Orthogonalisation in Krylov subspace methods for model order reduction

Orthogonalisation in Krylov subspace methods for model order reduction

P.J. Heres[1] and W.H.A. Schilders[2]

[1]) Eindhoven University of Technology, Department of Mathematics and Computer Science, PO Box 513, 5600 MB Eindhoven, The Netherlands

p.j.heres@tue.nl

[2]) Philips Research Laboratories Eindhoven, The Netherlands

Introduction

The modelling of the EM behaviour of electronic structures nowadays involves a broad frequency range and coupling of analog and digital behaviour. Much research and increasing computational resources enabled the designers in the past decades to simulate complicated and large structures. One of the approaches to make this modelling feasible is Model Order Reduction. In this approach one tries to capture the essential features of a large model, into a smaller, a more easy to handle model. A wide range of different techniques has been proposed and investigated in the last few decades. Especially Krylov-subspace methods have proved themselves to be very suitable for this area of application (eg. [2], [4], [6] and [8]). Many of these methods guarantee preservation of passivity, which makes them even more interesting.

However, implementing the methods straightforwardly is not enough to make them applicable for real-life applications. In order to make the methods accurate, efficient and suitable for large systems, extra attention and mathematical knowledge is needed. In this paper we will focus on the orthogonalisation of the Krylov space, which is seen to be of importance. Special attention is paid to the orthogonalisation of a Block Krylov space. Also some directions to cheaply avoid parts of the redundancy in the Krylov space methods are pointed out in this paper.

Krylov subspace methods

Modelling of an electronic structure can lead to a Differential Algebraic Equation (DAE), which form now on will be considered in this form:

$$\begin{aligned}(\mathbf{C}\frac{d}{dt} + \mathbf{G})\mathbf{x}(t) &= \mathbf{B}_i\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{B}_o^T\mathbf{x}(t),\end{aligned} \tag{1}$$

where $\mathbf{C} \in \mathbb{R}^{n\times n}$,$\mathbf{G} \in \mathbb{R}^{n\times n}$, $\mathbf{B}_i \in \mathbb{R}^{n\times p}$ and $\mathbf{B}_o \in \mathbb{R}^{n\times p}$ In the very common case that $\mathbf{C}$ is singular this model is not an ODE, but a DAE. The models we consider here can be derived in several ways. It can for instance be a transmission line model, a PEEC model or an FDTD model with spatial discretizations. In general the matrices $\mathbf{G}$ and $\mathbf{C}$ are real and constant in time.

This system of equations can be transformed to the frequency domain with a Laplace transform:

$$\begin{aligned}(s\mathbf{C} + \mathbf{G})\mathbf{X}(s) &= \mathbf{B}_i\mathbf{U}(s) \\ \mathbf{Y}(s) &= \mathbf{B}_o^T\mathbf{X}(s)\end{aligned} \tag{2}$$

When the state space vector in frequency domain $\mathbf{X}(s)$ is eliminated, a transfer function is obtained:

$$\mathbf{H}(s) = \mathbf{B}_o^T(\mathbf{G} + s\mathbf{C})^{-1}\mathbf{B}_i, \tag{3}$$

$\mathbf{H}(s) \in \mathbb{C}^{\,p\times p}$. This transfer function gives a direct relation between input and output of the system and is therefore a compact description of the system behaviour in the frequency domain.

Model Order Reduction methods attempt to approximate the behaviour of the system with a smaller model. A Krylov-subspace method generates a Krylov subspace based on some input matrix $\mathbf{B}$ and some generating matrix $\mathbf{A}$:

$$\mathcal{K}_q(\mathbf{B}, \mathbf{A}) = [\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^q\mathbf{B}] \tag{4}$$

The actual definition of $\mathbf{B}$ and $\mathbf{A}$ depends on the method of choice. For instance, in the method Laguerre-SVD [4] for some choice of $\alpha \in \mathbb{R}$, the input matrix is defined as:

$$(\mathbf{G} + \alpha\mathbf{C})^{-1}\mathbf{B}_i \tag{5}$$

and the generating matrix is:

$$(\mathbf{G} + \alpha\mathbf{C})^{-1}(\mathbf{G} - \alpha\mathbf{C}) \tag{6}$$

In general, for the basis of the Krylov space, say $\mathbf{V}$, the following basic property holds:

$$\mathbf{A}\mathbf{V}_m = \mathbf{V}_{m+1}\mathbf{H} \text{ for all } m, \tag{7}$$

for some matrix $\mathbf{H}$. Here the notation $\mathbf{A}_m$ means the first m columns of the matrix $\mathbf{A}$.

In a next step the system matrices are projected onto an orthonormal basis of the Krylov space. This can be done explicitly; the matrices of the reduced system are then defined as:

$$\begin{aligned} \mathbf{G}_q = \mathbf{V}^T\mathbf{G}\mathbf{V} \quad \mathbf{C}_q = \mathbf{V}^T\mathbf{C}\mathbf{V} \\ \mathbf{B}_{iq} = \mathbf{V}^T\mathbf{B}_i \quad \mathbf{B}_{oq} = \mathbf{V}^T\mathbf{B}_o \end{aligned}$$

If the dimensions of the space are smaller than the dimensions of the original system, an order reduction is achieved. Some methods, like [7] make use of the matrix $\mathbf{H}$ as defined in (7). The projection is then implicit. Others define two Krylov spaces [2], which are orthogonal with respect to each other. Other details about Krylov subspace methods can be found in [4], [6] and [8] and many other papers.

Orthogonalisation

The columns in the Krylov space

$$\mathcal{K}_q(\mathbf{b}, \mathbf{A}) = [\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^q\mathbf{b}] \tag{8}$$

gradually converge to the dominant eigenvector of the matrix $\mathbf{A}$, i.e. the eigenvector of $\mathbf{A}$ associated to the largest eigenvalue. This causes the Krylov space to be very ill-conditioned. Next to that, it becomes hard to calculate an accurate orthogonal basis of this space, because the columns become similar to each other. If the orthogonalisation is done after the generation of the space, as proposed in the Laguerre-SVD method [4], the convergence of the method stagnates. We advocate here to orthogonalizes during the generation of the columns. In that case more directions than only the dominant eigenvector can be calculated accurately and severe numerical artefacts are avoided. We therefore propose to orthogonalize the newly generated vectors

immediately after generation. We have been using Modified Gram-Schmidt for this and in there we orthogonalize against all previously generated vectors. After the newly generated columns are made orthogonal with respect to all previously generated columns, they are normalized. This procedure costs some computation time, but the accuracy of the method is drastically increased in all directions. Also numerical artefacts are avoided.

Next to this, we propose to apply a second refinement on the orthogonalisation, in order to ensure orthogonality up to the machine precision. This is needed in some critical problems, to ensure the preservation of stability during time domain simulations of the reduced model.

Block Arnoldi Orthogonalisation

When a system has more than one, say p ports, $\mathbf{B}_i$ has more has than one column:

$$\mathbf{B}_i = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p] \tag{9}$$

For this system a *Block* Krylov space is built:

$$\mathcal{K}_q(\mathbf{B}_i, \mathbf{A}) = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_p, \mathbf{A}\mathbf{b}_1, \ldots, \mathbf{A}\mathbf{b}_p, \ldots, \mathbf{A}^q\mathbf{b}_1, \ldots, \mathbf{A}^q\mathbf{b}_p] \tag{10}$$

One can imagine that the size of the Krylov space grows with p and so the approximation will be larger if the number of ports grows. Orthogonalisation and normalization in a Block Krylov space can be done in several orders. For instance, one can add columns to the space one column at the time, or one can add them in blocks. We state that in this case it is important to preserve the basic property of a Krylov space given in (7). If this property is violated, the generated approximation can be totally wrong. In experiments we saw that for a corrupted Krylov space, already for very small Krylov spaces of 8 columns, the transfer function of the approximation differed dramatically from the original function. The order of orthogonalisation in the Block Arnoldi Algorithm, as proposed in PRIMA [6] is seen as a right order to orthogonalize a Block Krylov subspace. Here, we also applied a second orthogonalisation step, to ensure exact orthogonality.

The Block Arnoldi algorithm, to generate a Block Krylov space for Laguerre-SVD, looks like this:

Solve $\mathbf{V}_1$ from $(\mathbf{G} + \alpha\mathbf{C})\mathbf{V}_1 = \mathbf{B}$
$\mathbf{V}_1\mathbf{R} = qr(\mathbf{V}_1)$

for $j = 1...q-1$
  Solve $\mathbf{W}$ from $(\mathbf{G}+\alpha\mathbf{C})\mathbf{W} = (\mathbf{G}-\alpha\mathbf{C})\mathbf{B}$
  for $i = 1...j$
    $\mathbf{H}_{ij} = \mathbf{V}_i^T\mathbf{W}$
    $\mathbf{W} = \mathbf{W} - \mathbf{V}_i\mathbf{H}_{ij}$
  end
  for $j = 1...j$
    $\mathbf{\Theta} = \mathbf{V}_i^T\mathbf{W}$
    $\mathbf{W} = \mathbf{W} - \mathbf{V}_i\mathbf{\Theta}$
    $\mathbf{H}_{ij} = \mathbf{H}_{ij} + \mathbf{\Theta}$
  end
  $\mathbf{V}_{j+1}\mathbf{H}_{i+1,j} = qr(\mathbf{W})$
end
$\mathbf{V}_{tot} = [\mathbf{V}_1, ..., \mathbf{V}_q]$

Redundancy

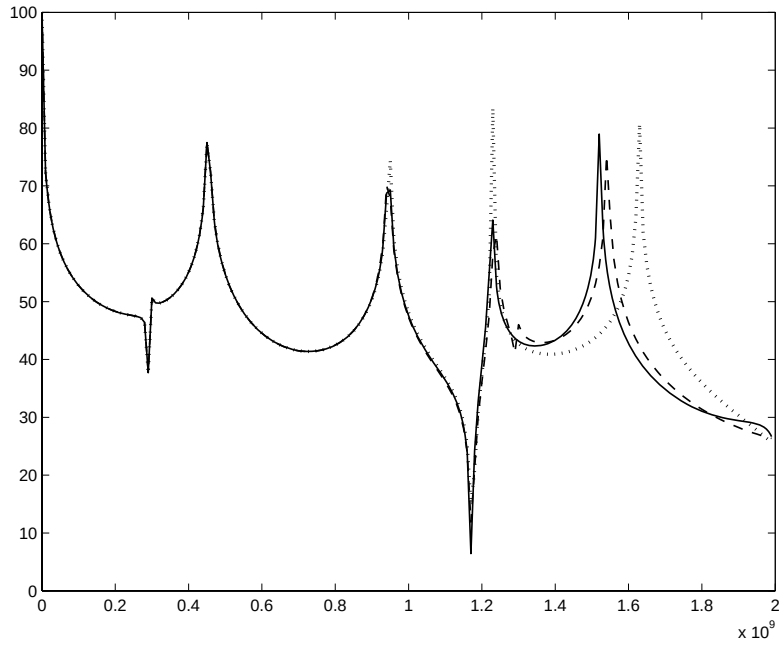
Krylov-subspace methods are known for their redundancy. The method is relatively cheap, but it can contain a lot of information which is not really needed for an accurate approximation. This is even worse if one realizes that there is no known error bound for Arnoldi methods: Easily too large approximations are generated. But even if we were able to stop in time, the Block structure of the Krylov space leads to redundant approximation. Many authors proposed therefore a combination of a Krylov-subspace method with another method, to form a two-step method. In that approach, first a course approximation is calculated with a cheap Krylov-subspace method. In a second step the order of this approximation is decreased by a more expensive but more controllable method like a Truncated Balanced Realization method [5] or by Proper Orthogonal Decomposition [1]. In our research we discovered that a lot can already be done, very cheaply, during the first run of the Krylov-subspace method.

If a Block Krylov-space method is to be generated, it can occur that one of the columns in a new block is almost zero or almost completely spanned by the other columns in the block. In that case we want to stop iterating with this columns, while proceeding with the others. Simply removing information from the space we project on, can lead to the same problems we saw with careless orthogonalisation. With a modified way to calculate a QR-decomposition in the Block Arnoldi Algorithm we are now able to stop

iterating with any wanted column, at any wanted time, because still the basic property of Krylov spaces holds for this algorithm. Details can be found in [3].

Results

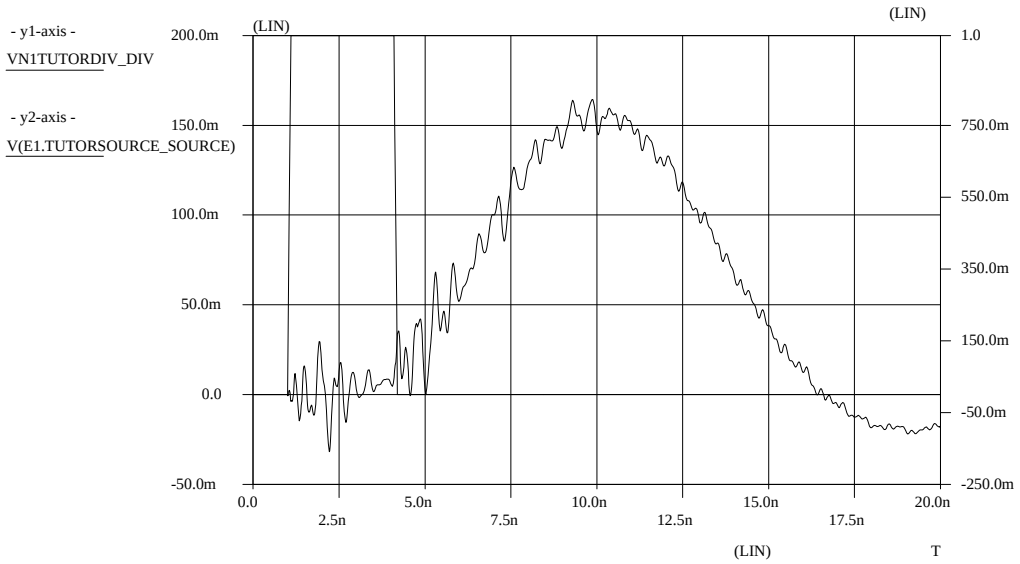
For example, we tested the proposed algorithm on a MNA formulation of an RLC-circuit. The formulation consisted of matrices with size 695. The system has 11 input/ouput ports. We generated a reduced model with 7 iteration of the Block Arnoldi algorithm. In the standard algorithm this leads to a 77-sized system. Columns with norm smaller than the tolerance 10^{-12} were removed. Then in the 2-nd and 3-th iteration a column is removed and the total system size is eventually 66. The approximation, however, is identical to the approximation of size 77, generated by the ordinary PRIMA algorithm. In the picture below, the magnitude of the (1,2) entry of the transfer function of this system (dashed line) is compared with the transfer function of the system of the same size, but generated by ordinary PRIMA (dotted line) and with the transfer function of the full system (solid line). The transfer functions are plotted for values of the frequency ranging from 0 to 2 GHz.



We see that the approximation of the system where the redundant columns are removed, forms a better approximation of the original transfer function than an approximation of the same size, but without removal of redundant columns.

Apart from the removal of columns, we also propose a way to remove unwanted poles from the system, without distroying the Krylov space property. This can be done by an eigendecomposition. The reduced system is reasonably small to make the calculation of a full eigendecomposition feasible. This decomposition gives us direct access to the poles of the reduced system and the associated residues.

The most important reason to implement Krylov subspace methods was their preservation of stability and passivity. This makes stable time domain analysis of very large models of real-life electronic structures possible. The preservation of stability is shown by an example of a model of the printed circuit board, in the following picture. The input is a very steap input pulse with a rise-time of 100 ps.



Conclusions

We have shown that, to be able to apply Krylov subspace methods for Model Order Reduction to large real-life problems, extra effort is needed. Firstly, the accuracy of the method can be improved by orthogonalisation during

the generation of the Krylov space. The Block Arnoldi algorithm is one way to do the orthogonlisation in a correct way. This orthogonalisation is sometimes needed twice. Further, converged columns can be removed during the orthogonalisation step. This can be done without violating the basic Krylov subspace properties. The proposed removal makes the reduced models smaller and therefore less redundant.

All these improvements can be implemented easily in existing methods. This all makes the application of existing methods to large real-life problems feasible.

References

[1] G. Berkooz, P. Holmes, and J.L. Lumley. Turbulence, coherent structures, dynamical systems and symmetry. Technical report, Cambrigde Monographs on Mechanics, 1996.

[2] Peter Feldmann and Roland W. Freund. Efficient Linear Circuit Analysis by Padé Approximation via the Lanczos Process. *IEEE Trans. Computer-Aided Design*, 14:137–158, 1993.

[3] P.J. Heres and W.H.A. Schilders. Deflation of Converged Columns in Krylov Subspace Methods. *To be published*, 2004.

[4] Luc Knockaert and Daniel De Zutter. Passive Reduced Order Multiport Modeling: The Padé-Arnoldi-SVD Connection. *Int. J. Electronics and Communications*, 53:254–260, 1999.

[5] Bruce. C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Trans. Automatic Control*, AC-26(1):17–31, 1981.

[6] A. Odabasioglu and M. Celik. PRIMA: Passive Reduced-order Interconnect Macromodeling Algorithm. *IEEE. Trans. Computer-Aided Design*, 17(8):645–654, August 1998.

[7] L. M. Silveira, M. Kamon, and White. J. Efficient Reduced-Order Modeling of Frequency-Dependent coupling inductances associated with 3-D interconnect structures. In *Design Automation Conference*, pages 376–380, 1995.

[8] D. Skoogh. A Rational Krylov Method for Model Order Reduction. Technical Report 1998-47, Department of Mathematics, Chalmers Unversity of Technology and the university of Goteborg, Goteborg, 41296, Sweden, 1998.