Pseudospectra and Linearization Techniques of Rational Eigenvalue Problems

Axel Torshage

May 2013

Masters’s thesis 30 Credits
Umeå University
Supervisor - Christian Engström
Abstract

This thesis concerns the analysis and sensitivity of nonlinear eigenvalue problems for matrices and linear operators. The first part illustrates that lack of normality may result in catastrophic ill-conditioned eigenvalue problem. Linearization of rational eigenvalue problems for both operators over finite and infinite dimensional spaces are considered. The standard approach is to multiply by the least common denominator in the rational term and apply a well known linearization technique to the polynomial eigenvalue problem. However, the symmetry of the original problem is lost, which may result in a more ill-conditioned problem. In this thesis, an alternative linearization method is used and the sensitivity of the two different linearizations are studied. Moreover, this work contains numerically solved rational eigenvalue problems with applications in photonic crystals. For these examples the pseudospectra is used to show how well-conditioned the problems are which indicates whether the solutions are reliable or not.
# CONTENTS

1 Introduction 3
   1.1 History of spectra and pseudospectra .......................... 7

2 Normal and almost normal matrices 9
   2.1 Perturbation of normal Matrix ................................ 13

3 Linearization of Eigenvalue problems 16
   3.1 Properties of the finite dimension case ......................... 18
   3.2 Finite and infinite physical system ............................ 23
      3.2.1 Non-damped polynomial method ........................... 24
      3.2.2 Non-damped rational method ............................. 25
      3.2.3 Damped case ............................................ 27
      3.2.4 Damped polynomial method ............................... 28
      3.2.5 Damped rational method ................................. 29
   3.3 Photonic Crystals ............................................. 36
      3.3.1 Hilbert space ........................................... 37
      3.3.2 TE-waves ............................................... 40
      3.3.3 TM-waves ............................................... 46
      3.3.4 Properties of TM and TE waves ........................... 49

4 Numerical examples of the TM-wave 54
   4.1 Discontinuous Galerkin ....................................... 55
   4.2 Eigenvalues of TM-waves ..................................... 56
   4.3 Pseudospectra of TM-waves .................................. 59

5 Conclusions 65

6 Appendix 68
   6.1 Proofs ..................................................... 68
6.1.1 Proof of Theorem 27 ................................. 68
6.1.2 Proof of Lemma 28 ................................. 71
6.1.3 Proof of Lemma 33 ................................. 73
6.1.4 Proof of Lemma 43 ................................. 76
6.2 Matlab code ........................................ 80
   6.2.1 TM_Poly_linearization_No_damping.m .......... 80
   6.2.2 TM_Poly_linearization_damping.m .......... 83
   6.2.3 TM_Rat_linearization_No_damping.m .......... 86
   6.2.4 TM_Rat_linearization_damping.m .......... 88
   6.2.5 function [L,lambda,s] = lowerbound(A) .......... 90
   6.2.6 function U = upperbound(A) ................. 91
1. Introduction

This thesis addresses computational aspects of different ways of solving eigenvalue problems. The study has two major parts: One part concerning linearization of non-linear eigenvalue problems and one concerning pseudospectrum on non-normal matrices and operators obtained from the linearizations. Linearization is here used with a different meaning than it usually has, i.e. approximating the problem by a similar linear or piecewise linear function. By linearization we mean rewriting a problem, in our case an eigenvalue problem, into an equivalent linear problem. It is immediate to see the use of linearization as we know that many non-linear problems are hard to solve. The use and definition of pseudospectra is harder to understand, so we have to start by the spectra.

Definition 1. The spectra or spectrum of a matrix $A \in \mathbb{C}^{n \times n}$ is the same as the set of eigenvalues, defined by

$$\sigma(A) = \{ \lambda \in \mathbb{C} : \det(A - \lambda I) = 0 \}.$$  

From this definition it’s apparent that the spectra is somewhat binary as it only says for what values $\lambda$ we have that $\det(A - \lambda I) = 0$, and where it is non-zero, it may not give a good description of the matrix behavior as completely different matrices can share eigenvalues. The spectra doesn’t tell us anything of how close to zero $\det(A - \lambda I)$ is for a given $\lambda \in \mathbb{C}$. We can have that $A - \lambda I$ is so close to singular that it is hard to work with, or that $\lambda$ easily mistaken for an eigenvalue, even though it is not. For example: If $A$ is non-singular the inverse of $A$, can be computed theoretically, still if $A$ is close to singular the computed $A^{-1}$ might be completely wrong as computational errors easily arise for such matrices. The idea of pseudospectra is simply to compute $|\det(A - \lambda I)|$ and see when it is large and thus be able to expect bad behavior of the matrix and hence also get information on how it behaves in general. Instead of just invertible or not invertible $A - \lambda I$ could be described by more or less invertible. The formal definition of the pseudospectrum or rather $\epsilon-$pseudospectrum is given below.
Definition 2. Let $A$ be a matrix in $\mathbb{C}^{n \times n}$, and $\|\cdot\|$ the given matrix norm. Then the $\epsilon$-pseudospectrum $\sigma_\epsilon(A)$, is the set of $\lambda \in \mathbb{C}$ such that:

$$\|(A - \lambda I)^{-1}\| > \epsilon^{-1}$$

or equivalently the set of all $\lambda \in \mathbb{C}$ such that

$$\lambda \in \sigma(T + E) \text{ for some matrix } E \in \mathbb{C}^{n \times n} \text{ with } \|E\| < \epsilon.$$ 

Both these definitions are equivalent, using the notation $\|(A - \lambda I)^{-1}\| = \infty$ meaning that $\lambda \in \sigma(A)$, [3, page 31]. It can be said that the pseudospectra is defined in a corresponding way for bounded operators over Hilbert spaces, we won’t, however, define this properly as the problems with more general operators will be approximated by matrices before computing the pseudospectra. To be able to express the necessary for pseudospectrum we will start by a simple example that shows that the spectrum is not always enough. Suppose we want to compute the spectrum of the $7 \times 7$ matrix, known as the Godunov matrix, with entries as:

$$A = \begin{bmatrix}
289 & 2064 & 336 & 128 & 80 & 32 & 16 \\
1152 & 30 & 1312 & 512 & 288 & 128 & 32 \\
-29 & -2000 & 756 & 384 & 1008 & 224 & 48 \\
512 & 128 & 640 & 0 & 640 & 512 & 128 \\
1053 & 2256 & -504 & -384 & -756 & 800 & 208 \\
-287 & -16 & 1712 & -128 & 1968 & -30 & 2032 \\
-2176 & -287 & -1565 & -512 & -541 & -1152 & -289 \\
\end{bmatrix}. \quad (1.1)$$

Computing this by hand will give us a seventh-degree polynomial equation so it is probably not possible to find the eigenvalues analytically, thus it serves as a test case for numerical method, for example matlab’s eig-function, the QZ algorithm from LAPACK [1]. Applying this function to $A$ give the numerical eigenvalues

$$\lambda = \begin{cases}
-3.9411 + 1.0302i \\
-3.9411 - 1.0302i \\
4.2958 \\
-0.7563 + 2.6110i \\
-0.7563 - 2.6110i \\
2.5494 + 1.7181i \\
2.5494 - 1.7181i
\end{cases}.$$

As can be seen the imaginary parts are in most of the numerically approximated eigenvalues notably nonzero. It’s possible to prove that, the actual eigenvalues
Figure 1.1: All seven solutions of the characteristic polynomial can be seen in this plot.
Figure 1.2: The level curves of $|\det (A - \lambda I)|$ in the complex region close to the origin. The real eigenvalues are $0, \pm 1, \pm 2, \pm 4$ and this can be seen in the characteristic polynomial, depicted in Figure 1.1.

Now this only says that matlab’s eig-function is not suitable to solve this problem, even though it gives a hint that it might be hard to approximate the eigenvalues and get an accurate result. We will get back to this example in the following chapters but, the reason to why this problem occur is connected to the fact that $A$ has a high departure from normality. Non-normal matrices are in some cases very sensitive to perturbation and (1.1) is constructed to illustrate that this can cause numerical methods to fail. How could we now have been able to foreseen that $A$’s eigenvalues might have been hard to numerically approximate? The answer lies (partly) in the pseudospectra. As can be seen in Figure 1.2
the matrix $A - \lambda I$ has a determinant that is very close to 0 for $\lambda$ in a large area everywhere near the eigenvalues and thus there is a large subset of the complex plane where the eigenvalues might be located. The lines denote in what area $A - \lambda I$ has a determinant that has absolute value less than $10^x$, where $x$ is the given number in the sidebar. We can see that inside the seven headed area we have that $\det (A - \lambda I) < 10^{-13}$ and thus very close to 0. It’s also worth noting that the eigenvalues computed by matlab’s eig-function are not similar to those indicated in the picture of pseudospectra, so depending on what method used we get different numerically computed eigenvalues. One might say that the pseudospectra provides an idea of how trustworthy eigenvalues and eigenvectors are when computed numerically.

1.1. History of spectra and pseudospectra

The history given here is mainly a summary of [3, page 6-7 and 41-42]. Much of the theory concerning eigenvalues and eigenfunctions or eigenvectors origins in the nineteenth century where mathematicians like Joseph Fourier and Siméon Denis Poisson provided the basis of the subject, later to be improved by the work of David Hilbert, Erik Ivar Fredholm and other mathematicians in the early twentieth century. The early concern about eigenvalue problems were closely connected to eigenvalues of functions, differential equations and vibration problems. The use of spectral theory was then increased as it was shown that quantum mechanics could be explained by eigenvectors and eigenfunctions. The eigenvalues showed to be of great use as it led to diagonalization of matrices and describing behavior of vibrations of waves in a certain medium. The frequency near an eigenvalue will result in a resonance, that’s how a bell can get the desired ring. The latter will be of interest in this thesis as we’ll work on problems inherited from light waves equations. Even though definitions of a mathematical object defined similarly as the pseudospectra can be dated back to at least 1967 by an unpublished thesis by J. M. Varah [4, page 41], the general interest for the subject started as computers became increasingly important for computing eigenvalue problem. As we know computers rely much on approximations and eigenvalue computations are no exception, some method to be able to find how reliable the approximated eigenvalues were, was needed. Thus the interest for pseudospectra arose and the year 1992 [3, page 46] is noted as the year when the idea of pseudospectra were formally defined as it is today. Pseudospectra has since then been used to determine how well-conditioned non-normal eigenvalue problems are in both theoretical and
physical interpretations, some of which we will see later.
2. **Normal and almost normal matrices**

This chapter describe ways of discussing normality and non-normality of matrices without using pseudospectra as a tool. We will notice that non-normal eigenvalue problems will be hard to solve numerically. In this chapter $A \in \mathbb{C}^{n \times n}$ where $n \in \mathbb{N}$ if nothing else is indicated.

**Definition 3.** The set of singular values to a matrix $A$ is defined as

$$S(A) = \{ \sqrt{\sigma(A^*A)} \}.$$  

**Definition 4.** The spectral norm for matrices $A$ is defined as

$$\|A\|_2 = \sup_{\|x\|=1} (\|Ax\|_2).$$

where $x \in \mathbb{C}^n$ and $\|Ax\|_2$ is the standard metric for vectors of length $n$.

**Definition 5.** The Eulerian norm is defined as

$$\|A\|_e = \left( \sum s_i^2 \right)^{\frac{1}{2}},$$

where $s_i$ denotes the singular values of $A$.

**Definition 6.** A matrix $A$ is normal if

$$A^*A = AA^*.$$  

This section concerns an abstract way of determining how "close" to normal a particular matrix is. Any matrix $A$ can be written as

$$A = UTU^*,$$  \hspace{1cm} (2.1)
where $T$ is upper triangular and $U$ is a unitary matrix [6]. In general, for a given $A$, more than one $T$ and $U$ satisfies (2.1). However, in each case we can write

$$T = D + M,$$

where $D$ is the diagonal matrix with the same entries as $T$ in the main diagonal:

$$T = \begin{bmatrix}
d_1 & m_{12} & \cdots & m_{1n} \\
\vdots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & m_{(n-1)n} \\
d_n & & & d_n
\end{bmatrix}.$$

Thus $M$ only have entries different from zero right of the main diagonal. Let $v$ denote some given matrix norm, then the departure of $v-$normality, [7, page 27], of $A$, is defined as

$$\Delta_v (A) = \inf (\|M\|_v), \quad (2.2)$$

where inf is taken over the set of possible $T$. For a normal matrix $A$, $\Delta_v (A) = 0$, since $A$ is normal it is unitary diagonalizable due to the spectral theorem. Thus we have that $M = 0$ for $T = D$ and $\inf (\|M\|) = 0$. To be able to see how the condition of the problem depends on the departure of $v-$normality gives, one can use following theorem:

**Definition 7.** Given two norms $\|\cdot\|_v$ and $\|\cdot\|_w$ working on a vectorspace $V$, we say that $v$ majorizes $w$ if

$$\|u\|_v \geq \|u\|_w \quad \text{for all } u \in V.$$

**Theorem 8.** [7, page 32] Suppose $v$ is a norm that majorizes the spectral norm and $A \in \mathbb{C}^{n \times n}$, then for all matrices $E \in \mathbb{C}^{n \times n}$,

$$\max_{1 \leq i \leq n} \left( \min_{1 \leq j \leq n} (|\lambda_i - \mu_j|) \right) \leq \frac{y}{g} \|E\|_v, \quad (2.3)$$

where $y = \Delta_v (A) / \|E\|_v$ and $g$ is the unique positive solution to

$$g^1 + g^2 + \ldots + g^n = y. \quad (2.4)$$

Furthermore $\{\mu\}_{i=1}^n$ is the set of eigenvalues to the perturbed matrix $A + E$ and $\{\lambda_i\}_{i=1}^n$ is the set of eigenvalues to $A$. 
Theorem 9. [7, page 32] For an arbitrary matrix $A \in \mathbb{R}^{n \times n}$ we have that
\[
\Delta_2(A) \leq \Delta_e(A) \leq \sqrt[4]{\frac{n^3 - n}{12}} \sqrt{\|A^* A - A A^*\|_e}, \tag{2.5}
\]

This theorem gives an upper bound for the value of (2.2) with the Eulerian norm for a matrix $A$ thus also for the spectral norm as it is majored by the Eulerian norm, $\|A\|_2 \leq \|A\|_e$ for all $A$. We also have the following theorem:

Theorem 10. [8, page 32] For an arbitrary matrix $A$ we have that
\[
\min_{\pi} \left( \max_{1 \leq i \leq n} \left( |s_i(A) - |\lambda_{\pi(i)}(A)|| \right) \right) \leq \Delta_2(A) \leq \Delta_e(A), \tag{2.6}
\]

where $\pi$ is a labeling of the eigenvalues.

This theorem gives a lower bound for the spectral-norm and thus also for the Eulerian norm as the Eulerian norm majorizes the spectral norm. Combining these results we have a span where we can expect the departure of $v-$normality of $A$ for both the Eulerian and the spectral norm. It can easily be seen that for a normal matrix $A$, both the lower bound (2.6) and the upper bound (2.5) are equal to 0. These bounds are now computed for the Godunov matrix given in (1.1). We are using matlab to compute these values, the complete routines are given in the appendix. The upper bound (2.5) is easy to calculate and even for large matrices due to only low order calculations. The lower bound is somewhat harder to calculate for large dimensions $n$, this is due to the number of different labelling of the eigenvalues. The number of permutations are $n!$, so in our case $7! = 5040$. From (2.5) we get the upper bound in this case
\[
\Delta_e(A) \leq 13148,
\]

and form (2.6) we have the lower bound
\[
4318 \leq \Delta_2(A).
\]

thus we have
\[
4318 \leq \Delta_2(A) \leq \Delta_e(A) \leq 13148. \tag{2.7}
\]

Which is a really large bound for the departure of normality from $A$ just as we suspected after noting the large pseudospectra in Fig 1.2. It’s also important to
note that both these bounds use approximated values of \(s_i\) and the lower bound also use \(\lambda_i\), as these are not known on forehand. Since these bounds states how hard the \(\lambda_i\) are to approximate, they might be somewhat inexact. Especially the lower bound as \(\lambda_i\) might be hard to approximate. With a high lower bound (2.6), however, we cannot guarantee that the approximated \(\lambda_i\) are close to the real values. Now if using (2.3) to see how good bound for the distraction of the spectra a disturbance gives when we have the lower bound of the departure of \(v\)–normality given in (2.7). We can suppose \(\|E\|\) is a very small number and thus \(g(y) \approx y^1/n\) since \(y\) is large. The bound (2.3) gives approximately

\[
\max_{1 \leq i \leq n} \left( \min_{1 \leq j \leq n} (|\lambda_i - \mu_j|) \right) \leq y^{n-1} \|E\|_2.
\]

In this particular case we have from the lower bound of (2.7) that \(y = 4318/\|E\|_2\), and thus approximately

\[
\max_{1 \leq i \leq n} \left( \min_{1 \leq j \leq n} (|\lambda_i - \mu_j|) \right) \leq 1306 \|E\|_2^{1/2}.
\]

for small values \(\|E\|\). This result says that the spectra of \(A\) might change greatly with a small disturbance \(E\), by a factor up to 1306 \(\|E\|_2^{3/2}\), thus we can suspect that the numerically obtained eigenvalues are not to be trusted. In this thesis we will, however, mostly use the pseudospectra to show if an eigenvalue problem is hard to approximate solutions to or if we can trust the values we have obtained, but this is closely related to that normality of a matrix. In [9, page 186] we have an approximated condition number to the computation of eigenvalue \(\lambda\) of a matrix \(A\) as

\[
k = \frac{1}{y^*x},
\]

where \(x\) is the right unit eigenvector and \(y\) is the left unit eigenvector to eigenvalue \(\lambda\), the right eigenvector is the standard eigenvector while the left eigenvector is the eigenvector of \(A^*\). It can easily be seen that for a normal matrix these vectors coincide as \(A = A^*\) and thus \(v = 1\). When \(x\) and \(y\) approach orthogonality, however, the condition number \(k\) grows infinitely large. This method is somewhat flawed as it has the eigenvectors as inputs, generally the exact value of these vectors are not known, and if the approximated eigenvalues are wrong they might generate wrong eigenvectors as well and thus give a wrong value \(k\).
2.1. Perturbation of normal Matrix

Definition 11. A matrix $A \in \mathbb{C}^{n \times n}$ is Hermitian if

$$A = A^*.$$ 

It follows trivially that all Hermitian matrices are normal, thus a matrix is well-conditioned if it is "close" to Hermitian. This section, is dedicated to illustrate how much a disturbance of an Hermitian matrix $A'$ affects an eigenvalue problem’s solutions.

Proposition 12. [9, page 212-213] Given a Hermitian matrix $A' \in \mathbb{C}^{n \times n}$ and a matrix $E \in \mathbb{C}^{n \times n}$ such that $\|E\|_v < \epsilon$, where $v$ is a given matrix norm that majorises the spectral norm. Then for each $\lambda \in \sigma (A)$ we have that $|\text{Im} (\lambda)| < \epsilon$, where $A = A' + E$.

Proof. Let $\lambda \in \sigma (A')$ then we have that $\lambda = \mu + i\nu$ for some $\mu, \nu \in \mathbb{R}$. By normalization there exist some eigenvector $y$ corresponding to $\lambda$ such that $\|y\| = 1$ and thus

$$Ay = \lambda y = (\mu + i\nu) y.$$ 

Multiplying by the conjugate transpose of $y$ gives

$$y^*Ay = (\mu + i\nu) \|y\|^2 = \mu + i\nu. \quad (2.8)$$

We now investigate $y^*A^*$;

$$y^*A^* = (Ay)^* = ((\mu + i\nu) y)^* = (\mu - i\nu) y^*.$$ 

Multiplying by $y$ we have

$$y^*A^*y = (\mu - i\nu) \|y\|^2 = \mu - i\nu. \quad (2.9)$$

Finally by combining (2.8) and (2.9) it follows that

$$y^*Ay - y^*A^*y = y^* (A - A^*) y = 2\nu i,$$

using $A' + E = A$ this equals

$$y^* (A' + E - (A' + E)^*) y = y^* (A' + E - A' - E^*) y = y^* (E - E^*) y.$$
This can be evaluated to

\[ y^* (E - E^*) y = 2\nu i \implies \frac{y^* (E - E^*) y}{2i} = v. \]

Using the definition of Im \((E)\) we have

\[ y^* \text{Im} (E) y = v, \]

the property \(\|y\| = 1\) yields, since \(v\) majorizes the spectral norm:

\[ |v| = |y^* \text{Im} (E) y| \leq \|\text{Im} (E)\|_v < \epsilon. \]

\[ \square \]

**Definition 13.** A matrix \(A\) is skew-Hermitian if

\[-A = A^*.\]

It follows that any skew-Hermitian matrix is normal as well, the use of their normality, however, will not be investigated in this thesis. The following proposition is new.

**Proposition 14.** Given a Hermitian matrix \(A' \in \mathbb{C}^{n \times n}\) and a skew-Hermitian matrix \(E \in \mathbb{C}^{n \times n}\), Then for each matrix norm \(v\) that majorises the spectral norm; for each \(\lambda \in \sigma (A)\) we have that \(|\text{Re} (\lambda)| \leq \||A'||_v\), where \(A = A' + E\).

**Proof.** Just as in Theorem 12 we can obtain (2.8), (2.9) and from them it follows that

\[ y^* A y + y^* A^* y = y^* (A + A^*) y = 2\mu. \]

Then using \(A' + E = A\) it follows that

\[ y^* (A' + E + (A' + E)^*) y = y^* (2A' + E + E^*) y = 2\mu, \]

since \(E\) is a skew-Hermitian matrix we have

\[ y^* A'y = \mu, \]

which leads to the inequality

\[ |\mu| = |y^* A'y| \leq \|A'||_v. \]
where \( \lambda' \) is the largest eigenvalue to \( A' \) and it’s real since \( A \) is symmetric. ■

Since all matrices \( A \) can be written on the form \( A' + E \) where \( A' \) is Hermitian and \( E \) is skew-Hermitian (14) gives that the real part of a matrix is always less or equal \( \|A'\|_e \) where \( A' \) is the Hermitian part of \( A \). There are some other estimates which we will not prove but barely state as they are all proved in [9, page 212-213] and thus not shown here.

**Theorem 15.** [9, page 212-213] Suppose we have a matrix \( A \in \mathbb{C}^{n\times n} \) such that \( A = A' + E \) where \( A' \in \mathbb{C}^{n\times n} \) is a Hermitian matrix and \( E \in \mathbb{C}^{n\times n} \). Let \( \{\mu_j + \nu_j i\}_{j=1}^n \) denote the set of eigenvalues to \( A \). Then

\[
\sqrt{\sum_{j=1}^n v_j^2} \leq \|\text{Im} (E)\|_e.
\]

**Theorem 16.** [9, page 212-213] Suppose we have a matrix \( A \in \mathbb{C}^{n\times n} \) such that \( A = A' + E \) where \( A' \in \mathbb{C}^{n\times n} \) is a Hermitian matrix and \( E \in \mathbb{C}^{n\times n} \). Let \( \{\mu_j + \nu_j i\}_{j=1}^n \) denote the set of eigenvalues to \( A \) and let \( \{\lambda_j\}_{j=1}^n \) denote the set of eigenvalues to \( A' \). Then, by some ordering of \( \lambda_i \)

\[
\sqrt{\sum_{j=1}^n (\mu_j - \lambda_j)^2} \leq \|\text{Re} (E)\|_e + \sqrt{\|\text{Im} (E)\|_e^2 - \sum_{j=1}^n v_j^2}.
\]

**Theorem 17.** [9, page 212-213] Suppose we have a matrix \( A \in \mathbb{C}^{n\times n} \) such that \( A = A' + E \) where \( A' \in \mathbb{C}^{n\times n} \) is a Hermitian matrix and \( E \in \mathbb{C}^{n\times n} \). Let \( \{\mu_j + \nu_j i\}_{j=1}^n \) denote the set of eigenvalues to \( A \) and let \( \{\lambda_j\}_{j=1}^n \) denote the set of eigenvalues to \( A' \). Then, by some ordering of \( \lambda_i \)

\[
\sqrt{\sum_{j=1}^n |(\mu_j + i\nu) - \lambda_j|^2} \leq \|E\|_e.
\]

These estimates can be of interest if we know how to choose \( E \) to make \( A' \) Hermitian and still have a small norm of \( E \). Then we can either approximate the complex eigenvalues of \( A \) by those of \( A' \) as they are much easier to compute since \( A \) is normal, or just see if the computed eigenvalues of \( A \) is good estimates because if they differ from those of \( A' \) more than \( \|E\|_e \) these theorems states that we must have computed them wrong. The last estimate gives a complete estimate for the change of all eigenvalues, which can be useful as it gives an area around each eigenvalue for the corresponding Hermitian matrix where the eigenvalues of the disturbed matrix can be located.
3. Linearization of Eigenvalue Problems

The most commonly known eigenvalue problem is the linear eigenvalue problem for matrices; for $A \in \mathbb{C}^{n \times n}$ find $\lambda$ such that

$$L(\lambda) u := (A - \lambda I) u = 0,$$  \hspace{1cm} (3.1)

for some $u \in \mathbb{C}^n$. This is a linear eigenvalue problem but it can still be hard to solve as we saw in (1.1). In (3.1) $L(\lambda)$ is a $n \times n$-matrix valued function in variable $\lambda$, as it for each value of $\lambda$ gives $n \times n$-matrix as output

$$L : \mathbb{C} \to \mathbb{C}^{n \times n}.$$

Eigenvalue problems, however, doesn’t have to be linear as (3.1), they can also define matrix valued functions that are not linear. For example $L(\lambda)$ can have quadratic or rational dependance of variable $\lambda$. A large number of quadratic eigenvalue problems can be found in [11], where the following problem is studied: Find scalars $\lambda$ and vectors $x$ and $y$, such that:

$$\begin{align*}
(\lambda^2 M + \lambda C + K) x &= 0, \\
y^* (\lambda^2 M + \lambda C + K) &= 0,
\end{align*}$$  \hspace{1cm} (3.2)

where $M, C, K \in \mathbb{C}^{n \times n}$. This particular example is inherited from describing the resonance phenomena of the Millennium footbridge. One interesting property is that a quadratic eigenvalue problem can have as many as $2n$ eigenvalues instead of the regular $n$ for linear eigenvalue problems. This can be seen from the fact that a second degree polynomial usually has two roots. It might also be worth noting that the matrix valued functions in (3.2) is

$$L(\lambda) := \lambda^2 M + \lambda C + K.$$

This chapter discusses one important class of eigenvalue problem, the rational eigenvalue problem, in different forms and how to linearize in each case. The
problem of rational eigenvalue problems can be found in [10, page 259-260.] and is a connected problem to the quadratic eigenvalue problems due to the fact that one can always obtain polynomial eigenvalue problem from a rational by multiplying by the denominator even though as we will see in this chapter is not generally the best way. First we discuss the finite (matrix case) without and with damping and then the case with operators over infinite dimensional spaces. We will use two different linearizations of the problems and then compare properties of the formulations. Our main linearization is the same for finite dimensional matrices and bounded operators over infinite dimensional spaces. The linearization origins from [10], where the spectral properties of this linearization is investigated. In all linearizations of rational eigenvalue problems it’s understood that the linearization is equivalent for $\lambda$ when the denominator is nonzero even if it not said explicitly for each case. This is a necessary for the problem to be defined for that value of $\lambda$.

We are given the rational eigenvalue problem

$$T(\lambda) u = (L(\lambda) - B(D - \lambda I)^{-1}B^*) u = 0,$$

(3.3)

$$L(\lambda) = A - \lambda C,$$

(3.4)

where $A$ and $C : H_1 \to H_1$, $B : H_2 \to H_1$ and $D : H_2 \to H_2$, are bounded linear operators and $H_1, H_2$ are Hilbert spaces, finite- or infinite dimensional. The idea is to find $\lambda$ such that there exist a nonzero vector $u \in H_1$ such that (3.3) is satisfied. $A$ and $C$ will for us always be positive definite operators and $D = c_1 I$ where $I$ is the identity operator and $c_1 > 0$, this is however not important for the linearization.

**Proposition 18.** Suppose that we are given the rational eigenvalue problem defined as (3.3) and (3.4) then the set $\sigma(T) = \sigma(T') \setminus \sigma(D)$ where $T'$ is defined as

$$T'(\lambda) := \begin{bmatrix} A & B \\ B^* & D \end{bmatrix} - \lambda \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix},$$

(3.5)

while the corresponding eigenvectors $u$ to $T$’s eigenvalues are the upper part of the eigenvectors to (3.5).

**Proof.** Rewriting equations (3.5) as

$$\begin{bmatrix} Au + Bv \\ B^*u + Dv \end{bmatrix} = \begin{bmatrix} \lambda Cu \\ \lambda v \end{bmatrix},$$
which for $\lambda$ is equivalent to

\[
\begin{bmatrix}
Au - \lambda C u + B v \\
B^* u
\end{bmatrix} = \begin{bmatrix}
0 \\
-(D - \lambda) v
\end{bmatrix}.
\]

Since $\lambda \notin \sigma(D)$, the inverse of the operator $-(D - \lambda)$ exists from which it follows

\[
\begin{bmatrix}
L(\lambda) u + B v \\
-(D - \lambda)^{-1} B^* u
\end{bmatrix} = \begin{bmatrix}
0 \\
v
\end{bmatrix}.
\] (3.6)

$L(\lambda) u$ is defined by (3.4), inserting the second row into the first row gives

\[
(L(\lambda) - B (D - \lambda)^{-1} B^*) u = 0.
\]

It is from this construction obvious that the eigenvectors coincide as well. ■

### 3.1. Properties of the finite dimension case

The finite dimensional case is of course more simple than the case when we have the more general operators working on infinite dimensional spaces. We’ll start out with the case when there is no damping, (the equations that is described are inherited from physical frequency problems and the sought eigenvalue is thus named $\omega$ or $\omega^2 = \lambda$) this will be discussed more in the next part of this chapter.

For the finite dimensional case that can be described by real matrices, the equation (3.3) reads, we have that $H_1 = \mathbb{R}^n$ and $H_2 = \mathbb{R}^{n^2}$ and $A, B, C, D$ are some known linear transformations on these spaces. We can directly use the linearization (3.5) as the main way of investigating this problem.

First we study the possibility that eigenvectors of (3.3) span $\mathbb{R}^n$, this is always an interesting property for an eigenvalue problem. We study the case where $A$ is symmetric, $C$ is the identity matrix and $D = c_1 I$ for some real $c_1$. The linearization (3.5) then becomes

\[
\begin{bmatrix}
A & B \\
B^* & c_1 I
\end{bmatrix}
\begin{bmatrix}
u \\
v
\end{bmatrix} = \lambda
\begin{bmatrix}
u \\
v
\end{bmatrix}.
\] (3.7)

The following theorem is new:

**Theorem 19.** Suppose that we have the system (3.7) and that

\[
\sigma(A) \subset (c_1, \infty).
\] (3.8)
Then each eigenvalue $\lambda$ of (3.7), for which the upper part $u$ of the eigenvector
$$\begin{bmatrix} u \\ v \end{bmatrix}$$
is not the zero-vector is an eigenvalue of (3.3).

**Proof.** We need to show that $c_1$ is not an eigenvalue to (3.7) and then it follows from Proposition 18 (since $c_1$ is the only eigenvalue to $D$) that the eigenvalues of (3.7) are eigenvalues to (3.3). We preform a proof by contradiction. Suppose that $c_1$ is an eigenvalue with the corresponding eigenvector $\begin{bmatrix} u \\ v \end{bmatrix}$, such that $u \neq 0$ and
$$\begin{bmatrix} A & B \\ B^* & D \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = c_1 \begin{bmatrix} u \\ v \end{bmatrix}.$$ 

Since $D = c_1 I$ it follows that
$$B^* u = 0, \quad (3.9)$$
$$\enneq (A - c_1) u + B v = 0, \quad (3.10)$$
we now multiply the equation with $u^*$ from the left to obtain from (3.10)
$$u^* (A - c_1) u + u^* B v = 0.$$

Using the adjoint properties yields
$$u^* (A - c_1) u + (B^* u)^* v = 0,$$
and using (3.9) it follows that
$$u^* (A - c_1) u + (0)^* v = u^* (A - c_1) u = 0.$$

From (3.8) it follows that $(A - c_1)$ is a positive definite matrix. Since $u$ is not the zero-vector it follows that $u^* (A - c_1) u > 0$ thus this leads to a contradiction. This leads to that $c_1$ is not an eigenvalue of (3.7), Proposition 18 then gives us that each eigenvalue of (3.7) is a eigenvalue of (3.3). $\blacksquare$

In a similar way the same result can also be obtain if we have the property
$$\sigma (A) \subset (-\infty, c_1). \quad (3.11)$$

instead of (3.8) but the proof is omitted. The only difference is that $(A - c_1)$ is negative definite instead of positive definite.
Corollary 20. Assume the system (3.7) and either property (3.8) or (3.11) and that $A \in \mathbb{R}^{n \times n}$ is Hermitian. Then the upper part of the eigenvectors $U_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix}$ in (3.7) span $\mathbb{R}^n$.

Proof. Since $\begin{bmatrix} A & B \\ B^* & D \end{bmatrix}$ is symmetric, this regular eigenvalue problem has orthonormal eigenvectors $U_i$ such that $\text{span} \{ U_1, ..., U_{n+n_2} \} = \mathbb{R}^{n+n_2}$ due to the spectral theorem. Thus $\text{span} \{ u_1, ..., u_{n+n_2} \} = \mathbb{R}^n$ so if all these eigenvectors are eigenvectors to the original system $\{ u_i \}$ spans $\mathbb{R}^n$. However they might have the corresponding eigenvalue $\lambda = c_1$ which means that $\lambda \in \sigma(D)$ and Proposition 18 does not apply.

Thus we show that the eigenvectors $\begin{bmatrix} u \\ v \end{bmatrix}$ where $u$ is not the zero vector can’t have $c_1$ as the corresponding eigenvalue. If $u$ is the zero vector it doesn’t matter if $c_1$ is the eigenvalue as $u$ doesn’t span any dimension in $\mathbb{R}^n$. Thus suppose $u$ is not the zero vector then it follows from Theorem 19 that the corresponding eigenvalue is an eigenvalue to (3.3) and thus $c_1$ is not an eigenvalue. But then all nonzero vectors $\{ u_i \}$ are eigenvectors to (3.3) and they span $\mathbb{R}^n$.

We will give an example where this property is not true, suppose we have that $\sigma(A) \subset (-\infty, \infty)$, instead of (3.8) or (3.11). An easy counterexample can be found by choosing $n = 3$, $c_1 = 1$ and $n_2 = 2$, and then choose $A$ and $B$ as

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$ 

$D$ is $I$ and (3.7) becomes

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} u \\ v \end{bmatrix},$$

and then we have that the eigenvalues is 1 and 2 and 1 has multiplicity 4 but 1 is not an eigenvalue of the original system since $\lambda = c_1$ is not defined in (3.3). Thus the only eigenvalues we have in the original system is 2 and only one eigenvector $u$
to eigenvalue 2. Inspired by Theorem 19, we want to be able to construct a matrix $A_0$ from $A$ such that we don’t get any new eigenvalues but so that either property (3.8) or (3.11) is valid. This will make it certain that we have a complete set of eigenvectors. Assume we have $m$ eigenvalues $\tilde{\lambda}_i$ to a symmetric matrix $A$ and a constant $c_1$ such that

$$\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_N \leq c_1 < \tilde{\lambda}_{N+1} \leq \cdots \leq \tilde{\lambda}_m,$$  \hspace{1cm} (3.13)

for some $1 \leq N < m$. Let

$$v = \text{span}\{v_1, \ldots, v_N\},$$  \hspace{1cm} (3.14)

where $v_i$ is the eigenvector corresponding to the $i$-th eigenvalue. Moreover define $P$ to be the orthogonal projection on $v$, and finally define

$$A_0 = A(I - P) + \tilde{\lambda}_{N+1}P.$$  \hspace{1cm} (3.15)

**Proposition 21.** Suppose (3.13), (3.14) and (3.15). Then

$$\sigma(A_0) = \bigcup_{i=N+1}^{m} \{\tilde{\lambda}_i\}.$$  

**Proof.** Let $u$ be an eigenvector to $A_0$ then for some eigenvalue $\lambda$ from (3.15)

$$\left(A(I - P) + \tilde{\lambda}_{N+1}P\right)u = \lambda u.$$  \hspace{1cm} (3.16)

since $A = A^*$ the spectral theorem states that $A$ has a orthonormal basis of eigenvectors $\{v_1, \ldots, v_m\}$. We then have that

$$u = \sum_{i=1}^{m} k_i v_i,$$

thus from (3.16)

$$\left(A(I - P) + \tilde{\lambda}_{N+1}P\right)\sum_{i=1}^{m} k_i v_i = \lambda \sum_{i=1}^{m} k_i v_i.$$  

This can be expanded to

$$A \sum_{i=1}^{m} k_i v_i - AP \sum_{i=1}^{m} k_i v_i + \tilde{\lambda}_{N+1}P \sum_{i=1}^{m} k_i v_i = \lambda \sum_{i=1}^{m} k_i v_i.$$  \hspace{1cm} (3.17)
Since the eigenvectors are orthonormal and $P$ is the orthogonal projection on $v$ it follows that

$$P \sum_{i=1}^{m} k_i v_i = \sum_{i=1}^{N} k_i v_i.$$  

Thus the left hand side of (3.17) can be written in the form

$$A \sum_{i=1}^{m} k_i v_i - AP \sum_{i=1}^{m} k_i v_i + \tilde{\lambda}_{N+1} P \sum_{i=1}^{m} k_i v_i =$$

$$\sum_{i=1}^{m} k_i Av_i - \sum_{i=1}^{N} k_i Av_i + \sum_{i=1}^{N} k_i \tilde{\lambda}_{N+1} v_i,$$

and since $v_i$ are eigenvectors to $A$ we have

$$\sum_{i=1}^{m} k_i \tilde{\lambda}_i v_i - \sum_{i=1}^{N} k_i \tilde{\lambda}_i v_i + \sum_{i=1}^{N} k_i \tilde{\lambda}_{N+1} v_i = \lambda \sum_{i=1}^{m} k_i v_i.$$  

By simplifying it follows that

$$\sum_{i=N+1}^{m} k_i \tilde{\lambda}_i v_i + \sum_{i=1}^{N} k_i \tilde{\lambda}_{N+1} v_i = \lambda \sum_{i=1}^{m} k_i v_i. \quad (3.18)$$

The idea is now to find $m$ eigenvectors and their corresponding eigenvalues. One can easily see that for $k_i = 1$ and all other $k_j = 0$ for $1 \leq i \leq N+1$ we have that (3.18) implies

$$\tilde{\lambda}_{N+1} v_i = \lambda v_i,$$

and thus we have $N + 1$ eigenvectors $\{v_i\}_{i=1}^{N+1}$ to the eigenvalue $\lambda = \lambda_{N+1}$. For $N + 2 \leq i \leq m$, it follows that by choosing $k_i = 1$ and all other $k_j = 0$ yields in (3.18)

$$\tilde{\lambda}_i v_i = \lambda v_i,$$

thus we have another $m - (N + 1)$ eigenvectors with eigenvalues $\tilde{\lambda}_i$ for $N + 2 \leq i \leq m$. But then we have found $m$ eigenvectors and since they are linearly independent they gives us all eigenvalues, namely

$$\sigma (A_0) = \bigcup_{i=N+1}^{m} \{\tilde{\lambda}_i\}.$$
We can also define
\[ A_0 = AP + \tilde{\lambda}_N (I - P). \] (3.19)

where the eigenvalues have the values
\[ 0 < \tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_N < c_1 \leq \tilde{\lambda}_{N+1} \leq \cdots \leq \tilde{\lambda}_m, \]

(the difference from (3.13) is the inequalities \( \tilde{\lambda}_N < c_1 \leq \tilde{\lambda}_{N+1} \) as we want \( \tilde{\lambda}_N < c_1 \)). We then have a similar property with a similar proof, all steps are carried out in the same way but in this case the result is

\[ \sigma (A_0) = \bigcup_{i=0}^{N} \{ \tilde{\lambda}_i \}. \]

In all cases where \( \sigma (A) \neq \{c_1\} \) at least one of these definitions of \( A_0 \) can be used, often both, and in both cases \( c_1 \) is not in the spectrum interval of \( A_0 \).

**Corollary 22.** Suppose the we have a matrix \( A_0 \) defined as (3.15) or as (3.19). Then \( A_0 \) if used as the symmetric matrix \( A \) in equation (3.7) we have the property that the upper parts of the eigenvectors span \( \mathbb{R}^n \).

**Proof.** From Proposition 21 we have that \( \sigma (A_0) \in (c_1, \infty) \) or \( \sigma (A_0) \in (0, c_1) \). Corollary 20 then proves the Corollary.

**3.2. Finite and infinite physical system**

When discussing damping in our eigenvalue problems, the terminology comes from one physical application of these equations. In these eigenvalue problems concerning waves, the eigenvalues \( \lambda \) equals \( \omega^2 \), where \( \omega \) is the frequency of the wave. Now if we have a problem as in (3.3) this system comes from a situation when the waves are non-damped. In a damped system we will also have a negative imaginary term depending on \( \omega \). The negative term results in that the waves looses energy. Theoretically we could have a positive instance of \( \omega \) or damping but this is rarely the case when observing a real physical problem as it would mean that the total energy is increasing. We will start with the non-damped equation, and extend to the case when we have a damping. In both the non-damped and the damped cases, we preform the linearization in two different ways, and compare the different methods.
3.2.1. Non-damped polynomial method

Consider the non-damped non-linear eigenvalue problem given as

\[ Gu - \lambda \alpha_1 M_1 u - \lambda \left( \alpha_2 + \frac{\xi^2}{\eta^2 - \lambda} \right) M_2 u = 0, \quad (3.20) \]

Where \( G, M_1, M_2 \in \mathbb{R}^{n \times n} \) are matrices and \( \alpha_1, \alpha_2, \xi \) and \( \eta \) are real positive constants, we want to find the eigenfrequency \( \omega \), since there only are instances of \( \omega^2 = \lambda \) we find \( \omega \) by finding \( \lambda \). \( M_2 \) also have the property that \( M_2 = diag \left( 0, \widehat{M}_2 \right) \) where \( \widehat{M}_2 \in \mathbb{R}^{n_2 \times n_2} \) is positive definite, and \( M_1 + M_2 \) is positive definite. One possible, intuitive, idea is to multiply by the denominator \( \eta^2 - \lambda \) to obtain the problem

\[ \left( Gu - \lambda \alpha_1 M_1 u - \lambda \left( \alpha_2 + \frac{\xi^2}{\eta^2 - \lambda} \right) M_2 u \right) (\eta^2 - \lambda) = 0. \]

Which is equivalent to the polynomial equation

\[ Gu \left( \eta^2 - \lambda \right) - \lambda \alpha_1 M_1 u \left( \eta^2 - \lambda \right) - \lambda \left( \alpha_2 \left( \eta^2 - \lambda \right) + \xi^2 \right) M_2 u = 0. \]

By rearranging the terms we have

\[ \eta^2 Gu - \lambda \left( G + \eta^2 \alpha_1 M_1 \right) u + \lambda^2 \alpha_1 M_1 u - \lambda \left( \alpha_2 \eta^2 + \xi^2 \right) M_2 u + \lambda^2 \alpha_2 M_2 u = 0, \]

and finally

\[ \eta^2 Gu - \lambda \left( G + \eta^2 \alpha_1 M_1 + \alpha_2 \eta^2 M_2 + \xi^2 M_2 \right) u + \lambda^2 \left( \alpha_1 M_1 + \alpha_2 M_2 \right) u = 0. \]

Defining the \( n \times n \) matrices

\[ A' = \eta^2 G, \quad (3.21) \]

\[ B' = - \left( G + \eta^2 \alpha_1 M_1 + \alpha_2 \eta^2 M_2 + \xi^2 M_2 \right), \]

\[ C' = - \left( \alpha_1 M_1 + \alpha_2 M_2 \right). \]

The system then becomes

\[ A' u + \lambda B' u - \lambda^2 C' u = 0. \quad (3.22) \]

By defining

\[ v = \lambda u, \]
the linearization follows as
\[
\begin{bmatrix}
0 & I \\
A' & B'
\end{bmatrix}
\begin{bmatrix}
u \\
v
\end{bmatrix} = \lambda \begin{bmatrix}
I & 0 \\
0 & C'
\end{bmatrix}
\begin{bmatrix}
u \\
v
\end{bmatrix}.
\] (3.23)

The eigenvalues from (3.22) coincides with those of (3.20). The system we obtain is however a generalized eigenvalue problem for $2n \times 2n$ matrices so this method increases the computation time to that of solving linear eigenvalue problems of size $2n$. Another important property that can be noted is that the matrix on the left hand side is not symmetric and thus not normal in general, this may result in the possibility that the eigenvalues are hard to compute numerically. Therefore it might be good to use another linearization method to avoid this, e.g the method in the previous chapter.

3.2.2. Non-damped rational method

We will now do a more complicated approach, but it will give us an eigenvalue problem of less dimension than the one obtained in (3.23). Furthermore the matrix in the left-hand side will be normal, the latter of these properties are the more important. To be able to do this (3.20) must be written as (3.3).

**Lemma 23.** Given the linear system (3.20). Then a linearization of the system is
\[
\begin{bmatrix}
\hat{G} & B^* \\
B & D
\end{bmatrix}
\begin{bmatrix}
u \\
v
\end{bmatrix} = \lambda \begin{bmatrix}
\hat{M} & 0 \\
0 & I
\end{bmatrix}
\begin{bmatrix}
u \\
v
\end{bmatrix},
\] (3.24)

where $\hat{G} = G + \xi^2 M_2$, $\hat{M} = \alpha_1 M_1 + \alpha_1 M_2$ and $B = \xi \eta F$, $F \in \mathbb{R}^{n^2 \times n}$ is the matrix such that $M_2 = F^* F$.

**Proof.** (3.20) can be written as
\[
Gu - \lambda (\alpha_1 M_1 + \alpha_1 M_2) u - \frac{\xi^2}{\eta^2 - \lambda} M_2 u = 0,
\]
or equally
\[
Gu - \lambda (\alpha_1 M_1 + \alpha_1 M_2) u - \frac{\xi^2 (\lambda - \eta^2)}{\eta^2 - \lambda} M_2 u - \frac{\xi^2 \eta^2}{\eta^2 - \lambda} M_2 u = 0.
\]
Finally, we get
\[
(G + \xi^2 M_2) u - \lambda (\alpha_1 M_1 + \alpha_1 M_2) u - \frac{\xi^2 \eta^2}{\eta^2 - \lambda} M_2 u = 0.
\]
Setting $G + \xi^2 M_2 = \hat{G}$ and $\alpha_1 M_1 + \alpha_1 M_2 = \hat{M}$ we have the system
\[
\hat{G}u - \lambda \hat{M}u - \frac{\xi^2 \eta^2}{\eta^2 - \lambda} M_2 u = 0,
\]
where $M_2 = \text{diag}\left(0, \hat{M}_2\right)$, and $\hat{M}_2 \in \mathbb{R}^{n_2 \times n_2}$, $n_2 \leq n$ and $\hat{M}_2$ is positive definite and real. Hence $\hat{M}_2$ can, due to the Cholesky decomposition theorem, be decomposed as $\hat{M}_2 = \hat{F}^* \hat{F}$ where $\hat{F} \in \mathbb{R}^{n_2 \times n_2}$. Defining $F$ as the $n_2 \times n$ matrix such that
\[
F = \begin{bmatrix} 0 & \hat{F} \end{bmatrix}.
\]
From this definition
\[
F^* F = \begin{bmatrix} 0 & \hat{F}^* \end{bmatrix} \begin{bmatrix} 0 & \hat{F} \end{bmatrix},
\]
and thus
\[
\begin{bmatrix} 0 & 0 \\ 0 & \hat{F}^* \hat{F} \end{bmatrix} = M_2.
\]
The problem can then be written as
\[
\hat{G}u - \lambda \hat{M}u - F^* \frac{\xi^2 \eta^2}{\eta^2 - \lambda} F u = 0,
\]
setting $B = \xi \eta F$ and $D = \eta^2 I$ and defining $v$ as
\[
v = - (D - \lambda)^{-1} Bu,
\]
leads to the equivalent linear system
\[
\begin{bmatrix} \hat{G} & B^* \\ B & D \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} \hat{M} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.
\]

The two problems (3.20) and (3.24) are thus equivalent which follows from Proposition 18.

The advantage with this method is that the obtained linear matrix eigenvalue problem has the dimension is $n + n_2 \leq 2n$ so it has lower dimension of the system than (3.23). It also is symmetric and thus normal which makes the problem well conditioned. It can easily be seen that this eigenvalue problem is similar to that of (3.7) and the lemmas and theorems above apply to this equation too.
3.2.3. Damped case

Now we can continue with the more complicated problem, where the eigenvalue problem has a damping term. We will notice that the linearization is different in this case. The problem is exactly the same as (3.20) except for a new term in the rational part. This equation, which can be found in even more general cases, see [12, page 184], can be written as

\[ T\left(\sqrt{\lambda}\right) u := Gu - \lambda \alpha_1 M_1 u - \lambda \left( \alpha_2 + \frac{\xi^2}{\eta^2 - \lambda - i\gamma\sqrt{\lambda}} \right) M_2 u = 0, \quad (3.25) \]

where the new parameter \( \gamma \in \mathbb{R} \) induces a damping. In this case it’s more convenient to try to find eigenvalues \( \omega = \sqrt{\lambda} \) instead of \( \omega^2 = \lambda \), it’s therefore better to use the notation \( \omega \) instead of \( \sqrt{\lambda} \), as it is \( \omega \) that are interesting in the first place. We start with a theorem concerning the eigenvalues of (3.25), the proof needs the following definitions:

**Definition 24.** The numerical range for an operator-valued function \( T(\omega) : H_1 \rightarrow H_1 \) is defined as

\[ W(T) := \{ \omega \in \mathbb{C} | (T(\omega)u, u) = 0 \text{ for some } u \in H_1 \text{ where } \|u\| = 1 \}. \]

**Definition 25.** The numerical range for an bounded operator \( T : H_1 \rightarrow H_1 \) is defined as

\[ W(T) := \{ \omega \in \mathbb{C} | (Tu, u) = \omega \text{ for some } u \in H_1 \text{ where } \|u\| = 1 \}. \]

**Lemma 26.** Given the system defined by (3.25) and assume that \( G \) is positive definite and \( M_1, M_2 \) are positive semi-definite matrices. Furthermore assume that the constants \( \gamma, \alpha_1, \alpha_2, \xi, \eta \) are strictly positive. Then \( \sigma(T) \subset W(T) \).

**Proof.** Let \( \omega = i \) in the operator valued function,

\[ T(i) = G + \alpha_1 M_1 + \left( \alpha_2 + \frac{\xi^2}{\eta^2 + 1 + \gamma} \right) M_2. \quad (3.26) \]

Computing the numerical range for this particular \( \omega \), we have

\[ W(T(i)) = \{ u^* (T(i)) u | u \in \mathbb{C}^n, \|u\| = 1 \}. \]

Using that the matrix given by \( T(i) \) (3.26) is positive definite (which follows easily from the given properties) we have that \( 0 \notin W(T(i)) \). The theorem [13, page 139] states that

\( \sigma(T) \subset W(T) \).

\[ \blacksquare \]
**Theorem 27.** Suppose we are given the system (3.25) and assume that $G$ is positive definite and $M_1, M_2$ are positive semi-definite matrices. Furthermore assume that the constants $\gamma, \alpha_1, \alpha_2, \xi, \eta$ are strictly positive. Then $\text{Im} \left( \sigma (T) \right) < 0$.

The proof of this theorem is long and rather technical and is thus placed in the appendix.

### 3.2.4. Damped polynomial method

As before we start by the more intuitive method to linearize the problem by multiplying with the denominator to obtain a polynomial equation from (3.25). Using that

$$
\frac{\xi^2 \omega^2}{\eta^2 - \omega^2 - i\gamma \omega} = -\xi^2 + \frac{\xi^2 (\eta^2 - i\gamma \omega)}{\eta^2 - \omega^2 - i\gamma \omega},
$$

(3.25) can be written as

$$
\left( (G + \xi^2 M_2) - \omega^2 (\alpha_1 M_1 + \alpha_2 M_2) - \left( \frac{\xi^2 (\eta^2 - i\gamma \omega)}{\eta^2 - \omega^2 - i\gamma \omega} \right) M_2 \right) u = 0. \quad (3.27)
$$

To simplify the notation we set $\overset{\sim}{G} = G + \xi^2 M_2$ and $M = \alpha_1 M_1 + \alpha_2 M_2$. The equation (3.27) can then be written as

$$
\left( \overset{\sim}{G} - \omega^2 M - \left( \frac{\xi^2 (\eta^2 - i\gamma \omega)}{\eta^2 - \omega^2 - i\gamma \omega} \right) M_2 \right) u = 0. \quad (3.28)
$$

Multiplication by the denominator gives

$$
(\eta^2 - \omega^2 - i\gamma \omega) \left( \overset{\sim}{G} - \omega^2 M - \left( \frac{\xi^2 (\eta^2 - i\gamma \omega)}{\eta^2 - \omega^2 - i\gamma \omega} \right) M_2 \right) u = 0,
$$

which is equivalent to

$$
\left( (\eta^2 - \omega^2 - i\gamma \omega) \overset{\sim}{G} - (\eta^2 \omega^2 - \omega^4 - i\gamma \omega^3) M - \xi^2 (\eta^2 - i\gamma \omega) M_2 \right) u = 0.
$$

Collecting terms after dependence of $\omega$ yields

$$
\left( (\eta^2 \overset{\sim}{G} - \xi^2 \eta^2 M_2) - (i\gamma \overset{\sim}{G} - i\gamma \xi^2 M_2) \omega
\right.
\left. - (\overset{\sim}{G} + \eta^2 M) \omega^2 + i\gamma M \omega^3 + M \omega^4 \right) u = 0. \quad (3.29)
$$
Define the constants

\[
A' = -\left(\eta^2 \hat{G} - \xi^2 \eta^2 M_2\right), \\
B' = i\gamma \hat{G} - i\xi \eta^2 M_2, \\
C' = \left(\hat{G} + \eta^2 M\right), \\
D' = -i\gamma M.
\]

(3.30)

It follows that (3.29) can be written as

\[
(-A' - B'\omega - C'\omega^2 - D'\omega^3 + M\omega^4) u = 0,
\]

or

\[
(A' + B'\omega + C'\omega^2 + D'\omega^3) u = M\omega^4 u.
\]

We now denote new vectors as

\[
v = \omega u, \\
w = \omega v, \\
z = \omega w.
\]

(3.31)

Hence obtaining the linear system as

\[
\begin{bmatrix}
0 & I & 0 & 0 \\
0 & 0 & I & 0 \\
0 & 0 & 0 & I \\
A' & B' & C' & D'
\end{bmatrix}
\begin{bmatrix}
u \\
v \\
w \\
z
\end{bmatrix}
= \omega
\begin{bmatrix}
I & 0 & 0 & 0 \\
0 & I & 0 & 0 \\
0 & 0 & I & 0 \\
0 & 0 & 0 & M
\end{bmatrix}
\begin{bmatrix}
u \\
v \\
w \\
z
\end{bmatrix}.
\]

(3.32)

This follows from that standard linearization method exactly as in (3.23).

3.2.5. Damped rational method

We are now interested to obtain a different linearization with inspiration from (3.7) hoping that the dimension of the linear system will be less and also that the resulting matrix will be close normal for small damping coefficients \(\gamma\). To obtain this linearization we start by the equivalent system (3.28). The rational linearization of this damped problem will be a lot more technical than that of the undamped case.
To convert this problem to a linear problem, define $A$ and $E$ as the two-by-two matrices as
\[
A = \frac{2\theta}{\xi^2} \begin{bmatrix} \theta - \frac{1}{2} i \gamma & 0 \\ 0 & \theta + \frac{1}{2} i \gamma \end{bmatrix}, \quad E = \frac{2\theta}{\xi^2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},
\] (3.33)
and the vector $b$ as
\[
b = \begin{bmatrix} \theta - \frac{1}{2} i \gamma \\ \theta + \frac{1}{2} i \gamma \end{bmatrix},
\] (3.34)
where
\[
\theta = \sqrt{\eta^2 - \frac{1}{4} \gamma^2}.
\] (3.35)

**Lemma 28.** [12, page 185] The system (3.28) is equivalent to
\[
\left( \hat{G} - \omega^2 M - b^T (A - \omega E)^{-1} b M_2 \right) u = 0,
\] (3.36)
where $A, E$ and $b$ are defined as (3.33),(3.34) and (3.35).

The proof which only relies on computations is long and has been put in the appendix. The last part of getting a linear problem is noting that $M_2 = \text{diag} \left( 0, \widetilde{M}_2 \right)$ where $\widetilde{M}_2 \in \mathbb{R}^{n_2 \times n_2}$, $n_2 \leq n$ and $\widetilde{M}_2$ is positive definite. Thus $M_2$ can be written as $F^T F$ for some $F \in \mathbb{R}^{n_2 \times N}$, just as it was done in the non-damped case. By some further definitions we obtain the next lemma,

\[
\mathcal{A} = A \otimes I_{N_2}, \\
\mathcal{B} = b \otimes F, \\
\mathcal{E} = E \otimes I_{N_2},
\] (3.37)
where $\otimes$ denotes the Kronecker product.

**Lemma 29.** [12, page 185] The system (3.36) is equivalent to
\[
\left( \hat{G} - \omega^2 M - \mathcal{B}^T (\mathcal{A} - \omega \mathcal{E})^{-1} \mathcal{B} \right) u = 0,
\] (3.38)
where $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{E}$ are defined as (3.37).
Proof. Once again the proof is done by straight forward computations. From (3.37) it follows
\[ B^T (A - \omega E)^{-1} B = (b \otimes F)^T (A \otimes I_{N_2} - \omega E \otimes I_{N_2})^{-1} b \otimes F. \]

Using (3.34) yields
\[ (b \otimes F)^T ((A - \omega E) \otimes I_{N_2})^{-1} = \begin{bmatrix} b_1 F^T \\ b_2 F \end{bmatrix}^T ((A - \omega E) \otimes I_{N_2})^{-1}. \tag{3.39} \]

The matrix \((A - \omega E) \otimes I_{N_2}\) is diagonal since \(A, E\) and \(I\) are diagonal matrices. Therefore \(((A - \omega E) \otimes I_{N_2})^{-1}\) is a diagonal matrix,
\[ ((A - \omega E) \otimes I_{N_2})^{-1} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}, \]
for some diagonal matrices \(c_1, c_2 \in \mathbb{R}^{n_2 \times n_2}\). Since \(((A - \omega E) \otimes I_{N_2})^{-1}\) is a diagonal matrix it follows from (3.39) that
\[ (b \otimes F)^T ((A - \omega E) \otimes I_{N_2})^{-1} = \begin{bmatrix} b_1 F^T c_1 & b_2 F^T c_2 \end{bmatrix}. \]

Then this gives
\[ (b \otimes F)^T ((A - \omega E) \otimes I_{N_2})^{-1} (b \otimes F) = \begin{bmatrix} b_1 F^T c_1 \\ b_2 F^T c_1 \end{bmatrix} b \otimes F, \]
which is
\[ \begin{bmatrix} b_1 F^T c_1 & b_2 F^T c_2 \end{bmatrix} \begin{bmatrix} b_1 F \\ b_2 F \end{bmatrix} = b_1 F^T c_1 F b_1 + b_2 F^T c_2 F b_2. \]

We can complete the proof by showing that from (3.39)
\[ b_1 c_1 b_1 M_2 + b_2 c_2 b_2 M_2 = b^T ((A - \omega E) \otimes I_{N_2})^{-1} b M_2, \]
and then it follows
\[ \left( \tilde{G} - \omega^2 M - B^T (A - \omega E)^{-1} B \right) u = \left( \tilde{G} - \omega^2 M - b^T (A - \omega E)^{-1} b M_2 \right) u. \]
To complete the linearization of the problem, introduce vectors $v, w$ defined as

$$v = - (A - \omega E)^{-1} B u,$$

(3.40)

and

$$w = \omega u.$$

(3.41)

From them obtaining the system

$$\begin{bmatrix}
\hat{G} & B^T & 0 \\
B & A & 0 \\
0 & 0 & I
\end{bmatrix}
\begin{bmatrix}
u \\
v \\
w
\end{bmatrix}
= \omega
\begin{bmatrix}
0 & 0 & M \\
0 & E & 0 \\
I & 0 & 0
\end{bmatrix}
\begin{bmatrix}
u \\
v \\
w
\end{bmatrix},

(3.42)

that is equivalent to the system (3.25) for $\eta^2 - \omega^2 - i\gamma \omega \neq 0$. This method is harder to understand but the size of the resulting eigenvalue problem is a $(2n_2 + 2n) \times (2n_2 + 2n)$ matrix, which is lower degree than that of problem (3.32) that was of order $4n \times 4n$, $n_2$ is the length of $\hat{M}_2$. we also have that the matrix on the left hand side is normal if $\gamma = 0$ and for low values $\gamma$ it is close to symmetric and thus close to normal, this follows from that the disturbance is small. We now show the equivalence, suppose $\omega$ is an eigenvalue to the system (3.28) with corresponding eigenvector $[u, v, w]^T$. It then follows from (3.42) that

$$\begin{bmatrix}
\hat{G}u + B^Tv \\
Bu + Av \\
w
\end{bmatrix}
= \begin{bmatrix}
\omega M w \\
\omega E v \\
\omega u
\end{bmatrix},

and since $\det (A - \omega E) \neq 0$ the identity $Bu + Av = \omega E v \iff Bu = -(A - \omega E)v$ holds, and the equivalent system

$$\begin{bmatrix}
\hat{G}u + B^Tv \\
v \\
w
\end{bmatrix}
= \begin{bmatrix}
\omega M w \\
-(A - \omega E)^{-1}Bu \\
\omega u
\end{bmatrix},

is obtained. Thus (3.40) and (3.41) holds and by using the definitions of $v$ and $w$, (3.38) is equivalent to (3.42). By Lemma 28 and Lemma 29 this is equivalent to (3.25) as well. This problem can now be solved by a standard method both when calculating spectra and pseudospectra. To make the computations easier, however, another approach can be used to compute the spectra. To calculate the solution as easy as possible we note that if we choose to calculate the eigenvalues by taking the inverse of the matrix in the right-hand side, the problem of inverting
a matrix of order $2n+2n_2$ as well as performing a matrix multiplication of the same order will be encountered. Furthermore computing a regular eigenvalue problem of order $2n+2n_2$ is required. Therefore it is interesting to reduce the numbers of calculations. By instead taking the inverse of $\omega$ on both sides and take the inverse of the left-hand side, the equation follows as

$$\frac{1}{\omega} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \hat{G} & B^T & 0 \\ B & A & 0 \\ 0 & 0 & I \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 & M \\ 0 & \mathcal{E} & 0 \\ I & 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}. \quad (3.43)$$

The idea is that the inverse of this matrix is more easily obtained for general inputs.

$$\begin{bmatrix} \hat{G} & B^T & 0 \\ B & A & 0 \\ 0 & 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} \hat{G} & B^T \\ B & A \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 & M \\ 0 & \mathcal{E} & 0 \\ I & 0 & 0 \end{bmatrix},$$

thus the inverse of the two by two block matrix $\begin{bmatrix} \hat{G} & B^T \\ B & A \end{bmatrix}$ instead.

**Lemma 30.** [12, page 186] Given the matrix

$$\begin{bmatrix} \hat{G} & B^T \\ B & A \end{bmatrix},$$

and that $A$ is invertible it can be Frobenius-Schur factorized to

$$\begin{bmatrix} I & B^T A^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{G} - B^T A^{-1} B \\ 0 & A \end{bmatrix} \begin{bmatrix} I & 0 \\ A^{-1} B & I \end{bmatrix}.$$

**Proof.** The proof is simple and only by calculation, since $A$ is invertible the expression exists and we get

$$\begin{bmatrix} I & B^T A^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{G} - B^T A^{-1} B \\ 0 & A \end{bmatrix} \begin{bmatrix} I & 0 \\ A^{-1} B & I \end{bmatrix} = \begin{bmatrix} \hat{G} - B^T A^{-1} B & B^T \\ 0 & A \end{bmatrix} \begin{bmatrix} I & 0 \\ A^{-1} B & I \end{bmatrix}. $$
and then
\[
\begin{bmatrix}
\hat{G} - B^T A^{-1} B & B^T \\
0 & A
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
A^{-1} B & I
\end{bmatrix}
= \begin{bmatrix}
\hat{G} - B^T A^{-1} B + B^T A^{-1} B & B^T \\
A A^{-1} B & A
\end{bmatrix}
= \begin{bmatrix}
\hat{G} & B^T \\
B & A
\end{bmatrix}.
\]

Thus the property
\[
\begin{bmatrix}
\hat{G} & B^T \\
B & A
\end{bmatrix}^{-1}
= \left(\begin{bmatrix}
I & B^T A^{-1} \\
0 & I
\end{bmatrix}
\begin{bmatrix}
\hat{G} - B^T A^{-1} B & 0 \\
0 & A
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
A^{-1} B & I
\end{bmatrix}\right)^{-1},
\]
can be obtained. Using identity concerning \(\hat{G} - B^T A^{-1} B\) from (3.37) and the definition of \(\hat{G}\) yields
\[
\hat{G} - B^T A^{-1} B = G + \xi^2 M_2 - (b \otimes F)^T \cdot (A \otimes I_{N_2})^{-1} (b \otimes F). \tag{3.44}
\]

From Lemma (29) it easily follows that \((b \otimes F)^T \cdot (A \otimes I_{N_2})^{-1} (b \otimes F) = b^T A^{-1} b M_2\), using this and the definitions (3.33) and (3.34) in (3.44), it follows that
\[
\hat{G} - B^T A^{-1} B = G + \xi^2 M_2 - \left[\begin{array}{cc}
\theta - \frac{1}{2} i \gamma & \theta + \frac{1}{2} i \gamma \\
\theta - \frac{1}{2} i \gamma & 0
\end{array}\right] \cdot
\begin{bmatrix}
2 \theta \\
\xi^2
\end{bmatrix}
\begin{bmatrix}
\theta - \frac{1}{2} i \gamma & 0 \\
0 & \theta + \frac{1}{2} i \gamma
\end{bmatrix}
\begin{bmatrix}
\theta - \frac{1}{2} i \gamma \\
0
\end{bmatrix}
M_2.
\]

By simplification
\[
= G + \xi^2 M_2 - \frac{\xi^2}{2 \theta} \left[\begin{array}{cc}
\theta - \frac{1}{2} i \gamma & \theta + \frac{1}{2} i \gamma \\
0 & \theta + \frac{1}{2} i \gamma
\end{array}\right] \begin{bmatrix}
\frac{1}{\theta - \frac{1}{2} i \gamma} & 0 \\
0 & \frac{1}{\theta + \frac{1}{2} i \gamma}
\end{bmatrix}
\begin{bmatrix}
\theta - \frac{1}{2} i \gamma \\
0
\end{bmatrix}
M_2;
\]

and finally
\[
= G + \xi^2 M_2 - \frac{\xi^2}{2 \theta} \left[\begin{array}{c}
1 \\
1
\end{array}\right]
\begin{bmatrix}
\theta - \frac{1}{2} i \gamma & \theta + \frac{1}{2} i \gamma
\end{bmatrix}
M_2 =
= G + \xi^2 M_2 - \frac{\xi^2}{2 \theta} \left[\begin{array}{c}
\theta - \frac{1}{2} i \gamma + \theta + \frac{1}{2} i \gamma
\end{array}\right] M_2 = G.
\]

Thus using previous Lemma
\[
\begin{bmatrix}
\hat{G} & B^T \\
B & A
\end{bmatrix}^{-1} = 
\begin{bmatrix}
I & B^T A^{-1} \\
0 & I
\end{bmatrix}
\begin{bmatrix}
G & 0 \\
0 & A
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
A^{-1} B & I
\end{bmatrix}^{-1} = 
\begin{bmatrix}
G & 0 \\
0 & A
\end{bmatrix}^{-1}
\begin{bmatrix}
I & B^T A^{-1} \\
0 & I
\end{bmatrix}^{-1}.
\]

To be able to find these inverses we note that
\[
\begin{bmatrix}
I & 0 \\
A^{-1} B & I
\end{bmatrix}
\begin{bmatrix}
I & 0 \\
-A^{-1} B & I
\end{bmatrix} = 
\begin{bmatrix}
A^{-1} B & -A^{-1} B & I
\end{bmatrix} = I,
\]
and
\[
\begin{bmatrix}
I & B^T A^{-1} \\
0 & I
\end{bmatrix}
\begin{bmatrix}
I & -B^T A^{-1} \\
0 & I
\end{bmatrix} = 
\begin{bmatrix}
I & B^T A^{-1} - B^T A^{-1} I
\end{bmatrix} = I.
\]

Inserting the inverses yields
\[
\begin{bmatrix}
\hat{G} & B^T \\
B & A
\end{bmatrix}^{-1} = 
\begin{bmatrix}
I & 0 \\
-A^{-1} B & I
\end{bmatrix}
\begin{bmatrix}
G^{-1} & 0 \\
0 & A^{-1}
\end{bmatrix}
\begin{bmatrix}
I & -B^T A^{-1} \\
0 & I
\end{bmatrix}.
\]

Returning to the compact form of the equation
\[
\begin{bmatrix}
\hat{G} & B^T \\
B & A
\end{bmatrix}^{-1} = 
\begin{bmatrix}
G^{-1} & 0 \\
-A^{-1} B G^{-1} & A^{-1}
\end{bmatrix}
\begin{bmatrix}
I & -B^T A^{-1} \\
0 & I
\end{bmatrix},
\]
preforming the matrix multiplication to obtain
\[
\begin{bmatrix}
\hat{G} & B^T \\
B & A
\end{bmatrix}^{-1} = 
\begin{bmatrix}
G^{-1} & -G^{-1} B^T A^{-1} \\
-A^{-1} B G^{-1} & A^{-1} B G^{-1} B^T A^{-1} + A^{-1}
\end{bmatrix}.
\]

Finally the less computational demanding \( \Phi \), using (3.43) is
\[
\Phi = 
\begin{bmatrix}
G^{-1} & -G^{-1} B^T A^{-1} & 0 \\
-A^{-1} B G^{-1} & A^{-1} B G^{-1} B^T A^{-1} + A^{-1} & 0 \\
0 & I & \mathcal{E}
\end{bmatrix}
\begin{bmatrix}
0 & 0 & M \\
0 & \mathcal{E} & 0 \\
I & 0 & 0
\end{bmatrix},
\]
and in compact form
\[
\Phi = 
\begin{bmatrix}
0 & -G^{-1} B^T A^{-1} \mathcal{E} & G^{-1} M \\
0 & (A^{-1} B G^{-1} B^T A^{-1} + A^{-1}) \mathcal{E} & -A^{-1} B G^{-1} M \\
I & 0 & \mathcal{E}
\end{bmatrix}.
\] (3.45)
To obtain the eigenvalues $\omega$ we now only have to find the eigenvalues of $\Phi$ and invert them, this approach was also used in [12, page 186-187]. In the numerical examples we will, however, not use this method when computing the eigenvalues $\omega$, they will be computed straight from (3.42) by the standard approach of inverting the right hand side.

### 3.3. Photonic Crystals

We have until now only considered finite dimensional spaces where the eigenvalue problems can be expressed by the means of matrices. We will now move on to the more general case with operators on infinite dimensional Hilbert space. The definition of the spectra of a bounded linear operator is analogous with the definition of the spectra of a matrix.

**Definition 31.** Let $T$ be a bounded linear operator on a Hilbert space $X$ over $\mathbb{C}$. Let $I$ be the identity operator in $X$, then the spectrum of $T$, $\sigma(T)$, are the $\lambda \in \mathbb{C}$ such that $T - \lambda I$ doesn’t have a inverse that is a bounded linear operator.

By [5, page 84] we have that if $T - \lambda I$ is invertible then this inverse is linear, from theorem page [5, page 286] it also follows that this inverse also is bounded. Hence the the spectrum consist entirely of the values $\lambda$ for which $T - \lambda I$ is not a bijective operator. In the operator case we will however have different classifications of the points in the spectrum according to their properties.

The interest for the equation we will look at now comes from an application of photonic crystals, where the wave eigenfrequency is sought. The bounded set over which the problem will be defined is denoted $\Omega \subset \mathbb{R}^2$.

In for example nano-opticts [17] there are periodic structures made from two or more materials, with different electrodynamic properties, called photonic crystals. To solve Maxwell’s equations for photonic crystals in terms of Floquet-Bloch waves, the solutions depends on the frequency $\omega$ of the waves [18]. The electromagnetic wave can be decomposed in the transverse electric (TE) polarized waves and the transverse magnetic (TM) polarized waves. According to [19] the TE-waves can be described by the spectra of operator valued function:

$$T_{TE}(\omega) := -\nabla \cdot \left( \frac{1}{\epsilon(x, \omega)} \nabla \right) - \omega^2, \ x \in \mathbb{R}^2.$$
While the TM-waves are described by the spectra of operator valued function:

$$T_{TM}(\omega) := -\Delta - \omega^2 \epsilon(x, \omega), \quad x \in \mathbb{R}^2.$$ 

We are interested in periodic solutions to these equations so we do the ansatz that the eigenvectors are $E_3(x) = e^{ik \cdot x} u(x)$ and $H_3(x) = e^{ik \cdot x} u(x)$ where $u$ is periodic in our domain $\Omega$. We have that $k$ can be any vector in $\mathbb{R}^2$, however if any component of $k$ is outside of the range of $[-\pi, \pi]$ this could always be included in the Fourier series of $u(x)$ due to periodicity. Thus we only have to investigate the $k$ in the subdomain $[-\pi, \pi]^2 \subset \mathbb{R}^2$. This leads to that

$$\sigma(T_{TE}) = \bigcup_{k \in [-\pi, \pi]^2} \sigma \left( -\nabla_k \cdot \left( \frac{1}{\epsilon(x, \omega)} \nabla_k \right) - \omega^2 \right) \quad (3.46)$$

and

$$\sigma(T_{TM}) = \bigcup_{k \in [-\pi, \pi]^2} \sigma \left( -\nabla_k \cdot \nabla_k - \omega^2 \epsilon(x, \omega) \right). \quad (3.47)$$

Hence the spectra of the operator valued functions $T_{TE}$ and $T_{TM}$ can be found as the unions over all possible $k$ values as described in (3.46) and (3.47). Experience say that all eigenvalues are obtained by computing for $k$ along the boundary of the triangle with vertices $(0,0), (\pi,0)$ and $(\pi,\pi)$, this can be seen in ([16, pages 16-19]). Now if there exist some frequencies $\omega$ that are not eigenvalues to the problem for any $k \in [\pi, \pi]^2$, these values of $\omega$ will never have a resonance response in this domain. This leads to that for these frequencies the domain will work as a nano-mirror, since the waves is not able to travel through the matter. One can therefore choose components such that the wanted frequencies are reflected while the other frequencies are not. The problem of finding the reflected frequencies is hence equivalent of solving the nonlinear eigenvalueproblems (3.46) and (3.47).

### 3.3.1. Hilbert space

We now define a space to the eigenvalue problem from where we accept our vectors. The sought eigenvectors $u : \Omega \to \mathbb{C}$ are periodic in $\Omega$ and can thus described with Fourier coefficients, in this thesis we have that $\Omega = [0,1]^2$. For that domain $\Omega$ the Fourier coefficients will have the apperance:

$$u(x) = \sum_{n \in \mathbb{Z}^2} \hat{u}(n) e^{2\pi i n \cdot x}, \quad \text{for } x \in \Omega. \quad (3.48)$$
since \( u(x) \) is periodic in \([0,1]^2\). Given this representation, define the Hilbert space as

\[
H^1(\Omega) = \left\{ u \in L^2(\Omega) : \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) |\hat{u}(n)|^2 < \infty \right\},
\]

with the inner product

\[
(u, v) = \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) \hat{u}(n) \hat{v}(n)^*.
\]

From the inner product it follows that the norm is indeed defined as in (3.49). First, however, we have to verify that (3.50) is a proper inner product.

**Proposition 32.** (3.50) is an inner product

**Proof.** (3.50) is conjugate symmetric for all \( u, v \in H^1(\Omega) \):

\[
(u, v) = \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) \hat{u}(n) \hat{v}(n)^* = \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) (\hat{u}(n)^* \hat{v}(n))^* = \left(\sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) \hat{u}(n)^* \hat{v}(n)\right)^* = (v, u)^*.
\]

Linear in the first argument:

\[
(u + v, w) = \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) (\hat{u}(n) + \hat{v}(n)) \hat{w}(n)^* = \\
\sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) \hat{u}(n)^* \hat{w}(n) + \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) \hat{v}(n)^* \hat{w}(n) = (u, w) + (v, w),
\]

\[
(au, v) = \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) a \hat{u}(n) \hat{v}(n)^* = a \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) \hat{u}(n)^* \hat{v}(n) = a (u, v).
\]

Positive definite \((u, u) \geq 0\) and \((u, u) = 0 \implies u = 0\):

\[
(u, u) = \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) \hat{u}(n)^* \hat{u}(n) = \sum_{n \in \mathbb{Z}^2} \left(1 + |2\pi n|^2\right) \left|\hat{u}(n)\right|^2 \geq 0,
\]

and if \((u, u) = 0\) then \(|\hat{u}(n)| = 0\) for all \(n\) and from the definition (3.48) \(u = 0\). ■

This inner product yields two important lemmas of boundness.
Lemma 33. Provided the Hilbert space (3.49) with inner product (3.50), then

\[
\left( \int_\Omega |\nabla u|^2 \, dx \right)^{\frac{1}{2}} \leq |\Omega|^{\frac{1}{2}} \|u\|_{H^1},
\]

for all \( u \in H^1(\Omega) \).

The proof is put in the appendix due to its length.

Lemma 34. Provided the Hilbert space (3.49) with inner product (3.50), then

\[
\left( \int_\Omega |u|^2 \, dx \right)^{\frac{1}{2}} \leq |\Omega|^{\frac{1}{2}} \|u\|_{H^1},
\]

for all \( u \in H^1(\Omega) \).

Proof. Using the Fourier series (3.49) yields

\[
\left( \int_\Omega |u|^2 \, dx \right)^{\frac{1}{2}} = \left( \int_\Omega \sum_{n \in \mathbb{Z}^2} |\hat{u}(n) e^{2\pi i n \cdot x}|^2 \, dx \right)^{\frac{1}{2}},
\]

and using the norm for complex numbers, we have since the integrand is constant in \( x \)

\[
\left( \int_\Omega |u|^2 \, dx \right)^{\frac{1}{2}} = \left( |\Omega| \sum_{n \in \mathbb{Z}^2} |\hat{u}(n)|^2 \right)^{\frac{1}{2}}.
\]

Finally this can be written as

\[
\left( \int_\Omega |u|^2 \, dx \right)^{\frac{1}{2}} \leq \left( |\Omega| \sum_{n \in \mathbb{Z}^2} |\hat{u}(n)|^2 + |\hat{u}(n)|^2 |2\pi n|^2 \right)^{\frac{1}{2}} = |\Omega|^{\frac{1}{2}} \|u\|_{H^1},
\]

where we used the norm given by (3.49).

The Lemmas 33 and 34 will be crucial when showing boundness of operators, which is of importance due to the use of Riesz representations theorem.
3.3.2. TE-waves

We are given the set disjunction of $\Omega$ as $\Omega_1 \cup \Omega_2$ where $\Omega_1 \cap \Omega_2 = 0$ and use the notation $\nabla_k = \nabla + ik$, for the by $k \in [-\pi, \pi]^2$ shifted gradient operators. In this thesis we have that

\[
\begin{align*}
\Omega &= [0, 1]^2, \\
\Omega_2 &= [0.1, 0.9]^2, \\
\Omega_1 &= \Omega \setminus \Omega_2.
\end{align*}
\]

The nonlinear eigenvalue problem is then defined as.

\[
T(\omega) u := -\nabla_k \cdot \left( \frac{1}{\epsilon(x, \omega)} \nabla_k u \right) - \omega^2 u = 0, \tag{3.51}
\]

see (3.46). Where $\omega$ is the sought eigenvalue, $k \in [-\pi, \pi]^2$ is some known parameter, and $u(x) \in H^1(\Omega)$ is a periodic eigenfunction in $\Omega$. The problem is now to find solutions to (3.51) provided that

\[
\epsilon(x, \omega) = \epsilon_1 \chi_1(x) + \epsilon_2(\omega) \chi_2(x), \tag{3.52}
\]

where $\epsilon_1 > 0$ is constant and $\epsilon_2(\omega)$ is a function in $\omega$. $\chi_1$ and $\chi_2$ are the indicator functions defined as usual

\[
\chi_1 = \begin{cases} 1 & \text{for } x \in \Omega_1, \\
0 & \text{for } x \in \Omega_2, \end{cases} \quad \chi_2 = \begin{cases} 0 & \text{for } x \in \Omega_1, \\
1 & \text{for } x \in \Omega_2. \end{cases}
\]

We can write (3.52) as

\[
\frac{1}{\epsilon(x, \omega)} = \frac{1}{\epsilon_1 \chi_1(x) + \epsilon_2(\omega) \chi_2(x)},
\]

and then

\[
\frac{1}{\epsilon(x, \omega)} = \frac{1}{\epsilon_1} \chi_1 + \frac{1}{\epsilon_2(\omega)} \chi_2(x). \tag{3.53}
\]

$\epsilon_2(\omega)$ defined as the rational expression:

\[
\epsilon_2(\omega) := 1 + \frac{b}{c - i\gamma \omega - \omega^2} = \frac{b + c - i\gamma \omega - \omega^2}{c - i\gamma \omega - \omega^2}, \tag{3.54}
\]

where $c > 0$ and $\gamma \geq 0$. The $\gamma$ term determines damping and $\gamma = 0$ will result in a non-damped problem. Aiming to preform a linearization of this problem to express
as a linear eigenvalue problem we write it on weak form. Multiplying by any vector \( v^* \in H^1(\Omega) \), and integrate over \( \Omega \) yields
\[
\int_{\Omega} -\nabla_k \cdot \left( \frac{1}{\epsilon(x, \omega)} \nabla_k u \right) v^* \, dx = \int_{\Omega} \omega^2 u v^* \, dx.
\]
Using the definition of \( \nabla_k \) to obtain
\[
\int_{\Omega} -\nabla \cdot \left( \frac{1}{\epsilon(x, \omega)} (\nabla + ik) u \right) v^* - ik \cdot \left( \frac{1}{\epsilon(x, \omega)} (\nabla + ik) u \right) v^* \, dx = \int_{\Omega} \omega^2 u v^* \, dx.
\]
Rearranging the terms and integrating by parts yields
\[
- \int_{\partial\Omega} n \cdot \left( \frac{1}{\epsilon(x, \omega)} (\nabla + ik) u \right) v^* \, dx + \int_{\Omega} \left( \frac{1}{\epsilon(x, \omega)} (\nabla + ik) u \right) \cdot \nabla v^* - \left( \frac{1}{\epsilon(x, \omega)} (ik \cdot \nabla - |k|^2) u \right) v^* \, dx = \int_{\Omega} \omega^2 u v^* \, dx.
\]
The function \( v \in H^1(\Omega) \) is periodic on the boundary, follows from the definition (3.49) and this yields that
\[
\int_{\Omega} \left( \frac{1}{\epsilon(x, \omega)} \right) \nabla u \cdot \nabla v^* \, dx + i \int_{\Omega} \frac{1}{\epsilon(x, \omega)} uk \cdot \nabla v^* \, dx
- i \int_{\Omega} \left( \frac{1}{\epsilon(x, \omega)} k \cdot \nabla u \right) v^* \, dx + \int_{\Omega} \left( \frac{1}{\epsilon(x, \omega)} |k|^2 u \right) v^* \, dx = \int_{\Omega} \omega^2 u v^* \, dx. \quad (3.55)
\]
(3.53) says \( \epsilon(x, \omega) \) is piecewise constant in space, so when moving the gradient in the weak formulation \( \frac{1}{\epsilon(x, \omega)} \) can be seen as a constant, it is periodic on the boundary on both \( \Omega_1 \) and \( \Omega_2 \). Integrate by parts to obtain
\[
- i \int_{\Omega} \left( \frac{1}{\epsilon(x, \omega)} k \cdot \nabla u \right) v^* \, dx = i \int_{\Omega} \frac{1}{\epsilon(x, \omega)} uk \cdot \nabla v^* \, dx - i \int_{\partial\Omega} \frac{1}{\epsilon(x, \omega)} n \cdot ku v^* \, dx.
\]
Due to the periodic boundary condition, (3.55) gives
\[
\int_{\Omega} \frac{1}{\epsilon(x, \omega)} \nabla u \cdot \nabla v^* \, dx + 2i \int_{\Omega} \frac{1}{\epsilon(x, \omega)} uk \cdot \nabla v^* \, dx + \int_{\Omega} \frac{1}{\epsilon(x, \omega)} |k|^2 u v^* \, dx
= \omega^2 \int_{\Omega} uv^* \, dx. \quad (3.56)
\]
(3.53) can then from identity (3.56) be used to obtain
\[
\int_{\Omega_1} \frac{1}{\epsilon_1} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega_1} \frac{1}{\epsilon_1} uk \cdot \nabla v^* dx + \int_{\Omega_1} \frac{1}{\epsilon_1} |k|^2 uv^* dx + \\
\int_{\Omega_2} \frac{c - i\gamma \omega - \omega^2}{b + c - i\gamma \omega - \omega^2} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega_2} \frac{c - i\gamma \omega - \omega^2}{b + c - i\gamma \omega - \omega^2} uk \cdot \nabla v^* dx \\
+ \int_{\Omega_2} \frac{c - i\gamma \omega - \omega^2}{b + c - i\gamma \omega - \omega^2} |k|^2 uv^* dx = \omega^2 \int_{\Omega} uv^* dx. \quad (3.57)
\]

The last expression (3.57) can be written as
\[
\frac{1}{\epsilon_1} \left( \int_{\Omega_1} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega_1} uk \cdot \nabla v^* dx + \int_{\Omega_1} |k|^2 uv^* dx \right) + \\
\frac{c - \gamma i\omega - \omega^2}{b + c - \gamma i\omega - \omega^2} \left( \int_{\Omega_2} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega_2} uk \cdot \nabla v^* dx \\
+ \int_{\Omega_2} |k|^2 uv^* dx \right) = \omega^2 \int_{\Omega} uv^* dx. \quad (3.58)
\]

Define the sesquilinear forms on \( H^1(\Omega) \times H^1(\Omega) \) as
\[
s_1[u, v] = \int_{\Omega_1} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega_1} uk \cdot \nabla v^* dx + \int_{\Omega_1} |k|^2 uv^* dx, \quad (3.59)
\]
\[
s_2[u, v] = \int_{\Omega_2} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega_2} uk \cdot \nabla v^* dx + \int_{\Omega_2} |k|^2 uv^* dx,
\]
\[
m[u, v] = \int_{\Omega} uv^* dx.
\]

**Proposition 35.** The sesquilinear form \( s_2[u, v] \) from (3.59) is Hermitian

**Proof.**
\[
s_2[u, v]^* = \left( \int_{\Omega_2} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega_2} uk \cdot \nabla v^* dx + \int_{\Omega_2} |k|^2 uv^* dx \right)^*,
\]
which can be written as
\[
s_2[u, v]^* = \int_{\Omega_2} (\nabla u \cdot \nabla v^*)^* dx - 2i \int_{\Omega_2} (uk \cdot \nabla v^*)^* dx + \int_{\Omega_2} (|k|^2 uv^*)^* dx.
\]
The conjugate property yields

\[ s_2 [u, v]^* = \int_{\Omega_2} \nabla u^* \cdot \nabla v^* dx - 2i \int_{\Omega_2} u^* k \cdot \nabla v^* dx + \int_{\Omega_2} |k|^2 u^* v^* dx. \]

The last step is by integrate the second term by part and using that \( v \) is periodic on \( \Omega \) to obtain

\[ s_2 [u, v]^* = \int_{\Omega_2} \nabla u^* \cdot \nabla v^* dx + 2i \int_{\Omega_2} \nabla u \cdot kv^* dx + \int_{\Omega_2} |k|^2 u^* v^* dx, \]

thus

\[ s_2 [u, v]^* = s_2 [v, u]. \]

This property is of interest as we will later use this property in the linearization. With these definitions, (3.58) can be written as

\[ \frac{1}{\epsilon_1} s_1 [u, v] + \frac{c - i\gamma \omega - \omega^2}{b + c - \omega^2 - i\gamma \omega} s_2 [u, v] = \omega^2 m [u, v], \]

or

\[ \frac{1}{\epsilon_1} s_1 [u, v] + s_2 [u, v] - \frac{b}{b + c - \omega^2 - i\gamma \omega} s_2 [u, v] = \omega^2 m [u, v]. \]

By setting \( \frac{1}{\epsilon_1} s_1 [u, v] + s_2 [u, v] = s_3 [u, v] \) and \( b + c = c_1 \), the following system is obtained:

\[ s_3 [u, v] - \frac{b}{c_1 - \omega^2 - i\gamma \omega} s_2 [u, v] = \omega^2 m [u, v]. \] (3.60)

We now show that the sesquilinear forms \( s_2, s_3 \) and \( m \) defined in (3.59) is bounded in \( H^1 (\Omega) \).

\[ |s_2 [u, v]| = \left| \int_{\Omega_2} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega_2} uk \cdot \nabla v^* dx + \int_{\Omega_2} |k|^2 uv^* dx \right|. \]

The triangle inequality yields

\[ |s_2 [u, v]| \leq \int_{\Omega_2} |\nabla u \cdot \nabla v^*| dx + 2 \int_{\Omega_2} |uk \cdot \nabla v^*| dx + |k|^2 \int_{\Omega_2} |uv^*| dx. \]
Using the Cauchy-Schwartz inequality for $L^2$ functions, works as $H^1(\Omega) \subset L^2(\Omega)$,
\[
|s_2[u, v]| \leq \left( \int_{\Omega_2} |\nabla u|^2 \, dx \cdot \int_{\Omega_2} |\nabla v|^2 \, dx \right)^{\frac{1}{2}} + |k|^2 \left( \int_{\Omega_2} |u|^2 \, dx \cdot \int_{\Omega_2} |v|^2 \, dx \right)^{\frac{1}{2}} + 2 |k| \left( \int_{\Omega_2} |u|^2 \, dx \int_{\Omega_2} |\nabla v|^2 \, dx \right)^{\frac{1}{2}}. \tag{3.61}
\]

Lemma 33 and Lemma 34 together with (3.61), since the integrands are positive we can integrate over the whole area to be able to use the norms as upper bounds, yields
\[
|s_2[u, v]| \leq |\Omega| \|u\| \|v\| + |k|^2 |\Omega| \|u\| \|v\| + 2 |k| |\Omega| \|u\| \|v\|.
\]
Equivalently
\[
|s_2[u, v]| \leq |\Omega| (|k|^2 + 2 |k| + 1) \|u\| \|v\|,
\]
and thus $s_2[u, v]$ is bounded. Boundness of $s_3[u, v]$ follows similarly by showing that $s_1[u, v]$ is bounded, thus $s_3[u, v]$ is so as well. From (3.59) we have
\[
|s_1[u, v]| = \left| \int_\Omega \nabla u \cdot \nabla v^* \, dx + 2i \int_\Omega uk \cdot \nabla v^* \, dx + \int_\Omega |k|^2 uv^* \, dx \right|,
\]
by triangle inequality
\[
|s_1[u, v]| \leq \left| \int_\Omega \nabla u \cdot \nabla v^* \, dx \right| + 2 \left| \int_\Omega uk \cdot \nabla v^* \, dx \right| + \left| \int_\Omega |k|^2 uv^* \, dx \right|.
\]
Using the continuous version of the triangle inequality yields
\[
|s_1[u, v]| \leq \int_\Omega |\nabla u \cdot \nabla v^*| \, dx + 2 \int_\Omega |uk \cdot \nabla v^*| \, dx + \int_\Omega \left| |k|^2 uv^* \right| \, dx,
\]
and by Cauchy–Schwarz we have
\[
\leq \left( \int_{\Omega_1} |\nabla u|^2 \, dx \int_{\Omega_1} |\nabla v|^2 \, dx \right)^{\frac{1}{2}} + 2 |k| \left( \int_{\Omega_1} |u|^2 \, dx \int_{\Omega_1} |\nabla v|^2 \, dx \right)^{\frac{1}{2}} + |k|^2 \left( \int_{\Omega_1} |u|^2 \int_{\Omega_1} |v|^2 \, dx \right)^{\frac{1}{2}}. \tag{3.62}
\]
Using Lemma 33 and Lemma 34 we have

\[ |s_1[u, v]| \leq |\Omega| \|u\| \|v\| + 2 |k| |\Omega| \|u\| \|v\| + |k|^2 |\Omega| \|u\| \|v\| = |\Omega| (|k|^2 + 2 |k| + 1) \|u\| \|v\|.\]

From the definition of \( s_3[u, v] \) it follows,

\[ |s_3[u, v]| \leq \frac{1}{\epsilon_1} |s_1[u, v]| + |s_2[u, v]| = |\Omega| \left( 1 + \frac{1}{\epsilon_1} \right) (|k|^2 + 2 |k| + 1) \|u\| \|v\|,\]

and since \( \epsilon_1 > c_2 \), \( |s_3[u, v]| \) is bounded. We finally show that \( m[u, v] \) is bounded, using (3.59),

\[ |m[u, v]| = \left| \int_{\Omega} uv^* dx \right| \leq \int_{\Omega} |uv^*| dx,\]

and Cauchy-Schwarz inequality gives that

\[ |m[u, v]| \leq \left( \int_{\Omega} |u|^2 dx \int_{\Omega} |v|^2 dx \right)^{\frac{1}{2}}.\]

Using Lemma 34 boundness follows as

\[ |m[u, v]| \leq |\Omega| \|u\| \|v\|.\]

\( s_2[u, v], s_3[u, v] \) and \( m[u, v] \) are thus bounded sesquilinear forms, sesquilinearity follows trivially from the linearity of the integral operator. Now use Riesz-representation theorem to be able to write

\[ (\widetilde{B}_{TE} u, v) = s_2[u, v], \quad (A_{TE} u, v) = s_3[u, v], \quad (C_{TE} u, v) = m[u, v],\] (3.63) (3.64)

for some linear bounded operators \( A_{TE}, \widetilde{B}_{TE}, C_{TE} : H^1(\Omega) \to H^1(\Omega) \). Using these properties gives that (3.60) can be written as

\[ (A_{TE} u, v) - \frac{b}{c_1 - \omega^2 - i\gamma \omega} (\widetilde{B}_{TE} u, v) = \omega^2 (C_{TE} u, v),\]

for all \( u, v \in H^1(\Omega) \). Due to linearity of inner products this can be written as

\[ (A_{TE} u - \frac{b}{c_1 - \omega^2 - i\gamma \omega} \widetilde{B}_{TE} u - \omega^2 C_{TE} u, v) = 0.\]
We then have that as this should be satisfied for all $v$ especially for $H^1(\Omega) \ni v = Au - \frac{b}{c_1 - \lambda} \hat{B}u - \lambda Cu$, due inner product property $(v, v) = 0 \implies v = 0$ it follows that

$$Q_{TE}(\omega) := A_{TE}u - \frac{b}{c_1 - \omega^2 - i\gamma\omega} \hat{B}_{TE}u - \omega^2 C_{TE}u = 0, \quad (3.65)$$

or

$$A_{TE}u - \frac{b}{c_1 - \omega^2 - i\gamma\omega} \hat{B}_{TE}u = \omega^2 C_{TE}u.$$ 

This is a familiar eigenvalue problem recalling the finite problem (3.25). The problem defined as (3.51) has been rewritten as (3.65). However, we will leave this problem and return for further properties in Subsection 3.3.4.

### 3.3.3. TM-waves

This section concerns the related problem of TM-waves, where the eigenvalue problem is

$$T(\omega)u := -\nabla_k \cdot \nabla_k u - \omega^2 \epsilon(x, \omega) u = 0, \quad (3.66)$$

see (3.47). We use the same notation and Hilbert space $H^1(\Omega)$ as in (3.51). Multiplying by a test function $v^* \in H^1(\Omega)$ and integrate by parts to obtain a weak solution results in the equation

$$\int_{\Omega} \nabla_k u \cdot \nabla_k v^* dx - \int_{\Omega} \omega^2 \epsilon(x, \omega) uv^* dx = 0, \quad (3.67)$$

where $v^*$ is the conjugate of the function $v$. By integrating by parts and using the periodicity this can be written in the form

$$\int_{\Omega} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega} u k \cdot \nabla v^* dx + \int_{\Omega} |k|^2 uv^* dx - \int_{\Omega} \omega^2 \epsilon(x, \omega) uv^* dx = 0. \quad (3.68)$$

From the definition of $\epsilon$ (3.54),

$$\int_{\Omega} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega} u k \cdot \nabla v^* dx + \int_{\Omega} |k|^2 uv^* dx$$

$$- \int_{\Omega_1} \omega^2 \epsilon_1 uv^* dx - \int_{\Omega_2} \omega^2 \left(1 + \frac{b}{c_1 - \omega^2 - i\gamma\omega}\right) uv^* dx = 0.$$

Note that

$$\omega^2 \left(1 + \frac{b}{c - \omega^2 - i\gamma\omega}\right) = \left(\omega^2 + \frac{b\omega^2 - bc + i\gamma\omega}{c - \omega^2 - i\gamma\omega} + \frac{bc - ib\gamma\omega}{c - \omega^2 - i\gamma\omega}\right).$$
and thus
\[ \omega^2 \left( 1 + \frac{b}{c - \omega^2 - i\gamma\omega} \right) = \left( \omega^2 - b + \frac{bc - ib\gamma\omega}{c - \omega^2 - i\gamma\omega} \right). \]

Using this in the problem it follows that
\[ \int_\Omega \nabla u \cdot \nabla v^* dx + 2i \int_\Omega uk \cdot \nabla v^* dx + \int_\Omega |k|^2 uv^* dx 
- \omega^2 \epsilon_1 uv^* dx - \int_{\Omega_2} \left( \omega^2 - b + \frac{bc - ib\gamma\omega}{c - \omega^2 - i\gamma\omega} \right) uv^* dx = 0, \]
and by rearranging the terms
\[ \int_\Omega \nabla u \cdot \nabla v^* dx + 2i \int_\Omega uk \cdot \nabla v^* dx + \int_\Omega |k|^2 uv^* dx + \int_{\Omega_2} buv^* dx 
- \omega^2 \left( \int_{\Omega_1} \epsilon_1 uv^* dx + \int_{\Omega_2} uv^* dx \right) - \frac{bc - ib\gamma\omega}{c - \omega^2 - i\gamma\omega} \int_{\Omega_2} uv^* dx = 0. \]

Define the sesquilinear forms on \( H^1(\Omega) \times H^1(\Omega) \) as
\[ t_1 [u, v] := \int_\Omega \nabla u \cdot \nabla v^* dx + 2i \int_\Omega uk \cdot \nabla v^* dx + \int_\Omega |k|^2 uv^* dx + \int_{\Omega_2} buv^* dx, \]
(3.69)
\[ t_2 [u, v] := \int_{\Omega_1} \epsilon_1 uv^* dx + \int_{\Omega_2} uv^* dx, \]
\[ m_2 [u, v] := \int_{\Omega_2} uv^* dx. \]

**Proposition 36.** The sesquilinear form \( m_2 [u, v] \) from (3.59) is Hermitian.

**Proof.** Follows directly from the definition in (3.69) \( \blacksquare \)

This property is important in the numerical linearization of the problem. These forms yields the system
\[ t_1 [u, v] - t_2 [u, v] = \frac{bc - ib\gamma\omega}{c - \omega^2 - i\gamma\omega} m_2 [u, v]. \]
(3.70)

We can see that this is the same structure on this problem as in (3.59) so if these forms are bounded, they can be written as inner products and get a result similar
to that in (3.65). Now we have from the triangle- and Cauchy Schwarz inequalities
\[
|t_1[u,v]| \leq \sqrt{\int_\Omega |\nabla u|^2 \, dx} \sqrt{\int_\Omega |\nabla v|^2 \, dx} + 2|k| \sqrt{\int_\Omega |u|^2 \cdot dx} \sqrt{\int_\Omega |\nabla v|^2 \, dx}
+ |k|^2 \sqrt{\int_\Omega |u|^2 \, dx} \sqrt{\int_\Omega |v|^2 \, dx} + b \sqrt{\int_\Omega |u|^2 \, dx} \sqrt{\int_\Omega |v|^2 \, dx}.
\]

From Lemma 33, Lemma 34 we have
\[
\leq |\Omega| \|u\| \|v\| + 2|k| |\Omega| \|u\| \|v\| + |k|^2 |\Omega| \|u\| \|v\| + b |\Omega| \|u\| \|v\|
= |\Omega| (1 + 2|k| + |k|^2 + b) \|u\| \|v\|,
\]
thus \(t_1[u,v]\) is bounded. For \(t_2[u,v]\) it can be shown that
\[
|t_2[u,v]| \leq \epsilon_1 \sqrt{\int_\Omega |u|^2 \, dx} \sqrt{\int_\Omega |v|^2 \, dx} + \sqrt{\int_\Omega |u|^2 \, dx} \sqrt{\int_\Omega |v|^2 \, dx},
\]
and then
\[
\leq \epsilon_1 |\Omega| \|u\| \|v\| + |\Omega| \|u\| \|v\| = |\Omega| (\epsilon_1 + 1) \|u\| \|v\|.
\]
For \(m_2[u,v]\) we get that boundness follows as
\[
m_2[u,v] \leq \sqrt{\int_\Omega |u|^2 \, dx} \sqrt{\int_\Omega |v|^2 \, dx} \leq |\Omega| \|u\| \|v\|.
\]
All these operators are bounded and we can from Riesz-representation get an expression just as in (3.65):
\[
\begin{align*}
(\hat{B}_{TM} u, v) &= m_2[u,v], \\
(A_{TM} u, v) &= t_1[u,v], \\
(C_{TM} u, v) &= t_2[u,v].
\end{align*}
\]
Hence the rational problem can be written as
\[
Q_{TM}(\omega) u := A_{TM} u - \frac{b(c - i\gamma\omega)}{c - \omega^2 - i\gamma\omega} \hat{B}_{TM} u - \omega^2 C_{TM} u = 0,
\]
in the same way as (3.65).
3.3.4. Properties of TM and TE waves

We have in the two previous part found similar ways of expressing the TE-waves problem (3.51) and the TM-waves problem (3.66), now some properties of these two problems will be proven. It turns out that one property that can be proved for TM case with the presented method fails for the TE case, the result might however still be true. Since TM will behave more nicely this case will be considered first. The first important property is being a Fredholm operator, this is a very important property when preforming finite approximation of the problem as it gives us that no eigenvalue is of infinite multiplicity.

**Definition 37.** Given a linear operator \( T : H^1 \to H^2 \) where \( H^1 \) and \( H^2 \) are vector spaces.

\[
\ker (T) = \{ v \in H^1 : Tv = 0 \}
\]

**Definition 38.** Given a linear operator \( T : H^1 \to H^2 \) where \( H^1 \) and \( H^2 \) are vector spaces.

\[
\text{coker} (T) = \{ u \in H^2 : \forall v \in H^1 \, Tv \neq u \}
\]

**Definition 39.** An operator \( T \) is Fredholm if

\[
\dim (\ker (T)) < \infty
\]

and

\[
\dim (\text{coker} (T)) < \infty
\]

the Fredholm index is the number given by \( \dim (\ker (T)) - \dim (\text{coker} (T)) \).

There are different ways of classifying the different types of eigenvalues. One usual way is to divide them into the point spectrum \( \sigma_p \), which consists of the eigenvalues to the operator, the continuous spectrum \( \sigma_c \) and the residual spectrum \( \sigma_r \). They are disjoint and their union span the spectrum [5, page 371]. In this thesis we will however not use these classification due to the fact that if we can show that some \( \lambda \in \sigma_p \) we don’t know if \( \lambda I - T \) is a Fredholm operator, \( \sigma_p \) might have a infinite mutiplicity. Thus the classification used in this thesis states:

**Definition 40.** The set \( \sigma_{\text{ess}} (T) := \{ \lambda : \lambda I - T \text{ is not a Fredholm operator} \} \) is called the essential spectrum.

**Definition 41.** The set \( \sigma_{\text{discr}} (T) = \sigma (T) \setminus \sigma_{\text{ess}} (T) \) is called the discrete spectra.
It’s obvious that with this classification we have divided in the spectrum after what properties we want to use. It follows that $\lambda \in \sigma_{discr} (T)$ implies that $\lambda$ is an eigenvalue of finite multiplicity.

The TM-waves have been investigated in [16] so here the results are only stated.

**Theorem 42.** [16, page 9 & 12] Theorem 4.1 and Lemma 4.8. The operator valued function $Q_{TM} (\omega)$ defined as (3.73) has no real eigenvalues and is a Fredholm operator of index 0 for all $\omega$. Moreover, the spectrum consist of countably many isolated eigenvalues of finite multiplicity.

The TE-waves case is more carefully considered here, and thus the proof will be rather long. We now want to show that (3.65) and in the extension (3.51) are Fredholm operators of index 0 for each $\omega \in \mathbb{C}$.

**Lemma 43.** Suppose we have the operator valued function $Q_{TE} (\omega)$ defined as (3.65) then $\sigma (Q_{TE}) \subset W (Q_{TE})$.

We give the proof in the appendix.

**Theorem 44.** The operator valued function defined as (3.65) is Fredholm of index 0 for all $\omega$.

**Proof.** To prove this we start by showing that the operator $D : H^1 (\Omega) \to H^1 (\Omega)$ defined by,

$$(Du, u) := \int_\Omega \nabla u \cdot \nabla u^* dx,$$

is Fredholm of index 0, such an operator exist due to Riesz representantion theorem and that the integral is bounded. The first step is to show that the kernel has dimension 1.

$$(Du, u) := \int_\Omega \nabla u \cdot \nabla u^* dx = \int_\Omega \left| \nabla \sum_{n \in \mathbb{Z}^2} \hat{u} (n) e^{2\pi i n \cdot x} \right|^2 dx \in \mathbb{C},$$

from the definition the complex norm and the Fourier coefficients. Now if this should be equal to zero, it has to be equal to zero for all $x \in \Omega$ it now follows that

$$\nabla \sum_{n \in \mathbb{Z}^2} \hat{u} (n) e^{2\pi i n \cdot x} = 0 \implies u = \hat{u} (0),$$
since otherwise $|\nabla u| > 0$ for some $x$. So $u$ is a constant function, thus a basis for the null space of $(Du, u)$ is
\[ \text{span} \{1\}, \]
which yields
\[ \dim (\ker (D)) = 1. \]
The second step is computing the dimension of the co-kernel. Using the identity, $\text{coker} (D) = \ker (D^*)$ and that the adjoint operator to $D$ is $D$ as it is Hermitian. Thus we can once again note that
\[ (Du, u) = \int_{\Omega_2} \nabla u \cdot \nabla u^* dx = 0 \implies u = \hat{u}_0, \]
and it follows that $\dim (\text{coker} (D)) = 1$ and that $D$ is a Fredholm operator of index, $1 - 1 = 0$. The operators $A$, $\hat{B}$ and $C$ from (3.65) are generated from integrals of the same and lower order differentials as $D$. This results in that the sum is Fredholm of index 0 as well so $A$, $\hat{B}$ and $C$ are Fredholm of order 0 due to them being embedded in $D$. $Q_{TE} (\omega)$ is just a sum of $A$, $\hat{B}$ and $C$ depending on constant $\omega$ so $Q_{TE} (\omega)$ is Fredholm of index 0.

The Theorems 42 and 44 states thus that both the TM-wave problem represented by the operator valued function (3.73) and TE-wave problem represented by the operator valued function (3.65) are Fredholm of order 0. Still if we want to be able to go smoothly over to the finite dimension approximation we would like to know that the eigenvalues does not cluster around certain values. If we can show the requirements for the analytic Fredholm we know that we only have a discrete set of eigenvalues in $\mathbb{C}$ and hence we know that the eigenvalues does not cluster.

The analytic Fredholm theorem states, according to [14, page 266],

**Theorem 45.** The analytic Fredholm theorem: Suppose the mapping $B (\omega)$ is analytic and compact for each $\omega \in \mathbb{C}$ then, either
\[ (I - B (\omega))^{-1} \text{ exist for no } \omega \in \mathbb{C}, \]
or
\[ (I - B (\omega))^{-1} \text{ exist for all } \omega \in \mathbb{C}\setminus S \text{ where } S \text{ is a discrete complex set.} \]
The TM-wave eigenvalue problem (3.73) can be written as

\[
I - A_{TM}^{-1} \left( \frac{b(c - i\gamma \omega)}{c - \omega^2 - i\gamma \omega} \widehat{B}_{TM} - \omega^2 C_{TM} \right) \equiv 0, \tag{3.74}
\]

by taking the inverse operator of \(A_{TM}\). This inverse exists due to (3.71) and (3.69) and that \((A_{TM}u, u) \geq K \Vert u \Vert^2\) for some constant \(K > 0\), the last part follows from a proof similar to that of Lemma 43 in the appendix. The operators \(\widehat{B}_{TM}\) and \(C_{TM}\) are compact operators which follow trivially from the definitions (3.71) and (3.69). Hence

\[
\frac{b(c - i\gamma \omega)}{c - \omega^2 - i\gamma \omega} \widehat{B}_{TM} - \omega^2 C_{TM},
\]

is compact as well. \(A_{TM}^{-1}\) in linear and bounded since \(A_{TM}\) is linear and bounded. This leads to

\[
A_{TM}^{-1} \left( \frac{b(c - i\gamma \omega)}{c - \omega^2 - i\gamma \omega} \widehat{B}_{TM} - \omega^2 C_{TM} \right),
\]

is compact and thus is (3.74) on the desired form for applying the analytic Fredholm Theorem. From Theorem 42 it follows that (3.74) has no real eigenvalues, hence the second property of the analytic Fredholm Theorem is granted. (3.74) and in the extent (3.73) has separated eigenvalues which is good when we are applying numerical methods for solving the eigenvalue problems. The TE-waves case it we will not be able to use the same method. We can still do the same approach, however, the TE-wave eigenvalue problem (3.65) can be written as

\[
I - A_{TE}^{-1} \left( \frac{b}{c_1 - \omega^2 - i\gamma \omega} \widehat{B}_{TE} - \omega^2 C_{TE} \right) u = 0, \tag{3.75}
\]

by taking the inverse operator of \(A_{TE}\). This inverse exists due to (3.59) and (3.63) and that \((A_{TE}u, u) \geq K \Vert u \Vert^2\) for some constant \(K > 0\). In the TE-waves case, however, \(\widehat{B}_{TE}\) is not compact operator follows easily from (3.59) and (3.63). Thus

\[
A_{TE}^{-1} \left( \frac{b}{c_1 - \omega^2 - i\gamma \omega} \widehat{B}_{TE} - \omega^2 C_{TE} \right),
\]

is not compact if \(A_{TE}^{-1}\) is not a compact operator which is not proved nor disproved. This leads to that the analytic Fredholm Theorem can not be used to show that the set of eigenvalues are discrete with no cluster of eigenvalues, this may still be true, of course, but compactness of \(A_{TE}^{-1}\) is not investigated here. Since we didn’t
prove that we have a discrete set with eigenvalues to the TE-case we will not approximate the solutions by finite matrices. But we will still show the possible approach, for each matrix approximation we get the problem

\[ A_f u - B_f \frac{b}{c_1 - \omega^2 - i\gamma\omega} B_f^* u = \omega^2 C_f u, \]  

(3.76)

where \( B_f \in \mathbb{C}^{n \times n_2} \) is the Cholosky-complement matrix such that \( B_f^* B_f = \hat{B} \). Such a matrix \( B_f \) will exist due (3.63) and the fact that \( s_2 [u, v] \) is Hermitian from Proposition 35. The matrix will clearly be semi-positive definite as we have that

\[
s_2 [u, u] := \int_{\Omega_2} \nabla u \nabla u^* dx + 2i \int_{\Omega_2} uk \cdot \nabla u^* dx + \int_{\Omega_2} |k|^2 uu^* dx = \int_{\Omega_2} |\nabla_k u|^2 dx \geq 0.
\]

Another problem with the TE-waves is that the rational term in (3.76) does not reassemble that of (3.25), thus that linearization (3.42) will not be valid either. Therefore we will from now only work with the TM-waves. From (3.73) we approximate the operators by the matrices \( A_f, \hat{B}_f, C_f \in \mathbb{C}^{n \times n} \) (we use the same name of the matrices as the previous part as that approximation was only hypothetical). The problem can then be written as

\[ A_f u - \hat{B}_f \frac{b(c - i\gamma\omega)}{c - \omega^2 - i\gamma\omega} \hat{B}_f^* u = \omega^2 C_f u, \]  

(3.77)

where \( \hat{B}_f \in \mathbb{C}^{n \times n_2} \) is once again the operator such that \( \hat{B}_f^* \hat{B}_f = \hat{B} \). Existence is granted by (3.71). The form \( m_2 [u, v] \) is Hermitian from Proposition 36 and positive definite since

\[
m_2 [u, u] := \int_{\Omega_2} uu^* dx \geq 0.
\]

This gives us that numerical problem (3.77) is exactly on the form (3.25). It can thus be approximated by the linearization given there, if \( \gamma = 0 \), we can use the linearization given in (3.3) instead.
4. Numerical examples of the TM-wave

We will now, by computer assistance solve examples of the TM-case obtained from (3.66). In the examples we use the indata:

\[ \Omega = [0, 1]^2, \]
\[ a = 10, \]
\[ b = 5(2\pi)^2 \]
\[ c = (2\pi)^2 \]
\[ \epsilon_1 = 1. \]

The domain \( \Omega \) is divided in \( \Omega_1 \) and \( \Omega_2 \) as

\[ \Omega_2 = [0.1, 0.9]^2, \]
\[ \Omega_1 = \Omega \setminus \Omega_2. \]

The equation from (3.25) is

\[ T(\omega)u := Gu - \omega^2 \epsilon_1 M_1 u - \omega^2 \left( a + \frac{b}{c - \omega^2 - i\gamma\omega} \right) M_2 u = 0, \]

as can be seen this is not completely the same as (3.73) but for right choices of \( A_f, \tilde{B}_f \) and \( C_f \) they will be equivalent. The matrices \( G, M_1 \) and \( M_2 \) are given by the inner products (3.71) or more specifically they are according to (3.28), the unique matrices that satisfies

\[ A_f = G + bM_2, \]
\[ \tilde{B}_f = M_2, \]
\[ C_f = \epsilon_1 M_1 + aM_2. \]

The above backwards reasoning is due to that the used matrices \( G, M_1 \) and \( M_2 \) were computed in an earlier step of the linearization of the TM-wave problem.
The triangulation used for the weak forms is the easiest possible, and the order of the FEM polynomials is 8. The matrices $M_1$, $M_2$ and $G$ are not computed here but rather obtained from [12], the triangulation combined by the degree of the polynomial gives that the size of the matrices are $450 \times 450$. Before we show some examples, however, we will describe the FEM method used to obtain the these matrices as it is not the usual FEM. This due to properties that we want to have for the rational linearization.

### 4.1. Discontinuous Galerkin

In the rational linearization we don't want to use the standard FEM when we assemble the mass matrix $M_2$ since that would result in a matrix that is not necessary on the form $M_2 = \text{diag}(0, \widehat{M}_2)$ which is used both in (3.23) and (3.32). Thus the result if using rational linearization and a standard FEM would not decrease the size of the resulting system. It is good if the rank of $M_2$ is low as the system is of order $n + n_2$ in (3.23) and $2n + 2n_2$ in (3.32), where $n_2$ is $\text{rank } (M_2)$.

Therefore a different way of assembling the mass matrix $M_2$ is needed. We will use a Discontinuous Galerkin (DG) method from [12, page 182-183]. The idea of DG is to compute the $M_2$ from the weak form triangle-wise rather than by integrating over the whole domain. This will result in a discontinuous approximation of the solution. In our example, we approximate the TM-waves by $u_n$ a piecewise $p$-th degree polynomial, this yield the equation

$$-\nabla_k \cdot \nabla_k u_h = \omega^2 \varepsilon (x, \omega) u_h = 0.$$
The global solution can then be obtained as the direct sum of the solutions local over one triangle. This is written as [12, page 182-183],

\[
\int_{\Omega} \nabla u_h \cdot \nabla v_h^* dx - 2i \int_{\Omega} (k \cdot \nabla u_h) v_h^* dx + \int_{\Omega} |k|^2 u_h v_h^* dx - \\
\int_{\mathcal{E}} \{ \nabla u_h \} \cdot [v_h^*] dx - \int_{\mathcal{E}} \{ \nabla v_h \} \cdot [u_h^*] dx \\
+ \int_{\mathcal{E}} \beta [u_h] \cdot [v_h^*] dx = \omega^2 \int_{\Omega} \epsilon (x, \omega) u_h v_h^* dx,
\]

where \( \mathcal{E} \) is the set of edges in the triangulation and

\[
\{ \nabla w \} = \frac{w_+ + w_-}{2}, \quad [w] = w_+ n_+ + w_- n_-,
\]

are standard notations for jumps and in function value and mean of the gradient at the triangle edges. A numerical method on the equation above can be shown to be symmetric and stable if \( \beta \) is large enough, and we use

\[
\beta = 20 \frac{(p + 1)(p + 2) |\partial K|}{2 |K|},
\]

where \( |\partial K| \) is the length of the edge and \( |K| \) is the area of the smaller adjacent triangle. The equation

\[
Gu - \omega^2 \epsilon_1 M_1 + \omega^2 \epsilon_2 (\omega) M_2,
\]

follows from the definition of \( \epsilon (x, w) \) in (3.52). \( M_2 = \left(0, \widetilde{M}_2\right) \) and \( M_1 \left(0, \widetilde{M}_1\right) \), this is also due to (3.52), furthermore \( \widetilde{M}_2, \widetilde{M}_1 \) are positive definite matrices. Thus the linearization (3.42) or if non-damped (3.24) can be used and will give a relatively small solution since the rank of \( M_2 \) is low (the rank of \( M_1 \) is also low but this is not important for the linearization).

4.2. Eigenvalues of TM-waves

From the matrices \( G, M \) and \( M_2 \) obtained by the Discontinuous Galerkin method and the parameters (4.1), either a polynomial linearization or a rational linearization can be preformed. Only the values obtained by the polynomial linearization
are shown, the pseudospectra, however, is computed by both methods. To be able to get an overview of the results will we show first show the eigenvalues for each damping. The codes used to compute the linearization and solve the eigenvalue problems are put in the appendix.

First consider the non-damped case when $\gamma = 0$ and the polynomial linearization technique (3.23) is used. The linearization is given by matlab routine TM_Poly_linearization_No_damping.m. The routine uses 30 different values $k \in [-\pi, \pi]^2$ placed around the boundary of the triangle with vertices origin, $(\pi, 0)$ and $(\pi, \pi)$ in that order. For the TM-waves $G$ is affected by the value of $k$ this can be seen in (3.69). The labeling of the $x$-axis is due to that along each of the three edges, $x_1$ or $x_2$ varies by $\pi$, thus we have the interval $[0, 3\pi]$ for the eigenvalues.

![Eigenvalues for non-damped problem](image)

Figure 4.1: The eigenvalues for $k$ along the triangle. The horizontal lines denote the vertices of the triangle.

We have in Figure 4.1 not denoted the imaginary values as they are zero for a symmetric problem. We can see that for $\omega \approx 2.8$ we have a gap where we have no eigenvalues for any value of $k$ this will be of interest in the application of photonic crystals as it shows that those frequencies will be reflected, similar results will be seen for the damped systems as well. For the damped systems we also have an imaginary part which is also plotted. Figure 4.2 shows the result for damping coefficient $\gamma = 1$, the linearization method used is (3.32) written as matlab script TM_Poly_linearization_damping.m.
Figure 4.2: The eigenvalues for $\gamma = 1$, the imaginary part is actually negative but the absolute value is used to visually show its resemblance of the real part.

Figure 4.3 show the eigenvalues for damping $\gamma = 3$, the same function and linearization is used for all damped cases, the imaginary part is always shown as the absolute value. The least imaginary part is corresponding to the least real in each of the graphs. The other real and imaginary parts are paired accordingly. Figure 4.4 display the case when $\gamma = 9$ is the damping coefficient.

The last value of damping used is $\gamma = 13$, and seen Figure 4.5.

From these Figures one can see that the real part for the lower eigenvalues are not so affected by the damping while the imaginary part increases for high damping. This was expected as a higher damping coefficient gives a matrix with larger departure from normality follows from the linearization (3.42). One interesting thing is that for the real system the band gap was not particular affected by change of damping coefficient $\gamma$, at least for our values of damping. Thus for the choice (4.1) of parameters we have a reflection of waves with frequencies $\omega \approx 2.8$, this can be seen in all Figures.
4.3. Pseudospectra of TM-waves

To see how sensitive these solutions the pseudospectra is computed for \( k = \left( \frac{\pi}{100}, \frac{\pi}{100} \right) \) for each of the considered damping. To be able to more easily understand what the pseudospectra states we use a theorem from [15, page 272]. For any matrix \( A \in \mathbb{C}^{n \times n} \),

\[
\| (A - I\lambda)^{-1} \| \geq \frac{1}{\text{dist}(\lambda, \sigma(A))}, \quad \lambda \in \rho(A),
\]

where the equality is valid for normal matrices, \( \text{dist}(\lambda, \sigma(A)) \) denotes the distance between a point \( \lambda \) and the spectrum. Now if we have a normal matrix (4.2) says that the pseudospectra obtained will be circles with the eigenvalue in the center. If the matrix is non-normal, however, (4.2) only says that for each \( \epsilon > 0 \) the \( \epsilon \)-pseudospectra will be larger than the circle with radius \( \text{dist}(\lambda, \sigma(A)) \).

For a strongly non-normal case as (1.1) the pseudospectra can be large and thus result in eigenvalue problems hard to compute numerically. The linearization is done both by rational and polynomial approach, so we are able to compare the computed pseudospectra by the different methods and by damping. The
pseudospectra from the rational technique is to the right in each figure. For the non-damped case we use linearization (3.23) and linearization (3.24) implemented as TM_Rat_linearization_No_damping.m for rational. It’s important to remember that the plotted eigenvalues and spectra is for $\lambda = \omega^2$ not for $\omega$. 

Figure 4.4: The eigenvalues for $\gamma = 9$. 

```
We will get back to these pseudospectra later but first we show the corresponding eigenvalues for the damped case, it is worth noting that the eigenvalues are all along the real axis as they should be for a symmetric problem. For the damped
eigenvalue problems we use polynomial linearization (3.32) and the rational linearization (3.42). The latter is implemented in the matlab script TM_Rat_linearization_damping.m. For damping coefficient $\gamma = 1$ we have the pseudospectra below:

With damping coefficient 1 it is apparent that the pseudospectra is much larger for the problem. This, however, mostly depends on the more sensitive linearization method that we have to use for a damped problem. The non-damped eigenvalue problem, if been computed by the damped linearization method would have given a larger pseudospectrum as well. For damping coefficient $\gamma = 3$ we have the pseudospectra below:
An interesting behavior for some of the eigenvalues can be observed, they are "breaking free" from the line of eigenvalues as the damping increases. For damping coefficient $\gamma = 9$ this behavior has been obvious:

For the last investigated damping $\gamma = 13$ the eigenvalues as well as the pseudospectra is very different from the non-damped case visualized (4.3):

We have now seen how the eigenvalues in the neighborhood of the origin is affected by the damping and it is obvious that the pseudospectra is larger for high damping coefficients. We can also see that the rational method gives a smaller pseudospectra for almost all eigenvalues in each damping. Furthermore the matrix
is of less dimension which makes the computations less demanding. These pictures however only shows how pseudospectra behaves in a large area, what’s matter is to see how the pseudospectra behaves near eigenvalues. Therefore we have zoomed in some of the most problematic eigenvalues in the pseudospectra where $\gamma = 13$ to obtain the plots.

![Pseudospectra plots](image)

We can here see that even very close up the pseudospectra is not incomprehensible large, $|\det (A - \lambda I)^{-1}| << 10^{12}$, and that the eigenvalues are separated by the pseudospectra this means that they can be trusted. (4.8) also shows that the rational method is more reliable since the pseudospectra is larger for the polynomial method. The conclusions we can make is that in this case the pseudospectra grants that the eigenvalues are more reliable for a low damping $\gamma$ and the rational method. That the rational method gives more reliable solutions may be due to that it preserves the symmetry properties of the problem better. We can, however still see that the eigenvalues, almost, coincides between these two methods and that the pseudospectra is small for the polynomial linearization too. This tells us that the approximated eigenvalues in the previous subsection can be trusted. Notice that some eigenvalues imaginary part increased fast compared to others as the damping increased. The "moving" eigenvalues also begun to cluster around the imaginary axis and becoming more ill-conditioned, still not close to the pseudospectra of the $7 \times 7$ matrix in the introduction, the Godunov matrix defined as (1.1).

(4.8)
5. **Conclusions**

The idea of this thesis was to write mainly about non-linear eigenvalue problems, pseudospectra and the theory concerning those subjects. As the work continued, the focus became the linearization of the rational eigenvalue problem. While the linearization of simple non-damped case (3.3) was easy to understand rather right away, the linearization of the damped problem took much longer to comprehend. The Lemmas and Theorems for the non-damped case were completed in an early part of the writing process. The main source of the damped eigenvalue, [12], skipped many of the rather technical parts, for example the proofs Theorem 27 and Lemma 28, these proofs are included in this thesis and it took some time to work them through properly. When we moved on to the more general case of operators valued functions working on infinite dimensional spaces there were some initial problems with the concept. The investigated numerical example concerning the photonic crystals, were present during almost the whole time in the literature. Still they were introduced to the report in the very end when the complete theoretical background, given in this thesis, had been established. It can be seen as a good result that in the studied TM-case the rational linearization had a noticeable smaller pseudospectra compared to the polynomial technique. This since one of the major parts of this thesis was to develop and understand the rational linearization method. During this work a understanding of the non-linear eigenvalue problem and perhaps most of all pseudospectra and the approach of linearizing a problem, have been obtained.
BIBLIOGRAPHY


6. Appendix

6.1. Proofs

6.1.1. Proof of Theorem 27

Given the system (3.25) and that \( G > 0, M_1, M_2 \geq 0 \) and that \( \gamma, \alpha_1, \alpha_2, \xi, \eta > 0 \). Then \( \operatorname{Im}(\sigma(T)) < 0 \).

Proof. To show that the imaginary part of the spectrum is negative we only have to show that the numerical range has negative imaginary part due to Lemma 26. This can be shown by showing that for each \( \omega \) in (3.26),
\[
u^* T(\omega) u = 0 \text{ for some } u \in \mathbb{C}^n \implies \operatorname{Im}(\omega) < 0.\]

Assuming that for some \( \omega \), \( u^* T(\omega) u = 0 \) then the imaginary part must be zero too,
\[
\operatorname{Im}(u^* T(\omega) u) = 0,
\]

or
\[
\operatorname{Im}
\left(u^* \left(G - \omega^2 \alpha_1 M_1 - \omega^2 \left(\alpha_2 + \frac{\xi^2}{\eta^2 - \omega^2 - i\gamma \omega}\right) M_2\right) u\right) = 0.
\]

By multiplying in the vectors \( u \) and \( u^* \)
\[
\operatorname{Im}
\left(u^* Gu - \omega^2 \alpha_1 u^* M_1 u - \omega^2 \left(\alpha_2 + \frac{\xi^2}{\eta^2 - \omega^2 - i\gamma \omega}\right) u^* M_2 u\right) = 0.
\]

Using that \( G, M_1 \) and \( M_2 \) at least positive semi-definite (3.26) it follows that for some non-negative real numbers \( p, q \) and \( r \) we have,
\[
\operatorname{Im}
\left(p - \omega^2 \alpha_1 q - \omega^2 \left(\alpha_2 + \frac{\xi^2}{\eta^2 - \omega^2 - i\gamma \omega}\right) r\right) = 0.
\]

Then this can be written as, using that \( \omega = a + bi \), for some real \( a \) and \( b \) and that \( p \in \mathbb{R} \),
\[
\operatorname{Im}
\left(- (a + bi)^2 \alpha_1 q - (a + bi)^2 \left(\alpha_2 + \frac{\xi^2}{\eta^2 - (a + bi)^2 - i\gamma (a + bi)}\right) r\right) = 0,
\]
which is
\[
\text{Im} \left( - \left( a^2 - b^2 + 2abi \right) \left( \alpha_1 q + \alpha_2 r + \frac{\xi^2}{\eta^2 - (a + bi)^2 - i\gamma (a + bi) r} \right) \right) = 0.
\]

By calculating the multiplication explicitly
\[
\text{Im} \left( - \left( a^2 - b^2 \right) (\alpha_1 q + \alpha_2 r) - 2ab (\alpha_1 q + \alpha_2 r) i \\
- \left( a^2 - b^2 + 2abi \right) \left( \frac{\xi^2}{\eta^2 - (a + bi)^2 - i\gamma (a + bi) r} \right) \right) = 0.
\]

This can be written as, since \( a \) and \( b \), are real,
\[
- 2ab (\alpha_1 q + \alpha_2 r) + \\
\text{Im} \left( - \left( a^2 - b^2 + 2abi \right) \frac{\xi^2}{\eta^2 - (a + bi)^2 - i\gamma (a + bi) r} \right) = 0.
\]

Evaluating the denominator
\[
- 2ab (\alpha_1 q + \alpha_2 r) + \\
\text{Im} \left( - \left( a^2 - b^2 + 2abi \right) \frac{\xi^2}{\eta^2 - a^2 + b^2 + b\gamma - (2ab + a\gamma) i} \right) = 0.
\]

Multiplying by the complex conjugate to the denominator yields
\[
- 2ab (\alpha_1 q + \alpha_2 r) + \\
\xi^2 r \text{Im} \left( - \left( a^2 - b^2 + 2abi \right) \frac{(\eta^2 - a^2 + b^2 + b\gamma) + (2ab + a\gamma) i}{(\eta^2 - a^2 + b^2 + b\gamma)^2 + (2ab + a\gamma)^2} \right) = 0.
\]

Since the denominator is the sum of two real quadrates we can write it as some \( s > 0 \) and we evaluate
\[
- 2ab (\alpha_1 q + \alpha_2 r) + \frac{\xi^2 r}{s} \text{Im} \left( - \left( (a^2 - b^2) (\eta^2 - a^2 + b^2 + b\gamma) - 2ab (2ab + a\gamma) \right) \\
- \left( (a^2 - b^2) (2ab + a\gamma) + 2ab (\eta^2 - a^2 + b^2 + b\gamma) \right) i \right) = 0,
\]

and as we are only interested in the imaginary part it follows
\[
- 2ab (\alpha_1 q + \alpha_2 r) + \frac{\xi^2 r}{s} \text{Im} \left( - \left( (a^2 - b^2) (2ab + a\gamma) + 2ab (\eta^2 - a^2 + b^2 + b\gamma) \right) i \right) = 0.
\]
Evaluating yields

\[-2ab (\alpha_1 q + \alpha_2 r) - \frac{\xi^2 r}{s} (2a^3 b + a^3 \gamma - 2ab^3 - ab^2 \gamma + 2ab\eta^2 - 2a^3 b + 2ab^3 + 2ab^2 \gamma) = 0,\]

by collecting terms

\[-2ab (\alpha_1 q + \alpha_2 r) - \frac{\xi^2 r}{s} (a^3 \gamma + ab^2 \gamma + 2ab\eta^2) = 0,\]

or

\[a \left( 2b (\alpha_1 q + \alpha_2 r) + \frac{\xi^2 r}{s} (a^2 \gamma + b^2 \gamma + 2b\eta^2) \right) = 0.\]

There is a solution to this system when \(a = 0\). Now supposing \(a\) is non-zero yields

\[\left( 2b (\alpha_1 q + \alpha_2 r) + \frac{\xi^2 r}{s} (a^2 \gamma + b^2 \gamma + 2b\eta^2) \right) = 0,\]

and by a rearrangement

\[b \left( 2\alpha_1 q + 2\alpha_2 r + 2 \frac{\xi^2 r}{s} \eta^2 \right) + b^2 \left( \frac{\xi^2 r}{s} \gamma \right) + \frac{\xi^2 r}{s} a^2 \gamma = 0.\]

Since all constants are positive we must have that if \(b\) is a solution to this system then \(b < 0\) since otherwise \(LH > 0\). Now we only have to show that \(b < 0\) in the case when \(a = 0\) and then the Theorem follows. Suppose \(a = 0\) then compute the numerical range for the \(\omega = bi\) where \(b \in \mathbb{R}\) such that

\[u^* T (bi) u = 0,\]

now this gives

\[u^* \left( G + b^2 \alpha_1 M_1 + b^2 \left( \alpha_2 + \frac{\xi^2}{\eta^2 + b^2 + \gamma b} \right) M_2 \right) u = 0.\]

We have that

\[\left( G + b^2 \alpha_1 M_1 + b^2 \left( \alpha_2 + \frac{\xi^2}{\eta^2 + b^2 + \gamma b} \right) M_2 \right)\]

is positive definite for \(b > 0\) since all constants are positive and the matrices are positive definite. Thus \(b < 0\) if this is a solution to the equation. But then we have shown that the numerical range is a subset to the negative imaginary part and using Lemma 26 gives that the spectrum of \(T\) has a negative imaginary part as well. \(\blacksquare\)
6.1.2. Proof of Lemma 28

Lemma (28), The system (3.28) is equivalent to

\[
\left( \hat{G} - \omega^2 M - b^T (A - \omega E)^{-1} b M_2 \right) u = 0,
\]

(6.1)

where \( A, E \) and \( b \) are defined as (3.33), (3.34) and (3.35).

Proof. The proof is completely by straight-forward computation.

\[
b^T (A - \omega E)^{-1} b = b^T \left( \frac{2\theta}{\xi^2} \begin{bmatrix} \theta - \frac{1}{2}i\gamma & 0 & 0 \\ 0 & \theta + \frac{1}{2}i\gamma \end{bmatrix} - \omega \frac{2\theta}{\xi^2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \right)^{-1} b,
\]

this can be written as

\[
b^T (A - \omega E)^{-1} b = b^T \left( \frac{2\theta}{\xi^2} \begin{bmatrix} \theta - \frac{1}{2}i\gamma - \omega & 0 \\ 0 & \theta + \frac{1}{2}i\gamma + \omega \end{bmatrix} \right)^{-1} b.
\]

Since this is a diagonal matrix the inverse is easily found as the diagonal matrix

with the inverses of the diagonal entries. We therefore have

\[
b^T (A - \omega E)^{-1} b = \frac{\xi^2}{2\theta} b^T \begin{bmatrix} (\theta - \frac{1}{2}i\gamma - \omega)^{-1} & 0 \\ 0 & (\theta + \frac{1}{2}i\gamma + \omega)^{-1} \end{bmatrix} b.
\]

Using (3.34) it follows

\[
b^T (A - \omega E)^{-1} b = \frac{\xi^2}{2\theta} \begin{bmatrix} (\theta - \frac{1}{2}i\gamma - \omega)^{-1} & 0 \\ 0 & (\theta + \frac{1}{2}i\gamma + \omega)^{-1} \end{bmatrix} \begin{bmatrix} \theta - \frac{1}{2}i\gamma \\ 0 \end{bmatrix}.
\]

Evaluating gives

\[
b^T (A - \omega E)^{-1} b = \frac{\xi^2}{2\theta} \left( \frac{(\theta - \frac{1}{2}i\gamma)^2}{\theta - \frac{1}{2}i\gamma - \omega} + \frac{(\theta + \frac{1}{2}i\gamma)^2}{\theta + \frac{1}{2}i\gamma + \omega} \right).
\]

Now using (3.35) and computing the squares gives

\[
b^T (A - \omega E)^{-1} b = \frac{\xi^2}{2\sqrt{\eta^2 - \frac{1}{4}\gamma^2}} \left( \frac{\eta^2 - \frac{1}{4}\gamma^2 - i\gamma \sqrt{\eta^2 - \frac{1}{4}\gamma^2} - \frac{1}{4}\gamma^2}{\sqrt{\eta^2 - \frac{1}{4}\gamma^2} - \frac{1}{2}i\gamma - \omega} + \frac{\eta^2 - \frac{1}{4}\gamma^2 + i\gamma \sqrt{\eta^2 - \frac{1}{4}\gamma^2} - \frac{1}{4}\gamma^2}{\sqrt{\eta^2 - \frac{1}{4}\gamma^2} + \frac{1}{2}i\gamma + \omega} \right),
\]
this can be simplified to

\[
b^T (A - \omega E)^{-1} b = \frac{\xi^2}{2\sqrt{\eta^2 - \frac{1}{4}\gamma^2}} \left( \frac{\eta^2 - i\gamma \sqrt{\eta^2 - \frac{1}{4}\gamma^2 - \frac{1}{2}\gamma^2}}{\sqrt{\eta^2 - \frac{1}{4}\gamma^2 - \left( \frac{1}{2}i\gamma + \omega \right)}} + \frac{\eta^2 + i\gamma \sqrt{\eta^2 - \frac{1}{4}\gamma^2 - \frac{1}{2}\gamma^2}}{\sqrt{\eta^2 - \frac{1}{4}\gamma^2 + \left( \frac{1}{2}i\gamma + \omega \right)}} \right).
\]

Multiplying by the conjugate in the denominator yields

\[
b^T (A - \omega E)^{-1} b = \frac{\xi^2}{2\sqrt{\eta^2 - \frac{1}{4}\gamma^2}} \left( \frac{\left( \eta^2 - i\gamma \sqrt{\eta^2 - \frac{1}{4}\gamma^2 - \frac{1}{2}\gamma^2} \right) \left( \sqrt{\eta^2 - \frac{1}{4}\gamma^2 + \left( \frac{1}{2}i\gamma + \omega \right)} \right)}{\eta^2 - \frac{1}{4}\gamma^2 - \left( \frac{1}{2}i\gamma + \omega \right)^2} + \frac{\left( \eta^2 + i\gamma \sqrt{\eta^2 - \frac{1}{4}\gamma^2 - \frac{1}{2}\gamma^2} \right) \left( \sqrt{\eta^2 - \frac{1}{4}\gamma^2 - \left( \frac{1}{2}i\gamma + \omega \right)} \right)}{\eta^2 - \frac{1}{4}\gamma^2 - \left( \frac{1}{2}i\gamma + \omega \right)^2} \right).
\]

Evaluating the products

\[
b^T (A - \omega E)^{-1} b = \frac{\xi^2}{2\sqrt{\eta^2 - \frac{1}{4}\gamma^2}} \left( \frac{\eta^2 \sqrt{\eta^2 - \frac{1}{4}\gamma^2} - \gamma \left( \eta^2 - \frac{1}{4}\gamma^2 \right) + \eta^2 \left( \frac{1}{2}i\gamma + \omega \right)}{\eta^2 - \frac{1}{4}\gamma^2 - \left( \frac{1}{2}i\gamma + \omega \right)^2} - \frac{\gamma \sqrt{\eta^2 - \frac{1}{4}\gamma^2} \left( \frac{1}{2}i\gamma + \omega \right) - \frac{1}{2}\gamma^2 \sqrt{\eta^2 - \frac{1}{4}\gamma^2 - \frac{1}{2}\gamma^2} \left( \frac{1}{2}i\gamma + \omega \right)}{\eta^2 - \frac{1}{4}\gamma^2 - \left( \frac{1}{2}i\gamma + \omega \right)^2} + \frac{\eta^2 \sqrt{\eta^2 - \frac{1}{4}\gamma^2} + \gamma \left( \eta^2 - \frac{1}{4}\gamma^2 \right) - \eta^2 \left( \frac{1}{2}i\gamma + \omega \right)}{\eta^2 - \frac{1}{4}\gamma^2 - \left( \frac{1}{2}i\gamma + \omega \right)^2} - \frac{\eta^2 - \gamma \sqrt{\eta^2 - \frac{1}{4}\gamma^2} \left( \frac{1}{2}i\gamma + \omega \right) - \frac{1}{2}\gamma^2 \sqrt{\eta^2 - \frac{1}{4}\gamma^2 + \frac{1}{2}\gamma^2} \left( \frac{1}{2}i\gamma + \omega \right)}{\eta^2 - \frac{1}{4}\gamma^2 - \left( \frac{1}{2}i\gamma + \omega \right)^2} \right).
\]
We now only have to simplify,

\[ b^T (A - \omega E)^{-1} b = \frac{\xi^2}{2 \sqrt{\eta^2 - \frac{1}{4} \gamma^2}} \left( \frac{2 \eta^2 \sqrt{\eta^2 - \frac{1}{4} \gamma^2} - 2 i \gamma \sqrt{\eta^2 - \frac{1}{4} \gamma^2} \left( \frac{1}{2} i \gamma + \omega \right) - \gamma^2 \sqrt{\eta^2 - \frac{1}{4} \gamma^2}}{\eta^2 - \frac{1}{4} \gamma^2 - \left( \frac{1}{2} i \gamma + \omega \right)^2} \right), \]

and continue as

\[ b^T (A - \omega E)^{-1} b = \frac{\xi^2}{\eta^2 - \omega^2 - i \gamma \omega} \left( \frac{- \eta^2 + \frac{1}{2} \gamma^2 - i \gamma \omega - \frac{1}{2} \gamma^2}{\eta^2 - \frac{1}{4} \gamma^2 + \frac{1}{4} \gamma^2 - \omega^2 - i \gamma \omega} \right). \]

Finally

\[ b^T (A - \omega E)^{-1} b = \frac{\xi^2 (\eta^2 - i \gamma \omega)}{\eta^2 - \omega^2 - i \gamma \omega}, \]

thus

\[ \left( \hat{G} - \omega^2 M - \left( \frac{\xi^2 (\eta^2 - i \gamma \omega)}{\eta^2 - \omega^2 - i \gamma \omega} \right) M_2 \right) u = \left( \hat{G} - \omega^2 M - b^T (A - \omega E)^{-1} b M_2 \right) u. \]

6.1.3. Proof of Lemma 33

Provided the Hilbert space (3.49) with inner product (3.50), then we have the estimate

\[ \left( \int_{\Omega} |\nabla u|^2 \, dx \right)^{\frac{1}{2}} \leq |\Omega|^{\frac{1}{2}} \| u \|, \]

for all \( u \in H^1(\Omega) \).

**Proof.** We use the Fourier series of \( u \) as in (3.48) to obtain

\[ \left( \int_{\Omega} |\nabla u|^2 \, dx \right)^{\frac{1}{2}} = \left( \int_{\Omega} \left| \nabla \sum_{n \in \mathbb{Z}^2} \hat{u}(n) e^{2\pi i n \cdot x} \right|^2 \, dx \right)^{\frac{1}{2}}. \]
Since these series converge in $\Omega$ it is valid to move $\nabla$ inside the sum

$$= \left( \int_{\Omega_2} \left| \sum_{n \in \mathbb{Z}^2} \nabla \left( \hat{u}(n) e^{2\pi i n \cdot x} \right) \right|^2 \, dx \right)^{\frac{1}{2}} = \left( \int_{\Omega_2} \left| \sum_{n \in \mathbb{Z}^2} \hat{u}(n) (2\pi i n_1, 2\pi i n_2) e^{2\pi i n \cdot x} \right|^2 \, dx \right)^{\frac{1}{2}}.$$ 

Evaluate the norm gives

$$= \left( \int_{\Omega} \left( \sum_{n \in \mathbb{Z}^2} \hat{u}(n) (2\pi i n_1, 2\pi i n_2) e^{2\pi i n \cdot x} \right) \right)^{\frac{1}{2}} \left( \sum_{n \in \mathbb{Z}^2} \hat{u}(n) (2\pi i n_1, 2\pi i n_2) e^{2\pi i n \cdot x} \right)^{\frac{1}{2}},$$

or

$$= \left( \int_{\Omega} \left( \sum_{n \in \mathbb{Z}^2} \hat{u}(n) (2\pi i n_1, 2\pi i n_2) e^{2\pi i n \cdot x} \right) \right)^{\frac{1}{2}} \left( \sum_{n \in \mathbb{Z}^2} (\hat{u}(n))^* (-2\pi i n_1, -2\pi i n_2) e^{-2\pi i n \cdot x} \right) \, dx \right)^{\frac{1}{2}}.$$

Multiplying term-wise yields

$$= \left( \int_{\Omega} \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}(n)|^2 \left( (2\pi n_1)^2 + (2\pi n_2)^2 \right) \right)^{\frac{1}{2}} + \sum_{n \neq m \in \mathbb{Z}^2} \hat{u}(n) \hat{u}(m)^* \left( 4\pi^2 n_1 m_1 + 4\pi^2 n_2 m_2 \right) e^{2\pi i n \cdot x - 2\pi i m \cdot x} \right) \, dx \right)^{\frac{1}{2}}.$$
which can be simplified to, due to the identity \((2\pi n_1)^2 + (2\pi n_2)^2 = |2\pi n|^2\),

\[
\leq \left( \int_{\Omega} \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}(n)|^2 |2\pi n|^2 + \sum_{n \neq m \in \mathbb{Z}^2} |\hat{u}(n)|^2 |2\pi n|^2 \right) \sum_{n \neq m \in \mathbb{Z}^2} \hat{u}(n) \hat{u}(m)^* 4\pi^2 (n_1 m_1 + n_2 m_2) e^{2\pi i (n-m) \cdot x} \right)^{\frac{1}{2}}.
\]

It can be seen the first sum is independent of space variable \(x\) so it can written by letting \(|\Omega|\) denote the area of \(\Omega\)

\[
\left( |\Omega| \sum_{n \in \mathbb{Z}^2} |\hat{u}(n)|^2 |2\pi n|^2 + \left( \sum_{n \neq m \in \mathbb{Z}^2} \hat{u}(n) \hat{u}(m)^* 4\pi^2 (n_1 m_1 + n_2 m_2) \int_{\Omega} e^{2\pi i (n-m) \cdot x} dx \right) \right)^{\frac{1}{2}}.
\]

Since \(n - m \in \mathbb{Z}^2 \setminus \{(0,0)\}\) we have that \(\int_{\Omega} e^{2\pi i (n-m) \cdot x} dx = 0\) and thus we have

\[
= \left( |\Omega| \sum_{n \in \mathbb{Z}^2} |\hat{u}(n)|^2 |2\pi n|^2 \right)^{\frac{1}{2}}.
\]

From the norm given by (3.50) this is

\[
\leq \left( |\Omega| \sum_{n \in \mathbb{Z}^2} |\hat{u}(n)|^2 \right)^{\frac{1}{2}} \left( |2\pi n|^2 + 1 \right)^{\frac{1}{2}} = |\Omega|^{\frac{1}{2}} \|u\|.
\]

Thus the conclusion follows as

\[
\left( \int_{\Omega} |\nabla u|^2 \, dx \right)^{\frac{1}{2}} \leq |\Omega|^{\frac{1}{2}} \|u\|.
\]
6.1.4. Proof of Lemma 43

Suppose we have the operator valued function $Q_{TE}(\omega)$ defined as (3.65) then $\sigma(Q_{TE}) \subset W(Q_{TE})$.

**Proof.** We will now once again use the theorem that says that the if for some $\lambda_0$ zero is not in the closure of the numerical range we have that the spectrum is consisted in the numerical range. We choose to evaluate $\lambda_0 = i$ thus it is to show that

$$\inf_{\|u\|=1} |(T(i)u,u)| \neq 0.$$  

To be able to continue in a nice way we return to an earlier way of describing the operator $Q_{TE}(\omega)$ (3.56),

$$\inf_{\|u\|=1} \left| \int_{\Omega} \frac{1}{\epsilon(x,i)} \nabla u \cdot \nabla v^* dx + 2i \int_{\Omega} \frac{1}{\epsilon(x,i)} uk \cdot \nabla v^* dx + \int_{\Omega} \frac{1}{\epsilon(x,i)} |k|^2 uv^* dx + \int_{\Omega} uv^* dx \right| = 0 \quad (6.2)$$

The idea is to define a sequence $\{u^s\}$ such that $\|u^s\| = 1$ and $|(T(0)u,u)| \to 0$ and then show that such a sequence is not possible to construct. Using the Fourier expansion of $u^s$ yields

$$\int_{\Omega} \frac{1}{\epsilon(x,i)} \nabla \left( \sum_{n \in \mathbb{Z}^2} \hat{u}^s(n) e^{2\pi in \cdot x} \right) \cdot \nabla \left( \sum_{n \in \mathbb{Z}^2} \hat{u}^s(n) e^{2\pi in \cdot x} \right)^* dx$$

$$+ 2i \int_{\Omega} \frac{1}{\epsilon(x,i)} \left( \sum_{n \in \mathbb{Z}^2} \hat{u}^s(n) e^{2\pi in \cdot x} \right) k \cdot \nabla \left( \sum_{n \in \mathbb{Z}^2} \hat{u}^s(n) e^{2\pi in \cdot x} \right)^* dx$$

$$+ \int_{\Omega} \left( \frac{1}{\epsilon(x,i)} |k|^2 + a^2 \right) \left( \sum_{n \in \mathbb{Z}^2} \hat{u}^s(n) e^{2\pi in \cdot x} \right) \left( \sum_{n \in \mathbb{Z}^2} \hat{u}^s(n) e^{2\pi in \cdot x} \right)^* dx \to 0.$$

The products of the sums can be written as the as we do in the proof of Lemma 33, we take the gradient and divide the products of the sums into one part consisting of one sum where the pairs of integer match and one where they don’t match.
The system then follows as

$$
\int_\Omega \frac{1}{\epsilon(x,i)} \left( \sum_{n \in \mathbb{Z}^2} |2\pi n|^2 |\hat{u}^s(n)|^2 \right) + \\
\left( \sum_{m \neq n \in \mathbb{Z}^2} 4\pi^2 m \cdot n \hat{u}^s(n) \hat{u}^s(m) e^{2\pi i (n-m) \cdot x} \right) dx \\
- 2 \left( \int_\Omega \frac{1}{\epsilon(x,i)} \left( \sum_{n \in \mathbb{Z}^2} 2\pi k \cdot n |\hat{u}^s(n)|^2 \right) \right) - \\
\left( \sum_{m \neq n \in \mathbb{Z}^2} 2\pi k \cdot m \hat{u}^s(n) \hat{u}^s(m) e^{2\pi i (n-m) \cdot x} \right) dx \\
+ \int_\Omega \left( \frac{1}{\epsilon(x,i)} |k|^2 + 1 \right) \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s(n)|^2 \right) \\
+ \left( \sum_{m \neq n \in \mathbb{Z}^2} \hat{u}^s(n) \hat{u}^s(m) e^{2\pi i (n-m) \cdot x} \right) dx \rightarrow 0.
$$

When integrating over the whole domain we have that second sums in each row cancel due to that the terms are periodic. Thus it follows

$$
\int_\Omega \frac{1}{\epsilon(x,i)} \left( \sum_{n \in \mathbb{Z}^2} |2\pi n|^2 |\hat{u}^s(n)|^2 \right) dx - 2 \int_\Omega \frac{1}{\epsilon(x,i)} \left( \sum_{n \in \mathbb{Z}^2} 2\pi k \cdot n |\hat{u}^s(n)|^2 \right) dx \\
+ \int_\Omega \left( \frac{1}{\epsilon(x,i)} |k|^2 + a^2 \right) \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s(n)|^2 \right) dx \rightarrow 0.
$$

This gives the identity

$$
\int_\Omega \frac{1}{\epsilon(x,i)} \left( \sum_{n \in \mathbb{Z}^2} |2\pi n|^2 |\hat{u}^s(n)|^2 \right) dx - 2 \int_\Omega \frac{1}{\epsilon(x,i)} \left( \sum_{n \in \mathbb{Z}^2} 2\pi k \cdot n |\hat{u}^s(n)|^2 \right) dx \\
+ \int_\Omega \left( \frac{1}{\epsilon(x,i)} |k|^2 + a^2 \right) \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s(n)|^2 \right) dx \rightarrow 0.
$$
We rearrange and put expressions depending on $\epsilon$ in the same integral and sum, to obtain

$$
\int_{\Omega} \frac{1}{\epsilon (x, i)} \left( \sum_{n \in \mathbb{Z}^2} |2\pi n|^2 |\hat{u}^s (n)|^2 - 2 \left(2\pi k \cdot n |\hat{u}^s (n)|^2 + |k|^2 |\hat{u}^s (n)|\right) \right) dx
$$

$$
+ a^2 \int \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s (n)|^2 \right) dx \to 0.
$$

This can be written as

$$
\int_{\Omega} \frac{1}{\epsilon (x, i)} \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s (n)|^2 \left(2\pi n|^2 - 2 (2\pi k \cdot n) + |k|^2 \right) \right) dx
$$

$$
+ a^2 \int \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s (n)|^2 \right) dx \to 0.
$$

Using the quadratic rule,

$$
\int_{\Omega} \frac{1}{\epsilon (x, i)} \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s (n)|^2 |2\pi n - k|^2 \right) dx + \int \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s (n)|^2 \right) dx \to 0.
$$

Now writing out the $\epsilon (x, i)$ explicitly from (3.52) it follows

$$
\epsilon (x, i) = \epsilon_1 \chi_1 (x) + \frac{b + c - \gamma i - i^2}{c - i^2 - i^2 \omega} \chi_2 (x),
$$

or

$$
\epsilon (x, i) = \epsilon_1 \chi_1 (x) + \frac{b + c + \gamma + 1}{c + \gamma + 1} \chi_2 (x).
$$

This gives us that there is a real number and $c_2 \leq \frac{1}{\epsilon (x, i)} \leq c_3$ since $\epsilon (x, i)$ only have two values both positive, we can always choose $c_2$ such that $c_2 \leq 1$. From this it follows that

$$
\left| \int_{\Omega} \frac{1}{\epsilon (x, i)} \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s (n)|^2 |2\pi n - k|^2 \right) dx + \int \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s (n)|^2 \right) dx \right|
$$

$$
\geq c_2 |\Omega| \left| \left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s (n)|^2 \left(|2\pi n - k|^2 + 1\right) \right) \right| \to 0.
$$
We now want to show that no sequence \( \{u^s\} \) approaches zero we must thus show that this is larger than some constant \( d > 0 \) for all \( u^s \) in the sequence. \( k \in \mathbb{R}^2 \) is a fixed number so there are only a finite number of \( n \in \mathbb{Z}^2 \) such that \( |2\pi n - k| \) is less than \( |\pi n| \) thus the sum can be written as two sums over two different sets,

\[
N_1 = \{ n \in \mathbb{Z}^2 : |2\pi n - k| < |\pi n| \},
\]

and

\[
N_2 = \mathbb{Z}^2 \setminus N_1.
\]

Now from (6.3) it follows that

\[
\left( \sum_{n \in \mathbb{Z}^2} |\hat{u}^s(n)|^2 \left( |2\pi n - k|^2 + 1 \right) \right) = \\
\left( \sum_{n \in N_1} |\hat{u}^s(n)|^2 \left( |2\pi n - k|^2 + 1 \right) \right) + \left( \sum_{n \in N_2} |\hat{u}^s(n)|^2 \left( |2\pi n - k|^2 + 1 \right) \right).
\]

We now estimate both these sum to obtain, using the property of \( N_2 \),

\[
\geq \left( \sum_{n \in N_1} |\hat{u}^s(n)|^2 \right) + \left( \sum_{n \in N_2} |\hat{u}^s(n)|^2 \left( |\pi n|^2 + 1 \right) \right). \quad (6.4)
\]

Since \( N_1 \) only have a finite number of elements there is some \( n_0 \in N_1 \) such that we \( |n_0| \geq |n| \) for all \( n \in N_1 \) thus the first sum in (6.4) can be written as.

\[
\left( \sum_{n \in N_1} |\hat{u}^s(n)|^2 \right) = \frac{|2\pi n_0|^2 + 1}{|2\pi n_0|^2 + 1} \left( \sum_{n \in N_1} |\hat{u}^s(n)|^2 \right).
\]

This can be written as, using the maximum property of \( n \),

\[
\frac{1}{|2\pi n_0|^2 + 1} \left( \sum_{n \in N_1} |\hat{u}^s(n)|^2 \left( |2\pi n_0|^2 + 1 \right) \right) \geq \\
\frac{1}{|2\pi n_0|^2 + 1} \left( \sum_{n \in N_1} |\hat{u}^s(n)|^2 \left( |2\pi n|^2 + 1 \right) \right). \quad (6.5)
\]
The second sum in (6.4) yields

$$\left( \sum_{n \in N_2} |\widehat{u}^s(n)|^2 \left( |\pi n|^2 + 1 \right) \right) = \left( \sum_{n \in N_2} |\widehat{u}^s(n)|^2 \left( \frac{|2\pi n|^2}{4} + 1 \right) \right),$$

and we have the estimate

$$\geq \left( \sum_{n \in N_2} |\widehat{u}^s(n)|^2 \left( \frac{|2\pi n|^2}{4} + \frac{1}{4} \right) \right) = \frac{1}{4} \left( \sum_{n \in N_2} |\widehat{u}^s(n)|^2 \left( |2\pi n|^2 + 1 \right) \right). \quad (6.6)$$

Combining (6.4), (6.5) and (6.6), it follows that

$$\left( \sum_{n \in \mathbb{Z}^2} |\widehat{u}^s(n)|^2 \left( |2\pi n - k|^2 + 1 \right) \right) \geq d \left( \sum_{n \in N} |\widehat{u}^s(n)|^2 \left( |2\pi n|^2 + 1 \right) \right).$$

where \( d = \min \left( \frac{1}{|2\pi n_0|^2 + 1}, \frac{1}{4} \right) > 0 \). Using the norm of \( u \) to obtain

$$= d \| u \| = d.$$

Thus we have in (6.3) that

$$\inf_{\| u \| = 1} |\langle T (-1) u, u \rangle| \geq c_2 d > 0,$$

and no sequence with the desired properties can be found, thus \( 0 \) is not in the numerical range and the spectrum is a subset of the numerical range due to [13, page 139].

### 6.2. Matlab code

#### 6.2.1. TM_Poly_linearization_No_damping.m

close all

```matlab
clear all
load Cedric_Fig5_p8_ndof450
```

```matlab
% Eigenvalue problem in \( \omega^2 \) with two real parameters k1,k2
```
\( (G(k_1, k_2) - \omega^2 \cdot \text{eps1}(\omega)) \cdot \text{MI} - \omega^2 \cdot \text{eps2} \cdot \text{MO} \cdot x = 0 \)

\( \text{eps1}(\omega) = a + b/(c - \omega^2) \)

\( G \geq 0, \text{ with } G > 0 \text{ if } k_1 \neq 0 \text{ and } k_2 \neq 0, \ G^{*} = G \)

\( \text{MO} \geq 0, \ \text{MI} = 0, \ \text{MO} + \text{MI} > 0 \)

---

Data as in

Permittivity in the outer subdomain

\( a = 10; \)
\( b = 5 \cdot (2\pi)^2; \)
\( c = 1 \cdot (2\pi)^2; \)

Permittivity in the inner subdomain

\( \text{eps2} = 1; \)

---

Solution with linearization

\( A_2 = a \cdot \text{MO} + \text{eps2} \cdot \text{MI}; \)
\( A_1 = -G - (c \cdot a + b) \cdot \text{MO} - c \cdot \text{eps2} \cdot \text{MI}; \)
\( A_0 = c \cdot G; \)

\( A_2 > 0, A_1 < 0, A_0 = 0 \)

---

The spectrum is real and we have no associated vectors

\( A_2 = a \cdot \text{MO} + \text{eps2} \cdot \text{MI}; \)

---

Preprocessing

\( \text{MO} = (\text{MO} + \text{MO}')/2; \)
\( \text{MI} = (\text{MI} + \text{MI}')/2; \)
\( \text{M} = \text{MO} + \text{MI}; \)

k vectors

\( N = 10; \)
\( k_{1v0} = \pi \cdot \text{linspace}(0.01, 0.99, N); \)
\( k_{2v0} = \text{zeros}(1, N); \)
\( k_{1v1} = \pi \cdot \text{ones}(1, N); \)
\( k_{2v1} = \pi \cdot \text{linspace}(0.01, 0.99, N); \)
k1v2 = pi*pi*linspace(0.01,0.99,N);
k2v2 = pi*pi*linspace(0.01,0.99,N);
k1v=[k1v0;k1v1;k1v2];
k2v=[k2v0;k2v1;k2v2];

% reset output data
%omega0 = [];
%kplot = [];
ndof = size(M,1);
O = sparse(ndof,ndof);
id = speye(ndof,ndof);
nev=5;
sigma=0; %c-c/1000;
omega0 = [];
kplot = [];
for n=1:3
    for kk = 1:N
        k1=k1v(n,kk);
        k2=k2v(n,kk);
        display(sprintf([' Computing eigenvalues for '...
            ' k1 = %f, k2 = %f ...
        ],k1,k2));
        G = ATM - 2i*(k1*(MDxIn+MDxOut) + k2*(MDyIn+MDyOut)) + (k1*k1 + k2*k2)*M;
        A1=-G-(c*a+b)*MO-c*eps2*MI;
        A0=c*G;
        VL = [O id; -A0 -A1];
        HL = [id O; O A2];
        egvalue = eigs(VL,HL,nev,sigma);
        omega0 = [omega0; sqrt(egvalue)];
        kplot = [kplot; repmat(k1v(1, kk) + (n-1)*pi, length(egvalue), 1)];
    end
end
plot(kplot,real(omega0),' r*' )
hold on
x=[pi pi];
y=[min(real(omega0)) max(real(omega0))];
plot(x,y,'k-','LineWidth',2)
x=[2*pi 2*pi];
plot(x,y,'k-','LineWidth',2)
x=[3*pi 3*pi];
plot(x,y,'k-','LineWidth',2)
ylabel('
\lambda')
axis tight

6.2.2. TM_Poly_linearization_damping.m

close all
clear all
load Cedric_Fig5_p8_ndof450

%Eigenvalue problem in $\omega^2$ with two real parameters $k1,k2$

%(G(k1,k2)-$\omega^2$*eps1($\omega$))*MI-$\omega^2$*eps2*MO)*x=0

%eps1($\omega$)=a+b/($c-\omega^2-gamma*1i\omega$)

%G>0, with G>0 if k1 neq 0 and k2 neq 0, G^{-}=G
%MO>0, MI>=0, MO+MI>0

%Data as in
% Permittivity in the outer subdomain
a=10;
b=5*(2*pi)^2;
c=1*(2*pi)^2;
% Permittivity in the inner subdomain
eps2=1;
% damping in the inner subdomain
gamma=input('Damping? ');

%Solution with linearization
%A4=a*MO+eps2*MI;
%A3=-gamma*1i*(a*MO+eps2*MI);
%A2=(G+b*MO)+(c*(a*MO+eps2*MI));
%A1=gamma*1i*((G+b*MO)-b*MI)
%A0=-c*((G+b*MO)-b*MO);

% %A2>0, A1<0, A0>=0
%
%------------------------------------------------------------------------
MO = (MO + MO')/2;
MI = (MI + MI')/2;
M = MO + MI;
A4=a*MO+eps2*MI;
% preprocessing
% k vectors
N=10;
k1v0 = pi*linspace(0.01,0.99,N);
k2v0 =zeros(1,N);
k1v1 = pi*ones(1,N);
k2v1 = pi*linspace(0.01,0.99,N);
k1v2 =pi-pi*linspace(0.01,0.99,N);
k2v2 =pi-pi*linspace(0.01,0.99,N);
k1v=[k1v0;k1v1;k1v2];
k2v=[k2v0;k2v1;k2v2];
% reset output data
%omega0 = [];
kplot = [];
ndof =size(M,1);
O = sparse(ndof,ndof);
id=speye(ndof,ndof);
nev=10;
sigma=0; %c-c/1000;
omega0 = [];
kplot = [];
for n=1:3
for kk = 1:N
k1=k1v(n,kk);
k2=k2v(n,kk);
display(sprintf([' Computing eigenvalues for ' ...
k1 = %f, k2 = %f ...

G = ATM - 2i*(k1*(MDxIn+MDxOut) + k2*(MDyIn+MDyOut)) + (k1*k1 + k2*k2)*M;
A3=-gamma*1i*(a*MO+eps2*MI);
A2=(G+b*MO)+(c*(a*MO+eps2*MI));
A1=gamma*1i*(G);
A0=-c*G;
VL = [O id O O; O O id O; O O O id; A0 A1 A2 A3];
HL = [id O O O; O id O O; O O id O; O O O A4];
egvalue = eigs(VL,HL,nev,sigma);
omega0 = [omega0; egvalue];
kplot = [kplot; repmat(k1v(1, kk) + (n-1)*pi, length(egvalue), 1)];
ylabel('Imag(\omega)')
axis tight
set(gcf,'NextPlot','add');
axes;

h = title(['Eigenvalues of the problem when the damping coefficient is ', num2str(gamma)]);
set(gca,'Visible','off');
set(h,'Visible','on');

AA=full(HL\VL);

6.2.3. TM_Rat_linearization_No_damping.m

close all
clear all
load Cedric_Fig5_p8_ndof450

%— — — — — — — — — — — — — — — — — — — — — —
%Eigenvalue problem in \omega^2 with two real parameters k1,k2
%
%(G(k1,k2)-\omega^2*eps1(\omega^2)*MI-\omega^2*eps2*MO)*x=0
%
%eps1(\omega)=a+b/(c-\omega^2)
%
%G\ge0, with G>0 if k1\neq0 and k2\neq0, G^{*}=G
%MO\ge0, MI\ge0, MO+MI>0
%
%— — — — — — — — — — — — — — — — — — — — — —
%Data as in
% Permittivity in the outer subdomain
a=10;
b=5*(2*pi)^2;
c=1*(2*pi)^2;
% Permittivity in the inner subdomain
eps2=1;
%
%Solution with linearization
%
%G2=G+b*MO
%Mbar=a*MO+eps2*MI;
%B'*B=MO
%D=c*eye(n2);
%
%
The spectrum is real and we have no associated vectors
%
Mbar=a*MO+eps2*MI;
% preprocessing
MO = (MO + MO')/2;
MI = (MI + MI')/2;
M = MO + MI;
% k vectors
N=10;
k1v0 = pi*linspace(0.01,0.99,N);
k2v0 =zeros(1,N);
k1v1 = pi*ones(1,N);
k2v1 = pi*linspace(0.01,0.99,N);
k1v2 =pi-pi*linspace(0.01,0.99,N);
k2v2 =pi-pi*linspace(0.01,0.99,N);
k1v=[k1v0;k1v1;k1v2];
k2v=[k2v0;k2v1;k2v2];
% reset output data
%omega0 = [];
kplot = [];
nzidx = find(any(MO));
m = length(nzidx);
S2 = chol(MO(nzidx, nzidx));
n1=length(MO);
n2=length(S2);
F=[zeros(n2,n1-n2),S2];
D=c*eye(n2);
B=sqrt(b*c)*F;
ndof =size(M,1);
O = sparse(n2,n1);
id=speye(n2,n2);
nev=5;
**6.2.4. TM_Rat_linearization_damping.m**

```matlab
sigma=0; %c-c/1000;
k1=k1v(3,N);
k2=k2v(3,N);
G = ATM - 2i*(k1*(MDxIn+MDxOut) + k2*(MDyIn+MDyOut)) + (k1*k1 + k2*k2)*M;
G2=G+b*MO;
VL = [G2 B'; B D];
HL = [Mbar O'; O id];
egvalue = eigs(VL,HL,nev,sigma);
omega0 = sqrt(egvalue);
```

close all

clear all

load Cedric_Fig5_p8_ndof450

%— — — — — — — — — — — — — — — — — — — — — —
%Eigenvalue problem in \omega^2 with two real parameters k1,k2
%
%(G(k1,k2)-\omega^2*eps1(\omega))*MO-\omega^2*eps2*M1)*x=0
%
%eps1(\omega)=a+b/(c-\omega^2-gamma*1i*\omega)
%
%G>=0, with G>0 if k1\neq 0 and k2\neq 0, G^{*}=G
%MO>=0, MI>=0, MO+MI>0
%
%— — — — — — — — — — — — — — — — — — — — — —
%Data as in
% Permittivity in the outer subdomain
a=10;
b=5*(2*pi)^2;
c=1*(2*pi)^2;
% Permittivity in the inner subdomain
eps2=1;
% damping in the outer subdomain
gamma=input('Damping? ');
%---------------------------------------------------
```
%Solution with rational linearization

MO = (MO + MO')/2;
MI = (MI + MI')/2;
M = MO + MI;
Mbar=a*MO+eps2*MI;

% preprocessing
% k vectors
N=10;
k1v0 = pi*linspace(0.01,0.99,N);
k2v0 =zeros(1,N);
k1v1 = pi*ones(1,N);
k2v1 = pi*linspace(0.01,0.99,N);
k1v2 =pi-pi*linspace(0.01,0.99,N);
k2v2 =pi-pi*linspace(0.01,0.99,N);
k1v=[k1v0;k1v1;k1v2];
k2v=[k2v0;k2v1;k2v2];
% reset output data
%omega0 = [];
%kplot = [];
om=sqrt(c-(gamma^2)/4);
A=(2*om/b)*[om-(1/2)*1i*gamma 0; 0 om+(1/2)*1i*gamma];
E=(2*om/b)*[1 0; 0 -1];
b2=[om-(1/2)*1i*gamma; om+(1/2)*1i*gamma];
nzidx = find(any(MO));
m = length(nzidx);
S2 = chol(MO(nzidx, nzidx));
n1=length(MO);
n2=length(S2);
F=[zeros(n2,n1-n2),S2];
In2=eye(n2);
A=kron(A,In2);
B=kron(b2,F);
E=kron(E,In2);
% This function gives a lower bound for the departure from normality
% for a given matrix A in the spectral norm and thus also in the
% Eulerian norm
% Code by Torshage Programming inc
n=length(A);
P=perms(1:n);
m=length(P);
lambda=eig(A);
s=sqrt(eig(A'*A));
L=inf;
for i=1:m
    L2=max(abs(s-abs(lambda(P(i,:)))));
    L=min(L,L2);
end

6.2.5. function [L,lambda,s] = lowerbound(A)
6.2.6. function U = upperbound(A)

% This function gives an upper bound for the departure from normality
% for a given matrix A in the Eulerian norm and thus also in the
% spectral norm
% Code by Torshage Programming Inc
n = length(A);
B = A*A - A'*A;
U = sqrt(sqrt((n^3-n)/12))*sqrt(norm(B,'fro'));
end