Numerical Analysis - Part II

Anders C. Hansen

Lecture 20

Iterative methods for linear algebraic systems

Minimization of quadratic function

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

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

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 $\mathbf{d}^{(k)} = D^{-1}(\mathbf{b} - A\mathbf{x}^{(k)})$.
For solving $A\mathbf{x} = \mathbf{b}$ 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

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

since the minimizer is clearly the exact solution. Here, $\|\mathbf{y}\|_A := (A\mathbf{y}, \mathbf{y})^{1/2} := \sqrt{\mathbf{y}^T A\mathbf{y}}$ 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 $\mathbf{x}^{(k)}$ and the exact solution $\mathbf{x}^*$. Thus, for a symmetric positive definite $A > 0$, we choose an iterative method that provides the descent condition

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

Minimization of quadratic function

An equivalent approach is to minimize the quadratic function

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

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

Quadratic function – Jacobi and Gauss–Seidel

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

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

where for Gauss-Seidel: $C = D > 0$, $\mathbf{y}^{(k)} := (L_0 + D)^{-1} A\mathbf{e}^{(k)}$;

and for Jacobi: $C = 2D - A > 0$, $\mathbf{y}^{(k)} := D^{-1} A\mathbf{e}^{(k)}$.

A-orthogonal projection

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

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

Lemma 1

The minimizer in (2) is given by the formula

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

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

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

This gives expression for α_k in (3). □

The steepest descent method

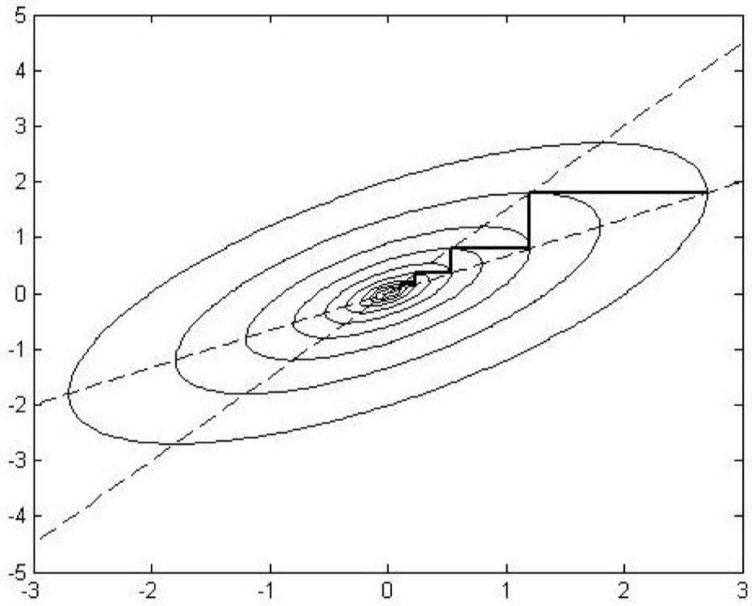
The steepest descent method: This method takes $\mathbf{d}^{(k)} = -\nabla F_1(\mathbf{x}^{(k)}) = \mathbf{b} - A\mathbf{x}^{(k)}$ for every k, the reason being that, locally, the negative gradient of a quadratic function shows the direction of the (locally) steepest descent at a given point. Thus, the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

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

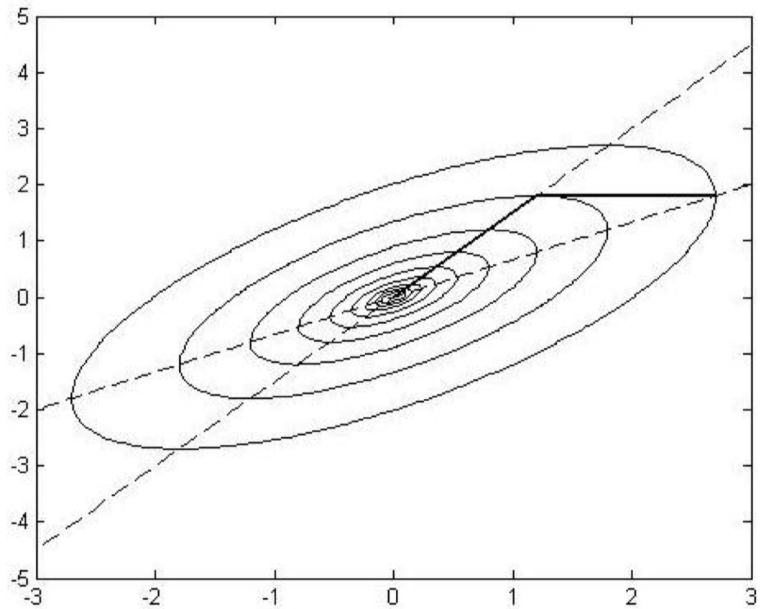
It can be proved that the sequence $(\mathbf{x}^{(k)})$ converges to the solution $\mathbf{x}^*$ of the system $A\mathbf{x} = \mathbf{b}$ as required, but usually the speed of convergence is rather slow.

The reason is that the iteration (4) decreases the value of $F(\mathbf{x}^{(k+1)})$ locally, relatively to $F(\mathbf{x}^{(k)})$, but the global decrease, with respect to $F(\mathbf{x}^{(0)})$, 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

Definition 2 (Conjugate directions)

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

Theorem 3 (Non-examinable)

Given $A \in \mathbb{R}^{n\times n}$, $A > 0$, *let* $\{\mathbf{d}^{(k)}\}_{k=0}^{n-1}$ *be a set of the conjugate directions, i.e.,* $(A\mathbf{d}^{(k)}, \mathbf{d}^{(i)}) = 0$ *for* $i < k$. *Then the value of* $F(\mathbf{x}^{(m+1)})$ *obtained through step-by-step minimization for each* $k = 0..m$ *as described in (3) coincides with the minimum of* $F(\mathbf{y})$ *taken over all* $\mathbf{y} = \mathbf{x}^{(0)} + \sum_{k=0}^{m} c_k \mathbf{d}^{(k)}$ *simultaneously, namely*

$$\arg \min_{c_0,\ldots,c_m} F(\mathbf{y}) = \mathbf{x}^{(m+1)} = \mathbf{x}^{(0)} + \sum_{k=0}^{m} \alpha_k \mathbf{d}^{(k)} .$$

Proof. Again, it is clear geometrically that the minimal A-distance between the exact solution $\mathbf{x}^*$ and the points $\mathbf{y}$ on the plane $\mathcal{P} := \{\mathbf{x}^{(0)} + \sum_{k=0}^{m} c_k \mathbf{d}^{(k)} : c_k \in \mathbb{R}\}$ is attained when $\mathbf{x}^{(m+1)} \in \mathcal{P}$ is the A-orthogonal projection of $\mathbf{x}^*$ onto $\mathcal{P}$, i.e.,

$$\arg\min_{\mathbf{y}\in\mathcal{P}} F(\mathbf{y}) = \mathbf{x}^{(m+1)} \quad \Leftrightarrow \quad \mathbf{x}^* - \mathbf{x}^{(m+1)} \perp_A \{\mathbf{d}^{(k)}\}_{k=0}^{m}.$$

It can be shown then, that (for conjugate $\{\mathbf{d}^{(k)}\}$) the latter conditions provide expressions for α_k as given in (3). □

Conjugate gradient - Warm up

So, if a sequence $(\mathbf{d}^{(k)})$ 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 $\{\mathbf{r}_0, A\mathbf{r}_0, A^2\mathbf{r}_0, ..., A^{n-1}\mathbf{r}_0\}$ with $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$. 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 $A\mathbf{x} = \mathbf{b}$ with symmetric positive definite A to any other matrices by a simple trick. Suppose we want to solve $B\mathbf{x} = \mathbf{c}$, 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 $A = B^T B$, $\mathbf{b} = B^T \mathbf{c}$ and then solving $A\mathbf{x} = \mathbf{b}$ 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 $\mathbf{x}^{(0)}$, set $\mathbf{d}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$;

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

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{d}^{(k)}, \quad \text{with}$$

$$\alpha_k := \{\, \mathbf{r}^{(k+1)} \perp \mathbf{d}^{(k)} \} = \frac{(\mathbf{r}^{(k)}, \mathbf{d}^{(k)})}{(A\mathbf{d}^{(k)}, \mathbf{d}^{(k)})}, \quad k \geq 0\,. \tag{5}$$

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

$$\mathbf{d}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{d}^{(k)}, \quad \text{with}$$

$$\beta_k := \{\, \mathbf{d}^{(k+1)} \perp A\mathbf{d}^{(k)} \} = -\frac{(\mathbf{r}^{(k+1)}, A\mathbf{d}^{(k)})}{(\mathbf{d}^{(k)}, A\mathbf{d}^{(k)})}, \quad k \geq 0\,. \tag{6}$$

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

$$\operatorname{span}\{\mathbf{r}^{(i)}\}_{i=0}^{m} = \operatorname{span}\{\mathbf{d}^{(i)}\}_{i=0}^{m} = \operatorname{span}\{A^i\mathbf{r}^{(0)}\}_{i=0}^{m}\,.$$

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

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

Proof. We use induction on $m \geq 0$, the assertions being trivial for $m = 0$, since $\mathbf{d}^{(0)} = \mathbf{r}^{(0)}$ 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 (6)

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

readily implies that equivalence of the spaces spanned by $(\mathbf{r}^{(i)})_0^k$ and $(\mathbf{d}^{(i)})_0^k$, is preserved when k is increased to $k + 1$. Similarly, from $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{d}^{(k)}$ in (5), and from the inductive assumption $\mathbf{r}^{(k)}, \mathbf{d}^{(k)} \in \mathrm{span}\{A^i\mathbf{r}^{(0)}\}_{i=0}^{k}$, it follows that $\mathbf{r}^{(k+1)} \in \mathrm{span}\{A^i\mathbf{r}^{(0)}\}_{i=0}^{k+1}$. To see that $A^{k+1}\mathbf{r}^{(0)} \in \mathrm{span}\{\mathbf{r}^{(i)}\}_{i=0}^{k+1}$, since $\alpha_k \neq 0$, the claim follows by (5) if $\mathbf{d}^{(k)}$ has a non-zero component from $A^k\mathbf{r}^{(0)}$, and if not the claim follows from the induction hypothesis.

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

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

We have $\mathbf{r}^{(k+1)} \perp \mathbf{d}^{(k)}$ by the definition of α_k in (5), so we need

$$\mathbf{r}^{(k+1)} \overset{(5)}{=} \mathbf{r}^{(k)} - \alpha_k A\mathbf{d}^{(k)} \perp \mathbf{d}^{(i)} \quad \text{for} \quad i < k\,,$$

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

Proof. Cont. (3) It remains to justify (3), namely that $\mathbf{d}^{(k+1)}$ defined in (6) satisfies

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

The value of β_k in (6) is defined to give $\mathbf{d}^{(k+1)} \perp A\mathbf{d}^{(k)}$, so we need

$$\mathbf{d}^{(k+1)} \overset{(6)}{=} \mathbf{r}^{(k+1)} + \beta_k \mathbf{d}^{(k)} \perp A\mathbf{d}^{(i)} \quad \text{for} \quad i < k\,.$$

By the inductive hypothesis $\mathbf{d}^{(k)} \perp A\mathbf{d}^{(i)}$, hence it remains to establish that $\mathbf{r}^{(k+1)} \perp A\mathbf{d}^{(i)}$ for $i < k$. Now, the formula (5) yields $A\mathbf{d}^{(i)} = (\mathbf{r}^{(i)} - \mathbf{r}^{(i+1)})/\alpha_i$, therefore we require the conditions $\mathbf{r}^{(k+1)} \perp (\mathbf{r}^{(i)} - \mathbf{r}^{(i+1)})$ for $i < k$, and they are a consequence of the assertion (2) for $m = k + 1$ obtained previously. □

Termination property

Corollary 6 (A termination property)

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

Proof. Assertion (2) of Theorem 5 states that residuals $(\mathbf{r}^{(k)})_{k\geq 0}$ 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 $\operatorname{span}\{A^i\mathbf{r}_0\}_{i=0}^{n-1}$, their number is bounded by the dimension of that space. □

Definition 7 (The Krylov subspaces)

Let A be an $n \times n$ matrix, $\mathbf{v} \in \mathbb{R}^n$ nonzero, and $m \in \mathbb{N}$. The linear space $K_m(A, \mathbf{v}) := \operatorname{span}\{A^i \mathbf{v}\}_{i=0}^{m-1}$ 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 $\mathbf{v}$,

$$\dim K_m(A, \mathbf{v}) \leq s \quad \forall\, m\,. \tag{7}$$

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

Proof. Inequality (7) is true not just for positive definite $A > 0$, but for any A with n linearly independent eigenvectors $(\mathbf{u}_i)$. Indeed, in that case one can expand $\mathbf{v} = \sum_{i=1}^{n} a_i \mathbf{u}_i$, and then group together eigenvectors with the same eigenvalues: for each λ_ν we set $\mathbf{w}_\nu = \sum_{k=1}^{m_\nu} a_{i_k} \mathbf{u}_{i_k}$ if $A\mathbf{u}_{i_k} = \lambda_\nu \mathbf{u}_{i_k}$. Then

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

hence $A^i \mathbf{v} = \sum_{\nu=1}^{s} c_\nu \lambda_\nu^i \mathbf{w}_\nu$, thus for any m we get $K_m(A, \mathbf{v}) \subseteq \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s\}$, and that proves (7). By Corollary 6, the number of iteration in CGM is bounded by $\dim K_m(A, \mathbf{r}^{(0)})$, hence the final conclusion. □

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 (5)-(6) as follows:

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

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

Here, for β, we used in (a) the fact that $A\mathbf{d}^{(k)}$ is a multiple of $\mathbf{r}^{(k+1)} - \mathbf{r}^{(k)}$ by (5), and in (b) orthogonality of $\mathbf{r}^{(k+1)}$ to both $\mathbf{r}^{(k)}, \mathbf{d}^{(k)}$ proved in Theorem 5(2). Then, for both β and α, we used in (c) the property $(\mathbf{d}^{(k)}, \mathbf{r}^{(k)}) = \|\mathbf{r}^{(k)}\|^2$ which follows from (6) with index $k+1$, taking in account orthogonality $\mathbf{r}^{(k+1)} \perp \mathbf{d}^{(k)}$.
Secondly, we let $\mathbf{x}^{(0)}$ be the zero vector.

Standard form of the conjugate gradient method

Here it is.

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

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

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

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

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

(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 $\mathbf{v}^{(k)} = A\mathbf{d}^{(k)}$. Thus the conjugate gradient algorithm is highly suitable when most of the elements of A are zero, i.e. when A is *sparse*.