THE CONJUGATE GRADIENT METHOD

MARTIN H. GUTKNECHT

mhg@sam.math.ethz.ch

http://www.sam.math.ethz.ch/~mhg/

SEMINAR FOR APPLIED MATHEMATICS

ETH ZURICH

PART I OF SHORT COURSE AT EMS WORKSHOP '99

To solve: **linear system**

$$\boxed{\mathbf{Ax} = \mathbf{b}}$$

where $\mathbf{A}$ is $N \times N$, nonsingular, large, and sparse.

large: say, $500 \leq N \leq 10'000'000$

sparse: most elements are zero;
say, 5 to 50 nonzero elements per row

Such systems appear in most applications of scientific computing.

$$\boxed{\text{PDEs}} \quad \overset{\text{discretization/linearization}}{\Longrightarrow} \quad \boxed{\begin{array}{c} \mathbf{Ax} = \mathbf{b} \\ \text{or} \\ \mathbf{Ax} = \lambda\mathbf{x} \end{array}}$$

Often, in such problems most computer time is spent for repeatedly solving the linear system or the eigenvalue problem, respectively.

Classical, standard way to solve $\mathbf{Ax} = \mathbf{b}$:

$\boxed{\textbf{Gauss elimination} \approx \textbf{LU-decomposition: } \mathbf{A} = \mathbf{LU}}$

Cost:	$\sim \frac{2}{3}N^3$ FLOPs,	in general
	$\sim 2b^2N$ FLOPs	if $\mathbf{A}$ banded, bandwidth $b \ll N$
	reduction by $\sim 50\%$	if $\mathbf{A}$ sym. pos. def.

There are variations of Gauss elimination for sparse linear systems, so-called **sparse direct solvers**.

But the more $\mathbf{A}$ differs from a narrowly banded matrix, the less effective they are due to **fill-in**.

If original physical problem is 2- or 3-dimensional, $\mathbf{A}$ is far from being narrowly banded.

A mathematician's approach to solving $\mathbf{Ax} = \mathbf{b}$ iteratively: rewrite system in fixed point form

$$\boxed{\mathbf{x} = \mathbf{Bx} + \mathbf{b}} \quad \text{with } \mathbf{B} :\equiv \mathbf{I} - \mathbf{A}\,,$$

and apply **Picard iteration** or **fixed point iteration**: starting with some $\mathbf{x}_0$, we compute

$$\boxed{\mathbf{x}_{n+1} := \mathbf{Bx}_n + \mathbf{b}} \tag{1}$$

$\mathbf{x}_n$: nth **iterate** = approximation of fixed point $\mathbf{x}_\star$

If $a_{k,k} = 1$ $(\forall k)$, *i.e.*, $b_{k,k} = 0$ $(\forall k)$, then this is also called **Jacobi iteration**.

Convergence of (1):

$$\|\mathbf{B}\| < 1 \quad \Longrightarrow \quad \mathbf{x}_n \to \mathbf{x}_\star\,.$$

More exactly,

$$\mathbf{x}_n \to \mathbf{x}_\star \quad \Longleftrightarrow \quad \rho(\mathbf{B}) < 1\,.$$

If this is satisfied in practice, then typically only with $\rho(\mathbf{B})$ nearly 1, so that convergence is very slow.

The simplest application:

$$\begin{aligned} u'' &= f \quad \text{on } (0,1) \\ u(0) = u(1) &= 0 \end{aligned}$$

The difference method leads to the linear system

$$\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_n \end{bmatrix} \frac{1}{N^2}$$

whose matrix has the eigenvalues

$$\lambda_j = 2 - 2\cos\frac{j\pi}{N+1} = 4\sin^2\frac{j\pi}{2(N+1)} .$$

Thus, when we divide every equation by 2, so that $a_{k,k} = 1$ and $b_{k,k} = 0$, then

$$\rho(\mathbf{B}) = \cos\frac{\pi}{N+1} .$$

For example, when $N = 100$, then $\rho(\mathbf{B}) = 0.9995\,1628...$

In order to reduce the error by a factor 10, we need 4760 iterations!

To check convergence, *i.e.*, to check the 'quality' of the approximation, we need nth **residual**

$$\mathbf{r}_n :\equiv \mathbf{b} - \mathbf{A}\mathbf{x}_n = \mathbf{B}\mathbf{x}_n + \mathbf{b} - \mathbf{x}_n$$

since we cannot compute the nth **error**

$$\mathbf{f}_n :\equiv \mathbf{x}_n - \mathbf{x}_\star .$$

Note that

$$\mathbf{r}_n = -\mathbf{A}(\mathbf{x}_n - \mathbf{x}_\star) = -\mathbf{A}\mathbf{f}_n .$$

We then can rewrite the Picard iteration as

$$\boxed{\mathbf{x}_{n+1} := \mathbf{x}_n + \mathbf{r}_n}$$

and see that

$$\boxed{\mathbf{r}_{n+1} := \mathbf{r}_n - \mathbf{A}\mathbf{r}_n} \quad [= \mathbf{B}\mathbf{r}_n] .$$

Thus,

$$\mathbf{r}_n \in \operatorname{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^n\mathbf{r}_0\} \equiv: \mathcal{K}_{n+1}\,.$$

$\mathcal{K}_n = \mathcal{K}_n(\mathbf{A}, \mathbf{r}_0)$: nth **Krylov (sub)space** generated from $\mathbf{r}_0$ by $\mathbf{A}$.

By induction we easily see that

$$\boxed{\mathbf{x}_{n+1} - \mathbf{x}_0 \in \mathcal{K}_{n+1}} \tag{2}$$

and

$$\boxed{\mathcal{K}_{n+1} = \operatorname{span}\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_n\}}$$

Condition (2) characterizes a **Krylov space method**.

Alternative denominations for Krylov space methods are:

- **semiiterative methods**,
- **polynomial acceleration methods**,
- **polynomial preconditioners**,
- **Krylov space solvers**.

Problem. For given $\mathbf{A}$ and $\mathbf{r}_0$, let $\bar{\nu}$ be the maximum dimension of the Krylov subspaces $\mathcal{K}_n$. Show that

$$\bar{\nu} = \min\ \{n :\ \dim \mathcal{K}_n = \dim \mathcal{K}_{n+1}\}\ .$$

$\bar{\nu}$ is called the **grade of A with respect to** $\mathbf{r}_0$. ∘

Problem. Show that the dimension $\bar{\nu}$ defined in the previous problem satisfies

$$\bar{\nu} = \min\ \left\{n :\ \mathbf{A}^{-1}\mathbf{r}_0 \in \mathcal{K}_n\right\}.$$

Consequently, $\bar{\nu}$ is at most equal to the degree of the minimum polynomial of $\mathbf{A}$.
(Hint: Project $\mathbf{A}$ into the invariant subspace $\mathcal{K}_{\bar{\nu}}$. Then apply the Cayley-Hamilton theorem, which says that the minimum polynomial of a matrix is the monic polynomial of lowest degree that annihilates that matrix.) ∘

We look for

- fast convergence
- only one or two matrix-vector products (MV) per step
- small memory requirement ($\leadsto$ short recurrences for $\mathbf{x}_n$ and $\mathbf{r}_n$)

If the residuals are linearly independent, then — in theory — a Krylov space solver must converge in at most N steps. ($\leadsto$ **finite termination property**).

We can make the residuals orthogonal by combining

$$\boxed{\text{Krylov space generation + Gram-Schmidt orthogonalization}}$$

For this, in the nth step, we orthogonalize $\mathbf{Ar}_n$ with respect to $\mathcal{K}_{n+1}$, *i.e.*, with respect to $\mathbf{r}_0$, …, $\mathbf{r}_n$.

Starting from $\mathbf{r}_0$ we let, for $n = 0, 1, 2, \ldots$,

$$\boxed{\mathbf{r}_{n+1} := (\mathbf{Ar}_n - \mathbf{r}_n\, \eta_{n,n} - \mathbf{r}_{n-1}\, \eta_{n-1,n} - \cdots - \mathbf{r}_0\, \eta_{0,n}) \,/\eta_{n+1,n}}$$

with

$$\eta_{k,n} :\equiv \frac{\langle \mathbf{r}_k, \mathbf{Ar}_n \rangle}{\langle \mathbf{r}_k, \mathbf{r}_k \rangle} \quad (k = 0, \ldots, n),$$

$\eta_{n+1,n}$ determines the length of $\mathbf{r}_{n+1}$. If we choose

$$\eta_{n+1,n} :\equiv -\sum_{k=0}^{n} \eta_{k,n}\,,$$

then it is easy to verify that the recursion

$$\boxed{\mathbf{x}_{n+1} := -\,(\mathbf{r}_n + \mathbf{x}_n\, \eta_{n,n} + \mathbf{x}_{n-1}\, \eta_{n-1,n} + \cdots + \mathbf{x}_0\, \eta_{0,n}) \,/\eta_{n+1,n}}$$

is consistent with the residual relationship

$$\mathbf{r}_n = \mathbf{b} - \mathbf{Ax}_n\,.$$

These two recursions define the ORTHORES version of the **generalized conjugate gradient (GCG) method**.

GCG is applicable to nonsymmetric matrices, but it may break down, because some $\eta_{n+1,n}$ may be 0.

Moreover, in general, it has long recursions.

Now assume that $\mathbf{A}$ is Hemitian positive definite (Hpd).

Then, if $k < n-1$,

$$\begin{aligned} \langle \mathbf{r}_k, \mathbf{A}\mathbf{r}_n \rangle &= \langle \mathbf{A}\mathbf{r}_k, \mathbf{r}_n \rangle \\ &= \langle \mathbf{r}_{k+1}\eta_{k+1,k} + \mathbf{r}_k\eta_{k,k} + \cdots + \mathbf{r}_0\eta_{0,k}, \mathbf{r}_n \rangle \\ &= 0 \end{aligned}$$

and, thus $\eta_{k,n} = 0 \quad \text{if } k < n-1\,.$

This leads to the three-term version of the **conjugate gradient (CG) method**:

ALGORITHM 1 (ORES FORM OF THE CG METHOD).
For solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ *with an Hpd matrix* $\mathbf{A}$, *choose an initial approximation* $\mathbf{x}_0$, *and let* $\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\beta_0 := 0$, *and* $\delta_0 := ||\mathbf{r}_0||^2$. *Then, for* $n = 0, 1, \ldots$, *compute*

$$\begin{aligned} &\text{(3a)} & \alpha_n &:= ||\mathbf{r}_n||^2_{\mathbf{A}} \,/\, \delta_n\,, \\ &\text{(3b)} & \beta_{n-1} &:= \gamma_{n-1}\delta_n/\delta_{n-1} \quad (\textit{if } n > 0), \\ &\text{(3c)} & \gamma_n &:= -\alpha_n - \beta_{n-1}\,, \\ &\text{(3d)} & \mathbf{r}_{n+1} &:= (\mathbf{A}\mathbf{r}_n - \mathbf{r}_n\alpha_n - \mathbf{r}_{n-1}\beta_{n-1})/\gamma_n\,, \\ &\text{(3e)} & \mathbf{x}_{n+1} &:= -(\mathbf{r}_n + \mathbf{x}_n\alpha_n + \mathbf{x}_{n-1}\beta_{n-1})/\gamma_n\,. \\ &\text{(3f)} & \delta_{n+1} &:= ||\mathbf{r}_{n+1}||^2, \end{aligned}$$

If $||\mathbf{r}_{n+1}|| \leq \mathrm{tol}$, *the algorithm terminates and* $\mathbf{x}_{n+1}$ *is the solution.*

Minimizing the error in the $\mathbf{A}$-norm (assuming $\mathbf{A}$ is Hpd) or the residual in the $\mathbf{A}^{-1}$-norm means minimizing a quadratic function:

$$||\mathbf{x}_n - \mathbf{x}_\star||^2_{\mathbf{A}} = ||\mathbf{A}\mathbf{x}_n - \mathbf{b}||^2_{\mathbf{A}^{-1}} = ||\mathbf{r}_n||^2_{\mathbf{A}^{-1}} = 2\,\Psi(\mathbf{x}_n),$$

where

$$\Psi(\mathbf{x}) := \tfrac{1}{2}\mathbf{x}^\star\mathbf{A}\mathbf{x} - \Re(\mathbf{b}^\star\mathbf{x}) + \tfrac{1}{2}\mathbf{b}^\star\mathbf{A}^{-1}\mathbf{b}.$$

Ψ is a convex function when $\mathbf{A}$ is Hpd, and its gradient is then

$$\nabla\Psi(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b} = -\mathbf{r}\,,$$

where $\mathbf{r}$ is the residual corresponding to $\mathbf{x}$. Hence,

$$\mathbf{x} \text{ minimizer of } \Psi \iff \nabla\Psi(\mathbf{x}) = \mathbf{o} \iff \mathbf{A}\mathbf{x} = \mathbf{b}.$$

The interpretation of the solution of a linear system as the minimizer of quadratic minimization problem suggests to try to find this minimizer by the **method of steepest descent**.

In the nth step, we proceeed from $\mathbf{x}_n$ in the opposite direction of the gradient, that is in the direction of $\mathbf{r}_n$.

The steplength is then determined by line search, *i.e.*, by a one-dimensional minimization of the residual on the line

$$\omega_n \mapsto \mathbf{x}_n + \mathbf{r}_n \omega_n \,. \tag{4}$$

The minimum is at

$$\omega_n :\equiv \frac{\langle \mathbf{r}_n, \mathbf{r}_n \rangle}{\langle \mathbf{r}_n, \mathbf{A}\mathbf{r}_n \rangle} \,. \tag{5}$$

The steepest descent method differs from the fixed point iteration only in the (optimal) choice of the step length.

It may converge slowly even if N is only 2.

Proof of (5). (4) implies $\mathbf{r}_{n+1} = \mathbf{r}_n - \mathbf{A}\mathbf{r}_n\omega_n$, so that

$$\begin{aligned} \|\mathbf{r}_{n+1}\|^2_{\mathbf{A}^{-1}} &= \|\mathbf{r}_n\|^2_{\mathbf{A}^{-1}} - 2\langle \mathbf{r}_n, \mathbf{A}\mathbf{r}_n \rangle_{\mathbf{A}^{-1}} \, \omega_n + \|\mathbf{A}\mathbf{r}_n\|^2_{\mathbf{A}^{-1}} \, \omega_n^2 \\ &= \|\mathbf{r}_n\|^2_{\mathbf{A}^{-1}} - 2\langle \mathbf{r}_n, \mathbf{r}_n \rangle \, \omega_n + \langle \mathbf{r}_n, \mathbf{A}\mathbf{r}_n \rangle \, \omega_n^2 \,. \end{aligned}$$

As a function of ω_n, this expression has a vanishing derivative (and is thus minimal) when (5) holds. □

The case $N = 2$ suggest to choose **conjugate**, *i.e.*, **A-orthogonal** direction vectors $\mathbf{v}_n$:

$$\boxed{\mathbf{v}_n^\star \mathbf{A} \mathbf{v}_k = 0 \,, \quad k = 0, \ldots, n-1.} \tag{6}$$

This can be seen to imply that the minimization problem for a method based on

$$\boxed{\mathbf{x}_{n+1} := \mathbf{x}_n + \mathbf{v}_n \omega_n}$$

decouples:

$$\Psi(\mathbf{x}_{n+1}) = \Psi(\mathbf{x}_n) - \Re(\omega_n \mathbf{v}_n^\star \mathbf{r}_0) + \tfrac{1}{2}|\omega_n|^2 \mathbf{v}_n^\star \mathbf{A} \mathbf{v}_n \,.$$

The problem of finding the minimum with respect to the search directions $\mathbf{v}_0, \ldots, \mathbf{v}_n$ decouples into the one of minimizing with respect to $\mathbf{v}_0, \ldots, \mathbf{v}_{n-1}$ (which by induction we may assume to yield $\mathbf{x}_n$) and the one-dimensional minimization (line search) with respect $\mathbf{v}_n$.

Conjugate direction methods in general, as well as the special case of the conjugate gradient method treated here are due to Hestenes and Stiefel (1952).

The conjugate gradient (CG) method is the the conjugate direction method that chooses the direction vectors $\mathbf{v}_0, \ldots, \mathbf{v}_n$ such that they span the Krylov space $\mathcal{K}_{n+1}(\mathbf{A}, \mathbf{r}_0)$:

$$\mathrm{span}(\mathbf{v}_0, \ldots, \mathbf{v}_n) = \mathcal{K}_{n+1}(\mathbf{A}, \mathbf{r}_0) \quad (\forall n).$$

This leads to the (standard) **Hestenes-Stiefel** or **coupled two-term** formulation of the conjugate gradient (CG) method:

Algorithm 2 (OMin form of the CG method). *For solving* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *choose an initial approximation* $\mathbf{x}_0$, *and let* $\mathbf{v}_0 := \mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\delta_0 := ||\mathbf{r}_0||^2$, *and* $\delta_0' := ||\mathbf{v}_0||^2_{\mathbf{A}}$.
Then, for $n = 0, 1, 2, \ldots$, *compute*

$$\omega_n := \delta_n / \delta_n' , \tag{7a}$$
$$\mathbf{r}_{n+1} := \mathbf{r}_n - \mathbf{A}\mathbf{v}_n \omega_n , \tag{7b}$$
$$\mathbf{x}_{n+1} := \mathbf{x}_n + \mathbf{v}_n \omega_n , \tag{7c}$$
$$\delta_{n+1} := ||\mathbf{r}_{n+1}||^2, \tag{7d}$$
$$\psi_n := -\delta_{n+1} / \delta_n , \tag{7e}$$
$$\mathbf{v}_{n+1} := \mathbf{r}_{n+1} - \mathbf{v}_n \psi_n , \tag{7f}$$
$$\delta_{n+1}' := ||\mathbf{v}_{n+1}||^2_{\mathbf{A}}. \tag{7g}$$

If $||\mathbf{r}_{n+1}|| \le \mathrm{tol}$ *the process terminates and* $\mathbf{x}_{n+1}$ *is the solution.*

Properties of the CG method:

$$\mathbf{r}_n, \mathbf{v}_n \in \mathcal{K}_{n+1} , \quad \mathbf{x}_n - \mathbf{x}_0 \in \mathcal{K}_n , \tag{8}$$

where

$$\mathcal{K}_n := \mathcal{K}_n(\mathbf{A}, \mathbf{r}_0) := \mathrm{span}\left(\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{n-1}\mathbf{r}_0\right)$$

is the nth **Krylov (sub)space** generated by $\mathbf{A}$ from $\mathbf{r}_0$.

$$\dim \mathcal{K}_n = \begin{cases} n & \text{if } n \le \bar{\nu} =: \bar{\nu}(\mathbf{r}_0, \mathbf{A}) , \\ \bar{\nu} & \text{if } n \ge \bar{\nu} =: \bar{\nu}(\mathbf{r}_0, \mathbf{A}) . \end{cases}$$

$\bar{\nu}(\mathbf{r}_0, \mathbf{A})$ is called **grade of $\mathbf{r}_0$ with respect to $\mathbf{A}$**.

$\{\mathbf{r}_n\}_{n=0}^{\bar{\nu}-1}$ is an **orthogonal basis** of $\mathcal{K}_{\bar{\nu}}$:

$$\boxed{\langle \mathbf{r}_m, \mathbf{r}_n \rangle = \begin{cases} 0 & \text{if } m \ne n, \\ \delta_n \ne 0 & \text{if } m = n. \end{cases}} \tag{9}$$

$\{\mathbf{v}_n\}_{n=0}^{\bar{\nu}-1}$ is a **conjugate basis** of $\mathcal{K}_{\bar{\nu}}$:

$$\boxed{\langle \mathbf{v}_m, \mathbf{A}\mathbf{v}_n \rangle = \begin{cases} 0 & \text{if } m \ne n, \\ \delta_n' \ne 0 & \text{if } m = n. \end{cases}} \tag{10}$$

The nth residual $\mathbf{r}_n$ is shortest possible in the $\mathbf{A}^{-1}$-norm:

$$\boxed{||\mathbf{r}_n||_{\mathbf{A}^{-1}} = \min_{\mathbf{r}\in\mathcal{K}_{n+1}} ||\mathbf{r}||_{\mathbf{A}^{-1}}}$$

The nth error $\mathbf{x}_n - \mathbf{x}_\star$ is shortest possible in the $\mathbf{A}$-norm:

$$\boxed{||\mathbf{x}_n - \mathbf{x}_\star||_{\mathbf{A}} = \min_{\mathbf{x}\in\mathbf{x}_0+\mathcal{K}_n} ||\mathbf{x} - \mathbf{x}_\star||_{\mathbf{A}}}$$

Associated **Galerkin condition**:

$$\boxed{\mathcal{K}_n \perp \mathbf{r}_n \in \mathcal{K}_{n+1}} \tag{11}$$

Shorthand notation: matrix relations
Let

$$\mathbf{R}_m := [\ \mathbf{r}_0 \quad \mathbf{r}_1 \quad \cdots \quad \mathbf{r}_{m-1}\ ], \qquad \mathbf{V}_m := [\ \mathbf{v}_0 \quad \mathbf{v}_1 \quad \cdots \quad \mathbf{v}_{m-1}\ ],$$

and

$$\mathbf{U}_m := \begin{bmatrix} 1 & \psi_0 & & & \\ & 1 & \psi_1 & & \\ & & 1 & \ddots & \\ & & & \ddots & \psi_{m-2} \\ & & & & 1 \end{bmatrix},$$

$$\underline{\mathbf{L}}_m^\circ := \begin{bmatrix} \omega_0^{-1} & & & & \\ -\omega_0^{-1} & \omega_1^{-1} & & & \\ & -\omega_1^{-1} & \omega_2^{-1} & & \\ & & \ddots & \ddots & \\ & & & -\omega_{m-2}^{-1} & \omega_{m-1}^{-1} \\ & & & & -\omega_{m-1}^{-1} \end{bmatrix},$$

then the coupled two-term CG recursions can be gathered into

$$\boxed{\mathbf{R}_m = \mathbf{V}_m\mathbf{U}_m\,, \quad \mathbf{A}\mathbf{V}_m = \mathbf{R}_{m+1}\underline{\mathbf{L}}_m^\circ \ \ (m \le \bar{\nu})} \tag{12}$$

$\underline{\mathbf{L}}_m^\circ$ is an 'extended' lower bidiagonal matrix.

Eliminating the direction vectors contained in $\mathbf{V}_m$ yields

$$\text{(13)} \qquad \boxed{\mathbf{A}\mathbf{R}_m = \mathbf{R}_{m+1}\underline{\mathbf{T}}_m^\circ \ \ (m \le \bar{\nu})}$$

where

$$\text{(14)} \qquad \boxed{\underline{\mathbf{T}}_m^\circ :\equiv \underline{\mathbf{L}}_m^\circ \mathbf{U}_m}$$

is an 'extended' tridiagonal matrix:

$$\underline{\mathbf{T}}_m^\circ :\equiv \left[\begin{array}{c} \mathbf{T}_m^\circ \\ \hline \gamma_{m-1}\mathbf{l}_m^\top \end{array}\right] = \begin{bmatrix} \alpha_0 & \beta_0 & & & \\ \gamma_0 & \alpha_1 & \beta_1 & & \\ & \gamma_1 & \alpha_2 & \ddots & \\ & & \ddots & \ddots & \beta_{m-2} \\ & & & \gamma_{m-2} & \alpha_{m-1} \\ & & & & \gamma_{m-1} \end{bmatrix}.$$

(13) is a matrix representation of the three-term CG recursions.

$\underline{\mathbf{T}}_m^\circ$ has column sums 0.

Assume

$$\mathbf{A}\mathbf{Y}_n = \mathbf{Y}_{n+1}\underline{\mathbf{T}}_n$$

with

$$\mathbf{Y}_n := [\ \mathbf{y}_0 \ \ \mathbf{y}_1 \ \ \cdots \ \ \mathbf{y}_{n-1}\ ], \qquad \mathbf{Y}_n^\star\mathbf{Y}_n = \mathbf{I}_n$$

and

$$\mathbf{x}_n - \mathbf{x}_0 \in \mathcal{K}_n = \operatorname{span}\{\mathbf{y}_0, \dots, \mathbf{y}_{n-1}\}, \qquad \mathbf{r}_0 = \mathbf{y}_0\rho_0\,,.$$

We can write, with some **coordinate vector $\mathbf{k}_n$**,

$$\text{(15)} \qquad \mathbf{x}_n = \mathbf{x}_0 + \mathbf{Y}_n\mathbf{k}_n\,, \quad \mathbf{r}_n = \mathbf{r}_0 - \mathbf{A}\mathbf{Y}_n\mathbf{k}_n\,.$$

When $\mathbf{A}$ is Hpd, minimization of the error

$$\text{(16)} \quad ||\underbrace{\mathbf{x}_n - \mathbf{x}_\star}_{\text{error}}||_{\mathbf{A}}^2 = ||\underbrace{\mathbf{x}_0 - \mathbf{x}_\star}_{\text{initial error}} + \mathbf{Y}_{n-1}\mathbf{k}_n||_{\mathbf{A}}^2 =: 2\,\Phi(\mathbf{k}_n),$$

yields in view of

$$\text{(17)} \qquad \nabla\Phi(\mathbf{k}_n) = \mathbf{Y}_n^\star\mathbf{A}(\mathbf{x}_n - \mathbf{x}_\star) = \mathbf{Y}_n^\star\mathbf{r}_n$$

for CG the **Galerkin condition**

$$\text{(18)} \qquad \boxed{\mathbf{r}_n^\star\mathbf{Y}_n = \mathbf{o}^\star} \quad \text{or} \quad \boxed{\mathbf{r}_n \perp \mathcal{K}_n}$$

which implies that the residuals are orthogonal.

Hence, **minimization of the error in the A-norm requires that the residuals be orthogonal**.

Likewise, minimization of the residual

$$||\mathbf{r}_n||^2 = ||\mathbf{r}_0 - \mathbf{A}\mathbf{Y}_n\mathbf{k}_n||^2 = 2\,\widehat{\Phi}(\mathbf{k}_n), \tag{19}$$

yields in view of

$$\nabla\widehat{\Phi}(\mathbf{k}_n) = \mathbf{Y}_n^\star\mathbf{A}^\star\mathbf{r}_n \tag{20}$$

the **Galerkin condition**

$$\boxed{\mathbf{r}_n^\star\mathbf{A}\mathbf{Y}_n = \mathbf{o}^\star} \quad \text{or} \quad \boxed{\mathbf{r}_n \perp \mathbf{A}\mathcal{K}_n} \tag{21}$$

which implies that the residuals are **A-orthogonal** or so-called **conjugate**.

Hence, **minimization of the residual in the 2-norm requires that the residuals be conjugate**.

This argument applies also to non-Hermitian $\mathbf{A}$.

It is the basis of the **conjugate residual (CR) method** or **minimal residual method** for Hermitian $\mathbf{A}$ (a special algorithm is called MINRES),

and of the **generalized conjugate residual (GCR) method** or **generalized minimal residual method** for non-Hermitian $\mathbf{A}$ (a special algorithm is called GMRES).

Summary of CG **features:**

- CG = **Gram-Schmidt** + **Krylov space extension**
- **short recurrences**: only the orthogonality of $\mathbf{r}_{n+1}$ w.r.t. $\mathbf{r}_n$ and $\mathbf{r}_{n-1}$ is enforced; the orthogonality w.r.t. $\mathbf{r}_0, \ldots, \mathbf{r}_{n-2}$ is inherited
- the residuals [errors] are minimal in the $\mathbf{A}^{-1}$-norm [$\mathbf{A}$-norm]

Conjugate residual (CR) or **minimal residual** (MINRES) **method**:

- mathematically analogous to CG, except that inner product $\langle ., . \rangle$ is replaced by $\langle ., . \rangle_{\mathbf{A}}$
- still short recurrences
- the residuals are minimal in the Euclidean norm
- some versions work for indefinite Hermitian matrices; in particular, MINRES due to (Paige and Saunders, 1975).

Generalized conjugate gradient method for non-Hermitian systems:

- many generalizations: *e.g.*,
  - Generalized conjugate gradient (GCR) method,
  - Arnoldi = full orthogonalization method (FOM),
  - generalized minimal residual (GMRES) method,
  - biconjugate gradient (BICG) method,
  - other Lanczos-type methods.
- minimization of the residual in some norm is incompatible with short recurrences (Faber and Manteuffel, 1984; Faber and Manteuffel, 1987; Voevodin, 1983).

The **Lanczos-type methods** including BICG feature short recurrences, but no minimization property.

For an overview of Lanczos-type methods see (Gutknecht, 1997).

References

V. Faber and T. Manteuffel (1984), 'Necessary and sufficient conditions for the existence of a conjugate gradient method', **SIAM J. Numer. Anal. 21**, 352–362. .

V. Faber and T. Manteuffel (1987), 'Orthogonal error methods', **SIAM J. Numer. Anal. 24**, 170–187.

M. H. Gutknecht (1997), 'Lanczos-type solvers for nonsymmetric linear systems of equations', **Acta Numerica 6**, 271–397.

M. R. Hestenes and E. Stiefel (1952), 'Methods of conjugate gradients for solving linear systems', **J. Res. Nat. Bureau Standards 49**, 409–435.

C. C. Paige and M. A. Saunders (1975), 'Solution of sparse indefinite systems of linear equations', **SIAM J. Numer. Anal. 12**, 617–629.

V. V. Voevodin (1983), 'The question of non-self-adjoint extension of the conjugate gradients method is closed', **U.S.S.R. Comput. Maths. Math. Phys. 23**(2), 143–144.