Science & Technology Facilities Council

Technical Report

An analysis of GMRES worst case convergence

Mario Arioli

December 11, 2009

RAL-TR-2009-020

Library and Information Services
SFTC Rutherford Appleton Laboratory
Harwell Science and Innovation Campus
Didcot
OX11 0QX
UK
Tel: +44 (0)1235 445384
Fax: +44(0)1235 446403
Email: library@rl.ac.uk

ISSN 1358-6254

RAL-TR-2009-020

An analysis of GMRES worst case convergence

M. Arioli[1]

Krylov space methods minimizing the 2-norm of the residual, such as GMRES, used in solving a linear system with an unsymmetric matrix of order $n \times n$ can present pathological cases where the convergence will be achieved only after $n-1$ steps. Here we will characterize the class of real matrices for which a starting point inducing this worst case convergence exists always

Keywords: GMRES convergence, Lie Group.

AMS(MOS) subject classifications: 65F05, 65F50, 65F10, 65G50

Current reports available by anonymous ftp to ftp.numerical.rl.ac.uk in directory pub/reports.
[1] mario.arioli@stfc.ac.uk Rutherford Appleton Laboratory,
The work was supported by EPSRC grant GR/S42170/01.

Computational Science and Engineering Department
Atlas Centre
Rutherford Appleton Laboratory
Oxon OX11 0QX

December 11, 2009

Contents

1 Introduction

Given the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1.1}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is unsymmetric, the GMRES algorithm computes an approximation of the solution. The analysis presented here is related to the Arnoldi process [6] and in particular to the GMRES algorithm and its variants such as Flexible GMRES, see [7, 5, 6].

GMRES builds an upper Hessenberg matrix $\mathbf{H}$ starting from an initial vector $\mathbf{x}^{(0)}$ computing $\tilde{\mathbf{b}} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$ and following the rules

$$\begin{cases} \mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\mathbf{H}_k \\ \mathbf{Q}_k^T\tilde{\mathbf{b}} = \mathbf{e}_1^k\ ||\tilde{\mathbf{b}}||_2 \end{cases} \tag{1.2}$$

where $\mathbf{H}_k \in \mathbb{R}^{(k+1)\times k}$ is an upper Hessenberg matrix. The algorithm stops if the least-squares problem

$$\min_{\mathbf{y}} ||\mathbf{e}_1||\tilde{\mathbf{b}}||_2 - \mathbf{H}_k\mathbf{y}||_2 \tag{1.3}$$

is less or equal to a chosen threshold η (normally $\eta \approx \sqrt{\varepsilon}$ with ε machine precision).

In particular, we are interested in identifying the matrices and the starting points for which convergence will be achieved only after $n-1$ steps for a reasonably small threshold η. Owing to the free choice of the starting point, we can reduce ourselves, without loss of generality, to the problem of identifying the matrices where there exists a right-hand side $\mathbf{b}$ such that the starting point $\mathbf{x}^{(0)} = 0$ will imply convergence after $n-1$ steps. Moreover, it is possible to assume that $\mathbf{det}(\mathbf{A}) > 0$. The negative determinant case can be reduced to the positive multiplying by -1 the first row in $\mathbf{A}$ and the first entry in $\mathbf{b}$.

The whole process can be split into two phases. First, the Arnoldi process is equivalent to applying a Gram-Schmidt procedure to the matrix

$$\mathbf{B} = [\mathbf{b}; \mathbf{A}] \in \mathbb{R}^{n\times(n+1)}$$

in order to compute the upper triangular matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ with positive entries on the diagonal and the matrix $\mathbf{Q}$ such that

$$\mathbf{B} = \mathbf{Q}[\mathbf{R}; \mathbf{q}] = \mathbf{Q}\left[\ \beta\mathbf{e}_1; \mathbf{H}\ \right].$$

where $\beta = ||\mathbf{b}||_2$ Then, $\mathbf{x}$ is approximated by $\mathbf{Q}_k\mathbf{y}$ with $\mathbf{y}$ solution of (1.3).

The computation of y is also performed in two stages. The Givens (or the Householder) algorithm computes elementary rotations (or reflections in the case of Householder) $\mathbf{G}^{(i)}$ in order to reduce the matrix $\mathbf{H}$ to the upper triangular form $\mathbf{U}$. Here we choose the Householder form in order to avoid complications connected with sign choices in the following:

$$\mathbf{G}^{(i)} = \begin{bmatrix} \mathbf{I}_{i-1} & & & \\ & c_i & s_i & \\ & s_i & -c_i & \\ & & & \mathbf{I}_{n-i-1} \end{bmatrix} \quad i = 1, \ldots, n$$

$$c_i = \frac{(h)_{i,i}}{\sqrt{(h)_{i,i}^2 + (h)_{i+1,i}^2}}) \qquad \text{and} \qquad s_i = \frac{(h)_{i+1,i}}{\sqrt{(h)_{i,i}^2 + (h)_{i+1,i}^2}}$$

Then, we compute

$$\mathbf{z}^{(k)} = \beta \prod_{i=1}^{k} \mathbf{G}^{(i)} \mathbf{e}_1 = \begin{bmatrix} \mathbf{z}_1^{(k)} \\ z_k^{(k)} \end{bmatrix},$$

and solve the system

$$\mathbf{U}\mathbf{y} = \mathbf{z}_1^{(k)}. \tag{1.4}$$

We must immediately note that

$$|| \beta \mathbf{e}_1 - \mathbf{H}_k \mathbf{y} || = |z_k|. \tag{1.5}$$

Moreover, we have that

$$z_i^{(k)} = \beta \prod_{j=1}^{i} s_{j-1} c_i \qquad i = 1, \ldots, k \qquad s_0 = 1 \tag{1.6}$$

$$z_k^{(k)} = \beta \prod_{j=1}^{k} s_j \tag{1.7}$$

Therefore, if $|s_i| < 1$ for all i then the vector $\mathbf{z}$ entries are decreasing in absolute values when i increases. However, the convergence can be extremely slow when all the $s_i = \mathcal{O}(1-\zeta)$ with $\zeta \ll 1$ and in the worst case $n-1$ steps are required. Finally, we point out that in this case $\mathbf{H} \in \mathbb{R}^{n \times n}$, (1.3) is consistent, and the residual is zero. In the following, we will assume that $||\mathbf{b}|| = 1$ for the sake of simplicity and without loss of generality, because our sole interest is to seek the direction of $\mathbf{b}$.

2 Convergence problems for GMRES

Therefore, the Arnoldi algorithm applied to the matrix $\mathbf{A}$ computes, starting with the vector $\mathbf{q}_1 = \mathbf{b}$ and in exact arithmetic, an orthonormal matrix $\mathbf{Q}$ and an upper Hessenberg $\mathbf{H}$ with entries $h_{i+1,i} \geq 0$ $i = 1, \ldots, n$. In particular, $\mathbf{q}_1$ is the first column of $\mathbf{Q}$.

The decomposition

$$\mathbf{A}\mathbf{Q} = \mathbf{Q}\mathbf{H} \tag{2.8}$$

is one of many possible decompositions that can be computed changing the initial vector $\mathbf{q}_1$.

Let $\mathbf{E}_k \in \mathbb{R}^{n \times k}$ be the matrix of the first k column of the $n \times n$ indentity. The GMRES method can be seen as a truncation of (2.8)

$$\mathbf{A}\mathbf{Q}\mathbf{E}_k = \mathbf{Q}\mathbf{H}\mathbf{E}_K = \mathbf{Q}_{k+1}\mathbf{H}_k. \tag{2.9}$$

From the previous analysis it is straightforward to see that the GMRES residual can stagnate if, and only if, the first row of $\mathbf{H}$ has its first $n-1$ entries equal to zero or very small in absolute values. In this case all the residuals will be equal or very close to the norm of $\mathbf{q}_1$ until step n when the final residual collapses to zero if $\mathbf{A}$ is non singular.

Reversely, if the residual at step k does not decrease then the value of s_k in the Givens matrix $\mathbf{G}^{(k)}$ will be equal to 1 or to $1 - \mathcal{O}(\zeta_k)$, where $\zeta_k \leq \zeta \ll 1$. In this second case, the matrix $\mathbf{G}^{(k)}$ is a perturbation of the permutation matrix swapping rows k and $k+1$:

$$\mathbf{G}^{(k)} = (1 - \mathcal{O}(\zeta_k)) \begin{bmatrix} \mathbf{I}_{k-1} & & & \\ & 0 & 1 & \\ & 1 & 0 & \\ & & & \mathbf{I}_{n-k-1} \end{bmatrix} + \mathcal{O}(\zeta_k^{1/2}) \begin{bmatrix} 0_{k-1} & & & \\ & 1 & 0 & \\ & 0 & -1 & \\ & & & 0_{n-k-1} \end{bmatrix} \tag{2.10}$$

The value of $s_k = 1$ if and only if $h_{kk} = 0$ and $h_{k+1,k} > 0$, thus, if all residuals are equals the first $n-1$ entries of the first row of $\mathbf{H}$ must be zero. Thus, the reverse order product of the $\mathbf{G}^{(k)}$

$$\mathbf{G} = \prod_{k=1}^{n} \mathbf{G}^{(n-k)}, \tag{2.11}$$

is the perturbation of the matrix $\mathbf{P}$, where $\mathbf{P}$ is the circulant shifting permutation matrix such that

$$\begin{aligned} \mathbf{P}\mathbf{e}_i &= \mathbf{e}_{i+1} \qquad i = 1, \ldots, n-1 \\ \mathbf{P}\mathbf{e}_n &= (-1)^{n-1}\mathbf{e}_1, \end{aligned}$$

i.e. $\mathbf{G} = \mathbf{P} + \boldsymbol{\Upsilon}$ with $||\boldsymbol{\Upsilon}||_2 \leq \mathcal{O}(\zeta^{1/2})$, and $\mathbf{det}(\mathbf{P}) = 1$.

Moreover, the following relations hold:

$$\left.\begin{aligned} &\mathbf{H} = \mathbf{G}\mathbf{U} \\ &\mathbf{A} = \mathbf{Q}\mathbf{H}\mathbf{Q}^T = \mathbf{Q}\mathbf{G}\mathbf{U}\mathbf{Q}^T = \mathbf{Q}\mathbf{P}\mathbf{U}\mathbf{Q}^T + \mathbf{Q}\boldsymbol{\Upsilon}\mathbf{U}\mathbf{Q}^T \\ &\boldsymbol{\Upsilon}\mathbf{U} \qquad \text{is upper Hessenberg.} \end{aligned}\right\} \tag{2.12}$$

Finally, it is possible, without loss of generality, to assume that the sign of $\mathbf{U}_{nn}$ is positive, owing to the assumption $\mathbf{det}(\mathbf{A}) > 0$.

Remark 1. *Owing to the previous discussion, the assumption* $\mathbf{det}(\mathbf{A}) > 0$ *can be easily removed by re-defining* $\mathbf{P}$. *If the* $\mathbf{det}(\mathbf{A}) < 0$ *then* $\mathbf{U}_{nn} < 0$. *However, in this case we can define* $\mathbf{P}\mathbf{e}_n = (-1)^{n-1}\mathbf{e}_1 sign(\mathbf{U}_{nn})$ *and assume* $\mathbf{U}_{nn} > 0$.

From the above discussion, observing that $||\mathbf{A}||_2 = ||\mathbf{U}||_2$, we can conclude that an initial $\mathbf{q}_1$ that gives the worst case convergence for $\tilde{\mathbf{A}} = \mathbf{Q}\mathbf{P}\mathbf{U}\mathbf{Q}^T$ will produces a very slow convergence also for $\mathbf{A}$. Therefore, in the following section, we will focus only on the worst convergence case.

3 GMRES worst case convergence

From the results of Section 2, we introduce the manifold

$$\mathfrak{M} = \left\{\mathbf{A} : \mathbf{A} = \mathbf{Q}\mathbf{P}\mathbf{U}\mathbf{Q}^T | \mathbf{Q}^T\mathbf{Q} = \mathbf{I}, \mathbf{U}_{i,j} = 0 \; i < j, \; \mathbf{U}_{ii} > 0 \; i = 1, \ldots, n\right\}$$

as the manifold of all the matrices for which there exists a vector $\mathbf{q}_1$ such that if we apply GMRES to the system

$$\mathbf{A}\mathbf{x} = \mathbf{q}_1$$

we will have convergence only after $n-1$ steps with no decrease of the residual at each step. This can be seen as a generalization of the result presented in [3] (see example C page 788).

Owing to the presence in (2.8) of both $\mathbf{Q}$ and its transpose we can assume that $\mathbf{Q} \in \mathsf{SO}(n)$ [1, 4] the special Lie group of orthogonal matrices. Moreover, $\mathbf{U} \in \mathsf{U}(n)$ where $\mathsf{U}(n)$ is the Lie Group of the upper triangular matrices with positive diagonal entries. The group U can be decomposed as the Cartesian product of $\mathsf{D}(n) \times \mathsf{Hei}(n)$ where $\mathsf{D}(n)$ is the group of the positive definite diagonal matrices and $\mathsf{Hei}(n)$ is the Heisenberg group. Following [4, 1], we can also write each matrix $\mathbf{A}$ in $\mathfrak{M}$ as

$$\mathbf{A} = e^{\mathbf{S}}\mathbf{P}e^{\mathbf{W}}e^{\mathbf{V}}e^{-\mathbf{S}} \tag{3.13}$$

with $\mathbf{S} \in \mathfrak{so}(n)$ i.e $\mathbf{S}^T = -\mathbf{S}$ ($\mathbf{S}$ skew-symmetric) where $\mathfrak{so}(n)$ is the Lie algebra of the Lie group $\mathsf{SO}(n)$, $\mathbf{W} \in \mathfrak{d}(n)$ where $\mathfrak{d}(n)$ is the Lie algebra of the diagonal matrices, and $\mathbf{V}$ a strictly upper triangular matrix $\mathbf{V} \in \mathfrak{hei}(n)$ the Heisenberg Lie algebra.

3.1 A density result

Taking into account that $\mathfrak{so}(n)$, $\mathfrak{d}(n)$, and $\mathfrak{hei}(n)$ are finite dimensional linear vector spaces with dimensions

$$\left.\begin{array}{rcl} \mathbf{dim}(\mathfrak{so}(n)) & = & \dfrac{(n-1)n}{2} \\ \mathbf{dim}(\mathfrak{d}(n)) & = & n \\ \mathbf{dim}(\mathfrak{hei}(n)) & = & \dfrac{(n-1)n}{2} \end{array}\right\} \tag{3.14}$$

then $\mathfrak{U} = \mathfrak{so}(n) \times \mathfrak{d}(n) \times \mathfrak{hei}(n)$ is an open subset of $\mathbb{R}^{n^2}$. The manifold $\mathfrak{M}$ can be seen as the image of the map

$$\mathbf{\Phi} : \mathfrak{U} \longrightarrow \mathbb{R}^{n^2} \qquad \mathbf{\Phi}\big(\mathbf{S}, \mathbf{W}, \mathbf{V}\big) = e^{\mathbf{S}}\mathbf{P}e^{\mathbf{W}}e^{\mathbf{V}}e^{-\mathbf{S}}.$$

In particular, $\mathbf{\Phi} \in C^\infty$ i.e. is a smooth map. Its image does not contain the symmetric matrices. Moreover, the symmetric part of of $\mathbf{A}$ is equal to

$$\mathbf{Q}\big(\mathbf{P}\mathbf{U} + \mathbf{U}^T\mathbf{P}^T\big)\mathbf{Q}^T$$

i.e. it is similar to a matrix having entry in position $(1,1)$ equal to zero. Therefore, the image of $\mathbf{\Phi}$ does not contains the matrices with a positive definite symmetric part.

From Sard Theorem[1] [8, 2] it follows that the map $\mathbf{\Phi}$ is a diffeomorphism almost everywhere in $\mathbb{R}^{n^2}$, i.e. the critical points of $\mathbf{\Phi}$ are a small set. In particular, when in Sard Theorem $m = p$ for a regular point y ($x \ni \Psi^{-1}(y)$ are non critical) the set $\Psi^{-1}(y)$ reduces to a single point. Therefore, the map $\mathbf{\Phi}\big(\mathbf{S}, \mathbf{W}, \mathbf{V}\big)$ is a diffeomorphism almost everywhere, and then it is invertible and injective almost everywhere. Owing to the bijective correspondence between the Lie group and the Lie algebras above, the function $\mathbf{\Phi}\big(\mathbf{S}, \mathbf{W}, \mathbf{V}\big)$ is also a function between $\mathsf{SO}(n) \times \mathsf{U}(n)$ into $\mathbb{R}^{n^2}$, i.e. $\mathbf{\Phi}\big(\mathbf{S}, \mathbf{W}, \mathbf{V}\big) = \tilde{\mathbf{\Phi}}\big(\mathbf{Q}, \mathbf{U}\big)$ where $\mathbf{Q} \in \mathsf{SO}(n)$ and $\mathbf{U} \in \mathsf{U}(n)$. In the following section, the set of the critical values of this map is characterized.

[1] **Sard Theorem**

Theorem 3.1. *Let $\Psi : U \longrightarrow \mathbb{R}^m$ be a smooth map, defined on an open set $U \subset \mathbb{R}^p$, and let*

$$C = \big\{x \in U \mid \mathbf{rank}(d\Psi_x) < m\big\}.$$

Then the image $\Psi(C) \subset \mathbb{R}^m$ has Lebesgue measure zero.

3.2 The Jacobian of $\tilde{\mathbf{\Phi}}(\mathbf{Q},\mathbf{U})$

Taking into account that every matrix $\mathbf{Q}\in\mathsf{SO}(n)$ is a diffeomorphic function of a $\mathbf{S}\in\mathfrak{so}(n)$, the following Lemma is straightforward

Lemma 3.1. *Let* $\mathbf{Q}\in\mathsf{SO}(n)$ *and* $\mathbf{E}_{ij}=\mathbf{e}_i\mathbf{e}_j^T-\mathbf{e}_j\mathbf{e}_i^T$ *the natural basis of the skew-symmetric matrices vector space, i.e.* $so\ni\mathbf{S}=\sum_{ij}\sigma_{ij}\mathbf{E}_{ij},\quad \sigma_{ij}\in I\!R$. *Then*

$$\mathbf{Q}^T\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}}\in\mathfrak{so},\qquad \textit{and}\qquad \frac{\partial\mathbf{Q}}{\partial\sigma_{ij}}\mathbf{Q}^T\in\mathfrak{so}.$$

Proof.

$$0=\frac{\partial(\mathbf{Q}^T\mathbf{Q})}{\partial\sigma_{ij}}=\frac{\partial\mathbf{Q}^T}{\partial\sigma_{ij}}\mathbf{Q}+\mathbf{Q}^T\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}}=(\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}})^T\mathbf{Q}+\mathbf{Q}^T\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}},$$

$$0=\frac{\partial(\mathbf{Q}\mathbf{Q}^T)}{\partial\sigma_{ij}}=\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}}\mathbf{Q}^T+\mathbf{Q}\frac{\partial\mathbf{Q}^T}{\partial\sigma_{ij}}=\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}}\mathbf{Q}^T+\mathbf{Q}(\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}})^T.$$

□

From the linearity in $\mathbf{U}$ of $\tilde{\mathbf{\Phi}}$ it follows that the

$$\mathbf{rank}(d(\tilde{\mathbf{\Phi}}))\geq\frac{n(n+1)}{2}.$$

In particular, the Jacobin of $\tilde{\mathbf{\Phi}}$ can be expanded taking into consideration that $\mathbf{Q}=\mathbf{Q}(\mathbf{S})$ where $\mathbf{S}$ is a skew-symmetric matrix with $n(n-1)/2$ parameters σ_{ij}, i.e.

$$\mathbf{S}=\sum_{i\neq j}\beta_{ij}\mathbf{E}_{ij},$$

and that

$$\mathbf{U}=\sum_{k\leq p}u_{kp}\mathbf{e}_k\mathbf{e}_p^T.$$

Then the Jacobian expression can be derived from

$$\begin{aligned}\frac{\partial\tilde{\mathbf{\Phi}}}{\partial\sigma_{ij}} &= \frac{\partial\mathbf{Q}}{\partial\sigma_{ij}}\mathbf{P}\mathbf{U}\mathbf{Q}^T-\mathbf{Q}\mathbf{P}\mathbf{U}\mathbf{Q}^T\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}}\mathbf{Q}^T=\mathbf{Q}\big[\mathbf{Q}^T\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}},\mathbf{P}\mathbf{U}\big]\mathbf{Q}^T\\ \frac{\partial\tilde{\mathbf{\Phi}}}{\partial u_{kp}} &= \mathbf{Q}\mathbf{P}\big(\mathbf{e}_k\mathbf{e}_p^T\big)\mathbf{Q}^T;\end{aligned}\tag{3.15}$$

where $\big[\mathbf{X},\mathbf{Y}\big]=\mathbf{X}\mathbf{Y}-\mathbf{Y}\mathbf{X}$ is the Lie bracket. Thus, the Jacobin matrix of $\tilde{\mathbf{\Phi}}\in\mathbb{R}^{n^2\times n^2}$ can be written as

$$\mathbf{J}(\tilde{\mathbf{\Phi}})=(\frac{\partial\tilde{\mathbf{\Phi}}}{\partial u_{kp}};\frac{\partial\tilde{\mathbf{\Phi}}}{\partial\sigma_{ij}})\big|_{i\neq j;k\leq p}\qquad i,j,k,p\in\{1,\ldots,n\}$$

where the columns of $\mathbf{J}(\tilde{\mathbf{\Phi}})$ are expressed as $n\times n$ matrices.

$\mathbf{J}(\tilde{\mathbf{\Phi}})$ will not be full rank if there exists a set of real values $x_{ij)}$ and y_{kp} not all identically zero such that

$$\sum_{j\neq j}x_{ij}\mathbf{Q}\big[\mathbf{Q}^T\frac{\partial\mathbf{Q}}{\partial\sigma_{ij}},\mathbf{P}\mathbf{U}\big]\mathbf{Q}^T+\sum_{k\leq p}y_{kp}\mathbf{Q}\mathbf{P}\big(\mathbf{e}_k\mathbf{e}_p^T\big)\mathbf{Q}^T=\mathbf{0}.\tag{3.16}$$

From (3.16) and Lemma 3.1, and from the linear vector space properties of $\mathfrak{so}(n)$ it follows that $\mathbf{rank}(\mathbf{J}(\tilde{\mathbf{\Phi}})) < n^2$ if, and only if, there exists a skew-symmetric matrix $\tilde{\mathbf{S}} \in \mathfrak{so}(n)$, an upper triangular matrix $\tilde{\mathbf{U}}$ with positive diagonal entries, and an upper triangular matrix $\tilde{\mathbf{V}}$ such that

$$\mathbf{Q}\Big(\big[\tilde{\mathbf{S}}, \mathbf{P}\tilde{\mathbf{U}}\big] + \mathbf{P}\tilde{\mathbf{V}}\Big)\mathbf{Q}^T = \mathbf{0}. \tag{3.17}$$

Equation (3.17), has non trivial solutions. If, $\mathbf{V} = \mathbf{0}$, then (3.17) implies that $\mathbf{S}$ and $\mathbf{PU}$ commute. From the properties of the skew-symmetric matrices it follows that $\mathbf{U} = \lambda\mathbf{I}$, $\lambda \in \mathbb{R} \setminus \{0\}$. An $\mathbf{S} = \mathbf{P}^T\mathbf{S}\mathbf{P}$ will then satisfy (3.17). Furthermore, it is easy to see that in the previous case, the image of $\mathbf{\Phi}\big(\mathbf{Q}, \mathbf{I}\big)$ and of all $\mathbf{\Phi}\big(\mathbf{Q}\mathbf{P}^k, \mathbf{I}\big)$ will collapse to the same value.

In particular, (3.17) will be satisfied, if the strict lower triangular part of $\mathbf{P}^T\mathbf{S}\mathbf{P}\mathbf{U} - \mathbf{U}\mathbf{S}$ has all zero entries, owing to the possibility of choosing $\mathbf{V}$ equal to the upper triangular part.

3.3 An algebraic geometry point of view

We describe here a simple characterization that defines a necessary and sufficient condition such that $\mathbf{A} \in \mathfrak{M}$.

Theorem 3.2. *Let $\mathbf{A} \in I\!R^{n^2}$ non singular. Let*

$$f(\mathbf{x}) = \begin{bmatrix} \mathbf{x}^T\mathbf{A}\mathbf{x} \\ \mathbf{x}^T\mathbf{A}^2\mathbf{x} \\ \vdots \\ \mathbf{x}^T\mathbf{A}^{n-1}\mathbf{x} \end{bmatrix}; \qquad \mathfrak{V} = \big\{\mathbf{x} \neq 0; f(\mathbf{x}) = 0\big\}.$$

Then

$$\mathbf{A} \in \mathfrak{M} \Longleftrightarrow \mathfrak{V} \neq \emptyset.$$

Proof. Without loss of generality, we assume that $||\mathbf{b}|| = 1$ and $\mathbf{q}_1 = \mathbf{b}$.

$$\Longrightarrow$$

We have that

$$\mathbf{Q}^T\mathbf{b} = \mathbf{e}_1.$$

Then

$$\mathbf{A}\mathbf{b} = \mathbf{Q}\mathbf{P}\mathbf{U}\mathbf{e}_1 = \mathbf{Q}\mathbf{P}\mathbf{e}_1\xi_2 = \mathbf{Q}\mathbf{e}_2\xi_2,$$

and, thus,

$$\mathbf{b}^T\mathbf{A}\mathbf{b} = \mathbf{e}_1^T\mathbf{e}_2 = 0.$$

Moreover, we have

$$\mathbf{b}^T\mathbf{A}^i\mathbf{b} = \mathbf{e}_1^T\big(\prod_{k=1}^{i}\mathbf{P}\mathbf{U}\big)\mathbf{e}_1 = \mathbf{e}_1^T\big(\sum_{j=2}^{i}\mathbf{e}_j\xi_j\big) = 0.$$

$$\Longleftarrow$$

We will proceed by induction on k. The Gram-Schmidt process in exact arithmetic applied within GMRES can be used to prove that under the Hypothesis of the theorem there exists a $\mathbf{q}_1 = \mathbf{b}$ such that $\mathbf{A} \in \mathfrak{M}$. The first step is to compute $\mathbf{q}_2$ from $\mathbf{q}_1$:

$$\mathbf{q}_2 = \big(\mathbf{A}\mathbf{q}_1 - \mathbf{q}_1\mathbf{q}_1^T\mathbf{A}\mathbf{q}_1\big)\big/||\mathbf{A}\mathbf{q}_1 - \mathbf{q}_1\mathbf{q}_1^T\mathbf{A}\mathbf{q}_1||.$$

Thus, from the hypothesis, it follows that

$$\mathbf{q}_2 = \mathbf{A}\mathbf{q}_1 / ||\mathbf{A}\mathbf{q}_1||,$$

and

$$\mathbf{q}_1^T \mathbf{A}\mathbf{q}_2 = \mathbf{q}_1^T \mathbf{A}^2 \mathbf{q}_1 = 0.$$

Moreover, if $n > 3$ then we have

$$\mathbf{q}_1^T \mathbf{A}^i \mathbf{q}_2 = \mathbf{q}_1^T \mathbf{A}^{i+1} \mathbf{q}_1 = 0 \quad \forall i \leq n-2.$$

Let us assume that

$$\mathbf{q}_1^T \mathbf{A}^i \mathbf{q}_j = 0 \quad \forall i \leq n-j \quad \forall j \in [2, \ldots, k]. \tag{3.18}$$

Then, we have that

$$\mathbf{q}_{k+1} = \frac{\left(\mathbf{I} - \sum_{j=1}^{k} \mathbf{q}_j \mathbf{q}_j^T\right)\mathbf{A}\mathbf{q}_k}{\alpha}$$

with $\alpha = ||\left(\mathbf{I} - \sum_{j=1}^{k} \mathbf{q}_j \mathbf{q}_j^T\right)\mathbf{A}\mathbf{q}_k||$. Thus, it follows from (3.18 that

$$\mathbf{q}_1^T \mathbf{A}^i \mathbf{q}_{k+1} = \mathbf{q}_1^T \mathbf{A}^{i+1} \mathbf{q}_k - \sum_{j=1}^{k} \mathbf{q}_1^T \mathbf{A}^i \mathbf{q}_j \mathbf{q}_j^T \mathbf{A}\mathbf{q}_k = 0 \quad \forall i \leq n-(k+1).$$

□

The proof is constructive and if $\mathfrak{V} = \emptyset$ but if, given $f_k(\mathbf{x})$

$$f_k(\mathbf{x}) = \begin{bmatrix} \mathbf{x}^T \mathbf{A}\mathbf{x} \\ \mathbf{x}^T \mathbf{A}^2 \mathbf{x} \\ \vdots \\ \mathbf{x}^T \mathbf{A}^{k-1} \mathbf{x} \end{bmatrix},$$

$\mathfrak{V}_k = \left\{\mathbf{x} \neq 0; f_k(\mathbf{x}) = 0\right\} \neq \emptyset$, then there exists $\mathbf{b} \neq 0$ such that the residual during the GMRES algorithm will not decrease for $k-1$ steps.

Corollary 3.1. *Let K be the degree of the minimal polynomial of $\mathbf{A}$ and $\mathbf{A}$ non singular. Then*

$$\mathfrak{V}_k = \emptyset \qquad \forall k \geq K$$

Proof. Let $\wp(\mathbf{x}) \in \mathcal{P}^{(K)}$ the minimal polynomial of degree K such that

$$\wp(\mathbf{A}) = 0.$$

Thus, we have

$$\mathbf{x}^T \wp(\mathbf{A})\mathbf{x} = 0.$$

If $f_j(\mathbf{x}) = 0$ for $j > 0$ all $\mathbf{x}^T \mathbf{A}^j \mathbf{x} = 0$ and then

$$\mathbf{x}^T \mathbf{x} \det(\mathbf{A}) = 0.$$

□

Corollary 3.1 can be used to prove that if we have preconditioned (1.1) such that the preconditioned matrix has multiple eigenvalues then GMRES will converge in the worst case after a number of steps equal to the number of distinct eigenvalues.

Finally, for the classes of symmetric or skew-symmetric matrices, and of matrices with symmetric part positive or negative definite there does not exist a $\mathbf{b} \neq 0$ corresponding to the worst case convergence of GMRES.

4 Conclusions

We have characterized the class of matrices for which a starting point for GMRES for which convergence will be achieved after $n-1$ steps does exist. Despite its purely theoretical nature, the result shows the necessity of linking any result on the GMRES rate of convergence to the starting point and to the right-hand side of (1.1).

Finally, the discussion in Section **??** shows that the worst case pathological behaviour can also occurs in a neighbouring of the worst starting point, i.e. when the first row of $\mathbf{H}$ has small entries.

References

[1] B. C. Hall, *Lie Groups, Lie Algebras, and Representations. An Elementary Introduction*, GTM 22, Springer, New York, USA, second edition ed., 2004.

[2] J. W. Milnor, *Topology from the differentiable viewpoint*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, USA, 1997 reprint of the 1965 original ed., 1965.

[3] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?*, SIAM Journal on Matrix Analysis and Applications, 13 (1994), pp. 778–795.

[4] W. Rossmann, *Lie Groups. An Introduction Through Linear Groups*, Oxford Graduated Texts in Mathematichs, 5, Oxford University Press, Oxford, UK, 2002.

[5] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Stat. Comput., 14 (1993), pp. 461–469.

[6] Y. Saad, *Iterative Methods for Sparse Linear Systems Second Edition*, Society for Industrial and Applied Mathematics, 2003.

[7] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[8] A. Sard, *The measure of the critical values of differentiable maps*, Bull. Amer. Math. Soc., 48 (1942), pp. 883–890.