CG and GMRES iterations

From Southwell to Conjugate Gradient iterations

Let $\mathbf{e}_k$ be the error at step k of an iterative method in approximating the solution $\mathbf{x}$ of a linear system $A\mathbf{x} = \mathbf{b}$, i.e. $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k$. Choose a vector $\mathbf{d}_k \neq \mathbf{0}$ and set $\mathbf{e}_{k+1} = \mathbf{e}_k - \omega \mathbf{d}_k$. Let H be a positive definite matrix and consider the inner product $(\mathbf{u}, \mathbf{v})_H = \mathbf{u}^T H \mathbf{v}$. Then the value of ω for which $\|\mathbf{e}_{k+1}\|_H$ is minimum is

$$\omega = \omega_k = \frac{(\mathbf{e}_k, \mathbf{d}_k)_H}{\|\mathbf{d}_k\|_H^2}.$$

($\|\mathbf{e}_{k+1}\|_H^2 = \|\mathbf{e}_k\|^2 - 2\omega(\mathbf{e}_k, \mathbf{d}_k)_H + \omega^2 \|\mathbf{d}_k\|_H^2$). So, we have the iterative scheme

$$\mathbf{x}_0 \in \mathbb{R}^n, \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \frac{(\mathbf{e}_k, \mathbf{d}_k)_H}{\|\mathbf{d}_k\|_H^2} \mathbf{d}_k, \quad k = 0, 1, 2, \ldots . \tag{it}$$

Note that each of the three conditions:

1) $\|\mathbf{e}_{k+1}\|_H$ minimum,
2) $(\mathbf{e}_{k+1}, \mathbf{d}_k)_H = 0$,
3) $F(\mathbf{x}_k + \omega \mathbf{d}_k)$ minimum, $F(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T H \mathbf{y} - \mathbf{y}^T H A^{-1} \mathbf{b}$

yields the value $\omega = \omega_k$, i.e. such conditions are equivalent. (Note that, since $F(\mathbf{y} + \mathbf{z}) = F(\mathbf{y}) + \mathbf{z}^T H(\mathbf{y} - A^{-1}\mathbf{b}) + \frac{1}{2}\mathbf{z}^T H \mathbf{z}$, the vector $A^{-1}\mathbf{b}$ is the global minimum for F, and the contours of F are neighborhoods of $A^{-1}\mathbf{b}$ in the metric induced by the norm $\|\cdot\|_H$). Moreover, for $\omega = \omega_k$ we have

4) $\|\mathbf{e}_{k+1}\|_H^2 = \|\mathbf{e}_k\|_H^2 - \|\omega_k \mathbf{d}_k\|_H^2$,
5) $\lim_k \|\mathbf{e}_k\|_H = l_{\{\mathbf{d}_k\}, H} \geq 0$,
6) $\lim_{k \to +\infty} \|\omega_k \mathbf{d}_k\|_H = 0$,
7) let $\mathbf{v}_i \neq \mathbf{0}$, $i = 1, \ldots, r$; if $\mathbf{d}_k = \mathbf{v}_{k \bmod r + 1}$ then $\lim_{k \to +\infty} \frac{|(\mathbf{e}_{k+1}, \mathbf{v}_i)_H|}{\|\mathbf{v}_i\|_H} = 0$.

Suitable choice of H and $\{\mathbf{d}_k\}$ make $l_{\{\mathbf{d}_k\}, H} = 0$, i.e. make (it) convergent to $\mathbf{x} = A^{-1}\mathbf{b}$. In the following, $\mathbf{r}_k$ denotes the vector $\mathbf{b} - A\mathbf{x}_k = A\mathbf{e}_k$.

Choice $H = A^T A$

$$\mathbf{x}_0 \in \mathbb{R}^n, \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\mathbf{r}_k^T A \mathbf{d}_k}{\|A\mathbf{d}_k\|_2^2} \mathbf{d}_k, \quad k = 0, 1, 2, \ldots .$$

Choose $\mathbf{d}_k$ equal to the s_kth canonical basis vector ($s_k \in \{1, \ldots, n\}$). For the sake of simplicity, call $A^{(j)}$ the jth column of A, $j = 1, \ldots, n$. Then we obtain the algorithm:

```
compute $\|A^{(j)}\|_2$, $j = 1, \ldots, n$;
$\mathbf{x}_0 \in \mathbb{R}^n$, $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$. For $k = 0, 1, 2 \ldots$
compute $\mathbf{r}_k^T A$
choose $s_k \in \{1, 2, \ldots, n\}$
$(\mathbf{x}_{k+1})_j = (\mathbf{x}_k)_j$, $j \neq s_k$                                  (S, GS)
$(\mathbf{x}_{k+1})_{s_k} = (\mathbf{x}_k)_{s_k} + \frac{(\mathbf{r}_k^T A)_{s_k}}{\|A^{(s_k)}\|_2^2}$
$\mathbf{r}_{k+1} = \mathbf{r}_k - \frac{(\mathbf{r}_k^T A)_{s_k}}{\|A^{(s_k)}\|_2^2} A^{(s_k)}$
```

Note that, both the choices s_k such that

$$\frac{|(\mathbf{r}_k^T A)_{s_k}|}{\|A^{(s_k)}\|_2} \geq \frac{|(\mathbf{r}_k^T A)_j|}{\|A^{(j)}\|_2}, \quad \forall j$$

(Southwell) and $s_k = k \bmod n + 1$ (Gauss-Seidel) yield $\frac{|\mathbf{r}_k^T A^{(j)}|}{\|A^{(j)}\|_2} \to 0$, $\forall j$ (use 6) and 7), respectively), i.e. the residuals $\mathbf{r}_k$ converge to the null vector, since the columns of A are linearly independent.

Thus the Southwell and Gauss-Seidel variants of the above algorithm can solve any linear system $A\mathbf{x} = \mathbf{b}$, by requiring for each step only the computation of $\mathbf{r}_k^T A$.

Choose $\mathbf{d}_k = \mathbf{r}_k$. Then we have the *variable-step* Richardson-Euler method

$$\mathbf{x}_0 \in \mathbb{R}^n, \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\mathbf{r}_k^T A \mathbf{r}_k}{\|A\mathbf{r}_k\|_2^2}\mathbf{r}_k, \quad k = 0, 1, 2, \dots . \tag{RE}$$

whose convergence is assured if the symmetric part of A is positive definite (by 6), in fact, $\frac{\|\mathbf{r}_k\|^2_{A_S}}{\|A\|_2\|\mathbf{r}_k\|_2} \to 0$).

Choose $\mathbf{d}_k = A^T\mathbf{r}_k$. Then we obtain an algorithm always convergent

$$\mathbf{x}_0 \in \mathbb{R}^n, \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\mathbf{r}_k^T A A^T \mathbf{r}_k}{\|AA^T\mathbf{r}_k\|_2^2} A^T\mathbf{r}_k, \quad k = 0, 1, 2, \dots , \tag{G}$$

since the result 6), in this case, implies $\frac{\|A^T\mathbf{r}_k\|_2}{\|A\|_2} \to 0$. The required computations are $A^T\mathbf{r}_k$ and $A(A^T\mathbf{r}_k)$ each step. Note that this method is nothing else that Gradient method (see below Choice $H = A$) applied to the system $A^T A\mathbf{x} = A^T\mathbf{b}$. One can obviously apply CG method to the same system and obtain an algorithm whose rate of convergence can be much faster than (G) convergence rate in many cases (see below the theory on CG method).

Choice $H = I$

For $H = I$ the iterative scheme (it) becomes:

$$\mathbf{x}_0 \in \mathbb{R}^n, \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \frac{(\mathbf{x} - \mathbf{x}_k)^T \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{d}_k}\mathbf{d}_k, \quad k = 0, 1, 2, \dots .$$

Choose $\mathbf{d}_k = A^T\mathbf{r}_k$. Then, by 6), we have $\frac{\|\mathbf{r}_k\|_2}{\|A^T\|_2} \to 0$, that is, the x_k converge to $\mathbf{x}$ under the only assumption $\det(A) \neq 0$. Here is the corresponding algorithm:

$$\begin{array}{l} \mathbf{x}_0 \in \mathbb{R}^n, \quad \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0. \\ \text{For } k = 0, 1, 2, \dots \\ \mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\mathbf{r}_k^T \mathbf{r}_k}{\|A^T\mathbf{r}_k\|_2^2} A^T\mathbf{r}_k, \\ \mathbf{r}_{k+1} = \mathbf{r}_k - \frac{\mathbf{r}_k^T \mathbf{r}_k}{\|A^T\mathbf{r}_k\|_2^2} A(A^T\mathbf{r}_k), \end{array} \tag{M}$$

each step of which requires the computations $A^T\mathbf{r}_k$ and $A(A^T\mathbf{r}_k)$.

In general, all the above methods (it) may have a low rate of convergence in solving $A\mathbf{x} = \mathbf{b}$. For example, for the iterations generated by (G) we know that

$\|\mathbf{x}-\mathbf{x}_k\|_{A^TA} \leq ((\mu_2(A^TA)-1)/(\mu_2(A^TA)+1))^k\|\mathbf{x}-\mathbf{x}_0\|_{A^TA}$, and it is not difficult to prove that $\mu_2(A^TA)=\mu_2(A)^2$. However, the convergence rate of (it) could be improved by applying it to $P^{-1}A\mathbf{x}=P^{-1}\mathbf{b}$ for a suitable choice of P $n\times n$ non singular. One could set, for instance, $P^{-1}=C_{A^TA}^{-1}A^T$ where C_{A^TA} is some approximation of A^TA (of course such approximation should be defined without computing the entries of A^TA; may be one could set $C_{A^TA}=C_A^TC_A$ where C_A is some approximation of A). This procedure could improve the rate of convergence at least of (G) and (GC).

Question: can a suitable choice of P improve the rate of convergence of the methods (S),(GS),(M). If yes, then (S) and (GS) would be competitive in solving generic large linear systems, because of their low complexity per step.

Choice $H=A$ (allowed if A is positive definite).

It is simple to prove that also for $H=A$ the choice $\mathbf{d}_k$ equal to the s_kth canonical basis vector ($s_k\in\{1,\dots,n\}$) with s_k such that

$$\frac{|(\mathbf{r}_k)_{s_k}|}{|a_{s_ks_k}|}\geq\frac{|(\mathbf{r}_k)_j|}{|a_{jj}|},\quad\forall j$$

(Southwell) and $s_k=k\bmod n+1$ (Gauss-Seidel) yields vectors $\mathbf{x}_k$ convergent to $\mathbf{x}=A^{-1}\mathbf{b}$. Note that in the second case ($s_k=k\bmod n+1$), n iterations are equivalent to one iteration of the well known *stationary* Gauss-Seidel method; it follows that the latter method converges when applied to solve positive definite linear systems (recall that the other well known stationary method, Jacobi, has not such feature).

G method

Assume A, in the system $A\mathbf{x}=\mathbf{b}$ we have to solve, positive definite. Then the choices $H=A$ and $\mathbf{d}_k=\mathbf{r}_k=\mathbf{b}-A\mathbf{x}_k$ yield $l_{\{\mathbf{d}_k\},H}=0$ (use 6)). The method so obtained is called steepest descent (or Gradient) since $\mathbf{d}_k=\mathbf{r}_k=-\nabla F(\mathbf{x}_k)$ and F decreases along $-\nabla F(\mathbf{x}_k)$ more rapidly than in any other direction (in a neighborhood of $\mathbf{x}_k$). However, in general the contours of F are far from being spheres, so the steepest descent direction (which is orthogonal to such contours) is far from pointing to $A^{-1}\mathbf{b}$. In particular, from 4) and the Kantorovich inequality,

$$1\leq\frac{\mathbf{z}^TA\mathbf{z}\mathbf{z}^TA^{-1}\mathbf{z}}{(\mathbf{z}^T\mathbf{z})^2}\leq\frac{(\lambda_{\max}+\lambda_{\min})^2}{4\lambda_{\max}\lambda_{\min}}$$

($\lambda_{\max}=\max\lambda(A)$, $\lambda_{\min}=\min\lambda(A)$), we have the following result

$$\frac{\|\mathbf{x}-\mathbf{x}_k\|_A}{\|\mathbf{x}-\mathbf{x}_0\|_A}\leq\left(\frac{\frac{\lambda_{\max}}{\lambda_{\min}}-1}{\frac{\lambda_{\max}}{\lambda_{\min}}+1}\right)^k$$

that states that G can be very slow when $\frac{\lambda_{\max}}{\lambda_{\min}}>>1$.

CG method

Assume A, the coefficient matrix of our system $A\mathbf{x}=\mathbf{b}$, positive definite (recall that A and $\mathbf{b}$ are real). In the general scheme choose $H=A$, $\mathbf{d}_0=\mathbf{r}_0=\mathbf{b}-A\mathbf{x}_0$, $\mathbf{d}_k=\mathbf{r}_k+\beta_{k-1}\mathbf{d}_{k-1}$, $k=1,2,\dots$, ($\mathbf{r}_k=\mathbf{b}-A\mathbf{x}_k$) where β_{k-1} is such that

$$(\mathbf{d}_k,\mathbf{d}_{k-1})_A=0$$

($\mathbf{d}_k$ *conjugate to* $\mathbf{d}_{k-1}$).

Note that for $n = 2$ the choice of β_0 so that $(\mathbf{d}_1, \mathbf{d}_0)_A = 0$ makes the search direction $\mathbf{d}_1 = \mathbf{r}_1 + \beta_0 \mathbf{d}_0$ pointing to the center of the contours of F, and thus makes $\mathbf{x}_2 = A^{-1}\mathbf{b}$, i.e. we have convergence in two steps.

Here below is the CG algorithm when n is generic:

```
$\mathbf{x}_0 \in \mathbb{R}^n,\ \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0,\ \mathbf{d}_0 = \mathbf{r}_0.$
For $k = 0, 1, \dots, \{$
$\tau_k = \frac{\mathbf{d}_k^T \mathbf{r}_k}{\mathbf{d}_k^T A \mathbf{d}_k}$
$\mathbf{x}_{k+1} = \mathbf{x}_k + \tau_k \mathbf{d}_k$
$\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1} = \mathbf{r}_k - \tau_k A \mathbf{d}_k$
$\beta_k = -\frac{\mathbf{r}_{k+1}^T A \mathbf{d}_k}{\mathbf{d}_k^T A \mathbf{d}_k}$
$\mathbf{d}_{k+1} = \mathbf{r}_{k+1} + \beta_k \mathbf{d}_k$
$\}$
```

known as Conjugate Gradient (CG) algorithm.

We stop here the discussion about the general iterations (it), in order to investigate CG in detail.

First note that

$$0 = (\mathbf{x} - \mathbf{x}_{k+1})^T H \mathbf{d}_k = (\mathbf{x} - \mathbf{x}_{k+1})^T A \mathbf{d}_k = \mathbf{r}_{k+1}^T \mathbf{d}_k,\ \mathbf{r}_{k+1}^T \mathbf{d}_{k+1} = \|\mathbf{r}_{k+1}\|_2^2.$$

As a consequence, if at step s we have $\mathbf{r}_s = \mathbf{b} - A\mathbf{x}_s \neq \mathbf{0}$, then $\mathbf{d}_s \neq \mathbf{0}$, τ_s is well defined and not zero ...: the algorithm works.

If $\mathbf{r}_0, \dots, \mathbf{r}_{m-1}$ are non null and $\mathbf{r}_m = \mathbf{0}$, then $\beta_{m-1} = 0$, $\mathbf{d}_m = \mathbf{0}$, τ_m cannot be defined, but it doen't matter since $\mathbf{x}_m = A^{-1}\mathbf{b}$. This hypothesis is effectively verified, in fact there exists $m \leq n =$ the order of A, such that $\mathbf{r}_m = \mathbf{0}$ (see below).

Alternative expressions for τ_k and β_k hold:

$$\tau_k = \frac{\mathbf{d}_k^T \mathbf{r}_k}{\mathbf{d}_k^T A \mathbf{d}_k} = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{d}_k^T A \mathbf{d}_k}$$

($\mathbf{d}_k = \mathbf{r}_k + \beta_{k-1}\mathbf{d}_{k-1}$, $\mathbf{r}_k^T \mathbf{d}_{k-1} = 0$),

$$\begin{array}{rcl} \beta_k & = & -\frac{\mathbf{r}_{k+1}^T A \mathbf{d}_k}{\mathbf{d}_k^T A \mathbf{d}_k} = -\frac{\mathbf{r}_{k+1}^T \tau_k^{-1}(\mathbf{r}_k - \mathbf{r}_{k+1})}{\mathbf{d}_k^T \tau_k^{-1}(\mathbf{r}_k - \mathbf{r}_{k+1})} \\ & = & \frac{-\mathbf{r}_{k+1}^T \mathbf{r}_k + \mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{d}_k^T \mathbf{r}_k} = \frac{\mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{r}_k}. \end{array}$$

The latter identity uses the result

$$\mathbf{r}_{k+1}^T \mathbf{r}_k = 0$$

(residual at step $k+1$ is orthogonal to residual at step k, exactly as in the Gradient method) which is not obvious:

$$\begin{array}{rcl} \mathbf{r}_{k+1}^T \mathbf{r}_k & = & \mathbf{r}_{k+1}^T(\mathbf{d}_k - \beta_{k-1}\mathbf{d}_{k-1}) = -\beta_{k-1}\mathbf{r}_{k+1}^T \mathbf{d}_{k-1} \\ & = & -\beta_{k-1}(\mathbf{r}_k - \tau_k A \mathbf{d}_k)^T \mathbf{d}_{k-1} = \beta_{k-1}\tau_k \mathbf{d}_k^T A \mathbf{d}_{k-1} = 0. \end{array}$$

First main result: If $\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_p$ are non null, then

$$\mathbf{d}_l^T A \mathbf{d}_j = 0, \quad \mathbf{r}_l^T \mathbf{r}_j = 0, \quad 0 \leq j < l \leq p.$$

That is, each new residual (search direction) is orthogonal (conjugate) to all previous residuals (search directions). As a consequence, the residual $\mathbf{r}_m$ must be null for some $m \leq n$, or, equivalently, CG finds the solution of $A\mathbf{x} = \mathbf{b}$ in at most n steps.

Proof. ...

A useful representation of the residuals. If $\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{k-1}$ are non null, then there exist polynomials $s_k(\lambda), q_k(\lambda)$ such that

$$\mathbf{r}_k = s_k(A)\mathbf{r}_0, \quad \mathbf{d}_k = q_k(A)\mathbf{r}_0,$$
$$s_k(\lambda) = (-1)^k \tau_0\tau_1 \cdots \tau_{k-1}\lambda^k + \ldots + 1, \quad \tau_0\tau_1 \cdots \tau_{k-1} \neq 0.$$

Proof (by induction). The equality $\mathbf{r}_0 = s_0(A)\mathbf{r}_0$ holds if $s_0(\lambda) = 1$; $\mathbf{d}_0 = q_0(A)\mathbf{r}_0$ holds if $q_0(\lambda) = 1$. Moreover,

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \tau_k A\mathbf{d}_k = s_k(A)\mathbf{r}_0 - \tau_k A q_k(A)\mathbf{r}_0 = s_{k+1}(A)\mathbf{r}_0$$

if $s_{k+1}(\lambda) = s_k(\lambda) - \tau_k \lambda q_k(\lambda)$, and

$$\mathbf{d}_{k+1} = \mathbf{r}_{k+1} + \beta_k \mathbf{d}_k = s_{k+1}(A)\mathbf{r}_0 + \beta_k q_k(A)\mathbf{r}_0 = q_{k+1}(A)\mathbf{r}_0$$

if $q_{k+1}(\lambda) = s_{k+1}(\lambda) + \beta_k q_k(\lambda)$. Finally, since

$$s_{k+1}(\lambda) = s_k(\lambda) - \tau_k \lambda(s_k(\lambda) + \beta_{k-1}q_{k-1}(\lambda)),$$

the coefficient of λ^{k+1} in $s_{k+1}(\lambda)$ is $-\tau_k$ times the coefficient of λ^k in $s_k(\lambda)$. Thus, by the inductive assumption, it must be $(-1)^{k+1}\tau_0\tau_1 \cdots \tau_{k-1}\tau_k$. Also, the coefficient of λ^0 in $s_{k+1}(\lambda)$ is equal to the coefficient of λ^0 in $s_k(\lambda)$, which is 1 by the inductive assumption.

Second main result: $\mathbf{r}_k = \mathbf{0}$ for some $k \leq \#\{\text{distinct eigenvalues of } A\}$.

Proof. Let $\mu_1, \mu_2, \ldots, \mu_m$ be the distinct eigenvalues of A ($m \leq n =$ order of A). Assume that CG requires more than m steps to converge. So, the vectors $\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_m$ are non null, and, by the First main result, orthogonal ($\Rightarrow$ linearly independent). Let V be an orthonormal matrix whose columns are eigenvectors of A, thus $V^T = V^{-1}$ and $AV = VD$ for D diagonal with the eigenvalues of A as diagonal entries. Observe that there is a degree-m polynomial which is null in A,

$$\prod_{j=1}^{m}(A - \mu_j I) = \prod_{j=1}^{m}(VDV^T - \mu_j I) = \prod_{j=1}^{m} V(D - \mu_j I)V^T = V \prod_{j=1}^{m}(D - \mu_j I)V^T = 0.$$

As a consequence the matrices $A^0 = I$, A, ..., A^m are linearly dependent. But this implies that the dimension of the space

$$\mathbb{K}_{m+1}(\mathbf{r}_0) = \text{Span}\,\{\mathbf{r}_0, A\mathbf{r}_0, \ldots, A^m\mathbf{r}_0\} = \text{Span}\,\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_m\}$$

is smaller than $m + 1$, which is absurd. It follows that one of the vectors $\mathbf{r}_i$, $i = 0, \ldots, m$, must be null.

Let Π_k^1 be the set of all polynomials of degree exactly k whose graphic pass through $(0, 1)$. We now see that the polynomial $s_k(\lambda)$ in the expression $\mathbf{r}_k = s_k(A)\mathbf{r}_0$ is a very particular polynomial in the class Π_k^1: it makes the norm of the vector $p_k(A)\mathbf{r}_0$, $p_k \in \Pi_k^1$, minimum (for a suitable choice of the norm). This result let us give estimates of the rate of convergence of CG, as precise as good is the knowledge about the location of the eigenvalues of A. For example, if it is known that the eigenvalues of A *cluster* around 1, then CG must converge with a superlinear rate of convergence (see below).

Notice that $\mathbf{r}_k = s_k(A)\mathbf{r}_0 = \mathbf{r}_0 + \hat{\mathbf{h}}_k$, for a particular vector $\hat{\mathbf{h}}_k$ in the space $\mathcal{M} = \text{Span}\,\{A\mathbf{r}_0, A^2\mathbf{r}_0, \dots, A^k\mathbf{r}_0\}$. Take a generic vector $\mathbf{h}_k$ in this space. Then

$$\begin{aligned}\|\mathbf{r}_0 + \mathbf{h}_k\|^2_{A^{-1}} &= \|\mathbf{r}_0 + \hat{\mathbf{h}}_k + \mathbf{h}_k - \hat{\mathbf{h}}_k\|^2_{A^{-1}} \\ &= \|\mathbf{r}_0 + \hat{\mathbf{h}}_k\|^2_{A^{-1}} + \|\mathbf{h}_k - \hat{\mathbf{h}}_k\|^2_{A^{-1}} + 2(\mathbf{r}_0 + \hat{\mathbf{h}}_k, \mathbf{h}_k - \hat{\mathbf{h}}_k)_{A^{-1}}.\end{aligned}$$

Now observe that the latter inner product is null, in fact, for $j = 0, \dots, k-1$, $0 = \mathbf{r}_k^T\mathbf{r}_j = \mathbf{r}_k^T A^{-1} A\mathbf{r}_j = (\mathbf{r}_k, A\mathbf{r}_j)_{A^{-1}}$, that is, $\mathbf{r}_k$ is A^{-1}-orthogonal to the space $\text{Span}\,\{A\mathbf{r}_0, A\mathbf{r}_1, \dots, A\mathbf{r}_{k-1}\}$, but this space is exactly $\mathcal{M}$. The thesis follows since $\mathbf{h}_k - \hat{\mathbf{h}}_k \in \mathcal{M}$. So we have:

$$\|\mathbf{r}_0 + \mathbf{h}_k\|^2_{A^{-1}} = \|\mathbf{r}_0 + \hat{\mathbf{h}}_k\|^2_{A^{-1}} + \|\mathbf{h}_k - \hat{\mathbf{h}}_k\|^2_{A^{-1}} \geq \|\mathbf{r}_0 + \hat{\mathbf{h}}_k\|^2_{A^{-1}}.$$

In other words,

$$\begin{aligned}\|\mathbf{r}_k\|^2_{A^{-1}} &= \|\mathbf{r}_0 + \hat{\mathbf{h}}_k\|^2_{A^{-1}} = \min\{\|\mathbf{r}_0 + \mathbf{h}_k\|^2_{A^{-1}} : \mathbf{h}_k \in \mathcal{M}\} \\ &= \min\{\|p_k(A)\mathbf{r}_0\|^2_{A^{-1}} : p_k \in \Pi_k^1\}.\end{aligned} \tag{m}$$

Comparison with GMRES. Notice that for any $\mathbf{h}_k \in \mathcal{M}$ we have

$$\mathbf{r}_0 + \mathbf{h}_k = \mathbf{b} - A(\mathbf{x}_0 + \mathbf{z}),\ \mathbf{z} = -A^{-1}\mathbf{h}_k \in \mathbb{K}_k(\mathbf{r}_0) = \text{Span}\,\{\mathbf{r}_0, A\mathbf{r}_0, \dots, A^{k-1}\mathbf{r}_0\}.$$

Thus, the vector $\mathbf{x}_k$ generated by the CG method is of type $\mathbf{x}_0 + \hat{\mathbf{z}}$ where $\hat{\mathbf{z}}$ solves the problem

$$\|\mathbf{b} - A(\mathbf{x}_0 + \hat{\mathbf{z}})\|_{A^{-1}} = \min\{\|\mathbf{b} - A(\mathbf{x}_0 + \mathbf{z})\|_{A^{-1}} : \mathbf{z} \in \mathbb{K}_k(\mathbf{r}_0)\} \tag{p}$$

($\mathbb{K}_k(\mathbf{r}_0)$ is known as Krylov space). GMRES is a method able to solve $A\mathbf{x} = \mathbf{b}$ in at most n steps under the only assumption $\det(A) \neq 0$. (Like CG, GMRES in order to be competitive must be used as an iterative method, i.e. less than n steps must be sufficient to give a good approximation of $\mathbf{x}$). In the k-th step of GMRES it is defined a vector $\mathbf{x}_k$ of type $\mathbf{x}_0 + \hat{\mathbf{z}}$ where $\hat{\mathbf{z}}$ solves exactly the problem (p) but the norm involved is the euclidean one. So, CG is a minimal residual algorithm different from $\text{GMRES}|_{A\,pd}$.

It is easy to see that the condition (m) can be rewritten as follows:

$$\|\mathbf{x} - \mathbf{x}_k\|^2_A = \min_{p_k \in \Pi_k^1} \|p_k(A)(\mathbf{x} - \mathbf{x}_0)\|^2_A.$$

Now we give a bound for the quantity $\|p_k(A)(\mathbf{x} - \mathbf{x}_0)\|^2_A$, $p_k \in \Pi_k^1$, which can be evaluated if (besides $A, \mathbf{b}$) also some information about the location of the eigenvalues λ_i of A is given. Let $\mathbf{v}_i \neq \mathbf{0}$ be such that $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$, $\mathbf{v}_i^T\mathbf{v}_j = \delta_{ij}$. Then

$$\begin{aligned}\|p_k(A)(\mathbf{x} - \mathbf{x}_0)\|^2_A &= (\mathbf{x} - \mathbf{x}_0)^T A p_k(A)^2(\mathbf{x} - \mathbf{x}_0) \\ &= (\textstyle\sum \alpha_i\mathbf{v}_i)^T \sum \alpha_i A p_k(A)^2\mathbf{v}_i \\ &= (\textstyle\sum \alpha_i\mathbf{v}_i)^T \sum \alpha_i A p_k(\lambda_i)^2\mathbf{v}_i \\ &= (\textstyle\sum \alpha_i\mathbf{v}_i)^T \sum \alpha_i \lambda_i p_k(\lambda_i)^2\mathbf{v}_i \\ &= \textstyle\sum \alpha_i^2\lambda_i p_k(\lambda_i)^2 \leq \max_i |p_k(\lambda_i)|^2\|\mathbf{x} - \mathbf{x}_0\|^2_A.\end{aligned}$$

So, we obtain the following

Third main result: If $\mathbf{x}_k$ is the k-th vector generated by CG when applied to solve the pd linear system $A\mathbf{x} = \mathbf{b}$, then

$$\|\mathbf{x}-\mathbf{x}_k\|_A^2 = \min_{p_k \in \Pi_k^1} \|p_k(A)(\mathbf{x}-\mathbf{x}_0)\|_A^2 \leq \max_i |p_k(\lambda_i)|^2 \|\mathbf{x}-\mathbf{x}_0\|_A^2, \quad \forall\, p_k \in \Pi_k^1.$$

So, if $S \subset \mathbb{R}$, $p_k \in \Pi_k^1$, $M_k \in \mathbb{R}$ are known such that $\lambda_i \in S\ \forall\, i$ and $|p_k(\lambda)| \leq M_k$ $\forall\, \lambda \in S$, then $\|\mathbf{x}-\mathbf{x}_k\|_A \leq M_k \|\mathbf{x}-\mathbf{x}_0\|_A$.

Let us see two applications of the latter result. As consequences of the first application we observe that CG (considered as an iterative method) has a linear rate of convergence, is in general faster than G, and is competitive (f.i. with direct methods) if $\lambda_{\max}$ and $\lambda_{\min}$ are comparable. However, as a consequence of the second application, the latter condition is not necessary: the rate of convergence of CG remains high (so, CG remains competitive) if most of the eigenvalues are in $[\lambda_{\min}, \hat{\lambda}]$ with $\lambda_{\min}$ and $\hat{\lambda}$ comparable. Further useful applications of the Third main result hold. In particular, as a consequence of one of these (see below), it can be stated that CG has a superlinear rate of convergence if most of the eigenvalues of A are in the interval $S = [1-\varepsilon, 1+\varepsilon]$.

(1)

$$S = [\lambda_{\min}, \lambda_{\max}],\ \ p_k(x) = \frac{T_k\left(\frac{\lambda_{\max}+\lambda_{\min}-2x}{\lambda_{\max}-\lambda_{\min}}\right)}{T_k\left(\frac{\lambda_{\max}+\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}\right)} \ \ \Rightarrow$$

$$\|\mathbf{x}-\mathbf{x}_k\|_A < 2\left(\frac{\sqrt{\mu_2(A)}-1}{\sqrt{\mu_2(A)}+1}\right)^k \|\mathbf{x}-\mathbf{x}_0\|_A,\ \ \mu_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

(2)

$$S = [\lambda_{\min}, \hat{\lambda}] \cup \{\lambda_i : \lambda_i > \hat{\lambda}\},\ \ r_{\hat{\lambda}} = \#\{i : \lambda_i > \hat{\lambda}\},$$

$$p_k(x) = \Pi_{i:\, \lambda_i > \hat{\lambda}} \left(1 - \frac{x}{\lambda_i}\right) \frac{T_{k-r_{\hat{\lambda}}}\left(\frac{\hat{\lambda}+\lambda_{\min}-2x}{\hat{\lambda}-\lambda_{\min}}\right)}{T_{k-r_{\hat{\lambda}}}\left(\frac{\hat{\lambda}+\lambda_{\min}}{\hat{\lambda}-\lambda_{\min}}\right)} \ \ \Rightarrow$$

$$\|\mathbf{x}-\mathbf{x}_k\|_A < 2\left(\frac{\sqrt{\hat{\lambda}/\lambda_{\min}}-1}{\sqrt{\hat{\lambda}/\lambda_{\min}}+1}\right)^{k-r_{\hat{\lambda}}} \|\mathbf{x}-\mathbf{x}_0\|_A, \quad k \geq r_{\hat{\lambda}}.$$

The applications (1) and (2) of the Third main result suggest an idea. When $\lambda_{\min}$ and $\lambda_{\max}$ are not comparable and the eigenvalues of A are uniformly distributed in the interval $[\lambda_{\min}, \lambda_{\max}]$ (in this case all n steps of CG are required in order to give a good approximation of $\mathbf{x}$), replace the given system $A\mathbf{x} = \mathbf{b}$ with an equivalent system $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$, $\tilde{A} = E^{-1}AE^{-T}$, $\tilde{\mathbf{x}} = E^T\mathbf{x}$, $\tilde{\mathbf{b}} = E^{-1}\mathbf{b}$, $\det(E) \neq 0$, where the matrix E is such that $\mu_2(\tilde{A}) < \mu_2(A)$ and has one of the following properties

- $\mu_2(\tilde{A}) << \mu_2(A)$
- $\tilde{A}$ has much less distinct eigenvalues than A

- $\tilde{A}$ has the eigenvalues much more clustered (around 1) than A

Then apply CG to $\tilde{A}\tilde{x} = \tilde{\mathbf{b}}$.

If such matrix E can be found, then the pd matrix $P = EE^T$ is said *preconditioner*.

Note that $E^{-T}\tilde{A}E^T = P^{-1}A$, so one could look directly for a pd matrix P such that the (real positive) eigenvalues of $P^{-1}A$ have the required properties. For example, in order to obtain something of type $P^{-1}A \approx I$ (which would result in a very high increase of the CG rate of convergence) one could choose P as an approximation $\mathcal{A}$ of A. We shall see that applying CG to $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ requires, for each step, a surplus of computation: solve a system of type $P\mathbf{z} = \mathbf{h}_k$. This computation must not make CG slow, in other words P must be a lower complexiy matrix than A. Also notice that E_1 and E_2, $E_1 \neq E_2$, $E_1E_1^T = E_2E_2^T$, define matrices $\tilde{A}_1 = E_1^{-1}AE_1^{-T}$ and $\tilde{A}_2 = E_2^{-1}AE_2^{-T}$, $\tilde{A}_1 \neq \tilde{A}_2$, with the same spectrum. For this reason one prefers to call preconditioner P instead of E.

A final remark. The vector $\mathbf{x} = A^{-1}\mathbf{b}$ we are looking for is also the minimum point of the function $F(\mathbf{z}) = \frac{1}{2}\mathbf{z}^TA\mathbf{z} - \mathbf{z}^T\mathbf{b}$. Analogously, $\tilde{\mathbf{x}} = \tilde{A}^{-1}\tilde{\mathbf{b}}$ is the minimum point of the function $\tilde{F}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T\tilde{A}\mathbf{z} - \mathbf{z}^T\tilde{\mathbf{b}}$. The preconditioning technique replaces the (sections of the) contours of F with the more spherical (sections of the) contours of $\tilde{F}$, and this results in a more efficient minimization when using gradient-type methods.

Let us write the preconditioned version of the CG algorithm, well defined once that A, $\mathbf{b}$ and the preconditioner P are given.

Let us apply CG to the system $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$:

```
$\tilde{\mathbf{x}}_0 \in \mathbb{R}^n,\ \tilde{\mathbf{r}}_0 = \tilde{\mathbf{b}} - \tilde{A}\tilde{\mathbf{x}}_0,\ \tilde{\mathbf{d}}_0 = \tilde{\mathbf{r}}_0.$
For $k = 0, 1, \ldots, \{$
$\tilde{\tau}_k = \frac{\tilde{\mathbf{r}}_k^T\tilde{\mathbf{r}}_k}{\tilde{\mathbf{d}}_k^T\tilde{A}\tilde{\mathbf{d}}_k}$
$\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{x}}_k + \tilde{\tau}_k\tilde{\mathbf{d}}_k$
$\tilde{\mathbf{r}}_{k+1} = \tilde{\mathbf{b}} - \tilde{A}\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{r}}_k - \tilde{\tau}_k\tilde{A}\tilde{\mathbf{d}}_k$
$\tilde{\beta}_k = \frac{\tilde{\mathbf{r}}_{k+1}^T\tilde{\mathbf{r}}_{k+1}}{\tilde{\mathbf{r}}_k^T\tilde{\mathbf{r}}_k}$
$\tilde{\mathbf{d}}_{k+1} = \tilde{\mathbf{r}}_{k+1} + \tilde{\beta}_k\tilde{\mathbf{d}}_k$
$\}$
```

Note that the convergence rate of the sequence $\{\tilde{\mathbf{x}}_k\}$ can be evaluated by using the following results

$$\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_k\|_{\tilde{A}} < 2\left(\frac{\sqrt{\mu_2(\tilde{A})} - 1}{\sqrt{\mu_2(\tilde{A})} + 1}\right)^k \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_0\|_{\tilde{A}}, \quad \mu_2(\tilde{A}) = \frac{\tilde{\lambda}_{\max}}{\tilde{\lambda}_{\min}},$$

$$\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_k\|_{\tilde{A}} < 2\left(\frac{\sqrt{\tilde{\hat{\lambda}}/\tilde{\lambda}_{\min}} - 1}{\sqrt{\tilde{\hat{\lambda}}/\tilde{\lambda}_{\min}} + 1}\right)^{k - r_{\tilde{\hat{\lambda}}}} \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_0\|_{\tilde{A}}, \quad k \geq r_{\tilde{\hat{\lambda}}} :$$

if $\mu_2(\tilde{A}) << \mu_2(A)$ or $\tilde{A}$ has most of the eigenvalues $\tilde{\lambda}_i$ in $[\tilde{\lambda}_{\min}, \tilde{\hat{\lambda}}]$ and $\tilde{\hat{\lambda}}/\tilde{\lambda}_{\min} << \lambda_{\max}/\lambda_{\min}$, then $\tilde{\mathbf{x}}_k \to \tilde{\mathbf{x}} = E^T\mathbf{x}$ with a greater rate than $\mathbf{x}_k \to \mathbf{x}$.

Now we obtain each row of the preconditioned CG method. Define $\mathbf{x}_k = E^{-T}\tilde{\mathbf{x}}_k$, $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$, and $\mathbf{d}_k = E^{-T}\tilde{\mathbf{d}}_k$. Then

$$\begin{aligned} \tilde{\mathbf{r}}_k &= \tilde{\mathbf{b}} - \tilde{A}\tilde{\mathbf{x}}_k = E^{-1}\mathbf{b} - E^{-1}AE^{-T}(E^T\mathbf{x}_k) \\ &= E^{-1}\mathbf{r}_k = E^T E^{-T} E^{-1}\mathbf{r}_k = E^T\mathbf{h}_k, \mathbf{h}_k = P^{-1}\mathbf{r}_k, \end{aligned}$$

$$\begin{aligned} \tilde{\mathbf{r}}_k^T\tilde{\mathbf{r}}_k &= \mathbf{r}_k^T E^{-T}E^{-1}\mathbf{r}_k = \mathbf{r}_k^T\mathbf{h}_k, \\ \tilde{\mathbf{d}}_k^T\tilde{A}\tilde{\mathbf{d}}_k &= \tilde{\mathbf{d}}_k^T E^{-1}AE^{-T}\tilde{\mathbf{d}}_k = \mathbf{d}_k^T A\mathbf{d}_k. \end{aligned}$$

Thus

$$\tilde{\tau}_k = \frac{\mathbf{r}_k^T\mathbf{h}_k}{\mathbf{d}_k^T A\mathbf{d}_k}. \qquad \text{(row1)}$$

Moreover, we have

$$\begin{aligned} &\tilde{\mathbf{x}}_{k+1} = E^T\mathbf{x}_{k+1} = E^T\mathbf{x}_k + \tilde{\tau}_k E^T\mathbf{d}_k \Rightarrow \\ &\mathbf{x}_{k+1} = \mathbf{x}_k + \tilde{\tau}_k\mathbf{d}_k, \end{aligned} \qquad \text{(row2)}$$

$$\begin{aligned} &\tilde{\mathbf{r}}_{k+1} = E^{-1}\mathbf{r}_{k+1} = E^{-1}\mathbf{r}_k - \tilde{\tau}_k E^{-1}AE^{-T}E^T\mathbf{d}_k \Rightarrow \\ &\mathbf{r}_{k+1} = \mathbf{r}_k + \tilde{\tau}_k A\mathbf{d}_k, \end{aligned} \qquad \text{(row3)}$$

$$\tilde{\beta}_k = \frac{\mathbf{r}_{k+1}^T\mathbf{h}_{k+1}}{\mathbf{r}_k^T\mathbf{h}_k} \qquad \text{(row4)}$$

(row3.5: $\mathbf{h}_{k+1} = P^{-1}\mathbf{r}_{k+1}$),

$$\begin{aligned} &\tilde{\mathbf{d}}_{k+1} = E^T\mathbf{d}_{k+1} = E^T\mathbf{h}_{k+1} + \tilde{\beta}_k E^T\mathbf{d}_k \Rightarrow \\ &\mathbf{d}_{k+1} = \mathbf{h}_{k+1} + \tilde{\beta}_k\mathbf{d}_k. \end{aligned} \qquad \text{(row5)}$$

Finally, in order to initialize the algorithm, set:

$$\begin{aligned} &\mathbf{x}_0 = E^{-T}\tilde{\mathbf{x}}_0,\ \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0, \\ &\mathbf{d}_0 = E^{-T}\tilde{\mathbf{d}}_0 = E^{-T}\tilde{\mathbf{r}}_0 = E^{-T}E^T\mathbf{h}_0 = \mathbf{h}_0. \end{aligned} \qquad \text{(row0)}$$

Regarding the convergence rate of the sequence $\{\mathbf{x}_k\}$, generated by the algorithm row0 and, for $k = 0, 1, \ldots$, rows$1, 2, 3, 3.5, 4, 5$, note that

$$\begin{aligned} \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}\|_{\tilde{A}}^2 &= (\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}})^T\tilde{A}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}) \\ &= (E^T\mathbf{x}_k - E^T\mathbf{x})^T E^{-1}AE^{-T}(E^T\mathbf{x}_k - E^T\mathbf{x}) \\ &= (\mathbf{x}_k - \mathbf{x})^T A(\mathbf{x}_k - \mathbf{x}) = \|\mathbf{x}_k - \mathbf{x}\|_A^2. \end{aligned}$$

Thus the bounds for $\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_k\|_{\tilde{A}}$ obtained above, can be rewritten as follows

$$\begin{aligned} &\frac{\|\mathbf{x}_k - \mathbf{x}\|_A}{\|\mathbf{x}_0 - \mathbf{x}\|_A} \leq 2\left(\frac{\sqrt{\mu_2(\tilde{A})}-1}{\sqrt{\mu_2(\tilde{A})}+1}\right)^k, \quad \mu_2(\tilde{A}) = \frac{\tilde{\lambda}_{\max}}{\tilde{\lambda}_{\min}}, \\ &\frac{\|\mathbf{x}_k - \mathbf{x}\|_A}{\|\mathbf{x}_0 - \mathbf{x}\|_A} \leq 2\left(\frac{\sqrt{\hat{\tilde{\lambda}}/\tilde{\lambda}_{\min}}-1}{\sqrt{\hat{\tilde{\lambda}}/\tilde{\lambda}_{\min}}+1}\right)^{k-r_{\hat{\tilde{\lambda}}}}, \quad k \geq r_{\hat{\tilde{\lambda}}}. \end{aligned}$$

Why clustering around 1 *is good*

Let A be a p.d. matrix and ε, $0 < \varepsilon < 1$, be fixed.

Denote by λ_j^ε the eigenvalues of A outside the interval $[1-\varepsilon, 1+\varepsilon]$ and by r_ε the number of such eigenvalues. Set $S = [1-\varepsilon, 1+\varepsilon] \cup \{\lambda_j^\varepsilon\}$ and let p_q be the polynomial

$$p_q(\lambda) = \prod_{\lambda_j^\varepsilon} \left(1 - \frac{\lambda}{\lambda_j^\varepsilon}\right) \frac{T_{q-r_\varepsilon}((1-\lambda)/\varepsilon)}{T_{q-r_\varepsilon}(1/\varepsilon)}, \ q \geq r_\varepsilon$$

where $T_k(x)$ denotes the chebycev polynomial of degree k. $((b+a-2\lambda)/(b-a) = (1-\lambda)/\varepsilon$, $(b+a)/(b-a) = 1/\varepsilon$, if $a = 1-\varepsilon$, $b = 1+\varepsilon$). Notice that S is a set containing all the eigenvalues of A, and p_q has exactly degree q and $p_q(0) = 1$. Then one can say that if $\mathbf{x}_q$ is the q-th vector generated by the CG method when solving $A\mathbf{x} = \mathbf{b}$, then

$$\|\mathbf{x} - \mathbf{x}_q\|_A \leq (\max_{\lambda \in S} |p_q(\lambda)|)\|\mathbf{x} - \mathbf{x}_0\|_A. \tag{bound}$$

This bound for $\|\mathbf{x} - \mathbf{x}_q\|_A$ allows a better evaluation of the CG rate of convergence with respect to the well known bound

$$\|\mathbf{x} - \mathbf{x}_q\|_A \leq 2\left(\frac{\sqrt{\mu_2(A)}-1}{\sqrt{\mu_2(A)}+1}\right)^q \|\mathbf{x} - \mathbf{x}_0\|_A, \quad \mu_2(A) = \frac{\max \lambda(A)}{\min \lambda(A)} \tag{wkbound}$$

in case it is known that most of (almost all) the eigenvalues of A are in some interval $[1-\varepsilon, 1+\varepsilon]$ where ε is small (almost zero).

If, moreover, the $n \times n$ linear system $A\mathbf{x} = \mathbf{b}$ can be seen as one of a sequence of increasing order linear systems, with the property that $\forall\, \varepsilon > 0\ \exists\, k_\varepsilon$, n_ε such that for all $n > n_\varepsilon$ outside $[1-\varepsilon, 1+\varepsilon]$ fall no more than n_ε eigenvalues of A, then (bound) allows to prove the superlinear convergence of CG.

(Note that in general CG has a linear rate of convergence, as a consequence of (wkbound)).

Let us prove these assertions, by evaluating $\max_{\lambda \in S} |p_q(\lambda)|$.

$$\begin{array}{rcl}\max_{\lambda\in S}|p_q(\lambda)| & = & \max_{\lambda\in[1-\varepsilon,1+\varepsilon]}|p_q(\lambda)| \\ & \leq & (\max_{\dots}\prod_{\lambda_j^\varepsilon}\left|1-\frac{\lambda}{\lambda_j^\varepsilon}\right|)(\max_{\dots}\left|\frac{T_{q-r_\varepsilon}((1-\lambda)/\varepsilon)}{T_{q-r_\varepsilon}(1/\varepsilon)}\right|) \\ & = & (\max_{\dots}\prod_{\lambda_j^\varepsilon}\left|1-\frac{\lambda}{\lambda_j^\varepsilon}\right|)\frac{1}{T_{q-r_\varepsilon}(1/\varepsilon)}.\end{array}$$

Now first notice that

$$T_{q-r_\varepsilon}\left(\frac{1}{\varepsilon}\right) = T_{q-r_\varepsilon}\left(\frac{\frac{1+\varepsilon}{1-\varepsilon}+1}{\frac{1+\varepsilon}{1-\varepsilon}-1}\right) > \frac{1}{2}\left(\frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}+1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}-1}\right)^{q-r_\varepsilon}.$$

Then denote by $\hat{\lambda}_j^\varepsilon$ those eigenvalues λ_j^ε satisfying the inequalities

$$\lambda_j^\varepsilon < 1-\varepsilon, \quad \lambda_j^\varepsilon < \frac{1}{2}(1+\varepsilon)$$

and observe that

$$\begin{array}{rcl}\max_{\lambda\in[1-\varepsilon,1+\varepsilon]}\prod_{\lambda_j^\varepsilon}\left|1-\frac{\lambda}{\lambda_j^\varepsilon}\right| & \leq & \max_{\dots}\prod_{\hat{\lambda}_j^\varepsilon}\left|1-\frac{\lambda}{\lambda_j^\varepsilon}\right| \\ & = & \prod_{\hat{\lambda}_j^\varepsilon}\left(\frac{1+\varepsilon}{\hat{\lambda}_j^\varepsilon}-1\right).\end{array}$$

So, we have

$$\begin{array}{rcl}
\max_{\lambda\in S}|p_q(\lambda)| & \leq & \prod_{\hat{\lambda}_j^\varepsilon}\left(\frac{1+\varepsilon}{\hat{\lambda}_j^\varepsilon}-1\right)2\left(\frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}-1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}+1}\right)^{q-r_\varepsilon} \\
 & \leq & 2\left(\frac{1+\varepsilon}{\min\lambda(A)}-1\right)^{\#\hat{\lambda}_j^\varepsilon}\left(\frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}-1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}+1}\right)^{q-r_\varepsilon} \\
 & \approx & \left(\frac{1+\varepsilon}{\min\lambda(A)}-1\right)^{\#\hat{\lambda}_j^\varepsilon}\frac{\varepsilon^q}{\varepsilon^{r_\varepsilon}2^{q-r_\varepsilon-1}},
\end{array}$$

where in the latter approximation we have used the following Taylor expansion

$$f(\varepsilon)=\frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}-1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}+1}=\frac{\varepsilon}{2}+\frac{\varepsilon^2}{2}f''(0)+\ldots.$$

From CG to GMRES

The minimal property satisfied by the residual $\mathbf{r}_k=\mathbf{b}-A\mathbf{x}_k$ at the kth iteration of the CG method can be rewritten in a way that allows us to compare the $\mathbf{x}_k$ generated by CG with the approximation $\mathbf{x}_k$ defined by GMRES, which is a method sharing CG properties (convergence in at most the number of distict eigenvalues of A; fast convergence if the eigenvalues of A are clustered), but working for any linear system $A\mathbf{x}=\mathbf{b}$ (i.e. not limited to positive definite ones). However, we immediately underline that GMRES is not as cheap as CG.

Note that, for any $p_k\in\Pi_k^1$,

$$\begin{array}{rcl}
p_k(A)\mathbf{r}_0 & = & \mathbf{r}_0+\alpha_1A\mathbf{r}_0+\ldots+\alpha_kA^k\mathbf{r}_0 \\
 & = & \mathbf{b}-A\mathbf{x}_0+\alpha_1A\mathbf{r}_0+\ldots+\alpha_kA^k\mathbf{r}_0 \\
 & = & \mathbf{b}-A(\mathbf{x}_0-\alpha_1\mathbf{r}_0-\ldots-\alpha_kA^{k-1}\mathbf{r}_0) \\
 & = & \mathbf{b}-A(\mathbf{x}_0+\mathbf{z}), \\
\multicolumn{3}{l}{\mathbf{z}\in\ \mathrm{Span}\,\{\mathbf{r}_0,\,A\mathbf{r}_0,\ldots,\,A^{k-1}\mathbf{r}_0\}.}
\end{array}$$

Thus the result

$\mathbf{x}_k$ of CG is such that
$\|\mathbf{r}_k\|_{A^{-1}}=\|s_k(A)\mathbf{r}_0\|_{A^{-1}}=\min_{p_k\in\Pi_k^1}\|p_k(A)\mathbf{r}_0\|_{A^{-1}}$

can be rewritten as follows

$\mathbf{x}_k$ of CG is equal to $\mathbf{x}_0+\tilde{\mathbf{z}}$ with
$\tilde{\mathbf{z}}$ such that $\|\mathbf{b}-A(\mathbf{x}_0+\tilde{\mathbf{z}})\|_{A^{-1}}=\min_{\mathbf{z}\in\mathbb{K}_k}\|\mathbf{b}-A(\mathbf{x}_0+\mathbf{z})\|_{A^{-1}}$,
where $\mathbb{K}_k=$ Span $\{\mathbf{r}_0,\,A\mathbf{r}_0,\ldots,\,A^{k-1}\mathbf{r}_0\}$.

Note that $\mathbb{K}_k=\mathbb{K}_k(\mathbf{r}_0)$ is a Krylov space and has dimension at most k.

Now, consider a generic linear system $A\mathbf{x}=\mathbf{b}$, where A is assumed non singular. Then the approximation of $\mathbf{x}=A^{-1}\mathbf{b}$ proposed by GMRES at step k is defined as follows:

$\mathbf{x}_k$ of GMRES is equal to $\mathbf{x}_0+\tilde{\mathbf{z}}$ with
$\tilde{\mathbf{z}}$ such that $\|\mathbf{b}-A(\mathbf{x}_0+\tilde{\mathbf{z}})\|_2=\min_{\mathbf{z}\in\mathbb{K}_k}\|\mathbf{b}-A(\mathbf{x}_0+\mathbf{z})\|_2$,
where $\mathbb{K}_k=$ Span $\{\mathbf{r}_0,\,A\mathbf{r}_0,\ldots,\,A^{k-1}\mathbf{r}_0\}$.

So, in case A is positive definite, the “only” difference between GMRES and CG is in the fact that in GMRES the euclidean norm of the residual is minimized, whereas in CG the norm used is $\|\cdot\|_{A^{-1}}$.

The cost of the computation of $\tilde{\mathbf{z}}$ in GMRES grows with k. For this reason, 1) GMRES is not competitive with CG when A is positive definite; 2) when A is not positive definite, GMRES is more efficient than other linear system solvers only if A has suitable spectral properties (a “small” number of distinct eigenvalues, eigenvalues clustering, etc), otherwise GMRES must be modified to be competitive (restarted GMRES, etc).

Performing the kth iteration of GMRES

First we have to find an orthonormal basis $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ of $\mathbb{K}_k$, by using the *Arnoldi procedure*, so that $\tilde{\mathbf{z}} \in \mathbb{K}_k$ ($\mathbf{z} \in \mathbb{K}_k$) can be represented as $\tilde{\mathbf{z}} = V_k\tilde{\mathbf{y}}$, $\tilde{\mathbf{y}} \in \mathbb{R}^k$ ($\mathbf{z} = V_k\mathbf{y}$, $\mathbf{y} \in \mathbb{R}^k$), where V_k is the $n \times k$ matrix whose columns are $\mathbf{v}_1, \dots, \mathbf{v}_k$.

Let us find such basis. Let $\mathbf{x}_0 \in \mathbb{R}^n$ and set $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$.

If $\mathbf{r}_0 \neq \mathbf{0}$, then set

$$\begin{array}{l} \mathbf{v}_1 = \frac{1}{\|\mathbf{r}_0\|_2}\mathbf{r}_0,\ \hat{\mathbf{v}}_2 = A\mathbf{v}_1 - h_{11}\mathbf{v}_1, \\ h_{11} = (A\mathbf{v}_1, \mathbf{v}_1),\ h_{21} = \|\hat{\mathbf{v}}_2\|_2. \end{array}$$

Note that $(\hat{\mathbf{v}}_2, \mathbf{v}_1) = 0$.

If $\hat{\mathbf{v}}_2 \neq \mathbf{0}$ ($h_{21} \neq 0$), then set

$$\begin{array}{l} \mathbf{v}_2 = \frac{1}{h_{21}}\hat{\mathbf{v}}_2,\ \hat{\mathbf{v}}_3 = A\mathbf{v}_2 - h_{12}\mathbf{v}_1 - h_{22}\mathbf{v}_2, \\ h_{12} = (A\mathbf{v}_2, \mathbf{v}_1),\ h_{22} = (A\mathbf{v}_2, \mathbf{v}_2),\ h_{32} = \|\hat{\mathbf{v}}_3\|_2. \end{array}$$

Note that $(\hat{\mathbf{v}}_3, \mathbf{v}_1) = (\hat{\mathbf{v}}_3, \mathbf{v}_2) = 0$. ...

If $\hat{\mathbf{v}}_m \neq \mathbf{0}$ ($h_{mm-1} \neq 0$), then set

$$\begin{array}{l} \mathbf{v}_m = \frac{1}{h_{mm-1}}\hat{\mathbf{v}}_m, \quad \hat{\mathbf{v}}_{m+1} = A\mathbf{v}_m - h_{1m}\mathbf{v}_1 - h_{2m}\mathbf{v}_2 \ldots - h_{mm}\mathbf{v}_m, \\ h_{1m} = (A\mathbf{v}_m, \mathbf{v}_1),\ h_{2m} = (A\mathbf{v}_m, \mathbf{v}_2),\ \ldots \\ \ldots,\ h_{mm} = (A\mathbf{v}_m, \mathbf{v}_m),\ h_{m+1m} = \|\hat{\mathbf{v}}_{m+1}\|_2. \end{array}$$

Note that $(\hat{\mathbf{v}}_{m+1}, \mathbf{v}_1) = (\hat{\mathbf{v}}_{m+1}, \mathbf{v}_2) = \ldots = (\hat{\mathbf{v}}_{m+1}, \mathbf{v}_m) = 0$.

If $\hat{\mathbf{v}}_{m+1} \neq \mathbf{0}$ ($h_{m+1m} \neq 0$), then set $\mathbf{v}_{m+1} = \frac{1}{h_{m+1m}}\hat{\mathbf{v}}_{m+1}$ ...

Observe that the vectors $\mathbf{v}_k$ defined as above satisfy the vectorial identities

$$\begin{array}{l} A\mathbf{v}_1 = h_{11}\mathbf{v}_1 + h_{21}\mathbf{v}_2 \\ A\mathbf{v}_2 = h_{12}\mathbf{v}_1 + h_{22}\mathbf{v}_2 + h_{32}\mathbf{v}_3 \\ \ldots \\ A\mathbf{v}_m = h_{1m}\mathbf{v}_1 + h_{2m}\mathbf{v}_2 + \ldots + h_{mm}\mathbf{v}_m + h_{m+1m}\mathbf{v}_{m+1} \end{array}$$

or, equivalently, the matrix identity

$$AV_m = V_{m+1}\tilde{H}_m = V_mH_m + \mathbf{v}_{m+1}[0\ \cdots\ 0\, h_{m+1m}],$$

where $V_m = \Big[\mathbf{v}_1\,\mathbf{v}_2\ \cdots\ \mathbf{v}_m\Big]$,

$$H_m = \begin{bmatrix} h_{11} & h_{12} & \cdot & \cdot & h_{1m} \\ h_{21} & h_{22} & \cdot & & \cdot \\ & h_{32} & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & h_{m-1m} \\ & & & h_{mm-1} & h_{mm} \end{bmatrix}, \quad \tilde{H}_m = \begin{bmatrix} H_m \\ [\ 0 \quad \cdot \quad 0 \quad h_{m+1,m}\ ] \end{bmatrix}.$$

Note that H_m is a Hessenberg matrix.

It is clear that

$$\begin{aligned}
&\mathbf{r}_0 \neq \mathbf{0} \;\Rightarrow\; \mathbb{K}_1(\mathbf{r}_0) = \mathbb{K}_1(\mathbf{v}_1) = \text{ Span}\,\{\mathbf{v}_1\} \\
&\mathbf{r}_0, \hat{\mathbf{v}}_2 \neq \mathbf{0} \;\Rightarrow\; \mathbb{K}_2(\mathbf{r}_0) = \mathbb{K}_2(\mathbf{v}_1) = \text{ Span}\,\{\mathbf{v}_1, A\mathbf{v}_1\} = \text{ Span}\,\{\mathbf{v}_1, \mathbf{v}_2\} \\
&\mathbf{r}_0, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_m \neq \mathbf{0} \;\Rightarrow\; \mathbb{K}_m(\mathbf{r}_0) = \mathbb{K}_m(\mathbf{v}_1) = \text{ Span}\,\{\mathbf{v}_1, A\mathbf{v}_1, \ldots, A^{m-1}\mathbf{v}_1\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad = \text{ Span}\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}
\end{aligned}$$

i.e. $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a well defined orthonormal basis for $\mathbb{K}_m(\mathbf{r}_0)$, provided that $\hat{\mathbf{v}}_1 := \mathbf{r}_0$, $\hat{\mathbf{v}}_2, \ldots,$ $\hat{\mathbf{v}}_m \neq \mathbf{0}$ ($h_{10} := \|\hat{\mathbf{v}}_1\|_2$, $h_{21}, \ldots,$ $h_{mm-1} \neq 0$).

Now let us proceed by rewriting the function $\|\mathbf{b} - A(\mathbf{x}_0 + \mathbf{z})\|_2$ of $\mathbf{z} \in \mathbb{K}_k$, so that its minimum value and the point $\tilde{\mathbf{z}} \in \mathbb{K}_k$ where such value is assumed, are defined more explicitly and become computable.

In order to do that we change variable, i.e. we set $\mathbf{z} = V_k\mathbf{y}$, $\mathbf{y} \in \mathbb{R}^k$. Then

$$\|\mathbf{b} - A(\mathbf{x}_0 + \mathbf{z})\|_2^2 = \|\mathbf{b} - A(\mathbf{x}_0 + V_k\mathbf{y})\|_2^2 = \|\mathbf{r}_0 - AV_k\mathbf{y})\|_2^2 = \ldots$$

Let M_{n-k} be a $n \times (n-k)$ matrix whose columns are orthonormal, each other and with the columns of V_k, and assume its first column equal to $\mathbf{v}_{k+1}$:

$$\begin{aligned}
\ldots &= \| \begin{bmatrix} V_k^* \\ M_{n-k}^* \end{bmatrix} \mathbf{r}_0 - \begin{bmatrix} V_k^* \\ M_{n-k}^* \end{bmatrix} AV_k\mathbf{y}\|_2^2 \\
&= \|V_k^*\mathbf{r}_0 - V_k^*AV_k\mathbf{y}\|_2^2 + \|M_{n-k}^*\mathbf{r}_0 - M_{n-k}^*AV_k\mathbf{y}\|_2^2 = \ldots
\end{aligned}$$

The equality $AV_k = V_kH_k + \mathbf{v}_{k+1}[0 \cdots 0\, h_{k+1,k}]$ holds:

$$\ldots = \| \begin{bmatrix} \|\mathbf{r}_0\|_2 \\ 0 \\ \cdot \\ 0 \end{bmatrix} - H_k\mathbf{y}\|_2^2 + \| - M_{n-k}^*AV_k\mathbf{y}\|_2^2 = \| \begin{bmatrix} \|\mathbf{r}_0\|_2 \\ 0 \\ \cdot \\ 0 \end{bmatrix} - H_k\mathbf{y}\|_2^2 + y_k^2h_{k+1,k}^2$$

$$= \| \begin{bmatrix} \|\mathbf{r}_0\|_2 \\ 0 \\ \cdot \\ 0 \\ 0 \end{bmatrix} - \tilde{H}_k\mathbf{y}\|_2^2 = \ldots$$

Assume that we know Q_k real unitary and R_k upper triangular such that $Q_k^TH_k = R_k$. By using these matrices one can construct at a low cost Q_{k+1} real unitary and R_{k+1} upper triangular such that $Q_{k+1}^TH_{k+1} = R_{k+1}$. In fact, set

$$\tilde{Q}_k^T = \begin{bmatrix} I & & \\ & \alpha & -\beta \\ & \beta & \alpha \end{bmatrix} \begin{bmatrix} Q_k^T & \mathbf{0} \\ & \\ \mathbf{0}^T & 1 \end{bmatrix}, \; \alpha^2 + \beta^2 = 1. \tag{Q}$$

Then

$$\tilde{Q}_k^T\tilde{H}_k = \begin{bmatrix} I & & \\ & \alpha & -\beta \\ & \beta & \alpha \end{bmatrix} \begin{bmatrix} R_k \\ \\ 0 \cdots 0\; h_{k+1k} \end{bmatrix}.$$

Now choose α,β such that

$$\begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} \begin{bmatrix} [R_k]_{kk} \\ h_{k+1,k} \end{bmatrix} = \begin{bmatrix} \pm\sqrt{[R_k]_{kk}^2 + h_{k+1,k}^2} \\ 0 \end{bmatrix}$$

i.e.

$$\alpha = \alpha_k = \frac{\pm[R_k]_{kk}}{\sqrt{[R_k]_{kk}^2 + h_{k+1,k}^2}}, \ \beta = \beta_k = \frac{\mp h_{k+1,k}}{\sqrt{[R_k]_{kk}^2 + h_{k+1,k}^2}} \qquad (\alpha\beta)$$

by choosing the upper sign if $[R_k]_{kk} \geq 0$ and the lower sign if $[R_k]_{kk} < 0$ (the reason for that will be clear below). Then

$$\tilde{Q}_k^T \tilde{H}_k = \begin{bmatrix} R_k + \begin{bmatrix} & \\ & \gamma \end{bmatrix} \\ 0 \ \cdots \ \cdots \ 0 \end{bmatrix}, \ \ \gamma = \gamma_k = \pm\sqrt{[R_k]_{kk}^2 + h_{k+1,k}^2} - [R_k]_{kk}, \qquad (\gamma)$$

$$\tilde{Q}_k^T H_{k+1} = \tilde{Q}_k^T \left[\begin{array}{c|c} \tilde{H}_k & \begin{bmatrix} h_{1,k+1} \\ \cdot \\ h_{k+1,k+1} \end{bmatrix} \end{array} \right] = \left[\begin{array}{c|c} R_k + \begin{bmatrix} & \\ & \gamma \end{bmatrix} & \tilde{Q}_k^T \begin{bmatrix} h_{1,k+1} \\ \cdot \\ h_{k+1,k+1} \end{bmatrix} \\ 0 \ \cdots \ \cdots \ 0 & \end{array} \right].$$

The latter matrix is upper triangular. Call it R_{k+1} and set $Q_{k+1} = \tilde{Q}_k$. Then $Q_{k+1}^T H_{k+1} = R_{k+1}$.

$$\ldots = \|Q_{k+1}^T \begin{bmatrix} \|\mathbf{r}_0\|_2 \\ 0 \\ \cdot \\ 0 \end{bmatrix} - Q_{k+1}^T \tilde{H}_k \mathbf{y}\|^2 = \|Q_{k+1}^T \begin{bmatrix} \|\mathbf{r}_0\|_2 \\ 0 \\ \cdot \\ 0 \end{bmatrix} - \begin{bmatrix} R_k + \begin{bmatrix} & \\ & \gamma \end{bmatrix} \\ 0 \ \cdots \ \cdots \ 0 \end{bmatrix} \mathbf{y}\|_2^2$$

$$= \|I_k^1 Q_{k+1}^T \begin{bmatrix} \|\mathbf{r}_0\|_2 \\ 0 \\ \cdot \\ 0 \end{bmatrix} - (R_k + \begin{bmatrix} & \\ & \gamma \end{bmatrix})\mathbf{y}\|_2^2 + \|\mathbf{r}_0\|_2^2 [Q_{k+1}^T]_{k+1,1}^2.$$

It follows that, if $\mathbf{r}_k$ is the residual at step k of the GMRES method, then

$$\|\mathbf{r}_k\|_2^2 = \|\mathbf{b} - A(\mathbf{x}_0 + \tilde{\mathbf{z}})\|_2^2 = \min_{\mathbf{z}\in\mathbb{K}_k} \|\mathbf{b} - A(\mathbf{x}_0 + \mathbf{z})\|_2^2 = \|\mathbf{r}_0\|_2^2 [Q_{k+1}^T]_{k+1,1}^2,$$

where

$$\tilde{\mathbf{z}} = V_k \tilde{\mathbf{y}}, \ (R_k + \begin{bmatrix} & \\ & \gamma \end{bmatrix})\tilde{\mathbf{y}} = I_k^1 Q_{k+1}^T \begin{bmatrix} \|\mathbf{r}_0\|_2 \\ 0 \\ \cdot \\ 0 \end{bmatrix},$$

i.e. $\tilde{\mathbf{z}}$ can be computed by solving an upper triangular linear system of k linear equations and by performing a matrix-vector product involving a $n \times k$ matrix. However, one can do such operations, i.e. *compute* $\mathbf{x}_k$, only when the inequality $\|\mathbf{r}_k\| < TOL\|\mathbf{r}_0\|$ is satisfied, a condition that can be checked by using the following formula

$$\frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_0\|^2} = \prod_{j=1}^{k} \frac{h_{j+1,j}^2}{[R_j]_{jj}^2 + h_{j+1,j}^2}.$$

Proof: $[Q_{k+1}^T]_{k+1,1} = \beta_k [Q_k^T]_{k1} \ \Rightarrow$

$$\|\mathbf{r}_k\|_2^2 = \beta_k^2 \|\mathbf{r}_{k-1}\|_2^2 = \frac{h_{k+1,k}^2}{[R_k]_{kk}^2 + h_{k+1,k}^2} \|\mathbf{r}_{k-1}\|_2^2. \qquad \text{(r)}$$

So, we have the GMRES algorithm:

Choose $\mathbf{x}_0 \in \mathbb{R}^n$ and set $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$.

For $m = 0, 1, \ldots$:

```
Compute $h_{m+1,m} = \|\hat{\mathbf{v}}_{m+1}\|$ ($\hat{\mathbf{v}}_1 = \mathbf{r}_0$).
If $h_{m+1,m} \neq 0$, THEN {
    if $\frac{\|\mathbf{r}_m\|}{\|\mathbf{r}_0\|} < TOL$ THEN {
    compute $\mathbf{x}_m$ and stop ($\mathbf{x}_m \approx A^{-1}\mathbf{b}$),
    } ELSE { define $Q_{m+1}^T$ from $Q_m^T$ (except when $m = 0$: $Q_1^T = 1$),
    set $\mathbf{v}_{m+1} = \frac{1}{h_{m+1,m}}\hat{\mathbf{v}}_{m+1}$, compute $A\mathbf{v}_{m+1}$, $h_{j,m+1} = (A\mathbf{v}_{m+1}, \mathbf{v}_j)$,
    $j = 1, \dots, m+1$, $\hat{\mathbf{v}}_{m+2} = A\mathbf{v}_{m+1} - h_{1,m+1}\mathbf{v}_1 \dots - h_{m+1,m+1}\mathbf{v}_{m+1}$;
    define $R_{m+1}$ from $R_m$ (except when $m = 0$: $R_1 = h_{11}$) }
} ELSE compute $\mathbf{x}_m$ and stop ($\mathbf{x}_m = A^{-1}\mathbf{b}$).
```

Exercise. Count the number of of memory allocations needed to implement the above algorithm. Count the number of arithmetic operations required to do k steps.

The first step in detail

Given $\mathbf{x}_0 \in \mathbb{R}^n$, let us perform the first step of the GMRES method.

Set $k = 1$, i.e. $\mathbf{x}_1 = \mathbf{x}_0 + V_1\tilde{\mathbf{y}} = \mathbf{x}_0 + \mathbf{v}_1\tilde{y}_1 = \mathbf{x}_0 + \frac{1}{\|\mathbf{r}_0\|}\mathbf{r}_0\tilde{y}_1$ with $\tilde{y}_1 \in \mathbb{R}$ such that

$$(R_1 + \gamma_1)\tilde{y}_1 = I_1^1 Q_2^T \begin{bmatrix} \|\mathbf{r}_0\| \\ 0 \end{bmatrix}.$$

Since $H_1 = h_{11} = (A\mathbf{v}_1, \mathbf{v}_1) = \frac{\mathbf{r}_0^T A\mathbf{r}_0}{\|\mathbf{r}_0\|^2}$ and $1 \cdot h_{11} = h_{11}$, we have $Q_1^T = 1$ and $R_1 = [R_1]_{11} = h_{11}$.

Then $\hat{\mathbf{v}}_2 = A\mathbf{v}_1 - h_{11}\mathbf{v}_1$, $h_{21} = \|\hat{\mathbf{v}}_2\|$,

$$Q_2^T = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} \begin{bmatrix} Q_1^T & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix},$$
$$\alpha = \alpha_1 = \tfrac{\pm h_{11}}{\sqrt{h_{11}^2 + h_{21}^2}}, \quad \beta = \beta_1 = \tfrac{\mp h_{21}}{\sqrt{h_{11}^2 + h_{21}^2}},$$

$\gamma_1 = \pm\sqrt{h_{11}^2 + h_{21}^2} - h_{11}$ (choose the upper sign if $h_{11} > 0$ and the lower sign if $h_{11} < 0$). So, the system becomes

$$\pm\sqrt{h_{11}^2 + h_{21}^2}\tilde{y}_1 = I_1^1 \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} \begin{bmatrix} \|\mathbf{r}_0\| \\ 0 \end{bmatrix} = \alpha\|\mathbf{r}_0\|$$

Thus, $\tilde{y}_1 = \|\mathbf{r}_0\|\frac{h_{11}}{h_{11}^2 + h_{21}^2}$, and $\mathbf{x}_1 = \mathbf{x}_0 + \frac{h_{11}}{h_{11}^2 + h_{21}^2}\mathbf{r}_0$.

But $h_{21}^2 = (A\mathbf{v}_1 - h_{11}\mathbf{v}_1)^T(A\mathbf{v}_1 - h_{11}\mathbf{v}_1) = \dots = \mathbf{v}_1^T A^T A\mathbf{v}_1 - (\mathbf{v}_1^T A\mathbf{v}_1)^2$, so $h_{11}^2 + h_{21}^2 = \mathbf{v}_1^T A^T A\mathbf{v}_1 = \frac{\|A\mathbf{r}_0\|^2}{\|\mathbf{r}_0\|^2}$. It follows that

$$\mathbf{x}_1 = \mathbf{x}_0 + \frac{\mathbf{r}_0^T A\mathbf{r}_0}{\|A\mathbf{r}_0\|^2}\mathbf{r}_0,$$

i.e. $\mathbf{x}_1$ coincides with the approximation generated by one application of step-variable Richardson-Euler method. We could foresee this remark since

$\mathbf{x}_1$ of GMRES is equal to $\mathbf{x}_0 + \tilde{\mathbf{z}}$ with
$\tilde{\mathbf{z}}$ such that $\|\mathbf{b} - A(\mathbf{x}_0 + \tilde{\mathbf{z}})\|_2 = \min_{\mathbf{z}\in\mathbb{K}_1} \|\mathbf{b} - A(\mathbf{x}_0 + \mathbf{z})\|_2$,
where $\mathbb{K}_1 = \text{Span}\,\{\mathbf{r}_0\}$

is exactly the same condition required on $\mathbf{x}_1 = \mathbf{x}_0 + \omega\mathbf{r}_0$, $\omega \in \mathbb{R}$, by RE.

Note that if $\mathbf{r}_0 \neq \mathbf{0}$, then $R_1 + \gamma_1 = \pm\sqrt{h_{11}^2 + h_{21}^2} \neq 0$, i.e. the first step of GMRES is well defined. Let us show this fact. If $h_{21} \neq 0$ then the thesis is true. Assume $h_{21} = \|\hat{\mathbf{v}}_2\| = 0$. If $h_{11} = 0$ too, then $\mathbf{0} = \hat{\mathbf{v}}_2 = A\mathbf{v}_1 - h_{11}\mathbf{v}_1 = A\mathbf{v}_1$, with $\mathbf{v}_1 \neq \mathbf{0}$; this implies $\det(A) = 0$, which is absurd. So, if $h_{21} = 0$, h_{11} must be nonzero, and the proof of the fact is completed.

Note also that, from

$$\|\mathbf{r}_1\|_2^2 = \beta_1^2\|\mathbf{r}_0\|_2^2 = \frac{h_{2,1}^2}{[R_1]_{11}^2 + h_{2,1}^2}\|\mathbf{r}_0\|_2^2,$$

and the previous reasonings, it follows that 1) if $h_{21} = 0$, then $h_{11} = [R_1]_{11}$ must be nonzero and the the initial residual is canceled by the first step of GMRES. 2) if $h_{21} \neq 0$, then the norm of the initial residual is reduced by the first step of GMRES, unless the choice of the initial guess is so unlucky that $h_{11} = (A\mathbf{v}_1, \mathbf{v}_1) = 0$; in such case, the norm of the residual remains unchanged (stagnation phenomenon).

Convergence in at most n steps

A simple investigation let us observe that the upper triangular matrices R_k generated by the GMRES algorithm have the form

$$R_k = \begin{bmatrix} \sigma_1\sqrt{[R_1]_{11}^2 + h_{21}^2} & * & \cdots & * & * \\ & \sigma_2\sqrt{[R_2]_{22}^2 + h_{32}^2} & & & * \\ & & \ddots & & \vdots \\ & & & \sigma_{k-1}\sqrt{[R_{k-1}]_{k-1k-1}^2 + h_{kk-1}^2} & * \\ & & & & [R_k]_{kk} \end{bmatrix}$$

where $\sigma_j = 1$ if $[R_j]_{jj} \geq 0$ and $\sigma_j = -1$ if $[R_j]_{jj} < 0$. Also recall that the coefficient matrix of the system we have to solve to compute $\mathbf{x}_k$ is

$$R_k + \begin{bmatrix} & \\ & \gamma_k \end{bmatrix}, \; \gamma_k = \sigma_k\sqrt{[R_k]_{kk}^2 + h_{k+1,k}^2} - [R_k]_{kk}.$$

where $\sigma_k = 1$ if $[R_k]_{kk} \geq 0$ and $\sigma_k = -1$ if $[R_k]_{kk} < 0$.

Now assume that at a certain step k we have $h_{21}, \ldots, h_{k,k-1}$ nonzero (so $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are orthonormal, i.e. $V_k^*V_k = I$), but $h_{k+1k} = 0$ (so $\hat{\mathbf{v}}_{k+1} = \mathbf{0}$, and $A\mathbf{v}_k$ is a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_k$). Note that there must exists $k \leq n$ for which this happens. Then necessarily $[R_k]_{kk} \neq 0$ (see below). As a consequence all diagonal entries of R_k are nonzero, and $\gamma_k = \sigma_k|[R_k]_{kk}| - [R_k]_{kk} = 0$. Thus the coefficient matrix, at such step k, must be equal to R_k and non singular. It follows that $\mathbf{x}_k = \mathbf{x}_0 + V_k\tilde{\mathbf{y}}$ is well defined, and by (r) the residual in $\mathbf{x}_k$ must be null ($h_{k+1,k} = 0!$), that is, $\mathbf{x}_k = A^{-1}\mathbf{b}$.

Let us show that $[R_k]_{kk} \neq 0$. Assume $[R_k]_{kk} = 0$. Then, since $R_k = Q_k^T H_k$, the matrix H_k must be singular. This fact with the identity $AV_k = V_kH_k$ ($h_{k+1,k} = 0!$) implies $AV_k\mathbf{e}_i = \sum_{j\neq i} AV_k\mathbf{e}_j$ for some i, and thus $\det(A) = 0$ (otherwise the columns of V_k should be linearly dependent which is impossible since $V_k^*V_k = I$). It follows that $[R_k]_{kk}$ must be nonzero.

Stagnation phenomenon

Assume $\mathbf{r}_0 \neq \mathbf{0}$ and set $\mathbf{v}_1 = \frac{\mathbf{r}_0}{\|\mathbf{r}_0\|}$. Assume that the value of h_{11} for which $\hat{\mathbf{v}}_2 = A\mathbf{v}_1 - h_{11}\mathbf{v}_1$ is orthogonal to $\mathbf{v}_1$ turns out to be zero, i.e.

$$h_{11} = [R_1]_{11} = (A\mathbf{v}_1, \mathbf{v}_1) = 0,$$

but $\hat{\mathbf{v}}_2 = A\mathbf{v}_1 - h_{11}\mathbf{v}_1 = A\mathbf{v}_1 \neq \mathbf{0}$, so that $h_{21} = \|\hat{\mathbf{v}}_2\| \neq 0$. Then the first step of GMRES is well defined ($R_1 + \gamma_1 = \sigma_1\sqrt{h_{21}^2} = \sigma_1 h_{21} \neq 0$), and $\|\mathbf{r}_1\|^2 = \frac{h_{21}^2}{0+h_{21}^2}\|\mathbf{r}_0\|^2 = \|\mathbf{r}_0\|^2$.

Assume also that $[R_2]_{22} = 0$ but $h_{32} = \|\hat{\mathbf{v}}_3\| \neq 0$ ($\hat{\mathbf{v}}_3 = A\mathbf{v}_2 - h_{12}\mathbf{v}_1 - h_{22}\mathbf{v}_2 \neq \mathbf{0}$). From the equality $R_2 = Q_2^T H_2$ it follows that

$$\begin{aligned} 0 &= |\det(\begin{bmatrix} \sigma_1 h_{21} & * \\ 0 & [R_2]_{22} = 0 \end{bmatrix})| = |\det(\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix})| \\ &= |\det(\begin{bmatrix} 0 & h_{12} \\ h_{21} & h_{22} \end{bmatrix})| = |h_{21}h_{12}| \end{aligned}$$

$\Rightarrow h_{12} = (A\mathbf{v}_2, \mathbf{v}_1) = (\frac{1}{\|A\mathbf{v}_1\|}A^2\mathbf{v}_1, \mathbf{v}_1) = 0$. Moreover,

$$R_2 + \begin{bmatrix} & \\ & \gamma_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 h_{21} & * \\ 0 & \sigma_2 h_{32} \end{bmatrix}$$

is non singular, so the second step of GMRES is well defined, and $\|\mathbf{r}_2\|^2 = \frac{h_{32}^2}{0+h_{32}^2}\|\mathbf{r}_1\|^2 = \|\mathbf{r}_1\|^2$. ...

Question: for any $k \leq n-1$ is it possible to introduce a residual $\mathbf{r}_0$ such that GMRES iterations yield

$$\begin{array}{l} 0 = h_{11} = \ldots = h_{1k} \\ 0 = [R_1]_{11} = \ldots = [R_k]_{kk} \\ 0 = (A\mathbf{r}_0, \mathbf{r}_0) = (A^2\mathbf{r}_0, \mathbf{r}_0) = \ldots = (A^k\mathbf{r}_0, \mathbf{r}_0) \\ h_{21}, \ldots, h_{k+1,k} \text{ all nonzero} \end{array}$$

? (for $k = n$ it cannot be possible because $h_{n+1,n}$ must be zero, or, equivalently, $A\mathbf{v}_n$ must be a linear combination of $\mathbf{v}_1, \ldots, \mathbf{v}_n$).

If yes, then we would have stagnation for k steps:

$$\|\mathbf{r}_k\| = \ldots = \|\mathbf{r}_1\| = \|\mathbf{r}_0\|.$$

For instance, if such situation occurs for $k = n-1$, then the method works well only in last step, in fact $\mathbf{r}_n$ must be necessarily null by the convergence in at most n steps property ($[R_n]_{nn}$ must be nonzero and $h_{n+1,n}$ must be zero).

Question: investigate stagnation that begins at the kth step ($k < n$): $[R_j]_{jj} \neq 0$ ($h_{j+1j} \neq 0$), $j = 1, \ldots, k-1$, $[R_k]_{kk} = 0$ ($\Rightarrow h_{k+1k} \neq 0$), $\|\mathbf{r}_k\| = \|r_{k-1}\| < \|\mathbf{r}_{k-2}\| < \ldots < \|\mathbf{r}_0\|$.