

FAST FREQUENCY SWEEP MODEL ORDER
REDUCTION OF POLYNOMIAL MATRIX EQUATIONS
RESULTING FROM FINITE ELEMENT
DISCRETIZATIONS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Rodney Daryl Slone, M.S.

* * * * *

The Ohio State University

2002

Dissertation Committee:

Prof. R. Lee, Co-Adviser

Prof. J.-F. Lee, Co-Adviser

Prof. P. H. Pathak

Prof. B. Sandstede

Approved by

Co-Adviser

Co-Adviser

Department of Electrical
Engineering

© Copyright by
Rodney Daryl Slone
2002

ABSTRACT

The frequency domain finite element method (FEM) results in matrix equations that have polynomial dependence (or transcendental dependence which can be written as a polynomial via a Taylor series) on the frequency of excitation. For a wide-band fast frequency sweep technique based on a moment-matching model order reduction (MORE) process, researchers generally take one of two approaches. The first is to linearize the polynomial dependence (which will either limit the bandwidth of accuracy or require the introduction of extra degrees of freedom) and then use a well-conditioned Krylov subspace technique such as the projection via Arnoldi (PVA) or the Padé via Lanczos (PVL) processes. The second approach is to work directly with the polynomial matrix equation and use one of the available, but ill-conditioned, asymptotic waveform evaluation (AWE) methods. For large-scale FEM simulations, introducing extra degrees of freedom, and therefore increasing the length of the MORE vectors and the amount of memory required, is not desirable; therefore, the first approach is not alluring. On the other hand, an ill-conditioned AWE process is unattractive. This dissertation presents two MORE techniques for polynomial matrix equations. The first, an automated multipoint Galerkin AWE (MGAWA) process, is capable of producing a reduced order model (ROM) with a relatively small subspace. The second novel process presented, well-conditioned AWE (WCAWE), is capable of producing an accurate, robust, wide-band simulation with just one expansion point.

These novel processes are able to circumvent the problematic issues that arise from the traditional PVA, PVL or AWE techniques. First, these novel processes do not require any additional unknowns and can operate directly on the polynomial matrix equation. Second, these processes are wide-band, and in the case of WCAWE, very well-conditioned even for a large approximation order. Along with the presentation of these algorithms, numerical examples modeled using the FEM are given throughout the work to illustrate their accuracy, efficiency and robustness. Finally, this dissertation closes with a detailed description of many possible areas of further research including an extension of the methods to a block and/or multivariable versions, and applications of the methods to problems in which the system matrix has exponential variations in the ROM varying parameter.

To the incessant search of awareness through clearer thinking

ACKNOWLEDGMENTS

First and foremost, the author would like to give thanks to God through Jesus Christ for everything.

The author is grateful to his co-advisers, Professor Robert Lee and Professor Jin-Fa Lee; the former for giving him the freedom to work on the research that the author found interesting, and the latter for encouraging him to think more intensely about some research issues. In addition, he is appreciative of their suggestions of how to direct this work. The author would also like to thank his examination committee, including Professor Prabhakar Pathak and Professor Bjorn Sandstede, for their invaluable service.

Furthermore, the author wishes to extend his gratitude for the Litton Fellowship and to Ansoft Corporation for their fellowship. Financial support was also provided by Nichols Research Corporation under contract NRC-CR-96-0013 and by Defense Service Office National Laboratory of Singapore under contract DSO/C/99134/0.

Finally, the author thanks everyone with whom he has interacted (family, friends, acquaintances) that has either helped, attempted to help, or unknowingly encouraged him, or has contributed to his personal development and/or present position in life.

VITA

January 13, 1975 Born in Martin, Floyd County, KY

December 1996 B.S. double major in Electrical
Engineering and Mathematics,
University of Kentucky,
Lexington, KY

August 1997 M.S. Electrical Engineering,
University of Kentucky,
Lexington, KY

September 1997 - present Graduate Research Associate,
ElectroScience Laboratory,
The Ohio State University,
Columbus, OH

PUBLICATIONS

Research Publications

W. T. Smith, R. D. Slone and S. K. Das, "Recent Progress in Reduced-Order Modeling of Electrical Interconnects Using Asymptotic Waveform Evaluation and Padé Approximation Via the Lanczos Process," invited paper for the *Applied Computational Electromagnetics Society Newsletter*, vol. 12, no. 2, pp. 46-71, July 1997.

Z. Bai, R. D. Slone, W. T. Smith and Q. Ye, "Error Bound for Reduced System Model by Padé Approximation Via the Lanczos Process," *IEEE Trans. Comput.-Aided Des. Integrated Circuits and Syst.*, vol. 18, no. 2, pp. 133-141, Feb. 1999.

R. D. Slone and R. Lee, "Applying Padé via Lanczos to the finite element method for electromagnetic radiation problems," *Radio Science*, vol. 35, no. 2, pp. 331-340, Mar.-Apr. 2000.

R. D. Slone, R. Lee and J. F. Lee, "Multipoint Galerkin Asymptotic Waveform Evaluation for Model Order Reduction of Frequency Domain FEM Electromagnetic Radiation Problems," *IEEE Trans. Antennas and Propagat.*, vol. 49, no. 10, pp. 1504-1513, Oct. 2001.

R. D. Slone, J. F. Lee and R. Lee, "A comparison of some model order reduction techniques," accepted for publication in *Electromagnetics*.

R. D. Slone, J. F. Lee and R. Lee, "Automating Multipoint Galerkin AWE for a FEM Fast Frequency Sweep," *IEEE Trans. Magn.*, vol. 38, no. 2, pp. 637-640, Mar. 2002.

FIELDS OF STUDY

Major Field: Electrical Engineering

Studies in:

Electromagnetics	Prof. Robert Lee
Controls	Prof. Hooshang Hemami
Mathematics	Prof. Bogdan M. Baishanski

TABLE OF CONTENTS

	Page
Abstract	ii
Dedication	iv
Acknowledgments	v
Vita	vi
List of Tables	xi
List of Figures	xii
Chapters:	
1. Introduction	1
2. Problem statement and illustrative examples	10
2.1 Mathematical statement of MORE problem	10
2.2 Numerical examples used in this study	12
2.2.1 TM_z antenna radiation with an ABC	12
2.2.2 TM_z radiation with anisotropic, dispersive PML	15
2.2.3 TE_z scattering from a material cylinder	18
2.2.4 Three-dimensional tangentially continuous vector FEM	20
3. Classical MORE techniques	26
3.1 Krylov subspace techniques for linear equations	26
3.1.1 Projection via Arnoldi (PVA) review	29
3.1.2 Padé via Lanczos (PVL) review	31

3.2	Moment matching techniques for polynomial equations	32
3.2.1	Asymptotic waveform evaluation (AWE) review	32
3.2.2	Galerkin AWE (GAWE) review	34
3.3	Numerical comparisons	35
4.	Multipoint Galerkin asymptotic waveform evaluation (MGAWWE)	41
4.1	General MGAWWE information	41
4.2	Determining the orders of the subspaces q_v	44
4.3	Using the relative residual to automate MGAWWE	46
4.4	Numerical examples: initial examinations	48
4.5	Numerical examples: further investigations	59
4.5.1	The breakeven point as a function of tol_1	61
4.5.2	MGAWWE versus rational polynomial interpolation	64
4.5.3	Adaptively choosing evaluation frequencies	68
5.	Well-conditioned asymptotic waveform evaluation (WCAWE)	71
5.1	Motivation	71
5.2	The connection between classical MORE techniques	73
5.3	The WCAWE moment-matching process	76
5.3.1	A broadband, moment-matching process	76
5.3.2	Significance of the \mathbf{U} coefficients	77
5.3.3	The WCAWE algorithm	79
5.4	Numerical examples	81
6.	Summary, conclusions and future study	92
6.1	Summary of the findings and conclusions drawn	92
6.2	Future study	94
Appendices:		
A.	Matrix Padé via Lanczos algorithm	96
B.	Proof that (3.21) satisfies (3.32)	100
C.	Counterexamples to show vectors from (5.1) do not match moments	103
C.1	Case one: right hand side linear or higher in σ	104

C.2 Case two: constant right hand side and matrix quadratic or higher in σ	104
D. Proof to show vectors from (5.7) match moments	106
E. Choosing \mathbf{U} in WCAWE to produce Arnoldi vectors	116
Bibliography	119

LIST OF TABLES

Table	Page
2.1 Coefficient values in regions of anisotropic, dispersive PML.	17
4.1 Reduced order model characteristics selected by the MGAWÉ process for the horn antenna.	48
4.2 Reduced order model characteristics selected by the MGAWÉ process for the example in subsection 2.2.2.	50
4.3 Reduced order model characteristics selected by the MGAWÉ process for the example in subsection 2.2.3.	52
4.4 Reduced order model characteristics selected by the MGAWÉ process for the low pass filter.	54
4.5 Reduced order model characteristics selected by the MGAWÉ process for the bowtie antenna.	57
4.6 Reduced order model characteristics selected by the MGAWÉ process for the band pass filter.	59
4.7 The breakeven point as a function of tol_1 for the low pass filter. . . .	61

LIST OF FIGURES

Figure	Page
2.1 FEM problem domain with an ABC.	13
2.2 Geometry of the horn antenna.	14
2.3 FEM problem domain for PEC backed PML.	15
2.4 Interface between elements.	17
2.5 S_{11} for the low pass filter.	22
2.6 S_{12} for the low pass filter.	23
2.7 S_{11} for the bowtie antenna.	24
2.8 S_{12} for the band pass filter.	25
3.1 Impedance calculated for the horn antenna using an LU decomposition, AWE, GAWE and MPVL.	36
3.2 Relative error in the solution vector for the example in subsection 2.2.2 using AWE, GAWE and PVA.	38
3.3 Relative error in the solution vector for the example in subsection 2.2.3 using AWE, GAWE and PVA.	40
4.1 Impedance calculated for the horn antenna using an LU decomposition and MGAW.	49
4.2 Relative error in the solution vector for the example in subsection 2.2.2 using MGAW.	51

4.3	Relative error in the solution vector for the example in subsection 2.2.3 using MGAWE.	53
4.4	MGAWE solution of S_{11} for the low pass filter.	55
4.5	MGAWE solution of S_{12} for the low pass filter.	56
4.6	MGAWE solution of S_{11} for the bowtie antenna.	58
4.7	MGAWE solution of S_{12} for the band pass filter.	60
4.8	MGAWE solution of S_{11} for the low pass filter for various tol_1 values.	62
4.9	MGAWE solution of S_{12} for the low pass filter for various tol_1 values.	63
4.10	Rational polynomial interpolation of S_{11} for the low pass filter.	65
4.11	Rational polynomial interpolation of S_{12} for the low pass filter.	66
4.12	Rational polynomial interpolation of S_{11} for the bowtie antenna.	67
4.13	Pole distribution in the s plane for the horn antenna.	69
4.14	Zoom plot for the impedance of the horn antenna.	70
5.1	Relative error in the solution vector for the example in subsection 2.2.3 using PVA and WCAWE.	82
5.2	Condition number versus order n for AWE and WCAWE methods on the bowtie antenna.	84
5.3	WCAWE solution of S_{11} for the bowtie antenna.	85
5.4	Condition number versus order n for AWE and WCAWE methods on the low pass filter.	87
5.5	WCAWE solution of S_{12} for the low pass filter.	88
5.6	Frequency band convergence versus GAWE and WCAWE iterations for the horn antenna with $s_0 = j2\pi 700\text{MHz}$	90

5.7	Frequency band convergence versus GAWE and WCAWE iterations for the horn antenna with $s_0 = j2\pi 715\text{MHz}$	91
D.1	Reordering the data access.	111

CHAPTER 1

INTRODUCTION

Model order reduction (MORe) is a process in which the number of unknowns (order) of a mathematical representation (model) of a problem of interest is decreased. The mathematical representation of the system with the smaller number of unknowns is known as the reduced order model (ROM). Although this work is concerned with applications in computational electromagnetics, a review of MORe through the late 1980's can be found in [1] and the references contained therein.

There are many reasons that may motivate the application of a MORe procedure. All are ultimately related to obtaining a faster simulation time. However, the way this speedup is obtained can be different from one type of application of a MORe procedure to another. A few of the most important reasons for applying MORe are outlined below.

One reason to use MORe is for macromodeling. In macromodeling a part of the system that does not change is separated from the rest of the system and then simplified to a ROM. Then the ROM is rejoined to the rest of the system and solved. Macromodeling can be used as part of an optimization problem, or when the original system contains both a large linear (on which the MORe is applied) and small non-linear part. An example of macromodeling is when an electronic circuit is separated

into the linear subsystem and a nonlinear subsystem. MORE is applied to the linear part, and then the resulting ROM is reconnected to the nonlinear part and the system is then solved using a circuit simulator such as SPICE. Since the linear part contains fewer unknowns, the entire simulation (including the MORE procedure to create the ROM) can be faster than if SPICE were applied to the original, large system. A recent paper addressing macromodeling is [2].

A second area to which MORE is applicable is domain decomposition problems. This area is much the same as macromodeling except instead of separating a part of the system that does not change from the rest of the system, the system is divided into several subdomains. Then MORE is applied to each of the subdomains and all the resulting ROMs are rejoined and solved. Domain decomposition MORE can be especially useful if either the ratio of the number of unknowns on the boundaries that join the regions to the total number of unknowns is very small or if there are several subdomains that are repeated throughout the system. An example of this case in electromagnetism is a problem whose original domain is inside a waveguide. Some recent work in the area of domain decomposition is given in [3].

A third reason to use MORE is for fast sweep capability. In a fast sweep problem, some parameter on which the original problem depends is variable and it is desired to solve this problem as the parameter changes from some initial value through many intermediate values to some final value. One reason why the solution may be required at many intermediate values is to reduce the risk of failing to capture a region that is rapidly varying, for example, near a resonance. When MORE is applied it is essential that the dominant characteristics of the original system with respect to the varying parameter are captured in the ROM so the solution to the ROM is accurate. In

addition, since the ROM has fewer unknowns than the original system, it is more computationally efficient to solve the ROM for many different values of the varying parameter than to solve the original system. Some examples of fast sweep problems in electromagnetism are fast frequency sweep problems (when the varying parameter is the frequency of excitation) and fast angle sweep problems, where it is desired to investigate how the solution to the original problem changes with the incident angle of the electromagnetic wave.

Now that the reasons for applying MORE have been established and the type of general approach is determined, it is necessary to choose some method to create the original mathematical model of the problem of interest. Some of the popular modeling methods that have been used in conjunction with MORE are the method of moments (MoM), a finite difference method, and the finite element method (FEM). Once a modeling method has been chosen and applied to create the original, large order mathematical model, a MORE technique must be chosen and applied. A list of some MORE techniques follows.

Model based parameter estimation (MBPE) [4] is one of the first discussions of MORE in computational electromagnetics. In that work, MBPE is used to create a fast frequency sweep ROM of a wire antenna modeled using MoM from a modified Numerical Electromagnetics Code. The ROM in [4] is created by matching a rational polynomial to available data from the MoM code. In [5], MBPE is extended to create a simultaneous fast angle and frequency sweep ROM for a wire antenna modeled using MoM. However, their MBPE implementation has a drawback in that “the frequency and spatial domain sampling points and the Padé and polynomial function interpolation orders were chosen experimentally by varying the parameter values and

selecting those which produced the best results” [5]. Therefore, the procedure is not automated and essentially requires running several simulations and then choosing the best solution as the final answer.

A second MORE technique is the spectral Lanczos decomposition method (SLDM). The SLDM is first presented in [6] where it is used to solve three dimensional problems in the time and frequency domains. In [6] Yee’s grid is used to discretize the problem and create the original mathematical model. However, only diffusion behavior of the equations is assumed. In [7] the FEM is used to model the problem domain and no assumption is made about the displacement currents being negligible. However, the applicability of the method in [7] is limited to closed domain regions instead of being able to handle both open and closed domain problems.

A third type of MORE procedure is the Padé approximation via the Lanczos process (PVL). Although the Lanczos algorithm first appears in [8] in 1950, it was not used for MORE until 1994 [9]. The next year, PVL appears in [10] and almost instantly becomes popular. Initially PVL was applied in the circuit analysis community, but later was adopted by researchers in computational electromagnetics. An example of applying PVL to open domain geometries modeled using a finite difference method is given in [11], and a FEM model of an open domain problem is solved using PVL in [12]. In addition, in [13] an adaptive Lanczos-Padé sweep (ALPS) [14] variant of PVL is used to simulate electromagnetics problems modeled using the boundary element method. However, all of these methods suffer from an inherent limitation of PVL which requires the original mathematical model to be a linear function of the ROM varying parameter.

Another type of MORE procedure is based on the Arnoldi process [15]. Like PVL, Arnoldi requires the matrix system to be a linear function of the ROM varying parameter, and the right hand side to be constant. Therefore, to apply Arnoldi to a polynomial matrix equation, the original system must be linearized by introducing extra degrees of freedom. However, in [16] a method is shown which will allow the inverse operator that must be applied to be of the same dimension as the original system, even though the final system must be expanded and linearized. Nevertheless, the Arnoldi vectors must be of the dimension of the expanded, linearized system. Since the memory required to store the ROM vectors can be greater than the memory required to store the sparse matrices, the Arnoldi method may not be practical for large scale computations where memory is an issue.

Another popular MORE technique is asymptotic waveform evaluation (AWE). Unlike PVL and Arnoldi, AWE does not require the original model to be a linear function of the ROM varying parameter. AWE is introduced in [17] to perform MORE on circuit analysis problems. After the introduction of AWE, it was found that it is possible to extend AWE to a more accurate MORE technique if complex frequency hopping (CFH) is employed. In [18] one type of CFH is used in which approximate solutions computed from different expansion points are considered independently. In [19] another type of CFH is shown in which some system poles, computed from different expansion points, are considered simultaneously. However, the system poles themselves are computed using information from only one expansion point. Unfortunately, the CFH algorithms are somewhat difficult to automate and often require user supervision to ensure accuracy. Furthermore, MBPE [4] can be considered to

be a multipoint AWE in that the coefficients of the rational polynomial can be calculated by considering multiple expansion points simultaneously. At any rate, after its introduction in circuit analysis, AWE was then applied in computational electromagnetics, just like PVL. For example, AWE is applied for fast frequency sweep of open domain geometries modeled using an electric field integral equation in [20] and on a combined-field integral equation in [21]. The MoM modeling procedure is also used in conjunction with AWE in [22] where a fast angle sweep is performed for radar cross section calculations. However, since AWE matches moments in a Taylor series, and since integral equation formulations result in transcendental functions of the ROM varying parameter, it is unclear how many terms must be kept in the Taylor series expansion of the transcendental functions to ensure an accurate solution. On the other hand, recall that it is not required to use MoM to model problems that are to be reduced with AWE. In [23] both open and closed domain problems for fast frequency sweep are modeled using a finite difference method and solved using AWE. Of course, it is also possible to use the FEM to model the problem domain. In [24] closed domains are modeled and solved using the FEM and AWE. Finally, open domain problems modeled with FEM are solved for a fast frequency sweep using AWE in [25] and [26].

After discussing all of these MORE techniques and some of the areas in which they have been applied, one may ask why a new MORE technique is needed. An answer to that question can be found by asking, “What is wrong with the MORE techniques that already exist?” Consider the two most popular MORE techniques, PVL and AWE. Although PVL has a large bandwidth of accuracy with respect to the ROM varying parameter, the technique is very limited in applicability. This

limitation arises not only from PVL requiring the original mathematical model to be a linear function of the ROM varying parameter as discussed earlier, but also from the fact shown in [27] that the number of moments matched (mm) in PVL after q iterations¹ is related to the number of independent excitations (i.e. inputs, denoted as p) and number of unknowns at which the solution is desired in the model (i.e. outputs, denoted as o) by the expression $mm = \lfloor q/o \rfloor + \lfloor q/p \rfloor$. On the other hand, AWE suffers no degradation in moment matching power with increasing o and/or p . In addition, as pointed out earlier, AWE does not require the original model to be a linear function of the ROM varying parameter. However, AWE suffers from a small bandwidth of accuracy with respect to the ROM varying parameter. These problems are addressed in [28] where the authors suggest that the problem with AWE is that the process is not Galerkin. Therefore, in [28] “AWE is subjected to a Galerkin treatment (weighted residual)” and is named Galerkin AWE (GAWE). An extension of GAWE to a multipoint treatment (MGAWE) is given in [29].

After identifying a reason for applying MORE, choosing a method to create the original mathematical model of the problem of interest and selecting a MORE technique, it is necessary to consider some issues involved with the application of the MORE technique. For example, for open region problems modeled using finite methods it is necessary to terminate the problem domain with a boundary truncation scheme such as an absorbing boundary condition (ABC), a perfectly matched layer (PML), or a PML backed by an ABC (ABC-PML). In [30] a finite difference grid is terminated with a PML and the model is reduced using PVL. However, the method by which the PML is treated requires introducing an auxiliary variable and thereby

¹Assuming no deflation.

increases the number of unknowns that must be solved for. In [12] a model, resulting when a FEM mesh is terminated with an ABC, is reduced, again using PVL. However, [12] also has a drawback in that to make the FEM model conform to the frequency requirement of PVL the number of unknowns in the model must be doubled. In [25] the AWE technique is used for the MORE of a problem modeled using the FEM with mesh truncation performed by each one of the methods: ABC, PML, and ABC-PML. However, the PML implemented in [25] is non-causal. Dispersive PML is considered in [31], where the FEM is used to model the domain and AWE is used for MORE. The frequency variation used in [31] requires a fifth degree polynomial. In [29] it is shown how to implement dispersive PML in conjunction with the FEM and GAWE requiring only a polynomial frequency variation up to and including fourth order.

Although much work has been done in the area of MORE, there are many issues that are yet to be addressed. Some of these issues are discussed in the remainder of this document and possible approaches to them are disclosed. The remainder of this document is organized as follows.

In chapter 2 a statement of the MORE problem with respect to a fast frequency sweep is given. In addition, several numerical examples, which are used throughout the remainder of this work, are introduced. Then, in chapter 3, several classical MORE techniques are reviewed and compared. The benefits and disadvantages of each of the methods are illustrated. Next, chapter 4 introduces the multipoint Galerkin AWE (MGAWA) method. In chapter 4, several practical implementation issues are addressed to automate the process so no user supervision is required. Then chapter 5 shows a new well-conditioned method for computing a basis for the AWE moment-matching subspace. Finally, in chapter 6, a summary of this research is presented

along with the conclusions drawn and the author's suggested areas of future investigation.

CHAPTER 2

PROBLEM STATEMENT AND ILLUSTRATIVE EXAMPLES

2.1 Mathematical statement of MORE problem

Assume that a modeling procedure, such as the finite element method (FEM), has been applied to a problem of interest, and a matrix system of equations has been obtained. In particular, consider the matrix equation

$$\mathbf{A}(s)\mathbf{X}(s) = \mathbf{B}(s) \tag{2.1}$$

where $\mathbf{A}(s) \in \mathbb{C}^{N \times N}$ is the system matrix, $\mathbf{B}(s) \in \mathbb{C}^{N \times p}$ (where p is the number of *independent* inputs that can inject excitations into the FEM mesh) is the right hand side matrix, $\mathbf{X}(s) \in \mathbb{C}^{N \times p}$ is the solution matrix, $s = j\omega$ and $\omega = 2\pi f$ for the frequency of excitation f . Since the mapping from f to s is injective, the notation $\mathbf{X}(f)$ will replace $\mathbf{X}(s)$ when it is convenient to do so. To illustrate the ideas, assume that it is desired to use MORE for a fast frequency sweep (FFS) in which (2.1) will be solved for f_{num} values of f ranging from $f_1 = f_{min}$ to $f_{f_{num}} = f_{max}$. The straightforward way to solve this problem is to compute

$$\mathbf{X}(f_u) = \mathbf{A}(f_u)^{-1}\mathbf{B}(f_u) \quad \text{for } u = 1, 2, \dots, f_{num}, \tag{2.2}$$

where $\mathbf{A}(f_u)^{-1}$ is performed efficiently by computing, for example, LU decompositions or preconditioners for the conjugate gradient method. Note that a total of f_{num} different decompositions or preconditioners are necessary. However, a FFS MORE procedure only requires computing a few decompositions or preconditioners; in fact, the number required is equal to the number of expansion points (num_pts) used. MORE techniques are computationally efficiency because $num_pts \ll f_{num}$; using only this information, they attempt to accurately extrapolate and interpolate the solution at all f_u . Details common to most MORE techniques are further outlined below.

Definition 2.1 Define the set $\nu = \{1, 2, \dots, num_pts\}$, and let

$$\sigma_{v_u} = j2\pi f_u - s_{0_v} \quad \text{for } v \in \nu, u = 1, 2, \dots, f_{num} \quad (2.3)$$

and

$$\sigma_v = j2\pi f - s_{0_v} \quad \text{for } v \in \nu \quad (2.4)$$

where s_{0_v} is the location of the v th expansion point. □

Using a finite order matrix Taylor series, (2.1) can be rewritten as

$$\sum_{i=0}^{a_1} (\sigma^i \mathbf{A}_i) \mathbf{X}(f) = \sum_{k=0}^{b_1} \sigma^k \mathbf{B}_k. \quad (2.5)$$

Note that in (2.5) \mathbf{A}_i and \mathbf{B}_k have already been shifted to the coordinate system at s_{0_v} . If a_1 and b_1 are chosen large enough so no significant higher order \mathbf{A}_i and/or \mathbf{B}_k term is truncated, any σ_v and corresponding set of \mathbf{A}_i and \mathbf{B}_k can be used in (2.5) as long as consistency is maintained. Otherwise, given some value f_u of f , let σ depend

on the s_{0_ν} that is “closest” to $j2\pi f_u$. More precisely, given some value f_u of f , let

$$\sigma = \sigma_{\min\{\xi \in \nu : |\sigma_{\xi_u}| = \min_{v \in \nu} (|\sigma_{v_u}|)\}}. \quad (2.6)$$

This σ and the corresponding \mathbf{A}_i and \mathbf{B}_k should be used in (2.5) to give a more accurate evaluation at the particular f_u in question.

Finally, assume there are $o \leq N$ unknowns of interest that are desired as outputs. Let $\mathbf{L} \in \mathbb{C}^{N \times o}$ be the matrix that selects these o outputs, and let $\mathbf{H}(f) \in \mathbb{C}^{o \times p}$ be the solution matrix, where

$$\mathbf{H}(f) = \mathbf{L}^T \mathbf{X}(f). \quad (2.7)$$

2.2 Numerical examples used in this study

2.2.1 TM_z antenna radiation with an ABC

Consider the TM_z case for the generalized 2-D Helmholtz equation applied to an antenna radiation problem,

$$\left(\frac{\partial}{\partial x} \frac{1}{\mu} \frac{\partial}{\partial x} + \frac{\partial}{\partial y} \frac{1}{\mu} \frac{\partial}{\partial y} + \omega^2 \epsilon \right) E_z = j\omega J_z, \quad (2.8)$$

where μ and ϵ are isotropic and non-dispersive. The angular frequency is given by ω , and the problem is driven by an electric current source J_z . The problem domain of interest is shown in Figure 2.1, where Ω is the FEM region and $\partial\Omega$ is the outer boundary of the FEM region. Application of the method of weighted residuals gives

$$\begin{aligned} \int \int_{\Omega} \frac{1}{\mu} \frac{\partial E_z}{\partial x} \frac{\partial \phi_t}{\partial x} + \frac{1}{\mu} \frac{\partial E_z}{\partial y} \frac{\partial \phi_t}{\partial y} - \omega^2 \epsilon E_z \phi_t \, dS \\ = -j\omega \int \int_{\Omega} J_z \phi_t \, dS + \int_{\partial\Omega} \frac{1}{\mu} \phi_t \frac{\partial E_z}{\partial n} \, d\ell, \end{aligned} \quad (2.9)$$

where ϕ_t ($t = 1, 2, \dots, N$) are the weighting functions.

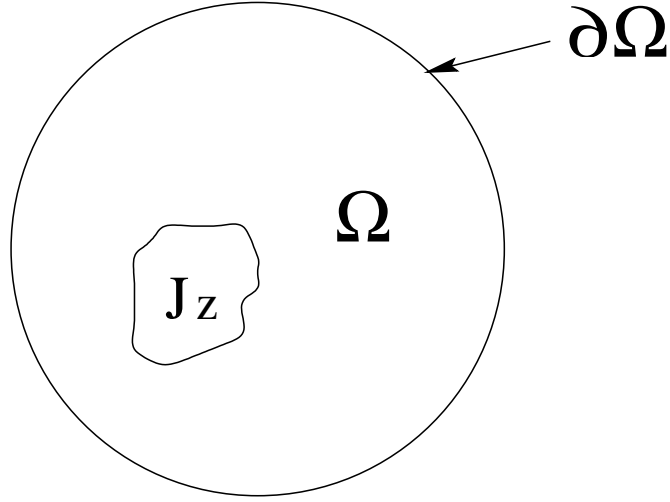


Figure 2.1: FEM problem domain with an ABC.

In order to solve (2.9) the normal derivative of E_z ($\partial E_z / \partial n$) must be specified on $\partial\Omega$. Assuming that $\partial\Omega$ is in the shape of a circle, one can apply the Bayliss Gunzburger Turkel (BGT) absorbing boundary condition (ABC) [32] to the outer boundary. The first-order BGT ABC is given by

$$\left(\frac{\partial}{\partial \rho} + j\beta + \frac{1}{2\rho} \right) E_z = 0, \quad (2.10)$$

where ρ is the distance of the boundary from the center of the mesh, $\hat{\rho}$ is the direction normal to $\partial\Omega$ and β is the wave number. Substitution of (2.10) into (2.9) leads to

$$\begin{aligned} \int \int_{\Omega} \frac{1}{\mu} \frac{\partial E_z}{\partial x} \frac{\partial \phi_t}{\partial x} + \frac{1}{\mu} \frac{\partial E_z}{\partial y} \frac{\partial \phi_t}{\partial y} - \omega^2 \epsilon E_z \phi_t \, dS + \int_{\partial\Omega} \frac{1}{\mu} \phi_t \left(j \frac{\omega}{c} + \frac{1}{2\rho} \right) E_z \, dl \\ = -j\omega \int \int_{\Omega} J_z \phi_t \, dS. \end{aligned} \quad (2.11)$$

Note that (2.11) is composed of terms which depend on polynomial orders of ω , so the terms in (2.11) with common polynomial orders can be grouped together. Also, E_z can be expanded in terms of the FEM basis functions with the number of

basis functions being equal to the number of weighting functions, resulting in $N \times N$ matrices. The final equation is of the form

$$(\mathbf{A}_0 + s\mathbf{A}_1 + s^2\mathbf{A}_2) \mathbf{x}(s) = s\mathbf{b}_1, \quad (2.12)$$

where a nonzero s_{0_v} has not yet been chosen, $p = 1$ because there is a single excitation vector, and \mathbf{b}_1 and each \mathbf{A}_i matrix is independent of s .

Example 1: A model of a two-dimensional horn antenna (whose diagram is shown in Figure 2.2) is created using the above approach. The dimension of the resulting \mathbf{A}_i

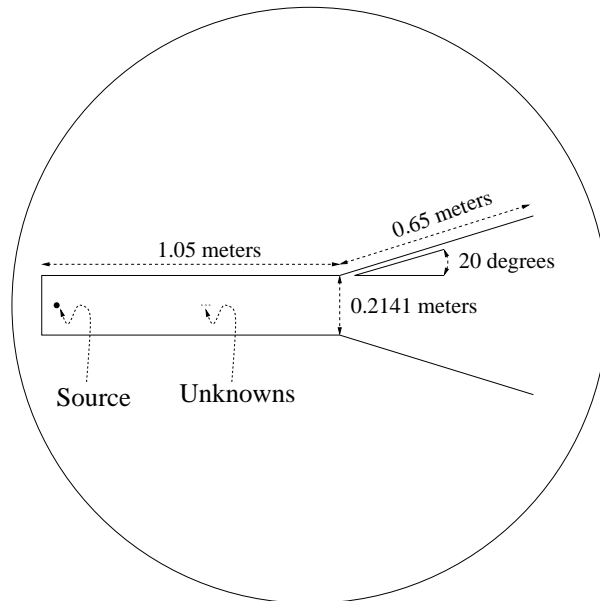


Figure 2.2: Geometry of the horn antenna.

matrices is $N = 3438$. It is desired to find the wave impedance inside the horn from 500MHz to 1.5GHz, which can be done by using \mathbf{L} to select the 3 outputs indicated as unknowns in Figure 2.2.

2.2.2 TM_z radiation with anisotropic, dispersive PML

Again consider the TM_z case for the generalized 2-D Helmholtz equation applied to an antenna radiation problem where this time the permeability (μ) and permittivity (ϵ) are anisotropic and dispersive,

$$\left(\frac{\partial}{\partial x} \frac{1}{\mu_y} \frac{\partial}{\partial x} + \frac{\partial}{\partial y} \frac{1}{\mu_x} \frac{\partial}{\partial y} + \omega^2 \epsilon_z \right) E_z = j\omega J_z. \quad (2.13)$$

The problem domain of interest is shown in Figure 2.3, where Ω is the FEM region,

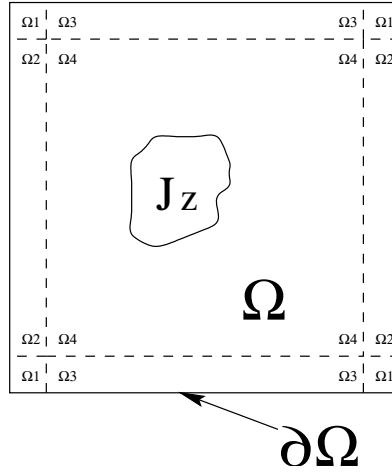


Figure 2.3: FEM problem domain for PEC backed PML.

which is the union of the regions Ω_m for $m = 1 \dots 4$, and $\partial\Omega$ is the perfect electric conductor (PEC) outer boundary of the FEM region. The region between $\partial\Omega$ and the parallel dashed region contains anisotropic, dispersive perfectly matched layer (PML) material. In particular, let

$$\mu_x = \frac{\alpha_y - j\frac{\beta_y}{\omega}}{\alpha_x - j\frac{\beta_x}{\omega}} \mu_0, \quad \mu_y = \frac{\alpha_x - j\frac{\beta_x}{\omega}}{\alpha_y - j\frac{\beta_y}{\omega}} \mu_0, \quad \epsilon_z = \left(\alpha_x - j\frac{\beta_x}{\omega} \right) \left(\alpha_y - j\frac{\beta_y}{\omega} \right) \epsilon_0 \epsilon_r$$

where

$$\alpha_x, \alpha_y = \begin{cases} \alpha & \text{if in PML with dissipation desired in } x \text{ and/or } y, \\ 1 & \text{if in isotropic material} \end{cases}$$

and likewise

$$\beta_x, \beta_y = \begin{cases} \beta & \text{if in PML with dissipation desired in } x \text{ and/or } y, \\ 0 & \text{if in isotropic material.} \end{cases}$$

Apply the method of weighted residuals (and multiply through by $(\alpha - j\frac{\beta}{\omega})\omega^2$) to obtain

$$\begin{aligned} & \int \int_{\Omega} \frac{(\alpha_y - j\frac{\beta_y}{\omega})(\alpha - j\frac{\beta}{\omega})\omega^2}{(\alpha_x - j\frac{\beta_x}{\omega})\mu_0} \frac{\partial E_z}{\partial x} \frac{\partial \phi_t}{\partial x} + \frac{(\alpha_x - j\frac{\beta_x}{\omega})(\alpha - j\frac{\beta}{\omega})\omega^2}{(\alpha_y - j\frac{\beta_y}{\omega})\mu_0} \frac{\partial E_z}{\partial y} \frac{\partial \phi_t}{\partial y} \\ & - \left(\alpha_x - j\frac{\beta_x}{\omega}\right) \left(\alpha_y - j\frac{\beta_y}{\omega}\right) \left(\alpha - j\frac{\beta}{\omega}\right) \omega^4 \epsilon_0 \epsilon_r E_z \phi_t dS = - \int \int_{\Omega} \left(\alpha - j\frac{\beta}{\omega}\right) j\omega^3 J_z \phi_t dS \\ & + \int_{\partial\Omega} \hat{n} \cdot \left\{ \left[\left(\alpha - j\frac{\beta}{\omega}\right) \frac{\omega^2}{\mu_y} \frac{\partial E_z}{\partial x} \phi_t \right] \hat{x} + \left[\left(\alpha - j\frac{\beta}{\omega}\right) \frac{\omega^2}{\mu_x} \frac{\partial E_z}{\partial y} \phi_t \right] \hat{y} \right\} dl \quad (2.14) \end{aligned}$$

where \hat{n} is the unit normal to $\partial\Omega$ and ϕ_t ($t = 1, 2, \dots, N$) are again the weighting functions. Enforce the boundary conditions on tangential H at the interface between elements as shown in Figure 2.4 to give

$$\frac{1}{\mu_{x_1}} \frac{\partial E_{z_1}}{\partial y} = -j\omega H_{x_1} = -j\omega H_{x_2} = \frac{1}{\mu_{x_2}} \frac{\partial E_{z_2}}{\partial y},$$

and

$$\frac{1}{\mu_{y_1}} \frac{\partial E_{z_1}}{\partial x} = -j\omega H_{y_1} = -j\omega H_{y_2} = \frac{1}{\mu_{y_2}} \frac{\partial E_{z_2}}{\partial x}.$$

Therefore, all contributions from inter-element boundaries cancel each other. In addition, since $\phi_t = 0$ on PEC (i.e. do not test there), the term in (2.14) that contains $\int_{\partial\Omega} dl$ is equal to zero for all inter-element and outer boundaries. Equation

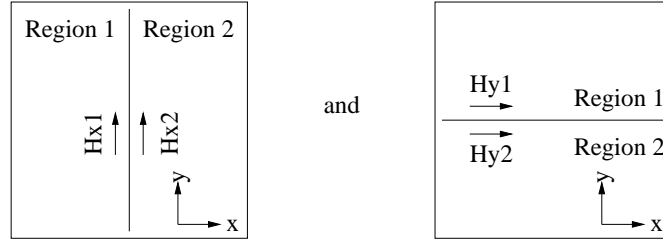


Figure 2.4: Interface between elements.

(2.14) now becomes

$$\int \int_{\Omega} C_1(\omega) \frac{1}{\mu_0} \frac{\partial E_z}{\partial x} \frac{\partial \phi_t}{\partial x} + C_2(\omega) \frac{1}{\mu_0} \frac{\partial E_z}{\partial y} \frac{\partial \phi_t}{\partial y} - C_3(\omega) \epsilon_0 \epsilon_r E_z \phi_t dS = - \int \int_{\Omega} C_4(\omega) j J_z \phi_t dS \quad (2.15)$$

where $C_j(\omega)$ for $j = 1 \dots 4$ is given in Table 2.1 for each region Ω_m for $m = 1 \dots 4$.

Now (2.15) is composed of terms which depend on polynomial orders of ω , and the common order terms are again grouped together. Expanding E_z in terms of FEM basis functions gives a final equation of the form

$$\sum_{i=0}^4 (s^i \bar{\mathbf{A}}_i) \mathbf{x}(s) = \sum_{k=0}^3 s^k \bar{\mathbf{b}}_k. \quad (2.16)$$

	Ω_1	Ω_2	Ω_3	Ω_4
α_x	α	α	1	1
α_y	α	1	α	1
β_x	β	β	0	0
β_y	β	0	β	0
C_1	$\alpha\omega^2 - j\beta\omega$	ω^2	$\alpha^2\omega^2 - j2\alpha\beta\omega - \beta^2$	$\alpha\omega^2 - j\beta\omega$
C_2	$\alpha\omega^2 - j\beta\omega$	$\alpha^2\omega^2 - j2\alpha\beta\omega - \beta^2$	ω^2	$\alpha\omega^2 - j\beta\omega$
C_3	$\alpha\omega^4 - j3\alpha^2\beta\omega^3 - 3\alpha\beta^2\omega^2 + j\beta^3\omega$	$\alpha^2\omega^4 - j2\alpha\beta\omega^3 - \beta^2\omega^2$	$\alpha^2\omega^4 - j2\alpha\beta\omega^3 - \beta^2\omega^2$	$\alpha\omega^4 - j\beta\omega^3$
C_4	$\alpha\omega^3 - j\beta\omega^2$	$\alpha\omega^3 - j\beta\omega^2$	$\alpha\omega^3 - j\beta\omega^2$	$\alpha\omega^3 - j\beta\omega^2$

Table 2.1: Values for α_x , α_y , β_x , β_y and $C_j(\omega)$ for $j = 1 \dots 4$ for each region Ω_m for $m = 1 \dots 4$.

where, again, $p = 1$ and the matrices $\bar{\mathbf{A}}_i$ for $i = 0 \dots 4$ and vectors $\bar{\mathbf{b}}_k$ for $k = 0 \dots 3$ are all independent of s . To shift (2.16) to some nonzero s_{0_v} , substitute $s = \sigma + s_{0_v}$ into (2.16) and collect common terms of σ . The result is

$$\sum_{i=0}^4 (\sigma^i \mathbf{A}_i) \mathbf{x}(f) = \sum_{k=0}^3 \sigma^k \mathbf{b}_k \quad (2.17)$$

where

$$\mathbf{A}_0 = \bar{\mathbf{A}}_0 + s_{0_v} \bar{\mathbf{A}}_1 + s_{0_v}^2 \bar{\mathbf{A}}_2 + s_{0_v}^3 \bar{\mathbf{A}}_3 + s_{0_v}^4 \bar{\mathbf{A}}_4 \quad (2.18)$$

$$\mathbf{A}_1 = \bar{\mathbf{A}}_1 + 2s_{0_v} \bar{\mathbf{A}}_2 + 3s_{0_v}^2 \bar{\mathbf{A}}_3 + 4s_{0_v}^3 \bar{\mathbf{A}}_4$$

$$\mathbf{A}_2 = \bar{\mathbf{A}}_2 + 3s_{0_v} \bar{\mathbf{A}}_3 + 6s_{0_v}^2 \bar{\mathbf{A}}_4$$

$$\mathbf{A}_3 = \bar{\mathbf{A}}_3 + 4s_{0_v} \bar{\mathbf{A}}_4$$

$$\mathbf{A}_4 = \bar{\mathbf{A}}_4$$

$$\mathbf{b}_0 = \bar{\mathbf{b}}_0 + s_{0_v} \bar{\mathbf{b}}_1 + s_{0_v}^2 \bar{\mathbf{b}}_2 + s_{0_v}^3 \bar{\mathbf{b}}_3$$

$$\mathbf{b}_1 = \bar{\mathbf{b}}_1 + 2s_{0_v} \bar{\mathbf{b}}_2 + 3s_{0_v}^2 \bar{\mathbf{b}}_3$$

$$\mathbf{b}_2 = \bar{\mathbf{b}}_2 + 3s_{0_v} \bar{\mathbf{b}}_3$$

$$\mathbf{b}_3 = \bar{\mathbf{b}}_3.$$

Example 2: A model of a material cylinder illuminated by a uniform electric line source is created using the above method. The order of the resulting \mathbf{A}_i matrices is $N = 4734$. The solution is computed up to 500MHz, which is the frequency where the edge length of a side of an element is about 1/20 of a wavelength.

2.2.3 TE_z scattering from a material cylinder

For this type of example, the \mathbf{A}_i matrices can be found as shown in the previous subsections, but special care must be given to the right hand side because it contains

exponential terms of s . The right hand side, however, can be expanded into a Taylor series around s_{0_v} . The form of an entry in the right hand side looks like

$$(s\bar{C}_1 + \bar{C}_2) e^{s\bar{C}_3} \quad (2.19)$$

where

$$\bar{C}_1 = \sqrt{\mu\epsilon} (1 - \cos(\phi)), \quad (2.20)$$

$$\bar{C}_2 = \frac{1}{2\rho}, \quad (2.21)$$

$$\bar{C}_3 = -\sqrt{\mu\epsilon}\rho \cos(\phi), \quad (2.22)$$

ϕ is the angle difference between the location of the finite element under consideration and the incoming uniform plane wave, and ρ is the distance of the boundary from the center of the mesh. The Taylor series for (2.19) is

$$\sum_{k=0}^{\infty} \frac{\sigma^k}{k!} \left[k\bar{C}_1\bar{C}_3^{k-1}e^{s_{0_v}\bar{C}_3} + (s_{0_v}\bar{C}_1 + \bar{C}_2)\bar{C}_3^k e^{s_{0_v}\bar{C}_3} \right]. \quad (2.23)$$

Example 3: A TE_z uniform plane wave is scattered from a material cylinder and modeled using the approach outlined above. The outer boundary of the FEM mesh is treated with an absorbing boundary condition and it is found that the Taylor series expansion introduces insignificant error into the solution if $b_1 = 10$ (that is, eleven terms are used with powers ranging from σ^0 to σ^{10}). Therefore, the matrix system can be written as

$$\sum_{i=0}^2 (\sigma^i \mathbf{A}_i) \mathbf{x}(f) = \sum_{k=0}^{10} \sigma^k \mathbf{b}_k \quad (2.24)$$

where, as always, it is the case that \mathbf{A}_i and \mathbf{b}_k must be written so they are not functions of s . The solution to this problem is then computed up to 500MHz, which is again the frequency where an element's average edge length is about $\lambda/20$. For this example, $N = 1276$.

2.2.4 Three-dimensional tangentially continuous vector FEM

The differential form of Maxwell's Equations in the frequency domain are

$$\nabla \times \mathbf{E} = -j\omega[\bar{\mu}] \cdot \mathbf{H} \quad (2.25)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + j\omega[\bar{\epsilon}] \cdot \mathbf{E} \quad (2.26)$$

where $[\bar{\epsilon}]$ and $[\bar{\mu}]$ are tensors. Substituting (2.26) into the curl of (2.25) gives the the vector Helmholtz equation

$$\nabla \times \left([\bar{\mu}]^{-1} \cdot \nabla \times \mathbf{E} \right) - \omega^2 [\bar{\epsilon}] \cdot \mathbf{E} + j\omega \mathbf{J} = \mathbf{0}. \quad (2.27)$$

Then, for any vector function \mathbf{F} ,

$$\iiint_{\Omega} \left\{ \nabla \times \left([\bar{\mu}]^{-1} \cdot \nabla \times \mathbf{E} \right) - \omega^2 [\bar{\epsilon}] \cdot \mathbf{E} + j\omega \mathbf{J} \right\} \cdot \mathbf{F} dV = 0. \quad (2.28)$$

Assuming that $\nabla \times \mathbf{F}$ exists, then by Green's first identity one obtains

$$\begin{aligned} \iiint_{\Omega} (\nabla \times \mathbf{F}) \cdot [\bar{\mu}]^{-1} \cdot (\nabla \times \mathbf{E}) - \omega^2 \mathbf{F} \cdot [\bar{\epsilon}] \cdot \mathbf{E} dV \\ + \oint_S \left(\hat{\mathbf{n}} \times \left([\bar{\mu}]^{-1} \cdot \nabla \times \mathbf{E} \right) \right) \cdot \mathbf{F} dS = -j\omega \iiint_{\Omega} \mathbf{F} \cdot \mathbf{J} dV \end{aligned} \quad (2.29)$$

which reduces to

$$\begin{aligned} \iiint_{\Omega} (\nabla \times \mathbf{F}) \cdot [\bar{\mu}]^{-1} \cdot (\nabla \times \mathbf{E}) - \omega^2 \mathbf{F} \cdot [\bar{\epsilon}] \cdot \mathbf{E} dV \\ - j\omega \oint_S (\hat{\mathbf{n}} \times \mathbf{H}) \cdot \mathbf{F} dS = -j\omega \iiint_{\Omega} \mathbf{F} \cdot \mathbf{J} dV. \end{aligned} \quad (2.30)$$

Now assuming that \mathbf{H} is continuous across inter-element boundaries, and that the outer boundary is truncated with PEC backed PML, one obtains

$$\iiint_{\Omega} (\nabla \times \mathbf{F}) \cdot [\bar{\mu}]^{-1} \cdot (\nabla \times \mathbf{E}) - \omega^2 \mathbf{F} \cdot [\bar{\epsilon}] \cdot \mathbf{E} dV = -j\omega \iiint_{\Omega} \mathbf{F} \cdot \mathbf{J} dV. \quad (2.31)$$

To ensure that the assumptions that were made (that is, $\nabla \times \mathbf{F}$ exists and \mathbf{H} is continuous across inter-element boundaries) are satisfied, the finite element basis functions will be chosen to be a curl-conforming, tangentially continuous vector basis. In particular, the FEM basis functions used are those shown in [33]. Once the FEM matrix is assembled, it is of the form

$$\mathbf{A}(s)\mathbf{x}(s) = \mathbf{b}(s) \quad (2.32)$$

which can be put into the form

$$\sum_{i=0}^{a_1} (\sigma^i \mathbf{A}_i) \mathbf{x}(f) = \sum_{k=0}^{b_1} \sigma^k \mathbf{b}_k. \quad (2.33)$$

by interpolation. More precisely, the reason that (2.32) is not quadratic is because the PML materials are dispersive; therefore, once (2.32) is formed, only negligible error is introduced by dropping the higher order σ terms. As will be shown in the following examples, using $a_1 = b_1 = 2$ and interpolating at f_{min} , $(f_{min} + f_{max})/2$ and f_{max} seems to be sufficient for many problems of interest.

Example 4: A microwave low pass filter is discretized and the curl-conforming, tangentially continuous vector basis functions are used to create the FEM model. The model contains 26044 unknowns and is simulated from 2GHz to 20GHz. Figures 2.5 and 2.6 show the S parameters for this microwave device for both the solution to the original equation (2.32) and the quadratic approximation (2.33).

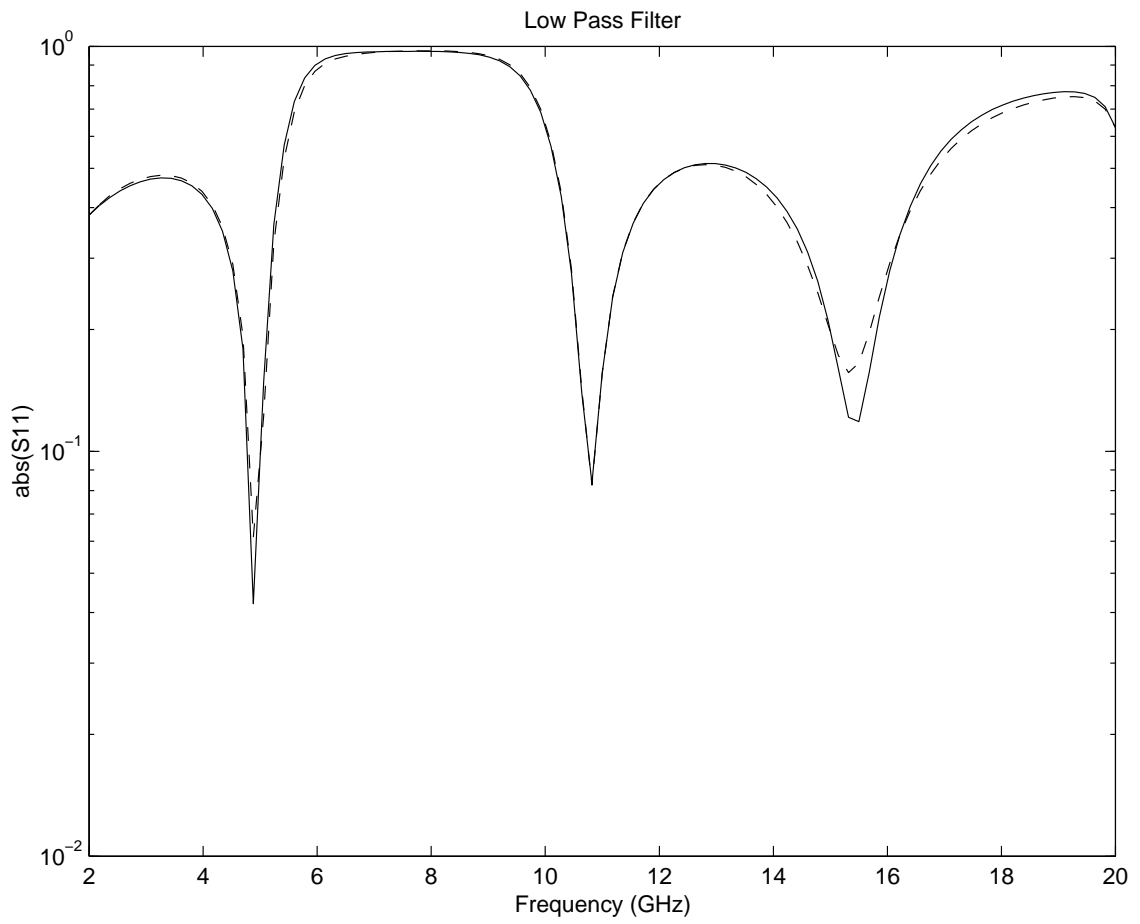


Figure 2.5: S_{11} for the low pass filter. Solid \rightarrow solution to (2.32), dash-dash \rightarrow solution to quadratic approximation.

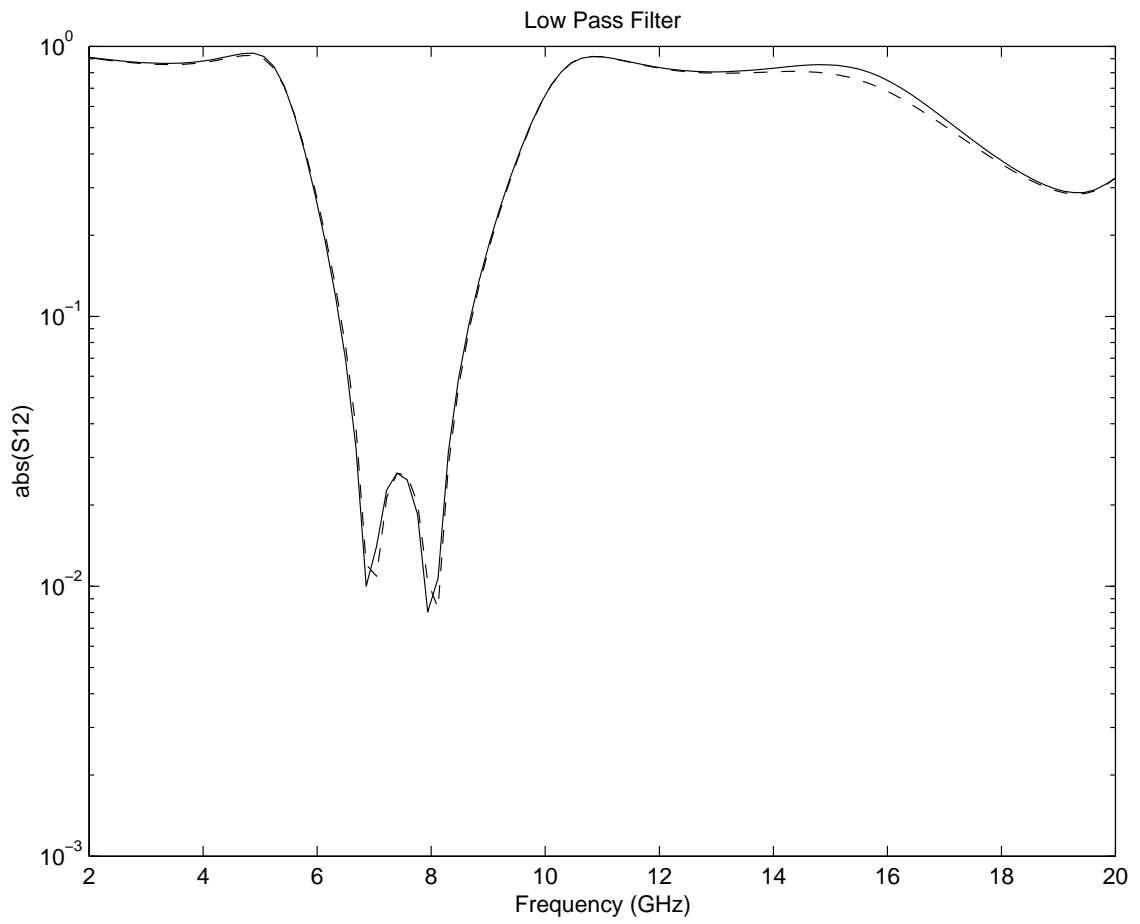


Figure 2.6: S_{12} for the low pass filter. Solid \rightarrow solution to (2.32), dash-dash \rightarrow solution to quadratic approximation.

Example 5: This numerical example is a broadband bowtie antenna which is placed on a half-sphere absorber and simulated from 500MHz to 5GHz. A total of 884670 unknowns is used to model the geometry. Figure 2.7 shows the input S_{11} for this antenna for both equation (2.32) and the quadratic approximation (2.33). In this case, for the scale shown, the curves are indistinguishable.

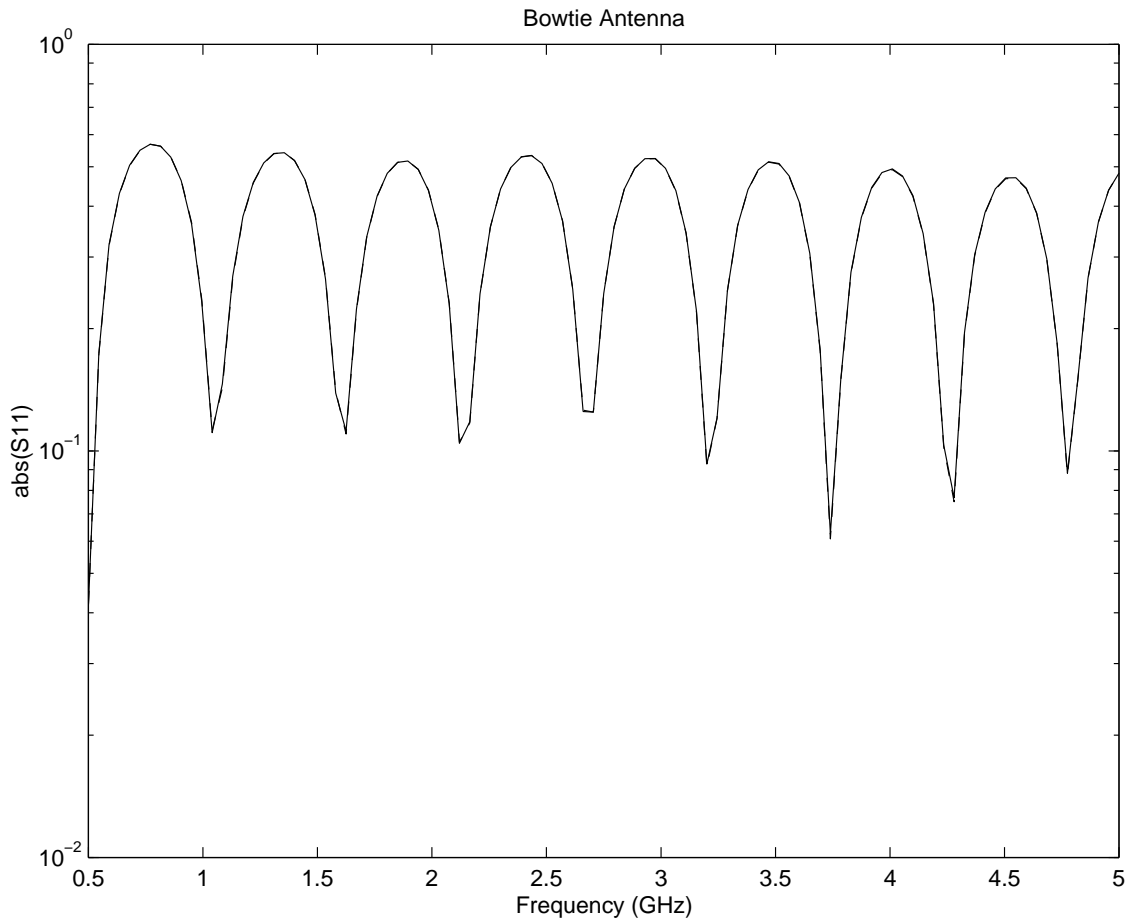


Figure 2.7: S_{11} for the bowtie antenna. Solid \rightarrow solution to (2.32), dash-dash \rightarrow solution to quadratic approximation.

Example 6: In this numerical example, a band pass filter is simulated from 3.9GHz to 4.1GHz. There are 905892 unknowns in this model. Figure 2.8 shows the magnitude of S_{12} versus frequency in the specified bandwidth for both equation (2.32) and the quadratic approximation (2.33). Again, for the scale shown, there is no difference in the solutions.

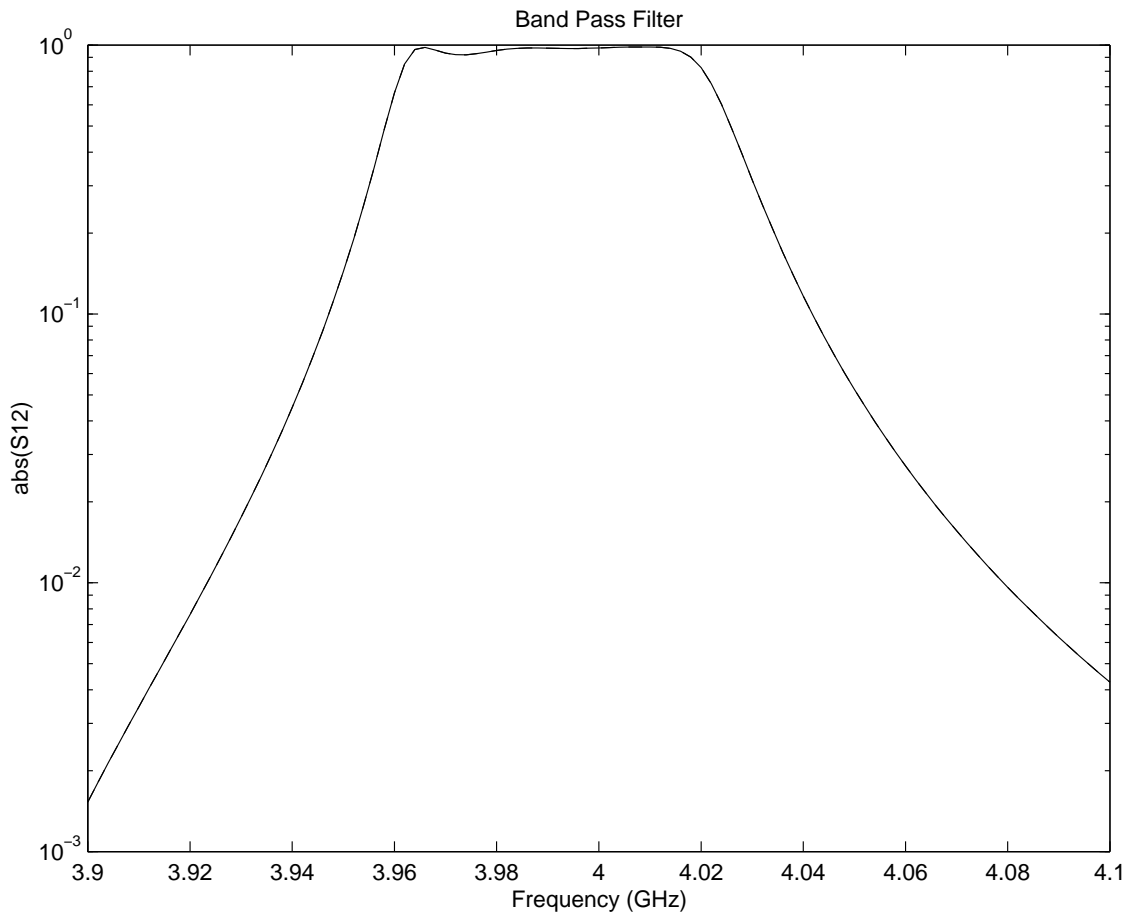


Figure 2.8: S_{12} for the band pass filter. Solid \rightarrow solution to (2.32), dash-dash \rightarrow solution to quadratic approximation.

CHAPTER 3

CLASSICAL MORE TECHNIQUES

3.1 Krylov subspace techniques for linear equations

This section covers MORE techniques that operate on matrix equations that are linear in the MORE parameter σ , and that have constant right hand sides. There are two major techniques that are commonly used for this type of problem. They are the Lanczos [8] and the Arnoldi [15] processes, which were popularized in the MORE community by the works [9, 10, 34, 35]. Although both techniques produce Krylov subspaces in the reduction process, only the Lanczos process produces Padé approximants. Nevertheless, there is no consensus on which technique is superior.

Starting from (2.5) with the output equation (2.7) and assuming a single excitation vector ($p = 1$) one obtains

$$\sum_{i=0}^{a_1} (\sigma^i \mathbf{A}_i) \mathbf{x}(f) = \sum_{k=0}^{b_1} \sigma^k \mathbf{b}_k \quad (3.1)$$

and

$$\mathbf{h}(f) = \mathbf{L}^T \mathbf{x}(f). \quad (3.2)$$

Next, define the following quantities.

Definition 3.1 For any integer i , let

$$\delta_{i0} = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

□

Definition 3.2 Let $c_1 = \max(a_1, b_1)$. For any integer i such that $0 \leq i \leq c_1$ let the matrices

$$\mathbf{M}_i = \begin{bmatrix} \mathbf{A}_i & \mathbf{b}_i \\ \mathbf{0}_{1 \times N} & -\delta_{i0} \end{bmatrix} \quad (3.3)$$

where \mathbf{M}_i is an $(N + 1) \times (N + 1)$ complex matrix and $\mathbf{A}_i = \mathbf{0}_{N \times N}$ for $i > a_1$ or $\mathbf{b}_i = \mathbf{0}_{N \times 1}$ for $i > b_1$. □

Definition 3.3 Let r be some positive integer. Then \mathbf{e}_r is a vector with all entries equal to zero except the r th entry which is equal to 1. The length of \mathbf{e}_r conforms to the matrix that operates on it. □

Now using the above definitions, (3.1) becomes

$$\sum_{i=0}^{c_1} (\sigma^i \mathbf{M}_i) \bar{\mathbf{x}}(f) = \mathbf{e}_{N+1} \quad (3.4)$$

where

$$\bar{\mathbf{x}}(f) = \begin{bmatrix} \mathbf{x}(f) \\ -1 \end{bmatrix}. \quad (3.5)$$

Note that (3.4) has a constant right hand side.

It is now possible to use the “Local Approximations” given in section V of [16] to linearize (3.4) with respect to σ . Unlike other linearization techniques (such as the one shown in [12]), the brilliant work [16] permits the expanded, linearized system to

be written in such a way as to allow the order of the matrix, whose inverse is used as an operator, to remain at $N + 1$. The result is

$$(\mathbf{C} - \sigma \mathbf{D}) \mathbf{z}(f) = \mathbf{y} \quad (3.6)$$

with

$$\mathbf{h}(f) = \bar{\mathbf{L}}^T \mathbf{z}(f) \quad (3.7)$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{M}_0 & \mathbf{M}_1 & \mathbf{M}_2 & \cdots & \mathbf{M}_{c_1-1} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & -\mathbf{M}_{c_1} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix},$$

$$\mathbf{z}(f) = \begin{bmatrix} \bar{\mathbf{x}}(f) \\ \mathbf{z}^{(2)} \\ \mathbf{z}^{(3)} \\ \vdots \\ \mathbf{z}^{(c_1)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{e}_{N+1} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \quad \text{and} \quad \bar{\mathbf{L}} = \begin{bmatrix} \mathbf{L} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$$

Next, it can be shown that

$$\mathbf{C}^{-1} = \begin{bmatrix} \mathbf{M}_0^{-1} & -\mathbf{M}_0^{-1}\mathbf{M}_1 & -\mathbf{M}_0^{-1}\mathbf{M}_2 & \cdots & -\mathbf{M}_0^{-1}\mathbf{M}_{c_1-1} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}. \quad (3.8)$$

Multiplying both sides of (3.6) by \mathbf{C}^{-1} gives

$$(\mathbf{I} - \sigma \mathbf{C}^{-1} \mathbf{D}) \mathbf{z}(f) = \mathbf{C}^{-1} \mathbf{y} \quad \text{with} \quad \mathbf{h}(f) = \bar{\mathbf{L}}^T \mathbf{z}(f) \quad (3.9)$$

where

$$\mathbf{C}^{-1}\mathbf{D} = \begin{bmatrix} -\mathbf{M}_0^{-1}\mathbf{M}_1 & -\mathbf{M}_0^{-1}\mathbf{M}_2 & -\mathbf{M}_0^{-1}\mathbf{M}_3 & \cdots & -\mathbf{M}_0^{-1}\mathbf{M}_{c_1} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix}$$

and $\mathbf{C}^{-1}\mathbf{y} = \begin{bmatrix} \mathbf{M}_0^{-1}\mathbf{e}_{N+1} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$ (3.10)

Now, either the projection via Arnoldi (PVA) [16] or the matrix Padé via Lanczos (MPVL) [27] process can be applied to (3.9) since the equation is linear in σ and has a constant excitation vector.

3.1.1 Projection via Arnoldi (PVA) review

The governing equations for the Arnoldi process on the matrix $\mathbf{C}^{-1}\mathbf{D}$ with starting vector

$$\mathbf{z}_1 = \mathbf{C}^{-1}\mathbf{y}/\|\mathbf{C}^{-1}\mathbf{y}\| \quad (3.11)$$

for q steps are

$$\mathbf{C}^{-1}\mathbf{D}\mathbf{Z}_q = \mathbf{Z}_q\mathbf{U}_{\mathbf{H}q} + \mathbf{U}_{\mathbf{H}}(q+1, q)\mathbf{z}_{q+1}\mathbf{e}_q^T \quad (3.12)$$

and

$$\mathbf{Z}_q^H\mathbf{Z}_q = \mathbf{I}_q \quad (3.13)$$

where \mathbf{Z}_n is the collection of the n Arnoldi vectors \mathbf{z}_r for $r = 1, 2, \dots, n$ and $\mathbf{U}_{\mathbf{H}n}$ is a complex $n \times n$ upper Hessenberg matrix. These governing equations give rise to algorithm 3.1.

Algorithm 3.1 (q steps of the Arnoldi process)

for $n = 1, 2, \dots, q$ do

 set $\mathbf{z}_{n+1} = \mathbf{C}^{-1}\mathbf{D}\mathbf{z}_n$

 for $j = 1, 2, \dots, n$ do

 set $\mathbf{U}_{\mathbf{H}}(j, n) = \mathbf{z}_j^H \mathbf{z}_{n+1}$

 set $\mathbf{z}_{n+1} = \mathbf{z}_{n+1} - \mathbf{U}_{\mathbf{H}}(j, n)\mathbf{z}_j$

 endfor

 set $\mathbf{U}_{\mathbf{H}}(n+1, n) = \|\mathbf{z}_{n+1}\|$

 set $\mathbf{z}_{n+1} = \mathbf{z}_{n+1}/\mathbf{U}_{\mathbf{H}}(n+1, n)$

endfor

□

Once \mathbf{Z}_q is computed, $\mathbf{z}(f)$ can be approximated by

$$\mathbf{z}(f) \approx \mathbf{Z}_q \mathbf{g}_{\mathbf{A}_q}(f) \quad (3.14)$$

where $\mathbf{g}_{\mathbf{A}_q}(f) \in \mathbb{C}^{q \times 1}$ are the frequency dependent weighting coefficients for the Arnoldi basis vectors \mathbf{Z}_q that span the Krylov subspace. Substitute (3.14) into the first equation in (3.9) and perform a Galerkin test to give

$$\mathbf{Z}_q^H (\mathbf{I} - \sigma \mathbf{C}^{-1}\mathbf{D}) \mathbf{Z}_q \mathbf{g}_{\mathbf{A}_q}(f) = \mathbf{Z}_q^H \mathbf{C}^{-1}\mathbf{y} \quad (3.15)$$

which, from the Arnoldi process governing equations (3.12) and (3.13), results in

$$\mathbf{g}_{\mathbf{A}_q}(f) = (\mathbf{I} - \sigma \mathbf{U}_{\mathbf{H}_q})^{-1} \mathbf{Z}_q^H \mathbf{C}^{-1}\mathbf{y}. \quad (3.16)$$

Now note that $\mathbf{Z}_q^H \mathbf{C}^{-1}\mathbf{y} = \mathbf{e}_1 \|\mathbf{C}^{-1}\mathbf{y}\|$ from the way \mathbf{z}_1 was chosen in (3.11). Therefore,

$$\mathbf{z}(f) \approx \mathbf{z}_q(f) = \mathbf{Z}_q (\mathbf{I} - \sigma \mathbf{U}_{\mathbf{H}_q})^{-1} \mathbf{e}_1 \|\mathbf{C}^{-1}\mathbf{y}\| \quad \text{with} \quad \mathbf{h}_q(f) = \bar{\mathbf{L}}^T \mathbf{z}_q(f). \quad (3.17)$$

3.1.2 Padé via Lanczos (PVL) review

In [27] a matrix-PVL (MPVL) algorithm for multiple starting vectors is reported which has the ability to produce simultaneously output for many different unknowns. In general, assume that there are p inputs and o outputs to the matrix system. Then the governing equation for the iterative MPVL process for q steps with exact deflation and no look-ahead are

$$\mathbf{C}^{-1}\mathbf{D}\Psi_q = \Psi_q\mathbf{T}_q + \underbrace{\left[\mathbf{0} \dots \mathbf{0} \right]}_{q-p_c} \underbrace{\left[\hat{\psi}_{q+1} \dots \hat{\psi}_{q+p_c} \right]}_{p_c} \quad (3.18)$$

and

$$(\mathbf{C}^{-1}\mathbf{D})^T \Xi_q = \Xi_q \tilde{\mathbf{T}}_q + \underbrace{\left[\mathbf{0} \dots \mathbf{0} \right]}_{q-o_c} \underbrace{\left[\hat{\xi}_{q+1} \dots \hat{\xi}_{q+o_c} \right]}_{o_c} \quad (3.19)$$

where Ψ_n and Ξ_n are the collection of the n right and left Lanczos vectors ψ_r and ξ_r for $r = 1, 2, \dots, n$, \mathbf{T}_n and $\tilde{\mathbf{T}}_n$ are banded $n \times n$ matrices which have the same eigenvalues, and the integers p_c and o_c are the current right and left block sizes. Furthermore, the right and left vector spaces Ψ_n and Ξ_n are constructed to be biorthogonal. These governing equations give rise to algorithm A.1 shown in appendix A, and result in the approximation

$$\mathbf{H}(f) \approx \mathbf{H}_q(f) = \boldsymbol{\eta}^T (\mathbf{I} - \sigma \mathbf{T}_q)^{-1} \boldsymbol{\rho} \quad (3.20)$$

where $\boldsymbol{\eta} \in \mathbb{C}^{q \times o}$, $\mathbf{T}_q \in \mathbb{C}^{q \times q}$ and $\boldsymbol{\rho} \in \mathbb{C}^{q \times p}$ are generated during the execution of algorithm A.1. As shown in [27], assuming no deflation, the number of moments matched in (3.20) is $\lfloor q/o \rfloor + \lfloor q/p \rfloor$.

3.2 Moment matching techniques for polynomial equations

This section covers MORE techniques that operate on matrix equations and right hand sides that contain polynomial variations in the MORE parameter σ . The major technique in this area is asymptotic waveform evaluation (AWE) [17, 36]. A newer version of AWE is the Galerkin AWE (GAWE) [37] which, unlike AWE, does not form Padé approximants. However, both techniques match moments, and it will be shown that the matrix $\mathbf{W}_q \in \mathbb{C}^{N \times q}$ where $\mathbf{W}_q = [\mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_q]$ with

$$\begin{aligned} \mathbf{w}_1 &= \mathbf{A}_0^{-1} \mathbf{b}_0 \\ \mathbf{w}_2 &= \mathbf{A}_0^{-1} (\mathbf{b}_1 - \mathbf{A}_1 \mathbf{w}_1) \\ \mathbf{w}_3 &= \mathbf{A}_0^{-1} (\mathbf{b}_2 - \mathbf{A}_1 \mathbf{w}_2 - \mathbf{A}_2 \mathbf{w}_1) \\ &\vdots \\ \mathbf{w}_q &= \mathbf{A}_0^{-1} \left(\mathbf{b}_{q-1} - \sum_{m=1}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{w}_{q-m} \right) \end{aligned} \tag{3.21}$$

and $\mathbf{b}_k = \mathbf{0}$ for $k > b_1$ plays a critical role in these processes.

3.2.1 Asymptotic waveform evaluation (AWE) review

Starting from equation (3.1) and expanding the unknown solution vector $\mathbf{x}(f)$ into a Taylor series around s_0 gives

$$\mathbf{x}(f) = \sum_{n=0}^{\infty} \sigma^n \mathbf{m}_n \tag{3.22}$$

where each of the \mathbf{m}_n is a N -vector moment. To obtain a Padé approximant of order Q , there must be $2Q$ moments generated which means that moments up to and including order $2Q - 1$ must be generated. Substituting (3.22) into (3.1) and

performing moment matching gives

$$\mathbf{m}_n = \mathbf{w}_{n+1} \quad \text{for } 0 \leq n \leq q-1 \quad (3.23)$$

where $q = 2Q$ and \mathbf{w}_n is given in (3.21). Then let

$$\mathbf{h}_n = \mathbf{L}^T \mathbf{m}_n \quad (3.24)$$

and denote the r th entry of the o -vector \mathbf{h}_n as η_n^r . Finally, for each of the o unknowns desired as outputs, let $r = 1, 2, \dots, o$ and form the Padé approximant by finding c_t^r for $t = 0, 1, \dots, Q-1$ and d_u^r for $u = 1, 2, \dots, Q$ from

$$\frac{\sum_{t=0}^{Q-1} \sigma^t c_t^r}{1 + \sum_{u=1}^Q \sigma^u d_u^r} = \sum_{n=0}^{q-1} \sigma^n \eta_n^r \quad (3.25)$$

which requires solving the system

$$\begin{bmatrix} \eta_0^r & \eta_1^r & \eta_2^r & \cdots & \eta_{Q-1}^r \\ \eta_1^r & \eta_2^r & \eta_3^r & \cdots & \eta_Q^r \\ \eta_2^r & \eta_3^r & \eta_4^r & \cdots & \eta_{Q+1}^r \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \eta_{Q-1}^r & \eta_Q^r & \eta_{Q+1}^r & \cdots & \eta_{2Q-2}^r \end{bmatrix} \begin{bmatrix} d_Q^r \\ d_{Q-1}^r \\ d_{Q-2}^r \\ \vdots \\ d_1^r \end{bmatrix} = - \begin{bmatrix} \eta_Q \\ \eta_{Q+1} \\ \eta_{Q+2} \\ \vdots \\ \eta_{2Q-1} \end{bmatrix} \quad (3.26)$$

and

$$\begin{aligned} c_0^r &= \eta_0^r \\ c_1^r &= \eta_1^r + d_1^r \eta_0^r \\ c_2^r &= \eta_2^r + d_2^r \eta_0^r + d_1^r \eta_1^r \\ &\vdots \\ c_{Q-1}^r &= \eta_{Q-1}^r + \sum_{i=1}^{Q-1} d_{Q-i}^r \eta_{i-1}^r. \end{aligned} \quad (3.27)$$

Then the r th desired output of the o -vector $\mathbf{h}(f)$, denoted by $h^r(f)$, is given by

$$h^r(f) \approx h_q^r(f) = \frac{\sum_{t=0}^{Q-1} \sigma^t c_t^r}{1 + \sum_{u=1}^Q \sigma^u d_u^r}. \quad (3.28)$$

3.2.2 Galerkin AWE (GAWE) review

Start from equation (3.1) and assume there is a collection of q linearly independent N -vectors $\bar{\mathbf{w}}_n$ and q scalars $\gamma_n(f)$ for $n = 1, 2, \dots, q$. Let $\overline{\mathbf{W}}_q = [\bar{\mathbf{w}}_1 \bar{\mathbf{w}}_2 \dots \bar{\mathbf{w}}_q]$ and define a q -vector $\mathbf{g}_q(f)$ such that the n th component in $\mathbf{g}_q(f)$ is $\gamma_n(f)$. The quantities $\bar{\mathbf{w}}_n$ and $\gamma_n(f)$ are chosen such that the approximation

$$\mathbf{x}(f) \approx \mathbf{x}_q(f) = \overline{\mathbf{W}}_q \mathbf{g}_q(f) = \sum_{n=1}^q \bar{\mathbf{w}}_n \gamma_n(f) \quad (3.29)$$

minimizes the residual

$$\mathbf{r}_q(f) = \sum_{i=0}^{a_1} (\sigma^i \mathbf{A}_i) \sum_{n=1}^q \bar{\mathbf{w}}_n \gamma_n(f) - \sum_{k=0}^{b_1} \sigma^k \mathbf{b}_k \quad (3.30)$$

in the sense that if $\mathbf{r}_q(f)$ is expressed in a Taylor series as

$$\mathbf{r}_q(f) = \sum_{l=0}^{\infty} \sigma^l \mathbf{r}_q^l \quad (3.31)$$

then

$$\mathbf{r}_q^l = \mathbf{0} \quad \text{for} \quad l = 0 \dots q - 1 \quad (3.32)$$

and

$$\mathbf{r}_q(f) \perp \overline{\mathbf{W}}_q. \quad (3.33)$$

A proof is given in appendix B which shows that choosing $\overline{\mathbf{W}}_q = \mathbf{W}_q$ from (3.21) satisfies (3.32). Of course, in practice the vectors $\bar{\mathbf{w}}_n$ for $n = 1 \dots q$ are actually chosen to be an orthonormal basis for the space \mathbf{W}_q . Another difference between AWE and GAWE is that instead of performing a Padé approximation, $\mathbf{g}_q(f)$ is made to satisfy (3.33) and is found from

$$\mathbf{g}_q(f) = \left(\sum_{i=0}^{a_1} \sigma^i \overline{\mathbf{W}}_q^T \mathbf{A}_i \overline{\mathbf{W}}_q \right)^{-1} \left(\sum_{k=0}^{b_1} \sigma^k \overline{\mathbf{W}}_q^T \mathbf{b}_k \right). \quad (3.34)$$

Then

$$\mathbf{h}(f) \approx \mathbf{h}_q(f) = \mathbf{L}^T \overline{\mathbf{W}}_q \mathbf{g}_q(f). \quad (3.35)$$

3.3 Numerical comparisons

Example 1: The horn antenna described in subsection 2.2.1 is simulated using the MPVL, AWE and GAWE techniques (PVA was not used for this example because \mathbf{L} extracts only three unknowns and so MPVL will match more moments per iteration than Arnoldi). A total of 300 iterations were performed for MPVL, and 20 iterations were performed for both AWE and GAWE. All techniques used a single expansion point corresponding to 1GHz. Figure 3.1 shows the impedance calculated using each method, along with the exact solution to the matrix equation computed using an LU decomposition. It is clear that MPVL is accurate in a much wider bandwidth than either AWE or GAWE. However, note that even with 300 iterations (which is 15 times larger than either the AWE or the GAWE subspace generated) MPVL is still not totally indistinguishable throughout the entire simulated band. Therefore, this example indicates that a practical MORE solution methodology probably should be a multipoint technique, unless switching expansion points is very computationally expensive.

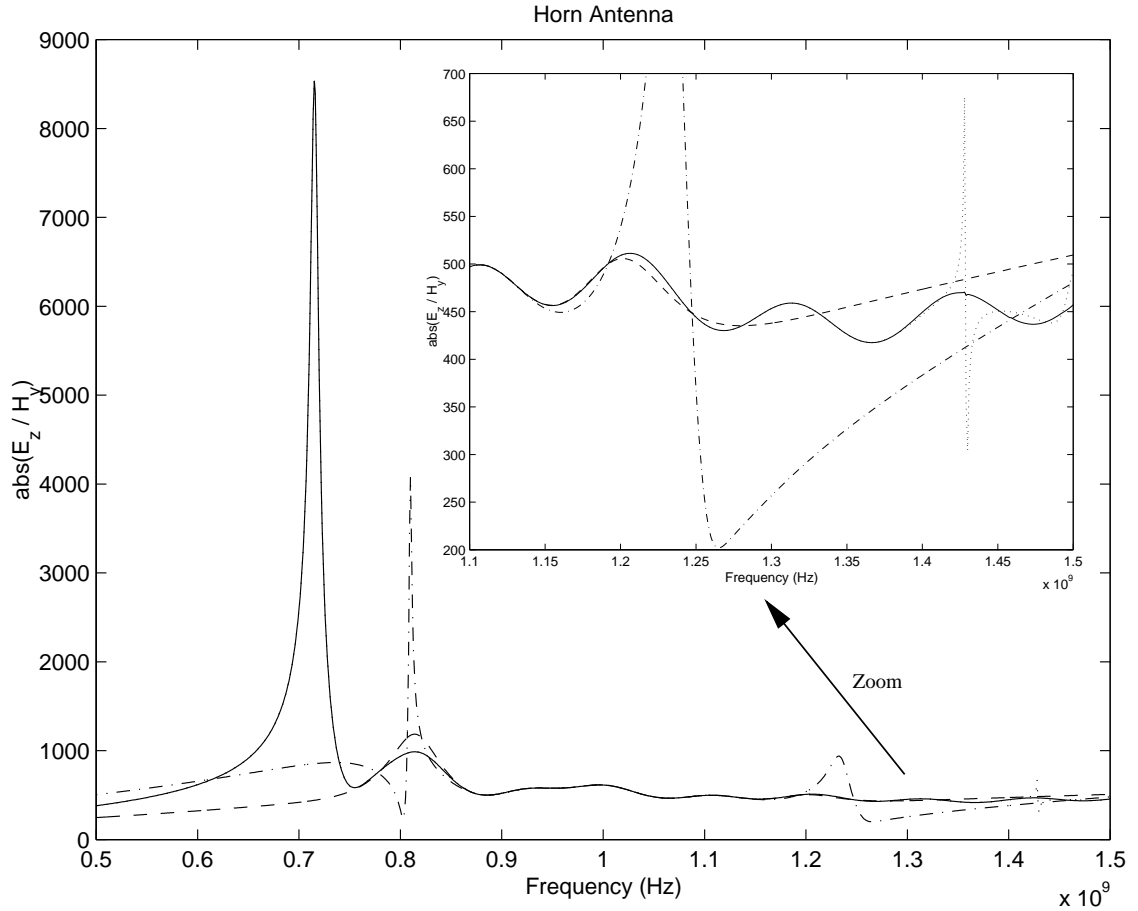


Figure 3.1: Impedance calculated for the horn antenna. Solid \rightarrow **LU** response, dash-dot \rightarrow **AWE** response, dash-dash \rightarrow **GAWE** response, dot-dot \rightarrow **MPVL** response (almost indistinguishable from **LU** response except for a slight deviation in zoom mode near $f = 1.425 \times 10^9$ Hz).

Example 2: The TM_z radiation problem terminated with an anisotropic, dispersive PML discussed in subsection 2.2.2 is simulated using the PVA, AWE and GAWE techniques. Since the entire solution vector was calculated with $\mathbf{L} = \mathbf{I}$, MPVL was not used in this simulation. A total of 100 iterations were performed for PVA, and 30 iterations were performed for both AWE and GAWE; all the methods used a single expansion point corresponding to 250MHz. Figure 3.2 shows the relative error (measured with the 1-norm) in the solution vector for each of the methods, that is,

$$\frac{\|\mathbf{h}_q(f) - \mathbf{h}(f)\|_1}{\|\mathbf{h}(f)\|_1} \quad (3.36)$$

with $\mathbf{L} = \mathbf{I}$ so $\mathbf{h}(f) = \mathbf{x}(f)$. Of course, as expected, the error increases for frequencies further removed from the expansion point. Note that although PVA is much more accurate, it used many more iterations. Again, as in the first example, a practical MORE solution methodology probably should be a multipoint technique, unless (as previously noted) switching expansion points is very computationally expensive.

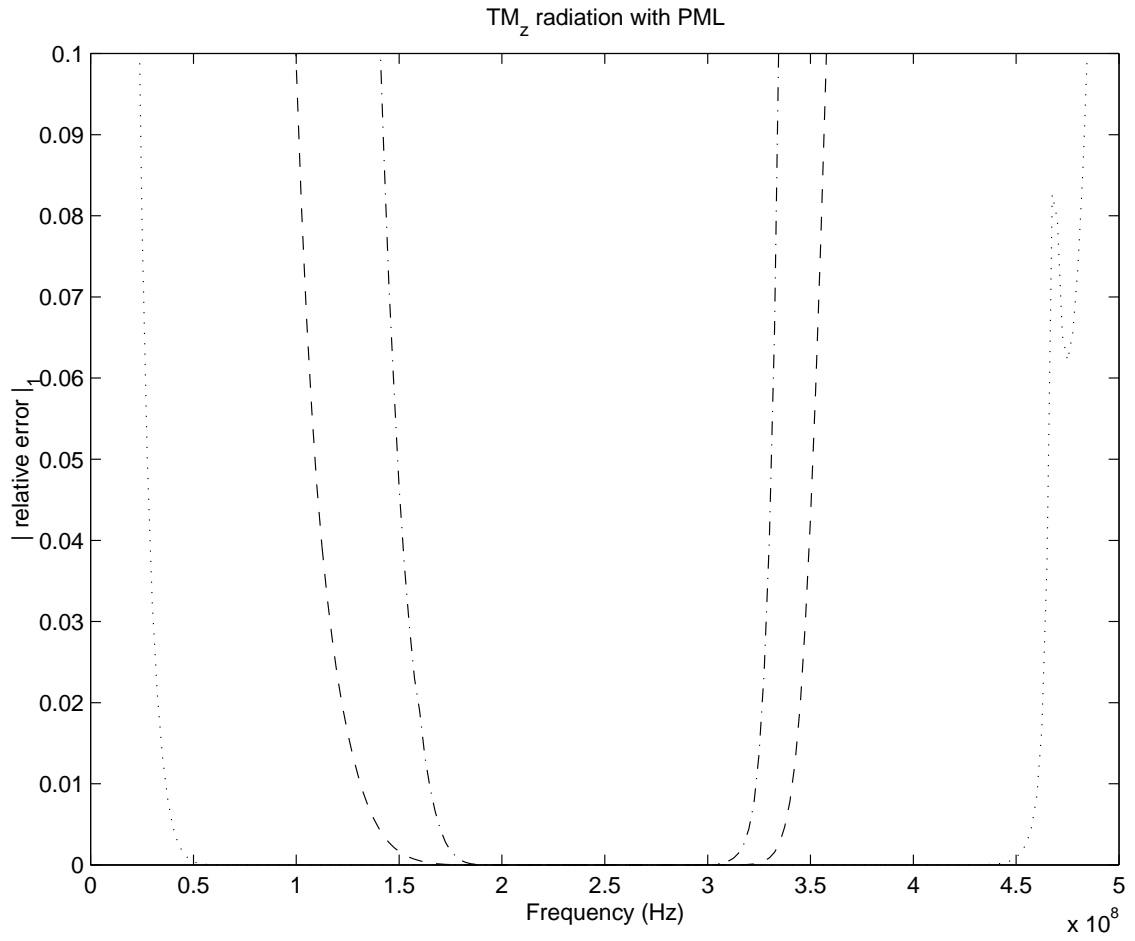


Figure 3.2: Relative error in the solution vector for the example in subsection 2.2.2. Dash-dot \rightarrow AWE method, dash-dash \rightarrow GAWE method, dot-dot \rightarrow PVA method.

Example 3: The TE_z scattering problem from subsection 2.2.3 is solved. Again, the PVA, AWE and GAWE techniques were used with $\mathbf{L} = \mathbf{I}$. For each technique a total of 30 iterations were performed with the expansion point chosen to correspond to 250MHz. Figure 3.3 shows the relative error (3.36) in the solution vector for each method with $\mathbf{L} = \mathbf{I}$. Since a small number of iterations were used for each method, it is not surprising that the Arnoldi method is not significantly more accurate than the other methods. Of course, for more iterations the Arnoldi method would not stagnate like the AWE and GAWE methods would. However, as pointed out in the previous two examples, a practical MORE solution methodology probably should be a multipoint technique. Therefore, in the following chapter, a multipoint technique will be presented. Even though the single point Arnoldi method is more accurate than GAWE for a large number of iterations, the multipoint technique will be based on GAWE. This is because of three reasons. First, for a small number of iterations, GAWE is essentially as accurate as PVA. Second, for a multipoint technique, only a small number of iterations will be performed at each expansion point. Third, GAWE requires less memory to store the vectors than PVA (because, unless the higher order terms are truncated, the system must be expanded and linearized for PVA).

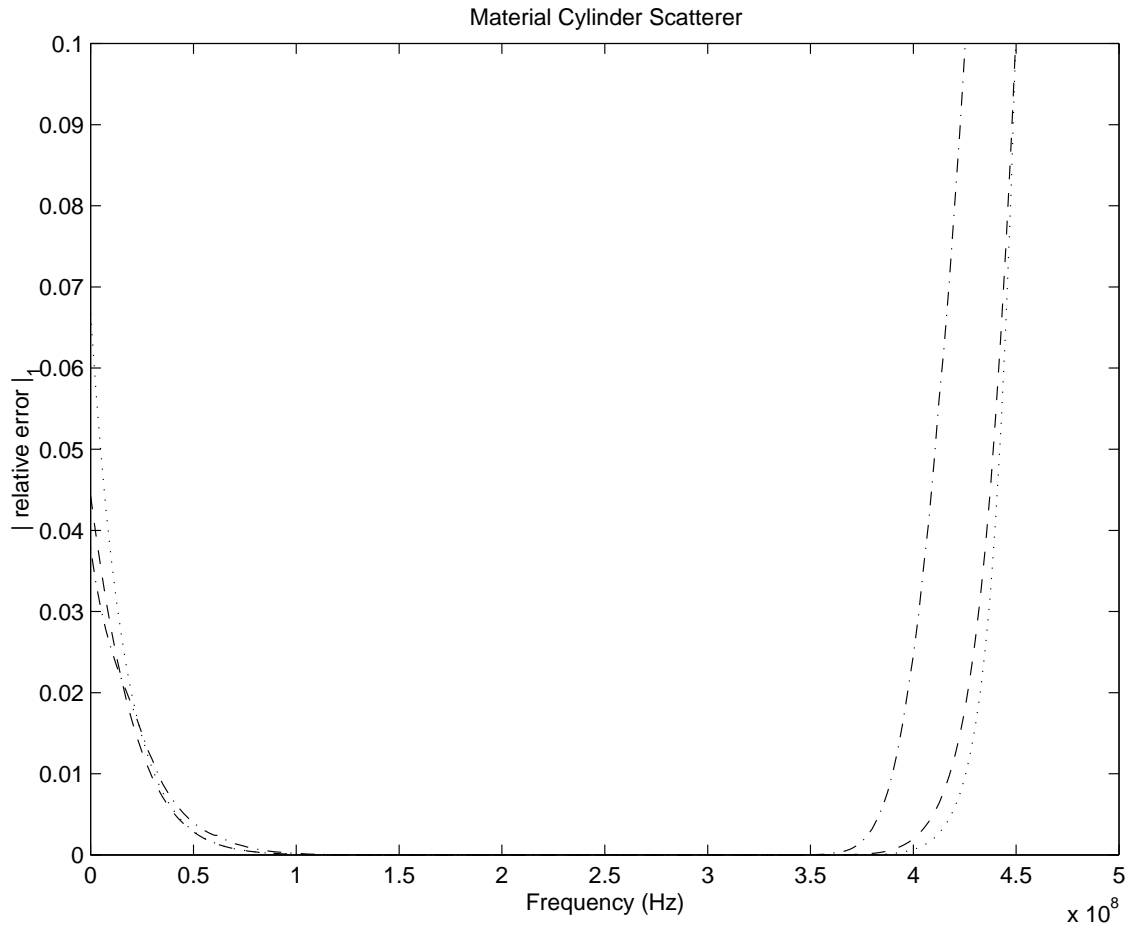


Figure 3.3: Relative error in the solution vector for the example in subsection 2.2.3. Dash-dot \rightarrow AWE method, dash-dash \rightarrow GAWE method, dot-dot \rightarrow PVA method.

CHAPTER 4

MULTIPOINT GALERKIN ASYMPTOTIC WAVEFORM EVALUATION (MGAWE)

4.1 General MGAWE information

In chapter 3 several single expansion point MORE techniques were presented. Although the Lanczos and Arnoldi methods are more broadband than the AWE methods, they require the polynomial matrix equation to be linearized with respect to σ , which results in ROM vectors that are of length $c_1(N+1)$ instead of the original length N . In MORE simulations involving sparse matrices, the memory required to store the ROM vectors can be greater than the memory required to store the system matrices. Therefore, expanding and linearizing the equations and applying Lanczos or Arnoldi is not an option; as a result, this chapter is devoted to developing a broadband, multipoint version of the GAWE technique.

Several practical implementation issues must be addressed to make the MGAWE technique viable. These issues include: how many expansion points to use, where to pick them, and how large the approximation order at each expansion point should

become before the iterative process is terminated and the MGAWF solution is declared to have converged to the true solution. These issues have already been addressed for some other MORE techniques. For example, in [18] and [19] the issues are addressed for AWF by considering complex frequency hopping (CFH). In addition, [10, 26, 38, 39] address one or more of these issues for PVL.

Recall equations (2.3) and (2.4) from Definition 2.1. They are

$$\sigma_{v_u} = j2\pi f_u - s_{0_v} \quad \text{for } v \in \nu, u = 1, 2, \dots, f_{num}$$

and

$$\sigma_v = j2\pi f - s_{0_v} \quad \text{for } v \in \nu$$

where $\nu = \{1, 2, \dots, num_pts\}$, num_pts is the total number of expansion points (section 4.3 shows how to automatically determine the exact value for num_pts) and s_{0_v} is the location of the v th expansion point (sections 4.2 and 4.3 show how to automatically determine this location). Again assuming $p = 1$, (2.5) and (2.7) can be rewritten as

$$\sum_{i=0}^{a_1} (\sigma^i \mathbf{A}_i) \mathbf{x}(f) = \sum_{k=0}^{b_1} \sigma^k \mathbf{b}_k \quad \text{and} \quad \mathbf{h}(f) = \mathbf{L}^T \mathbf{x}(f) \quad (4.1)$$

where σ is as given in (2.6) with lax restrictions such as those in (2.5).

As in GAWF, to solve the above problem assume there is a collection of q linearly independent N -vectors $\bar{\mathbf{w}}_n$ and q scalars $\gamma_n(f)$, but now let

$$q = \sum_{v=1}^{num_pts} q_v, \quad (4.2)$$

where q_v is the order of the approximation generated at s_{0_v} (section 4.2 shows how to automatically determine the size of each q_v), and $n = 1, 2, \dots, q$. As before, choose

$\bar{\mathbf{w}}_n$ and $\gamma_n(f)$ so the approximation

$$\mathbf{x}(f) \approx \mathbf{x}_q(f) = \overline{\mathbf{W}}_q \mathbf{g}_q(f) = \sum_{n=1}^q \bar{\mathbf{w}}_n \gamma_n(f) \quad (4.3)$$

minimizes the residual

$$\mathbf{r}_q(f) = \sum_{i=0}^{a_1} (\sigma^i \mathbf{A}_i) \sum_{n=1}^q \bar{\mathbf{w}}_n \gamma_n(f) - \sum_{k=0}^{b_1} \sigma^k \mathbf{b}_k \quad (4.4)$$

in the sense that if $\mathbf{r}_q(f)$ is expressed in a Taylor series as

$$\mathbf{r}_q(f) = \sum_{l=0}^{\infty} \sigma_v^l \mathbf{r}_{q_v}^l \quad (4.5)$$

then

$$\begin{aligned} \mathbf{r}_{q_1}^l &= \mathbf{0} & \text{for } l = 0 \dots q_1 - 1 \\ \mathbf{r}_{q_2}^l &= \mathbf{0} & \text{for } l = 0 \dots q_2 - 1 \\ &\vdots \\ \mathbf{r}_{q_{num_pts}}^l &= \mathbf{0} & \text{for } l = 0 \dots q_{num_pts} - 1 \end{aligned} \quad (4.6)$$

and

$$\mathbf{r}_q(f) \perp \overline{\mathbf{W}}_q. \quad (4.7)$$

At the v th expansion point, the q_v vectors shown in (3.21) (with \mathbf{A}_i and \mathbf{b}_k corresponding to σ_v) will satisfy the v th equation in (4.6). To satisfy all the equations in (4.6), choose the vectors $\bar{\mathbf{w}}_n$ for $n = 1, 2, \dots, q$ to be an orthonormal basis for the space

$$\mathbf{W}_{q_1} \cup \mathbf{W}_{q_2} \cup \dots \cup \mathbf{W}_{q_{num_pts}} \quad (4.8)$$

where, again, \mathbf{W}_{q_v} is given by (3.21) with careful attention given to the fact that for each v the set \mathbf{A}_i and \mathbf{b}_k must correspond to σ_v . Then, as before, $\mathbf{g}_q(f)$ is found to satisfy (4.7), that is

$$\mathbf{g}_q(f) = \left(\sum_{i=0}^{a_1} \sigma^i \overline{\mathbf{W}}_q^T \mathbf{A}_i \overline{\mathbf{W}}_q \right)^{-1} \left(\sum_{k=0}^{b_1} \sigma^k \overline{\mathbf{W}}_q^T \mathbf{b}_k \right) \quad (4.9)$$

where the requirements on σ , \mathbf{A}_i and \mathbf{b}_k used in (4.9) are the same as in (4.1), and are therefore less stringent than the requirements in (4.8). Finally, as before,

$$\mathbf{h}(f) \approx \mathbf{h}_q(f) = \mathbf{L}^T \overline{\mathbf{W}}_q \mathbf{g}_q(f). \quad (4.10)$$

4.2 Determining the orders of the subspaces q_v

Although the MGAW algorithm can initiate with any number of initial expansion points, the author suggests that the process start with only one. Of course, more can be added if and when they are needed. Next, this expansion point's location must be specified. If it is chosen as suggested in [26] (where AWE is used) then it will be chosen in the right half plane² so the Taylor series, which is used to generate a Padé approximation, is accurate in a wider bandwidth. However, details in [38] (where MORE is performed by PVL) suggest that convergence should be accelerated; therefore the expansion point should be chosen near the $j\omega$ axis, with a slightly negative real part. However, in [29] the expansion points are chosen *on* the $j\omega$ axis. Since the latter approach will be beneficial in section 4.3, expansion points in this study are constrained to the $j\omega$ axis in the range

$$j2\pi f_{min} \leq s_{0_v} \leq j2\pi f_{max}, \quad (4.11)$$

²A different definition for s is given in [26] which maps the right half plane to the lower half plane.

the initial expansion point is chosen at the closest evaluation point of f to the center of the bandwidth of interest, that is

$$s_{0_1} = j2\pi f_{[(1+f_{num})/2]}, \quad (4.12)$$

and all additional expansion points that are used are chosen at

$$s_{0_v} = j2\pi f_{[w_l l_e + w_r r_e]} \quad (4.13)$$

where l_e and r_e are the numbers associated with the frequencies located at the left and right endpoints of the sub-band of interest, and w_l and w_r are the weights associated with each (see (4.21)).

After s_{0_v} is chosen and the MORE iterative process continues to increase the ROM size, there must be some way to determine when the process has essentially extracted all the worthwhile information from s_{0_v} (this answers the question of how large q_v should be at s_{0_v}); then either the entire process should stop if the solution has converged, or it should find and jump to $s_{0_{v+1}}$ (see section 4.3).

One way to determine the approximation order q_v is to stop the iteration when there is going to be a \mathbf{w}_{n+1} vector generated that is mostly contained in the space $\overline{\mathbf{W}}_n$. This will occur when the iterative process starts to stagnate because no new useful information will be contained in \mathbf{w}_{n+1} . A similar idea [26] is to monitor the projection of

$$\mathbf{y}_{n+1} = \mathbf{b}_n - \sum_{m=1}^{\min(a_1, n)} \mathbf{A}_m \mathbf{w}_{n+1-m}. \quad (4.14)$$

onto the space $\mathbf{A}_0 \overline{\mathbf{W}}_n$. Note that there is no additional computational cost required to form \mathbf{y}_{n+1} because it must be generated anyway for use in (3.21) through the

equation

$$\mathbf{w}_{n+1} = \mathbf{A}_0^{-1}(\mathbf{y}_{n+1}). \quad (4.15)$$

To determine if $\bar{\mathbf{w}}_{n+1}$ should be generated, consider how much of \mathbf{y}_{n+1} is contained in the space $\mathbf{A}_0 \bar{\mathbf{W}}_n$, that is, form

$$\mathbf{y}_{n+1}^{\parallel} = \mathbf{A}_0 \bar{\mathbf{W}}_n \left(\bar{\mathbf{W}}_n^T \mathbf{A}_0 \bar{\mathbf{W}}_n \right)^{-1} \bar{\mathbf{W}}_n^T \mathbf{y}_{n+1} \quad (4.16)$$

(again note that the majority of the computations required in (4.16) must be done anyway for use in (4.9)), define

$$\text{coeff} = \frac{\|\mathbf{y}_{n+1} - \mathbf{y}_{n+1}^{\parallel}\|}{\|\mathbf{y}_{n+1}\|} \quad (4.17)$$

and declare that s_{0_v} is exhausted

$$\text{if } (\text{coeff} \leq \text{tol}_1) \quad (4.18)$$

for some tolerance value tol_1 .

4.3 Using the relative residual to automate MGAWE

After extracting all worthwhile information from s_{0_v} , one must determine if the solution has converged or if $s_{0_{v+1}}$ is necessary (and if so, where it should be located). To resolve these issues, consider the following procedure using section 4.2 and the relative residual

$$rr_n(f) = \frac{\|\mathbf{r}_n(f)\|_{\infty}}{\left\| \sum_{k=0}^{b_1} \sigma^k \mathbf{b}_k \right\|_{\infty}}. \quad (4.19)$$

After s_{0_1} is chosen, generate $\bar{\mathbf{W}}_{q_1}$ and also, for all necessary f_u , generate $\mathbf{g}_{q_1}(f_u)$ and $rr_{q_1}(f_u)$. The f_u corresponding to points of s adjacent to s_{0_1} are marked as

converged for $\mathbf{h}_{q_1}(f_u)$

$$\text{if } (rr_{q_1}(f_u) < tol_2) \quad (4.20)$$

for some tolerance value tol_2 . If some f_u is marked as converged, then f_u corresponding to the next farthest point of s from s_{0_1} is checked. This continues until (4.20) is not satisfied for some f_u . If neither f_{min} nor f_{max} is marked as converged, then there are two unconverged regions, one on each side of s_{0_1} , from which the next expansion point, s_{0_2} , can be chosen. Pick s_{0_2} in the unconverged region that has the widest bandwidth. After s_{0_2} is chosen, generate $\overline{\mathbf{W}}_n$, $\mathbf{g}_n(f_u)$ and $rr_n(f_u)$ as required (where $n = q_1 + q_2$ in this case). Continue to pick expansion points and divide and test the unconverged regions until all values of f_u are marked as converged; then assign the current value of v to num_pts .

Once an unconverged region is selected in which the next s_{0_v} will be located, exactly where in the region should s_{0_v} be placed? That is, what should the values for the endpoint weights w_l and w_r from (4.13) be? Although it may seem that s_{0_v} should be located as close to the middle of the region as possible, this is not the case if one of the region's endpoints is $j2\pi f_{min}$ exclusive or $j2\pi f_{max}$. If the region contains neither of these extrema, then there is an expansion point on each side of the region; therefore $rr_n(f)$ is likely to be smaller. To compensate for a region that contains one of the two extrema, the expansion point should be biased closer to the extremum. Therefore, to locate s_{0_v} for $v \geq 2$, use (4.13) with

$$\text{if } (l_e == 1) \quad w_l = 3/4 \text{ and } w_r = 1/4 \quad (4.21)$$

$$\text{else if } (r_e == f_{num}) \quad w_l = 1/4 \text{ and } w_r = 3/4$$

$$\text{else} \quad w_l = 1/2 \text{ and } w_r = 1/2.$$

4.4 Numerical examples: initial examinations

The automated MGAWÉ process outlined in this chapter is used to simulate all the numerical examples discussed in chapter 2. For each example in this section, $tol_1 = 10^{-6}$ and $tol_2 = 10^{-2}$. In addition, for each example the breakeven point is given. The breakeven point is defined as the number of frequency point solutions which could be carried out using a traditional point-by-point sweep in the amount of time required by the entire MORE process. Of course, this definition of the breakeven point is unfortunately a function of the underlying solver used for the simulation.

Example 1: The horn antenna described in subsection 2.2.1 is simulated. The MGAWÉ process selected the number of expansion points, their locations, and the subspace order generated at each of them as shown in Table 4.1. Figure 4.1 shows the impedance calculated by MGAWÉ along with the exact solution to the matrix equation computed using an LU decomposition. For this example, the MGAWÉ process is seen to be accurate because the responses shown in Figure 4.1 are indistinguishable. In addition, MGAWÉ is computationally efficient; the breakeven point is 13.

	location ($j2\pi$ MHz)	q_v
s_{0_1}	1000	13
s_{0_2}	590	22
s_{0_3}	1440	11

Table 4.1: Reduced order model characteristics selected by the MGAWÉ process for the horn antenna.

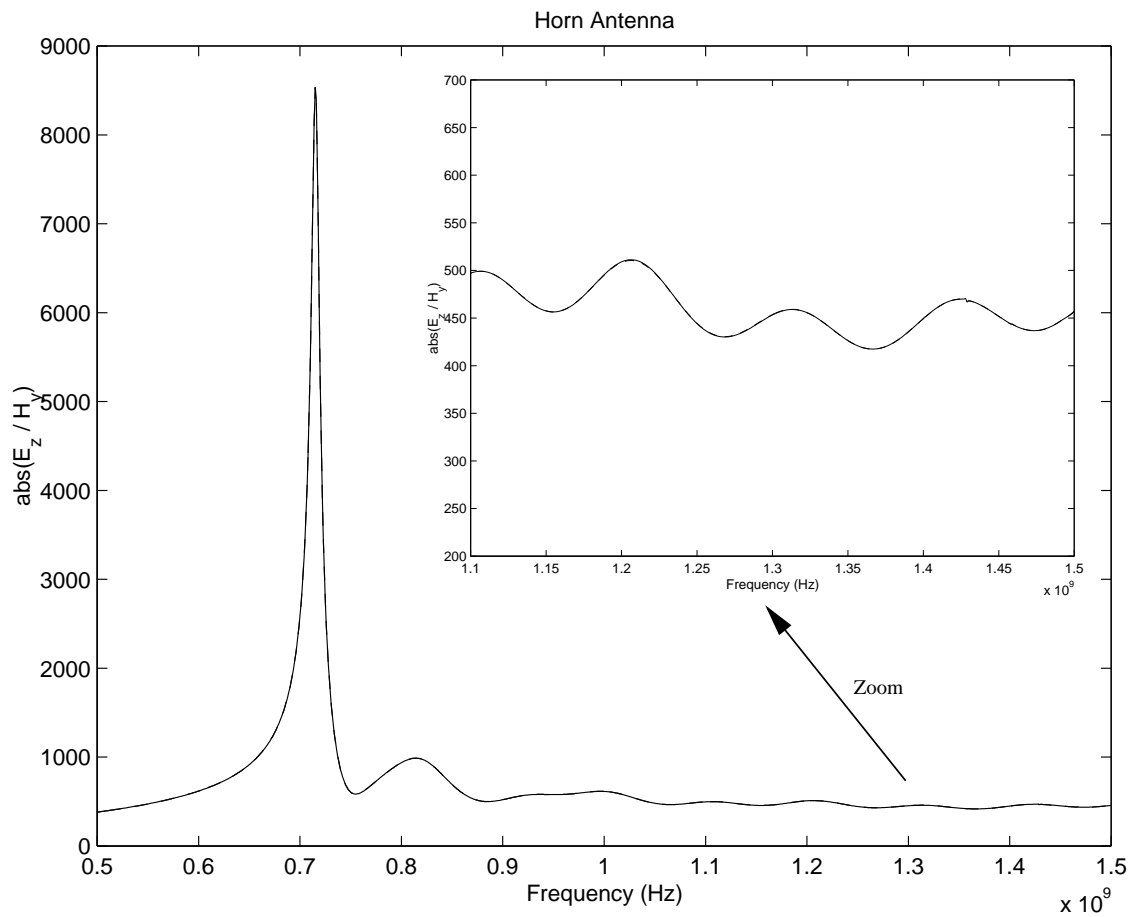


Figure 4.1: Impedance calculated for the horn antenna. Solid \rightarrow LU response, dash-dash \rightarrow MGAWF response (indistinguishable). Compare to Figure 3.1.

Example 2: The TM_z radiation problem terminated with an anisotropic, dispersive PML discussed in subsection 2.2.2 is simulated. Table 4.2 shows the quantities automatically selected by the MGAWE process. In addition, Figure 4.2 shows the relative error (3.36) in the solution vector computed by MGAWE with $\mathbf{L} = \mathbf{I}$. Compare Figure 4.2 to Figure 3.2, and notice the difference in the scale of the dependent variable. For this example, the time required for the MGAWE simulation results in a breakeven point of 7.

	location ($j2\pi$ MHz)	q_v
s_{0_1}	252.5	16
s_{0_2}	40	19
s_{0_3}	462.5	10
s_{0_4}	2.5	11
s_{0_5}	380	13

Table 4.2: Reduced order model characteristics selected by the MGAWE process for the example in subsection 2.2.2.

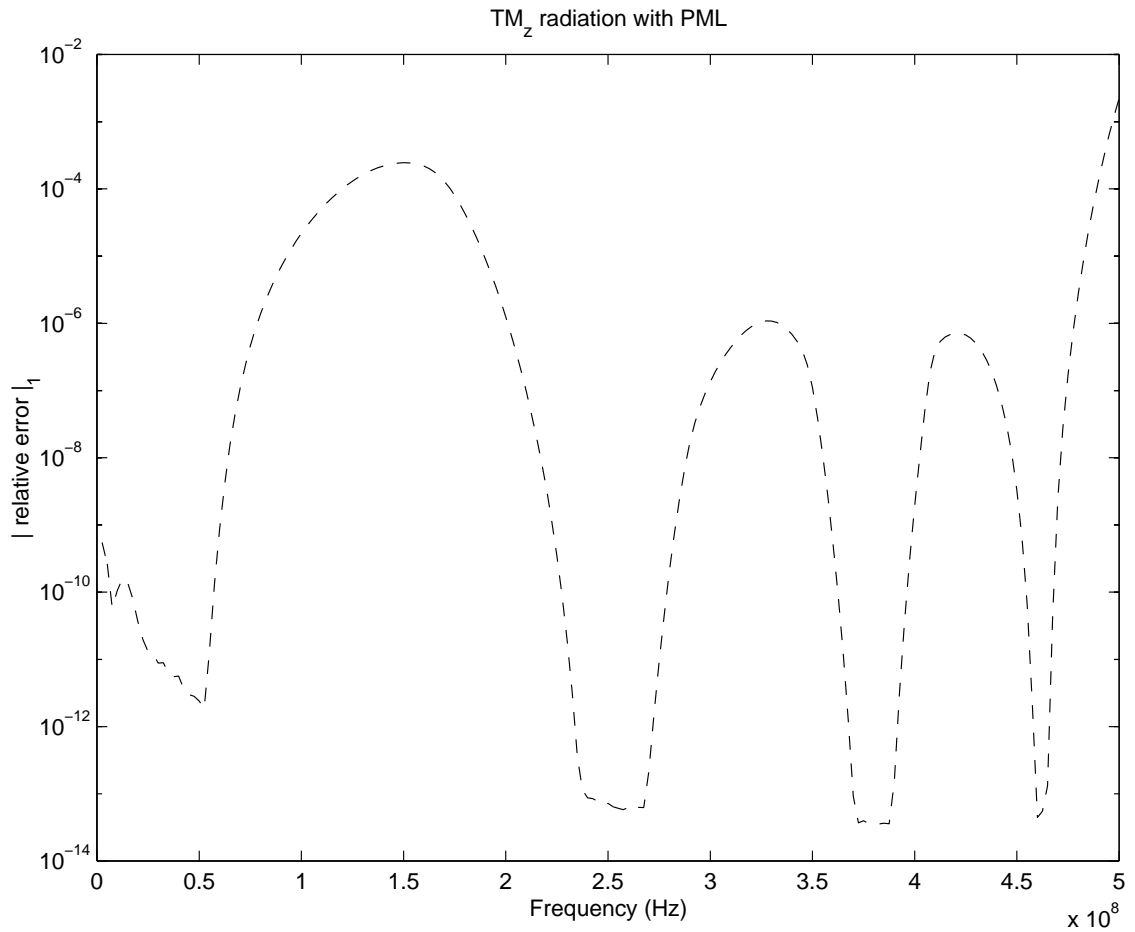


Figure 4.2: Relative error in the solution vector for the example in subsection 2.2.2. Dash-dash \rightarrow MGAW method (compare the scale to that in Figure 3.2).

Example 3: The TE_z scattering problem from subsection 2.2.3 is solved. Table 4.3 shows the quantities automatically chosen by the MGAWE process. Furthermore, Figure 4.3 (which is analogous to Figure 3.3) shows the relative error (3.36) in the solution vector computed by MGAWE with $\mathbf{L} = \mathbf{I}$. The breakeven point for MGAWE in this example is 8.

	location ($j2\pi$ MHz)	q_v
s_{0_1}	252.5	13
s_{0_2}	35	13
s_{0_3}	472.5	11

Table 4.3: Reduced order model characteristics selected by the MGAWE process for the example in subsection 2.2.3.

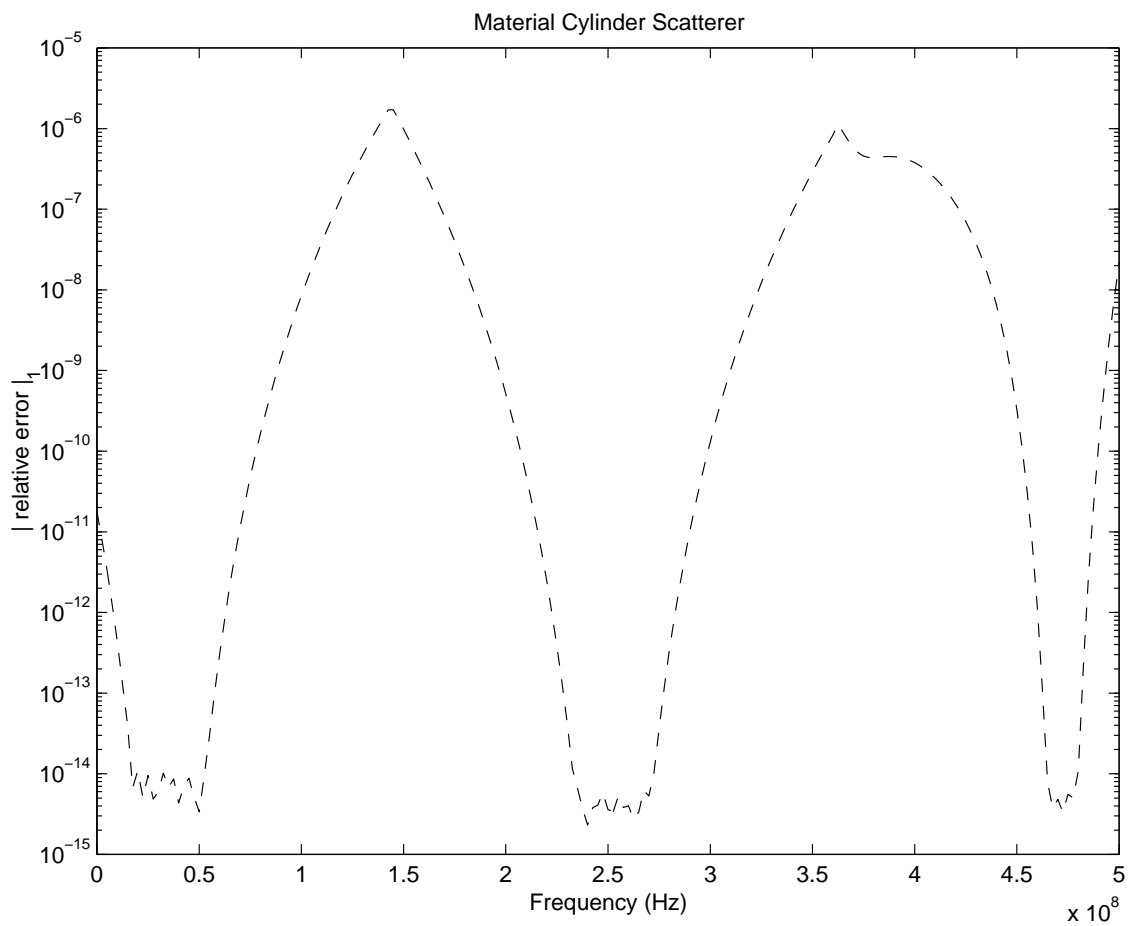


Figure 4.3: Relative error in the solution vector for the example in subsection 2.2.3. Dash-dash \rightarrow MGAW method (compare the scale to that in Figure 3.3).

Example 4: The low pass filter described in subsection 2.2.4 is simulated. Table 4.4 shows the quantities chosen by MGAWE for this example. Figures 4.4 and 4.5 show the magnitude of S_{11} and S_{12} calculated using MGAWE along with the numerically exact solution to the quadratic equation (2.33). As can be seen from these figures, MGAWE essentially adds no additional error once the approximate system (2.33) is obtained. The breakeven point for MGAWE in this example is 77. The major reason that this breakeven point is so much larger than those observed in examples 1-3 is that for this example an iterative solver was used instead of a direct solver. Furthermore, preconditioning the matrix for this example is not much more computationally expensive than using the iterative solver to find a solution to a linear system of equations. This is because the number of unknowns for this example is relatively small compared to the other three dimensional examples to follow (recall the data given in subsection 2.2.4) in which the breakeven point will be much smaller. An additional (minor) reason that the breakeven point is so large for this example is that the iterative solver used for MGAWE produces vectors with a residual of $tol_1/10 = 10^{-7}$, but when used to solve (2.33) the tolerance is only set to 10^{-4} . This example will be investigated further in section 4.5.

	location ($j2\pi$ GHz)	q_v
s_{0_1}	11	13
s_{0_2}	18.74	23
s_{0_3}	3.44	27

Table 4.4: Reduced order model characteristics selected by the MGAWE process for the low pass filter.

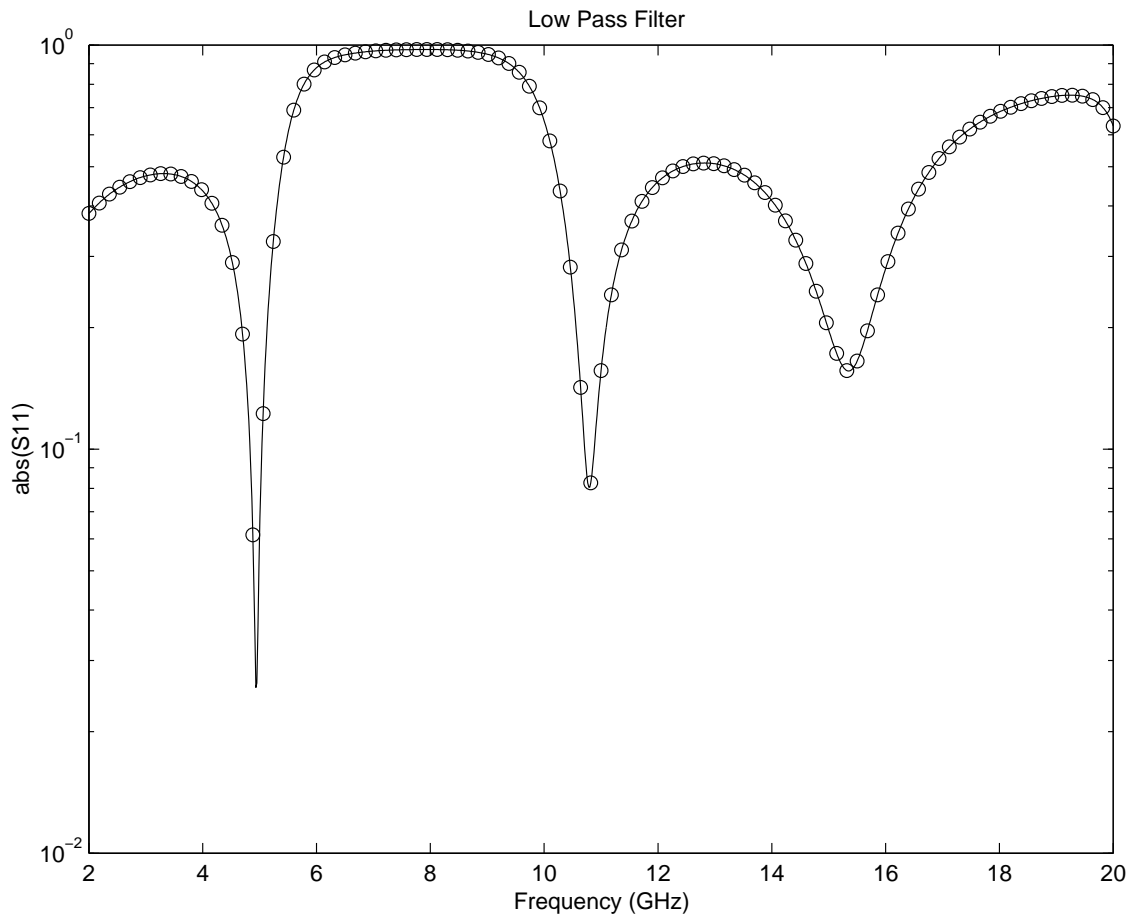


Figure 4.4: S_{11} for the low pass filter. Circles \rightarrow solution to (2.33), solid \rightarrow MGAWF solution.

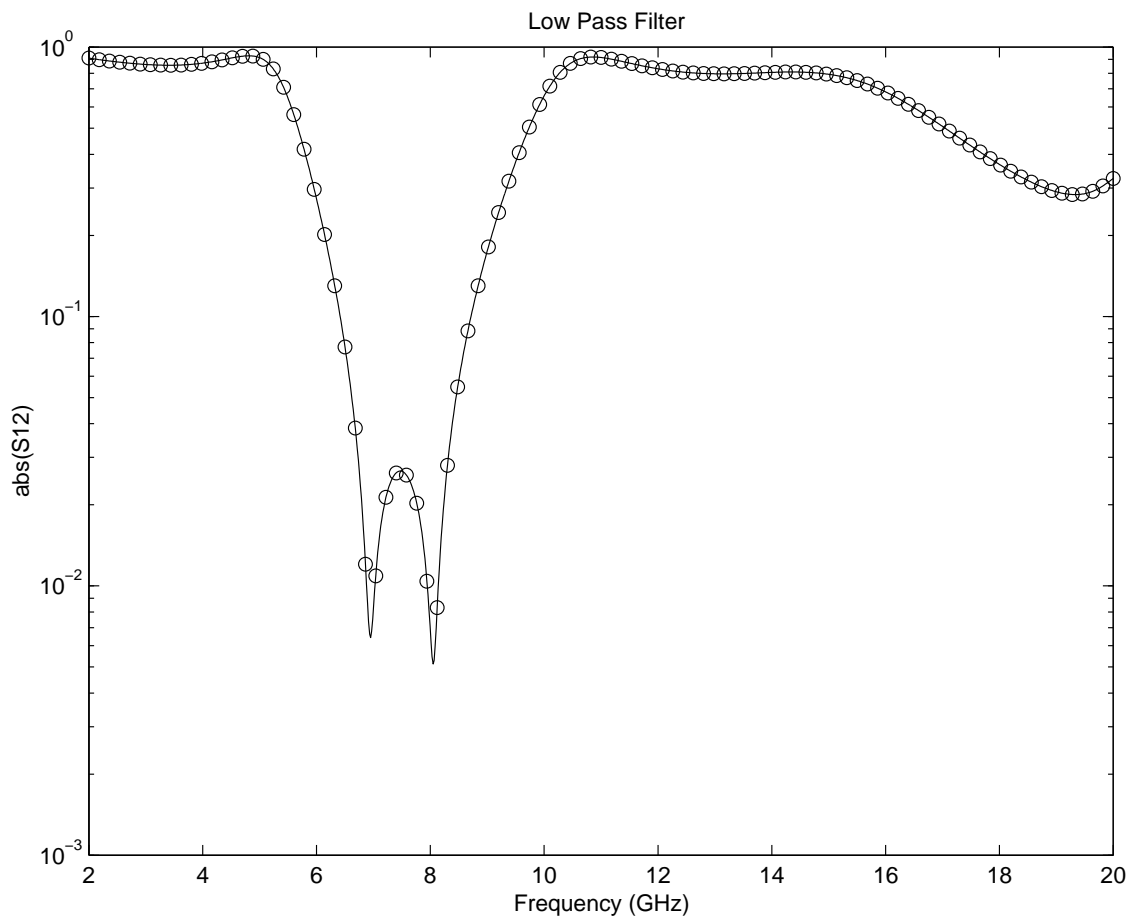


Figure 4.5: S_{12} for the low pass filter. Circles \rightarrow solution to (2.33), solid \rightarrow MGAWE solution.

Example 5: The broadband bowtie antenna described in subsection 2.2.4 is solved using MGAWE. The quantities chosen by MGAWE for this example are shown in Table 4.5. Figure 4.6 shows the input S parameter to the antenna computed using MGAWE and the exact solution to (2.33). The breakeven point for this example is 13.

	location ($j2\pi$ GHz)	q_v
s_{0_1}	2.75	15
s_{0_2}	4.64	20
s_{0_3}	0.86	17
s_{0_4}	3.74	13

Table 4.5: Reduced order model characteristics selected by the MGAWE process for the bowtie antenna.

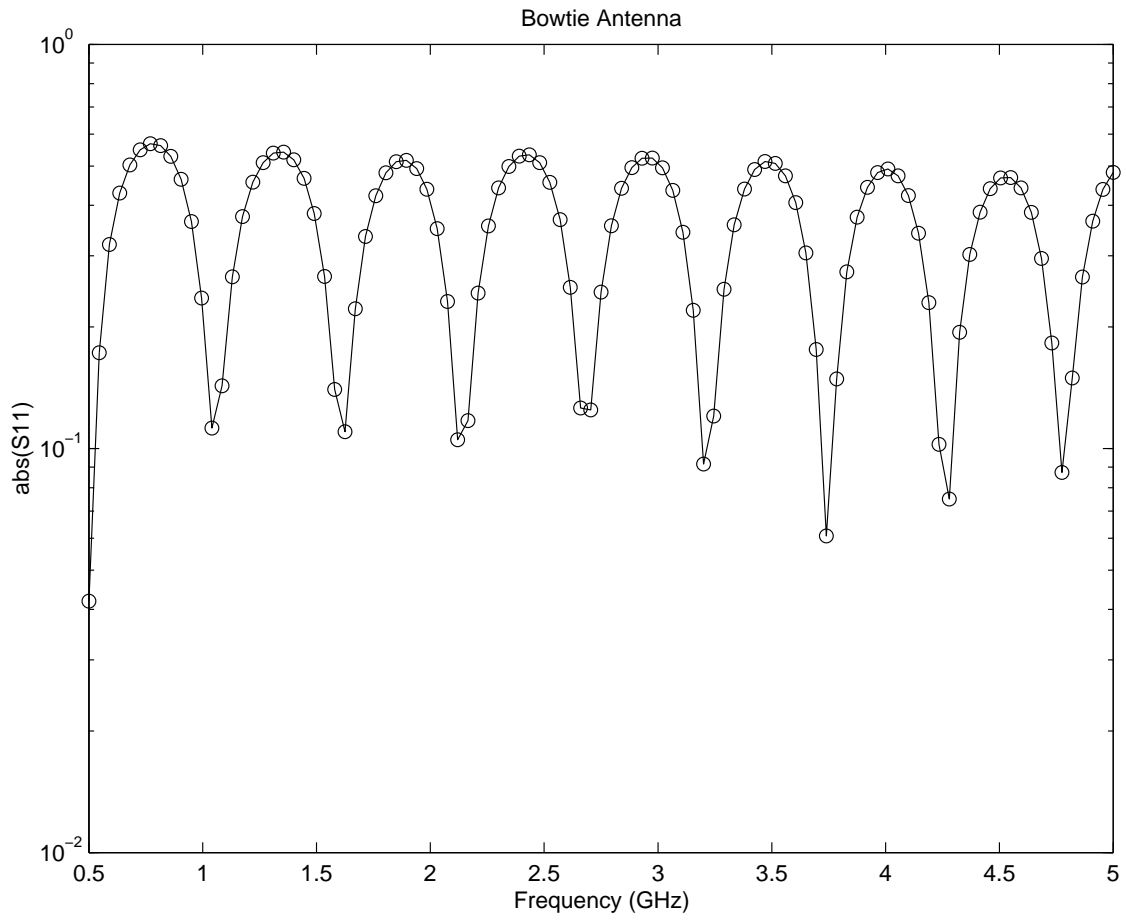


Figure 4.6: S_{11} for the bowtie antenna. Circles \rightarrow solution to (2.33), solid \rightarrow MGAWF solution.

Example 6: The band pass filter from subsection 2.2.4 is simulated using MGAWE and compared to the exact solution of (2.33). Table 4.6 shows the MGAWE quantities chosen. Figure 4.7 shows S_{12} for this filter. The breakeven point for this filter is 8.

	location ($j2\pi$ GHz)	q_v
s_{0_1}	4	8

Table 4.6: Reduced order model characteristics selected by the MGAWE process for the band pass filter.

4.5 Numerical examples: further investigations

Section 4.4 covered many issues involved with a MGAWE MORE solution procedure such as choosing the number of expansion points, as well as their locations and their subspace orders. In addition, the resulting breakeven point for the computation time of the entire MGAWE process was observed for each example. Nevertheless, several other issues still need to be addressed. These issues include: reducing the breakeven point (in particular for the low pass filter example), comparing the efficiency and robustness of MGAWE to a rational polynomial interpolation scheme, and using the reduced order model to adaptively choose the frequency points at which the ROM should be evaluated and plotted.

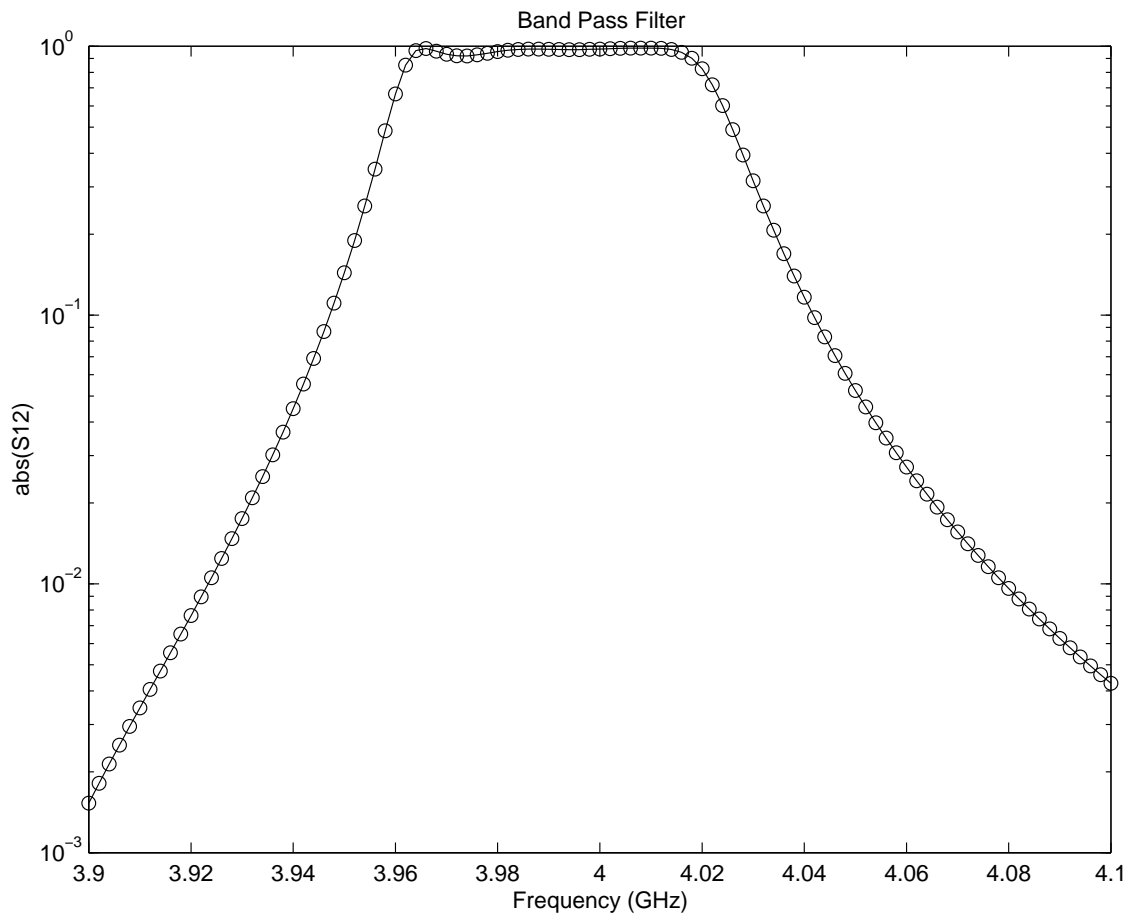


Figure 4.7: S_{12} for the band pass filter. Circles \rightarrow solution to (2.33), solid \rightarrow MGAWE solution.

4.5.1 The breakeven point as a function of tol_1

The breakeven points for all the examples in section 4.4 were acceptably small except for the low pass filter example. As discussed in that example, the major reason is that the number of unknowns for that example is so small that computing the matrix preconditioner is not much more computationally expensive than using the iterative solver to find a solution to a linear system of equations. Since the number of unknowns is fixed, the minor reason given in the low pass filter example will be investigated. Recall that the iterative solver used for MGAWWE produces vectors with a residual of $tol_1/10$, while the tolerance used to solve (2.33) is set to 10^{-4} . In Table 4.7 the effect on the breakeven point of relaxing tol_1 is shown. In Figures 4.8 and 4.9 the S_{11} and S_{12} values are shown for the low pass filter for all values of tol_1 shown in Table 4.7. As can be seen from the figures, no drop in accuracy is observed since the MGAWWE process automatically compensates for changes in tol_1 by using more expansion points if necessary (see Table 4.7, column 2). This numerical example also illustrates the robustness of the MGAWWE algorithm.

tol_1	num_pts	$\sum_{v=1}^{num_pts} q_v$	breakeven point
10^{-6}	3	63	77
10^{-5}	3	55	64
10^{-4}	3	46	50
10^{-3}	4	48	44
10^{-2}	4	35	38
10^{-1}	5	27	31

Table 4.7: The breakeven point as a function of tol_1 for the low pass filter.

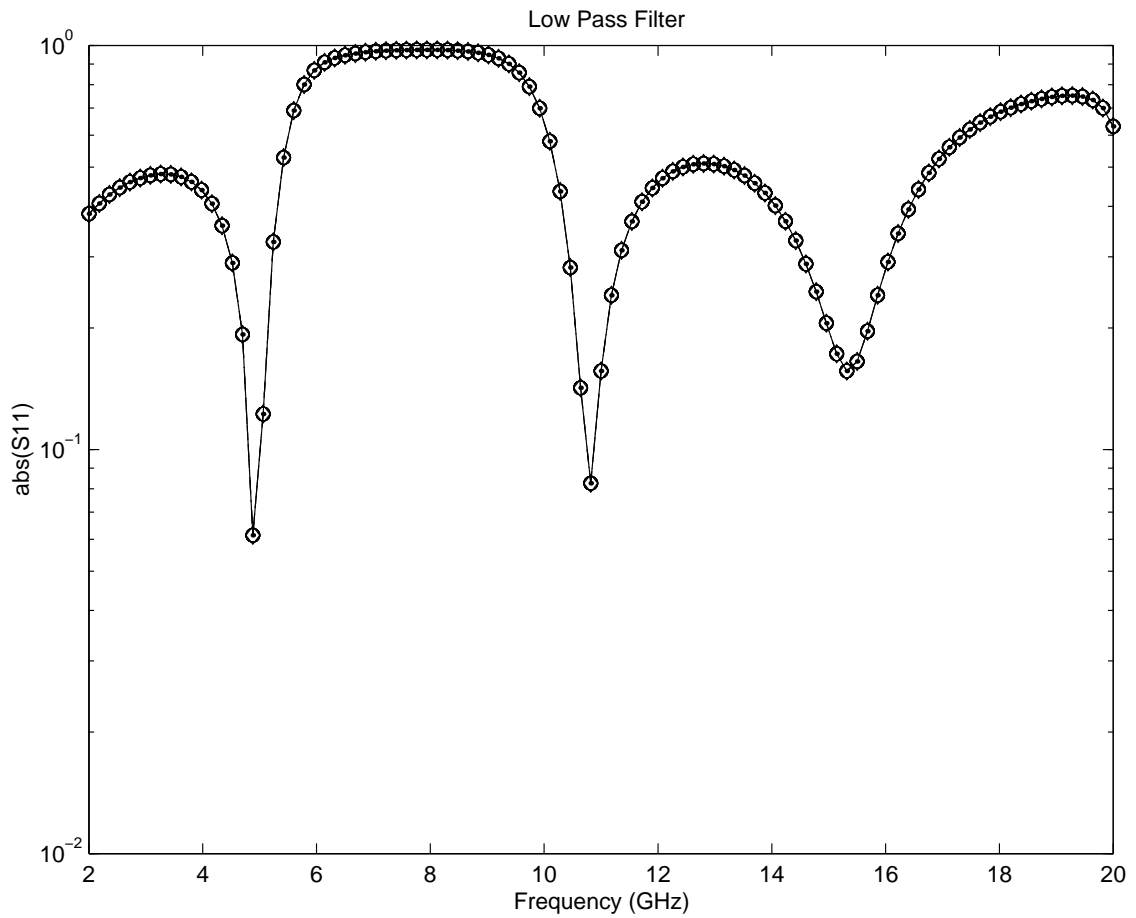


Figure 4.8: S_{11} for the low pass filter. Circles \rightarrow solution to (2.33), diamonds \rightarrow MGAWF with $tol_1 = 10^{-6}$, points \rightarrow MGAWF with $tol_1 = 10^{-5}$, solid \rightarrow MGAWF with $tol_1 = 10^{-4}$, dash-dash \rightarrow MGAWF with $tol_1 = 10^{-3}$, dash-dot \rightarrow MGAWF with $tol_1 = 10^{-2}$, dot-dot \rightarrow MGAWF with $tol_1 = 10^{-1}$.

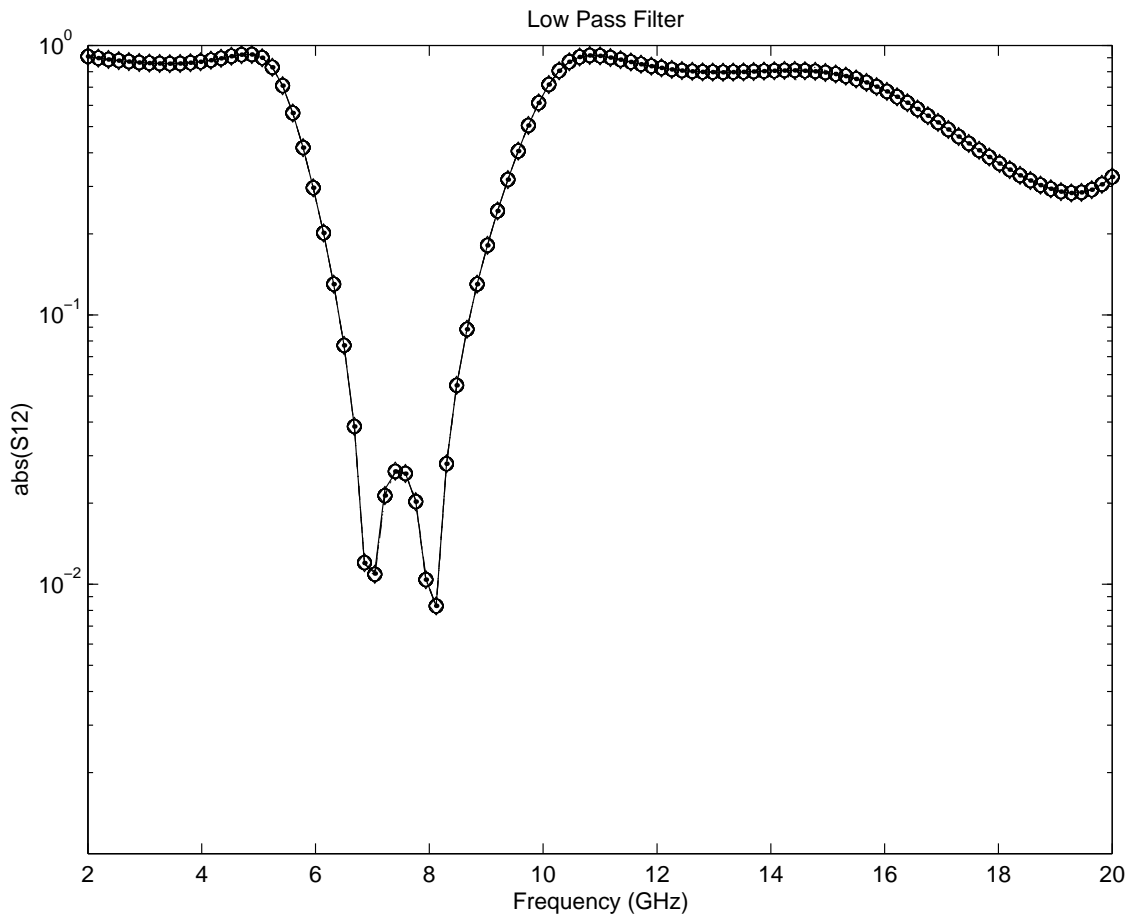


Figure 4.9: S_{12} for the low pass filter. Circles \rightarrow solution to (2.33), diamonds \rightarrow MGAWF with $tol_1 = 10^{-6}$, points \rightarrow MGAWF with $tol_1 = 10^{-5}$, solid \rightarrow MGAWF with $tol_1 = 10^{-4}$, dash-dash \rightarrow MGAWF with $tol_1 = 10^{-3}$, dash-dot \rightarrow MGAWF with $tol_1 = 10^{-2}$, dot-dot \rightarrow MGAWF with $tol_1 = 10^{-1}$.

4.5.2 MGAWE versus rational polynomial interpolation

The accuracy, efficiency, and robustness of the MGAWE process has been observed in the previous numerical examples. However, given the breakeven points realized for those examples, it is natural to ask how MGAWE compares to a simple fitting scheme such as a rational polynomial interpolation. To this end, the function **ratint** found in section 3.2 of reference [40] is used for the low pass filter and bowtie antenna examples. The driver routine used to call **ratint** picks points at which to evaluate the solution for inclusion in the approximation. These points are picked at the maximum of the error estimate returned by **ratint**; the driver continues to pick additional points until the error estimate is less than $tol_2 = 10^{-2}$ throughout the specified bandwidth.

In Figure 4.10 and 4.11 the rational polynomial interpolation scheme is shown to be very accurate for the low pass filter example. In addition, the rational polynomial interpolation scheme is also more efficient than MGAWE because the number of evaluations required (which is comparable to the breakeven point for MGAWE) is 16 for S_{11} of the low pass filter and 14 for S_{12} of the same filter. However, the question of the robustness of the rational polynomial interpolation still needs to be addressed. To this end, consider Figure 4.12 where the rational polynomial interpolation scheme is used to compute the approximation to S_{11} for the bowtie antenna. In this case note that rational polynomial interpolation prematurely terminates; therefore it is not very robust. Furthermore, the number of evaluations performed by rational polynomial interpolation for the bowtie antenna example is 15, which is greater than the breakeven point of 13 required by MGAWE to compute the approximation shown in Figure 4.6. Therefore, MGAWE seems to be a more desirable MORE method than rational polynomial interpolation for two reasons. The first (and most important)

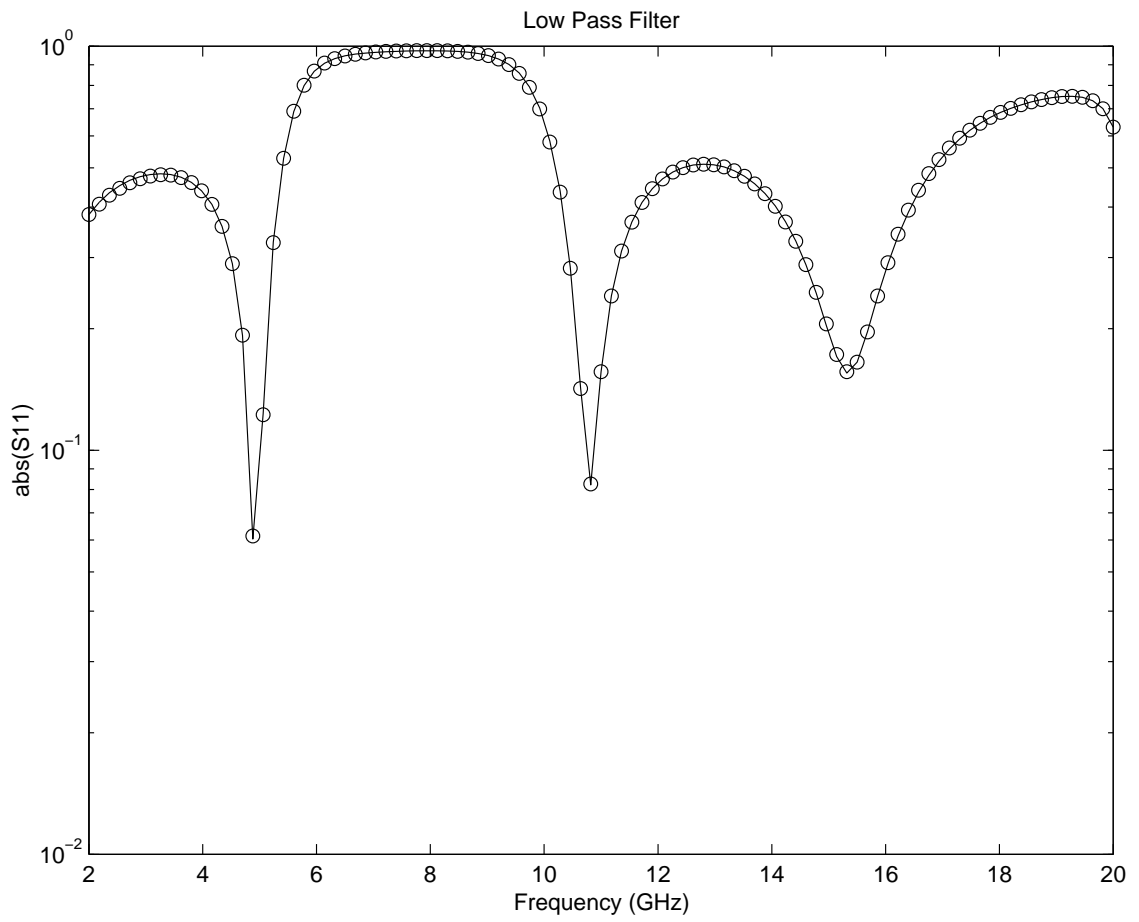


Figure 4.10: S_{11} for the low pass filter. Circles \rightarrow solution to (2.33), solid \rightarrow rational polynomial interpolation.

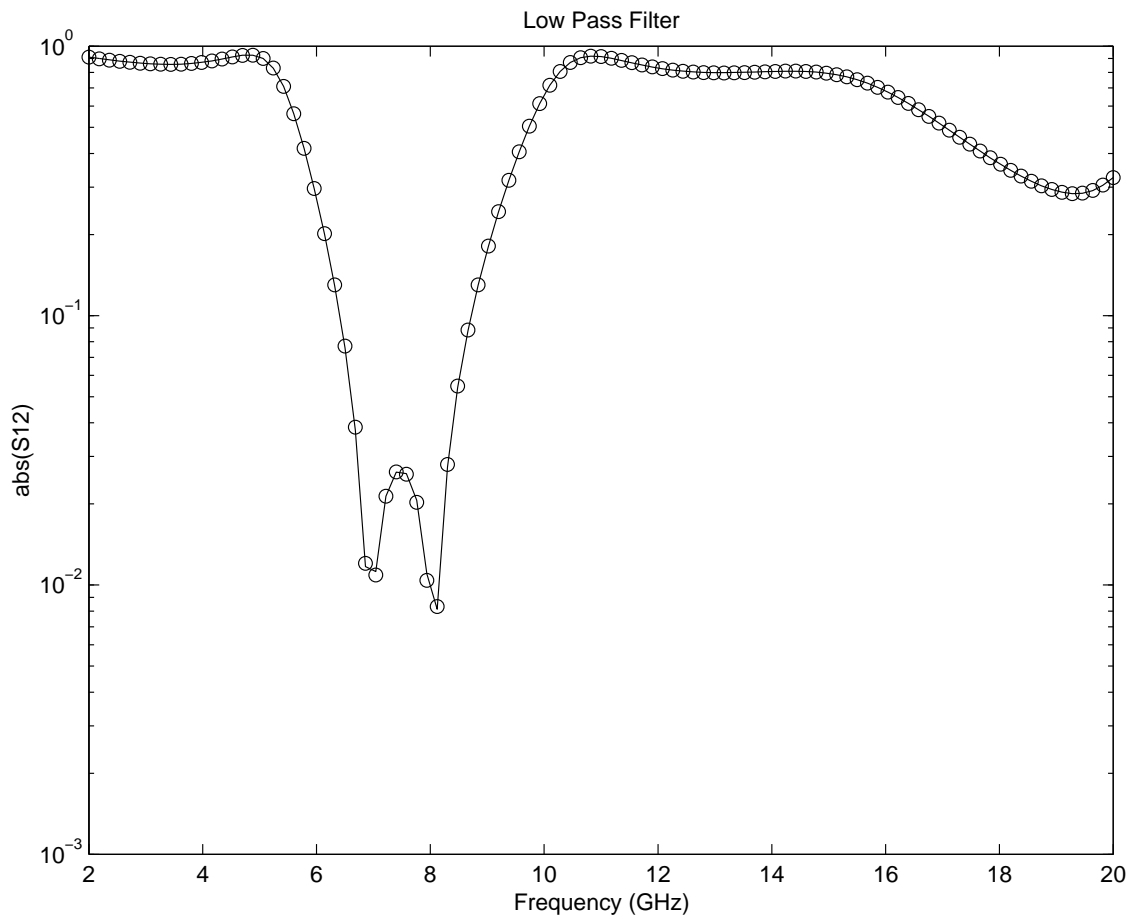


Figure 4.11: S_{12} for the low pass filter. Circles \rightarrow solution to (2.33), solid \rightarrow rational polynomial interpolation.

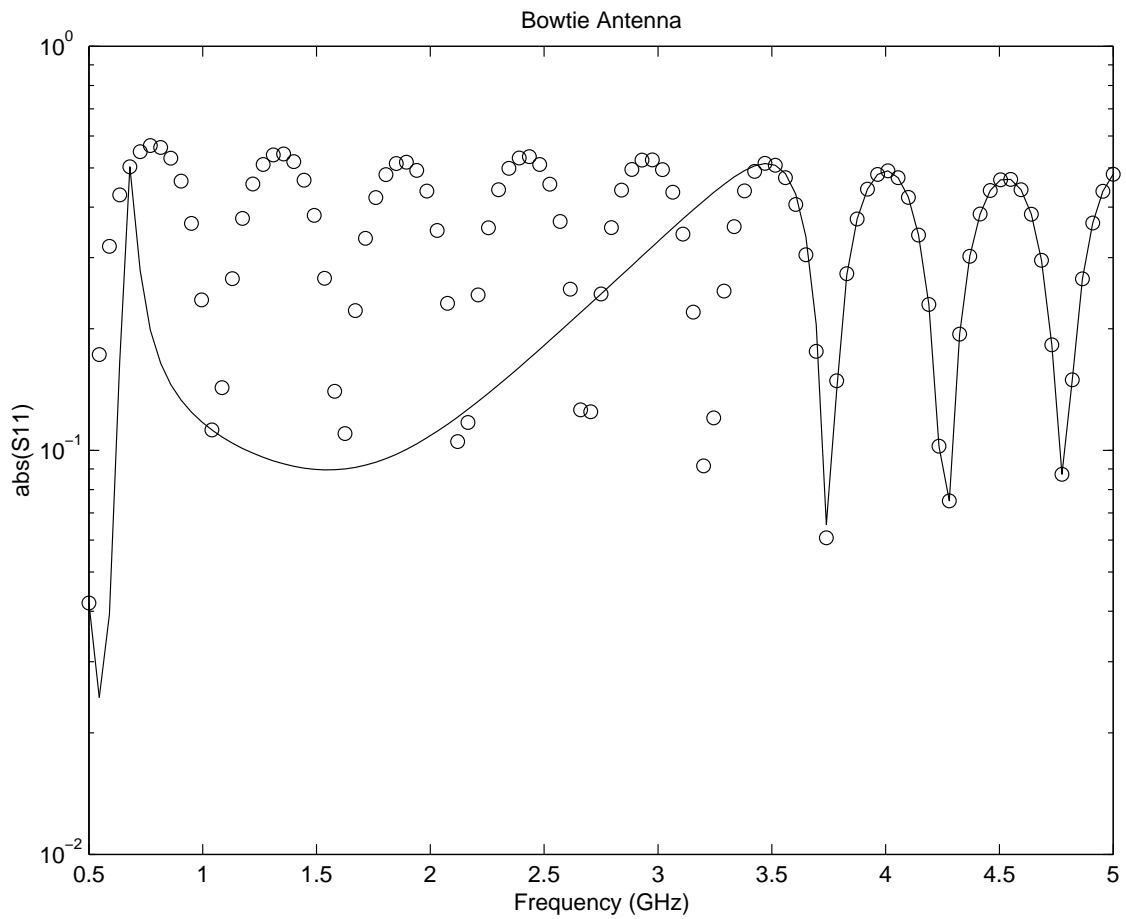


Figure 4.12: S_{11} for the bowtie antenna. Circles \rightarrow solution to (2.33), solid \rightarrow rational polynomial interpolation (which fails in this case).

reason is that rational polynomial interpolation is not necessarily robust. The second reason is that in some cases MGAWF may be somewhat more efficient than rational polynomial interpolation.

4.5.3 Adaptively choosing evaluation frequencies

Once a reduced order model is created, it is very inexpensive to evaluate the solution at any given f . However, regardless of how small the discretization is between f_u and f_{u+1} from (2.2), it is possible that a resonance exists between these points of evaluation. On the other hand, once the iterative MORE process terminates, the ROM is an accurate representation of the original system; the pole distribution of the ROM should approximate the pole distribution of the original system. Therefore, the points at which the ROM should be evaluated and plotted can be made a function of the pole distribution of the ROM. This is shown for the horn antenna described in subsection 2.2.1. In Figure 4.13 the pole distribution in the complex s plane is given for $e^{j\omega t}$ time convention. These poles are then used to adaptively choose the points at which the ROM should be evaluated for plotting purposes. Although overall the resulting plot (not shown) would look like the plot shown in Figure 4.1, some details can be lost unless the adaptive plotting scheme is implemented. This is illustrated in Figure 4.14 where it is obvious that choosing equally spaced ($f_{u+1} - f_u = 1\text{MHz}$) evaluation points can (and in this case does) miss information in a rapidly varying frequency domain plot. This is just another advantage of a MORE process; the readily available information contained in the ROM pole distribution allows for the creation of an adaptive evaluation/plotting scheme.

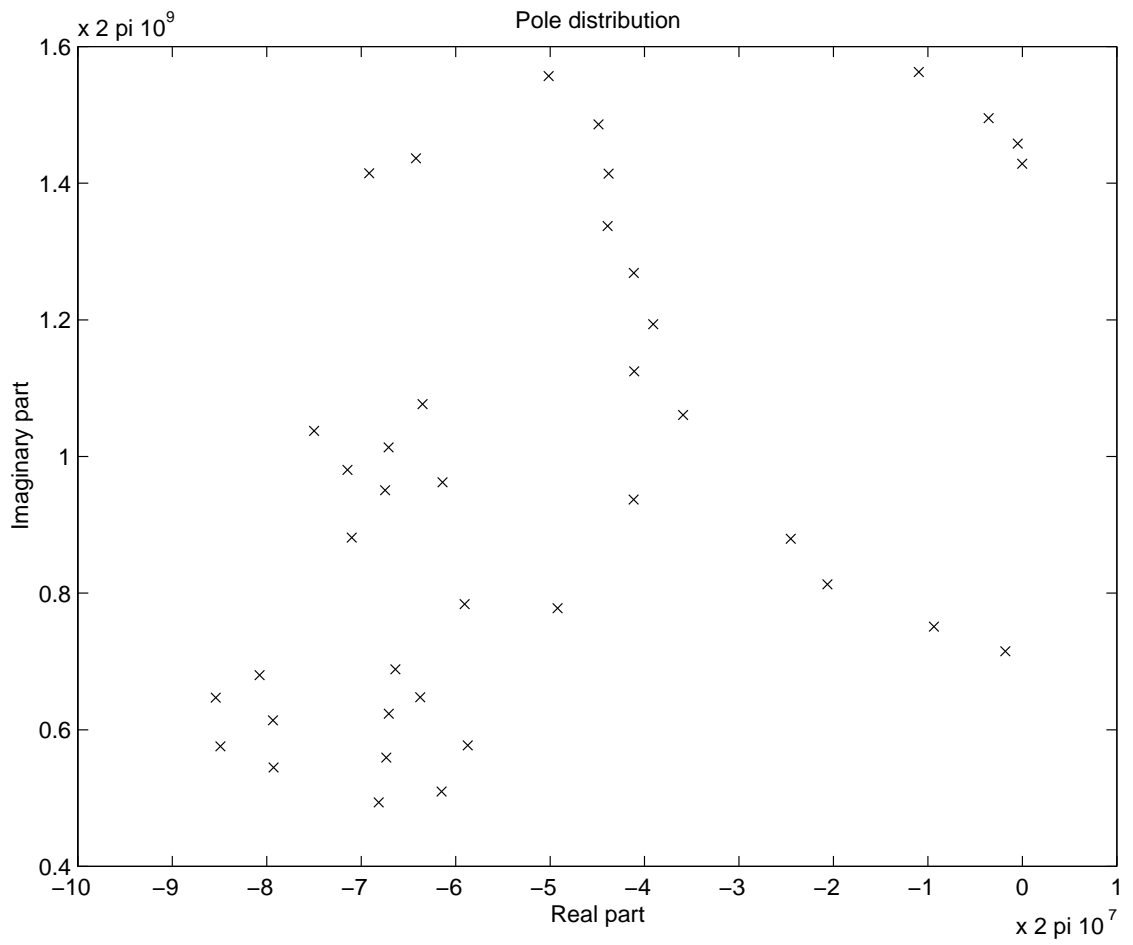


Figure 4.13: Pole distribution in the s plane for the horn antenna. $\times \rightarrow$ poles.

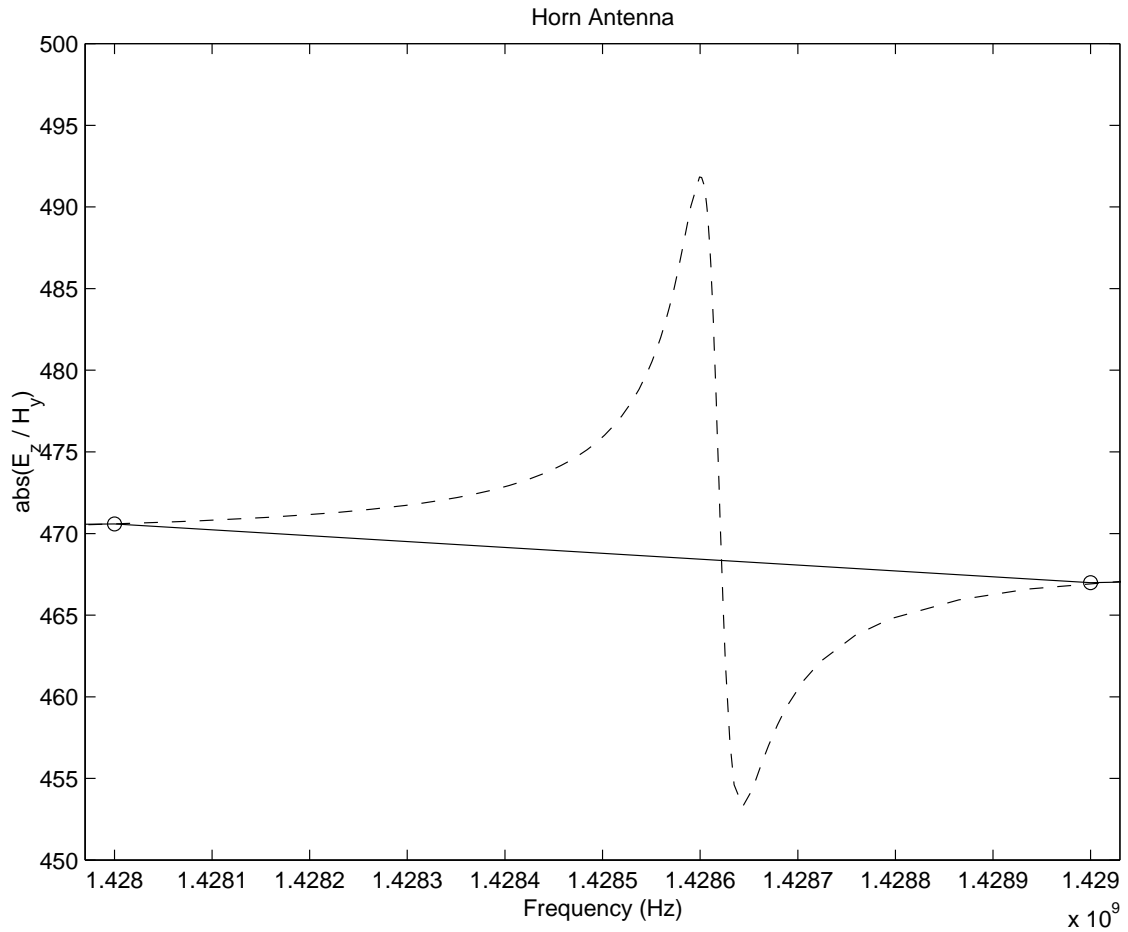


Figure 4.14: Zoom plot for the impedance of the horn antenna using equally spaced points versus adaptively choosing points from the ROM pole distribution. Solid with circles \rightarrow 2 of the 1001 equally spaced evaluation points from Figure 4.1, dash-dash \rightarrow evaluation points adaptively chosen from ROM pole distribution from Figure 4.13.

CHAPTER 5

WELL-CONDITIONED ASYMPTOTIC WAVEFORM EVALUATION (WCAWE)

5.1 Motivation

In chapter 4 the wide-band MGAWE technique was presented. MGAWE obtains its wide-band convergence by utilizing multiple GAWE approximations. However, if any of the GAWE approximations were considered individually, it would only be slightly more wide-band than an AWE approximation of the same order at that expansion point. This is because both AWE and GAWE build their moment-matching subspace \mathbf{W}_q through the ill-conditioned vector-generating process found in (3.21). More specifically, AWE uses \mathbf{W}_q directly to obtain the moments given in (3.23), and GAWE orthonormalizes \mathbf{W}_q into the basis $\overline{\mathbf{W}}_q$ which is then used in (3.29) and (3.34).

In an attempt to increase the bandwidth of accuracy for GAWE, a natural question to ask is “Why wait until \mathbf{W}_q is generated before orthonormalizing the vectors into

$\overline{\mathbf{W}}_q$?" In other words, modify the process (3.21) to define the new process

$$\begin{aligned}
\widehat{\mathbf{w}}_1 &= \mathbf{A}_0^{-1} \mathbf{b}_0 \\
\widehat{\mathbf{w}}_2 &= \mathbf{A}_0^{-1} (\mathbf{b}_1 - \mathbf{A}_1 \widehat{\mathbf{w}}_1) \\
&\vdots \\
\widehat{\mathbf{w}}_q &= \mathbf{A}_0^{-1} \left(\mathbf{b}_{q-1} - \sum_{m=1}^{\min(a_1, q-1)} \mathbf{A}_m \widehat{\mathbf{w}}_{q-m} \right)
\end{aligned} \tag{5.1}$$

where, at step n , $\widehat{\mathbf{w}}_n$ is immediately orthonormalized against $\overline{\mathbf{W}}_{n-1}$ to form $\widehat{\mathbf{w}}_n$ before $\widehat{\mathbf{w}}_{n+1}$ is generated. Then replace the subspace $\overline{\mathbf{W}}_q$ in GAWE with this new subspace $\overline{\widehat{\mathbf{W}}}_q$.

The author has observed that using the vectors from (5.1) in a GAWE process can, in certain situations, exhibit a much wider bandwidth of convergence than using the vectors from (3.21). **However, the use of (5.1) is strongly discouraged for two reasons.**³ The first reason is it has been found that using (5.1) can fail for certain problems. The second reason is that, in general, vectors from (5.1) do not match moments. This will be shown in appendix C for two cases: the first case is when the right hand side of (3.1) is of linear or higher order in σ (that is, $b_1 \geq 1$), and the other case is when $b_1 = 0$ and $a_1 \geq 2$. Of course, for the case $b_1 = 0$ and $a_1 = 1$, (5.1) is valid; it is just the Arnoldi process [15].

In summary, although orthonormalizing the vectors \mathbf{w}_n from (3.21) against \mathbf{W}_{n-1} before \mathbf{w}_{n+1} is generated is a wonderful idea, there is a problem in ensuring that the resulting subspace matches moments. The correct process will be shown to be the

³The author believes these reasons are actually related. He believes that a method based on (5.1) will fail for any problem in which there is a “significant” portion of the solution vector that is contained in the moment-matching space that (5.1) fails to capture. A rigorous proof of this conjecture seems unessential since the alternative, correct WCAWE is available and therefore the existence of (5.1) becomes a moot point.

moment-matching WCAWE process in section 5.3. However, to understand WCAWE it is first necessary to understand the link between the AWE vectors in (3.21) and the power method applied to (3.9). This link will be shown in section 5.2.

5.2 The connection between classical MORE techniques

Although the link between AWE and PVL for linear matrix equations in infinite precision arithmetic has been known for about a decade, the link between AWE and Krylov subspace techniques for matrix equations with a polynomial dependence on the MORE parameter was elusive. Furthermore, this link for polynomial systems (which will be shown in theorem 5.1) proves to be essential in understanding the WCAWE process that follows in section 5.3. However, before stating theorem 5.1 it is necessary to give the following definition.

Definition 5.1 (Power method applied to (3.9)) Let $\check{\check{\mathbf{w}}}_1 = \mathbf{C}^{-1}\mathbf{y}$ where $\mathbf{C}^{-1}\mathbf{y}$ is given in (3.10). Then for all $n \geq 1$ let $\check{\check{\mathbf{w}}}_{n+1} = \mathbf{C}^{-1}\mathbf{D}\check{\check{\mathbf{w}}}_n$. Furthermore, let $\check{\check{\mathbf{w}}}_n$ be the first $N + 1$ entries of the vector $\check{\check{\mathbf{w}}}_n$, and let \mathbf{w}_n be the first N entries of $\check{\check{\mathbf{w}}}_n$. Finally, let $\check{\check{\mathbf{w}}}_{n(j)}$ be the $1 + (j - 1)(N + 1)$ to the $j(N + 1)$ entries of $\check{\check{\mathbf{w}}}_n$. These quantities are shown in (5.2).

$$\check{\check{\mathbf{w}}}_n = \left[\begin{array}{c} \check{\check{\mathbf{w}}}_{n(1)} \\ \check{\check{\mathbf{w}}}_{n(2)} \\ \vdots \\ \check{\check{\mathbf{w}}}_{n(c_1)} \end{array} \right] \left\{ \begin{array}{c} \uparrow \\ N + 1 \\ \downarrow \end{array} \right\} = \check{\check{\mathbf{w}}}_n = \left[\begin{array}{c} \vdots \\ \vdots \end{array} \right] \left\{ \begin{array}{c} \uparrow \\ N \\ \downarrow \end{array} \right\} = \mathbf{w}_n. \quad (5.2)$$

□

Now the crucial theorem 5.1 can be given. It provides the link between AWE and the Krylov subspace generated by the power method when applied to (3.9) with the matrix $\mathbf{C}^{-1}\mathbf{D}$ given as shown in (3.10).

Theorem 5.1 (Link between AWE and the power method on (3.9)) For any integer $1 \leq n \leq q$ the vector \mathbf{w}_n from definition 5.1 is exactly the same as the vector $\check{\mathbf{w}}_n$ from (3.21), even in finite precision arithmetic.

Proof: Start with definition 5.1 and show that (3.21) results.

First notice from definition 5.1 and the structure of $\mathbf{C}^{-1}\mathbf{D}$ given in (3.10) that $\check{\check{\mathbf{w}}}_{n+1(j+1)} = \check{\check{\mathbf{w}}}_{n(j)}$ for $n \geq 1$ and $1 \leq j \leq c_1 - 1$. In addition, from (3.3) note that the $(N+1)$ th entry of \mathbf{M}_i times any vector of length $N+1$ is always zero for $i \geq 1$. This fact coupled with the structure of the last row of \mathbf{M}_0 ensures that the $(N+1)$ th entry of $\check{\check{\mathbf{w}}}_n$ is zero for all $n \geq 2$. This, together with the fact $\check{\check{\mathbf{w}}}_{n+1(j+1)} = \check{\check{\mathbf{w}}}_{n(j)}$, means the last entry of $\check{\check{\mathbf{w}}}_{n(j)} = 0$ for all $n \neq j$.

After establishing the above facts consider $\check{\check{\mathbf{w}}}_1 = \mathbf{C}^{-1}\mathbf{y}$, so $\check{\mathbf{w}}_1 = \mathbf{M}_0^{-1}\mathbf{e}_{N+1}$ and all other entries of $\check{\check{\mathbf{w}}}_1$ are zero. From the structure of the last row of \mathbf{M}_0 , the $(N+1)$ th entry of $\check{\mathbf{w}}_1$ is -1 . Therefore, $\mathbf{A}_0\mathbf{w}_1 - \mathbf{b}_0 = 0$, that is, $\mathbf{w}_1 = \mathbf{A}_0^{-1}\mathbf{b}_0$, which is the first equation in (3.21).

For $\check{\check{\mathbf{w}}}_2$, note that $\check{\check{\mathbf{w}}}_{2(2)} = \check{\mathbf{w}}_1$ and $\check{\mathbf{w}}_2 = -\mathbf{M}_0^{-1}\mathbf{M}_1\check{\mathbf{w}}_1$. Keeping in mind that the $(N+1)$ th entry of $\mathbf{M}_1\check{\mathbf{w}}_1 = 0$, it is then clear that $\mathbf{w}_2 = -\mathbf{A}_0^{-1}(\mathbf{A}_1\mathbf{w}_1 - \mathbf{b}_1) = \mathbf{A}_0^{-1}(\mathbf{b}_1 - \mathbf{A}_1\mathbf{w}_1)$ which is the second equation in (3.21).

For $\check{\check{\mathbf{w}}}_3$, note that it is the case that $\check{\check{\mathbf{w}}}_{3(3)} = \check{\check{\mathbf{w}}}_{2(2)} = \check{\mathbf{w}}_1$ as well as $\check{\check{\mathbf{w}}}_{3(2)} = \check{\mathbf{w}}_2$. In addition, $\check{\mathbf{w}}_3 = -\mathbf{M}_0^{-1}(\mathbf{M}_1\check{\mathbf{w}}_2 + \mathbf{M}_2\check{\mathbf{w}}_1)$. Therefore, $\mathbf{w}_3 = -\mathbf{A}_0^{-1}(\mathbf{A}_1\mathbf{w}_2 + \mathbf{A}_2\mathbf{w}_1 - \mathbf{b}_2) = \mathbf{A}_0^{-1}(\mathbf{b}_2 - \mathbf{A}_1\mathbf{w}_2 - \mathbf{A}_2\mathbf{w}_1)$ which is the third equation in (3.21).

In general, for $\check{\check{\mathbf{w}}}_q$, note that

$$\check{\mathbf{w}}_q = -\mathbf{M}_0^{-1} \sum_{m=1}^{\min(c_1, q-1)} \mathbf{M}_m \check{\check{\mathbf{w}}}_{q-1(m)} = -\mathbf{M}_0^{-1} \sum_{m=1}^{\min(c_1, q-1)} \mathbf{M}_m \check{\mathbf{w}}_{q-m}. \quad (5.3)$$

where the $(N + 1)$ th entry of $\check{\mathbf{w}}_q$ is zero because of (3.3). In addition, the last entry of each of the $\check{\mathbf{w}}_{q-m}$ is zero except for $\check{\mathbf{w}}_1$. Therefore,

$$\mathbf{w}_q = \begin{cases} -\mathbf{A}_0^{-1} \left(-\mathbf{b}_{q-1} + \sum_{m=1}^{q-1} \mathbf{A}_m \mathbf{w}_{q-m} \right) & \text{if } q - 1 \leq c_1 \\ -\mathbf{A}_0^{-1} \left(\sum_{m=1}^{c_1} \mathbf{A}_m \mathbf{w}_{q-m} \right) & \text{otherwise} \end{cases} \quad (5.4)$$

where $\mathbf{A}_m = \mathbf{0}$ for all $m > a_1$ and $\mathbf{b}_{q-1} = \mathbf{0}$ if $q - 1 > b_1$. Note (5.4) can be written more succinctly as

$$\mathbf{w}_q = \begin{cases} -\mathbf{A}_0^{-1} \left(-\mathbf{b}_{q-1} + \sum_{m=1}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{w}_{q-m} \right) & \text{if } q - 1 \leq c_1 \\ -\mathbf{A}_0^{-1} \left(\sum_{m=1}^{a_1} \mathbf{A}_m \mathbf{w}_{q-m} \right) & \text{otherwise} \end{cases} \quad (5.5)$$

where $\mathbf{b}_{q-1} = \mathbf{0}$ if $q - 1 > b_1$. Finally, the most concise way to write (5.5) is

$$\mathbf{w}_q = \mathbf{A}_0^{-1} \left(\mathbf{b}_{q-1} - \sum_{m=1}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{w}_{q-m} \right) \quad (5.6)$$

where $\mathbf{b}_{q-1} = \mathbf{0}$ if $q - 1 > b_1$, which is the last equation in (3.21). \square

It is well known that the Arnoldi method is superior to the power method for generating a Krylov subspace for $\mathbf{C}^{-1}\mathbf{D}$ and $\mathbf{C}^{-1}\mathbf{y}$ given in (3.10) because the Arnoldi method is better conditioned. In addition, since theorem 5.1 shows that the power method on the linearized system (3.9) is equivalent to the AWE technique applied to (3.1) which results in the process (3.21), it is natural to ask if it is possible to modify the AWE vector generating process (3.21) to obtain a well-conditioned method for generating a basis for \mathbf{W}_q that is directly applicable to (3.1) without linearization, yet has the superior convergence properties of the Arnoldi method. The answer to this question is given in section 5.3.

5.3 The WCAWE moment-matching process

5.3.1 A broadband, moment-matching process

To be able to maintain a moment matching process and simultaneously orthonormalize (or even orthogonalize) \mathbf{w}_n from (3.21) against \mathbf{W}_{n-1} before \mathbf{w}_{n+1} is generated, some correction terms must be introduced into (3.21). The resulting vector generating process is

$$\begin{aligned}
 \tilde{\mathbf{v}}_1 &= \mathbf{A}_0^{-1} \mathbf{b}_0 \\
 \tilde{\mathbf{v}}_2 &= \mathbf{A}_0^{-1} (\mathbf{b}_1 \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(2, 1) \mathbf{e}_1 - \mathbf{A}_1 \mathbf{v}_1) \\
 &\vdots \\
 \tilde{\mathbf{v}}_q &= \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \mathbf{A}_1 \mathbf{v}_{q-1} \right. \\
 &\quad \left. - \sum_{m=2}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{V}_{q-m} \mathbf{P}_{\mathbf{U}_2}(q, m) \mathbf{e}_{q-m} \right).
 \end{aligned} \tag{5.7}$$

where \mathbf{e}_r is given in definition 3.3, and $\tilde{\mathbf{V}}_n$ and \mathbf{V}_n are related by an $n \times n$ upper-triangular, nonsingular matrix \mathbf{U} (which can be, but does not have to be, chosen as the coefficients in a modified Gram-Schmidt process to orthonormalize \mathbf{V}_n) by the equation

$$\mathbf{V}_n = \tilde{\mathbf{V}}_n \mathbf{U}^{-1}. \tag{5.8}$$

Furthermore, the correction terms $\mathbf{P}_{\mathbf{U}_w}(n, m)$ are defined in appendix D on page 107. In addition, the proof that the vectors in (5.7) with the relationship (5.8) match moments can also be found in appendix D.

Just as the Arnoldi process provides a much more well-conditioned method for generating a Krylov subspace than the power method, using (5.7) is a much more well-conditioned (and therefore a more broadband) method for generating the moment matching vector subspace than using (3.21). This will be shown in the numerical examples in section 5.4.

5.3.2 Significance of the \mathbf{U} coefficients

Note that no constraints have been placed on the \mathbf{U} matrix except that it is upper triangular and nonsingular. This freedom to choose \mathbf{U} can be exploited to show that WCAWE is actually a generalization of both the AWE and Arnoldi processes. In particular, if \mathbf{U} is chosen as the identity matrix, then it is trivial to see that the WCAWE vectors $\tilde{\mathbf{v}}_n$ from (5.7) reduce to the AWE vectors \mathbf{w}_n from (3.21). On the other hand, in appendix E it is shown that it is possible to choose \mathbf{U} in such a way that the Arnoldi vectors for the expanded, linearized system (3.9) can be produced from the well-conditioned vectors (5.7).

Neither of these choices for \mathbf{U} is used in this work. Of course, on one hand the desire to avoid the ill-conditioned AWE vectors is clear. On the other hand, choosing \mathbf{U} in such a way that the Arnoldi vectors can be produced is not only very complicated, but also not necessarily the best choice. As will be seen in the numerical examples section, WCAWE with \mathbf{U} chosen as the modified Gram-Schmidt coefficients required to orthonormalize \mathbf{V}_n gives a more accurate solution than the PVA process on the expanded, linearized matrix described in section 3.1.1. This is because PVA orthonormalizes the vectors in the space $\mathbb{C}^{c_1(N+1)}$ while WCAWE orthonormalizes the vectors in the space \mathbb{C}^N . For the case of PVA, the leading order terms that are operated on

by $\mathbf{M}_0^{-1}\mathbf{M}_1$ in the matrix $\mathbf{C}^{-1}\mathbf{D}$ in (3.10) are not orthonormal even though the entire \mathbf{z}_n vectors from algorithm 3.1 are. To see this for a simple example, consider the case where $c_1 = 2$ and $N = 2$ and the matrices are real. Then PVA orthonormalizes in the space \mathbb{R}^6 and WCAWE orthonormalizes in the space \mathbb{R}^2 . Assume the first PVA vector is

$$\mathbf{z}_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad (5.9)$$

and the second PVA vector (before orthonormalization) is given to be

$$\begin{bmatrix} 1 \\ 0.1 \\ -1 \\ 0 \\ 1 \\ -1 \end{bmatrix} \quad (5.10)$$

which orthonormalized against \mathbf{z}_1 results in

$$\mathbf{z}_2 = \frac{1}{\sqrt{4.01}} \begin{bmatrix} 1 \\ 0.1 \\ -1 \\ 0 \\ 1 \\ -1 \end{bmatrix}. \quad (5.11)$$

Now the projection of \mathbf{z}_1 and \mathbf{z}_2 onto the leading order subspace \mathbb{R}^2 by the mapping

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (5.12)$$

gives

$$\mathbf{z}_{p1} = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{z}_{p2} = \frac{1}{\sqrt{4.01}} \begin{bmatrix} 1 \\ 0.1 \end{bmatrix}. \quad (5.13)$$

The angle between $\mathbf{z}_{\mathbf{p}_1}$ and $\mathbf{z}_{\mathbf{p}_2}$ is about 5.71° . Of course if the orthonormalization for PVA had been performed in the leading subspace \mathbb{R}^2 then

$$\mathbf{z}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{z}_2 = \begin{bmatrix} 0 \\ 1 \\ -20 \\ -10 \\ 0 \\ -20 \end{bmatrix} \quad (5.14)$$

so in this case

$$\mathbf{z}_{\mathbf{p}_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{z}_{\mathbf{p}_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \quad (5.15)$$

Essentially, PVA produces $\mathbf{z}_{\mathbf{p}_1}$ and $\mathbf{z}_{\mathbf{p}_2}$ as given in (5.13) while WCAWE produces them as given in (5.15). Therefore, in this work \mathbf{U} for WCAWE will not be chosen as the coefficients in appendix E which produce Arnoldi vectors, but rather as the modified Gram-Schmidt coefficients required to orthonormalize \mathbf{V}_n . Note that this is similar to the way in which Krylov vectors are orthogonalized in algorithm 2 in the work [41].

5.3.3 The WCAWE algorithm

Consider algorithm 5.1. It computes, for the matrix equation given in (3.1), the q th WCAWE approximation $\mathbf{x}_q(f)$ with \mathbf{U} chosen as the modified Gram-Schmidt coefficients required to orthonormalize \mathbf{V}_n . Note that in algorithm 5.1 the $q \times q$ matrices $\mathbf{V}_q^T \mathbf{A}_i \mathbf{V}_q$ only need to be computed once. This is where the significant computational saving are obtained: each iteration of the last **for** loop requires the inverse of a $q \times q$ matrix instead of the $N \times N$ matrix $\mathbf{A}(f)$ where $q \ll N$.

Algorithm 5.1 (WCAWE process with modified Gram-Schmidt)

$$\tilde{\mathbf{v}}_1 = \mathbf{A}_0^{-1} \mathbf{b}_0$$

$$\mathbf{U}_{[1,1]} = \|\tilde{\mathbf{v}}_1\|$$

$$\mathbf{v}_1 = \tilde{\mathbf{v}}_1 \mathbf{U}_{[1,1]}^{-1}$$

for $n = 2, 3, \dots, q$ do

$$\tilde{\mathbf{v}}_n = \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, n-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(n, m) \mathbf{e}_{n-m}) - \mathbf{A}_1 \mathbf{v}_{n-1} \right. \\ \left. - \sum_{m=2}^{\min(a_1, n-1)} \mathbf{A}_m \mathbf{V}_{n-m} \mathbf{P}_{\mathbf{U}_2}(n, m) \mathbf{e}_{n-m} \right)$$

for $\alpha = 1, 2, \dots, n-1$ do

$$\mathbf{U}_{[\alpha, n]} = \mathbf{v}_\alpha^H \tilde{\mathbf{v}}_n$$

$$\tilde{\mathbf{v}}_n = \tilde{\mathbf{v}}_n - \mathbf{U}_{[\alpha, n]} \mathbf{v}_\alpha$$

endfor

$$\mathbf{U}_{[n, n]} = \|\tilde{\mathbf{v}}_n\|$$

$$\mathbf{v}_n = \tilde{\mathbf{v}}_n \mathbf{U}_{[n, n]}^{-1}$$

endfor

for any desired f in the range $f_{min} \leq f \leq f_{max}$ do

$$\sigma = j2\pi f - s_0$$

$$\mathbf{g}_q(f) = \left(\sum_{i=0}^{a_1} \sigma^i \mathbf{V}_q^T \mathbf{A}_i \mathbf{V}_q \right)^{-1} \left(\sum_{k=0}^{b_1} \sigma^k \mathbf{V}_q^T \mathbf{b}_k \right)$$

$$\mathbf{x}_q(f) = \mathbf{V}_q \mathbf{g}_q(f)$$

endfor

□

5.4 Numerical examples

Example 1: The TE_z scattering problem from subsection 2.2.3 is solved. Since this is a small example (1276 unknowns), the problem can be expanded and linearized so the Arnoldi method discussed in subsection 3.1.1 can be used to find the solution vector with $\mathbf{L} = \mathbf{I}$. In addition, the WCAWE method discussed in this chapter was also applied to find $\mathbf{x}_q(f)$. For each technique a total of 30 iterations were performed with an expansion point corresponding to 250MHz. In Figure 5.1 the relative errors (3.36) in the solution vector (with $\mathbf{L} = \mathbf{I}$) for both PVA and WCAWE are shown.

In this simulation, all that is desired is for WCAWE to maintain essentially the same accuracy as the Arnoldi process, and then to claim superiority from the fact that WCAWE does not require the matrix equation to be expanded and linearized (and therefore requires less memory to store the q ROM vectors). However, as shown in Figure 5.1, the new WCAWE method has a smaller relative error; therefore, as alluded to in subsection 5.3.2, the WCAWE method (with \mathbf{U} chosen as the modified Gram-Schmidt coefficients required to orthonormalize \mathbf{V}_n) is not only as accurate as PVA, but is even more accurate.

Finally, recall that this is a small example with only 1276 unknowns. In the example to follow PVA will not be applicable because it will not be possible to store q ROM vectors of doubled length ($c_1 = 2$) for that particular simulation.

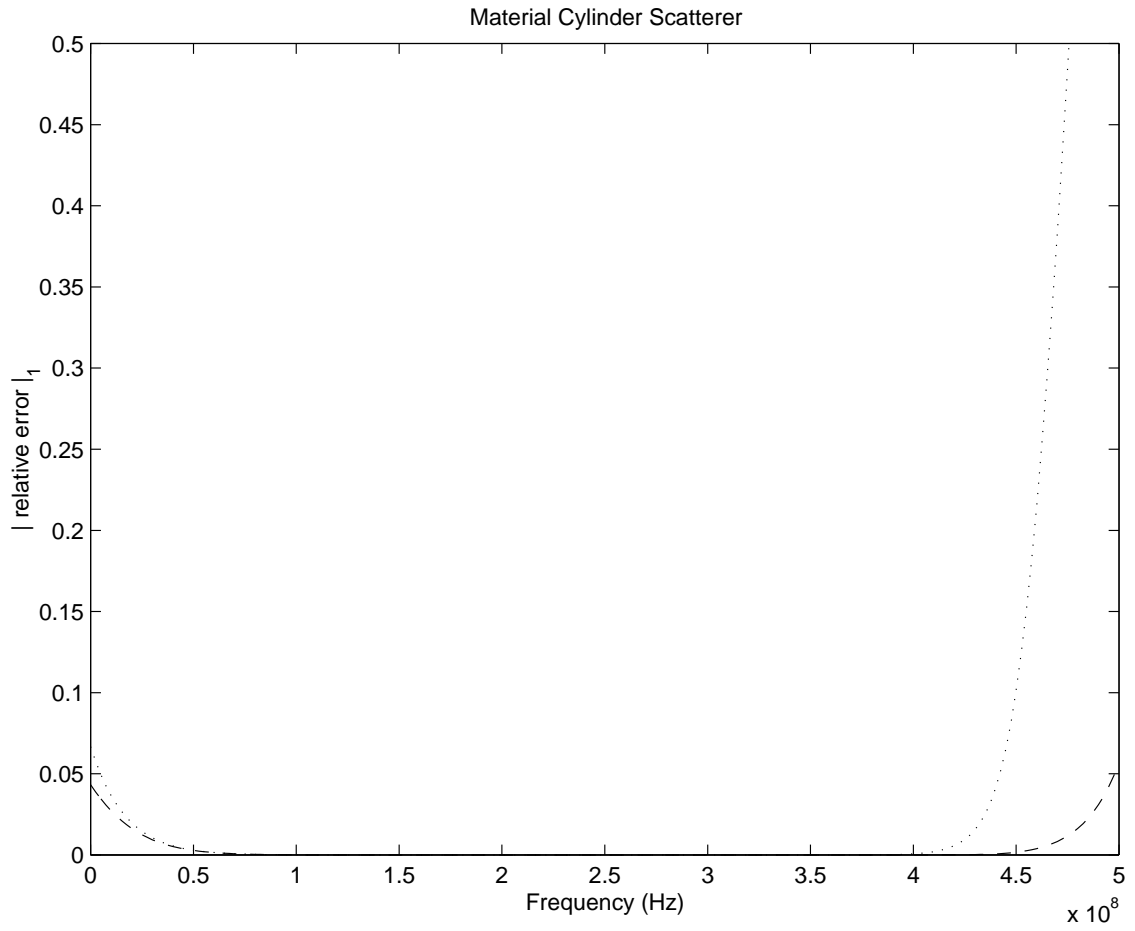


Figure 5.1: Relative error in the solution vector for the example in subsection 2.2.3 with $q = 30$. Dot-dot \rightarrow PVA method, dash-dash \rightarrow WCAWE method.

Example 2: The broadband bowtie antenna described in subsection 2.2.4 is solved using WCAWE and compared to GAWE from subsection 3.2.2 with an expansion point corresponding to 2750MHz. Figure 5.2 shows the condition number for the traditional AWE vectors ($\mathbf{W}_n^T \mathbf{W}_n$ from (3.21)) along with the new WCAWE vectors ($\tilde{\mathbf{V}}_n^T \tilde{\mathbf{V}}_n$ from (5.7)) for $n = 1, 2, \dots, q$. As can be seen, WCAWE is indeed much more well-conditioned.

To illustrate the accuracy of the new WCAWE method, Figure 5.3 shows S_{11} for the input to the antenna, where the circles are the exact solution of the original FEM system (2.32), the dash-dot curve is the classical GAWE solution method in subsection 3.2.2, and the dash-dash curve is the WCAWE solution method discussed in this chapter. As can be observed, the WCAWE method is accurate throughout the entire simulated band.

Note that since this example is so large (884670 unknowns), the Arnoldi solution method from subsection 3.1.1 can not be used for comparison since expanding and linearizing the system would result in over 1.75 million unknowns, and the memory required for $q = 110$ vectors of that length would be greater than the available resources.

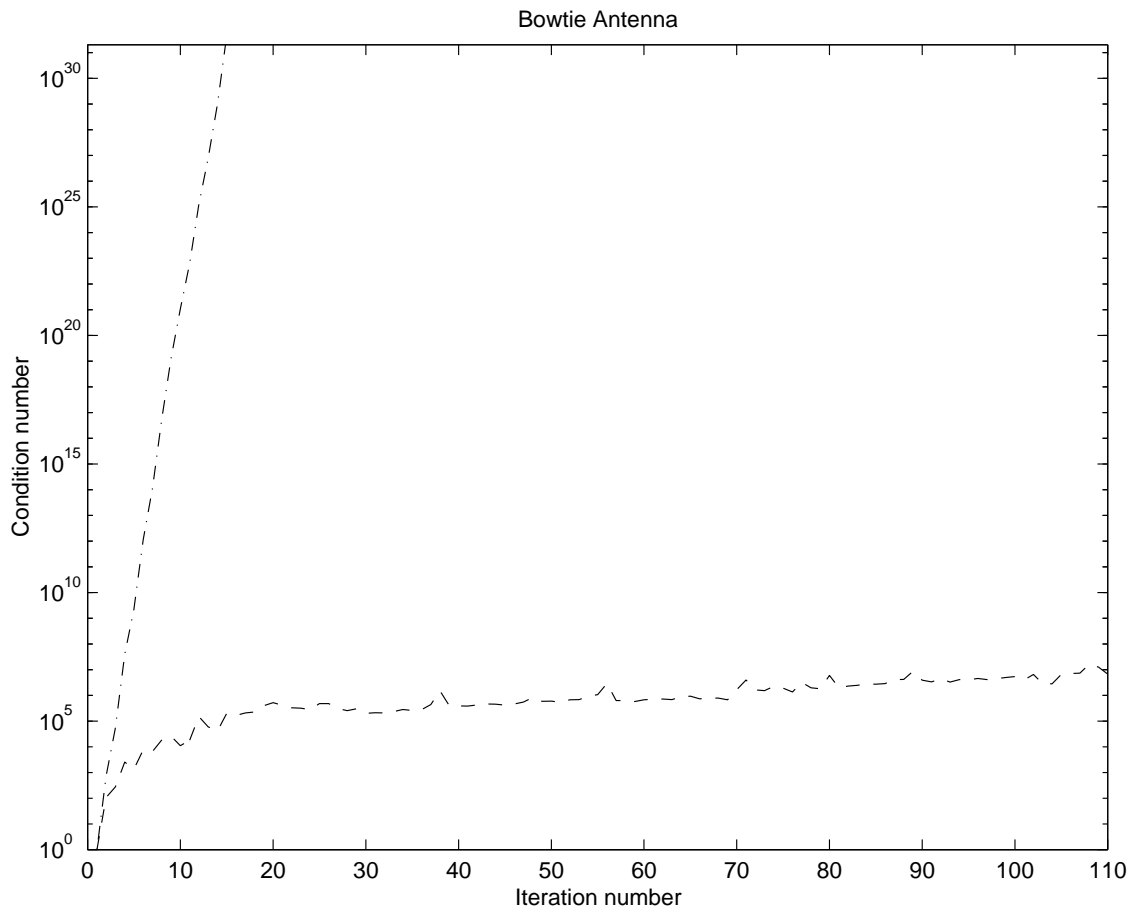


Figure 5.2: Condition number versus order n . Dash-dot \rightarrow AWE (and GAWE) method, dash-dash \rightarrow WCAWE method.

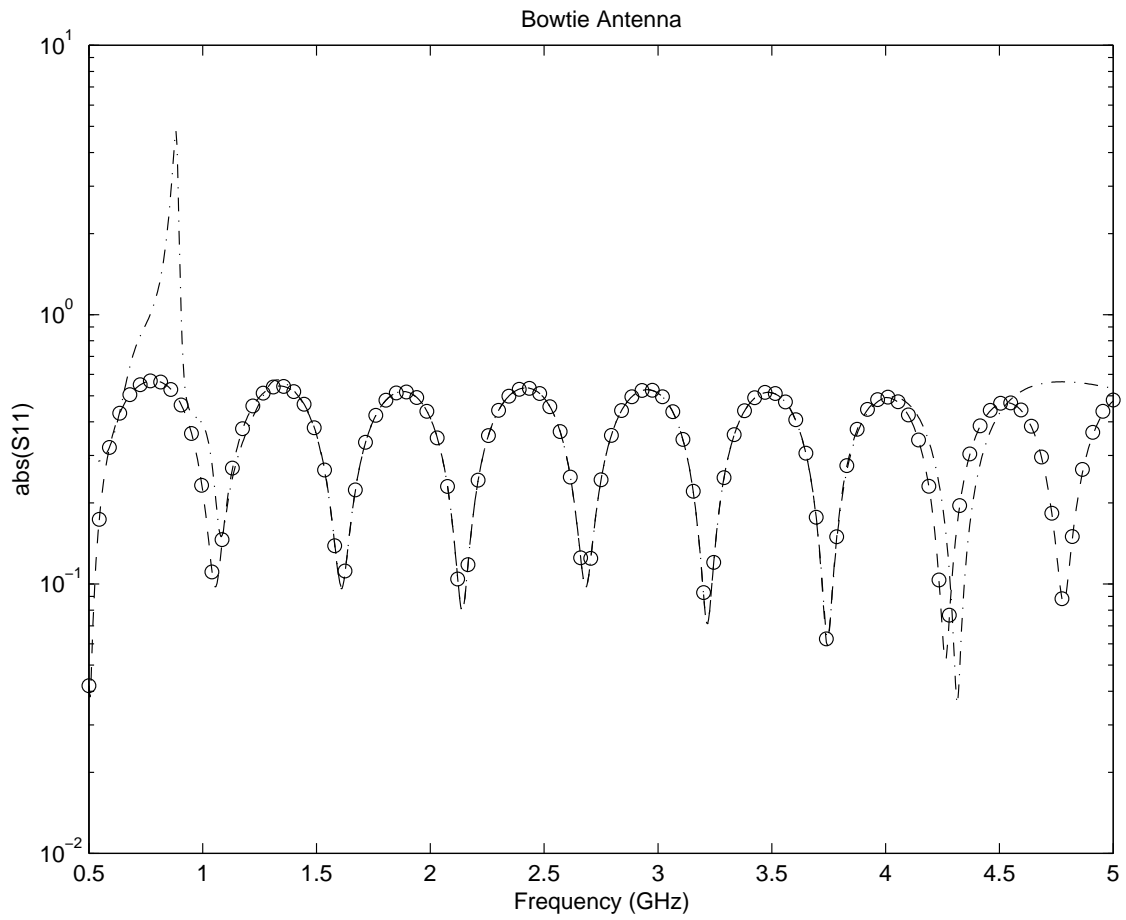


Figure 5.3: S_{11} for the bowtie antenna with $q = 110$. Circles \rightarrow solution to (2.32), dash-dot \rightarrow GAWE method, dash-dash \rightarrow WCAWE method.

Example 3: The low pass filter described in subsection 2.2.4 is simulated using WCAWE and compared to GAWE from subsection 3.2.2. The expansion points are chosen to correspond to 11GHz. Figure 5.4 is synonymous to Figure 5.2. Again, notice how much more well-conditioned the vectors generated for the ROM subspace are for the WCAWE process (5.7) than they are for the AWE process (3.21).

Figure 5.5 shows S_{12} for this example. As before, the circles are the exact solution of the original FEM system (2.32), the dash-dot curve is the classical GAWE solution method from subsection 3.2.2, and the dash-dash curve is the WCAWE solution method. As before, the WCAWE method is again accurate throughout the entire simulated band, while the GAWE method is not.

Since the band pass filter described in subsection 2.2.4 is successfully simulated by MGAWE with only one expansion point in section 4.4 (that is, GAWE is successful in that simulation), there is no reason to simulate that example in this section with WCAWE.

After establishing the accuracy obtainable with a single WCAWE expansion point, attention will now shift to the robustness of the WCAWE method. This will be shown in the numerical example to follow.

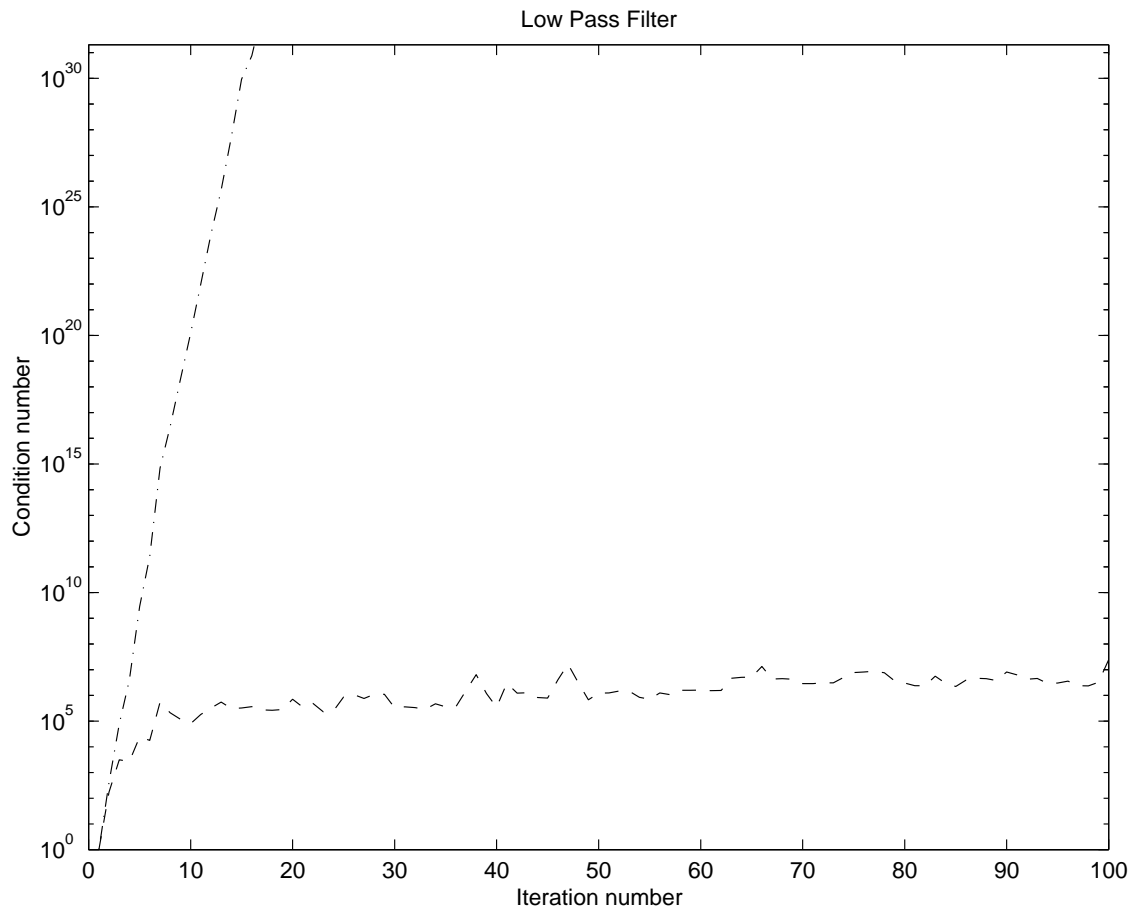


Figure 5.4: Condition number versus order n . Dash-dot \rightarrow AWE (and GAWE) method, dash-dash \rightarrow WCAWE method.

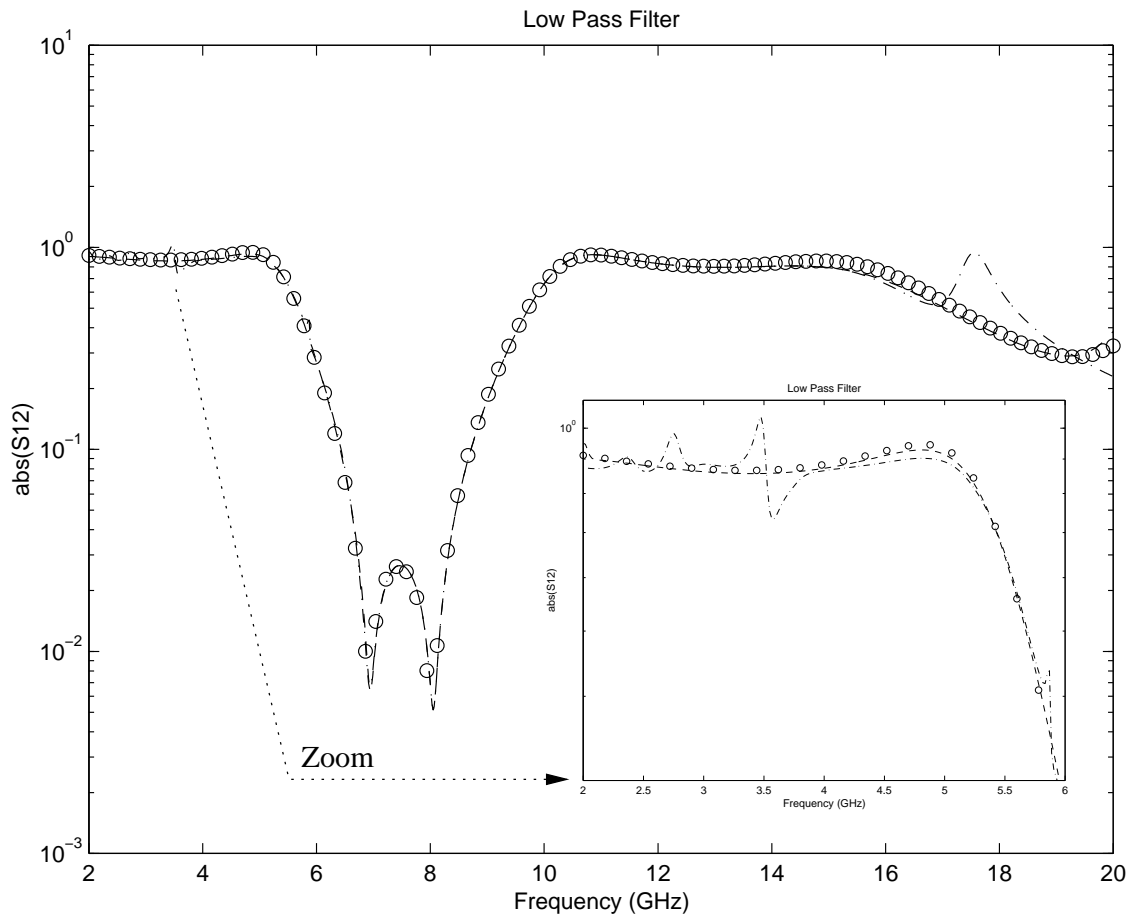


Figure 5.5: S_{12} for the low pass filter with $q = 100$. Circles \rightarrow solution to (2.32), dash-dot \rightarrow GAWE method, dash-dash \rightarrow WCAWE method.

Example 4: The horn antenna described in subsection 2.2.1 is simulated using GAWE from subsection 3.2.2 as well as WCAWE. Two different simulations are performed on the horn: one from $f_{min} = 400\text{MHz}$ with $f_{max} = 1000\text{MHz}$ and the other with $f_{min} = 430.013\text{MHz}$ and $f_{max} = 1000\text{MHz}$. For the first simulation, the midpoint of the band where the expansion point is located is $s_0 = j2\pi 700\text{MHz}$ and for the second simulation $s_0 = j2\pi 715.0065\text{MHz}$, which is very close to one of the poles in Figure 4.13 located at $j2\pi (715.006398 + j1.80)\text{MHz}$. The idea is that a method which is not very robust will stagnate more quickly (that is, the condition number will grow more rapidly) when the expansion point moves closer to a singularity such as a pole. The results of this investigation are shown in Figure 5.6 for $s_0 = j2\pi 700\text{MHz}$ and in Figure 5.7 for $s_0 = j2\pi 715.0065\text{MHz}$. In the low frequency region the solution is easy to capture because the field is evanescent there (the waveguide feeding the horn antenna is designed for cutoff to be at about 700MHz). Therefore, concentrate on the high frequency region above 700MHz. Indeed, GAWE proves to not be very robust because the iterative process stagnates (the unconverged frequency region stops shrinking) after 17 iterations for Figure 5.6 (when the expansion point is further from the pole), but stagnates after only 8 iterations for Figure 5.7 (when the expansion point is closer to the pole). However, in both cases WCAWE shows it is a robust method because it does not stagnate; the location of the expansion point with respect to poles does not adversely effect convergence of the method.

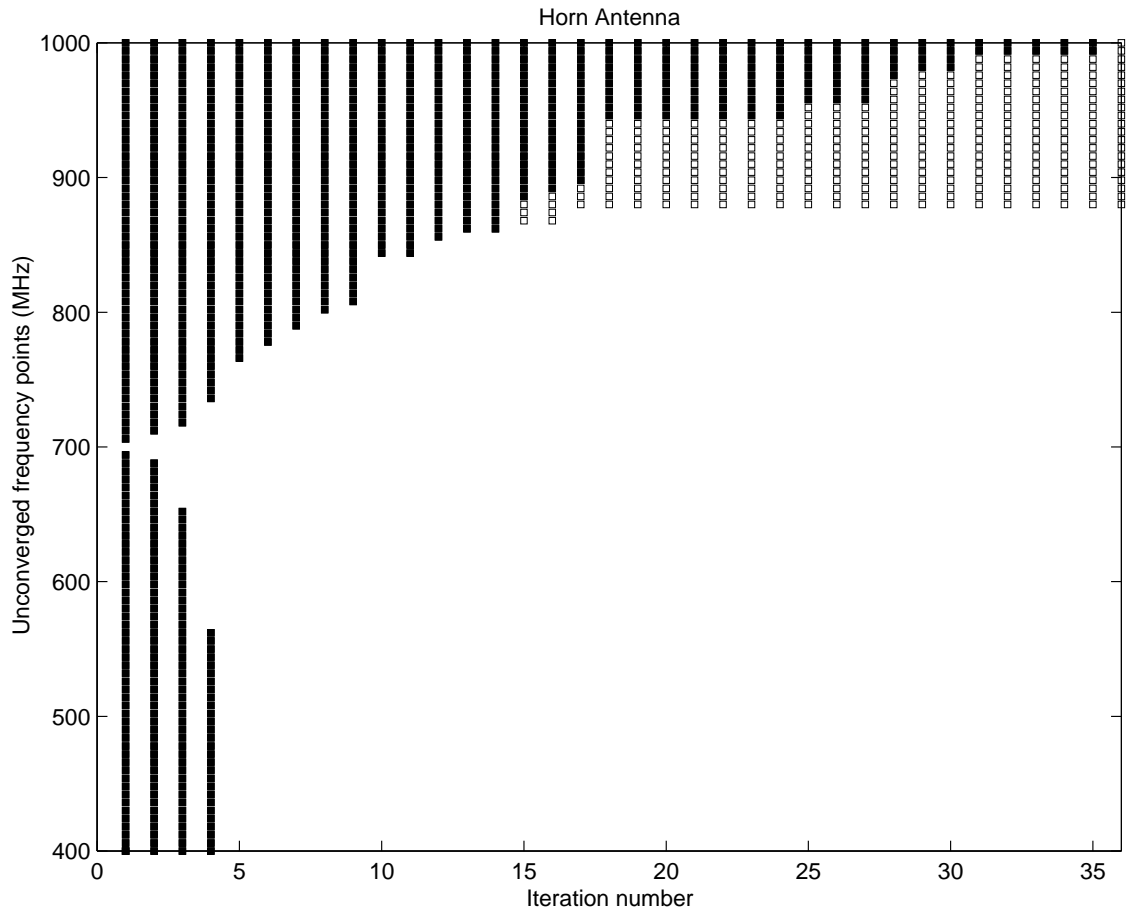


Figure 5.6: Frequency band convergence versus GAW and WCAWE iterations for the horn antenna with expansion points corresponding to 700MHz. Open squares → unconverged GAW points, closed squares → both GAW and WCAWE unconverged points.

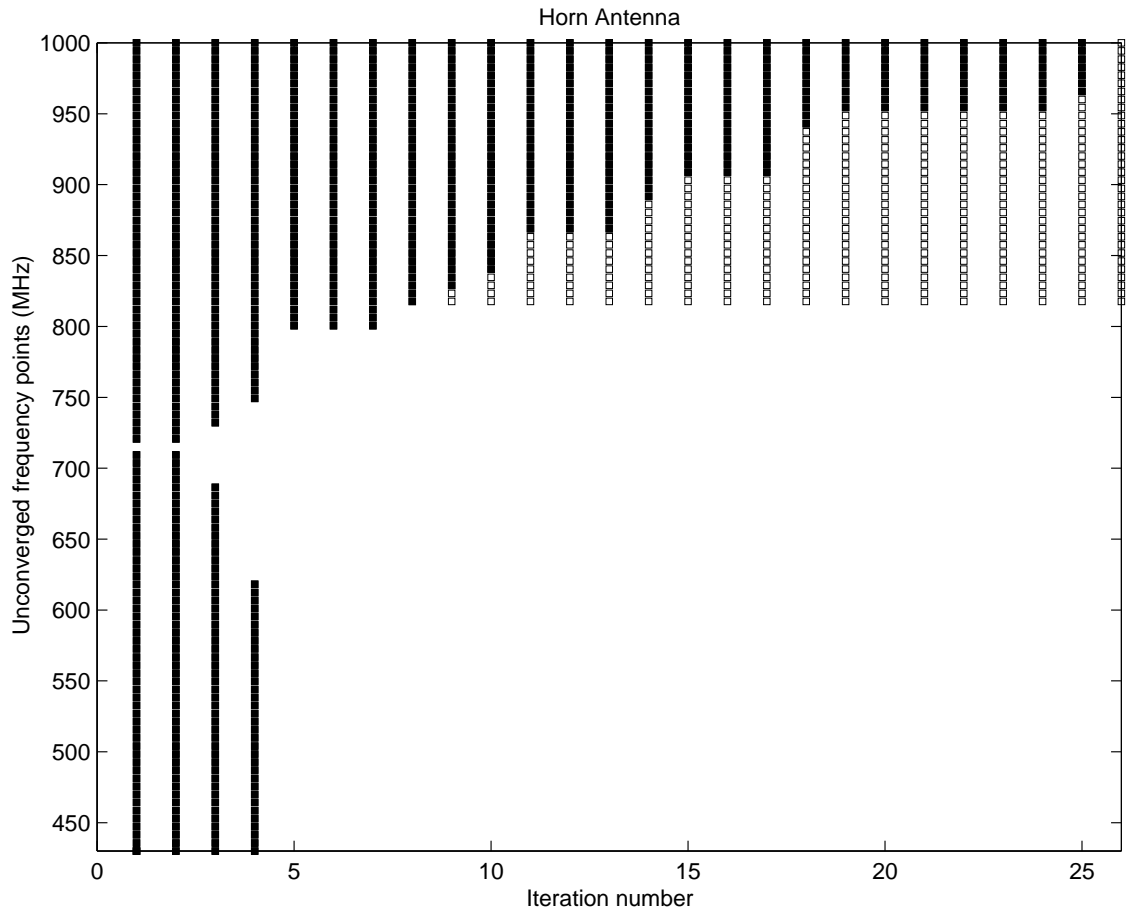


Figure 5.7: Frequency band convergence versus GAWE and WCAWE iterations for the horn antenna with expansion points corresponding to approximately 715MHz. Open squares → unconverged GAWE points, closed squares → both GAWE and WCAWE unconverged points.

CHAPTER 6

SUMMARY, CONCLUSIONS AND FUTURE STUDY

6.1 Summary of the findings and conclusions drawn

In this dissertation a broad scope of MORE is given in the context of the literature review in the beginning of chapter 1. Then a more detailed discussion of the aspect of MORE known as fast parameter sweeping is given to set the stage for the remainder of this work.

In chapter 2 the mathematical description of a fast frequency sweep MORE problem is outlined. Then several examples are given that are used throughout the remainder of this text. These examples include both two and three dimensional geometries, and computational domains terminated with either absorbing boundary conditions or anisotropic, dispersive perfectly matched layers. Furthermore, both radiation as well as scattering problems are solved, and it is illustrated how a FFS MORE technique must treat each of these cases differently for the right hand sides. Finally, simulations are performed for computing the S parameters of microwave devices.

The discussion in chapter 3 centers around classical MORE techniques. However, before discussing these classical solution methodologies, it is shown how to write a polynomial matrix equation with a polynomial forcing vector as a linear matrix

equation with a constant right hand side. Next two Krylov subspace techniques, the Arnoldi and Lanczos processes, are reviewed. Following the Krylov subspace techniques some moment-matching methods that are directly applicable to polynomial systems are shown. These methods include both AWE and its Galerkin treatment, GAWE. Finally, chapter 3 closes with a comparison of these classical MORE techniques.

In chapter 4 the automated MGAWWE process is presented. To automate the process, it is shown how to determine the number of expansion points needed, where they should be located, and how large the subspace order should be at each of them. These details are then followed by several numerical examples that illustrate the accuracy and efficiency of the MGAWWE method. After the initial numerical investigation, a deeper look is taken at the issues of how the computational efficiency depends on a tolerance value in the process, how MGAWWE compares to a rational polynomial interpolation with respect to robustness as well as efficiency, and how it is important to choose the evaluation frequencies as a function of the pole distribution of the system.

In chapter 5 the WCAWE method is presented. In the beginning of the chapter motivation is given for the development of the technique as well as the disclosure of a related technique that should be avoided. Then the connection between the AWE method on a polynomial system and the Krylov subspace created by a power method on the expanded, linearized system is given. This connection is critical in understanding the underlying process followed in the creation of the WCAWE algorithm. After presenting the WCAWE process along with a discussion of the significance of the mapping matrix \mathbf{U} , several numerical examples are given that illustrate the accuracy and robustness of the process.

In summary, for polynomial systems, the MGAWE process appears to be most useful when there is not a large computational overhead associated with switching expansion points. On the other hand, the WCAWE process would appear to be preferable when switching expansion points is undesirable (for example, when a direct matrix solver such as an **LU** decomposition is used, or when computing the preconditioner for an iterative solver is a significant computational expense).

6.2 Future study

One area of future study that is proposed is to extend the MGAWE and WCAWE processes to block and/or multivariable versions. For MGAWE, this extension should be straightforward, but for WCAWE with dispersive right hand sides there are difficult issues that will arise. Suppose there are two vectors in the block right hand side. If one wishes to orthonormalize, during the subspace generation, vectors seeded by the first vector against vectors seeded by the second vector then it appears there will be extreme difficulty. This is related to the idea of orthonormalizing, during the subspace generation, vectors from a Krylov subspace for one matrix against vectors from a Krylov subspace for a different matrix. The reason the matrices would be different can be seen in definition 3.2 where the matrix \mathbf{M}_i depends on the vector \mathbf{b}_i .

Nevertheless, a block version extension of either MGAWE or WCAWE would mean that the method could handle a radiation problem in which the excitation contains a few elements that are out of phase and/or are non-uniformly excited. In addition, it would allow the analysis of multiport devices. On the other hand, a multivariable extension would mean that the processes could perform fast sweep MORE in both the frequency and angle domains simultaneously.

The final proposed area of future study is to solve a problem in which $\mathbf{A}(s)$ contains exponential variations of the ROM varying parameter. An example of this type of problem is a fast frequency sweep for an infinite antenna array. In this example $\mathbf{A}(s)$ will contain exponential frequency variations because of the periodic boundary conditions. In this case it may be advantageous to keep a_1 and b_1 small because of storage requirements. Using an iterative solver means that even though many derivatives may not be generated at a particular expansion point, more expansion points can be evaluated without being concerned about the need to perform a new \mathbf{LU} decomposition each time the expansion point changes (as long as the computation of the preconditioner is not prohibitive). Furthermore, applying MGAW or WCAWE to problems modeled using the method of moments will also encounter $\mathbf{A}(s)$ matrices that contain exponential variations in s .

Finally, WCAWE needs some more research in a couple of practical implementation issue areas. These include finding a robust, efficient termination scheme to find the value needed for q , and finding a way to decrease the high number that can be necessary for q to obtain a wide-band response with just one expansion point.

APPENDIX A

MATRIX PADÉ VIA LANCZOS ALGORITHM

The following Matrix Padé via Lanczos (MPVL) algorithm is taken from [27]. It uses exact deflation and no look-ahead on the matrix $\mathbf{C}^{-1}\mathbf{D}$ with the right starting vectors \mathbf{y}_i for $i = 1, 2, \dots, p$ and left starting vectors \mathbf{l}_i for $i = 1, 2, \dots, o$.

Algorithm A.1 (q steps of the MPVL process)

```

for  $i = 1, 2, \dots, p$  set  $\boldsymbol{\psi}_i = \mathbf{y}_i$ 
for  $i = 1, 2, \dots, o$  set  $\boldsymbol{\xi}_i = \mathbf{l}_i$ 
set  $p_c = p$  and  $o_c = o$ 
for  $n = 1, 2, \dots, q$  do
  while  $\|\boldsymbol{\psi}_n\| = 0$  do
    if  $p_c = 1$  then stop
    for  $i = n, n + 1, \dots, n + p_c - 2$  set  $\boldsymbol{\psi}_i = \boldsymbol{\psi}_{i+1}$ 
    set  $p_c = p_c - 1$ 
  endwhile
  while  $\|\boldsymbol{\xi}_n\| = 0$  do
    if  $o_c = 1$  then stop
    for  $i = n, n + 1, \dots, n + o_c - 2$  set  $\boldsymbol{\xi}_i = \boldsymbol{\xi}_{i+1}$ 
    set  $o_c = o_c - 1$ 
  endwhile
  set  $\mu = n - p_c$  and  $\phi = n - o_c$ 
  set  $t_{n,\mu} = \|\boldsymbol{\psi}_n\|$  and  $\tilde{t}_{n,\phi} = \|\boldsymbol{\xi}_n\|$ 
  set  $\boldsymbol{\psi}_n = \boldsymbol{\psi}_n/t_{n,\mu}$  and  $\boldsymbol{\xi}_n = \boldsymbol{\xi}_n/\tilde{t}_{n,\phi}$ 

```

```

set  $\delta_n = \boldsymbol{\xi}_n^T \boldsymbol{\psi}_n$ 
set  $\boldsymbol{\psi} = \mathbf{C}^{-1} \mathbf{D} \boldsymbol{\psi}_n$ 
set  $i_v = \max\{1, n - o_c\}$ 
for  $i = i_v, i_v + 1, \dots, n - 1$  do
  if  $i = n - o_c$  then
    set  $t_{i,n} = \tilde{t}_{n,i} \delta_n / \delta_i$ 
  else
    set  $t_{i,n} = \boldsymbol{\xi}_i^T \boldsymbol{\psi} / \delta_i$ 
  endif
  set  $\boldsymbol{\psi} = \boldsymbol{\psi} - \boldsymbol{\psi}_i t_{i,n}$ 
endfor
set  $\boldsymbol{\psi}_{n+p_c} = \boldsymbol{\psi}$ 
set  $\boldsymbol{\xi} = (\mathbf{C}^{-1} \mathbf{D})^T \boldsymbol{\xi}_n$ 
set  $i_w = \max\{1, n - p_c\}$ 
for  $i = i_w, i_w + 1, \dots, n - 1$  do
  if  $i = n - p_c$  then
    set  $\tilde{t}_{i,n} = t_{n,i} \delta_n / \delta_i$ 
  else
    set  $\tilde{t}_{i,n} = \boldsymbol{\xi}^T \boldsymbol{\psi}_i / \delta_i$ 
  endif
  set  $\boldsymbol{\xi} = \boldsymbol{\xi} - \boldsymbol{\xi}_i \tilde{t}_{i,n}$ 
endfor
set  $\boldsymbol{\xi}_{n+o_c} = \boldsymbol{\xi}$ 
if  $\delta_n = 0$  then stop
for  $i = n - p_c + 1, n - p_c + 2, \dots, n$  do
  if  $i \leq 0$  or  $i = n$  then
    set  $t_{n,i} = \boldsymbol{\xi}_n^T \boldsymbol{\psi}_{p_c+i} / \delta_n$ 
  else
    set  $t_{n,i} = \tilde{t}_{i,n} \delta_i / \delta_n$ 
  endif
  set  $\boldsymbol{\psi}_{p_c+i} = \boldsymbol{\psi}_{p_c+i} - \boldsymbol{\psi}_n t_{n,i}$ 
endfor

```

```

for  $i = n - o_c + 1, n - o_c + 2, \dots, n$  do
  if  $i \leq 0$  then
    set  $\tilde{t}_{n,i} = \boldsymbol{\xi}_{o_c+i}^T \boldsymbol{\psi}_n / \delta_n$ 
  else
    set  $\tilde{t}_{n,i} = t_{i,n} \delta_i / \delta_n$ 
  endif
  set  $\boldsymbol{\xi}_{o_c+i} = \boldsymbol{\xi}_{o_c+i} - \boldsymbol{\xi}_n \tilde{t}_{n,i}$ 
endfor

if  $n \leq p_c$  then
  for  $i = n - p_c + p, n - p_c + p + 1, \dots, p$  do
    set  $\rho_{n,i} = t_{n,i-p}$ 
  endfor
  set  $p_1 = p_c$ 
endif

if  $n \leq o_c$  then
  for  $i = n - o_c + o, n - o_c + o + 1, \dots, o$  do
    set  $\eta_{n,i} = \tilde{t}_{n,i-o}$ 
  endfor
  set  $o_1 = o_c$ 
endif

endfor

for  $i = 1, 2, \dots, q$  do
  for  $j = 1, 2, \dots, q$  do
    set  $T_{i,j} = t_{i,j}$ 
  endfor
endfor

for  $i = p_1 + 1, p_1 + 2, \dots, q$  do
  for  $j = 1, 2, \dots, p$  do
    set  $\rho_{i,j} = 0$ 
  endfor
endfor

```

```
for  $i = o_1 + 1, o_1 + 2, \dots, q$  do
  for  $j = 1, 2, \dots, o$  do
    set  $\eta_{i,j} = 0$ 
  endfor
endfor
for  $i = 1, 2, \dots, q$  do
  for  $j = 1, 2, \dots, o$  do
    set  $\eta_{i,j} = \delta_i \eta_{i,j}$ 
  endfor
endfor
```

□

APPENDIX B

PROOF THAT (3.21) SATISFIES (3.32)

To start the proof, for the step $n = 1$, assume $\mathbf{x}(f) = \mathbf{0}$ which then results in $\mathbf{r}_0(f) = -\sum_{k=0}^{b_1} \sigma^k \mathbf{b}_k$. Therefore, $\mathbf{r}_0^0 = -\mathbf{b}_0$. Choose $\mathbf{A}_0 \bar{\mathbf{w}}_1$ in the direction (or -1 times the direction) of \mathbf{r}_0^0 , so there is a component of the solution that can span that part of the residual, thus forcing this component of the residual to zero as required in the definition involving $\mathbf{r}_q^l = \mathbf{0}$ for $l = 0 \dots q-1$ as given in (3.32). Therefore, choose

$$\bar{\mathbf{w}}_1 = \mathbf{A}_0^{-1} \mathbf{b}_0 \tag{B.1}$$

and let $\bar{\mathbf{W}}_1 = [\bar{\mathbf{w}}_1]$. Now assume that $\mathbf{x}(f) = \bar{\mathbf{W}}_1 \mathbf{g}_1(f) = \bar{\mathbf{w}}_1 \gamma_1(f)$, so one can find $\mathbf{r}_1(f) = \sum_{i=0}^{a_1} (\sigma^i \mathbf{A}_i) \bar{\mathbf{w}}_1 \gamma_1(f) - \sum_{k=0}^{b_1} \sigma^k \mathbf{b}_k$. Therefore, $\mathbf{r}_1^0 = \mathbf{A}_0 \bar{\mathbf{w}}_1 \gamma_1^0 - \mathbf{b}_0 = \mathbf{0}$ if $\gamma_1^0 = 1$.

For the step $n = 2$, again assume $\mathbf{x}(f) = \bar{\mathbf{W}}_1 \mathbf{g}_1(f) = \bar{\mathbf{w}}_1 \gamma_1(f)$, which now results in $\mathbf{r}_1^1 = \mathbf{A}_0 \bar{\mathbf{w}}_1 \gamma_1^1 + \mathbf{A}_1 \bar{\mathbf{w}}_1 \gamma_1^0 - \mathbf{b}_1 = \mathbf{A}_0 \bar{\mathbf{w}}_1 \gamma_1^1 + \mathbf{A}_1 \bar{\mathbf{w}}_1 - \mathbf{b}_1$. To find $\bar{\mathbf{w}}_2$, consider \mathbf{r}_1^1 . Choose $\mathbf{A}_0 \bar{\mathbf{w}}_2$ in the direction of \mathbf{r}_1^1 so there is a component of the solution that can span that part of the residual, thus forcing this component of the residual to zero as required in the definition involving \mathbf{r}_q^l given in (3.32). Note that it is the case that $-\mathbf{A}_0^{-1} \mathbf{r}_1^1 = -\bar{\mathbf{w}}_1 \gamma_1^1 + \mathbf{A}_0^{-1} (\mathbf{b}_1 - \mathbf{A}_1 \bar{\mathbf{w}}_1)$. Since $\bar{\mathbf{W}}_1$ already spans $\bar{\mathbf{w}}_1$, there is no need to include $\bar{\mathbf{w}}_1 \gamma_1^1$ when constructing $\bar{\mathbf{w}}_2$. Therefore, choose

$$\bar{\mathbf{w}}_2 = \mathbf{A}_0^{-1} (\mathbf{b}_1 - \mathbf{A}_1 \bar{\mathbf{w}}_1) \tag{B.2}$$

and let $\overline{\mathbf{W}}_2 = [\overline{\mathbf{w}}_1 \overline{\mathbf{w}}_2]$. Now that a better approximation for $\mathbf{x}(f)$ is available, assume $\mathbf{x}(f) = \overline{\mathbf{W}}_2 \mathbf{g}_2(f) = \overline{\mathbf{w}}_1 \gamma_1(f) + \overline{\mathbf{w}}_2 \gamma_2(f)$. Following the same procedure as before, obtain $\mathbf{r}_2^0 = \mathbf{A}_0 \overline{\mathbf{w}}_1 \gamma_1^0 + \mathbf{A}_0 \overline{\mathbf{w}}_2 \gamma_2^0 - \mathbf{b}_0$. Therefore, if γ_2^0 is chosen to be zero, then $\mathbf{g}_2^0 = [\gamma_1^0 \ 0]^T = [1 \ 0]^T$ and $\mathbf{r}_2^0 = \mathbf{r}_1^0 = \mathbf{0}$. In addition, it is now clearly the case that $\mathbf{r}_2^1 = \mathbf{A}_0 \overline{\mathbf{W}}_2 \mathbf{g}_2^1 + \mathbf{A}_1 \overline{\mathbf{W}}_2 \mathbf{g}_2^0 - \mathbf{b}_1 = \mathbf{A}_0 \overline{\mathbf{W}}_2 \mathbf{g}_2^1 + \mathbf{A}_1 \overline{\mathbf{w}}_1 - \mathbf{b}_1$. Note that if \mathbf{g}_2^1 is chosen to be $\mathbf{g}_2^1 = [0 \ 1]^T$ then $\mathbf{r}_2^1 = \mathbf{0}$.

For the step $n = 3$, again assume $\mathbf{x}(f) = \overline{\mathbf{W}}_2 \mathbf{g}_2(f) = \overline{\mathbf{w}}_1 \gamma_1(f) + \overline{\mathbf{w}}_2 \gamma_2(f)$. To find $\overline{\mathbf{w}}_3$, consider \mathbf{r}_2^2 . Choose $\mathbf{A}_0 \overline{\mathbf{w}}_3$ in the direction of \mathbf{r}_2^2 so there is a component of the solution that can span that part of the residual, again forcing this component of the residual to zero as required in the above definition. Now note that it is the case that $-\mathbf{A}_0^{-1} \mathbf{r}_2^2 = -\overline{\mathbf{W}}_2 \mathbf{g}_2^2 + \mathbf{A}_0^{-1} (\mathbf{b}_2 - \mathbf{A}_1 \overline{\mathbf{W}}_2 \mathbf{g}_2^1 - \mathbf{A}_2 \overline{\mathbf{W}}_2 \mathbf{g}_2^0)$. But $\mathbf{g}_2^1 = [0 \ 1]^T$ and $\mathbf{g}_2^0 = [1 \ 0]^T$. In addition, since $\overline{\mathbf{W}}_2$ is already spanned, there is no need to consider $\overline{\mathbf{W}}_2 \mathbf{g}_2^2$ when constructing $\overline{\mathbf{w}}_3$. Therefore, choose

$$\overline{\mathbf{w}}_3 = \mathbf{A}_0^{-1} (\mathbf{b}_2 - \mathbf{A}_1 \overline{\mathbf{w}}_2 - \mathbf{A}_2 \overline{\mathbf{w}}_1) \quad (\text{B.3})$$

and let $\overline{\mathbf{W}}_3 = [\overline{\mathbf{w}}_1 \overline{\mathbf{w}}_2 \overline{\mathbf{w}}_3]$. Now assume $\mathbf{x}(f) = \overline{\mathbf{W}}_3 \mathbf{g}_3(f)$. Again following the same procedure, $\mathbf{r}_3^0 = \mathbf{A}_0 \overline{\mathbf{W}}_3 \mathbf{g}_3^0 - \mathbf{b}_0$, and so $\mathbf{r}_3^0 = \mathbf{r}_1^0 = \mathbf{0}$ if $\gamma_3^0 = 0$. In addition, it is now the case that $\mathbf{r}_3^1 = \mathbf{A}_0 \overline{\mathbf{W}}_3 \mathbf{g}_3^1 + \mathbf{A}_1 \overline{\mathbf{W}}_3 \mathbf{g}_3^0 - \mathbf{b}_1$. If $\gamma_3^1 = 0$ then $\mathbf{r}_3^1 = \mathbf{r}_2^1 = \mathbf{0}$. Moreover, $\mathbf{r}_3^2 = \mathbf{A}_0 \overline{\mathbf{W}}_3 \mathbf{g}_3^2 + \mathbf{A}_1 \overline{\mathbf{W}}_3 \mathbf{g}_3^1 + \mathbf{A}_2 \overline{\mathbf{W}}_3 \mathbf{g}_3^0 - \mathbf{b}_2$ and since $\mathbf{g}_3^0 = [1 \ 0 \ 0]^T$ and $\mathbf{g}_3^1 = [0 \ 1 \ 0]^T$ if \mathbf{g}_3^2 is chosen to be $\mathbf{g}_3^2 = [0 \ 0 \ 1]^T$ then $\mathbf{r}_3^2 = \mathbf{A}_0 \overline{\mathbf{w}}_3 + \mathbf{A}_1 \overline{\mathbf{w}}_2 + \mathbf{A}_2 \overline{\mathbf{w}}_1 - \mathbf{b}_2$ which gives $\mathbf{r}_3^2 = \mathbf{A}_0 \mathbf{A}_0^{-1} (\mathbf{b}_2 - \mathbf{A}_1 \overline{\mathbf{w}}_2 - \mathbf{A}_2 \overline{\mathbf{w}}_1) + \mathbf{A}_1 \overline{\mathbf{w}}_2 + \mathbf{A}_2 \overline{\mathbf{w}}_1 - \mathbf{b}_2 = \mathbf{0}$.

For the step $n = 4 \dots q$, again assume $\mathbf{x}(f) = \overline{\mathbf{W}}_{n-1} \mathbf{g}_{n-1}(f)$. To find $\overline{\mathbf{w}}_n$, consider \mathbf{r}_{n-1}^{n-1} . Choose $\mathbf{A}_0 \overline{\mathbf{w}}_n$ in the direction of \mathbf{r}_{n-1}^{n-1} so there is a component of the solution that can span that part of the residual, again forcing this component of the residual to zero.

Note that $-\mathbf{A}_0^{-1}\mathbf{r}_{n-1}^{n-1} = -\overline{\mathbf{W}}_{n-1}\mathbf{g}_{n-1}^{n-1} + \mathbf{A}_0^{-1}\left(\mathbf{b}_{n-1} - \sum_{m=1}^{\min(a_1, n-1)} (\mathbf{A}_m \overline{\mathbf{W}}_{n-1} \mathbf{g}_{n-1}^{n-m-1})\right)$ where $\mathbf{b}_{n-1} = \mathbf{0}$ if $n-1 > b_1$. But \mathbf{g}_{n-1}^{n-m-1} for $m = 1, 2, \dots, n-1$ has already been chosen such that the only nonzero entry is in position $n-m$. In addition, since $\overline{\mathbf{W}}_{n-1}$ is already spanned, there is no need to consider $\overline{\mathbf{W}}_{n-1}\mathbf{g}_{n-1}^{n-1}$ when constructing $\overline{\mathbf{w}}_n$. Therefore, choose

$$\overline{\mathbf{w}}_n = \mathbf{A}_0^{-1} \left(\mathbf{b}_{n-1} - \sum_{m=1}^{\min(a_1, n-1)} (\mathbf{A}_m \overline{\mathbf{w}}_{n-m}) \right) \quad (\text{B.4})$$

where again $\mathbf{b}_{n-1} = \mathbf{0}$ if $n-1 > b_1$. Now let $\overline{\mathbf{W}}_n = [\overline{\mathbf{w}}_1 \dots \overline{\mathbf{w}}_n]$ and furthermore assume $\mathbf{x}(f) = \overline{\mathbf{W}}_n \mathbf{g}_n(f)$. As before, $\mathbf{r}_n^0 \dots \mathbf{r}_n^{n-2} = \mathbf{0}$ if $\gamma_n^0 \dots \gamma_n^{n-2}$ are all chosen to be zero. Now, $\mathbf{r}_n^{n-1} = \sum_{m=0}^{n-1} (\mathbf{A}_m \overline{\mathbf{W}}_n \mathbf{g}_n^{n-m-1}) - \mathbf{b}_{n-1}$ and if \mathbf{g}_n^{n-1} is chosen such that the only nonzero entry is a 1 in the last position, then $\mathbf{r}_n^{n-1} = \mathbf{0}$ as required.

Choosing $\overline{\mathbf{W}}_q$ as shown in equations (B.1) through (B.4) satisfies (3.32). Note, however, that equations (B.1) through (B.4) are the same as (3.21). Therefore, (3.21) satisfies (3.32).

APPENDIX C

COUNTEREXAMPLES TO SHOW VECTORS FROM (5.1) DO NOT MATCH MOMENTS

Recall that the vectors from (3.21) are

$$\begin{aligned}
 \mathbf{w}_1 &= \mathbf{A}_0^{-1} \mathbf{b}_0 \\
 \mathbf{w}_2 &= \mathbf{A}_0^{-1} (\mathbf{b}_1 - \mathbf{A}_1 \mathbf{w}_1) \\
 \mathbf{w}_3 &= \mathbf{A}_0^{-1} (\mathbf{b}_2 - \mathbf{A}_1 \mathbf{w}_2 - \mathbf{A}_2 \mathbf{w}_1) \\
 &\vdots \\
 \mathbf{w}_q &= \mathbf{A}_0^{-1} \left(\mathbf{b}_{q-1} - \sum_{m=1}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{w}_{q-m} \right)
 \end{aligned}$$

and that the vectors from (5.1) are

$$\begin{aligned}
 \widehat{\mathbf{w}}_1 &= \mathbf{A}_0^{-1} \mathbf{b}_0 \\
 \widehat{\mathbf{w}}_2 &= \mathbf{A}_0^{-1} (\mathbf{b}_1 - \mathbf{A}_1 \widehat{\mathbf{w}}_1) \\
 \widehat{\mathbf{w}}_3 &= \mathbf{A}_0^{-1} (\mathbf{b}_2 - \mathbf{A}_1 \widehat{\mathbf{w}}_2 - \mathbf{A}_2 \widehat{\mathbf{w}}_1) \\
 &\vdots \\
 \widehat{\mathbf{w}}_q &= \mathbf{A}_0^{-1} \left(\mathbf{b}_{q-1} - \sum_{m=1}^{\min(a_1, q-1)} \mathbf{A}_m \widehat{\mathbf{w}}_{q-m} \right)
 \end{aligned}$$

where $\mathbf{b}_k = \mathbf{0}$ for $k > b_1$. If the subspace $\widehat{\mathbf{W}}_q$ matches moments, then it must be the case that

$$\text{span}(\widehat{\mathbf{W}}_n) = \text{span}(\mathbf{W}_n) \quad \forall n \ni 1 \leq n \leq q. \quad (\text{C.1})$$

However, since $\widehat{\mathbf{W}}_q$ is generated from $\widehat{\mathbf{W}}_q$ by an orthonormalization process, it is always the case that

$$\text{span}(\widehat{\mathbf{W}}_n) = \text{span}(\widehat{\mathbf{W}}_n) \quad \forall n \ni 1 \leq n \leq q. \quad (\text{C.2})$$

Therefore, the requirement given in (C.1) is equivalent to the following requirement:

$$\text{span}(\widehat{\mathbf{W}}_n) = \text{span}(\mathbf{W}_n) \quad \forall n \ni 1 \leq n \leq q. \quad (\text{C.3})$$

The following counterexamples show (C.3) does not hold for the cases outlined.

C.1 Case one: right hand side linear or higher in σ

The requirement given in (C.3) can fail to be true for values of n as low as 2. For example, although $\mathbf{w}_1 = \widehat{\mathbf{w}}_1$, assume $\|\widehat{\mathbf{w}}_1\| = u_{11} \neq 1$. Then $\widetilde{\mathbf{w}}_1 = \mathbf{w}_1/u_{11}$ and so $\widehat{\mathbf{w}}_2 = \mathbf{A}_0^{-1}(\mathbf{b}_1 - \mathbf{A}_1\mathbf{w}_1/u_{11}) \notin \text{span}(\mathbf{W}_2)$. Therefore, $\text{span}(\widehat{\mathbf{W}}_2) \neq \text{span}(\mathbf{W}_2)$ so in general (5.1) does not match moments for $b_1 \geq 1$.

C.2 Case two: constant right hand side and matrix quadratic or higher in σ

To see that (5.1) fails to match moments for this case, it is necessary to generate vectors up to $n = 3$. As before, $\mathbf{w}_1 = \widehat{\mathbf{w}}_1$ and again assume $\|\widehat{\mathbf{w}}_1\| = u_{11} \neq 1$ so $\widetilde{\mathbf{w}}_1 = \mathbf{w}_1/u_{11}$. But now $n = 2$ works: $\widehat{\mathbf{w}}_2 = -\mathbf{A}_0^{-1}\mathbf{A}_1\mathbf{w}_1/u_{11} = \mathbf{w}_2/u_{11}$. So $\text{span}(\widehat{\mathbf{W}}_2) = \text{span}(\mathbf{W}_2)$. Nevertheless, for $n = 3$ problems arise. Note that $\widetilde{\mathbf{w}}_2$

is formed by orthonormalizing $\widehat{\mathbf{w}}_2$ against $\widehat{\mathbf{w}}_1$, that is, $\widehat{\mathbf{w}}_2 = (\widehat{\mathbf{w}}_2 - u_{12}\widehat{\mathbf{w}}_1)/u_{22} = (\mathbf{w}_2 - u_{12}\mathbf{w}_1)/(u_{11}u_{22})$. Now

$$\begin{aligned}
\widehat{\mathbf{w}}_3 &= -\mathbf{A}_0^{-1} (\mathbf{A}_1\widehat{\mathbf{w}}_2 + \mathbf{A}_2\widehat{\mathbf{w}}_1) & (C.4) \\
&= -\mathbf{A}_0^{-1} (\mathbf{A}_1(\mathbf{w}_2 - u_{12}\mathbf{w}_1)/(u_{11}u_{22}) + \mathbf{A}_2\mathbf{w}_1/u_{11}) \\
&= -\mathbf{A}_0^{-1} (\mathbf{A}_1\mathbf{w}_2/u_{22} + \mathbf{A}_2\mathbf{w}_1)/u_{11} - u_{12}\mathbf{w}_2/(u_{11}u_{22}) \\
&\notin \text{span}(\mathbf{W}_3).
\end{aligned}$$

Therefore (5.1) does not match moments for the case $b_1 = 0$ with $a_1 \geq 2$.

As a final note, it may appear that the process might work if one tried to only orthogonalize instead of orthonormalize. However, considering $n = 4$ will show that orthogonalization also will fail to match moments.

APPENDIX D

PROOF TO SHOW VECTORS FROM (5.7) MATCH MOMENTS

Before proving WCAWE matches moments, several useful facts and definitions are given. Each small fact is given without proof since they are easy to establish. In the following, the matrix \mathbf{U} is the upper-triangular, nonsingular matrix from (5.8) that maps one vector basis to another. Of course, when one basis happens to be orthonormal, \mathbf{U} will be modified Gram-Schmidt coefficients.

Fact D.1 Let \mathbf{U}_1 and \mathbf{U}_2 be upper triangular matrices. Then the product $\mathbf{U}_1\mathbf{U}_2$ is an upper triangular matrix. Furthermore, \mathbf{U}_1 is nonsingular if and only if all its diagonal entries are nonzero. \square

Definition D.1 Given an $n \times n$ matrix \mathbf{U} and four integers i_1 , i_2 , j_1 , and j_2 such that $1 \leq i_1 \leq i_2 \leq n$ and $1 \leq j_1 \leq j_2 \leq n$, let $\mathbf{U}_{[i_1, j_1]}$ be the entry in \mathbf{U} at the intersection of row i_1 and column j_1 . Furthermore, let $\mathbf{U}_{[i_1: i_2, j_1: j_2]}$ denote the block matrix extracted from \mathbf{U} starting from row i_1 and going through row i_2 from columns j_1 through j_2 .

Fact D.2 Let \mathbf{U} be an $n \times n$ upper triangular, nonsingular matrix. Then \mathbf{U}^{-1} is upper triangular. Furthermore, for any integers j_1 and j_2 such that $1 \leq j_1 \leq j_2 \leq n$

the equality $(\mathbf{U}^{-1})_{[j_1:j_2, j_1:j_2]} = (\mathbf{U}_{[j_1:j_2, j_1:j_2]})^{-1}$ holds.

Remark: Any entry in the block $(\mathbf{U}^{-1})_{[j_1:j_2, j_1:j_2]}$ is independent of entries in \mathbf{U} outside the block $\mathbf{U}_{[j_1:j_2, j_1:j_2]}$. Therefore, if a small upper triangular matrix is embedded in a larger upper triangular matrix, then the block (at the position of the embedded matrix) in the inverse of the large matrix is just the inverse of the small matrix before embedding. This fact is needed to establish facts D.3 and D.4, and the notation $\mathbf{U}_{[j_1:j_2, j_1:j_2]}^{-1}$ will be used throughout the remainder of this paper without considering if the inverse is taken before or after the block is selected. \square

Definition D.2 Let \mathbf{U} be an $n \times n$ upper triangular, nonsingular matrix. For some integers m , η and w where $w = 1$ or 2 and $w \leq m < \eta \leq n$, define the $(\eta - m) \times (\eta - m)$ matrix

$$\mathbf{P}_{\mathbf{U}_w}(\eta, m) = \prod_{t=w}^m \mathbf{U}_{[t:\eta-m+t-1, t:\eta-m+t-1]}^{-1} \quad (\text{D.1})$$

where $\prod_{t=1}^2 \mathbf{U}_t^{-1} = \mathbf{U}_1^{-1} \mathbf{U}_2^{-1}$.

Remark: First notice that $\mathbf{P}_{\mathbf{U}_w}(\eta, m)$ is a function of the integers η and m . Furthermore, $\mathbf{P}_{\mathbf{U}_w}(\eta, m)$ is just a composition of many blocks extracted from the mapping \mathbf{U} . In definition D.4 it will be obvious that $\mathbf{P}_{\mathbf{U}_w}(\eta, m)$ is the matrix that tracks the mappings from one vector space to another for the higher order terms in the WCAWE process. \square

Fact D.3 Assume the integers α , $\bar{\alpha}$ and γ satisfy $1 \leq \gamma < \min(\alpha, \bar{\alpha})$. Then for all integers j_1 and j_2 which satisfy $1 \leq j_1, j_2 \leq \min(\alpha, \bar{\alpha}) - \gamma$ the equality

$$\mathbf{e}_{j_1}^T \mathbf{P}_{\mathbf{U}_1}(\alpha, \gamma) \mathbf{e}_{j_2} = \mathbf{e}_{j_1}^T \mathbf{P}_{\mathbf{U}_1}(\bar{\alpha}, \gamma) \mathbf{e}_{j_2} \quad (\text{D.2})$$

holds.

Remark: This fact follows from definition D.2 along with facts D.1 and D.2; its physical significance is as follows. Consider some vector space that has a subspace that is growing larger (for example, growing from α to $\bar{\alpha}$) with some iterative process. Also consider two bases for this subspace. At each iteration a new vector is added to each basis. This new vector may immediately be replaced with a linear combination of itself with the previous vectors, but the previous vectors can not be modified (this makes \mathbf{U} upper triangular). Fact D.3 says the mapping $\mathbf{P}_{\mathbf{U}_1}$ between the first $\alpha - 1$ vectors of these two bases for iterations $\bar{\alpha} > \alpha$ is exactly the same mapping that existed between these $\alpha - 1$ vectors at iteration α . Fact D.3 is needed to help prove theorem D.2. \square

Fact D.4 Let \mathbf{U} be the matrix in definition D.2. For some integers β, m, η and w where $w = 1$ or $2, w \leq m < \eta \leq n$ and $1 \leq \beta \leq n - 1$ then for all integers j_1 and j_2 such that $\beta \leq j_1, j_2 \leq \eta - m$ the equality

$$\mathbf{e}_{j_1}^T \mathbf{P}_{\mathbf{U}_w}(\eta, m) \mathbf{e}_{j_2} = \mathbf{e}_{j_1 - \beta + 1}^T \prod_{t=w}^m \mathbf{U}_{[t+\beta-1:\eta-m+t-1, t+\beta-1:\eta-m+t-1]}^{-1} \mathbf{e}_{j_2 - \beta + 1} \quad (\text{D.3})$$

holds.

Remark: This fact also follows from definition D.2 along with facts D.1 and D.2. Just like the remark after fact D.2, all this fact says is an entry in the inverse of some upper triangular matrix (formed from the product of upper triangular matrices \mathbf{U} with embedded blocks) can be found without considering entries in the \mathbf{U} matrices outside those blocks. This fact is needed for the proof of theorem D.2. \square

Definition D.3 Let \mathbf{X} be a $n \times n$ upper triangular matrix whose entries are

$$\mathbf{X}_{[j_1, j_2]} = \begin{cases} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(j_2, j_1 - 1) \mathbf{e}_{j_2 - j_1 + 1} & \text{for } 2 \leq j_1 \leq j_2 \leq n \\ 1 & \text{for } j_1 = j_2 = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (\text{D.4})$$

Furthermore, \mathbf{X} is nonsingular since \mathbf{U} is nonsingular.

Remark: The \mathbf{X} matrix is actually never computed, but is necessary to facilitate the proof of theorem D.3. The physical significance of \mathbf{X} is that $\mathbf{X}_{[1:n-1, n]}$ are the coefficients of the first $n - 1$ AWE vectors that are implicitly removed from the n th AWE vector (which is implicitly scaled by $\mathbf{X}_{[n, n]}$) to generate the n th vector in the novel well-conditioned process that is presented in definition D.4. The key is that the vector is generated with these components already removed and the scaling already applied, instead of requiring these processes to be performed post vector generation. This makes the numerical properties of the novel WCAWE algorithm far superior to AWE. \square

Theorem D.2 (Result used in the proof of theorem D.3) Let \mathbf{X} be the matrix in definition D.3. Then for some integers β , m and η where $1 \leq m < \eta \leq n$ and $1 \leq \beta \leq \eta - m$

$$\mathbf{X}_{[\beta, 1: \eta - m]} \mathbf{P}_{\mathbf{U}_1}(\eta, m) \mathbf{e}_{\eta - m} = \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(\eta, m + \beta - 1) \mathbf{e}_{\eta - m - \beta + 1}. \quad (\text{D.5})$$

Proof: This is an equality between scalars. Let $\alpha_\beta = \mathbf{X}_{[\beta, 1: \eta - m]} \mathbf{P}_{\mathbf{U}_1}(\eta, m) \mathbf{e}_{\eta - m}$. Since the proof is trivial for $\beta = 1$ (because the first row of \mathbf{X} is \mathbf{e}_1^T from definition D.3), consider the case $2 \leq \beta \leq \eta - m$. Expand \mathbf{X} into a summation; from definition D.3 note $\mathbf{X}_{[\beta, 1: \beta - 1]} = \mathbf{0}$ so

$$\alpha_\beta = \sum_{r=\beta}^{\eta - m} \mathbf{X}_{[\beta, r]} \mathbf{e}_r^T \mathbf{P}_{\mathbf{U}_1}(\eta, m) \mathbf{e}_{\eta - m}.$$

Now use definition D.3 to write the scalar $\mathbf{X}_{[\beta,r]}$ as

$$\alpha_\beta = \sum_{r=\beta}^{\eta-m} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(r, \beta-1) \mathbf{e}_{r-\beta+1} \mathbf{e}_r^T \mathbf{P}_{\mathbf{U}_1}(\eta, m) \mathbf{e}_{\eta-m}$$

and use fact D.3 with $\alpha = r$, $\bar{\alpha} = \eta - m$, $\gamma = \beta - 1$, $j_1 = 1$ and $j_2 = r - \beta + 1$ to obtain

$$\alpha_\beta = \sum_{r=\beta}^{\eta-m} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(\eta - m, \beta - 1) \mathbf{e}_{r-\beta+1} \mathbf{e}_r^T \mathbf{P}_{\mathbf{U}_1}(\eta, m) \mathbf{e}_{\eta-m}.$$

Now use fact D.4 with $j_1 = r$ and $j_2 = \eta - m$ to obtain

$$\alpha_\beta = \sum_{r=\beta}^{\eta-m} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(\eta-m, \beta-1) \mathbf{e}_{r-\beta+1} \mathbf{e}_r^T \prod_{t=1}^m \mathbf{U}_{[t+\beta-1:\eta-m+t-1, t+\beta-1:\eta-m+t-1]}^{-1} \mathbf{e}_{\eta-m-\beta+1}$$

which can be simplified by noting that $\sum_{r=\beta}^{\eta-m} \mathbf{e}_{r-\beta+1} \mathbf{e}_r^T$ is an identity matrix:

$$\alpha_\beta = \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(\eta - m, \beta - 1) \prod_{t=1}^m \mathbf{U}_{[t+\beta-1:\eta-m+t-1, t+\beta-1:\eta-m+t-1]}^{-1} \mathbf{e}_{\eta-m-\beta+1}.$$

Now invoke definition D.2 to give

$$\alpha_\beta = \mathbf{e}_1^T \prod_{t=1}^{\beta-1} \mathbf{U}_{[t:\eta-m-\beta+1+t-1, t:\eta-m-\beta+1+t-1]}^{-1} \prod_{t=1}^m \mathbf{U}_{[t+\beta-1:\eta-m+t-1, t+\beta-1:\eta-m+t-1]}^{-1} \mathbf{e}_{\eta-m-\beta+1}.$$

Carefully note the products can be written compactly (after shifting the second t index) as

$$\alpha_\beta = \mathbf{e}_1^T \prod_{t=1}^{m+\beta-1} \mathbf{U}_{[t:\eta-m-\beta+1+t-1, t:\eta-m-\beta+1+t-1]}^{-1} \mathbf{e}_{\eta-m-\beta+1}.$$

Finally, use definition D.2 again to give $\alpha_\beta = \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(\eta, m + \beta - 1) \mathbf{e}_{\eta-m-\beta+1}$. \square

Fact D.5 Consider some function $f(m, \beta)$ which is summed over the data points shown on the left in Figure D.1. Further assume the sum is reordered as shown on the right of Figure D.1. Then

$$\sum_{m=1}^{\min(a_1, q-1)} \sum_{\beta=1}^{q-m} f(m, \beta) = \sum_{j_1=1}^{q-1} \sum_{j_2=1}^{\min(a_1, j_1)} f(j_2, j_1 - j_2 + 1). \quad (\text{D.6})$$

Remark: This fact is needed in the proof of theorem D.3 so some quantities can be accessed in a way that makes them more easily identifiable. \square

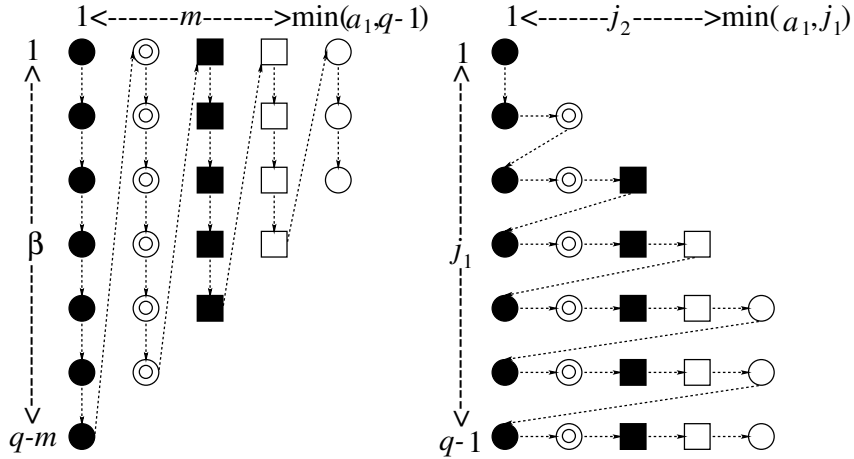


Figure D.1: Reordering the data access.

Fact D.6 Let \mathbf{U} be the matrix in definition D.2. For some integers m and η where $2 \leq m < \eta \leq n$ the equality

$$\mathbf{U}_{[1:\eta-m, 1:\eta-m]}^{-1} \mathbf{P}_{\mathbf{U}_2}(\eta, m) = \mathbf{P}_{\mathbf{U}_1}(\eta, m) \quad (\text{D.7})$$

holds.

Remark: This fact follows trivially from definition D.2; it is needed to prove theorem D.3. \square

Definition D.4 (WCAWE vectors for (3.1)) Let \mathbf{U} be the $q \times q$ matrix in definition D.2, $\tilde{\mathbf{V}}_q$ be the collection of q N -vectors $\tilde{\mathbf{v}}_1$ through $\tilde{\mathbf{v}}_q$, and \mathbf{V}_q be the collection of q N -vectors \mathbf{v}_1 through \mathbf{v}_q where $\mathbf{V}_q = \tilde{\mathbf{V}}_q \mathbf{U}^{-1}$ and

$$\begin{aligned}
\tilde{\mathbf{v}}_1 &= \mathbf{A}_0^{-1} \mathbf{b}_0 & (D.8) \\
\tilde{\mathbf{v}}_2 &= \mathbf{A}_0^{-1} (\mathbf{b}_1 \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(2, 1) \mathbf{e}_1 - \mathbf{A}_1 \mathbf{v}_1) \\
\tilde{\mathbf{v}}_3 &= \mathbf{A}_0^{-1} (\mathbf{b}_1 \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(3, 1) \mathbf{e}_2 + \mathbf{b}_2 \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(3, 2) \mathbf{e}_1 - \mathbf{A}_1 \mathbf{v}_2 - \mathbf{A}_2 \mathbf{V}_1 \mathbf{P}_{\mathbf{U}_2}(3, 2) \mathbf{e}_1) \\
&\vdots \\
\tilde{\mathbf{v}}_q &= \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \mathbf{A}_1 \mathbf{v}_{q-1} \right. \\
&\quad \left. - \sum_{m=2}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{V}_{q-m} \mathbf{P}_{\mathbf{U}_2}(q, m) \mathbf{e}_{q-m} \right).
\end{aligned}$$

Remark: This definition for the vectors is the same as given in (5.7) and (5.8). \square

Fact D.7 Since \mathbf{U} is nonsingular, $\text{span}(\mathbf{V}_q) = \text{span}(\tilde{\mathbf{V}}_q)$. \square

Theorem D.3 (Result used in the proof of theorem D.4) Let \mathbf{W}_q be as given in (3.21), \mathbf{X} as given in definition D.3 and $\tilde{\mathbf{V}}_q$ as given in definition D.4. Then $\tilde{\mathbf{V}}_q = \mathbf{W}_q \mathbf{X}_{[1:q, 1:q]}$ and therefore $\text{span}(\tilde{\mathbf{V}}_q) = \text{span}(\mathbf{W}_q)$.

Proof: The proof is by induction.

Clearly $\tilde{\mathbf{v}}_1 = \mathbf{w}_1 \mathbf{X}_{[1,1]}$. Therefore, $\text{span}(\tilde{\mathbf{v}}_1) = \text{span}(\mathbf{w}_1)$.

For the case $q = 2$,

$$\begin{aligned}
\tilde{\mathbf{v}}_2 &= \mathbf{A}_0^{-1} (\mathbf{b}_1 \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(2, 1) \mathbf{e}_1 - \mathbf{A}_1 \mathbf{v}_1) = \mathbf{A}_0^{-1} (\mathbf{b}_1 \mathbf{U}_{[1,1]}^{-1} - \mathbf{A}_1 \tilde{\mathbf{v}}_1 \mathbf{U}_{[1,1]}^{-1}) \\
&= \mathbf{A}_0^{-1} (\mathbf{b}_1 - \mathbf{A}_1 \mathbf{w}_1) \mathbf{U}_{[1,1]}^{-1} = \mathbf{w}_2 \mathbf{U}_{[1,1]}^{-1} = \mathbf{w}_2 \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(2, 1) \mathbf{e}_1 = \mathbf{w}_2 \mathbf{X}_{[2,2]}.
\end{aligned}$$

Therefore, $\tilde{\mathbf{V}}_2 = \mathbf{W}_2 \mathbf{X}_{[1:2, 1:2]}$ and $\text{span}(\tilde{\mathbf{V}}_2) = \text{span}(\mathbf{W}_2)$.

Now assume $\tilde{\mathbf{V}}_{q-1} = \mathbf{W}_{q-1} \mathbf{X}_{[1:q-1, 1:q-1]}$ and so $\text{span}(\tilde{\mathbf{V}}_{q-1}) = \text{span}(\mathbf{W}_{q-1})$. It will be shown that $\tilde{\mathbf{V}}_q = \mathbf{W}_q \mathbf{X}_{[1:q, 1:q]}$. Note

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \mathbf{A}_1 \mathbf{v}_{q-1} - \sum_{m=2}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{V}_{q-m} \mathbf{P}_{\mathbf{U}_2}(q, m) \mathbf{e}_{q-m} \right)$$

but $\mathbf{V}_{q-m} = \tilde{\mathbf{V}}_{q-m} \mathbf{U}_{[1:q-m, 1:q-m]}^{-1}$ and furthermore by assumption it is the case that $\tilde{\mathbf{V}}_{q-m} = \mathbf{W}_{q-m} \mathbf{X}_{[1:q-m, 1:q-m]}$ so

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \mathbf{A}_1 \mathbf{W}_{q-1} \mathbf{X}_{[1:q-1, 1:q-1]} \mathbf{U}_{[1:q-1, 1:q-1]}^{-1} \mathbf{e}_{q-1} - \sum_{m=2}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{W}_{q-m} \mathbf{X}_{[1:q-m, 1:q-m]} \mathbf{U}_{[1:q-m, 1:q-m]}^{-1} \mathbf{P}_{\mathbf{U}_2}(q, m) \mathbf{e}_{q-m} \right).$$

Now use fact D.6 so to obtain

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \mathbf{A}_1 \mathbf{W}_{q-1} \mathbf{X}_{[1:q-1, 1:q-1]} \mathbf{U}_{[1:q-1, 1:q-1]}^{-1} \mathbf{e}_{q-1} - \sum_{m=2}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{W}_{q-m} \mathbf{X}_{[1:q-m, 1:q-m]} \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m} \right).$$

From definition D.2 note that $\mathbf{U}_{[1:q-1, 1:q-1]}^{-1} = \mathbf{P}_{\mathbf{U}_1}(q, 1)$ so \mathbf{A}_1 can be absorbed into the summation to give

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \sum_{m=1}^{\min(a_1, q-1)} \mathbf{A}_m \mathbf{W}_{q-m} \mathbf{X}_{[1:q-m, 1:q-m]} \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m} \right).$$

Now note that $\mathbf{W}_{q-m} = \sum_{\beta=1}^{q-m} \mathbf{w}_\beta \mathbf{e}_\beta^T$, so

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \sum_{m=1}^{\min(a_1, q-1)} \mathbf{A}_m \sum_{\beta=1}^{q-m} \mathbf{w}_\beta \mathbf{e}_\beta^T \mathbf{X}_{[1:q-m, 1:q-m]} \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m} \right).$$

Next, use the fact that $\mathbf{e}_\beta^T \mathbf{X}_{[1:q-m, 1:q-m]} = \mathbf{X}_{[\beta, 1:q-m]}$ to obtain

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \sum_{m=1}^{\min(a_1, q-1)} \mathbf{A}_m \sum_{\beta=1}^{q-m} \mathbf{w}_\beta \mathbf{X}_{[\beta, 1:q-m]} \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m} \right).$$

Now use theorem D.2 to give

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \sum_{m=1}^{\min(a_1, q-1)} \sum_{\beta=1}^{q-m} \mathbf{A}_m \mathbf{w}_\beta \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m + \beta - 1) \mathbf{e}_{q-m-\beta+1} \right).$$

Now use fact D.5 to reorder the data access and obtain

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{m=1}^{\min(b_1, q-1)} (\mathbf{b}_m \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, m) \mathbf{e}_{q-m}) - \sum_{j_1=1}^{q-1} \sum_{j_2=1}^{\min(a_1, j_1)} \mathbf{A}_{j_2} \mathbf{w}_{j_1-j_2+1} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, j_1) \mathbf{e}_{q-j_1} \right)$$

which, with a change of variables of m to r , j_1 to r and j_2 to m , can be rewritten as

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{r=1}^{\min(b_1, q-1)} (\mathbf{b}_r \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, r) \mathbf{e}_{q-r}) - \sum_{r=1}^{q-1} \sum_{m=1}^{\min(a_1, r)} \mathbf{A}_m \mathbf{w}_{r-m+1} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, r) \mathbf{e}_{q-r} \right).$$

Factor out the term $\mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, r) \mathbf{e}_{q-r}$ to give

$$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1} \left(\sum_{r=1}^{\min(b_1, q-1)} \mathbf{b}_r - \sum_{r=1}^{q-1} \sum_{m=1}^{\min(a_1, r)} \mathbf{A}_m \mathbf{w}_{r-m+1} \right) \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, r) \mathbf{e}_{q-r}.$$

Now identify each r term from 1 to $q - 1$ and, for each term, use (3.21) to give

$$\tilde{\mathbf{v}}_q = \sum_{r=1}^{q-1} \mathbf{w}_{r+1} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, r) \mathbf{e}_{q-r} = \sum_{r=2}^q \mathbf{w}_r \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(q, r-1) \mathbf{e}_{q-r+1} = \sum_{r=2}^q \mathbf{w}_r \mathbf{X}_{[r,q]}.$$

Finally, note that $\mathbf{X}_{[1,q]} = 0$ and obtain

$$\tilde{\mathbf{v}}_q = \mathbf{W}_q \mathbf{X}_{[1:q,q]}.$$

Therefore, $\tilde{\mathbf{V}}_q = \mathbf{W}_q \mathbf{X}_{[1:q,1:q]}$ and so $\text{span}(\tilde{\mathbf{V}}_q) = \text{span}(\mathbf{W}_q)$. \square

Theorem D.4 (Main result) The space \mathbf{V}_q in definition D.4 matches moments.

Proof: This proof follows trivially from the well known fact that the AWE vector space \mathbf{W}_q matches moments, along with fact D.7 and theorem D.3.

Remark: The proof of this theorem hinges on the fact that $\text{span}(\mathbf{V}_n) = \text{span}(\mathbf{W}_n)$ for all $1 \leq n \leq q$. This does not (and should not) say that $\mathbf{v}_n = \mathbf{w}_n$ for any $2 \leq n \leq q$. In fact, this is why the numerical properties of WCAWE is superior to AWE: the n th vector can be constructed to contain essentially only new information. Finally, again recall that the vectors in definition D.4 are the same as the equations given in (5.7) and (5.8). \square

APPENDIX E

CHOOSING \mathbf{U} IN WCAWE TO PRODUCE ARNOLDI VECTORS

Before showing how to choose \mathbf{U} to produce the Arnoldi vectors it is necessary to give the following definition.

Definition E.1 Let j_1 and j_2 be integers that satisfy $1 \leq j_1 \leq ((N + 1)(c_1 - 1) + 1)$ and $j_2 = j_1 + N - 1$. Then $\mathbf{E}(j_1 : j_2)$ is the $c_1(N + 1) \times N$ matrix that has exactly N nonzero entries. All of these nonzero entries are 1, and they are located such that $\mathbf{E}_{[j_1:j_2,1:N]} = \mathbf{I}_{N \times N}$. □

Now the choice for \mathbf{U} which will produce the Arnoldi vectors can be given:

Algorithm E.1 (Choosing \mathbf{U} in WCAWE to produce Arnoldi vectors)

After $\tilde{\mathbf{v}}_n$ is generated and before $\tilde{\mathbf{v}}_{n+1}$ is generated

let $\alpha = 1$

$$\mathbf{U}_{[1,n]} = \mathbf{v}_1^H \tilde{\mathbf{v}}_n$$

$$\tilde{\mathbf{v}}_n = \tilde{\mathbf{v}}_n - \mathbf{U}_{[1,n]} \mathbf{v}_1$$

let $\alpha = 2$

$$\mathbf{U}_{[2,n]} = \mathbf{v}_2^H \tilde{\mathbf{v}}_n + \left(\mathbf{v}_1 \mathbf{P}_{\mathbf{U}_2}(3, 2) \right)^H \mathbf{v}_{n-1} + \left(-\mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(3, 1) \mathbf{e}_2 \right)^H \mathbf{U}_{[1,n]} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}_1}(2, 1) \mathbf{e}_1$$

$$\tilde{\mathbf{v}}_n = \tilde{\mathbf{v}}_n - \mathbf{U}_{[2,n]} \mathbf{v}_2$$

for $\alpha = 3, 4, \dots, n-1$ do

$$\begin{aligned} \mathbf{U}_{[\alpha,n]} = & \mathbf{v}_\alpha^H \tilde{\mathbf{v}}_n + \left(\mathbf{V}_{\alpha-1} \mathbf{P}_{\mathbf{U}2}(\alpha+1, 2) \mathbf{e}_{\alpha-1} \right)^H \left(\mathbf{v}_{n-1} - \sum_{j=2}^{\alpha-1} \mathbf{U}_{[j,n]} \mathbf{V}_{j-1} \mathbf{P}_{\mathbf{U}2}(j+1, 2) \mathbf{e}_{j-1} \right) \\ & + \sum_{m=2}^{\min(\alpha-1, c_1-1)} \left(\left(\mathbf{V}_{\alpha-m} \mathbf{P}_{\mathbf{U}2}(\alpha+1, m+1) \mathbf{e}_{\alpha-m} \right)^H \right. \\ & \left. \left(\mathbf{V}_{n-m} \mathbf{P}_{\mathbf{U}2}(n, m) \mathbf{e}_{n-m} - \sum_{j=m+1}^{\alpha-1} \mathbf{U}_{[j,n]} \mathbf{V}_{j-m} \mathbf{P}_{\mathbf{U}2}(j+1, m+1) \mathbf{e}_{j-m} \right) \right) \\ & + \left(-\mathbf{e}_1^T \mathbf{P}_{\mathbf{U}1}(\alpha+1, 1) \mathbf{e}_\alpha \right)^H \left(\sum_{j=1}^{\alpha-1} \mathbf{U}_{[j,n]} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}1}(j+1, 1) \mathbf{e}_j \right) \\ & + \sum_{m=2}^{\min(\alpha-1, c_1)} \left(-\mathbf{e}_1^T \mathbf{P}_{\mathbf{U}1}(\alpha+1, m) \mathbf{e}_{\alpha-m+1} \right)^H \\ & \left(\mathbf{e}_1^T \mathbf{P}_{\mathbf{U}1}(n, m-1) \mathbf{e}_{n-m+1} + \sum_{j=m}^{\alpha-1} \mathbf{U}_{[j,n]} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}1}(j+1, m) \mathbf{e}_{j-m+1} \right) \end{aligned}$$

$$\tilde{\mathbf{v}}_n = \tilde{\mathbf{v}}_n - \mathbf{U}_{[\alpha,n]} \mathbf{v}_\alpha$$

endfor

$$\begin{aligned} \mathbf{U}_{[n,n]} = & \left\| \mathbf{E}(1 : N) \tilde{\mathbf{v}}_n + \mathbf{E}(N+2 : 2N+1) \left(\mathbf{v}_{n-1} - \sum_{j=2}^{n-1} \mathbf{U}_{[j,n]} \mathbf{V}_{j-1} \mathbf{P}_{\mathbf{U}2}(j+1, 2) \mathbf{e}_{j-1} \right) \right. \\ & + \sum_{m=2}^{\min(n-1, c_1-1)} \mathbf{E}(m(N+1)+1 : m(N+1)+N) \\ & \left(\mathbf{V}_{n-m} \mathbf{P}_{\mathbf{U}2}(n, m) \mathbf{e}_{n-m} - \sum_{j=m+1}^{n-1} \mathbf{U}_{[j,n]} \mathbf{V}_{j-m} \mathbf{P}_{\mathbf{U}2}(j+1, m+1) \mathbf{e}_{j-m} \right) \\ & + \mathbf{e}_{N+1} \sum_{j=1}^{n-1} \mathbf{U}_{[j,n]} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}1}(j+1, 1) \mathbf{e}_j \\ & \left. + \sum_{m=2}^{\min(n, c_1)} \mathbf{e}_{m(N+1)} \left(\mathbf{e}_1^T \mathbf{P}_{\mathbf{U}1}(n, m-1) \mathbf{e}_{n-m+1} + \sum_{j=m}^{n-1} \mathbf{U}_{[j,n]} \mathbf{e}_1^T \mathbf{P}_{\mathbf{U}1}(j+1, m) \mathbf{e}_{j-m+1} \right) \right\| \end{aligned}$$

$$\mathbf{v}_n = \tilde{\mathbf{v}}_n \mathbf{U}_{[n,n]}^{-1} \quad \square$$

Now the \mathbf{V}_q vectors, created from this choice of \mathbf{U} , can be used to produce the Arnoldi vectors for the expanded, linearized system given in (3.9). If \mathbf{z}_n is the n th

Arnoldi vector, then

$$\mathbf{z}_n = \mathbf{E}(1 : N)\mathbf{v}_n + \sum_{m=2}^{\min(n, c_1)} \mathbf{E}((m-1)(N+1)+1:(m-1)(N+1)+N)\mathbf{V}_{n-m+1}\mathbf{P}_{\mathbf{U}_2}(n+1, m)\mathbf{e}_{n-m+1} - \sum_{m=1}^{\min(n, c_1)} \mathbf{e}_{m(N+1)}\mathbf{e}_1^T\mathbf{P}_{\mathbf{U}_1}(n+1, m)\mathbf{e}_{n-m+1}. \quad (\text{E.1})$$

BIBLIOGRAPHY

- [1] C. de Villemagne and R. E. Skelton, "Model reductions using a projection formulation," *Int. J. Control*, vol. 46, no. 6, pp. 2141-2169, 1987.
- [2] A. C. Cangellaris and L. Zhao, "Model order reduction techniques for electromagnetic macromodelling based on finite methods," *Int. J. Numer. Model.*, vol. 13, no. 2-3, pp. 181-197, Mar.-Jun. 2000.
- [3] J. Rubio, J. Arroyo and J. Zapata, "SFELP—An Efficient Methodology for Microwave Circuit Analysis," *IEEE Trans. Microwave Theory Tech.*, vol. 49, no. 3, pp. 509-516, Mar. 2001.
- [4] G. J. Burke, E. K. Miller, S. Chakrabarti and K. Demarest, "Using model-based parameter estimation to increase the efficiency of computing electromagnetic transfer functions," *IEEE Trans. Magn.*, vol. 25, no. 4, pp. 2807-2809, Jul. 1989.
- [5] D. H. Werner and R. J. Allard, "The simultaneous interpolation of antenna radiation patterns in both the spatial and frequency domains using model-based parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 48, no. 3, pp. 383-392, Mar. 2000.
- [6] V. Druskin and L. Knizhnerman, "Spectral approach to solving three-dimensional Maxwell's diffusion equations in the time and frequency domains," *Radio Sci.*, vol. 29, no. 4, pp. 937-953, Jul.-Aug. 1994.
- [7] M. R. Zunoubi, K. C. Donepudi, J. M. Jin and W. C. Chew, "Efficient time-domain and frequency-domain finite-element solution of Maxwell's equations using spectral Lanczos decomposition method," *IEEE Trans. Microwave Theory Tech.*, vol. 46, no. 8, pp. 1141-1149, Aug. 1998.
- [8] C. Lanczos, "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators," *J. Res. Nat. Bur. Stand.*, vol. 45, pp. 255-282, 1950.
- [9] K. Gallivan, E. Grimme and P. Van Dooren, "Asymptotic waveform evaluation via a Lanczos method," *Appl. Math. Lett.*, vol. 7, no. 5, pp. 75-80, 1994.

- [10] P. Feldmann and R. W. Freund, "Efficient linear circuit analysis by Padé approximation via the Lanczos Process," *IEEE Trans. Comput.-Aided Des. Integrated Circuits and Syst.*, vol. 14, no. 5, pp. 639-649, May 1995.
- [11] A. C. Cangellaris and L. Zhao, "Passive reduced-order modeling of electromagnetic systems," *Comput. Methods Appl. Mech. Engrg.*, vol. 169, no. 3-4, pp. 345-358, Feb. 1999.
- [12] R. D. Slone and R. Lee, "Applying Padé via Lanczos to the finite element method for electromagnetic radiation problems," *Radio Sci.*, vol. 35, no. 2, pp. 331-340, Mar.-Apr. 2000.
- [13] J. E. Bracken, D.-K. Sun and Z. Cendes, "Characterization of electromagnetic devices via reduced-order models," *Comput. Methods Appl. Mech. Engrg.*, vol. 169, no. 3-4, pp. 311-330, Feb. 1999.
- [14] D.-K. Sun, J.-F. Lee and Z. Cendes, "ALPS – A New Fast Frequency-Sweep Procedure for Microwave Devices," *IEEE Trans. Microwave Theory Tech.*, vol. 49, no. 2, pp. 398-402, Feb. 2001.
- [15] W. E. Arnoldi, "The principle of minimized iterations in the solution of the matrix eigenvalue problem," *Quart. Appl. Math.*, vol. 9, pp. 17-29, 1951.
- [16] J. Cullum, A. Ruehli and T. Zhang, "A Method for Reduced-Order Modeling and Simulation of Large Interconnect Circuits and its Application to PEEC Models with Retardation," *IEEE Trans. Circuits Syst.-II*, vol. 47, no. 4, pp. 261-273, Apr. 2000.
- [17] L. T. Pillage and R. A. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Trans. Comput.-Aided Des. Integrated Circuits and Syst.*, vol. 9, no. 4, pp. 352-366, Apr. 1990.
- [18] R. Sanaie, E. Chiprout, M. S. Nakhla and Q. J. Zhang, "A fast method for frequency and time domain simulation of high speed VLSI interconnects," *IEEE Trans. Microwave Theory Tech.*, vol. 42, no. 12, pp. 2562-2571, Dec. 1994.
- [19] E. Chiprout and M. S. Nakhla, "Analysis of interconnect networks using complex frequency hopping (CFH)," *IEEE Trans. Comput.-Aided Des. Integrated Circuits and Syst.*, vol. 14, no. 2, pp. 186-200, Feb. 1995.
- [20] C. J. Reddy, M. D. Deshpande, C. R. Cockrell and F. B. Beck, "Fast RCS computation over a frequency band using method of moments in conjunction with asymptotic waveform evaluation technique," *IEEE Trans. Antennas Propagat.*, vol. 46, no. 8, pp. 1229-1233, Aug. 1998.

- [21] D. Jiao, X. Y. Zhu and J. M. Jin, "Fast and accurate frequency-sweep using asymptotic waveform evaluation and the combined-field integral equation," *Radio Sci.*, vol. 34, no. 5, pp. 1055-1063, Sep.-Oct. 1999.
- [22] Y. E. Erdemli, J. Gong, C. J. Reddy and J. L. Volakis, "Fast RCS pattern fill using AWE technique," *IEEE Trans. Antennas Propagat.*, vol. 46, no. 11, pp. 1752-1753, Nov. 1998.
- [23] M. Li, Q.-J. Zhang and M. Nakhla, "Finite Difference solution of EM fields by asymptotic waveform techniques," *IEE Proc., H, Microw. Antennas Propag.*, vol. 143, no. 6, pp. 512-520, Dec. 1996.
- [24] M. A. Kolbehdari and M. S. Nakhla, "Reduced-order method for dielectric resonators using FEM and CFH," *COMPEL-The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 18, no. 1, pp. 84-94, 1999.
- [25] J. P. Zhang and J. M. Jin, "Preliminary Study of AWE Method for FEM Analysis of Scattering Problems," *Microw. Opt. Technol. Lett.*, vol. 17, no. 1, pp. 7-12, Jan. 1998.
- [26] X. M. Zhang and J.-F. Lee, "Application of the AWE method with the 3-D TVFEM to model spectral responses of passive microwave components," *IEEE Trans. Microwave Theory Tech.*, vol. 46, no. 11, pp. 1735-1741, Nov. 1998.
- [27] R. W. Freund, "Reduced-order modeling techniques based on Krylov subspaces and their use in circuit simulation," *Numer. Anal. Manuscr.*, 98-3-02, Bell Lab., Murray Hill, N.J., Feb. 1998.
- [28] J.-F. Lee and D.-K. Sun, "Fast Spectral Response Calculations for Passive MMICs using a Novel Galerkin Asymptotic Wave Evaluation (GAWE) Method," working papers, 1999.
- [29] R. D. Slone, R. Lee and J.-F. Lee, "Multipoint Galerkin Asymptotic Waveform Evaluation for Model Order Reduction of Frequency Domain FEM Electromagnetic Radiation Problems," *IEEE Trans. Antennas Propagat.*, vol. 49, no. 10, pp. 1504-1513, Oct. 2001.
- [30] L. Zhao and A. C. Cangellaris, "Reduced-Order Modeling of Electromagnetic Field Interactions in Unbounded Domains Truncated by Perfectly Matched Layers," *Microw. Opt. Technol. Lett.*, vol. 17, no. 1, pp. 62-66, Jan. 1998.
- [31] M. Kuzuoglu and R. Mittra, "Finite element solution of electromagnetic problems over a wide frequency range via the Padé approximation," *Comput. Methods Appl. Mech. Engrg.*, vol. 169, no. 3-4, pp. 263-277, Feb. 1999.

- [32] A. Bayliss and E. Turkel, "Radiation boundary conditions for wave-like equations," *Commun. Pure and Appl. Math.*, vol. 33, pp. 707-725, 1980.
- [33] D.-K. Sun, J.-F. Lee and Z. Cendes, "Construction of nearly orthogonal Nedelec bases for rapid convergence with multilevel preconditioned solvers," *SIAM J. Sci. Comput.*, vol. 23, no. 4, pp. 1053-1076, 2001.
- [34] A. Odabasioglu, M. Celik, and L. T. Pileggi, "PRIMA: passive reduced-order interconnect macromodeling algorithm," *IEEE Trans. Comput.-Aided Des. Integr. Circ. Syst.*, vol. 17, no. 8, pp. 645-654, Aug. 1998.
- [35] L. M. Silveira, M. Kamon, I. Elfadel and J. White, "A coordinate-transformed Arnoldi algorithm for generating guaranteed stable reduced-order models of RLC circuits," *Comput. Methods Appl. Mech. Engrg.*, vol. 169, no. 3-4, pp. 377-389, Feb. 1999.
- [36] J. E. Bracken, V. Raghavan and R. A. Rohrer, "Interconnect Simulation with Asymptotic Waveform Evaluation (AWE)," *IEEE Trans. Circ. Syst.-I*, vol. 39, no. 11, pp. 869-878, Nov. 1992.
- [37] R. D. Slone, J. F. Lee, and R. Lee, "A comparison of some model order reduction techniques" accepted for publication in *Electromagnetics*, 2002.
- [38] R. D. Slone, "Removing the Frequency Restriction on Padé Via Lanczos With an Error Bound," *IEEE Int. Symp. Electromagn. Compat. Proc.*, vol. 1, pp. 202-207, Aug. 1998.
- [39] Z. Bai, R. D. Slone, W. T. Smith and Q. Ye, "Error Bound for Reduced System Model by Padé Approximation Via the Lanczos Process," *IEEE Trans. Comput.-Aided Des. Integrated Circuits Syst.*, vol. 18, pp. 133-141, Feb. 1999.
- [40] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C*. Cambridge University Press, 1992.
- [41] T.-J. Su and R. R. Craig, Jr., "Krylov Vector Methods for Model Reduction and Control of Flexible Structures," *Control and Dynamic Systems*, vol. 54, pp. 449-481, 1992.