Numerical Analysis - Part II

Anders C. Hansen

Lecture 19

Iterative methods for linear algebraic systems

Minimization of quadratic function

The methods we considered so far for solving $A\boldsymbol{x} = \boldsymbol{b}$, namely Jacobi, Gauss-Seidel, and those with relaxation, fit into the scheme

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + c_k \boldsymbol{d}^{(k)}\,,$$

where we were aimed at getting $\rho(H) < 1$ for the iteration matix H. Say, for Jacobi with relaxation, we set $c_k = \omega$ and $\boldsymbol{d}^{(k)} = D^{-1}(\boldsymbol{b} - A\boldsymbol{x}^{(k)})$.
For solving $A\boldsymbol{x} = \boldsymbol{b}$ with a (positive definite) matrix $A > 0$, there is a different approach to constructing good iterative methods. It is based on succesive minimization of the quadratic function

$$F(\boldsymbol{x}^{(k)}) := \|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|_A^2 = \|\boldsymbol{e}^{(k)}\|_A^2\,,$$

since the minimizer is clearly the exact solution. Here, $\|\boldsymbol{y}\|_A := (A\boldsymbol{y}, \boldsymbol{y})^{1/2} := \sqrt{\boldsymbol{y}^T A\boldsymbol{y}}$ is a Euclidean-type distance which is well-defined for $A > 0$.

Minimization of quadratic function

So, at each step k, we are decreasing the A-distance between $\boldsymbol{x}^{(k)}$ and the exact solution $\boldsymbol{x}^*$. Thus, for a symmetric positive definite $A > 0$, we choose an iterative method that provides the descent condition

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + c_k \boldsymbol{d}^{(k)} \quad \Rightarrow \quad F(\boldsymbol{x}^{(k+1)}) < F(\boldsymbol{x}^{(k)}) \,. \tag{1}$$

Minimization of quadratic function

An equivalent approach is to minimize the quadratic function

$$F_1(\boldsymbol{x}) = \tfrac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b}\,,$$

which attains its minimum when $\nabla F_1(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b} = 0$, and which does not involve the unknown $\boldsymbol{x}^*$. It is easy to check that $F_1(\boldsymbol{x}) = \frac{1}{2}F(\boldsymbol{x}) - \frac{1}{2}c$, where $c = \boldsymbol{x}^{*T}A\boldsymbol{x}^*$ is a constant independent of k, hence equivalence.

Both the Jacobi and the Gauss–Seidel methods satisfy (1), precisely

$$(A\boldsymbol{e}^{(k+1)}, \boldsymbol{e}^{(k+1)}) = (A\boldsymbol{e}^{(k)}, \boldsymbol{e}^{(k)}) - (C\boldsymbol{y}^{(k)}, \boldsymbol{y}^{(k)}) < (A\boldsymbol{e}^{(k)}, \boldsymbol{e}^{(k)}),$$

$$\begin{aligned} \text{where for Gauss-Seidel:} \quad & C = D > 0, && \boldsymbol{y}^{(k)} := (L_0 + D)^{-1} A\boldsymbol{e}^{(k)}; \\ \text{and for Jacobi:} \quad & C = 2D - A > 0, && \boldsymbol{y}^{(k)} := D^{-1} A\boldsymbol{e}^{(k)}. \end{aligned}$$

A-**orthogonal projection method:** Next, we strengthen the descent condition (1), namely given $\boldsymbol{x}^{(k)}$ and some $\boldsymbol{d}^{(k)}$ (called a *search direction*), we will seek $\boldsymbol{x}^{(k+1)}$ from the set of vectors on the line $\ell = \{\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}\}_{\alpha \in \mathbb{R}}$ such that it makes the value of $F(\boldsymbol{x}^{(k+1)})$ not just smaller than $F(\boldsymbol{x}^{(k)})$, but as small as possible (with respect to this set), namely

$$\boldsymbol{x}^{(k+1)} := \arg\min_{\alpha} F(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) . \tag{2}$$

A-orthogonal projection

Lemma 1

The minimizer in (2) is given by the formula

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}, \qquad \alpha_k = \frac{(\boldsymbol{r}^{(k)}, \boldsymbol{d}^{(k)})}{(A\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(k)})}. \tag{3}$$

This choice of α_k is referred to as exact line search.

Proof. From the definition of F, it follows that in (2) we should choose the point $\boldsymbol{x}^{(k+1)} \in \ell$ that minimizes the A-distance between $\boldsymbol{x}^*$ and the points $\boldsymbol{y} \in \ell$. Geometrically, it is clear that the minimum occurs when $\boldsymbol{x}^{(k+1)}$ is the A-orthogonal projection of $\boldsymbol{x}^*$ onto the line $\ell = \{\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}\}$, i.e., when

$$\begin{aligned} \boldsymbol{x}^* - \boldsymbol{x}^{(k+1)} \perp_A \boldsymbol{d}^{(k)} \quad &\Rightarrow \quad A(\boldsymbol{x}^* - \boldsymbol{x}^{(k+1)}) \perp \boldsymbol{d}^{(k)} \\ &\Rightarrow \quad \boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k A \boldsymbol{d}^{(k)} \perp \boldsymbol{d}^{(k)} . \end{aligned}$$

This gives expression for α_k in (3). $\square$

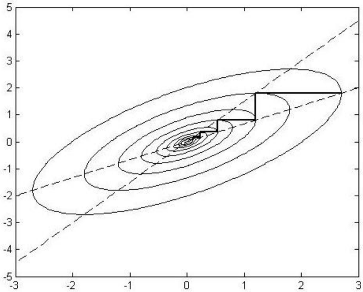
The steepest descent method: This method takes $\boldsymbol{d}^{(k)} = -\nabla F_1(\boldsymbol{x}^{(k)}) = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ for every k, the reason being that, locally, the negative gradient of a quadratic function shows the direction of the (locally) steepest descent at a given point. Thus, the iterations have the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k(\boldsymbol{b} - A\boldsymbol{x}^{(k)}), \qquad k \geq 0\,. \tag{4}$$

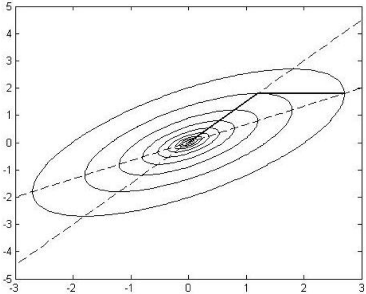
It can be proved that the sequence $(\boldsymbol{x}^{(k)})$ converges to the solution $\boldsymbol{x}^*$ of the system $A\boldsymbol{x} = \boldsymbol{b}$ as required, but usually the speed of convergence is rather slow.

The reason is that the iteration (4) decreases the value of $F(\boldsymbol{x}^{(k+1)})$ locally, relatively to $F(\boldsymbol{x}^{(k)})$, but the global decrease, with respect to $F(\boldsymbol{x}^{(0)})$, is often not that large. The use of *conjugate directions* provides a method with a global minimization property.

Steepest descent and conjugate gradient



(a) Worst case scenario of steepest descent



(b) Conjugate gradient method applied to the same problem as in (a)

Conjugate directions

Let's revisit equation (3) for a general direction $\boldsymbol{d}$ (i.e., not necessarily equal to the negative gradient). Assume $\boldsymbol{x} = \boldsymbol{x}^{(k)}$, and let $\boldsymbol{e}^{(k)} = \boldsymbol{x}^* - \boldsymbol{x}^{(k)}$ be the error and $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)} = A\boldsymbol{e}^{(k)}$ be the residual. Then we can write $\langle \boldsymbol{r}^{(k)}, \boldsymbol{d} \rangle = \langle \boldsymbol{e}^{(k)}, \boldsymbol{d} \rangle_A$, and so for a general search direction $\boldsymbol{d}$ with an exact line search, the iterate takes the form $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \frac{\langle \boldsymbol{e}^{(k)}, \boldsymbol{d} \rangle_A}{\langle \boldsymbol{d}, \boldsymbol{d} \rangle_A} \boldsymbol{d}$. By substracting $\boldsymbol{x}^*$, the iterates in terms of the error $\boldsymbol{e}^{(k+1)}$ are given by:

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{e}^{(k)} - \frac{\langle \boldsymbol{e}^{(k)}, \boldsymbol{d} \rangle_A}{\langle \boldsymbol{d}, \boldsymbol{d} \rangle_A} \boldsymbol{d}. \tag{5}$$

Geometrically, this means that $\boldsymbol{e}^{(k+1)}$ is the projection of $\boldsymbol{e}^{(k)}$ onto the hyperplane that is A-orthogonal to $\boldsymbol{d}$, i.e., we have

$$\langle \boldsymbol{e}^{(k+1)}, \boldsymbol{d} \rangle_A = 0. \tag{6}$$

Conjugate directions

Definition 2 (Conjugate directions)

The vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ are *conjugate* with respect to a symmetric positive definite matrix A if they are nonzero and A-orthogonal: $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_A := \langle \boldsymbol{u}, A\boldsymbol{v} \rangle = 0$.

Theorem 3

Let $\boldsymbol{d}^{(0)}, \boldsymbol{d}^{(1)}, \ldots, \boldsymbol{d}^{(n-1)}$ be n nonzero pairwise conjugate directions, and consider the sequence of iterates

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}, \qquad \alpha_k = \frac{\langle \boldsymbol{r}^{(k)}, \boldsymbol{d}^{(k)} \rangle}{\langle \boldsymbol{d}^{(k)}, A\boldsymbol{d}^{(k)} \rangle}.$$

Let $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ be the residual. Then for each $k = 1, \ldots, n$, $\boldsymbol{r}^{(k)}$ is orthogonal to $\mathrm{span}\{\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(k-1)}\}$. In particular $\boldsymbol{r}^{(n)} = 0$.

Proof. Since $\boldsymbol{r}^{(k)} = A\boldsymbol{e}^{(k)}$, it suffices to show that $\boldsymbol{e}^{(k)}$ is A-orthogonal to $\mathrm{span}\{\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(k-1)}\}$. The proof is by induction on k. For $k = 0$ there is nothing to prove. Assume the statement is true for $k \geq 0$, and consider the equation (5) (with $\boldsymbol{d} = \boldsymbol{d}^{(k)}$). From the induction hypothesis, and the fact that the $\boldsymbol{d}^{(i)}$ are pairwise conjugate directions, we see that $\boldsymbol{e}^{(k+1)}$ is A-orthogonal to $\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(k-1)}$. Furthermore, we have already seen in (6) that $\langle \boldsymbol{e}^{(k+1)}, \boldsymbol{d}^{(k)} \rangle_A = 0$. Thus this shows that $\boldsymbol{e}^{(k+1)}$ is A-orthogonal to $\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(k)}$ as desired. $\square$

So, if a sequence $(\boldsymbol{d}^{(k)})$ of conjugate directions is at hands, we have an iterative procedure with good approximation properties.
The (A-orthogonal) basis of conjugate directions is constructed by A-orthogonalization of the sequence $\{\boldsymbol{r}_0, A\boldsymbol{r}_0, A^2\boldsymbol{r}_0, ..., A^{n-1}\boldsymbol{r}_0\}$ with $\boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0$. This is done in the way similar to orthogonalization of the monomial sequence $\{1, x, x^2, ..., x^{n-1}\}$ using a recurrence relation.

Remark 4

It is possible to extend the methods for solving $A\boldsymbol{x} = \boldsymbol{b}$ with symmetric positive definite A to any other matrices by a simple trick. Suppose we want to solve $B\boldsymbol{x} = \boldsymbol{c}$, where $B \in \mathbb{R}^{n\times n}$ is nonsingular. We can convert the above system to the symmetric and positive definite setting by defining $A = B^T B$, $\boldsymbol{b} = B^T \boldsymbol{c}$ and then solving $A\boldsymbol{x} = \boldsymbol{b}$ with the conjugate gradient algorithm (or any other method for positive definite A).

The conjugate gradient method

Here it is.

(A) For any initial vector $\boldsymbol{x}^{(0)}$, set $\boldsymbol{d}^{(0)} = \boldsymbol{r}^{(0)} = \boldsymbol{b} - A\boldsymbol{x}^{(0)}$;

(B) For $k \geq 0$, calculate $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}$ and the residual

$$\begin{aligned} &\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k A\boldsymbol{d}^{(k)}, \quad \text{with} \\ &\qquad \alpha_k := \{\, \boldsymbol{r}^{(k+1)} \perp \boldsymbol{d}^{(k)} \} = \frac{(\boldsymbol{r}^{(k)}, \boldsymbol{d}^{(k)})}{(A\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(k)})}\,, \quad k \geq 0\,. \end{aligned} \tag{7}$$

(C) For the same k, the next conjugate direction is the vector

$$\begin{aligned} &\boldsymbol{d}^{(k+1)} = \boldsymbol{r}^{(k+1)} + \beta_k \boldsymbol{d}^{(k)}, \quad \text{with} \\ &\qquad \beta_k := \{\, \boldsymbol{d}^{(k+1)} \perp A\boldsymbol{d}^{(k)} \} = -\frac{(\boldsymbol{r}^{(k+1)}, A\boldsymbol{d}^{(k)})}{(\boldsymbol{d}^{(k)}, A\boldsymbol{d}^{(k)})}, \quad k \geq 0\,. \end{aligned} \tag{8}$$

Theorem 5 (Properties of CGM)

For every $m \geq 0$, the conjugate gradient method has the following properties.

(1) *The linear space spanned by the residuals $\{\boldsymbol{r}^{(i)}\}$ is the same as the linear space spanned by the conjugate directions $\{\boldsymbol{d}^{(i)}\}$ and it coincides with the space spanned by $\{A^i \boldsymbol{r}^{(0)}\}$:*

$$\mathrm{span}\{\boldsymbol{r}^{(i)}\}_{i=0}^{m} = \mathrm{span}\{\boldsymbol{d}^{(i)}\}_{i=0}^{m} = \mathrm{span}\{A^i \boldsymbol{r}^{(0)}\}_{i=0}^{m} .$$

(2) *The residuals satisfy the orthogonality conditions: $(\boldsymbol{r}^{(m)}, \boldsymbol{r}^{(i)}) = (\boldsymbol{r}^{(m)}, \boldsymbol{d}^{(i)}) = 0$ for $i < m$.*

(3) *The directions are conjugate (A-orthogonal): $(\boldsymbol{d}^{(m)}, \boldsymbol{d}^{(i)})_A = (\boldsymbol{d}^{(m)}, A\boldsymbol{d}^{(i)}) = 0$ for $i < m$.*

Proof. We use induction on $m \geq 0$, the assertions being trivial for $m = 0$, since $\boldsymbol{d}^{(0)} = \boldsymbol{r}^{(0)}$ and (2)-(3) are void. Therefore, assuming that the assertions are true for some $m = k$, we ask if they remain true when $m = k + 1$.

(1) Formula (8)

$$\boldsymbol{d}^{(k+1)} = \boldsymbol{r}^{(k+1)} + \beta_k \boldsymbol{d}^{(k)}$$

readily implies that equivalence of the spaces spanned by $(\boldsymbol{r}^{(i)})_0^k$ and $(\boldsymbol{d}^{(i)})_0^k$, is preserved when k is increased to $k+1$. Similarly, from $\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k A \boldsymbol{d}^{(k)}$ in (7), and from the inductive assumption $\boldsymbol{r}^{(k)}, \boldsymbol{d}^{(k)} \in \mathrm{span}\{A^i \boldsymbol{r}^{(0)}\}_{i=0}^k$, it follows that $\boldsymbol{r}^{(k+1)} \in \mathrm{span}\{A^i \boldsymbol{r}^{(0)}\}_{i=0}^{k+1}$. To see that $A^{k+1}\boldsymbol{r}^{(0)} \in \mathrm{span}\{\boldsymbol{r}^{(i)}\}_{i=0}^{k+1}$, since $\alpha_k \neq 0$, the claim follows by (7) if $\boldsymbol{d}^{(k)}$ has a non-zero component from $A^k \boldsymbol{r}^{(0)}$, and if not the claim follows from the induction hypothesis.

Proof. Cont. (2) Turning to assertion (2), we need $\boldsymbol{r}^{(k+1)} \perp \boldsymbol{r}^{(i)}$ for $i \leq k$, which by (1) is equivalent to

$$\boldsymbol{r}^{(k+1)} \perp \boldsymbol{d}^{(i)} \quad \text{for} \quad i \leq k\,.$$

We have $\boldsymbol{r}^{(k+1)} \perp \boldsymbol{d}^{(k)}$ by the definition of α_k in (7), so we need

$$\boldsymbol{r}^{(k+1)} \overset{(7)}{=} \boldsymbol{r}^{(k)} - \alpha_k A\boldsymbol{d}^{(k)} \perp \boldsymbol{d}^{(i)} \quad \text{for} \quad i < k\,,$$

and this follow from the inductive assumptions $\boldsymbol{r}^{(k)} \perp \boldsymbol{d}^{(i)}$ and $A\boldsymbol{d}^{(k)} \perp \boldsymbol{d}^{(i)}$.

Proof. Cont. (3) It remains to justify (3), namely that $\boldsymbol{d}^{(k+1)}$ defined in (8) satisfies

$$\boldsymbol{d}^{(k+1)} \perp A\boldsymbol{d}^{(i)} \quad \text{for} \quad i \leq k\,.$$

The value of β_k in (8) is defined to give $\boldsymbol{d}^{(k+1)} \perp A\boldsymbol{d}^{(k)}$, so we need

$$\boldsymbol{d}^{(k+1)} \overset{(8)}{=} \boldsymbol{r}^{(k+1)} + \beta_k \boldsymbol{d}^{(k)} \perp A\boldsymbol{d}^{(i)} \quad \text{for} \quad i < k\,.$$

By the inductive hypothesis $\boldsymbol{d}^{(k)} \perp A\boldsymbol{d}^{(i)}$, hence it remains to establish that $\boldsymbol{r}^{(k+1)} \perp A\boldsymbol{d}^{(i)}$ for $i < k$. Now, the formula (7) yields $A\boldsymbol{d}^{(i)} = (\boldsymbol{r}^{(i)} - \boldsymbol{r}^{(i+1)})/\alpha_i$, therefore we require the conditions $\boldsymbol{r}^{(k+1)} \perp (\boldsymbol{r}^{(i)} - \boldsymbol{r}^{(i+1)})$ for $i < k$, and they are a consequence of the assertion (2) for $m = k + 1$ obtained previously. $\square$

Termination property

Corollary 6 (A termination property)

If the conjugate gradient method is applied in exact arithmetic, then, for any $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, termination occurs after at most n iterations. More precisely, termination occurs after at most s iterations, where $s = \dim \operatorname{span}\{A^i \boldsymbol{r}_0\}_{i=0}^{n-1}$ (which can be smaller than n).

Proof. Assertion (2) of Theorem 5 states that residuals $(\boldsymbol{r}^{(k)})_{k\geq 0}$ form a sequence of mutually orthogonal vectors in $\mathbb{R}^n$, therefore at most n of them can be nonzero. Since they also belong to the space $\mathrm{span}\{A^i \boldsymbol{r}_0\}_{i=0}^{n-1}$, their number is bounded by the dimension of that space. $\square$

Definition 7 (The Krylov subspaces)

Let A be an $n \times n$ matrix, $\boldsymbol{v} \in \mathbb{R}^n$ nonzero, and $m \in \mathbb{N}$. The linear space $K_m(A, \boldsymbol{v}) := \mathrm{span}\{A^i \boldsymbol{v}\}_{i=0}^{m-1}$ is called the *m-th Krylov subspace* of $\mathbb{R}^n$.

Theorem 8 (Number of iterations in CGM)

Let $A > 0$, and let s be the number of its distinct eigenvalues. Then, for any $\boldsymbol{v}$,

$$\dim K_m(A, \boldsymbol{v}) \leq s \quad \forall\, m\,. \tag{9}$$

Hence, for any $A > 0$, the number of iterations of the CGM for solving $A\boldsymbol{x} = \boldsymbol{b}$ is bounded by the number of distinct eigenvalues of A.

Proof. Inequality (9) is true not just for positive definite $A > 0$, but for any A with n linearly independent eigenvectors $(\boldsymbol{u}_i)$. Indeed, in that case one can expand $\boldsymbol{v} = \sum_{i=1}^{n} a_i \boldsymbol{u}_i$, and then group together eigenvectors with the same eigenvalues: for each λ_ν we set $\boldsymbol{w}_\nu = \sum_{k=1}^{m_\nu} a_{i_k} \boldsymbol{u}_{i_k}$ if $A\boldsymbol{u}_{i_k} = \lambda_\nu \boldsymbol{u}_{i_k}$. Then

$$\boldsymbol{v} = \sum_{\nu=1}^{s} c_\nu \boldsymbol{w}_\nu\,, \qquad c_\nu \in \{0,1\}\,,$$

hence $A^i\boldsymbol{v} = \sum_{\nu=1}^{s} c_\nu \lambda_\nu^i \boldsymbol{w}_\nu$, thus for any m we get $K_m(A,\boldsymbol{v}) \subseteq \mathrm{span}\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_s\}$, and that proves (9). By Corollary 6, the number of iteration in CGM is bounded by $\dim K_m(A, \boldsymbol{r}^{(0)})$, hence the final conclusion. $\square$

Remark 9

Theorem 8 shows that, unlike other iterative schemes, the conjugate gradient method is both iterative and direct: each iteration produces a reasonable approximation to the exact solution, and the exact solution itself will be recovered after n iterations at most.

Simplifying the CGM-algorithm

We now simplify and reformulate the CGM-algorithm.
Firstly, we rewrite expressions for the parameters α_k and β_k in (7)-(8) as follows:

$$\alpha_k = \frac{(\boldsymbol{r}^{(k)}, \boldsymbol{d}^{(k)})}{(\boldsymbol{d}^{(k)}, A\boldsymbol{d}^{(k)})} \overset{(c)}{=} \frac{\|\boldsymbol{r}^{(k)}\|^2}{(\boldsymbol{d}^{(k)}, A\boldsymbol{d}^{(k)})} > 0\,,$$

$$\beta_k = -\frac{(\boldsymbol{r}^{(k+1)}, A\boldsymbol{d}^{(k)})}{(\boldsymbol{d}^{(k)}, A\boldsymbol{d}^{(k)})} \overset{(a)}{=} -\frac{(\boldsymbol{r}^{(k+1)}, \boldsymbol{r}^{(k+1)} - \boldsymbol{r}^{(k)})}{(\boldsymbol{d}^{(k)}, \boldsymbol{r}^{(k+1)} - \boldsymbol{r}^{(k)})} \overset{(b)}{=} \frac{\|\boldsymbol{r}^{(k+1)}\|^2}{(\boldsymbol{d}^{(k)}, \boldsymbol{r}^{(k)})} \overset{(c)}{=} \frac{\|\boldsymbol{r}^{(k+1)}\|^2}{\|\boldsymbol{r}^{(k)}\|^2} > 0\,.$$

Here, for β, we used in (a) the fact that $A\boldsymbol{d}^{(k)}$ is a multiple of $\boldsymbol{r}^{(k+1)} - \boldsymbol{r}^{(k)}$ by (7), and in (b) orthogonality of $\boldsymbol{r}^{(k+1)}$ to both $\boldsymbol{r}^{(k)}, \boldsymbol{d}^{(k)}$ proved in Theorem 5(2). Then, for both β and α, we used in (c) the property $(\boldsymbol{d}^{(k)}, \boldsymbol{r}^{(k)}) = \|\boldsymbol{r}^{(k)}\|^2$ which follows from (8) with index $k+1$, taking in account orthogonality $\boldsymbol{r}^{(k+1)} \perp \boldsymbol{d}^{(k)}$.
Secondly, we let $\boldsymbol{x}^{(0)}$ be the zero vector.

Standard form of the conjugate gradient method

Here it is.

(1) Set $k = 0$, $\boldsymbol{x}^{(0)} = 0$, $\boldsymbol{r}^{(0)} = \boldsymbol{b}$, and $\boldsymbol{d}^{(0)} = \boldsymbol{r}^{(0)}$;

(2) Calculate the matrix-vector product $\boldsymbol{v}^{(k)} = A\boldsymbol{d}^{(k)}$ and $\alpha_k = \|\boldsymbol{r}^{(k)}\|^2/(\boldsymbol{d}^{(k)}, \boldsymbol{v}^{(k)}) > 0$;

(3) Apply the formulae $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}$ and $\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k \boldsymbol{v}^{(k)}$;

(4) Stop if $\|\boldsymbol{r}^{(k+1)}\|$ is acceptably small;

(5) Set $\boldsymbol{d}^{(k+1)} = \boldsymbol{r}^{(k+1)} + \beta_k \boldsymbol{d}^{(k)}$, where $\beta_k = \|\boldsymbol{r}^{(k+1)}\|^2/\|\boldsymbol{r}^{(k)}\|^2 > 0$;

(6) Increase $k \to k + 1$ and go back to (2).

Standard form of the conjugate gradient method

The total work is dominated by the number of iterations, multiplied by the time it takes to compute $\boldsymbol{v}^{(k)} = A\boldsymbol{d}^{(k)}$. Thus the conjugate gradient algorithm is highly suitable when most of the elements of A are zero, i.e. when A is *sparse*.