \alpha_k p_k$

4: $\beta_{k+1} = \frac{\langle \nabla f(x^{k+1}), \nabla f(x^{k+1}) - \nabla f(x^k) \rangle}{\langle \nabla f(x^k), \nabla f(x^k) \rangle}$

5: $p_{k+1} = -\nabla f(x^{k+1}) + \beta_{k+1} p_k$

6: **end for**

Output: x^K

Conjugate gradients method for general problems

- Generalizations work well, but the guarantees in theory are far from optimistic.

Conjugate gradients method for general problems

- Generalizations work well, but the guarantees in theory are far from optimistic.
- It is better to do «restarts» sometimes. In this case, «restarts» involve sometimes taking $\beta_k = 0$, forgetting history. **Question:** which method iterates then?

Conjugate gradients method for general problems

- Generalizations work well, but the guarantees in theory are far from optimistic.
- It is better to do «restarts» sometimes. In this case, «restarts» involve sometimes taking $\beta_k = 0$, forgetting history. **Question:** which method iterates then? Gradient descent.

Conjugate gradients method for general problems

- Generalizations work well, but the guarantees in theory are far from optimistic.
- It is better to do «restarts» sometimes. In this case, «restarts» involve sometimes taking $\beta_k = 0$, forgetting history. **Question:** which method iterates then? Gradient descent.
- Suitable as a «starter» method, by which from an initial unknown point, we can get close, but not exactly to the solution.