Solution of large linear systems by means of iterative methods

Alberto Tibaldi

October 22, 2013

Contents

Chapter 1

Conjugate gradient method

In these notes we are going to discuss the solution of large linear systems using advanced iterative methods. Therefore, the objective is to solve a matrix equation like:

$$\underline{\underline{M}}\,\underline{x} = \underline{h}$$

where $\underline{\underline{M}}$ is a square matrix, so $\underline{\underline{M}} \in \mathbb{R}^{n,n}$; then, $\underline{x}, \underline{h} \in \mathbb{R}^n$; this introduces some simplifications, which can be used to reduce the number of terms in the computations; however, this last hypothesis is not strictly mandatory.

The idea behind the method introduced in this section is to recall the convex paraboloid of the two-dimensional space: to have a situation such that

$$z = \underline{x}^{\mathrm{T}}\,\underline{\underline{M}}\,\underline{x}$$

has a unique branch, exactly like a parabola; indeed, everything is positive, owing to the fact that all the eigenvalues are positive. The fact to have a positive definite operator allow us to use the following idea: the solution of the linear system can be found minimizing a certain functional $J(\underline{x})$ of the solution $\underline{x}$: $\underline{x}$ represents the set of parameters for which J is minimum, in some sense. We are going to find a proper functional to be minimized.

The idea behind this minimization requires to have a matrix which is symmetric, positive-definite, in order to be hermitian and so to build $J(\underline{x})$ with the standard variational principle. To this aim, let us start from the initial system:

$$\underline{\underline{M}}\,\underline{x} = \underline{h}$$

it is possible to multiply both members times the adjoint of $\underline{\underline{M}}$, identified by $\underline{\underline{M}}^{\mathrm{a}}$; this is the transposed of the matrix with the elements equal to the complex conjugates of the elements of the previous matrix:

$$\underline{\underline{M}}^{\mathrm{a}} = (\underline{\underline{M}}^{*})^{\mathrm{T}}$$

so:

$$\underline{\underline{M}}^{\mathrm{a}}\,\underline{\underline{M}}\,\underline{x} = \underline{\underline{M}}^{\mathrm{a}}\,\underline{h}$$

Then, it is possible to define $\underline{\underline{A}}$ and $\underline{b}$ as:

$$\underline{\underline{A}} = \underline{\underline{M}}^{\mathrm{a}}\,\underline{\underline{M}}$$

and

$$\underline{b} = \underline{\underline{M}}^{\mathrm{a}}\,\underline{h}$$

so the system becomes:

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

this is called **associated linear system**. The difference between the previous system and this new system is that, in this case, $\underline{\underline{A}}$ is Hermitian, and therefore it is positive-definite and symmetric, so it will be possible to apply for it the variational principle. Let $\underline{r}$ be the residue of the **associated** linear system:

$$\underline{r} = \underline{b} - \underline{\underline{A}}\,\underline{x}$$

instead, let us define the residual of the original system $\overline{r} = \underline{h} - \underline{\underline{M}}\,\underline{x}$; it is possible to find the expression of the functional which has to be minimized, in order to solve the system, by applying some manipulation to the euclidean norm of this residual; by this way, we are finding the solution of the system as the vector such that the euclidean norm of the difference between $\underline{\underline{M}}\,\underline{x}$ and the known term is minimum:

$$
\begin{aligned}
||\overline{r}||_2 &= \langle \overline{r}, \overline{r} \rangle = \left\langle \underline{\underline{M}}\,\underline{x} - \underline{h}, \underline{\underline{M}}\,\underline{x} - \underline{h} \right\rangle = \\
&= \left\langle \underline{\underline{M}}\,\underline{x}, \underline{\underline{M}}\,\underline{x} \right\rangle - \left\langle \underline{h}, \underline{\underline{M}}\,\underline{x} \right\rangle - \left\langle \underline{\underline{M}}\,\underline{x}, \underline{h} \right\rangle + \langle \underline{h}, \underline{h} \rangle = \\
&= \left\langle \underline{\underline{M}}^{\mathrm{a}} \underline{\underline{M}}\,\underline{x}, \underline{x} \right\rangle - \left\langle \underline{\underline{M}}^{\mathrm{a}} \underline{h}, \underline{x} \right\rangle - \left\langle \underline{x}, \underline{\underline{M}}^{\mathrm{a}} \underline{h} \right\rangle + \langle \underline{h}, \underline{h} \rangle = \\
&= \left\langle \underline{\underline{A}}\,\underline{x}, \underline{x} \right\rangle - \langle \underline{b}, \underline{x} \rangle - \langle \underline{x}, \underline{b} \rangle + \langle \underline{h}, \underline{h} \rangle = \\
&\triangleq J(\underline{x}) + c
\end{aligned}
$$

Where:

$$
c \triangleq \langle \underline{h}, \underline{h} \rangle
$$

and:

$$
J(\underline{x}) = \left\langle \underline{\underline{A}}\,\underline{x}, \underline{x} \right\rangle - \langle \underline{b}, \underline{x} \rangle - \langle \underline{x}, \underline{b} \rangle
$$

Since $\underline{b}$ is related to $\underline{h}$ and $\underline{x}$ is real, it is possible to write:

$$
J(\underline{x}) = \left\langle \underline{\underline{A}}\,\underline{x}, \underline{x} \right\rangle - 2 \langle \underline{b}, \underline{x} \rangle
$$

We just proved that minimizing J corresponds to minimizing the euclidean norm of the residue, which is the difference of the $\underline{\underline{A}}$ matrix applied to the solution $\underline{x}$ with the known term. Therefore, thanks to the idea which have been previously discussed concerning the fact that the form is positive-definite symmetric. Then, it makes sense to compute the gradient of this functional, in order to find its global and unique minimum:

$$
\nabla J(\underline{x}) \Longrightarrow \frac{\mathrm{d}J(\underline{x})}{\mathrm{d}x_i} = \hat{\mathrm{e}}_i^{\mathrm{T}}\, \underline{\underline{A}}\,\underline{x} + \underline{x}^{\mathrm{T}} \underline{\underline{A}}\,\hat{\mathrm{e}}_i - 2\hat{\mathrm{e}}_i\, \underline{b}
$$

let us consider the term $\underline{x}^{\mathrm{T}}\,\underline{\underline{A}}\,\hat{\mathrm{e}}_i$: the left product $\underline{x}^{\mathrm{T}}\underline{\underline{A}}$ is a vector, and, of all its components, only the i-th one will survive, owing to the fact that $\hat{\mathrm{e}}_i$ is the unit vector of the canonical basis (all zeros except the i-th component). The same applies to the $\hat{\mathrm{e}}_i \underline{\underline{A}}\,\underline{x}$ term; but, since $\underline{\underline{A}}$ is symmetric, it is possible to say that:

$$
\hat{\mathrm{e}}_i^{\mathrm{T}}\, \underline{\underline{A}}\,\underline{x} = \underline{x}^{\mathrm{T}} \underline{\underline{A}}\,\hat{\mathrm{e}}_i
$$

indeed, thanks to the canonical basis unit vector, the components are not *mixed*. So:

$$(\nabla J(\underline{x}))_i = 2\hat{\mathrm{e}}_i^{\mathrm{T}}\,\underline{\underline{A}}\,\underline{x} - 2\hat{\mathrm{e}}_i\,\underline{b} \Longrightarrow \nabla J(\underline{x}) = 2\left(\underline{\underline{A}}\,\underline{x} - \underline{b}\right) = -2\underline{r}(\underline{x})$$

where $\underline{r}$ is the residual of the **associated** Hermitian linear system.

There is an alternative point of view for the interpretation of the functional $J(\underline{x})$: let us consider the following quadratic functional:

$$E(\underline{x}) = \left\langle \underline{\underline{A}}(\overline{\underline{x}} - \underline{x}), \overline{\underline{x}} - \underline{x} \right\rangle$$

where $\overline{\underline{x}}$ is the actual solution of the linear system; by defining the error functional with respect to the solution, $\underline{e}(\underline{x})$, as:

$$\underline{e}(\underline{x}) = \overline{\underline{x}} - \underline{x}$$

we have:

$$E(\underline{x}) = \left\langle \underline{\underline{A}}\underline{e}(\underline{x}), \underline{e}(\underline{x}) \right\rangle = ||\underline{e}(\underline{x})||^2_{\underline{\underline{A}}}$$

this quantity is a norm, because it satisfies all the properties of a norm. Moreover, there is a relation between the residual and the error; indeed, if we write the residual for the associate system, we have:

$$\underline{r}(\underline{x}) = \underline{b} - \underline{\underline{A}}\,\underline{x} = \underline{\underline{A}}\,\overline{\underline{x}} - \underline{\underline{A}}\,\underline{x} = \underline{\underline{A}}(\overline{\underline{x}} - \underline{x}) = \underline{\underline{A}}\underline{e}(\underline{x})$$

so, using the relationships just reported:

$$\left\langle \underline{\underline{A}}\,\underline{e}(\underline{x}), \underline{e}(\underline{x}) \right\rangle = \left\langle \underline{r}, \underline{\underline{A}}^{-1}\underline{r} \right\rangle$$

Since $\underline{\underline{A}}$ is positive-definite symmetric, also $\underline{\underline{A}}^{-1}$ is and therefore:

$$\left\langle \underline{r}, \underline{\underline{A}}^{-1}\underline{r} \right\rangle = ||\underline{r}||^2_{\underline{\underline{A}}^{-1}}$$

This show us that $E(\underline{x})$ is either the norm of the error of the norm of the residual, using two different definitions of norms (more precisely, with the dual norm); all these norms are simply weighted versions of the L^2 norm; indeed, if $\underline{\underline{A}} = \underline{\underline{I}}$, both these norms are reduced to a standard euclidean norm. In other words, what we have found is a relationship between the norm found starting from $\underline{\underline{A}}$ or $\underline{\underline{A}}^{-1}$.

Furthermore, it is possible to show that $E(\underline{x})$ and $J(\underline{x})$ are the same, in two different reference systems; indeed:

$$E(\underline{x}) = \left\langle \underline{\underline{A}}(\overline{x} - \underline{x}), \overline{x} - \underline{x} \right\rangle = \left\langle \underline{\underline{A}}\,\underline{x}, \underline{x} \right\rangle - 2\left\langle \underline{\underline{A}}\,\underline{x}, \overline{x} \right\rangle + \left\langle \underline{\underline{A}}\,\overline{x}, \overline{x} \right\rangle =$$
$$= \left\langle \underline{\underline{A}}\,\overline{x}, \overline{x} \right\rangle + J(\underline{x})$$

the first term is only a translation, while the second term is exactly $J(\underline{x})$. Therefore, we proved that finding the minimum of E is exactly the same of finding the minimum of J, so of the residual; indeed, it is like working on the domain instead that on the image of $\underline{\underline{M}}$, because:

$$\underline{r}(\underline{x}) = \underline{\underline{A}}\,\underline{e}(\underline{x})$$

Moreover, since the only difference between the two functional is a translation, which is a constant term, it can be found that:

$$\nabla J(\underline{x}) = \nabla E(\underline{x}) = -2\underline{r}(\underline{x})$$

Now, let $\underline{x}_k$ be a tentative of solution of our system at a certain k-th step; our objective is to move towards the real solution of the system; let $\underline{p}_k$ be the direction in which we want to move to get closer to the solution, and let α_k be the weight with which we want to approach the solution (in other words, α_k measures the width of the step); then, we can define the approximation of the solution at the $k+1$-th step as:

$$\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{p}_k$$

the term $\alpha_k \underline{p}_k$ is called **step**, because it represents the step from the k-th solution towards the $k+1$-th one.

As criterion for the selection of $\underline{p}_k$ and α_k, we desire to have that the functional evaluated in $\underline{x}_{k+1}$ is lower than the value of the functional evaluated at $\underline{x}_k$:

$$E(\underline{x}_{k+1}) < E(\underline{x}_k)$$

How can we find the α_k parameter? Let us see:

$$
\begin{aligned}
E(\underline{x}_{k+1}) &= \left\langle \underline{\underline{A}}(\overline{x} - \underline{x}_{k+1}), \overline{x} - \underline{x}_{k+1} \right\rangle = \\
&= \left\langle \underline{\underline{A}}\, \underline{x}_{k+1}, \underline{x}_{k+1} \right\rangle - 2 \left\langle \underline{x}_{k+1}, \underline{\underline{A}}\, \overline{x} \right\rangle + \left\langle \underline{\underline{A}}\, \overline{x}, \overline{x} \right\rangle = \\
&= \left\langle \underline{\underline{A}}\, \underline{x}_k, \underline{x}_k \right\rangle + 2\alpha_k \left\langle \underline{\underline{A}}\, \underline{x}_k, \underline{p}_k \right\rangle + \alpha_k^2 \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle + \\
&- 2 \left\langle \underline{x}_k, \underline{\underline{A}}\, \overline{x} \right\rangle - 2\alpha_k \left\langle \underline{p}_k, \underline{\underline{A}}\, \overline{x} \right\rangle + \left\langle \underline{\underline{A}}\, \overline{x}, \overline{x} \right\rangle = \\
&= \left\langle \underline{\underline{A}}(\overline{x} - \underline{x}_k), \overline{x} - \underline{x}_k \right\rangle - 2\alpha_k \left\langle \underline{\underline{A}}\, \underline{p}_k, \overline{x} - \underline{x}_k \right\rangle + \alpha_k^2 \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle = \\
&= \left\langle \underline{\underline{A}} \underline{e}(\underline{x}_k), \underline{e}(\underline{x}_k) \right\rangle - 2\alpha_k \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{e}(\underline{x}_k) \right\rangle + \alpha_k^2 \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle
\end{aligned}
$$

since the matrix $\underline{\underline{A}}$ is Hermitian, it is possible to move it from one side to another of the inner products without any additional operation on it. The objective is to find the α_k for which this E is minimum; using calculus, the minimum is a stationary point; for the considerations of the starting part of the chapter, we have only one stationary point. Therefore:

$$
\frac{\mathrm{d}E}{\mathrm{d}\alpha_k} = -2 \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{e}(\underline{x}_k) \right\rangle + 2\alpha_k \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle = 0
$$

this is satisfied if:

$$
\alpha_k = \frac{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{e}(\underline{x}_k) \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle} = \frac{\left\langle \underline{p}_k, \underline{r}(\underline{x}_k) \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle}
$$

This is the optimal step size.

It can be shown that, for all possible $\underline{p}_k$, the residual of the approxmation at the $k+1$ step is orthogonal to the one at the k step, if the step size α_k is the optimal one, calculated in the previous expression:

$$
\left\langle \underline{r}_{k+1}, \underline{r}_k \right\rangle \triangleq \left\langle \underline{r}(\underline{x}_{k+1}), \underline{r}(\underline{x}_k) \right\rangle = 0
$$

Let us prove this statement:

$$
\underline{r}_{k+1} = \underline{b} - \underline{\underline{A}}\, \underline{x}_{k+1} = \underline{b} - \underline{\underline{A}}\, \underline{x}_k - \alpha_k \underline{\underline{A}}\, \underline{p}_k = \underline{r}_k - \alpha_k \underline{\underline{A}}\, \underline{p}_k
$$

so:

$$\left\langle \underline{r}_k - \alpha_k \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle = \left\langle \underline{r}_k, \underline{p}_k \right\rangle - \alpha_k \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle =$$

$$= \left\langle \underline{r}_k, \underline{p}_k \right\rangle - \frac{\left\langle \underline{p}_k, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle} \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle = 0$$

Now, in order to add more ideas, let us manipulate the expression of the functional $E(\underline{x}_{k+1})$:

$$E(\underline{x}_{k+1}) = \langle \underline{\underline{A}}(\overline{x} - \underline{x}_{k+1}), \overline{x} - \underline{x}_{k+1} \rangle =$$

$$= \langle \underline{\underline{A}}(\overline{x} - \underline{x}_k), \overline{x} - \underline{x}_k \rangle - 2\alpha_k \left\langle \underline{\underline{A}}\, \underline{p}_k, \overline{x} - \underline{x}_k \right\rangle + \alpha_k^2 \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle =$$

$$= \langle \underline{\underline{A}}(\overline{x} - \underline{x}_k), \overline{x} - \underline{x}_k \rangle - 2\alpha_k \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{e}_k \right\rangle + \alpha_k^2 \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle$$

and:

$$\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{e}_k \right\rangle = \left\langle \underline{p}_k, \underline{r}_k \right\rangle$$

If we put the expression of the optimal step size instead of α_k, we obtain:

$$E(\underline{x}_{k+1}) = E(\underline{x}_k) - 2\frac{\left\langle \underline{p}_k, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle} \left\langle \underline{p}_k, \underline{r}_k \right\rangle + \left(\frac{\left\langle \underline{p}_k, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle} \right)^2 \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle =$$

$$= E(\underline{x}_k) - \frac{\left\langle \underline{p}_k, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle} = \left[\underline{\underline{I}} - \frac{\left\langle \underline{p}_k, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle E(\underline{x}_k)} \right] E(\underline{x}_k)$$

This last passage is not very formal since we are bringing vectors at the denominator, but since this is true for each component, we are considering this notation with this meaning. Let us further rewrite these expressions:

$$E(\underline{x}_k) = \left\langle \underline{\underline{A}}\, \underline{e}_k, \underline{e}_k \right\rangle = \langle \underline{r}_k, \underline{e}_k \rangle = \left\langle \underline{r}_k, \underline{\underline{A}}^{-1} \underline{r}_k \right\rangle$$

so, substituting:

$$E(\underline{x}_k) - \frac{\left\langle \underline{p}_k, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle} = \left[\underline{\underline{I}} - \frac{\left\langle \underline{p}_k, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle \left\langle \underline{r}_k, \underline{\underline{A}}^{-1} \underline{r}_k \right\rangle} \right] E(\underline{x}_k)$$

This is a recursive expression; therefore, it can be applied for every k. This equations states that for each $\underline{p}_k$ which is **not orthogonal** to the residue $\underline{r}_k$, it is possible to have a reduction of the functional.

1.1 Steepest descent method

Now, let us focus on $\underline{p}_k$: the most natural direction which someone can think is the one opposite to the **gradient of the functional**; indeed, the gradient is also the vector of maximum variation of a quantity. Therefore, the first method which will be realized is the **steepest descent method**, also called **gradient method**; this method consists of choosing:

$$\underline{p}_k = \underline{r}_k$$

if we use this approach, recalling the previous results, we obtain:

$$\alpha_k = \frac{||\underline{r}_k||^2}{\left\langle \underline{\underline{A}}\, \underline{r}_k, \underline{r}_k \right\rangle}$$

and:

$$E(\underline{x}_{k+1}) = \left[\underline{\underline{I}} - \frac{||\underline{r}_k||^4}{\left\langle \underline{\underline{A}}\, \underline{r}_k, \underline{r}_k \right\rangle \left\langle \underline{\underline{A}}^{-1} \underline{r}_k, \underline{r}_k \right\rangle} \right] E(\underline{x}_k)$$

By using the Kantorovich inequality, which is not trivial to be proved (and therefore it will not be discussed in this document), it is possible to show that the second term depends only on the conditioning number of the matrix $\kappa(\underline{\underline{A}})$; indeed:

$$E(\underline{x}_{k+1}) = \left[\frac{\kappa(\underline{\underline{A}}) - 1}{\kappa(\underline{\underline{A}}) + 1} \right]^k E(\underline{x}_0)$$

where:

$$\kappa(\underline{\underline{A}}) = \frac{\lambda_{\max}(\underline{\underline{A}})}{\lambda_{\min}(\underline{\underline{A}})}$$

and $\lambda_{\max}(\underline{\underline{A}})$, $\lambda_{\min}(\underline{\underline{A}})$ are the maximum and the minimum eigenvalues of the matrix $\underline{\underline{A}}$ respectively.

What we proved now is that the number of iterations required to achieve a good result in terms of norm of the residual (which has to be minimized) is strongly dependent on the conditioning number of the matrix of the system $\underline{\underline{A}}$; if this number equals 1 (best possible case), we can converge to the solution with one iteration; however, if κ is very large, it is necessary to perform a very high number of iterations to solve the system. To sum up, if the condition number of the matrix is small, this method is very good; otherwise, it would be better to find something else.

So, question is: is $\underline{p}_k = \underline{r}_k$ the best possible choice? Is it possible to do something better? Let us use:

$$\underline{p}_k = \underline{r}_k + \beta_k \underline{p}_{k-1}$$

What does it means to do this? Well, we proved that $\underline{r}_k$ and $\underline{p}_{k-1}$ are always orthogonal, if we use the optimal step size α_k; no information has been provided, right now, about β_k; so, which is the optimal value for β_k ? Obviously, the one which provide us with the largest reduction of the error.

Let us try to study the quantity:

$$\left\langle \underline{p}_k, \underline{r}_k \right\rangle = \langle \underline{r}_k, \underline{r}_k \rangle + \beta_k \left\langle \underline{p}_{k-1}, \underline{r}_k \right\rangle = \langle \underline{r}_k, \underline{r}_k \rangle$$

since we just recalled the fact that the residue of the k-th step is orthogonal to the $k-1$ direction; therefore, this quantity does not depend on β_k, and it can not be used to evaluate it.

On the other hand, let us recall:

$$E(\underline{x}_{k+1}) = \left[\underline{\underline{I}} - \frac{\left\langle \underline{p}_k, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle \left\langle \underline{r}_k, \underline{\underline{A}}^{-1} \underline{r}_k \right\rangle} \right] E(\underline{x}_k)$$

playing with β_k it is necessary to make the term $\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle$ as small as possible, in order to minimize the term inside the square brackets. So:

$$\left\langle \underline{\underline{A}}\,\underline{p}_k, \underline{p}_k \right\rangle = \left\langle \underline{\underline{A}}\,\underline{r}_k, \underline{r}_k \right\rangle + 2\beta_k \left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{r}_k \right\rangle + \beta_k^2 \left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{p}_{k-1} \right\rangle$$

Just like before, let us calculate the derivative of this term with respect to β_k:

$$\frac{\mathrm{d}\left\langle \underline{\underline{A}}\,\underline{p}_k, \underline{p}_k \right\rangle}{\mathrm{d}\beta_k} = 2\left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{r}_k \right\rangle + 2\beta_k \left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{p}_{k-1} \right\rangle = 0$$

so:

$$\beta_k = -\frac{\left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{p}_{k-1} \right\rangle}$$

The method based on this step is called **conjugate gradient method**. The reason why this name has been chosen is now motivated; since:

$$\left\langle \underline{\underline{A}}\,\underline{p}_k, \underline{p}_{k-1} \right\rangle = 0$$

we have that the descent direction is orthogonal in a product which depends also on $\underline{\underline{A}}$. Let us prove this:

$$\begin{aligned} &\left\langle \underline{\underline{A}}\,\underline{r}_k, \underline{p}_{k-1} \right\rangle + \beta_k \left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{p}_{k-1} \right\rangle = \\ &= \left\langle \underline{\underline{A}}\,\underline{r}_k, \underline{p}_{k-1} \right\rangle - \frac{\left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{p}_{k-1} \right\rangle} \left\langle \underline{\underline{A}}\,\underline{p}_{k-1}, \underline{p}_{k-1} \right\rangle = 0 \end{aligned}$$

This proves the previous statement.

This has interesting effects: supposing to be in $\mathbb{R}^2$, considering contour lines, they are ellipses:

Where the axes of these ellipses are related to the largest and smallest eigenvalues of the matrix $\underline{\underline{A}}$. If we arrived at a certain point through the step $\underline{p}_k$, then we have to choose a descent direction $\underline{r}_{k+1}$; what is done in the conjugate gradient method is to take an $\underline{\underline{A}}$-orthogonal direction, instead of a direction orthogonal with respect to the previous step.

The idea of the gradient method was based on the fact that if the contour lines are circles, the method is extremely efficient; instead, in this method,

we are taking into account, by including the effect of $\underline{\underline{A}}$ in the calculation of the step, the fact that the contour lines are ellipses; the gradient method fails when the contour lines are very different from circles (*i.e.* when the minimum and the maximum eigenvalues are very different), while in this case we are somehow taking this into account.

If we use this method, we obtain:

$$E(\underline{x}_{k+1} \leq 4 \left[\frac{\sqrt{\kappa(\underline{\underline{A}}) - 1}}{\sqrt{\kappa(\underline{\underline{A}}) + 1}} \right] E(\underline{x}_k)$$

this is much better than before!

It is possible to prove that the sequence of the residuals that we produce is orthogonal; indeed:

$$\underline{r}_{k+1} = \underline{b} - \underline{\underline{A}}\, \underline{x}_{k+1} = \underline{b} - \underline{\underline{A}}\, \underline{x}_k - \alpha_k \underline{\underline{A}}\, \underline{p}_k = \underline{r}_k - \alpha_k\, \underline{\underline{A}}\, \underline{p}_k$$

so, using this relationship:

$$\begin{aligned}
\left\langle \underline{r}_{k+1}, \underline{r}_k \right\rangle &= \langle \underline{r}_k, \underline{r}_k \rangle - \alpha_k \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{r}_k \right\rangle = \\
&= \langle \underline{r}_k, \underline{r}_k \rangle - \frac{\left\langle \underline{p}_k, \underline{r}_k \right\rangle}{\left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{p}_k \right\rangle} \left\langle \underline{\underline{A}}\, \underline{p}_k, \underline{r}_k \right\rangle = \\
&= \langle \underline{r}_k, \underline{r}_k \rangle - \left\langle \underline{p}_k, \underline{r}_k \right\rangle + \beta_k \left\langle \underline{\underline{A}}\, \underline{p}_k \cdot \underline{p}_{k-1} \right\rangle = \langle \underline{r}_k, \underline{r}_k \rangle - \left\langle \underline{p}_k, \underline{r}_k \right\rangle = \\
&= \langle \underline{r}_k, \underline{r}_k \rangle - \langle \underline{r}_k, \underline{r}_k \rangle + \beta_k \left\langle \underline{p}_{k-1}, \underline{r}_k \right\rangle = 0
\end{aligned}$$

in these steps we exploited the fact that $\underline{p}_k$ is $\underline{\underline{A}}$-orthogonal to $\underline{p}_{k-1}$, and the fact that $\left\langle \underline{r}_{k+1}, \underline{p}_k \right\rangle = 0$, as we proved in the previous part, concerning generic gradient methods (and this is valid also here).

Using similar steps it is possible to prove that the conjugate gradient method exhibits several properties:

- the residual of the following iteration equals the one of the present one (just proved):

$$\left\langle \underline{r}_{k+1}, \underline{r}_k \right\rangle = 0$$

- all the residuals are orthogonal:

$$\langle \underline{r}_k, \underline{r}_i \rangle = 0, \forall i = 0, 1, ..., k-1$$

- the $\underline{p}_k$ steps are 2-by-2 orthogonal:

$$\left\langle \underline{p}_k, \underline{\underline{A}}\, \underline{p}_i \right\rangle = 0, \forall i = 0, 1, ..., k-1$$

- The space spanned by the residuals is the same space spanned by the steps:

$$\text{span}\left(\{\underline{r}_0, \underline{r}_1, ..., \underline{r}_k\}\right) = \text{span}\left(\left\{\underline{p}_0, \underline{p}_1, ..., \underline{p}_k\right\}\right)$$

Let us focus on this last point: we stated that $\underline{p}_1$ is orthogonal to $\underline{\underline{A}}\, \underline{p}_0$, that $\underline{p}_2$ is orthogonal to $\underline{\underline{A}}\, \underline{p}_1$, and so on; it is possible to prove that:

$$\text{span}\left(\{\underline{r}_0, \underline{r}_1, ..., \underline{r}_k\}\right) = \text{span}\left(\left\{\underline{r}_0, \underline{\underline{A}}\, \underline{r}_0, \underline{\underline{A}}^2 \underline{r}_0, \underline{\underline{A}}^3 \underline{r}_0, ..., \underline{\underline{A}}^{k-1} \underline{r}_0\right\}\right) = \mathcal{K}_k\left\{\underline{\underline{A}}, \underline{r}_0\right\}$$

where $\mathcal{K}_k\left\{\underline{\underline{A}}, \underline{r}_0\right\}$ is the Krylov space generated by the span of $\underline{\underline{A}}$ and the initial $\underline{r}_0$.

All these properties have relevant consequences on the method; indeed it is possible to prove, thanks to these considerations, that:

$$E(\underline{x}_k) \leq E(\underline{x})$$

where:

$$\underline{x} \in \underline{x}_0 + \mathcal{K}_k\left\{\underline{\underline{A}}, \underline{r}_0\right\}$$

this means that $\underline{x}_k$ provides the minimum of $E(\underline{x}_k)$, in the Krylov subspace; then, at each iteration we increase the dimension of the Krylov space, and this means that we are searching for the minimum of a bigger space; after n iterations, the Krylov space equals $\mathbb{R}^n$; in other words, after at most n iterations, we have the exact solution, since the minimum becomes the minimum in $\mathbb{R}^n$.

Since all residuals are orthogonal, we are sure that the dimension of the space is increasing every time, at each iteration.

1.1.1 Final notes

If we want to use the conjugate gradient algorithm as an iterative method, it is necessary to introduce a stopping criterion, aimed at establishing when the desired tolerance is achieved. Let t be this value of tolerance; then, a criterion may be:

$$\frac{\left|\left|\underline{r}_{k+1}\right|\right|}{||\underline{b}||} \leq t$$

Indeed, it is possible to prove that this quantity is related to the error:

$$\frac{\left|\left|\underline{x}_k - \underline{x}_{k+1}\right|\right|}{||b||} \leq \kappa(\underline{\underline{A}}) \frac{\left|\left|\underline{r}_{k+1}\right|\right|}{||b||}$$

so, if the error can be controlled by this, if we require that this quantity is smaller than a certain constant, we know that the error, keeping into account the conditioning number κ, will be controlled by this.

Chapter 2

Notes on Krylov subspaces

In the previous chapter we introduced the idea of Krylov subspace, but we discussed any detail concerning it. In this chapter we are going to introduce some notes concerning these spaces and some additional ideas.

Let us consider the definition of a Krylov space:

$$\mathcal{K}_m \left\{ \underline{\underline{A}}, \underline{v} \right\} = \text{span} \left\{ \underline{v}, \underline{\underline{A}}\, \underline{v}, \underline{\underline{A}}^2 \underline{v}, ..., \underline{\underline{A}}^{m-1} \underline{v} \right\}$$

It is important to remark that the dimension of this space in general **is different from** m: indeed, it is not possible to know if these vectors are linearly dependent. For instance, if $\underline{\underline{A}}$ is a symmetric positive-definite matrix, it is possible to prove that:

$$\dim \left\{ \mathcal{K}_m \left\{ \underline{\underline{A}}, \underline{v} \right\} \right\} = m$$

this occurs, for instance, in conjugate gradient; however, we are going to generalize these ideas, in order to introduce the GMRES method.

Let μ_v be the **degree of the Krylov space**, which is the dimension of the largest Krylov space that we can get:

$$\mu_v \Longrightarrow \mathcal{K}_{\mu_v} = \text{span} \left\{ \underline{v}, \underline{\underline{A}}\, \underline{v}, \underline{\underline{A}}^2 \underline{v}, ..., \underline{\underline{A}}^{\mu_v - 1} \underline{v} \right\}$$

this means that if we add other vectors to this space starting from this construction, it will surely be linearly dependent on the others.

Basically, the Krylov space is the space of vectors $\underline{x} \in \mathbb{R}^n$ such that:

$$\underline{x} = p_{m-1}(\underline{\underline{A}}) \underline{v}$$

where $p_{m-1}(\underline{\underline{A}})$ is the polynomial of degree $m-1$ of the matrix $\underline{\underline{A}}$, applied to the vector $\underline{v}$. This idea can be exploited in order to find a basis for this Krylov space. Since the most attractive bases are the ones built with orthonormal elements, we may try to consider $\underline{v}_1$ parallel to $\underline{r}_0$, and to apply the Gram-Schmidt algorithm to find an orthonormal set of vectors. Surely, we can write:

$$\underline{v}_1 = \frac{\underline{r}_0}{||\underline{r}_0||}$$

so:

$$\langle \underline{v}_1, \underline{v}_1 \rangle = \left\langle \frac{\underline{r}_0}{||\underline{r}_0||}, \frac{\underline{r}_0}{||\underline{r}_0||} \right\rangle = \left\langle \frac{||\underline{r}_0||^2}{||\underline{r}_0||^2} \right\rangle = 1$$

then, we have to build a vector orthogonal to $\underline{v}_1$; to this aim, let us consider:

$$\underline{v} = \underline{\underline{A}}\,\underline{r}_0 - \left\langle \underline{\underline{A}}\,\underline{r}_0, \underline{v}_1 \right\rangle \underline{v}_1$$

then, this vector is surely orthogonal to $\underline{v}_1$; indeed:

$$\begin{aligned} \langle \underline{v}_1, \underline{v} \rangle &= \left\langle \underline{v}_1, \underline{\underline{A}}\,\underline{r}_0 - \left\langle \underline{\underline{A}}\,\underline{r}_0, \underline{v}_1 \right\rangle \underline{v}_1 \right\rangle = \left\langle \underline{v}_1, \underline{\underline{A}}\,\underline{r}_0 \right\rangle - \left\langle \underline{\underline{A}}\,\underline{r}_0, \underline{v}_1 \right\rangle \langle \underline{v}_1, \underline{v}_1 \rangle = \\ &= \left\langle \underline{v}_1, \underline{\underline{A}}\,\underline{r}_0 \right\rangle - \left\langle \underline{\underline{A}}\,\underline{r}_0, \underline{v}_1 \right\rangle = 0 \end{aligned}$$

otherwise, this vector should equal 0; this occurs in the case in which the maximum dimension of the Krylov space is 1. Now, let us define $\underline{v}_2$ as the normalized version of $\underline{v}$:

$$\underline{v}_2 = \frac{\underline{v}}{||\underline{v}||} = \frac{\underline{\underline{A}}\,\underline{r}_0 - \left\langle \underline{\underline{A}}\,\underline{r}_0, \underline{v}_1 \right\rangle \underline{v}_1}{\left|\left| \underline{\underline{A}}\,\underline{r}_0 - \left\langle \underline{\underline{A}}\,\underline{r}_0, \underline{v}_1 \right\rangle \underline{v}_1 \right|\right|}$$

Then, it is possible to define a new $\underline{v}$ removing from $\underline{\underline{A}}^2\underline{r}_0$ the previous vectors projected on $\underline{\underline{A}}^2\underline{v}_0$, and then $\underline{v}_3$ normalizing this:

$$\underline{v}_3 = \frac{\underline{\underline{A}}^2\underline{r}_0 - \left\langle \underline{\underline{A}}^2\underline{r}_0, \underline{v}_1 \right\rangle \underline{v}_1 - \left\langle \underline{\underline{A}}^2\underline{r}_0, \underline{v}_2 \right\rangle \underline{v}_2}{\left|\left| \underline{\underline{A}}^2\underline{r}_0 - \left\langle \underline{\underline{A}}^2\underline{r}_0, \underline{v}_1 \right\rangle \underline{v}_1 - \left\langle \underline{\underline{A}}^2\underline{r}_0, \underline{v}_2 \right\rangle \underline{v}_2 \right|\right|}$$

and so on.

This procedure is very similar to the Gram-Schmidt algorithm.

At most, this procedure has to be applied for all the m vectors of the set; actually, this has to be applied until one of these $\underline{v}_i$ vectors become zero; indeed, up to this point, all vectors are surely linearly independent on the previous ones.

Let us consider the following orthonormalization procedure:

```
b=norm(r0)
v(1)=r0/b

for j=1:m
for i=1:n
h(i,j) = innerproduct(A v(j),v(i))
end
w(j) = A v(j) - sum(h(i,j)*v(i),i=1:j)
h(j+1,j)=norm(w(j))
if h(j+1,j)==0 break
v(j+1)=w(j)/h(j+1,j)
end
```

This algorithm is not Gram-Schmidt: indeed, $\underline{\underline{A}}^2$, $\underline{\underline{A}}^3$ and so on are never evaluated; instead, we are considering the following space:

$$\text{span}\left\{\underline{v}_1, \underline{v}_2, \underline{v}_3, ...\right\} = \text{span}\left\{\underline{v}_1, \underline{\underline{A}}\,\underline{v}_1, \underline{\underline{A}}\,\underline{v}_2, ...\right\}$$

so, $\underline{v}_2 = \underline{\underline{A}}\,\underline{v}_1$, $\underline{v}_3 = \underline{\underline{A}}\,\underline{v}_2$, and so on.

We are going to prove that the space spanned by these vectors actually equals the Krylov subspace. To this aim, it is necessary to explain the algorithm, and to verify that each $\underline{v}_j$ can be written as:

$$\underline{v}_j = p_{j-1}(\underline{\underline{A}})\underline{r}_0, \quad \forall j = 1...m$$

Let β be:

$$\beta = ||\underline{r}_0||$$

just like in the algorithm; then, the statement is trivially verified for $\underline{v}_1$:

$$\underline{v}_1 = \frac{1}{\beta}\hat{\mathrm{r}}_0 = \frac{1}{\beta}\,\underline{\underline{A}}^0\,\underline{r}_0$$

In order to prove the statement $\forall j$, we want to apply an induction procedure; let us guess that induction holds for $\underline{v}_i$:

$$\underline{v}_i = p_{i-1}(\underline{\underline{A}})\,\underline{r}_0, i = 1, ..., j-1$$

this is the induction hypothesis; therefore, following the algorithm, let:

$$h_{ij} = \left\langle \underline{\underline{A}}\,\underline{v}_j, \underline{v}_i \right\rangle$$

so, let us define $\underline{w}_j$ as the $j+1$-th non-normalized vector:

$$\underline{w}_j = \underline{\underline{A}}\,\underline{v}_j - \sum_{i=1}^{j} h_{i,j}\,\underline{v}_i$$

this means that $\underline{w}_j$ is defined as $\underline{\underline{A}}$, applied to $\underline{v}_j$, where we subtract all the vectors with components equal to $\left\langle \underline{\underline{A}}\,\underline{v}_j, \underline{v}_i \right\rangle$; this is the main difference from Gram-Schmidt, since in all these operations, only $\underline{\underline{A}}$ is applied, not $\underline{\underline{A}}^i$.

Now, we obtained the next vector, but we still have to normalize it; therefore, it is possible to define:

$$h_{j+1,j} = \left|\left|\underline{w}_j\right|\right|$$

so:

$$\underline{v}_{j+1} = \frac{\underline{w}_j}{h_{j+1,j}}$$

if we consider instead of j the index $j-1$, this transforms to:

$$\underline{v}_j = \frac{\underline{w}_{j-1}}{h_{j,j-1}}$$

which means that:

$$\underline{w}_{j-1} = \underline{v}_j h_{j,j-1}$$

now, substituting $j \to j-1$ also in the definition of $\underline{w}_j$, we can obtain:

$$\underline{w}_{j-1} = \underline{\underline{A}}\,\underline{v}_{j-1} - \sum_{i=1}^{j-1} h_{i,j-1}\underline{v}_i$$

but:

$$\underline{w}_{j-1} = h_{j,j-1}\underline{v}_j$$

so,

$$h_{j,j-1}\underline{v}_j = \underline{\underline{A}}\,\underline{v}_{j-1} - \sum_{i=1}^{j-1} h_{i,j-1}\underline{v}_i$$

so, substituting the induction hypothesis in the two members of the right-hand side, it is possible to obtain:

$$h_{j,j-1}\underline{v}_j = \underline{\underline{A}}\,p_{j-2}(\underline{\underline{A}})\,\underline{r}_0 - \sum_{i=1}^{j-2} h_{i,j-1}p_{i-1}(\underline{\underline{A}})\,\underline{r}_0$$

since in the left-hand side we have $\underline{v}_j$ times some coefficient, and in the right-hand side we have a polynomial in $\underline{\underline{A}}$ of degree p_{j-1}, we proved that $\underline{v}_j$ is a polynomial of degree $j-1$ in $\underline{\underline{A}}$, applied to $\underline{r}_0$.

This means that the basis built using this algorithm actually is a basis for the Krylov subspace. This is the **Arnoldi's method**.

Now, since we verified that these vectors are a basis, it is necessary to understand better how to work with them, to find some relationships between these vectors; indeed, what is the vector $\underline{\underline{A}}\,\underline{v}_m$? In order to represent a generic vector, let us define the matrix $\underline{\underline{V}}_m$ as:

$$\underline{\underline{V}}_m = \begin{bmatrix}\underline{v}_1 & \underline{v}_2 & \dots & \underline{v}_m\end{bmatrix}$$

then, it is possible to write $\underline{\underline{A}}\,\underline{m}$ as the matrix $\underline{\underline{V}}_m$ multiplied times $\underline{\underline{H}}$, where $\underline{\underline{H}}$ is the matrix which has as elements the h_{ij} which we were discussing during the definition of the Arnoldi method. During this definition, in the algorithm, we wrote:

$$\underline{w}_j = \underline{\underline{A}}\,\underline{v}_j - \sum_{i=1}^{j} h_{i,j}\underline{v}_i$$

so, if $m = j$, isolating $\underline{v}_j$:

$$\underline{\underline{A}}\,\underline{v}_m = \sum_{i=1}^{m} h_{i,m}\underline{v}_i + \underline{w}_m$$

by defining $\underline{H}_m$ as $\underline{\underline{H}}(:, m)$, which means considering the entire m-th column of the matrix $\underline{\underline{H}}$, this last expression can be written as:

$$\underline{\underline{A}}\,\underline{v}_m = \underline{\underline{V}}_m\,\underline{H}_m + \underline{w}_m$$

but, in the Arnoldi method, we also wrote:

$$\underline{w}_m = h_{m+1,m}\underline{v}_{m+1}$$

so:

$$\underline{\underline{A}}\,\underline{v}_m = \underline{\underline{V}}_m\,\underline{H}_m + h_{m+1,m}\underline{v}_{m+1}$$

We have that $\underline{\underline{H}}_m \in \mathbb{R}^{m,m}$; however, starting from $\underline{\underline{H}}_m$, it is possible to modify this matrix, in order to obtain a Hessenberg matrix $\overline{\underline{\underline{H}}}_m$. A Hessenberg matrix is a matrix which is *almost triangular*, since including the terms $h_{m+1,m}$ means to add the lower diagonal of the matrix; this diagonal goes from the element $h_{2,1}$, to the element $h_{m+1,m}$; this means that, to obtain this matrix, it is necessary to introduce all these terms on the lower diagonal and to add another row to the matrix, which have only zeros except for the last term, which is $h_{m+1,m}$, which equals $||\underline{w}_m||$, as we discussed in the introduction to the Arnoldi's method. So, at this point:

$$\overline{\underline{\underline{H}}}_m = \begin{bmatrix} & & \\ & \underline{\underline{H}}_m & \\ & & \\ \hline 0 & 0 \quad 0 \quad ... & ||\underline{w}_m|| \end{bmatrix}$$

So,

$$\overline{\underline{\underline{H}}}_m \in \mathbb{R}^{m+1,m}$$

therefore, this matrix is not a Hessenberg matrix, since Hessenberg matrices are square. Using this definition, it is possible to write more compactly the last expression, putting all the columns of $\underline{\underline{H}}$ instead that only the m-th one:

$$\underline{\underline{A}}\,\underline{\underline{V}}_m = \underline{\underline{V}}_{m+1}\overline{\underline{\underline{H}}}_m$$

It is possible now to multiply both these terms times $\underline{\underline{V}}_m^{\mathrm{T}}$:

$$\underline{\underline{V}}_m^{\mathrm{T}} \underline{\underline{A}}\, \underline{v}_m = \underline{\underline{V}}_m^{\mathrm{T}} \underline{\underline{V}}_{m+1} \overline{\underline{\underline{H}}}_m$$

Let us work on $\underline{\underline{V}}_m^{\mathrm{T}} \underline{\underline{V}}_{m+1}$: the left matrix is the transposed of $\underline{\underline{V}}_m$, which, we remark, is the matrix which has as columns the vectors $\underline{v}_i$, defined with the Arnoldi's method; since we proved that these vectors are equivalent to the ones built by means of the Gram-Schmidt orthonormalization, these columns are 2-by-2 orthonormal. Moreover, $\underline{\underline{V}}_m^{\mathrm{T}}$ is a $m \times m$ matrix, while $\underline{\underline{V}}_{m+1}$ is a $(m+1) \times (m+1)$ matrix; therefore, their product will be a $m \times (m+1)$ matrix. Moreover, owing to the orthonormality of the columns, it is possible to say that the left $m \times m$ part of the matrix will be the identity. Then, the product of the $\underline{\underline{V}}_m^{\mathrm{T}}$ matrix with the last column returns a column of zeros, since $\underline{v}_{m+1}$ is orthonormal to all the rows of $\underline{V}_m^{\mathrm{T}}$; so:

$$\underline{\underline{V}}_m^{\mathrm{T}} \underline{\underline{V}}_{m+1} = \left[\begin{array}{ccc|c} & & & 0 \\ & \underline{\underline{I}} & & 0 \\ & & & 0 \\ & & & 0 \\ & & & 0 \\ & & & 0 \end{array}\right]$$

this matrix has to be multiplied times $\overline{\underline{\underline{H}}}_m$; the effect of the last column of zeros is to *kill* the last row of $\overline{\underline{\underline{H}}}_m$; the result is:

$$\underline{\underline{V}}_m^{\mathrm{T}} \underline{\underline{V}}_{m+1} = \left[\begin{array}{ccc|c} & & & 0 \\ & \underline{\underline{I}} & & 0 \\ & & & 0 \\ & & & 0 \\ & & & 0 \\ & & & 0 \end{array}\right] \overline{\underline{\underline{H}}}_m = \underline{\underline{H}}_m$$

therefore:

$$\underline{\underline{V}}_m^{\mathrm{T}} \underline{\underline{V}}_{m+1} \overline{\underline{\underline{H}}}_m = \underline{\underline{H}}_m$$

and so, finally:

$$\underline{\underline{V}}_m^{\mathrm{T}} \underline{\underline{A}}\, \underline{v}_m = \underline{\underline{H}}_m$$

This is all the theory concerning the Krylov spaces which is needed in order to formulate the GMRES method.

Chapter 3

Generalized Minimum RESidual method (GMRES)

3.1 Algorithm

GMRES is a generalization of the conjugate gradient method, and for this method we will consider a generic matrix $\underline{\underline{A}}$, without having any hypothesis on its symmetry or positive-definition; given

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

the only requirement for $\underline{\underline{A}}$ is to be non-singular.

Given a guess solution $\underline{x}_0$, we want to build $\underline{x}_m$ such that:

$$\underline{x}_m \in \underline{x}_0 + \mathcal{K}_m\left\{\underline{\underline{A}}, \underline{r}_0\right\}$$

asking as objective the minimization of the norm of the residual:

$$\min_{\underline{x}_m \in \underline{x}_0 + \mathcal{K}_m\left\{\underline{\underline{A}}, \underline{r}_0\right\}} \left|\left|\underline{b} - \underline{\underline{A}}\,\underline{x}_m\right|\right|$$

In the previous section we have found a base for the Krylov subspaces; let $\underline{y}_m$ be a vector with some elements; a generic vector in the Krylov subspace will be $\underline{\underline{V}}_m\,\underline{y}_m$; so:

$$\underline{x}_m = \underline{x}_0 + \underline{\underline{V}}_m \underline{y}_m$$

therefore, it is possible to calculate the residual corresponding to this $\underline{y}_m$:

$$R(\underline{y}_m) = \left|\left|\underline{b} - \underline{\underline{A}}\,\underline{x}_m\right|\right|^2 = \left|\left|\underline{b} - \underline{\underline{A}}\,\underline{x}_0 - \underline{\underline{A}}\,\underline{\underline{V}}_m\underline{y}_m\right|\right|^2 = \left|\left|\underline{r}_0 - \underline{\underline{A}}\,\underline{\underline{V}}_m\underline{y}_m\right|\right|^2$$

where it is possible to define a relationship between the residual at the zeroth-iteration and the first vector of the basis of the Krylov subspace:

$$\underline{r}_0 = \beta\underline{v}_1$$

so:

$$\left|\left|\underline{r}_0 - \underline{\underline{A}}\,\underline{\underline{V}}_m\underline{y}_m\right|\right|^2 = \left|\left|\beta\underline{v}_1 - \underline{\underline{A}}\,\underline{\underline{V}}_m\underline{y}_m\right|\right|^2$$

but we proved that:

$$\underline{\underline{A}}\,\underline{\underline{V}}_m = \underline{\underline{V}}_{m+1}\overline{\underline{\underline{H}}}_m$$

therefore:

$$= \left|\left|\beta\underline{v}_1 - \underline{\underline{V}}_{m+1}\overline{\underline{\underline{H}}}_m\underline{y}_m\right|\right|^2$$

but:

$$\underline{v}_1 = \underline{\underline{V}}_{m+1}\hat{\mathrm{e}}_1$$

where $\hat{\mathrm{e}}_1$ is the first vector of the canonical basis; so:

$$\left|\left|\beta\underline{\underline{V}}_{m+1}\hat{\mathrm{e}}_1 - \underline{\underline{V}}_{m+1}\overline{\underline{\underline{H}}}_m\underline{y}_m\right|\right|^2 = \left|\left|\underline{\underline{V}}_{m+1}\left(\beta\hat{\mathrm{e}}_1 - \overline{\underline{\underline{H}}}_m\underline{y}_m\right)\right|\right|^2$$

since $\underline{\underline{V}}_{m+1}$ is an orthonormal matrix, it is possible to use the following property of norms:

$$\left|\left|\underline{\underline{V}}_{m+1}\left(\beta\hat{\mathrm{e}}_1 - \overline{\underline{\underline{H}}}_m\underline{y}_m\right)\right|\right|^2 = \left|\left|\beta\hat{\mathrm{e}}_1 - \overline{\underline{\underline{H}}}_m\underline{y}_m\right|\right|^2$$

The application of an orthonomal matrix to a vector does not change the norm of this vector (but only its direction). Let us prove this fact: given a vector $\underline{x}$ and a vector $\underline{y} = \underline{\underline{Q}}\,\underline{x}$, where $\underline{\underline{Q}}$ is an orthonormal matrix, it is known from matrix theory that:

$$\underline{\underline{Q}}^{-1} = \underline{\underline{Q}}^{\mathrm{T}}$$

so:

$$\underline{\underline{Q}}^{\mathrm{T}}\,\underline{\underline{Q}} = \underline{\underline{Q}}\,\underline{\underline{Q}}^{\mathrm{T}} = \underline{\underline{I}}$$

therefore:

$$||\underline{x}||^2 = \underline{x}^{\mathrm{T}}\underline{x} = \underline{y}^{\mathrm{T}}\,\underline{\underline{Q}}\,\underline{\underline{Q}}^{\mathrm{T}}\underline{y} = \underline{y}^{\mathrm{T}}\underline{y} = \left|\left|\underline{y}\right|\right|^2$$

this proves that orthonormal matrices are associated to pure rotations (isometries) of the vector, not to modifications of its length.

Now, let us apply the QR factorization to the matrix $\overline{\underline{\underline{H}}}_m$; from this decomposition we obtain an orthonomal matrix $\underline{\underline{Q}}$ and an upper triangular matrix $\underline{\underline{R}}$:

$$\overline{\underline{\underline{H}}}_m = \underline{\underline{Q}}\,\underline{\underline{R}}$$

where:

$$\underline{\underline{Q}} \in \mathbb{R}^{m+1,m+1}$$

and:

$$\underline{\underline{R}} \in \mathbb{R}^{m+1,m}$$

where $\underline{\underline{R}}$ is an upper triangular matrix and the last row has only zeros.

Now, by computing this factorization, it is possible to obtain:

$$\left|\left|\beta\hat{\mathrm{e}}_1 - \overline{\underline{\underline{H}}}_m\underline{y}_m\right|\right|^2 = \left|\left|\beta\hat{\mathrm{e}}_1 - \underline{\underline{Q}}\,\underline{\underline{R}}\,\underline{y}_m\right|\right|^2$$

The euclidean norm of a vector is the sum of the squares of its components; since $\underline{y}_m \in \mathbb{R}^m$, and $\underline{\underline{Q}}\,\underline{\underline{R}} \in \mathbb{R}^{m+1,m}$, $\underline{\underline{Q}}\,\underline{\underline{R}}\,\underline{y}_m \in \mathbb{R}^{m+1}$; so, this vector can be decomposed in two parts: the first m coefficients of this vector are called $\underline{r}$; therefore, $\underline{r} \in \mathbb{R}^m$; then, the last component is called r_{m+1}; therefore:

$$\left|\left|\beta\hat{\mathrm{e}}_1 - \underline{\underline{Q}}\,\underline{\underline{R}}\,\underline{y}_m\right|\right|^2 = ||\underline{r}||^2 + |r_{m+1}|^2$$

to use this idea, let us apply the following trick:

$$\left|\left|\beta\hat{\mathrm{e}}_1 - \underline{\underline{Q}}\,\underline{\underline{R}}\,\underline{y}_m\right|\right|^2 = \left|\left|\underline{\underline{Q}}\left(\underline{\underline{Q}}^{\mathrm{T}}\beta\hat{\mathrm{e}}_1 - \underline{\underline{R}}\,\underline{y}_m\right)\right|\right|^2 = \left|\left|\underline{\underline{Q}}^{\mathrm{T}}\beta\hat{\mathrm{e}}_1 - \underline{\underline{R}}\,\underline{y}_m\right|\right|^2$$

Owing to the fact that $\underline{\underline{R}}$ has the last row equal to zero, it is possible to decouple the following terms:

$$\left|\left|\underline{\underline{Q}}^{\mathrm{T}}\beta\hat{\mathrm{e}}_1 - \underline{\underline{R}}\,\underline{y}_m\right|\right|^2 = \left|\left|\underline{\underline{Q}}^{\mathrm{T}}\beta\hat{\mathrm{e}}_1\Big|_{1\ldots m} - \underline{\underline{R}}\,\underline{y}_m\right|\right|^2 + \left|\left|\underline{\underline{Q}}^{\mathrm{T}}\beta\hat{\mathrm{e}}_1\Big|_{m+1}\right|\right|^2$$

So, we proved that:

$$R(\underline{y}_m) = \left|\left|\underline{\underline{Q}}^{\mathrm{T}}\beta\hat{\mathrm{e}}_1\Big|_{1\ldots m} - \underline{\underline{R}}\,\underline{y}_m\right|\right|^2 + \left|\left|\underline{\underline{Q}}^{\mathrm{T}}\beta\hat{\mathrm{e}}_1\Big|_{m+1}\right|\right|^2$$

it is possible to play with $\underline{y}_m$ in order to minimize this residual, but we can not vanish the $\left|\left|\underline{\underline{Q}}^{\mathrm{T}}\beta\hat{\mathrm{e}}_1\Big|_{m+1}\right|\right|^2$ component; the best that we can obtain is:

$$\underline{\underline{R}}\,\underline{y}_m = \beta\underline{\underline{Q}}^{\mathrm{T}}\hat{\mathrm{e}}_1\Big|_{1\ldots m}$$

This is a linear system, where the unknown is $\underline{y}_m$, but this is very easy to be solved, since $\underline{\underline{R}}$ is an upper triangular matrix. Let us define the following vectors:

$$\underline{\gamma}_{1\ldots m} = \beta\underline{\underline{Q}}^{\mathrm{T}}\hat{\mathrm{e}}_1$$

and let $\underline{\gamma}_{m+1}$ be the last component.

3.2 Efficient QR factorization: Givens matrices

How can we compute the $\underline{\underline{Q}}\,\underline{\underline{R}}$ factorization? The key ingredients to perform these factorizations are **Givens matrices**; in $\mathbb{R}^2$, the Givens matrix $\underline{\underline{G}}(\vartheta)$ is a matrix of type:

$$\underline{\underline{G}}(\vartheta) = \begin{bmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{bmatrix}$$

by applying $\underline{\underline{G}}$ to a vector in $\mathbb{R}^2$, we obtain the same vector, rotated of ϑ in clockwise sense (or a rotation of the system in counterclockwise sense).

The matrix $\underline{\underline{G}}$ is orthogonal, and its inverse is simply the inverse rotation. Why have we defined this matrix, and how can we use it? Well, let us consider an unknown vector $\underline{x}$ as:

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

now, let us apply $\underline{\underline{G}}$ to this vector, and let us consider, as known term, a vector which has a vanishing second component:

$$\underline{\underline{G}}\,\underline{x} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$$

where:

$$\alpha = \sqrt{x_1^2 + x_2^2}$$

this means that the application of $\underline{\underline{G}}$ has not changed the length of $\underline{x}$, but the rotation is vanishing the second component. Which is the ϑ which puts the x axis parallel to the vector?

$$\begin{bmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}$$

so:

$$\begin{cases} x_1 \cos\vartheta + x_2 \sin\vartheta = \alpha \\ -x_1 \sin\vartheta + x_2 \cos\vartheta = 0 \end{cases}$$

From the second equation:

$$x_1 = x_2 \frac{\cos\vartheta}{\sin\vartheta}$$

substituting in the first one:

$$x_2 \frac{\cos\vartheta}{\sin\vartheta} \cos\vartheta + x_2 \sin\vartheta = \alpha$$

which becomes:

$$x_2 \left(\cos^2\vartheta + \sin^2\vartheta\right) = \alpha \sin\vartheta$$

so:

$$\sin\vartheta = \frac{x_2}{\alpha}$$

similarly we can show that:

$$\cos\vartheta = \frac{x_1}{\alpha}$$

This was the $\mathbb{R}^2$ case; now, let us consider a more general case: $\underline{\underline{G}} \in \mathbb{R}^{p,p}$. In this case, the matrix $\underline{\underline{G}}$ is the identity matrix from the components $1...p-2$, and then for the $p-1$ to p components there is the 2×2 Givens matrix; the two ending rows and columns, except for the 2×2 block, are filled with zeros. Considering this matrix, the objective may be to obtain:

$$\underline{\underline{G}}\,\underline{x} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ 0 \end{bmatrix}$$

every number should be different from zero, except the last component:

$$\underline{\underline{G}} = \left[\begin{array}{cc|cc} & & & \\ & \underline{\underline{I}} & & \\ & & & \\ \hline & & \cos\vartheta & \sin\vartheta \\ 0 & & -\sin\vartheta & \cos\vartheta \end{array}\right]$$

so, this is a rotation in the plane of the p-th and the $p-1$-th components, and the identity out of it. So:

$$\begin{cases} x_{p-1}\cos\vartheta + x_p \sin\vartheta = \sqrt{x_{p+1}^2 + x_p^2} \\ -x_{p-1}\sin\vartheta + x_p \cos\vartheta = 0 \end{cases}$$

and, since the remaining part of the system is the identity, the remaining components of the solution, from 1 to $p-2$, are equal to the known term.

Why are we focusing on this idea? Taking one coefficient and erase it? Well, the idea is to apply this to the Hessenberg matrix: indeed, if we want

to perform a QR factorization on the Hessenberg matrix, the $\underline{\underline{R}}$ matrix has to be triangular; however, Hessenberg matrix is almost triangular, so it is just necessary to modify it a little, in order to remove these terms and modify their neighbors; this can be done using Givens matrices, and, by producing this $\underline{\underline{R}}$, the $\underline{\underline{Q}}$ matrix should come as well. The matrix $\underline{\underline{\overline{H}}}_m$ has a form like:

$$\underline{\underline{\overline{H}}}_m = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & h_{1,4} & h_{1,5} & ... \\ h_{2,1} & h_{2,2} & h_{2,3} & h_{2,4} & h_{2,5} & ... \\ 0 & h_{3,2} & h_{3,3} & h_{3,4} & h_{3,5} & ... \\ 0 & 0 & h_{3,4} & h_{4,4} & h_{4,5} & ... \\ 0 & 0 & 0 & h_{5,4} & h_{5,5} & ... \\ 0 & 0 & 0 & 0 & h_{6,5} & ... \end{bmatrix}$$

In order to obtain an upper triangular matrix starting from this one, it is necessary to multiply it times a certain $\underline{\underline{G}}_i$, which is a matrix modifying only two rows of $\underline{\underline{\overline{H}}}_m$, in order to erase $h_{2,1}$ ahd modifying all the remaining components of the first two lines. This matrix can be called $\underline{\underline{G}}_{2,1}$, since it acts in order to erase $h_{2,1}$, and it equals the identity in the remaining lines: this is basically a 2×2 matrix; then, we will have a Hessenberg matrix with $h_{2,1} = 0$; then, a 3×3 matrix, called $\underline{\underline{G}}_{3,2}$ is applied to erase the element $h_{3,2}$.

Every time that we apply this operation, we are basically peforming a matrix product; so, we have something like:

$$\underline{\underline{\overline{H}}}_m = \underline{\underline{G}}_{2,1} \underline{\underline{\overline{H}}}^{(1)}_m = \underline{\underline{G}}_{3,2} \underline{\underline{G}}_{2,1} \underline{\underline{\overline{H}}}^{(2)}_m = ...$$

the product of all the $\underline{\underline{G}}_i$ is a matrix $\underline{\underline{Q}}_m$, which is proved to be **orthogonal**; indeed, it comes from the product of orthonormal matrices, and therefore it is orthonormal too. Furthermore, if the operation is completed, $\underline{\underline{\overline{H}}}_m$ is transformed in a triangular matrix, $\underline{\underline{R}}$; so, we obtained:

$$\underline{\underline{\overline{H}}}_m = \underline{\underline{Q}}_m \underline{\underline{R}}$$

this is the QR factorization of the Hessenberg matrix, obtained in an efficient way, exploiting the properties of the matrix and of the Givens rotation matrices.

3.3 Additional remarks

GMRES is an iterative method, and therefore the way in which this method has been explained may be misleading: the correct way to proceed is not to build the entire Hessenberg matrix and then modify it and check the residual: since the method is iterative, it is a good idea to check the residual every time, in order to understand if it is possible to stop the algorithm and to obtain a good accuracy. The matrix $\overline{\underline{\underline{H}}}_m$ is not therefore available after one iteration: it is build step by step.

After the first step, we have $\overline{\underline{\underline{H}}}_1$, which is just a vector of two components; therefore, the QR factorization for this matrix is just the 2×2 Givens matrix:

$$\begin{bmatrix} \cos\vartheta_1 & \sin\vartheta_1 \\ -\sin\vartheta_1 & \cos\vartheta_1 \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{21} \end{bmatrix} = \begin{bmatrix} h_{11}^{(1)} \\ 0 \end{bmatrix}$$

After one iteration, this is the result; then, it is possible to evaluate the residual as:

$$||\underline{r}_1|| = -\beta \sin\vartheta$$

where $\beta = ||\underline{r}_0||$.

If the solution after one iteration is not accurate enough, it is possible to go on, and to build the second term; in this case, it will be:

$$\overline{\underline{\underline{H}}}_2^{(0)} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ 0 & h_{32} \end{bmatrix}$$

actually, exploiting orthonormality and so on, it is possible to multiply this times the matrix:

$$\begin{bmatrix} \cos\vartheta_1 & \sin\vartheta_1 & 0 \\ -\sin\vartheta_1 & \cos\vartheta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

this is the extension of the matrix found in the previous step, with the identity at the end; by this way,

$$\overline{\underline{\underline{H}}}_2^{(1)} = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} \\ 0 & h_{22}^{(1)} \\ 0 & h_{32} \end{bmatrix}$$

this matrix acts only on the first two rows; here:

$$h_{12}^{(1)} = h_{12}\cos\vartheta_1 + h_{22}\sin\vartheta_1$$

and

$$h_{22}^{(1)} = -h_{12}\sin\vartheta_1 + h_{22}\cos\vartheta_1$$

Now, we want to eliminate alto the h_{32} coefficient; by this way, it is necessary to multiply $\overline{\underline{\underline{H}}}_2^{(1)}$ times another 3×3 matrix, this time unknown, since we haven't worked on ϑ_2; we have:

$$\overline{\underline{\underline{H}}}_2^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\vartheta_2 & \sin\vartheta_2 \\ 0 & -\sin\vartheta_2 & \cos\vartheta_2 \end{bmatrix} \overline{\underline{\underline{H}}}_2^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\vartheta_2 & \sin\vartheta_2 \\ 0 & -\sin\vartheta_2 & \cos\vartheta_2 \end{bmatrix} \begin{bmatrix} \cos\vartheta_1 & \sin\vartheta_1 & 0 \\ -\sin\vartheta_1 & \cos\vartheta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ 0 & h_{32} \end{bmatrix}$$

therefore, we obtained, for the second step:

$$\underline{\underline{Q}}^{\mathrm{T}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\vartheta_2 & \sin\vartheta_2 \\ 0 & -\sin\vartheta_2 & \cos\vartheta_2 \end{bmatrix} \begin{bmatrix} \cos\vartheta_1 & \sin\vartheta_1 & 0 \\ -\sin\vartheta_1 & \cos\vartheta_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \cos\vartheta_1 & \sin\vartheta_1 & 0 \\ -\cos\vartheta_2\sin\vartheta_1 & \cos\vartheta_2\cos\vartheta_1 & \sin\vartheta_2 \\ \sin\vartheta_2\sin\vartheta_1 & -\sin\vartheta_2\cos\vartheta_1 & \cos\vartheta_2 \end{bmatrix}$$

and:

$$\underline{\underline{R}} = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} \\ 0 & h_{22}^{(2)} \\ 0 & 0 \end{bmatrix}$$

where:

$$h_{22}^{(2)} = h_{22}^{(1)}\cos\vartheta_2 + h_{32}^{(1)}\sin\vartheta_2$$

this is the QR decomposition for the second iteration. The interesting thing is in the fact that, going on with the iterations, it is possible to upgrade the QR decomposition starting from the one obtained at the previous iterations.

The residual is:

$$\underline{r}_2 = -\beta\underline{\underline{Q}}^{\mathrm{T}}\hat{\mathrm{e}}_1\Big|_{2+1}$$

the unit vector $\hat{\mathrm{e}}_1$ selects the first column only, and erases all the remaining ones; $m = 2$, so $2 + 1 = 3$: the residual is:

$$\underline{r}_2 = \beta(\underline{\underline{Q}}^{\mathrm{T}})_{3,1} = \beta \sin\vartheta_2 \sin\vartheta_1$$

but we can recall that:

$$-\beta \sin\vartheta_1 = \underline{r}_1$$

so:

$$\underline{r}_1 = -\underline{r}_1 \sin\vartheta_2$$

This is a **general rule**; indeed, at each i-th iteration:

$$||\underline{r}_i|| = \big|\big|-\sin\vartheta_i \underline{r}_{i-1}\big|\big|$$

With this residual we can understand if the i-th iteration is sufficient. Moreover, it is not necessary to explicitly calculate the solution to understand if it is accurate: the evaluation of the residual does not require the evaluation of $\underline{y}_m$. Given the tolerance t, if:

$$\frac{||\underline{r}_m||}{||\underline{b}||} \leq t$$

then, $\underline{y}_m$ can be calculated as the solution of the system:

$$\underline{\underline{R}}\,\underline{y}_m = \beta \underline{\underline{Q}}_m^{\mathrm{T}}\,\hat{\mathrm{e}}_1$$

which is the triangular system, very easy to be solved. The matrix $\underline{\underline{R}}$ is the result of the application of the set of Givens matrices to the Hessenberg matrix; then:

$$\underline{x}_m = \underline{x}_0 + \underline{\underline{V}}_m\,\underline{y}_m$$

therefore, at each iteration, it is necessary to store the modified Hessenberg matrix and the components of the Krylov subspace basis. If the vector $\underline{v}_i$ is never zero, the Krylov subspace will equal $\mathbb{R}^n$, and the solution is exact, since it is the global minimum on $\mathbb{R}^n$.

3.3.1 Final considerations

This procedure suggests that, with respect to the conjugate gradient, we have a problem: $\underline{\underline{V}}$, which is a full matrix, has to be stored, just like $\overline{\underline{\underline{H}}}_m$; this operation can be memory consuming.

Another possibility for the method to fail is in the Gram-Schmidt algorithm; indeed, this can be unstable, and so the quality of our basis may be bad; if the basis is ill-defined, everything will be affected; this usually occurs if the system matrix $\underline{\underline{A}}$ is ill-conditioned. This introduces the following subject, which is **preconditioning**, which will be quickly discussed.

Before discussing preconditioners, it is necessary to point out that there are different formulations for GMRES, using different ideas; an idea is to use **restarting**. With restarting methods, we fix *a priori* the maximum dimension of the Krylov subspace to a certain N_{dim}; then, after N_{dim} iterations, we compute the solution and we use this as initial guess for a new cycle of N_{dim} iterations; this solution saves memory, since we are fixing the maximum dimension of the matrix; on the other hand, we lose the convergence properties of the method, since we are getting closer and closer to the solution, without using the method in the "natural way".

Chapter 4

Preconditioning

As usual, our objective is to solve:

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

but in this case, we consider to have, for hypothesis, an ill-conditioned matrix:

$$\kappa(\underline{\underline{A}}) \gg 1$$

Now, a question: is there some matrix $\underline{\underline{M}}^{-1}$ such that, if we multiply both members times it, we get a new system with better conditioning number?

$$\underline{\underline{M}}^{-1}\underline{\underline{A}}\,\underline{x} = \underline{\underline{M}}^{-1}\underline{b}$$

defining:

$$\underline{\underline{\tilde{A}}} \triangleq \underline{\underline{M}}^{-1}\underline{\underline{A}}$$

and

$$\underline{\tilde{b}} = \underline{\underline{M}}^{-1}\underline{b}$$

we want:

$$\kappa(\underline{\underline{\tilde{A}}}) \ll \kappa(\underline{\underline{A}})$$

Obviously, the best possible choice is:

$$\underline{\underline{M}}^{-1} = \underline{\underline{A}}^{-1}$$

indeed, by this way, $\tilde{A} = \underline{\underline{I}}$, and the conditioning number equals one: the system is automatically solved. But this is not reasonable: we want to solve a linear system with efficient methods, and inverting a matrix to obtain a preconditioner is not reasonable. But this idea is useful, because it suggests that $\underline{\underline{M}}$ should approximate somehow the matrix $\underline{\underline{A}}$, and that, given:

$$\underline{\underline{M}}\,\underline{y} = \underline{c}$$

the solution of this system should be easy. If $\underline{\underline{M}}$ satisfies these two properties, it is a good preconditioner.

Different kind of preconditioners can be applied; an alternative to what we discussed up to this moment, which were **left preconditioners**, are **right preconditioners**, which are matrices multiplied by right; from:

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

let us define

$$\underline{u} = \underline{\underline{M}}\,\underline{x}$$

so:

$$\underline{\underline{A}}\,\underline{\underline{M}}^{-1}\,\underline{u} = \underline{b}$$

and:

$$\tilde{A} = \underline{\underline{A}}\,\underline{\underline{M}}^{-1}$$

in this case, the right hand side is the same; the residual is:

$$\underline{r} = \underline{b} - \underline{\underline{\tilde{A}}}\,\underline{u} = \underline{b} - \underline{\underline{A}}\,\underline{x}$$

so the residual, in this case, is not changing; this is very important in CG or in GMRES, because, by this way, the stopping criterion remains the same, for both preconditioned or not residuals! Therefore, even if left preconditioners are easier to be applied, they are not always the best choice.

We stated that the preconditioning matrix has to be, in some sense, an approximation of the original matrix; therefore, a good class of preconditioners are **incomplete factorizations**. Let $\underline{\underline{A}}$ be a matrix; we know that it is possible to compute the LU factorization, and we know that, if we know this factorization, the solution of the system become easy. But the computation

of the exact factorization is quite complicated, and requires a lot of calculations, so we don't want to compute the exact $\underline{\underline{L}}$ and $\underline{\underline{U}}$ matrices, but we want to obtain something which give, **as product**, something which approximates the system matrix $\underline{\underline{A}}$:

$$\underline{\underline{L}}_{\text{incomplete}}\underline{\underline{U}}_{\text{incomplete}} = \underline{\underline{M}} \sim \underline{\underline{A}}$$

So: how to compute these matrices? It is possible to find in literature the 0-fill in algorithm, and, sometimes, this is working. However, if we accept more elements, the preconditioner improves; depending on the fill in, we can obtain better preconditioners. Usually, a moderate fill-in is sufficient.

What is usually done for the calculation of the ILU (Incomplete LU) is to put a threshold, and if the computed value is too low, we neglect it.

An alternative is the Choleski factorization.

Another alternative, which is quite complicated but very good, is to consider the following matrix norm:

$$\left|\left|\underline{\underline{I}} - \underline{\underline{M}}^{-1}\underline{\underline{A}}\right|\right|$$

in this case, some elements of $\underline{\underline{M}}^{-1}$ are fixed, and then it is possible to try to minimize this norm, changing the values of $\underline{\underline{M}}$; this is much more complicated, but it is also a very good preconditioner.