SANDIA REPORT

SAND2009-6862
Unlimited Release
Printed October 2009

Final Report LDRD Project 105816: Model Reduction of Large Dynamic Systems with Localized Nonlinearities

Daniel J. Segalman, Clark. R. Dohrmann, Richard L. Lehoucq, and Ulrich L. Hetmaniuk

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2009-6862
Unlimited Release
Printed October 2009

Final Report LDRD Project 105816: Model Reduction of Large Dynamic Systems with Localized Nonlinearities

D. J. Segalman
Computational Structural Mechanics & Applications Organization

C. R. Dohrmann
Advanced Structural Dynamics Organization

R. B. Lehoucq
Applied Mathematics & Applications Organization

Sandia National Laboratories, PO Box 5800, Albuquerque, NM 87185-0557

&

U. L. Hetmaniuk
University of Washington, Box 352420, Seattle, WA 98195-2420

Abstract

Advanced computing hardware and software written to exploit massively parallel architectures greatly facilitate the computation of extremely large problems. On the other hand, these tools, though enabling higher fidelity models, have often resulted in much longer run-times and turn-around-times in providing answers to engineering problems. The impediments include smaller elements and consequently smaller time steps, much larger systems of equations to solve, and the inclusion of nonlinearities that had been ignored in days when lower fidelity models were the norm. The research effort reported focuses on the accelerating the analysis process for structural dynamics though combinations of model reduction and mitigation of some factors that lead to over-meshing.

Acknowledgment

Though there are only four authors of this report, many people have contributed the work presented here. These include people who have come in an out of various components of this work over the last three years. Specific people are mentioned within the chapters describing each of the three research directions. Additionally, the authors wish to express their gratitude to their many coworkers who maintain an environment supportive of explorations such as this.

Contents

List of Figures

List of Tables

Chapter 1

Overview and Introduction

Advanced computing hardware and software written to exploit massively parallel architectures greatly facilitate the computation of extremely large problems. On the other hand, these tools, though enabling higher fidelity models, have often resulted in much longer run-times and turn-around-times in providing answers to engineering problems. The impediments include smaller elements and consequently smaller time steps, much larger systems of equations to solve, and the inclusion of nonlinearities that had been ignored in days when lower fidelity models were the norm.

In this section, we discuss each of these classes of difficulties and attempt to foreshadow the research reported further below to address those difficulties.

Incommensurate Meshes One of the major impediments of creating finite element meshes for large structures is the need to have commensurate meshes at the interfaces; nodes from each side must coincide with nodes from the other (See Figure 1.1). This not only requires much forethought in the meshing of each unit, but it places complex constraints on the meshing process.

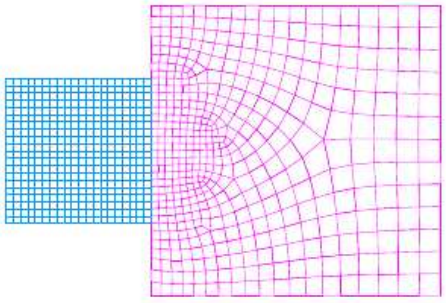

Figure 1.1. Creating conformal meshes integrating two features is relatively easy, but connection of multiple features in three dimensions can be unreasonably tedious and may delay initiation of calculations by months.

The same problem arises in the context of component mode synthesis (CMS). In this approach, one performs a model reduction on each substructure of the system, introducing modal degrees of freedom, retaining nodes on the boundaries, and resolving away internal nodes. This approach has many advantages, including the common practice of having subsystems manufactured by different contractors and having each manufacturer provide a CMS model for his own substructure. It is the duty of the system integrator to assemble the mathematical subsystems into a model for the full system. In order to avoid large loss of accuracy, the nodes of the subsystem models most coincide.

Of course, one can connect dissimilar meshes using master-slave constraints, but such an approach may result in significant loss of accuracy near interfaces - undermining the purpose of using fine meshes for fidelity. Another, competing approach of joining dissimilar meshes is that of *mortar methods.* These methods do preserve accuracy, but their practical implementation for 3-D applications requires significant coding effort for every element type.

An accurate and efficient method for connecting dissimilar meshes is presented in Chapter 2. This a novel, efficient, and highly usable method of joining disparate meshes is presented. The ramifications will be great both in terms of model reduction, but also in terms of efficiency in effective mesh generation.

Scalable Component Mode Synthesis As we find ourselves in a situation where it is possible to employ fine enough meshes to capture structural features with great fidelity, we also find ourselves with numerical problems having far more degrees of freedom than are actually necessary to capture strain fields and other features that define problem mechanics.

The resulting numerical problems are sufficiently challenging that solving them cannot be routine. What should be routine calculations can become *capability tests* of large massively parallel computers.

A traditional method of model reduction that does have the capability to retain model fidelity (at least of linear problems) is that of component mode synthesis, discussed above. One of the limitations of current use of CMS was also discussed above. Another limitation is that if one employs CMS to create more and more subregions of model components to achieve better and better geometric fidelity, the problem size becomes dominated by the nodal degrees of freedom at the interfaces of the subregions. The number of degrees of freedom and their coupling become major impediments to computational scaling.

A novel approach to mitigating the computational limitations of Component Mode Synthesis is developed and presented in Chapter 3.

Spatially Distributed Nonlinearities Very fine meshing of finite element models is generally expected to converge to an accurate result so long as the structure consists of a monolithic piece of metal. Predictive modeling of the dynamics of real structures - accounting

for such things as bolted connections, compression fits, and other joints - must account for the nonlinear frictional mechanics at interfaces. Those nonlinear mechanics result in hysteresis behavior such as shown in Figure 1.2.

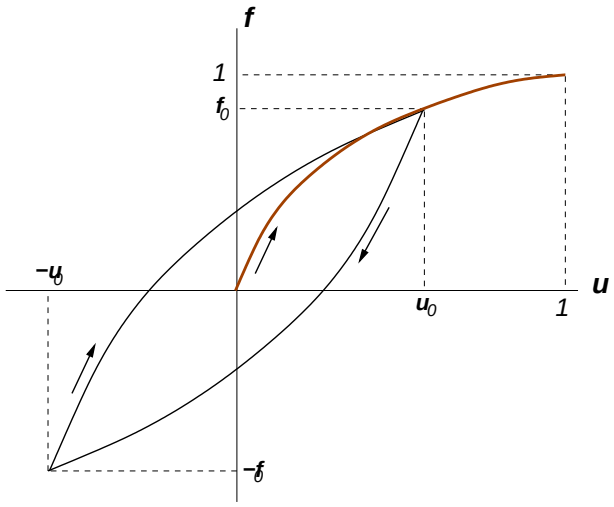

Figure 1.2. Frictional interfaces between components of the structure generate hysteresis loops such as this. The sharp discontinuity in tangent stiffness with load reversals causes significant numerical difficulty.

The discontinuous changes in tangent stiffness with load reversal introduces creates a sharp nonlinearity in the equations of motion. Numerical solution of the resulting nonlinear algebraic equations requires Newton iteration using very small time steps. The large number of equations to be solved at each iteration and the very small time steps that must be employed make simulation of the dynamics of real structures prohibitively difficult.

A model reduction technique that not only reduces the number of degrees of freedom, but also enables the use of much larger time steps is developed in Chapter 4. This reduced order modeling strategy has the added advantage of yielding cleaner and more helpful results than are obtained with when the original problem is solved directly.

Combination of Methods Each of these three approaches can be employed with the others to make possible important calculations related to the weapons-related mission. For instance, when externally applied loads are known in only a statistical manner, credibility of calculations on the accelerations that will be seen by components distributed through the structure will require many case calculations. This can take place only if the kinds of calculations that now spend weeks in queue and days on a supercomputer can be run orders of magnitude faster. It is the intent of the mutually complementary strategies whose investigation is presented in this report to make such rapid simulations possible.

There is more discussion of the integration of these three approaches in the Conclusions portion of this report.

Chapter 2

Stabilized Tied Contact

Clark R. Dohrmann

2.1 Chapter Abstract

In this study we present a simple method for improving the accuracy of the classic method of tied contact for connecting dissimilar finite element meshes. The method augments the standard constraint equations of tied contact with a set of discrete springs located on mesh interfaces. Although the approach can be viewed as a penalty method, there is no need to select a penalty parameter. Moreover, the method avoids the calculation of complicated surface integrals typically required by mortar or Nitche type methods. Numerical results show that the method improves both the local and global rates of convergence of tied contact while not incurring significant computational cost.

2.2 Introduction

One of the first, and perhaps simplest, methods for connecting dissimilar finite element meshes is to constrain each node on a designated slave side of a mesh interface directly to the nearest point on a designated master side of the interface. This classic method of connecting meshes is known by a variety of names including the node-to-surface approach, multipoint constraint (MPC) approach, and permanent glued contact [32], but we will use the designation tied contact [13] here. Because of its simplicity, tied contact is available in a variety of commercial and research finite element codes. One of the attractive features of tied contact is that it avoids the need to introduce Lagrange multipliers to enforce continuity at the interface. Because of the local nature of the tied contact constraints, there is also no significant loss of matrix sparsity when dependent (slave) degrees of freedom (dofs) are eliminated.

Although relatively simple to understand and implement, tied contact also has its shortcomings. It is known for elasticity problems that stresses near mesh interfaces may not

converge to exact solutions with mesh refinement when tied contact is used. Consequently, global rates of convergence can be smaller than those for similar meshes without interfaces. The purpose of this study is to investigate a simple fix to improve the accuracy of tied contact while retaining much of its appealing simplicity.

Modeling complicated structures with a single finite element mesh can be a very difficult if not impossible task. Even when a single person develops a system model, it may be convenient to mesh different parts independently without the restriction of having conforming interfaces. By conforming we mean that the nodal locations and interpolation functions on both sides of an interface are the same. For these reasons and others, many references on connecting dissimilar finite element meshes can be found in the literature. A recent article with a discussion of several different methods is given in [35].

Significant efforts have been made over the years to develop mesh connecting methods that pass the engineering patch test. That is, when two meshes are connected by a candidate method, the finite element solution is exact for boundary conditions corresponding to a constant state a stress. Examples of two different methods with an emphasis on satisfying the patch test can be found in [45] and [36], but there are others as well. We comment at the outset that the simplicity of the present method comes with what some may perceive a cost. Namely, it does not pass the patch test. Nevertheless, the method converges with mesh refinement and satisfies the patch test in the asymptotic limit. A companion theoretical study of the method is given in [14].

An earlier method which retains the standard tied contact constraints, like the present method, while also satisfying first-order patch tests for both planar and curved interfaces is given in [15]. The approach is somewhat complicated and requires element matrices on the slave sides of interfaces to be modified. We consider the present method to be simpler, and have also observed better rates of convergence for meshes of quadratic finite elements. In the interest of brevity, we will limit the discussion of other mesh connecting approaches to mortar [8] and Nitche [6] type methods in the next two paragraphs. A discussion of other methods can be found in [35] and the references therein.

As with tied contact, the constraint equations for mortar methods allow one to solve for slave dofs in terms of master dofs, and then eliminate these slave dofs; otherwise, one needs to solve a saddle-point system of equations which is indefinite. For the Lagrange multiplier bases of earlier mortar methods, each slave dof could depend on every dof on the master side of the interface. Thus, elimination of the slave dofs could lead to complete loss of sparsity in the stiffness matrix for the master dofs. More recent dual mortar methods [44, 38] lead to constraint equations more like those of tied contact, but obtaining these constraint equations may require the calculation of complicated interface integrals, especially in three dimensions. Some care is also needed in choosing the Lagrange multiplier basis for mortar methods to ensure that an inf-sup condition for convergence is satisfied.

In contrast to tied contact and mortar methods, Nitche type methods do no involve any constraint equations. Rather, a discrete bilinear form is employed which is closely related to a primal formulation of a discontinuous Galerkin method (cf. Method IP in Table 3.2 of

[1]). Like mortar methods, however, they require the calculation of surface integrals for 3D problems. The complexity of calculating surface integrals for mortar and Nitche methods is recognized as a practical concern [29, 18, 21, 19]. We note also that Nitche methods require a parameter that must be chosen with some care. If the parameter is too small, then the method is unstable and will not converge. If the parameter is too large, then mesh interfaces can be overly stiff and lead to significant loss of accuracy.

The basic idea of the present method is to augment the constraint equations of tied contact with a set of discrete springs located on mesh interfaces. Thus, following elimination of all slave dofs, the stiffness matrix equals that for classic tied contact plus a *stabilization* matrix associated with the springs. We note that a fully symmetric method with no distinction between master and slave sides of interfaces is possible by replacing all constraint equations with springs, but we do not investigate this variant here.

The organization of the paper is summarized as follows. First, we review the classic method of tied contact for connecting meshes in §2.3, and then present its stabilized form in §2.4. We then present a variety of numerical examples in §2.5 demonstrating both the improved accuracy of stabilized tied contact and the small affect of the stabilization on iterative solver performance. Some closing remarks are made in §2.6.

2.3 Classic Tied Contact

After imposing essential and natural boundary conditions, the finite element equations of equilibrium for linear elastostatics can be expressed as

$$\underbrace{\begin{bmatrix} K_{ss} & K_{sm} & K_{sr} \\ K_{ms} & K_{mm} & K_{mr} \\ K_{rs} & K_{rm} & K_{rr} \end{bmatrix}}_{K} \underbrace{\begin{bmatrix} u_s \\ u_m \\ u_r \end{bmatrix}}_{u} = \underbrace{\begin{bmatrix} f_s \\ f_m \\ f_r \end{bmatrix}}_{f}, \tag{2.1}$$

where K is the assembled stiffness matrix, u is the displacement vector, and f is the force vector. The subscript s is for the *slave* dofs to be eliminated by the tied contact constraints. Likewise, the subscript m is for the *master* dofs involved in the tied contact constraints. Finally, the subscript r refers to the *remaining* dofs. After the removal of any redundant constraints, the tied contact constraint equations can be expressed concisely as

$$u_s = C u_m, \tag{2.2}$$

where C is the constraint matrix for the slave dofs. We thus have

$$u = \underbrace{\begin{bmatrix} C & 0 \\ I & 0 \\ 0 & I \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} u_m \\ u_r \end{bmatrix}}_{u_i}, \tag{2.3}$$

where I is an identity matrix of appropriate dimension and the subscript i refers to *independent* dofs. Substitution of (2.3) into (2.1) and premultiplication by T^T (the transpose of T) gives the standard reduced equilibrium equations

$$K_i u_i = f_i, \tag{2.4}$$

where

$$K_i = T^T K T \quad \text{and} \quad f_i = T^T f.$$

Once u_i is obtained by solving (2.4), u_s can then be recovered from (2.2).

2.3.1 Tied Contact Constraints

Each row of the constraint matrix C in (2.2) corresponds to a specific degree of freedom of a specific node on the slave side of a mesh interface. If a mesh interface is curved or the master and slave sides of an interface are not coincident for some other reason, then each slave node may not be initially located on the master side of the interface. In such cases it is common practice to initially move, in a stress-free state, each slave node onto the master surface prior to applying any constraint equations or loads. Doing so ensures that there is no strain energy when the model is deformed into the shape of a rigid body mode for floating structures without any essential boundary conditions. We will henceforth assume that all slave nodes are thus initially positioned on the master surface.

For purposes of discussion, we will consider a 3D elasticity problem in which each node has 3 degrees of freedom before boundary conditions are applied. By assumption, each slave node is coincident with a point on the face of one or more elements on the master side of the interface. Let (ξ_1, ξ_2) denote the element coordinates for one such face. For isoparametric elements we have

$$x_j(\xi_1, \xi_2) = \sum_{k \in \mathscr{F}} x_{jk} \phi_k(\xi_1, \xi_2), \tag{2.5}$$

where $x_j(\xi_1, \xi_2)$ is the j-coordinate of the point on the face with element coordinates (ξ_1, ξ_2), and $\mathscr{F}$ is the set of node numbers for the face. Further, x_{jk} and ϕ_k are the j-coordinate and shape function, respectively, of node k of the element face. Letting x_{js} denote the j-coordinate of slave node s, we can determine the element coordinates (ξ_{1s}, ξ_{2s}) associated with node s by minimizing the squared distance function

$$d^2 = \sum_{j=1}^{3} (x_j - x_{js})^2$$

using Newton's method. Indeed, this approach can also be used to identify the locations to move slave nodes so that they are initially on a master surface. The row of the constraint equations in (2.2) corresponding to the displacement in direction j of slave node s then reads

$$u_{js} = \sum_{k \in \mathscr{F}} \phi_k(\xi_{1s}, \xi_{2s}) u_{jk},$$

where u_{jk} is the displacement in direction j of node k.

2.3.2 Mortar Method Connection

Classic tied contact can be understood as a mortar method for a specific Lagrange multiplier basis. Specifically, this basis consists of Dirac delta functions centered at the slave nodes. To illustrate, let v_m and v_s denote scalar fields on the master and slave sides of an interface Γ. The mortar constraint equation associated with shape function λ_m of the Lagrange multiplier basis is given by

$$\int_\Gamma (v_s - v_m)\lambda_m \, dx = 0.$$

By choosing λ_m to be a Dirac delta function at slave node s, we find that $v_s|_{\boldsymbol{x}_s} = v_m|_{\boldsymbol{x}_s}$, where $\boldsymbol{x}_s$ is the position of s. That is, the value of the field v_s on the slave side of the interface is constrained to be the same as the value of v_m on the master side of the interface at the location of s. For mortar methods, the Lagrange multiple basis is chosen so that constant functions can be approximated exactly on the interface. That is, there exist α_m such that $\sum_m \alpha_m \lambda_m = 1$ on Γ. The Lagrange multiplier basis for tied contact does not have this property. Thus, one should not expect tied contact to have the same convergence properties as mortar methods.

2.3.3 Potential Shortcomings

Here we use a simple 2D plane stress example to illustrate some of the potential shortcomings of tied contact. Let u_j and x_j denote the displacement and spatial coordinate, respectively, for direction j of an orthogonal coordinate system. The boundary conditions and exact solution for the example are given by

$$u_1(0,x_2) = 0, \quad u_2(0,0) = 0, \quad \sigma_{11}(2,x_2) = -Ex_2$$

and

$$u_1(x_1,x_2) = -x_1x_2, \quad u_2(x_1,x_2) = (x_1^2 + \nu x_2^2)/2,$$

where E is Young's modulus, ν is the Poisson ratio, and σ_{11} is the axial stress.

In order to measure the accuracy of finite element solutions, we define displacement and energy norms of a vector function $\boldsymbol{u} = (u_1, \ldots, u_d)$ as

$$\|\boldsymbol{u}\|_0 = \left(\sum_{j=1}^{d} \int_\Omega u_j^2 \, dx \right)^{1/2} \quad \text{and} \quad \|\boldsymbol{u}\|_1 = \left(\sum_{j=1}^{d} \int_\Omega \nabla u_j \cdot \nabla u_j \, dx \right)^{1/2}, \tag{2.6}$$

where Ω is the problem domain and d is the spatial dimension. Let $\boldsymbol{u}_e = (u_{e1}, \ldots, u_{ed})$ denote the exact solution to an elasticity problem. For sufficiently smooth exact solutions, we expect from finite element theory [43] that

$$\|\boldsymbol{u} - \boldsymbol{u}_e\|_0 \le Ch^{q+1} \quad \text{and} \quad \|\boldsymbol{u} - \boldsymbol{u}_e\|_1 \le Ch^q, \tag{2.7}$$

where C is a constant, h is the diameter of the largest element in the mesh, and q is the polynomial degree of the finite elements. We call $\|\boldsymbol{u} - \boldsymbol{u}_e\|_0$ and $\|\boldsymbol{u} - \boldsymbol{u}_e\|_1$ the displacement error and energy error, respectively.

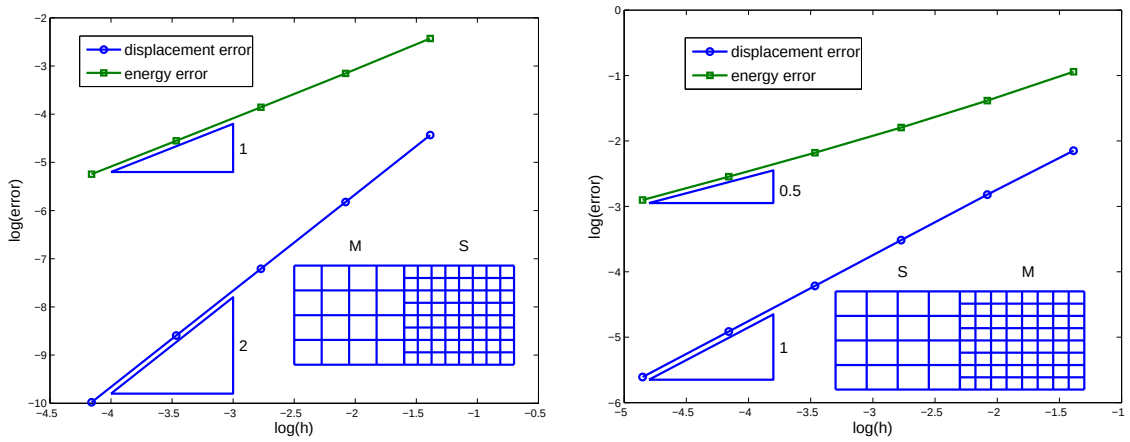

Figure 2.1. Classic tied contact results for 1:2 master-slave mesh transition example in §2.3.3 (left) and 1:2 slave-master mesh transition example in §2.3.3 (right).

1:2 master-slave mesh transition

Consider the two unit square meshes of 4-node quadrilateral (QUAD4) elements shown in Figure 2.1 with a 1:2 transition at the interface. The coarser side of the interface (left) is designated as master and the more refined side (right) as slave. In this case, classic tied contact works perfectly fine because no gaps or overlaps can develop at the interface as the structure deforms. In other words, the displacement is continuous at the interface and we have a conforming finite element method. The displacement and energy errors shown in Figure 2.1-left as a function of element length h are consistent with the theoretical convergence estimates in (2.7).

1:2 slave-master mesh transition

We repeat the previous example, but now the master and slave sides are reversed. That is, the master side of the interface (right) is now twice as refined as the slave side (left). Simply by reversing the choice of master and slave interfaces, we see in Figure 2.1-right that the observed rates of convergence are about half of the theoretical estimates in (2.7). A theory explaining this reduced rate of convergence is provided in [14]. Compared with the first example in §2.3.3, only about one half of the nodes on the right side of the interface are connected to the left side. This comparison clearly demonstrates the benefits in the conventional wisdom of choosing the master side of an interface coarser than the slave side.

The normal stress σ_{22} and shear stress σ_{12} are both zero for the exact solution. The maximum absolute value of these two stresses at element integration points are shown in

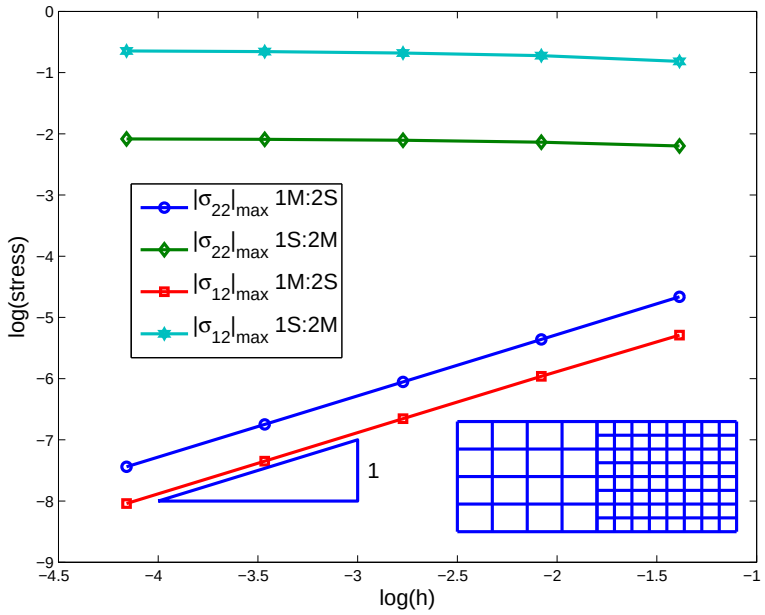

Figure 2.2. Classic tied contact stresses for a 1:2 mesh transition. The notations 1M:2S and 1S:2M indicate that the master side of the interface is on the left and right, respectively. The stresses σ_{22} and σ_{12} are both zero for the exact solution, and the maximum absolute value of the axial stress σ_{11} is 1. When the right side of the interface is chosen as master (1S:2M), the maximum stresses appear to be converging to $|\sigma_{22}|_{\max} \approx 0.13$ and $|\sigma_{12}|_{\max} \approx 0.53$ rather than 0.

Figure 2.2 for the two cases of the master side of the interface on the left and on the right. As expected, these two stresses decrease with mesh refinement when the left side of the interface is chosen as master. In stark contrast, these stresses do not converge to zero when the right side of the interface is chosen as master. Closer examination of the numerical results shows that these stresses are largest in elements directly adjacent to the interface and appear to be converging to $|\sigma_{22}|_{\max} \approx 0.13$ and $|\sigma_{12}|_{\max} \approx 0.53$ rather than 0. We note $|\sigma_{11}|_{\max} = 1$ for the exact solution.

Care should be taken when choosing the master and slave sides of an interface for tied contact. To illustrate a possible dilemma, consider a hypothetical case in which only half of the left side of the interface is coarser than the right side. The question then becomes which side of the interface should be chosen as master? One could break the interface into two parts to address this problem, but that would require additional work on the part of a structural analyst and could be prone to errors. In the next section we present a stabilized form of tied contact which address these and other concerns surrounding the use of classic tied contact.

2.4 Stabilized Tied Contact

Recall for classic tied contact that continuity across a mesh interface is only enforced at the locations of the slave nodes. Thus, except for special cases, continuity will not necessarily hold at all other locations on the interface. The basic idea of stabilized tied contact is to penalize any discontinuities that may occur at a prescribed set of other locations.

As was done in the previous section, we consider a 3D elasticity problem for purposes of discussion. Consider a point p on a master surface, and let (ξ_{1p}, ξ_{2p}) denote the element coordinates of p associated with a face on the master surface containing p. Similar to (2.5), we have

$$x_{jp} = \sum_{k \in \mathscr{F}_p} x_{jk} \phi_k(\xi_{1p}, \xi_{2p}), \tag{2.8}$$

where x_{jp} is the j-coordinate of p and $\mathscr{F}_p$ is the set of nodes for the element face associated with p. Next, let $\hat{p}$ be a point on the slave surface closest to p, and let $(\xi_{1\hat{p}}, \xi_{2\hat{p}})$ denote the element coordinates of $\hat{p}$ for a face on the slave surface containing $\hat{p}$. Thus,

$$x_{j\hat{p}} = \sum_{k \in \mathscr{F}_{\hat{p}}} x_{jk} \phi_k(\xi_{1\hat{p}}, \xi_{2\hat{p}}). \tag{2.9}$$

The present goal is to develop an expression for the strain energy of a spring associated with points p and $\hat{p}$. This development is straightforward when p and $\hat{p}$ are initially coincident, but there is a slight complication when they are not, e.g., for curved interfaces. This complication is related to the need to have zero strain energy when a structure without any essential boundary conditions is deformed into the shape of a rigid body mode with a rotational component. We note that a related issue is also present for mortar methods [38].

Consistent with the isoparametric formulations in (2.8) and (2.9), we have

$$u_p = \sum_{k \in \mathscr{F}_p} u_k \phi_k(\xi_{1p}, \xi_{2p}) \quad \text{and} \quad u_{\hat{p}} = \sum_{k \in \mathscr{F}_{\hat{p}}} u_k \phi_k(\xi_{1\hat{p}}, \xi_{2\hat{p}}),$$

where $u_p = [u_{1p}\, u_{2p}\, u_{3p}]^T$. If p and $\hat{p}$ are initially coincident or the issue regarding rigid body modes is deemed unimportant, then we define the gap

$$g_{p1} := u_p - u_{\hat{p}} = \sum_{k \in \mathscr{F}_p} u_k \phi_k(\xi_{1p}, \xi_{2p}) - \sum_{k \in \mathscr{F}_{\hat{p}}} u_k \phi_k(\xi_{1\hat{p}}, \xi_{2\hat{p}}) = C_{p1}^T u, \tag{2.10}$$

where C_{p1} is a sparse matrix with three rows and n columns, where n is the dimension of the displacement vector u in (2.1). For the more general case, let $\mathscr{E}_p$ denote the set of node numbers for the element E_p which includes the face containing p. Again, for an isoparametric formulation, we have

$$x_j(\eta_1, \eta_2, \eta_3) = \sum_{k \in \mathscr{E}_p} x_{jk} \phi_k(\eta_1, \eta_2, \eta_3),$$

where the shape function ϕ_k for node k depends on three element coordinates rather than just two. We next determine the element coordinates $(\eta_{1\hat{p}}, \eta_{2\hat{p}}, \eta_{3\hat{p}})$ such that

$$x_j(\eta_{1\hat{p}}, \eta_{2\hat{p}}, \eta_{3\hat{p}}) = x_{j\hat{p}}.$$

Again, Newton's method can be used to determine these element coordinates, and the initial guess for the solution could be the element coordinates of point p. Introducing

$$\tilde{u}_{\hat{p}} = \sum_{k \in \mathscr{E}_p} u_k \phi_k(\eta_{1\hat{p}}, \eta_{2\hat{p}}, \eta_{3\hat{p}}),$$

we define the gap

$$g_{p2} := \tilde{u}_{\hat{p}} - u_{\hat{p}} = \sum_{k \in \mathscr{E}_p} u_k \phi_k(\eta_{1\hat{p}}, \eta_{2\hat{p}}, \eta_{3\hat{p}}) - \sum_{k \in \mathscr{F}_{\hat{p}}} u_k \phi_k(\xi_{1\hat{p}}, \xi_{2\hat{p}}) = C_{p2} u. \tag{2.11}$$

If points p and $\hat{p}$ are initially coincident, then $C_{p2} = C_{p1}$. We note that $\tilde{u}_{\hat{p}}$ can be viewed as the displacement of a point in the extension of element E_p that initially has the same position as $\hat{p}$. In the sequel, we will use C_p to denote either C_{p1} or C_{p2}, depending on the choice of gap function g_p.

We next introduce the stabilization term

$$U_s = \sum_{p \in \mathscr{M}} g_p^T K_p g_p = u^T S u,$$

where $\mathscr{M}$ is the set of all points on master surfaces used in the stabilization and K_p is a symmetric matrix. Recipes for both $\mathscr{M}$ and K_p will be given shortly. With reference to (2.10) and (2.11), we see that the stabilization matrix S is given by

$$S = \sum_{p \in \mathscr{M}} C_p^T K_p C_p. \tag{2.12}$$

Replacing K with $K + S$ in the reduced equilibrium equations (2.4), we obtain

$$K_{is} u_i = f_i, \tag{2.13}$$

where

$$K_{is} = T^T (K + S) T.$$

Thus, stabilized tied contact simplifies to the classic form for $S = 0$.

2.4.1 Selection of Point Set $\mathscr{M}$

A suitable choice for the point set $\mathscr{M}$ in the case of lowest-order (linear) finite elements is simply all nodes on master surfaces with nonempty projections onto adjacent slave surfaces; we will denote this point set as $\mathscr{M}_0$. For quadratic elements, it may happen that use of $\mathscr{M}_0$ results in no stabilization at all, i.e., $K_{is} = K_i$. For example, consider the mesh in Figure 2.3 where both the master (left) and slave (right) sides of the mesh are discretized using 8-node quadrilateral elements. For the 2:3 transition at the interface shown, it turns out that $g_p = 0$ for all 9 nodes on the master surface.

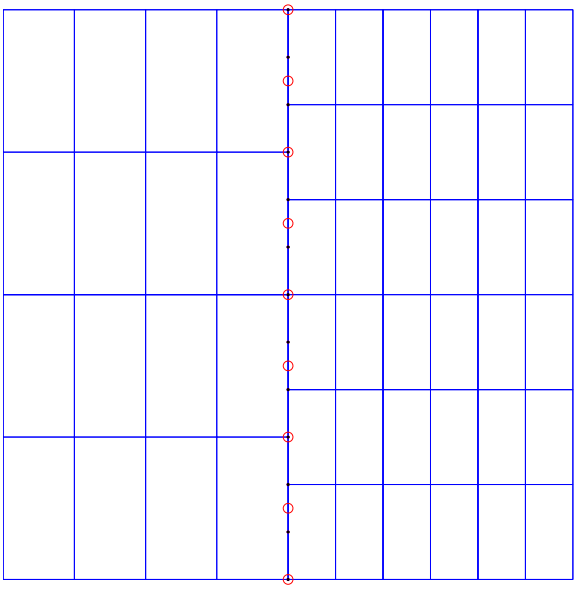

Figure 2.3. 8-node quad element meshes with master and slave nodes appearing as circles and dots, respectively.

Let η_1 denote the parent element coordinate for a one-dimensional element (edge). Similarly, let (η_1, η_2) denote the parent element coordinates for a two-dimensional element (face). We assume the standard practice for edges that $\eta_1 \in [-1,1]$. Similarly, for quadrilateral faces we assume $\eta_1 \in [-1,1]$ and $\eta_2 \in [-1,1]$. Finally, for triangular faces we assume $\eta_1 \in [0,1]$ and $\eta_2 \in [0,1]$, where $\eta_1 + \eta_2 \leq 1$. We say that an edge or face on the master side of an interface is active if all its nodes are included in $\mathscr{M}_0$.

For quadratic elements in 2D, we augment $\mathscr{M}_0$ with points at $\eta_1 = -1/2$ and $\eta_1 = 1/2$ for each active edge on the master side of an interface. Thus, two extra points are added to $\mathscr{M}_0$ for each active edge. For quadratic elements in 3D, the element coordinates of the additional points for each active face on the master side of an interface are shown in Table 2.1.

2.4.2 Selection of K_p

With reference to Figure 2.4, consider a one-dimensional model of a bar with elastic modulus E, cross sectional area A, and length L. The bar is discretized into finite elements of length h, and its left end is fixed while its right end is subjected to an axial load P. We split the bar in half as shown in the bottom half of the figure, and reconnect the coincident nodes in the middle with a spring of stiffness βk, where β is a dimensionless parameter and $k = EA/h$ is the stiffness of the node on the left (master) side.

Table 2.1. Extra points for faces of 3D quadratic elements. The extra point at $\eta_1 = 0$ and $\eta_2 = 0$ only applies to 8-node quadrilateral faces.

quad face		tria face	
η_1	η_2	η_1	η_2
-1/2	-1/2	1/3	1/3
1/2	-1/2	1/6	1/6
1/2	1/2	2/3	1/6
-1/2	1/2	2/3	1/6
0	-1/2	5/12	1/6
1/2	0	5/12	5/12
0	1/2	1/6	5/12
-1/2	0		
0	0		

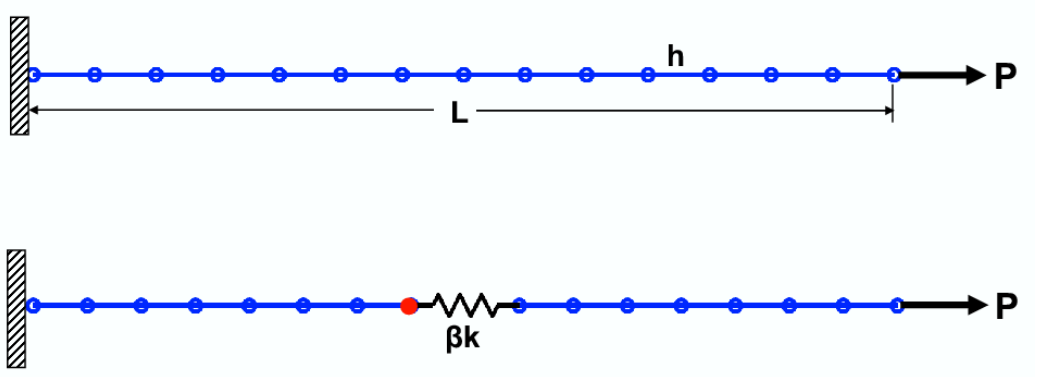

Figure 2.4. One-dimensional bar example to motivate recipe for K_p.

From basic strength of materials, we know that the exact tip deflection of the original bar is given by $\delta = PL/(AE)$, whereas the tip deflection of the bar with a spring in the middle is $\delta(\beta) = [1+(1/\beta)(h/L)]\delta$. We see that $\delta(\beta) = \delta$ as $\beta \to \infty$, but we do not want to pick β too large for 2D and 3D applications. Otherwise, mesh interfaces would be too stiff. Rather, the goal is to choose β just large enough so that the convergence rates of the finite elements are retained. Let q denote the degree of the finite element, e.g., $q = 1$ for linear and $q = 2$ for quadratic elements. Requiring the relative error of $\delta\beta$ to be of order $(h/L)^{q+1}$ gives

$$(1/\beta)(h/L) = (h/L)^{q+1} \quad \Longrightarrow \quad \beta = (L/h)^q.$$

Motivated by this development, K_p is given in general by

$$K_p = (H_p/h_p)^q k_p, \tag{2.14}$$

where H_p is the diameter of the master surface containing p, h_p is a length associated with p, and k_p is a stiffness matrix associated with p. We note that it is not important to use an exact value for H_p. For example, H_p could be chosen as two times the largest distance from the centroid of all nodes on a master surface to any node on this surface.

It now only remains to specify h_p and k_p. Let h_k denote the average diameter of all element faces on a master surface that contain node k. With reference to (2.8), h_p is given by

$$h_p = \sum_{k \in \mathscr{F}_p} h_k \phi_k(\xi_{1p}, \xi_{2p}).$$

Similarly, the stiffness matrix k_p is given by

$$k_p = \sum_{k \in \mathscr{F}_p} k_k \phi_k(\xi_{1p}, \xi_{2p}),$$

where k_k is the stiffness matrix for node k. One could argue that the method is parameter free since all its terms are clearly defined. Alternatively, one has the option to scale K_p in (2.14) by a fixed amount so that the method can also be viewed as having an adjustable parameter. No such scaling is used in the numerical examples.

One practical concern with any stabilized method is the effect of the stabilization on the condition number of the resulting linear system of equations. Let $\|K_p\|$ denote the 2-norm of K_p. Although the ratio $\alpha = \|K_p\|/\|k_p\|$ increases with mesh refinement, it is not likely to be too large in practice. For example, consider a uniform refinement in which all element diameters are reduced by a factor of 2. If such a refinement is done three times, then α would only increase by a factor of 8^q. Thus, even for quadratic elements, α would only increase by a factor of 64. Numerical results are provided in §2.5.3 which compare the performance of an iterative solver for both classic and stabilized tied contact. As will be seen, the stabilization does not cause any significant reduction in solver performance.

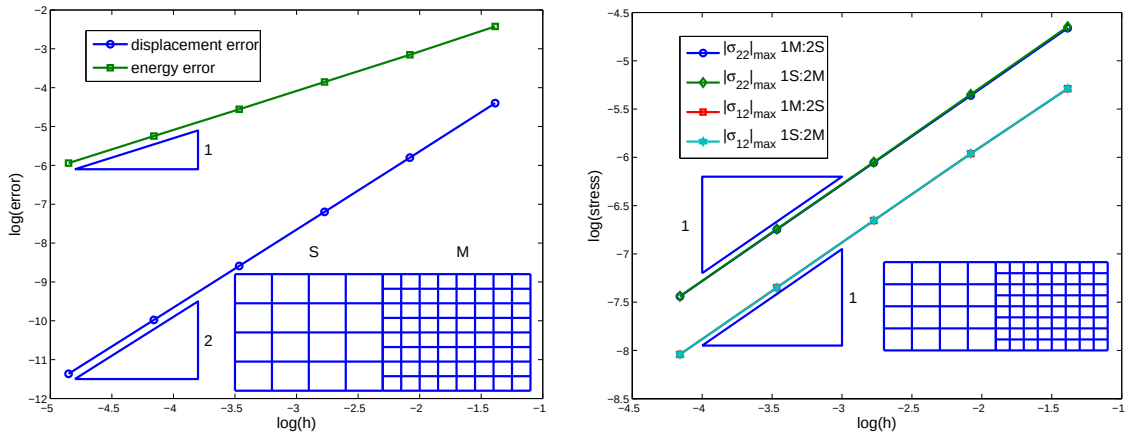

Figure 2.5. Stabilized tied contact results for 1:2 mesh transition example in §2.3.3.

2.5 Numerical Examples

2.5.1 Convergence Tests

2D linear elements: plane stress bending

We return to the example in §2.3.3 where the global convergence rates for classic tied contact were observed to be about half of those for the finite elements used in the mesh. Recall also that stresses near the interface did not converge to the exact values even for very refined meshes. The counterparts of Figure 2.1-right and Figure 2.2 for stabilized tied contact are shown in Figure 2.5. Notice that the global convergence rates are the same as those for the finite elements of the mesh, and the stresses converge to their exact values with mesh refinement. Interestingly, the convergence of stresses shown in Figure 2.5-right are nearly identical to those of the conforming method in which the left side if the mesh interface is chosen as master.

3D linear elements: bending

Results for the 3D counterpart of the example in §2.3.3 are shown in Figure 2.6 for meshes of 8-node hexahedral (HEX8) elements and a variety of mesh transitions. Here the essential boundary conditions are chosen as

$$u_1(0,x_2,x_3)=0, \quad u_2(0,0,0)=0, \quad u_3(0,0,0)=0 \quad u_2(0,0,1)=0,$$

and we replace the natural boundary condition with $u_1(2,x_2,x_3)=-2x_3$. Notice in all cases that the convergence rates are consistent with those of finite element meshes without

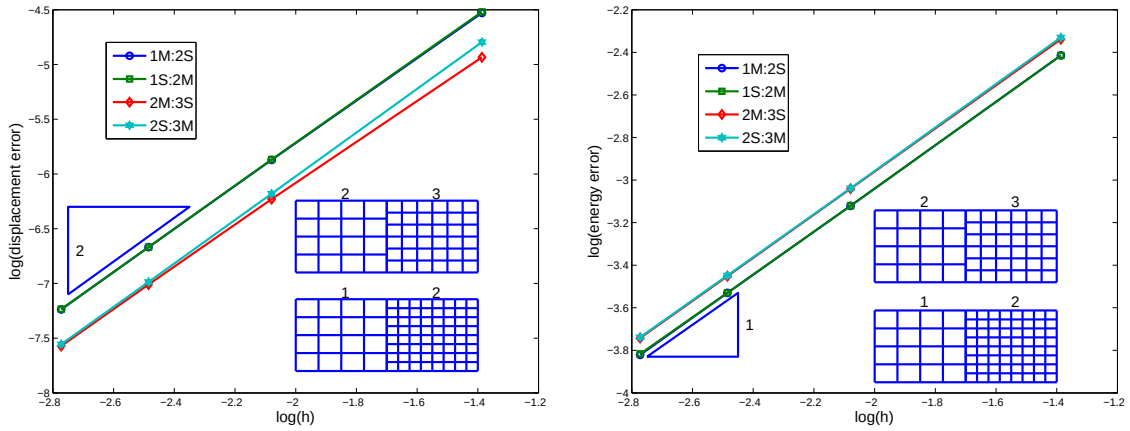

Figure 2.6. Stabilized tied contact results for 1:2 and 2:3 mesh transitions for linear HEX8 elements.

interfaces. Moreover, there is only a slight dependence of the global errors measures on the choice of master and slave sides of the interface.

3D quadratic elements

The final convergence test is for meshes of 20-node hexahedral (HEX20) elements. As in the previous example, the domains for the left and right meshes are unit cubes. The exact solution for the problem is given by

$$\begin{aligned} u_1(x_1,x_2,x_3) &= (1-\cos\pi x_1)(1-\cos 2\pi x_2)(1-\cos 2\pi x_3), \\ u_2(x_1,x_2,x_3) &= 0, \quad u_3(x_1,x_2,x_3) = 0, \end{aligned}$$

and the essential boundary conditions are

$$u_1(0,x_2,x_3) = u_2(0,x_2,x_3) = u_3(0,x_2,x_3) = 0.$$

Body forces corresponding to the exact solution are also applied. Notice in Figure 2.7 that the convergence rates are no less than those for quadratic elements, and the global error measures depend only slightly on the choice of master and slave surfaces.

2.5.2 Nonstructured Interface

Here we repeat the bending example of §2.5.1, but for the meshes shown in Figure 2.8. Again, all stress components are zero for the exact solution except for σ_{11} which has a

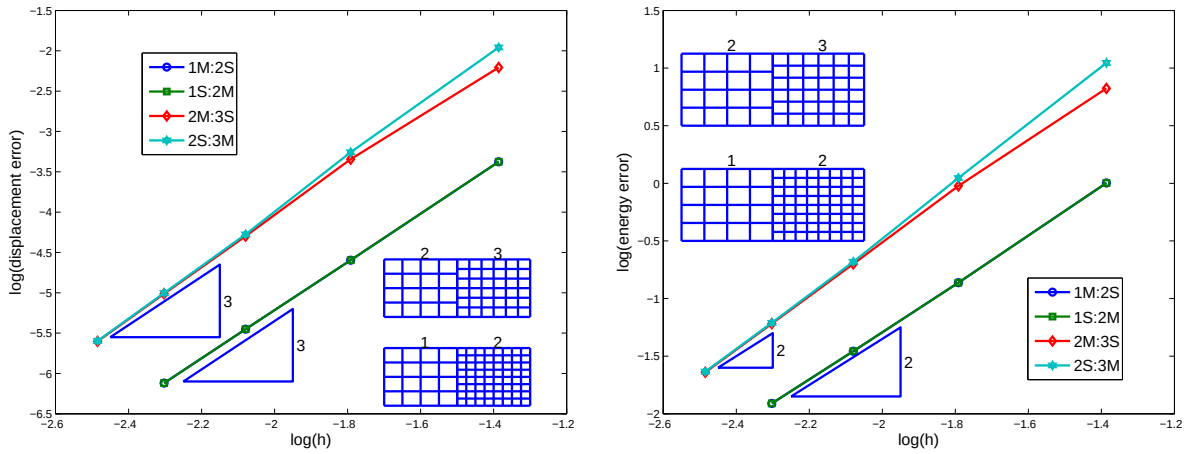

Figure 2.7. Stabilized tied contact results for 1:2 and 2:3 mesh transitions for quadratic HEX20 elements.

maximum absolute value of approximately 1. Results are shown in Table 2.2 for both stabilized and classic tied contact for meshes of HEX8 and HEX20 elements. We see again that the stress results for classic tied contact are very sensitive to the choice of master and slave surfaces, whereas those for stabilized tied contact are not. The results for stabilized tied contact using HEX20 elements are noticeably better than those for HEX8 elements, but the same is not true for classic tied contact. In all cases the error in the transverse tip displacement is less than one percent. We note that the magnitude of stress errors for classic tied contact in the S:M case become larger as the mesh on the right (master) side is refined while keeping the mesh on the left the same.

2.5.3 Iterative Solver Performance

We next investigate the effects of stabilization on the performance of an iterative solver for 3D elasticity problems. The three different meshes of linear HEX8 elements used in this example are shown in Figure 2.9. The meshes (not shown) of quadratic HEX20 elements have half the number of elements in each coordinate direction. The specific problem solved is the one in §2.5.1. For the iterative solver, the problem domain is decomposed into 16 subdomains, and a domain decomposition preconditioner [16] is used together with the conjugate gradient algorithm to solve the linear systems to a relative residual tolerance of 10^{-8}.

Table 2.3 reports on iterative solver performance for both classic and stabilized tied contact. The column headings ndof, #iter, and cond refer to the number of unknowns, the number of iterations, and condition number estimates from conjugate gradient iterations of the preconditioned equations, respectively. Solution times are also reported for a direct

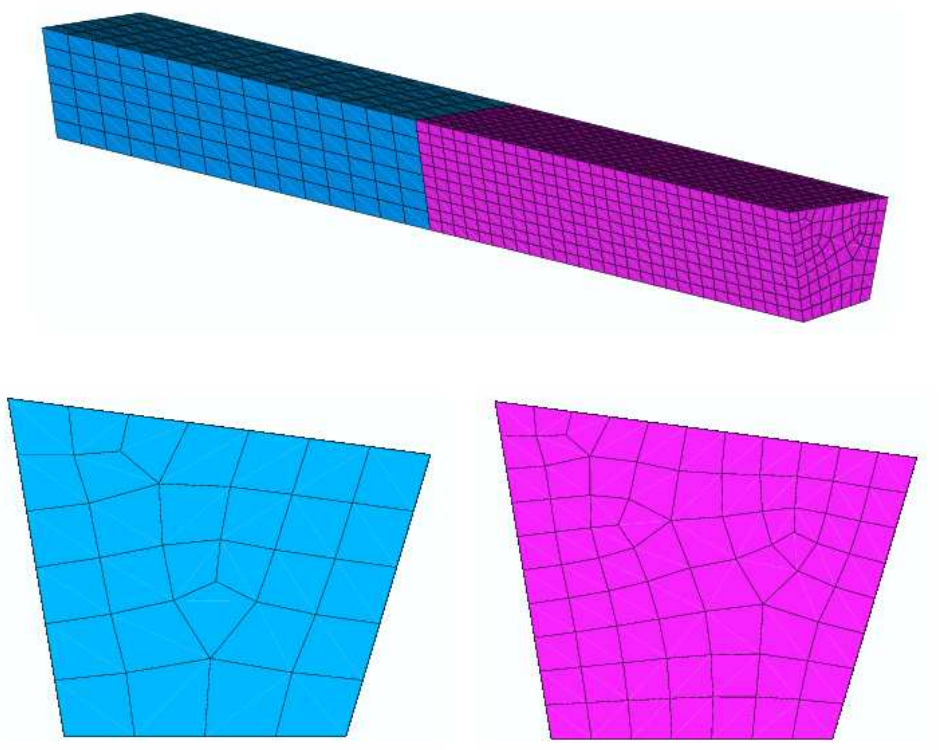

Figure 2.8. Meshes for the example in §2.5.2.

Table 2.2. Tied contact results for example in §2.5.2. The designations M:S and S:M are for the master side of the interface on the left and right, respectively. The relative error in the transverse displacement of a point on the top surface at the end of the beam is denote by e_{tip}.

	HEX8 results				HEX20 results			
	classic		stabilized		classic		stabilized	
	M:S	S:M	M:S	S:M	M:S	S:M	M:S	S:M
$\lvert\sigma_{22}\rvert_{max}$	0.015	0.075	0.017	0.021	0.012	0.063	0.0011	0.0013
$\lvert\sigma_{33}\rvert_{max}$	0.014	0.059	0.019	0.028	0.012	0.10	0.0009	0.0010
$\lvert\sigma_{12}\rvert_{max}$	0.025	0.15	0.007	0.009	0.012	0.18	0.0017	0.0009
$\lvert\sigma_{23}\rvert_{max}$	0.004	0.023	0.009	0.011	0.006	0.055	0.0007	0.0009
$\lvert\sigma_{31}\rvert_{max}$	0.018	0.21	0.011	0.013	0.024	0.17	0.0010	0.0012
e_{tip}	-0.007	-0.007	-0.007	-0.007	0.0002	0.0002	0.0002	0.0002

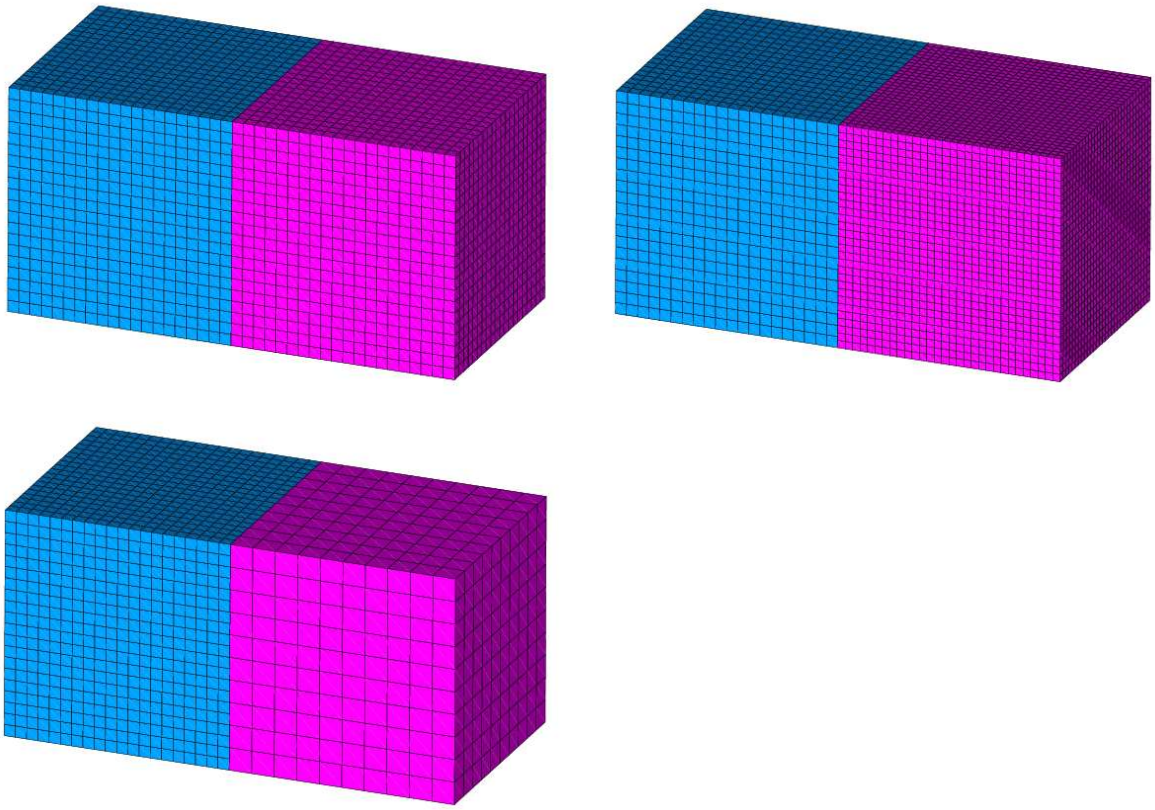

Figure 2.9. Meshes for problems in §2.5.3.

sparse Cholesky solver [33]. It is clear from Table 2.3 that the stabilization does not have a significant effect on the performance of either the direct or iterative solvers. In addition, the iterative solver is faster than the direct one for all problems except the smallest.

2.6 Conclusions

A stabilized form of the classic tied contact method for connecting finite element meshes was presented and observed to have the following features:

1. Improved accuracy and less sensitivity to master-slave designations
2. Optimal convergence rates of underlying finite elements retained
3. Parameter-free
4. Simple physical interpretation based on springs at mesh interfaces

Table 2.3. Solver results for example in §2.5.1 and the meshes shown in Figure 2.9. Solution times for direct and iterative solvers are in seconds.

20x20x20 HEX8 elements on left									
		classic tied contact				stabilized tied contact			
	ndof	direct	iterative	#iter	cond	direct	iterative	#iter	cond
2M:2S	53,358	43	18	19	6.7				
2M:3S	112,868	232	56	23	8.8	228	57	24	8.9
2M:1S	30,848	13	8	18	5.6	13	8	19	6.2
10x10x10 HEX20 elements on left									
2M:2S	28,058	17	13	17	5.3				
2M:3S	58,443	91	41	19	6.5	90	40	22	6.8
2M:1S	16,423	6	7	19	6.7	6	8	20	6.7

5. Computation of potentially complicated surface integrals avoided
6. Iterative solver performance not affected significantly for example 3D problems
7. Can reuse existing computational geometry algorithms for classic tied contact

The stabilization has the effect of adding a positive semidefinite matrix to the stiffness matrix for classic tied contact. Thus, the finite element model for stabilized tied contact is no less stiff than its classic counterpart. We note in closing that the results presented thus far are very encouraging, but a more complete assessment of the method will require further numerical studies for nonplanar interfaces and nonlinear geometric and material behavior.

Chapter 3

Scalable Component Mode Synthesis

U. L. Hetmaniuk
Department of Applied Maths, University of Washington,
Box 352420, Seattle, WA 98195-2420,
(hetmaniu@u.washington.edu)
and
R. B. Lehoucq
Organization 1414, Sandia National Laboratories

3.1 Chapter Abstract

The goal of the work presented here is to introduce basis functions for the finite element discretization of a second order linear elliptic operator with rough or highly oscillating coefficients. The proposed basis functions are inspired by the classic idea of component mode synthesis and exploit an orthogonal decomposition of the trial subspace to minimize the energy. Numerical experiments illustrate the effectiveness of the proposed basis functions.

3.2 Introduction

The finite element solution of the partial differential equation

$$\begin{cases} -\nabla\cdot(c(\mathbf{x})\nabla u(\mathbf{x})) &= f(\mathbf{x}) & \text{in } \Omega, \\ u &= 0 & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

has been the subject of much research. Difficulties arise when the coefficient c associated with the second order linear elliptic operator is rough or highly oscillating so that a standard application of the finite element method necessitates a highly refined mesh. An important task is to define an appropriate approximation space that has knowledge of the coefficient

c, followed by an adroit choice of basis functions, for example functions of local support. These functions give rise to an effective finite element method when a reasonably implemented algorithm with acceptable performance and sufficient accuracy results. Babuška, Caloz, and Osborn [3] denote such finite element methods *special*.

The goal of our paper is to determine a conforming approximation space of functions for the finite element solution of (3.1). In contrast to other approaches, we exploit the fact that the solution u of (3.1) solves the minimization problem

$$\arg \min_{v \in H_0^1(\Omega)} \left(\frac{1}{2} \int_\Omega c(\mathbf{x}) |\nabla v(\mathbf{x})|^2 \, d\mathbf{x} - \int_\Omega f(\mathbf{x}) v(\mathbf{x}) \, d\mathbf{x} \right) \tag{3.2}$$

and therefore is the minimum energy solution. This energy principle represents an intrinsic metric for comparing the quality of approximations to the solution of (3.1). Our procedure is based upon the classic idea of component mode synthesis (CMS), introduced in [24, 12] Starting from a partition of the domain Ω, component mode synthesis methods exploit an orthogonal decomposition of $H_0^1(\Omega)$ to solve the optimality system associated with (3.2). Motivated by this orthogonal decomposition, we develop a conforming finite dimensional approximation space. We contrast our CMS-based approach with the multiscale finite element method (MsFEM) [17] and draw a relationship with the generalized finite element method (GFEM) [5]. We argue that our approach does not fit exactly into the framework of generalized finite element methods (in contrast to MsFEM). We demonstrate the efficacy of our CMS-based approach through a suite of careful numerical experiments.

3.2.1 Notation and assumptions

We quickly review our use of standard notation. Let Ω be a two- or three-dimensional domain with Lipschitz boundary $\partial\Omega$ and so let $H^1(\Omega)$ denote a Sobolev space of order 1; let $H_0^1(\Omega)$ denote a subspace of $H^1(\Omega)$ consisting of functions that vanish on $\partial\Omega$. Let the norm and inner product on $H^1(\Omega)$ and $L^2(\Omega)$ be given by $\|\cdot\|_1, (\cdot,\cdot)_1$, and $\|\cdot\|, (\cdot,\cdot)$, respectively. Let

$$a(u,v) = \int_\Omega c(\mathbf{x}) \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) d\mathbf{x}, \tag{3.3}$$

denote the bilinear form induced by (3.1). We suppose that $a(\cdot,\cdot)$ is coercive,

$$\exists\ \alpha > 0,\ 0 < \alpha \|v\|_1^2 \leq a(v,v), \quad \forall\ v \in H_0^1(\Omega), \tag{3.4}$$

and continuous,

$$\exists\ \gamma > 0,\ a(v,w) \leq \gamma \|v\|_1 \, \|w\|_1 \quad \forall\ v,w \in H_0^1(\Omega). \tag{3.5}$$

We rewrite (3.2) as

$$\arg \min_{v \in H_0^1(\Omega)} \left(\frac{1}{2} a(v,v) - (f,v) \right), \tag{3.6}$$

and the associated optimality system is the variational formulation of (3.1), e.g. given $f \in L^2(\Omega)$, find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = (f,v) \quad \forall v \in H_0^1(\Omega). \tag{3.7}$$

We refer to the solutions of (3.1), (3.2), and (3.7) as equivalent in a formal sense. Our approach is not restricted to (3.1). Other coercive and continuous bilinear forms a can be considered, such as elastostatics.

3.3 Component mode synthesis

We review the classical technique of component mode synthesis [24, 12] from an abstract perspective. Partition the domain Ω into J non intersecting subdomains $\Omega_j, j = 1, \cdots, J$, that share the common interface Γ; see figure 3.1 for the case of 4 subdomains.

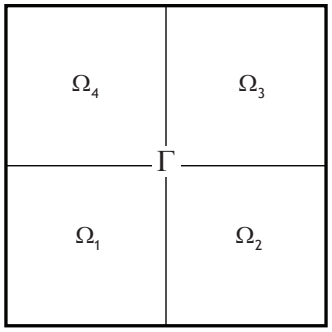

Figure 3.1. The domain Ω partitioned four subdomains.

Let V_{Ω_j} be the subspace of local functions that are nonzero in Ω_j and are trivially extended throughout Ω,

$$V_{\Omega_j} = \{v \in H_0^1(\Omega) : v|_{\Omega \setminus \Omega_j} = 0\}. \tag{3.8}$$

We remark that any member function of V_{Ω_j} has a zero trace on the boundary $\partial\Omega$ and on the interface Γ. Let V_Γ be the subspace of harmonic extensions of trace functions on Γ,

$$V_\Gamma = \{E_\Omega \tau \in H_0^1(\Omega) : \tau \in H_{00}^{1/2}(\Gamma)\}, \tag{3.9}$$

where $H_{00}^{1/2}(\Gamma)$ denotes the trace space of $H_0^1(\Omega)$ on Γ and the harmonic extension E_Ω of $\tau \in H_{00}^{1/2}(\Gamma)$ solves the minimization problem

$$\inf_{v \in H_0^1(\Omega)} a(v,v) \quad \text{subject to} \quad v|_\Gamma = \tau.$$

We remark that the harmonic extension E_Ω satisfies also

$$\left\{\begin{array}{rcl} -\nabla\cdot(c(\mathbf{x})\nabla E_\Omega\tau(\mathbf{x})) &=& 0 \quad \text{in } \Omega_j, \text{ for all } j, \\ E_\Omega\tau &=& \tau \quad \text{on } \Gamma, \\ E_\Omega\tau &=& 0 \quad \text{on } \partial\Omega. \end{array}\right. \tag{3.10}$$

This property indicates that functions in V_Γ are governed by the underlying partial differential equation. Note that any non-zero member function of V_Γ has a non-zero trace on Γ. The spaces V_Γ and V_{Ω_j} contain the components of the solution among, and within the subdomains, respectively, associated with a rough or highly oscillating coefficient c.

A key result is the orthogonal decomposition

$$H_0^1(\Omega) = \left(\bigoplus_{j=1}^{J} V_{\Omega_j}\right) \oplus V_\Gamma. \tag{3.11}$$

Although not often stated in this form, this is a well-known result, at the heart of the analysis and development of domain decomposition methods for elliptic partial differential equations [39], and modern component mode synthesis methods [9, 7].

The decomposition (3.11) is orthogonal with respect to the inner product $a(\cdot,\cdot)$ because

$$a(v_i, v_j) = 0, \; \forall\, v_i \in V_{\Omega_i}, \; \forall\, v_j \in V_{\Omega_j}, \; (i \neq j), \tag{3.12a}$$

$$a(v_i, v_\Gamma) = 0, \; \forall\, v_i \in V_{\Omega_i}, \; \forall\, v_\Gamma \in V_\Gamma. \tag{3.12b}$$

The former equality follows because the supports of the two functions v_i and v_j are disjoint. The latter equality follows by definition of the harmonic extension (3.10).

The decomposition (3.11) also implies that

$$\min_{v\in H_0^1(\Omega)} \left(\frac{1}{2}a(v,v) - (f,v)\right) = \sum_{j=1}^{J} \min_{v\in V_{\Omega_j}} \left(\frac{1}{2}a(v,v) - (f,v)\right) + \min_{v\in V_\Gamma} \left(\frac{1}{2}a(v,v) - (f,v)\right). \tag{3.13}$$

The solution of (3.7) is the sum of J local functions, respectively in V_{Ω_1}, $\cdots$, V_{Ω_J}, and a function of V_Γ, *i.e.*

$$u = u_1 + \cdots + u_J + u_\Gamma, \tag{3.14}$$

where u_j and u_Γ minimizes the energy in V_{Ω_j} and V_Γ, respectively. The local function $u_j \in V_{\Omega_j}$ satisfies

$$a(u_j, v) = (f, v), \qquad \forall v \in V_{\Omega_j}, \tag{3.15}$$

and is also the orthogonal projection of u onto V_{Ω_j}. The function $u_\Gamma \in V_\Gamma$ satisfies

$$a(u_\Gamma, v) = (f, v), \qquad \forall v \in V_\Gamma, \tag{3.16}$$

and is also the orthogonal projection of u onto V_Γ.

The orthogonal decomposition of the solution given by (3.14) explains that the purpose of $u_\Gamma \in V_\Gamma$ is to couple the J subdomain solutions u_j. Component mode synthesis is

thus defined where components from the $J+1$ subspaces are synthesized to approximate a function over Ω.

An approximating subspace consistent with the decomposition (3.11) arises from selecting a subset of eigenmodes[1] for $a(\cdot,\cdot)$ in the subspaces V_{Ω_j} and V_Γ. To build this approximating subspace, we introduce two different sets of eigenvalue problems. First, we define J *fixed-interface* eigenvalue problems: Find $(z_{*,j},\lambda_{*,j}) \in V_{\Omega_j} \times \mathbb{R}$ such that

$$a(z_{*,j},v) = \lambda_{*,j}(z_{*,j},v) \quad \forall v \in V_{\Omega_j}, \tag{3.17}$$

and, then, the *coupling* eigenvalue problem: Find $(z_{*,\Gamma},\lambda_{*,\Gamma}) \in V_\Gamma \times \mathbb{R}$ such that

$$a(z_{*,\Gamma},v) = \lambda_{*,\Gamma}(z_{*,\Gamma},v) \quad \forall v \in V_\Gamma. \tag{3.18}$$

Note that the only differences between these two eigenvalue problems are the approximating spaces V_{Ω_j} and V_Γ. Because a member of V_Γ is determined by its trace on Γ, the coupling eigenvalue problem (3.18) can be equivalently expressed as follows: Find $(\tau_*,\lambda_{*,\Gamma}) \in H_{00}^{1/2}(\Gamma) \times \mathbb{R}$ such that

$$a(E_\Omega \tau_*, E_\Omega \eta) = \lambda_{*,\Gamma}(E_\Omega \tau_*, E_\Omega \eta) \quad \forall \eta \in H_{00}^{1/2}(\Gamma). \tag{3.19}$$

We assume that the eigenvalues $\{\lambda_{i,j}\}_{i=1}^\infty$ and $\{\lambda_{i,\Gamma}\}_{i=1}^\infty$ are ordered into nondecreasing sequences and that the eigenmodes $z_{*,j}$ and $z_{*,\Gamma}$ are normalized for the L^2 inner product.

The fixed-interface and coupling eigenmodes can then be employed to expand the source term f and the solution u of (3.2)

$$u = \sum_{j=1}^{J} \sum_{i=1}^{\infty} \frac{(f,z_{i,j})}{\lambda_{i,j}} z_{i,j} + \sum_{i=1}^{\infty} \frac{(f,z_{i,\Gamma})}{\lambda_{i,\Gamma}} z_{i,\Gamma}. \tag{3.20}$$

We define the finite-dimensional subspace

$$V_{CMS} = \left(\bigoplus_{j=1}^{J} \operatorname{span}\{z_{i,j}; 1 \le i \le I_j\} \right) \oplus \operatorname{span}\{z_{i,\Gamma}; 1 \le i \le I_\Gamma\}, \tag{3.21}$$

where I_j and I_Γ are non-negative integers. The approximate solution u_{CMS} satisfies

$$a(u_{CMS},v) = (f,v), \quad \forall v \in V_{CMS}, \tag{3.22}$$

and is given by the truncated series

$$u_{CMS} = \sum_{j=1}^{J} \sum_{i=1}^{I_j} \frac{(f,z_{i,j})}{\lambda_{i,j}} z_{i,j} + \sum_{i=1}^{I_\Gamma} \frac{(f,z_{i,\Gamma})}{\lambda_{i,\Gamma}} z_{i,\Gamma}. \tag{3.23}$$

[1]The *natural* choice of eigenmodes is frequent in structural analysis and optimal, among subspaces with the same dimension, in terms of n-widths (see [2, Theorem 5.1]).

The following energy estimate easily follows

$$a(u-u_{CMS},u-u_{CMS}) \leq \sum_{j=1}^{J} \frac{1}{\lambda_{I_j+1,j}} \sum_{i=I_j+1}^{\infty} (f,z_{i,j})^2 + \frac{1}{\lambda_{I_\Gamma+1,\Gamma}} \sum_{i=I_\Gamma+1}^{\infty} (f,z_{i,\Gamma})^2. \quad (3.24)$$

This energy estimate indicates that an accurate approximation of u is obtained when fixed-interface eigenmodes and coupling modes are combined in the approximation subspace.

When the approximation subspace V_{CMS} does not contain any fixed-interface mode (*i.e.* $V_{CMS} \subset V_\Gamma$), the energy norm of the error becomes

$$a(u-u_{CMS},u-u_{CMS}) = \sum_{j=1}^{J} a(u_j,u_j) + \sum_{i=I_\Gamma+1}^{\infty} \frac{(f,z_{i,\Gamma})^2}{\lambda_{i,\Gamma}}. \quad (3.25)$$

Unless all the local solutions $u_j \in V_{\Omega_j}$ are zero, the error $u-u_{CMS}$ cannot converge to zero as $I_\Gamma \to \infty$. The components u_j satisfy also

$$a(u_j,u_j) = \int_{\Omega_j} f u_j \leq \|f\|_{L^2(\Omega_j)} \|u_j\|_{L^2(\Omega_j)} \leq C \mathrm{diam}(\Omega_j) \|f\|_{L^2(\Omega_j)} \|\nabla u_j\|_{L^2(\Omega_j)}, \quad (3.26)$$

where we used the Cauchy-Schwarz and the Poincaré inequalities in succession. Coercivity (3.4) of the bilinear form $a(\cdot,\cdot)$ then results in

$$a(u_j,u_j) \leq \frac{C}{\alpha} \mathrm{diam}^2(\Omega_j) \|f\|^2_{L^2(\Omega_j)}. \quad (3.27)$$

When the components u_j are non-zero on a partition $\mathscr{T}$, the functions u_j may not be negligible. But, when the partition is refined, the subdomains Ω_j and their diameters, $\mathrm{diam}(\Omega_j)$, both decrease. So the error $u-u_{CMS}$ can converge to zero with $V_{CMS} \subset V_\Gamma$ as the partition is refined.

On the other hand, when the approximation subspace V_{CMS} does not contain any coupling modes (*i.e.* $V_{CMS} \cap V_\Gamma = \{0\}$), the energy norm of the error becomes

$$a(u-u_{CMS},u-u_{CMS}) = \sum_{j=1}^{J} \sum_{i=I_j+1}^{\infty} \frac{(f,z_{i,j})^2}{\lambda_{I_j+1,j}} + a(u_\Gamma,u_\Gamma). \quad (3.28)$$

Unless the coupling function u_Γ is zero (or the trace of u on Γ is zero), the error $u-u_{CMS}$ cannot converge to zero when all the indices I_j go to infinity. Contrary to the previous case, refining the partition would make the interface Γ larger and so would not decrease $a(u_\Gamma,u_\Gamma)$.

Consequently, combining (or synthesizing) functions from both V_{Ω_j} and V_Γ into the approximation subspace V_{CMS} is a strategy that can lead to an accurate approximation of u on a coarse partition $\mathscr{T}$.

3.4 New special finite element method

Motivated by the orthogonal decomposition (3.11), our goal is to determine a finite-dimensional subspace V_{ACMS} of $H_0^1(\Omega)$ spanned by basis functions of local support and that approximates V_{CMS} (3.21). The eigenmodes in V_{Ω_j} have, by construction, local support but the coupling modes in V_Γ have typically global support in Ω. So we propose to select basis functions of local support from the subspaces V_{Ω_j} and V_Γ.

To simplify the presentation, we assume that $\Omega = (0,1)\times(0,1)$ and that $\mathscr{T}$ is a partition of Ω into rectangles Ω_j. The interface Γ is the union of all the interior edges between two rectangles. We remind the reader that the subspace V_{Ω_j}, defined by (3.8), contain functions of zero trace on Γ and can only hold information on the subdomain Ω_j. Functions of V_Γ (3.9) are governed by the underlying partial differential equation because they are harmonic extensions in Ω of trace functions on Γ. They satisfy the boundary value problem (3.10).

The conforming discretization space we propose is consistent with the decomposition (3.11) and the basis functions have local support. With the partition $\mathscr{T}$, we define the subspace

$$V_{ACMS} := \left(\bigoplus_{j=1}^{J} \operatorname{span}\{z_{1,j}\}\right) \oplus \left[\left(\bigoplus_{P\in\Omega} \operatorname{span}\{\varphi_P\}\right) \oplus \left(\bigoplus_{e\subset\Omega} \operatorname{span}\{\psi_e\}\right)\right], \tag{3.29}$$

where $z_{1,j}$ is the first *fixed-interface* mode (3.17) in Ω_j and the letter A in *ACMS* stands for approximate. Note that the vertices P and the edges e are taken in the interior of Ω. The Dirichlet boundary condition is built into V_{ACMS}.

For any interior point P of the partition $\mathscr{T}$, φ_P belongs to V_Γ and is a harmonic extension satisfying

$$\left\{\begin{array}{rcll} -\nabla\cdot(c(\mathbf{x})\nabla\varphi_P(\mathbf{x})) &=& 0 & \text{in } \Omega_j,\\ \varphi_P &=& 0 & \text{on } \partial\Omega,\\ \varphi_P &\neq& 0 & \text{on } \Gamma,\\ \varphi_P(P') &=& \delta_{P,P'}, & \end{array}\right. \tag{3.30}$$

for any element Ω_j, where $\delta_{P,P'}$ is the Kronecker delta function. On Γ, we select a trace for φ_P that has local support along the boundaries of elements sharing the vertex P. The resulting function φ_P will also have as support the elements sharing the point P. On a horizontal edge $[x_L,x_P]\times\{y_P\}$, the trace for φ_P is defined by

$$\varphi_P(x,y_P) = \left(\int_{x_L}^{x} \frac{ds}{c(s,y_P)}\right) \Big/ \left(\int_{x_L}^{x_P} \frac{ds}{c(s,y_P)}\right) \quad \forall x\in[x_L,x_P]. \tag{3.31}$$

Along a vertical edge, a similar definition is used.[2] Figure 3.2 plots an example of trace for φ_P. Note that the trace is piecewise monotonic along the edges.

[2] Hou and Wu [23, Section 2.2] proposed the two-dimensional trace (3.31) in their MsFEM-O approach. This trace is motivated by one-dimensional problems for which Babuška and Osborn [4] recommended the local approximation, $\operatorname{span}\left\{1, \int_{x_0}^{x} \frac{ds}{c(s)}\right\}$ instead of $\operatorname{span}\{1,x\}$.

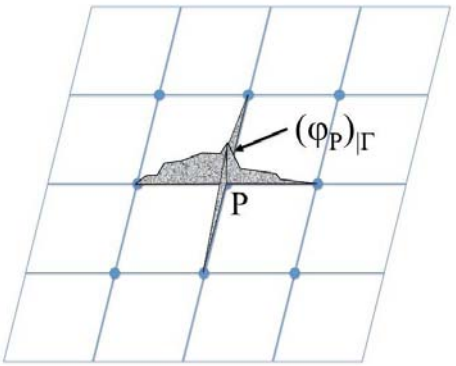

Figure 3.2. Trace of φ_P along Γ for a domain partitioned into 16 subdomains

The function ψ_e, where e is an interior edge, belongs also to V_Γ and is the harmonic extension of $\tau_e \in H_{00}^{1/2}(\Gamma)$, whose support is the edge, e, between two elements. The trace function τ_e is the first eigenmode for the *coupling* mode problem:

$$a(E_\Omega \tau_e, E_\Omega \eta) = \lambda (E_\Omega \tau_e, E_\Omega \eta), \quad \forall\ \eta \in H_{00}^{1/2}(\Gamma) \text{ such that } \mathrm{supp}(\eta) \subset e. \tag{3.32}$$

An example for $\tau_e = (\psi_e)_{|\Gamma}$ is given in figure 3.3. The function ψ_e satisfies also

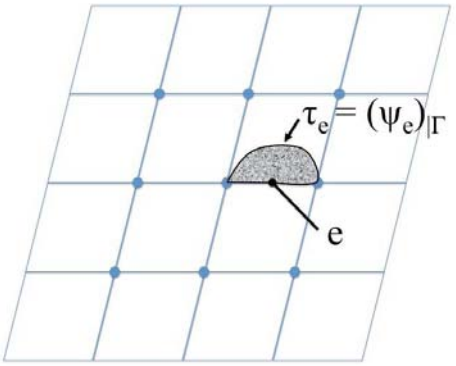

Figure 3.3. Example of a local coupling mode along an interior edge e.

$$\left\{ \begin{array}{rcll} -\nabla \cdot (c(\mathbf{x}) \nabla \psi_e(\mathbf{x})) & = & \lambda \psi_e & \text{in } \Omega_j, \\ \psi_e & = & 0 & \text{on } \partial\Omega, \\ \psi_e & = & \tau_e & \text{on } \Gamma, \end{array} \right. \tag{3.33}$$

for any element Ω_j.

In summary, the conforming finite-dimensional subspace $V_{ACMS} \subset H_0^1(\Omega)$ exploits the orthogonal decomposition (3.11) for incorporating information on the variational form $a(\cdot,\cdot)$. The subspace V_{ACMS} contains information within subdomains Ω_j via the first fixed-interface mode. The functions φ_P and ψ_e carry information among four and two subdomains, respectively. These three special basis functions have local support. The generalization of V_{ACMS} to triangular cells is straightforward.

The special basis functions $z_{1,j}$, φ_P, and ψ_e are obtained numerically. They are computed via a finite element discretization within each element Ω_j. Local problems are solved to obtain the functions $z_{1,j}$, φ_P, and ψ_e (which can be done in parallel). In a second step, a global problem is solved to compute the approximate solution u_{ACMS} in V_{ACMS}. Further details are given in section 3.6. Before presenting the numerical experiments, we discuss other choices of finite-dimensional approximation subspaces.

Remark 1. *By introducing subdomains, the cost of computing eigenmodes in V_{Ω_j} is tractable. However, computing the coupling eigenmodes (3.19) associated with V_{CMS} is nontrivial because a generalized eigenvalue problem composed of Schur and mass complement operators represents a significant computation; see the survey paper [22] for details.*

3.5 Relationship to other approximating methods

Numerous choices of basis functions are possible for defining a finite dimensional subspace of $H_0^1(\Omega)$. Babuška, Caloz, and Osborn [3] use the phrase *special finite elements* to denote finite element methods (FEM) that employ basis functions that, for instance, incorporate specialized knowledge of the partial differential operator. Many methods have been proposed to incorporate relevant information into the special basis functions; for instance the generalized FEM (GFEM) [5] and the multiscale FEM (MsFEM) [17]. The purpose of this section is to compare the special finite element introduced in section 3.4 for the solution of (3.1) with the classical FEM, MsFEM, and GFEM. We only consider comparisons with conforming finite element methods and with methods that do not lead to modifications of the variational formulation, e.g. the bilinear and linear forms of (3.7) are not modified. For instance, MsFEM with oversampling is a nonconforming finite element method [17, p.23] and the recent multiscale framework presented by Nolen, Papanicolaou and Pironneau [34] modifies the variational formulation.

3.5.1 Classical FEM

The standard nodal linear finite element method (Q1) defines an approximation subspace V_{Q1}

$$V_{Q1} := \operatorname{span}\{N_P;\ P \in \mathscr{T}\}, \tag{3.34}$$

where N_P is the bilinear nodal shape function for an interior point P. When c is a constant, for any interior point P, the associated nodal shape function N_P belongs to V_Γ because N_P satisfies

$$\begin{cases} -\Delta N_P &= 0 \quad \text{in } \Omega_j, \text{ for all } j, \\ N_P &\neq 0 \quad \text{on } \Gamma, \\ N_P &= 0 \quad \text{on } \partial\Omega. \end{cases} \tag{3.35}$$

Therefore V_{Q1} is a finite-dimensional subspace of V_Γ that is orthogonal the subspaces V_{Ω_j}.

However, when c is not equal to a constant, the approximation subspace V_{Q1} is no longer a subspace of V_Γ. For any interior point P, the nodal shape function N_P is not a member of V_Γ because N_P is no longer an harmonic extension, *i.e.*

$$\nabla \cdot (c(\mathbf{x})\nabla N_p(\mathbf{x})) \neq 0 \quad \text{in } \Omega_j, \tag{3.36}$$

when Ω_j intersects the support of N_P. The nodal shape function N_P is not a member of Ω_j either because its trace on Γ is non zero. Therefore the nodal shape function N_P has nonzero components in V_Γ and some V_{Ω_j} in stark contrast to φ_P defined by (3.30).

3.5.2 MsFEM

The MsFEM of Hou and Wu [23] selects basis functions exclusively from V_Γ. A MsFEM basis function φ_P is defined by (3.30) and its trace along the interface Γ. This choice leads to the approximating subspace

$$V_{MsFEM} := \bigoplus_{P\in\Omega} \text{span}\{\varphi_P\} \subset V_\Gamma \subset H_0^1(\Omega), \tag{3.37}$$

When c is constant, the MsFEM is equivalent to the linear finite element method, e.g., $V_{MsFEM} = V_{Q1}$. When c is not equal to a constant, V_{MsFEM} is no longer equal to V_{Q1} but remains a subspace of V_Γ. The orthogonal decomposition (3.11) indicates that MsFEM is a generalization of the linear finite element method for a nonconstant coefficient c because $V_{MsFEM} \subset V_\Gamma$. When the partition $\mathscr{T}$ is coarse, components in V_{Ω_j} of the solution u are not computed by V_{MsFEM} and this error may limit the accuracy of the computed solution in V_{MsFEM}. This limitation is also explained by the error analysis (3.25)–(3.27) that results from the absence of components in V_{Ω_j}. To remove this limitation and decrease the error, a partition finer than $\mathscr{T}$ needs to be used.

The MsFEM-O [3] arises when φ_P is the same harmonic extension used in V_{ACMS}, defined by (3.30) and the trace (3.31). On the other hand, the MsFEM-L results when the trace of φ_P on Γ is set equal to the trace of N_p. In an attempt to mitigate the resonance effect that arises when using MsFEM-L and MsFEM-O, MsFEM-os-L introduces oversampling; see Efendiev and Hou [23, Section 2.3] for a discussion. However, oversampling leads

[3] Hou and Wu points that O indicates the oscillatory boundary condition defining φ_P. However, φ_P is not oscillatory because its trace is monotonic on each edge. This monotonicity arises because the coercivity of a implies that the coefficient c is positive.

to discontinuous basis functions and, hence, a nonconforming finite element method. In contrast, the method proposed in section 3.4 is conforming.

3.5.3 GFEM

The GFEM space is formally defined by

$$V_{GFEM} := \left\{ \sum_{j=1}^{N} \phi_j \xi_j : \ \xi_j \in S_j, \quad \sum_{j=1}^{N} \phi_j = 1 \text{ on } \Omega, \quad \Omega = \bigcup_j \omega_j, \quad \phi_j = 0 \text{ on } \Omega \setminus \omega_j, \ j = 1, \ldots, N \right\},$$

where the patches ω_1, $\cdots$, ω_N are open sets. The finite dimensional space S_j contains functions ξ_j defined on ω_j,

$$S_j = \text{span}\left\{ \xi_{i,j} \in H^1(\omega_j); \ \xi_{i,j} = 0 \text{ on } \overline{\omega}_j \cap \partial\Omega \right\}, \tag{3.38}$$

such that the functions ξ_j approximate well, on ω_j, the solution u with respect to the energy norm. The functions $\{\phi_j\}$ form a *partition of unity* on Ω. Their role is to paste together the local approximation functions, $\xi_j \in S_j$, to form global approximation functions that are conforming, *i.e.* $\phi_j \xi_j$ will belong to $H_0^1(\Omega)$. If, in addition, the functions ϕ_j and their gradients $\nabla\phi_j$ are uniformly bounded, Babuška, Banerjee, and Osborn[5] prove convergence estimates for GFEM. Note that their proof can give suboptimal convergence rates (see [5, p. 88-89]). In order to show that a special finite element method is a GFEM, we need to exhibit patches $\{\omega_j\}$, the partition of unity $\{\phi_j\}$, and subspaces S_1, $\cdots$, S_N.

We now establish a relationship between V_{ACMS} and V_{GFEM} in two steps. We first demonstrate that MsFEM is a generalized finite element method[4]. Second, we show that V_{ACMS} is a proper subspace of a GFEM subspace.

Consider the functions φ_P defined by (3.30). The definition extends easily to the case where the vertex P belongs to $\partial\Omega$. Based on the choice of trace function (3.31), the functions $\{\varphi_P\}$ satisfy

$$\sum_{P \in \overline{\Omega}} \varphi_P(\mathbf{x}) = 1, \quad \forall\, \mathbf{x} \in \Omega$$

(see also Hou and Wu [23, p. 173]). Therefore, the shape functions $\{\varphi_P\}$ form a partition of unity on Ω. We can select the family $\{\phi_j\}$ to be the family $\{\varphi_P\}$ and the patches $\{\omega_j\}$ to be the support of the shape functions φ_P. Introduce the finite dimensional subspace S_j,

$$S_j = \begin{cases} \{0\} & \text{when } \overline{\omega}_j \cap \partial\Omega \neq \emptyset, \\ \text{span}\{1\} & \text{otherwise.} \end{cases} \tag{3.39}$$

The space S_{MsFEM},

$$S_{MsFEM} = \text{span}\left\{ \phi_j \xi_j; \text{ where } \xi_j \in S_j \text{ defined by (3.39)}, \ j = 1, \cdots, N \right\},$$

[4]To the best of our knowledge, this relation between MsFEM and GFEM is new.

is a generalized finite element approximation space. By construction, this space is equal to V_{MsFEM}, defined by (3.37). So MsFEM is a generalized finite element method where the local approximation functions ξ_j are constant and where the partition of unity functions ϕ_j are harmonic extensions. This particular choice of partition of unity is unusual because the partition of unity involves the partial differential equation.

Next, for the space V_{ACMS}, the partition of unity $\{\varphi_P\}$ and the patches $\{\omega_j\}$ are retained. Introduce the local approximating subspace S_j,

$$S_j = \begin{cases} \{0\} \oplus \text{span}\{\psi_e;\ e \subset \overline{\omega}_j \cap \Omega\} \oplus \text{span}\{z_{1,k};\ \Omega_k \subset \omega_j\} & \text{when } \overline{\omega}_j \cap \partial\Omega \neq \emptyset, \\ \text{span}\{1\} \oplus \text{span}\{\psi_e;\ e \subset \overline{\omega}_j\} \oplus \text{span}\{z_{1,k};\ \Omega_k \subset \omega_j\} & \text{otherwise.} \end{cases} \tag{3.40}$$

The space S_{ACMS},

$$S_{ACMS} = \text{span}\left\{\phi_j \xi_j;\ \text{where } \xi_j \in S_j \text{ defined by (3.40)},\ j = 1, \cdots, N\right\},$$

is a generalized finite element approximation space where the local approximation functions are the constant, the edge-based functions ψ_e, and the fixed-interface modes $z_{1,*}$. V_{ACMS} is a subspace of S_{ACMS} because the partition of unity property implies

$$z_{1,k} = \sum_{P \in \overline{\Omega}_k} \varphi_P z_{1,k} \quad \text{and} \quad \psi_e = \sum_{P \in \overline{\Omega}_k;\ e \cap \overline{\Omega}_k \neq \emptyset} \varphi_P \psi_e. \tag{3.41}$$

However, V_{ACMS} is different from S_{ACMS} because the dimension of S_{ACMS} is larger than the dimension of V_{ACMS}. For example, in S_{ACMS}, the functions $\{\varphi_P z_{1,1}\}_{P \in \overline{\Omega}_1}$ are linearly independent while the definition for V_{ACMS} contains only one instance of $z_{1,1}$. Our proposed special finite element method is a proper subspace of S_{ACMS} and does not appear to be equivalent to a generalized finite element method. Consequently, the GFEM theory does not apply directly to V_{ACMS} (in contrast to MsFEM). Note that because the functions ψ_e and $z_{1,*}$ belong to $H_0^1(\Omega)$ by construction, V_{ACMS} does not require any pasting for these functions.

3.6 Numerical Experiments

We present a series of numerical experiments using our CMS-inspired special finite element method introduced in section (3.4). We first discuss aspects associated with the computations. The first set of experiments is on the Laplace equation. The second and third sets of experiments are on (3.1) with a nontrivial coefficient c. All three cases compare the proposed special FEM with MsFEM and with CMS. The second set of experiments also illustrates the effect of the fixed interface modes and of the trace functions defining φ_P.

3.6.1 Practical remarks

In this section, we discuss practical aspects for the numerical experiments. First we give details on obtaining the basis functions $z_{1,*}$, φ_P, and ψ_e and on assembling the resulting

stiffness matrix. We describe the sparsity of the stiffness matrix. Finally, we describe how the approximate solutions are compared.

Computation of basis functions

Let $\mathscr{T}_n$ be a partition of $\Omega = (0,1)\times(0,1)$ with n square elements per direction and a uniform mesh size $h = 1/n$. To compute the special shape functions $z_{1,*}$, φ_P, and ψ_e, each element is divided into $m \times m$ square elements with $h_f = h/m$. The local submeshes are conforming among elements.

We use piecewise bilinear elements to compute the special shape functions by solving local problems. For the functions φ_P, we solve approximately the problem (3.30). This solution is local to an element Ω_j and the corresponding linear system is of dimension $(m-1)^2$. For the fixed-interface modes $z_{1,j}$, we solve approximately (3.17). The corresponding discrete eigenproblem is local to Ω_j and of dimension $(m-1)^2$. The first eigenmode is computed with a *direct* solver. For an edge-based function ψ_e, we solve approximately (3.32). Recall that, for any function η supported on an edge e between the elements Ω_1 and Ω_2, its harmonic extension $E_\Omega\eta$ has support in $\overline{\Omega}_1 \cup \overline{\Omega}_2$, and has the discrete representation

$$\mathbf{E}\eta = \begin{bmatrix} -\mathbf{K}_{11}^{-1}\mathbf{K}_{1e} \\ -\mathbf{K}_{22}^{-1}\mathbf{K}_{2e} \\ \mathbf{I} \end{bmatrix} \eta,$$

where η is the discrete representation of η. $\mathbf{K}_{11}$ and $\mathbf{K}_{22}$ are the local stiffness matrices in, respectively, Ω_1 and Ω_2. Then we compute the first eigenmode for the pencil

$$\left(\begin{bmatrix} -\mathbf{K}_{11}^{-1}\mathbf{K}_{1e} \\ -\mathbf{K}_{22}^{-1}\mathbf{K}_{2e} \\ \mathbf{I} \end{bmatrix}^T \begin{bmatrix} \mathbf{K}_{11} & \mathbf{0} & \mathbf{K}_{1e} \\ \mathbf{0} & \mathbf{K}_{22} & \mathbf{K}_{2e} \\ \mathbf{K}_{1e}^T & \mathbf{K}_{2e}^T & \mathbf{K}_{ee} \end{bmatrix} \begin{bmatrix} -\mathbf{K}_{11}^{-1}\mathbf{K}_{1e} \\ -\mathbf{K}_{22}^{-1}\mathbf{K}_{2e} \\ \mathbf{I} \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} -\mathbf{K}_{11}^{-1}\mathbf{K}_{1e} \\ -\mathbf{K}_{22}^{-1}\mathbf{K}_{2e} \\ \mathbf{I} \end{bmatrix}^T \begin{bmatrix} \mathbf{M}_{11} & \mathbf{0} & \mathbf{M}_{1e} \\ \mathbf{0} & \mathbf{M}_{22} & \mathbf{M}_{2e} \\ \mathbf{M}_{1e}^T & \mathbf{M}_{2e}^T & \mathbf{M}_{ee} \end{bmatrix} \begin{bmatrix} -\mathbf{K}_{11}^{-1}\mathbf{K}_{1e} \\ -\mathbf{K}_{22}^{-1}\mathbf{K}_{2e} \\ \mathbf{I} \end{bmatrix} \right)$$

or, equivalently, the pencil of the Schur and mass complements (of dimension $m-1$).

The assembly of the global stiffness matrix and the right-hand side vector requires the computation of the volume integrals, for example,

$$\int_{\Omega_1} c(\mathbf{x})\nabla\varphi_P(\mathbf{x})\cdot\nabla\psi_e(\mathbf{x})d\mathbf{x}, \tag{3.42}$$

on $\mathscr{T}_h$. We exploit the expression of φ_P and ψ_e on the submesh contained in $\overline{\Omega}_1$

$$\varphi_P = \sum_{P_f\in\overline{\Omega}_1} \chi_{P_f} N_{P_f} \quad \text{and} \quad \psi_e = \sum_{P_f\in\overline{\Omega}_1} \xi_{P_f} N_{P_f} \tag{3.43}$$

where N_{P_f} is the piecewise bilinear shape function for the point P_f on the submesh contained in $\overline{\Omega}_1$. Using the stiffness matrix $\mathbf{K}_f$ computed on the submesh, we write

$$\int_{\Omega_1} c(\mathbf{x})\nabla\varphi_P(\mathbf{x})\cdot\nabla\psi_e(\mathbf{x})d\mathbf{x} = \left[(\chi_{P_f})_{P_f\in\overline{\Omega}_1}\right]^T \mathbf{K}_f \left[(\xi_{P_f})_{P_f\in\overline{\Omega}_1}\right] \tag{3.44}$$

The other volume integrals are computed similarly.

Sparsity of stiffness matrix

The approximate solution to (3.7) in a finite-dimensional subspace will be obtained by solving a linear system with a direct solver. Table 3.1 lists information about the linear system for the different approximation subspaces. With bilinear finite elements, the sub-

Subspace	Matrix Dimension	Matrix Non-Zeros
V_{Q1}	$(n-1)^2$	$\approx 9(n-1)^2$
$V_{MsFEM-O}$	$(n-1)^2$	$\approx 9(n-1)^2$
V_{ACMS}	$(2n-1)^2$	$\approx 12(2n-1)^2$
V_{CMS}	$(2n-1)^2$	$= (2n-1)^2$

Table 3.1. Matrix dimensions and non-zeros for different special finite element methods

space V_{Q1} has $(n-1)^2$ degrees of freedom and, asymptotically, 9 non-zero entries per row. The subspace $V_{MsFEM-O}$ generates a matrix with the same dimension and the same sparsity pattern. For our proposed special finite element method, the subspace V_{ACMS} (3.29) has n^2 fixed interface modes, $(n-1)^2$ functions φ_P, and $2n(n-1)$ edge functions. The dimension of V_{ACMS} is

$$n^2 + (n-1)^2 + 2n(n-1) = (n+n-1)^2 = (2n-1)^2.$$

With V_{ACMS}, the stiffness matrix contains a diagonal block for the n^2 fixed interface modes. A row associated with φ_P (respectively ψ_e) has at most 21 (resp. 13) non-zero entries. So an estimate for the number of non-zeros is

$$1\times n^2 + 21\times(n-1)^2 + 13\times 2n(n-1) \approx \left(\frac{1}{4}+\frac{21}{4}+\frac{26}{4}\right)\times(2n-1)^2 = 12\times(2n-1)^2.$$

For the sake of comparison, we use also the subspace V_{CMS} (3.21) with 1 fixed-interface mode per element and $(2n-1)^2-n^2$ coupling modes. The dimension of V_{CMS} is also $(2n-1)^2$. The resulting linear system will be diagonal. We emphasize that V_{CMS} is not practical because it demands a large number of global coupling eigenmodes whose computations are daunting. However, for the numerical experiments, we will compute these global coupling eigenmodes accurately as a basis for comparison.

Metric for comparing approximate solutions

Recall that the solution u solves the minimization problem (3.2) and therefore is the minimum energy solution. The energy,

$$\mathscr{E}(v) = \frac{1}{2}\int_\Omega c(\mathbf{x})|\nabla v(\mathbf{x})|^2\,d\mathbf{x} - \int_\Omega f(\mathbf{x})v(\mathbf{x})\,d\mathbf{x} = \frac{1}{2}a(v,v) - (f,v),$$

represents an intrinsic metric for comparing the quality of approximations to the solution u. Between two approximate solutions, the one with lowest energy is the most accurate one.

Computing the difference between the energy of the computed solution and the energy of the exact solution u is equivalent to computing the norm of the error for the inner product $a(\cdot,\cdot)$. Indeed, we have

$$\begin{aligned}\left(\frac{1}{2}a(u_{Q1},u_{Q1}) - (f,u_{Q1})\right) - \left(\frac{1}{2}a(u,u) - (f,u)\right) &= \frac{1}{2}a(u_{Q1},u_{Q1}) - (f,u_{Q1}) + \frac{1}{2}a(u,u)\\ &= \frac{1}{2}\left(a(u_{Q1},u_{Q1}) - 2(f,u_{Q1}) + a(u,u)\right)\\ &= \frac{a(u-u_{Q1},u-u_{Q1})}{2}\end{aligned}$$

when u_{Q1} is the approximate solution computed on V_{Q1} and where we used

$$a(u,u_{Q1}) = (f,u_{Q1}) \quad \text{and} \quad a(u,u) = (f,u)$$

(from (3.7)). This difference of energies is an intrinsic metric for comparing the quality of approximations. When the exact solution u is not explicitly known, approximating the minimal energy,

$$\mathscr{E}^* = \frac{1}{2}a(u,u) - (f,u) = -\frac{a(u,u)}{2} = -\frac{(f,u)}{2}, \tag{3.45}$$

is simpler than extrapolating the exact solution. In the numerical experiments, we compute the energy differences.

3.6.2 Experiments with the Laplace equation

Consider the problem

$$\left\{\begin{array}{rcll} -\Delta u &=& f & \text{on } \Omega\\ u &=& 0 & \text{in } \partial\Omega\end{array}\right. \tag{3.46}$$

We choose $f(x,y) = 2x(1-x) + 2y(1-y)$ such that the exact solution u is $x(1-x)y(1-y)$.

Introduce a mesh $\mathscr{T}_n$ composed of squares with uniform mesh size $h = 1/n$. $\mathscr{T}_n$ contains n^2 elements, $(n-1)^2$ interior points, and $2n(n-1)$ interior edges. We compare the accuracy of computed solutions when using different finite-dimensional subspaces.

Figure 3.4 plots convergence curves for the difference of energies, which is proportional to the H^1 semi-norm of the error, in terms of the number of degrees of freedom. The number of degrees of freedom is, indeed, more relevant than the mesh size h or the number of elements per direction n. As highlighted in Table 3.1, the considered approximation subspaces have different dimensions on the same mesh $\mathscr{T}_n$. As expected, the bilinear finite

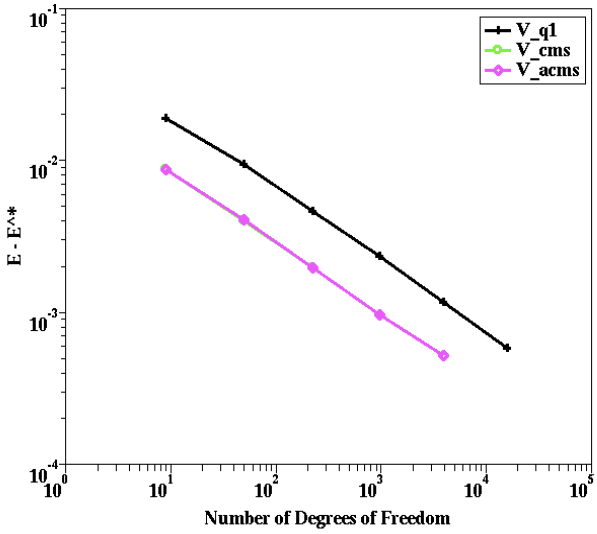

Figure 3.4. Comparison of special finite element methods for problem (3.46).

element has a convergence rate proportional to h or inversely proportional to the square root of the total number of degrees of freedom. The curves for V_{CMS} and V_{ACMS} are indistinguishable, indicating that the basis functions in $V_{ACMS} \cap V_\Gamma$ with local support approximate well the subspace spanned by the global eigenmodes for the Schur and mass complements.

For a fixed number of degrees of freedom, the approximate solutions computed in the subspaces V_{CMS} and V_{ACMS} are more accurate than in the subspace V_{Q1}. To reach a fixed level of accuracy for this problem, V_{CMS} and V_{ACMS} require 5 times less degrees of freedom than V_{Q1}.

For the curves in Figure 3.4, the special basis functions $z_{1,*}$, φ_P, and ψ_e, were approximated with 16×16 bilinear finite elements in a square element of $\mathscr{T}_n$, *i.e.* $h_f = h/16$. Figure 3.5 illustrates the convergence of the energy $\mathscr{E}$ for the subspace V_{ACMS} with a fixed mesh size h as $m = h/h_f$ increases. A ratio of $m = 16$ is sufficient to compute numerically the special basis functions.

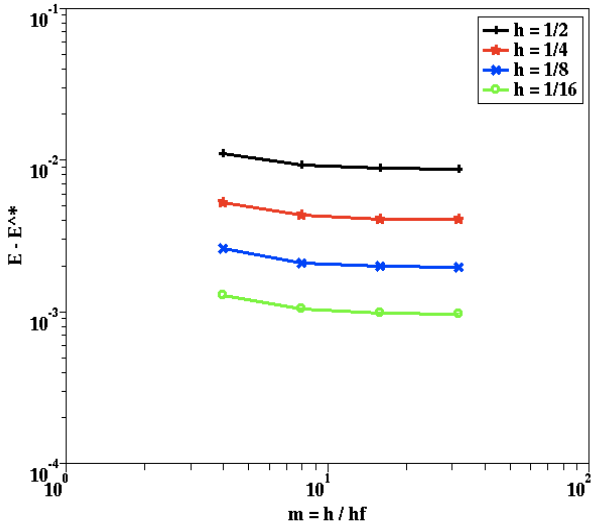

Figure 3.5. Effect of subcell mesh size to compute basis functions of V_{ACMS} for problem (3.46).

3.6.3 Experiments with a varying coefficient

Consider the problem

$$\begin{cases} -\nabla\cdot\left(\dfrac{1}{1.2+\cos(32\pi x(1-x)y(1-y))}\nabla u(x,y)\right) &= f \quad \text{on } \Omega \\ u &= 0 \quad \text{in } \partial\Omega \end{cases} \tag{3.47}$$

We choose $f(x,y) = 64\pi\,[x(1-x)+2y(1-y)]$ such that the exact solution u is

$$u(x,y) = (1.2\times 32\pi)x(1-x)y(1-y)+\sin(32\pi x(1-x)y(1-y)).$$

Note that the coefficient c oscillates while the source term f does not.

Introduce a mesh $\mathscr{T}_n$ composed of squares with uniform mesh size $h = 1/n$. $\mathscr{T}_n$ contains n^2 elements, $(n-1)^2$ interior points, and $2n(n-1)$ interior edges.

Convergence plots

We compare the accuracy of computed solutions using the finite-dimensional subspaces V_{Q1}, $V_{MsFEM-O}$, V_{ACMS}, and V_{CMS}. Since the coefficient c is varying, the subspace $V_{MsFEM-O}$ is different from the subspace V_{Q1}.

Figure 3.6 plots convergence curves for the difference of energies, which is proportional to the energy norm of the error, in terms of the number of degrees of freedom. For this

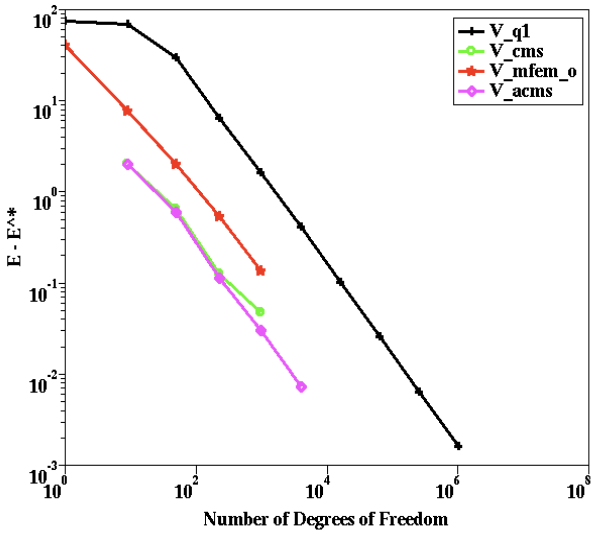

Figure 3.6. Comparison of special finite element methods for problem (3.47).

problem, the value for $\mathscr{E}^*$ is

$$\mathscr{E}^* = -132.67094817007. \tag{3.48}$$

As expected, the bilinear finite element has a convergence rate proportional to h^2 or inversely proportional to the total number of degrees of freedom. The curves for V_{CMS} and V_{ACMS} are aligned, indicating again that the local basis functions in V_{ACMS} approximate well the subspace spanned by the global eigenmodes for the Schur and mass complements. For a fixed number of degrees of freedom, the approximate solutions computed in V_{CMS} and in V_{ACMS} are the most accurate followed by the subspace $V_{MsFEM-O}$. The approximate solution in V_{Q1} is the least accurate. The solution from V_{ACMS} is more accurate than the solution from $V_{MsFEM-O}$, highlighting the importance of the edge functions ψ_e and the fixed-interface modes $z_{1,j}$. To reach a fixed level of accuracy for this problem,

- $V_{MsFEM-O}$ requires 15 times less degrees of freedom than V_{Q1},
- V_{ACMS} requires almost 55 times less degrees of freedom than V_{Q1}.

Table 3.2 compares the approximations obtained with the subspaces V_{Q1}, $V_{MsFEM-O}$, and V_{ACMS}. For this example, the subspaces $V_{MsFEM-O}$ and V_{ACMS} generate good approxi-

h	Matrix Dimension			Matrix Non-Zeros			$\mathscr{E} - \mathscr{E}^*$		
	V_{Q1}	$V_{MsFEM-O}$	V_{ACMS}	V_{Q1}	$V_{MsFEM-O}$	V_{ACMS}	V_{Q1}	$V_{MsFEM-O}$	V_{ACMS}
1/2	1	1	9	1	1	25	75.1	41.0	2.02
1/4	9	9	49	49	49	361	69.3	7.64	0.60
1/8	49	49	225	361	361	2,185	29.9	2.02	0.11
1/16	225	225	961	1,849	1,849	10,441	6.42	0.53	0.03
1/32	961	961	3,969	8,289	8,289	45,385	1.63	0.13	0.007

Table 3.2. Matrix dimension, matrix non-zeros, and energy error for different special finite element methods

mations of u on meshes that are too coarse for V_{Q1} or the piecewise linear interpolation. The subspaces V_{Q1} and $V_{MsFEM-O}$ generate matrices with the same dimensions, the same sparsity patterns, and an average of 9 non-zero entries per row. The subspace V_{ACMS} generates a matrix with an average of 12 non-zero entries per row. For $h = 1/2$, the subspace V_{ACMS} reaches a level of accuracy that the subspace V_{Q1} reaches when h is close to $1/30$. This ratio of 15 between the mesh sizes corresponds to a factor 55 for the degrees of freedom. Between $V_{MsFEM-O}$ and V_{ACMS}, the subspace V_{ACMS} uses 4 times less degrees of freedom than $V_{MsFEM-O}$ that would correspond to a ratio of 4 between the mesh sizes.

For the curves in Figure 3.6, the special basis functions were approximated with, at least, 32×32 bilinear finite elements in any square element of $\mathscr{T}_n$, *i.e.* $h_f \leq h/32$. Figure 3.7 illustrates the convergence of the energy for the subspace V_{ACMS} with a fixed mesh size h as $m = h/h_f$ increases. For this problem, a ratio of $m = 32$ appears sufficient to compute numerically the special basis functions. Further analysis is required to define a priori rules for choosing m; see, for instance Brezzi and Marini [10] for a study on two-level methods.

Impact of basis functions ψ_e and $z_{1,*}$

The error bound (3.24) and the discussion at the end of section 3.3 highlight the importance of approximating the components of u in V_{Ω_j} and in V_Γ. Failure to do so might require a finer partition $\mathscr{T}$ and a larger number of degrees of freedom in order to reach a prescribed level of accuracy, as implied by the results of Table 3.2. In the next experiment, we emphasize the importance of the functions φ_P, ψ_e, and $z_{1,*}$. We compute approximate solutions with the following finite-dimensional subspaces:

- $V_{MsFEM-O} = \text{span}(\varphi_P;\ \text{vertex } P \in \Omega)$;
- $V_{MsFEM-O-INT} = V_{MsFEM-O} \oplus \text{span}(z_{1,j};\ 1 \leq j \leq J)$;
- $V_{MsFEM-O-EDGE} = V_{MsFEM-O} \oplus \text{span}(\psi_e;\ \text{edge } e \subset \Omega)$;

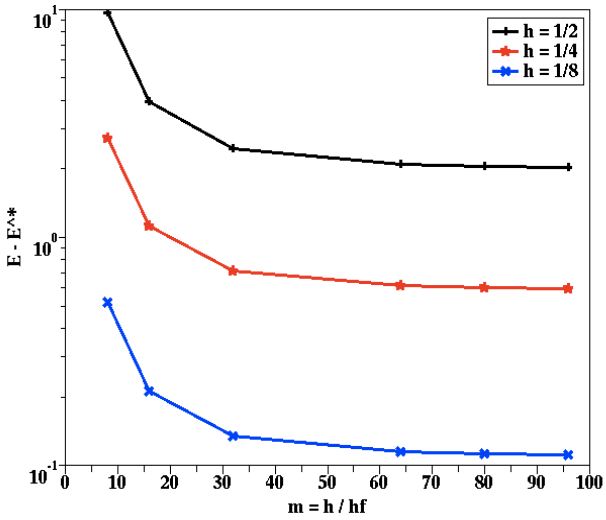

Figure 3.7. Effect of subcell mesh size to compute basis functions of V_{ACMS} for problem (3.47).

- $V_{ACMS} = V_{MsFEM-O} \oplus \text{span}(\psi_e;\ \text{edge } e \subset \Omega) \oplus \text{span}(z_{1,j};\ 1 \leq j \leq J)$.

Note that $V_{MsFEM-O}$ and $V_{MsFEM-O-EDGE}$ are proper subspaces of V_Γ while $V_{MsFEM-O-INT}$ and V_{ACMS} have components in V_{Ω_j} and V_Γ.

Figure 3.8 plots convergence curves for the energy difference in terms of the number of degrees of freedom. The curves for $V_{MsFEM-O}$, for $V_{MsFEM-O-INT}$, for $V_{MsFEM-O-EDGE}$, and for V_{ACMS} were computed on the same set of partitions. Recall that, on a given partition $\mathscr{T}$, all these subspaces have different dimension. For a fixed level of accuracy, the approximate solution in V_{Q1} requires the largest number of degrees of freedom, followed by the approximation in $V_{MsFEM-O-EDGE}$, in $V_{MsFEM-O}$, in $V_{MsFEM-O-INT}$, and in V_{ACMS}. The subspaces $V_{MsFEM-O}$ and $V_{MsFEM-O-EDGE}$ approximate only the component of the solution u in V_Γ. Their respective convergence curves indicate that adding more basis functions in V_Γ does not improve the accuracy per degree of freedom because these subspaces do not approximate the components in V_{Ω_j}. On the other hand, adding the first fixed-interface eigenmodes to $V_{MsFEM-O}$ improves the accuracy per degree of freedom. Indeed, the subspace $V_{MsFEM-O-INT}$ approximates now all the components of u. Incorporating all the functions ϕ_P, ψ_e, and $z_{1,j}$ in V_{ACMS} gives the best accuracy per degree of freedom among all the subspaces.

We emphasize that, on a given partition $\mathscr{T}$, the subspace V_{ACMS} is larger than $V_{MsFEM-O}$ and computes a more accurate approximation to u. However, the gain in accuracy is so

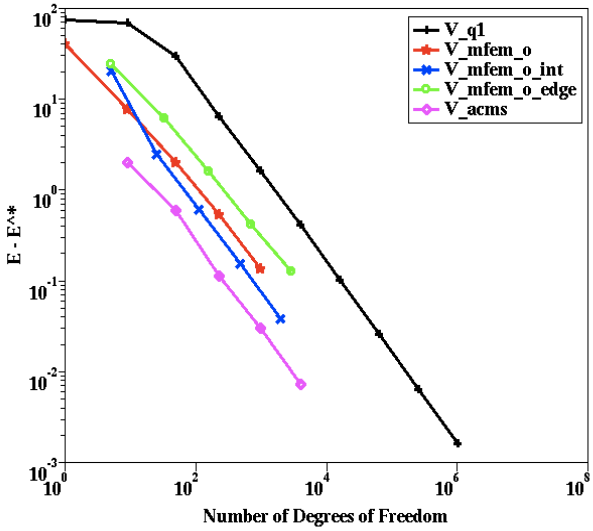

Figure 3.8. Comparison of subspaces motivated by the decomposition (3.11) for problem (3.47).

large that the accuracy with $V_{MsFEM-O}$ on $\mathscr{T}$ is reached with a subspace V_{ACMS} built on a partition coarser than $\mathscr{T}$. For this example, the subspace V_{ACMS} reaches the same level of accuracy than the subspace $V_{MsFEM-O}$ with 4 times less degrees of freedom. This ratio of 4 in the number of degrees of freedom translates into a coarser partition with a mesh size smaller by a factor 4.

Impact of choice for the trace of φ_P

When building the approximating subspace V_{ACMS}, the definition of functions φ_P requires a choice of traces on Γ. Even though the functions φ_P still reside in V_Γ, different traces on Γ result in different approximating subspaces. For example, we could use the functions φ_P^L satisfying the boundary value problem (3.30) and having the same trace on Γ as the bilinear shape function N_P (the piecewise linear variation on Γ is indicated by the superscript L). Figure 3.9 plots such a trace for φ_P^L.

Figure 3.10 plots convergence curves for the energy difference in terms of the number of degrees of freedom for solutions computed with the subspaces V_{ACMS} and V_{ACMS-L}. The subspace V_{ACMS-L} differs only from V_{ACMS} by the replacement of the functions φ_P with φ_P^L. Note that when c is constant, the subspaces V_{ACMS-L} and V_{ACMS} are equal. The approximation with V_{ACMS-L} appears to require a finer mesh to reach the asymptotic regime. Before reaching its asymptotic regime, the curve for V_{ACMS-L} exhibits a bump. The curves

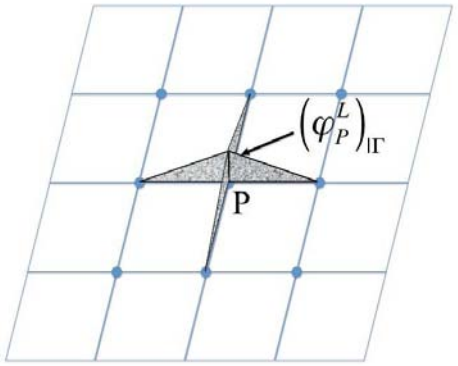

Figure 3.9. Trace of φ_P^L along Γ for a domain partitioned into 16 subdomains

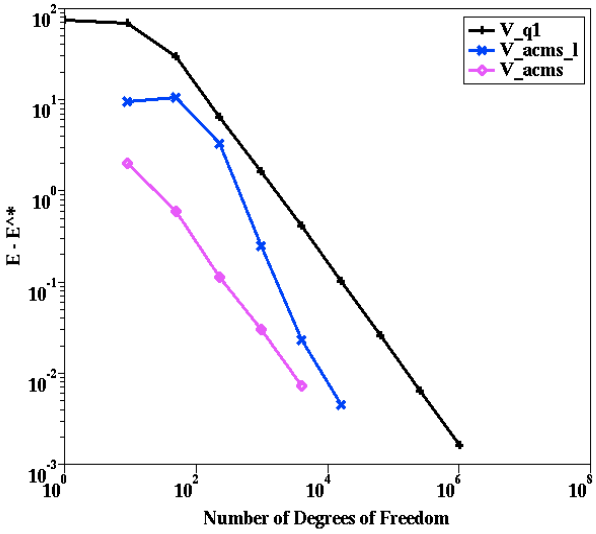

Figure 3.10. Comparison of two choices for the functions φ_P when solving problem (3.47).

for V_{ACMS-L} and V_{ACMS} are different. But the curve for V_{ACMS-L} appears to reach asymptotically the curve for V_{ACMS}, which would be consistent with the case where c is constant.

This experiment highlights the importance of choosing an appropriate trace on Γ for the function φ_P in order to preserve the property that the subspace V_{ACMS} approximates well the subspace V_{CMS}.

3.6.4 Experiments with another varying coefficient

Finally, consider the problem

$$\begin{cases} -\nabla\cdot(c(\mathbf{x})\nabla u(\mathbf{x})) &= f(\mathbf{x}) \quad \text{on } \Omega, \\ u &= 0 \quad \text{in } \partial\Omega. \end{cases} \tag{3.49}$$

We choose $f=-1$ and the scalar coefficient c

$$c(x,y)=\frac{2+1.8\sin(25\pi x)}{2+1.8\cos(25\pi y)}+\frac{2+\sin(25\pi y)}{2+1.8\sin(25\pi x)}. \tag{3.50}$$

This example was studied in the paper [23].

On a mesh $\mathscr{T}_n$ made of squares with uniform mesh size $h=1/n$, we compare the accuracy of computed solutions when using the finite-dimensional subspaces V_{Q1}, $V_{MsFEM-O}$, V_{ACMS}, and V_{CMS}. Figure 3.11 plots convergence curves for half the energy norm of the error in terms of the number of degrees of freedom. The reference energy $\mathscr{E}^*$,

$$\mathscr{E}^*=-0.004717883361515083, \tag{3.51}$$

is computed by Richardson extrapolation based on energies computed with bi-quadratic finite elements and with quintic finite elements using COMSOL Multiphysics[5].

All the methods have a convergence rate inversely proportional to the total number of degrees of freedom. For a fixed number of degrees of freedom, the approximate solution computed in V_{CMS} is the most accurate followed by the subspaces V_{ACMS} and $V_{MsFEM-O}$. The approximate solution in V_{Q1} is the least accurate. Here the curves for V_{CMS} and V_{ACMS} are different. The approximation with V_{ACMS} appears to require a finer mesh to reach the asymptotic regime. Before reaching its asymptotic regime, the curve for V_{ACMS} exhibits a bump. This bump seems similar to the one for V_{ACMS-L}, described in section 3.6.3. It was removed when the functions φ_P^L were replaced by the functions φ_P. This experiment suggests that, for this example, the current choice of trace on Γ for φ_P might not be optimal. Further analysis is required to find a different choice of trace functions that would allow V_{ACMS} to attain its asymptotic regime with fewer degrees of freedom.

[5] Version 3.5a, see www.comsol.com

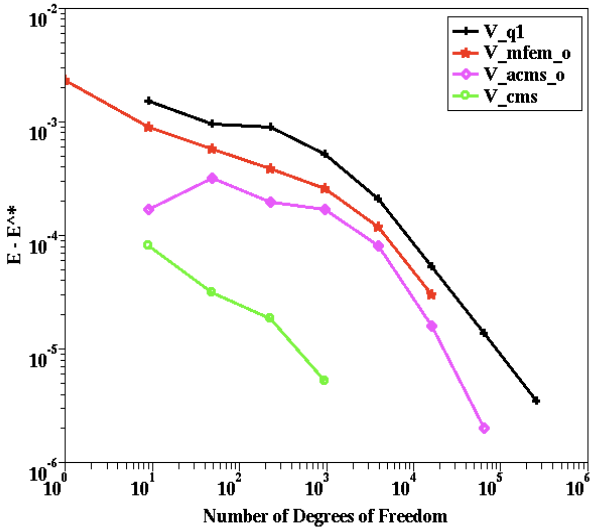

Figure 3.11. Comparison of special finite element methods for problem (3.49).

3.7 Conclusions

We have presented a new conforming special finite element method. The approach is based on the classical idea of component mode synthesis and exploits a $H_0^1(\Omega)$ orthogonal decomposition. Fixed-interface eigenmodes, vertex-based harmonic extensions, and edge-based modes define the approximating subspace V_{ACMS}. We illustrated theoretically and numerically the importance of the three types of functions to obtain an accurate approximate solution. On academic examples, the new approximation subspace is, for the same number of degrees of freedom, more accurate than the bilinear finite element and the multiscale finite element method.

Chapter Acknowledgements

The authors acknowledge the useful comments from the anonymous referees and Dan Segalman of Sandia National Laboratories. They also thank Prof. J. Osborn (U. Maryland) and Prof. U. Banerjee (Syracuse U.) for enlightening discussions about the generalized finite element method and the *n*-width.

U. L. Hetmaniuk was supported in part by the Laboratory Directed Research and Development program at Sandia National Laboratories.

Chapter 4

Method of Discontinuous Basis Functions

Daniel J. Segalman, Organization 1525
and Ulrich . L. Hetmaniuk, University of Washington

4.1 Introduction

Though various Sandia programs have access to massively parallel computers and finite element code that can employ many processors simultaneously, the resulting numerical predictions often are difficult to interpret physically. This difficulty is particularly frustrating in the area of structural dynamics where modal analysis has historically been a major tool both for calculation and for interpretation. Once the nonlinearity of the structure has been acknowledged, modal analysis no longer applies and the remaining tools are awkward to apply and without intuitively obvious physical meaning. The need to address this issue motivated the LDRD funding that made the work discussed in the following possible.

The method presented here provides a partial resurrection of modal analysis in the context of nonlinear structures whose nonlinearity is local in nature. Though for problems of large size or complexity it is still be necessary to employ large computing resources in order to exploit the method presented here, two major advantages are gained:

1. The results are presented in terms of modal coordinates so that often the predictions lend themselves to direct physical interpretation.

2. The reduced order system runs so quickly that many calculations over long periods of time can be run casually. Force boundary conditions can be changed and the system can be recalculated with minimal difficulty or additional computer resources.

The most straight-forward approach to model reduction for nonlinear systems is that of employing assumed modes in a Galerkin formulation. (A good discussion on Galerkin

methods can be found in [27].) This is the approach most often used in problems of modest and diffuse nonlinearity and it is often very successful. As one would expect, the success of a Galerkin approach depends largely on whether the set of basis functions employed span the space of the full solution. We shall see that in cases of localized nonlinearities, it is necessary to include within the basis functions ones that can accommodate the locality of the nonlinearity. Examples are provided.

The initial portion of this report employs mathematical quantities specific to interfaces. How one evaluates those quantities is discussed in a following section.The presentation that follows focuses specifically on problems of structural dynamics, but one anticipates that these techniques could be applied to model reduction of other classes of problem characterized by local nonlinearity.

4.2 Formulation

4.2.1 Galerkin formulation for a System of Localized Nonlinearity

Here we assume a discretization has already been performed - probably by a fine level Galerkin finite element process. The governing equation now has the following nonlinear differential-algebraic form:

$$M\ddot{u} + C\dot{u} + \hat{K}u + \sum_j f_j(s_j, \{\zeta_j\})F^j = f_x(t) \tag{4.1}$$

Above M is the mass matrix and C is the damping matrix, the f_j are (nonlinear) forces acting between node pair j of the system and f_x is the vector of external loads. The nonlinear interface force f_j is a function of the distance s_j between node pair j and of state variables $\{\zeta_j\}$ that evolve along with s_j. The matrix $\hat{K}$ captures the linear elasticity of the rest of the structure; it is the stiffness matrix of a conventional finite element code, where the nonlinear interfaces are ignored.

The vector F^j captures the direction of forces between the node pair j (system degrees of freedom j_1 and j_2) and is related to nodal kinematics by

$$F_k^j = \partial s_j / \partial u_k \tag{4.2}$$

where u_k is the k^{th} degree of freedom of the finite element discretization.

The Galerkin procedure begins with some assumed deformation modes $\{y_k\}$ so that the kinematics of the problem can be approximated by

$$u(t) = a_k(t)\, y_k \tag{4.3}$$

The coefficients $\{a_k(t)\}$ are referred to as generalized coordinates. Here and in the following, summation on repeated indices is assumed.

The next step is to assert that the residual is orthogonal to the each of the assumed deformation modes:

$$y_n^T M y_k \ddot{a}_k + y_n^T C y_k \dot{a}_k + y_n^T \hat{K} y_k a_k + f_j(s_j, \{\zeta_j\}) y_n^T F^j = y_n^T f_x(t) \tag{4.4}$$

for each y_n. This is simplified as

$$\tilde{M}\ddot{a} + \tilde{C}\dot{a} + \tilde{K}a + f_j(s_j, \{\zeta_j\})\tilde{F}^j = \tilde{f}_x(t) \tag{4.5}$$

Ideally, one can obtain adequate solutions to the nonlinear system with far fewer basis functions y_n than the degrees of freedom of the original finite element formulation. The success of a Galerkin approach generally hinges on the appropriate choice of basis functions.

4.2.2 Reference Linear System

The first basis functions that come to mind are the eigen modes of a reference linear system.

Say that at small loads, our interface forces can be approximated as

$$df_j(s_j, \{\zeta_j\}) = \left.\frac{\partial f_j(s, \{\zeta_j = 0\})}{\partial s}\right|_{s=0} ds_j = k_j\, ds = k_j F^j\, du \tag{4.6}$$

In that range of small loads, the governing equation (Eq. 4.1) becomes

$$M\ddot{u} + C\dot{u} + K_0 u = f_x(t) \tag{4.7}$$

where

$$K_0 = \hat{K} + \sum_j k_j F^j F^{j^T} \tag{4.8}$$

The use of a subset of the eigen modes of the reference linear system (RLS) in the linear system itself is the familiar modal truncation. Modal truncation of a RLS is illustrated on the structure depicted in Figure 4.1. The eleven unit masses of this system are connected by springs of unit stiffness. An external triangularly shaped impulse of duration equal to one quarter of the longest period is applied to the mass at the free end. It is primarily the first mode that is excited, so one expects modal truncation to serve as a good approximation to the full system. Indeed Figure 4.2 shows modal truncation to be quite adequate for this problem. In this figure and in other kinetic energy plots, the legend refers to the envelopes of the kinetic energies.

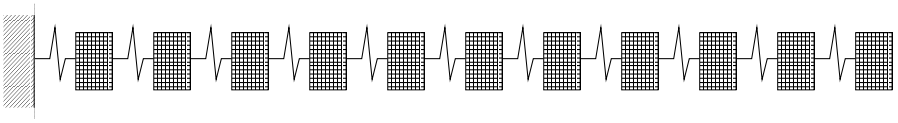

Figure 4.1. For purposes of illustration, we consider this simple system of eleven unit masses connected in a series manner to ground by a system of unit springs.

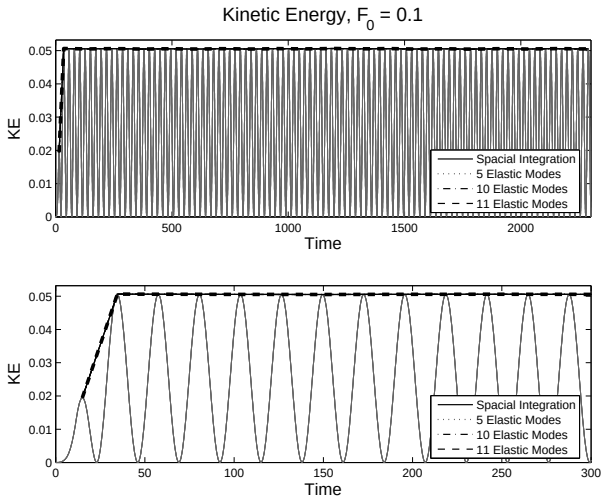

Figure 4.2. The response of the system shown in Figure 4.1 is calculated by the numerical solution for full spacial system (eleven degrees of freedom) and by several levels of modal truncation. In this and in similar plots, the legend refers to envelopes of the kinetic energy curves.

4.2.3 Galerkin Solution, Modal Truncation, and Slightly Nonlinear Systems

Let's now consider a system that is just slightly nonlinear. We supplement the unit spring between the 5^{th} and 6^{th} masses of the structure in Figure 4.1 with a slightly nonlinear spring so that the net force between the masses is

$$f(s) = K_1 s + K_2 s^3 \tag{4.9}$$

where $K_1 = 1$ and $K_2 = 50$. This structure is shown symbolically in Figure 4.3. Here we consider a very low amplitude (peak force $F_0 = 0.05$) externally applied impulse so that only a little of the nonlinearity is manifest (see Figure 4.4 for the force displacement plot.). Here the full solution of the nonlinear system of eleven differential algebraic equations is our truth model.

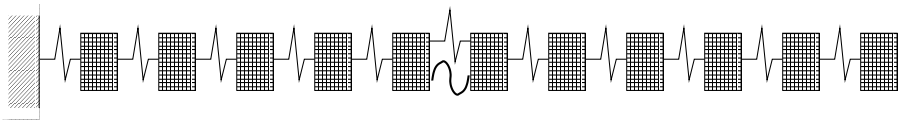

Figure 4.3. We consider this simple eleven unit masses connected in a series manner to ground by a system of unit spring. Additionally we place a cubic spring between the 5^{th} and 6^{th} masses of the system.

We use eigen modes of the reference linear system as basis functions in the Galerkin formulation. Examination of the kinetic energies predicted by our truth model and the reduced models are shown in Figure 4.5. The good agreement between the predicted kinetic energies argues that for this case, a Galerkin procedure using eigen modes of a RLS can yield good approximation.

Another indication of the adequacy of the RLS modes to capture the response of the slightly nonlinear system is a comparison of the singular value decomposition (SVD) modes of the solution of the full nonlinear system with the RLS eigen modes. (The SVD method identifies correlations among degrees of freedom. A good discussion on using SVD to explore the properties of nonlinear systems can be found in [28].) Figure 4.6 shows the first SVD mode and the first RLS eigen mode to be nearly identical. Figure 4.7 shows that it is only the first SVD mode that plays a significant role in the response to the low amplitude triangular impulse.

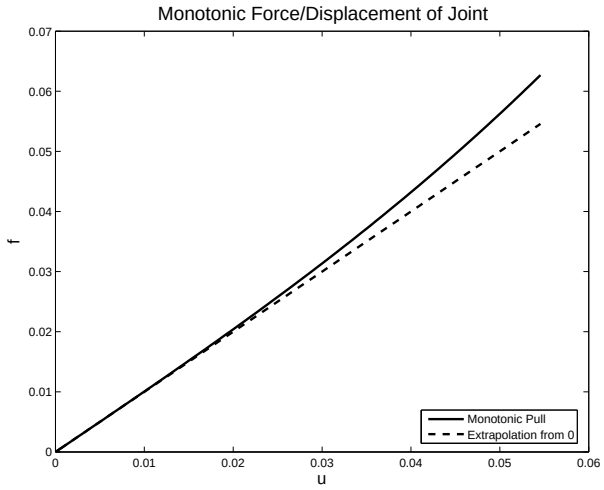

Figure 4.4. The response of a system with a small cubic nonlinearity appears almost linear so long as the excitations are also small.

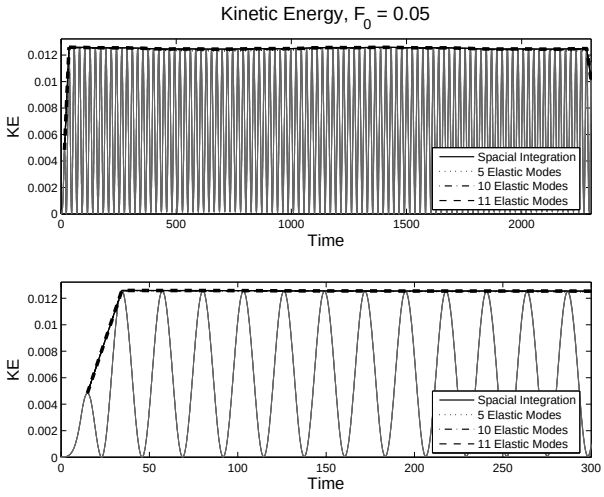

Figure 4.5. The kinetic energy of the system with a small cubic nonlinearity resulting from a triangularly shaped impulse. The Galerkin solution employing various numbers of eigen modes of the reference linear system provides a reasonably good approximation to this slightly nonlinear system.

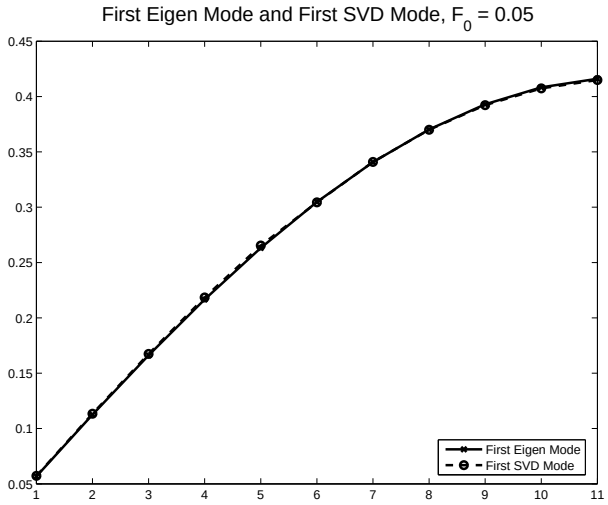

Figure 4.6. The response of a system with a small cubic nonlinearity is explored the the singular value decomposition (SVD) modes of the history of the full nonlinear system. Shown here are the first SVD mode of the fully history and the first eigen mode of the RLS. For this small nonlinearity, both modes are almost identical.

4.2.4 Problems of Larger Nonlinearity

Problems of even large nonlinearity are often quite amendable to Galerkin approximation employing modes of neighboring linear systems. Generally those successes are ones where the nonlinearity is diffused smoothly through a significant part of the structure. We show in this section examples of problems where the nonlinearity is very local in nature and the eigen modes of a reference linear system are a less adequate basis.

Consider an eleven-element nonlinear structure identical to that discussed above (Figure 4.3), but subject to a higher amplitude triangular impulse. The force-displacement curve of the parallel linear and cubic springs is shown in Figure 4.8 where significant nonlinearity is observed.

This problem is much less amenable to Galerkin solution using the RLS eigen modes. Figure 4.9 shows the kinetic energy of the system over time predicted by the full nonlinear solution and by various levels of Galerkin approximation. Not only is approximation by five modes inadequate, but approximation by even ten modes results in significant error. Only when the number of modes is equal to the total number of physical degrees of freedom of the system does the kinetic energy predicted by a modal approximation match that of the full nonlinear solution.

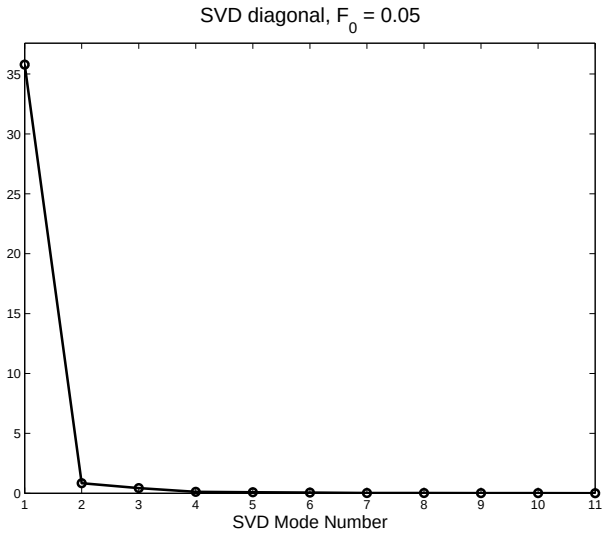

Figure 4.7. The response of a system with a small cubic nonlinearity is explored the the singular value decomposition (SVD) modes of the history of the full nonlinear system. The relative role of each SVD mode in the history is shown here. Only the first such mode is significant.

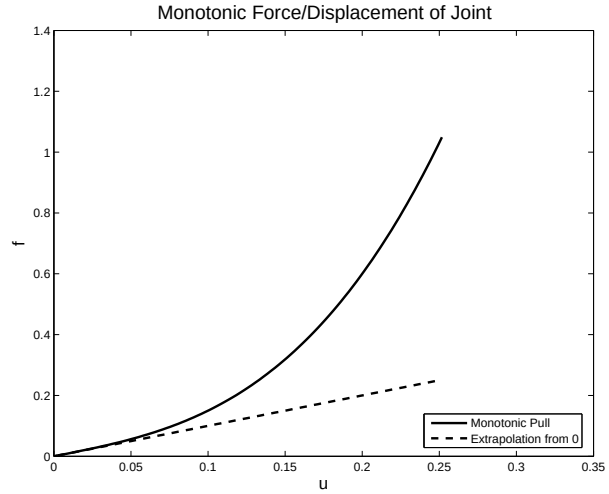

Figure 4.8. The response of a system with a cubic nonlinearity appears extremely nonlinear when the excitations are large. In this case the peak excitation is 0.5.

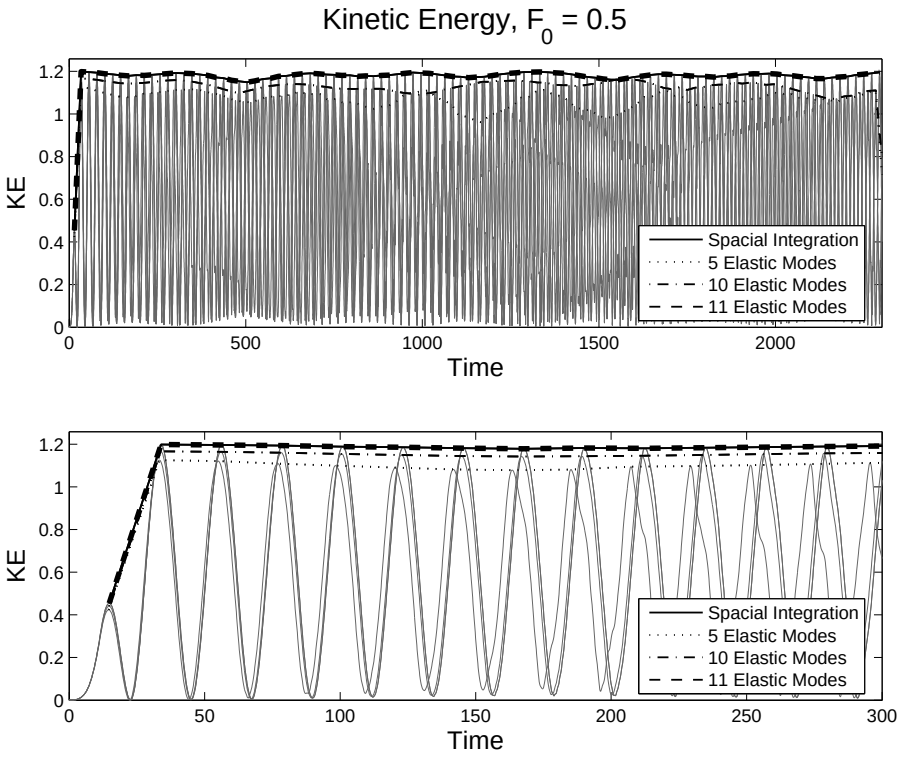

Figure 4.9. The kinetic energy of the system with a large cubic nonlinearity resulting from a triangularly shaped impulse. The Galerkin solution employing various numbers of eigen modes of the reference linear system does not provide a good approximation to this nonlinear system unless the number of modes equals the total number of degrees of freedom of the physical system.

The source of the difficulty of performing model reduction of systems of local nonlinearity is suggested in Figure 4.10 where the first SVD mode of the full nonlinear solution and the first eigen mode of the RLS are shown. We see an apparent discontinuity in the SVD mode. Because there is a stiffening spring between the 5^{th} and 6^{th} masses, there is less deformation there than in the corresponding mode of the RLS. One should not be surprised that attempting to capture this apparent discontinuity with a sum of modes of the RLS would result in a Gibbs' type phenomenon requiring a very large number of modes.

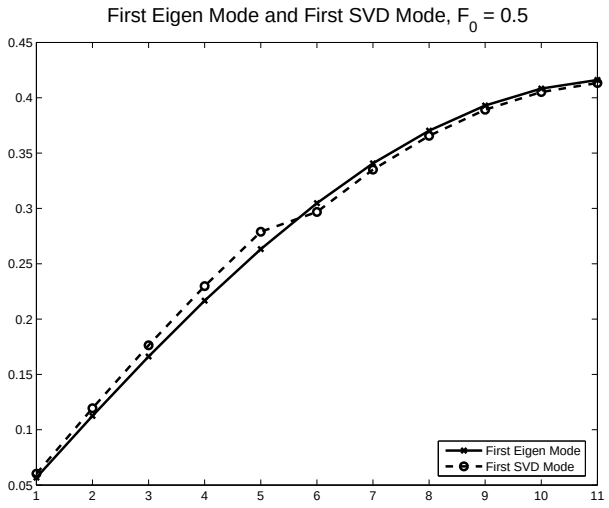

Figure 4.10. The response of a system with a large cubic nonlinearity is explored the the singular value decomposition (SVD) modes of the history of the full nonlinear system. Shown here are the first SVD mode of the fully history and the first eigen mode of the RLS. For this large nonlinearity, the modes show a marked difference in the location of the nonlinear spring. Because there is a stiffening spring between the 5^{th} and 6^{th} masses, the SVD mode shows less deformation at that location than is the case of the linear eigen mode.

4.2.5 Joint Modes

One anticipates that discontinuities such as that illustrated in the SVD mode of the above problem will be characteristic of systems with local stiffness nonlinearities. So one should expect to encounter convergence issues when using approximation by modes of a reference linear system. One well-known approach to accommodating analogous problems in Fourier analysis is to subtract out the discontinuity; that is to augment the basis functions with a simple function sharing the discontinuity of the function to be approximated.

Two candidate classes of basis function were examined for this study. The first is one associated with eigen vector sensitivity analysis and the second is one associated with the static response of the RLS to self-equilibrating loads.

Eigen Vector Sensitivities

The term eigen vector sensitivity can mean the sensitivity of eigen vectors of a matrix system to small perturbations of those matrices or it may refer to sensitivities of the eigenvectors of a mechanical system to physical parameters of that system. The concepts presented here fit into both categories.

Consider a mechanical system with a nonlinear but differentiable connection between degrees of freedom x_{J_1} and x_{J_2} and that the tangent stiffness at zero load of that connection is k_J. As before, the stiffness matrix for that reference linear system is K_0 and the mass matrix is M. We select an eigen mode, V_m^J, of the reference linear system that causes significant deformation at connection J.

Consider also another linear system that differs from the reference linear system only in the stiffness at connection J, where the stiffness is $k_J + \delta k_J$. The m^{th} eigen mode of this perturbed system is $V_m^J + \delta V_m^J$ and the sensitivity of the m^{th} eigen mode with respect to stiffness at the connection is

$$\hat{V}_m^J = \delta V_m^J / \delta k_J \tag{4.10}$$

Because the eigenvectors of the perturbed systems differ from those of the reference system primarily in the displacement across the connection, sensitivity vectors will manifest a discontinuity at the connection location. The reasoning presented above to explain the slow convergence of a Galerkin procedure using only the eigen modes of the RLS would argue that the discontinuity found in these perturbed eigen modes could make them valuable in accelerating the convergence of the Galerkin process. The sensitivity of the first eigen mode of our example system with respect to the stiffness of the connection between the 5^{th} and 6^{th} masses is shown in Figure 4.11.

The strategy proposed above is illustrated in Figure 4.12. Here we see that a Galerkin basis that consists of the first four eigen modes of the RLS and the sensitivity mode presented in Figure 4.11 almost exactly captures the kinetic energy predicted by the full non-

linear solution. In fact, even the use of just one eigen mode along with the sensitivity mode does a pretty good job of predicting the kinetic energy (Figure 4.13).

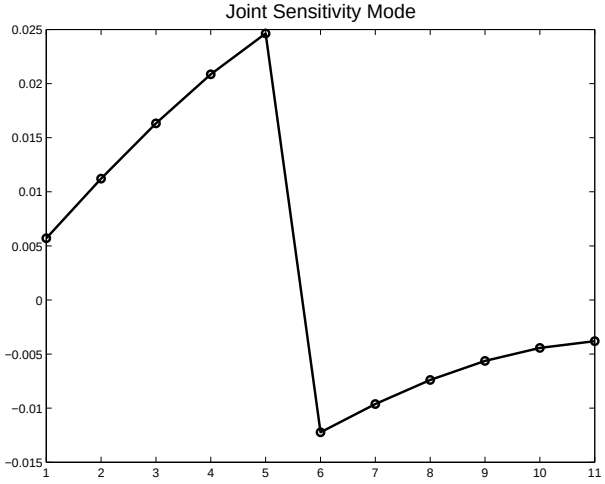

Figure 4.11. The sensitivity of the first eigen mode of the reference linear system with respect to stiffness at the location of the nonlinear spring manifests a discontinuity at that location.

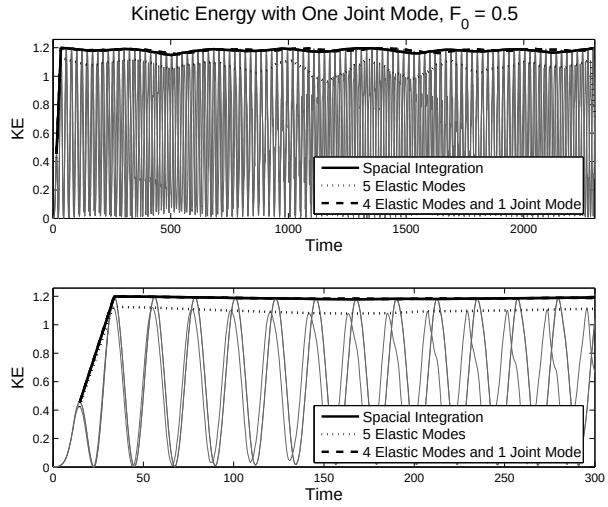

Figure 4.12. Convergence of the Galerkin procedure is greatly enhance when the basis includes an eigen mode sensitivity vector. In this case there are 4 eigen modes of the reference linear system and one eigen mode sensitivity vector.

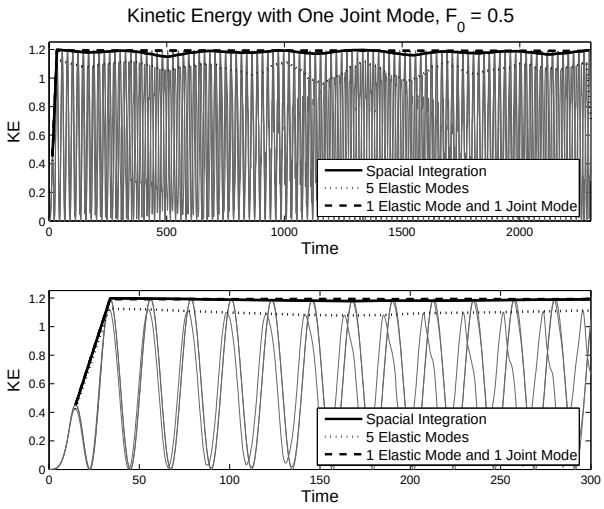

Figure 4.13. Convergence of the Galerkin procedure is greatly enhance when the basis includes an eigen mode sensitivity vector. In this case there are 1 eigen mode of the reference linear system and one eigen mode sensitivity vector.

Milman-Chu Modes

In addressing the optimal selection of dampers for linear systems, Milman and Chu ([11], [31])introduced basis functions obtained by solving the statics problem of self-equilibrating loads acting between the degrees of freedom where the linear damper was intended. A character of these basis functions is that they have a discontinuity at the location of that connection.

Milman and Chu referred to their basis functions as Ritz vectors. Because this term is so general as to be unhelpful in the context of the work reported here, we refer to their basis functions as Milman-Chu vectors.

The Milman-Chu vector for our reference linear system is shown in Figure 4.14, where that anticipated discontinuity is manifest. In Figures 4.15 and 4.16 we see that the Milman-Chu vectors perform almost identically as the eigen mode sensitivity vectors in accelerating the convergence of the Galerkin procedure. A major advantage of the Milman-Chu modes over the eigen mode sensitivities is that they can be calculated much more economically. Since the Milman-Chu (M-C) vectors perform as well as the eigen mode sensitivity vectors, they are used exclusively in the following.

In the calculations presented, the M-C vectors were made orthonormal with respect to the mass matrix to each of the eigen modes employed. This in no way changes the configuration space available to the Galerkin algorithm, but it makes interpretation of the

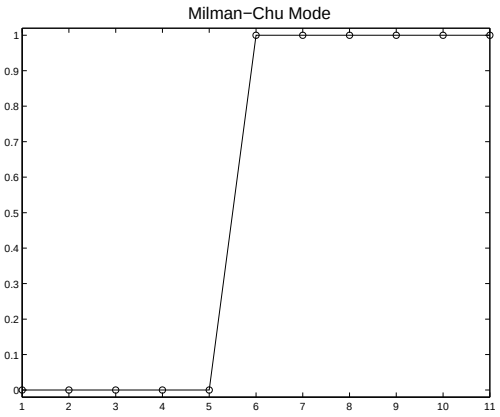

Figure 4.14. The Milman-Chu mode is the solution to a statics problem. It also has the discontinuity that is desired at the location of the local nonlinearity, but it is computed much economically than is the eigen mode sensitivity.

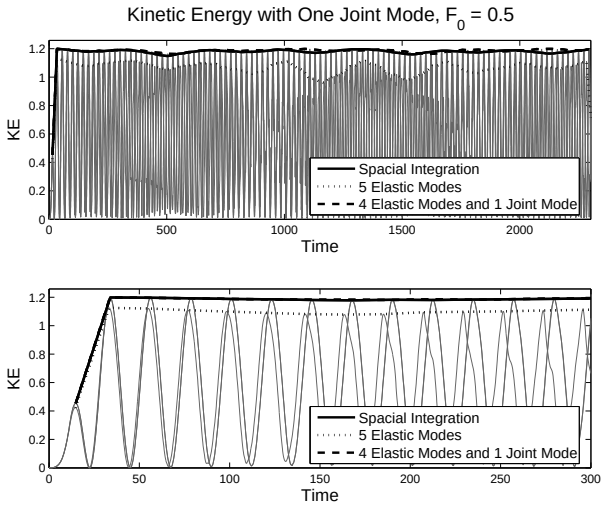

Figure 4.15. Convergence of the Galerkin procedure is greatly enhanced when the basis includes an Milman-Chu vector. In this case there are 4 eigen modes of the reference linear system and one Milman-Chu vector.

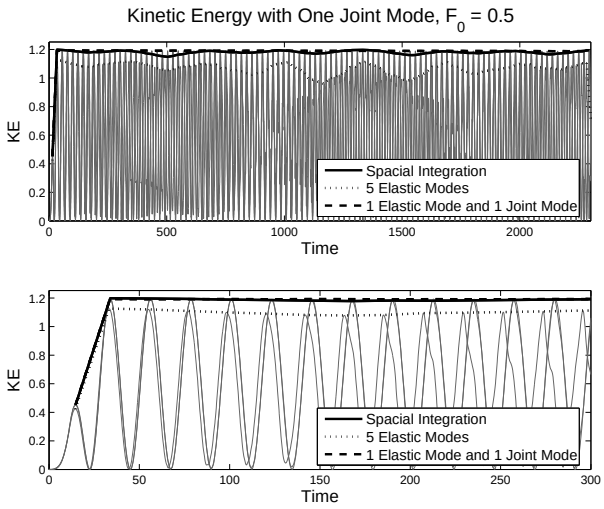

Figure 4.16. Convergence of the Galerkin procedure is greatly enhanced when the basis includes an Milman-Chu vector. In this case there are 4 eigen modes of the reference linear system and one Milman-Chu vector.

calculated generalized coordinates easier.

4.3 Model Reduction for a Structure Containing a Mechanical Joint

The major source of nonlinearity in structural dynamics is the localized frictional slip processes at interfaces in mechanical joints. The two important qualitative properties of mechanical joints - softening and dissipation - are illustrated in Figures 4.17 and 4.18. With respect to the first figure, one sees that under small load, the force-displacement curve appears nearly linear, though there is some amount of micro-slip and dissipation taking place even there. At larger loads, the force-displacement curve begins to level off and at very high loads macro-slip takes place and the tangent stiffness goes to zero. The second figure shows the power-law relationship between the amplitude of oscillatory load and the dissipation per cycle that is commonly seen experimentally over large load ranges. Mechanical joints manifest very little rate dependence.

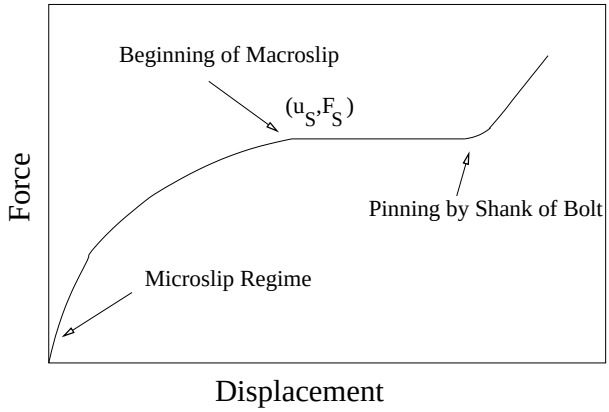

Figure 4.17. Mechanical joints manifest small regions of micro-slip where force-displacement appears linear, though some amount of dissipation accompanies any load. As the load increases, the tangent stiffness decreases until macro-slip initiates.

The usual process of dealing with the presence of joints in structural dynamics is to represent the joint compliances by tunable springs and to represent the joint dissipation by modal damping. The resulting tuned linear models are of course of little value except at the excitation amplitudes at which the structure is calibrated. A discussion of the limitation of such approaches can be found in [41].

A interesting feature of mechanical joints that increases their interest in the world of nonlinear model reduction beyond their practical importance is the intrinsic path dependence to their force-displacement properties. This feature is referred to as non-locality in the sense that the full state of the system cannot be known solely from the current values of kinematic variables and their rates. The nonlocality would appear to proscribe rigorous

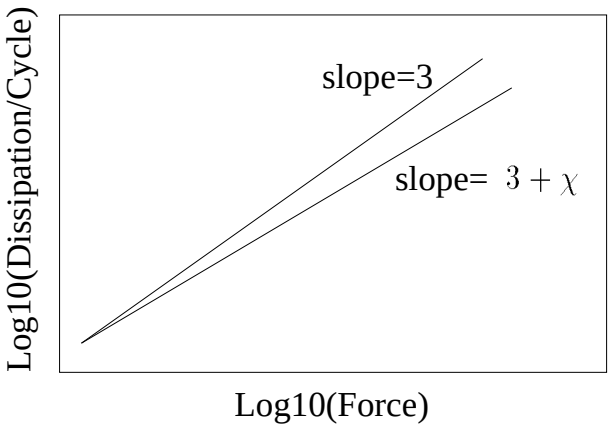

Figure 4.18. When mechanical joints are subject to oscillatory loads, the energy dissipation per cycle appears to increase with load amplitude in a power-law manner. In the above, χ is a number such that $(-1 < \chi \leq 0)$.

application of a number of otherwise powerful mathematical tools - including the use of nonlinear normal modes.

We shall introduce a particular constitutive model for joints so that we may explore the model reduction technique of this report in context of problems of practical importance.

4.3.1 Whole-Joint Approximation

Modeling the complexity of interface mechanics in the midst of structural dynamics calculations would be impractical. A tractable approach involves the introduction of a class of approximation that reduces the complexity of the contact problem to a small number of scalar constitutive equation. This approximation constrains the kinematics of all degrees of freedom on each side of the contact patch to a single kinematic variable. Corresponding kinematic variables on opposite sides of the interface are connected by a single scalar constitutive equation each. Such approximations are called "whole-joint" models. The whole-joint approximation currently employed in Sandia codes imposes multi-point constraints to cause surface nodes on each side of the interface to be constrained rigidly to a centralized node on that surface (Figure 4.19). In the absence of more complete knowledge of joint physics, the constitutive equations for the six relative degrees of freedom are treated as being independent.

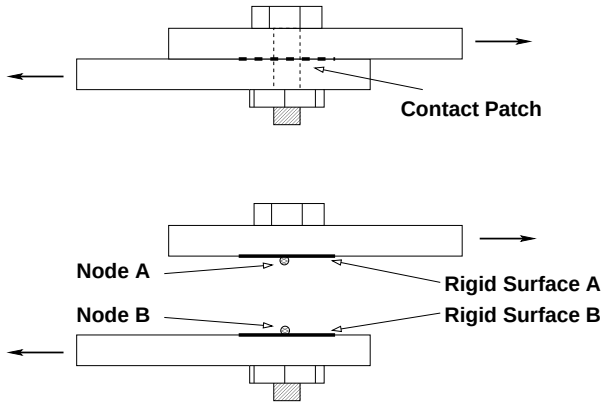

Figure 4.19. The mathematical complexity of the joint is simplified by approximating the whole interface by a single scalar constitutive equation for each of the six relative degrees of freedom. In the illustration shown here all of the nodes on each side of the interface are held rigid and connected to a single joint.

4.3.2 The Four-Parameter Iwan Model

A constitutive equation consistent with the more important qualitative joint behavior observed experimentally is the four-parameter Iwan model discussed in [40] and [42]. This model is an instance of Iwan's parallel-series configuration ([25], [26]) represented graphically in Figure 4.20 showing a continuum of Jenkins elements. All the spring stiffnesses are identical, so the model response is determined entirely by the population density ρ of Jenkins elements of given slider strengths ϕ. The mathematics of such models is discussed in depth in the papers of the above four citations.

The 4-parameter Iwan model is defined as follows:

$$\rho(\phi) = R\phi^{\chi}\left[H(\phi) - H(\phi - \phi_{\max})\right] + S\,\delta(\phi - \phi_{\max}) \tag{4.11}$$

where $H()$ is the Heaviside step function and the the process for finding parameters R, S, χ, and $\phi_{\max}$ is found in [40] and [42] . Values of $-1 < \chi < 0$ results in power-law exponents of $3+\chi$. The general form of this 4-parameter distribution is shown in Figure 4.21. A major deficiency of the above set of parameters is the fractional dimensions of R and S so an alternate and preferred set of parameters (F_S, K_T, χ, and β) for this model have been developed [42]. In the following we use values

$F_S = 1.0$	the force that initiates macro-slip.
$K_T = 1.0$	joint stiffness in the regime of small load.
$\chi = -0.5$	the dimensionless strength of the singularity at zero.
$\beta = 2$	a dimensionless parameter having to do with the shape of the dissipation curve.

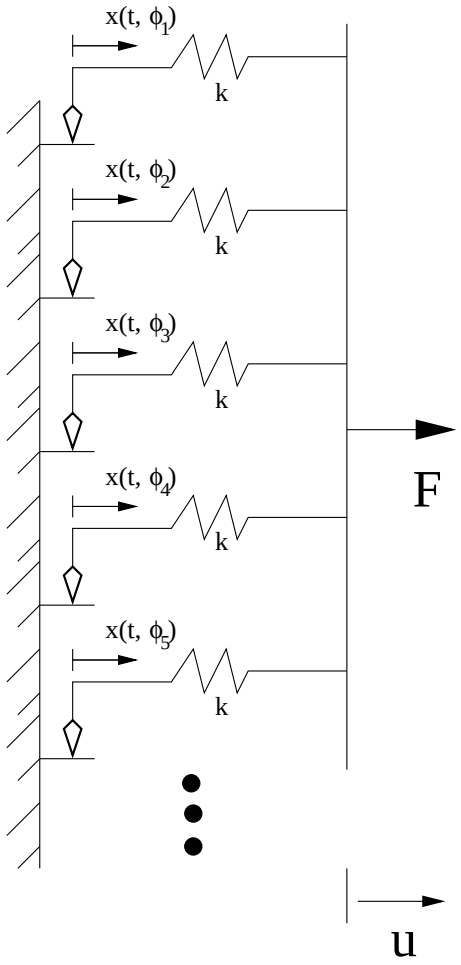

Figure 4.20. The parallel series Iwan model consists of a continuum of Jenkins elements. All the spring stiffnesses are identical, so the model response is determined entirely by the population density of Jenkins elements of given slider strengths.

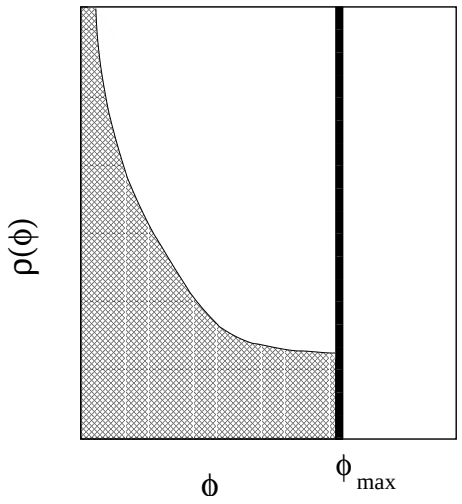

Figure 4.21. The four-parameter Iwan model predicts the correct qualitative behavior of mechanical joints.

4.3.3 Numerical Results for Problems of Micro-Slip

We first examine the results of a numerical experiment associated with the structure shown in Figure 4.3 but where the cubic nonlinearity is replaced with the Iwan model described above. In this case the amplitude of the triangular pulse is $F_0 = 0.5$ - just one half of the break-free force F_S of the joint. Referring to Figure 4.22, we see that a Galerkin solution using the first five eigen modes of the reference linear system does a very poor job of capturing the kinetic energy of this system, but a Galerkin solution using the first three elastic modes and a joint mode (Milman-Chu) performs very well.

We also see that once the excitation is complete, the energy of the system continuously declines because of the hysteretic nature of the joint. This joint damping plays a major role in mitigating the shock that weapons systems can experience in a hostile environment.

It is natural at this point to ask the question: if the joint mode is necessary to capture the mechanics of the system, how does it change the kinematics that would be observed from the transient solution. This question is addressed with reference to Figure 4.23 where we see that the generalized accelerations seem to be dominated by the first mode - as one would expect. One order of magnitude lower are the kinematics of the second mode and the kinematics of the third mode and the joint mode are an order of magnitude yet smaller. Each of the modes is mass normalized, so the contributions to the physical accelerations are roughly proportional to the generalized accelerations.

Recall also that in these simulations the Milman-Chu mode that is employed has been made orthogonal to the elastic modes with respect to the mass and stiffness matrices. The

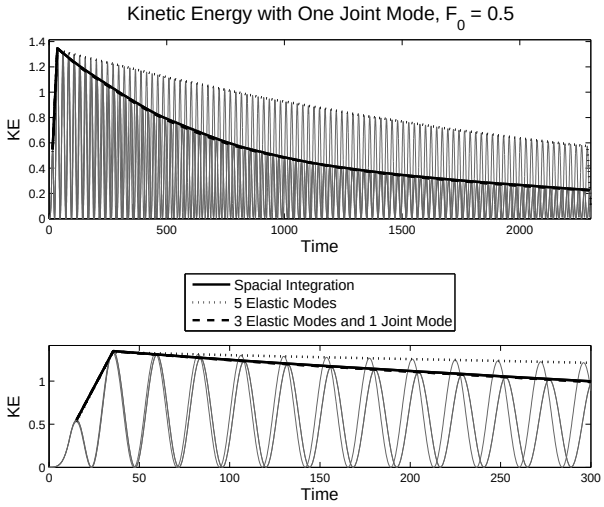

Figure 4.22. Convergence of the Galerkin procedure is greatly enhanced by the presence of Milman-Chu vector in this problem involving the structure shown in Figure 4.3, $F_0 = 0.5$, and a non-linear Iwan joint model. In this case there are 3 eigen modes of the reference linear system and one Milman-Chu vector.

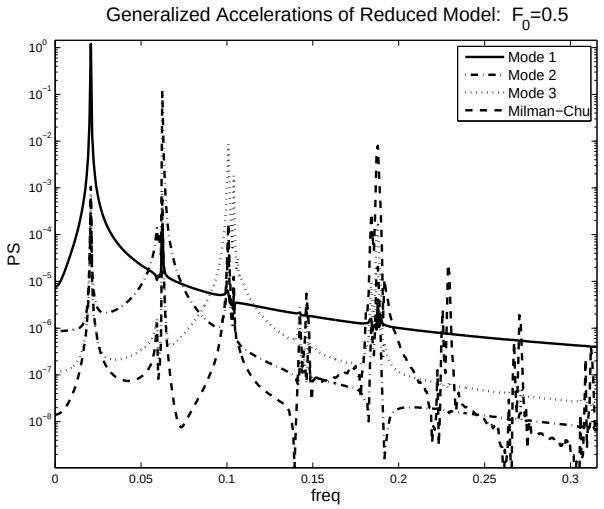

Figure 4.23. Though the presence of the joint mode among the basis vectors of the Galerkin calculation greatly accelerates convergence, the amplitude of the generalized acceleration associated with that vector is actually fairly small in this problem.

coupling evidenced by the peaks of the generalized acceleration associated with the M-C mode occurring at frequencies of the peaks of the generalized accelerations of the elastic modes is purely a nonlinear effect. Because the Milman-Chu mode has been made orthogonal to only the first three elastic modes, it carries some part of the shapes of higher modes and the peaks of the corresponding generalized accelerations at higher frequencies are reflective of modes at those frequencies.

In the following simulations we highlight both the softening and dissipative features of mechanical joints through simulations of base excitation experiments. Again, we consider an eleven mass system with the nonlinear element placed between the fifth and sixth masses. This configuration is illustrated in Figure 4.24.

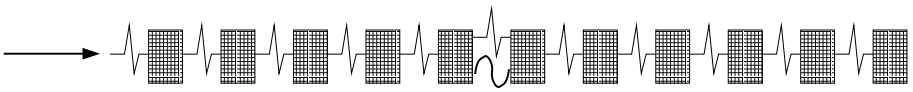

Figure 4.24. An eleven-mass system with a nonlinear joint excited at its base.

In these experiments, we employ an impulse of a sort that is increasingly popular in base excitation experiments - the Morlet wavelet of frequency 4:

$$f(t) = A_0 \cos(\omega 2\pi t/\tau) \exp\left(\frac{(2\pi t/\tau)^2}{2}\right) \tag{4.12}$$

where A_0 is the peak amplitude to be obtained, τ is the period of the frequency to be excited, and ω defines the shape of the wavelet. In the experiments presented here $\omega = 4$ and the characteristic shape is shown in Figure 4.25.

In the first set of numerical experiments the max impulse is set to $F_0 = 0.005$ and Figure 4.26 shows that the resulting joint force calculated from the full spatial solution stays well below the break-free force F_0 of the joint. The corresponding portion of the monotonic force-displacement curve for the joint is shown along with the tangent stiffness at zero load in Figure 4.27. One expects such excitations to cause very little nonlinear response in the joint.

Indeed Figure 4.28 shows that the Galerkin solution employing eigen modes of the reference linear system generates a very good approximation for the kinetic energy of the jointed system subject to a small amplitude impulse. There is so little nonlinearity at the joints that the linear eigen modes do a good job of spanning the configurations taken on by the jointed structure. The presence of the joint is indicated only by the decrease in system energy over time due to dissipation in the joint.

One could also anticipate the adequacy of the eigen modes of the RLS in solving this low amplitude problem by consideration of Figures 4.29 and 4.30 obtained from the full

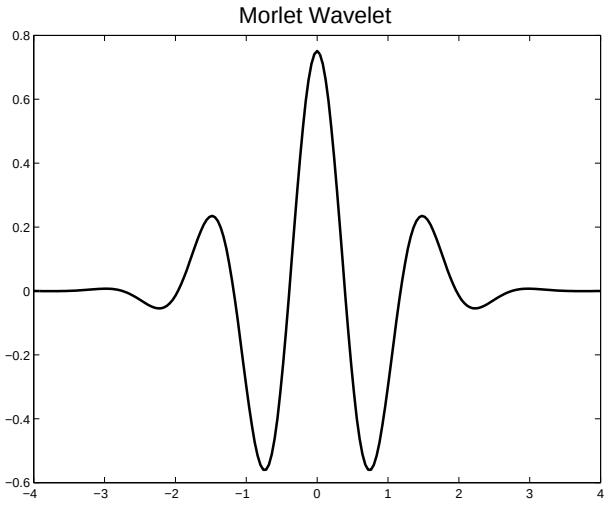

Figure 4.25. The Morlet wavelet with $\omega = 4$.

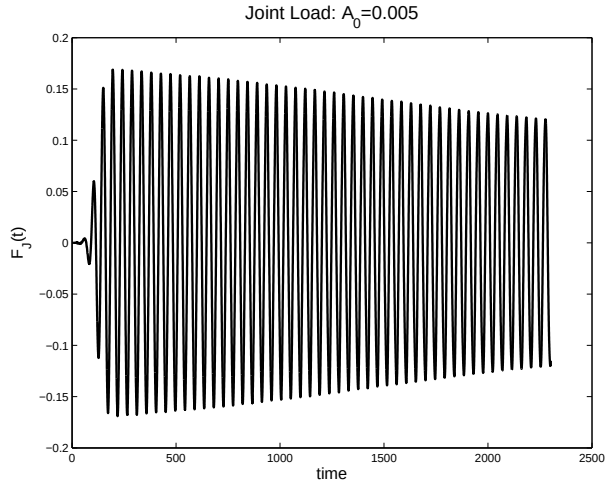

Figure 4.26. The force history of the joint resulting from a very low amplitude ($A_0 = 0.005$) base excitation.

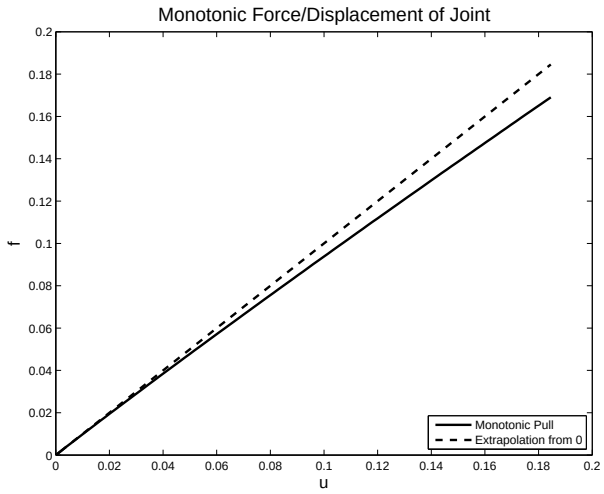

Figure 4.27. The force history of the joint resulting from a very low amplitude ($A_0 = 0.005$) base excitation corresponds to the above portion of the monotonic force-displacement curve for the joint. Also shown is the tangent stiffness at zero load.

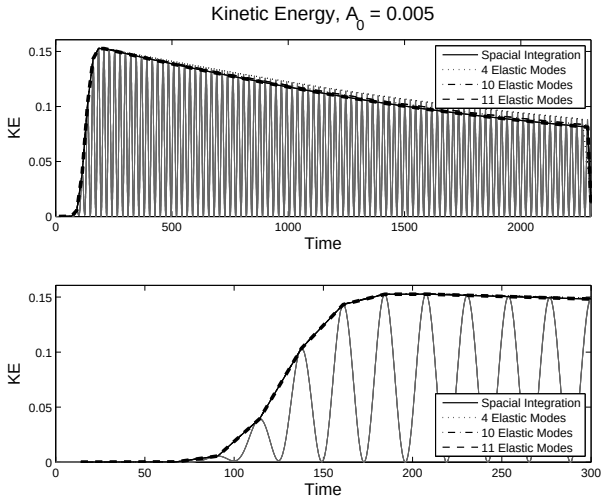

Figure 4.28. The Galerkin solution employing eigen modes of the reference linear system generates a very good approximation for the kinetic energy of the jointed system subject to a small amplitude impulse ($A_0 = 0.005$).

nonlinear spatial solution. The first shows that the nonlinear system response is limited to resonance of just the first natural frequency of the reference linear system and the second shows that the first SVD mode of the numerical solution is almost identical to the first eigen mode of the RLS.

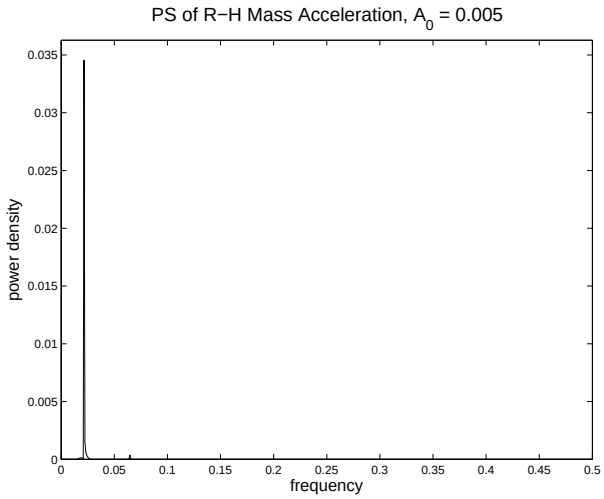

Figure 4.29. When subject to very small amplitude excitation ($A_0 = 0.005$), the system responds with a nearly monochromatic response at the frequency of excitation - which was tuned to the first natural frequency of the reference linear system.

Interestingly Figure 4.31 shows that even for the nearly linear case discussed in this section, the use of a single Milman-Chu joint mode greatly increases convergence of the reduced order model to the solution of the full system.

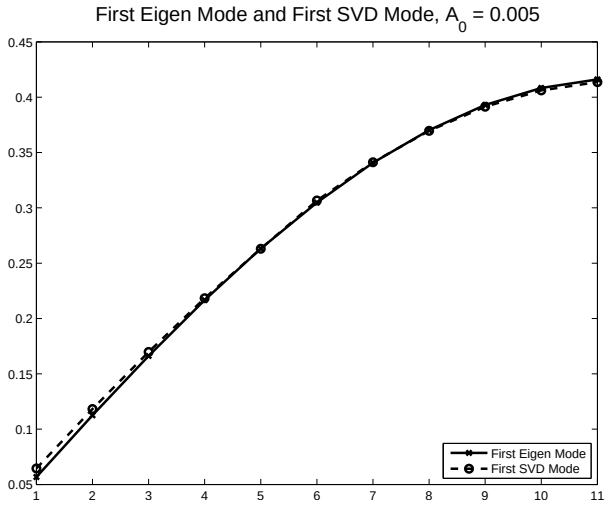

Figure 4.30. The SVD of the full nonlinear spacial solution and the first eigen-mode of the reference linear system are nearly identical when the system is subject to a very low amplitude ($A_0 = 0.005$) base excitation.

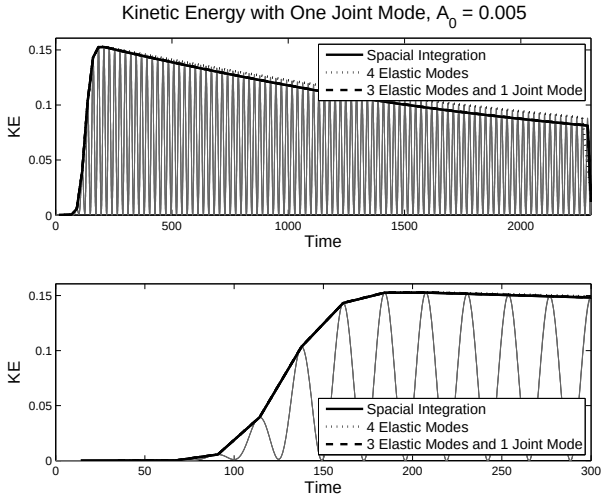

Figure 4.31. Even when the system is subject to a very low amplitude ($A_0 = 0.005$) base excitation, the use of a Milman-Chu joint mode makes a noticeable improvement in convergence.

When the amplitude of the base excitation is raised to $A_0 = 0.02$, the elastic modes are a much less satisfactory basis of modeling the more nonlinear system. Figure 4.32 shows that three elastic modes augmented with a Milman-Chu mode provide a much better basis for modeling vibration ring-down in this problem than are six elastic modes. Though not shown here, the peak joint force encountered in this simulation is approximately 70% the break-free force F_S.

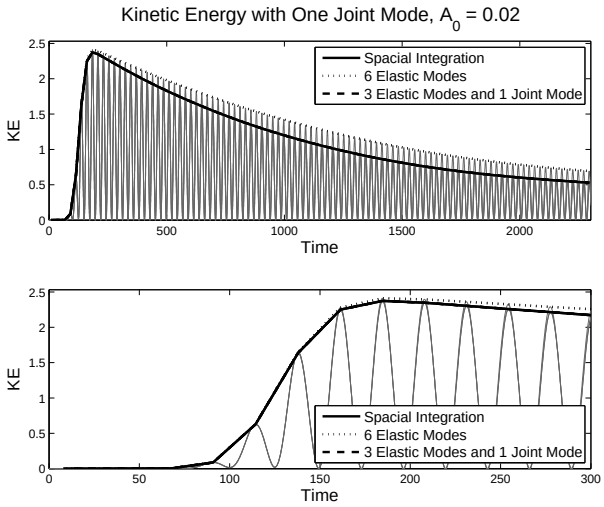

Figure 4.32. At a higher level of base excitation ($A_0 = 0.02$), the use of a Milman-Chu joint mode makes a more noticeable improvement in convergence.

The generalized accelerations shown in Figure 4.33 show behavior similar to that presented in Figure 4.23. Again, we see that though the joint mode is necessary for capturing the correct mechanics in this problem, it mode does not make a strong appearance in the structural kinematics.

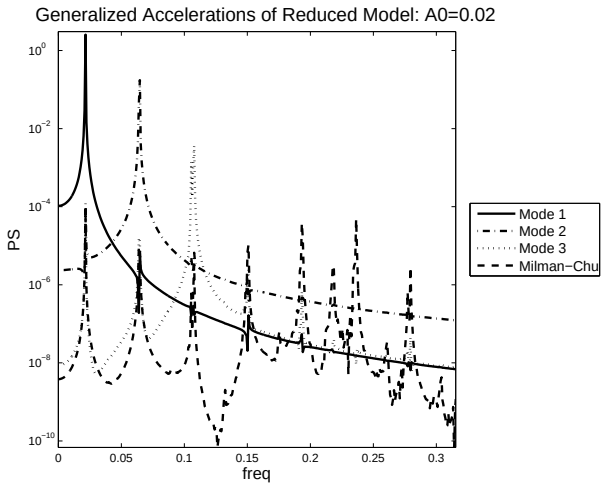

Figure 4.33. As was the case in the resonance calculation of Figure 4.23, though the presence of the joint mode among the basis vectors of the Galerkin calculation greatly accelerates convergence, the amplitude of the generalized acceleration associated with that vector is actually fairly small in this base excitation problem ($A_0 = 0.02$).

4.3.4 Numerical Results for Problems of Macro-Slip

A Galerkin solution using just the eigen modes of the reference linear system is dramatically less successful for problems where applied loads approach or exceed the break-free force F_0 of the joint.

In the cases considered here, the amplitude of input wavelet is 0.05 and the resulting joint load history predicted by the full nonlinear spatial solution is shown in Figure 4.34. We see here that the joint is brought into macro-slip and peak force levels in the joint are saturated at F_S until enough energy has dissipated that the system loads on the joint drop to lower levels. This force history on the joint corresponds to the monotonic force-displacement curve shown in Figure 4.35, where the strong nonlinearity is illustrated by its contrast to the zero-load tangent stiffness curve.

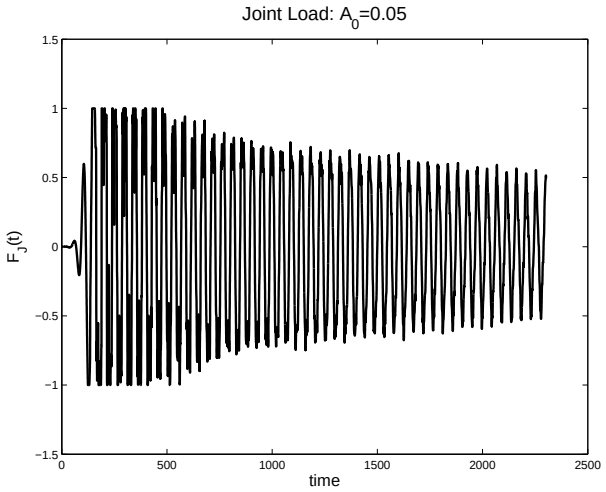

Figure 4.34. When the system is subject to a high amplitude ($A_0 = 0.05$) base excitation, the joint is brought into macro-slip and force levels in the joint are saturated at F_S.

The nonlinearity that the joint lends to the system dynamics is suggested by the plots of the first SVD mode of the full nonlinear spatial solution and the first eigen mode of the reference linear system in Figure 4.36. Note that these curves are quite different in the vicinity of the joint.

Another indication of the strong nonlinearity of this system is shown in Figure 4.37. Here we see that the acceleration of the right hand mass of the system contains not only components at the frequency of the excitation (which was tuned to the first natural frequency of the reference linear system), but also many higher frequency components.

Given the above, it should be no surprise that Galerkin solution using just the eigen

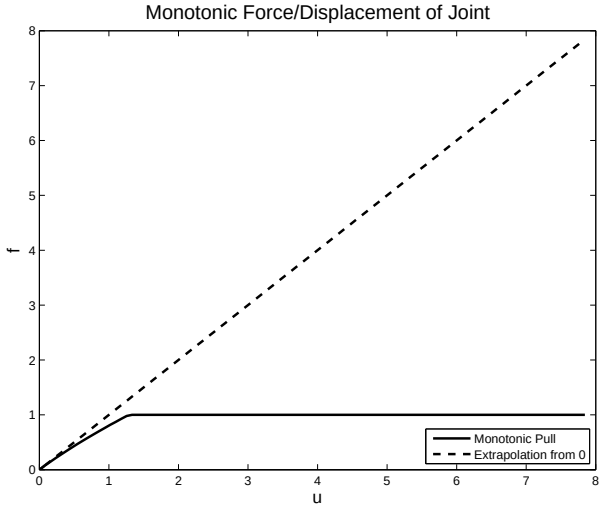

Figure 4.35. The force history of the joint resulting from a high amplitude ($A_0 = 0.05$) base excitation corresponds to the above portion of the monotonic force-displacement curve for the joint. Also shown is the tangent stiffness at zero load. The nonlinearity manifest at these force levels is large.

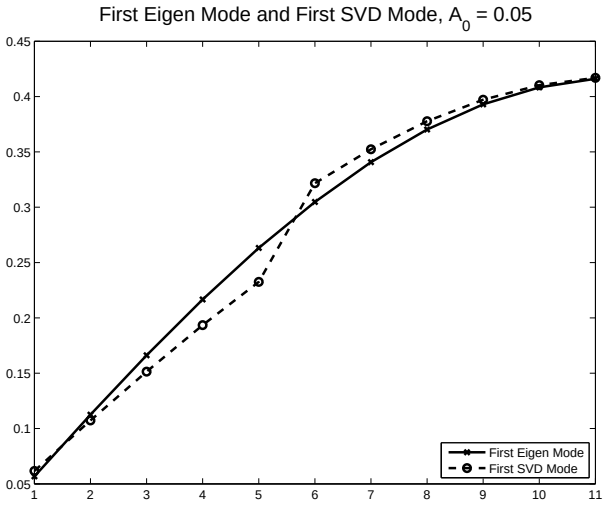

Figure 4.36. The first SVD mode of the full nonlinear spacial solution and the first eigen mode of the reference linear system are quite different in the vicinity of the joint when the system is subject to a high amplitude ($A_0 = 0.05$ base excitation.

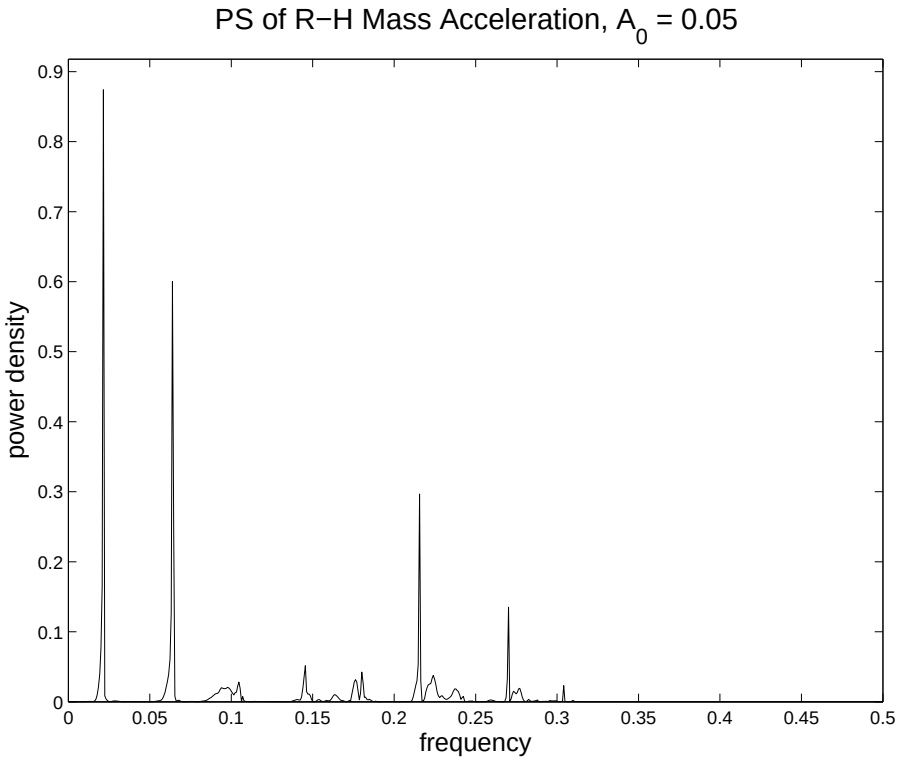

Figure 4.37. Macro-slip causes frequency responses of the structure that well above that of the base excitation - which was tuned to the first resonance of the reference linear system.

modes of the reference linear system demonstrates very poor performance. In Figure 4.38 we see that even with ten elastic modes, the kinetic energy is approximated very poorly despite the fact that those modes correspond to much higher frequencies in the linear system than are indicated in Figure 4.37. The transitions to macro-slip in this problem appear to be responsible for the transfer of energy from the low excitation frequency to much higher frequencies. As expected, when sufficient modes to reach the frequency response of the corresponding linear system are augmented by a joint mode, the system is modeled much better (Figure 4.39). Figure 4.40 shows that the augmented basis set captures the correct character of the magnitude of the Fourier transform of acceleration of the right-most mass, while the approximate solution that employs only the elastic eigen modes leaves too much energy at the lower frequencies.

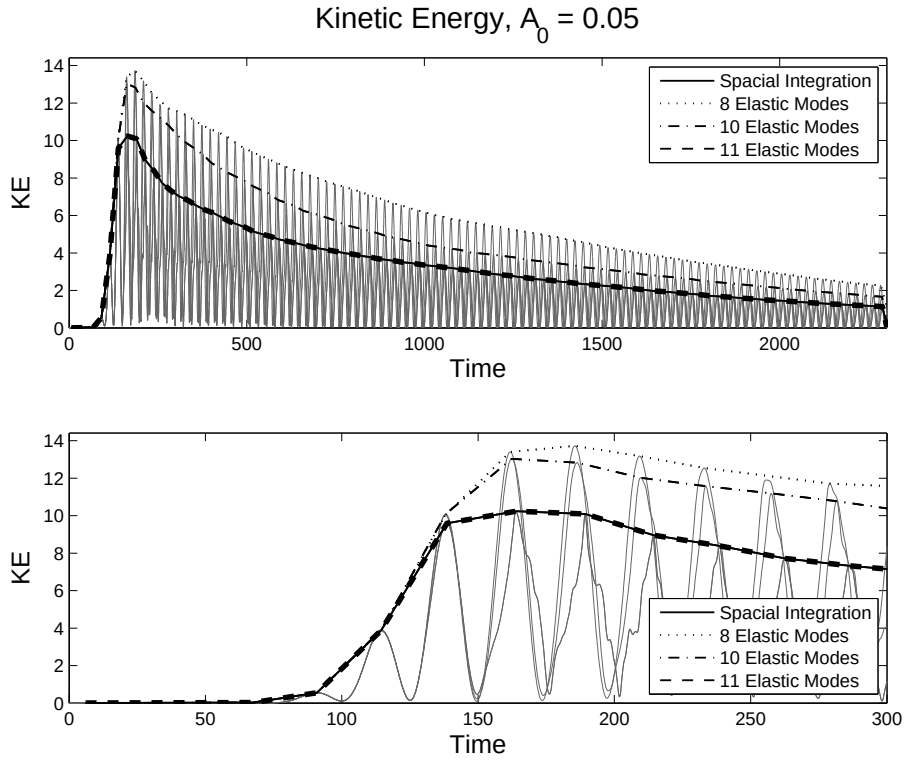

Figure 4.38. The Galerkin solution employing eigen modes of the reference linear system generates a very poor approximation for the kinetic energy of the jointed system subject to a large amplitude impulse.

That the displacement across the joint in macro slip can be an appreciable part of the overall kinematics is evidenced in Figure 4.41 where the generalized acceleration associated with the joint mode is comparable with that of the first linear vibration mode.

An interesting result is found when a “ruthlessly reduced” model is employed. When

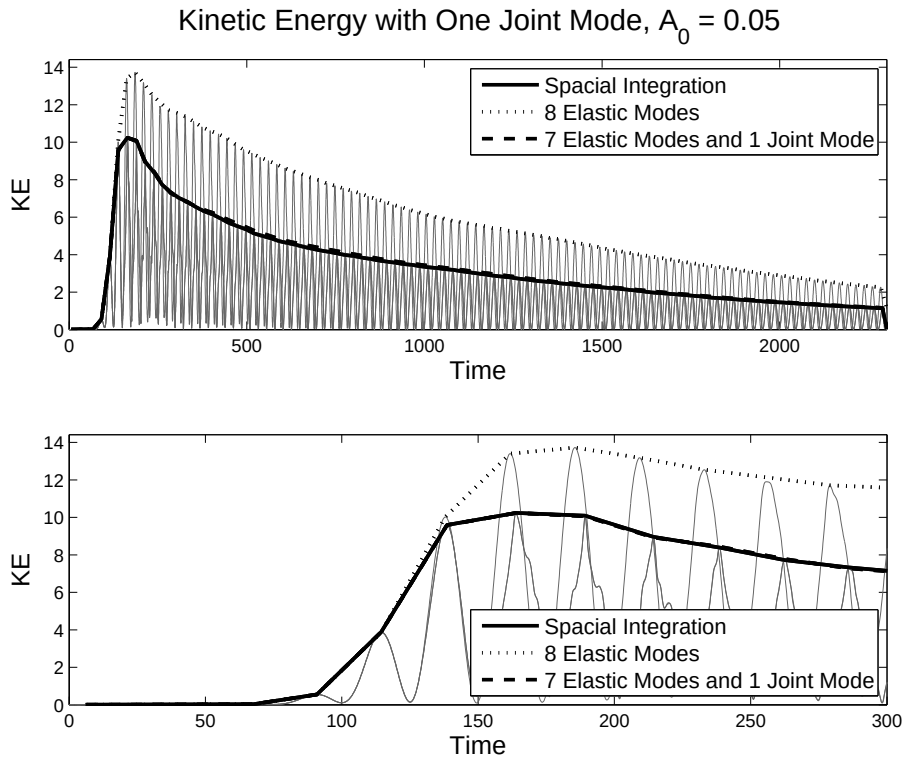

Figure 4.39. The Galerkin solution employing seven eigen modes of the reference linear system augmented by one joint mode generates approximation for the kinetic energy of the jointed system subject to a large amplitude impulse. A large number of elastic modes are necessary to capture the high frequency response of the systems. The necessity of including the joint mode is illustrated by comparison to the prediction resulting from use of eight elastic modes.

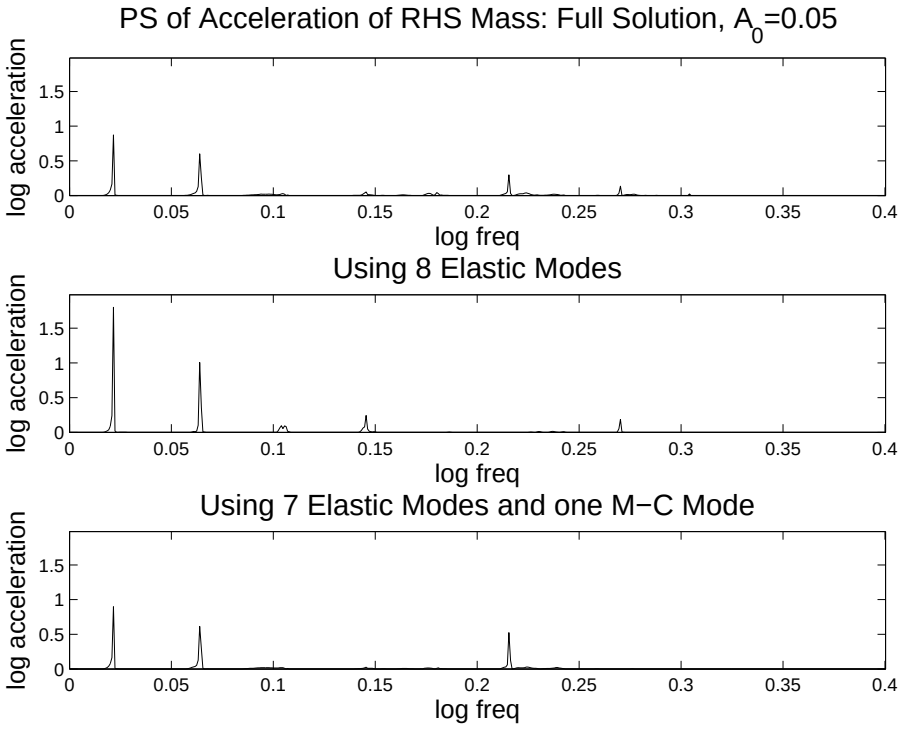

Figure 4.40. Comparison of the acceleration power spectra for the right most mass for the full spacial solution and the two reduced order solutions illustrates how resolution of joint kinematics is necessary to capture the energy shift from low frequencies to high.

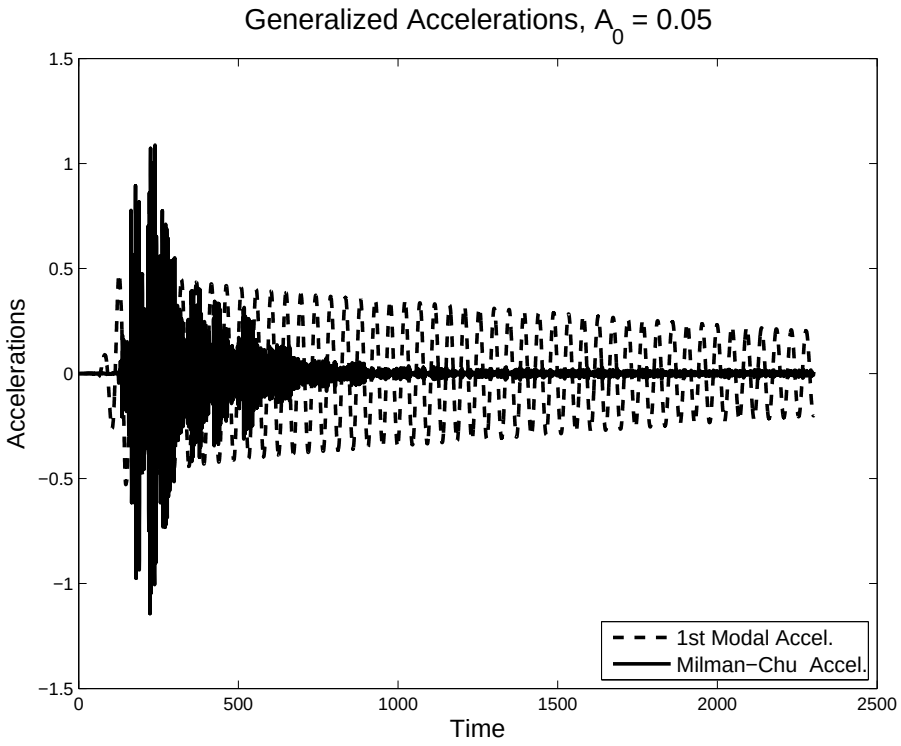

Figure 4.41. In this problem of macro-slip the generalized acceleration of the joint coordinate is no longer small.

this large amplitude experiment, resulting in high frequency components in the structural response, is approximated by three elastic eigen modes and one joint mode, the kinetic energy predicted (Figure 4.42) is in noticeable error. However, the error is substantially less than that of the Galerkin solution where eight elastic eigen modes but no joint mode are employed.

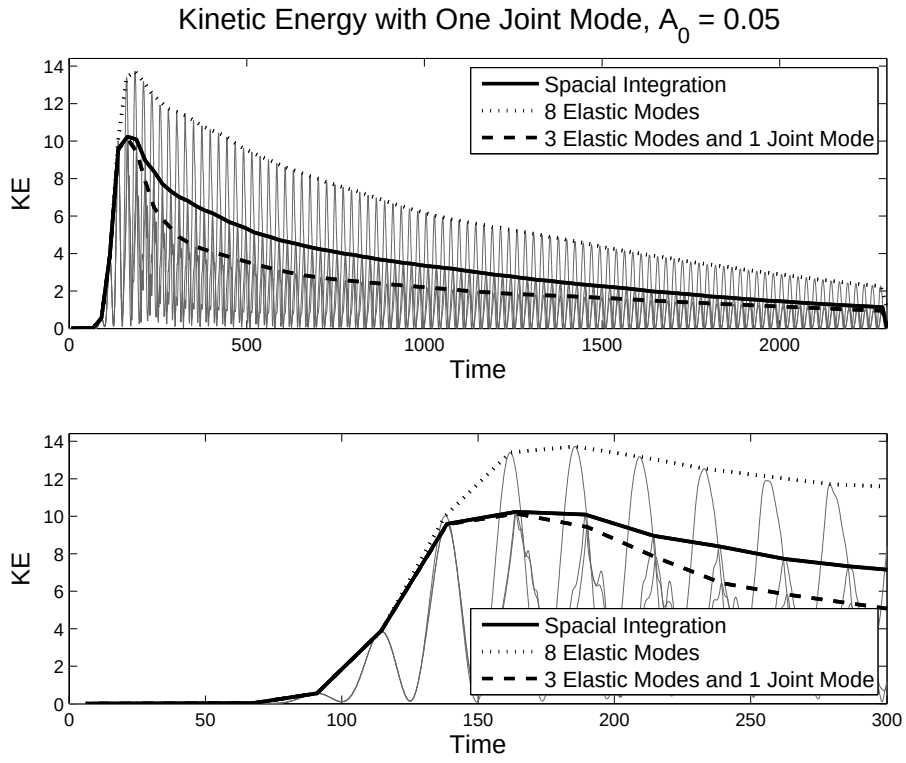

Figure 4.42. A "ruthlessly reduced" analysis using only three elastic eigen modes and one joint mode results in noticeable error in the kinetic energy, but substantially less error than an analysis using twice the number of elastic eigen modes and no joint mode.

Particularly intriguing is the predicted acceleration of the right-most mass. In Figure 4.43 the accelerations predicted by the very reduced model appear as though they were the full spatial solution as seen through a low-pass filter. That hypothesis is tested by comparing the full solution and the very reduced solution when both are sent through a low-pass filter. (Sixth order Butterworth filter with cut-off frequency 0.03). The results shown in Figure 4.44 do argue that for the special case of these Iwan joint models, a low-order model for the full non-linear structure seems to capture the low frequency response of the structure reasonably well. Why this reduced order model works so well for these joint models is not entirely clear at this time, though one would have every reason to believe that such fortuitous results would not occur for a rate-type nonlinearity.

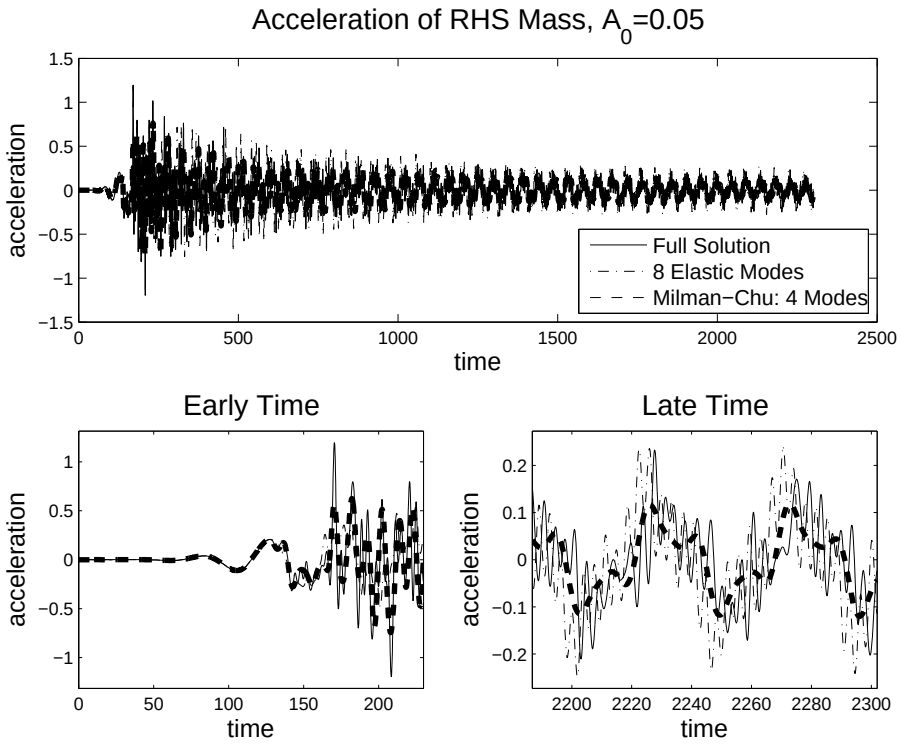

Figure 4.43. A “ruthlessly reduced” analysis using only three elastic eigen modes and one joint mode results results in accelerations of the right most mass that have the appearance of a low-pass filter of the full spacial solution.

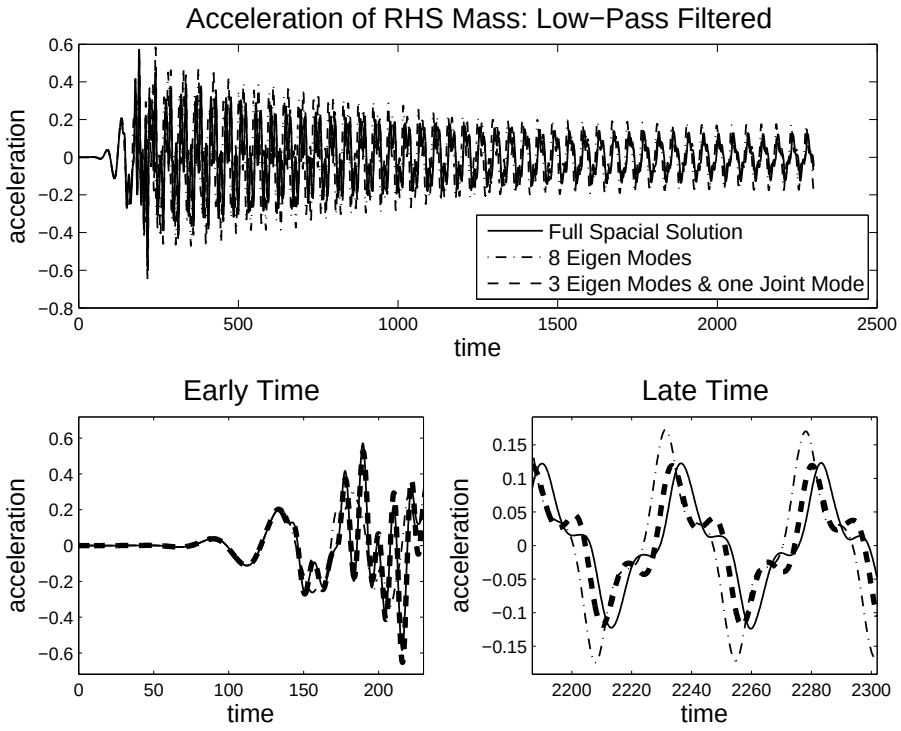

Figure 4.44. When the "ruthlessly reduced" analysis using only three elastic eigen modes and one joint mode and the full spacial solution are seen through a low pass filter, they appear very similar.

4.4 Implementation in the Context of Finite Element Analysis

The tools presented in this report are intended to facilitate predictive structural dynamic simulation, and that means integration into finite element analysis. Most of the necessary components for large scale analyzes of jointed structures are available in standard finite element packages, including eigen analysis and elastic static analysis. Additionally, the techniques presented here seem to be complementary to other model reduction methods - component mode synthesis especially.

4.4.1 Automatic Determination of Special Vectors and Matrices

The only quantities whose construction is not obvious are the vector F_j aligned along the joint j, the tangent stiffness matrix $K_N(\{s_j\})$, and the Milman-Chu vectors $y_{MC,j}$.

Algebraic Vector F_j

In the simplest case, when the joint is aligned in a principle direction, the vectors F_j are constructed by putting a 1 in the entry associated with the degree of freedom of the first joint node and the joint direction and a -1 in the entry associated with the degree of freedom of the second joint node and that joint direction.

On the other hand, the task is more difficult when the joint is aligned in a local coordinate system and here we discuss a formal strategy for deducing F_j in such general circumstances. Recall that K_0 is the stiffness of the reference linear system, where each joint j is represented by a properly oriented spring of stiffness k_j. Let's define K_j to be the corresponding stiffness matrix when the equivalent spring for joint j is replaced by a spring of stiffness $k_j + \delta k_j$. The difference matrix $\delta K_j = K_j - K_0$ will have nonzero entries only for degrees of freedom associated with the nodes associated with that joint. For an extensional joint (or a torsional joint), those entries map to a six by six matrix $K_{6,j}$ which has only one nonzero eigen value. Let $\hat{F}_j$ be the corresponding eigen vector, normalized so that $\hat{F}_j^T \hat{F}_j = 2$ and let F_j be that vector mapped back to the full system.

Gradient Matrix K_N

In solving the nonlinear dynamic equations numerically one often employs methods such as Newton iteration which require taking the gradient of all terms in the governing equation with respect to all the kinematic variables. The relevant gradient of the joint terms are neatly merged to those of the stiffness matrix of the reference linear system in the

following manner

$$K_N = K_0 + \sum_{\text{joints } j} G_j \left(\frac{d f_j(s)}{ds} - k_j \right) \tag{4.13}$$

where

$$G_j = F_j F_j^T \tag{4.14}$$

Milman-Chu Algebraic Vector $y_{MC,j}$

The Milman-Chu vector associated with joint j is easily constructed by performing a static analysis associated with applying equal and opposite loads on the joint nodes in the direction of the joint alignment while applying no loads at other joints or external boundaries.

4.4.2 How Many Modes?

The number of elastic modes and Milman-Chu modes necessary for application to a particular problem can be estimated in a manner similar to that employed in modal truncation of linear systems. In the simplest implementation, one employs all elastic modes corresponding to frequencies below an appropriately chosen cut-off frequency and one uses a Milman-Chu mode for each joint degree of freedom.

4.5 Employment in Conjunction with Component Mode Synthesis

The model reduction method presented here addresses difficulties particularly associated with local nonlinearities. It is consistent with other model reduction methods - the method of component mode synthesis in particular.

Consider a structure $\mathscr{B}$ consisting of a number of substructures $\mathscr{B}_k$ with joint models connecting some of the interface degrees of freedom. The kinematics of each substructure is characterized by the values of interface degrees of freedom $\{u_{k,n}\}$ and modal degrees of freedom $\{\phi_{k,n}\}$. The development of the reduced order model proceeds much as discussed earlier in this report:

- Eigen analysis is performed on the linearized component mode representation for $\mathscr{B}$.
- The Milman-Chu vector is calculated by placing self equilibrating loads on nodes on the interface between substructures and performing a system level statics solution.
- The numerical results are in terms of vectors whose support is the whole structure.

4.6 Nonlinear Normal Modes

Because Nonlinear Normal Modes (NNMs) have proven helpful in understanding the dynamics of many nonlinear systems, it is helpful to discuss the approach presented here in terms of NNMs in the hope of expanding the utility of NNMs in structural dynamics.

In the sense of Preisach et al. [37] the existence of a nonlinear normal mode is equivalent to the assertion that the deformation field at any time can be expressed

$$u(t) = A(t)\,y_1 + \sum_{n=2}^{N} P_n(A,\dot{A})\,y_n \tag{4.15}$$

where vectors $\{y_n\}$ are a displacement basis for the structure, A is a periodic function of time, and the coefficient functions P_n are characteristic of the system. Usually, all basis vectors are chosen to be the eigen modes of the reference linear system.

The similarity of equations 4.3 and 4.15 and the numerical calculations of the previous section permit us to make the following assertion: *For structures containing localized nonlinearities, unless the set of basis functions is selected to include some with the appropriate discontinuities, expansions such as the above cannot converge to the true solution.*

On the other hand, the analysis technique explored in the previous chapter does employ basis vectors with the appropriate discontinuity so it might be profitable to see if that technique yields solutions containing any of the character of nonlinear normal modes. We consider "ruthlessly-reduced" cases using as basis vectors just the first eigen mode of the reference linear system and a Milman-Chu vector. We perform simulations for $F_0 = 0.05$, $F_0 = 0.1$, and $F_0 = 0.5$.

The nonlinear normal mode expression corresponding to our two-basis element Galerkin formulation is

$$h(t) = A(t)y + f(A(t))w \tag{4.16}$$

In the context of Equation 4.3, $a_1 = A(t)$ and the assertion of this being a nonlinear normal mode would be

$$a_2(t) = f(A,\dot{A}) \tag{4.17}$$

We examine the results of our transient numerical calculation to assess if Eq. 4.17 is satisfied. Figure 4.45 shows plots of $a_2(t)$ vs $a_1(t)$ for each of our three cases. These plots are obscured by gray clouds of higher frequency components. When the phase space $(a_1, \dot{a}_1)$ is distributed into bins (a_1 x $\dot{a}_1$ into 15x5 bins) and the values of a_2 in each bin in averaged, the points shown in dark squares results. The lack of scatter in these dots indicates the absence of velocity dependence - as one would expect for a perfectly elastic system.

The averaged results of the previous figure are plotted together in Figure 4.46 and indeed the data are consistent with an assertion that a_2 is some function of a_1. This is the the assertion of the manifestation of a nonlinear normal mode.

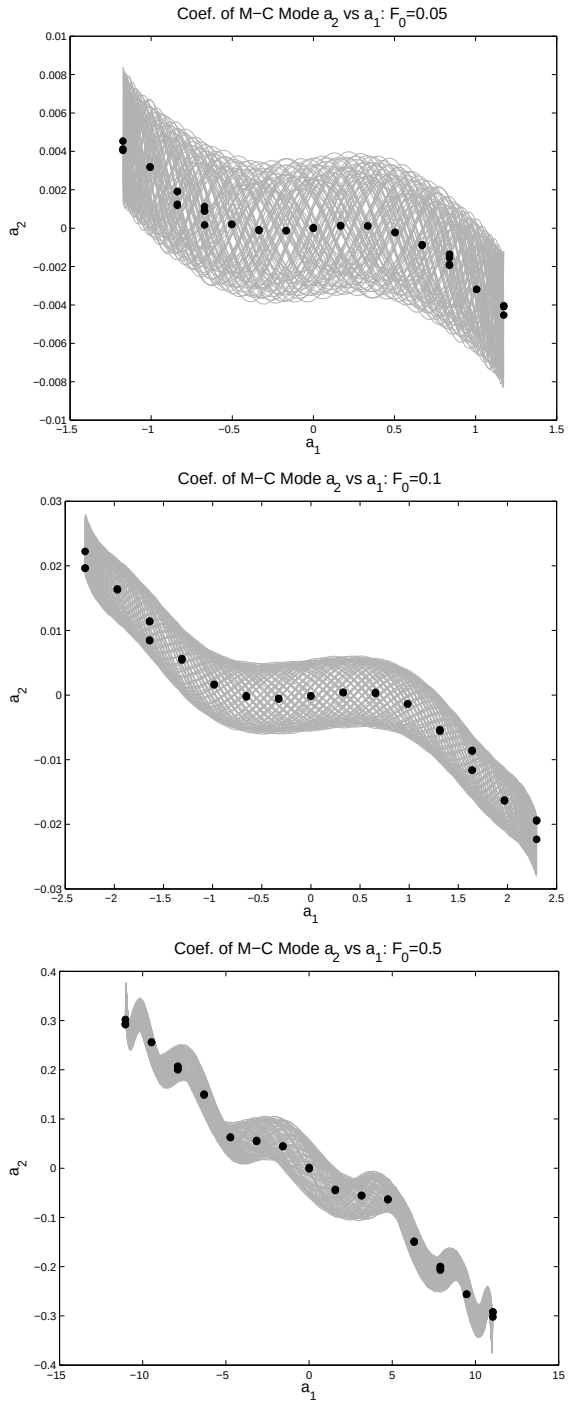

Figure 4.45. When the coefficient for the second generalized coordinate (Milman-Chu) is plotted against the first, clouds (gray) associated with higher frequency result. When that the phase space $(a_1, \dot{a}_1)$ is distributed into bins and the values of a_2 in each bin are averaged, the points shown in dark squares results. The lack of scatter in these dots indicates the absence of velocity dependence - as it should.

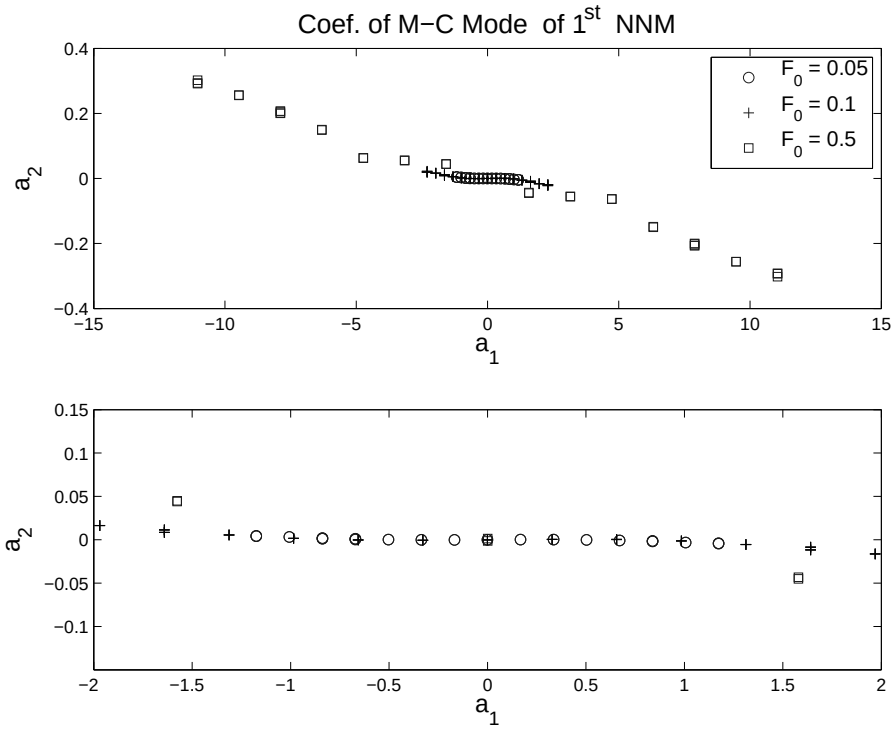

Figure 4.46. When averaged mapping of the coefficient for the second generalized coordinate (Milman-Chu) against the first we see a pattern suggestive of the existence of a nonlinear normal mode.

Let's now see how consistent these results are with a simple nonlinear normal mode calculation. Recall that the joint mode (Milman-Chu in this case) is made orthonormal with respect to the mass matrix to the linear eigen modes that employed. Say that $y = y_1$ is the first eigen mode of the RLS and w is the M-C mode made orthonormal to y with respect to the mass matrix:

$$y^T M y = 1 \quad y^T M w = 0 \quad w^T M w = 1 \tag{4.18}$$

Contraction of these vectors with the stiffness matrix yields the Rayleigh quotients

$$y^T K y = \omega_0^2 \quad w^T K w = \hat{\omega}^2 \tag{4.19}$$

where ω_0 is the natural frequency of the first eigen mode of the RLS and $\hat{\omega}$ is a number greater than or equal to the second natural frequency of the RLS. (This is the minimax principle. [30])

We do not know much about the evolution of this NNM, but we can assert that the system is conservative so that the kinetic energy when the system velocities are maximum equals the strain energy when the system displacements are greatest. The maximum kinetic energy is

$$\begin{aligned} \mathrm{KE}_{\max} &= \tfrac{1}{2}\dot{A}_{\mathrm{m}}^2 \left(y + f'(A)w\right)^T M \left(y + f'(A)w\right) && (4.20) \\ &\approx \tfrac{1}{2}\omega_0^2 A_{\mathrm{m}}^2 \left(1 + f'(A)^2\right) && (4.21) \end{aligned}$$

where A_{m} is the maximum value taken on by $A(t)$ during the cycle, $\dot{A}_{\mathrm{m}}$ is the maximum value taken on by $\dot{A}(t)$ during the cycle, $f'(A) = df(A)/dA$, and we have assumed that $\dot{A}_{\mathrm{m}} = \omega_0 A_{\mathrm{m}}$.

The maximum strain energy is

$$\begin{aligned} \mathrm{SE}_{\max} = \frac{1}{2}&\left(A_{\mathrm{m}} y + f(A_{\mathrm{m}})w\right)^T K \left(A_{\mathrm{m}} y + f(A_{\mathrm{m}})w\right) \\ &+ N\left(A_{\mathrm{m}}\,\delta y^* + f(A_{\mathrm{m}})\,\delta w^*\right) \end{aligned} \tag{4.22}$$

where N is the nonlinear part of the strain energy at the joint. In the case of our cubic spring,

$$N(\delta u) = K_2 \delta u^4/4 \tag{4.23}$$

The maximum strain energy is now

$$\mathrm{SE}_{\max} = \frac{1}{2}\omega_0^2 A_{\mathrm{m}}^2 + \frac{1}{2}\hat{\omega}^2 f(A_{\mathrm{m}})^2 + N\left(A_{\mathrm{m}}\,\delta y^* + f(A_{\mathrm{m}})\,\delta w^*\right) \tag{4.24}$$

where δy^* and δw^* are the displacements across the nonlinear joint of y and w respectively and N is the strain energy associated with the essential nonlinearity. When we equate the two energies (Equations 4.21 and 4.24, we obtain an equation for f'

$$\omega_0^2 A_{\mathrm{m}}^2 f'(A)^2 w = \hat{\omega}^2 f(A_{\mathrm{m}})^2 + 2N\left(A_{\mathrm{m}}\delta y^* + f(A_{\mathrm{m}})\delta w^*\right) \tag{4.25}$$

which can be solved numerically for $f(A)$. These results are shown in Figure 4.47 along with the data shown previously in Figure 4.46. We see a strong similarity between the NNM prediction and the values deduced from post-processing of simulations and we also see systematic differences that can be attributed to the ambitious effort to represent the system dynamics with just two basis vectors.

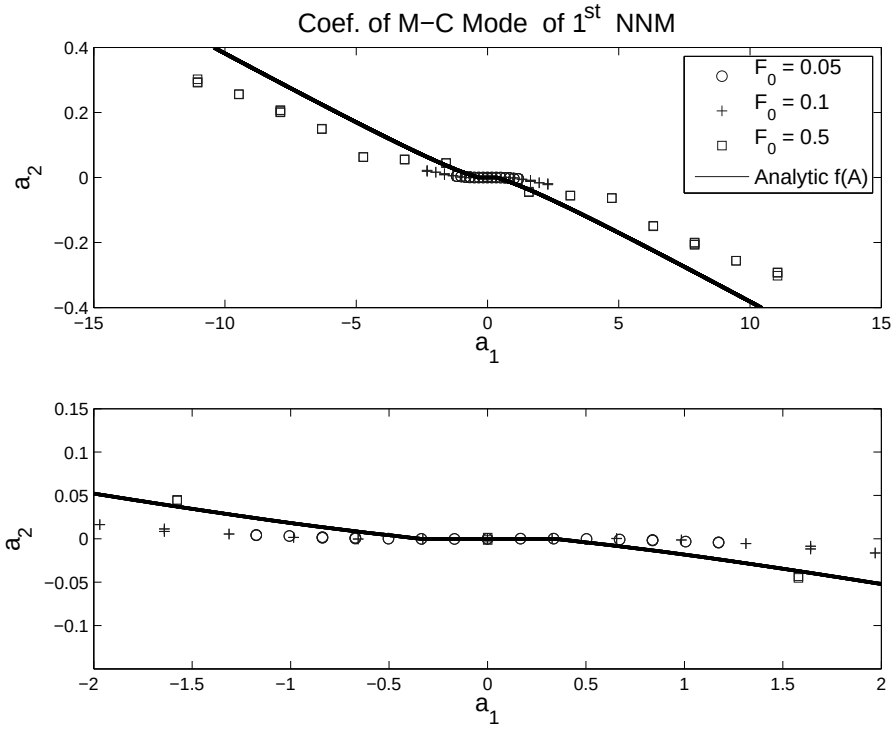

Figure 4.47. When one assumes that a nonlinear normal mode exists and can be represented by the first eigenmode and a Milman-Chu mode, energy methods permit the estimation of the dependence of the second generalized coordinate as a function of the first.

4.7 Implementation in Finite Element Analysis

In a series of calculations reported in [20] Griffith and Segalman performed finite element analysis on two structures to explore the utility of the method of discontinuous basis functions in true computational problems.

The object in Figure 4.48, containing two Iwan joints was subject to a uniform traction in the y direction on the free side of the structure as noted in the figure modulated by a triangular pulse of 1e-4 second duration. Each joint is capable of deformation in only the indicated x direction. This system contains 722 nodes or 2166 total degrees of freedom; however, due to boundary and MPC constraints the model possesses only 1803 active degrees of freedom. We consider the analysis of this 1803 degree of freedom model to be the full order system. Geometry and material parameters are presented in [20].

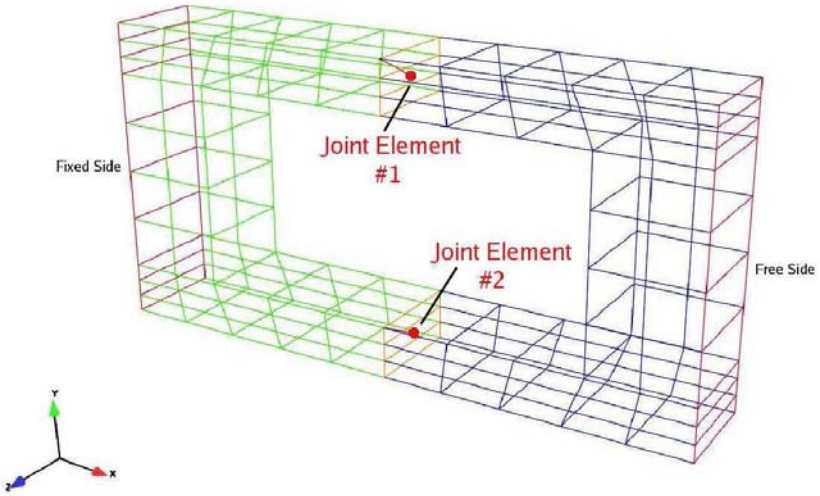

Figure 4.48. Mesh for two-joint structure.

Cases of two very different load amplitude were examined, but the large load is of greater interest as stronger nonlinearities and longer compute times are involved. In each case, three analyses were performed:

1. Transient analysis of the full nonlinear finite element model. This serves as a truth model.
2. Transient analysis of a nonlinear Galerkin model using twenty eigen modes of the reference linear system. In the following, we refer to these analyses of the modally truncated system as the reference model reduction.

3. Transient analysis of a nonlinear reduced model using eighteen eigen modes of the reference linear system and one joint mode appropriate for each of the two system joints. (The first eighteen eigen modes of the reference linear system include all those with frequencies below 20 kHz.)

The relative compute time between the full finite element analysis (in Salinas) and the Matlab calculations using the MDBF for this simple problem is shown in Table 4.1.

Table 4.1. Timing Summary for Two Joint Structure Nonlinear Model Reduction.

	Full System(sec) 1803 DOF	Reduced System(sec) 20 DOF
Nominal Load	426.5	0.4
10x Nominal	2281.2	0.4

Much more demanding calculations were performed on the mesh of the jointed structure shown in Figure 4.49.

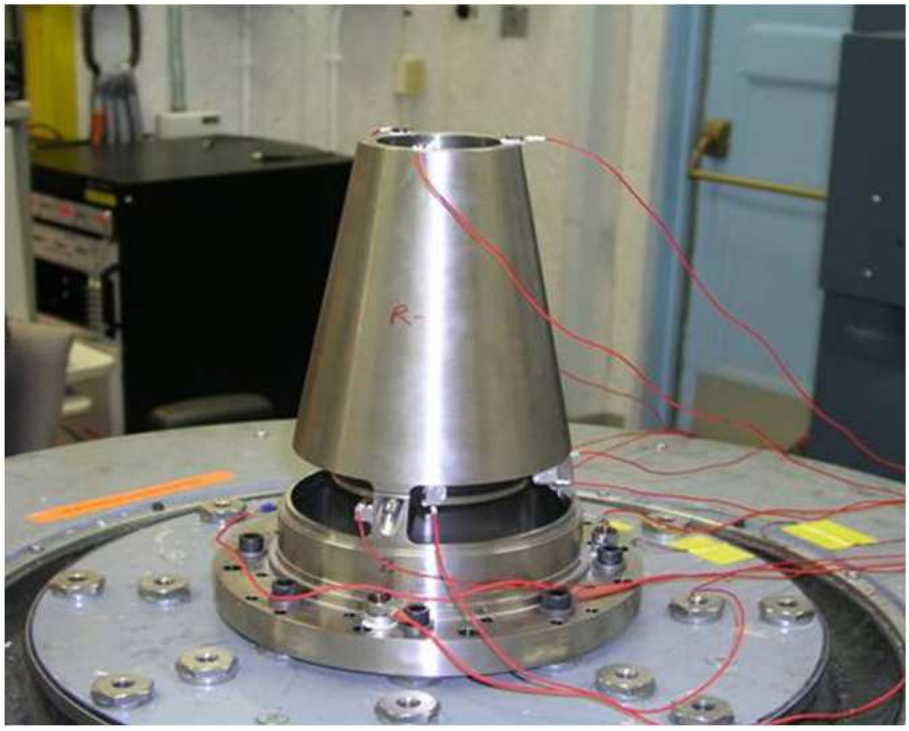

Figure 4.49. Mass mock used to test three transient analysis methods.

This object is subject to a base excitation of the form shown in Figure 4.50 and of sufficient amplitude to cause the joints on the joints almost to be sufficient to initiate macro-slip.

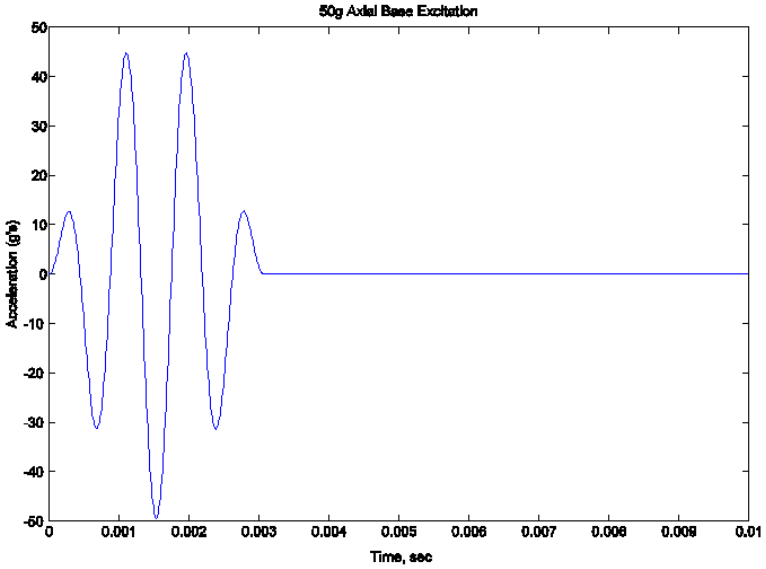

Figure 4.50. Base excitation imposed on the object of Figure 4.49.

Three methods of analysis were employed:

1. Full transient finite element analysis.
2. Reduction of the degrees of freedom of each monolithic substructure by Component Mode Synthesis and while solving the nonlinear joint equations connecting them.
3. Galerkin calculation using the first 15 eigen modes of the linearized structure.
4. Calculation using the method of discontinuous basis functions employing the lowest twelve natural modes and a Milman-Chu mode each joint degree of freedom.

The relative computational efficiency of these methods is shown in Table 4.2.

Table 4.2. Timing Summary for Mock AF&F Structure Nonlinear Reduced Models.

Model	No. DOF	CPU Time (sec)
CMS	117	36.3
Reference Reduced	15	2.8
Augmented Reduced	15	9.3
Full Salinas	206,343	40 hours (approximate)

Several observations should be made:

- One sees extraordinary increases in computational efficiency moving from the full finite element model to the CMS model. The method of discontinuous basis functions provides solutions in a quarter the time required by CMS. Solution using eigen modes along is faster still. Not all of these methods yield solutions of comparable quality.
- Not shown here, but presented in [20] , the results full finite element analysis manifest a number of spurious spikes. These are a result of the sharp nonlinearity of the joint model. (Physical joints have these sharp nonlinearities, but they also have additional properties that largely suppress such spikes.)
- The CMS analysis has similar spikes, but it can accommodate larger time steps than are possible in stable analysis of the full finite element analysis.
- Analysis using only eigen modes predicts ring-down very poorly. This is because the kinematics in the neighborhoods of the joints is not captured at all well and dissipation is grossly under-predicted.
- Predictions of the method of discontinuous basis functions predicts the experimental results better than any of the other analysis methods

On should note that the reduced order models discussed in this section do require the use of massively parallel computers to calculate matrices that are requisite for their employment.

We next consider a much larger problem. The structure shown in Figure 4.51 has about six million degrees of freedom. The five components are held together by a number Iwan joints and the base is subject to a prescribed acceleration.

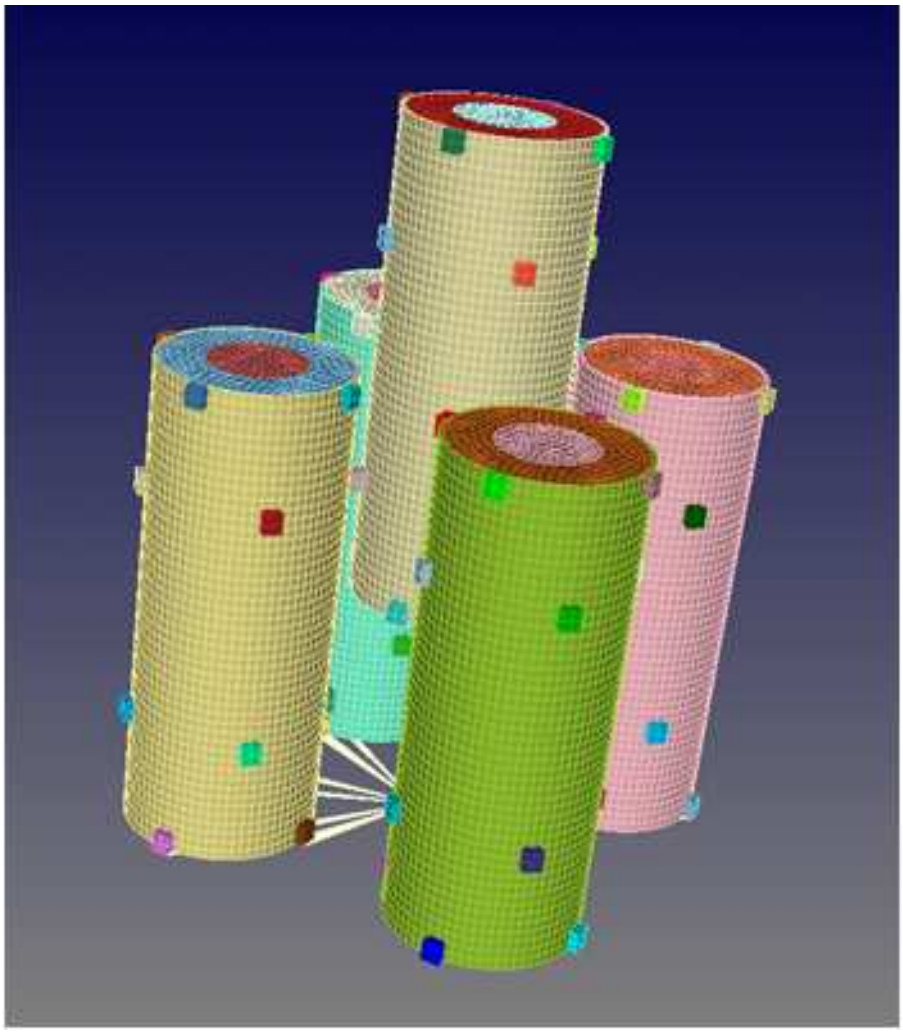

Figure 4.51. A very large structure mesh for testing model reduction.

The relative efficiency of the full finite element analysis and of calculation using the reduced order modeling of this chapter was explored by Mathew Brake and is indicated in Table 4.3. One notes that the time required for full finite element analysis is at best measured in hundreds of hours while the corresponding compute times when the reduced order model is employed is measured in minutes.

Table 4.3. Timing summary for dynamic analysis using very large mesh.

Time Step (sec)	Finite Element CPU Time (sec)	ROM CPU Time
1E-4	Canceled after 200 Hours	312 minutes
2E-4	Canceled after 200 Hours	156 minutes
4E-4	122.5 Hours	79 minutes
8E-4	Unstable	39 minutes
2E-3	Unstable	19.5 minutes
4E-3	Unstable	Unstable

This new ability to do capacity computing enables us to do types of analysis that are critical to the SNL mission. An example is that of mechanical systems such as discussed here, but with consideration of the statistical variability of joint properties and of load amplitudes. For instance, Fourier transform amplitudes are shown in Figure 4.52 for several hundred cases of excitation of the system considered here. (Principle component analysis of a small number of known joint parameter sets was employed to generate a very large pool of plausible parameter sets.) A very large number of transient dynamic simulations enables a statical analysis of component vulnerability.

4.8 Convergence of the Method of Discontinous Basis Functions

Many engineering systems are made of structures containing mechanical joints. These joints are nonlinear components that are spatially localized. Even though localized nonlinearities constitute a small part of the structure, the dynamic response of the structure is completely nonlinear and the analysis of the dynamic behavior is different from the analysis of linear structures. Assuming the knowledge of a realistic joint model, numerical simulations for structures with joints are still challenging. The direct integration for the resulting equations of motion is the simplest method. But it is computationally expensive. Several efforts have been made to extend techniques of linear structural dynamics to the analysis of systems with localized nonlinearities. How to best exploit these tools for solving the resulting nonlinear equations is still to be determined.

Above sections introduce a reduced order model for structures with localized nonlin-

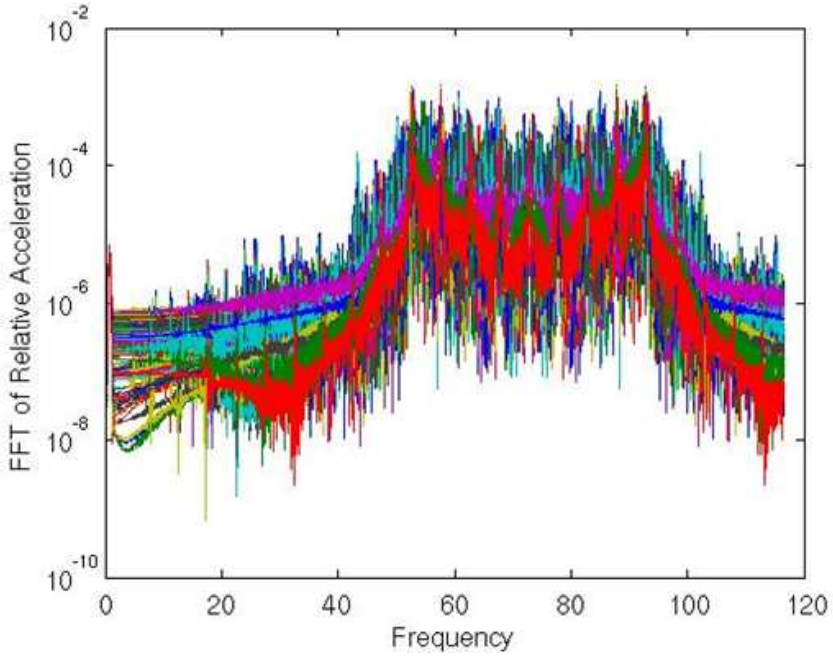

Figure 4.52. The many Fourier transforms that have been calculated from the very reduced model.

earities. The method combines eigenmodes of a reference linear system with functions having appropriate discontinuities at the locations of nonlinearity. Numerical experiments illustrate that the solution of the global nonlinear model is well approximated by a linear combination of eigenmodes and discontinuous modes. However, the mathematical analysis of the method is still to be developed and open questions remain.

So far, the number of elastic eigenmodes and discontinuous joint modes has been estimated in a heuristic manner. One employs all elastic modes corresponding to frequencies below a heuristically chosen cutoff frequency and one joint mode for each joint degree of freedom. In order to have confidence in the accuracy of the resulting analysis, quantifying the modal truncation error and the effect of discontinuous functions is important. A posteriori error analysis is critical for this quantification. The objective of this study is to develop a posteriori error estimators for this model reduction. We consider estimation for global norms of the error. We study first the static problem and, then, analyze the dynamic problem. In both cases, numerical experiments illustrate the numerical efficiency of the proposed estimators.

4.8.1 Static nonlinear problem

In this section, we study the static problem. After reviewing the formulation, we describe some properties of the nonlinear problem. For the reduced order model proposed in this chapter, we discuss an a posteriori error estimator.

Formulation

Consider a nonlinear problem composed of masses connected by springs. Between springs p and $p+1$, a cubic nonlinear spring is inserted. The left end of the system is fixed while the other end is free. The potential energy of the system is

$$\mathscr{E}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u} + \frac{1}{4}k_2(u_p - u_{p+1})^4 - \mathbf{u}^T\mathbf{f} \tag{4.26}$$

where $\mathbf{K}$ is the matrix associated with the system of linear springs and k_2 is a stiffness constant for the nonlinear spring.

We remark that the potential energy from the nonlinear spring is

$$\frac{1}{4}k_2(u_p - u_{p+1})^4 = \frac{1}{4}k_2(\mathbf{u}^T\mathbf{d})^4 \tag{4.27}$$

where the vector $\mathbf{d}$ is

$$\mathbf{d} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The stationary point to the energy $\mathscr{E}$ will satisfy

$$\mathbf{K}\mathbf{u} + \mathbf{N}(\mathbf{u}) = \mathbf{f} \tag{4.28}$$

where $\mathbf{N}$ is a localized nonlinearity of the form

$$\mathbf{N}(\mathbf{u}) = \alpha(\mathbf{u}^T\mathbf{d})\mathbf{d} \tag{4.29}$$

with $\alpha(x) = k_2 x^3$ and

$$\mathbf{d} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Interesting property for Newton nonlinear solver

Consider the problem (4.28) and its solution with the Newton nonlinear algorithm. Starting with a zero initial guess ($\mathbf{u}_0 = \mathbf{0}$), we have

$$\mathbf{r}_0 = \mathbf{f} - \mathbf{K}\mathbf{u}_0 - \mathbf{N}(\mathbf{u}_0) = \mathbf{f}.$$

The next iterate $\mathbf{u}_1$ is defined by

$$\left[\mathbf{K} + \frac{\partial \mathbf{N}}{\partial \mathbf{u}}(\mathbf{u}_0)\right](\mathbf{u}_1 - \mathbf{u}_0) = \mathbf{r}_0.$$

The derivative for $\mathbf{N}$ is

$$\frac{\partial \mathbf{N}}{\partial \mathbf{u}}(\mathbf{u}) = \alpha'(\mathbf{d}^T\mathbf{u})\mathbf{d}\mathbf{d}^T = 3k_2(\mathbf{d}^T\mathbf{u})^2\mathbf{d}\mathbf{d}^T.$$

Note that this derivative is of rank 1. Consequently, we can write the inverse matrix for

$$\mathbf{K} + \alpha'(\mathbf{d}^T\mathbf{u})\mathbf{d}\mathbf{d}^T,$$

which is a rank one pertubation of the stiffness matrix $\mathbf{K}$. With the Sherman-Morrison formula, its inverse satisfies

$$\left[\mathbf{K} + \alpha'(\mathbf{d}^T\mathbf{u})\mathbf{d}\mathbf{d}^T\right]^{-1} = \mathbf{K}^{-1} - \frac{\alpha'(\mathbf{d}^T\mathbf{u})}{1+\alpha'(\mathbf{d}^T\mathbf{u})\mathbf{d}^T\mathbf{K}^{-1}\mathbf{d}}\mathbf{K}^{-1}\mathbf{d}\mathbf{d}^T\mathbf{K}^{-1}. \tag{4.30}$$

The new iterate $\mathbf{u}_1$ will satisfy

$$\mathbf{u}_1 = \mathbf{u}_0 + \left[\mathbf{K} + \alpha'(\mathbf{d}^T\mathbf{u}_0)\mathbf{d}\mathbf{d}^T\right]^{-1}\mathbf{r}_0 = \mathbf{K}^{-1}\mathbf{f} + \delta_1\mathbf{K}^{-1}\mathbf{d}$$

where δ_1 is a scalar number (where we used the Sherman-Morrison formula (4.30)).

The new residual $\mathbf{r}_1$ is

$$\mathbf{r}_1 = \mathbf{f} - \mathbf{K}\mathbf{u}_1 - \mathbf{N}(\mathbf{u}_1) = \mathbf{f} - \mathbf{f} - \delta_1\mathbf{d} - \alpha(\mathbf{d}^T\mathbf{u}_1)\mathbf{d} = \beta_1\mathbf{d}$$

where β_1 is a scalar number. The new iterate $\mathbf{u}_2$ now becomes

$$\begin{aligned}\mathbf{u}_2 = \mathbf{u}_1 + \left[\mathbf{K} + \alpha'(\mathbf{d}^T\mathbf{u}_1)\mathbf{d}\mathbf{d}^T\right]^{-1}\mathbf{r}_1 &= \mathbf{K}^{-1}\mathbf{f} + \delta_1\mathbf{K}^{-1}\mathbf{d} + \tilde{\delta}_1\mathbf{K}^{-1}\mathbf{d} \\ &= \mathbf{K}^{-1}\mathbf{f} + \delta_2\mathbf{K}^{-1}\mathbf{d}\end{aligned}$$

where δ_2 is a scalar number. It is easy to generalize for any iterate $\mathbf{u}_n$ and any residual $\mathbf{r}_n$.

Proposition 4.8.1. *When* $\mathbf{u}_0 = \mathbf{0}$ *and we are solving*

$$\mathbf{K}\mathbf{u} + \mathbf{N}(\mathbf{u}) = \mathbf{f} \tag{4.31}$$

with the Newton algorithm, the first residual $\mathbf{r}_0$ *is equal to* $\mathbf{f}$*. Then every iterate* $\mathbf{u}_{n+1}$ *satisfies*

$$\mathbf{u}_{n+1} = \mathbf{K}^{-1}\mathbf{f} + \delta_{n+1}\mathbf{K}^{-1}\mathbf{d} \tag{4.32}$$

and every residual $\mathbf{r}_{n+1}$ *is aligned with the vector* $\mathbf{d}$*, i.e.*

$$\mathbf{r}_{n+1} = \beta_{n+1}\mathbf{d}. \tag{4.33}$$

Remark. If we know the vectors $\mathbf{f}$, $\mathbf{K}^{-1}\mathbf{f}$, and $\mathbf{K}^{-1}\mathbf{d}$, then we could exploit the equations (4.32) and (4.33) by computing only the scalars δ_{n+1} and β_{n+1}. This property could speed up the nonlinear solver.

Remark. When we are working with a reduced space spanned by $\mathbf{V}$, the nonlinear problem becomes

$$\mathbf{V}^T\mathbf{K}\mathbf{V}\mu + \mathbf{V}^T\mathbf{N}(\mathbf{V}\mu) = \mathbf{V}^T\mathbf{f} \tag{4.34}$$

or

$$\mathbf{V}^T\mathbf{K}\mathbf{V}\mu + \tilde{\mathbf{N}}(\mu) = \mathbf{V}^T\mathbf{f}$$

where the nonlinear function $\tilde{\mathbf{N}}$ satisfies

$$\tilde{\mathbf{N}}(\mu) = \alpha(\mu^T\mathbf{V}^T\mathbf{d})\mathbf{V}^T\mathbf{d}.$$

Starting from a zero initial guess $\mu_0 = \mathbf{0}$, we have $\rho_0 = \mathbf{V}^T\mathbf{f}$ and the iterate μ_{n+1} will satisfy

$$\mu_{n+1} = (\mathbf{V}^T\mathbf{K}\mathbf{V})^{-1}\mathbf{V}^T\mathbf{f} + \gamma_{n+1}(\mathbf{V}^T\mathbf{K}\mathbf{V})^{-1}\mathbf{V}^T\mathbf{d}$$

and the residual ρ_{n+1} will be

$$\rho_{n+1} = \zeta_{n+1}\mathbf{V}^T\mathbf{d}.$$

A posteriori error estimation on reduced space

We denote by $(\phi_i, \theta_i)_{1\leq i\leq N}$ the eigenmodes of $\mathbf{K}$ such that the eigenvalues are ordered in a non-decreasing way (N is the dimension of $\mathbf{K}$). Consider the subspace

$$\mathbf{V}_n = [\mathbf{K}^{-1}\mathbf{d}, \phi_1, \cdots, \phi_{n-1}],$$

where we solve approximately the nonlinear problem (4.28). The objective of this section is to estimate the error between the reduced solution in V_n and the exact solution in $\mathbb{R}^N$. Before studying an a posteriori error estimator, we state a result satisfied by the solution in the reduced space.

Interesting property of the approximate solution

Proposition 4.8.2. *The solution* $\mathbf{u}_n$ *in the reduced space* V_n *for the nonlinear problem* (4.28) *satisfies*

$$\mathbf{d}^T\mathbf{u} = \mathbf{d}^T\mathbf{u}_n, \quad \forall n > 0, \tag{4.35}$$

where $\mathbf{u}$ *denotes the solution for* (4.28) *in the whole space.*

Proof. To prove this result, we write the orthogonality of the residual with the subspace $\mathbf{V}_n$. We have

$$\mathbf{z}_n^T\left(\mathbf{f} - \mathbf{N}(\mathbf{u}_n) - \mathbf{K}\mathbf{u}_n\right) = 0 \quad \forall \mathbf{z}_n \in V_n$$

and

$$\mathbf{z}_n^T\left(\mathbf{K}\mathbf{u} + \mathbf{N}(\mathbf{u}) - \mathbf{N}(\mathbf{u}_n) - \mathbf{K}\mathbf{u}_n\right) = 0 \quad \forall \mathbf{z}_n \in V_n.$$

We can select $\mathbf{z}_n = \mathbf{K}^{-1}\mathbf{d}$ and we obtain

$$\mathbf{d}^T\left(\mathbf{u} - \mathbf{u}_n\right) + \mathbf{d}^T\mathbf{K}^{-1}\left(\mathbf{N}(\mathbf{u}) - \mathbf{N}(\mathbf{u}_n)\right) = 0.$$

Using the special form for $\mathbf{N}$, we have

$$\mathbf{N}(\mathbf{u}) - \mathbf{N}(\mathbf{u}_n) = \mathbf{d}\left(\alpha(\mathbf{u}^T\mathbf{d}) - \alpha(\mathbf{u}_n^T\mathbf{d})\right)$$

and

$$\alpha(\mathbf{u}^T\mathbf{d}) - \alpha(\mathbf{u}_n^T\mathbf{d}) = k_2\left(\mathbf{u}^T\mathbf{d} - \mathbf{u}_n^T\mathbf{d}\right)\left((\mathbf{u}^T\mathbf{d})^2 + (\mathbf{u}^T\mathbf{d}) \times (\mathbf{u}_n^T\mathbf{d}) + (\mathbf{u}_n^T\mathbf{d})^2\right).$$

Next we write

$$\mathbf{d}^T\left(\mathbf{u} - \mathbf{u}_n\right) \times \left[1 + k_2\mathbf{d}^T\mathbf{K}^{-1}\mathbf{d}\left((\mathbf{u}^T\mathbf{d})^2 + (\mathbf{u}^T\mathbf{d}) \times (\mathbf{u}_n^T\mathbf{d}) + (\mathbf{u}_n^T\mathbf{d})^2\right)\right] = 0.$$

Notice that

$$(\mathbf{u}^T\mathbf{d})^2 + (\mathbf{u}^T\mathbf{d}) \times (\mathbf{u}_n^T\mathbf{d}) + (\mathbf{u}_n^T\mathbf{d})^2 = \left(\mathbf{u}^T\mathbf{d} + \frac{1}{2}\mathbf{u}_n^T\mathbf{d}\right)^2 + \frac{3}{4}(\mathbf{u}_n^T\mathbf{d})^2 \geq 0.$$

Since $k_2 \geq 0$ and $\mathbf{d}^T\mathbf{K}^{-1}\mathbf{d} \geq 0$, we obtain

$$\left[1 + k_2\mathbf{d}^T\mathbf{K}^{-1}\mathbf{d}\left((\mathbf{u}^T\mathbf{d})^2 + (\mathbf{u}^T\mathbf{d}) \times (\mathbf{u}_n^T\mathbf{d}) + (\mathbf{u}_n^T\mathbf{d})^2\right)\right] > 0.$$

and, consequently,

$$\mathbf{d}^T(\mathbf{u} - \mathbf{u}_n) = 0.$$

□

The result does not depend on the vectors used to enrich the subspace V_n (here the eigenvectors), on the right hand side, nor on the value of k_2 (as long as $k_2 > 0$). It indicates that the reduced space will immediately capture the correct value for $\mathbf{d}^T\mathbf{u}_n$. A practical consequence is that we do not need to update the Jacobian matrix for the nonlinear spring when $n > 1$.

Remark. The result holds for nonlinearities of the form

$$\mathbf{N}(\mathbf{u}) = \alpha(\mathbf{u}^T\mathbf{d})\mathbf{d} \tag{4.36}$$

where the function α satisfies

$$\frac{\alpha(x) - \alpha(y)}{x - y} \geq 0. \tag{4.37}$$

Remark. A similar result holds for nonlinearities of the form

$$\mathbf{N}(\mathbf{u}) = \sum_{j=1}^{J} \alpha_j(\mathbf{u}^T\mathbf{d}_j)\mathbf{d}_j \tag{4.38}$$

when the reduced space contains the directions $\mathbf{K}^{-1}\mathbf{d}_1, \cdots, \mathbf{K}^{-1}d_J$ and with the assumptions that the functions α_j satisfy

$$\frac{\alpha_j(x) - \alpha_j(y)}{x - y} \geq 0 \tag{4.39}$$

and the vectors $\mathbf{d}_j$ verify

$$\mathbf{d}_i^T\mathbf{K}^{-1}\mathbf{d}_j = 0 \quad \text{for all } i \neq j. \tag{4.40}$$

A posteriori error estimator We derive a simple a posteriori error estimator for the nonlinear problem. We have

$$\mathbf{y}^T\mathbf{K}(\mathbf{u} - \mathbf{u}_n) = \mathbf{y}^T[\mathbf{K}\mathbf{u} - \mathbf{K}\mathbf{u}_n] \tag{4.41}$$

$$\mathbf{y}^T\mathbf{K}(\mathbf{u} - \mathbf{u}_n) = \mathbf{y}^T[\mathbf{f} - \mathbf{N}(\mathbf{u}) - \mathbf{K}\mathbf{u}_n] \tag{4.42}$$

$$\mathbf{y}^T\mathbf{K}(\mathbf{u} - \mathbf{u}_n) = \mathbf{y}^T[\mathbf{f} - \mathbf{N}(\mathbf{u}_n) - \mathbf{K}\mathbf{u}_n] + \mathbf{y}^T[\mathbf{N}(\mathbf{u}_n) - \mathbf{N}(\mathbf{u})] \tag{4.43}$$

Note that the result

$$\mathbf{d}^T\mathbf{u} = \mathbf{d}^T\mathbf{u}_n$$

implies that the second term satisfies $\mathbf{y}^T[\mathbf{N}(\mathbf{u}_n)-\mathbf{N}(\mathbf{u})]=0$.

We assume that the approximate solution $\mathbf{u}_n$ satisfies the relation

$$\mathbf{z}_n^T[\mathbf{f}-\mathbf{N}(\mathbf{u}_n)-\mathbf{K}\mathbf{u}_n]=0. \tag{4.44}$$

(because $\mathbf{u}_n$ comes from a reduced subspace V_n). Then we have

$$\mathbf{y}^T\mathbf{K}(\mathbf{u}-\mathbf{u}_n)=(\mathbf{y}-\mathbf{z}_n)^T[\mathbf{f}-\mathbf{N}(\mathbf{u}_n)-\mathbf{K}\mathbf{u}_n]. \tag{4.45}$$

So we get

$$\mathbf{y}^T\mathbf{K}(\mathbf{u}-\mathbf{u}_n)\le\|\mathbf{y}-\mathbf{z}_n\|_2\,\|\mathbf{f}-\mathbf{N}(\mathbf{u}_n)-\mathbf{K}\mathbf{u}_n\|_2\,,\quad\forall\,\mathbf{z}_n\in V_n. \tag{4.46}$$

Introducing the projection into $V_n^\perp$, $\mathbf{P}_{V_n^\perp}$, we write

$$\mathbf{y}^T\mathbf{K}(\mathbf{u}-\mathbf{u}_n)\le\left\|\mathbf{P}_{V_n^\perp}\mathbf{y}\right\|_2\|\mathbf{f}-\mathbf{N}(\mathbf{u}_n)-\mathbf{K}\mathbf{u}_n\|_2\,. \tag{4.47}$$

Recall that

$$\sup_{\mathbf{y}\neq\mathbf{0}}\frac{\mathbf{y}^T\mathbf{K}(\mathbf{u}-\mathbf{u}_n)}{\sqrt{\mathbf{y}^T\mathbf{K}\mathbf{y}}}=\sqrt{(\mathbf{u}-\mathbf{u}_n)^T\mathbf{K}(\mathbf{u}-\mathbf{u}_n)}.$$

We obtain

$$\sqrt{(\mathbf{u}-\mathbf{u}_n)^T\mathbf{K}(\mathbf{u}-\mathbf{u}_n)}\qquad\le\qquad\sup_{\mathbf{y}\neq\mathbf{0}}\sqrt{\frac{\mathbf{y}^T\mathbf{P}_{V_n^\perp}^T\mathbf{P}_{V_n^\perp}\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}\,\|\mathbf{f}-\mathbf{K}\mathbf{u}_n-\mathbf{N}(\mathbf{u}_n)\|_2\,. \tag{4.48}$$

Proposition 4.8.3. *The constant satisfies*

$$\frac{1}{\sqrt{\theta_{n+1}}}\le\sup_{\mathbf{y}\in V}\sqrt{\frac{\mathbf{y}^T\mathbf{P}_{V_n^\perp}^T\mathbf{P}_{V_n^\perp}\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}\le\frac{1}{\sqrt{\theta_n}} \tag{4.49}$$

Proof. First we prove the upper bound

$$\sup_{\mathbf{y}\neq\mathbf{0}}\sqrt{\frac{\mathbf{y}^T\mathbf{P}_{V_n^\perp}^T\mathbf{P}_{V_n^\perp}\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}\le\frac{1}{\sqrt{\theta_n}}.$$

For $n=1$, we have $\mathbf{V}_n=[\mathbf{K}^{-1}\mathbf{d}]$. $\mathbf{P}_{V_1^\perp}$ is a projection. So we have

$$\mathbf{y}^T\mathbf{P}_{V_1^\perp}^T\mathbf{P}_{V_1^\perp}\mathbf{y}=\left\|\mathbf{P}_{V_1^\perp}\mathbf{y}\right\|_2^2\le\|\mathbf{y}\|_2^2=\mathbf{y}^T\mathbf{y}$$

and

$$\sup_{\mathbf{y}\neq\mathbf{0}}\sqrt{\frac{\mathbf{y}^T\mathbf{P}_{V_1^\perp}^T\mathbf{P}_{V_1^\perp}\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}\le\sup_{\mathbf{y}\neq\mathbf{0}}\sqrt{\frac{\mathbf{y}^T\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}.$$

Recall that

$$\theta_1\le\inf_{\mathbf{y}\neq0}\frac{\mathbf{y}^T\mathbf{K}\mathbf{y}}{\mathbf{y}^T\mathbf{y}}$$

and

$$\sup_{\mathbf{y}\neq 0}\frac{\mathbf{y}^T\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}=\left(\inf_{\mathbf{y}\neq 0}\frac{\mathbf{y}^T\mathbf{K}\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\right)^{-1}.$$

So we obtain

$$\sup_{\mathbf{y}\neq\mathbf{0}}\sqrt{\frac{\mathbf{y}^T\mathbf{P}_{V_1^\perp}^T\mathbf{P}_{V_1^\perp}\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}\leq\frac{1}{\sqrt{\theta_1}}.$$

Consider now the case where $n>1$. We decompose the vector $\mathbf{d}$ on the basis of eigenvectors

$$\mathbf{d}=\sum_{i=1}^{N}d_i\phi_i$$

(where N is the dimension of the matrix $\mathbf{K}$) and introduce the vector $\tilde{\mathbf{d}}$

$$\tilde{\mathbf{d}}=\sum_{i=n}^{N}d_i\phi_i.$$

Note that

$$\mathbf{K}^{-1}\tilde{\mathbf{d}}=\sum_{i=n}^{N}\frac{d_i}{\theta_i}\phi_i$$

and

$$\phi_i^T\mathbf{K}^{-1}\tilde{\mathbf{d}}=0\quad\text{for } i<n.$$

The subspace V_n satisfies

$$\operatorname{span}\left(\mathbf{K}^{-1}\mathbf{d},\phi_1,\ldots,\phi_{n-1}\right)=\operatorname{span}\left(\mathbf{K}^{-1}\tilde{\mathbf{d}},\phi_1,\ldots,\phi_{n-1}\right).$$

but the basis for the right hand side is orthogonal. So we can decompose the projection $\mathbf{P}_{V_n^\perp}$ as follows

$$\mathbf{P}_{V_n^\perp}=\mathbf{P}_{[\mathbf{K}^{-1}\tilde{\mathbf{d}}]^\perp}\mathbf{P}_{[\phi_1,\ldots\phi_{n-1}]^\perp}.$$

So we have

$$\mathbf{y}^T\mathbf{P}_{V_n^\perp}^T\mathbf{P}_{V_n^\perp}\mathbf{y}=\left\|\mathbf{P}_{[\mathbf{K}^{-1}\tilde{\mathbf{d}}]^\perp}\mathbf{P}_{[\phi_1,\ldots\phi_{n-1}]^\perp}\mathbf{y}\right\|_2^2\leq\left\|\mathbf{P}_{[\phi_1,\ldots\phi_{n-1}]^\perp}\mathbf{y}\right\|_2^2$$

and

$$\sup_{\mathbf{y}\neq\mathbf{0}}\sqrt{\frac{\mathbf{y}^T\mathbf{P}_{V_n^\perp}^T\mathbf{P}_{V_n^\perp}\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}\leq\sup_{\mathbf{y}\neq\mathbf{0}}\sqrt{\frac{\mathbf{y}^T\mathbf{P}_{[\phi_1,\ldots\phi_{n-1}]^\perp}^T\mathbf{P}_{[\phi_1,\ldots\phi_{n-1}]^\perp}\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}=\frac{1}{\sqrt{\theta_n}}.$$

Next we need to prove the lower bound

$$\frac{1}{\sqrt{\theta_{n+1}}}\leq\sup_{\mathbf{y}\neq\mathbf{0}}\sqrt{\frac{\mathbf{y}^T\mathbf{P}_{V_n^\perp}^T\mathbf{P}_{V_n^\perp}\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}$$

Consider a vector $\mathbf{y}$ such that $\mathbf{d}^T\mathbf{K}^{-T}\mathbf{y} = 0$. Using an eigendecomposition of $\mathbf{y}$, the ratio becomes

$$\sqrt{\frac{\sum_{i=n}^{N} y_i^2}{\sum_{i=1}^{N} \theta_i y_i^2}}.$$

The result would hold if we have

$$\sum_{i=1}^{N} \theta_i y_i^2 \leq \sum_{i=n}^{N} \theta_{n+1} y_i^2$$

or

$$\sum_{i=1}^{n-1} \theta_i y_i^2 + (\theta_n - \theta_{n+1}) y_n^2 + 0 \times y_{n+1}^2 + \sum_{i=n+2}^{N} (\theta_i - \theta_{n+1}) y_i^2 \leq 0 \tag{4.50}$$

Since the eigenvalues θ_i are positive and ordered in a non-decreasing fashion, only the coefficients of y_n^2 and y_{n+1}^2 are negative or zero. If we select a vector $\mathbf{y}$ in $\left(\text{span}\left\{\mathbf{K}^{-1}\mathbf{d}\right\}\right)^{\perp} \cap \text{span}\left(\phi_n, \phi_{n+1}\right)$, then the bound (4.50) holds, which proves

$$\frac{1}{\sqrt{\theta_{n+1}}} \leq \sup_{\mathbf{y} \neq \mathbf{0}} \sqrt{\frac{\mathbf{y}^T \mathbf{P}_{V_n^{\perp}}^T \mathbf{P}_{V_n^{\perp}} \mathbf{y}}{\mathbf{y}^T \mathbf{K} \mathbf{y}}}$$

Note that the intersection $\left(\text{span}\left\{\mathbf{K}^{-1}\mathbf{d}\right\}\right)^{\perp} \cap \text{span}\left(\phi_n, \phi_{n+1}\right)$ contains a vector different from $\mathbf{0}$ because the sum of their dimensions is greater than N. □

On the subspaceV_n, an a posteriori error estimator for the $\mathbf{K}$-norm of the error is

$$\frac{1}{\sqrt{\theta_n}} \left\| \mathbf{f} - \mathbf{K}\mathbf{u}_n - \mathbf{N}(\mathbf{u}_n) \right\|_2. \tag{4.51}$$

Numerical experiment on efficiency To assess the efficiency of the estimator (4.51), we consider a system composed of 21 unit masses connected by springs of unit stiffness. Between springs 10 and 11, a cubic nonlinear spring is inserted. The left end of the system is fixed while the other end is free. The source term is

$$\mathbf{f} = f_0 [1, \cdots, 1]^T. \tag{4.52}$$

The estimator is

$$\frac{1}{\sqrt{\theta_n}} \left\| \mathbf{f} - \mathbf{K}\mathbf{u}_n - \mathbf{N}(\mathbf{u}_n) \right\|_2. \tag{4.53}$$

In Figure 4.53, we plot the ratio

$$\frac{\frac{1}{\sqrt{\theta_n}} \left\| \mathbf{f} - \mathbf{K}\mathbf{u}_n - \mathbf{N}(\mathbf{u}_n) \right\|_2}{\sqrt{(\mathbf{u} - \mathbf{u}_n)^T \mathbf{K} (\mathbf{u} - \mathbf{u}_n)}}. \tag{4.54}$$

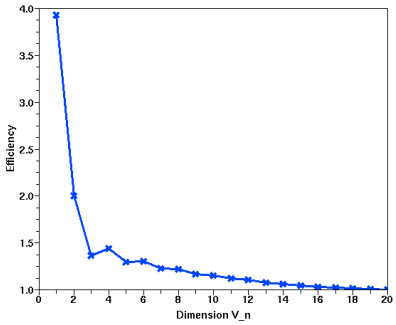

Figure 4.53. Efficiency with a subspace composed of $\mathbf{K}^{-1}\mathbf{d}$ and of eigenmodes

For this example, the estimator is fairly accurate for all values of n. For high values of n, the a posteriori estimator tends asymptotically towards the $\mathbf{K}$-norm of the error

$$\sqrt{(\mathbf{u}-\mathbf{u}_n)^T\mathbf{K}(\mathbf{u}-\mathbf{u}_n)}. \tag{4.55}$$

In Figure 4.54, we plot the ratio

$$\frac{1}{\sqrt{\theta_n}}\Big/\sup_{\mathbf{y}\neq\mathbf{0}}\sqrt{\frac{\mathbf{y}^T\mathbf{P}_{V_n^\perp}^T\mathbf{P}_{V_n^\perp}\mathbf{y}}{\mathbf{y}^T\mathbf{K}\mathbf{y}}}.$$

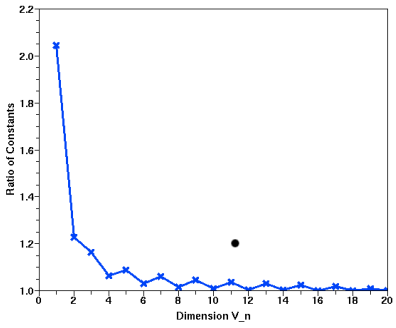

Figure 4.54. Comparison of constants with a subspace composed of $\mathbf{K}^{-1}\mathbf{d}$ and eigenmodes

The ratio is always greater than 1 because we have an upper bound. For this example,

the upper bound

$$\sup_{\mathbf{y}\neq\mathbf{0}} \sqrt{\frac{\mathbf{y}^T \mathbf{P}_{V_n^\perp}^T \mathbf{P}_{V_n^\perp} \mathbf{y}}{\mathbf{y}^T \mathbf{K} \mathbf{y}}} \leq \frac{1}{\sqrt{\theta_n}}$$

is sharp.

4.8.2 Dynamic nonlinear problem

Consider a nonlinear problem composed of masses connected by springs. Several nonlinear springs are inserted between linear springs. The left end of the system is fixed while the other end is free. We study now the dynamic problem. The semi-discrete undamped equations of motion are

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} + \sum_{j=1}^{J} n_j(\mathbf{u}^T \mathbf{d}_j)\mathbf{d}_j = \mathbf{f}. \tag{4.56}$$

Gronwall inequalities

For the sake of completeness, we recall some variants of the Gronwall lemma that will be useful in the following analysis.

Lemma 4.8.4. *Let $\alpha \in (0,1)$ and $C > 0$. Consider two continuous positive functions $\phi(t)$ and $m(t)$ such that*

$$\forall t \in [0,T], \quad \phi(t) \leq C + \int_0^t m(s)\phi(s)^\alpha ds. \tag{4.57}$$

Then we have, for all $t \in [0,T]$,

$$\phi(t) \leq \left\{ C^{1-\alpha} + (1-\alpha) \int_0^t m(s) ds \right\}^{\frac{1}{1-\alpha}}. \tag{4.58}$$

Lemma 4.8.5. *Let $\alpha \in (0, \frac{1}{2}]$. Consider two continuous positive functions $m(t)$ and $p(t)$ on $[0,T]$. Denote ϕ a differentiable positive function on $[0,T]$ whose derivative is continuous and satisfying*

$$\forall t \in [0,T], \quad \phi'(t) \leq m(t)\phi(t)^\alpha + p(t)\phi(t). \tag{4.59}$$

Then we have, for all $t \in [0,T]$,

$$\phi(t) \leq \left[\phi(0)^{1-\alpha} e^{(1-\alpha)\int_0^t p(s)ds} + (1-\alpha) \int_0^t m(s) e^{(1-\alpha)\int_s^t p(\tau)d\tau} ds \right]^{\frac{1}{1-\alpha}}. \tag{4.60}$$

Proof. If ϕ is zero on $[0,T]$, then the result holds. Denote $G = \phi^\alpha$. We have

$$\phi' = \frac{1}{\alpha} G' G^{\frac{1-\alpha}{\alpha}}$$

and

$$\frac{1}{\alpha}G'G^{\frac{1-\alpha}{\alpha}} \leq mG + pG^{\frac{1}{\alpha}}.$$

Consider I a sub-interval of $[0,T]$ where ϕ is non-zero. Then we obtain on I, after dividing by G,

$$\frac{1}{\alpha}G'G^{\frac{1}{\alpha}-2} - pG^{\frac{1}{\alpha}-1} \leq m$$

and

$$\left(\frac{1}{\alpha}-1\right)G'G^{\frac{1}{\alpha}-2} - \alpha\left(\frac{1}{\alpha}-1\right)pG^{\frac{1}{\alpha}-1} \leq \alpha\left(\frac{1}{\alpha}-1\right)m.$$

The inequality remains true for any t in $[0,T]$ because the left hand side is equal to 0 on $[0,T]\backslash I$. The left hand side is an exact derivative

$$\frac{d}{dt}\left(G(t)^{\frac{1}{\alpha}-1}e^{-(1-\alpha)\int_{t_0}^{t}p(s)ds}\right) = \left[\left(\frac{1}{\alpha}-1\right)G'G^{\frac{1}{\alpha}-2} - (1-\alpha)\,pG^{\frac{1}{\alpha}-1}\right]e^{-(1-\alpha)\int_{t_0}^{t}p(s)ds}.$$

Integrating the previous inequality, we obtain

$$G(t)^{\frac{1}{\alpha}-1}e^{-(1-\alpha)\int_0^t p(s)ds} - G(0)^{\frac{1}{\alpha}-1} \leq (1-\alpha)\int_0^t m(s)e^{-(1-\alpha)\int_0^s p(\tau)d\tau}ds$$

and

$$G(t)^{\frac{1}{\alpha}-1} \leq G(0)^{\frac{1}{\alpha}-1}e^{(1-\alpha)\int_0^t p(s)ds} + (1-\alpha)\int_0^t m(s)e^{(1-\alpha)\int_s^t p(\tau)d\tau}ds.$$

We get

$$\phi(t) \leq \left[\phi(0)^{1-\alpha}e^{(1-\alpha)\int_0^t p(s)ds} + (1-\alpha)\int_0^t m(s)e^{(1-\alpha)\int_s^t p(\tau)d\tau}ds\right]^{\frac{1}{1-\alpha}}.$$

□

Analysis

In this section, we perform the analysis of the semi-discrete equations of motion. In particular, we study stability bounds, uniqueness of the solution, and a priori error estimates.

Stability estimates

Theorem 4.8.6. *Consider the semi-discrete undamped equations of motion*

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} + \sum_{j=1}^{J} n_j(\mathbf{u}^T\mathbf{d}_j)\mathbf{d}_j = \mathbf{f}. \tag{4.61}$$

We assume that each function n_j has a positive primitive. Then, for any time $t \in [0,T]$, we have

$$\sqrt{\frac{1}{2}\dot{\mathbf{u}}^T\mathbf{M}\dot{\mathbf{u}} + \frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u} + \sum_{j=1}^{J} N_j(\mathbf{u}^T\mathbf{d}_j)} \leq \sqrt{E_0} + \frac{1}{\sqrt{2}}\int_0^t \|\mathbf{f}(s)\|_{\mathbf{M}^{-1}}\, ds \tag{4.62}$$

where E_0 is the initial energy

$$E_0 = \frac{1}{2}\mathbf{v}_0^T\mathbf{M}\mathbf{v}_0 + \frac{1}{2}\mathbf{u}_0^T\mathbf{K}\mathbf{u}_0 + \sum_{j=1}^{J} N_j(\mathbf{u}_0^T\mathbf{d}_j). \tag{4.63}$$

Remark. Theorem 4.8.6 still holds with a weaker assumption on the functions n_j:

$$\frac{1}{2}\dot{\mathbf{u}}^T\mathbf{M}\dot{\mathbf{u}} + \frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u} + \sum_{j=1}^{J} N_j(\mathbf{u}^T\mathbf{d}_j) \geq 0, \quad \forall t \in [0,T]. \tag{4.64}$$

Proof. We start by multiplying the equations of motion with the vector $\dot{\mathbf{u}}^T$. Notice that we have

$$\dot{\mathbf{u}}^T\mathbf{M}\ddot{\mathbf{u}} = \frac{1}{2}\frac{d}{dt}\left(\dot{\mathbf{u}}^T\mathbf{M}\dot{\mathbf{u}}\right) \quad \text{and} \quad \dot{\mathbf{u}}^T\mathbf{K}\mathbf{u} = \frac{1}{2}\frac{d}{dt}\left(\mathbf{u}^T\mathbf{K}\mathbf{u}\right)$$

and

$$n_j(\mathbf{u}^T\mathbf{d}_j)\dot{\mathbf{u}}^T\mathbf{d}_j^T = \frac{d}{dt}\left(N_j(\mathbf{u}^T\mathbf{d}_j)\right).$$

Denote

$$E(t) = \frac{1}{2}\dot{\mathbf{u}}^T\mathbf{M}\dot{\mathbf{u}} + \frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u} + \sum_{j=1}^{J} N_j(\mathbf{u}^T\mathbf{d}_j).$$

Then we have

$$\frac{dE}{dt}(t) = \dot{\mathbf{u}}^T\mathbf{f} \leq \sqrt{\dot{\mathbf{u}}^T\mathbf{M}\dot{\mathbf{u}}}\sqrt{\mathbf{f}^T\mathbf{M}^{-1}\mathbf{f}} \leq \sqrt{2E(t)}\sqrt{\mathbf{f}^T\mathbf{M}^{-1}\mathbf{f}}.$$

After integration, we obtain

$$E(t) \leq E(0) + \int_0^t \|\mathbf{f}(s)\|_{\mathbf{M}^{-1}}\sqrt{2E(s)}ds$$

We conclude by using the Gronwall inequality from Lemma 4.8.4. □

Corollary 4.8.7. *Consider the semi-discrete undamped equations of motion*

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} + \sum_{j=1}^{J} n_j(\mathbf{u}^T\mathbf{d}_j)\mathbf{d}_j = \mathbf{f}. \tag{4.65}$$

We assume that each function n_j has a positive primitive. Then, for any time $t \in [0,T]$, we have

$$\|\dot{\mathbf{u}}\|_{\mathbf{M}} \leq \sqrt{2E_0} + \int_0^t \|\mathbf{f}(s)\|_{\mathbf{M}^{-1}}\, ds \tag{4.66}$$

$$\|\mathbf{u}\|_{\mathbf{K}} \leq \sqrt{2E_0} + \int_0^t \|\mathbf{f}(s)\|_{\mathbf{M}^{-1}}\, ds \tag{4.67}$$

$$\sqrt{\sum_{j=1}^{J} N_j(\mathbf{u}^T\mathbf{d}_j)} \leq \sqrt{2E_0} + \int_0^t \|\mathbf{f}(s)\|_{\mathbf{M}^{-1}}\, ds \tag{4.68}$$

$$\|\mathbf{u}\|_{\mathbf{M}} \leq \|\mathbf{u}_0\|_{\mathbf{M}} + t\sqrt{2E_0} + \int_0^t (t-s)\,\|\mathbf{f}(s)\|_{\mathbf{M}^{-1}}\, ds \tag{4.69}$$

where E_0 is the initial energy

$$E_0 = \frac{1}{2}\mathbf{v}_0^T\mathbf{M}\mathbf{v}_0 + \frac{1}{2}\mathbf{u}_0^T\mathbf{K}\mathbf{u}_0 + \sum_{j=1}^{J} N_j(\mathbf{u}_0^T\mathbf{d}_j). \tag{4.70}$$

Proof. The first three inequalities are straightforward because the energy is bounding the left hand sides. The last estimate is obtained by writing

$$\mathbf{u}(t) = \mathbf{u}_0 + \int_0^t \dot{\mathbf{u}}(s)ds$$

and using the bound for $\|\dot{\mathbf{u}}\|_{\mathbf{M}}$. □

Corollary 4.8.8. *Consider the semi-discrete undamped equations of motion*

$$\mathbf{V}^T\mathbf{M}\mathbf{V}\ddot{\mu} + \mathbf{V}^T\mathbf{K}\mathbf{V}\mu + \sum_{j=1}^{J} n_j(\mu^T\mathbf{V}^T\mathbf{d}_j)\mathbf{V}^T\mathbf{d}_j = \mathbf{V}^T\mathbf{f}. \tag{4.71}$$

We assume that each function n_j has a positive primitive. Then, for any time $t \in [0,T]$, we have

$$\sqrt{\frac{1}{2}\dot{\mu}^T\mathbf{V}^T\mathbf{M}\mathbf{V}\dot{\mu} + \frac{1}{2}\mu^T\mathbf{V}^T\mathbf{K}\mathbf{V}\mu + \sum_{j=1}^{J} N_j(\mu^T\mathbf{V}^T\mathbf{d}_j)} \leq \sqrt{E_0} + \frac{1}{\sqrt{2}}\int_0^t \left\|\mathbf{V}^T\mathbf{f}(s)\right\|_{\mathbf{M}^{-1}} ds \tag{4.72}$$

where E_0 is the initial energy

$$E_0 = \frac{1}{2}\mathbf{v}_0^T\mathbf{V}^T\mathbf{M}\mathbf{V}\mathbf{v}_0 + \frac{1}{2}\mathbf{u}_0^T\mathbf{V}^T\mathbf{K}\mathbf{V}\mathbf{u}_0 + \sum_{j=1}^{J} N_j(\mathbf{u}_0^T\mathbf{V}^T\mathbf{d}_j). \tag{4.73}$$

Uniqueness Next we study the uniqueness of the solution.

Theorem 4.8.9. *Consider the semi-discrete undamped equations of motion*

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} + \sum_{j=1}^{J} n_j(\mathbf{u}^T\mathbf{d}_j)\mathbf{d}_j = \mathbf{f} \tag{4.74}$$

with $\mathbf{u}(0) = \mathbf{u}_0$ *and* $\dot{\mathbf{u}}(0) = \dot{\mathbf{u}}_0$. *If the solution* $\mathbf{u}$ *exists, it is unique.*

Proof. Consider two solutions $\mathbf{u}$ and $\mathbf{v}$ in $[0,T]$. We start by writing the equation satisfied by $\mathbf{u}-\mathbf{v}$,

$$\mathbf{M}(\ddot{\mathbf{u}} - \ddot{\mathbf{v}}) + \mathbf{K}(\mathbf{u} - \mathbf{v}) = \mathbf{f} - \sum_{j=1}^{J} n_j(\mathbf{u}^T\mathbf{d}_j)\mathbf{d}_j - \mathbf{f} + \sum_{j=1}^{J} n_j(\mathbf{v}^T\mathbf{d}_j)\mathbf{d}_j$$

We multiply this equation with $\dot{\mathbf{u}} - \dot{\mathbf{v}}$ to obtain

$$\frac{d}{dt}\left[\frac{1}{2}(\dot{\mathbf{u}}-\dot{\mathbf{v}})^T\mathbf{M}(\dot{\mathbf{u}}-\dot{\mathbf{v}}) + \frac{1}{2}(\mathbf{u}-\mathbf{v})^T\mathbf{K}(\mathbf{u}-\mathbf{v})\right] = \sum_{j=1}^{J}\left[n_j(\mathbf{v}^T\mathbf{d}_j) - n_j(\mathbf{u}^T\mathbf{d}_j)\right](\dot{\mathbf{u}}-\dot{\mathbf{v}})^T\mathbf{d}_j$$

and

$$\frac{d}{dt}\left[\frac{1}{2}(\dot{\mathbf{u}}-\dot{\mathbf{v}})^T\mathbf{M}(\dot{\mathbf{u}}-\dot{\mathbf{v}}) + \frac{1}{2}(\mathbf{u}-\mathbf{v})^T\mathbf{K}(\mathbf{u}-\mathbf{v})\right] \leq \sum_{j=1}^{J}\left|n_j(\mathbf{v}^T\mathbf{d}_j) - n_j(\mathbf{u}^T\mathbf{d}_j)\right| \|\dot{\mathbf{u}}-\dot{\mathbf{v}}\|_{\mathbf{M}} \left\|\mathbf{d}_j\right\|_{\mathbf{M}^{-1}}$$

Corollary 4.8.7 implies that $\mathbf{u}$ and $\mathbf{v}$ are bounded for $t \in [0,T]$. Since the functions n_j are locally Lipschitz continuous, there exists a constant C_0 depending on T, $\mathbf{f}$, and the initial conditions $(\mathbf{u}_0, \dot{\mathbf{u}}_0)$ such that

$$\left|n_j(\mathbf{u}^T\mathbf{d}_j) - n_j(\mathbf{v}^T\mathbf{d}_j)\right| \leq C_0\left|\mathbf{u}^T\mathbf{d}_j - \mathbf{v}^T\mathbf{d}_j\right| \leq C_0\|\mathbf{u}-\mathbf{v}\|_{\mathbf{K}}\left\|\mathbf{d}_j\right\|_{\mathbf{K}^{-1}}. \tag{4.75}$$

Denote

$$\tilde{E}(t) = \frac{1}{2}(\dot{\mathbf{u}}-\dot{\mathbf{v}})^T\mathbf{M}(\dot{\mathbf{u}}-\dot{\mathbf{v}}) + \frac{1}{2}(\mathbf{u}-\mathbf{v})^T\mathbf{K}(\mathbf{u}-\mathbf{v}).$$

We obtain

$$\frac{d\tilde{E}}{dt} \leq 2C_0\tilde{E}\sum_{j=1}^{J}\left\|\mathbf{d}_j\right\|_{\mathbf{K}^{-1}}\left\|\mathbf{d}_j\right\|_{\mathbf{M}^{-1}}$$

which implies that

$$E(t) \leq E(0)e^{2tC_0\Sigma_{j=1}^{J}\left\|\mathbf{d}_j\right\|_{\mathbf{K}^{-1}}\left\|\mathbf{d}_j\right\|_{\mathbf{M}^{-1}}}$$

Assuming that $\tilde{E}(0) = 0$, we obtain that $\tilde{E}$ is zero at all time and that the solutions $\mathbf{u}$ and $\mathbf{v}$ are equal. □

Error estimates

Theorem 4.8.10. *Consider the semi-discrete undamped equations of motion*

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} + \sum_{j=1}^{J} n_j(\mathbf{u}^T\mathbf{d}_j)\mathbf{d}_j = \mathbf{f} \tag{4.76}$$

and the reduced-order model

$$\mathbf{V}^T\mathbf{M}\mathbf{V}\ddot{\mu} + \mathbf{V}^T\mathbf{K}\mathbf{V}\mu + \sum_{j=1}^{J} n_j(\mu^T\mathbf{V}^T\mathbf{d}_j)\mathbf{V}^T\mathbf{d}_j = \mathbf{V}^T\mathbf{f}. \tag{4.77}$$

We assume that the columns of $\mathbf{V}$ *contain the vectors* $\mathbf{K}^{-1}\mathbf{d}_j$ *and that the functions* n_j *are locally Lipschitz continuous. Denote* μ_{exact} *the projection of* $\mathbf{u}$ *into the span of* $\mathbf{V}$ *for the* $\mathbf{K}$*-inner product,*

$$\mathbf{V}^T\mathbf{K}(\mathbf{u} - \mathbf{V}\mu_{exact}) = 0 \quad \Rightarrow \quad \mathbf{V}^T\mathbf{K}\mathbf{V}\mu_{exact} = \mathbf{V}^T\mathbf{K}\mathbf{u}. \tag{4.78}$$

Then there exists a constant $C > 0$ *such that, for any time* $t \in [0,T]$*, we have*

$$\begin{aligned}\sqrt{\|\dot{\mu}(t) - \dot{\mu}_{exact}(t)\|^2_{\mathbf{V}^T\mathbf{M}\mathbf{V}} + \|\mu(t) - \mu_{exact}(t)\|^2_{\mathbf{V}^T\mathbf{K}\mathbf{V}}} \\ \leq e^{Ct}\sqrt{\|\dot{\mu}(0) - \dot{\mu}_{exact}(0)\|^2_{\mathbf{V}^T\mathbf{M}\mathbf{V}} + \|\mu(0) - \mu_{exact}(0)\|^2_{\mathbf{V}^T\mathbf{K}\mathbf{V}}} \\ + \int_0^t \|\ddot{\mathbf{u}} - \mathbf{V}\ddot{\mu}_{exact}\|_{\mathbf{M}} e^{C(t-s)} ds\end{aligned}$$

Proof. We start by writing the equation satisfied by $\mu - \mu_{exact}$,

$$\begin{aligned}&\mathbf{M}\mathbf{V}(\ddot{\mu} - \ddot{\mu}_{exact}) + \mathbf{K}\mathbf{V}(\mu - \mu_{exact}) \\ &\qquad = \mathbf{M}\mathbf{V}\ddot{\mu} + \mathbf{K}\mathbf{V}\mu + \sum_{j=1}^{J} n_j(\mu^T\mathbf{V}^T\mathbf{d}_j)\mathbf{d}_j - \mathbf{f} \\ &\qquad\quad + \mathbf{M}(\ddot{\mathbf{u}} - \mathbf{V}\ddot{\mu}_{exact}) + \mathbf{K}(\mathbf{u} - \mathbf{V}\mu_{exact}) \\ &\qquad\qquad + \sum_{j=1}^{J} \left[n_j(\mathbf{u}^T\mathbf{d}_j) - n_j(\mu^T\mathbf{V}^T\mathbf{d}_j)\right]\mathbf{d}_j\end{aligned}$$

We multiply this equation with $(\dot{\mu} - \dot{\mu}_{exact})^T\mathbf{V}^T$ to obtain

$$\begin{aligned}&\frac{d}{dt}\left[\frac{1}{2}(\dot{\mu} - \dot{\mu}_{exact})^T\mathbf{V}^T\mathbf{M}\mathbf{V}(\dot{\mu} - \dot{\mu}_{exact}) + \frac{1}{2}(\mu - \mu_{exact})^T\mathbf{V}^T\mathbf{K}\mathbf{V}(\mu - \mu_{exact})\right] \\ &\qquad = (\dot{\mu} - \dot{\mu}_{exact})^T\mathbf{V}^T\mathbf{M}(\ddot{\mathbf{u}} - \mathbf{V}\ddot{\mu}_{exact}) \\ &\qquad\quad + \sum_{j=1}^{J} \left[n_j(\mathbf{u}^T\mathbf{d}_j) - n_j(\mu^T\mathbf{V}^T\mathbf{d}_j)\right](\dot{\mu} - \dot{\mu}_{exact})^T\mathbf{V}^T\mathbf{d}_j\end{aligned}$$

because μ is the solution of the reduced-order problem and μ_{exact} is the projection of $\mathbf{u}$ into the span of $\mathbf{V}$ for the $\mathbf{K}$-inner product. The Cauchy-Schwarz inequality gives

$$\begin{aligned}\frac{d}{dt}\left[\frac{1}{2}(\dot{\mu}-\dot{\mu}_{exact})^T\mathbf{V}^T\mathbf{M}\mathbf{V}(\dot{\mu}-\dot{\mu}_{exact})+\frac{1}{2}(\mu-\mu_{exact})^T\mathbf{V}^T\mathbf{K}\mathbf{V}(\mu-\mu_{exact})\right]\\ \leq \sqrt{(\dot{\mu}-\dot{\mu}_{exact})^T\mathbf{V}^T\mathbf{M}\mathbf{V}(\dot{\mu}-\dot{\mu}_{exact})}\,\|\ddot{\mathbf{u}}-\mathbf{V}\ddot{\mu}_{exact}\|_{\mathbf{M}}\\ +\sum_{j=1}^{J}\left|n_j(\mathbf{u}^T\mathbf{d}_j)-n_j(\mu^T\mathbf{V}^T\mathbf{d}_j)\right|\sqrt{(\dot{\mu}-\dot{\mu}_{exact})^T\mathbf{V}^T\mathbf{M}\mathbf{V}(\dot{\mu}-\dot{\mu}_{exact})}\,\left\|\mathbf{d}_j\right\|_{\mathbf{M}^{-1}}\end{aligned}$$

Corollary 4.8.7 implies that $\mathbf{u}$ and $\mathbf{V}\mu$ are bounded for $t\in[0,T]$. Since the functions n_j are locally Lipschitz continuous, there exists a constant C_0 depending on T, $\mathbf{f}$, and the initial conditions $(\mathbf{u}_0,\mathbf{v}_0)$ such that

$$\left|n_j(\mathbf{u}^T\mathbf{d}_j)-n_j(\mu^T\mathbf{V}^T\mathbf{d}_j)\right|\leq C_0\left|\mathbf{u}^T\mathbf{d}_j-\mu^T\mathbf{V}^T\mathbf{d}_j\right|. \tag{4.79}$$

The columns of $\mathbf{V}$ contain the vectors $\mathbf{K}^{-1}\mathbf{d}_j$ and we have

$$\mathbf{u}^T\mathbf{d}_j=\mu_{exact}^T\mathbf{V}^T\mathbf{d}_j.$$

We obtain

$$\begin{aligned}\frac{d}{dt}\left[\frac{1}{2}\|\dot{\mu}(t)-\dot{\mu}_{exact}(t)\|^2_{\mathbf{V}^T\mathbf{M}\mathbf{V}}+\frac{1}{2}\|\mu(t)-\mu_{exact}(t)\|^2_{\mathbf{V}^T\mathbf{K}\mathbf{V}}\right]\\ \leq \sqrt{(\dot{\mu}-\dot{\mu}_{exact})^T\mathbf{V}^T\mathbf{M}\mathbf{V}(\dot{\mu}-\dot{\mu}_{exact})}\,\|\ddot{\mathbf{u}}-\mathbf{V}\ddot{\mu}_{exact}\|_{\mathbf{M}}\\ +C_0\sum_{j=1}^{J}\sqrt{\|\dot{\mu}(t)-\dot{\mu}_{exact}(t)\|^2_{\mathbf{V}^T\mathbf{M}\mathbf{V}}}\sqrt{\|\mu(t)-\mu_{exact}(t)\|^2_{\mathbf{V}^T\mathbf{K}\mathbf{V}}}\left\|\mathbf{d}_j\right\|_{\mathbf{K}^{-1}}\left\|\mathbf{d}_j\right\|_{\mathbf{M}^{-1}}\end{aligned}$$

We conclude by using the Gronwall inequality from Lemma 4.8.5 with $\alpha=1/2$ and

$$m(t)=\sqrt{2}\,\|\ddot{\mathbf{u}}-\mathbf{V}\ddot{\mu}_{exact}\|_{\mathbf{M}}\ \text{ and }\ p(t)=2C_0\sum_{j=1}^{J}\left\|\mathbf{d}_j\right\|_{\mathbf{K}^{-1}}\left\|\mathbf{d}_j\right\|_{\mathbf{M}^{-1}}.$$

□

Corollary 4.8.11. *Under the assumptions of Theorem 4.8.10, we have, for any time $t\in[0,T]$,*

$$\begin{aligned}\sqrt{\|\dot{\mathbf{u}}(t)-\mathbf{V}\dot{\mu}(t)\|^2_{\mathbf{M}}+\|\mathbf{u}(t)-\mathbf{V}\mu(t)\|^2_{\mathbf{K}}}\\ \leq e^{Ct}\sqrt{\|\dot{\mu}(0)-\dot{\mu}_{exact}(0)\|^2_{\mathbf{V}^T\mathbf{M}\mathbf{V}}+\|\mu(0)-\mu_{exact}(0)\|^2_{\mathbf{V}^T\mathbf{K}\mathbf{V}}}\\ +\|\dot{\mathbf{u}}(t)-\mathbf{V}\dot{\mu}_{exact}(t)\|_{\mathbf{M}}+\|\mathbf{u}(t)-\mathbf{V}\mu_{exact}(t)\|_{\mathbf{K}}\\ +\int_0^t\|\ddot{\mathbf{u}}-\mathbf{V}\ddot{\mu}_{exact}\|_{\mathbf{M}}\,e^{C(t-s)}ds\end{aligned}$$

A posteriori error estimator

In this section, we study an a posteriori error estimator. This work was made in collaboration with Mikala Johnson, graduate student in the Department of Applied Mathematics at University of Washington.

Theorem 4.8.12. *Consider the semi-discrete undamped equations of motion*

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} + \sum_{j=1}^{J} n_j(\mathbf{u}^T\mathbf{d}_j)\mathbf{d}_j = \mathbf{f} \tag{4.80}$$

and the reduced-order model

$$\mathbf{V}^T\mathbf{M}\mathbf{V}\ddot{\mu} + \mathbf{V}^T\mathbf{K}\mathbf{V}\mu + \sum_{j=1}^{J} n_j(\mu^T\mathbf{V}^T\mathbf{d}_j)\mathbf{V}^T\mathbf{d}_j = \mathbf{V}^T\mathbf{f}. \tag{4.81}$$

We assume that the columns of $\mathbf{V}$ *contain the vectors* $\mathbf{K}^{-1}\mathbf{d}_j$ *and that the functions* n_j *are locally Lipschitz continuous. Then there exists a constant* $C > 0$ *such that, for any time* $t \in [0,T]$*, we have*

$$\begin{aligned}\sqrt{\|\dot{\mathbf{u}}(t) - \mathbf{V}\dot{\mu}(t)\|_{\mathbf{M}}^2 + \|\mathbf{u}(t) - \mathbf{V}\mu(t)\|_{\mathbf{K}}^2} \\ \leq \sqrt{\|\dot{\mathbf{u}}(0) - \mathbf{V}\dot{\mu}(0)\|_{\mathbf{M}}^2 + \|\mathbf{u}(0) - \mathbf{V}\mu(0)\|_{\mathbf{K}}^2}\, e^{Ct\sum_{j=1}^{J}\|\mathbf{d}_j\|_{\mathbf{K}^{-1}}\|\mathbf{d}_j\|_{\mathbf{M}^{-1}}} \\ + \int_0^t \|\mathbf{r}\|_{\mathbf{M}^{-1}} e^{C(t-s)\sum_{j=1}^{J}\|\mathbf{d}_j\|_{\mathbf{K}^{-1}}\|\mathbf{d}_j\|_{\mathbf{M}^{-1}}} ds\end{aligned}$$

where the residual vector $\mathbf{r}$ *is defined by*

$$\mathbf{r} = \mathbf{f} - \mathbf{M}\mathbf{V}\ddot{\mu} - \mathbf{K}\mathbf{V}\mu - \sum_{j=1}^{J} n_j(\mu^T\mathbf{V}^T\mathbf{d}_j)\mathbf{d}_j$$

Proof. Without any loss of generality, we consider only one nonlinearity. Recall the exact equation of motion:

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} + n(\mathbf{u}^T\mathbf{d})\mathbf{d} = \mathbf{f} \tag{4.82}$$

and the equation of the reduced order model:

$$\mathbf{V}^T\mathbf{M}\mathbf{V}\ddot{\mu} + \mathbf{V}^\mathbf{T}\mathbf{K}\mathbf{V}\ddot{\mu} + n(\mu^T\mathbf{V}^T\mathbf{d})\mathbf{V}^\mathbf{T}\mathbf{d} = \mathbf{V}^\mathbf{T}\mathbf{f} \tag{4.83}$$

Then the residual is given by:

$$\mathbf{r} = \mathbf{f} - \mathbf{M}\mathbf{V}\ddot{\mu} - \mathbf{K}\mathbf{V}\ddot{\mu} - n(\mu^T\mathbf{V}^T\mathbf{d})\mathbf{d} \tag{4.84}$$

Replacing the expression for $\mathbf{f}$ in equation (4.84) with the expression in (4.82):

$$\mathbf{r} = \mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} + n(\mathbf{u}^T\mathbf{d})\mathbf{d} - \mathbf{M}\mathbf{V}\ddot{\mu} - \mathbf{K}\mathbf{V}\ddot{\mu} - n(\mu^T\mathbf{V}^T\mathbf{d})\mathbf{d},$$

$$\mathbf{f}-\mathbf{MV}\ddot{\mu}-\mathbf{KV}\ddot{\mu}-n(\mu^T\mathbf{V}^T\mathbf{d})\mathbf{d}= \\ \mathbf{M}\ddot{\mathbf{u}}+\mathbf{Ku}+n(\mathbf{u}^T\mathbf{d})\mathbf{d}-\mathbf{MV}\ddot{\mu}-\mathbf{KV}\ddot{\mu}-n(\mu^T\mathbf{V}^T\mathbf{d})\mathbf{d}$$

and

$$\mathbf{M}(\ddot{\mathbf{u}}-\mathbf{V}\ddot{\mu})+\mathbf{K}(\mathbf{u}-\mathbf{V}\mu)= \\ \mathbf{f}-\mathbf{MV}\ddot{\mu}-\mathbf{KV}\mu-n(\mu^T\mathbf{V}^T\mathbf{d})\mathbf{d}-\left[n(\mathbf{u}^T\mathbf{d})-n(\mu^T\mathbf{V}^T\mathbf{d})\right]\mathbf{d}.$$

Multiplying the above equation by $(\dot{\mathbf{u}}-\mathbf{V}\dot{\mu})^T$ to obtain an exact derivative on the left hand side:

$$\frac{d}{dt}\left[\frac{1}{2}\dot{\mathbf{e}}^T\mathbf{M}\dot{\mathbf{e}}+\frac{1}{2}\mathbf{e}^T\mathbf{Ke}\right]=(\dot{\mathbf{u}}-\mathbf{V}\dot{\mu})^T\mathbf{r}+\left[n(\mu^T\mathbf{V}^T\mathbf{d})-n(\mathbf{u}^T\mathbf{d})\right](\dot{\mathbf{u}}-\mathbf{V}\dot{\mu})^{\mathbf{T}}\mathbf{d} \tag{4.85}$$

Let $R(t)=\left[\frac{1}{2}\dot{\mathbf{e}}^T\mathbf{M}\dot{\mathbf{e}}+\frac{1}{2}\mathbf{e}^T\mathbf{Ke}\right]$, bound each term of the right hand side separately as a function $R(t)$ starting with the first term, letting $(\dot{\mathbf{u}}-\mathbf{V}\dot{\mu})=\dot{\mathbf{e}}$, then utilizing the Cauchy-Schwarz inequality

$$\left|\dot{\mathbf{e}}^T\mathbf{r}\right|\leq\|\dot{\mathbf{e}}\|_{\mathbf{M}}\|\mathbf{r}\|_{\mathbf{M}^{-1}}$$

Next, we note that the quantity $\|\dot{\mathbf{e}}\|_{\mathbf{M}}$ can be bounded as:

$$\|\dot{\mathbf{e}}\|_{\mathbf{M}}=\sqrt{\dot{\mathbf{e}}^T\mathbf{M}\dot{\mathbf{e}}}\leq\sqrt{\dot{\mathbf{e}}^T\mathbf{M}\dot{\mathbf{e}}+\mathbf{e}^T\mathbf{Ke}}=\sqrt{2R(t)}$$

since $\mathbf{e}^T\mathbf{Ke}\geq 0$. Thus, for the first term of equation (4.85) we find:

$$(\dot{\mathbf{u}}-\mathbf{V}\dot{\mu})^T\mathbf{r}=\dot{\mathbf{e}}^T\mathbf{r}\leq\sqrt{2R(t)}\|\mathbf{r}\|_{\mathbf{M}^{-1}} \tag{4.86}$$

The second part of the second term of equation (4.85) is bounded as was done before except this time instead of **r** we now have **d**:

$$(\dot{\mathbf{u}}-\mathbf{V}\dot{\mu})^T\mathbf{d}=\dot{\mathbf{e}}^T\mathbf{d}\leq\sqrt{2R(t)}\|\mathbf{d}\|_{\mathbf{M}^{-1}} \tag{4.87}$$

Next we must bound the first part of the second term assuming Lipschitz continuity of the nonlinear function:

$$\begin{aligned}|n(\mu^T\mathbf{V}^T\mathbf{d})-n(\mathbf{u}^T\mathbf{d})|&\leq C|(\mu^T\mathbf{V}^T-\mathbf{u}^T)\mathbf{d}| \\ &\leq C|\mathbf{e}^T\mathbf{d}| \\ &\leq C\|\mathbf{e}\|_{\mathbf{K}}\|\mathbf{d}\|_{\mathbf{K}^{-1}} \\ &\leq C\|\mathbf{d}\|_{\mathbf{K}^{-1}}\sqrt{2R(t)}\end{aligned} \tag{4.88}$$

where C is the Lipschitz constant. Putting all three bounds together (4.86,4.87,4.88) into equation (4.85) we obtain:

$$\begin{aligned}\frac{dR(t)}{dt}&\leq\sqrt{2R(t)}\|\mathbf{r}\|_{\mathbf{M}^{-1}}+\sqrt{2R(t)}\|\mathbf{d}\|_{\mathbf{M}^{-1}}C\|\mathbf{d}\|_{\mathbf{K}^{-1}}\sqrt{2R(t)} \\ &\leq\sqrt{2R(t)}\|\mathbf{r}\|_{\mathbf{M}^{-1}}+2C\|\mathbf{d}\|_{\mathbf{K}^{-1}}\|\mathbf{d}\|_{\mathbf{M}^{-1}}R(t)\end{aligned}$$

To this equation we are able to apply the Gronwall inequality where $m(t) = \sqrt{2}\,\|\mathbf{r}\|_{\mathbf{M}^{-1}}$, $p(t) = 2C\,\|\mathbf{d}\|_{\mathbf{K}^{-1}}\,\|\mathbf{d}\|_{\mathbf{M}^{-1}}$, and $\alpha = 1/2$:

$$R(t)^{\frac{1}{2}} \leq R^{\frac{1}{2}}(0)\exp(\frac{1}{2}\int_0^t 2C\,\|\mathbf{d}\|_{\mathbf{K}^{-1}}\,\|\mathbf{d}\|_{\mathbf{M}^{-1}}) + \frac{1}{2}\int_0^t \sqrt{2}\,\|\mathbf{r}\|_{\mathbf{M}^{-1}}\exp(\frac{1}{2}\int_s^t 2C\,\|\mathbf{d}\|_{\mathbf{K}^{-1}}\,\|\mathbf{d}\|_{\mathbf{M}^{-1}})ds$$

Note that $p(t)$ does not depend on t so the integrals can be simplified, and also note that $R(t)$ is the expression for the energy of the error in both the linear acceleration and linear spring-stiffness terms:

$$\frac{1}{2}\left[\dot{\mathbf{e}}^T\mathbf{M}\dot{\mathbf{e}} + \mathbf{e}^T\mathbf{K}\mathbf{e}\right] \leq [R(0)^{\frac{1}{2}}\exp(C\,\|\mathbf{d}\|_{\mathbf{K}^{-1}}\,\|\mathbf{d}\|_{\mathbf{M}^{-1}}\,t) + \frac{1}{\sqrt{2}}\int_0^t \|\mathbf{r}\|_{\mathbf{M}^{-1}}\exp(C\,\|\mathbf{d}\|_{\mathbf{K}^{-1}}\,\|\mathbf{d}\|_{\mathbf{M}^{-1}}\,(t-s))ds]^2 \tag{4.89}$$

□

Numerical efficiency of estimator The results in figure (4.55) were compiled using the solutions obtained from solving the problem in equation (4.83) with the local cubic nonlinearity,

$$n(\mathbf{u}^T\mathbf{d})\mathbf{d} = (\mathbf{u}^T\mathbf{d})^3\mathbf{d},$$

on the subspaces of exact eigenvectors of $\mathbf{M}$ and $\mathbf{K}$ with the augmenting vector $\mathbf{K}^{-1}\mathbf{d}$. The Δt used for the numerical solutions was $\Delta t_{max}/128$, while the timestep for the "true" solution was taken to be $\Delta t_{max}/2048$. The "true" solution was obtained by directly integrating the full nonlinear problem (eq. 4.82). In addition, the Lipschitz constant was approximated as

$$C = \max_{(\mathbf{u},t)\in D} |3(\mathbf{u}^T\mathbf{d})^2|.$$

For this particular formulation of the problem a couple simplifications of equation (4.89) can be made. First the initial error is zero, $R(0) = 0$, since $\mathbf{u}_0 = \dot{\mathbf{u}}_0 = \mathbf{V}\mu = \mathbf{V}\dot{\mu} = 0$. Also, it is possible to separate the term,

$$\exp(C\,\|\mathbf{d}\|_{\mathbf{K}^{-1}}\,\|\mathbf{d}\|_{\mathbf{M}^{-1}}\,t),$$

from the square:

$$\dot{\mathbf{e}}^T\mathbf{M}\dot{\mathbf{e}} + \mathbf{e}^T\mathbf{K}\mathbf{e} \leq \exp(2C\,\|\mathbf{d}\|_{\mathbf{K}^{-1}}\,\|\mathbf{d}\|_{\mathbf{M}^{-1}}\,t)\times [\int_0^t \|\mathbf{r}\|_{\mathbf{M}^{-1}}\exp(-C\,\|\mathbf{d}\|_{\mathbf{K}^{-1}}\,\|\mathbf{d}\|_{\mathbf{M}^{-1}}\,s)ds]^2$$

Clearly this estimator radically overestimates the error for $t > 4$ or so, and it grows exponentially with the length of integration. However, it does not ever underestimate the error, even at the beginning of the time integration.

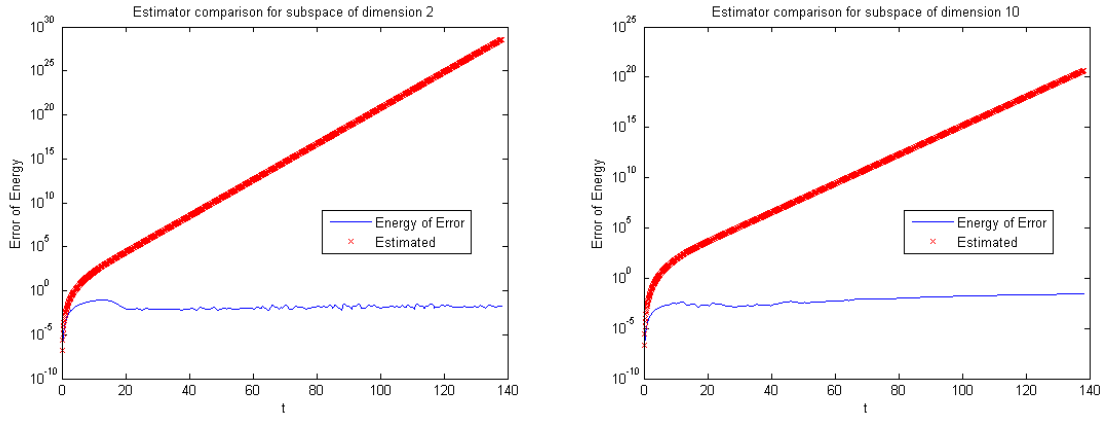

Figure 4.55. Comparison of Gronwall-derived error estimator and computed error as a function of time for various dimension subspaces including the augmenting vector $\mathbf{K}^{-1}\mathbf{d}$.

4.9 Conclusion

We have demonstrated the feasibility of achieving reduced models for systems with localized nonlinearities by augmenting the most natural basis functions with vectors that have appropriate discontinuities at the locations of nonlinearity. This assures that the kinematics necessary to couple the nonlinearity to the rest of the structural response are present.

An important observation - regardless of what analogies one wishes to make - is that it is because of the mechanical coupling of those joint modes with the eigen modes of the reference linear structure that those eigen modes satisfactory describe the nonlinear structural dynamics.

Though this model reduction appears to work well for both large and small structural loads, a few words are appropriate about how the results manifest themselves for the cases of small loads. It is observed in the examples shown here that for such cases the amplitude of the generalized coefficient for the discontinuous basis function is always very low compared to those of the first several elastic eigen modes. The augmenting mode serves the purpose of coupling the joint mechanics into the dynamics of the other modes. In doing so it does not change the characteristic mode shapes as seen by SVD, it just provides modest nonlinear damping. In these ways, the apparent modal response of the reduced nonlinear systems appears very much like that found in a modal lab for real structures: one sees apparently linear modes, except for nonlinear damping of each mode.

When this technique is employed in finite element analysis of jointed structures - espe-

cially when Component Mode Synthesis is employed along the way - dramatic increases in computing efficiency can be achieved. This new ability to do capacity computing enables us to do types of analysis that are critical to the SNL mission, such as when variability in joint properties of applied loads must be considered in a statistical analysis of component vulnerability.

Finally, original work by Ulrich Hetmaniuk demonstrates a theoretical basis for the quality of approximation demonstrated by the method of discontinuous basis functions exploratory computations.

4.10 Chapter Acknowledgments

The authors thanks Thomas Burton of New Mexico State University for alerting the author to the papers of Milman and Chu as well as for several helpful suggestions for exploration of the methods presented here. Thanks are also due to the authors' colleagues Todd Griffith, James Lauffer, Garth Reese, and Matt Hopkins many helpful suggestions.

Of course the authors thank Daniel (Todd) Griffith and Mathew Brake for their generosity of time and effort in performing the finite element calculations presented in this chapter. A detailed description of Todd Griffith's calculations is presented in [20] and Matt Brakes calculations will be the focus of a future Sandia report.

Chapter 5

Conclusion

Each of the model reduction methods presented in this report has the capacity to reduce compute times for realistic problems by orders of magnitude. Taken together, and integrated so much as possible, they can change the manner in which real problems of value to the SNL mission can be addressed. Instead of doing one or two simulations of a large system, we shall be able to do *many* such calculations accounting for uncertainties in loads, boundary conditions, and interface parameters. It will be possible to make meaningful probabilistic statements about system performance.

The first step in this acceleration of analysis is provides by stabilized tied contact. The analyst if freed to mesh each component of a structure independent of others, so long as the interfaces have sufficient fidelity to capture the physics of the problems. Mesh A of Figure 5.1 illustrates how when meshing the left and right blocks, one must specify the locations of the nodes on the interface for each to align with those specified for the other. The advantages of stabilized tied contact are illustrated in Mesh B, The blocks on the left and right are meshed independently of each other. The block on the left has a fine mesh appropriate to the load distribution anticipate for it. The block on the right is meshed in two regions: the region on the left is course but suitable for the strain gradients anticipated while the region on the left of that block is fine enough to accommodate the strain gradients associated with the corners of the interfaces of the two blocks. In Mesh B, stabilized tied contact is used in both interfaces.

One can integrate stabilized tied contact and scalable component mode synthesis. Consider the shape function of Figure 3.9 imposed between adjacent substructures. In that figure, one assumes the the nodal arrangements are identical across the interface. That condition is not necessary at all. One may connect the two interfaces via stabilized tied contact and then solving the specialized local problems (Equations 3.30 and 3.17).

A further integration of these methods can be employed in the integration of CMS models for components provided by different sources. Much current project management energy goes into seeing that the nodal configurations of these CMS models is consistent. This problem can be mitigated using the stabilized tied contact method provided the surface shape functions for each substructure model is available. (See Figure 5.2.)

The combination of stabilized tied contact and scalable component mode synthesis so reduces the size to the underlying model that it is now tractable to perform the system

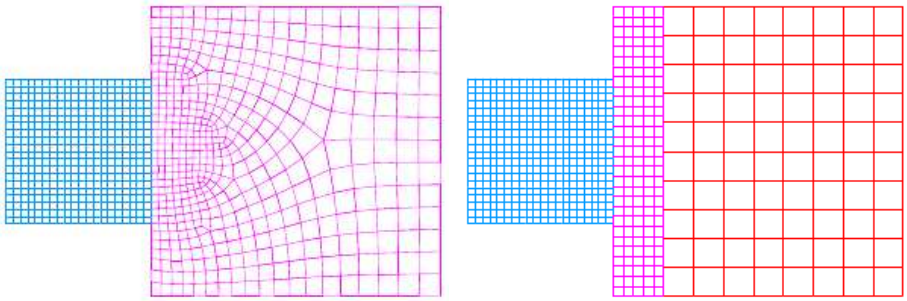

Figure 5.1. Creating conformal mesh integrating two features requires coordinated meshing of each. Stabilized tied contact makes it possible to mesh each independently.

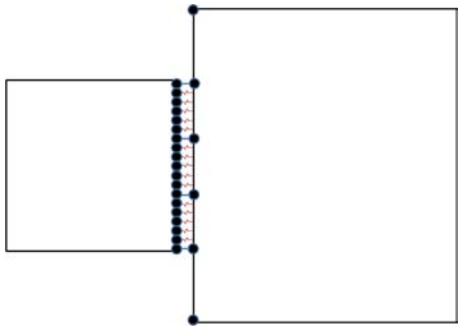

Figure 5.2. Stabilized tied contact can also be used to connect substructures, each of which is modeled by CMS, so long as shape functions for the surfaces are given.

eigen analysis and and the many quasi-static analyses necessary for the use of the method of discontinuous basis functions. This nonlinear transient analysis can then be performed using even fewer degrees of freedom. The large time steps possible with this method and the small problem size make it possible to perform the number of analyses to account for system variabilities.

References

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM Journal on Numerical Analysis*, 39(5):1749–1779, 2002.

[2] I. Babuška, U. Banerjee, and J. Osborn. On principles for the selection of shape functions for the generalized finite element method. *Comput. Methods Appl. Mech. Engrg.*, 191:5595–5629, 2002.

[3] I. Babuška, G. Caloz, and J. Osborn. Special finite element methods for a class of second order elliptic problems with rough coefficients. *SIAM J. Numer. Anal.*, 31(4):945–981, 1994.

[4] I. Babuška and J. E. Osborn. Generalized finite element methods: Their performance and their relation to mixed methods. *SIAM J. Numer. Anal.*, 20(3):510–536, 1983.

[5] T. Barth, P. Bochev, M. Gunzburger, and J. Shadid. A taxonomy of consistently stabilized finite element methods for the Stokes problem. *SIAM Journal on Scientific Computing*, 25(5):1585–1607, 2004.

[6] R. Becker, P. Hansbo, and R. Stenberg. A finite element method for domain decomposition with non-matching grids. *Mathematical Modeling and Numerical Analysis*, 37(2):209–225, 2003.

[7] J. K. Bennighof and R. B. Lehoucq. An automated multilevel substructuring method for eigenspace computation in linear elastodynamics. *SIAM J. Sci. Comput.*, 25(6):2084–2106, 2004.

[8] C. Bernardi, Y. Maday, and A. T. Patera. *A new nonconforming approach to domain decomposition: the mortar element method*. in Nonlinear Partial Differential Equations and Their Applications, H. Brezis and J. L. Lions eds. Longman Scientific & Technical, 1994, pp. 13–51.

[9] F. Bourquin. Component mode synthesis and eigenvalues of second order operators: Discretization and algorithm. *Mathematical Modelling and Numerical Analysis*, 26:385–423, 1992.

[10] F. Brezzi and L. Marini. Augmented spaces, two-level methods, and stabilizing subgrids. *Int. J. Numer. Meth. Fluids*, 40:31–46, 2002.

[11] C. Chu and M. H. Milman. Eigenvalue error analysis of viscously damped structures using a Ritz reduction method. *AIAA Journal*, 30:2935–2944, December 1992.

[12] R. R. Craig, Jr. and M. C. C. Bampton. Coupling of substructures for dynamic analysis. *AIAA Journal*, 6(7):1313–1319, 1968.

[13] Dassault Systemes Simulia Corp. *Abaqus Analysis User's Manual, Version 6.8*.

[14] C. R. Dohrmann. Analysis of stabilized tied contact. *in preparation*, 2009.

[15] C. R. Dohrmann, S. W. Key, and M. W. Heinstein. Methods for connecting dissimilar three-dimensional finite element meshes. *International Journal for Numerical Methods in Engineering*, 47:1057–1080, 2000.

[16] C. R. Dohrmann and O. B. Widlund. Hybrid domain decomposition algorithms for compressible and nearly incompressible elasticity. Technical Report TR2008-919, Department of Computer Science, New York University, 2009. to appear in International Journal for Numerical Methods in Engineering.

[17] Y. Efendiev and T. Hou. *Multiscale Finite Element Methods: Theory and Applications*, volume 4 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer New York, first edition, 2009.

[18] S. Falletta. The approximate integration in the mortar method constraint. In O. B. Widlund and D. E. Keyes, editors, *Lecture Notes in Computational Science and Engineering*, volume 55, pages 555–563, Berlin, 2008. Springer.

[19] M. J. Gander and C. Japhet. An algorithm for non-matching grid projections with linear complexity. In M. Bercovier, M. Gander, R. Kornhuber, and O. Widlund, editors, *Proceeding of the 18th International Conference on Domain Decomposition Methods*, pages 185–192, Jerusalem, Israel, 2008.

[20] D. Todd Griffith and Daniel J. Segalman. Finite element calculations illustrating a method of model reduction for the dynamics of structures with localized nonlinearities. Technical Report SAND2006-5843, Sandia National Laboratories, October 2006.

[21] P. Hansbo, C. Lovadian, I. Perugia, and G. Sangalli. A Lagrange multiplier method for the finite element solution of elliptic interface problems using non-matching meshes. *Numerische Mathematik*, 100:91–115, 2005.

[22] U. Hetmaniuk and R. B. Lehoucq. Multilevel methods for eigenspace computations in structural dynamics. In *Domain Decomposition Methods in Science and Engineering*, volume 55 of *Lecture Notes in Computational Science and Engineering*, pages 103–114. Springer-Verlag, 2007.

[23] T. Hou and X. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134:169–189, 1997.

[24] W. C. Hurty. Vibrations of structural systems by component-mode synthesis. *Journal of the Engineering Mechanics Division, ASCE*, 86:51–69, 1960.

[25] W. D. Iwan. Distributed-element model for hysteresis and its steady-state dynamic response. *ASME Journal of Applied Mechanics*, 33(4):893–900, Dec. 1966.

[26] W. D. Iwan. On a class of models for yielding behavior of continuous and composite systems. *ASME Journal of Applied Mechanics*, 34(3):612–617, 1967.

[27] L. V. Kantorovich and V. I. Krylov. *Approximate Methods of Higher Analysis*. Interscience, New York, 1958. (translated from Russian).

[28] G. Kerschen, J. C. Golinval, A. F. Vakakis, and L. A. Bergman. The method of proper orthogonal decomposition for dynamical characterization and order reduction of mechanical systems: An overview. *Nonlinear Dynamics*, 41(1-3):147 – 169, August 2005.

[29] Y. Maday, F. Rapetti, and B. I Wohlmuth. The influence of quadrature formulas in 2d and 3d mortar element methods. In L. F. Pavarino and A. Toselli, editors, *Lecture Notes in Computational Science and Engineering*, volume 23, pages 203–221. Springer, 2002.

[30] S. G. Mikhlin. *Mathematical Physics: an Advanced Course*. American Elsevier Publishing Company, Inc., New York, 1970.

[31] M. H. Milman and C. C. Chu. Optimization methods for passive damper placement and tuning. *Journal of Guidance, Control, and Dynamics*, 17(4):848 – 56, Jul-Aug 1994.

[32] MSC Software Corporation. *MD Nastran R3 Release Guide*, 2008.

[33] E. G. Ng and B. W. Peyton. Block sparse Cholesky algorithms on advanced uniprocessor computers. *SIAM Journal on Scientific Computing*, 14(5):1034–1056, 1993.

[34] J. Nolen, G. Papanicolaou, and O. Pironneau. A framework for adaptive multiscale methods for elliptic problems. *Multiscale Modeling & Simulation*, 7(1):171–196, 2008.

[35] A Pantano and R. C. Averill. A penalty-based interface technology for coupling independently modeled 3D finite element meshes. *Finite Elements in Analysis and Design*, 43:271–286, 2007.

[36] K. C. Park, C. A. Felippa, and G. Rebel. A simple algorithm for lacalized construction of non-matching structural interfaces. *International Journal for Numerical Methods in Engineering*, 53:2117–2142, 2002.

[37] E. Pesheck, C. Pierre, and S. W. Shaw. New galerkin-based approach for accurate nonlinear normal modes through invariant manifolds. *Journal of Sound and Vibration*, 249(5):971 – 993, Jan 2002.

[38] M. A. Puso. A 3D mortar method for solid mechanics. *International Journal for Numerical Methods in Engineering*, 59:315–336, 2004.

[39] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, UK, first edition, 1999.

[40] Daniel J. Segalman. A four-parameter Iwan model for lap-type joints. Technical Report SAND2002-3828, Sandia National Laboratories, 2002.

[41] Daniel J. Segalman. Modelling joint friction in structural dynamics. *Structural Control and Health Monitoring*, 13(1):430–453, 2005.

[42] Daniel Joseph Segalman. A four-parameter Iwan model for lap-type joints. *ASME Journal of Applied Mechanics*, 72(5):752–760, September 2005.

[43] Gilbert Strang and George J. Fix. *An Analysis of the Finite Element Method*. Wellesley-Cambridge Press, Wellesley, MA 02181 USA, 1988.

[44] B. I. Wohlmuth. A mortar finite element method using dual spaces for the Lagrange multiplier. *SIAM Journal on Numerical Analysis*, 38(3):989–1012, 2000.

[45] G. Zavaries and L. De Lorenzis. A modified node-to-segment algorithms passing the contact patch test. *International Journal for Numerical Methods in Engineering*, 79:379–416, 2009.

DISTRIBUTION:

5 Ulrich L. Hetmaniuk
Department of Applied Maths,
University of Washington, Box 352420,
Seattle, WA 98195-2420

5	MS 0346	Dohrmann, Clark R. , 01523
1	MS 0346	Brake, Mathew R., 01526
1	MS 0346	Baca, Tom, 01523
5	MS 0557	Segalman, Daniel J., 01525
1	MS 0557	Redmond, James M., 01525
1	MS 0557	Griffith, Todd, 01523
1	MS 1318	Ken Alvin , 01414
5	MS 1320	Lehoucq, Richard, 1414
1	MS 0899	Technical Library, 9536 (electronic)
1	MS 0123	D. Chavez, LDRD Office, 1011

Sandia National Laboratories