

Numerical Methods in Scientific Computing

NGSSC

Uppsala University

January 2012

Iterative Solution methods

Basic Iterative Solution methods

Basic Iterative Solution methods

Introduction:

The ideas to use iterative methods for solving linear systems of equations go back to Gauss (1823), Liouville (1837) and Jacobi (1845).

After deriving an iterative procedure, in 1823, Gauss has written in a letter the following:

"... You will hardly eliminate directly anymore, at least not when you have more than two unknowns. The indirect method can be pursued while half asleep or while thinking about other things."



Introduction:

Before considering iterative solution methods for **linear** systems of equations, we recall how do we solve **nonlinear** problems

Let $f(x) = 0$ have to be solved and $f(x)$ is a nonlinear function in x .

The usual way to approach the problem is:

$$F(x) \equiv x - f(x).$$

If x^* is the solution of $f(x) = 0$, then x^* is a stationary point for

$$x = F(x). \tag{1}$$

Then we proceed with finding the stationary point for (1) and this is done **iteratively**, namely,


$$x^{(k+1)} = F(x^{(k)}), k = 0, 1, \dots, x^{(0)} \text{ given.} \tag{2}$$

Convergence of the fixed point iteration:

For any initial guess $x^{(0)}$, there exists a unique fixed point x^* for $F(x)$,
 $x^* = \lim_{k \rightarrow \infty} x^{(k)}$ **if and only if** F is a contracting mapping, i.e.

$$\|F(x) - F(y)\| \leq q\|x - y\|$$

for some $q \in (0, 1)$.



Fixed point for linear problems:

Let now $f(\mathbf{x}) \equiv A\mathbf{x} - \mathbf{b}$ be linear. We use the same framework:

$$\begin{aligned} F(\mathbf{x}) &= \mathbf{x} - (A\mathbf{x} - \mathbf{b}) \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - (A\mathbf{x}^{(k)} - \mathbf{b}) = \mathbf{x}^{(k)} + \mathbf{r}^{(k)} \end{aligned}$$

where $\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$ is called the **residual** at iteration k .

In this way we obtain the simplest possible iterative scheme to solve

$$A\mathbf{x} = \mathbf{b},$$

namely,

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - (A\mathbf{x}^{(k)} - \mathbf{b}), \quad k = 0, 1, \dots \\ \mathbf{x}^{(0)} &\text{ given.} \end{aligned}$$

Simple iteration

For many reasons the latter form of the simple iteration is replaced by

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \tau \mathbf{r}^{(k)}, \quad (3)$$

where τ is some properly chosen method parameter.

Relation (3) defines the so-called *stationary basic iterative method of first kind*.

Stationary iterative methods ...

If we permit τ to change from one iteration to the next, we get

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \tau_k \mathbf{r}^{(k)}, \quad (4)$$

which latter defines the so-called
non-stationary basic iterative method of first kind.



So far τ and τ_k are some scalars. Nothing prevents us to replace the method parameter by some matrix, however, if this would improve the convergence of the iterative method.

(cont)

Nothing prevents us to replace the method parameter by some matrix, however, if this would improve the convergence of the iterative method. Thus, we can consider

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + C^{-1}(\mathbf{b} - A\mathbf{x}^{(k)}) \\ \text{or} \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + C^{-1}\mathbf{r}^{(k)},\end{aligned}\tag{5}$$

It is easy to see that we obtain (5) by replacing $A\mathbf{x} = \mathbf{b}$ with

$$C^{-1}A\mathbf{x} = C^{-1}\mathbf{b}$$

and use the simple iteration framework. In this case the iterative scheme takes the form

$$\begin{aligned}C\mathbf{d}^{(k)} &= \mathbf{r}^{(k)}, \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \mathbf{d}^{(k)}\end{aligned}\tag{6}$$

The scheme (6) has in general a higher computational complexity than (4), since a solution of a system with the matrix C is required at each iteration.

Concerns:

C1 Does the iteration process converge to the solution, i.e. does $\mathbf{x}^{(k)} \rightarrow \mathbf{x}^*$?

C2 If 'yes', how fast does it converge?

The number of iterations it needed for the iterative method to converge with respect to some convergence criterion, is a function of the properties of A . For instance, $it = it(n)$, where n is the size of A . If it turns out that $it = O(n^2)$, we haven't gained anything compared to the direct solution methods.

The best one can hope for is to get $it \leq Const$, where $Const$ is independent of n . Since the the computational complexity of one iteration is in many cases proportional to n (for sparse matrices, for instance) then the complexity of the whole solution process will be

$$O(n).$$

C3 Is the method robust with respect to the method parameters (τ, τ_k) ?

Concerns (cont.):

- C4 Is the method robust with respect to various problem parameters?

$$A = A(\rho, \nu, E, \dots)$$

- C5 When we are using the scheme $C^{-1}A\mathbf{x} = C^{-1}\mathbf{b}$, it must be easy to solve systems with C .

- C6 Is the method parallelizable?

Parallelization aspects become more and more important since n is XXL.

Concerns (cont.):

Suppose the method converges to the exact solution \mathbf{x}^* .

Then more questions arise:

C7 When do we stop the iterations?

→ We want $\|\mathbf{x}^* - \mathbf{x}^{(k)}\| \leq \varepsilon$ but \mathbf{x}^* is not known.

→ What about checking on $\mathbf{r}^{(k)}$?

→ Is it enough to have $\|\mathbf{r}^{(k)}\| \leq \tilde{\varepsilon}$?

Will the latter guarantee that $\|\mathbf{x}^* - \mathbf{x}^{(k)}\| \leq \varepsilon$?

Denote $\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$ (the error at iteration k). Then

$$\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)} = A(\mathbf{x}^* - \mathbf{x}^{(k)}) = A\mathbf{e}^{(k)}.$$

In other words $\mathbf{e}^{(k)} = A^{-1}\mathbf{r}^{(k)}$.

Scenario: Suppose $\|A^{-1}\| = 10^8$ and $\tilde{\varepsilon} = 10^{-4}$. Then

$$\|\mathbf{e}^{(k)}\| \leq \|A^{-1}\| \|\mathbf{r}^{(k)}\| \leq 10^4, \quad \text{which is not very exiting.}$$

Example: Discrete Laplace Δ_h^5 :

$$\|A^{-1}\| \approx \lambda_{min} = \frac{1}{2}(\pi h)^2 \approx 10^4 \text{ for } h = 10^{-2}.$$

Concerns (cont.):

- C8 How do we measure (estimate) the convergence rate?
- C9 How do we find good method parameters (τ, τ_k, C) , which will speed up the convergence?



We start our considerations with [C9].

Choosing C :

Intuitively, C has to do something with A .

Note that if $C = A$, then $C^{-1} = A^{-1}$ and we will get convergence in one step!

However, the computational effort to construct A^{-1} is higher than to use a direct solution method.

We try the following choice. Consider the following so-called **splitting** of A ,

$$A = C - R,$$

where C is nonsingular and R can be seen as an error matrix.

Choosing C (cont.)

Then $C^{-1}A = C^{-1}(C - R) = I - C^{-1}R = I - B$.

- The matrix $B = C^{-1}R$ is referred to as the **iteration matrix**.
- $\|B^m\|$ is the **convergence factor for m steps**
- $(\|B^m\|)^{1/m}$ is called the **average convergence factor**.

Equivalent formulation using the splitting:

Using the splitting $A = C - R$ we obtain the following equivalent form of the iterative procedure:

$$\begin{aligned} A = C - R &\longrightarrow R = C - A \\ \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + C^{-1}(\mathbf{b} - A\mathbf{x}^{(k)}) \\ &= \mathbf{x}^{(k)} + C^{-1}\mathbf{b} - C^{-1}(C - R)\mathbf{x}^{(k)} \\ &= C^{-1}\mathbf{b} + C^{-1}R\mathbf{x}^{(k)} \end{aligned}$$

$$C\mathbf{x}^{(k+1)} = R\mathbf{x}^{(k)} + \mathbf{b} \tag{7}$$

The matrix C is called a **preconditioner** to A . Its general purpose is to improve the properties of A in order to achieve a better (faster) convergence of the method.

A general convergence result:

Theorem 1 The sequence $\{\mathbf{x}^{(k)}\}$ from $C\mathbf{x}^{(k+1)} = R\mathbf{x}^{(k)} + \mathbf{b}$ converges to the solution \mathbf{x}^* of $A\mathbf{x} = \mathbf{b}$ for any initial guess $\mathbf{x}^{(0)}$ if and only if there holds

$$\rho(B) \equiv \rho(C^{-1}R) < 1$$

where $\rho(\dots)$ denotes the spectral radius.

Proof Let $\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$, $A = C - R$. Then

$$\left. \begin{aligned} C\mathbf{x}^* &= R\mathbf{x}^* + \mathbf{b} \\ C\mathbf{x}^{(k)} &= R\mathbf{x}^{(k-1)} + \mathbf{b} \end{aligned} \right] -$$
$$C\mathbf{e}^{(k)} = R\mathbf{e}^{(k-1)}$$
$$\mathbf{e}^{(k)} = B\mathbf{e}^{(k-1)} = B^2\mathbf{e}^{(k-2)} = \dots = B^k\mathbf{e}^{(0)}.$$

If $\rho(C^{-1}R) < 1$ then $\lim_{k \rightarrow \infty} B^k = 0$ and $\mathbf{e}^{(k)} \rightarrow 0$. ■

If $\rho(C^{-1}R) \geq 1$:

Let $\lambda_i = \text{eig}(B)$ and $\rho(B) = |\lambda_j|$ i.e., λ_j is the eigenvalue of B , such that $\rho(B) = |\lambda_j|$. Let \mathbf{v}^j be the corresponding eigenvector.

Then $(\mathbf{v}^j)^m = B^m \mathbf{v}^j = \lambda_j^m \mathbf{v}^j \not\rightarrow 0$.

$$\mathbf{e}^{(0)} = \sum_{k=1}^n \beta_k \mathbf{v}^k = \beta_j \mathbf{v}^j + \dots$$

$$B^m \mathbf{e}^{(0)} = \tilde{\beta}_j B^m \mathbf{v}^j + \dots$$

and at least one component of $\mathbf{e}^{(m)}$ does not converge to zero.

Remarks on the proof:

The basic argument in the latter proof is that if $\rho(B) < 1$ then $B^k \rightarrow 0$. This can be shown in the following way.

Lemma 1 *Let T be a nonsingular matrix and let $\|\mathbf{x}\|_T = \|T\mathbf{x}\|_\infty$. Let*

$\|A\|_T = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_T}{\|\mathbf{x}\|_T}$ be the induced matrix norm. Then.

(a) $\|A\|_T = \|TAT^{-1}\|_\infty$

(b) *For any $\varepsilon > 0$ and matrix A , there exists a nonsingular matrix T such that*
 $\|A\|_T \leq \rho(A) + \varepsilon$.

In other words, there exist matrix norms, which are arbitrary close to the spectral radius of a given matrix.

Remarks on the proof:

Proof (a) $\|\mathbf{x}\|_T$ is a vector norm and T^{-1} exists.

$$\begin{aligned}\|A\|_T &= \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_T}{\|\mathbf{x}\|_T} = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|TA\mathbf{x}\|_\infty}{\|T\mathbf{x}\|_\infty} \\ &= \sup_{\mathbf{y} \neq \mathbf{0}} \frac{\|TAT^{-1}\mathbf{y}\|_\infty}{\|\mathbf{y}\|_\infty} = \|TAT^{-1}\|_\infty.\end{aligned}$$



Proof (cont.)

(b) We use Schur's lemma: There exists a unitary matrix U , such that

$$UAU^{-1} = W = \begin{bmatrix} w_{11} & * & * & \cdots & * \\ 0 & w_{22} & * & \cdots & * \\ & & \ddots & \cdots & \vdots \\ & & & & w_{nn} \end{bmatrix},$$

where $w_{ii} = \lambda_i \in S(A)$; $S(A)$ denotes the spectrum of A . Let $\delta > 0$ and define $D = D(\delta) = \text{diag}\{\delta^{-1}, \delta^{-2}, \dots, \delta^{-n}\}$. Then DWD^{-1} is also upper triangular

$$\text{and } (DWD^{-1})_{ij} = \begin{cases} 0, & j < i \\ w_{ii}, & j = i \\ w_{ij}\delta^{j-1}, & j > i. \end{cases}$$

$$\Rightarrow \|DWD^{-1}\|_{\infty} \leq \max_i \left\{ |w_{ii}| + n \max_{j>i} |w_{ij}| \delta^{j-1} \right\}.$$

Proof (cont.)

We see that for any given $\varepsilon > 0$ we can choose $\delta > 0$ small enough so that

$$\|DWD^{-1}\|_{\infty} \leq \rho(A) + \varepsilon$$

$$n \max_{j>i} |w_{ij} \delta^{j-1}| < \varepsilon. \text{ Hence, } \|A\|_T = \|TAT^{-1}\|_{\infty} = \|DU AU^{-1} D^{-1}\|_{\infty} \\ = \|DWD^{-1}\|_{\infty} \leq \rho(A) + \varepsilon.$$

(for $T = DU$, nonsingular).

Convergence (cont.)

Lemma 2 For any square matrix there holds

(a) $\lim_{k \rightarrow \infty} A^k = 0 \Leftrightarrow \rho(A) < 1,$

(b) If $\rho(A) < 1$ then $(I - A)^{-1} = I + A + A^2 + \dots$ is convergent.

Proof (a)' \Rightarrow ': If $\rho(A) < 1$ then choose $\varepsilon > 0$: $\rho(A) + \varepsilon < 1$. Then there exists a nonsingular T (which depends on A), such that $\|A\|_T \leq \rho(A) + \varepsilon < 1$.

$$\Rightarrow \|A^k\|_T \leq \|A\|_T^k \rightarrow 0 \Rightarrow \lim_{k \rightarrow \infty} A^k = 0$$

(a)' \Leftarrow ': If $\lim_{k \rightarrow \infty} A^k = 0$, let $\{\lambda, \mathbf{v}\}$ be an eigensolution of A , then

$\lambda^k \mathbf{v} = A^k \mathbf{v} \rightarrow 0$. This is true for all eigenvalues, thus $\rho(A) = \max |\lambda| < 1$.

(b) $(I - A)(I + A + A^2 + \dots) = I - A^{k+1}$. If $\rho(A) < 1$ then $A^k \rightarrow 0 \Rightarrow$ (b) follows. ■

Rate of convergence

Theorem 1 shows both convergence and rate of convergence ($\mathbf{e}^{(k)} = B^k \mathbf{e}^{(0)}$). The latter is difficult to compute. Also the convergence may not be monotone.

Theorem 2 Consider $C\mathbf{x}^{(k+1)} = R\mathbf{x}^{(k)} + \mathbf{b}$, $B = C^{-1}R$ and let $\rho(B) < 1$. Then

$$\|\mathbf{x}^* - \mathbf{x}^{(k)}\| \leq \frac{\|B\|}{1 - \|B\|} \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

Rate of convergence, cont.

Proof $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = B(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})$ and
 $\mathbf{x}^{(k+m+1)} - \mathbf{x}^{(k+m)} = B^{m+1}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})$. We have

$$\|\mathbf{x}^{(k+s)} - \mathbf{x}^{(k)}\| = \left\| \sum_{j=0}^{s-1} (\mathbf{x}^{(k+j+1)} - \mathbf{x}^{(k+j)}) \right\| \leq \\ \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| + \|\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}\| + \dots$$

Therefore

$$\|\mathbf{x}^{(k+m+1)} - \mathbf{x}^{(k+m)}\| \leq \sum_{j=0}^m \|B^j\| \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| = \frac{\|B\| - \|B\|^{m+1}}{1 - \|B\|} \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$

We let now $m \rightarrow \infty$, i.e., $\mathbf{x}^{k+m} \rightarrow \mathbf{x}^*$, $\|B\|^m \rightarrow 0$.

$$\Rightarrow \|\mathbf{x}^* - \mathbf{x}^{(k)}\| \leq \frac{\|B\|}{1 - \|B\|} \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|.$$



Stopping tests:

Theorem 2 can be used to get information whether the iteration error $\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$ is small enough.

In practice, most used stopping tests are:

(S1) $\|\mathbf{r}^{(k)}\| \leq \varepsilon$, residual based, absolute

Stopping tests:

Theorem 2 can be used to get information whether the iteration error $\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$ is small enough.

In practice, most used stopping tests are:

(S1) $\|\mathbf{r}^{(k)}\| \leq \varepsilon$, residual based, absolute

(S2) $\|\mathbf{r}^{(k)}\| \leq \varepsilon \|\mathbf{r}^{(0)}\|$, residual based, relative

Stopping tests:

Theorem 2 can be used to get information whether the iteration error $\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$ is small enough.

In practice, most used stopping tests are:

(S1) $\|\mathbf{r}^{(k)}\| \leq \varepsilon$, residual based, absolute

(S2) $\|\mathbf{r}^{(k)}\| \leq \varepsilon \|\mathbf{r}^{(0)}\|$, residual based, relative

(S3) $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \varepsilon$

Stopping tests:

Theorem 2 can be used to get information whether the iteration error $\mathbf{e}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$ is small enough.

In practice, most used stopping tests are:

- (S1) $\|\mathbf{r}^{(k)}\| \leq \varepsilon$, residual based, absolute
- (S2) $\|\mathbf{r}^{(k)}\| \leq \varepsilon \|\mathbf{r}^{(0)}\|$, residual based, relative
- (S3) $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \leq \varepsilon$
- (S4) $\|\mathbf{x}^* - \mathbf{x}^{(k)}\| \leq \varepsilon_0 \|\mathbf{x}^* - \mathbf{x}^{(0)}\|$.

If the latter is wanted, then we must check on (S3) and choose ε such that $\varepsilon \leq \frac{\|B\|}{1-\|B\|} \varepsilon_0 \|\mathbf{x}^* - \mathbf{x}^{(0)}\|$.

Either estimate of $\|A^{-1}\|$ or of $\|B\|$ is required.

Choices of the matrix C

Choice 'J'

Let $A = D - L - U$, where D is diagonal, U is strictly upper triangular and L is strictly lower triangular.

Let $C \equiv D$, $R = L + U$. The iterative scheme is known as **Jacobi** iteration:

$$D\mathbf{x}^{(k+1)} = (L + U)\mathbf{x}^{(k)} + \mathbf{b}$$

$$\text{Entry-wise } x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{i \neq j} a_{ij} x_j \right) ..$$

For the method to converge: $B = D^{-1}(L + U)$

$$\rho(B) \leq \|D^{-1}(L + U)\|_{\infty} = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^n \left| \frac{a_{ij}}{a_{ii}} \right|$$

We want $\rho(B) < 1$. One class of matrices, for which Jacobi method converges is when A is strictly diagonally dominant.

Choices of the matrix C

- Choice GS-B Choose $C \equiv D - U$, $R = L$

Backward Gauss-Seidel

$$(D - U)\mathbf{x}^{(k+1)} = L\mathbf{x}^{(k)} + \mathbf{b}$$

Choices of the matrix C

- Choice GS-B Choose $C \equiv D - U$, $R = L$

Backward Gauss-Seidel $(D - U)\mathbf{x}^{(k+1)} = L\mathbf{x}^{(k)} + \mathbf{b}$

- Choice GS-F Choose $C \equiv D - L$, $R = U$

Forward Gauss-Seidel $(D - L)\mathbf{x}^{(k+1)} = U\mathbf{x}^{(k)} + \mathbf{b}$

Choices of the matrix C

- Choice GS-B Choose $C \equiv D - U$, $R = L$

Backward Gauss-Seidel $(D - U)\mathbf{x}^{(k+1)} = L\mathbf{x}^{(k)} + \mathbf{b}$

- Choice GS-F Choose $C \equiv D - L$, $R = U$

Forward Gauss-Seidel $(D - L)\mathbf{x}^{(k+1)} = U\mathbf{x}^{(k)} + \mathbf{b}$

- G-S is convergent for s.p.d. matrices.

Choices of the matrix C

- Choice GS-B Choose $C \equiv D - U, R = L$

Backward Gauss-Seidel $(D - U)\mathbf{x}^{(k+1)} = L\mathbf{x}^{(k)} + \mathbf{b}$

- Choice GS-F Choose $C \equiv D - L, R = U$

Forward Gauss-Seidel $(D - L)\mathbf{x}^{(k+1)} = U\mathbf{x}^{(k)} + \mathbf{b}$

- G-S is convergent for s.p.d. matrices.

- make it more fancy: $A = D - L - U$. Then

$$\begin{aligned}\omega A &= \omega D - \omega L - \omega L + D - D \leftarrow \text{overrelaxation} \\ &= (D - \omega L) - (\omega U + (1 - \omega)D)\end{aligned}$$

Choose $C \equiv D - \omega L, R = \omega U + (1 - \omega)D$:

SOR $(D - \omega L)\mathbf{x}^{(k+1)} = [\omega U + (1 - \omega)D]\mathbf{x}^{(k)} + \omega \mathbf{b}$

SOR - back to 1940

One can see SOR as a generalization of G-S ($\omega = 1$). Rewrite

$$(D - \omega L)\mathbf{x}^{(k+1)} = [\omega U + (1 - \omega)D]\mathbf{x}^{(k)} + \omega \mathbf{b}$$

$$\text{as } \left(\frac{1}{\omega}D - L\right)\mathbf{x}^{(k+1)} = \left[\left(\frac{1}{\omega} - 1\right)D + U\right]\mathbf{x}^{(k)} + \mathbf{b}$$

For the iteration matrix $B_\omega = \left(\frac{1}{\omega}D - L\right)^{-1} \left[\left(\frac{1}{\omega} - 1\right)D + U\right]$

One can show that $\rho(B_\omega) < 1$ for $0 < \omega < 2$. Furthermore, there is an optimal value of ω , for which $\rho(B_\omega)$ is minimized:

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(\hat{B})^2}}, \quad \hat{B} = I - D_A^{-1}A.$$

SOR - cont.

Rate of convergence: Let $\lambda_i = \text{eig}(B_\omega)$.

$$\left| \prod_{i=1}^n \lambda_i \right| = |\det((1 - \omega)I + \omega D^{-1}U)| = |1 - \omega|^n. \Rightarrow \text{at least one } \lambda_i \geq |1 - \omega|.$$

$$\Rightarrow \rho(B_\omega) \geq |1 - \omega|.$$

We want $\rho(B_\omega) < 1$, i.e. $|1 - \omega| \leq \rho(B_\omega) < 1, \Rightarrow 0 < \omega < 2$.

Splittings of A

Let $A, C, R \in \mathbb{R}^{n \times n}$ and consider $A = C - R$. A splitting of A is called

- **regular** if C is monotone and $R \geq 0$ (elementwise)
- **weak regular** if C is monotone and $C^{-1}R \geq 0$
- **nonnegative** if C^{-1} exists and $C^{-1}R \geq 0$
- **convergent** if $\rho(C^{-1}R) < 1$.

Recall: A matrix is called *monotone* if $A\mathbf{x} > 0$ implies $\mathbf{x} > 0$.

Theorem: A - monotone $\Leftrightarrow A^{-1} \geq 0$.

Shortly on projectors

Projectors and properties

Definitions:

Consider C^n and a mapping $P : C^n \rightarrow C^n$.

P is called a *projector* if $P^2 = P$ (i.e. P is idempotent).

If P is a projector, then $I - P$ is also such:

$$(I - P)^2 = I - 2P + P^2 = I - P.$$

$\mathcal{N}(P) = \{\mathbf{x} \in C^n : P\mathbf{x} = 0\}$ (null space (kernel) of P)

$\mathcal{R}(P) = \{A\mathbf{x} : \mathbf{x} \in C^n\}$ (range of P).

A subspace S is called *invariant under a square matrix A* whenever $AS \in S$.

Properties:

P1: $\mathcal{N}(P) \cap \mathcal{R}(P) = \{0\}$ Indeed,

if $\mathbf{x} \in \mathcal{R}(P) \Rightarrow \exists \mathbf{y} : \mathbf{y} = P\mathbf{x} \Rightarrow P\mathbf{y} = P^2\mathbf{x} = P\mathbf{x} \Rightarrow \mathbf{y} = \mathbf{x} \Rightarrow \mathbf{x} = P\mathbf{x}$

If $\mathbf{x} \in \mathcal{N}(P) \Rightarrow P\mathbf{x} = 0 \Rightarrow \mathbf{x} = P\mathbf{x} \Rightarrow \mathbf{x} = 0$.

P2: $\mathcal{N}(P) = \mathcal{R}(I - P)$

$\mathbf{x} \in \mathcal{N}(P) \Rightarrow P\mathbf{x} = 0$. Then $\mathbf{x} = I\mathbf{x} - P\mathbf{x} = (I - P)\mathbf{x}$.

$\mathbf{y} \in \mathcal{R}(I - P) \Rightarrow \mathbf{y} = (I - P)\mathbf{y} \Rightarrow P\mathbf{y} = 0$.

P3: $C^n = \mathcal{R}(P) \oplus \mathcal{N}(P)$.

P4: Given two subspaces K and L of same dimension m , the following two conditions are mathematically equivalent:

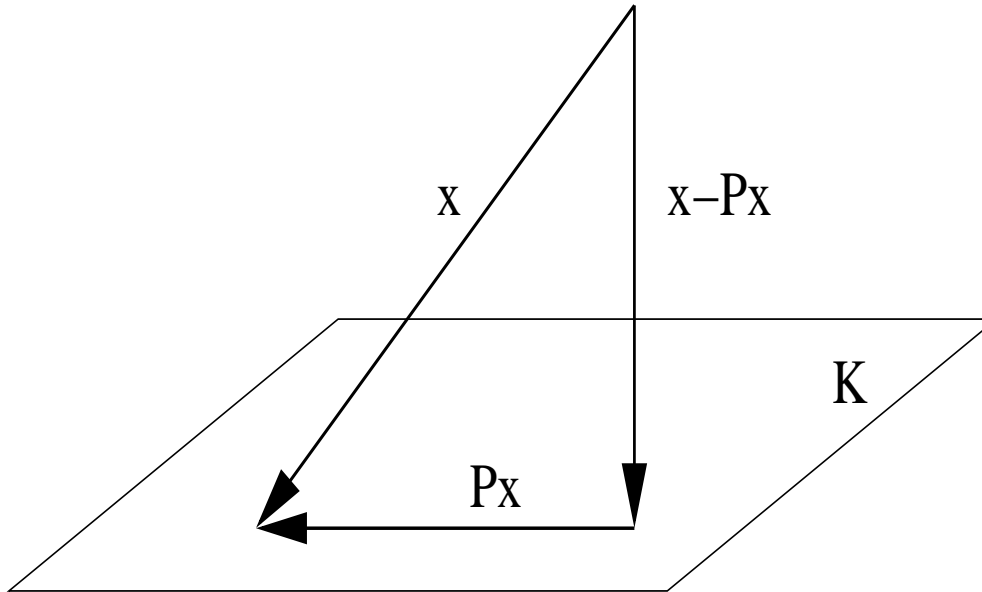
- (i) No nonzero vector in K is orthogonal to L
- (ii) $\forall \mathbf{x} \in C^n \exists$ unique vector $\mathbf{y} : \mathbf{y} \in K, \mathbf{x} - \mathbf{y} \in L$.

Proof (i) \Rightarrow (ii): $K \cap L^\perp = \{\emptyset\} \Rightarrow C^n = K \oplus L^\perp \Rightarrow \forall \mathbf{x} \in C^n : \mathbf{x} = \mathbf{y} + \mathbf{z}$,
where $\mathbf{y} \in K$ and $\mathbf{z} \in L^\perp$. Thus, $\mathbf{z} = \mathbf{x} - \mathbf{y} \Rightarrow$ (ii). ■

P5: *Orthogonal and oblique projectors*

P is *orthogonal* if $\mathcal{N}(P) = \mathcal{R}(P)^\perp$. Otherwise P is *oblique*.

Thus, if P is orthogonal onto K , then $P\mathbf{x} \in K$ and $(I - P)\mathbf{x} \perp K$. Equivalently, $((I - P)\mathbf{x}, \mathbf{y}) = 0, \forall \mathbf{y} \in K$.



Properties (cont.)

P6: If P is orthogonal, then $\|P\| = 1$.

Proof $\mathbf{x} = P\mathbf{x} + (I - P)\mathbf{x} = \mathbf{y} - \mathbf{z}$.

Then

$$(\mathbf{y}, \mathbf{z}) = 0 : (P\mathbf{x}, (I - P)\mathbf{x}) = (P\mathbf{x}, \mathbf{x}) - (P\mathbf{x}, P\mathbf{x}) = (P\mathbf{x}, \mathbf{x}) - (P\mathbf{x}, \mathbf{x}) = 0.$$

$$\Rightarrow \|\mathbf{x}\|_2^2 = \|P\mathbf{x}\|_2^2 + \|(I - P)\mathbf{x}\|_2^2$$

$$\Rightarrow \|\mathbf{x}\|_2^2 \geq \|P\mathbf{x}\|_2^2 \Rightarrow \frac{\|P\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \leq 1, \forall \mathbf{x} \in C^n.$$

However, for $\tilde{\mathbf{x}} \in \mathcal{R}(P)$ there holds $\frac{\|P\tilde{\mathbf{x}}\|_2^2}{\|\tilde{\mathbf{x}}\|_2^2} = 1$. Thus, $\|P\| = 1$. ■

P7: Any orthogonal projector has only two eigenvalues 0 and 1. Any vector from $\mathcal{R}(P)$ is an eigenvector to $\lambda = 1$. Any vector from $\mathcal{N}(P)$ is an eigenvector to $\lambda = 0$.

Theorem 3 Let P be orthogonal onto K . Then for any vector $\mathbf{x} \in C^n$ there holds

$$\min_{\mathbf{y} \in K} \|\mathbf{x} - \mathbf{y}\|_2 = \|\mathbf{x} - P\mathbf{x}\|_2. \quad (8)$$

Proof For any $\mathbf{y} \in K$, $P\mathbf{x} - \mathbf{y} \in K$, $P\mathbf{x} \in K$, $(I - P)\mathbf{x} \perp K$

$$\|\mathbf{x} - \mathbf{y}\|_2^2 = \|(\mathbf{x} - P\mathbf{x}) + (P\mathbf{x} - \mathbf{y})\|_2^2 =$$

$$\|\mathbf{x} - P\mathbf{x}\|_2^2 + \|P\mathbf{x} - \mathbf{y}\|_2^2 + 2(\mathbf{x} - P\mathbf{x}, P\mathbf{x} - \mathbf{y}) = \|\mathbf{x} - P\mathbf{x}\|_2^2 + \|P\mathbf{x} - \mathbf{y}\|_2^2.$$

Therefore, $\|\mathbf{x} - \mathbf{y}\|_2^2 \geq \|\mathbf{x} - P\mathbf{x}\|_2^2 \forall \mathbf{y} \in K$ and the minimum is reached for $\mathbf{y} = P\mathbf{x}$. ■

Corollary 1 Let $K \subset C^n$ and $\mathbf{x} \in C^n$ be given. Then $\min_{\mathbf{y}} \|\mathbf{x} - \mathbf{y}\|_2 = \|\mathbf{x} - \mathbf{y}^*\|_2$

is equivalent to $\mathbf{y}^* \in K$ and $\mathbf{x} - \mathbf{y}^* \perp K$.

Iterative solution methods

- Steepest descent
- conjugate gradient method (CG)
- Generalized conjugate gradient method (GCG)
- ORTHOMIN
- Minimal residual method (MINRES)
- Generalized minimal residual method (GMRES)
- Lanczos method
- Arnoldi method
- Orthogonal residual method (ORTHORES)
- Full orthogonalization method (FOM)
- Incomplete orthogonalization method (IOM)

Iterative solution methods

- SYMMLQ
- Biconjugate gradient method (BiCG)
- BiCGStab
- Conjugate gradients squared (CGS)
- Minimal residual method (MR)
- Quasiminimal residual method
- ...

Projection-based iterative methods

General framework – projection methods

Want to solve $\mathbf{b} - A\mathbf{x} = \mathbf{0}$, $\mathbf{b}, \mathbf{x} \in R^n$, $A \in R^{n \times n}$

Instead, choose two subspaces $L \subset R^n$ and $K \subset R^n$ and

* find $\tilde{\mathbf{x}} \in \mathbf{x}^{(0)} + K$, such that $\mathbf{b} - A\tilde{\mathbf{x}} \perp L$

K - search space

L - subspace of constraints

* - basic projection step

The framework is known as Petrov-Galerkin conditions.

There are two major classes of projection methods:

- orthogonal - if $K \equiv L$,
- oblique - if $K \neq L$.

Notations:

$$\tilde{\mathbf{x}} = \mathbf{x}^0 + \delta \text{ - } (\delta \text{ - correction})$$

$$\mathbf{r}^0 = \mathbf{b} - A\mathbf{x}^0 \text{ (}\mathbf{r}^0 \text{ - residual)}$$

* find $\delta \in K$, such that $\mathbf{r}^0 - A\delta \perp L$

Matrix formulation

Choose a basis in K and L : $V = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ and $W = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m\}$.

Then, $\tilde{\mathbf{x}} = \mathbf{x}^0 + \delta = \mathbf{x}^0 + V\mathbf{y}$ for some $\mathbf{y} \in \mathbb{R}^m$.

The orthogonality condition can be written as

$$(**) \quad \boxed{W^T(\mathbf{r}^0 - AV\mathbf{y})}$$

which is exactly the Petrov-Galerkin condition.

From (**) we get

$$\begin{aligned} W^T \mathbf{r}^0 &= W^T AV\mathbf{y} \\ \mathbf{y} &= (W^T AV)^{-1} W^T \mathbf{r}^0 \\ \tilde{\mathbf{x}} &= \mathbf{x}^0 + V(W^T AV)^{-1} W^T \mathbf{r}^0 \end{aligned}$$

In practice, $m < n$, even $m \ll n$, for instance, $m = 1$.

The matrix $W^T AV$ will be small and, hopefully, with a nice structure.

!!! $W^T AV$ should be invertible.

A prototype projection-based iterative method:

Given $\mathbf{x}^{(0)}$; $\mathbf{x} = \mathbf{x}^{(0)}$
Until convergence do:
 Choose K and L
 Choose basis V in K and W in L
 Compute $\mathbf{r} = \mathbf{b} - A\mathbf{x}$
 $\mathbf{y} = (W^T A V)^{-1} W^T \mathbf{r}$
 $\mathbf{x} = \mathbf{x} + V\mathbf{y}$

Degrees of freedom: m, K, L, V, W .

Clearly, if $K \equiv L$, then $V = W$.

Plan:

- (1) Consider two important cases: $L = K$ and $L = AK$
- (2) Make a special choice of K .

Property 1:

Theorem 4 Let A be square, $L = AK$. Then a vector $\tilde{\mathbf{x}}$ is an oblique projection on K orthogonally to AK with a starting vector \mathbf{x}^0 if and only if $\tilde{\mathbf{x}}$ minimizes the 2-norm of the residual over $\mathbf{x}^0 + K$, i.e.,

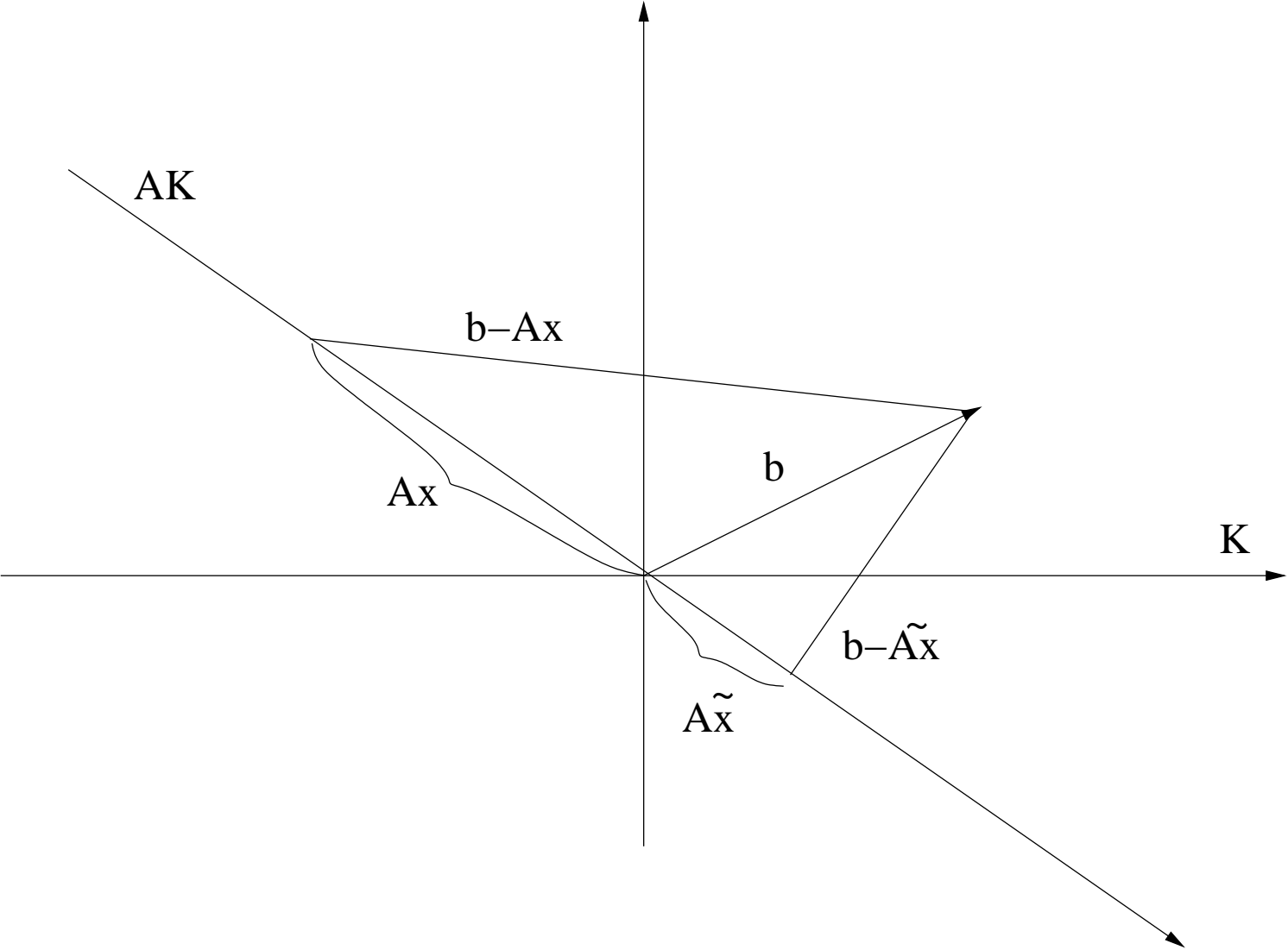
$$\|\mathbf{r} - A\tilde{\mathbf{x}}\|_2 = \min_{\mathbf{x} \in \mathbf{x}^0 + K} \|\mathbf{r} - A\mathbf{x}\|_2. \quad (9)$$

Thus, the residual decreases monotonically.

Referred to as *minimal residual methods*

CR, GCG, GMRES, ORTHOMIN

Property 1:



Property 2:

Theorem 5 Let A be symmetric positive definite, i.e., it defines a scalar product $(A\cdot, \cdot)$ and a norm $\|\cdot\|_A$. Let $L = K$, i.e., $\mathbf{r}^0 - A\tilde{\mathbf{x}} \perp K$. Then a vector $\tilde{\mathbf{x}}$ is an orthogonal projection onto K with a starting vector \mathbf{x}^0 if and only if it minimizes the A -norm of the error $\mathbf{e} = \mathbf{x}^* - \mathbf{x}$ over $\mathbf{x}^0 + K$, i.e.,

$$\|\mathbf{x}^* - \tilde{\mathbf{x}}\|_A = \min_{\mathbf{x} \in \mathbf{x}^0 + K} \|\mathbf{x}^* - \mathbf{x}\|_A. \quad (10)$$

The error decreases monotonically in the A -norm.
Error-projection methods.

Example: $m = 1$

Consider two vectors: \mathbf{d} and \mathbf{e} . Let $K = \text{span}\{\mathbf{d}\}$ and $L = \text{span}\{\mathbf{e}\}$.
Then $\tilde{\mathbf{x}} = \mathbf{x}^0 + \alpha\mathbf{d}$ ($\delta = \alpha\mathbf{d}$) and the orthogonality condition reads as:

$$\mathbf{r}^0 - A\delta \perp \mathbf{e} \Rightarrow (\mathbf{r}^0 - A\delta, \mathbf{e}) = 0 \Rightarrow \alpha(A\mathbf{d}, \mathbf{e}) = (\mathbf{r}^0, \mathbf{e}) \Rightarrow \alpha = \frac{(\mathbf{r}^0, \mathbf{e})}{(A\mathbf{d}, \mathbf{e})}.$$

If $\mathbf{d} = \mathbf{e}$ - Steepest Descent method (minimization on a line).

If we minimize over a plane - ORTHOMIN.

Choice of K :

$$K = \mathcal{K}^m(A, \mathbf{v}) = \{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \dots, A^{m-1}\mathbf{v}\}$$

Krylov subspace methods

- $L = K = \mathcal{K}^m(A, \mathbf{r}^0)$ and A spd \Rightarrow CG
- $L = AK = A\mathcal{K}^m(A, \mathbf{r}^0) \Rightarrow$ GMRES

Why are Krylov subspaces of interest?

How to construct a basis for \mathcal{K} ?

CG

Arnoldi's method for general matrices

Consider $\mathcal{K}^m(A, \mathbf{v}) = \{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \dots, A^{m-1}\mathbf{v}\}$, generated by some matrix A and vector \mathbf{v} .

1. Choose a vector \mathbf{v}_1 such that $\|\mathbf{v}_1\| = 1$
2. For $j = 1, 2, \dots, m$
3. For $i = 1, 2, \dots, j$
4. $h_{ij} = (A\mathbf{v}_j, \mathbf{v}_i)$
5. End
6. $\mathbf{w}_j = A\mathbf{v}_j - \sum_{i=1}^j h_{ij}\mathbf{v}_i$
7. $h_{j+1,j} = \|\mathbf{w}_j\|$
8. If $h_{j+1,j} = 0$, stop
9. $\mathbf{v}_{j+1} = \mathbf{w}_j/h_{j+1,j}$
10. End

The result of Arnoldi's process

• $V^m = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ is an orthonormal basis in $\mathcal{K}^m(A, \mathbf{v})$

• $AV^m = V^m H^m + \mathbf{w}_{m+1} \mathbf{e}_m^T$

$$\begin{array}{c}
 \boxed{\begin{array}{c} A \\ (n,n) \end{array}} * \boxed{\begin{array}{c} V^m \\ (n,m) \end{array}} = \boxed{\begin{array}{c} V^m \\ (n,m) \end{array}} * \boxed{\begin{array}{c} H^m \\ (m,m) \end{array}} + \begin{array}{c} \mathbf{w}^{m+1} \quad (\mathbf{e}^m)^T \\ | \quad \boxed{} \\ | \quad (1,m) \\ | \\ | \\ | \\ (n,1) \end{array}
 \end{array}$$

Arnoldi's process - example

$$H^3 = \begin{bmatrix} (A\mathbf{v}_1, \mathbf{v}_1) & (A\mathbf{v}_2, \mathbf{v}_1) & (A\mathbf{v}_3, \mathbf{v}_1) \\ \|\mathbf{w}_1\| & (A\mathbf{v}_2, \mathbf{v}_2) & (A\mathbf{v}_3, \mathbf{v}_2) \\ 0 & \|\mathbf{w}_2\| & (A\mathbf{v}_3, \mathbf{v}_3) \end{bmatrix}$$

Since $V^{m+1} \perp \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ then it follows that $(V^m)^T AV^m = H^m$.
 H^m is an upper-Hessenberg matrix.

Arnoldi's method for symmetric matrices

Let now A be real symmetric matrix. Then the Arnoldi method reduces to the Lanczos method.

Recall: $H^m = (V^m)^T A V^m$

If A is symmetric, then H^m must be symmetric too, i.e., H^m is three-diagonal

$$H^m = \begin{bmatrix} \gamma_1 & \beta_2 & & & \\ \beta_2 & \gamma_2 & \beta_3 & & \\ & & \ddots & & \\ & & & \beta_m & \gamma_m \end{bmatrix}$$

Thus, the vectors \mathbf{v}^j satisfy a three-term recursion:

$$\beta_{i+1} \mathbf{v}^{i+1} = A \mathbf{v}^i - \gamma_i \mathbf{v}^i - \beta_i \mathbf{v}^{i-1}$$

Lanczos algorithm to solve symmetric linear systems

Given: $\mathbf{x}^{(0)}$
Compute $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$, $\beta = \|\mathbf{r}^{(0)}\|$, $\mathbf{v}^1 = \mathbf{r}^{(0)} / \beta$
Set $\beta_1 = 0$ and $\mathbf{v}^0 = \mathbf{0}$
For $j = 1 : m$
 $\mathbf{w}^j = A\mathbf{v}^j - \beta_j \mathbf{v}^{j-1}$
 $\gamma_j = (\mathbf{w}^j, \mathbf{v}^j)$
 $\mathbf{w}^j = \mathbf{w}^j - \gamma_j \mathbf{v}^j$
 $\beta_{j+1} = \|\mathbf{w}^j\|_2$, if $\beta_{j+1} = 0$, go out of the loop
 $\mathbf{v}^{j+1} = \mathbf{w}^j / \beta_{j+1}$
End
Set $T_m = \text{tridiag}\{\beta_i, \gamma_i, \beta_{i+1}\}$
Compute $\mathbf{y}^m = T_m^{-1}(\beta \mathbf{e}_1)$
 $\mathbf{x}^m = \mathbf{x}^{(0)} + V^m \mathbf{y}^m$

Leads to three-term CG.

To solve, factor first $T_m = LL^T$ and then $\mathbf{x}^m = \mathbf{x}^{(0)} + V^m L^{-T} L^{-1} \beta \mathbf{e}_1$

Direct Lanczos: the factorization of T_m

The coefficients on the direct Lanczos algorithm correspond to the following factorization of T_m :

$$T_m = \begin{bmatrix} \gamma_1 & \beta_2 & & \\ \beta_2 & \gamma_2 & \beta_3 & \\ & & \ddots & \\ & \beta_m & & \gamma_m \end{bmatrix} = \begin{bmatrix} 1 & & & \\ \lambda_2 & 1 & & \\ & & \ddots & \\ & & & \lambda_m & 1 \end{bmatrix} \begin{bmatrix} \eta_1 & \beta_2 & & \\ & \eta_2 & \beta_3 & \\ & & \ddots & \\ & & & \eta_m \end{bmatrix}$$

Direct Lanczos

Instead of factorizing at the end, Gauss factorization without pivoting can be performed while constructing T .

Recall $\mathbf{x}^m = \mathbf{x}^0 + V^m L^{-T} L^{-1} \beta \mathbf{e}_1$ and let $G = V^m L^{-T}$ and $\mathbf{z} = L^{-1} \beta \mathbf{e}_1$

Compute $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$, $\xi_1 = \beta = \|\mathbf{r}^{(0)}\|$, $\mathbf{v}^1 = 1/\beta \mathbf{r}^{(0)}$

$$\lambda_1 = 1, \beta_1 = 1, \mathbf{g}^0 = \mathbf{0}$$

Set $\beta_1 = 0$ and $\mathbf{v}^0 = \mathbf{0}$

For $j = 1, 2, \dots$ until convergence

$$\mathbf{w} = A\mathbf{v}^j - \beta_j \mathbf{v}^{j-1}$$

$$\gamma_j = (\mathbf{w}, \mathbf{v}^j)$$

$$\text{if } j > 1, \lambda_j = \beta_j / \eta_{j-1}, \xi_j = -\lambda_j \xi_{j-1}$$

$$\eta_j = \gamma_j - \lambda_j \beta_j$$

$$\mathbf{g}^j = (\eta_j)^{-1} (\mathbf{v}^j - \beta_j \mathbf{g}^{j-1})$$

$$\mathbf{x}^j = \mathbf{x}^{j-1} + \xi_j \mathbf{g}^j, \text{ stop if convergence is reached}$$

$$\mathbf{w} = \mathbf{w} - \gamma_j \mathbf{v}^j$$

$$\beta_{j+1} = \|\mathbf{w}\|;$$

$$\mathbf{v}^{j+1} = \mathbf{w} / \beta_{j+1}$$

End

How to construct a basis for \mathcal{K} ?

CG

Arnoldi's method for general matrices

Consider $\mathcal{K}^m(A, \mathbf{v}) = \{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \dots, A^{m-1}\mathbf{v}\}$, generated by some matrix A and vector \mathbf{v} .

1. Choose a vector \mathbf{v}_1 such that $\|\mathbf{v}_1\| = 1$
2. For $j = 1, 2, \dots, m$
3. For $i = 1, 2, \dots, j$
4. $h_{ij} = (A\mathbf{v}_j, \mathbf{v}_i)$
5. End
6. $\mathbf{w}_j = A\mathbf{v}_j - \sum_{i=1}^j h_{ij}\mathbf{v}_i$
7. $h_{j+1,j} = \|\mathbf{w}_j\|$
8. If $h_{j+1,j} = 0$, stop
9. $\mathbf{v}_{j+1} = \mathbf{w}_j/h_{j+1,j}$
10. End

The algorithm breaks down in step j , i.e., $h_{j+1,j} = 0$, if and only if the minimal polynomial of A is of degree j .

The result of Arnoldi's process

• $V^m = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ is an orthonormal basis in $\mathcal{K}^m(A, \mathbf{v})$

• $AV^m = V^m H^m + \mathbf{w}_{m+1} \mathbf{e}_m^T$

$$\begin{array}{c} \boxed{A} \\ (n,n) \end{array} * \begin{array}{c} \boxed{V^m} \\ (n,m) \end{array} = \begin{array}{c} \boxed{V^m} \\ (n,m) \end{array} * \begin{array}{c} \boxed{H^m} \\ (m,m) \end{array} + \begin{array}{c} \boxed{\mathbf{w}^{m+1}} \\ (n,1) \end{array} * \begin{array}{c} \boxed{(\mathbf{e}^m)^T} \\ (1,m) \end{array}$$

Arnoldi's process - example

$$H^3 = \begin{bmatrix} (A\mathbf{v}_1, \mathbf{v}_1) & (A\mathbf{v}_2, \mathbf{v}_1) & (A\mathbf{v}_3, \mathbf{v}_1) \\ \|\mathbf{w}_1\| & (A\mathbf{v}_2, \mathbf{v}_2) & (A\mathbf{v}_3, \mathbf{v}_2) \\ 0 & \|\mathbf{w}_2\| & (A\mathbf{v}_3, \mathbf{v}_3) \end{bmatrix}$$

Since $V^{m+1} \perp \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ then it follows that $(V^m)^T AV^m = H^m$.
 H^m is an upper-Hessenberg matrix.

Arnoldi's method for symmetric matrices

Let now A be real symmetric matrix. Then the Arnoldi method reduces to the Lanczos method.

Recall: $H^m = (V^m)^T A V^m$

If A is symmetric, then H^m must be symmetric too, i.e., H^m is three-diagonal

$$H^m = \begin{bmatrix} \gamma_1 & \beta_2 & & & \\ \beta_2 & \gamma_2 & \beta_3 & & \\ & & \ddots & & \\ & & & \beta_m & \gamma_m \end{bmatrix}$$

Thus, the vectors \mathbf{v}^j satisfy a three-term recursion:

$$\beta_{i+1} \mathbf{v}^{i+1} = A \mathbf{v}^i - \gamma_i \mathbf{v}^i - \beta_i \mathbf{v}^{i-1}$$

Lanczos algorithm to solve symmetric linear systems

Given: $\mathbf{x}^{(0)}$
Compute $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$, $\beta = \|\mathbf{r}^{(0)}\|$, $\mathbf{v}^1 = \mathbf{r}^{(0)} / \beta$
Set $\beta_1 = 0$ and $\mathbf{v}^0 = \mathbf{0}$
For $j = 1 : m$
 $\mathbf{w}^j = A\mathbf{v}^j - \beta_j \mathbf{v}^{j-1}$
 $\gamma_j = (\mathbf{w}^j, \mathbf{v}^j)$
 $\mathbf{w}^j = \mathbf{w}^j - \gamma_j \mathbf{v}^j$
 $\beta_{j+1} = \|\mathbf{w}^j\|_2$, if $\beta_{j+1} = 0$, go out of the loop
 $\mathbf{v}^{j+1} = \mathbf{w}^j / \beta_{j+1}$
End
Set $T_m = \text{tridiag}\{\beta_i, \gamma_i, \beta_{i+1}\}$
Compute $\mathbf{y}^m = T_m^{-1}(\beta \mathbf{e}_1)$
 $\mathbf{x}^m = \mathbf{x}^{(0)} + V^m \mathbf{y}^m$

Leads to three-term CG.

To solve, factor first $T_m = LL^T$ and then $\mathbf{x}^m = \mathbf{x}^{(0)} + V^m L^{-T} L^{-1} \beta \mathbf{e}_1$

The CG method:

The CG algorithm using the above relations:

$$\text{Initialize: } \mathbf{r}^{(0)} = A\mathbf{x}^{(0)} - \mathbf{b}, \mathbf{g}^{(0)} = \mathbf{r}^{(0)}$$

For $k = 0, 1, \dots$, until convergence

$$\tau_k = \frac{(\mathbf{r}^{(k)}, \mathbf{r}^{(k)})}{(A\mathbf{g}^k, \mathbf{g}^{(k)})}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \tau_k \mathbf{g}^k$$

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} + \tau_k A\mathbf{g}^k$$

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

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

end



$\mathbf{r}^{(k)}$ – iteratively computed residuals

\mathbf{g}^k – search directions

Note: the coefficients β_k are different from those in the Lanczos method.

CG: computer implementation

```
x = x0
r = A*x-b
delta0 = (r,r)
g = -r
Repeat: h = A*g
        tau = delta0/(g,h)
        x = x + tau*g
        r = r + tau*h
        delta1 = (r,r)
        if delta1 <= eps, stop
        beta = delta1/delta0
        g = -r + beta*g
```

Optimality properties of the CG method

Opt1: Mutually orthogonal search directions: $(\mathbf{g}^{k+1}, A\mathbf{g}^j) = 0, j = 0, \dots, k$

Opt2: There holds $\mathbf{r}^{(k+1)} \perp K_m(A, \mathbf{r}^{(0)})$, i.e.,
 $(\mathbf{r}^{(k+1)}, A\mathbf{r}^{(k)}) = 0, j = 0, \dots, k$

Opt3: Optimization property: $\|\mathbf{r}^{(k)}\|$ smallest possible at any step, since CG minimizes the functional $f(\mathbf{x}) = 1/2(\mathbf{x}, A\mathbf{x}) - (\mathbf{x}, \mathbf{b})$

Opt4: $(\mathbf{e}^{(k+1)}, A\mathbf{g}^j) = (\mathbf{g}^{k+1}, A\mathbf{g}^j) = (\mathbf{r}^{(k+1)}, \mathbf{r}^{(k)}) = 0, j = 0, \dots, k$

Opt5: Finite termination property: there are n breakdowns of the CG algorithm.

Reasoning: if $\mathbf{g}^j = \mathbf{0}$ then τ_k is not defined. the vectors \mathbf{g}^j are computed from the formula $\mathbf{g}^k = \mathbf{r}^{(k)} + \beta_k \mathbf{g}^{k-1}$. Then

$$0 = (\mathbf{r}^{(k)}, \mathbf{g}^j) = -(\mathbf{r}^{(k)}, \mathbf{r}^{(k)}) + \beta_k \underbrace{(\mathbf{r}^{(k)}, \mathbf{g}^{k-1})}_0, \Rightarrow \mathbf{r}^{(k)} = \mathbf{0}, \text{ i.e., the}$$

solution is already found.

As soon as $\mathbf{x}^{(k)} \neq \mathbf{x}_{exact}$, then $\mathbf{r}^{(k)} \neq \mathbf{0}$ and then $\mathbf{g}^{k+1} \neq \mathbf{0}$.

However, we can generate at most n mutually orthogonal vectors in R^n , thus, CG has a finite termination property.



Convergence analysis

Convergence of the CG method

Theorem: In exact arithmetic, CG has the property that $\mathbf{x}_{exact} = \mathbf{x}^{(m)}$ for some $m \leq n$, where n is the order of A .

Rate of convergence of the CG method

Theorem: Let A is symmetric and positive definite.

Suppose that for some set S , containing all eigenvalues of A , for some polynomial $\tilde{P}(\lambda) \in \Pi_k^1$ and some constant M there holds $\max_{\lambda \in S} |\tilde{P}(\lambda)| \leq M$.

Then,

$$\|\mathbf{x}_{exact} - \mathbf{x}^{(k)}\|_A \leq M \|\mathbf{x}_{exact} - \mathbf{x}^{(0)}\|_A.$$

$$\|\mathbf{e}^k\|_A \leq 2 \left[\frac{\kappa(A) + 1}{\kappa(A) - 1} \right]^k \|\mathbf{e}^0\|_A$$

Rate of convergence (cont)

Repeat:

$$\|\mathbf{e}^k\|_A \leq 2 \left[\frac{\kappa(A) + 1}{\kappa(A) - 1} \right]^k \|\mathbf{e}^0\|_A$$

Seek now the smallest k , such that

$$\|\mathbf{e}^k\|_A \leq \varepsilon \|\mathbf{e}^0\|_A$$

$$\text{we want } \left(\frac{\kappa+1}{\kappa-1} \right)^k > \frac{2}{\varepsilon}$$

$$\Rightarrow k \ln \left(\frac{\kappa+1}{\kappa-1} \right) > \ln \left(\frac{2}{\varepsilon} \right)$$

$$\Rightarrow k > \ln \left(\frac{2}{\varepsilon} \right) / \ln \left(\frac{\kappa+1}{\kappa-1} \right)$$

$$\Rightarrow k > \frac{1}{2} \sqrt{\kappa} \ln \left(\frac{2}{\varepsilon} \right)$$



The GMRES method

Basic GMRES

Choose \mathbf{v}_1 to be the normalized $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$.

Any vector $\mathbf{x} \in \mathbf{x}_0 + K$ is of the form $\mathbf{x} = \mathbf{x}_0 + V_m\mathbf{y}$. Then

$$\begin{aligned}\mathbf{b} - A\mathbf{x} &= \mathbf{b} - A(\mathbf{x}_0 + V_m\mathbf{y}) \\ &= \mathbf{r}_0 - AV_m\mathbf{y} \\ &= \beta\mathbf{v}_1 - V_{m+1}\tilde{H}_m\mathbf{y} \\ &= V_{m+1}(\beta\mathbf{e}_1 - \tilde{H}_m\mathbf{y}).\end{aligned}$$

Since the columns of V_{m+1} are orthonormal, then

$$\|\mathbf{b} - A\mathbf{x}\|_2 = \|\beta\mathbf{e}_1 - \tilde{H}_m\mathbf{y}\|_2.$$

Basic GMRES

1. Compute $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$, $\beta = \|\mathbf{r}_0\|_2$ and $\mathbf{v}_1 = \mathbf{r}_0/\beta$
2. For $j = 1, 2, \dots, m$
3. Compute $\mathbf{w}_j = A\mathbf{v}_j$
4. For $i = 1, 2, \dots, j$
5. $h_{ij} = (\mathbf{w}_j, \mathbf{v}_i)$
6. $\mathbf{w}_j = \mathbf{w}_j - h_{ij}\mathbf{v}_i$
7. End
8. $h_{j+1,j} = \|\mathbf{w}_j\|_2$; if $h_{j+1,j} = 0$, set $m = j$, goto 11
9. $\mathbf{v}_{j+1} = \mathbf{w}_j/h_{j+1,j}$
10. End
11. Define the $(m+1) \times m$ Hessenberg matrix $\tilde{H}_m = \{h_{ij}\}$, $1 \leq i \leq m+1$, $1 \leq j \leq m$
12. Compute \mathbf{y}_m as the minimizer of $\|\beta\mathbf{e}_1 - \tilde{H}_m\mathbf{y}\|_2$ and $\mathbf{x}_m = \mathbf{0} + V_m\mathbf{y}_m$

GMRES:

- No breakdown of GMRES
- As m increases, storage and work per iteration increase fast. Remedies:
 - Restart (keep m constant)
 - Truncate the orthogonalization process
- The norm of the residual in the GMRES method is monotonically decreasing. However, the convergence may stagnate. The rate of convergence of GMRES cannot be determined so easy as that of CG.

References

- [1] O. Axelsson, *Iterative solution methods*, Cambridge Univ. Press, 1994.
- [2] G. Golub and C.F. van Loan, *Matrix computations*, The Johns Hopkins University Press, 1996 (Third edition).
- [3] A. Greenbaum, *Iterative methods for solving linear systems*, Frontiers in Applied Mathematics, SIAM, 1997.
- [4] A. Greenbaum, Estimating the attainable accuracy of recursively computed residual methods, *SIAM Journal on Matrix Analysis and Applications*, 18 (3), 1997, 535-551.
- [5] Y. Saad, *Iterative methods for sparse linear systems*, PWS Publishing Company, 1996.
- [6] L.N. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM. Philadelphia, 1997.
- [7] E.E. Tyrtyshnikov, *A brief introduction to Numerical Analysis*, Birkhäuser, 1997.