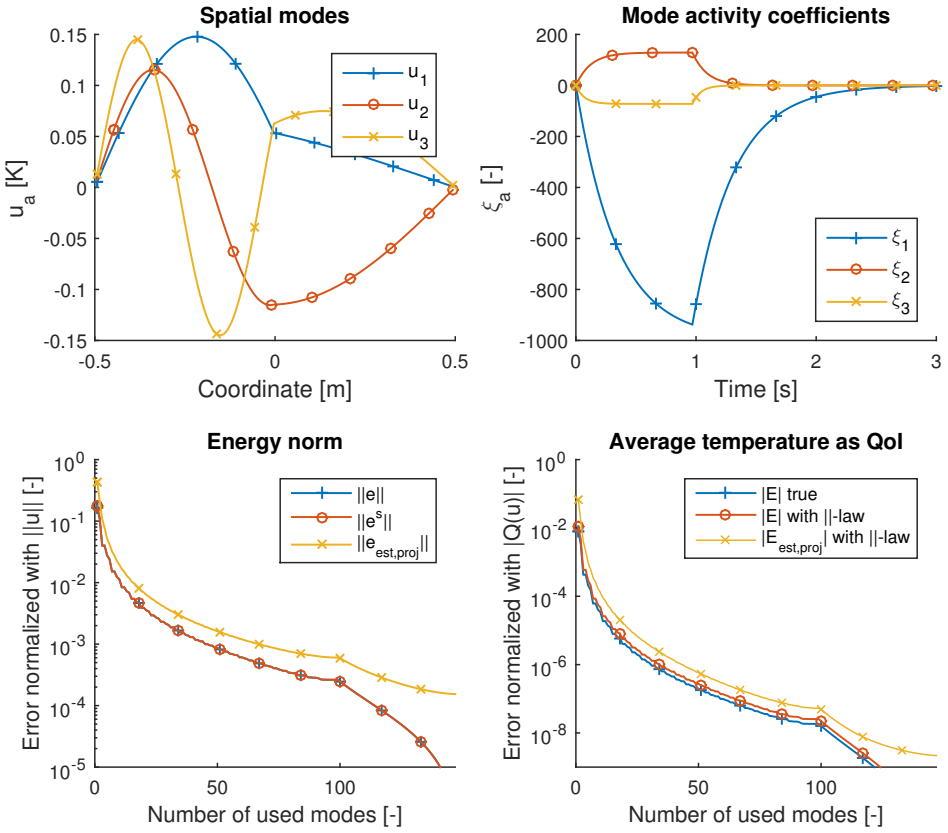
CHALMERS



Numerical Model Reduction with Error Estimation in Computational Homogenization of Transient Heat Flow

Master's thesis in Applied Mechanics

EMIL AGGESTAM

Department of Applied Mechanics
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2016

MASTER'S THESIS IN APPLIED MECHANICS

Numerical Model Reduction with Error Estimation in Computational Homogenization of Transient Heat Flow

EMIL AGGESTAM

Department of Applied Mechanics
Division of Material and Computational Mechanics
CHALMERS UNIVERSITY OF TECHNOLOGY

Göteborg, Sweden 2016

Numerical Model Reduction with
Error Estimation in Computational Homogenization of Transient Heat Flow
EMIL AGGESTAM

Master's thesis 2016:16
ISSN 1652-8557
Department of Applied Mechanics
Division of Material and Computational Mechanics
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: +46 (0)31-772 1000

Cover:
The front page figure is divided into four different sub-figures and is performed with a material that has a heterogeneous micro-structure. The figure is obtained from calculations on the micro-scale and the load case consist of a ramp of the estimated macro-scale temperature. The upper left sub-figure shows the first three eigenvectors that are used to estimate the spatial part of the modes. The upper right sub-figure shows the first three mode activity coefficients that, together with the spatial part of the modes, form the fluctuating solution inside the Representative Volume Element. The lower left sub-figure shows error estimates performed with the energy norm and the lower right sub-figure shows error estimates performed when the average temperature is the quantity of interest. The error estimates with the index $est, proj$ uses only the reduced modes, while the other error estimates uses all modes.

Chalmers Reproservice
Göteborg, Sweden 2016

Numerical Model Reduction with
Error Estimation in Computational Homogenization of Transient Heat Flow
Master's thesis in Applied Mechanics
EMIL AGGESTAM
Department of Applied Mechanics
Division of Material and Computational Mechanics
Chalmers University of Technology

Abstract

In his thesis, a multi-scale analysis of the transient heat flow problem is considered. By using First-order Computational Homogenization, applied on a Representative Volume Element, the micro-scale problem is successfully solved with Numerical Model Reduction. The computational cost of solving the micro-scale problem is reduced by reducing the number of linearly independent modes. The spatial part of the modes is estimated with Spectral Decomposition, which implies that the mode activity coefficients can be estimated by solving uncoupled ordinary differential equations.

An error analysis is performed in order to estimate the error that is introduced by reducing the number of modes. Error estimates are implemented that are based on either using all modes or only the reduced modes. The error is estimated with the energy norm as well as other quantities of interest such as the average temperature, the average heat flux in one direction and the final temperature. For the different quantities of interest, the error estimates are produced with the Galerkin orthogonality, together with either Cauchy–Schwarz inequality or the Parallelogram law. With these error estimates, the micro-scale problem can be solved as fast as possible by using the minimal number of modes that are needed to be within a given error tolerance.

The reduced model and the pertinent error estimates are verified for several test cases, where different materials and load cases are considered. These cases confirm (i) the gain in computational cost using the model reduction strategy and (ii) the robustness of the proposed error estimates.

Keywords: computational homogenization, transient heat flow, numerical model reduction, error analysis

Preface

The interest in multi-scale models has grown a lot over the past two decades. For non-linear and/or transient problems, nested finite element problems are typically obtained which are extremely computationally heavy. A theory to reduce the computational cost is to use Numerical Model Reduction (NMR). In this thesis, the theory behind NMR is further developed. Moreover, the thesis is also used as a partial fulfillment of the requirements for the MSc degree from Chalmers University of Technology.

The work was carried out during the period of January 2016 to June 2016 at the Department of Applied Mechanics, Division of Material and Computational Mechanics. Professor Kenneth Runesson and Professor Fredrik Larsson, both from the Department of Applied Mechanics, Division of Material and Computational Mechanics, were supervisors and Prof. Runesson was examiner as well.

Acknowledgements

I would like to give a special thanks to my supervisor Professor Fredrik Larsson for his vision and foresight regarding this project, and for always keeping his door open for questions and fruitful discussions. I would also like to thank my other supervisor and examiner Professor Kenneth Runesson. I wish to thank the colleagues at the Division of Dynamics and the Division of Material and Computational Mechanics. The work would not have been the same without the great atmosphere.

Finally, I wish to thank my fiancee Ida and my daughter Signe. Your love and support cannot be measured.

Nomenclature

Ω	Domain for the macro-scale problem
Γ	Boundary of the domain for the macro-scale problem
$\Omega_\square$	Domain for the micro-scale problem
$\Gamma_\square$	Boundary of the domain for the micro-scale problem
Γ_D	Dirichlet part of the boundary
Γ_N	Neumann part of the boundary
Π	Arbitrary potential
Π'	Directional derivative of an arbitrary potential
u	Temperature
$\bar{u}$	Macro-scale temperature
$\bar{\boldsymbol{g}}$	Gradient of the macro-scale temperature
u^M	Approximation of the macro-scale temperature within a Representative Volume Element
u^μ	Micro-scale temperature minus the estimated macro-scale temperature
δu	Test function
$\boldsymbol{x}$	Spatial coordinate
$\bar{\boldsymbol{x}}$	Center spatial location of a Representative Volume Element
t	Time
$\mathbb{E}$	Arbitrary vector space
$\mathbb{R}$	Space of all real numbers
$\mathbb{C}$	Space of all complex numbers
$\mathcal{C}$	Space of all continuous functions
$\mathcal{P}$	Space of all polynomials
$\mathbb{U}$	Sobolev space
$\mathbb{U}^0$	Sobolev space that is zero on the boundary of its domain
$\mathbb{U}_\square$	Sobolev space on the micro-scale
$\mathbb{U}_\square^0$	Sobolev space on the micro-scale that is zero on the boundary of its domain
$\mathbb{U}_\square^R$	$\mathbb{U}_\square$ for the reduced solution.
$\mathbb{U}_\square^{R,0}$	$\mathbb{U}_\square^0$ for the reduced solution.
$\mathcal{U}_\square$	Space-time Sobolev space on the micro-scale
$\mathcal{U}_\square^0$	Space-time Sobolev space on the micro-scale that is zero on the boundary of the spatial domain
$\mathcal{U}_\square^R$	$\mathcal{U}_\square$ for the reduced solution.
$\mathcal{U}_\square^{R,0}$	$\mathcal{U}_\square^0$ for the reduced solution.

$\mathbb{L}^2$	Space of square-integrable functions
$\boldsymbol{\nabla}$	Spatial differential operator
d_t	Derivative with respect to time
f	Heat source
$\boldsymbol{q}$	Heat flux field
Φ	Stored volume-specific internal energy
K	Conductivity
ρ	Density
c_p	Heat capacity at constant pressure
k	Thermal heat constant
$\boldsymbol{n}$	Normal
h	Neumann boundary value
g	Dirichlet boundary value
Φ_0	Initial condition
λ	Eigenvalue
$\boldsymbol{u}$	Eigenvector
ξ	Mode activity coefficient
$\underline{\boldsymbol{K}}$	Stiffness matrix
$\underline{\boldsymbol{M}}$	Mass matrix
$\boldsymbol{N}$	Base functions
$\underline{\boldsymbol{u}}$	Nodal values of u
$\delta\underline{\boldsymbol{u}}$	Nodal values of δu
e	Error between true and reduced solution
e^s	Symmetric representation of the true error
$Q_\square$	Quality of interest
E	Error for a quality of interest
$u^\star$	Solution to the dual problem
$u_R^\star$	Solution to the reduced dual problem
$e^\star$	Error of the dual problem
$e^{\star,s}$	Symmetric dual error
κ	Non-zero constant
$\in$	Symbol meaning in or contained
$\forall$	Symbol meaning for all

Contents

1 Introduction

This chapter gives a background and a problem description to the thesis, followed by the purpose and limitations. Thereafter the method is presented and finally the outline of the thesis is declared.

1.1 Background

During the last decades, the interest regarding mechanics of material on different length-scales has grown a lot [1]. The methods that are used to describe materials on different length-scales, called multi-scale methods, can work as a bridge between different scales and hence also be a bridge between mechanics of materials and material science [2]. Relationships derived with the multi-scale methods aims to predict properties on a seemingly smooth macro-scale by analyzing a heterogeneous micro-structure that might be multi-phased and/or anisotropic. The most well known method for finding such relationships is called Computational Homogenization and is nowadays well established, see [1–14].

Computational Homogenization analysis is typically produced on a so called Representative Volume Element (RVE). An RVE is a volume element that is large enough to be statistically representative, but still smaller than the smallest dimensions of the macro-scale [15]. This means that a seemingly homogeneous material on the macro-scale can be described by a heterogeneous RVE. The essence of the Computational Homogenization method is basically to solve a Boundary Value Problem (BVP) on an RVE. By solving the BVP local properties of the material at the macro-scale can be drawn. It exists other methods that can be used instead of the Computational Homogenization, e.g. the Taylor-Bishop-Hill estimates, asymptotic procedures and generalizations of self-consistent schemes [1, 16]. These methods are less established than the ordinary Computation Homogenization method, and therefore outside of the scope of this thesis.

1.2 Problem description

For simple problems, homogenization of the micro-structure can be used to obtain a closed-form constitutive relation of the macro-scale continuum [17]. However, for more complex problems a closed-form constitutive relation is not available and a FE solution is needed on both scales. In particular, for non-linear and/or transient problems each macro-scale quadrature point contains a unique FE-problem on the subscale [13]. This approach, known as FE^2, is tremendously computationally heavy for a fine macro-scale mesh. In order to reduce the computational cost of the problem, a reduced basis for the RVE-problem is usually introduced, called Numerical Model Reduction (NMR) [13, 18–22]. The disadvantage with NMR is that it introduces an error. In this thesis, the theory of NMR is used and extended by producing an a posteriori error estimate to the results.

1.3 Purpose

The purpose of this thesis is to reduce the computational cost of the RVE-problem by using Numerical Model Reduction and to apply a posteriori error estimation to the results.

1.4 Limitations

The study is mainly restricted as follows:

- Only the transient heat flow problem is considered. Moreover it is assumed that there is no heat source inside the Representative Volume Element.

- Linear heat flow is assumed. Other heat flow equations are not considered. This implies further that the Proper Orthogonal Decomposition method is not needed and the simpler Spectral Decomposition method can be used instead.
- More sophisticated Representative Volume Elements, such as the Statistical Volume Element and the Uncorrelated Volume Element, are not considered.
- The theory for the problem is derived in three spatial dimensions, but the numerical example is only implemented in one spatial dimension.
- The numerical example is implemented in MATLAB. Other programming languages are not considered in this thesis.
- Only First-order Computational Homogenization is applied. Higher-order Computational Homogenization, as well as other homogenization methods, are outside of the scope of this thesis.
- The boundary conditions on the RVE are restricted to the Dirichlet boundary condition.

1.5 Method

To tackle the problem description, analysis of similar previous studies needs to be done [23]. Therefore, the project work started with a literature study. In order to understand the most fundamental ideas behind multi-scale modeling, the Representative Volume Element and Computational Homogenization concepts were studied. Other considered theories were Spectral Decomposition, Proper Orthogonal Decomposition, Numerical Model Reduction, Functional analysis and the transient heat flow equation.

To process the knowledge from the literature study, the homogenization section (Chapter 2) was written simultaneously. Even more knowledge regarding the study was gained from weekly lectures given by the supervisors. From these lectures, the specific multi-scale problem for this thesis was introduced together with the concept of Numerical Model Reduction.

When the literature study was considered finished, a numerical implementation of the theory that was covered so far, started along with writing other parts of the thesis. Different loading scenarios were implemented to illustrate the results. When the micro-scale problem could be solved for an arbitrary number of modes, the theory behind error analysis was studied. Finally, with the theory for the error analysis in mind, proper error estimates of the result for different number of modes were incorporated in the numerical example.

1.6 Thesis outline

The thesis is divided into 7 different chapters. Chapter 2 presents the multi-scale problem and Chapter 3 shows how the micro-scale problem can be solved and how its computational cost can be reduced with Numerical Model Reduction. Chapter 4 presents the error analysis that is used to estimate the errors that are introduced by Numerical Model Reduction. The theory is verified in the numerical results, presented in Chapter 5. Finally, the summary and conclusions of the thesis can be found in Chapter 6 along with suggestions to further work in Chapter 7.

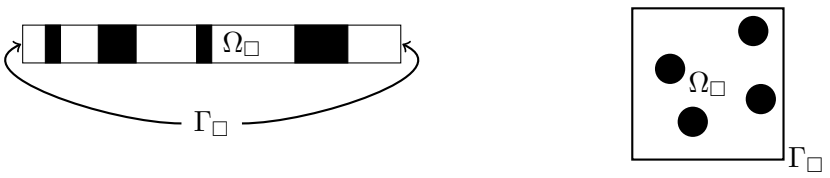
2 Homogenization of the transient heat flow problem

This chapter presents multi-scale modeling theory for the transient heat flow problem. The most basic theories that are needed to introduce the multi-scale problem are the concepts of a Representative Volume Element (RVE) and Computational Homogenization, which are given in Sections 2.1 and 2.2. The Hill-Mandel Macrohomogeneity condition, which is assumed to be valid in this thesis, is presented in Section 2.3. Section 2.4 presents the strong and weak form of the considered transient heat flow problem. With the concepts of an RVE and Computational Homogenization in mind, the macro- and micro-scale problem are defined in Section 2.5 and 2.6.

2.1 Representative volume element

In order to calculate micro-scale properties that are locally valid on the macro-scale, some kind of domain where the analysis can be done is needed [24]. The purpose of a Representative Volume Element (RVE), also called a representative elementary volume or unit cell, is to serve as such a domain. The RVE shall be representative independent of where it is spatially located and must hence catch all micro-scale properties such as grains, inclusions, voids, fibers, etc. [25]. All these heterogeneities introduce a lower limit of the length-scale of the RVE in order to make it statistically representative. But the RVE cannot be arbitrary large since it also shall serve as a volume element of continuum mechanics. The size of the RVE also affects the accuracy of the smoothness assumption in the Computation Homogenization procedure, see Section 2.2. For simplicity, the RVE is typically modeled as two different phases and the domain is a straight line in one dimension and a square in two dimensions, see Figure 2.1.

Throughout this thesis the properties of the traditional RVE is assumed to be valid. However, there are more general volume elements such are the Statistical Volume Element (SVE) and the Uncorrelated Volume Element (UVE). For the interested reader a short presentations of the SVE and UVE can be seen in Sections 2.1.1 and 2.1.2.



(a) *The one-dimensional model.* (b) *The two-dimensional model.*

Figure 2.1: *The figure shows how an RVE typically is modeled in one and two dimensions. The domain is denoted* $\Omega_\square$ *and the boundary* $\Gamma_\square$*. The RVE is assumed to consist of two different phases, which is indicated by the white and black areas.*

2.1.1 Statistical Volume Element

The criterion that the length-scale of the RVE needs to satisfy are not always possible to meet [9]. If there, for example, exists locally homogeneous or quasi-stationary random fields within the body, the traditional RVE concepts cannot be applied. A proposed way to solve this has been to make a large amount of scans of the micro-structure and from each scan compute an "RVE" that is valid in the exact position [26]. Even thought this method has qualified accuracy, it requires large amounts of both experimental data and pre-processing computations. In order to get a more computationally efficient approach, the concept of a Statistical Volume Element (SVE), also known as a stochastic volume element, is introduced. The purpose of the SVE is to capture randomness by producing simulation on a domain which is smaller than the traditional RVE, but still larger than the length-scale

of the micro-structure [27]. The simulations are performed by initially collecting micro-structural information at different macroscopic domains and from this predict spatial correlations between different random fields. Thanks to this procedure, the SVE can produce a more accurate prediction of material properties, which has been of great important in material science [28].

2.1.2 Uncorrelated Volume Element

Even thought the SVE can predict more accurate material properties compared to the RVE, it does not take the covariance between different micro-structures into account. Therefore, the Uncorrelated Volume Element (UVE) is introduced whose purpose is to produce its analysis on such a length-scale where the it can be assumed to be independent of its adjacent micro-structures [29]. The reason that the independence of adjacent micro-structure is an important assumption, is that the mathematical model becomes simpler. In some applications correlated volume elements introduces mesh dependency in the results and by using UVE, this dependency can be reduced.

2.2 Computational Homogenization

Computational Homogenization is a well established method used to combine features on different lengthscales of a material [1–14]. It is one of many multi-scale methods, which can be thought of as a bridge between mechanics of materials and material science [2]. The method is based on solving a Boundary Value Problem (BVP) on a Representative Volume Element (RVE) from which local properties of the material at the macro-scale can be drawn.

Let $\boldsymbol{x}$ denote the spatial coordinate, let $\bar{\boldsymbol{x}}$ denote the center point of a given RVE and let t denote time. The solution, u, to the BVP inside of the RVE can then be decomposed as

$$u(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) = u^M(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) + u^\mu(\bar{\boldsymbol{x}}, \boldsymbol{x}, t), \tag{2.1}$$

where u^M is a smooth approximation of the macro-scale solution and u^μ is a fluctuating term. If u^M is assumed to vary linearly, quadratic or cubic depends on what order of Computational Homogenization that is applied. For the First-order Computational Homogenization, u^M is assumed to be linear while u^M varies quadratically for the Second-order Computational Homogenization, see Sections 2.2.1 and 2.2.2.

From the decomposition of u into u^M and u^μ, the macro- and micro-scale can be combined in the following way. Kinematical quantities from the macro-scale are used to define u^M on the micro-scale. In the context of the transient heat equation, the needed quantities are typically the true macro-scale solution and its field gradient in the center point of the RVE. These quantities, that describes u^M, can then be used to define the BVP on the RVE. For simplicity, it is throughout this thesis assumed that the heat source is equal to zero and hence the macroscopic quantities serve as boundary condition to the BVP. Therefore, the BVP is well-defined and can be solved with, for example, finite elements. The solution to the BVP can then be used to draw conclusion regarding the macro-scale point in which the RVE lies by using standard mathematical averaging procedures.

Several properties in this thesis have both a macro-representation and a corresponding micro-representation. The relation between those are obtained from traditional volume averaging and denoted as follows. Let $\bar{\diamond}$ be the macroscopic representation of $\diamond$. The relation between those, for a given RVE, can then be defined as

$$\bar{\diamond} := \langle \diamond \rangle_\square := \frac{1}{|\Omega_\square|} \int_{\Omega_\square} \diamond \, \mathrm{d}\Omega, \tag{2.2}$$

and in particular

$$\bar{u} = \frac{1}{|\Omega_\square|} \int_{\Omega_\square} u \, \mathrm{d}\Omega. \tag{2.3}$$

The order of Computational Homogenization affects how well the approximated solution u^M reflects the true macro-scale solution $\bar{u}$. The difference $u^M - \bar{u}$ is generally non-zero inside of each RVE, which can be seen as a model error introduced by the Computational Homogenization method. Another property that affects how well u^M reflects $\bar{u}$ is the size of the RVE. The model error that is introduced from the regularity condition of u^M increases with the deviation from the center of the RVE. A larger RVE, introduces a larger distance from its center to the edge and therefore the regularity assumption becomes more severe.

2.2.1 First-order Computational Homogenization

The most common homogenization method was introduced for almost 30 years ago and is called First-order Computational Homogenization [2]. The key assumption is that the macro-scale field, in this thesis typically the temperature field, varies linear within a given RVE [15], see Figure 2.2. This means that u^M can be written as a first-order Taylor expansion of $\bar{u}$, i.e.

$$u^M(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) = \bar{u}(\bar{\boldsymbol{x}}, t) + \bar{\boldsymbol{g}}(\bar{\boldsymbol{x}}, t) \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}), \tag{2.4}$$

where $\bar{\boldsymbol{g}} = \boldsymbol{\nabla}\bar{u}$ is the macroscopic field gradient.

2.2.2 Higher-order Computational Homogenization

For materials with high gradients on the macro-scale, First-order Computational Homogenization is no longer appropriate [19, 30]. Since the macroscopic fields can vary rapidly, the linear assumption on u^M might introduce severe model errors. A way to solve this is to add additional terms from the Taylor expansion of $\bar{u}$ when u^M is defined. For the Second-order Computational Homogenization, u^M is defined as

$$u^M(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) = \bar{u}(\bar{\boldsymbol{x}}, t) + \bar{\boldsymbol{g}}(\bar{\boldsymbol{x}}, t) \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}) + (\boldsymbol{x} - \bar{\boldsymbol{x}}) \cdot \boldsymbol{\nabla}\bar{\boldsymbol{g}}(\bar{\boldsymbol{x}}, t) \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}). \tag{2.5}$$

Even though the result for Higher-order Computational Homogenization is more accurate, it is also more complicated to implement. This thesis is restricted to First-order Computational Homogenization.

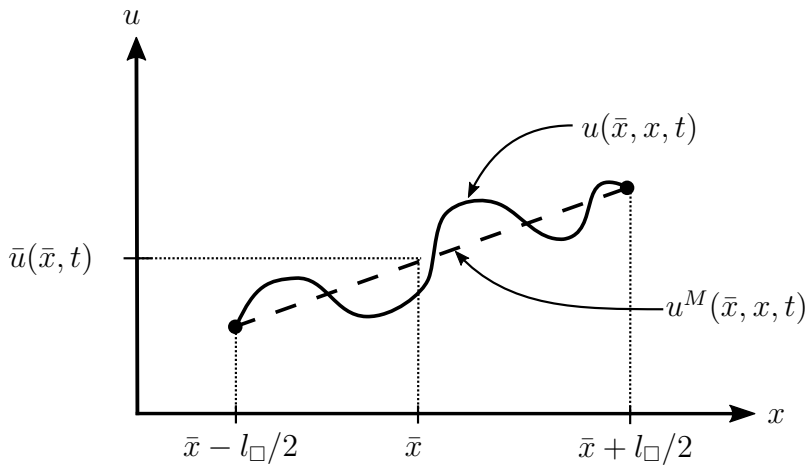


Figure 2.2: *The figure shows a visual interpretation of First-order Computational Homogenization in one dimension when it is assumed that $u^\mu = 0$ on the boundary (Dirichlet b.c).*

2.3 Hill-Mandel Macrohomogeneity condition

In order to derive the relations between the macro- and micro-scale, it is throughout this thesis assumed that the so-called Hill-Mandel Macrohomogeneity condition is fulfilled. The condition can be stated as follows and is illustrated in Figure 2.3. Consider an arbitrary potential $\Pi(u)$ that shall be minimized with respect to an arbitrary field variable u. By setting the directional derivative equals zero and restrict the test variable, δu, to be determined by the macro-scale, the expression $\Pi'(u\{u^M\}, \delta u^M) = 0$ is obtained[1]. This expression can be reached in two different ways, see Figure 2.3. The Hill-Mandel Macrohomogeneity condition says that the expression $\Pi'(u\{u^M\}, \delta u^M) = 0$ should be the same, independent if it is reached thought arrow (1) and (2) or (3) and (4).

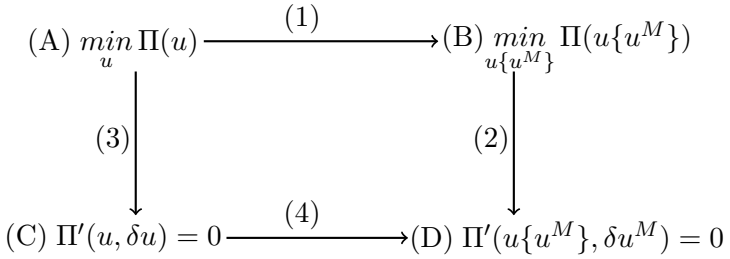


Figure 2.3: *The figure shows a schematic illustration of different ways of getting from statement (A) to statement (D). Path (A)-(B)-(D) denotes that the restriction of the test variable δu^M is applied prior to the directorial derivative concepts, while path (A)-(B)-(D) denotes the other way around.*

Note that $\Pi'(u\{u^M\}, \delta u^M)$ is the directional derivative of an arbitrary energy measure. The most common example when the Hill-Mandel Macrohomogeneity condition is applied, is when linear elasticity is considered. For this special case of the Hill-Mandel Macrohomogeneity condition, it can be shown that fulfilling the condition is equivalent with the condition that the virtual work on the macro-scale must be equal to the virtual work on the micro-scale [31].

2.4 Transient heat flow

In Section 2.4.1 the considered strong form is introduced and in Section 2.4.2 the corresponding variational form is derived. Simplifications, regarding the transient heat flow problem, that are assumed in this thesis is presented in Section 2.4.3.

2.4.1 Introducing the strong form

Let Ω be a spatial domain in which the transient heat flow equation is considered. For times between 0 and T, the strong form of the transient heat equation is given by [15]

$$\mathrm{d}_t \Phi + \boldsymbol{q} \cdot \boldsymbol{\nabla} = f \quad \text{in } \Omega \times (0, T], \tag{2.6}$$

where $\Phi = \Phi(u)$ is the stored volume-specific internal energy, $\boldsymbol{q}$ is the heat flux field, $\boldsymbol{\nabla}$ is the spatial gradient with respect to coordinate $\boldsymbol{x}$ in Ω and f is the volume-specific heat source within Ω. The boundary of Ω, denoted Γ, can be divided into one part where the temperature is prescribed and one part where the heat flux through the boundary is described. The part of the boundary with prescribed temperature, known as a Dirichlet boundary, is denoted Γ_D and similarly the part of the boundary with prescribed heat flux, known as a Neumann boundary, is denoted Γ_N. This implies

[1] The curly brackets indicate implicit functional dependence.

that the boundary can be decomposed as $\Gamma = \Gamma_D + \Gamma_N$. The boundary condition of the transient heat flow problem, together with an initial condition on Φ, can then be written as

$$\begin{cases} \boldsymbol{q} \cdot \boldsymbol{n} = h & \text{on } \Gamma_N \times (0,T], \\ u = g & \text{on } \Gamma_D \times (0,T], \\ \Phi = \Phi_0 & \text{in } \Omega\big|_{t=0}, \end{cases} \tag{2.7}$$

where $\boldsymbol{n}$ is the normal to Γ_N and h, g and Φ_0 are given constants.

2.4.2 Derivation of the variational form

Let $\mathbb{U}$ and $\mathbb{U}^0$ be two Sobolev spaces satisfying[2]

$$\begin{aligned} \mathbb{U} &= \Big\{ u : \int_\Omega \left(|u|^2 + |\boldsymbol{\nabla} u|^2 \right) \mathrm{d}\Omega \Big\}, \\ \mathbb{U}^0 &= \Big\{ u : \int_\Omega \left(|u|^2 + |\boldsymbol{\nabla} u|^2 \right) \mathrm{d}\Omega,\ u = 0 \text{ on } \Gamma_D \Big\}. \end{aligned} \tag{2.8}$$

Consider the strong form given in Equation (2.6) and let $u \in \mathbb{U}$. A step towards the equivalent variational form is obtained by multiplying the strong form with a test function, $\delta u \in \mathbb{U}^0$, and integrate over the domain as

$$\int_\Omega \mathrm{d}_t \Phi \delta u \, \mathrm{d}\Omega + \int_\Omega \boldsymbol{q} \cdot \boldsymbol{\nabla} \delta u \, \mathrm{d}\Omega = \int_\Omega f \delta u \, \mathrm{d}\Omega. \tag{2.9}$$

Integration by parts implies that

$$(\delta u \boldsymbol{q}) \cdot \boldsymbol{\nabla} = (\boldsymbol{\nabla} \delta u) \cdot \boldsymbol{q} + (\boldsymbol{q} \cdot \boldsymbol{\nabla}) \delta u. \tag{2.10}$$

Using this relation on the second term in Equation (2.9) gives that

$$\int_\Omega \mathrm{d}_t \Phi \delta u \, \mathrm{d}\Omega + \int_\Omega (\delta u \boldsymbol{q}) \cdot \boldsymbol{\nabla} - (\boldsymbol{\nabla} \delta u) \cdot \boldsymbol{q} \, \mathrm{d}\Omega = \int_\Omega f \delta u \, \mathrm{d}\Omega, \tag{2.11}$$

which is equivalent to

$$\int_\Omega \mathrm{d}_t \Phi \delta u \, \mathrm{d}\Omega - \int_\Omega (\boldsymbol{\nabla} \delta u) \cdot \boldsymbol{q} \, \mathrm{d}\Omega = \int_\Omega f \delta u \, \mathrm{d}\Omega - \int_\Omega (\delta u \boldsymbol{q}) \cdot \boldsymbol{\nabla} \, \mathrm{d}\Omega. \tag{2.12}$$

By using Gauss divergence theorem, the last term in Equation (2.12) can be written as

$$\int_\Omega (\delta u \boldsymbol{q}) \cdot \boldsymbol{\nabla} \, \mathrm{d}\Omega = \int_\Gamma \delta u \boldsymbol{q} \cdot \boldsymbol{n} \, \mathrm{d}\Gamma. \tag{2.13}$$

That the test function δu lies in $\mathbb{U}^0$ simplifies the expression ever further as

$$\int_\Gamma \delta u \boldsymbol{q} \cdot \boldsymbol{n} \, \mathrm{d}\Gamma = \int_{\Gamma_N} \delta u \boldsymbol{q} \cdot \boldsymbol{n} \, \mathrm{d}\Gamma =: \int_{\Gamma_N} \delta u h \, \mathrm{d}\Gamma, \tag{2.14}$$

where h is a constant. The weak form can now be written as: Find $u \in \mathbb{U}$ such that

$$\int_\Omega \mathrm{d}_t \Phi \delta u \, \mathrm{d}\Omega - \int_\Omega (\boldsymbol{\nabla} \delta u) \cdot \boldsymbol{q} \, \mathrm{d}\Omega = \int_\Omega f \delta u \, \mathrm{d}\Omega - \int_{\Gamma_N} \delta u h \, \mathrm{d}\Gamma \quad \forall \delta u \in \mathbb{U}^0. \tag{2.15}$$

[2]For information about Sobolev spaces, see Appendix A.

2.4.3 Simplification from the general transient heat flow

From the general weak form, given in Equation (2.15), some simplifications are applied. First of all it is assumed that there is no heat source within Ω, i.e. $f = 0$. Moreover it is also assumed that the stored internal energy $\Phi(u)$ varies linear with u, i.e. $\Phi(u) = cu$ where c is a constant. Physically, $c = \rho c_p$, where ρ is the density and c_p is the heat capacity at constant pressure. Finally, also Fourier's law is assumed to be valid for the heat flux field, i.e. $\boldsymbol{q}(u, \boldsymbol{\nabla} u) = -\boldsymbol{K} \cdot \boldsymbol{\nabla} u$, where $\boldsymbol{K}$ is the thermal conductivity. From these simplification, the weak form can be simplified into: Find $u \in \mathbb{U}$ such that

$$\int_\Omega c\dot{u}\delta u \,\mathrm{d}\Omega + \int_\Omega (\boldsymbol{\nabla}\delta u) \cdot \boldsymbol{K} \cdot (\boldsymbol{\nabla} u) \,\mathrm{d}\Omega = -\int_{\Gamma_N} \delta u h \,\mathrm{d}\Gamma \quad \forall \delta u \in \mathbb{U}^0, \tag{2.16}$$

where $\mathrm{d}_t u = \dot{u}$. By introducing

$$\begin{cases} \mathfrak{m}(u, \delta u) = \int_\Omega cu\delta u \,\mathrm{d}\Omega, \\ \mathfrak{a}(u, \delta u) = \int_\Omega (\boldsymbol{\nabla} u) \cdot \boldsymbol{K} \cdot (\boldsymbol{\nabla}\delta u) \,\mathrm{d}\Omega, \\ \mathfrak{l}(\delta u) = -\int_{\Gamma_N} \delta u h \,\mathrm{d}\Gamma, \end{cases} \tag{2.17}$$

the weak form can be written as: Find $u \in \mathbb{U}$ such that

$$\mathfrak{m}(\dot{u}, \delta u) + \mathfrak{a}(u, \delta u) = \mathfrak{l}(\delta u), \quad \forall \delta u \in \mathbb{U}^0, \tag{2.18}$$

with the initial condition

$$u\big|_{t=0} = \frac{\Phi_0}{c}. \tag{2.19}$$

2.5 Macro-scale problem

From the Computational Homogenization method it is shown that u^M consists of a truncated Taylor expansion of $\bar{u}$, see Section 2.2. Let $\mathcal{A} : \mathbb{R} \to \mathbb{R}$ be an operator satisfying

$$u^M(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) = \big(\mathcal{A}\bar{u}\big)(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) = \bar{u}(\bar{\boldsymbol{x}}, t) + \bar{\boldsymbol{g}}(\boldsymbol{x}, t) \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}). \tag{2.20}$$

The smoothness of u^M implies it lies in $\mathbb{U}$ and that it is possible to define a corresponding test function $\mathcal{A}\delta\bar{u} \in \mathbb{U}^0$. The regularity of u^M, together with the Hill-Mandel Macrohomogeneity condition, imply that Equation (2.16) can be rewritten as: Find $\bar{u} \in \mathbb{U}$ such that

$$\int_\Omega \Big[\frac{1}{|\Omega_\square|}\int_{\Omega_\square} c\dot{u}\mathcal{A}\delta\bar{u} \,\mathrm{d}\Omega + \frac{1}{|\Omega_\square|}\int_{\Omega_\square} (\boldsymbol{\nabla}\mathcal{A}\delta\bar{u}) \cdot \boldsymbol{K} \cdot (\boldsymbol{\nabla} u) \,\mathrm{d}\Omega\Big] \,\mathrm{d}\Omega = -\int_{\Gamma_N} \mathcal{A}\delta\bar{u} h \,\mathrm{d}\Gamma \quad \forall \delta u \in \mathbb{U}^0, \tag{2.21}$$

Let

$$\begin{cases} \mathfrak{m}_\square(u, v) = \frac{1}{|\Omega_\square|}\int_{\Omega_\square} cuv \,\mathrm{d}\Omega & \forall u, v \in \mathbb{U}, \\ \mathfrak{a}_\square(u, v) = \frac{1}{|\Omega_\square|}\int_{\Omega_\square} \big[\boldsymbol{\nabla} u\big] \cdot \boldsymbol{K} \cdot \big[\boldsymbol{\nabla} v\big] \,\mathrm{d}\Omega & \forall u, v \in \mathbb{U}. \end{cases} \tag{2.22}$$

The integrand of the first term can then be written as

$$\mathfrak{m}_\square(u, \mathcal{A}\delta\bar{u}) = \frac{1}{|\Omega_\square|}\int_{\Omega_\square} cu\big[\delta\bar{u} + \delta\bar{\boldsymbol{g}} \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}})\big] \,\mathrm{d}\Omega = \frac{1}{|\Omega_\square|}\int_{\Omega_\square} cu \,\mathrm{d}\Omega\delta\bar{u} + \frac{1}{|\Omega_\square|}\int_{\Omega_\square} cu(\boldsymbol{x} - \bar{\boldsymbol{x}}) \,\mathrm{d}\Omega \cdot \delta\bar{\boldsymbol{g}} \tag{2.23}$$

and introducing

$$\begin{cases} \frac{1}{|\Omega_\square|}\int_{\Omega_\square} cu \,\mathrm{d}\Omega\delta\bar{u} = \langle cu\rangle_\square \delta\bar{u} =: \bar{\Phi}\delta\bar{u} \\ \frac{1}{|\Omega_\square|}\int_{\Omega_\square} cu(\boldsymbol{x} - \bar{\boldsymbol{x}}) \,\mathrm{d}\Omega \cdot \delta\bar{\boldsymbol{g}} = \langle cu(\boldsymbol{x} - \bar{\boldsymbol{x}})\rangle_\square \cdot \delta\bar{\boldsymbol{g}} =: \bar{\bar{\boldsymbol{\Phi}}}\delta\bar{\boldsymbol{g}} \end{cases} \tag{2.24}$$

gives

$$\mathfrak{m}_{\square}(u, \mathcal{A}\delta\bar{u}) = \bar{\Phi}\delta\bar{u} + \bar{\bar{\boldsymbol{\Phi}}}\delta\bar{\boldsymbol{g}}. \tag{2.25}$$

In the same manner, the integrand of the second term can be written as

$$\begin{aligned}\mathfrak{a}_{\square}(u, \mathcal{A}\delta u) &= \frac{1}{|\Omega_{\square}|}\int_{\Omega_{\square}} \left[\boldsymbol{\nabla} u\right] \cdot \boldsymbol{K} \cdot \left[\boldsymbol{\nabla}\mathcal{A}\delta\bar{u}\right] \mathrm{d}\Omega\, \mathrm{d}\Omega = -\frac{1}{|\Omega_{\square}|}\int_{\Omega_{\square}} \boldsymbol{q} \cdot \delta\bar{\boldsymbol{g}}\, \mathrm{d}\Omega \\ &= -\frac{1}{|\Omega_{\square}|}\int_{\Omega_{\square}} \boldsymbol{q}\, \mathrm{d}\Omega \cdot \delta\bar{\boldsymbol{g}} = -\langle \boldsymbol{q}\rangle_{\square} \cdot \delta\bar{\boldsymbol{g}} =: -\bar{\boldsymbol{q}} \cdot \delta\bar{\boldsymbol{g}}.\end{aligned} \tag{2.26}$$

By assuming that the linear part of the Taylor expansion can be neglected on the boundary the Neumann boundary term can be written as

$$-\int_{\Gamma_N} \mathcal{A}\delta\bar{u}h\, \mathrm{d}\Gamma = -\int_{\Gamma_N} \delta\bar{u}h\, \mathrm{d}\Gamma. \tag{2.27}$$

The variational form can now be simplified to: Find $\bar{u} \in \mathbb{U}$ such that

$$\int_{\Omega} \left[\dot{\Phi}\delta\bar{u} + \dot{\bar{\bar{\boldsymbol{\Phi}}}} \cdot \delta\bar{\boldsymbol{g}} - \bar{\boldsymbol{q}} \cdot \delta\bar{\boldsymbol{g}}\right] \mathrm{d}\Omega = -\int_{\Gamma_N} \delta\bar{u}h\, \mathrm{d}\Gamma \quad \forall \delta\bar{u} \in \mathbb{U}^0. \tag{2.28}$$

Note that the fact that $\mathcal{A}\delta\bar{u} \in \mathbb{U}^0$ implies that $\delta\bar{u}, \delta\bar{\boldsymbol{g}} \in \mathbb{U}^0$ since $\mathcal{A}\delta\bar{u} = \bar{u} + \bar{\boldsymbol{g}} \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}})$.

Remark: By using Gauss divergence theorem and integration by parts, the variational form can be rewritten as: Find $\bar{u} \in \mathbb{U}$ such that

$$\int_{\Omega} \left[\dot{\bar{\Phi}} - \boldsymbol{\nabla} \cdot \dot{\bar{\bar{\boldsymbol{\Phi}}}} + \boldsymbol{\nabla} \cdot \bar{\boldsymbol{q}}\right] \delta\bar{u}\, \mathrm{d}\Omega = 0 \quad \forall \delta\bar{u} \in \mathbb{U}^0. \tag{2.29}$$

This shall be true for an arbitrary test function that lies in $\mathbb{U}^0$ on an arbitrary domain Ω. This implies that

$$\dot{\bar{\Phi}} - \boldsymbol{\nabla} \cdot \dot{\bar{\bar{\boldsymbol{\Phi}}}} + \boldsymbol{\nabla} \cdot \bar{\boldsymbol{q}} = 0, \tag{2.30}$$

which is the corresponding strong form of Equation (2.28).

2.6 Micro-scale problem (RVE-problem)

From Section 2.2 it is shown how the solution to the BVP inside each RVE can be decomposed as

$$u(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) = u^M(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) + u^{\mu}(\bar{\boldsymbol{x}}, \boldsymbol{x}, t). \tag{2.31}$$

By assuming that $u^{\mu}(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) = 0$ on the boundary of the RVE, the temperature is equal to the macro-scale temperature on the boundary. Since also First-order Computational Homogenization is assumed, see Section 2.2.1, the equation

$$u^M(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) = \bar{u}(\bar{\boldsymbol{x}}, t) + \bar{\boldsymbol{g}}(\boldsymbol{x}, t) \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}) \tag{2.32}$$

must hold. The purpose of the micro-scale problem is to, for given $\bar{u}$ and $\bar{\boldsymbol{g}}$, determine the three quantities $\boldsymbol{q}$, $\bar{\Phi}$ and $\bar{\bar{\boldsymbol{\Phi}}}$ defined in Section 2.5, see Figure 2.4.

Consider the two Sobolev spaces that define the test and solution space of the macro-problem given in Equation (2.8). In order to define the variational form of the micro-scale problem, the corresponding two spaces on the micro-scale need to be defined. Both the test and solution space must still be suitable Sobolev spaces in order for Lax-Milgrams theorem to hold[3]. Also the criteria

[3]For details about why Lax-Milgrams theorem needs to be fulfilled, see Appendix A.

Figure 2.4: *The figure shows a schematic flow of what variables that are needed and calculated in the micro-scale problem.*

that the test function must be zero on the boundary must hold. The new condition, compared to the spaces on the macro-scale is that the boundary on the solution space must fulfill conditions given by the macro-scale solution. The solution and test space for the micro-scale problem, can then be defined as

$$\begin{aligned}\mathbb{U}_\square &= \Big\{u:\ \int_\Omega \big(|u|^2 + |\boldsymbol{\nabla} u|^2\big)\,\mathrm{d}\Omega,\ u = u^M \text{ on } \Gamma_D\Big\},\\ \mathbb{U}^0_\square &= \Big\{u:\ \int_\Omega \big(|u|^2 + |\boldsymbol{\nabla} u|^2\big)\,\mathrm{d}\Omega,\ u = 0 \text{ on } \Gamma_D\Big\}.\end{aligned} \tag{2.33}$$

By using the variational form given in Equation (2.16), with the test function $\delta u \in \mathbb{U}^0_\square$, the equation

$$\frac{1}{|\Omega_\square|}\int_{\Omega_\square} \delta u c \dot{u}\,\mathrm{d}\Omega - \frac{1}{|\Omega_\square|}\int_{\Omega_\square} (\boldsymbol{\nabla}\delta u)\cdot \boldsymbol{q}\,\mathrm{d}\Omega = -\int_{\Gamma_N} \delta u h\,\mathrm{d}\Gamma, \quad \forall \delta u \in \mathbb{U}^0_\square \tag{2.34}$$

is obtained. For simplicity, h is assumed to be zero and therefore

$$\frac{1}{|\Omega_\square|}\int_{\Omega_\square} \delta u c \dot{u}\,\mathrm{d}\Omega - \frac{1}{|\Omega_\square|}\int_{\Omega_\square} (\boldsymbol{\nabla}\delta u)\cdot \boldsymbol{q}\,\mathrm{d}\Omega = 0, \quad \forall \delta u \in \mathbb{U}^0_\square. \tag{2.35}$$

The definitions of $\mathfrak{m}_\square$ and $\mathfrak{a}_\square$, given in Equation (2.22), imply that the equation can be written as

$$\mathfrak{m}_\square(\dot{u}, \delta u) + \mathfrak{a}_\square(u, \delta u) = 0, \quad \forall \delta u \in \mathbb{U}^0_\square. \tag{2.36}$$

By using the expression of u in Equation (2.31), the weak form of the micro-problem can be written as: Find $u^\mu \in \mathbb{U}_\square$ such that

$$\begin{aligned}\mathfrak{m}_\square(\dot{u}^\mu, \delta u) + \mathfrak{a}_\square(u^\mu, \delta u) &= -\mathfrak{m}_\square(\dot{u}^M, \delta u) - \mathfrak{a}_\square(u^M, \delta u)\\ &= -\mathfrak{m}_\square(\dot{\bar{u}}, \delta u) - \mathfrak{m}_\square(\dot{\bar{\boldsymbol{g}}}\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), \delta u) - \mathfrak{a}_\square(\bar{u}, \delta u) - \mathfrak{a}_\square(\bar{\boldsymbol{g}}\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), \delta u)\\ &= -\mathfrak{m}_\square(1, \delta u)\dot{\bar{u}} + \Big(\sum_{i=1}^{d} -\mathfrak{m}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), \delta u)\boldsymbol{e}_i\Big)\cdot\dot{\bar{\boldsymbol{g}}}\\ &\quad + \Big(\sum_{i=1}^{d} -\mathfrak{a}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), \delta u)\boldsymbol{e}_i\Big)\cdot\bar{\boldsymbol{g}}, \quad \forall \delta u \in \mathbb{U}^0_\square.\end{aligned} \tag{2.37}$$

Recall that the goal is to calculate $\boldsymbol{q}$, $\bar{\Phi}$ and $\bar{\bar{\boldsymbol{\Phi}}}$, which are given by

$$\begin{cases}\bar{\boldsymbol{q}} = \frac{1}{|\Omega_\square|}\int_{\Omega_\square} \boldsymbol{q}\,\mathrm{d}\Omega\\ \bar{\Phi} = \frac{1}{|\Omega_\square|}\int_{\Omega_\square} cu\,\mathrm{d}\Omega\\ \bar{\bar{\boldsymbol{\Phi}}} = \frac{1}{|\Omega_\square|}\int_{\Omega_\square} cu(\boldsymbol{x}-\bar{\boldsymbol{x}})\,\mathrm{d}\Omega.\end{cases} \tag{2.38}$$

From the weak form u^μ is calculated and $\boldsymbol{q}$, $\bar{\Phi}$ and $\bar{\bar{\boldsymbol{\Phi}}}$ can then be found from

$$\bar{\boldsymbol{q}} = \mathfrak{a}_\square(u, 1) = \mathfrak{a}_\square(u^M, 1) + \mathfrak{a}_\square(u^\mu, 1) = \boldsymbol{q}_{\bar{u}}\bar{u} + \boldsymbol{q}_{\bar{\boldsymbol{g}}}\cdot\bar{\boldsymbol{g}} + \mathfrak{a}_\square(u^\mu, 1), \tag{2.39}$$

$$\bar{\Phi}(u) = \mathfrak{m}_{\square}(cu, 1) = \mathfrak{m}_{\square}(cu^M, 1) + \mathfrak{m}_{\square}(cu^\mu, 1) = \bar{\Phi}_{\bar{u}}\bar{u} + \bar{\boldsymbol{\Phi}}_{\bar{\boldsymbol{g}}} \cdot \bar{\boldsymbol{g}} + \mathfrak{m}_{\square}(cu^\mu, 1), \tag{2.40}$$

and

$$\begin{aligned}\bar{\bar{\boldsymbol{\Phi}}}(u) &= \sum_{i=1}^{d} \mathfrak{m}_{\square}(u, \boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{e}_i = \sum_{i=1}^{d} \mathfrak{m}_{\square}(u^M, \boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{e}_i \\ &+ \sum_{i=1}^{d} \mathfrak{m}_{\square}(u^\mu, \boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{e}_i = \bar{\bar{\boldsymbol{\Phi}}}_{\bar{u}}\bar{u} + \bar{\bar{\boldsymbol{\Phi}}}_{\bar{\boldsymbol{g}}} \cdot \bar{\boldsymbol{g}} + \sum_{i=1}^{d} \mathfrak{m}_{\square}(u^\mu, \boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}))\boldsymbol{e}_i,\end{aligned} \tag{2.41}$$

where $\bar{\Phi}_{\bar{u}}$ is a given constant, $\boldsymbol{q}_{\bar{u}}$, $\bar{\boldsymbol{\Phi}}_{\bar{\boldsymbol{g}}}$ and $\bar{\bar{\boldsymbol{\Phi}}}_{\bar{u}}$ are given vectors and $\boldsymbol{q}_{\bar{\boldsymbol{g}}}$ respective $\bar{\bar{\boldsymbol{\Phi}}}_{\bar{\boldsymbol{g}}}$ are given matrices. Note that if u^μ is calculated, $\bar{\boldsymbol{q}}$, $\bar{\Phi}$ and $\bar{\bar{\boldsymbol{\Phi}}}$ can then be determined from Equations (2.39)-(2.41). Therefore, focus is this thesis is centered around calculating u^μ.

Remark: From the definition of the operation $\mathcal{A}$ in Section 2.5, Equation (2.31) can be written as

$$u(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) = \big(\mathcal{A}\bar{u}\big)(\bar{\boldsymbol{x}}, \boldsymbol{x}, t) + u^\mu(\bar{\boldsymbol{x}}, \boldsymbol{x}, t). \tag{2.42}$$

For the First-order Computational Homogenization, the conditions

$$\begin{cases} \bar{u} = \frac{1}{|\Omega_\square|} \int_{\Omega_\square} u \, \mathrm{d}\Omega \\ \bar{\boldsymbol{g}} = \frac{1}{|\Omega_\square|} \int_{\Omega_\square} \boldsymbol{\nabla} u \, \mathrm{d}\Omega = \frac{1}{|\Omega_\square|} \int_{\Gamma_\square} \boldsymbol{n} u \, \mathrm{d}\Gamma \end{cases} \tag{2.43}$$

must hold. Note that Gauss divergence theorem is applied to the second equation. Let $\mathcal{A}^\star : \mathbb{R} \to \mathbb{R}$ be an operator such that

$$\bar{u} = \mathcal{A}^\star u. \tag{2.44}$$

In order to fulfill Equation (2.43) the condition

$$\bar{u}\{u\} = \mathcal{A}^\star u = \mathcal{A}^\star(u^M + u^\mu) = \mathcal{A}^\star u^M + \mathcal{A}^\star u^\mu = \mathcal{A}^\star\mathcal{A}\bar{u} + \mathcal{A}^\star u^\mu \overset{!}{=} \bar{u} \tag{2.45}$$

must hold. Thus $\mathcal{A}^\star\mathcal{A}\bar{u} + \mathcal{A}^\star u^\mu = \bar{u}$ and the conditions

$$\begin{cases} \mathcal{A}^\star\mathcal{A} = \boldsymbol{I} \\ \mathcal{A}u^\mu = 0 \end{cases} \tag{2.46}$$

must hold thoughtout the derivations.

3 Numerical Model Reduction of the micro-scale problem

In this chapter, Numerical Model Reduction (NMR) is used to solve the micro-scale problem. Section 3.1 presents fundamental properties and assumptions in the NMR method. Section 3.2 shows how the micro-scale problem can be solved by letting the solution consist of a sum of modes in space and time. Finally, Section 3.3 shows how a reduced solution of the micro-scale problem can be calculated with NMR.

3.1 Numerical Model Reduction

In order to solve a FE^2-problem, it would be highly appreciated if the computational cost could be reduced. A starting point in order to solve this computational issue was introduced by Dvorak and Benveniste in 1992 [32]. In their work they introduced the so-called Transformation Field Analysis (TFA), where they showed that it is possible to reduce the number of macroscopic internal variables if it is assumed that the microscopic field of internal variables are piecewise uniform. The problem with TFA is that it cannot capture the material behavior of the plastic strain field without introducing a prohibitively high number of internal variables. In order to solve this issue, the Non-uniform Transformation Field Analysis (NTFA) was developed [33–35]. The key idea behind NTFA is that so-called inelastic modes $\boldsymbol{\epsilon}_\alpha(\boldsymbol{x})$ are introduced, which are spatially heterogeneous and time independent. By also introducing time dependent internal variables, called mode activity coefficients and denoted $\xi_\alpha(t)$, it is assumed that the plastic strain $\boldsymbol{\epsilon}^p(\boldsymbol{x},t)$ can be written as

$$\boldsymbol{\epsilon}^p(\boldsymbol{x},t) = \sum_{\alpha=1}^{N} \boldsymbol{\epsilon}_\alpha(\boldsymbol{x})\xi_\alpha(t), \tag{3.1}$$

where N is the number of inelastic modes. The method was further improved by Fritzen and Leuschner who extended the method such that more complex material models could be used [36]. Nowadays the method is well established, see [13, 18, 20, 22].

In this thesis, when the transient heat flow problem is considered, the corresponding decomposition is given by

$$u^\mu(\boldsymbol{x},t) = \sum_{a=1}^{N} u_a(\boldsymbol{x})\xi_a(t), \tag{3.2}$$

where u^μ is the fluctuating term on the micro-scale, see Section 2.2. The modes u_a need to be estimated in order to calculate ξ_a. This can be done with Spectral Decomposition or Proper Orthogonal Decomposition, which is presented in Sections 3.1.1 and 3.1.2.

The concept of Numerical Model Reduction (NMR) uses the same approach as NTFA, but are reducing the computational cost by not using all modes. This implies that the decomposition of the temperature can be written as

$$u^\mu(\boldsymbol{x},t) = \sum_{a=1}^{N_R} u_a(\boldsymbol{x})\xi_a(t), \tag{3.3}$$

where $1 \leq N_R < N$.

3.1.1 Spectral Decomposition

The key idea that is used in the Spectral Decomposition is that a diagonalizable matrix can be represented with its eigenvalues and eigenvectors [37]. Let $\boldsymbol{A}$ be a square matrix. An eigenvector, $\boldsymbol{u}$,

to $\boldsymbol{A}$ is a non-zero vector that satisfies the classical eigenvalue equation given by [38]

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}, \tag{3.4}$$

where λ is an eigenvalue to $\boldsymbol{A}$. The eigenvalue decomposition theorem claims that if $\boldsymbol{A}$ is a diagonalizable matrix, then $\boldsymbol{A}$ can be written as [39]

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}, \tag{3.5}$$

where $\boldsymbol{U}$ is a square matrix with the i:th column containing the i:th eigenvector and $\boldsymbol{\Lambda}$ is a diagonal matrix such that $\boldsymbol{\Lambda}_{ii} = \lambda_i$.

The classical eigenvalue problem can however be extended to the generalized eigenvalue problem given by

$$\boldsymbol{A}\boldsymbol{U} = \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{U}, \tag{3.6}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are two symmetric square matrices. The advantage with the generalized eigenvalue problem is that it works even if $\boldsymbol{B}$ is singular. In the special case when $\boldsymbol{B}$ is non-singular, the equation can be rewritten as

$$\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{\Lambda}\boldsymbol{U}, \tag{3.7}$$

which corresponds to the ordinary eigenvalue problem.

Since both $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices they can be diagonalized and the identities

$$\begin{cases} \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = \boldsymbol{\Lambda} & \text{(3.8a)} \\ \boldsymbol{U}^T\boldsymbol{B}\boldsymbol{U} = \boldsymbol{I} & \text{(3.8b)} \end{cases}$$

must hold, where $\boldsymbol{I}$ is the identity matrix. That the identities must hold can be motivated as follows. Multiply Equation (3.8b) with $\boldsymbol{\Lambda}$ gives $\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{B}\boldsymbol{U} = \boldsymbol{\Lambda}$. Using this identity into Equation (3.8a) gives

$$\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U} = \boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{B}\boldsymbol{U}. \tag{3.9}$$

Multiply both sides with $(\boldsymbol{U}^T)^{-1}$ from the right gives

$$\boldsymbol{A}\boldsymbol{U} = \boldsymbol{\Lambda}\boldsymbol{B}\boldsymbol{U}. \tag{3.10}$$

Equation (3.10) is equivalent to Equation (3.6), which is the generalized eigenvalue equation which needs to be fulfilled. Thus Equation (3.10) needs to be fulfilled and therefore the identities in Equation (3.8a) and (3.8b) must be fulfilled as well.

3.1.2 Proper Orthogonal Decomposition

In almost every field in modern science, simulations are performed that might be computationally heavy [40]. The purpose of Proper Orthogonal Decomposition (POD) is to reduce the computationally cost while still keep the essence of the simulation. It was originally introduced in 1901 by Karl Pearson in the famous article "On lines and planes of closest fit to systems of points in space" [41]. During the past century, the method has been evolved and in different research fields the method is slightly different due to the data it is applied to. For example, when the method is applied to large finite data sets it is known as Principal Component Analysis (PCA) [42], when working with distributed parameter system it is known as Karhunen–Loève Decomposition (KLD) [40] and when it is applied to non-squared matrices it is known as Singular Value Decomposition (SVD) [43].

3.2 Solving the micro-scale problem

As a starting point for the Numerical Model Reduction (NMR) of the transient heat flow problem, recall the weak form of the micro-scale problem given by Equation (2.37) which says that: Find $u^\mu \in \mathbb{U}_\square$ such that

$$\begin{aligned}\mathfrak{m}_\square(\dot{u}^\mu, \delta u) + \mathfrak{a}_\square(u^\mu, \delta u) = -\mathfrak{m}_\square(1, \delta u)\dot{\bar{u}} + \big(\sum_{i=1}^{d} -\mathfrak{m}_\square(\boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}), \delta u)\boldsymbol{e}_i\big) \cdot \dot{\bar{\boldsymbol{g}}} \\ + \big(\sum_{i=1}^{d} -\mathfrak{a}_\square(\boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}), \delta u)\boldsymbol{e}_i\big) \cdot \bar{\boldsymbol{g}}, \quad \forall \delta u \in \mathbb{U}^0_\square.\end{aligned} \tag{3.11}$$

The idea is to use the concepts behind NTFA, see Section 3.1, to assume that $u^\mu(\boldsymbol{x}, t)$ can be written as

$$u^\mu(\boldsymbol{x}, t) = \sum_{a=1}^{N} u_a(\boldsymbol{x})\tilde{\xi}_a(t) \tag{3.12}$$

for some number N. The corresponding test function can be written as

$$\delta u(\boldsymbol{x}, t) = \delta\Big(\sum_{a=1}^{N} u_a(\boldsymbol{x})\tilde{\xi}_a(t)\Big) = \sum_{a=1}^{N} \delta\big(u_a(\boldsymbol{x})\tilde{\xi}_a(t)\big) = \sum_{a=1}^{N} \tilde{\xi}_a(t)\delta u_a(\boldsymbol{x}) + \sum_{a=1}^{N} u_a(\boldsymbol{x})\delta\tilde{\xi}_a(t). \tag{3.13}$$

Since $u_a(\boldsymbol{x})$ is not time dependent, $\delta u_a(\boldsymbol{x}) = 0$ and therefore

$$\delta u(\boldsymbol{x}, t) = \sum_{a=1}^{N} u_a(\boldsymbol{x})\delta\tilde{\xi}_a(t). \tag{3.14}$$

In order to find $\tilde{\xi}_a(t)$, $u_a(\boldsymbol{x})$ needs to be estimated. This can be done with either Spectral Decomposition, see Section 3.1.1, or Proper Orthogonal Decomposition (POD), see Section 3.1.2. The main advantages with POD, compared to Spectral Decomposition, is that it can capture non-linear effects. Spectral Decomposition, on the other hand, works well for linear problems and all matrices are diagonalized automatically. Since only the linear transient heat flow problem is considered, Spectral Decomposition is applied. The spatial modes, $u_a(\boldsymbol{x})$, can then be estimated as the eigenvectors to the generalized eigenvalue problem

$$\lambda_a \mathfrak{m}_\square(u_a, \delta u) + \mathfrak{a}_\square(u_a, \delta u) = 0, \quad \forall \delta u \in \mathbb{U}_\square, \quad a = 1, 2, \ldots, N. \tag{3.15}$$

Equation (3.15) is given on its continuous form. In order to solve the micro-scale problem, the corresponding FE-discrete form is used. Therefore, N is replaced with a finite number N_h. In standard finite element fashion, the spaces $\mathbb{U}_{\square,h} \subset \mathbb{U}_\square$ and $\mathbb{U}^0_{\square,h} \subset \mathbb{U}^0_\square$ are introduced. This implies that the space discrete version of u and δu can be written as

$$\begin{cases} \mathbb{U}_{\square,h} \ni u_h = \sum_{a=1}^{N_h} u_{a,h}(\boldsymbol{x})\tilde{\xi}_a(t) \\ \mathbb{U}^0_{\square,h} \ni \delta u_h = \sum_{k=1}^{N_h} u_{a,h}(\boldsymbol{x})\delta\tilde{\xi}_a(t), \end{cases} \tag{3.16}$$

where

$$\begin{cases} \mathbb{U}_{\square,h} \ni u_{a,h}(\boldsymbol{x}) = \sum_{k=1}^{N_h} N_k(\boldsymbol{x})(\underline{\boldsymbol{u}}_a)_k \\ \mathbb{U}^0_{\square,h} \ni \delta u_{a,h}(\boldsymbol{x}) = \sum_{k=1}^{N_h} N_k(\boldsymbol{x})(\delta\underline{\boldsymbol{u}}_a)_k, \end{cases} \tag{3.17}$$

where N_k are the basis functions and $\underline{\boldsymbol{u}}_a$ respective $\delta\underline{\boldsymbol{u}}_a$ are the nodal vectors. $u_{a,h}$ can be estimated with the corresponding discrete generalized eigenvalue problem, which is given by

$$\delta\underline{\boldsymbol{u}}_a\Big[\lambda_a\underline{\boldsymbol{M}} + \underline{\boldsymbol{K}}\Big]\underline{\boldsymbol{u}}_a = 0, \quad a = 1, 2, \ldots, N_h, \quad \forall \delta\underline{\boldsymbol{u}}_a \in \mathbb{R}^{N_h}, \tag{3.18}$$

where $\underline{\boldsymbol{M}}$ is the mass matrix and $\underline{\boldsymbol{K}}$ is the stiffness matrix. Since this must hold for an arbitrary $\delta\underline{\boldsymbol{u}}_a \in \mathbb{R}^{N_h}$ the equation is simplified as

$$\Big[\lambda_a \underline{\boldsymbol{M}} + \underline{\boldsymbol{K}}\Big]\underline{\boldsymbol{u}}_a = 0, \quad a = 1, 2, \dots, N_h. \tag{3.19}$$

In the upcoming theory it is the discrete solution in space that is considered[1]. By using the decomposition of u, Equation (3.11) can be written as

$$\begin{aligned} \mathfrak{m}_\square\Big(\sum_{b=1}^{N} u_b\dot{\tilde{\xi}}_b, \sum_{a=1}^{N} u_a\delta\tilde{\xi}_a\Big) + \mathfrak{a}_\square\Big(\sum_{b=1}^{N} u_b\tilde{\xi}_b, \sum_{a=1}^{N} u_a\delta\tilde{\xi}_a\Big) = -\mathfrak{m}_\square\Big(1, \sum_{a=1}^{N} u_a\delta\tilde{\xi}_a\Big)\dot{\bar{u}} \\ +\big(\sum_{i=1}^{d} -\mathfrak{m}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), \sum_{a=1}^{N} u_a\delta\tilde{\xi}_a)\boldsymbol{e}_i\big)\cdot\dot{\bar{\boldsymbol{g}}} + \big(\sum_{i=1}^{d} -\mathfrak{a}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), \sum_{a=1}^{N} u_a\delta\tilde{\xi}_a\Big)\boldsymbol{e}_i\big)\cdot\bar{\boldsymbol{g}}. \end{aligned} \tag{3.20}$$

The linearity of $\mathfrak{m}_\square$ and $\mathfrak{a}_\square$ imply that

$$\begin{aligned} \sum_{a,b=1}^{N} \mathfrak{m}_\square\Big(u_b\dot{\tilde{\xi}}_b, u_a\delta\tilde{\xi}_a\Big) + \mathfrak{a}_\square\Big(u_b\tilde{\xi}_b, u_a\delta\tilde{\xi}_a\Big) = \sum_{a=1}^{N} -\mathfrak{m}_\square\Big(1, u_a\delta\tilde{\xi}_a\Big)\dot{\bar{u}} \\ +\big(\sum_{i=1}^{d} -\mathfrak{m}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), u_a\delta\tilde{\xi}_a)\boldsymbol{e}_i\big)\cdot\dot{\bar{\boldsymbol{g}}} + \big(\sum_{i=1}^{d} -\mathfrak{a}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), u_a\delta\tilde{\xi}_a\Big)\boldsymbol{e}_i\big)\cdot\bar{\boldsymbol{g}}. \end{aligned} \tag{3.21}$$

Since $\mathfrak{m}_\square$ and $\mathfrak{a}_\square$ only uses spatial integration, Equation (3.21) is simplified to

$$\begin{aligned} \sum_{a,b=1}^{N} \delta\tilde{\xi}_a\Big[\mathfrak{m}_\square(u_b, u_a)\dot{\tilde{\xi}}_b + \mathfrak{a}_\square(u_b, u_a)\tilde{\xi}_b\Big] = \sum_{a=1}^{N} \delta\tilde{\xi}_a\Big[-\mathfrak{m}_\square(1, u_a)\dot{\bar{u}} \\ +\big(\sum_{i=1}^{d} -\mathfrak{m}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), u_a)\boldsymbol{e}_i\big)\cdot\dot{\bar{\boldsymbol{g}}} + \big(\sum_{i=1}^{d} -\mathfrak{a}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), u_a\Big)\boldsymbol{e}_i\big)\cdot\bar{\boldsymbol{g}}\Big]. \end{aligned} \tag{3.22}$$

The equation must hold for all $\delta\tilde{\xi}_a \in \mathbb{L}^2$, which gives that[2]

$$\begin{aligned} \sum_{b=1}^{N} \mathfrak{m}_\square(u_a, u_b)\dot{\tilde{\xi}}_b + \mathfrak{a}_\square(u_a, u_b)\tilde{\xi}_b = -\mathfrak{m}_\square(1, u_a)\dot{\bar{u}} + \big(\sum_{i=1}^{d} -\mathfrak{m}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), u_a)\boldsymbol{e}_i\big)\cdot\dot{\bar{\boldsymbol{g}}} \\ + \big(\sum_{i=1}^{d} -\mathfrak{a}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), u_a\Big)\boldsymbol{e}_i\big)\cdot\bar{\boldsymbol{g}}, \quad a = 1, 2, \dots N. \end{aligned} \tag{3.23}$$

The orthogonal properties of the generalized eigenvalue problem implies that it is possible to choose the amplitudes of u_a and u_b such that

$$\mathfrak{m}_\square(u_a, u_b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \quad \text{and} \quad \mathfrak{a}_\square(u_a, u_b) = \begin{cases} \lambda_a & \text{if } a = b \\ 0 & \text{if } a \neq b. \end{cases} \tag{3.24}$$

This orthogonality implies that all coupling terms in the differential equations are equal to zero. This results in N independent ordinary differential equations given by

$$\begin{aligned} \dot{\tilde{\xi}}_a + \lambda_a\tilde{\xi}_a = -\mathfrak{m}_\square(1, u_a)\dot{\bar{u}} - \big(\sum_{i=1}^{d} \mathfrak{m}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), u_a)\boldsymbol{e}_i\big)\cdot\dot{\bar{\boldsymbol{g}}} \\ - \big(\sum_{i=1}^{d} \mathfrak{a}_\square(\boldsymbol{e}_i\cdot(\boldsymbol{x}-\bar{\boldsymbol{x}}), u_a\Big)\boldsymbol{e}_i\big)\cdot\bar{\boldsymbol{g}}, \quad a = 1, 2, \dots N. \end{aligned} \tag{3.25}$$

[1] In this thesis, only the discrete solution is used from now on and therefore the index h is dropped.

[2] For information of how the $\mathbb{L}^2$ space is defined, see Appendix A.

By setting

$$f_a(t) = -\mathfrak{m}_\square(1, u_a)\dot{\bar{u}} + \big(\sum_{i=1}^{d} -\mathfrak{m}_\square(\boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}), u_a)\boldsymbol{e}_i\big) \cdot \dot{\bar{\boldsymbol{g}}} + \big(\sum_{i=1}^{d} -\mathfrak{a}_\square(\boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}), u_a\Big)\boldsymbol{e}_i\big) \cdot \bar{\boldsymbol{g}}, \quad a = 1, 2, \dots N, \tag{3.26}$$

the equation can be written in a compact form as

$$\dot{\tilde{\xi}}_a + \lambda_a \tilde{\xi}_a = f_a(t), \quad a = 1, 2, \dots N, \tag{3.27}$$

which, in this thesis, is solved numerically with the Backward Euler method for each $a = 1, 2, \dots N$.

3.3 Implementing the reduced basis for the micro-problem

From Section 3.2 it is showed that u^μ is assumed to be given by

$$u^\mu(\boldsymbol{x}, t) = \sum_{a=1}^{N} u_a(\boldsymbol{x})\tilde{\xi}_a(t), \tag{3.28}$$

where $\tilde{\xi}_a(t)$ is calculated from N uncoupled ordinary differential equations given by

$$\dot{\tilde{\xi}}_a + \lambda_a \tilde{\xi}_a = f_a(t), \quad a = 1, 2, \dots N. \tag{3.29}$$

In Section 3.2 it is further derived how $u_a(\boldsymbol{x})$ can be estimated by using the generalized eigenvalue-problem. Hence all ingredients for calculating $u_a(\boldsymbol{x})$ and $\tilde{\xi}_a(t)$ are obtained and thus $u^\mu(\boldsymbol{x}, t)$ can be determined. As described in Section 1.1, the FE-problem is RVE dependent and has to be solved for each macro-scale quadrature point. Having a fine macro-scale mesh hence induces an unacceptable computational cost. A way to solve this computational issue is to reduce the computational cost of the micro-scale problem by reducing the basis of the solution space. This is done by using the generalized eigenvectors $u_a(\boldsymbol{x})$ corresponding to eigenvalues that is small enough.

Let N_R be the number of generalized eigenvectors that is used to span the reduced basis. The reduced solution of the micro-scale problem is then given by

$$u_R^\mu(\boldsymbol{x}, t) = \sum_{a=1}^{N_R} u_a(\boldsymbol{x})\tilde{\xi}_a(t). \tag{3.30}$$

The total temperature for the reduced problem is obtain by adding the estimated macro-temperature from the First-order Computational Homogenization, given in Equation (2.20), as

$$u_R(\boldsymbol{x}, t) = \bar{u}(t) + \bar{\boldsymbol{g}}(\boldsymbol{x}, t) \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}) + \sum_{a=1}^{N_R} u_a(\boldsymbol{x})\tilde{\xi}_a(t). \tag{3.31}$$

However, there exists a disadvantage with calculating $u_R(\boldsymbol{x}, t)$ in this way. Consider the ordinary differential equation from which $\tilde{\xi}_a(t)$ is calculated from

$$\begin{aligned} \dot{\tilde{\xi}}_a + \lambda_a \tilde{\xi}_a = &- \mathfrak{m}_\square(1, u_a)\dot{\bar{u}} - \big(\sum_{i=1}^{d} \mathfrak{m}_\square(\boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}), u_a)\boldsymbol{e}_i\big) \cdot \dot{\bar{\boldsymbol{g}}} \\ &- \big(\sum_{i=1}^{d} \mathfrak{a}_\square(\boldsymbol{e}_i \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}}), u_a\Big)\boldsymbol{e}_i\big) \cdot \bar{\boldsymbol{g}}, \quad a = 1, 2, \dots N_R. \end{aligned} \tag{3.32}$$

The last term in the right hand side is non-zero even for constant $\bar{\boldsymbol{g}}$ and, therefore, contributes to the stationary solution. This implies that in order to express the stationary solution, all modes are needed. Calculating the stationary solution to the problem is assumed to be a low computational cost compared to the computational cost of calculating the full base solution. Note that it is verified that this assumption is valid in Section 5.6. Let $u_{stat}^{(i)}$ satisfy the equation

$$\mathfrak{a}_{\square}(u_{stat}^{(i)}, \delta u) = 0, \quad \forall \delta u \in \mathbb{U}_{\square}^{0}. \tag{3.33}$$

By using $u_{stat}^{(i)}$ on the macro-level, $u_R(\boldsymbol{x}, t)$ can be written in an equivalent way as

$$u_R(\boldsymbol{x}, t) = \bar{u}(t) + \sum_{i=1}^{d} u_{stat}^{(i)} \boldsymbol{e}_i \cdot \bar{\boldsymbol{g}}_i(t) + \sum_{a=1}^{N_R} u_a(\boldsymbol{x}) \xi_a(t), \tag{3.34}$$

where $\xi_a(t)$ is the solution to the ordinary differential equation

$$\dot{\xi} + \lambda_a \xi = -\mathfrak{m}_{\square}(1, u_a)\dot{\bar{u}} - \big(\sum_{i=1}^{d} \mathfrak{m}_{\square}(u_{stat}^{(i)}, u_a)\boldsymbol{e}_i\big) \cdot \dot{\bar{\boldsymbol{g}}}, \quad a = 1, 2, \ldots N_R. \tag{3.35}$$

By expressing $u_R(\boldsymbol{x}, t)$ as in Equation (3.34), the stationary part of the solution is recovered by the first two terms and the transient part of the solution is covered by the sum. This implies that, independently of how many modes that are used, the stationary solution of $u(\boldsymbol{x}, t)$ is always captured.

4 Error analysis

In this chapter it is shown how the error, introduced by Numerical Model Reduction (NMR), can be estimated. Note that there are three different error sources:

1. Error introduced by space discretization.
2. Error introduced by time discretization.
3. Error introduced by NMR.

In this study it is only the error introduced by NMR that is analyzed. Therefore, the solution that is calculated with all modes is assumed to be the true solution. The error estimates that are performed can be divided into two different categories:

- Error estimates performed using all modes.
- Error estimates performed using only the reduced modes.

In order to estimate errors that span over both space and time, the space-time formulating of the problem is introduced, see Section 4.1. In Section 4.2 fundamental error estimate properties are defined along with several theorems that can be used for error estimates. Section 4.3 defines what quantities of interest that are considered in this thesis. In Section 4.4 it is derived how the theorems from Section 4.2 can be applied when all modes are used. Finally, Section 4.5 shows how similar error estimates can be performed when only the reduced modes are used.

4.1 Space-time formulation

In order to formulate the space-time variational form of the problem, recall Equation (2.36) which says that

$$\mathfrak{m}_\square(\dot{u}, \delta u) + \mathfrak{a}_\square(u, \delta u) = 0, \quad \forall \delta u \in \mathbb{U}_\square^0. \tag{4.1}$$

The goal with the space-time variational form is to define a variational form where the test function is integrated over both the spatial- and time domain. Let $\mathbb{U}_\square$ and $\mathbb{U}_\square^0$ be the same Sobolev spaces as defined in Equation (2.33) and let $\widetilde{\mathbb{U}}_\square$ be a the Sobolev space defined as[1]

$$\widetilde{\mathbb{U}}_\square = \Big\{ u : \int_\Omega \big(|u|^2 + |\boldsymbol{\nabla} u|^2\big)\, \mathrm{d}\Omega \Big\}. \tag{4.2}$$

The solution and test space can then be defined as

$$\begin{aligned} \mathcal{U}_\square &= \Big\{ u(\boldsymbol{x}, t) :\ u(\cdot, t) \in \mathbb{U}_\square(t),\ u(\boldsymbol{x}, \cdot) \in \widetilde{\mathbb{U}}_\square\big([0, T]\big) \Big\}, \\ \mathcal{U}_\square^0 &= \Big\{ u(\boldsymbol{x}, t) :\ u(\cdot, t) \in \mathbb{U}_\square^0(t),\ u(\boldsymbol{x}, \cdot) \in \widetilde{\mathbb{U}}_\square\big([0, T]\big) \Big\}, \end{aligned} \tag{4.3}$$

where $\mathbb{U}_\square(t)$ and $\mathbb{U}_\square^0(t)$ are the ordinary spaces $\mathbb{U}_\square$ and $\mathbb{U}_\square^0$ at the time t. The space-time variational form can then be formulated as follows. Find $u(\boldsymbol{x}, t) \in \mathcal{U}_\square$ such that

$$\int_0^T \mathfrak{m}_\square(\dot{u}, \delta u) + \mathfrak{a}_\square(u, \delta u)\, \mathrm{d}t + \mathfrak{m}_\square(u, \delta u)\big|_{t=0} = \int_0^T 0\, \mathrm{d}t + \mathfrak{m}_\square(u_0, \delta u)\big|_{t=0}, \quad \forall \delta u \in \mathcal{U}_\square^0, \tag{4.4}$$

[1] For more information about Sobolev spaces, see Appendix A.

where u_0 is the initial condition of the temperature. Define the two operators $A_\square : \mathbb{U}_\square \times \mathbb{U}^0_\square \to \mathbb{R}$ and $L_\square : \mathbb{U}^0_\square \to \mathbb{R}$ such that

$$\begin{aligned} A_\square(u, \delta u) &:= \int_0^T \left[\mathfrak{m}_\square(\dot{u}, \delta u) + \mathfrak{a}_\square(u, \delta u)\right] \mathrm{d}t + \mathfrak{m}_\square(u, \delta u)\big|_{t=0} \\ L_\square(\delta u) &:= \int_0^T 0 \, \mathrm{d}t + \mathfrak{m}_\square(u_0, \delta u)\big|_{t=0} = \mathfrak{m}_\square(u_0, \delta u)\big|_{t=0}. \end{aligned} \tag{4.5}$$

From the definition of $A_\square$ and $L_\square$ it can be seen that $A_\square$ is a bilinear form and $L_\square$ is a linear functional. By using the two operators, the variational form can be written as: Find $u(\boldsymbol{x}, t) \in \mathcal{U}_\square$ such that

$$A_\square(u, \delta u) = L_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.6}$$

Remark: By introducing the decomposition of u as

$$u = u^M + u^\mu = \bar{u}(t) + \sum_{i=1}^{d} u^{(i)}_{stat}(\boldsymbol{x}) \boldsymbol{e}_i \cdot \overline{\boldsymbol{g}}(t) + \sum_{a=1}^{N} u_a(\boldsymbol{x}) \xi_a(t), \tag{4.7}$$

the variational form can, in the same approach as for the ordinary variational form in Equations (3.20)-(3.23), be written as

$$\begin{aligned} \sum_{b=1}^{N} \int_0^T \mathfrak{m}_\square(u_a, u_b)\dot{\xi}_b + \mathfrak{a}_\square(u_a, u_b)\xi_b \, \mathrm{d}t + \mathfrak{m}_\square(u_a, u_b)\xi_b(0) = -\int_0^T \mathfrak{m}_\square(u_a, 1)\ddot{\bar{u}} \\ + \big(\sum_{i=1}^{d} \mathfrak{m}_\square(u_a, u^{(i)}_{stat})\boldsymbol{e}_i\big) \cdot \ddot{\overline{\boldsymbol{g}}} \, \mathrm{d}t + \mathfrak{m}_\square(u_0, u_a), \quad a = 1, 2, \ldots N. \end{aligned} \tag{4.8}$$

Let $\xi_{a,0} := \mathfrak{m}_\square(u_0, u_a)$. By using the orthogonal properties of u_a, all couplings terms are equal to zero and N independent equations are obtained as

$$\int_0^T \big(\dot{\xi}_a(t) + \lambda_a \xi_a(t)\big) \, \mathrm{d}t + \xi_a(0) = -\int_0^T \mathfrak{m}(u_a, 1)\ddot{\bar{u}} + \big(\sum_{i=1}^{d} \mathfrak{m}_\square(u_a, u^{(i)}_{stat})\boldsymbol{e}_i\big) \cdot \ddot{\overline{\boldsymbol{g}}} \, \mathrm{d}t + \xi_{a,0}, \quad a = 1, 2, \ldots N. \tag{4.9}$$

The corresponding ordinary differential equations are given by

$$\begin{cases} \dot{\xi}_a + \lambda_a \xi_a = -\mathfrak{m}(u_a, 1)\ddot{\bar{u}} - \big(\sum_{i=1}^{d} \mathfrak{m}_\square(u_a, u^{(i)}_{stat})\boldsymbol{e}_i\big) \cdot \ddot{\overline{\boldsymbol{g}}}, \quad a = 1, 2, \ldots N, \\ \xi_a(0) = \xi_{a,0}, \end{cases} \tag{4.10}$$

which, as expected, are the same differential equations as for the ordinary weak formulation, see Equation (3.35).

4.2 Fundamental theory and theorems used in error analysis

Section 4.2.1 presents theory for the exact error representation, and Section 4.2.2 defines the so-called symmetric error representation. How goal-oriented error analysis can be performed in order to estimate the error, in a given quantity of interest, is shown in Section 4.2.3. Finally, Section 4.2.4 presents how sharper error estimates can be obtained when the Galerkin orthogonality is used.

4.2.1 Exact error representation

Let $\mathcal{U}_\square^R \subset \mathcal{U}_\square$ and $\mathcal{U}_\square^{R,0} \subset \mathcal{U}_\square^0$ be two Sobolev spaces defined as

$$\begin{aligned} \mathcal{U}_\square^R &= \left\{ u \in \mathcal{U}_\square : u(\cdot, t) = Span\{u_a\}_{a=1}^{N_R} \right\} \\ \mathcal{U}_\square^{R,0} &= \left\{ u \in \mathcal{U}_\square^0 : u(\cdot, t) = Span\{u_a\}_{a=1}^{N_R} \right\}. \end{aligned} \tag{4.11}$$

In the same way as in Equation (3.34), define $u_R \in \mathcal{U}_\square^R$ as the reduced solution given by

$$u_R(\boldsymbol{x}, t) = \bar{u}(t) + \sum_{i=1}^{d} u_{stat}^{(i)} \boldsymbol{e}_i \cdot \bar{\boldsymbol{g}}_i(t) + \sum_{a=1}^{N_R} u_a(\boldsymbol{x}) \xi_a(t). \tag{4.12}$$

The error $e(\boldsymbol{x}, t) \in \mathcal{U}_\square^0$ can then be defined as

$$e(\boldsymbol{x}, t) := u(\boldsymbol{x}, t) - u_R(\boldsymbol{x}, t). \tag{4.13}$$

From the definition of $u(\boldsymbol{x}, t)$ and $u_R(\boldsymbol{x}, t)$, the error is given by

$$e(\boldsymbol{x}, t) = \sum_{a=N_R+1}^{N} u_a(\boldsymbol{x}) \xi_a(t). \tag{4.14}$$

Note that, in order to calculate this error, all modes from $N_R + 1$ to N are needed.

With the space-time formulation in mind, the residual for a given u_R is defined as

$$R_\square(\delta u) := L_\square(\delta u) - A_\square(u_R, \delta u), \quad \forall \delta u \in \mathcal{U}_\square^0. \tag{4.15}$$

The error needs to satisfy the error equation given by

$$A_\square(e, \delta u) = R_\square(\delta u), \quad \forall \delta u \in \mathcal{U}_\square^0. \tag{4.16}$$

This follows from the bilinearity of $A_\square$ since

$$A_\square(e, \delta u) = A_\square(u, \delta u) - A_\square(u_R, \delta u) = L(\delta u) - A_\square(u_R, \delta u) = R_\square(\delta u). \tag{4.17}$$

Note that if $\delta u \in \mathcal{U}_\square^{R,0} \subset \mathcal{U}_\square^0$, then $R_\square(\delta u) = 0$, and thus

$$A_\square(e, \delta u) = 0, \quad \forall \delta u \in \mathcal{U}_\square^{R,0}, \tag{4.18}$$

which is known as the Galerkin orthogonality. By assuming that $A_\square(v, v)$ is coercive for all $v \in \mathcal{U}_\square^0$ it is possible to define a norm on $\mathcal{U}_\square^0$ given by

$$\|v\| := \sqrt{A_\square(v, v)}. \tag{4.19}$$

This norm is in this thesis referred to as the energy norm.

4.2.2 Symmetrized error representation

In order to estimate the error, Cauchy-Schwarz inequality or the Parallelogram law can be used. In order to apply these theorems, the operator needs to be symmetric. Unfortunately, $A_\square$ is not symmetric but the fact that $A_\square$ is bilinear implies that a symmetric counterpart to $A_\square$ can be defined as

$$A_\square^s(u, v) := \frac{1}{2}\big[A_\square(u, v) + A_\square(v, u)\big]. \quad \forall u, v \in \mathcal{U}_\square^0. \tag{4.20}$$

The corresponding symmetric error, $e^s \in \mathcal{U}^0_\square$, can be calculated from the corresponding symmetric error equation given by

$$A^s_\square(e^s, \delta u) = R_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.21}$$

Let v be an arbitrary variable in $\mathcal{U}^0_\square$. The equality $A^s_\square(v,v) \equiv A_\square(v,v)$ must hold since

$$A^s_\square(v,v) := \frac{1}{2}\big(A_\square(v,v) + A_\square(v,v)\big) = A_\square(v,v). \tag{4.22}$$

This implies that the energy norm can be defined in an equivalent way as

$$\|v\| := \sqrt{A^s_\square(v,v)}. \tag{4.23}$$

Theorem 1. *If e^s is calculated from Equation* (4.21)*, then*

$$\|e\| \leq \|e^s\|. \tag{4.24}$$

Proof. In the case when $\|e\| = 0$, the theorem is fulfilled since $\|e^s\| \geq 0$. Now prove the theorem when $\|e\| > 0$. The definition of the energy norm gives that $\|e\|^2 = A^s_\square(e,e)$. The identity $A^s_\square(v,v) \equiv A_\square(v,v)$, for all $v \in \mathcal{U}^0_\square$, implies that energy norm of e can be written as

$$\|e\|^2 = A^s_\square(e,e) = A_\square(e,e) = R_\square(e) = A_\square(e^s, e), \tag{4.25}$$

where the two last steps follows from Equations (4.16) and (4.21). Cauchy-Schwarz inequality can now by applied to the symmetric operator $A^s_\square$ as

$$\|e\|^2 = A^s_\square(e^s, e) \leq \|e^s\|\|e\|. \tag{4.26}$$

Since $\|e\| > 0$, it is possible to divide with $\|e\|$ on both sides which implies that

$$\|e\| \leq \|e^s\|. \tag{4.27}$$

□

4.2.3 Goal-oriented error estimation

For a given problem there is often a given property that is of interest. Properties that are of interest might be the average displacement along a boundary, average stress or average temperature in a body. In order to estimate such properties the concept of goal-oriented error analysis is introduced. The key idea is to introduce a so-called quantity of interest, $Q_\square$, which is a linear functional from which a new error estimate can be defined as

$$E = Q_\square(e). \tag{4.28}$$

Changing how $Q_\square(e)$ is defined implies that different quantities of interest can be obtained. In order to estimate E, the dual problem is needed which is defined as: Find $u^\star \in \mathcal{U}^0_\square$ such that

$$A_\square(\delta u, u^\star) = Q_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.29}$$

From the definition of the dual problem it follows that E can be written as

$$E = Q_\square(e) = A_\square(e, u^\star) = R_\square(u^\star). \tag{4.30}$$

In order to estimate E from the dual solution, the symmetric representation of $u^\star$, denoted $u^{\star,s} \in \mathcal{U}^0_\square$, is needed. This quantity can be calculated from the weak form: Find $u^{\star,s} \in \mathcal{U}^0_\square$ such that

$$A^s_\square(\delta u, u^{\star,s}) = Q_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.31}$$

The error, E, can be estimated by the following theorem:

Theorem 2. *If e^s is calculated from Equation* (4.21) *and $u^{\star,s}$ from Equation* (4.31)*, then the identity*

$$|E| \leq \|e^s\| \|u^{\star,s}\| \tag{4.32}$$

must hold.

Proof. From Equations (4.28) and (4.31) it follows that

$$|E| = |Q_\square(e)| = |A^s_\square(e, u^{\star,s})|. \tag{4.33}$$

Cauchy-Schwarz theorem together with Theorem 1 concludes the proof since

$$|E| = |A^s_\square(e, u^{\star,s})| \leq \|e\| \|u^{\star,s}\| \leq \|e^s\| \|u^{\star,s}\|. \tag{4.34}$$

□

4.2.4 Sharper error estimates using the Galerkin orthogonality

The estimation of the error that is used in Theorem 2 never utilizes the Galerkin orthogonality property given in Equation (4.18). By introducing the approximate dual problem given by: Find $u^\star_R \in \mathcal{U}^R_\square$ such that

$$A_\square(\delta u, u^\star_R) = Q_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^{R,0}_\square, \tag{4.35}$$

the residual for the dual problem can be stated as

$$R^\star_\square(\delta u) := Q_\square(\delta u) - A_\square(\delta u, u^\star_R), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.36}$$

The corresponding error equation is given by: Find $u^\star - u^\star_R =: e^\star \in \mathcal{U}^0_\square$ such that

$$A_\square(\delta u, e^\star) = R^\star_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.37}$$

Finally, the symmetric error equation of the dual problem can be written as: Find $e^{\star,s} \in \mathcal{U}^0_\square$ such that

$$A^s_\square(\delta u, e^{\star,s}) = R^\star_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.38}$$

From these identities, the error can be estimated with the following theorems.

Theorem 3 (Cauchy-Schwarz bounds of the output)**.** *If e^s is solved from* (4.21) *and $e^{\star,s}$ is solved from* (4.38)*, then*

$$|E| \leq \|e^s\| \|e^{\star,s}\|. \tag{4.39}$$

Proof. By using the properties of the Galerkin orthogonality and the goal-oriented error estimate in Equation (4.30), the error can be written as

$$E = R_\square(u^\star) = R_\square(u^\star) - R_\square(u^\star_R) = R_\square(u^\star - u^\star_R) = R_\square(e^\star). \tag{4.40}$$

The symmetric error equation implies that

$$E = R_\square(e^\star) = A^s_\square(e^s, e^\star). \tag{4.41}$$

Cauchy-Schwarz theorem implies that

$$|E| = |A^s_\square(e^s, e^\star)| \leq \|e^s\| \|e^\star\|. \tag{4.42}$$

From Theorem 1 the relation $\|e^\star\| \leq \|e^{\star,s}\|$ follows which gives

$$|E| \leq \|e^s\| \|e^\star\| \leq \|e^s\| \|e^{\star,s}\|. \tag{4.43}$$

□

Theorem 4 (Parallelogram law bounds of the output)**.** *If e^s is solved from* (4.21)*, $e^{\star,s}$ is solved from* (4.38) *and $\kappa \neq 0$ is an arbitrary constant, then*

$$\begin{aligned} E &\leq \frac{1}{4}\|\kappa e^s + \frac{1}{\kappa}e^{\star,s}\|^2 \\ E &\geq -\frac{1}{4}\|\kappa e^s - \frac{1}{\kappa}e^{\star,s}\|^2. \end{aligned} \tag{4.44}$$

Proof. Consider the inequality

$$0 \leq \|\frac{1}{2}\big(\kappa e^s \pm \frac{1}{\kappa}e^{\star,s}\big) - \kappa e\|^2 = \frac{1}{4}\|\kappa e^s \pm \frac{1}{\kappa}e^{\star,s}\|^2 - \kappa^2 A^s_\square(e^s, e) \mp A^s_\square(e^{\star,s}, e) + \kappa^2\|e\|^2. \tag{4.45}$$

From the error equation for the symmetric and non-symmetric problem it follows that

$$A^s_\square(e^s, e) = R_\square(e) = A_\square(e, e) = A^s_\square(e, e) = \|e\|^2. \tag{4.46}$$

Hence the inequality can be simplified into

$$0 \leq \frac{1}{4}\|\kappa e^s \pm \frac{1}{\kappa}e^{\star,s}\|^2 - \kappa^2\|e\|^2 \mp A^s_\square(e^{\star,s}, e) + \kappa^2\|e\|^2 = \frac{1}{4}\|\kappa e^s \pm \frac{1}{\kappa}e^{\star,s}\|^2 \mp A^s_\square(e^{\star,s}, e) \tag{4.47}$$

The error equation from the dual symmetric problem and the definition of the dual residual gives

$$A^s_\square(e^{\star,s}, e) = A^s_\square(e, e^{\star,s}) = R^\star_\square(e) = Q_\square(e) - A_\square(e, u^\star_R). \tag{4.48}$$

But from the Galerkin orthogonality it follows that

$$A_\square(e, u^\star_R) = R_\square(u^\star_R) = 0 \tag{4.49}$$

since $u^\star_R \in \mathcal{U}^{R,0}_\square$, and therefore

$$A^s_\square(e^{\star,s}, e) = Q_\square(e) = E. \tag{4.50}$$

Thus

$$0 \leq \frac{1}{4}\|\kappa e^s \pm \frac{1}{\kappa}e^{\star,s}\|^2 \mp E, \tag{4.51}$$

which is equivalent to

$$\begin{aligned} E &\leq \frac{1}{4}\|\kappa e^s + \frac{1}{\kappa}e^{\star,s}\|^2 \\ E &\geq -\frac{1}{4}\|\kappa e^s - \frac{1}{\kappa}e^{\star,s}\|^2. \end{aligned} \tag{4.52}$$

□

When Theorem 4 is applied it is beneficial to set κ to the value that gives the error estimate that is as close as possible to the true error. This error estimate is obtained if

$$\kappa = \frac{\|e^{\star,s}\|}{\|e^s\|}, \tag{4.53}$$

see Theorem 5.

Theorem 5 (Optimal choice of κ)**.** *Assume that $\|e^s\|$ is positive. Then the error estimate with Theorem 4 is as close as possible to the true error if*

$$\kappa = \frac{\|e^{\star,s}\|}{\|e^s\|}. \tag{4.54}$$

Proof. The optimal value of κ is obtained if the norm from Theorem 4 is minimized, i.e.

$$\min_{\kappa\neq 0}\frac{1}{4}\|\kappa e^s \pm \frac{1}{\kappa}e^{\star,s}\|^2 = \min_{\kappa\neq 0}\frac{1}{4}\Big(\kappa^2\|e^s\|^2 \pm 2A^s_\square(e^s,e^{\star,s}) + \frac{1}{\kappa^2}\|e^{\star,s}\|^2\Big). \tag{4.55}$$

Differentiate with respect to κ and set the derivative equal to zero in order to find a stationary point.

$$\frac{\partial}{\partial\kappa}\frac{1}{4}\Big(\kappa^2\|e^s\|^2 \pm 2A^s_\square(e^s,e^{\star,s}) + \frac{1}{\kappa^2}\|e^{\star,s}\|^2\Big) = 0, \tag{4.56}$$

gives that

$$2\kappa\|e^s\|^2 - \frac{2}{\kappa^3}\|e^{\star,s}\|^2 = 0. \tag{4.57}$$

By solving the equation it can be found that

$$\kappa^2 = \frac{\|e^{\star,s}\|^2}{\|e^s\|^2}, \tag{4.58}$$

and since all norms are non-negative it can be concluded that

$$\kappa = \frac{\|e^{\star,s}\|}{\|e^s\|}. \tag{4.59}$$

This value of κ is a minimum since the second derivative is positive,

$$\frac{\partial^2}{\partial\kappa^2}\frac{1}{4}\|\kappa e^s + \frac{1}{\kappa}e^{\star,s}\|^2\Big|_{\kappa=\frac{\|e^{\star,s}\|}{\|e^s\|}} = \frac{1}{2}\|e^s\| + \frac{3}{2\kappa^4}\|e^{\star,s}\|^2\Big|_{\kappa=\frac{\|e^{\star,s}\|}{\|e^s\|}} = \frac{1}{2}\|e^s\| + \frac{3\|e^s\|^4}{2\|e^{\star,s}\|^2} > 0. \tag{4.60}$$

□

4.3 Considered quantities of interest

In order to keep a wide range of different quantities of interest, $Q_\square(u)$ is defined as

$$Q_\square(u) = \int_0^T \mathfrak{m}_\square(X,u) + \mathfrak{a}_\square(Y,u)\,\mathrm{d}t + \mathfrak{m}_\square(Z,u)\big|_{t=T}, \tag{4.61}$$

where X, Y and Z are arbitrary functions. The true error for the quantity of interest is given by

$$E = Q_\square(e) = \int_0^T \mathfrak{m}_\square(X,e) + \mathfrak{a}_\square(Y,e)\,\mathrm{d}t + \mathfrak{m}_\square(Z,e)\big|_{t=T}. \tag{4.62}$$

By choosing X, Y and Z to be specific functions, different quantities of interest are obtained.

Example 1 (*Average temperature*). If the quantity of interest should be the average temperature inside the domain, the quantity of interest must look like

$$Q_\square(u) = \frac{1}{T|\Omega_\square|}\int_0^T\int_{\Omega_\square} u\,\mathrm{d}\Omega\,\mathrm{d}t. \tag{4.63}$$

Lemma 1 (Average temperature)**.** *The expression of $Q_\square(u)$ in Equation* (4.63) *is obtained if*

$$\begin{cases} X = \frac{1}{cT} \\ Y = 0 \\ Z = 0. \end{cases} \tag{4.64}$$

Proof. The versatile quantity of interest is given by

$$Q_\square(u) = \int_0^T \mathfrak{m}_\square(X,u) + \mathfrak{a}_\square(Y,u)\,\mathrm{d}t + \mathfrak{m}_\square(Z,u)\big|_{t=T}. \tag{4.65}$$

In order to have the average temperature as quantity of interest, the identity

$$\frac{1}{T|\Omega_\square|}\int_0^T \int_{\Omega_\square} u\,\mathrm{d}\Omega\,\mathrm{d}t = \int_0^T \mathfrak{m}_\square(X,u) + \mathfrak{a}_\square(Y,u)\,\mathrm{d}t + \mathfrak{m}_\square(Z,u)\big|_{t=T} \tag{4.66}$$

must hold. In the left hand side there is no derivatives of the temperature and hence $Y = 0$. Also, there is no explicit boundary term, which implies that $Z = 0$. The definition of $\mathfrak{m}_\square$ gives that

$$\frac{1}{T|\Omega_\square|}\int_0^T \int_{\Omega_\square} u\,\mathrm{d}\Omega\,\mathrm{d}t = \frac{1}{|\Omega_\square|}\int_0^T \int_{\Omega_\square} cuX\,\mathrm{d}\Omega\,\mathrm{d}t. \tag{4.67}$$

This must hold for an arbitrary domain and final time T and thus

$$\frac{1}{T|\Omega_\square|}u = \frac{1}{|\Omega_\square|}cuX. \tag{4.68}$$

Solving X from the equation gives that

$$X = \frac{1}{cT}. \tag{4.69}$$

□

Example 2 (*Average heat flux in one direction*). If the quantity of interest should be the average heat flux in a specific direction, the quantity of interest must look like

$$Q_\square(u) = \frac{1}{T|\Omega_\square|}\int_0^T \int_{\Omega_\square} \boldsymbol{e}\cdot\boldsymbol{q}\,\mathrm{d}\Omega\,\mathrm{d}t, \tag{4.70}$$

where $\boldsymbol{q}$ is the heat flux and $\boldsymbol{e}$ is the unit vector in the desired direction.

Lemma 2 (Average heat flux in one direction)**.** *The expression of $Q_\square(u)$ in Equation (4.70) is obtained if*

$$\begin{cases} X = 0, \\ Y = \left(u^{(i)}_{stat} - \boldsymbol{x}\cdot\boldsymbol{e}_i\right)/T, \\ Z = 0. \end{cases} \tag{4.71}$$

Proof. The versatile quantity of interest is given by

$$Q_\square(u) = \int_0^T \mathfrak{m}_\square(X,u) + \mathfrak{a}_\square(Y,u)\,\mathrm{d}t + \mathfrak{m}_\square(Z,u)\big|_{t=T}, \tag{4.72}$$

which implies that

$$\frac{1}{T|\Omega_\square|}\int_0^T \int_{\Omega_\square} \boldsymbol{e}\cdot\boldsymbol{q}\,\mathrm{d}\Omega\,\mathrm{d}t \overset{!}{=} \int_0^T \mathfrak{m}_\square(X,u) + \mathfrak{a}_\square(Y,u)\,\mathrm{d}t + \mathfrak{m}_\square(Z,u)\big|_{t=T}. \tag{4.73}$$

Only the gradient of the temperature over the domain is needed and therefore $X = Z = 0$, which implies that

$$\frac{1}{T|\Omega_\square|}\int_0^T \int_{\Omega_\square} \boldsymbol{e}\cdot\boldsymbol{q}\,\mathrm{d}\Omega\,\mathrm{d}t = \int_0^T \mathfrak{a}_\square(Y,u)\,\mathrm{d}t. \tag{4.74}$$

The definition of $\mathfrak{a}_\square$, together with the definition of the heat flux, $\boldsymbol{q} := -\boldsymbol{K} \cdot \boldsymbol{\nabla} u$, implies that the left hand side can be rewritten as

$$\frac{1}{T|\Omega_\square|} \int_0^T \int_{\Omega_\square} \boldsymbol{e} \cdot \boldsymbol{q} \, \mathrm{d}\Omega \, \mathrm{d}t = -\frac{1}{T} \int_0^T \mathfrak{a}_\square(u, \boldsymbol{x} \cdot \boldsymbol{e}) \, \mathrm{d}t. \tag{4.75}$$

The symmetry and linearity of $\mathfrak{a}_\square$ implies that Equation (4.74) can be simplified into

$$\int_0^T \mathfrak{a}_\square(u, Y + \boldsymbol{x} \cdot \boldsymbol{e}/T) \, \mathrm{d}t = 0. \tag{4.76}$$

This must hold for an arbitrary final time T and thus

$$\mathfrak{a}_\square(u, Y + \boldsymbol{x} \cdot \boldsymbol{e}/T) = 0. \tag{4.77}$$

From the definition of $u_{stat}^{(i)}$, see Equation (3.33), it follows that

$$\mathfrak{a}_\square(u, u_{stat}^{(i)}/T) = 0. \tag{4.78}$$

From Equations (4.77) and (4.78) it follows that

$$Y + \boldsymbol{x} \cdot \boldsymbol{e}/T = u_{stat}^{(i)}/T, \tag{4.79}$$

which implies that

$$Y = \big(u_{stat}^{(i)} - \boldsymbol{x} \cdot \boldsymbol{e}\big)/T. \tag{4.80}$$

□

Example 3 (*Final temperature*). If the quantity of interest should be the final temperature, the quantity of interest must look like

$$Q_\square(u) = \frac{1}{|\Omega_\square|} \int_{\Omega_\square} u\big|_{t=T} \, \mathrm{d}\Omega. \tag{4.81}$$

Lemma 3 (Final temperature)**.** *The expression of $Q_\square(u)$ in Equation* (4.81) *is obtained if*

$$\begin{cases} X = 0, \\ Y = 0, \\ Z = \frac{1}{c}. \end{cases} \tag{4.82}$$

Proof. The versatile quantity of interest is given by

$$Q_\square(u) = \int_0^T \mathfrak{m}_\square(X, u) + \mathfrak{a}_\square(Y, u) \, \mathrm{d}t + \mathfrak{m}_\square(Z, u)\big|_{t=T}, \tag{4.83}$$

which implies that

$$\frac{1}{|\Omega_\square|} \int_{\Omega_\square} u\big|_{t=T} \, \mathrm{d}\Omega \stackrel{!}{=} \int_0^T \mathfrak{m}_\square(X, u) + \mathfrak{a}_\square(Y, u) \, \mathrm{d}t + \mathfrak{m}_\square(Z, u)\big|_{t=T}. \tag{4.84}$$

Only properties at $t = T$ are of interest and therefore $X = Y = 0$. The definition of $\mathfrak{m}_\square$ implies that

$$\frac{1}{|\Omega_\square|} \int_{\Omega_\square} u\big|_{t=T} \, \mathrm{d}\Omega = \frac{1}{|\Omega_\square|} \int_{\Omega_\square} cZu\big|_{t=T} \, \mathrm{d}\Omega. \tag{4.85}$$

This shall hold for an arbitrary spatial domain and therefore

$$u\big|_{t=T} = cZu\big|_{t=T}, \tag{4.86}$$

which implies that

$$Z = \frac{1}{c}. \tag{4.87}$$

□

4.4 Error estimates using all modes

In Section 4.2 it is shown how the error can be estimated with Theorems 1-4. In order to do so, it is stipulated that $\|e^s\|$, $\|u^{\star,s}\|$ and $\|e^{\star,s}\|$ need to be calculated, which is non-trivial. In Sections 4.4.1, 4.4.2 and 4.4.3 expressions of e^s, $u^{\star,s}$ and $e^{\star,s}$ are derived. Note that, in order to find exact expression of e^s, $u^{\star,s}$ and $e^{\star,s}$, all modes are needed. If these properties are calculated its corresponding energy norm can easily be found from

$$\|\cdot\| = \sqrt{A^s_\square(\cdot,\cdot)}. \tag{4.88}$$

The most important results are summarized in Box 4.4.1.

Box 4.4.1 (Error estimates using all modes)**.** *In order to apply Theorems 1-4, expressions of e^s, $u^{\star,s}$ and $e^{\star,s}$ are needed. e^s can be estimated as follows:*

$$\begin{cases} e^s = \sum_{a=1}^N u_a(\boldsymbol{x})\eta_a(t), \quad \forall t \in (0,T), \\ e^s\big|_{t=0} = 2u_0 - 2u_R\big|_{t=0}, \\ e^s\big|_{t=T} = 0, \end{cases} \tag{4.89}$$

where N is the number of modes, u_0 is the starting temperature, u_R is the reduced solution, u_a are the same spatial modes that are used to solve the micro-scale problem and

$$\eta_a = \begin{cases} 0 & a \le N_R \\ -\frac{1}{\lambda_a}\Big(\mathfrak{m}_\square(\ddot{\bar{u}} + \sum_{i=1}^d u^{(i)}_{stat}\boldsymbol{e}_i\cdot\ddot{\bar{\boldsymbol{g}}}_i, u_a) + \mathfrak{a}_\square(\bar{u} + \sum_{i=1}^d u^{(i)}_{stat}\boldsymbol{e}_i\cdot\bar{\boldsymbol{g}}_i, u_a)\Big) & a > N_R, \\ & \forall t \in (0,T),\ a = 1,2,\dots,N, \end{cases} \tag{4.90}$$

where N_R is the number of reduced modes. $u^{\star,s}$ can be estimated as follows:

$$\begin{cases} u^{\star,s} = \sum_{a=1}^N u_a(\boldsymbol{x})\xi^\star_a(t), \quad \forall t \in (0,T), \\ u^{\star,s}\big|_{t=0} = 0, \\ u^{\star,s}\big|_{t=T} = 2Z, \end{cases} \tag{4.91}$$

where Z is defined from the quantity of interest and

$$\xi^\star_a = \frac{1}{\lambda_a}\Big(\mathfrak{m}_\square(X,u_a) + \mathfrak{a}_\square(Y,u_a)\Big), \quad \forall t \in (0,T),\ a = 1,2,\dots,N, \tag{4.92}$$

where X and Y are defined from the quantity of interest. $e^{\star,s}$ can be estimated as follows:

$$\begin{cases} e^{\star,s} = \sum_{a=1}^N u_a(\boldsymbol{x})\eta^\star_a(t), \quad \forall t \in (0,T), \\ e^{\star,s}\big|_{t=0} = 0, \\ e^{\star,s}\big|_{t=T} = 2Z - 2u^\star_R\big|_{t=T}, \end{cases} \tag{4.93}$$

where

$$\eta^\star_a = \begin{cases} 0, & \forall t \in (0,T),\ a = 1,2,\dots,N_R. \\ \xi^\star_a, & \forall t \in (0,T),\ a = N_R+1, N_R+2,\dots,N. \end{cases} \tag{4.94}$$

4.4.1 Derivation of an expression of the symmetric error

In order to apply Theorems 1-4, e^s needs to be calculated. This can be done with Equation (4.21) which says that: Find $e^s \in \mathcal{U}^0_\square$ such that

$$A^s_\square(e^s, \delta u) = R_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.95}$$

Hence expressions of $A^s_\square(e^s, \delta u)$ and $R_\square(\delta u)$ are needed. From the definition of $A^s_\square(e^s, \delta u)$ it follows that

$$\begin{aligned} A^s_\square(e^s, \delta u) &= \frac{1}{2}\Big[A_\square(e^s, \delta u) + A_\square(\delta u, e^s)\Big] = \frac{1}{2}\Big[\int_0^T \mathfrak{m}_\square(\dot{e}^s, \delta u) + \mathfrak{a}_\square(e^s, \delta u)\,\mathrm{d}t \\ &\quad + \mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} + \int_0^T \mathfrak{m}_\square(\delta\dot{u}^s, e^s) + \mathfrak{a}_\square(\delta u, e^s)\,\mathrm{d}t + \mathfrak{m}_\square(\delta u, e^s)\big|_{t=0}\Big] \\ &= \frac{1}{2}\Big[\int_0^T \big(\mathfrak{m}_\square(\dot{e}^s, \delta u) + \mathfrak{m}_\square(\delta\dot{u}^s, e^s) + \mathfrak{a}_\square(e^s, \delta u) + \mathfrak{a}_\square(\delta u, e^s)\big)\,\mathrm{d}t + \\ &\quad + \mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} + \mathfrak{m}_\square(\delta u, e^s)\big|_{t=0}\Big]. \end{aligned} \tag{4.96}$$

The symmetry of $\mathfrak{a}_\square$ implies that

$$\begin{aligned} A^s_\square(e^s, \delta u) = \int_0^T \big[\frac{1}{2}\mathfrak{m}_\square(\dot{e}^s, \delta u) + \frac{1}{2}\mathfrak{m}_\square(\delta\dot{u}, e^s) + \mathfrak{a}_\square(e^s, \delta u)\big]\,\mathrm{d}t \\ + \frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} + \frac{1}{2}\mathfrak{m}_\square(\delta u, e^s)\big|_{t=0}. \end{aligned} \tag{4.97}$$

The identity $\int_0^T \dot{u}\,\mathrm{d}t = [u]_0^T = u(T) - u(0)$ can be used on the terms with $\mathfrak{m}_\square$ as

$$\int_0^T \mathfrak{m}_\square(\dot{e}^s, \delta u) + \mathfrak{m}_\square(\delta\dot{u}^s, e^s)\,\mathrm{d}t = \Big[\mathfrak{m}_\square(e^s, \delta u)\Big]_{t=0}^T = \mathfrak{m}_\square(e^s, \delta u)\big|_{t=T} - \mathfrak{m}_\square(e^s, \delta u)\big|_{t=0}. \tag{4.98}$$

By also using the symmetry of $\mathfrak{m}_\square$ implies that $A^s_\square(e^s, \delta u)$ can be written as

$$A^s_\square(e^s, \delta u) = \int_0^T \mathfrak{a}_\square(e^s, \delta u)\,\mathrm{d}t + \frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} + \frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=T}. \tag{4.99}$$

Now find an expression of the residual $R_\square(\delta u)$ as

$$\begin{aligned} R_\square(\delta u) &:= L_\square(\delta u) - A_\square(u_R, \delta u) = \mathfrak{m}_\square(u_0, \delta u)\big|_{t=0} \\ &\quad - \int_0^T \mathfrak{m}_\square(\dot{u}_R, \delta u) + \mathfrak{a}_\square(u_R, \delta u)\,\mathrm{d}t - \mathfrak{m}_\square(u_R, \delta u)\big|_{t=0} \\ &= -\int_0^T \big[\mathfrak{m}_\square(\dot{u}_R, \delta u) + \mathfrak{a}_\square(u_R, \delta u)\big]\,\mathrm{d}t + \mathfrak{m}_\square(u_0 - u_R, \delta u)\big|_{t=0}. \end{aligned} \tag{4.100}$$

The symmetric error can then be calculated as follows: Find $e^s \in \mathcal{U}^0_\square$ such that

$$\begin{aligned} &\int_0^T \mathfrak{a}_\square(e^s, \delta u)\,\mathrm{d}t + \frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} + \frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=T} \\ &= -\int_0^T \big[\mathfrak{m}_\square(\dot{u}_R, \delta u) + \mathfrak{a}_\square(u_R, \delta u)\big]\,\mathrm{d}t + \mathfrak{m}_\square(u_0 - u_R, \delta u)\big|_{t=0}, \quad \forall \delta u \in \mathcal{U}^0_\square. \end{aligned} \tag{4.101}$$

Equation (4.101) can be decomposed as

$$\begin{cases} \mathfrak{a}_\square(e^s, \delta u) = -\mathfrak{m}_\square(\dot{u}_R, \delta u) - \mathfrak{a}_\square(u_R, \delta u), \quad \forall \delta u \in \mathbb{U}^0_\square, \quad \forall t \in (0,T), & \text{(4.102a)} \\ \frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} = \mathfrak{m}_\square(u_0 - u_R, \delta u)\big|_{t=0}, \quad \forall \delta u \in \mathbb{U}^0_\square, & \text{(4.102b)} \\ \frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=T} = 0, \quad \forall \delta u \in \mathbb{U}^0_\square. & \text{(4.102c)} \end{cases}$$

From Equations (4.102b) and (4.102c) it can immediately be seen that $e^s\big|_{t=0} = 2u_0 - 2u_R\big|_{t=0}$ and $e^s\big|_{t=T} = 0$.

Now find e^s for times between 0 and T, i.e. solve Equation (4.102a). Assume that e^s can be written as

$$e^s(\boldsymbol{x}, t) = \sum_{a=1}^{N} u_a(\boldsymbol{x})\eta_a(t), \tag{4.103}$$

where $u_a(\boldsymbol{x})$ are the same spatial modes that are used to solve the micro-scale problem and $\eta_a(t)$ are unknown functions of time for all $a \in [1, 2, \ldots, N]$. The expression of δu, given in Equation (3.14), implies that Equation (4.102a) can be written as

$$\sum_{a,b=1}^{N} \delta\xi_b\Big(\eta_a \mathfrak{a}_\square(u_a, u_b)\Big) = -\sum_{b=1}^{N} \delta\xi_b\Big(\mathfrak{m}_\square(\dot{u}_R, u_b) + \mathfrak{a}_\square(u_R, u_b)\Big), \quad \forall t \in (0,T). \tag{4.104}$$

By using the expression of the reduced basis, given in Equation (3.34), the equation can be written as

$$\begin{aligned} \sum_{a,b=1}^{N} \delta\xi_b\Big(\eta_a \mathfrak{a}_\square(u_a, u_b)\Big) = -\sum_{b=1}^{N} \delta\xi_b\Big(&\mathfrak{m}_\square(\dot{\bar{u}} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \dot{\bar{\boldsymbol{g}}}_i, u_b) + \mathfrak{a}_\square(\bar{u} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \bar{\boldsymbol{g}}_i, u_b) \\ &+ \sum_{a=1}^{N_R} \mathfrak{m}_\square(u_a, u_b)\dot{\xi}_a + \mathfrak{a}_\square(u_a, u_b)\xi_a\Big), \quad \forall t \in (0,T). \end{aligned} \tag{4.105}$$

This must hold for all test functions $\delta\xi_b \in \mathbb{L}^2$ and therefore

$$\begin{aligned} \sum_{a=1}^{N} \Big(\eta_a \mathfrak{a}_\square(u_a, u_b)\Big) = &-\mathfrak{m}_\square(\dot{\bar{u}} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \dot{\bar{\boldsymbol{g}}}_i, u_b) - \mathfrak{a}_\square(\bar{u} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \bar{\boldsymbol{g}}_i, u_b) \\ &- \sum_{a=1}^{N_R} \big(\mathfrak{m}_\square(u_a, u_b)\dot{\xi}_a + \mathfrak{a}_\square(u_a, u_b)\xi_a\big), \quad \forall t \in (0,T),\ b = 1, 2, \ldots, N. \end{aligned} \tag{4.106}$$

The orthogonal properties of $\mathfrak{a}_\square$ and $\mathfrak{m}_\square$ give

$$\lambda_a\eta_a = \begin{cases} -\mathfrak{m}_\square(\dot{\bar{u}} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \dot{\bar{\boldsymbol{g}}}_i, u_a) - \mathfrak{a}_\square(\bar{u} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \bar{\boldsymbol{g}}_i, u_a) - \dot{\xi}_a - \lambda_a\xi_a, & a \leq N_R, \\ -\mathfrak{m}_\square(\dot{\bar{u}} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \dot{\bar{\boldsymbol{g}}}_i, u_a) - \mathfrak{a}_\square(\bar{u} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \bar{\boldsymbol{g}}_i, u_a), & a > N_R. \end{cases} \quad \forall t \in (0,T),\ a = 1, 2, \ldots, N. \tag{4.107}$$

For the used modes, the residual is equal to zero and therefore

$$\lambda_a\eta_a = \begin{cases} 0, & a \leq N_R, \\ -\mathfrak{m}_\square(\dot{\bar{u}} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \dot{\bar{\boldsymbol{g}}}_i, u_a) - \mathfrak{a}_\square(\bar{u} + \sum_{i=1}^{d} u^{(i)}_{stat}\boldsymbol{e}_i \cdot \bar{\boldsymbol{g}}_i, u_a), & a > N_R, \end{cases} \quad \forall t \in (0,T),\ a = 1, 2, \ldots, N, \tag{4.108}$$

and thus

$$\eta_a = \begin{cases} 0 & a \leq N_R \\ -\frac{1}{\lambda_a}\Big(\mathfrak{m}_\square(\dot{\bar{u}} + \sum_{i=1}^d u_{stat}^{(i)} \boldsymbol{e}_i \cdot \dot{\bar{\boldsymbol{g}}}_i, u_a) + \mathfrak{a}_\square(\bar{u} + \sum_{i=1}^d u_{stat}^{(i)} \boldsymbol{e}_i \cdot \bar{\boldsymbol{g}}_i, u_a)\Big) & a > N_R, \\ & \forall t \in (0,T),\ a = 1,2,\ldots,N. \end{cases} \tag{4.109}$$

This implies that the symmetric error can be calculated as

$$e^s(\boldsymbol{x},t) = \sum_{a=N_R+1}^{N} u_a(\boldsymbol{x})\eta_a(t), \tag{4.110}$$

where $\eta_a(t)$ is calculated from Equation (4.109).

4.4.2 Derivation of an expression of the symmetric dual solution

In order to apply Theorem 2, $\|u^{\star,s}\|$ is needed. $u^{\star,s} \in \mathcal{U}_\square^0$ can be obtain from the equation

$$A_\square^s(u^{\star,s}, \delta u) = Q_\square(\delta u), \quad \forall \delta u \in \mathcal{U}_\square^0. \tag{4.111}$$

Note that the left hand side is equal to the left hand side of Equation (4.21) with e^s replaced by $u^{\star,s}$. By using the same argumentation as in Section 4.4.1, $A_\square^s(u^{\star,s}, \delta u)$ can be written as

$$A_\square^s(u^{\star,s}, \delta u) = \int_0^T \mathfrak{a}_\square(u^{\star,s}, \delta u)\,\mathrm{d}t + \frac{1}{2}\mathfrak{m}_\square(u^{\star,s}, \delta u)\big|_{t=0} + \frac{1}{2}\mathfrak{m}_\square(u^{\star,s}, \delta u)\big|_{t=T}. \tag{4.112}$$

In order to find an expression of the right hand side of Equation (4.111), the function $Q_\square(\delta u)$ is defined according to Equation (4.61). Equation (4.111) can now be written as: Find $u^{\star,s} \in \mathcal{U}_\square^0$ such that

$$\begin{aligned} &\int_0^T \mathfrak{a}_\square(u^{\star,s}, \delta u)\,\mathrm{d}t + \frac{1}{2}\mathfrak{m}_\square(u^{\star,s}, \delta u)\big|_{t=0} + \frac{1}{2}\mathfrak{m}_\square(u^{\star,s}, \delta u)\big|_{t=T} \\ &= \int_0^T \mathfrak{m}_\square(X, \delta u) + \mathfrak{a}_\square(Y, \delta u)\,\mathrm{d}t + \mathfrak{m}_\square(Z, \delta u)\big|_{t=T}, \quad \forall \delta u \in \mathcal{U}_\square^0. \end{aligned} \tag{4.113}$$

Equation (4.113) can be decomposed as

$$\begin{cases} \mathfrak{a}_\square(u^{\star,s}, \delta u) = \mathfrak{m}_\square(X, \delta u) + \mathfrak{a}_\square(Y, \delta u), \quad \forall \delta u \in \mathbb{U}_\square^0, \quad \forall t \in (0,T), & \text{(4.114a)} \\ \frac{1}{2}\mathfrak{m}_\square(u^{\star,s}, \delta u)\big|_{t=0} = 0, \quad \forall \delta u \in \mathbb{U}_\square^0, & \text{(4.114b)} \\ \frac{1}{2}\mathfrak{m}_\square(u^{\star,s}, \delta u)\big|_{t=T} = \mathfrak{m}_\square(Z, \delta u)\big|_{t=T}, \quad \forall \delta u \in \mathbb{U}_\square^0. & \text{(4.114c)} \end{cases}$$

From Equations (4.114b) and (4.114c) it can be seen that $u^{\star,s}\big|_{t=0} = 0$ and $u^{\star,s}\big|_{t=T} = 2Z$.

Now find $u^{\star,s}$ for times between 0 and T, i.e. solve Equation (4.114a). Assume that $u^{\star,s}$ can be written as

$$u^{\star,s}(\boldsymbol{x},t) = \sum_{a=1}^{N} u_a(\boldsymbol{x})\xi_a^\star(t), \tag{4.115}$$

where $u_a(\boldsymbol{x})$ are the same spatial modes that are used to solve the micro-scale problem and $\xi_a^\star(t)$ are unknown functions of time for all $a \in [1,2,\ldots,N]$. The expression of $u^{\star,s}$ implies that the Equation (4.114a) can be expressed as

$$\sum_{a,b=1}^{N} \delta\xi_b\Big(\xi_a^\star \mathfrak{a}_\square(u_a, u_b)\Big) = \sum_{b=1}^{N} \delta\xi_b\Big(\mathfrak{m}_\square(X, u_b) + \mathfrak{a}_\square(Y, u_b)\Big), \quad \forall t \in (0,T). \tag{4.116}$$

This must hold for all test functions $\delta\xi_b \in \mathbb{L}^2$, which implies that

$$\sum_{a=1}^{N} \Big(\xi_a^\star \mathfrak{a}_\square(u_a, u_b)\Big) = \mathfrak{m}_\square(X, u_b) + \mathfrak{a}_\square(Y, u_b), \quad \forall t \in (0,T),\ b = 1,2,\dots,N. \tag{4.117}$$

The orthogonal properties of $\mathfrak{a}_\square$ gives

$$\lambda_a \xi_a^\star = \mathfrak{m}_\square(X, u_a) + \mathfrak{a}_\square(Y, u_a), \quad \forall t \in (0,T),\ a = 1,2,\dots,N, \tag{4.118}$$

and hence $\xi_a^\star$ can be calculated from

$$\xi_a^\star = \frac{1}{\lambda_a}\Big(\mathfrak{m}_\square(X, u_a) + \mathfrak{a}_\square(Y, u_a)\Big), \quad \forall t \in (0,T),\ a = 1,2,\dots,N. \tag{4.119}$$

When $\xi_a^\star$ is calculated, $u^{\star,s}$ can be found from Equation (4.115).

4.4.3 Derivation of an expression of the symmetric dual error

In order to apply Theorems 3 and 4, $e^{\star,s}$ needs to be calculated. The symmetric dual error is calculated from Equation (4.38) which says that: Find $e^{\star,s} \in \mathcal{U}_\square^0$ such that

$$A_\square^s(e^{\star,s}, \delta u) = R_\square^\star(\delta u), \quad \forall \delta u \in \mathcal{U}_\square^0. \tag{4.120}$$

By using the same argumentation as in Section 4.4.1, $A_\square^s(e^{\star,s}, \delta u)$ can be written as

$$A_\square^s(e^{\star,s}, \delta u) = \int_0^T \mathfrak{a}_\square(e^{\star,s}, \delta u)\,\mathrm{d}t + \frac{1}{2}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=0} + \frac{1}{2}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=T}. \tag{4.121}$$

The residual for the dual problem is defined as

$$R_\square^\star(\delta u) = Q_\square(\delta u) - A_\square(\delta u, u_R^\star). \tag{4.122}$$

Consistent with previous sections it is assumed that $Q(\delta u)$ can be written as

$$Q_\square(\delta u) = \int_0^T \mathfrak{m}_\square(X, \delta u) + \mathfrak{a}_\square(Y, \delta u)\,\mathrm{d}t + \mathfrak{m}_\square(Z, \delta u)\big|_{t=T}. \tag{4.123}$$

The definition of $A_\square$ gives

$$A_\square(\delta u, u_R^\star) = \int_0^T \mathfrak{m}_\square(\delta\dot{u}, u_R^\star) + \mathfrak{a}_\square(\delta u, u_R^\star)\,\mathrm{d}t + \mathfrak{m}_\square(\delta u, u_R^\star)\big|_{t=0}. \tag{4.124}$$

Partial integrating gives that

$$\begin{aligned}\int_0^T \mathfrak{m}_\square(\delta\dot{u}, u_R^\star)\,\mathrm{d}t &= \Big[\mathfrak{m}_\square(\delta u, u_R^\star)\Big]_{t=0}^T - \int_0^T \mathfrak{m}_\square(\delta u, \dot{u}_R^\star)\,\mathrm{d}t \\ &= -\int_0^T \mathfrak{m}_\square(\delta u, \dot{u}_R^\star)\,\mathrm{d}t + \mathfrak{m}_\square(\delta u, u_R^\star)\big|_{t=T} - \mathfrak{m}_\square(\delta u, u_R^\star)\big|_{t=0}.\end{aligned} \tag{4.125}$$

Inserting this expression into the expression of $A_\square(\delta u, u_R^\star)$ gives that

$$A_\square(\delta u, u_R^\star) = \int_0^T -\mathfrak{m}_\square(\delta u, \dot{u}_R^\star) + \mathfrak{a}_\square(\delta u, u_R^\star)\,\mathrm{d}t + \mathfrak{m}_\square(\delta u, u_R^\star)\big|_{t=T}. \tag{4.126}$$

By combining the expressions for $A^s_\square(\delta u, e^{\star,s})$, $Q_\square(\delta u)$ and $A_\square(\delta u, e^{\star,s})$, the weak form can be written as: Find $e^{\star,s} \in \mathcal{U}^0_\square$ such that

$$\begin{aligned}
&\int_0^T \mathfrak{a}_\square(e^{\star,s}, \delta u)\,\mathrm{d}t + \frac{1}{2}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=0} + \frac{1}{2}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=T} \\
&= \int_0^T \mathfrak{m}_\square(X, \delta u) + \mathfrak{a}_\square(Y, \delta u)\,\mathrm{d}t + \mathfrak{m}_\square(Z, \delta u)\big|_{t=T} \\
&+ \int_0^T \mathfrak{m}_\square(\delta u, \dot{u}^\star_R) - \mathfrak{a}_\square(\delta u, u^\star_R)\,\mathrm{d}t - \mathfrak{m}_\square(u^\star_R, \delta u)\big|_{t=T}, \quad \forall \delta u \in \mathcal{U}^0_\square.
\end{aligned} \tag{4.127}$$

Equation (4.127) can be decomposed as

$$\begin{cases}
\mathfrak{a}_\square(e^{\star,s}, \delta u) = \mathfrak{m}_\square(X, \delta u) + \mathfrak{a}_\square(Y, \delta u) \\
\qquad\qquad + \mathfrak{m}_\square(\delta u, \dot{u}^\star_R) - \mathfrak{a}_\square(\delta u, u^\star_R), \quad \forall \delta u \in \mathbb{U}^0_\square, \quad \forall t \in (0,T), & \text{(4.128a)} \\
\dfrac{1}{2}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=0} = 0, \quad \forall \delta u \in \mathbb{U}^0_\square, & \text{(4.128b)} \\
\dfrac{1}{2}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=T} = \mathfrak{m}_\square(Z - u^\star_R, \delta u)\big|_{t=T}, \quad \forall \delta u \in \mathbb{U}^0_\square. & \text{(4.128c)}
\end{cases}$$

From Equations (4.128b) and (4.128c) it can be seen that $e^{\star,s}\big|_{t=0} = 0$ and $e^{\star,s}\big|_{t=T} = 2Z - 2u^\star_R\big|_{t=T}$.

Now find $e^{\star,s}$ for times between 0 and T. Assume that $e^{\star,s}$ can be written as

$$e^{\star,s} = \sum_{a=1}^{N} u_a(\boldsymbol{x})\eta^\star_a(t), \tag{4.129}$$

where $u_a(\boldsymbol{x})$ are the same spatial modes that are used to solve the micro-scale problem and $\eta^\star_a(t)$ are unknown functions of time for all $a \in [1, 2, \ldots, N]$. The decomposition of $e^{\star,s}$ implies that Equation (4.128a) can be written as

$$\sum_{a,b=1}^{N} \delta\xi_b\Big(\eta^\star_a \mathfrak{a}_\square(u_a, u_b)\Big) = \sum_{b=1}^{N} \delta\xi_b\Big(\mathfrak{m}_\square(X, u_b) + \mathfrak{a}_\square(Y, u_b) + \sum_{a=1}^{N} \mathfrak{m}_\square(u_a, u_b)\dot{\xi}_{a,R} - \mathfrak{a}_\square(u_a, u_b)\xi_{a,R}\Big), \quad \forall t \in (0,T). \tag{4.130}$$

That this must hold for all $\xi_b \in \mathbb{L}^2$ gives that

$$\sum_{a=1}^{N} \eta^\star_a \mathfrak{a}_\square(u_a, u_b) = \mathfrak{m}_\square(X, u_b) + \mathfrak{a}_\square(Y, u_b) + \sum_{a=1}^{N} \mathfrak{m}_\square(u_a, u_b)\dot{\xi}^\star_{a,R} - \mathfrak{a}_\square(u_a, u_b)\xi^\star_{a,R}, \quad \forall t \in (0,T),\ b = 1, 2, \ldots N. \tag{4.131}$$

The orthogonal properties of $\mathfrak{a}_\square$ and $\mathfrak{m}_\square$ give that

$$\lambda_a \eta^\star_a = \mathfrak{m}_\square(X, u_a) + \mathfrak{a}_\square(Y, u_a) + \dot{\xi}^\star_{a,R} - \lambda_a \xi^\star_{a,R}, \quad \forall t \in (0,T),\ a = 1, 2, \ldots N, \tag{4.132}$$

and thus

$$\eta^\star_a = \frac{1}{\lambda_b}\Big(\mathfrak{m}_\square(X, u_a) + \mathfrak{a}_\square(Y, u_a) + \dot{\xi}^\star_{a,R} - \lambda_a \xi^\star_{a,R}\Big), \quad \forall t \in (0,T),\ a = 1, 2, \ldots N. \tag{4.133}$$

The expression of $\xi^\star_a$, see Equation (4.119), implies that

$$\eta^\star_a = \xi^\star_a + \frac{\dot{\xi}^\star_{a,R}}{\lambda_a} - \xi^\star_{a,R}, \quad \forall t \in (0,T),\ a = 1, 2, \ldots N. \tag{4.134}$$

Assume that N_R modes is used. Since $\dot{\xi}^\star_{a,R}$ and $\xi^\star_{a,R}$ lies in the reduced space they are equal to zero for a larger than N_R. Thus $\eta^\star_a$ can be written as

$$\eta^\star_a = \begin{cases} \xi^\star_a + \frac{\dot{\xi}^\star_{a,R}}{\lambda_a} - \xi^\star_{a,R}, & \forall t \in (0,T),\ a = 1,2,\dots,N_R. \\ \xi^\star_a, & \forall t \in (0,T),\ a = N_R+1, N_R+2,\dots,N. \end{cases} \tag{4.135}$$

But if the number of modes is equal or less to N_R, the residual must be equal to zero. The final expression of $\eta^\star_a$ can then be written as

$$\eta^\star_a = \begin{cases} 0, & \forall t \in (0,T),\ a = 1,2,\dots,N_R. \\ \xi^\star_a, & \forall t \in (0,T),\ a = N_R+1, N_R+2,\dots,N. \end{cases} \tag{4.136}$$

From $\eta^\star_a$, the symmetric dual error can be found from Equation (4.129).

4.5 Error estimates using only the reduced modes

One of the main theoretical goals of this thesis is to show how the error can be estimated when only the reduced modes are used. In Section 4.5.1 it is shown how this can be done with the energy norm and in Section 4.5.2 the corresponding expression for the goal-oriented approach is derived. In order to apply the Parallelogram law for the goal-oriented approach, further derivations are needed which are presented in Section 4.5.3.

4.5.1 Energy norm

In this section it is derived how the error analysis for the energy norm can be performed using only the reduced modes. The most central results are summarized in Box 4.5.1.

Box 4.5.1 (Error estimate using only the reduced modes with the energy norm)**.** *The error is defined from* $\|e\| := \sqrt{A_\square(e,e)} = \sqrt{A^s_\square(e,e)}$*. From Theorem 1 it follows that* $\|e\| \leq \|e^s\|$*. The so-called* $\mathfrak{a}$*-norm and* $\mathfrak{m}$*-norm are defined as* $\|\cdot\|_\mathfrak{a} = \sqrt{\mathfrak{a}(\cdot,\cdot)}$ *and* $\|\cdot\|_\mathfrak{m} = \sqrt{\mathfrak{m}(\cdot,\cdot)}$*. By using Theorem 6 it is derived that*

$$\begin{cases} \|e^s\|_\mathfrak{a} \leq \frac{\|r'_M\|_\mathfrak{m}}{\sqrt{\lambda_{N_R}}}, & \forall t \in (0,T) \\ \|e^s\|_\mathfrak{m}\big|_{t=0} \leq 2\|u'_0\|_\mathfrak{m} \\ \|e^s\|_\mathfrak{m}\big|_{t=T} = 0, \end{cases} \tag{4.137}$$

where λ_{N_R} *is the eigenvalue of the highest reduced mode,* $u'_0 = u_0 - \Pi_R u_0$*,* $r'_M = -r_M + \Pi_R r_M$ *and* $r_M = -\dot{\bar{u}} - \sum_{i=1}^d u^{(i)}_{stat}\dot{\bar{g}}_i$*. The energy norm of the error can then be estimated as*

$$\|e\| \leq \|e^s\| \leq \Big(\frac{1}{\lambda_{N_R}}\int_0^T \|r'_M\|^2_\mathfrak{m}\,\mathrm{d}t + 2\|u'_0\|^2_\mathfrak{m}\Big)^{1/2}. \tag{4.138}$$

In order to produce the error estimate when only the reduced modes are used, Cauchy-Schwarz inequality is applied and therefore a symmetric operator is needed. Thus it is the symmetric error that will be estimated which is defined from the symmetric error equation given by: Find $e^s \in \mathcal{U}^0_\square$ such that

$$A^s_\square(e^s, \delta u) = R_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.139}$$

As derived in Section 4.4.1, this equation can be written as

$$\begin{cases} \mathfrak{a}_\square(e^s, \delta u) = \mathfrak{r}_\square(\delta u), \quad \forall \delta u \in \mathbb{U}_\square^0, \quad \forall t \in (0,T), & \text{(4.140a)} \\ \frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} = \mathfrak{m}_\square(u_0 - u_R, \delta u)\big|_{t=0}, \quad \forall \delta u \in \mathbb{U}_\square^0, & \text{(4.140b)} \\ \frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=T} = 0, \quad \forall \delta u \in \mathbb{U}_\square^0, & \text{(4.140c)} \end{cases}$$

where $\mathfrak{r}_\square(\delta u) := -\mathfrak{m}_\square(\dot{u}_R, \delta u) - \mathfrak{a}_\square(u_R, \delta u)$. Define the $\mathfrak{a}$-norm as

$$\|\cdot\|_\mathfrak{a} := \sqrt{\mathfrak{a}_\square(\cdot,\cdot)}. \tag{4.141}$$

The symmetric error for a time $t \in (0,T)$, measured in the $\mathfrak{a}$-norm, can then be written as

$$\|e^s\|_\mathfrak{a}^2 = \mathfrak{a}_\square(e^s, e^s) = \mathfrak{r}_\square(e^s), \quad \forall t \in (0,T). \tag{4.142}$$

Hence the symmetric error can be estimated by estimating $\mathfrak{r}_\square(e^s)$. Define a new weak form given by: Find $r_M \in \mathbb{U}_\square$ such that

$$\mathfrak{m}_\square(r_M, \delta u) = \mathfrak{r}_\square(\delta u), \quad \forall \delta u \in \mathbb{U}_\square^0, \quad \forall t \in (0,T). \tag{4.143}$$

The following lemmas and theorem will prove how $\|e^s\|_\mathfrak{a}$ can be estimated with r_M.

Lemma 4. *Assume that*

$$\lambda_i \mathfrak{m}_\square(u_i, \delta u) = \mathfrak{a}_\square(u_i, \delta u), \quad \forall \delta u \quad \textit{and} \quad \forall t \in (0,T), \tag{4.144}$$

where $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_N$ *and* $1 \leq N_R < N$ *for the reduced basis* $\mathbb{U}_\square^R = \mathit{span}\{u_i\}_{i=1}^{N_R}$. *Let*

$$\Pi_R = \{\mathbb{U}_\square \to \mathbb{U}_\square^R : \ \mathfrak{m}_\square(\Pi_R u, \delta u) = \mathfrak{m}_\square(u, \delta u), \ \forall \delta u \in \mathbb{U}_\square^R\} \tag{4.145}$$

and assume that $u = \sum_{a=1}^N \xi_a u_a$. *Then*

$$\Pi_R u = \sum_{a=1}^{N_R} \xi_a u_a \quad \textit{and} \quad u - \Pi_R u = \sum_{a=N_R+1}^{N} \xi_a u_a, \tag{4.146}$$

for all $t \in (0,T)$.

Proof. Let $u_a \in \mathbb{U}_\square^R$ for all $a \leq N_R$. The linearity of $\mathfrak{m}_\square$ and the $\mathfrak{m}$-orthogonality give that

$$\mathfrak{m}_\square(\Pi_R u, u_a) = \mathfrak{m}_\square(\Pi_R \sum_{b=1}^N \xi_b u_b, u_a) = \sum_{b=1}^N \Pi_R \xi_b \mathfrak{m}_\square(u_b, u_a) = \Pi_R \xi_a \tag{4.147}$$

and similarly

$$\mathfrak{m}_\square(u, u_a) = \mathfrak{m}_\square(\sum_{b=1}^N \xi_b u_b, u_a) = \sum_{b=1}^N \xi_b \mathfrak{m}_\square(u_b, u_a) = \xi_a. \tag{4.148}$$

From the definition of Π_R it is known that $\mathfrak{m}_\square(\Pi_R u, u_a) = \mathfrak{m}_\square(u, u_a)$ for all $u_a \in \mathbb{U}_\square^R$ and thus

$$\Pi_R \xi_a = \xi_a \quad \text{if} \quad a \leq N_R. \tag{4.149}$$

$\Pi_R \xi_a$ projects ξ_a onto the $\mathbb{U}_\square^R$ space and therefore

$$\Pi_R \xi_a = 0 \quad \text{if} \quad a > N_R. \tag{4.150}$$

Equations (4.149) and (4.150) implies that

$$\Pi_R u = \sum_{a=1}^{N_R} \xi_a u_a \tag{4.151}$$

and thus

$$u - \Pi_R u = \sum_{a=1}^{N} \xi_a u_a - \sum_{a=1}^{N_R} \xi_a u_a = \sum_{a=N_R+1}^{N} \xi_a u_a. \tag{4.152}$$

□

Lemma 5. *Assume that Lemma 4 is fulfilled. Then*

$$\mathfrak{m}_\square(u - \Pi_R u, u - \Pi_R u) \leq \frac{1}{\lambda_{N_R}} \mathfrak{a}_\square(u - \Pi_R u, u - \Pi_R u), \quad \forall t \in (0, T). \tag{4.153}$$

Proof. From Lemma 4 it follows that

$$u - \Pi_a u = \sum_{a=N_R+1}^{N} \xi_a u_a. \tag{4.154}$$

Hence

$$\mathfrak{m}_\square(u - \Pi_R u, u - \Pi_R u) = \mathfrak{m}_\square\Big(\sum_{a=N_R+1}^{N} \xi_a u_a, \sum_{b=N_R+1}^{N} \xi_b u_b \Big) = \sum_{a,b=N_R+1}^{N} \xi_a \mathfrak{m}_\square(u_a, u_b) \xi_b. \tag{4.155}$$

With the $\mathfrak{m}$-orthogonality, the expression can be simplified into

$$\sum_{a,b=N_R+1}^{N} \xi_a \mathfrak{m}_\square(u_a, u_b) \xi_b = \sum_{a=N_R+1}^{N} \xi_a^2 = \sum_{a=N_R+1}^{N} \frac{\lambda_a}{\lambda_a} \xi_a^2 \leq \frac{1}{\underset{a>N_R}{min}\, \lambda_a} \sum_{a=N_R+1}^{N} \lambda_a \xi_a^2. \tag{4.156}$$

Using the $\mathfrak{a}$-orthogonality, and the fact that the eigenvalue sequence is an increasing sequence, give that

$$\mathfrak{m}_\square(u - \Pi_R u, u - \Pi_R u) \leq \frac{1}{\lambda_{N_R}} \sum_{a=N_R+1}^{N} \lambda_a \xi_a^2 = \frac{1}{\lambda_{N_R}} \mathfrak{a}_\square(u - \Pi_R u, u - \Pi_R u). \tag{4.157}$$

□

Theorem 6. *Assume that*

$$\mathfrak{a}_\square(e^s, \delta u) = \mathfrak{m}_\square(r_M, \delta u), \quad \forall \delta u \in \mathbb{U}_\square, \quad \forall t \in (0, T), \tag{4.158}$$

for some $r_M \in \mathbb{U}_\square$ and that the assumptions in Lemma 4 is fulfilled. Let $\| \cdot \|_{\mathfrak{m}} = \sqrt{\mathfrak{m}_\square(\cdot, \cdot)}$. Then

$$\|e^s\|_{\mathfrak{a}} \leq \frac{\|r_M\|_{\mathfrak{m}}}{\sqrt{\lambda_{N_R}}}, \quad \forall t \in (0, T). \tag{4.159}$$

Proof. The theorem is fulfilled in the trivial case when $\|e^s\|_{\mathfrak{a}} = 0$, since $\|r\|_{\mathfrak{m}}$ is non-negative and λ_{N_R} is positive. Now prove the theorem when $\|e^s\|_{\mathfrak{a}} > 0$. The definition of the energy norm gives that

$$\|e^s\|_{\mathfrak{a}}^2 = \mathfrak{a}_\square(e^s, e^s), \quad \forall t \in (0, T). \tag{4.160}$$

Cauchy-Schwarz inequality gives that

$$\|e^s\|_{\mathfrak{a}}^2 = \mathfrak{a}_\square(e^s, e^s) = \mathfrak{m}_\square(r_M, e^s) \leq \|r\|_{\mathfrak{m}}\|e^s\|_{\mathfrak{m}}, \quad \forall t \in (0,T). \tag{4.161}$$

The Galerkin orthogonality implies that $\Pi_R e^s = 0$ and hence

$$\|e^s\|_{\mathfrak{a}}^2 \leq \|r_M\|_{\mathfrak{m}}\|e^s - \Pi_R e^s\|_{\mathfrak{m}}, \quad \forall t \in (0,T). \tag{4.162}$$

Lemma 5 implies that

$$\|e^s\|_{\mathfrak{a}}^2 \leq \|r_M\|_{\mathfrak{m}}\|e^s - \Pi_R e^s\|_{\mathfrak{m}} \leq \|r_M\|_{\mathfrak{m}}\frac{\|e^s - \Pi_R e^s\|_{\mathfrak{a}}}{\sqrt{\lambda_{N_R}}} = \|r\|_{\mathfrak{m}}\frac{\|e^s\|_{\mathfrak{a}}}{\sqrt{\lambda_{N_R}}}, \quad \forall t \in (0,T). \tag{4.163}$$

It is possible to divide with $\|e^s\|_{\mathfrak{a}}$ since it is assumed that $\|e^s\|_{\mathfrak{a}} > 0$. Thus

$$\|e^s\|_{\mathfrak{a}} \leq \frac{\|r_M\|_{\mathfrak{m}}}{\sqrt{\lambda_{N_R}}}, \quad \forall t \in (0,T). \tag{4.164}$$

□

The total estimated error can be obtained by adding the contribution for all times as well as the boundary terms. The contribution at $t = 0$ can be estimated from Equation (4.140b) which says that

$$\frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} = \mathfrak{m}_\square(u_0 - u_R, \delta u)\big|_{t=0}. \tag{4.165}$$

The Galerkin orthogonality gives that

$$\begin{aligned}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} =& 2\mathfrak{m}_\square(u_0 - u_R, \delta u)\big|_{t=0} \\ =& 2\mathfrak{m}_\square(u_0 - u_R, \delta u - \Pi_R \delta u)\big|_{t=0} \\ =& 2\mathfrak{m}_\square(u_0 - u_R - \Pi_R u_0 + \Pi_R u_R, \delta u)\big|_{t=0}.\end{aligned} \tag{4.166}$$

The relation $u_R = \Pi_R u_R$ implies that

$$\begin{aligned}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=0} &= 2\mathfrak{m}_\square(u_0 - u_R - \Pi_R u_0 + \Pi_R u_R, \delta u)\big|_{t=0} \\ &= \mathfrak{m}_\square(2u_0 - 2\Pi_R u_0, \delta u)\big|_{t=0}.\end{aligned} \tag{4.167}$$

Set $\delta u\big|_{t=0} = e^s\big|_{t=0}$. Cauchy–Schwarz inequality gives

$$\begin{aligned}\|e^s\|_{\mathfrak{m}}^2\big|_{t=0} = \mathfrak{m}_\square(e^s, e^s)\big|_{t=0} &= \mathfrak{m}_\square(2u_0 - 2\Pi_R u_0, e^s)\big|_{t=0} \\ &\leq \sqrt{\mathfrak{m}_\square(2u_0 - 2\Pi_R u_0, 2u_0 - 2\Pi_R u_0)}\sqrt{\mathfrak{m}_\square(e^s, e^s)\big|_{t=0}} \\ &= \|2u_0 - 2\Pi_R u_0\|_{\mathfrak{m}}\|e^s\|_{\mathfrak{m}}\big|_{t=0}\end{aligned} \tag{4.168}$$

which implies that

$$\|e^s\|_{\mathfrak{m}}\big|_{t=0} \leq \|2u_0 - 2\Pi_R u_0\|_{\mathfrak{m}}. \tag{4.169}$$

Concerning the other boundary term, Equation (4.140c) says that

$$\frac{1}{2}\mathfrak{m}_\square(e^s, \delta u)\big|_{t=T} = 0 \tag{4.170}$$

which implies that

$$e^s\big|_{t=T} \equiv 0. \tag{4.171}$$

By adding the contributions for the boundary terms, together with the linearity of $\mathfrak{m}_\square$ and letting $u_0' := u_0 - \Pi_R u_0$ give

$$\frac{1}{2}\|e^s\|_{\mathfrak{m}}^2\big|_{t=0} + \frac{1}{2}\|e^s\|_{\mathfrak{m}}^2\big|_{t=T} \leq \frac{1}{2}\|2u_0 - 2\Pi_R u_0\|_{\mathfrak{m}}^2 = 2\|u_0'\|_{\mathfrak{m}}^2. \tag{4.172}$$

The total error can now be estimated with Equation (4.172) and Theorem 6 as

$$\begin{aligned}\|e^s\| &:= \sqrt{A_\square^s(e^s, e^s)} = \Big(\int_0^T \|e^s\|_{\mathfrak{a}}^2 \,\mathrm{d}t + \frac{1}{2}\|e^s\|_{\mathfrak{m}}^2\big|_{t=0} + \frac{1}{2}\|e^s\|_{\mathfrak{m}}^2\big|_{t=T}\Big)^{1/2} \\ &\leq \Big(\frac{1}{\lambda_{N_R}}\int_0^T \|r_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t + 2\|u_0'\|_{\mathfrak{m}}^2\Big)^{1/2}.\end{aligned} \tag{4.173}$$

The error can now be estimated if r_M is calculated. r_M is defined in Equation (4.143) which says that: Find $r_M \in \mathbb{U}_\square$ such that

$$\mathfrak{m}_\square(r_M, \delta u) = \mathfrak{r}_\square(\delta u), \quad \forall \delta u \in \mathbb{U}_\square. \tag{4.174}$$

The definition of $\mathfrak{r}_\square(\delta u)$ says that

$$\mathfrak{r}_\square(\delta u) = -\mathfrak{m}_\square(\dot{u}_R, \delta u) - \mathfrak{a}_\square(u_R, \delta u). \tag{4.175}$$

If the test function lies in the space of the used modes the residual is equal to zero, i.e.

$$\mathfrak{r}_\square(\delta u) = 0, \quad \text{if} \quad \delta u \in \mathbb{U}_\square^R. \tag{4.176}$$

If the test function does not lie in the space of the used modes, i.e. $\delta u \notin \mathbb{U}_\square^R$, then only the stationary mode is non-zero and

$$\begin{cases}\mathfrak{m}_\square(\dot{u}_R, \delta u) = \mathfrak{m}_\square(1, \delta u)\ddot{\bar{u}} + \sum_{i=1}^d \mathfrak{m}_\square(u_{stat}^{(i)}, \delta u)\ddot{\bar{g}}_i \\ \mathfrak{a}_\square(u_R, \delta u) = \mathfrak{a}_\square(1, \delta u)\dot{\bar{u}} + \sum_{i=1}^d \mathfrak{a}_\square(u_{stat}^{(i)}, \delta u)\dot{\bar{g}}_i = 0.\end{cases} \tag{4.177}$$

Hence

$$\mathfrak{r}_\square(\delta u) = -\mathfrak{m}_\square(1, \delta u)\ddot{\bar{u}} - \sum_{i=1}^d \mathfrak{m}_\square(u_{stat}^{(i)}, \delta u)\ddot{\bar{g}}_i = -\mathfrak{m}_\square\Big(\ddot{\bar{u}} + \sum_{i=1}^d u_{stat}^{(i)}\ddot{\bar{g}}_i, \delta u\Big). \tag{4.178}$$

From Equations (4.174) and (4.178) it can be concluded that

$$\mathfrak{m}_\square(r_M, \delta u) = -\mathfrak{m}_\square\Big(\ddot{\bar{u}} + \sum_{i=1}^d u_{stat}^{(i)}\ddot{\bar{g}}_i, \delta u\Big), \quad \forall \delta u \in \mathbb{U}_\square \tag{4.179}$$

and thus

$$r_M = -\ddot{\bar{u}} - \sum_{i=1}^d u_{stat}^{(i)}\ddot{\bar{g}}_i. \tag{4.180}$$

By using the Galerkin orthogonality a sharper error estimate of $\mathfrak{r}_\square(\delta u)$ can be obtained in the following way. $\mathfrak{r}_\square(\Pi_R \delta u) = 0$ implies that

$$\begin{aligned}\mathfrak{r}_\square(\delta u) &= \mathfrak{r}_\square(\delta u - \Pi_R \delta u) = \mathfrak{m}_\square(\delta u - \Pi_R \delta u, -r_M) \\ &= \mathfrak{m}_\square(\delta u, -r_M + \Pi_R r_M) =: \mathfrak{m}_\square(\delta u, r_M'),\end{aligned} \tag{4.181}$$

where

$$
\begin{aligned}
r'_M &= -r_M + \Pi_R r_M = \dot{\tilde{u}}1 + \sum_{i=1}^{d} u_{stat}^{(i)} \dot{\tilde{g}}_i - \sum_{a=1}^{N_R} \mathfrak{m}_\square(1, u_a) u_a(\boldsymbol{x}) \dot{\tilde{u}} - \mathfrak{m}_\square(u_{stat}^{(i)}, u_a) u_a(\boldsymbol{x}) \dot{\tilde{g}}_i \\
&= \Big(1 - \sum_{a=1}^{N_R} \mathfrak{m}_\square(1, u_a) u_a(\boldsymbol{x})\Big) \dot{\tilde{u}} + \sum_{i=1}^{d} \Big(u_{stat}^{(i)} - \sum_{a=1}^{N_R} \mathfrak{m}_\square(u_{stat}^{(i)}, u_a) u_a(\boldsymbol{x})\Big) \dot{\tilde{g}}_i.
\end{aligned}
\tag{4.182}
$$

Remark: Even though the error estimate is sharper when the projection is used, it is possible to define an error estimate of $\|e^s\|$ that does not use the Π_R-operator. The corresponding estimate of the error is then given by

$$
\|e^s\| \leq \Big(\frac{1}{\sqrt{\lambda_{N_R}}} \int_0^T \|r_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t + 2\|u_0 - u_R\big|_{t=0}\|_{\mathfrak{m}}^2\Big)^{1/2}. \tag{4.183}
$$

Since this error estimate never uses the projection properties, a large overestimate of the error is induced, see Section 5.4.

4.5.2 Goal-oriented approach

In this section it is shown how the goal-oriented approach can be applied to estimate the error for different quantities of interest using only the reduced modes. The most central results are summarized in Box 4.5.2.

Box 4.5.2 (Error estimate using only the reduced modes with Cauchy–Schwarz theorem)**.** *The error measured in a given quantity of interest is defined as* $E := Q_\square(e) = A_\square^s(e, u^\star)$*. By using Theorem 6 it is derived that*

$$
\begin{cases}
\|e^{\star,s}\|_{\mathfrak{a}} \leq \frac{1}{\sqrt{\lambda_{N_R}}} \|X'\|_{\mathfrak{m}} + \|Y'\|_{\mathfrak{a}}, \quad \forall t \in (0,T), \\
\|e^s\|_{\mathfrak{m}}\big|_{t=0} = 0, \\
\|e^s\|_{\mathfrak{m}}\big|_{t=T} \leq 2\|Z'\|_{\mathfrak{m}},
\end{cases}
\tag{4.184}
$$

where $X' = X - \Pi_R X$*,* $Y' = Y - \Pi_R Y$ *and* $Z' = Z - \Pi_R Z$*. The energy norm of the symmetric dual error can then be estimated as*

$$
\|e^{\star,s}\| \leq \Big(\int_0^T \big(\frac{1}{\sqrt{\lambda_{N_R}}} \|X'\|_{\mathfrak{m}} + \|Y'\|_{\mathfrak{a}}\big)^2 \,\mathrm{d}t + 2\|Z'\|_{\mathfrak{m}}^2\Big)^{1/2}. \tag{4.185}
$$

From Theorem 3 it is derived that

$$
|E| \leq \|e^s\| \|e^{\star,s}\|. \tag{4.186}
$$

Inserting the expressions of $\|e^s\|$ *and* $\|e^{\star,s}\|$ *from Box 4.5.1 and Equation* (4.185) *into the Equation* (4.186) *give the estimated expression of* $|E|$*.*

From Section 4.4.3 it is derived that

$$
\begin{cases}
\mathfrak{a}_\square(e^{\star,s}, \delta u) = \mathfrak{m}_\square(X, \delta u) + \mathfrak{a}_\square(Y, \delta u) \\
\qquad\qquad + \mathfrak{m}_\square(\delta u, \dot{u}_R^\star) - \mathfrak{a}_\square(\delta u, u_R^\star), \quad \forall \delta u \in \mathbb{U}_\square^0, \quad \forall t \in (0,T), & \text{(4.187a)} \\
\frac{1}{2}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=0} = 0, \quad \forall \delta u \in \mathbb{U}_\square^0, & \text{(4.187b)} \\
\frac{1}{2}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=T} = \mathfrak{m}_\square(Z - u_R^\star, \delta u)\big|_{t=T}, \quad \forall \delta u \in \mathbb{U}_\square^0. & \text{(4.187c)}
\end{cases}
$$

Set

$$\mathfrak{r}^\star_\square(\delta u) = \mathfrak{m}_\square(X, \delta u) + \mathfrak{a}_\square(Y, \delta u) + \mathfrak{m}_\square(\delta u, \dot{u}^\star_R) - \mathfrak{a}_\square(\delta u, u^\star_R). \tag{4.188}$$

Equation (4.187a), together with the linearity of symmetry of $\mathfrak{m}_\square$ and $\mathfrak{a}_\square$, imply that

$$\begin{aligned}\mathfrak{a}_\square(e^{\star,s}, \delta u) = \mathfrak{r}^\star_\square(\delta u) &= \mathfrak{m}_\square(X, \delta u) + \mathfrak{a}_\square(Y, \delta u) + \mathfrak{m}_\square(\delta u, \dot{u}^\star_R) - \mathfrak{a}_\square(\delta u, u^\star_R) \\ &= \mathfrak{m}_\square(X + \dot{u}^\star_R, \delta u) + \mathfrak{a}_\square(Y - u^\star_R, \delta u), \quad \forall \delta u \in \mathbb{U}^0_\square, \quad \forall t \in (0, T).\end{aligned} \tag{4.189}$$

Using the identity $\mathfrak{r}_\square(\Pi_R \delta u) = 0$ gives that

$$\begin{aligned}\mathfrak{a}_\square(e^{\star,s}, \delta u) &= \mathfrak{m}_\square(X + \dot{u}^\star_R, \delta u - \Pi_R \delta u) + \mathfrak{a}_\square(Y - u^\star_R, \delta u - \Pi_R \delta u) \\ &= \mathfrak{m}_\square(X + \dot{u}^\star_R - \Pi_R X - \Pi_R \dot{u}^\star_R, \delta u) + \mathfrak{a}_\square(Y - \Pi_R Y - u^\star_R + \Pi_R u^\star_R, \delta u) \\ &= \mathfrak{m}_\square(X - \Pi_R X, \delta u) + \mathfrak{a}_\square(Y - \Pi_R Y, \delta u), \quad \forall \delta u \in \mathbb{U}^0_\square, \quad \forall t \in (0, T).\end{aligned} \tag{4.190}$$

Setting $\delta u = e^{\star,s}$ together with Theorem 6 imply that

$$\begin{aligned}\|e^{\star,s}\|^2_{\mathfrak{a}} = \mathfrak{a}_\square(e^{\star,s}, e^{\star,s}) &= \mathfrak{m}_\square(X - \Pi_R X, e^{\star,s}) + \mathfrak{a}_\square(Y - \Pi_R Y, \delta u) \\ &\leq \|X - \Pi_R X\|_{\mathfrak{m}} \|e^{\star,s}\|_{\mathfrak{m}} + \|Y - \Pi_R Y\|_{\mathfrak{a}} \|e^{\star,s}\|_{\mathfrak{a}} \\ &\leq \frac{1}{\sqrt{\lambda_{N_R}}} \|X - \Pi_R X\|_{\mathfrak{m}} \|e^{\star,s}\|_{\mathfrak{a}} + \|Y - \Pi_R Y\|_{\mathfrak{a}} \|e^{\star,s}\|_{\mathfrak{a}}, \quad \forall t \in (0, T)\end{aligned} \tag{4.191}$$

and therefore

$$\|e^{\star,s}\|_{\mathfrak{a}} \leq \frac{1}{\sqrt{\lambda_{N_R}}} \|X - \Pi_R X\|_{\mathfrak{m}} + \|Y - \Pi_R Y\|_{\mathfrak{a}}, \quad \forall t \in (0, T). \tag{4.192}$$

Now estimate the contribution from the boundary terms. From Equation (4.187b) it follows that $\|e^{\star,s}\|\big|_{t=0} = 0$. The contribution at $t = T$ can be estimated from Equation (4.187c) as

$$\begin{aligned}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=T} &= 2\mathfrak{m}_\square(Z - u^\star_R, \delta u)\big|_{t=T} \\ &= 2\mathfrak{m}_\square(Z - u^\star_R, \delta u - \Pi_R \delta u)\big|_{t=T} \\ &= 2\mathfrak{m}_\square(Z - u^\star_R - \Pi_R Z + \Pi_R u^\star_R, \delta u)\big|_{t=T}.\end{aligned} \tag{4.193}$$

The relation $u^\star_R = \Pi_R u^\star_R$ implies that

$$\begin{aligned}\mathfrak{m}_\square(e^{\star,s}, \delta u)\big|_{t=T} &= 2\mathfrak{m}_\square(Z - u^\star_R - \Pi_R Z + \Pi_R u^\star_R, \delta u)\big|_{t=T} \\ &= \mathfrak{m}_\square(2Z - 2\Pi_R Z, \delta u)\big|_{t=T}.\end{aligned} \tag{4.194}$$

Set $\delta u\big|_{t=T} = e^{\star,s}\big|_{t=T}$. Cauchy–Schwarz inequality gives

$$\begin{aligned}\|e^{\star,s}\|^2_{\mathfrak{m}}\big|_{t=T} = \mathfrak{m}_\square(e^{\star,s}, e^{\star,s})\big|_{t=T} &= \mathfrak{m}_\square(2Z - 2\Pi_R Z, e^{\star,s})\big|_{t=T} \\ &\leq \sqrt{\mathfrak{m}_\square(2Z - 2\Pi_R Z, 2Z - 2\Pi_R Z)} \sqrt{\mathfrak{m}_\square(e^{\star,s}, e^{\star,s})\big|_{t=T}} \\ &= \|2Z - 2\Pi_R Z\|_{\mathfrak{m}} \|e^{\star,s}\|_{\mathfrak{m}}\big|_{t=T}\end{aligned} \tag{4.195}$$

which implies that

$$\|e^{\star,s}\|_{\mathfrak{m}}\big|_{t=T} \leq \|2Z - 2\Pi_R Z\|_{\mathfrak{m}}. \tag{4.196}$$

By adding the contributions for the boundary terms, together with the linearity of $\mathfrak{m}_\square$, give

$$\frac{1}{2}\|e^{\star,s}\|^2_{\mathfrak{m}}\big|_{t=0} + \frac{1}{2}\|e^{\star,s}\|^2_{\mathfrak{m}}\big|_{t=T} \leq \frac{1}{2}\|2Z - 2\Pi_R Z\|^2_{\mathfrak{m}} = 2\|Z - \Pi_R Z\|^2_{\mathfrak{m}}. \tag{4.197}$$

The total error can now be estimated as

$$\begin{aligned}\|e^{\star,s}\| &= \sqrt{A^s_\square(e^{\star,s},e^{\star,s})} = \Big(\int_0^T \|e^{\star,s}\|^2_{\mathfrak{a}}\,\mathrm{d}t + \frac{1}{2}\|e^{\star,s}\|^2_{\mathfrak{m}}\big|_{t=0} + \frac{1}{2}\|e^{\star,s}\|^2_{\mathfrak{m}}\big|_{t=T}\Big)^{1/2} \\ &\leq \Big(\int_0^T \big(\frac{1}{\sqrt{\lambda_{N_R}}}\|X-\Pi_R X\|_{\mathfrak{m}} + \|Y-\Pi_R Y\|_{\mathfrak{a}}\big)^2\,\mathrm{d}t + 2\|Z-\Pi_R Z\|^2_{\mathfrak{m}}\Big)^{1/2}.\end{aligned} \tag{4.198}$$

By letting $X' := X - \Pi_R X$, $Y' := Y - \Pi_R Y$ and $Z' := Z - \Pi_R Z$, Equation (4.198) can be written as

$$\|e^{\star,s}\| \leq \Big(\int_0^T \big(\frac{1}{\sqrt{\lambda_{N_R}}}\|X'\|_{\mathfrak{m}} + \|Y'\|_{\mathfrak{a}}\big)^2\,\mathrm{d}t + 2\|Z'\|^2_{\mathfrak{m}}\Big)^{1/2}. \tag{4.199}$$

Remark: Even though the estimate is sharper when the projection is used, it is possible to define an error estimate of $\|e^{\star,s}\|$ that does not use the Π_R-operator. The corresponding error estimate of the error is then given by

$$\|e^{\star,s}\| \leq \Big(\int_0^T \big(\frac{1}{\sqrt{\lambda_{N_R}}}\|X+\dot{u}^\star_R\|_{\mathfrak{m}} + \|Y-u^\star_R\|_{\mathfrak{a}}\big)^2\,\mathrm{d}t + 2\|Z-u^\star_R\big|_{t=T}\|^2_{\mathfrak{m}}\Big)^{1/2}. \tag{4.200}$$

Since this error estimate never uses the projection properties, a large overestimate of the error in induced and therefore this error estimate is never implemented in the numerical results.

4.5.3 Parallelogram law

In this section, detailed derivations of how E can be estimated with the Parallelogram law is presented. For the reader who is just interested in the most central equations and the most important results, see Box 4.5.3.

Box 4.5.3 (Error estimate using only the reduced modes with the Parallelogram law)**.** *The error measured in a given quantity of interest is defined as $E := Q_\square(e) = A^s_\square(e,u^\star)$. Theorem 4 implies that*

$$\begin{aligned}E &\leq \frac{1}{4}\|\kappa e^s + \frac{1}{\kappa}e^{\star,s}\|^2 \\ E &\geq -\frac{1}{4}\|\kappa e^s - \frac{1}{\kappa}e^{\star,s}\|^2.\end{aligned} \tag{4.201}$$

In the general case, E can be estimated using only the reduced modes as

$$\begin{cases} E \leq \displaystyle\int_0^T \Big[\frac{1}{4\lambda_{N_R}}\Big(\kappa^2\|r'_M\|^2_{\mathfrak{m}} + 2\mathfrak{m}_\square(r'_M,X') + \frac{1}{\kappa^2}\|X'\|^2_{\mathfrak{m}}\Big) \\ \quad + \dfrac{1}{2\kappa\sqrt{\lambda_{N_R}}}\|\kappa r'_M + \frac{1}{\kappa}X'\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}} + \frac{1}{\kappa^2}\|Y'\|^2_{\mathfrak{a}}\Big]\,\mathrm{d}t + \dfrac{\kappa^2}{2}\|u'_0\|^2_{\mathfrak{m}} + \dfrac{1}{2\kappa^2}\|Z'\|^2_{\mathfrak{m}}, \\ E \geq -\displaystyle\int_0^T \Big[\frac{1}{4\lambda_{N_R}}\Big(\kappa^2\|r'_M\|^2_{\mathfrak{m}} - 2\mathfrak{m}_\square(r'_M,X') + \frac{1}{\kappa^2}\|X'\|^2_{\mathfrak{m}}\Big) \\ \quad - \dfrac{1}{2\kappa\sqrt{\lambda_{N_R}}}\|\kappa r'_M + \frac{1}{\kappa}X'\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}} + \frac{1}{\kappa^2}\|Y'\|^2_{\mathfrak{a}}\Big]\,\mathrm{d}t - \dfrac{\kappa^2}{2}\|u'_0\|^2_{\mathfrak{m}} - \dfrac{1}{2\kappa^2}\|Z'\|^2_{\mathfrak{m}}. \end{cases} \tag{4.202}$$

*When the **average temperature** is considered, E is estimated with Equations* (4.233) *and*

(4.234). *From Theorem 7 it follows that the best error estimate is obtained if*

$$\kappa_{X',opt} = \left(\frac{\int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t} \right)^{1/4}. \tag{4.203}$$

With this value of κ, E can be estimated via the formulas

$$\begin{cases} E \leq \frac{1}{2\lambda_{N_R}} \int_0^T \mathfrak{m}_\square(r'_M, X') \,\mathrm{d}t + \frac{1}{2\lambda_{N_R}} \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2}, \\ E \geq \frac{1}{2\lambda_{N_R}} \int_0^T \mathfrak{m}_\square(r'_M, X') \,\mathrm{d}t - \frac{1}{2\lambda_{N_R}} \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2}. \end{cases} \tag{4.204}$$

When the ***average heat flux in one direction*** *is considered, E is estimated with Equations* (4.243) *and* (4.244). *Equation* (4.245) *gives that the optimal error estimate of E is obtained if*

$$\kappa_{Y',opt} = \left(\frac{\lambda_{N_R} \int_0^T \|Y'\|_{\mathfrak{a}}^2 \,\mathrm{d}t}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t} \right)^{1/4}. \tag{4.205}$$

With this value of κ, E can be estimated via the formulas

$$\begin{cases} E \leq \frac{1}{2\sqrt{\lambda_{N_R}}} \int_0^T \|r'_M\|_{\mathfrak{m}} \|Y'\|_{\mathfrak{a}} \,\mathrm{d}t + \frac{1}{2\sqrt{\lambda_{N_R}}} \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|Y'\|_{\mathfrak{a}}^2 \,\mathrm{d}t \Big)^{1/2}, \\ E \geq \frac{1}{2\sqrt{\lambda_{N_R}}} \int_0^T \|r'_M\|_{\mathfrak{m}} \|Y'\|_{\mathfrak{a}} \,\mathrm{d}t - \frac{1}{2\sqrt{\lambda_{N_R}}} \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|Y'\|_{\mathfrak{a}}^2 \,\mathrm{d}t \Big)^{1/2}. \end{cases} \tag{4.206}$$

When the ***final temperature*** *is considered, E is estimated with Equations* (4.249) *and* (4.250). *Equation* (4.251) *gives that the optimal error estimate of E is obtained if*

$$\kappa_{Z,opt} = \left(\frac{2\lambda_{N_R} \|Z'\|_{\mathfrak{m}}^2}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t} \right)^{1/4}. \tag{4.207}$$

With this value of κ, E can be estimated via the formulas

$$\begin{cases} E \leq \frac{1}{\sqrt{2}} \|Z'\|_{\mathfrak{m}}^2 \Big(\frac{1}{\lambda_{N_R}} \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2}, \\ E \geq -\frac{1}{\sqrt{2}} \|Z'\|_{\mathfrak{m}}^2 \Big(\frac{1}{\lambda_{N_R}} \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2}. \end{cases} \tag{4.208}$$

In order to apply the Parallelogram law using only the reduced modes, $\|\kappa e^s \pm \frac{1}{\kappa} e^{\star,s}\|$ needs to be estimated. Start with the symmetric error equation which is given by

$$A^s_\square(e^s, \delta u) = R_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.209}$$

The linearity of $A^s_\square$ implies that the equation can be written as

$$A^s_\square(\kappa e^s, \delta u) = \kappa R_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square, \tag{4.210}$$

where $\kappa \neq 0$. The same argumentation for the symmetric dual error equation gives that

$$A^s_\square(\frac{1}{\kappa} e^{\star,s}, \delta u) = \frac{1}{\kappa} R^\star_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.211}$$

Adding Equations (4.210) and (4.211) imply that

$$A^s_\square(\kappa e^s, \delta u) + A^s_\square(\frac{1}{\kappa} e^{\star,s}, \delta u) = \kappa R_\square(\delta u) + \frac{1}{\kappa} R^\star_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square. \tag{4.212}$$

Setting $e^{s,\kappa} := \kappa e^s + \frac{1}{\kappa} e^{\star,s}$ and once again using the linearity of $A^s_\square$ imply that

$$A^s_\square(e^{s,\kappa}, \delta u) = \kappa R_\square(\delta u) + \frac{1}{\kappa} R^\star_\square(\delta u), \quad \forall \delta u \in \mathcal{U}^0_\square, \tag{4.213}$$

and

$$\|\kappa e^s + \frac{1}{\kappa} e^{\star,s}\|^2 = \|e^{s,\kappa}\|^2 = A^s_\square(e^{s,\kappa}, e^{s,\kappa}) = \kappa R_\square(e^{s,\kappa}) + \frac{1}{\kappa} R^\star_\square(e^{s,\kappa}). \tag{4.214}$$

The expressions of $A^s_\square$, $R_\square$ and $R^\star_\square$ in Equations (4.99), (4.100) and (4.122) imply that

$$\begin{cases} \mathfrak{a}_\square(e^{s,\kappa}, \delta u) = -\kappa \mathfrak{m}_\square(\dot{u}_R, \delta u) - \kappa \mathfrak{a}_\square(u_R, \delta u) + \frac{1}{\kappa} \mathfrak{m}_\square(X, \delta u) + \frac{1}{\kappa} \mathfrak{a}_\square(Y, \delta u) \\ \qquad\qquad + \frac{1}{\kappa} \mathfrak{m}_\square(\delta u, \dot{u}^\star_R) - \frac{1}{\kappa} \mathfrak{a}_\square(\delta u, u^\star_R), \quad \forall \delta u \in \mathbb{U}^0_\square, \quad \forall t \in (0,T), & \text{(4.215a)} \\ \frac{1}{2} \mathfrak{m}_\square(e^{s,\kappa}, \delta u)\big|_{t=0} = \kappa \mathfrak{m}_\square(u_0 - u_R, \delta u)\big|_{t=0}, \quad \forall \delta u \in \mathbb{U}^0_\square, & \text{(4.215b)} \\ \frac{1}{2} \mathfrak{m}_\square(e^{s,\kappa}, \delta u)\big|_{t=T} = \frac{1}{\kappa} \mathfrak{m}_\square(Z - u^\star_R, \delta u)\big|_{t=T}, \quad \forall \delta u \in \mathbb{U}^0_\square. & \text{(4.215c)} \end{cases}$$

By using the definition of r_M, see Section 4.5.1, and the linearity and symmetry of $\mathfrak{m}_\square$ and $\mathfrak{a}_\square$ give that Equation (4.215a) can be written as

$$\begin{aligned} \mathfrak{a}_\square(e^{s,\kappa}, \delta u) &= \kappa \mathfrak{m}_\square(r_M, \delta u) + \frac{1}{\kappa}\big(\mathfrak{m}_\square(X + \dot{u}^\star_R, \delta u) + \mathfrak{a}_\square(Y - u^\star_R, \delta u)\big) \\ &= \mathfrak{m}_\square(\kappa r_M, \delta u) + \mathfrak{m}_\square\big(\frac{1}{\kappa}[X + \dot{u}^\star_R], \delta u\big) + \mathfrak{a}_\square\big(\frac{1}{\kappa}[Y - u^\star_R], \delta u\big), \quad \forall \delta u \in \mathbb{U}^0_\square, \quad \forall t \in (0,T). \end{aligned} \tag{4.216}$$

The identity $\mathfrak{m}_\square(\cdot, \Pi_R \delta u) = \mathfrak{a}_\square(\cdot, \Pi_R \delta u) = 0$ gives that

$$\begin{aligned} \mathfrak{a}_\square(e^{s,\kappa}, \delta u) &= \mathfrak{m}_\square(\kappa r_M, \delta u) + \mathfrak{m}_\square\big(\frac{1}{\kappa}[X + \dot{u}^\star_R], \delta u\big) + \mathfrak{a}_\square\big(\frac{1}{\kappa}[Y - u^\star_R], \delta u\big) \\ &= \mathfrak{m}_\square(\kappa r_M, \delta u - \Pi_R \delta u) + \mathfrak{m}_\square\big(\frac{1}{\kappa}[X + \dot{u}^\star_R], \delta u - \Pi_R \delta u\big) + \mathfrak{a}_\square\big(\frac{1}{\kappa}[Y - u^\star_R], \delta u - \Pi_R \delta u\big) \\ &= \mathfrak{m}_\square(\kappa[r_M - \Pi_R r_M], \delta u) + \mathfrak{m}_\square\big(\frac{1}{\kappa}\big[X + \dot{u}^\star_R - \Pi_R X - \Pi_R \dot{u}^\star_R\big], \delta u\big) \\ &+ \mathfrak{a}_\square(\frac{1}{\kappa}\big[Y - u^\star_R - \Pi_R Y + \Pi_R u^\star_R\big], \delta u) \\ &= \mathfrak{m}_\square(\kappa\big[r_M - \Pi_R r_M\big], \delta u) + \mathfrak{m}_\square\big(\frac{1}{\kappa}\big[X - \Pi_R X\big], \delta u\big) + \mathfrak{a}_\square\big(\frac{1}{\kappa}\big[Y - \Pi_R Y\big], \delta u)\big) \\ &= \mathfrak{m}_\square\big(\kappa r'_M + \frac{1}{\kappa} X', \delta u\big) + \mathfrak{a}_\square\big(\frac{1}{\kappa} Y', \delta u\big), \quad \forall \delta u \in \mathbb{U}^0_\square, \quad \forall t \in (0,T). \end{aligned} \tag{4.217}$$

Using the same argumentation on the boundary terms and letting $u'_0 := u_0 - \Pi_R u_0$ and $Z' := Z - \Pi_R Z$ imply that

$$\begin{cases} \mathfrak{a}_\square(e^{s,\kappa}, \delta u) = \mathfrak{m}_\square(\kappa r'_M + \frac{1}{\kappa} X', \delta u) + \mathfrak{a}_\square(\frac{1}{\kappa} Y', \delta u), \quad \forall \delta u \in \mathbb{U}^0_\square, \quad \forall t \in (0,T), & \text{(4.218a)} \\ \frac{1}{2} \mathfrak{m}_\square(e^{s,\kappa}, \delta u)\big|_{t=0} = \kappa \mathfrak{m}_\square(u'_0, \delta u)\big|_{t=0}, \quad \forall \delta u \in \mathbb{U}^0_\square, & \text{(4.218b)} \\ \frac{1}{2} \mathfrak{m}_\square(e^{s,\kappa}, \delta u)\big|_{t=T} = \frac{1}{\kappa} \mathfrak{m}_\square(Z', \delta u)\big|_{t=T}, \quad \forall \delta u \in \mathbb{U}^0_\square. & \text{(4.218c)} \end{cases}$$

In order to estimate $\|e^{s,\kappa}\|$, expressions of $\|e^{s,\kappa}\|_{\mathfrak{a}}$, $\|e^{s,\kappa}\|_{\mathfrak{m}}\big|_{t=0}$ and $\|e^{s,\kappa}\|_{\mathfrak{m}}\big|_{t=T}$ are needed. Equation (4.218a) and the definition of $\|e^{s,\kappa}\|_{\mathfrak{a}}$ gives

$$\begin{aligned}\|e^{s,\kappa}\|_{\mathfrak{a}}^2 &= \mathfrak{a}_{\square}(e^{s,\kappa}, e^{s,\kappa}) = \mathfrak{m}_{\square}(\kappa r_M' + \frac{1}{\kappa}X', e^{s,\kappa}) + \mathfrak{a}_{\square}(\frac{1}{\kappa}Y', e^{s,\kappa}) \\ &\leq \|\kappa r_M' + \frac{1}{\kappa}X', e^{s,\kappa}\|_{\mathfrak{m}}\|e^{s,\kappa}\|_{\mathfrak{m}} + \|\frac{1}{\kappa}Y'\|_{\mathfrak{a}}\|e^{s,\kappa}\|_{\mathfrak{a}}, \quad \forall t \in (0,T).\end{aligned} \tag{4.219}$$

Theorem 6 gives that

$$\|e^{s,\kappa}\|_{\mathfrak{a}}^2 \leq \|\kappa r_M' + \frac{1}{\kappa}X', e^{s,\kappa}\|_{\mathfrak{m}}\frac{\|e^{s,\kappa}\|_{\mathfrak{a}}}{\sqrt{\lambda_{N_R}}} + \|\frac{1}{\kappa}Y'\|_{\mathfrak{a}}\|e^{s,\kappa}\|_{\mathfrak{a}}, \quad \forall t \in (0,T), \tag{4.220}$$

and thus

$$\|e^{s,\kappa}\|_{\mathfrak{a}} \leq \frac{\|\kappa r_M' + \frac{1}{\kappa}X'\|_{\mathfrak{m}}}{\sqrt{\lambda_{N_R}}} + \|\frac{1}{\kappa}Y'\|_{\mathfrak{a}}, \quad \forall t \in (0,T). \tag{4.221}$$

In order to apply the Parallelogram law, an expression of $\|e^{s,\kappa}\|_{\mathfrak{a}}^2$ is needed.

$$\begin{aligned}\|e^{s,\kappa}\|_{\mathfrak{a}}^2 &\leq \Big(\frac{\|\kappa r_M' + \frac{1}{\kappa}X'\|_{\mathfrak{m}}}{\sqrt{\lambda_{N_R}}} + \|\frac{1}{\kappa}Y'\|_{\mathfrak{a}}\Big)^2 \\ &= \frac{\|\kappa r_M' + \frac{1}{\kappa}X'\|_{\mathfrak{m}}^2}{\lambda_{N_R}} + \frac{2}{\kappa\sqrt{\lambda_{N_R}}}\|\kappa r_M' + \frac{1}{\kappa}X'\|_{\mathfrak{m}}\|Y'\| + \frac{1}{\kappa^2}\|Y'\|_{\mathfrak{a}}^2 \\ &= \frac{1}{\lambda_{N_R}}\Big(\kappa^2\|r_M'\|_{\mathfrak{m}}^2 + 2\mathfrak{m}_{\square}(r_M', X') + \frac{1}{\kappa^2}\|X'\|_{\mathfrak{m}}^2\Big) \\ &+ \frac{2}{\kappa\sqrt{\lambda_{N_R}}}\|\kappa r_M' + \frac{1}{\kappa}X'\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}} + \frac{1}{\kappa^2}\|Y'\|_{\mathfrak{a}}^2, \quad \forall t \in (0,T).\end{aligned} \tag{4.222}$$

Now find expressions of the boundary terms. In Section 4.5 it is derived that

$$\|e^s\|_{\mathfrak{m}}\big|_{t=0} \leq \|2u_0 - 2\Pi_R u_0\|_{\mathfrak{m}} = 2\|u_0'\|_{\mathfrak{m}}, \tag{4.223}$$

which implies that

$$\|e^{s,\kappa}\|_{\mathfrak{m}}\big|_{t=0} \leq 2\kappa\|u_0'\|_{\mathfrak{m}}. \tag{4.224}$$

In a similar way it is derived in Section 4.5.2 that

$$\|e^{\star,s}\|_{\mathfrak{m}}\big|_{t=T} \leq \|2Z - 2\Pi_R Z\|_{\mathfrak{m}} = 2\|Z'\|_{\mathfrak{m}}. \tag{4.225}$$

and thus

$$\|e^{s,\kappa}\|_{\mathfrak{m}}\big|_{t=T} \leq \frac{2}{\kappa}\|Z'\|_{\mathfrak{m}}. \tag{4.226}$$

Inserting the derived expressions of $\|e^{s,\kappa}\|_{\mathfrak{a}}$, $\|e^{s,\kappa}\|_{\mathfrak{m}}\big|_{t=0}$ and $\|e^{s,\kappa}\|_{\mathfrak{m}}\big|_{t=T}$ into the definition of $\|e^{s,\kappa}\|$ gives that

$$\begin{aligned}\|e^{s,\kappa}\|^2 \leq \int_0^T &\Big[\frac{1}{\lambda_{N_R}}\Big(\kappa^2\|r_M'\|_{\mathfrak{m}}^2 + 2\mathfrak{m}_{\square}(r_M', X') + \frac{1}{\kappa^2}\|X'\|_{\mathfrak{m}}^2\Big) \\ &+ \frac{2}{\kappa\sqrt{\lambda_{N_R}}}\|\kappa r_M' + \frac{1}{\kappa}X'\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}} + \frac{1}{\kappa^2}\|Y'\|_{\mathfrak{a}}^2\Big]\,\mathrm{d}t + 2\kappa^2\|u_0'\|_{\mathfrak{m}}^2 + \frac{2}{\kappa^2}\|Z'\|_{\mathfrak{m}}^2.\end{aligned} \tag{4.227}$$

Theorem 4 implies that

$$E \leq \frac{1}{4}\|\kappa e^s + \frac{1}{\kappa}e^{\star,s}\|^2. \tag{4.228}$$

Inserting the expression of $\|e^{s,\kappa}\| = \|\kappa e^s + \frac{1}{\kappa}e^{\star,s}\|$ gives that

$$\begin{aligned} E \leq E_{est,pos} = \int_0^T \Big[\frac{1}{4\lambda_{N_R}}\Big(\kappa^2\|r'_M\|_{\mathfrak{m}}^2 + 2\mathfrak{m}_{\square}(r'_M, X') + \frac{1}{\kappa^2}\|X'\|_{\mathfrak{m}}^2\Big) \\ + \frac{1}{2\kappa\sqrt{\lambda_{N_R}}}\|\kappa r'_M + \frac{1}{\kappa}X'\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}} + \frac{1}{4\kappa^2}\|Y'\|_{\mathfrak{a}}^2\Big]\,\mathrm{d}t + \frac{\kappa^2}{2}\|u'_0\|_{\mathfrak{m}}^2 + \frac{1}{2\kappa^2}\|Z'\|_{\mathfrak{m}}^2. \end{aligned} \tag{4.229}$$

Equation (4.229) holds for all $\kappa \neq 0$. The optimal upper estimate of E can be found be setting κ equal to

$$\kappa_{pos} := \min_{\kappa\neq 0}\big(E_{est,pos}\big). \tag{4.230}$$

By using the same derivation for $\|\kappa e^s - \frac{1}{\kappa}e^{\star,s}\|$ it follows that

$$\begin{aligned} E \geq E_{est,neg} = -\int_0^T \Big[\frac{1}{4\lambda_{N_R}}\Big(\kappa^2\|r'_M\|_{\mathfrak{m}}^2 - 2\mathfrak{m}_{\square}(r'_M, X') + \frac{1}{\kappa^2}\|X'\|_{\mathfrak{m}}^2\Big) \\ - \frac{1}{2\kappa\sqrt{\lambda_{N_R}}}\|\kappa r'_M + \frac{1}{\kappa}X'\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}} + \frac{1}{4\kappa^2}\|Y'\|_{\mathfrak{a}}^2\Big]\,\mathrm{d}t - \frac{\kappa^2}{2}\|u'_0\|_{\mathfrak{m}}^2 - \frac{1}{2\kappa^2}\|Z'\|_{\mathfrak{m}}^2. \end{aligned} \tag{4.231}$$

Similar as for Equation (4.229), Equation (4.231) holds for all $\kappa \neq 0$. The optimal lower estimate of E can be found be setting κ equal to

$$\kappa_{neg} := \max_{\kappa\neq 0}\big(E_{est,neg}\big). \tag{4.232}$$

Solving Equations (4.230) and (4.232) is never done in the general case in this thesis. Instead explicit results for upper and lower error estimates are derived for each considered quantity of interest defined in Section 4.3. In the upcoming results, it is for simplicity assumed that the initial condition, u_0, is equal to zero.

Average temperature: In this case $Y' = Z' = 0$ which implies that Equation (4.229) is reduced to

$$E \leq \frac{1}{4\lambda_{N_R}}\int_0^T \Big(\kappa^2\|r'_M\|_{\mathfrak{m}}^2 + 2\mathfrak{m}_{\square}(r'_M, X') + \frac{1}{\kappa^2}\|X'\|_{\mathfrak{m}}^2\Big)\,\mathrm{d}t. \tag{4.233}$$

The corresponding lower limit of E, stated in Equation (4.231), is reduced to

$$E \geq -\frac{1}{4\lambda_{N_R}}\int_0^T \kappa^2\|r'_M\|_{\mathfrak{m}}^2 - 2\mathfrak{m}_{\square}(r'_M, X') + \frac{1}{\kappa^2}\|X'\|_{\mathfrak{m}}^2\,\mathrm{d}t. \tag{4.234}$$

Theorem 7. *Assume that $\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t \neq 0$. Then the optimal value of κ is equal to*

$$\kappa = \left(\frac{\int_0^T \|X'\|_{\mathfrak{m}}^2\,\mathrm{d}t}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t}\right)^{1/4} =: \kappa_{X',opt}. \tag{4.235}$$

Proof. The optimal value of κ can be found by minimizing the estimated norm of E as

$$\min_{\kappa\neq 0}\int_0^T \kappa^2\|r'_M\|_{\mathfrak{m}}^2 + \mathfrak{m}_{\square}(r'_M, X') + \frac{1}{\kappa^2}\|X'\|_{\mathfrak{m}}^2\,\mathrm{d}t. \tag{4.236}$$

Differentiate with respect to κ and set the expression equals zero imply that

$$\int_0^T 2\kappa\|r'_M\|_{\mathfrak{m}}^2 - \frac{2}{\kappa^3}\|X'\|_{\mathfrak{m}}^2\,\mathrm{d}t \stackrel{!}{=} 0. \tag{4.237}$$

Solving κ gives that

$$\kappa = \left(\frac{\int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t} \right)^{1/4}. \tag{4.238}$$

This value of κ is a minimum since the second derivative with respect to kappa is positive, i.e.

$$\frac{\partial^2}{\partial \kappa^2} \frac{1}{4} \|\kappa e^s + \frac{1}{\kappa} e^{\star,s}\|^2 \Big|_{\kappa=\kappa_{opt}} = \frac{1}{2} \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t + \frac{3}{2} \frac{\Big(\int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^2}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t} > 0. \tag{4.239}$$

□

Applying the optimal expression of κ implies that

$$\begin{aligned}
&\int_0^T \kappa^2 \|r'_M\|_{\mathfrak{m}}^2 + 2\mathfrak{m}_{\square}(r'_M, X') + \frac{1}{\kappa^2} \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big|_{\kappa=\kappa_{X',opt}} \\
&= \kappa^2 \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big|_{\kappa=\kappa_{opt}} + 2 \int_0^T \mathfrak{m}_{\square}(r'_M, X') \,\mathrm{d}t + \frac{1}{\kappa^2} \int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big|_{\kappa=\kappa_{X',opt}} \\
&= \left(\frac{\int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t} \right)^{1/2} \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t + 2 \int_0^T \mathfrak{m}_{\square}(r'_M, X') \,\mathrm{d}t + \left(\frac{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t}{\int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t} \right)^{1/2} \int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \\
&= \Big(\int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2} + 2 \int_0^T \mathfrak{m}_{\square}(r'_M, X') \,\mathrm{d}t + \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2} \\
&= 2 \int_0^T \mathfrak{m}_{\square}(r'_M, X') \,\mathrm{d}t + 2 \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2}
\end{aligned} \tag{4.240}$$

and similarly

$$\begin{aligned}
&\int_0^T \kappa^2 \|r'_M\|_{\mathfrak{m}}^2 - 2\mathfrak{m}_{\square}(r'_M, X') + \frac{1}{\kappa^2} \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big|_{\kappa=\kappa_{X',opt}} \\
&= -2 \int_0^T \mathfrak{m}_{\square}(r'_M, X') \,\mathrm{d}t + 2 \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2}.
\end{aligned} \tag{4.241}$$

The optimal error estimate of E can then be written as

$$\begin{cases}
E \leq \frac{1}{2\lambda_{N_R}} \int_0^T \mathfrak{m}_{\square}(r'_M, X') \,\mathrm{d}t + \frac{1}{2\lambda_{N_R}} \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2}, \\
E \geq \frac{1}{2\lambda_{N_R}} \int_0^T \mathfrak{m}_{\square}(r'_M, X') \,\mathrm{d}t - \frac{1}{2\lambda_{N_R}} \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t \int_0^T \|X'\|_{\mathfrak{m}}^2 \,\mathrm{d}t \Big)^{1/2}.
\end{cases} \tag{4.242}$$

Average heat flux: In this case $X' = Z' = 0$ which implies that Equation (4.229) is reduced to

$$E \leq \frac{1}{4} \int_0^T \frac{\kappa^2 \|r'_M\|_{\mathfrak{m}}^2}{\lambda_{N_R}} + \frac{2\|r'_M\|_{\mathfrak{m}} \|Y'\|_{\mathfrak{a}}}{\sqrt{\lambda_{N_R}}} + \frac{1}{\kappa^2} \|Y'\|_t^2 \,\mathrm{d}t. \tag{4.243}$$

The corresponding lower limit of E, stated in Equation (4.231), is reduced to

$$E \geq -\frac{1}{4} \int_0^T \frac{\kappa^2 \|r'_M\|_{\mathfrak{m}}^2}{\lambda_{N_R}} - \frac{2\|r'_M\|_{\mathfrak{m}} \|Y'\|_{\mathfrak{a}}}{\sqrt{\lambda_{N_R}}} + \frac{1}{\kappa^2} \|Y'\|_t^2 \,\mathrm{d}t. \tag{4.244}$$

Similar as for Theorem 7, the optimal value of κ is obtained if

$$\kappa = \left(\frac{\lambda_{N_R} \int_0^T \|Y'\|_{\mathfrak{a}}^2 \,\mathrm{d}t}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t} \right)^{1/4} =: \kappa_{Y',opt}. \tag{4.245}$$

Inserting the expression of $\kappa_{Y',opt}$ gives that

$$\begin{aligned}
&\int_0^T \frac{\kappa^2\|r'_M\|_{\mathfrak{m}}^2}{\lambda_{N_R}} + \frac{2\|r'_M\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}}}{\sqrt{\lambda_{N_R}}} + \frac{1}{\kappa^2}\|Y'\|_t^2\,\mathrm{d}t\Big|_{\kappa=\kappa_{Y',opt}} \\
&= \frac{\kappa^2}{\lambda_{N_R}}\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t\Big|_{\kappa=\kappa_{Y',opt}} + \frac{2}{\sqrt{\lambda_{N_R}}}\int_0^T \|r'_M\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}}\,\mathrm{d}t + \frac{1}{\kappa^2}\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t\Big|_{\kappa=\kappa_{Y',opt}} \\
&= \frac{1}{\lambda_{N_R}}\left(\lambda_{N_R}\frac{\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t}\right)^{1/2}\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t \\
&+ \frac{2}{\sqrt{\lambda_{N_R}}}\int_0^T \|r'_M\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}}\,\mathrm{d}t + \left(\frac{\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t}{\lambda_{N_R}\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t}\right)^{1/2}\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t \\
&= \frac{1}{\sqrt{\lambda_{N_R}}}\Big[\Big(\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t\Big)^{1/2} \\
&+ 2\int_0^T \|r'_M\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}}\,\mathrm{d}t + \Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t\Big)^{1/2}\Big] \\
&= \frac{2}{\sqrt{\lambda_{N_R}}}\int_0^T \|r'_M\|_{\mathfrak{a}}\|Y'\|_{\mathfrak{a}}\,\mathrm{d}t + \frac{2}{\sqrt{\lambda_{N_R}}}\Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t\Big)^{1/2}
\end{aligned} \tag{4.246}$$

and similarly

$$\begin{aligned}
&\int_0^T \frac{\kappa^2\|r'_M\|_{\mathfrak{m}}^2}{\lambda_{N_R}} - \frac{2\|r'_M\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}}}{\sqrt{\lambda_{N_R}}} + \frac{1}{\kappa^2}\|Y'\|_t^2\,\mathrm{d}t\Big|_{\kappa=\kappa_{Y',opt}} \\
&= -\frac{2}{\sqrt{\lambda_{N_R}}}\int_0^T \|r'_M\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}}\,\mathrm{d}t + \frac{2}{\sqrt{\lambda_{N_R}}}\Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t\Big)^{1/2}.
\end{aligned} \tag{4.247}$$

The optimal estimate of E can then be written as

$$\begin{cases}
E \le \frac{1}{2\sqrt{\lambda_{N_R}}}\int_0^T \|r'_M\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}}\,\mathrm{d}t + \frac{1}{2\sqrt{\lambda_{N_R}}}\Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t\Big)^{1/2}, \\
E \ge \frac{1}{2\sqrt{\lambda_{N_R}}}\int_0^T \|r'_M\|_{\mathfrak{m}}\|Y'\|_{\mathfrak{a}}\,\mathrm{d}t - \frac{1}{2\sqrt{\lambda_{N_R}}}\Big(\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t\int_0^T \|Y'\|_{\mathfrak{a}}^2\,\mathrm{d}t\Big)^{1/2}.
\end{cases} \tag{4.248}$$

Final temperature: In this case $X' = Y' = 0$ which implies that Equation (4.229) is reduced to

$$E \le \frac{1}{4}\Big(\int_0^T \frac{\kappa^2\|r'_M\|_{\mathfrak{m}}^2}{\lambda_{N_R}}\,\mathrm{d}t + \frac{2}{\kappa^2}\|Z'\|_{\mathfrak{m}}^2\Big). \tag{4.249}$$

The corresponding lower limit of E, stated in Equation (4.231), is reduced to

$$E \ge -\frac{\kappa^2}{4\lambda_{N_R}}\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t - \frac{1}{2\kappa^2}\|Z'\|_{\mathfrak{m}}^2. \tag{4.250}$$

Similar as for Theorem 7, the optimal value of κ is obtained if

$$\kappa = \left(\frac{2\lambda_{N_R}\|Z'\|_{\mathfrak{m}}^2}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t}\right)^{1/4} =: \kappa_{Z,opt}. \tag{4.251}$$

Inserting the expression of $\kappa_{Z',opt}$ gives that

$$\begin{aligned}
&\kappa^2 \int_0^T \frac{\|r'_M\|_{\mathfrak{m}}^2}{\lambda_{N_R}} \,\mathrm{d}t + \frac{2}{\kappa^2}\|u^{\star,\Pi}\|_{\mathfrak{m}}^2\Big|_{\kappa=\kappa_{Z,opt}} \\
&= \left(\frac{2\lambda_{N_R}\|Z'\|_{\mathfrak{m}}^2}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t}\right)^{1/2} \int_0^T \frac{\|r'_M\|_{\mathfrak{m}}^2}{\lambda_{N_R}} \,\mathrm{d}t + 2\left(\frac{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t}{2\lambda_{N_R}\|Z'\|_{\mathfrak{m}}^2}\right)^{1/2} \|Z'\|_{\mathfrak{m}}^2 \\
&= 2\sqrt{2}\|Z'\|_{\mathfrak{m}}^2 \Big(\frac{1}{\lambda_{N_R}} \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t\Big)^{1/2},
\end{aligned} \tag{4.252}$$

and similarly

$$-\kappa^2 \int_0^T \frac{\|r'_M\|_{\mathfrak{m}}^2}{\lambda_{N_R}} \,\mathrm{d}t - \frac{1}{\kappa^2}\|Z'\|_{\mathfrak{m}}^2\Big|_{\kappa=\kappa_{Z,opt}} = -2\sqrt{2}\|Z'\|_{\mathfrak{m}}^2 \Big(\frac{1}{\lambda_{N_R}} \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t\Big)^{1/2}. \tag{4.253}$$

The optimal error estimate of E can then be written as

$$\begin{cases}
E \leq \frac{1}{\sqrt{2}}\|Z'\|_{\mathfrak{m}}^2 \Big(\frac{1}{\lambda_{N_R}} \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t\Big)^{1/2}, \\
E \geq -\frac{1}{\sqrt{2}}\|Z'\|_{\mathfrak{m}}^2 \Big(\frac{1}{\lambda_{N_R}} \int_0^T \|r'_M\|_{\mathfrak{m}}^2 \,\mathrm{d}t\Big)^{1/2}.
\end{cases} \tag{4.254}$$

5 Numerical results

In this chapter, the numerical results are presented that is used to verify and show the applications of the theory. Section 5.1 presents the results from the Spectral Decomposition. Sections 5.2 and 5.3 present the results for two different canonical load cases given by the macro-scale. Sections 5.4 and 5.5 present the numerical results from the error analysis. Finally, Section 5.6 presents the computational advantages with Numerical Model Reduction.

5.1 Spectral Decomposition

The first step in order to find a reduced basis is to choose a proper base-reducing method. As discussed in Section 3.2, both Spectral Decomposition and Proper Orthogonal Decomposition can be used. Due to the linearity of the transient heat equation, the simpler Spectral Decomposition method is chosen and hence the spatial modes can be estimated as the eigenvectors from the equation

$$\lambda_a \mathfrak{m}_\square(u_a, \delta u) + \mathfrak{a}_\square(u_a, \delta u) = 0, \quad \forall \delta u \in \mathbb{U}_\square, \quad a = 1, 2, \ldots, N. \tag{5.1}$$

In the numerical example it is assumed that ρ and c_p are spatially constant and therefore the linear transient heat flow equation that can be written as

$$\rho c_p \dot{u} + \frac{\mathrm{d}}{\mathrm{d}x} K \frac{\mathrm{d}u}{\mathrm{d}x} = \dot{u} + \frac{\mathrm{d}}{\mathrm{d}x} \frac{K}{\rho c_p} \frac{\mathrm{d}u}{\mathrm{d}x} = \left\{ k := \frac{K}{\rho c_p} \right\} = \dot{u} + \frac{\mathrm{d}}{\mathrm{d}x} k \frac{\mathrm{d}u}{\mathrm{d}x} \stackrel{!}{=} 0. \tag{5.2}$$

Throughout the numerical results, the constants ρ, c_p and K are combined into one constant, k, named the thermal heat constant. The numerical calculations are applied to materials with both homogeneous and heterogeneous micro-structure. The results are presented for four different materials, see Table 5.1. For the one-dimensional transient heat flow, this is produced by changing the thermal heat constant, k, along the Representative Volume Element (RVE). Figure 5.1 shows how k varies within the RVE. As it can be seen in the figure, there are one homogeneous material and three heterogeneous materials.

Material 1	Homogeneous material, upper left in Figure 5.1.
Material 2	Heterogeneous spatial divided into two different phases, upper right in Figure 5.1.
Material 3	Heterogeneous spatial divided into two different phases, lower left in Figure 5.1
Material 4	Heterogeneous with spatial random k, lower right in Figure 5.1

Table 5.1: The table shows how the different materials are denoted.

For a specific distribution of the thermal heat constant, the generalized eigenvalue problem is solved. It is verified that the relations

$$\mathfrak{m}_\square(u_a, u_b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \quad \text{and} \quad \mathfrak{a}_\square(u_a, u_b) = \begin{cases} \lambda_a & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \tag{5.3}$$

holds, which indicates that the generalized eigenvalue problem is solved correctly.

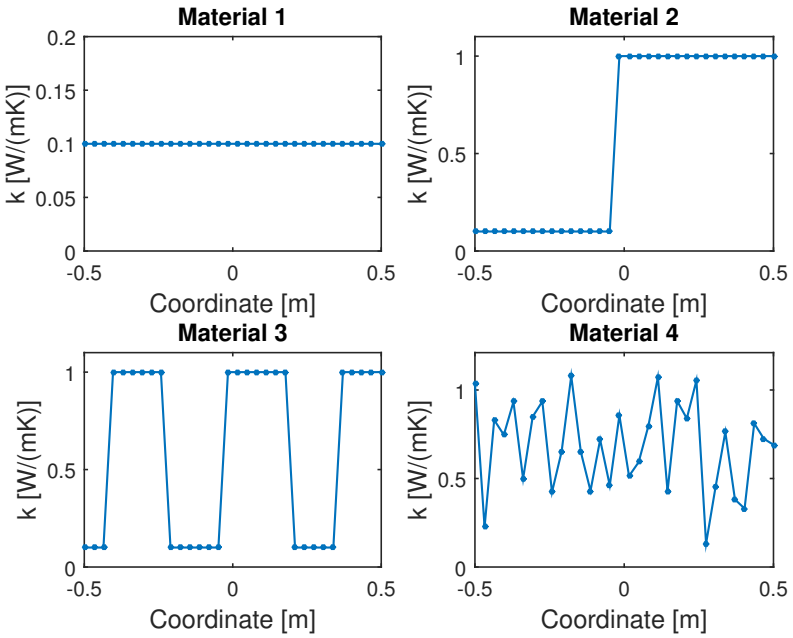


Figure 5.1: *The figure shows the four different cases of how the thermal heat constant, k, varies within the RVE. The upper left sub-figure shows a homogeneous material while the other sub-figures represents different kinds of heterogeneous materials. In the upper right and lower left sub-figures, k is either 0.1 $W/(mK)$ or 1 $W/(mK)$ depending on the spatial location. In the lower right sub-figure k is assigned to be a random number between 0.1 and 1.1 $W/(mK)$.*

For the different materials defined in Figure 5.1, the corresponding eigenvalues can be seen in Figure 5.2. In the figure the eigenvalues are sorted in increasing order and normalized with the smallest eigenvalue. From the figure it can be seen that the amplitude of the first eigenvalues increases rapidly. This indicates that only a few of the eigenvalues are needed to get accurate results since high eigenvalues induces that ξ becomes small fast, see Section 3.2. Another remark concerning the eigenvalues is that they are all real, which they indeed shall be since the stiffness and mass matrices are symmetric matrices.

Figure 5.3 shows the eigenvectors corresponding to the three smallest eigenvalues, for the different materials defined in Figure 5.1. As expected for the homogeneous material, the eigenvectors will be spatially sinusoidal with different frequencies, illustrated in the upper left sub-figure. A higher thermal heat constant implies that the eigenvectors varies slower with spatial location, which can be seen in the upper right and lower left figure. The discontinuities in the thermal heat constant implies that the eigenvectors will have discontinues derivatives. Physically, this implies that there will be a discontinuity in the heat flux in those points.

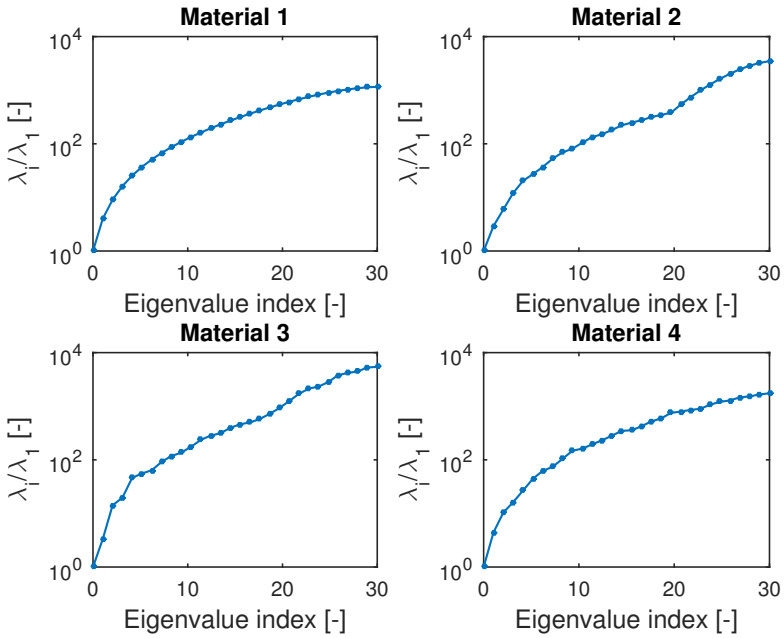


Figure 5.2: *The figure shows the eigenvalues for the different material given in Figure 5.1.*

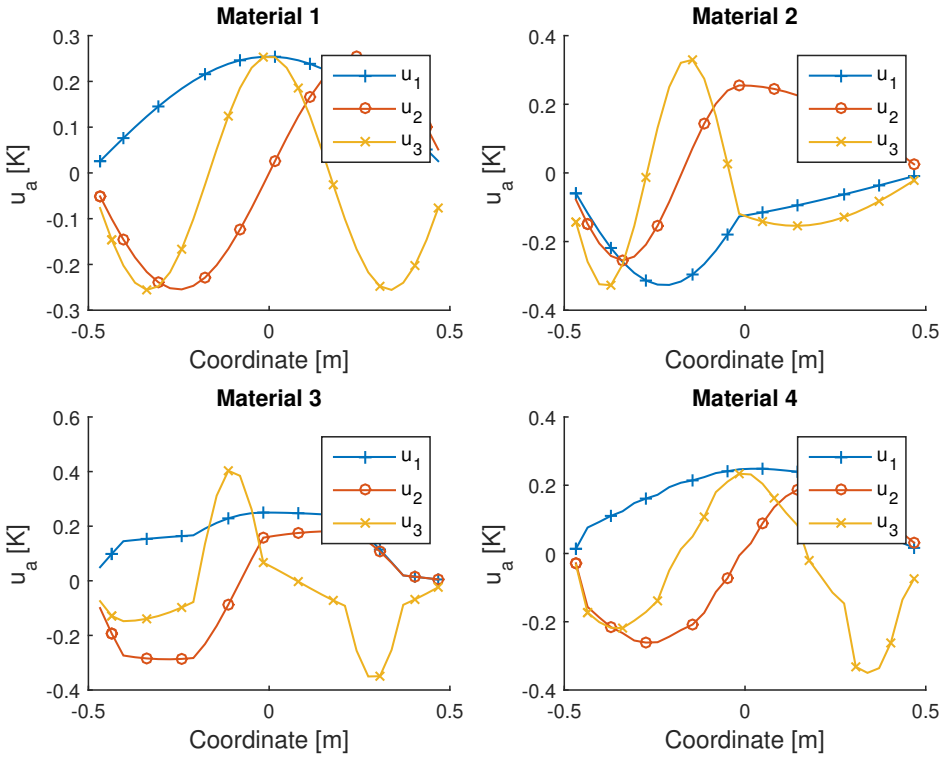


Figure 5.3: *The figure shows the first three eigenvectors for the four different materials given in Figure 5.1. As is can be seen in the figure, the eigenvectors for the homogeneous material will be spatially sinusoidal while the heterogeneous material will have discontinues derivatives in the points where the thermal heat constant is changed.*

5.2 Load case 1 - Ramp load of the macro-scale temperature

In order to evaluate the results, some canonical load cases are needed. For the RVE, this means that the results are calculated for some boundary loadings given by the macro-scale properties $\bar{u}$, $\dot{\bar{u}}$, $\bar{\boldsymbol{g}}$ and $\dot{\bar{\boldsymbol{g}}}$. Note that, the numerical example is just implemented in one spatial dimension which implies that $\bar{\boldsymbol{g}}$ and $\dot{\bar{\boldsymbol{g}}}$ will be scalars, from now on denoted $\bar{g}$ and $\dot{\bar{g}}$. The first load case that is presented consist of the case when $\bar{u}$ is initially a ramp and later in the simulation constant, and $\bar{g}$ equals zero, see Figure 5.4. This implies that $\dot{\bar{g}}$ is equal to zero and $\dot{\bar{u}}$ is during the ramp constant and after the ramp zero.

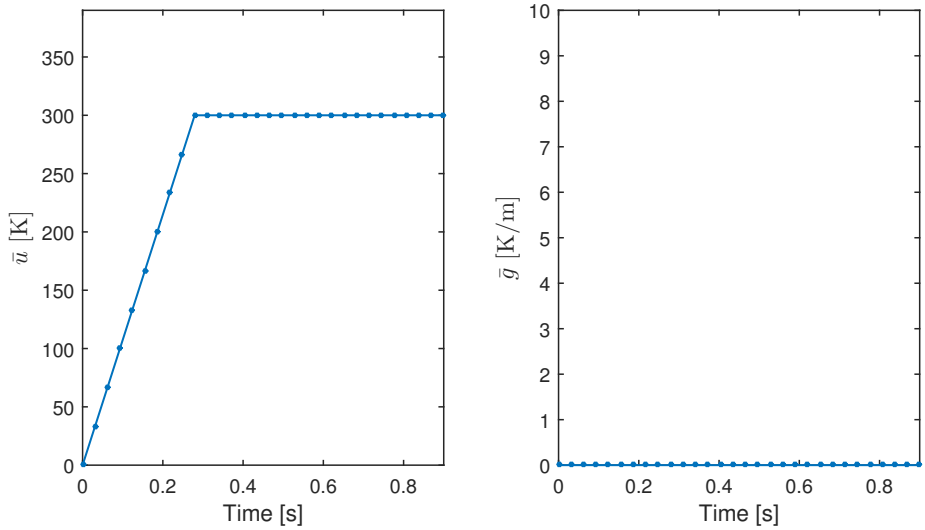


Figure 5.4: *The figure shows the first load case that is applied. As it can be seen in the figure $\bar{u}$ is initially a ramp and later in the simulation constant while $\bar{g}$ equals zero throughout the simulation.*

From the calculated eigenvalues in Section 5.1, the derived ordinary differential equations, see Equation (3.25), can be solved. Denote the mode activity coefficients as $\xi_a(t)$, $a = 1, \ldots, N$, where N is the number of modes. The first four mode activity coefficients $\xi_a(t)$, for respective material, are shown in Figure 5.5. As it can be seen in the figure, every even number of $\xi_a(t)$ is equal to zero for the homogeneous material. The reason is that the boundary condition implies that the solution around the center of the RVE must be symmetric. Recall that in order to get the full micro-scale solution the mode activity coefficient shall be multiplied with the corresponding eigenvector and be summarized, see Section 3.2, according to the equation

$$u^{\mu}(\boldsymbol{x}, t) = \sum_{a=1}^{N} u_a(\boldsymbol{x})\xi_a(t). \tag{5.4}$$

Since every even eigenvector is unsymmetrical, and hence not wanted in the symmetric solution for the homogeneous material, every even ξ_a must be equal to zero. For material 2, 3 and 4, which are heterogeneous, the solution is no longer symmetric around the center of the RVE, which implies that also the even numbers of ξ_a will give a contribution to the solution.

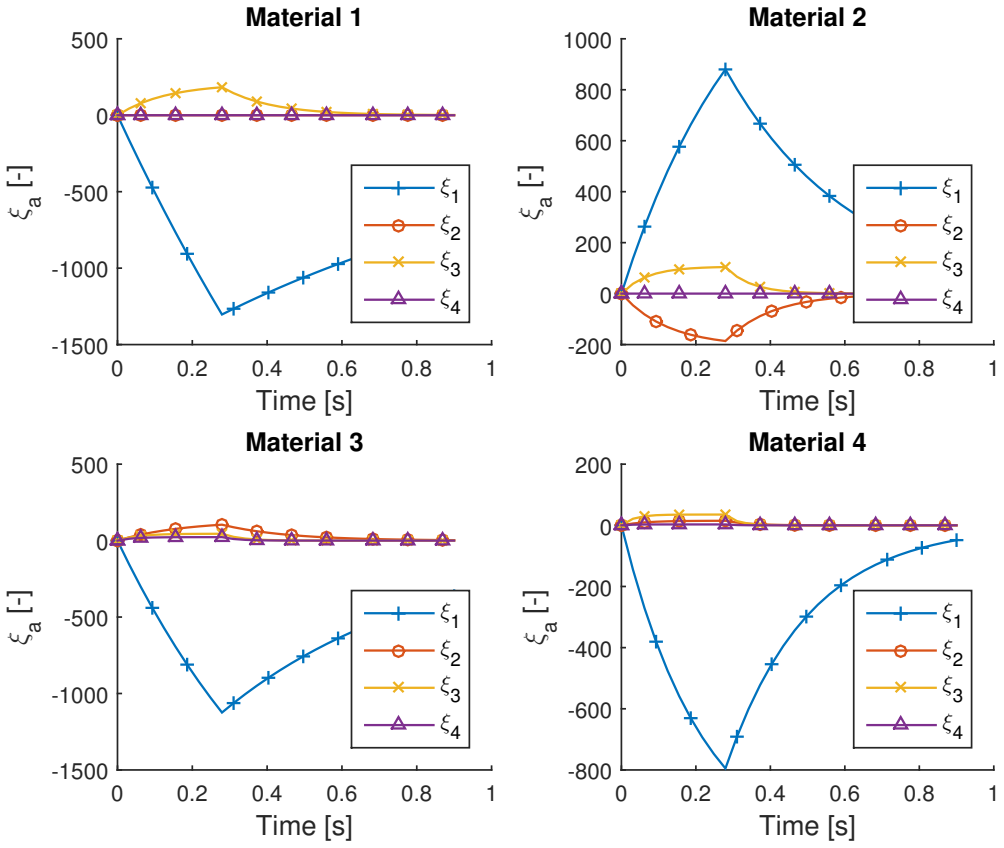


Figure 5.5: *The figure shows the first four mode activity coefficients for each of the considered materials.*

At first sight, it might look strange that ξ_1 is positive for material 2, while it is negative for material 1, 3 and 4. The explanation lies in Equation (5.4) and the fact that $\xi_a(t)$ is multiplied with the corresponding eigenvector. From Figure 5.3 it can be seen that the first eigenvector is positive in material 1, 3 and 4, while it is negative in material 2. This implies that the contribution from the first eigenvector and mode activity coefficient to the micro-solution is always negative.

From the eigenvectors and mode activity coefficients, the full micro-scale solution can be calculated. A typical solution to the RVE, for load case 1, is shown in Figure 5.6. The solutions in Figure 5.6 consist of the first three out of 30 modes. In order visualize the differences between the different materials better, the stationary solution is taken away. The solution for the different materials in this case is shown in Figure 5.7. From Figure 5.7 it can be seen that the transient solution varies a lot between the materials. In the figure is can be seen that the solution is indeed symmetric around the center of the RVE for the homogeneous material. For the second material, the amplitude of the solution is larger for negative coordinates in the RVE. The reason is that the thermal heat constant is smaller there. The shape of the micro-scale solutions are reflected by the eigenvectors for the specific material, see Figure 5.3.

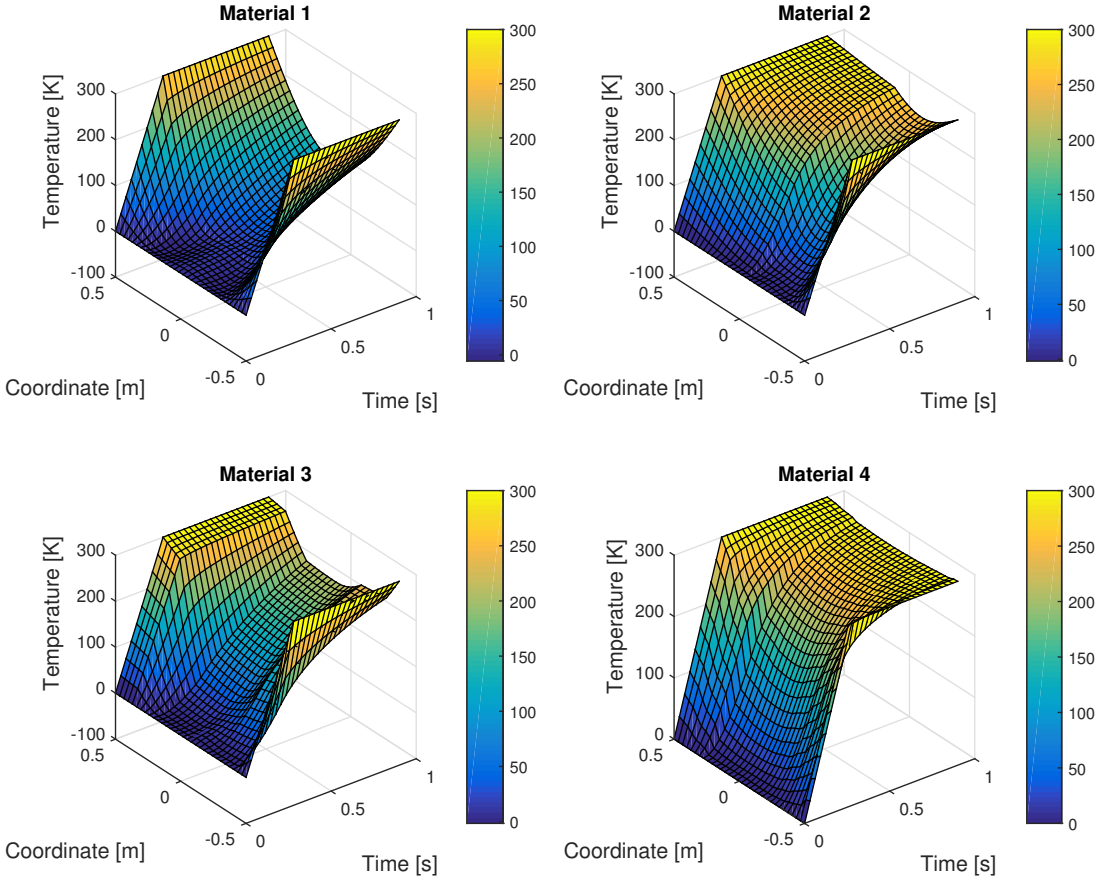


Figure 5.6: *The figure shows how the RVE solution typically looks like for the different materials. The figure is produced with 30 free spatial nodes, 30 time-steps and 3 modes for each material.*

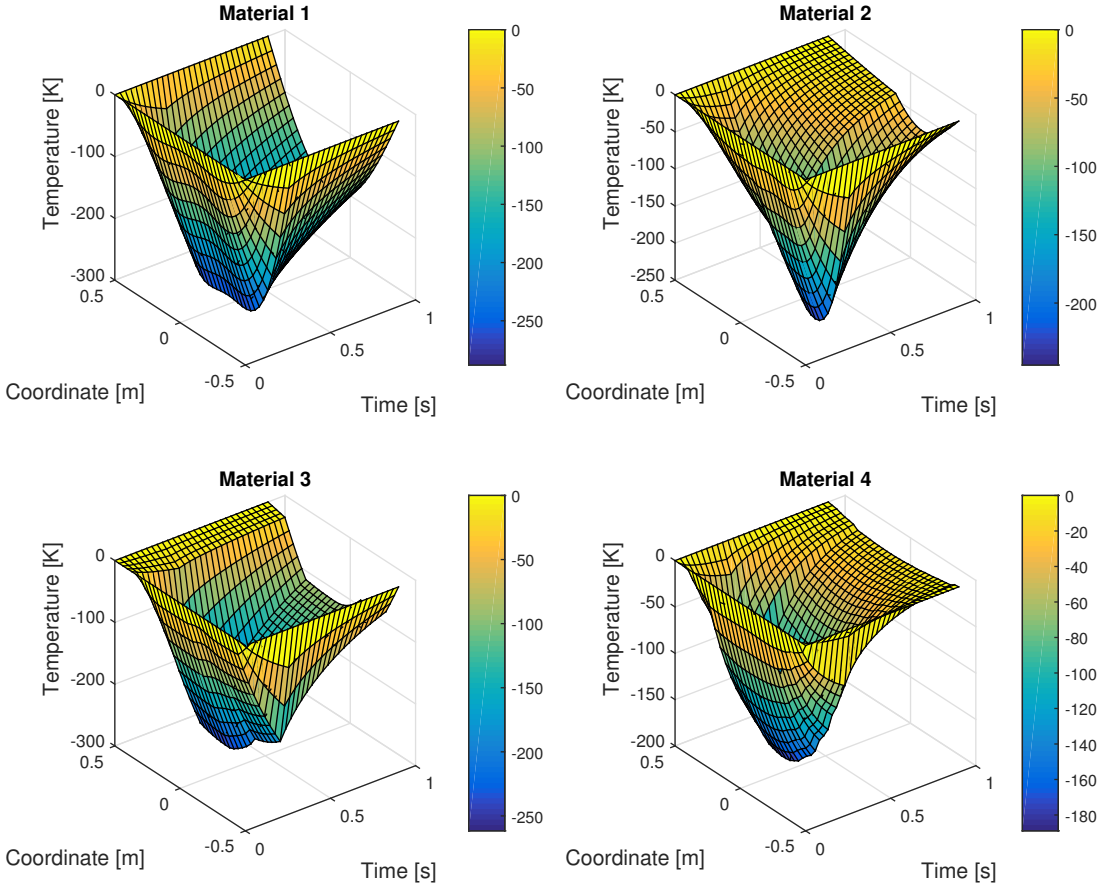


Figure 5.7: *The figure shows the micro-scale solution for the different materials when the stationary part of the solution is taken away. The figure is produced with 30 free spatial nodes, 30 time-steps and 3 modes for each material.*

One of the most interesting results is how accurate the reduced solution, compared to the full base solution, is. Figure 5.8 shows the maximal temperature deviation in one node between the reduced and full solution. From the figure it can be seen that material 1, at most basis, has larger deviations to its full base solution compared to material 4. It can also be seen that every even mode does not reduce the deviation for material 1. The reason is that material 1 is a homogeneous material and therefore the solution is symmetric around the RVE. This error estimate is not a good error estimate since all modes are needed to perform the error estimate. It is also just using information in one node and disregards the information from the rest of the nodes. For better error estimates, see Sections 5.4 and 5.5.

Remark: For the first load case, the stationary solution is captured in the macro-scale part of Equation (3.31). This implies that the results will be identical if the temperature is calculated with Equation (3.31) or (3.34).

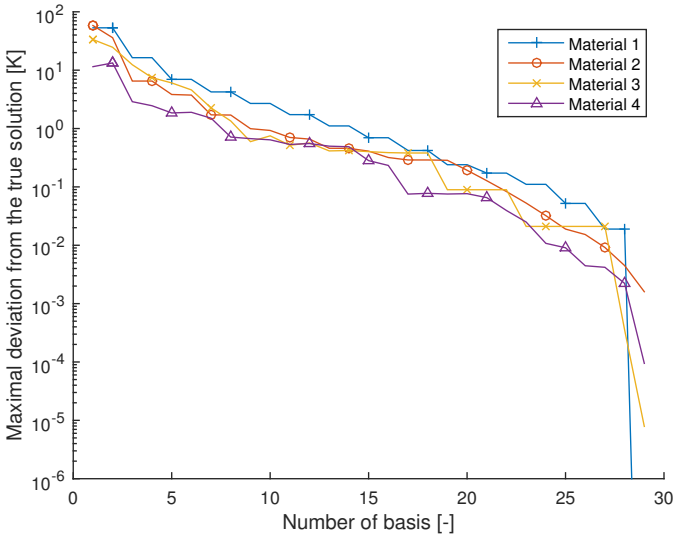


Figure 5.8: *The figure shows the maximal temperature deviation in one node between the reduced and full solution.*

5.3 Load case 2 - Ramp load of the gradient of the macro-scale temperature

In the second canonical load case $\bar{u}$ is zero, while $\bar{g}$ is a ramp. The load case is illustrated in Figure 5.9. By using the same approach as for the first load case, see Section 5.2, $\xi_a(t)$ are calculated, see Figure 5.10. From the figure it can be seen that every odd number of $\xi_a(t)$ is equal to zero for material 1. The reason is that material 1 is a homogeneous material and hence the solution must be purely unsymmetrical. In order to obtain a purely unsymmetrical solution, the odd eigenvectors must not give any contribution, see Figure 5.3. But as it can be seen in Figure 5.3, the eigenvectors are non-zero and thus all odd $\xi_a(t)$ must be equal to zero in order to cancel the effect of these eigenvectors.

Figure 5.11 shows a typical micro-scale solution. In this case, the simulation is performed on the different materials with 3 modes, 30 free spatial nodes and 30 time-steps. From the figure it can be seen that the homogeneous material is indeed purely unsymmetrical around the center on the RVE. In the same manner as for the first load case, it is easier to see the difference between the different materials if the stationary solution is taken away. Figure 5.12 shows only the transient solution for the different materials. Also in this figure, the unsymmetrical properties can be seen in material 1. For material 2 there are, similar as for the first load case, more variation of the temperature in the negative part of the RVE since the thermal heat constant is smaller there. From the figure it can also be seen that the randomness in the thermal heat constant in material 4 gives a non-smooth solution.

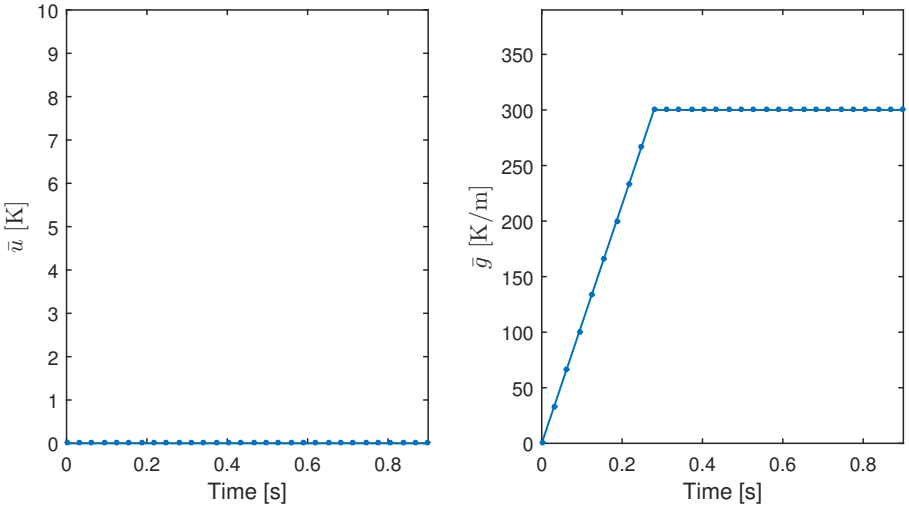


Figure 5.9: *The figure shows the second load case. As it can be seen in the figure $\bar{u}$ is equal to zero and $\bar{g}$ is initially a ramp load and later a constant.*

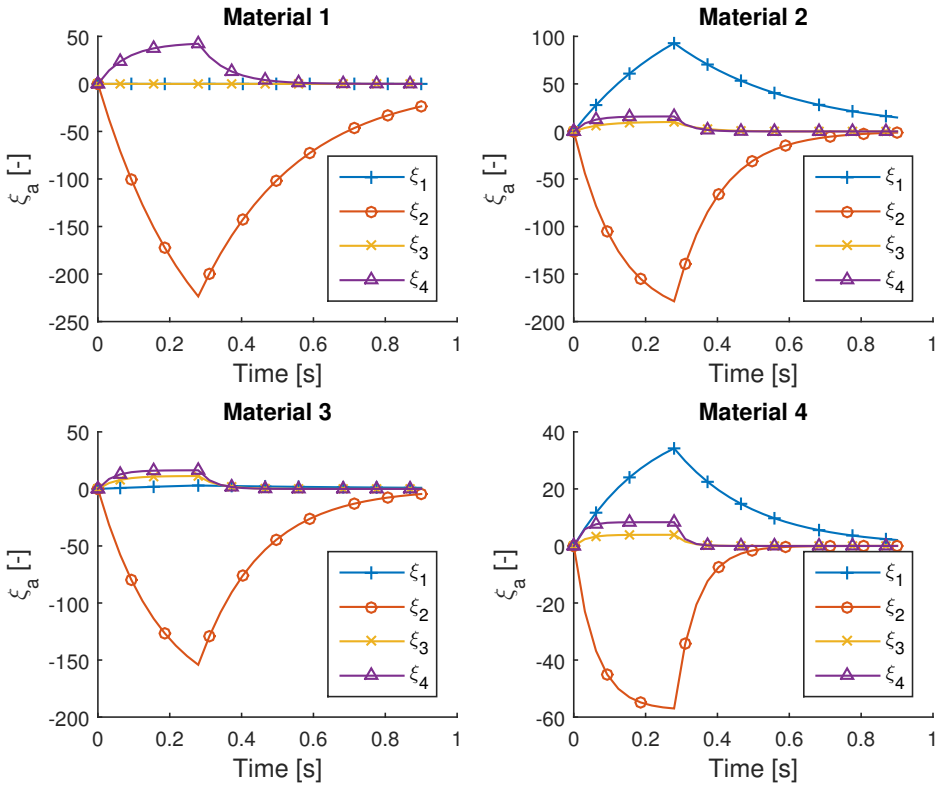


Figure 5.10: *The figure shows the first four mode activity coefficients for the different materials.*

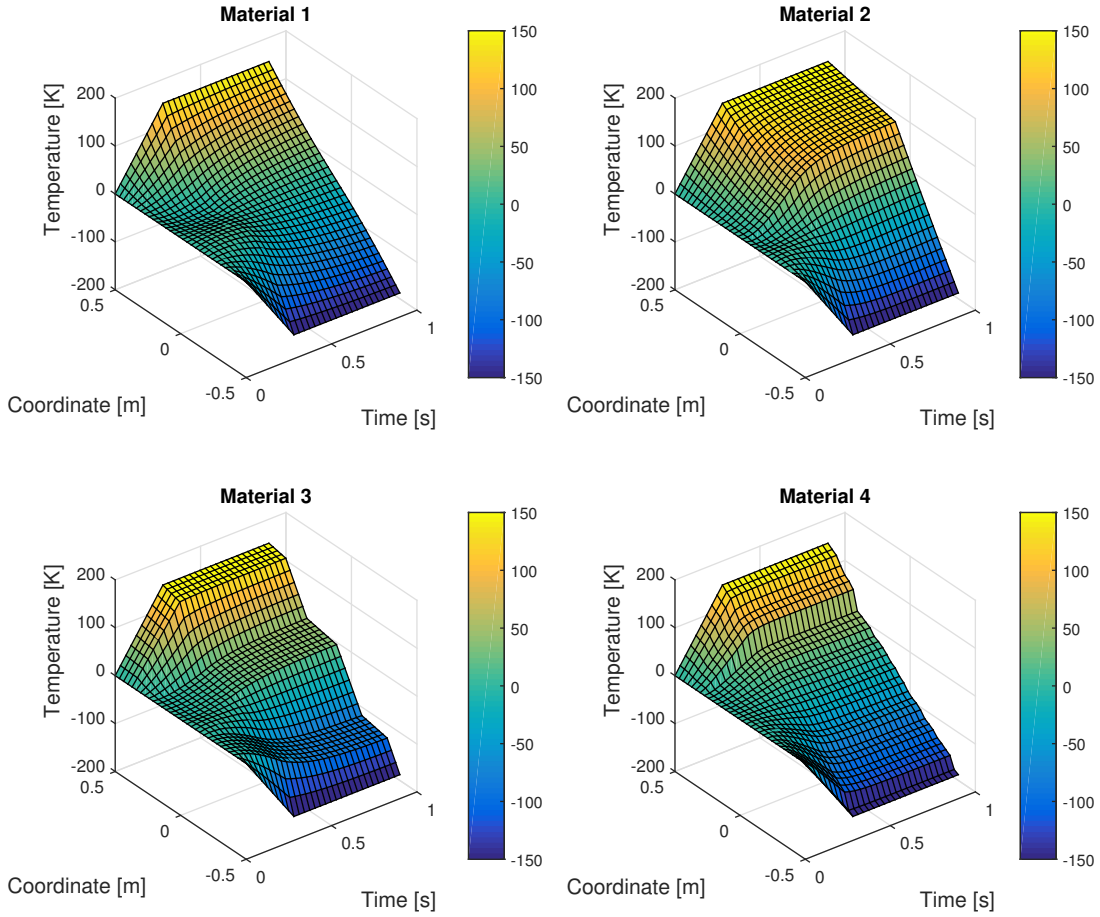


Figure 5.11: *The figure shows how the RVE solution typically looks like for the different materials. The figure is produced with 30 free spatial nodes, 30 time-steps and 3 modes for each material.*

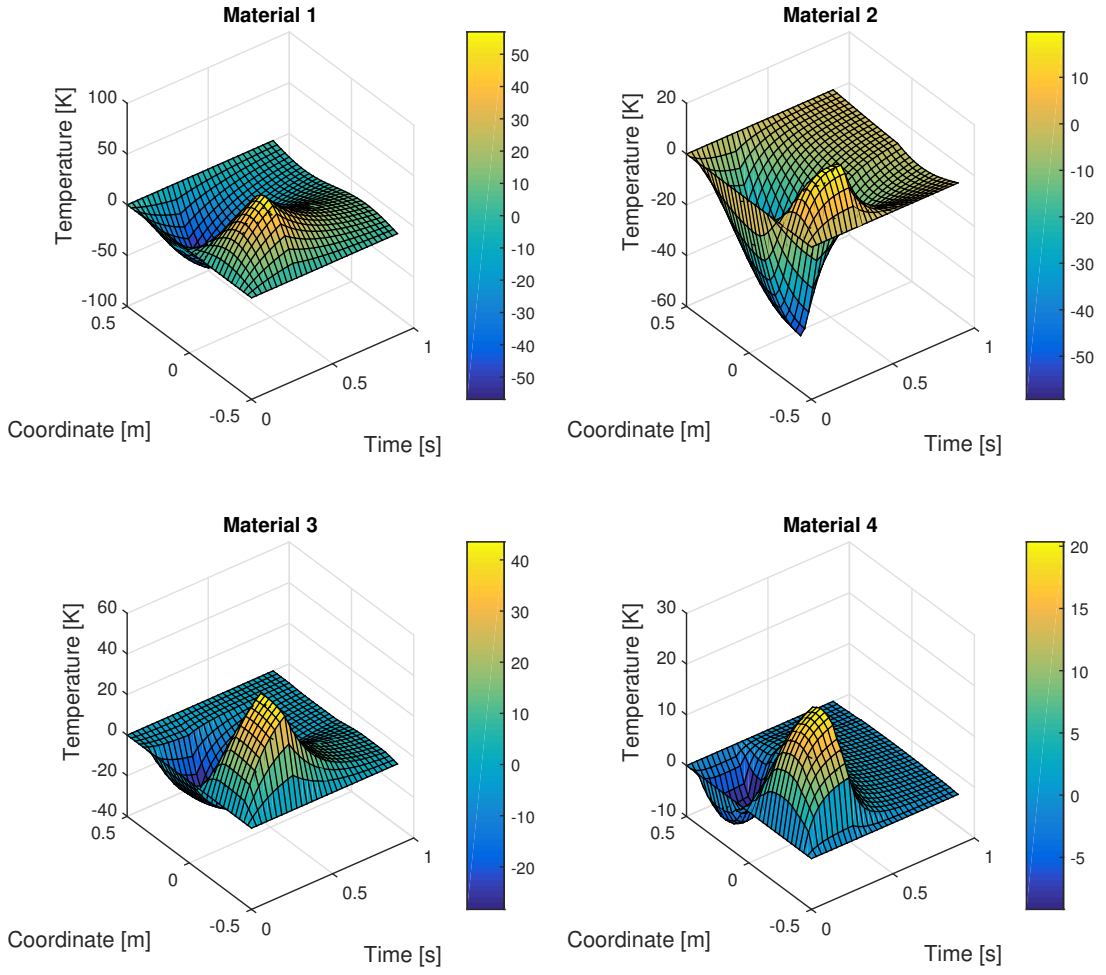


Figure 5.12: *The figure shows the solution of the micro-scale problem when the stationary solution is taken away. The figure is produced with 30 free spatial nodes, 30 time-steps and 3 modes for each material.*

Figure 5.13 shows the maximal temperature deviation in one node between the reduced and full solution. Similar as for the first load case, material 1 tends to have a larger deviation than material 4. For the second load case, the stationary solution is not always captured by the reduced solution for the heterogeneous materials if the temperature is calculated with Equation (3.31). If the temperature instead is calculated with Equation (3.34), the stationary solution will always be captured, independently of the number of modes. Thus, the error will be lower if the temperature is calculated with Equation (3.34), if the material is heterogeneous. The reason that the error estimate is the same for the homogeneous material depends on that the stationary solution is linear and can thus be captured by the macro-scale in Equation (3.31). An illustration of this can be seen in Figure 5.14. From the figure it can be seen that for the heterogeneous materials, the error is significantly lower if the temperature is calculated with Equation (3.34) compared to Equation (3.31).

Remark: Note that the boundary condition of the micro-scale problem can be more arbitrary compared to the load cases. Both $\bar{u}$ and $\bar{g}$ can be pre-multiplied with an arbitrary constant and then added together in order to get more versatile boundary condition on the RVE.

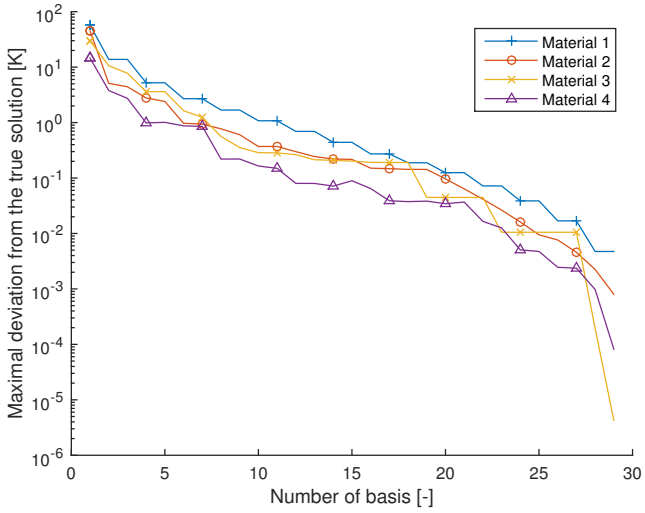


Figure 5.13: *The figure shows the maximal temperature deviation in one node between the reduced and full solution.*

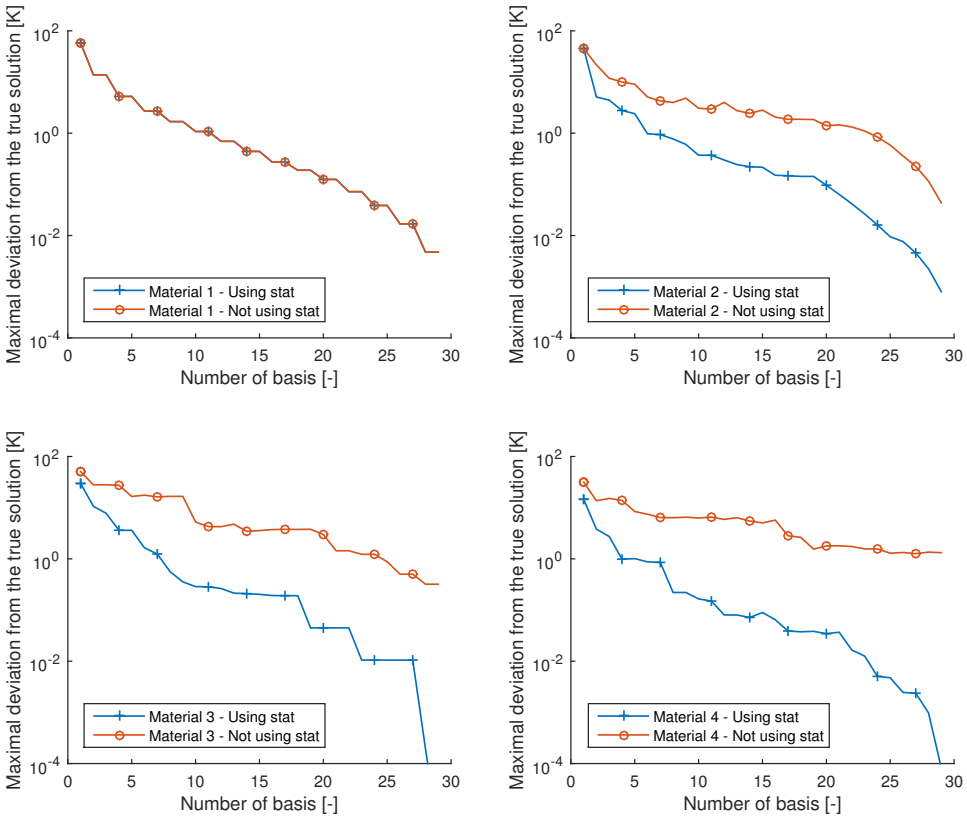


Figure 5.14: *The figure shows the maximal temperature deviation in one node between the reduced and full solution. The figure compare the error when the temperature is calculated with either Equation* (3.31) *or* (3.34).

5.4 Error estimates with the energy norm

One of the main purpose with this thesis is to estimate the error that is introduced by reduce the number of modes. This can be done in different ways, and the probably most simple error estimate is already introduced and presented in Figure 5.14. However, there exists several disadvantage with this simple error estimate. The probably biggest disadvantage is that in order to calculate the error estimate, all modes are needed. Other disadvantages is that knowledge from only one node is used as well as that a more versatile error estimate would be appreciated.

The goal with this section is to produce error estimates when the energy norm is used. In this section, the same 4 materials that are presented in Section 5.1 are used. The numerical results are performed on both of the different load cases. However, the results are similar for both load cases and therefore only the numerical results for load case 1 is presented in this section. The interested reader can find the numerical results for load case 2 in Appendix B.1.

By using the space-time formulation, given in Section 4.1, the energy norm is defined as

$$\|\cdot\| := \sqrt{A_{\square}(\cdot,\cdot)}. \tag{5.5}$$

In order to apply error estimates, Cauchy–Schwarz inequality or the Parallelogram law is suitable to use, but in order to use them the operator needs to be symmetric. Therefore, a symmetric counterpart to $A_{\square}$ is defined as

$$A^s_{\square}(u,v) := \frac{1}{2}\big[A_{\square}(u,v) + A_{\square}(v,u)\big], \quad \forall u,v \in \mathcal{U}_{\square}. \tag{5.6}$$

The true error is denoted e and the error that is calculated from the symmetric error equation is denoted e^s. From Theorem 1 it follows that $\|e\| \leq \|e^s\|$. For more details about the error equations and proof of the theorem, see Section 4.2. Even though $\|e^s\|$ is always larger than $\|e\|$, the difference is very small. In order to show the difference in a proper way, the quantity

$$\eta - 1 := \frac{\|e^s\|}{\|e\|} - 1 \tag{5.7}$$

is shown in Figure 5.15 as a function of the number of used modes for the different materials. From the figure it can be seen that $\|e^s\|$ is always larger than $\|e\|$ as well as that the difference between $\|e\|$ and $\|e^s\|$ becomes smaller and smaller when more modes are used.

In Figure 5.15, the last three modes are not presented. The reason that the last mode is not shown is because when all modes are used $\eta - 1 = 0$, which cannot be represented in a logarithmic figure. The reason that the second and third last mode are not used is from computationally accuracy effects. For specific setups of material, load case, number of modes and time-steps, the energy norm of e and e^s can be extremely small. For some setups, $\|e\|$ and $\|e^s\|$ become so small such that the second and third last modes lies in the same magnitude as the numerical accuracy of the results. This implies that the inequality $\|e\| \leq \|e^s\|$ is no longer guaranteed to hold independently of setup due to numerical errors. With the computational accuracy aspects in mind, the second and third last modes are thus not presented in the upcoming figures.

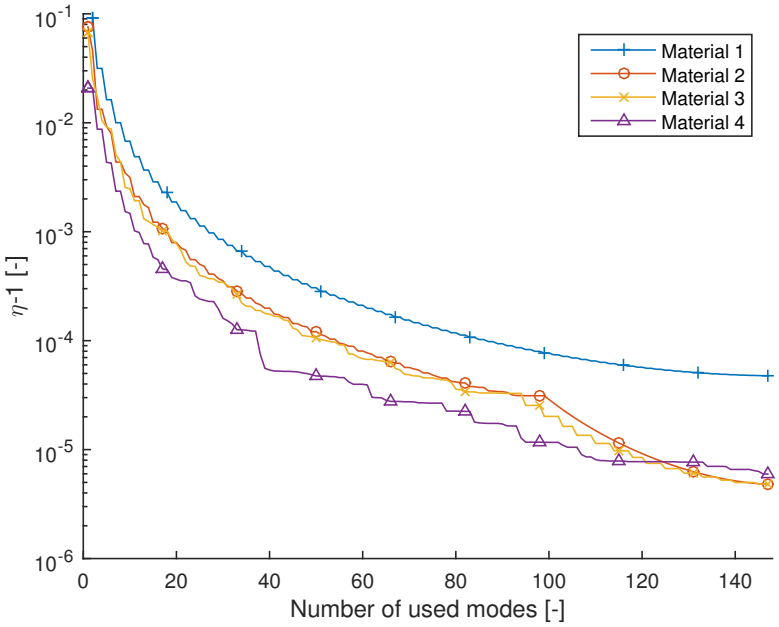


Figure 5.15: *The figure shows the quantity $\eta - 1$ for the energy norm as a function of the number of used modes for the different materials.*

By using the theory in Section 4.5.1, it is derived that the symmetric error can be estimated using only the reduced modes by the equation

$$\|e\| \leq \|e^s\| \leq \|e^s_{est}\| = \Big(\frac{1}{\sqrt{\lambda_{N_R}}} \int_0^T \|r_M\|^2_{\mathfrak{m}} \,\mathrm{d}t + 2\|u_0 - u_R\big|_{t=0}\|^2_{\mathfrak{m}}\Big)^{1/2}, \tag{5.8}$$

where λ_{N_R} is the eigenvalue for the highest reduced mode, u_R is the reduced solution, u_0 is the starting temperature and

$$r_M = -\dot{\bar{u}} - \sum_{i=1}^{d} u^{(i)}_{stat} \dot{\bar{g}}_i. \tag{5.9}$$

For simplicity, $u_0 = u_R\big|_{t=0} = 0$, which implies that

$$\|e^s_{est}\| = \Big(\frac{1}{\lambda_{N_R}} \int_0^T \|r_M\|^2_{\mathfrak{m}} \,\mathrm{d}t\Big)^{1/2}. \tag{5.10}$$

In Section 4.5.1, it is further derived that a better error estimates can be produced if r_M is replaced with

$$r'_M = -r_M + \Pi_R r_M = \Big(1 - \sum_{a=1}^{N_R} \mathfrak{m}_{\square}(1, u_a) u_a(\boldsymbol{x})\Big)\dot{\bar{u}} + \sum_{i=1}^{d} \Big(u^{(i)}_{stat} - \sum_{a=1}^{N_R} \mathfrak{m}_{\square}(u^{(i)}_{stat}, u_a) u_a(\boldsymbol{x})\Big)\dot{\bar{g}}_i. \tag{5.11}$$

This error estimate is denoted

$$\|e^s_{est,proj}\| = \Big(\frac{1}{\lambda_{N_R}} \int_0^T \|r'_M\|^2_{\mathfrak{m}} \,\mathrm{d}t\Big)^{1/2}. \tag{5.12}$$

Figure 5.16 gives a summary of the results when the energy norm is used. The difference between $\|e\|$ and $\|e^s\|$ cannot be seen in the figure because they are too close to each other. But Theorem 1

still hold and $\|e\| \leq \|e^s\|$, see Figure 5.15. Moreover it can be seen that both $\|e^s_{est}\|$ and $\|e^s_{est,proj}\|$ give an upper estimate of the error as expected. Note that the error estimates are normalized with $\|u\|$. This implies that 0.1 in the figure corresponds to 10% error and 0.01 corresponds to 1% error etc. From the figure it can, for example, be seen that when half of the modes are used for material 3, the true error is approximately 0.036% and the estimated error, that only the reduced modes, is 0.076%. Thus it can be guaranteed that the true error is smaller than 0.076% for material 3 when half of the modes are used without using knowledge from any higher modes. If 0.076% is a good enough error estimate depends on the applications. If a better guaranteed error limit is needed, measured in the energy norm, more modes have to be used.

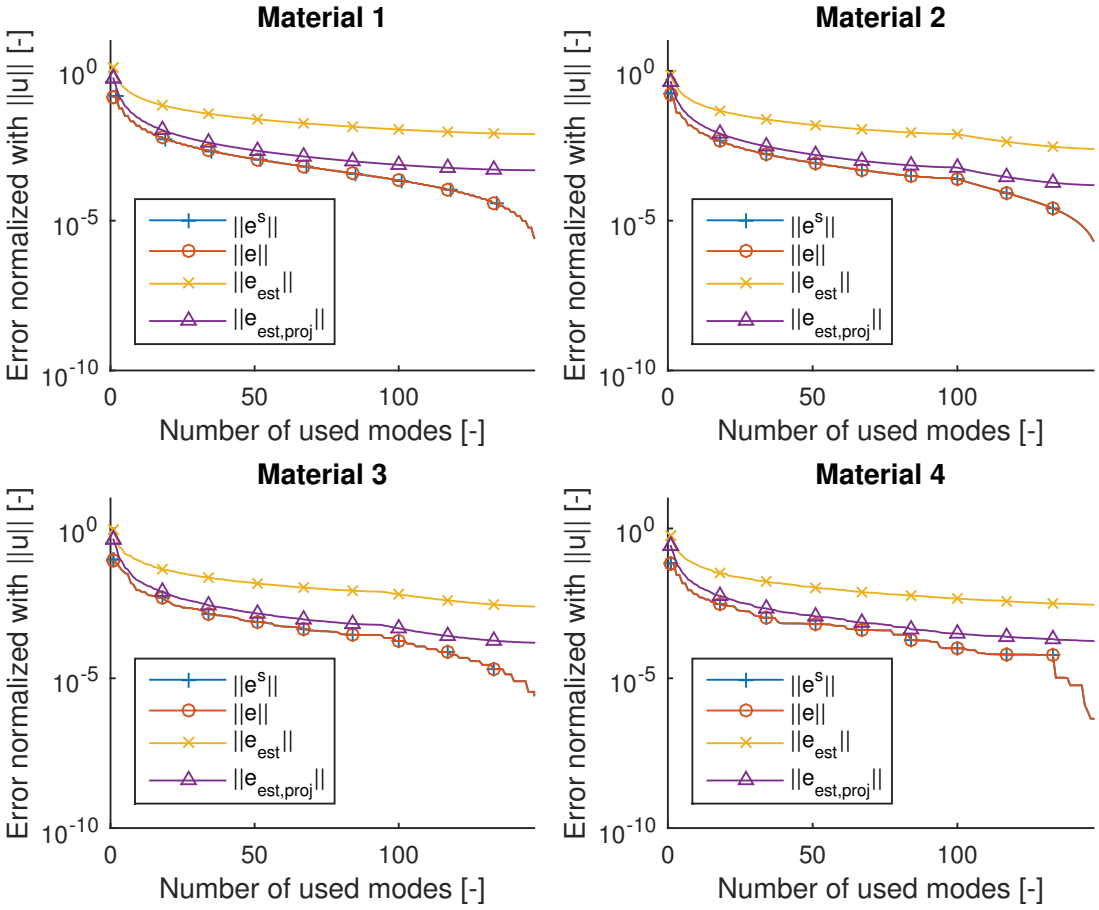


Figure 5.16: *The figure shows the true error and error estimates for the energy norm as a function of the number of used modes for the different materials.*

From Figure 5.16 it can be seen that the ratio between the true and estimated error is not constant with respect to the number of used modes. In order to visualize this phenomena the effectivity index, η, is shown in Figure 5.17, where

$$\eta^s := \frac{\|e^s\|}{\|e\|}, \quad \eta_{est} := \frac{\|e_{est}\|}{\|e\|} \quad \text{and} \quad \eta_{est,proj} := \frac{\|e_{est,proj}\|}{\|e\|}. \tag{5.13}$$

From the figure it can be seen that η^s is always close to 1, which implies that $\|e^s\|$ is always close to $\|e\|$. Since the true error is reduced a lot for high modes, while the estimated errors are almost constant, both η_{est} and $\eta_{est,proj}$ are larger for a large number of modes. This phenomena is probably

introduced during the discretization of the problem. The first two-third of the modes are probably similar to the continuous modes, both for the true and estimated errors. For the true error, the last third of the modes are probably mainly used to fit the discretization. The last third of the modes for the estimated error, on the other hand, just uses estimated modes that will not fit the discretization. Therefore, a larger overestimate of the results happens for the last third of the modes.

If this issue is introduced by the discretization of the problem, it should always be the last third of the modes that produces a larger overestimate of the error. Figure 5.18 and 5.19 shows the same results as Figures 5.16 and 5.17, but with 500 free spatial nodes instead of 150. From the figures it can be seen that the larger overestimate of the error still happens for the last third of the modes as expected.

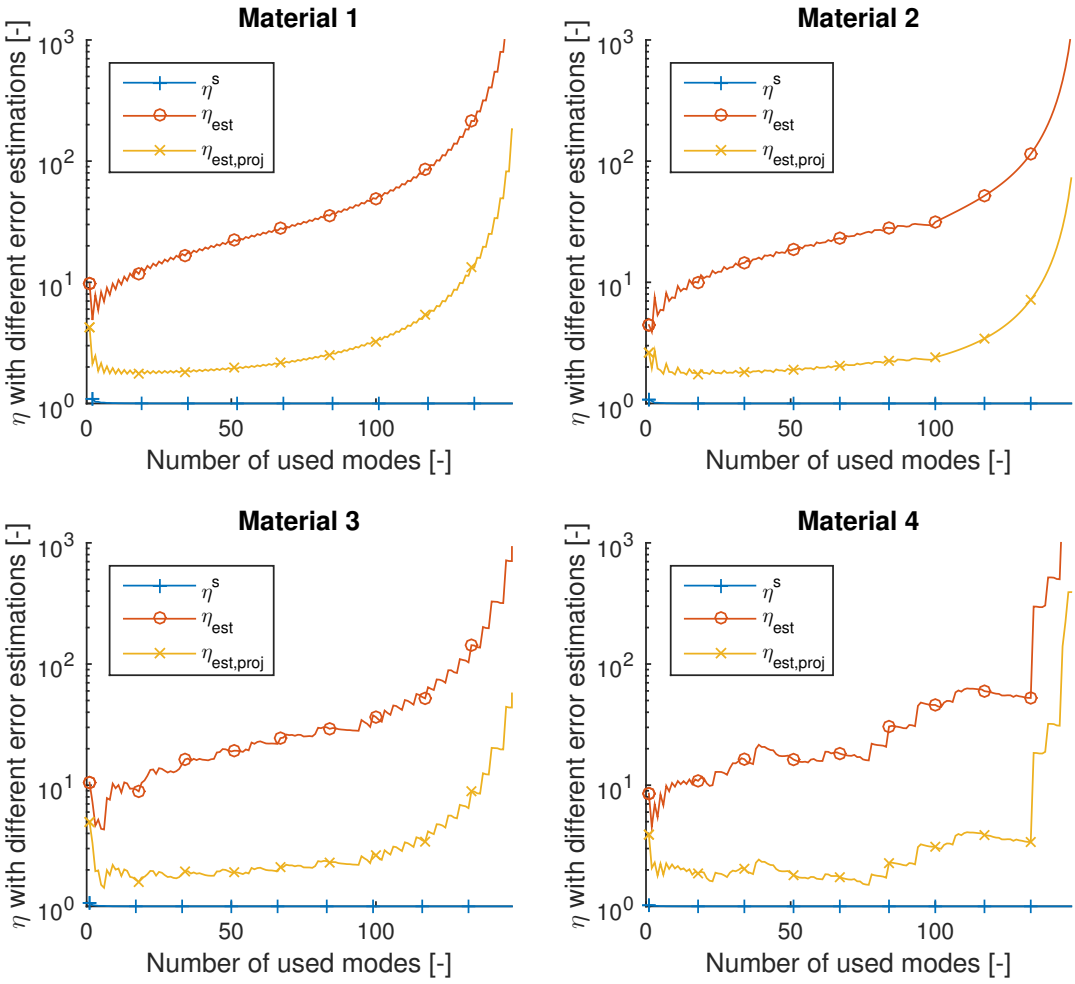


Figure 5.17: *The figure shows the effectivity index of the energy norm as a function of the number of used modes for the different materials.*

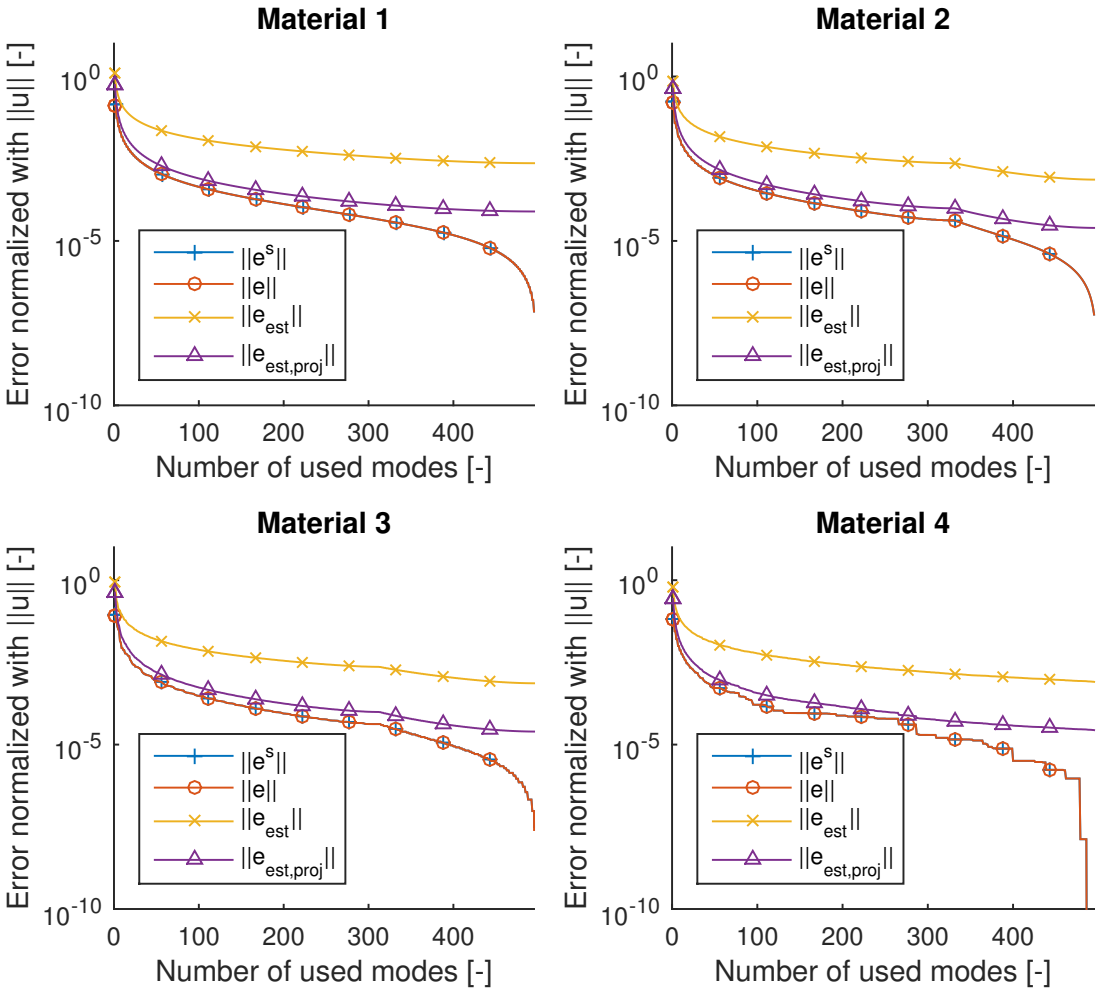


Figure 5.18: *The figure shows the same results as Figure 5.16, but with 500 free spatial nodes instead of 150. If the figure is compared to Figure 5.16, it can be seen that the shape the the true and estimated errors are similar.*

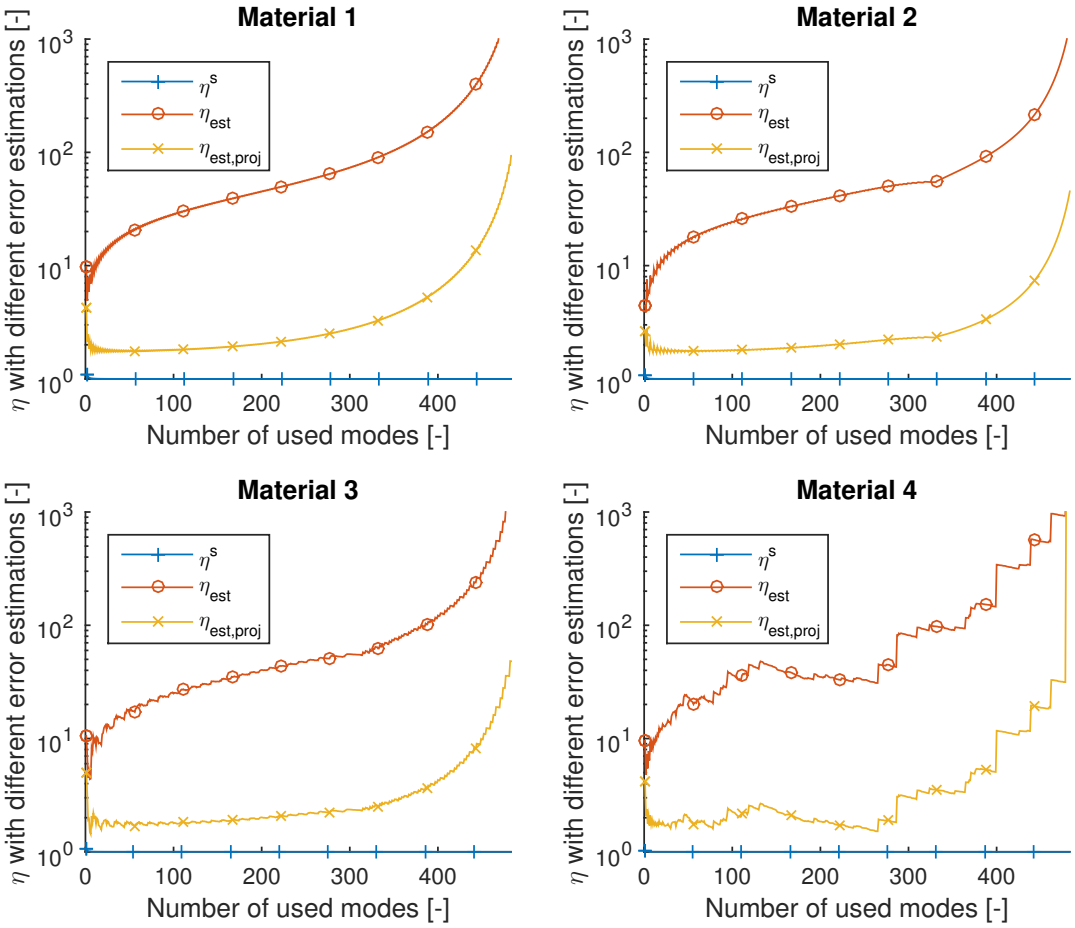


Figure 5.19: *The figure shows the same results as Figure 5.17, but with 500 free spatial nodes instead of 150. From the figure it can be seen that it is still the last third of the modes that produces a large value of* η_{est} *and* $\eta_{est,proj}$.

5.5 Goal-oriented error estimates

By using goal-oriented approaches, it is derived in Section 4.3 that is possible to estimate the error with different quantities of interest. In Section 4.4 it is shown how error estimates are performed when all modes are used and in Section 4.5 it is shown how error estimates are performed when only the reduced modes are used.

The results from these, more sophisticated, error estimates are divided into three different sections corresponding to the three different quantities of interest defined in Section 4.3. Section 5.5.1 presents the average temperature results, Section 5.5.2 presents the average heat flux in one direction results and Section 5.5.3 presents the final temperature results. The error estimates are performed on the same 4 materials that are presented in Section 5.1.

In this thesis, the quantities of interest that is considered can be written as

$$Q_{\square}(u) = \int_0^T \mathfrak{m}_{\square}(X,u) + \mathfrak{a}_{\square}(Y,u)\,\mathrm{d}t + \mathfrak{m}_{\square}(Z,u)\big|_{t=T}, \tag{5.14}$$

where X, Y and Z are arbitrary functions. For a given quantity of interest, the corresponding error is given by $E = Q_{\square}(e)$.

In Sections 4.2.3 and 4.2.4 it is derived how Theorems 2, 3 and 4 can be used to estimate $|E|$. From Theorem 2 it is derived that

$$|E| \leq \|e^s\| \|u^{\star,s}\|, \tag{5.15}$$

where $u^{\star,s}$ is the solution to the dual symmetric problem. This error estimate of E, never uses the Galerkin orthogonality, which induces a large overestimate of the error. However, the Galerkin orthogonality is applied in Theorem 3 which gives the error estimate

$$|E| \leq \|e^s\| \|e^{\star,s}\|. \tag{5.16}$$

The key in Theorems 2 and 3 are to use Cauchy–Schwarz inequality. From the Galerkin orthogonality, it follows that the error estimate from Theorem 3 will always be sharper than Theorem 2. Therefore, Theorem 2 is not implemented in the error estimates that only uses the reduced modes. Theorem 4 uses the Parallelogram law and therefore different error estimates of E is obtained dependent on its sign as

$$\begin{aligned} E &\leq \frac{1}{4}\|\kappa e^s + \frac{1}{\kappa} e^{\star,s}\|^2 \\ E &\geq -\frac{1}{4}\|\kappa e^s - \frac{1}{\kappa} e^{\star,s}\|^2. \end{aligned} \tag{5.17}$$

From Equation (5.17), it can be concluded that

$$|E| \leq \begin{cases} \frac{1}{4}\|\kappa e^s + \frac{1}{\kappa} e^{\star,s}\|^2, & \text{if } E \geq 0 \\ \frac{1}{4}\|\kappa e^s - \frac{1}{\kappa} e^{\star,s}\|^2, & \text{if } E < 0. \end{cases} \tag{5.18}$$

The disadvantage with directly using error estimates performed by Theorems 2, 3 and 4, is that all modes are needed. In order to find error estimates that are using only the reduced modes, $\|e^s\|$ and $\|e^{\star,s}\|$ needs to be estimated. How $\|e^s\|$ can be estimated is presented in Equation (5.12) and in Section 4.5.2 it is shown that and $\|e^{\star,s}\|$ can be estimated as

$$\|e^{\star,s}\| \leq \|e^{\star,s}_{est,proj}\| = \Big(\int_0^T \big(\frac{1}{\sqrt{\lambda_{N_R}}} \|X'\|_{\mathfrak{m}} + \|Y'\|_t \big)^2 \, \mathrm{d}t + 2\|Z'\|^2_{\mathfrak{m}} \Big)^{1/2}, \tag{5.19}$$

where λ_{N_R} is the eigenvalue corresponding to the highest used mode, $X' = X - \Pi_R X$, $Y' = Y - \Pi_R Y$ and $Z' = Z - \Pi_R Z$. By using these error estimates of $\|e^s\|$ and $\|e^{\star,s}\|$, $|E|$ can be estimated with Theorem 3 as

$$|E| \leq \|e^s_{est,proj}\| \|e^{\star,s}_{est,proj}\|. \tag{5.20}$$

In Section 4.5.3 it is derived how Theorem 4 can be applied using only the reduced modes as

$$\begin{cases} E \leq \displaystyle\int_0^T \Big[\frac{1}{4\lambda_{N_R}} \Big(\kappa^2 \|r'_M\|^2_{\mathfrak{m}} + 2\mathfrak{m}_{\square}(r'_M, X') + \frac{1}{\kappa^2}\|X'\|^2_{\mathfrak{m}} \Big) \\ \quad + \dfrac{1}{2\kappa\sqrt{\lambda_{N_R}}} \|\kappa r'_M + \frac{1}{\kappa} X'\|_{\mathfrak{m}} \|Y'\|_t + \frac{1}{\kappa^2}\|Y'\|^2_t \Big] \, \mathrm{d}t + \frac{\kappa^2}{2}\|u'_0\|^2_{\mathfrak{m}} + \frac{1}{2\kappa^2}\|Z'\|^2_{\mathfrak{m}}, \\ E \geq -\displaystyle\int_0^T \Big[\frac{1}{4\lambda_{N_R}} \Big(\kappa^2 \|r'_M\|^2_{\mathfrak{m}} - 2\mathfrak{m}_{\square}(r'_M, X') + \frac{1}{\kappa^2}\|X'\|^2_{\mathfrak{m}} \Big) \\ \quad - \dfrac{1}{2\kappa\sqrt{\lambda_{N_R}}} \|\kappa r'_M + \frac{1}{\kappa} X'\|_{\mathfrak{m}} \|Y'\|_t + \frac{1}{\kappa^2}\|Y'\|^2_t \Big] \, \mathrm{d}t - \frac{\kappa^2}{2}\|u'_0\|^2_{\mathfrak{m}} - \frac{1}{2\kappa^2}\|Z'\|^2_{\mathfrak{m}}, \end{cases} \tag{5.21}$$

where κ is an arbitrary non-zero constant. For optimal error estimates, the value of κ is different for different quantities of interest. For details about expressions of κ and the error estimates it induces, see Sections 5.5.1, 5.5.2 and 5.5.3.

5.5.1 Average temperature as quantity of interest

The first quantity of interest that is presented is the average temperature in the space-time domain, which is given by

$$Q_{\square}(u) = \frac{1}{T|\Omega_{\square}|} \int_0^T \int_{\Omega_{\square}} u \, \mathrm{d}\Omega \, \mathrm{d}t. \tag{5.22}$$

For this quantity of interest $Y = Z = 0$ and $X = \frac{1}{cT}$.

Numerical results are performed for both load cases. In the first load case, where the macro-scale temperature consists of a ramp, it is natural to use the average temperature as quantity of interest. Thus, the results from load case 1 are presented in this section. The results for load case 2 are similar, except for the homogeneous material which is symmetric and therefore has no error in average temperature, see Appendix B.2.

The results of the error estimates for Theorems 2, 3 and 4 are presented in Figure 5.20. Note that Theorem 2 gives a large overestimate since it never uses the Galerkin orthogonality. As it can be seen in the figure, Theorem 4 gives a slightly better error estimate than Theorem 3 since Theorem 4 distinguishes between different signs of E. Note that in order to produce these error estimates, all modes are needed.

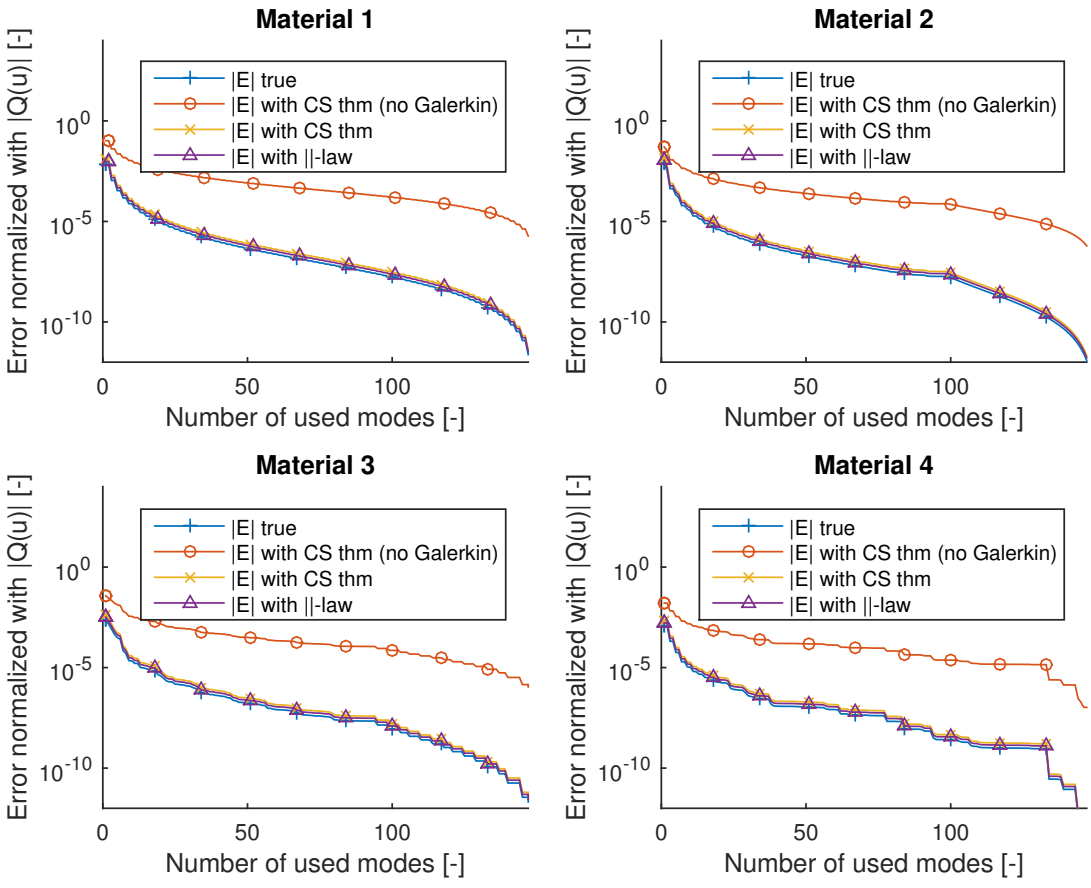


Figure 5.20: *The figure shows the true error and error estimates that are performed on the different materials with Theorems 2, 3 and 4. Note that all errors are normalized with* $|Q(u)|$.

For this specific quantity of interest, Equation (5.19) is reduced to

$$\|e^{\star,s}\| \leq \|e^{\star,s}_{est,proj}\| = \Big(\frac{1}{\lambda_{N_R}} \int_0^T \|X'\|_t^2 \, \mathrm{d}t\Big)^{1/2}. \tag{5.23}$$

With this error estimate, Theorem 3 can be applied which is shown in Figure 5.21. From the figure it can be seen that the ratio between the true and estimated errors grow a lot for the last third of the modes. The reason behind this probably depends on the discretization of the transient heat flow equation and is discussed in Section 5.4.

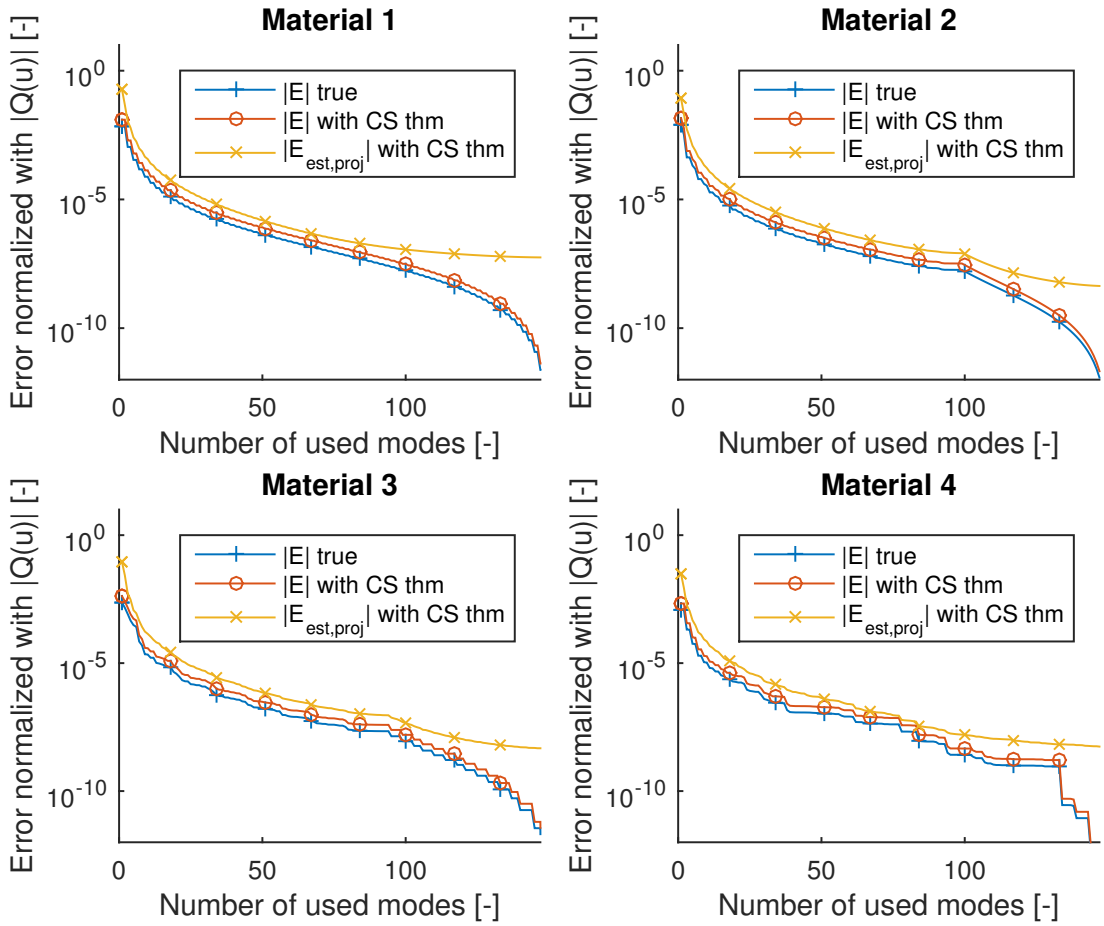


Figure 5.21: *The figure shows the true error and error estimates for the different materials when Theorem 3 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*

Also error estimates for Theorem 4, using only the reduced modes, are performed. When Theorem 4 is used, different error estimates are performed for different values on κ. In Section 4.5.3 it is derived that the optimal κ is obtained if

$$\kappa_{X',opt} = \left(\frac{\int_0^T \|X'\|_{\mathfrak{m}}^2 \, \mathrm{d}t}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \, \mathrm{d}t}\right)^{1/4}. \tag{5.24}$$

With this value of κ, E can be estimated as

$$|E| \leq |E_{est,proj}| = \begin{cases} \frac{1}{2\lambda_{N_R}} \int_0^T \mathfrak{m}_\square(r'_M, X')\,\mathrm{d}t + \frac{1}{2\lambda_{N_R}} \left(\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t \int_0^T \|X'\|_{\mathfrak{m}}^2\,\mathrm{d}t \right)^{1/2} & \text{if } E \geq 0 \\ \frac{1}{2\lambda_{N_R}} \int_0^T \mathfrak{m}_\square(r'_M, X')\,\mathrm{d}t - \frac{1}{2\lambda_{N_R}} \left(\int_0^T \|r'_M\|_{\mathfrak{m}}^2\,\mathrm{d}t \int_0^T \|X'\|_{\mathfrak{m}}^2\,\mathrm{d}t \right)^{1/2} & \text{if } E < 0, \end{cases} \tag{5.25}$$

The results when Theorem 4 is used can be seen in Figure 5.22.

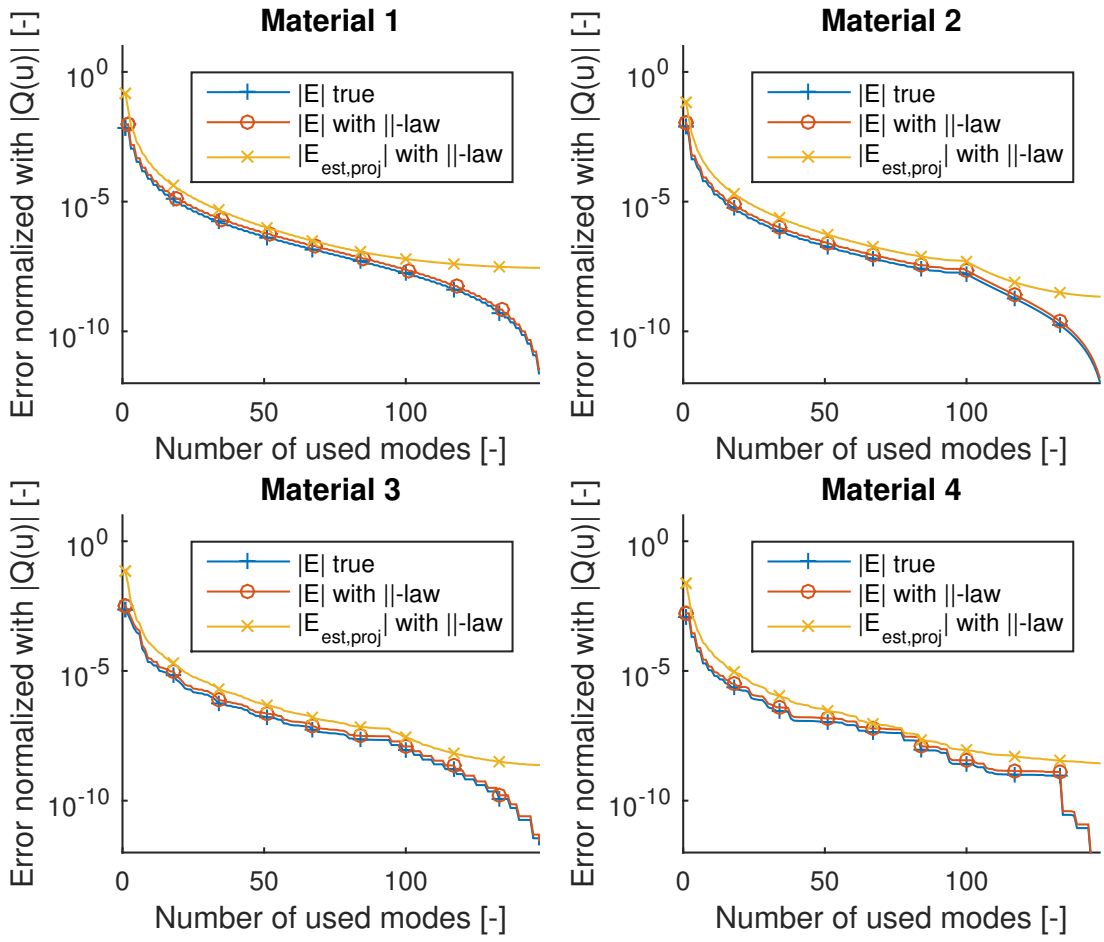


Figure 5.22: *The figure shows the true error and error estimates for the different materials when Theorem 4 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*

When Theorem 4 is applied, different error estimates are performed dependent on the sign of E. Figure 5.23 shows the upper and lower limit of E, performed with Theorem 4. From the figure it can be seen that the lower bound is smaller than the upper bound and thus the error span is reduced faster compared with Theorem 3.

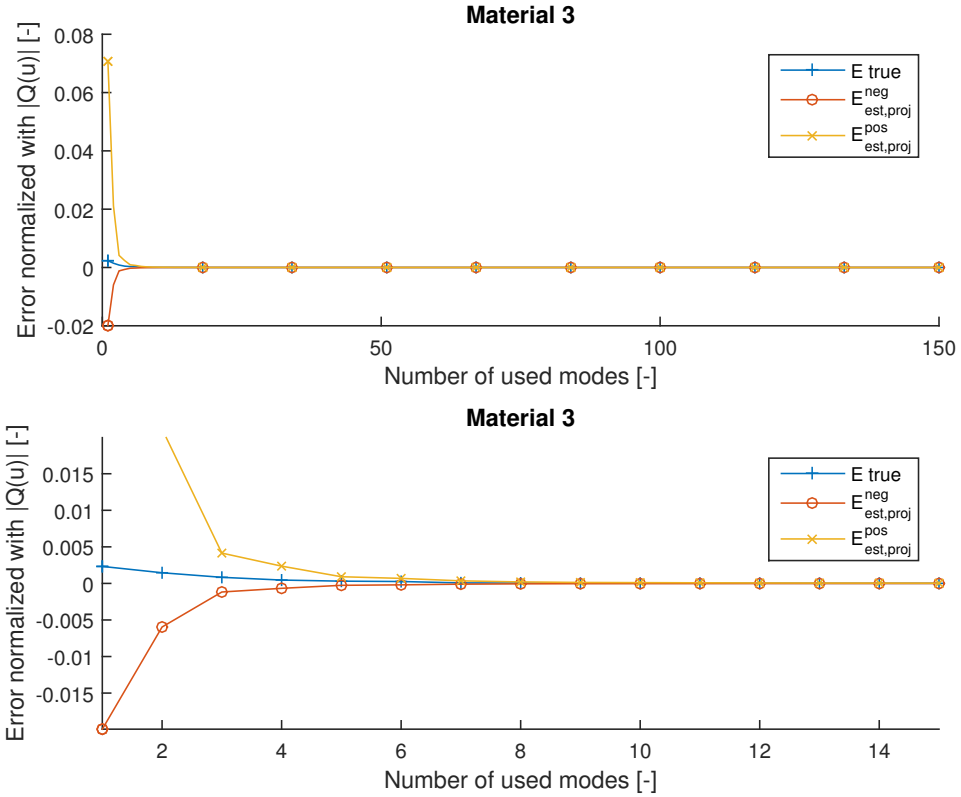


Figure 5.23: *The figure shows the true error and its upper and lower limits when Theorem 4 is applied.*

5.5.2 Average heat flux in one direction as quantity of interest

The second quantity of interest that is presented is the average heat flux in one direction. This quantity of interest is given by

$$Q_{\square}(u) = \frac{1}{T|\Omega_{\square}|} \int_0^T \int_{\Omega_{\square}} \boldsymbol{e} \cdot \boldsymbol{q} \, \mathrm{d}\Omega \, \mathrm{d}t. \tag{5.26}$$

Similar as for the first quantity of interest, numerical results are performed for both load cases. Using the average heat flux in one direction as quantity of interest is suitable for the second load case. Therefore, results for load case 2 are presented in this section. The results for the other load case are similar and the interested reader can find the results in Appendix B.3.

For the homogeneous material, the error is equal to zero since the heat flux is proportional to the temperature gradient. Therefore, the upcoming figures only shows result for material 2, 3 and 4. Figure 5.24 shows the true error and error estimates from Theorems 2, 3 and 4, normalized with $|Q(u)|$, as a function of number of used modes. Similar as for the average temperature, Theorem 4 gives a slightly better estimate than Theorem 3. The error estimate from Theorem 2 is not as good as the other error estimates since it does not use the Galerkin orthogonality. Note that the absolute value of the true error does not always go down if the number of modes is increased. The reason is that the modes are constructed in such a way that the energy norm will be minimized when additional modes are added. This implies that if the true error goes up or down with additional modes depends on what quantity of interest that is used.

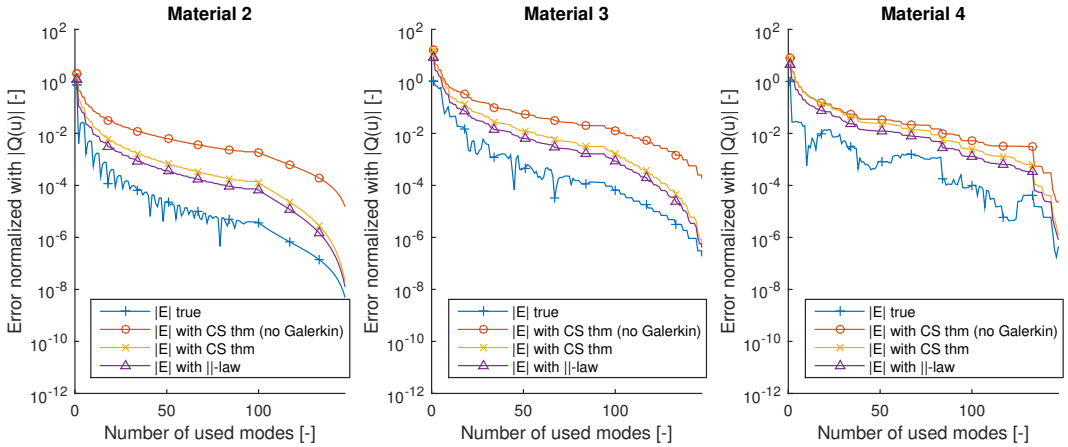


Figure 5.24: *The figure shows the true error and error estimates that are performed on the different materials with Theorems 2, 3 and 4. Note that in order to produce these error estimates all modes are needed.*

Also for this quantity of interest, error estimates using only the reduced modes are performed. For this specific quantity of interest, Equation (5.19) is reduced to

$$\|e^{\star,s}\| \leq \|e^{\star,s}_{est,proj}\| = \Big(\int_0^T \|Y'\|_t^2 \, \mathrm{d}t \Big)^{1/2}. \tag{5.27}$$

By using this error estimate of $\|e^{\star,s}\|$, Theorem 3 can be used for error estimates using only the reduced modes. The results can be seen in Figure 5.25.

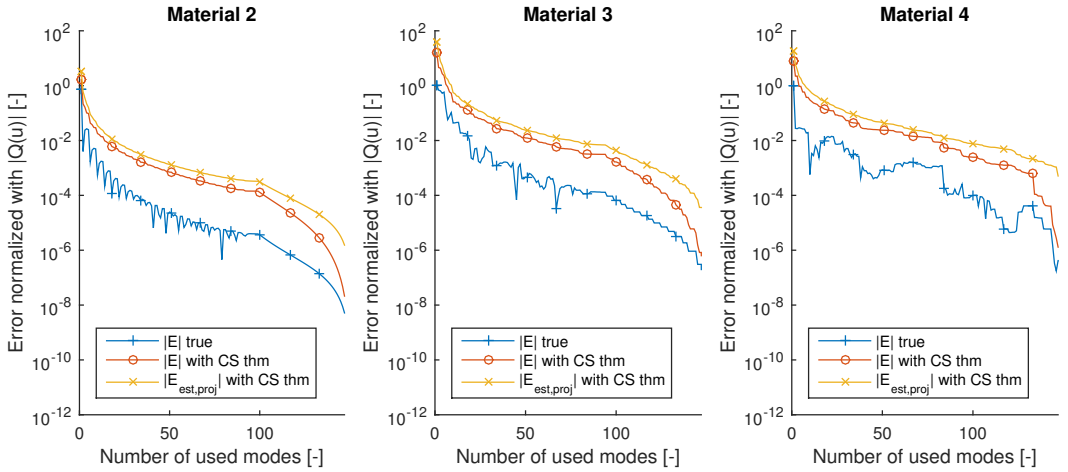


Figure 5.25: *The figure shows the true error and error estimates for the different materials when Theorem 3 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*

As derived in Section 4.5.3, a similar error estimate can be produced with Theorem 4. For this quantity of interest

$$\kappa_{Y',opt} = \left(\frac{\lambda_{N_R} \int_0^T \|Y'\|_t^2 \, \mathrm{d}t}{\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \, \mathrm{d}t} \right)^{1/4}. \tag{5.28}$$

With this value of κ, E can be estimated via the equations

$$\begin{cases} E \leq \frac{1}{2\sqrt{\lambda_{N_R}}} \int_0^T \|r'_M\|_{\mathfrak{m}} \|Y'\|_t \, \mathrm{d}t + \frac{1}{2\sqrt{\lambda_{N_R}}} \left(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \, \mathrm{d}t \int_0^T \|Y'\|_{\mathfrak{m}}^2 \, \mathrm{d}t \right)^{1/2}, \\ E \geq \frac{1}{2\sqrt{\lambda_{N_R}}} \int_0^T \|r'_M\|_{\mathfrak{m}} \|Y'\|_t \, \mathrm{d}t - \frac{1}{2\sqrt{\lambda_{N_R}}} \left(\int_0^T \|r'_M\|_{\mathfrak{m}}^2 \, \mathrm{d}t \int_0^T \|Y'\|_{\mathfrak{m}}^2 \, \mathrm{d}t \right)^{1/2}. \end{cases} \tag{5.29}$$

The results can be seen in Figure 5.26. In order to see how big the errors are, all error estimates are normalized with $|Q(u)|$. From the figure it can be determined that when half of the modes are used, the error is 0.026% from the true solution for material 3. The corresponding number when only the reduced modes are used for error estimate is 0.48%. In a real simulation, only the 0.48% error estimate is available since the other error estimates uses all modes. If maximal 0.48% error in average heat flux in a desired direction is good enough depends on the application the theory is applied to.

Finally, Figure 5.27 shows the upper and lower limit of E, performed with Theorem 4. Similar as for the first quantity of interest, the fact the upper and lower bounds are different implies that the error span is smaller compared with the results obtained with Theorem 3.

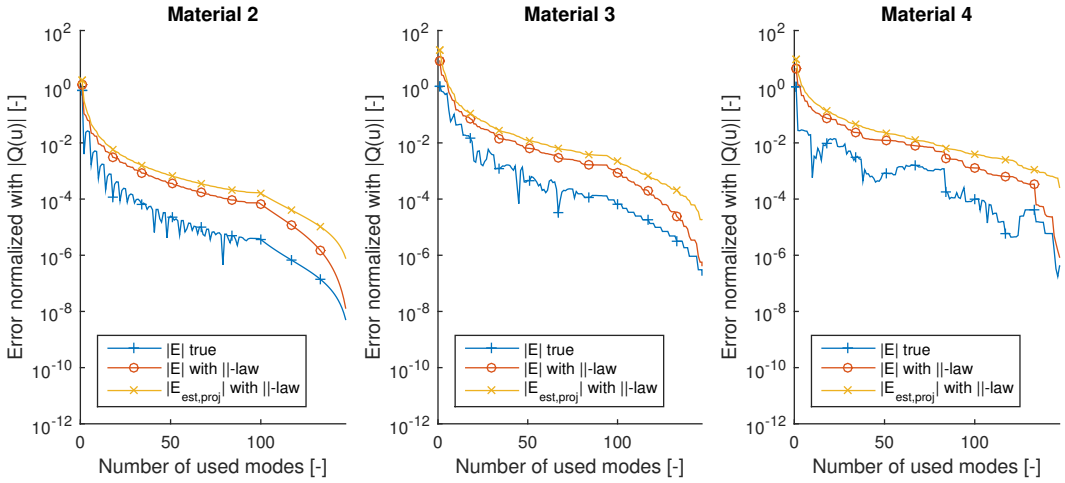


Figure 5.26: *The figure shows the true error and error estimates for the different materials when Theorem 4 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*

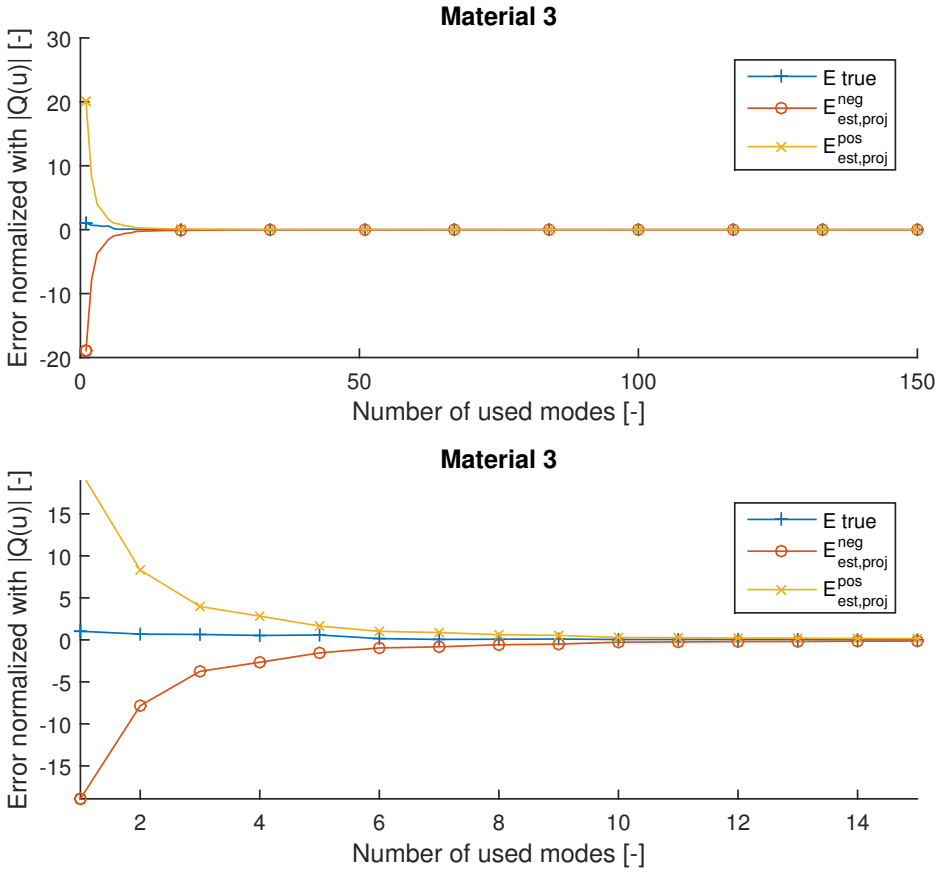


Figure 5.27: *The figure shows the true error and its upper and lower estimated limits when Theorem 4 is applied.*

5.5.3 Final temperature as quantity of interest

The last quantity of interest that is presented is the final temperature. In this case the quantity of interest is given by

$$Q_{\square}(u) = \frac{1}{|\Omega_{\square}|} \int_{\Omega_{\square}} u\big|_{t=T} \, \mathrm{d}\Omega. \tag{5.30}$$

The results for the different load cases are similar for this quantity of interest. The results for the load case that involves $\bar{u}$ are presented in this section, while the results for the load case that involves $\bar{g}$ are presented in Appendix B.4.

The true error and error estimates from Theorems 2, 3 and 4, when all modes are used, are shown in Figure 5.28. From the figure it can be seen that the true error becomes extremely small. The reason that the true error becomes so small depends on the quantity of interest. Since the stationary solution is captured by the macro-scale, only the transient solution will contribute to the error. But for load case 1, $\dot{\bar{u}}$ equals zero the last two-thirds of the simulated time, see Figure 5.4. This implies that the transient solution is almost 0 when $t = T$, and thus also the error. The reason that the estimated error does not follow the true error probably has to do with the symmetrization procedure introduced in order to use Cauchy–Schwarz inequality and the Parallelogram law, see Section 4.2.2.

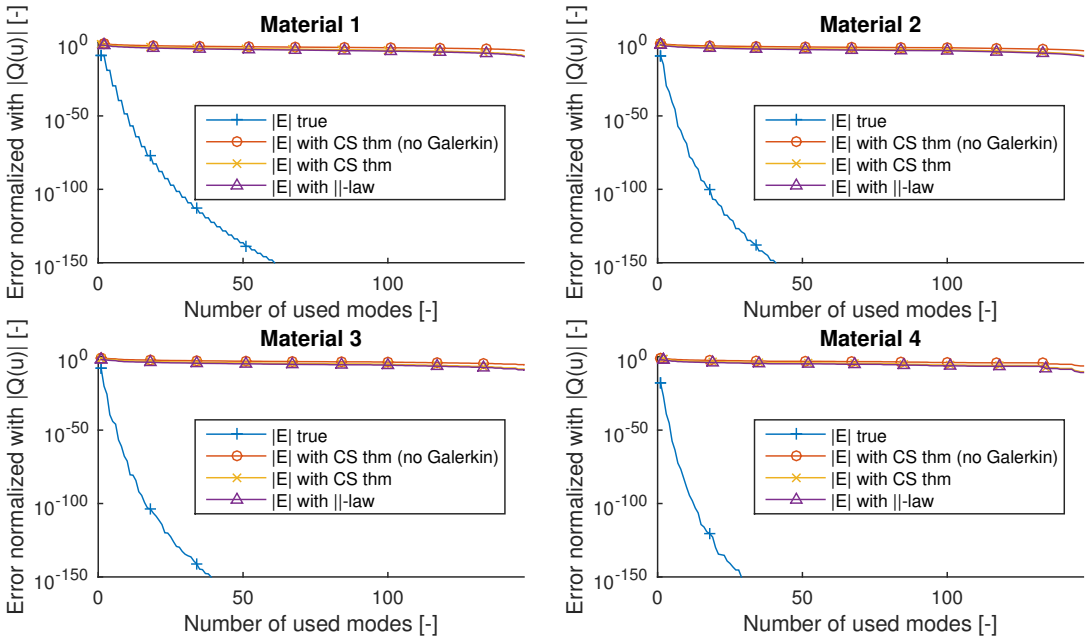


Figure 5.28: *The figure shows the true error and the error estimates that are using all modes when the final temperature is the quantity of interest. The results are produced with the first load case.*

In order to study a more interesting load case when the final temperature is the quantity of interest, the load case is slightly changed such that $\bar{u}$ is a ramp during the whole simulation. Figure 5.29 shows the error estimates that are obtained for this new load case. From the figure it can be seen that the overestimate is still large, but not as large as in Figure 5.28.

When Theorem 3 is applied, and only the reduced modes are used, Equation (5.19) is reduced to

$$\|e^{\star,s}\| \leq \|e^{\star,s}_{est,proj}\| = 2\|Z'\|_{\mathfrak{m}}. \tag{5.31}$$

The results from the error estimates with Theorem 3 is shown in Figure 5.30. Also similar results are perform with Theorem 4. For this quantity of interest,

$$\kappa_{Z,opt} = \left(\frac{2\lambda_{N_R}\|Z'\|^2_{\mathfrak{m}}}{\int_0^T \|r'_M\|^2_{\mathfrak{m}}\,\mathrm{d}t} \right)^{1/4} \tag{5.32}$$

which implies that

$$\begin{cases} E \leq \frac{1}{\sqrt{2}}\|Z'\|^2_{\mathfrak{m}}\left(\frac{1}{\lambda_{N_R}}\int_0^T \|r'_M\|^2_{\mathfrak{m}}\,\mathrm{d}t\right)^{1/2}, \\ E \geq -\frac{1}{\sqrt{2}}\|Z'\|^2_{\mathfrak{m}}\left(\frac{1}{\lambda_{N_R}}\int_0^T \|r'_M\|^2_{\mathfrak{m}}\,\mathrm{d}t\right)^{1/2}. \end{cases} \tag{5.33}$$

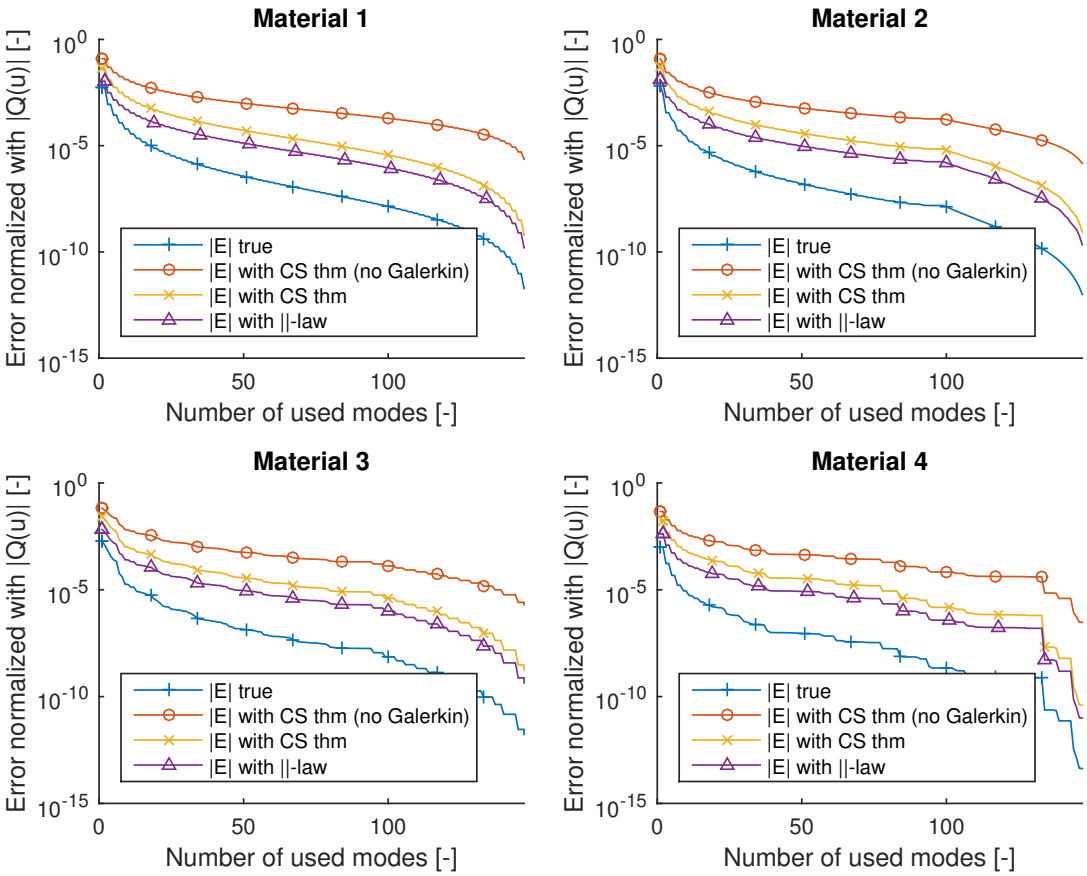


Figure 5.29: *The figure shows the same results as Figure 5.28 with the difference that the load case is changed. In this case $\bar{u}$ is a ramp during the whole simulation.*

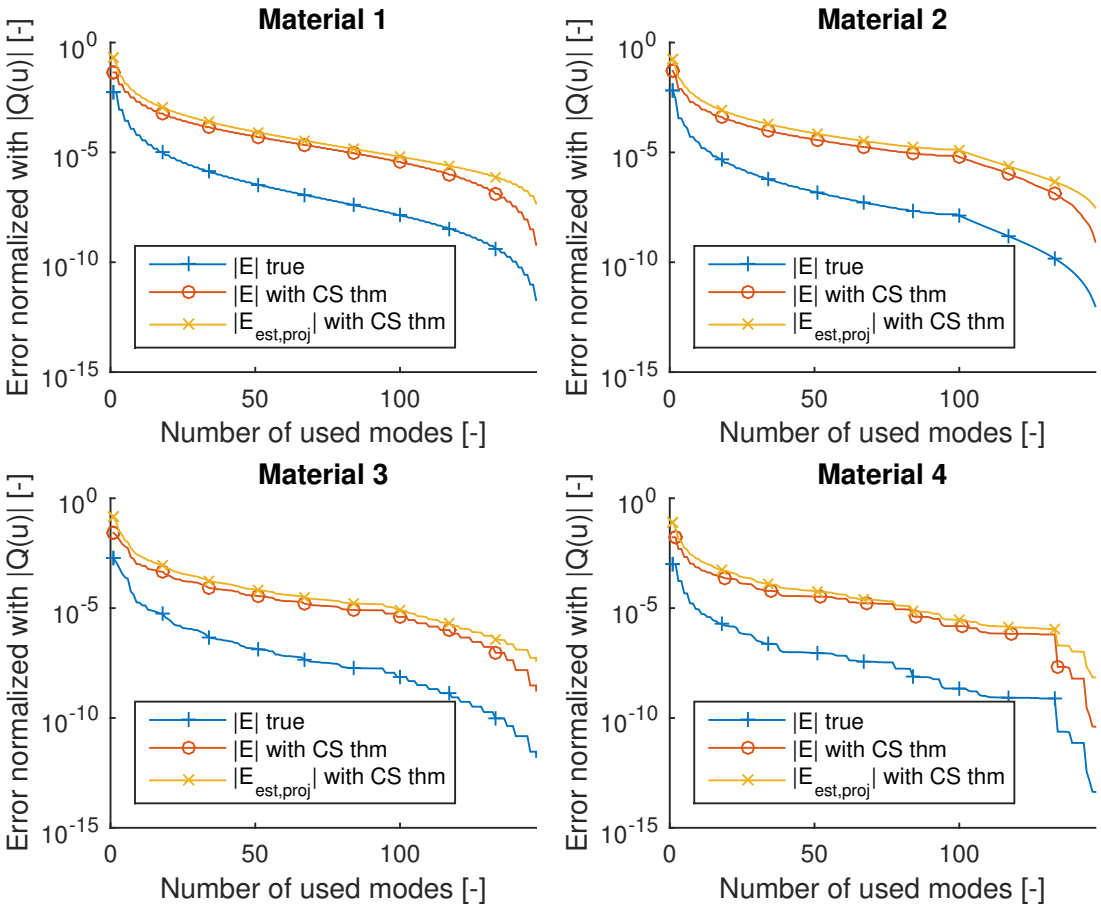


Figure 5.30: *The figure shows the true error and error estimates when Theorem 3 is applied as a function of different number of used modes.*

Figure 5.31 shows the results from Theorem 4. The error estimates are slightly better than Theorem 3, but there is still a large overestimate in all the error estimates. That the symmetrization of the operator for some quantities of interest introduce a large overestimate is one of the main weakness with this method. However, the estimated error when half of the modes are used in material 3 is 0.0015% from the true solution, which in many applications can be considered as good enough.

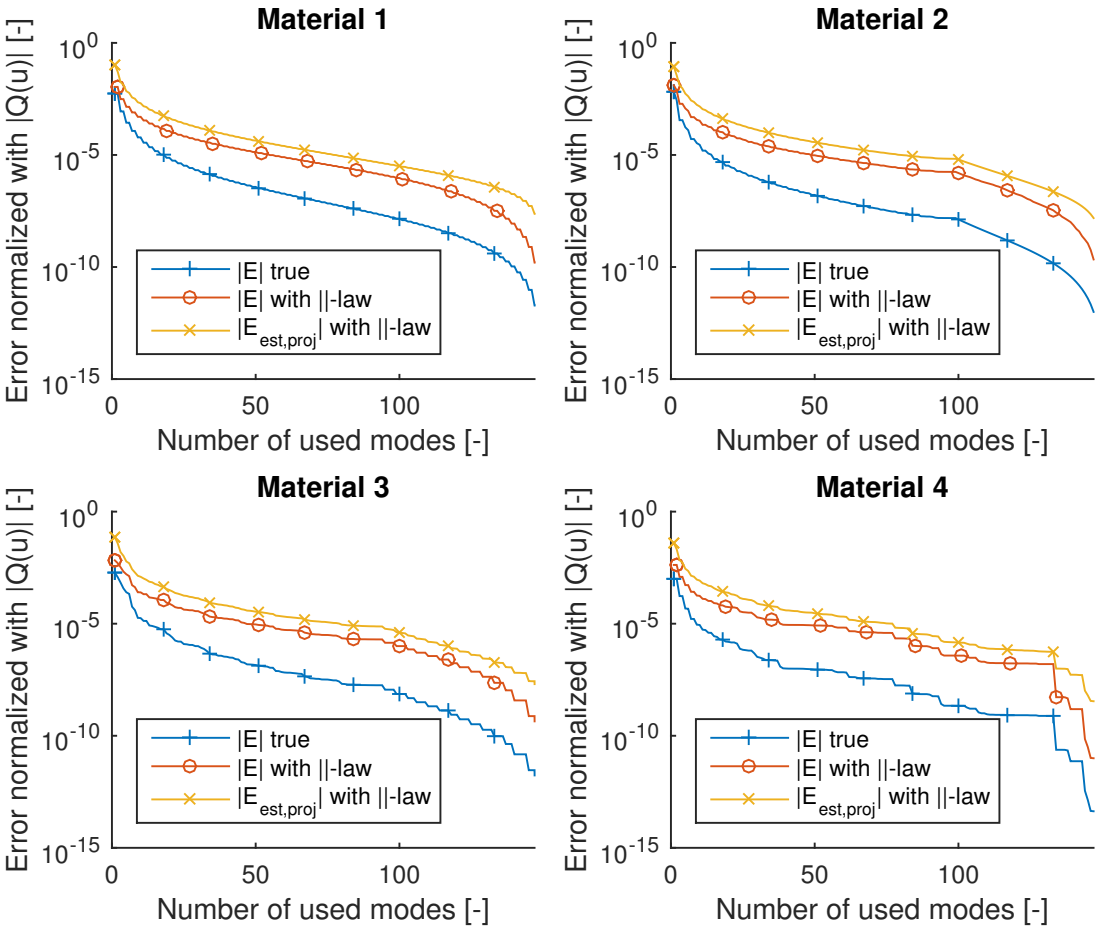


Figure 5.31: *The figure shows the true error and error estimates when Theorem 4 is applied as a function of different number of used modes.*

5.6 Gained computational time

The purpose of introducing Numerical Model Reduction (NMR) is to reduce the computational cost of solving the micro-scale problem. Figure 5.32 shows the computational cost as a function of number of used modes when 1000 free spatial nodes and 100 time-steps are used. In the figure it can be seen that solving the generalized eigenvalue problem is the most time consuming operation. It can also be seen that the computational cost of performing the error estimate is cheap, independently of number of used modes. The computational cost of solving the micro-scale problem, when the eigenvectors and eigenvalues are calculated, increases linearly when the number of modes are increased. Even though it increases linearly, the computational cost is always small compared to the computational cost of solving the generalized eigenvalue problem.

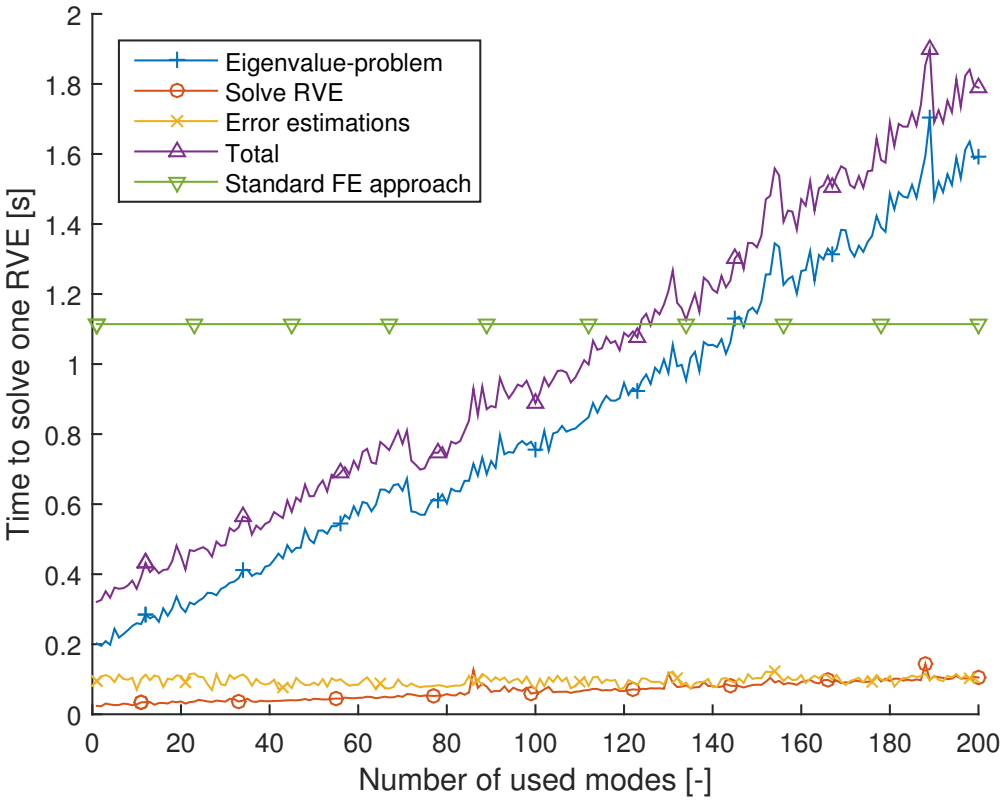


Figure 5.32: *The figure shows the computational cost of solving the micro-scale problem. The simulation is performed with 1000 free spatial nodes and 100 time-steps.*

The micro-scale problem is also solved with a more common Finite Element (FE) approach, in this thesis refereed to as the standard FE approach. In this method, the mass matrix $\underline{\boldsymbol{M}}$ and stiffness matrix $\underline{\boldsymbol{K}}$ are calculated in the same way as for the modal approach. From these matrices, the solution can be calculated as

$$\underline{\boldsymbol{M}}\dot{\boldsymbol{u}} + \underline{\boldsymbol{K}}\boldsymbol{u} = \underline{\boldsymbol{0}} \tag{5.34}$$

together with the same initial- and boundary conditions as for the modal approach. In this thesis, Equation (5.34) is solved with standard Backward Euler. The solution from the standard FE approach can also be seen in Figure 5.32. Since the standard FE approach does not use any modes, the computational cost is constant independently of the number of used modes. This implies that when a lot of modes are used, the standard FE approach is actually cheaper to compute, compared to the modal approach. In the example shown in Figure 5.32, it can be seen that when more than 120 modes are used, which corresponds to 12% of the free nodes, the standard FE approach is cheaper.

Remark: The computational results are performed with MATLAB. For the standard FE approach, MATLAB needs to solve a system of linear equation for each time-step. This is calculated with the \-command which is known as a fast operation in MATLAB [44]. The modal approach needs to find eigenvectors and eigenvalues and its computational cost can probably be reduced if the algorithm is implemented in, for example, C++. This implies that that level where the standard FE approach is faster than the modal approach, which in the presented figure is 12%, can probably be increased if the algorithm is implemented in C++.

6 Summary and Conclusions

The micro-scale problem has successfully been solved for the linear transient heat flow equation when First-order Computational Homogenization is applied on a Representative Volume Element (RVE). The solution consists of a number of linearly independent modes. Numerical Order Reduction is applied such that an arbitrary number of modes can be used, which induces a lower computational cost. It is derived, and also verified in the numerical results, that the stationary part of the solution can either be accounted for by the modes themselves, or pre-computed and implemented in the homogenization procedure such that the stationary solution is always captured independently of the number of modes.

The modes have been determined with Spectral Decomposition. Since the considered transfer heat flow equation is a linear equation, only linear features need to be captured by the modes, which makes Spectral Decomposition a natural choice. When Spectral Decomposition is applied to the transient heat equation, the spatial modes can be estimated as the eigenvectors to the generalized eigenvalue problem. A beautiful feature with Spectral Decomposition is that the activity mode coefficients, which usually are determined by a system of coupled Ordinary Differential Equations (ODEs), can be determined by solving a number of independently ODEs.

The theory is derived in three spatial dimensions, while the numerical example is only implemented in one spatial dimension. In the numerical example, four different micro-structures are studied that are models of homogeneous, two-phase and multiphase materials. All materials that have been implemented in the numerical examples, have been verified against different canonical load cases.

An error analysis has been derived and implemented in order to estimate the error that is introduced when the number of modes are reduced. Error estimates are performed with the classical energy norm. The space-time formulation of the transient heat flow problem consists partially of a non-symmetric operator. In order to apply Cauchy–Schwarz inequality for error estimation, a corresponding symmetric space-time formulation is defined. It is proven that the energy norm of symmetric error is always larger than the energy norm of the true error. Therefore, the symmetric space-time formulation can be used to estimate the true error measured with the energy norm.

In order to determine both the true and symmetric error, measured in the energy norm, all modes are needed. It is derived that Cauchy–Schwarz inequality can be applied to obtain upper and lower bounds of the symmetric error, and thus also the true error, using only the reduced modes. Hence the error estimates with Cauchy–Schwarz inequality can be applied for an arbitrary number of modes to estimate if the solution is close enough to the true solution. If the solution is good enough, the algorithm stops, and if not, additional modes are added to the solution and new error estimates are performed. Note that, how good an error estimate needs to be in order to be good enough depends on the application area.

The error is also estimated with different quantities of interest. For the different quantities of interest, the error estimates are produced with the Galerkin orthogonality, together with either Cauchy–Schwarz inequality or the Parallelogram law. Error estimates that are using either all modes or just the reduced modes are performed. Considered quantities of interest in this thesis are the average temperature, the average heat flux in one direction and the final temperature. For a given application, the best suited quantity of interest can be applied and just an error tolerance has to be defined. The proposed algorithm can then be used to construct the solution that consists of the minimal number of modes that are needed to be within the given error tolerance.

For all error estimates approaches, excepts the final temperature, the true error is close to the symmetric error. In these cases, the mayor part of the overestimate is performed with either Cauchy–Schwarz inequality or the Parallelogram law. For the final temperature the symmetric error is a large overestimate of the true error, depending on that only the transient solution will contribute to the error. Both the canonical load cases that are considered, have a transient solution which is

extremely small at the final time of the simulation. Therefore, also the true error is extremely small, which implies that the symmetric error induces a large overestimate of the results. In this thesis, another load case is tested when the final temperature is the quantity of interest, which has a larger transient solution at the final time of the simulation. For this new load case, the true error is increased and therefore the overestimate of the error is reduced. This discovery indicates how good the error estimate becomes, depends what load case that is applied. The estimated error, for the case when the overestimate is very large, is however between 0.01-1% depending on how many modes that are used, which in many applications can be considered as good enough.

The ratio between the true and estimated errors is calculated in order to investigate if a specific number of modes gives better error estimate. From this analysis, it can be concluded that the ratio grows rapidly for the last third of the used modes, independently of quantity of interest, number of time-steps and number of spatial nodes. The reason why the ratio grows for high modes probably has to do with the discretization of the transient heat flow equation. When the last third of the highest modes are added, the true error will decrease a lot since these modes are used to fit the given discretization. This discretization fit for the highest modes will never happen for the estimated error, since those modes are just estimated.

7 Further work

A key property that is used throughout this thesis is that the equation that describes the transient heat flow problem is linear. This implies that Spectral Decomposition is the natural choice for estimating the spatial modes. Spectral Decomposition can only capture linear behaviour and a further work is to derive a corresponding theory that captures non-linear behaviours as well. This can be done with Proper Orthogonal Decomposition (POD), but how POD shall be used to determine the spatial modes is non-trivial. In a recent study, so called transient training computations on the Representative Volume Element have been performed in order to find an expression of the spatial modes [13]. An outlook to this study can be to use a similar error analysis as in this thesis, combined with the ideas of training computations. If this is done also non-linear equations, such as non-linear heat flow equations and stress-strain equations, can be studied.

Another outlook from this thesis is to use Higher-order Computational Homogenization. As briefly discussed in Section 2.2.2, Higher-order Computational Homogenization is of special interest when there are high gradients in the macro-scale solution.

Even though several quantities of interest are considered in this thesis, other quantities of interest can be implemented. One example of a quantity of interest that can be of interest, that is outside the scope of this thesis, is the final heat flux in one direction.

As discussed in Chapter 5, the numerical results are only implemented in one spatial dimension. Since the theory is derived in three spatial dimensions, there is no theoretical limit to only one spatial dimension. Also, all numerical results are implemented in MATLAB, which is known to be a slow language compared to, for example, C++ (in most applications). A further work could be to implement the numerical results in three spatial dimensions and in another programming language. In this thesis, these numerical improvements have been down-prioritized in favor of a larger error analysis.

Bibliography

[1] Geers MGD, Kouznetsova VG, Brekelmans WAM. Computational homogenization. In: Pippan R, Gumbsch P, editors. Multiscale Modelling of Plasticity and Fracture by Means of Dislocation Mechanics. Springer Vienna; 2010. p. 327–394.

[2] Geers M, Kouznetsova V, Brekelmans W. Multi-scale computational homogenization: Trends and challenges. Journal of computational and applied mathematics. 2010;234(7):2175–2182.

[3] Fish J, Shek K, Pandheeradi M, Shephard MS. Computational plasticity for composite structures based on mathematical homogenization: Theory and practice. Computer Methods in Applied Mechanics and Engineering. 1997;148(1):53–73.

[4] Miehe C, Schröder J, Schotte J. Computational homogenization analysis in finite plasticity simulation of texture development in polycrystalline materials. Computer methods in applied mechanics and engineering. 1999;171(3):387–418.

[5] Zohdi T, Wriggers P. A model for simulating the deterioration of structural-scale material responses of microheterogeneous solids. Computer Methods in Applied Mechanics and Engineering. 2001;190(22):2803–2823.

[6] Terada K, Kikuchi N. A class of general algorithms for multi-scale analyses of heterogeneous media. Computer methods in applied mechanics and engineering. 2001;190(40):5427–5464.

[7] Miehe C, Koch A. Computational micro-to-macro transitions of discretized microstructures undergoing small strains. Archive of Applied Mechanics. 2002;72(4-5):300–317.

[8] Kouznetsova V, Geers MG, Brekelmans WM. Multi-scale constitutive modelling of heterogeneous materials with a gradient-enhanced computational homogenization scheme. International Journal for Numerical Methods in Engineering. 2002;54(8):1235–1260.

[9] Ostoja-Starzewski M. Material spatial randomness: From statistical to representative volume element. Probabilistic engineering mechanics. 2006;21(2):112–132.

[10] Zohdi TI, Wriggers P. An introduction to computational micromechanics. Springer Science & Business Media; 2008.

[11] Temizer I, Wriggers P. On the computation of the macroscopic tangent for multiscale volumetric homogenization problems. Computer Methods in Applied Mechanics and Engineering. 2008;198(3):495–510.

[12] Fritzen F. Microstructural modeling and computational homogenization of the physically linear and nonlinear constitutive behavior of micro-heterogeneous materials. vol. 1. KIT Scientific Publishing; 2011.

[13] Jänicke R, Larsson F, Runesson K, Steeb H. Numerical identification of a viscoelastic substitute model for heterogeneous poroelastic media by a reduced order homogenization approach. Computer Methods in Applied Mechanics and Engineering. 2016;298:108–120.

[14] Larsson F, Runesson K, Saroukhani S, Vafadari R. Computational homogenization based on a weak format of micro-periodicity for RVE-problems. Computer Methods in Applied Mechanics and Engineering. 2011;200:11–26.

[15] Larsson F, Runesson K, Su F. Variationally consistent computational homogenization of transient heat flow. International journal for numerical methods in engineering. 2010;81(13):1659–1686.

[16] Galvanetto U, Aliabadi MF. Multiscale modeling in solid mechanics: computational approaches. vol. 3. World Scientific; 2009.

[17] Jianmin Qu MC. Fundamentals of Micromechanics of Solids. John Wiley & Sons, Ltd; 2006.

[18] Fritzen F, Böhlke T. Reduced basis homogenization of viscoelastic composites. Composites Science and Technology. 2013;76:84–91.

[19] Kouznetsova V, Geers M, Brekelmans W. Multi-scale second-order computational homogenization of multi-phase materials: a nested finite element solution strategy. Computer Methods in Applied Mechanics and Engineering. 2004;193(48):5525–5550.

[20] Fritzen F, Hodapp M, Leuschner M. GPU accelerated computational homogenization based on a variational approach in a reduced basis framework. Computer Methods in Applied Mechanics and Engineering. 2014;278:186–217.

[21] Roussette S, Michel JC, Suquet P. Nonuniform transformation field analysis of elastic–viscoplastic composites. Composites Science and Technology. 2009;69(1):22–27.

[22] Ryckelynck D, Benziane DM. Multi-level a priori hyper-reduction of mechanical models involving internal variables. Computer Methods in Applied Mechanics and Engineering. 2010;199(17):1134–1142.

[23] Webster J, Watson RT. Analyzing the past to prepare for the future: Writing a. MIS quarterly. 2002;26(2):13–23.

[24] Drugan W, Willis J. A micromechanics-based nonlocal constitutive equation and estimates of representative volume element size for elastic composites. Journal of the Mechanics and Physics of Solids. 1996;44(4):497–524.

[25] Kanit T, Forest S, Galliet I, Mounoury V, Jeulin D. Determination of the size of the representative volume element for random composites: statistical and numerical approach. International Journal of solids and structures. 2003;40(13):3647–3679.

[26] Zhang J, Liu K, Luo C, Chattopadhyay A. Crack initiation and fatigue life prediction on aluminum lug joints using statistical volume element–based multiscale modeling. Journal of Intelligent Material Systems and Structures. 2013;24(17):2097–2109.

[27] Yin X, Chen W, To A, McVeigh C, Liu WK. Statistical volume element method for predicting microstructure–constitutive property relations. Computer methods in applied mechanics and engineering. 2008;197(43):3516–3529.

[28] Groeber M, Ghosh S, Uchic MD, Dimiduk DM. A framework for automated analysis and simulation of 3d polycrystalline microstructures.: Part 1: Statistical characterization. Acta Materialia. 2008;56(6):1257–1273.

[29] Sanei SHR, Fertig RS. Uncorrelated volume element for stochastic modeling of microstructures based on local fiber volume fraction variation. Composites Science and Technology. 2015;117:191–198.

[30] Lesičar T, Tonković Z, Sorić J. Second-Order Computational Homogenization Approach Using Higher-Order Gradients at Microlevel. In: Key Engineering Materials. vol. 665. Trans Tech Publ; 2015. p. 181–184.

[31] Gray WG, Schrefler BA, Pesavento F. The solid phase stress tensor in porous media mechanics and the Hill–Mandel condition. Journal of the Mechanics and Physics of Solids. 2009;57(3):539–554.

[32] Dvorak GJ, Benveniste Y. On transformation strains and uniform fields in multiphase elastic media. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. vol. 437. The Royal Society; 1992. p. 291–310.

[33] Michel JC, Suquet P. Nonuniform transformation field analysis. International journal of solids and structures. 2003;40(25):6937–6955.

[34] Michel JC, Suquet P. Computational analysis of nonlinear composite structures using the nonuniform transformation field analysis. Computer methods in applied mechanics and engineering. 2004;193(48):5477–5502.

[35] Fritzen F, Böhlke T. Three-dimensional finite element implementation of the nonuniform transformation field analysis. International Journal for Numerical Methods in Engineering. 2010;84(7):803–829.

[36] Fritzen F, Leuschner M. Reduced basis hybrid computational homogenization based on a mixed incremental formulation. Computer Methods in Applied Mechanics and Engineering. 2013;260:143–154.

[37] Eschmeier J, Putinar M. Spectral decompositions and analytic sheaves. 10. Oxford University Press; 1996.

[38] Sparr G. Linjär Algebra. Studentlitteratur AB; 1997.

[39] Hsu SB. Ordinary differential equations with applications. World scientific; 2006.

[40] Azam SE. Model Order Reduction of Dynamic Systems via Proper Orthogonal Decomposition. In: Online Damage Detection in Structural Systems. Springer; 2014. p. 57–86.

[41] Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901;2(11):559–572.

[42] Jolliffe I. Principal component analysis. Wiley Online Library; 2002.

[43] Klema VC, Laub AJ. The singular value decomposition: Its computation and some applications. Automatic Control, IEEE Transactions on. 1980;25(2):164–176.

[44] MathWorks. Systems of Linear Equations. `http://se.mathworks.com/help/matlab/math/systems-of-linear-equations.html#brs10rz-1`; 2016.

[45] Debnath L, Mikusinski P. Introduction to Hilbert Spaces with Applications. Elsevier Science; 2005.

[46] Lebedev LP, Vorovich II, Cloud MJ. Functional analysis in mechanics. Springer Science & Business Media; 2012.

[47] Johnson C. Numerical solution of partial differential equations by the finite element method. Courier Corporation; 2012.

A Basic functional analysis

In order to prove that the solution to the finite element problem is unique, the solution space and test space need to be defined properly. As it is shown is this section, the uniqueness is guaranteed if Lax-Milgrams Theorem is fulfilled. In order to do so, both the solution space and test space need to be Hilbert spaces.

In order to reach a proper definition of a Hilbert space, this chapter starts with definitions and examples of a vector space, Section A.1, a normed space, Section A.2, and a Banach space, Section A.3. From those, the Hilbert spaces are introduced in Section A.4. Section A.5 presents Lax-Milgrams Theorem and the important Sobolev spaces, which are subsets of the Hilbert spaces, are presented in Section A.6. Finally, Section A.7 gives proofs to several statements in Section A.3 and Section A.8 gives a proof of Lax-Milgrams Theorem.

A.1 Vector space

A vector space $\mathbb{E}$ is a set containing certain kinds of elements. In order to obtain a valid vector space, certain axioms needs to be fulfilled [45]. The formal definition of a vector space is defined as (quoted from [45]):

Definition 1 (Vector space)**.** *By a vector space we mean a nonempty set $\mathbb{E}$ with two operators:*

$(x, y) \mapsto x + y$ *from* $\mathbb{E} \times \mathbb{E}$ *into* $\mathbb{E}$ *called addition,*

$(\lambda, x) \mapsto \lambda x$ *from* $\mathbb{R} \times \mathbb{E}$ *into* $\mathbb{E}$ *called multiplication by scalar,*

such that the following conditions are satisfied for all $x, y, z \in \mathbb{E}$ and $\alpha, \beta \in \mathbb{F}$:

(a) $x + y = y + x$;

(b) $(x + y) + z = x + (y + z)$;

(c) For every $x, y \in \mathbb{E}$ there exists a $z \in \mathbb{E}$ such that $x + z = y$;

(d) $\alpha(\beta x) = (\alpha\beta)x$;

(e) $(\alpha + \beta)x = (\alpha x) + (\beta x)$;

(f) $\alpha(x + y) = \alpha x + \alpha y$;

(g) $1x = x$.

If $\mathbb{F} = \mathbb{R}$, then $\mathbb{E}$ is called a *real vector space* and if $\mathbb{F} = \mathbb{C}$, $\mathbb{E}$ is called a *complex vector space.*

Examples of spaces that fulfill the vector space definition is not limited by the common scalar fields $\mathbb{R}$ and $\mathbb{C}$. For example a corresponding vector space in N dimensions, can be defined as

$$\begin{aligned} \mathbb{R}^N &= \{(x_1, x_2, \ldots, x_N) : \ x_1, x_2, \ldots, x_N \in \mathbb{R}\}, \\ \mathbb{C}^N &= \{(z_1, z_2, \ldots, z_N) : \ z_1, z_2, \ldots, z_N \in \mathbb{C}\}. \end{aligned} \tag{A.1}$$

From the definition of a vector space, it is also possible to define function spaces. Let $\mathbb{X}$ be an arbitrary non-empty set and let $\mathbb{E}$ be a vector space. By letting $\mathbb{G}$ be a space of all functions from $\mathbb{X}$ to $\mathbb{E}$, the operations addition and multiplication of $\mathbb{G}$ can be defined as

$$\begin{aligned} (f + g)(x) &= f(x) + g(x), \\ (\lambda f)(x) &= \lambda f(x). \end{aligned} \tag{A.2}$$

If all the axioms (a)-(g) are fulfilled, the function space $\mathbb{G}$ is a vector space as well. By letting the non-empty set $\mathbb{X}$ be an open subset of $\mathbb{R}$, denoted Ω, the following function spaces are vector spaces as well:

$$\begin{aligned}\mathcal{C}(\Omega) &= \text{The space of all continuous function defined on } \Omega,\\ \mathcal{C}^k(\Omega) &= \text{The space of all continuous function defined on } \Omega \text{ with continuous}\\ &\quad\ \text{partial derivative of order } k,\\ \mathcal{P}(\Omega) &= \text{The space of all polynomials defined on } \Omega.\end{aligned}$$

Another family of vector spaces that are of strong importance in functional analysis are the l^p spaces, which are defined as follows (quoted from [45]):

Definition 2 (l^p-spaces)**.** *Denote by l^p, for $p \geq 1$, the space of all infinite sequences $(z_n)_{n=1}^{\infty}$ of complex numbers such that $\sum_{n=1}^{\infty} |z_n|^p < \infty$.*

A.2 Normed space

In basic calculus, the norm is defined as the length of a vector. The concept of a norm can be extended as an abstract generalization [45]. The definition of a norm is then given by:

Definition 3 (Norm)**.** *Let $\mathbb{E}$ be a vector space. Then $\|\cdot\| : \mathbb{E} \to [0, \infty)$ is a norm on $\mathbb{E}$ if:*

(a) $\|x\| = 0, \Rightarrow x = \mathbf{0}, \quad \forall x \in \mathbb{E}.$

(b) $\|\lambda x\| = |\lambda|\|x\|, \quad \forall \lambda \in \mathbb{R}, \ \forall x \in \mathbb{E}.$

(c) $\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in \mathbb{E}.$

Note that condition (c) implies that a norm is always non-negative since

$$0 = \|0\| = \|x - x\| \leq \|x\| + \|x\| = 2\|x\|. \tag{A.3}$$

An equivalent definition of the norm is to have equivalence is condition (a). The condition that $x = \mathbf{0}, \Rightarrow \|x\| = 0, \ \forall x \in \mathbb{E}$ follows however from condition (b) since

$$\|\mathbf{0}\| = \|0\mathbf{0}\| = |0|\|\mathbf{0}\| = 0\|\mathbf{0}\| = 0. \tag{A.4}$$

The most common norm is the *Euclidean norm*, which on $\mathbb{R}^N$ is defined as

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_N^2}, \quad x = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^{\mathbb{N}}. \tag{A.5}$$

From the definition of a vector space and the definition of a norm, it is now possible to give the definition of a normed space, which simply says that (quoted from [45]):

Definition 4 (Normed space)**.** *A vector space with a norm is called a* normed space*.*

A.3 Banach space

In order to reach the vector spaces that are used in the Finite Element Analysis (FEA), the definition and properties of a Banach space is required [46]. Before the definition of the Banach space is given, the definition of a Cauchy sequence is introduced (quoted from [45]):

Definition 5 (Cauchy sequence)**.** *A sequence of vectors $(x_n)_{n=1}^{\infty}$ in a normed space is called a Cauchy sequence if for every $\epsilon > 0$ there exists a number M such that $\|x_m - x_n\| < \epsilon$ for all $m, n > M$.*

From the concepts of a normed space and a Cauchy sequence, a Banach space can now be defined as (quoted from [45]):

Definition 6 (Banach space)**.** *A normed space* $\mathbb{E}$ *is called complete if every Cauchy sequence in* $\mathbb{E}$ *converges to an element in* $\mathbb{E}$*. A complete normed space is called a Banach space.*

The following normed spaces

1. $\mathbb{E} = \mathbb{R}, \quad \|\cdot\| = |\cdot|,$
2. $\mathbb{E} = l^p \ni \boldsymbol{x} = (x_1,_x 2, \dots, x_n, \dots), \quad \|x\|_l^p = \Big(\sum_{n=1}^N |x_n|^p\Big)^{1/p},$

are Banach spaces, while

3. $\mathbb{E} = \mathcal{C}^1([0,1]), \quad \|f\| = max_{x\in[0,1]}|f(x)|,$

is not. For proofs of these statements, see Appendix A.7.

One of the most important Banach spaces is the so called L^p-spaces. In order to define the L^p-spaces, the concepts of a Lebesgue integrable function and logically integrable functions are needed. The definition of a Lebesgue integrable function is given by (quoted from [45]):

Definition 7 (Lebesgue integrable function)**.** *A real valued function* f *defined on* $\mathbb{R}$ *is called a* Lebesgue integrable function *if there exists a sequence of step functions* $(f_n)_{n=1}^\infty$ *such that the following two conditions are satisfied:*

(a) $\sum_{n=1}^\infty \int |f_n| < \infty$*;*

(b) $f(x) = \sum_{n=1}^\infty f_n(x)$ *for every* $x \in \mathbb{R}$ *such that* $\sum_{n=1}^\infty |f_n(x)| < \infty$*.*

Throughout this thesis, the space of all Lebesgue integrable functions defined on $\mathbb{R}$ are denoted $L^1(\mathbb{R})$. The definition of a logically integrable function is given by (quoted from [45]):

Definition 8 (Logically integrable function)**.** *A function* f *defined on* $\mathbb{R}$ *is called logically integrable if the integral* $\int_a^b f$ *exists for every* $-\infty < a < b < \infty$*.*

The L^p-spaces can now be defined as (quoted from [45]):

Definition 9 ($L^p(\mathbb{R})$-space)**.** *For a real* $p > 1$*, by* $L^p(\mathbb{R})$ *we denote the space of all complex-valued logically integrable functions* f *such that* $|f|^p \in \mathbb{L}^1(\mathbb{R})$*.*

It can be proven that if the $L^p(\mathbb{R})$-space is equipped with the norm

$$\|f\|_p = \Big(\int |f|^p\Big)^{1/p}, \tag{A.6}$$

the $L^p(\mathbb{R})$-space is a Banach space. This can be done be first showing that $L^p(\mathbb{R})$ is a vector space, followed up by that $\|f\|_p$ is a norm in $L^p(\mathbb{R})$ and finally shows that the space is complete.

A.4 Hilbert space

In order to define a Hilbert space, the definition of an inner product space is needed which is a bilinear mapping defined as (quoted from [45]):

Definition 10 (Inner product space)**.** *Let* $\mathbb{E}$ *be a complex vector space. A mapping* $\langle\cdot,\cdot\rangle : \mathbb{E}\times\mathbb{E} \mapsto \mathbb{C}$ *is called an* inner product space *in* $\mathbb{E}$ *if for any* $x, y, z \in \mathbb{E}$ *and* $\alpha, \beta \in \mathbb{C}$ *the following condition are satisfied:*

(a) $\langle x, y\rangle = \overline{\langle y, x\rangle}$ *(the bar denotes the complex conjugate);*

(b) $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$*;*

(c) $\langle x, x\rangle \geq 0$*;*

(d) $\langle x, x\rangle = 0$ *implies* $x = 0$.

A vector space with an inner product is called an inner product space. A couple of examples of inner product spaces are

(A) $\mathbb{E} = l^2 \ni \boldsymbol{x} = (x_1, x_2, \dots, x_n, \dots), \quad \langle \boldsymbol{x}, \boldsymbol{y}\rangle_{l^2} = \sum_{n=1}^{\infty} x_n y_n$

(B) $\mathbb{E} = \mathcal{C}([0,1]), \quad \langle f, g\rangle_{L^2} = \int_0^1 f(x)g(x)\,\mathrm{d}x.$

For a given inner product space, its norm is defined by (quoted from [45]):

Definition 11 (Norm of an inner product space)**.** *By the norm in an* inner product space $\mathbb{E}$ *we mean the functional by* $\|x\| = \sqrt{\langle x, x\rangle}$.

From the definition of a Banach space and an inner product space, a Hilbert space can now be defined as (quoted from [45]):

Definition 12 (Hilbert space)**.** *A complete inner product space is called a Hilbert space.*

This means that a Hilbert space can be thought of as Banach space with an inner product equipped satisfying $\|x\| = \sqrt{\langle x, x\rangle}$. From the examples given as inner product spaces it can be seen that (A) is a Hilbert space, while (B) is not since $(\mathcal{C}([0,1]), \|\cdot\|_{L^2})$ is not a Banach space.

As introduced and mentioned in Section A.3, the $L^p(\mathbb{R})$ is one of the most important Banach spaces. It is however only the L^2-space, equipped with the inner product

$$\langle f, g\rangle = \int f\bar{g} \tag{A.7}$$

that also is a Hilbert space. In FEA context, functions are typically real-valued and defined in some domain Ω. By introducing a measure[1], the inner product becomes

$$\langle f, g\rangle = \int_\Omega fg\,\mathrm{d}\Omega \tag{A.8}$$

and the corresponding norm is given by

$$\|f\| = \sqrt{\langle f, f\rangle} = \Big(\int_\Omega |f|^2\,\mathrm{d}\Omega\Big)^{1/2}. \tag{A.9}$$

This implies that a real-valued function f will lie in the L^2-space if

$$\int_\Omega |f|^2\,\mathrm{d}\Omega < \infty \tag{A.10}$$

and hence functions in L^2 are sometimes called square-integrable functions.

[1]The theory behind a measure is outside of the scope of this thesis.

A.5 Lax-Milgrams Theorem

The probably most important theorem from functional analysis that is used in the FEA context is Lax Milgrams Theorem [47]. In order to define Lax Milgrams Theorem, the concept of an bilinear, bounded, and coercive operator is needed which can be defined as:

Definition 13 (Bilinear, bounded, coercive)**.** *Let* $(\mathbb{E}, \langle\cdot,\cdot\rangle)$ *be a Hilbert space. An operator* $\phi : \mathbb{E} \times \mathbb{E} \to \mathbb{R}$ *is called bilinear if*

$$\begin{cases} \phi(\alpha x + \beta y, x) = \alpha\phi(x, z) + \beta\phi(y, z), & \forall \alpha, \beta \in \mathbb{C}\ \forall x, y, z \in \mathbb{E} \\ \phi(x, \alpha y + \beta z) = \bar{\alpha}\phi(x, y) + \bar{\beta}\phi(x, z), & \forall \alpha, \beta \in \mathbb{C}\ \forall x, y, z \in \mathbb{E}. \end{cases} \tag{A.11}$$

ϕ *is called* bounded *if*

$$|\phi(x, y)| \leq M\|x\|\|y\|, \quad \forall x, y \in \mathbb{E},\ \textit{some } M > 0. \tag{A.12}$$

ϕ *is called* coercive *if*

$$\phi(x, x) \geq k\|x\|^2, \quad \forall x \in \mathbb{E},\ \textit{some } k > 0. \tag{A.13}$$

The theorem can now be stated as:

Theorem 8 (Lax-Milgrams Theorem)**.** *Let* $(\mathbb{E}, \langle\cdot,\cdot\rangle)$ *be a Hilbert space. Let* $\phi : \mathbb{E} \times \mathbb{E} \to \mathbb{R}$ *be a bilinear, bounded and coercive functional and let* $f : \mathbb{E} \to \mathbb{R}$ *be a bounded, linear functional. Then there exists a unique* $\widetilde{x}_f \in \mathbb{E} :\ \phi(x, \widetilde{x}_f) = f(x)\ \forall x \in \mathbb{E}$*.*

For a proof of Lax-Milgrams Theorem, see Appendix A.8. To put the theorem in the FEA context, x can be thought of as the test variable and $\widetilde{x}_f$ can be thought of as the solution variable. With this notation, the variational form can be written as: Find $\widetilde{x}_f \in \mathbb{E}$ such that

$$\phi(x, \widetilde{x}_f) = f(x)\ \forall x \in \mathbb{E}. \tag{A.14}$$

Hence Lax-Milgrams Theorem proves that the solution to the variational form is unique.

A.6 Sobolev space

One of the most important Hilbert spaces in the FEA context is the so called Sobolev spaces, which can be defined as follows [45]. Let Ω be an open set in $\mathbb{R}^N$. Denote by $\widetilde{\mathbb{U}}^m(\Omega)$, $m = 1, 2, \ldots$ the space of all real-valued functions $f \in \mathcal{C}^m(\Omega)$ such that $D^\alpha f \in L^2(\Omega)$ for all $|\alpha| \leq m$, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$, $\alpha_1, \alpha_2, \ldots, \alpha_N$ are non-negative integers, $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_N$ and

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_N^{\alpha_N}}. \tag{A.15}$$

Since $D^\alpha f \in L^2(\Omega)$ it follows that for every $f \in \widetilde{\mathbb{U}}^m(\Omega)$ the inequality equation

$$\int_\Omega \Big| \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_N^{\alpha_N}} \Big| \, \mathrm{d}\Omega < \infty \tag{A.16}$$

must hold. By also letting $g \in \mathcal{C}^m(\Omega)$ the inner product of the Sobolev space is defined by

$$\langle f, g \rangle = \int_\Omega \sum_{|\alpha| \leq m} D^\alpha f D^\alpha g \, \mathrm{d}\Omega. \tag{A.17}$$

In the special case when Ω is a one-dimensional interval, I, and $m = 1$, the Sobolev is given by

$$\widetilde{\mathbb{U}}^1(I) = \{f : \ f + \frac{\partial f}{\partial x} \in L^2(I)\}. \tag{A.18}$$

Letting $u, v \in \widetilde{\mathbb{U}}^1(I)$ gives that the inner product can be defined as

$$\langle u, v\rangle = \int_I u(x)v(x) + u'(x)v'(x)\,\mathrm{d}x \tag{A.19}$$

and hence the corresponding norm is given by

$$\|u\| = \Big(\int_I v(x)^2 + v'(x)^2\,\mathrm{d}x\Big)^{1/2} \tag{A.20}$$

A.7 Proofs regarding Banach spaces

From section A.3 it says that the following normed spaces

1. $\mathbb{E} = \mathbb{R}, \quad \|\cdot\| = |\cdot|$
2. $\mathbb{E} = l^p \ni \boldsymbol{x} = (x_{1,x}\,2, \ldots, x_n, \ldots), \quad \|x\|_l^p = \Big(\sum_{n=1}^N |x_n|^p\Big)^{1/p}$

are Banach spaces, while

3. $\mathbb{E} = \mathcal{C}^1([0,1]), \quad \|f\| = max_{x\in[0,1]}|f(x)|,$

is not. This section gives proofs to the statements.

Proof. 1. Fix a Cauchy sequence $(x_n)_{n=1}^\infty$ in $(\mathbb{R}, |\cdot|)$, i.e. $|x_n - x_m| \to 0, \ n, m \to \infty$.

First show that $\exists\ x \in \mathbb{R}$ such that $|x_n - x| \to 0, \ n \to \infty$. This can be done by first noting that $(x_n)_{n=1}^\infty$ is a bounded sequence since it is a Cauchy sequence, i.e. $\exists M > 0$ such that $-M < x_n < M$ for all n. Then

$$\lim_{n\to\infty} sup\, x_n \in \mathbb{R} \tag{A.21}$$

from the supremum axiom for $\mathbb{R}$. By setting

$$x := \lim_{n\to\infty} \sup_{k\geq n}\ x_k \tag{A.22}$$

it is proven that $|x_n - x| \to 0, \ n \to \infty$ since $\sup_{k\geq n}\ x_k$ is a decreasing sequence in x.

Now show that $x_n \to x$ in $(\mathbb{R}, |\cdot|)$. By using the definition of x, see Equation (A.22), the identities

$$\begin{cases} \forall \epsilon > 0\ \exists N :\ x_n < x + \epsilon\ \forall n \geq N, \\ \forall \epsilon > 0\ \forall N\ \exists\, n \geq N : x_n > x - \epsilon, \end{cases} \tag{A.23}$$

must hold. This implies that there exists a subsequence $(x_{n_k})_{k=1}^\infty$ of $(x_n)_{n=1}^\infty$ such that $x_{n_k} \to x$ in $(\mathbb{R}, |\cdot|)$. That all Cauchy sequence converges can now be shown from

$$\|x_n - x\| = \|x_n - x_{n_k} + x_{n_k} - x\| \leq \|x_n - x_{n_k}\| + \|x_{n_k} - x\| \to 0, n \to \infty, \tag{A.24}$$

and hence $(\mathbb{R}, |\cdot|)$ is a Banach space. □

Proof. 2. Fix a Cauchy sequence $(\boldsymbol{x}_n)_{n=1}^{\infty}$ in $(l^p, \|\cdot\|_{l^p})$. First show that there exist a limit point $\boldsymbol{x}$. Set $\boldsymbol{x_n} = (x_1^{(n)}, x_2^{(n)}, \dots, x_k^{(n)}, \dots)$, for $n = 1, 2 \dots$. Consider

$$\|\boldsymbol{x}_n - \boldsymbol{x}_m\|_{l^p} = \Big(\sum_{k=1}^{\infty} |x_k^{(n)} - x_k^{(m)}|^p\Big)^{1/p} \geq |x_{k_0}^{(n)} - x_{k_0}^{(m)}|, \quad \forall k_0 = 1, 2 \dots . \tag{A.25}$$

For all $k_0 = 1, 2, \dots$ $(x_{k_0}^{(n)})_{n=1}^{\infty}$ is a Cauchy sequence in $(\mathbb{R}, |\cdot|)$. But from the previous proof it is known that $(\mathbb{R}, |\cdot|)$ is a Banach space. This implies that $(x_{k_0}^{(n)})_{n=1}^{\infty}$ converges in $(\mathbb{R}, |\cdot|)$. Call this limit x_{k_0} and hence $x_{k_0}^{(n)} \to x_{k_0},\ n \to \infty\ \forall k_0 = 1, 2, \dots$.

Now show that the limit point lies in l^p. For a fix $\epsilon > 0$ there exists a $N > 0$ such that $\|\boldsymbol{x}_n - \boldsymbol{x}_m\|_{l^p} < \epsilon\ \forall n, m \geq N$. In particular

$$\sum_{k=1}^{K} |x_k^{(n)} - x_k^{(m)}|^p < \epsilon^p,\ \forall n, m \geq N,\ \forall K \in \mathbb{N}. \tag{A.26}$$

Letting $m \to \infty$ gives that

$$\sum_{k=1}^{K} |x_k^{(n)} - x_k|^p < \epsilon^p,\ \forall n \geq N,\ \forall K \in \mathbb{N}. \tag{A.27}$$

Minkovskis inequality says that: Let $p \geq 1$. If (x_n), $(y_n) \in l^p$, then

$$\Big(\sum_{k=1}^{\infty} |x_k + y_k|^p\Big)^{1/p} \leq \Big(\sum_{k=1}^{K} |x_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^{K} |y_k|^p\Big)^{1/p}. \tag{A.28}$$

Using Minkovskis inequality gives that

$$\Big(\sum_{k=1}^{K} |x_k|^p\Big)^{1/p} = \Big(\sum_{k=1}^{K} |x_k|^p - |x_k^{(n)}|^p + |x_k^{(n)}|^p\Big)^{1/p} \leq \Big(\sum_{k=1}^{K} |x_k - x_k^{(n)}|^p\Big)^{1/p} + \Big(\sum_{k=1}^{K} |x_k^{(n)}|^p\Big)^{1/p}. \tag{A.29}$$

Equation (A.27) implies that

$$\Big(\sum_{k=1}^{K} |x_k|^p\Big)^{1/p} \leq \Big(\sum_{k=1}^{K} |x_k - x_k^{(n)}|^p\Big)^{1/p} + \Big(\sum_{k=1}^{K} |x_k^{(n)}|^p\Big)^{1/p} \leq \epsilon + \|\boldsymbol{x}_n\|_{l^p} \leq \epsilon + \sup_n \|\boldsymbol{x}_n\|_{l^p} < \infty \tag{A.30}$$

But the final expression is independent of K. Therefore it is allowed to let K tend to infinity which implies that $\|\boldsymbol{x}\|_{l^p} < \infty$ and hence $\boldsymbol{x} \in l^p$. Equation (A.27) now implies that $\boldsymbol{x}_n \to \boldsymbol{x},\ n \to \infty$ in $(l^p, \|\cdot\|_{l^p})$. Hence all Cauchy sequences converge in $(l^p, \|\cdot\|_{l^p})$ and thus $(l^p, \|\cdot\|_{l^p})$ is a Banach space. □

Proof. 3. In order to show that $(\mathcal{C}^1([0,1]), max_{x\in[0,1]}|f(x)|)$ is not a Banach space it is enough to show that one Cauchy sequence does not converge. Consider $f(x) = |x - \frac{1}{2}| \notin \mathcal{C}^1([0,1])$. Set $f_n(x) = \sqrt{(x - \frac{1}{2})^2 + \frac{1}{n}}$, $n = 1, 2, \dots$. Then $f_n \in \mathcal{C}^1([0,1])$, $n = 1, 2, \dots$ and

$$0 \leq f_n(x) - f(x) = \sqrt{\big(x - \frac{1}{2}\big)^2 + \frac{1}{n}} - \Big|x - \frac{1}{2}\Big|, \tag{A.31}$$

for all $x \in [0,1]$. Hence $\|f_n - f\| \to 0,\ n \to \infty$. Thus $(f_n)_{n=1}^{\infty}$ is a Cauchy sequence in $(\mathcal{C}^1([0,1]), max_{x\in[0,1]}|f(x)|)$ but it does not converge in $(\mathcal{C}^1([0,1]), max_{x\in[0,1]}|f(x)|)$ since $f \notin \mathcal{C}^1([0,1])$. □

A.8 Proof of Lax-Milgrams Theorem

Theorem formulation: Let $(\mathbb{E}, \langle\cdot,\cdot\rangle)$ be a Hilbert space. Let $\phi : \mathbb{E}\times\mathbb{E} \to \mathbb{R}$ be a bilinear, bounded and coercive functional and let $f : \mathbb{E} \to \mathbb{R}$ be a bounded, linear functional. Then there exists a unique $\widetilde{x}_f \in \mathbb{E} :\ \phi(x, \widetilde{x}_f) = f(x)\ \forall x \in \mathbb{E}$.

Proof. Fix $y \in \mathbb{E}$. Define a mapping y_ϕ as

$$\mathbb{E} \ni x \overset{y_\phi}{\mapsto} \phi(x, y) \in \mathbb{R}. \tag{A.32}$$

$y_\phi : \mathbb{E} \to \mathbb{R}$ is a linear operator since

$$y_\phi(\alpha x + \beta z) = \phi(\alpha x + \beta z, y) = \alpha\phi(x, y) + \beta\phi(z, y) = \alpha y_\phi(x) + \beta y_\phi(z),\ \forall\alpha, \beta \in \mathbb{R},\ \forall x, y, z \in \mathbb{E}. \tag{A.33}$$

$y_\phi : \mathbb{E} \to \mathbb{R}$ is also a bounded operator since

$$|y_\phi(x)| = |\phi(x, y)| \le M\|y\|\|x\| = \widetilde{M}\|x\|,\ \text{some constants } M,\ \widetilde{M} < \infty,\ \forall x, y \in \mathbb{E}. \tag{A.34}$$

Hence y_ϕ is a bounded linear operator. Riesz Representation Theorem gives that there exists a unique $A(y) \in \mathbb{E}$, such that $y_\phi = \langle a, A(y)\rangle$ for all $x \in \mathbb{E}$. Now show that $A : \mathbb{E} \to \mathbb{E}$ is a bounded linear mapping. Consider

$$\begin{aligned}\langle x, A(\alpha y + \beta z)\rangle = \phi(x, \alpha y + \beta z) = \bar{\alpha}\phi(x, y) + \bar{\beta}\phi(x, z) = \\ = \bar{\alpha}\langle x, A(y)\rangle + \bar{\beta}\langle x, A(z)\rangle = \langle x, \alpha A(y)\rangle + \langle x, \beta A(z)\rangle.\end{aligned} \tag{A.35}$$

Take the left hand side subtracted with the right hand side gives

$$\langle x, A(\alpha y + \beta z) - \alpha A(y) - \beta A(z)\rangle = 0. \tag{A.36}$$

Set $x = A(\alpha y + \beta z) - \alpha A(y) - \beta A(z)\rangle$ implies that

$$\langle A(\alpha y + \beta z) - \alpha A(y) - \beta A(z), A(\alpha y + \beta z) - \alpha A(y) - \beta A(z)\rangle = \|A(\alpha y + \beta z) - \alpha A(y) - \beta A(z)\|^2 = 0 \tag{A.37}$$

which gives

$$A(\alpha y + \beta z) - \alpha A(y) - \beta A(z) = 0 \quad \Leftrightarrow \quad A(\alpha y + \beta z) = \alpha A(y) + \beta A(z) \tag{A.38}$$

and hence A is linear. Prove that A is bounded by considering

$$|\langle x, A(y)\rangle| = |\phi(x, y)| \le M\|x\|\|y\|, \text{some } M < \infty. \tag{A.39}$$

Set $x = A(y)$ implies that

$$|\langle A(y), A(y)\rangle| = \|A(y)\|^2 \le M\|A(y)\|\|y\| \tag{A.40}$$

which gives

$$\|A(y)\| \le M\|y\| \tag{A.41}$$

and hence A is bounded. In order to fulfill the proof by using Riesz Representation theorem once again, it needs to be proven that A is a $1-1$ and onto. Start by showing that A is $1-1$, i.e. $A(x_1) = A(x_2) \ \Rightarrow\ x_1 = x_2$. In order to prove this, use the coercivity of ϕ, i.e.

$$\phi(x, x) \ge k\|x\|^2, \quad \forall x \in \mathbb{E}, \quad \text{some } k > 0. \tag{A.42}$$

Hence

$$k\|x\|^2 \leq \langle x, A(x)\rangle \leq \|x\|\|A(x)\| \tag{A.43}$$

which gives that

$$\|x\| \leq \frac{1}{k}\|A(x)\|. \tag{A.44}$$

If $A(x_1) = A(x_2)$ the equation $A(x_1 - x_2) = \mathbf{0}$ must hold. The $1-1$ feature now follows since

$$\|x_1 - x_2\| \leq \frac{1}{k}\|A(x_1 - x_2)\| = \frac{1}{k}\|A(x_1) - A(x_2)\| = \mathbf{0}, \tag{A.45}$$

which gives $x_1 = x_2$.

Now show that A is onto, i.e. $\mathcal{R}(A) := \{A(x) : \ x \in \mathbb{E}\} \stackrel{!}{=} \mathbb{E}$. In order to prove this, first show that $\mathcal{R}(A)$ is a closed subset in $\mathbb{E}$. Pick $y_1, y_2 \in \mathcal{R}(A)$ and let α_1, α_2 be two scalars. Since $y_1, y_2 \in \mathcal{R}(A)$ it follows that there exists $x_1, x_2 \in \mathbb{E}$ such that $y_1 = A(x_1)$ and $y_2 = A(x_2)$. It also follows that $\alpha_1 y_1 + \alpha_2 y_2 \in \mathcal{R}(A)$ since

$$\alpha_1 y_1 + \alpha_2 y_2 = \alpha_1 A(x_1) + \alpha_2 A(x_2) = A(\alpha_1 x_1 + \alpha_2 x_2) \in \mathcal{R}(A). \tag{A.46}$$

By using this fact over and over again it is possible to add more and more y_n in the sequence and they will always lie in the range of A. In particular, it is possible to define a sequence $(y_n)_{n=1}^{\infty} \in \mathcal{R}(A)$. It is now proven that $\mathcal{R}(A)$ is closed if

$$y_n \to y \text{ in } \mathbb{E} \ \Rightarrow \ y \in \mathcal{R}(A). \tag{A.47}$$

Pick $(x_n)_{n=1}^{\infty} \in \mathbb{E}$ such that $A(x_n) = y_n$ for all n. By using equation (A.44) it follows that

$$\|x_n - x_m\| \leq \frac{1}{k}\|A(x_n) - A(x_m)\| = \frac{1}{k}\|y_n - y_m)\| \to 0, \quad n, m \to \infty. \tag{A.48}$$

Hence $(x_n)_{n=1}^{\infty}$ is a Cauchy sequence in $(\mathbb{E}, \|\cdot\|)$. The fact that $(\mathbb{E}, \|\cdot\|)$ is a Banach space implies that $(x_n)_{n=1}^{\infty}$ converges. Call this limit x. Since $x_n \to x$, together with the fact that A is continuous (this follows from that A is bounded and linear), it follow that

$$A(x_n) \to A(x), \quad \Leftrightarrow \quad y_n \to y \text{ in } \mathbb{E} \tag{A.49}$$

and hence $y \in \mathcal{R}(A)$ which implies that $\mathcal{R}(A)$ is a closed subset in $\mathbb{E}$. From the closeness of $\mathcal{R}(A)$ the Orthogonal Projection Theorem can be applied. Assume that $\mathcal{R}(A) \neq \mathbb{E}$. The Orthogonal Projection Theorem implies that it is possible to write $\mathbb{E}$ as the following decomposition:

$$\mathbb{E} = \mathcal{R}(A) \oplus \mathcal{R}(A)^{\perp}. \tag{A.50}$$

Now pick $z \in \mathcal{R}(A)\backslash\{0\}$. From the decomposition given by the Orthogonal Projection Theorem the identity

$$\langle z, A(x)\rangle = 0 \tag{A.51}$$

must hold for all $x \in \mathbb{E}$. In particular $\langle z, A(z)\rangle = 0$. The coercivity of ϕ gives a contradiction since

$$k\|z\|^2 \leq \phi(z, z) = \langle z, A(z)\rangle = 0, \quad \Rightarrow \quad z = \mathbf{0}. \tag{A.52}$$

Hence the assumption $\mathcal{R}(A) \neq \mathbb{E}$ cannot hold which implies that A is onto. Riesz Representation Theorem gives that there exists a unique $x_f \in \mathbb{E}$ such that

$$f(x) = \langle x, x_f\rangle, \quad \forall x \in \mathbb{E}. \tag{A.53}$$

Since A is $1-1$ and onto there exists a unique $\widetilde{x}_f \in \mathbb{E}$ such that $x_f = A(\widetilde{x}_f)$. This implies that

$$\langle x, x_f \rangle = \langle x, A(\widetilde{x}_f) \rangle = \phi(x, \widetilde{x}_f), \quad \forall x \in \mathbb{E}. \tag{A.54}$$

Equation (A.53) and Equation (A.54) give that there exists a unique $\widetilde{x}_f \in \mathbb{E}$ such that

$$f(x) = \phi(x, \widetilde{x}_f), \quad \forall x \in \mathbb{E}. \tag{A.55}$$

□

Helping theorems used in the proof of Lax-Milgrams Theorem are the Riesz Representation Theorem and the Orthogonal Projection Theorem which are formulated as:

Theorem 9 (Riesz Representation Theorem)**.** *Let $(\mathbb{E}, \langle \cdot, \cdot \rangle)$ be a Hilbert space and let $f : \mathbb{E} \to \mathbb{C}$ be a bounded linear functional. Then there exists a unique $x_f \in \mathbb{E}$ such that $f(x) = \langle x, x_f \rangle$ for all $x \in \mathbb{E}$.*

Theorem 10 (Orthogonal Projection Theorem)**.** *Let $(\mathbb{E}, \langle \cdot, \cdot \rangle)$ be a Hilbert space and let $\mathbb{S}$ be a closed subset in $\mathbb{E}$. Then $\mathbb{E} = \mathbb{S} \oplus \mathbb{S}^{\perp}$, i.e.*

$$\forall x \in \mathbb{E},\ \exists!\, y \in \mathbb{S},\ \exists!\, z \in \mathbb{S}^{\perp} :\ x = y + z. \tag{A.56}$$

B Additional numerical results from the error analysis

In this chapter additional figures from the error analysis are presented. Section B.1 presents results for the energy norm. Sections B.2, B.3 and B.4 presents additional numerical results when the quantity of interest is set to the average temperature, the average heat flux in one direction and the final temperature. Both for the energy norm and the quantities of interest, the numerical results for the different load cases are similar. Therefore, the description and interpretations of the figures is this chapter are only brief. For more detailed descriptions and interpretations, see Sections 5.4 and 5.5.

B.1 Energy norm

In this section the numerical results for the energy norm when the second load case is applied is presented. For more information about the similar figures that are produced for the first load case, see Section 5.4. Figure B.1 shows the quantity $\eta - 1$, which is defined as

$$\eta - 1 := \frac{\|e^s\|}{\|e\|} - 1, \tag{B.1}$$

as a function of the number of used modes. From the figure it can be seen that $\eta - 1$ is always positive and therefore the inequality $\|e^s\| \geq \|e\|$ holds.

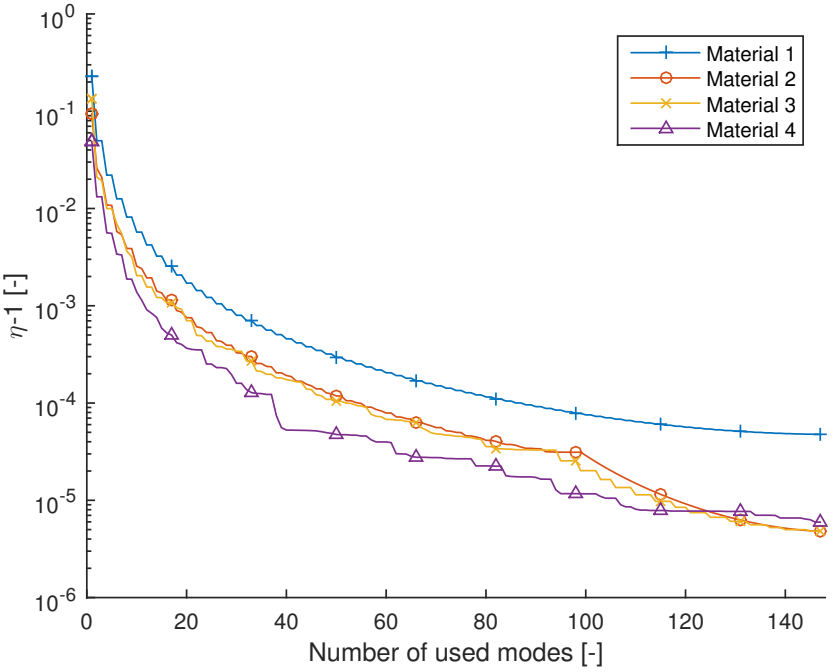


Figure B.1: *The figure shows the quantity $\eta - 1$ for the energy norm as a function of the number of used modes for the different materials.*

Figure B.2 shows the error estimate that are performed. From the figure it can be seen that $\|e^s\|$ and $\|e\|$ are close to each other and that both $\|e_{est}\|$ and $\|e_{est,proj}\|$ give an overestimate of the results. Figure B.3 shows the effectivity index as a function for the different error estimates, defined as

$$\eta^s := \frac{\|e^s\|}{\|e\|}, \quad \eta_{est} := \frac{\|e_{est}\|}{\|e\|} \quad \text{and} \quad \eta_{est,proj} := \frac{\|e_{est,proj}\|}{\|e\|}. \tag{B.2}$$

For more information about the figures, see Figures 5.16 and 5.17 which are the corresponding figures for the other load case.

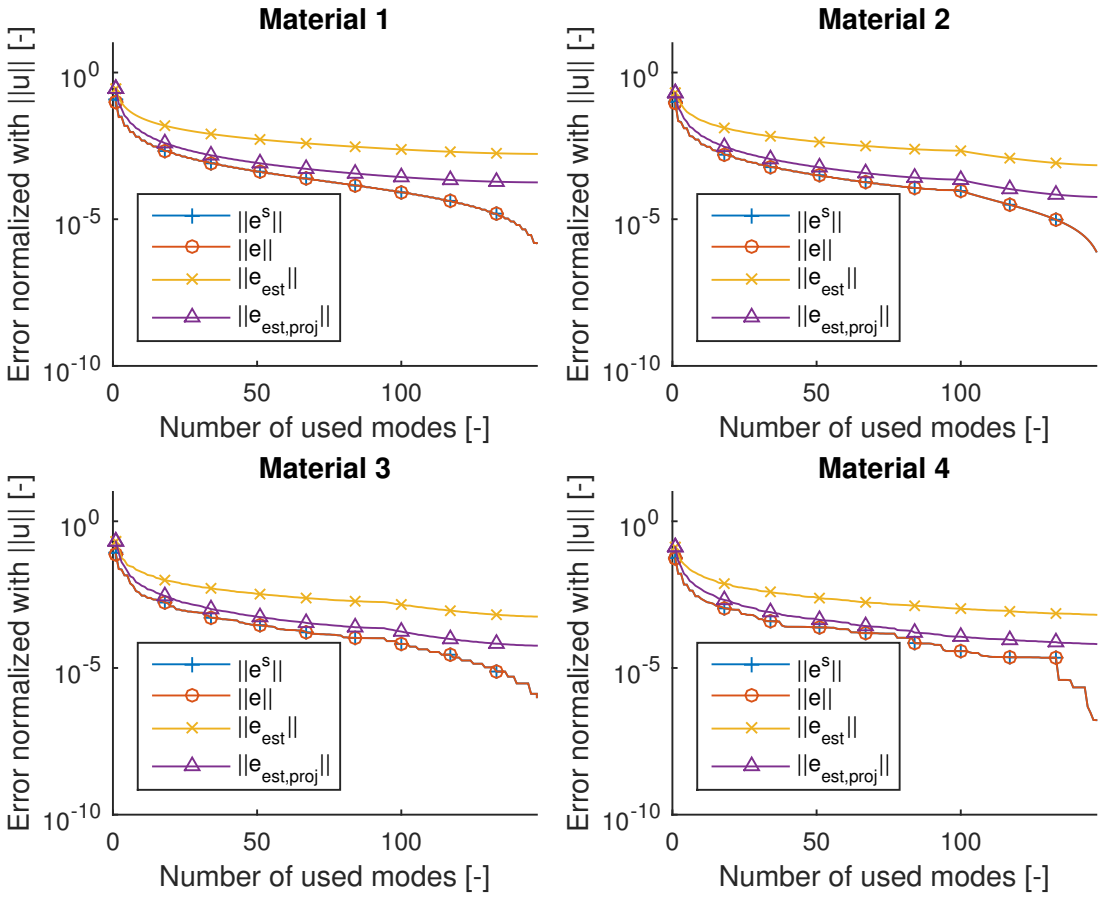


Figure B.2: *The figure shows the true error and error estimates for the energy norm as a function of the number of used modes for the different materials.*

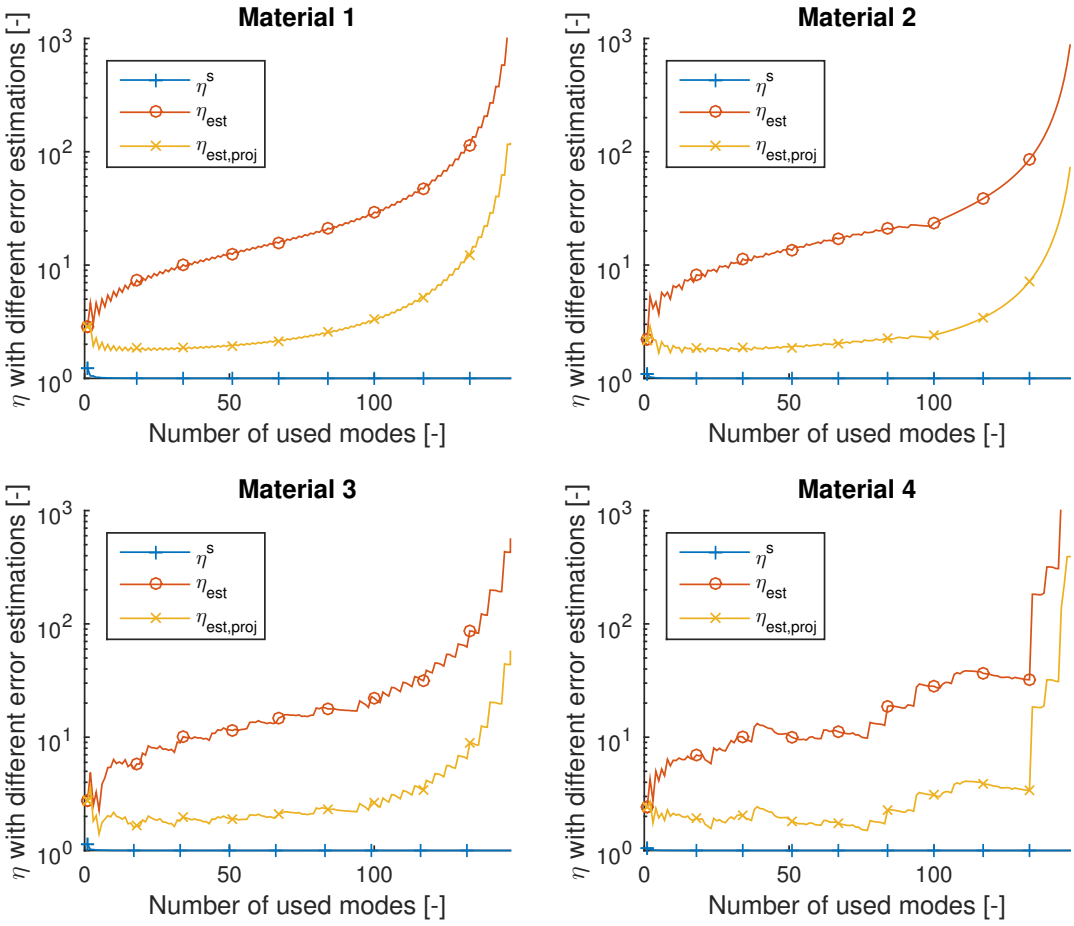


Figure B.3: *The figure shows the effectivity index for the energy norm as a function of the number of used modes for the different material.*

B.2 Average temperature as quantity of interest

In this section, the numerical results when the average temperature is the quantity of interest are presented for the second load case. Figure B.4 shows the true error and the error estimates when all modes are used. In the figure it can be seen that Theorem 2 induces a large overestimate since the Galerkin orthogonality is not used. Figure B.5 shows the results when Theorem 3 is applied and Figure B.6 shows the results when Theorem 4 is applied. Figure B.7 shows the upper and lower limit of the error that can be determined with Theorem 4. For information about how the figure can be interpreted, see Section 5.5.1 where similar figures are presented for the first load case.

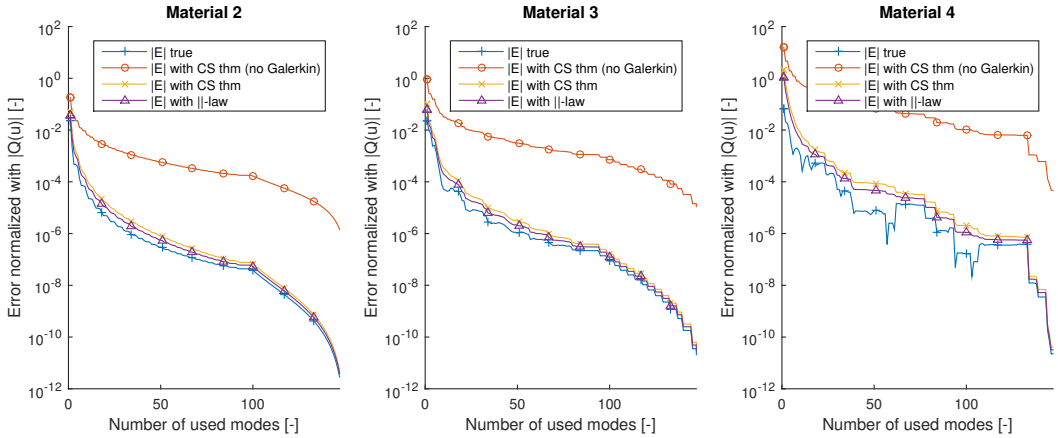


Figure B.4: *The figure shows the true error and error estimates that are performed on the different materials with Theorems 2, 3 and 4.*

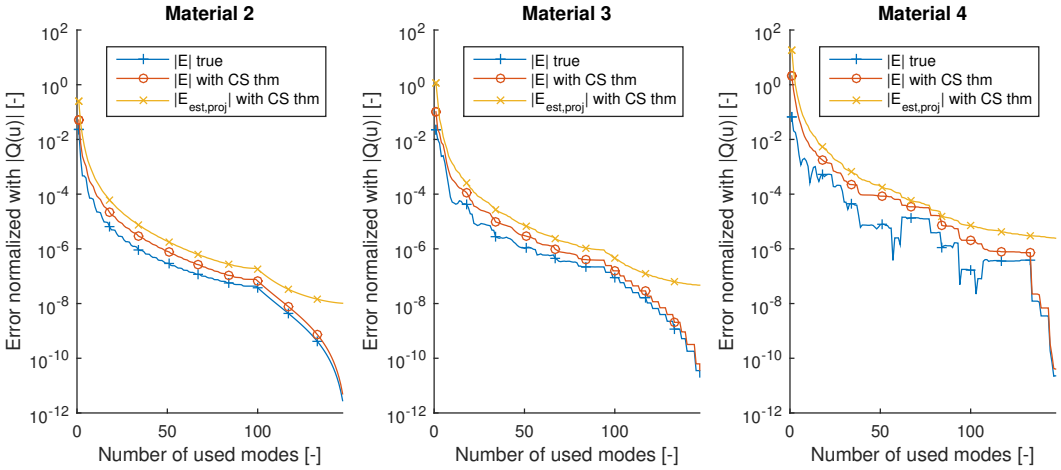


Figure B.5: *The figure shows the true error and error estimates for the different materials when Theorem 3 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*

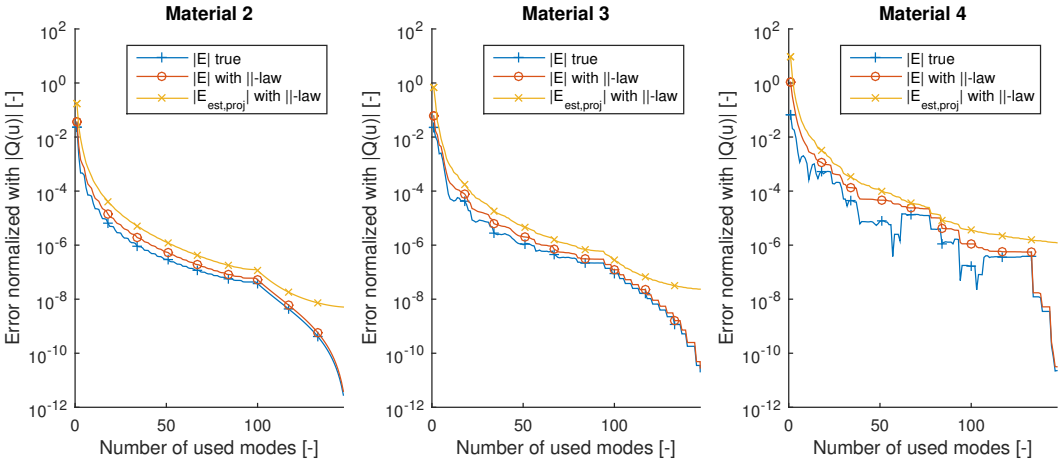


Figure B.6: *The figure shows the true error and error estimates for the different materials when Theorem 4 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*

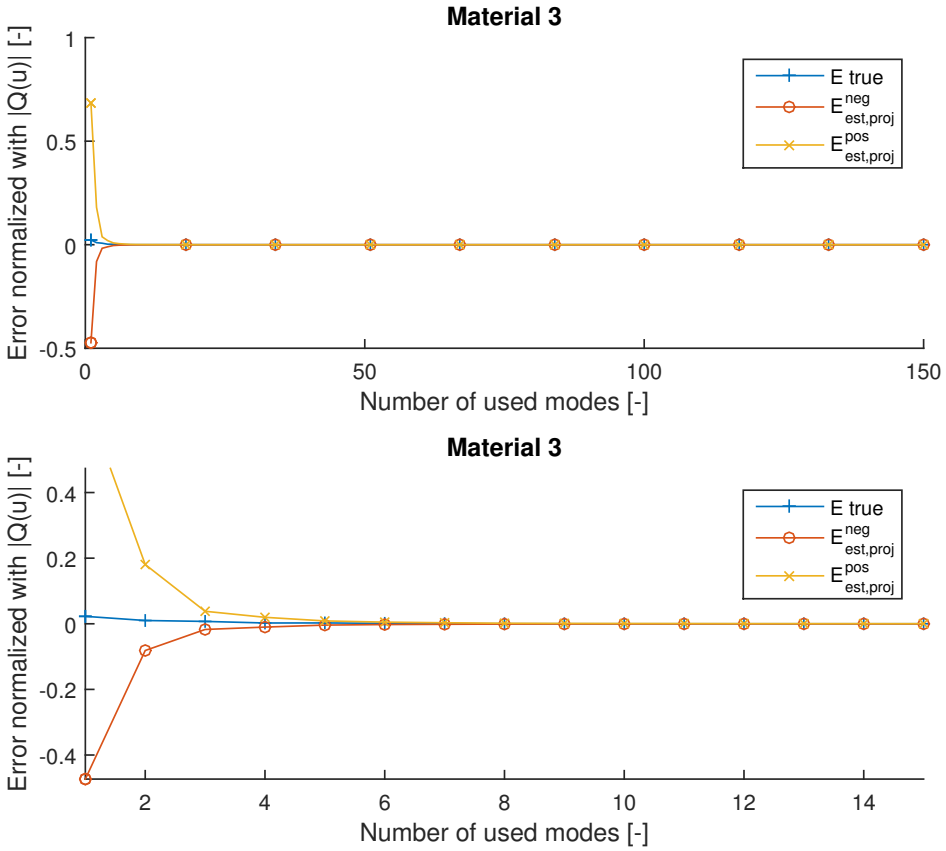


Figure B.7: *The figure shows the true error and its upper and lower limits when Theorem 4 is applied.*

B.3 Average heat flux in one direction as quantity of interest

In this section, the numerical results when the average heat flux in one direction is the quantity of interest are presented for the first load case. Figure B.8 shows the true error and the error estimates when all modes are used. In the figure it can be seen that Theorem 2 induces a large overestimate since the Galerkin orthogonality is not used. Figure B.9 shows the results when Theorem 3 is applied and Figure B.10 shows the results when Theorem 4 is applied. Finally, Figure B.11 shows the upper lower limit of the error that can be determined with Theorem 4. For information about how the figure can be interpreted, see Section 5.5.2 where similar figures are presented for the second load case.

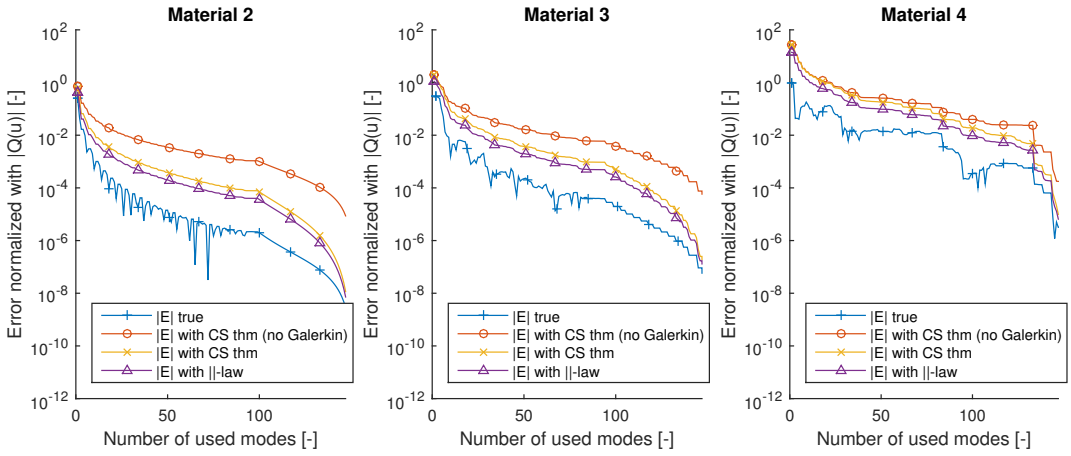


Figure B.8: *The figure shows the true error and error estimates that are performed on the different materials with Theorems 2, 3 and 4.*

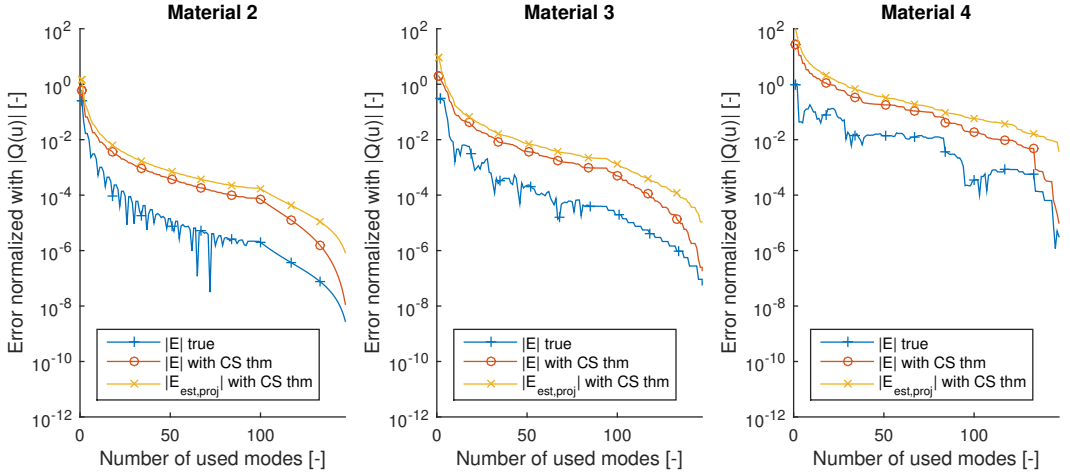


Figure B.9: *The figure shows the true error and error estimates for the different materials when Theorem 3 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*

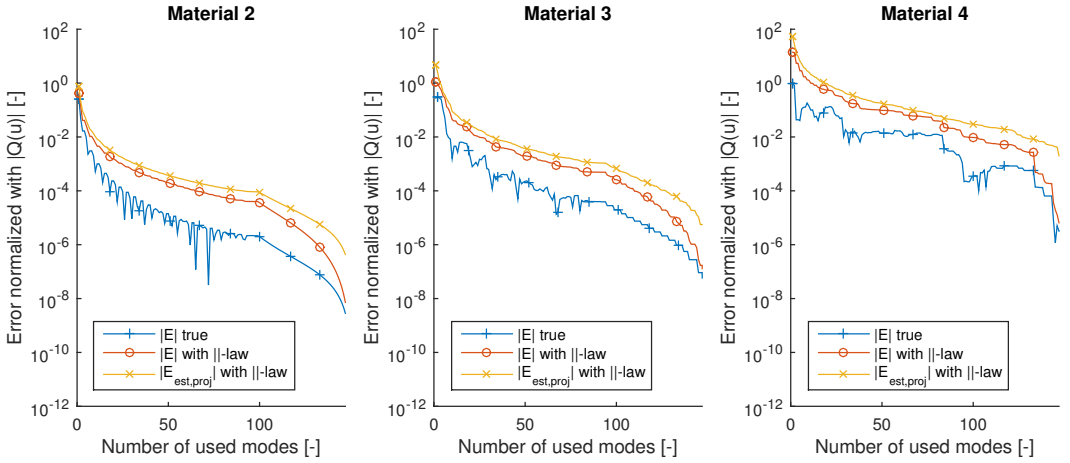


Figure B.10: *The figure shows the true error and error estimates for the different materials when Theorem 4 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*

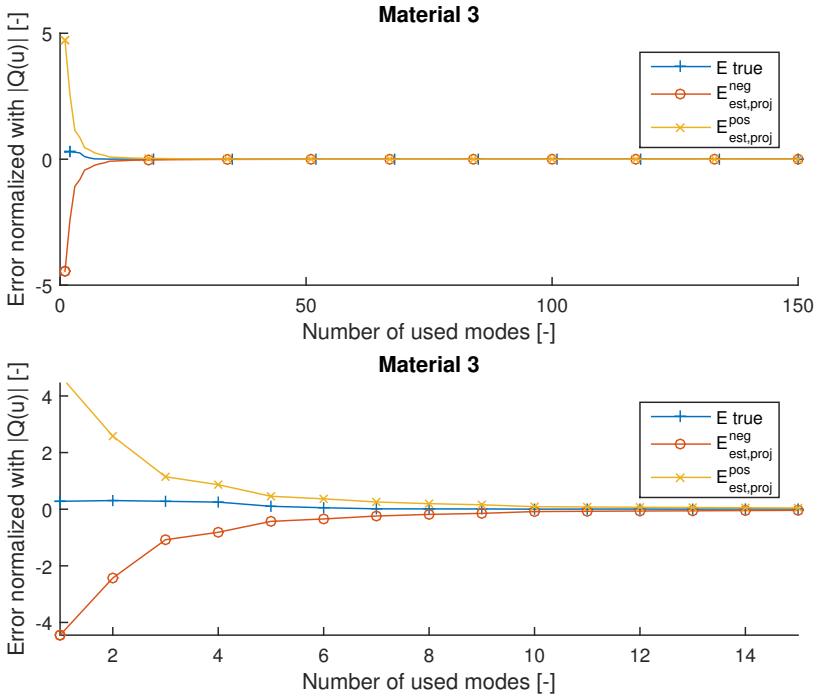


Figure B.11: *The figure shows the true error and its upper and lower limits when Theorem 4 is applied.*

B.4 Final temperature as quantity of interest

In this section, the numerical results when the final temperature is the quantity of interest are presented for the second load case. Figure B.12 shows the true error and error estimates that are using all modes. As it can be seen in the figure, the true error is extremely small. This depends on that $\bar{g}$ is constant for the last two-third if the simulation time. Therefore, the load case is modified such that $\bar{g}$ is a ramp during the whole simulation. The results for this modified load case can be seen in Figure B.13. Figure B.14 and B.15 shows the numerical results when Theorems 3 and 4 are applied. For information about how the figure can be interpreted, see Section 5.5.3 where the similar figures are presented for the first load case.

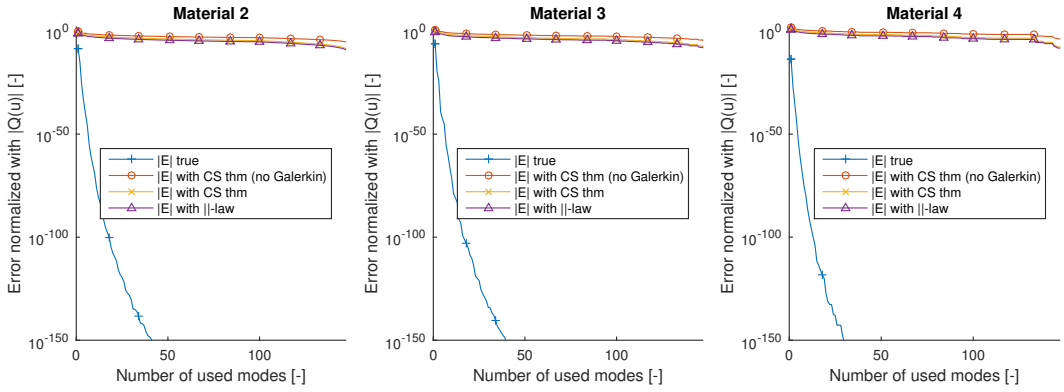


Figure B.12: *The figure shows the true error and the error estimates that are using all modes when the final temperature is the quantity of interest. The results are produced with the second load case.*

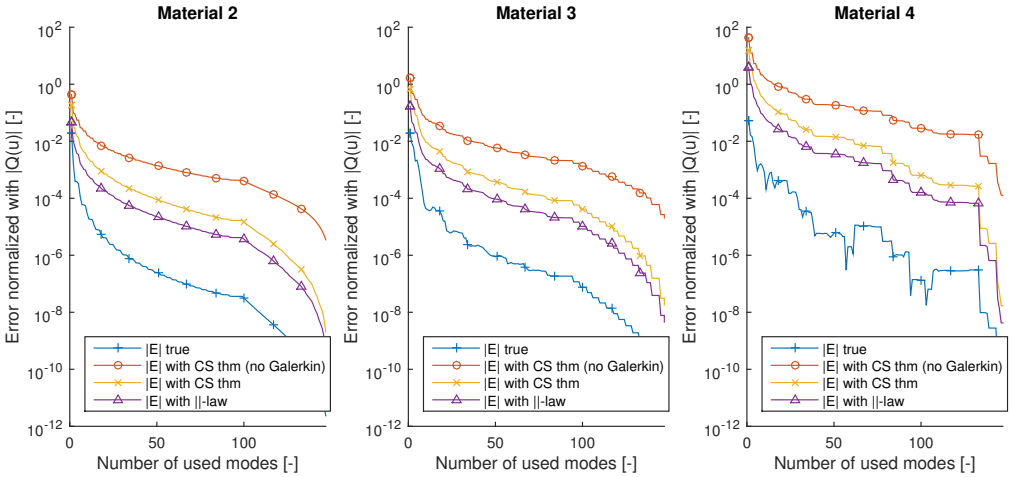


Figure B.13: *The figure shows the same results as Figure B.12 with the difference that the load case is changed. In this case $\bar{g}$ is a ramp during the whole simulation.*

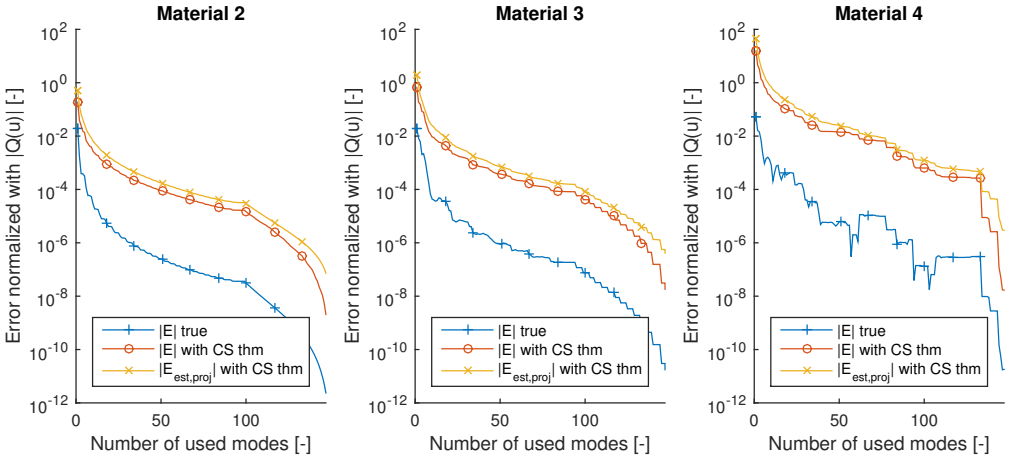


Figure B.14: *The figure shows the true error and error estimates for the different materials when Theorem 3 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*

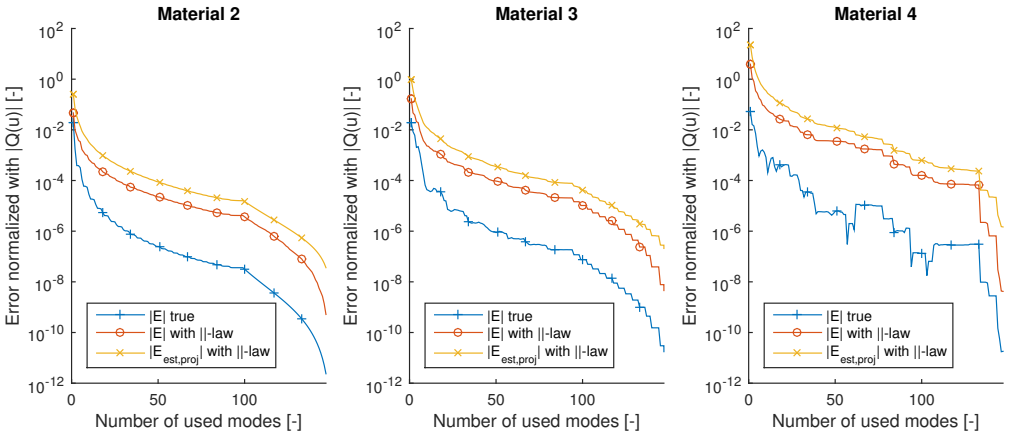


Figure B.15: *The figure shows the true error and error estimates for the different materials when Theorem 4 is applied. Note that $|E_{est,proj}|$ uses only the reduced modes.*