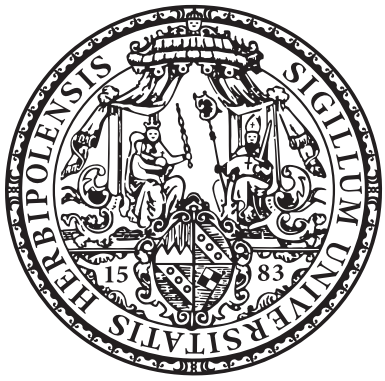
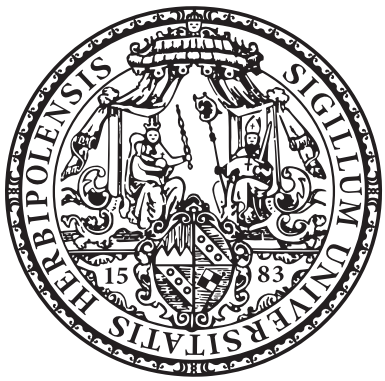
On algebraic aggregation methods in additive preconditioning

Michael Tichy



Dissertation
Department of Mathematics
University of Würzburg

On algebraic aggregation methods in additive preconditioning



Dissertation zur Erlangung
des naturwissenschaftlichen Doktorgrades
der Julius-Maximilians-Universität Würzburg

vorgelegt von

MICHAEL TICHY

aus

Mannheim

Eingereicht am: 22. November 2010

1. Gutachter: Prof. Dr. Manfred Dobrowolski, Universität Würzburg

2. Gutachter: Prof. Dr. Christian Klingenberg, Universität Würzburg

"*Cuiusvis hominis est errare,*
nullius nisi insipientis in errore perseverare.
(Jeder Mensch kann irren!
Unsinnige nur verharren im Irrtum!) "
(Marcus Tullius Cicero)

"*Was wir wissen, ist ein Tropfen;*
was wir nicht wissen, ein Ozean. "
(Sir Isaac Newton)

Acknowledgements

The doctoral thesis at hand is the result of my research during the time as a Ph.D. student at the Department of Mathematics at the University of Würzburg. Since I have worked on this project, this project has worked on me, too. Hence I am very happy that I had some people on my side who accompanied me during this time.
First of all, I would like to express my gratitude to my supervisor Prof. Dr. Manfred Dobrowolski. His door always was wide open for me and he was all ear for my problems and questions. Furthermore, there was enough freedom for me to go my own scientific way.
In addition to the thanks to my supervisor I want to express my thanks to some other persons for their support.
First, in this, is my family. I would like to thank my wife Diana, for going with me through all the ups and downs of the last years. My parents for generous financial support during my studies and my whole family for keeping my grounded. A very special thanks goes to my son Julius, for giving my a new angle of view.
Furthermore I would like to thank Dr. Ralf Winkler, Dr. Florian Möller for the helpful technical support and all my lectors for their patience.

Abstract: In the following dissertation we consider three preconditioners of algebraic multigrid type, though they are defined for arbitrary prolongation and restriction operators, we consider them in more detail for the aggregation method. The strengthened Cauchy-Schwarz inequality and the resulting angle between the spaces will be our main interests. For the problem of the one-dimensional convection we obtain perfect theoretical results. Although this is not the case for more complex problems, the numerical results we present will show that the modifications are also useful in these situation. Additionally, we will consider a symmetric problem in the energy norm and present a simple rule for algebraic aggregation.

Contents

1 Introduction

It is a quite old and simple, but even still interesting question how we can practically solve the system of linear equations

$$A\,u = f.$$

Since scientists from different disciplines use computers to solve mathematical models, the dimension of the systems of linear equations that can be solved exactly and in acceptable time is one of the limits for the complexity of the models and the precision of the conclusions they get. In particular, if the number of equations represents a number of gridpoints then the dependency of the precision on the size of the matrix is obvious. As the memory of computers grows and the processors become faster there is also a need for fast and robust solutions for systems of linear equations.

In the following section we will briefly sum up the popular methods to solve a system of linear equations. All the presented methods are based on the idea that we consider a partial differential equation (PDE) on a domain $\Omega \subset \mathbb{R}^2\,(\mathbb{R}^3)$ and that the system of linear equations results from the discretisation of the PDE. For some of the methods this is a necessary condition. For other methods this is only a motivation (black box solvers). The following dissertation belongs to the second kind of methods.
We will consider the papers [Van92], [Van95], [VBM96], [VBM01] and [VBT99] in more detail. These papers introduce and develop the smoothed aggregation method. This method has a similarity to this thesis. Of course we will present the differences of the ideas, too. Afterwards we will give a brief outline of this thesis.

The first idea we present is called the domain decomposition method. The idea is based on the notion that there is a domain Ω on which the continuous problem is defined. Thus the domain is simply decomposed in n subdomains $\Omega_i \subset \Omega$, $i = 1, \ldots n$. On each of these, a smaller system of linear equations results. It is obvious that these problems are easier to solve. But now the problem occurs that the right-hand side for several

of the smaller problems is influenced by the solution on the whole domain itself. Furthermore we have to compose the solutions on the different subdomains to a global solution. For both problems it is obvious that the more the solutions on the subdomains are influenced by each other, the bigger the problem is. Hence this is particularly problematic for elliptic problems.
In fact this is a quite old idea. The first time the idea was mentioned was in the 19th century in [Sch70]. Since that time there have been many evolutions and modifications for this method. Here we just mention the papers [BPS86], [BPS87], [BPS88] and [BPS89] as examples.
The first applications for the domain decomposition method on big domains are considered in [VAR60] in the 1960s. The first applications for elliptic problems are considered twenty years later in [BjH88], [BjW84], [Rou89], [Smi92] and [TRV91].

Another concept is what we call the multigrid method. It is basically motivated as follows: If we use an iterative solver for the system of linear equations (for example the Jacobi-method), and we consider the residual error $e^k = A\,u^k - f$ after k iterations, then we can decompose e^k into frequences. Even if the iterative method converges, there are frequences in which the error is reduced quite slowly by the iterative method. The idea is that on different grids, the error in different frequences shrinks fast. Hence we use the different grids and solve a linear system of equations on each of them. Again the problem occurs that we have to build a solution from the solutions on different grids.

In fact these methods can be split into two different types. First there are the multiplicative methods. In these methods we start on the finer grid with some steps of an iterative solver. Then the remaining error is mapped into a coarser space. On this space, a lower dimensional system of linear equations results which we have to solve. The right-hand side for this system results from the error which remained on the finer grid. We solve this lower dimensional system of linear equations and map the solution in the finer grid. As a last step, the mapped solution from the coarser grid is used to modify the current iteration on the finer grid to obtain a closer approximation to the solution on the finer grid.
The name of the method is inspired by the fact that there is a matrix representation for this algorithm which shows that the method is based on the multiplication of different solvers. We emphasize that as the remaining residuum on the finer grid is used as the right side in the coarser grid, the coarser grid system needs some information from current iteration on the finer system. This is obvious from the multiplicative represen-

tation of the algorithm we mentioned above.
The other idea is to use an additive method. In these methods we solve the equation, or a part of it, on different grids or different subspaces respectively. Thus we start by decomposing the right-hand side f of the system of linear equations which results in some subspaces. Then we use on each subspace an iterative method (or the exact inverse) to obtain an approximation for the solution. Afterwards we add the solutions of the subspaces to obtain a global solution. The name for these methods follows as the solvers on the different subspaces are linked additively.
If we compare this to the multiplicative methods, two things are obvious:
As the multiplicative methods use more information on coarser grids, these methods should need less iterations. As the additive methods on a grid need no information from another grid, these methods can be parallelised in a better way on a parallel computer. In this case, these methods should be faster.

The multigrid methods were first mentioned the 1960s in [Fed62], [Fed64], [KrD72], [Brk60] and [Bak66]. In the 1970s, important progress for these methods was made. In particular, the papers of A. Brandt [Brd73], [Brd77], [Brd82] and [McC87] have been important for the development of this method.

A popular evolution of the multigrid methods is the introduction of the hierarchical basis. The idea is to use on finer grids not the nodal basis but the basis of the coarser grid and some nodal basis functions for the finer grid. One of the most interesting aspects of this method is that the method can easily be formulated as a block iteration. This idea was first mentioned in [ZKGB82] and as a similar concept in [McR83]. The first analyses of this method were mainly influenced by H. Yserentant in [Yse83], [Yse85], [Yse86] and [Yse86a].

In our thesis, the preconditioners are based on the algebraic concept of multigrid methods (AMG). In contrast to other (geometrical) multigrids, in these methods the coarser grids are not defined by a geometrical structure. The coarser grids and coarser operators are simply calculated by the elements of A itself. Hence it is possible, but no longer necessary, to have a geometrical stucture on which the problem is based. According to this concept, the multigrid methods can be used as black box solvers.
This method was developed by Brandt, McCormick and Ruge in [BMR82a], [BMR82b] and [BMR84]. Further evolutions of this idea are for example presented in [Brd86], [Stu83] and [Bra95]. Up to now there are many modifications and applications for this

method. The presented dissertation is one of the modifications of this idea.

The smoothed aggregation (SA) is also one of the modifications of the AMG. As the SA is similar to our modifications, we will briefly introduce this idea:
If there is a multiplicative multigrid method used to solve a system of linear equations, then we obtain on each grid $j = 0, \ldots, J$ a system of linear equations

$$A_j x^{*,j} = f^j.$$

Hence we require a smoother on each grid. If the smoother follows from a splitting method, then one iteration is defined as

$$x^{k+1,j} := \widetilde{S}_j(x^{k,j}) = M_j\, x^{k,j} + N_j f^j$$

with an iteration matrix M_j. The error after k iterations is defined as

$$e^{k,j} = x^{*,j} - x^{k,j}.$$

Components of the error which are not effectively removable by smoothing, i.e.

$$M_j e^{k,j} \approx e^{k,j},$$

are called smooth components (or the algebraic smooth error). The idea of the SA method is to reduce the smooth components of an error we have on grid j on another grid. Thus the modification handles a similar problem as the multigrid method itself. In fact the modification should use the same property more effectively than the multigrid method itself.
The solution proposed in the SA-method is to modifiy the prolongator. Mostly $\hat{P}$ is the common aggregation prolongator. Then $P = S\,\hat{P}$ is used where S is a smoother. In the main idea of SA, the restriction operator follows from $R = P^T$ and the matrix A is symmetric positive definite. As the method deals with the algebraic smooth error, some knowledge about this is assumed. Hence we have a bit less the property of a black box method. Mostly, the smoother S is a polynomial in A and the coefficients of the polynomial are influenced by the eigenvalues of A. This makes it obvious that some knowledge about the algebraic smooth error (or the matrix A) is used.
The method was introduced by Vanek in [Van92] and [Van95] and developed in [VBM96], [VBM01], [BrV90], [KrV96] and [VBT99].

Now we want to highlight some aspects of the SA-method in relation to the present dissertation. First of all, the SA-method generally uses $P = R^T$, which is not necessary for our modifications. Furthermore, the SA-method is only introduced for multiplicative multigrid methods and it is based on the idea that iterative methods are used instead of the exact inverse. In particular, it is pointless to use the exact inverse on any other grid than the coarsest one in a multiplicative method. Our modifications are introduced while we use the exact inverse on some subspaces. For additive methods, this is a possibility to analyse the system and as for some subspaces iterative methods converge fast this is a reasonable model system. Additionally, we have mentioned that the most results in the SA-method are for symmetric matrices. However, we will consider the unsymmetric case more in-depth.

At last we will present two modifications of the SA-method. The first one is called the adaptive smoothed aggregation (αSA) and is introduced in [BFLMRC04]. The idea of this method is to drop the knowledge of the algebraic smooth error. This error is estimated by the algorithm itself. Hence we need less information on the linear system of equations to obtain a fast solution. Based on this modification, the method is again a bit more a black box solver.
The other evolution is introduced in [GJV08]. In this paper we have $R \neq P^T$ because the smoother is only used to modify the prolongation. Hence we have $P = S\,\widehat{P}$ and $R = \widehat{P}^T$. Hence it is quite similar to our modification.
The difference, however, is that A is symmetric in [GJV08] and that the idea of modification is still given by the algebraic smooth error. The main difference is that the smoother S is a polynomial in A. Hence S and A commutate. Thus the method is analysed based on the idea that we have

$$R\,A\,S\,P = R\,S^{1/2}\,A\,S^{1/2}\,P.$$

Now we will outline how the presented thesis is organized.

In the second chapter we will briefly introduce a symmetric and an unsymmetric model problem. These are both based on PDEs. Then we will introduce some operators which are important for the algebraic multigrid method. In particular, we will give a matrix representations of the abstract operators used in our theory. This is nothing new, but it is rarely written down.

In the third chapter we will introduce the three preconditioners we will analyse in this

paper. In doing so we will present them as two grid methods. For all of them we will prove some basic characterisations as for example a sufficient condition for their non singularity. Two of them ($C_{BPX}^{-1}, C_{DT}^{-1}$) will be analysed in more detail. We will reduce the estimations to one parameter, which is based on the strengthened Cauchy-Schwarz inequality. This is a common instrument to analyse multigrid methods. It is used for example in [AxB84], [BaD81] and [Bra81]. Afterwards it is easily possible to consider the behaviour of the operators concerning this constant. In particular, it is possible to compare the operators to each other. Of course, the detailed analysis is more complex if the number of grids is raised.

In the fourth chapter we will study some modifications of the prolongation (one sided modification) or of both, the prolongation and the restriction (two sided modification). The main result is that the results of the previous chapter still hold if we change the projection or the spaces in which we decompose an element $v \in V$.

The modifications are presented in the fifth chapter for the model problems we have presented in the second chapter. Thus they are based on PDEs. Especially for those systems which are only based on a convection, we obtain perfect results. This is particulary the case for the one dimensional convection. In the case of more than one dimension, the theoretical results belong to a condition which is hard to control in a numerical algorithm. Nevertheless, we will see that the results are also perfect in the two dimensional situation.

In the sixth and the seventh chapter we will present some aspects for the multigrid situation. In doing so the main interest is to obtain a condition concerning the non singularity of the preconditioners. This is done for the unmodified preconditioners in the sixth chapter and for the modifications in the seventh chapter.

In the eighth chapter we will consider the case of a symmetric matrix A for the unmodified preconditioners in more detail. The analysis of the DT-method in this situation will lead to the rule for aggregation we use in numerical examples. Furthermore, the quality of the preconditioner for a given system will be expressed by a constant which accentuates the black box character of the algebraic aggregation as used.

The ninth chapter is divided into two parts. In the first one we will briefly summarize some properties of matrices which are useful for iterative methods. Afterwards we

consider the properties that are maintained for the coarser operators. In the unmodified system, all properties remain true, but for the modified preconditioners, this is not generally the case. In the second part of the ninth chapter we will present numerical results for the different methods and modifications. The implementation is done in *FORTRAN 90*.

2 Definition of grids, function spaces and operators

2.1 Basics of finite elements

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain. Then we assume that we can decompose $\overline{\Omega}$ into closed triangles or squares Λ so, that they hold

$$\overline{\Omega} = \overline{\Omega}_h = \bigcup \Lambda$$

and fulfil the following condition R.

Condition R: The intersection of two different triangles (squares) is empty, a common edge or a common vertex. Each triangle (square) includes a circle of the radius $c_R \cdot h$ and is included in a circle of the radius $c_R \cdot h^{-1}$, where c_R does not depend on h.

As our finite elemente space we define

$$V := S_{0,h} = \{v_h \in C(\overline{\Omega}_h) \,:\, v_h|_\Lambda \text{ is linear and } v_h|_{\partial\Omega_h} = 0\} \tag{2.1}$$

$$\text{or} \quad V := S_{0,h} = \{v_h \in C(\overline{\Omega}_h) \,:\, v_h|_\Lambda \text{ is bilinear and } v_h|_{\partial\Omega_h} = 0\} \tag{2.2}$$

In this case the nodal base of V can be constructed as follows: Let $\mathcal{N}_1, \ldots, \mathcal{N}_n$ be the vertices or nodes of the triangles (squares) $\{\Lambda\}$ which are in the interior of Ω. Let then $\varphi_{h,i} \in V$ be the linear or bilinear functions for $i = 1, \ldots, n$ with

$$\varphi_i(\mathcal{N}_j) = \delta_{i,j},$$

where $\delta_{i,j}$ is the Kronecker δ. Thus, for $v \in V$ we have the unique representation

$$v(x) = \sum_{i=1}^{n} v(\mathcal{N}_i)\, \varphi_i(x).$$

Thus, the dimension of the space V is given by the number of vertices. We set for $i = 1, \ldots, n$ the unit vectors e_i for $(\varphi_i)_{i=1,\ldots,n}$ and represent v by the vector

$$v = (v(\mathcal{N}_1), \ldots, v(\mathcal{N}_n))\,.$$

2.2 Modell problems

In this section we will introduce some model problems given by partial differential equations. Then we will denote the stencils we get by the finite element method or the finite differences method for these problems. These stencils give us the structure of the matrices we will use as examples for our preconditioners.

Symmetric modell problem: As a first example let us consider the equation

$$-div(\alpha(x)\, grad\, u(x)) = f(x), \quad \forall x \in \Omega$$
$$u(x) = g(x), \quad \forall x \in \partial\Omega.$$

Furthermore, we assume

$$\alpha(x) = \begin{pmatrix} a(x) & 0 \\ 0 & b(x) \end{pmatrix} \tag{2.3}$$

with $a(x), b(x) \in C^1(\overline{\Omega})$, $a(x), b(x) > 0$ for all $x \in \Omega$ and $f(x) \in C(\overline{\Omega})$. The weak solution of this problem for a given $f \in L^2(\Omega)$ is given by a function $u \in H_0^{1,2}(\Omega)$ which, for all $\phi \in C_0^\infty(\Omega)$ fulfils the equation

$$\int_\Omega a(x)\, \frac{\partial u}{\partial x_1}\, \frac{\partial \phi(x)}{\partial x_1} + b(x)\, \frac{\partial u}{\partial x_2}\, \frac{\partial \phi(x)}{\partial x_2}\, dx = \int_\Omega f(x)\, \phi(x)\, dx.$$

We get the finite problem if we set $\phi_i \in V$ instead of $\phi \in C_0^\infty(\Omega)$. Since the matrix $\alpha(x)$ in (2.3) is symmetric this also holds for the stiffness matrix we get. This matrix is induced by the stencils

$$\begin{pmatrix} -\delta_{nw} & -\varepsilon_n & -\delta_{ne} \\ -\varepsilon_w & m & -\varepsilon_e \\ -\delta_{sw} & -\varepsilon_s & -\delta_{se} \end{pmatrix} \tag{2.4}$$

$$\text{with} \quad m = \varepsilon_w + \varepsilon_e + \varepsilon_n + \varepsilon_s + \delta_{nw} + \delta_{ne} + \delta_{se} + \delta_{sw}$$
$$\text{and} \quad \varepsilon_i > 0, \quad \text{for} \quad i = w, e, n, s$$
$$\delta_i > 0, \quad \text{for} \quad i = nw, ne, se, sw.$$

Furthermore the coefficients ε_i, δ_i are functions of $a(x), b(x)$.

If $a(x), b(x)$ are constant then we obtain for linear functions $\delta_i = 0$ for $i = nw, ne, se, sw$. For bilinear elements, this is the case if we approximate the integrals on the quadratic elements Λ of the area h^2 with the vetices x_i, $i = 1, \ldots, 4$ by

$$\int_\Lambda g(x)\,dx \approx I_\Lambda(g(x)) := h^2 \sum_{i=1}^{4} g(x_i).$$

In general, the values depend on the approximation of the integral. For the purpose of an example, it is sufficient to take the structure as given in the stencil (2.4).

Convections diffusion equation: As an unsymmetric example we consider the equation

$$b_1(x)\,\frac{\partial u}{\partial x_1} + b_2(x)\,\frac{\partial u}{\partial x_2} - \varepsilon \triangle u(x) = f \quad \forall x \in \Omega \tag{2.5}$$

$$u(x) = g(x) \quad \forall x \in \partial\Omega.$$

Thereby is $b \in C(\overline{\Omega})$ and $\varepsilon \in \mathbb{R}_+$. In this case we use the upwind method for finite differences for the discretization. Therewith we get for $\varepsilon > 0$ with

$$m = 4\varepsilon + |b_1|\,h + |b_2|\,h$$

the stencils

$$\begin{pmatrix} 0 & -\varepsilon & 0 \\ -b_1\,h - \varepsilon & m & -\varepsilon \\ 0 & -b_2\,h - \varepsilon & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & b_2\,h - \varepsilon & 0 \\ -b_1\,h - \varepsilon & m & -\varepsilon \\ 0 & -\varepsilon & 0 \end{pmatrix}$$

for $b_1, b_2 \geq 0$ for $b_1 \geq 0,\ b_2 < 0$

$$\begin{pmatrix} 0 & b_2\,h - \varepsilon & 0 \\ -\varepsilon & m & b_1\,h - \varepsilon \\ 0 & -\varepsilon & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & -\varepsilon & 0 \\ -\varepsilon & m & b_1 h - \varepsilon \\ 0 & -b_2\,h - \varepsilon & 0 \end{pmatrix}$$

for $b_1, b_2 < 0$ for $b_1 < 0,\ b_2 \geq 0$

2.3 Subspaces, prolongation and restriction

For a linear vector space V of dimension n we will define some subspaces V_i of the dimension $n_i < n$. We will do this by using restriction operators. That way we can

represent V_i by a n_i-dimensonal subspace of $\mathbb{R}^n$ or by $\mathbb{R}^{n_i}$. As we will reduce the dimension to solve linear equations, it is important for us to represent them by $\mathbb{R}^{n_i}$. However, to add elements of different subspaces, we need the representation for all of them as an element of $\mathbb{R}^n$. Thus we define restriction operators by matrices of the dimension $n_i \times n$. To represent an element $v_i \in V_i \equiv \mathbb{R}^{n_i}$ that is given as an element of $\mathbb{R}^{n_i}$ we need prolongation operators of the dimension $\mathbb{R}^{n \times n_i}$.

As we will also consider the preconditioners as black box methods for linear equation systems, we will give a definition of the subspaces and operators that has nothing to do with partial differential equations, finite elements, finite differences and so forth, we will also introduce the subspaces and operators we consider only by matrices.
Some operators will be introduced twice: By a definition for the image of the basis functions and by a matrix representation. If it is necessary, we will show that the given matrix is the representation of the respective operator.

2.3.1 Definition of restriction, prolongation and subspaces $V_i \subset V$.

For the space V that is given by

$$V =< \varphi_1, \ldots, \varphi_n >$$

with the basis functions φ_i, $i = 1, \cdots, n$ we set $V = V_J$ and generate recursive subspaces V_i, $i = 0, \ldots, J-1$ that fulfil

$$V_J \supset V_{J-1} \supset \cdots \supset V_0.$$

For the basis functions φ_i, $i = 1, \ldots, n$ of V, we set $\varphi_{J,i} = \varphi_i$, for $i = 1, \ldots, n_J = n$. Assume now that for $j \leq J$ and for all $j \leq k \leq J$, the spaces V_k are defined by the basis functions $\varphi_{k,i}$, $i = 1, \ldots, n_k$. Then we define a linear restriction operator R^j_{j-1} and so V_{j-1} is defined as

$$V_{j-1} = \left\langle R^j_{j-1}\varphi_{j,1}, \ldots, R^j_{j-1}\varphi_{j,n_j} \right\rangle.$$

Alternatively we can define basis functions $\varphi_{j-1,i}$ for $i = 1, \ldots, n_{j-1}$ by

$$\varphi_{j-1,i} := \sum_{s=1}^{n_j} r_{i,s}\varphi_{j,s}. \tag{2.6}$$

Then we set $V_{j-1} =< \varphi_{j-1,1}, \ldots, \varphi_{j-1,n_{j-1}}. >$

If we associate V_j with $\mathbb{R}^{n_j}$ and V_{j-1} with $\mathbb{R}^{n_{j-1}}$ we write $\widetilde{V}_j$ and $\widetilde{V}_{j-1}$ respectively. In this case, (2.6) gives a matrix representation for R^j_{j-1}, and it is $R^j_{j-1} \in \mathbb{R}^{n_{j-1} \times n_j}$. Since $< \varphi_{j-1,1}, \ldots, \varphi_{j-1,n_{j-1}} >$ is a basis of V_{j-1}, it is obvious that we get $rk(R^j_{j-1}) = n_{j-1}$. Further, we define for the prolongation also a linear operator $P^{j-1}_j : V_{j-1} \to V_j$ by the definition for the basis functions. Because of $V_{j-1} \subset V_j$, the identity is a quite common choice for P^{j-1}_j. If we use the vector representation by an element of $\widetilde{V}_{j-1} \equiv \mathbb{R}^{n_{j-1}}$ or $\widetilde{V}_j \equiv \mathbb{R}^{n_j}$ respectively for the elements $\widetilde{v}_{j-1} \in \widetilde{V}_{j-1}, \widetilde{v}_j \in \widetilde{V}_j$ we need the prolongation operator as a matrix $P^{j-1}_j \in \mathbb{R}^{n_j \times n_{j-1}}$. The simplest choice is to set $P^{j-1}_j = (R^j_{j-1})^T$. In this work we will only consider the situation

$$P^{j-1}_j := (R^j_{j-1})^T.$$

Hence we define the prolongation this way. As already mentioned in the introduction we will also consider situations in which we only modify the prolongation. But this will be obvious by another notation. At least P^{j-1}_j should fulfil $rk(P^{j-1}_j) = n_{j-1}$. For $P^{j-1}_j = (R^j_{j-1})^T$ this is obvious fulfilled.

If we have defined the operators R^j_{j-1} and P^{j-1}_j for $j = 1, \ldots, J$ then we can define the following operators for an easier notation:

1. Based on the definition of R^j_{j-1} for $j = 1, \ldots, J$ we define $R^l_k : V_l \to V_k$ for $0 \le k < l \le J$ by

   $$R^l_k := R^{k+1}_k \circ \cdots \circ R^l_{l-1}.$$

   In particular, we set for $j = 0, \ldots, J$

   $$R_j := R^J_j \quad \text{and} \quad R_J = I_n.$$

2. Based on the definition of P^{j-1}_j for $j = 1, \ldots, J$ we define for $P^k_l : V_k \to V_l$ $0 \le k < l \le J$ by

   $$P^k_l := P^{l-1}_l \circ \cdots \circ P^k_{k+1}.$$

   In particular, we set for $j = 0, \ldots, J$

   $$P_j := P^j_J \quad \text{and} \quad P_J = I_n.$$

At this point it is irrelevant whether the operators are identified by matrices or not. If we have the representation by matrices and we have a linear operator $A = A_J : V \to V$, we define coarser operators A_j for $j = 0, \ldots, J-1$ iteratively by

$$A_j = R_j^{j+1} A_{j+1} P_{j+1}^j. \tag{2.7}$$

This immediately implies

$$A_j = R_j A P_j$$

and, based on the assumption of $P_j = (R_j)^T$, this implies for $v_j = P_j \widetilde{v}_j$ and $w_j = P_j \widetilde{w}_j$

$$(A_j \widetilde{v}_j, \widetilde{w}_j) = (A P_j \widetilde{v}_j, P_j \widetilde{w}_j) = (A v_j, w_j).$$

If we interpret P_j as the identity and so $\widetilde{v}_j, \widetilde{w}_j$ and v_j, w_j as representations of the same elements in different spaces,the dot product is independent of the space in which we consider the elements.

Furthermore, we define the operators $Q_j, \widehat{Q}_j$ for $j = 0, \ldots, J-1$ by

$$Q_j : \widetilde{V}_{j+1} \to P_{j+1}^j(\widetilde{V}_j) \quad \text{with}$$

$$(\widetilde{v}_{j+1}, P_{j+1}^j \widetilde{v}_j) = (Q_j \widetilde{v}_{j+1}, P_{j+1}^j \widetilde{v}_j), \quad \text{for all} \quad \widetilde{v}_{j+1} \in \widetilde{V}_{j+1}, \widetilde{v}_j \in \widetilde{V}_j \tag{2.8}$$

$$\widehat{Q}_j : V_J \to V_j \quad \text{with}$$

$$(v_J, v_j) = (\widehat{Q}_j v_J, v_j), \quad \text{for all} \quad v_j \in V_j, v_J \in V_J. \tag{2.9}$$

Remark: 2.3.1. *With these definitions, it is obvious that $Q_j, \widehat{Q}_j$ are the orthogonal projections with respect to the inner products $(.,.)$.*

And at last we define the vector spaces $W_j^t, \widetilde{W}_j$ and W_j for $j = 1, \ldots, J$ with the basis $\varphi_{J,i}$, $i = 1, \ldots, n_J$ by

$$W_j^t := \left\langle (\widehat{Q}_j - \widehat{Q}_{j-1})\varphi_{J,1}, \ldots, (\widehat{Q}_j - \widehat{Q}_{j-1})\varphi_{J,n} \right\rangle$$

$$\widetilde{W}_j := \langle (I_j - Q_{j-1}) R_j \varphi_{J,1}, \ldots, (I_j - Q_{j-1}) R_j \varphi_{J,n} \rangle$$

$$W_j := \langle P_j (I_j - Q_{j-1}) R_j \varphi_{J,1}, \ldots, P_j (I_j - Q_{j-1}) R_j \varphi_{J,n} \rangle .$$

2.3.2 Basic results for matrices

First we define the i-th unitvector of $\mathbb{R}^{n_j}$ by e_i^j for $j = 0, \ldots, J$ and $i = 1, \ldots, n_j$ and the identity matrix of the dimension $n_j \times n_j$ by I_j.

With the definitions of the prolongation and restriction operators we will now assume the following characteristics for their matrix representations :

$$R_j^{j+1} \in \mathbb{R}^{n_j \times n_{j+1}}, \quad rk(R_j^{j+1}) = n_j$$
$$P_{j+1}^j = (R_j^{j+1})^T \in \mathbb{R}^{n_{j+1} \times n_j}, \quad rk(P_{j+1}^j) = n_j.$$

Hence follows immediately that

$$R_j^{j+k} \in \mathbb{R}^{n_j \times n_{j+k}}, \quad rk(R_j^{j+k}) = n_j$$
$$P_{j+k}^j = (R_j^{j+k})^T \in \mathbb{R}^{n_{j+k} \times n_j}, \quad rk(P_{j+k}^j) = n_j.$$

Further, we define the matrices $S_j \in \mathbb{R}^{n_j \times n_j}, \widehat{S}_j \in \mathbb{R}^{n_j \times n_j}$ for $j = 0, \ldots J$ by

$$S_j := (R_j^{j+1} P_{j+1}^j)^{-1} \quad \text{and} \quad \widehat{S}_j := (R_j P_j)^{-1}.$$

Therewith we obtain the follwing relation.

Remark: 2.3.2. *By the definitions for P_{j+1}^j, R_j^{j+1},*

$$(R_j^{j+1} P_{j+1}^j)^T = R_j^{j+1} P_{j+1}^j \quad \textit{and} \quad (R_j P_j)^T = R_j P_j.$$

holds for all $j = 0, \ldots, J$. Furthermore $R_j^{j+1} P_{j+1}^j$ and $R_j P_j$ are positive definite.

proof. First, just by the definition of $P_{j+1}^j = (R_j^{j+1})^T$ we obtain $P_j = R_j^T$. This implies

$$(R_j^{j+1} P_{j+1}^j)^T = (P_{j+1}^j)^T (R_j^{j+1})^T = R_j^{j+1} P_{j+1}^j$$
$$\text{and} \quad (R_j P_j)^T = P_j^T R_j^T = R_j P_j.$$

Further, for a $\widetilde{v}_j \in \mathbb{R}^{n_j}$ we get

$$(R_j^{j+1} P_{j+1}^j \widetilde{v}_j, \widetilde{v}_j) = (P_{j+1}^j \widetilde{v}_j, P_{j+1}^j \widetilde{v}_j) = \|P_{j+1}^j \widetilde{v}_j\|^2 \geq 0.$$

And as $P_{j+1}^j \in \mathbb{R}^{n_{j+1} \times n_j}$ has rank n_j, we have $\|P_{j+1}^j \widetilde{v}_j\|^2 = 0$ if and only if we have $\widetilde{v}_j = 0$. This proves that $R_j^{j+1} P_{j+1}^j$ is positive definite. By the same arguments, the propositions for $R_j P_j$ holds. □

From Remark 2.3.2 follows in particular $rk(R_j^{j+1} P_{j+1}^j) = rk(R_j P_j) = n_j$. Hence the operators $S_j, \widehat{S}_j$, $j = 0, \ldots, J$ are well posed. Further, we get the following basic propositions for them:

Remark: 2.3.3. *From the definitions for S_j, $\widehat{S}_j$ it follows that $S_j, \widehat{S}_j$ are symmetric and positive definite (s.p.d.) for $j = 0, \ldots, J$.*

proof. With Remark 2.3.2, these characteristics hold for S_j^{-1} and $\widehat{S}_j^{-1}$ respectivly. Hence the proposed characteristics follow for the operators $S_j, \widehat{S}_j$. □

Lemma: 2.3.4. *By the definitions of section 2.3*

1. *$Q_j = P_{j+1}^j S_j R_j^{j+1}$ and $\widehat{Q}_j = P_j \widehat{S}_j R_j$, holds for $j = 0, \ldots, J-1$.*
2. *The operators $(I_{j+1} - Q_j) : \widetilde{V}_{j+1} \to (P_{j+1}^j(\widetilde{V}_j))^\perp$ and $(I - \widehat{Q}_j) : V \to V_j^\perp$ are the orthogonal projection concerning the inner product $(.,.)$.*
3. *We have $Q_{j-1}\widetilde{v}_j = 0$ for an $\widetilde{v}_j \in \widetilde{V}_j$, if and only if we have $R_{j-1}^j \widetilde{v}_j = 0$. We have $\widehat{Q}_j v = 0$ for an $v \in V$, if and only if we have $R_j v_j = 0$.*
4. *$\widehat{Q}_j \widehat{Q}_k v = \widehat{Q}_k \widehat{Q}_j v = \widehat{Q}_j v$ holds for $j \leq k$.*

proof. 1. Because of the uniqueness of the orthogonal projection with respect to a given inner product, it is sufficient to prove that the operators $Q_j, \widehat{Q}_j$ define each an orthogonal projection. By the definition of S_j, we have for $j = 0, \ldots, J-1$

$$\begin{aligned}(Q_j)^2 &= P_{j+1}^j S_j R_j^{j+1} P_{j+1}^j S_j R_j^{j+1} \\ &= P_{j+1}^j (R_j^{j+1} P_{j+1}^j)^{-1} R_j^{j+1} P_{j+1}^j S_j R_j^{j+1} \\ &= P_{j+1}^j S_j R_j^{j+1} = Q_j.\end{aligned}$$

Hence Q_j is a projection. From the symmetry of S_j follows with $P_{j+1}^j = (R_j^{j+1})^T$

$$(Q_j)^T = (P_{j+1}^j S_j R_j^{j+1})^T = (R_j^{j+1})^T (S_j)^T (P_{j+1}^j)^T = P_{j+1}^j S_j R_j^{j+1} = Q_j.$$

This proves the orthogonality with respect to the inner product $(.,.)$. The results for $\widehat{Q}_j$ follow by the same arguments.

2. As Q_j and $\widehat{Q}_j$ are orthogonal projections this also holds for $(I_j - Q_j)$ and $(I - \widehat{Q}_j)$.

3. From the definition of Q_{j-1} as $Q_{j-1} = P_j^{j-1}\, S_{j-1}\, R_{j-1}^j$ it is obvious that $R_{j-1}^j \widetilde{v}_j = 0$ implies $Q_{j-1}\, \widetilde{v}_j = 0$. For the other implication we have that S_{j-1} is positive definite and P_j^{j-1} has rank n_{j-1}. For $w_j \neq 0$, we therefore obatin also $P_j^{j-1}\, S_{j-1}\, \widetilde{w}_j \neq 0$. So $Q_{j-1}\widetilde{v}_j = 0$ implies $R_{j-1}^j\, \widetilde{v}_j = 0$. The propositions for $\widehat{Q}_j$ and R_j follow again by the same arguments.

4. For $j \leq k$ we get

$$\begin{aligned}\widehat{Q}_j\, \widehat{Q}_k\, v &= P_j\, \widehat{S}_j\, R_j\, P_k\, \widehat{S}_k\, R_k\, v \\ &= P_j\, \widehat{S}_j\, R_j^k\, R_k\, P_k\, \widehat{S}_k\, R_k\, v \\ &= P_j\, \widehat{S}_j\, R_j^k\, R_k\, v = \widehat{Q}_j\, v.\end{aligned}$$

   By the same arguments follows $\widehat{Q}_k\, \widehat{Q}_j\, v = \widehat{Q}_j\, v$.

□

The fourth proposition of Lemma 2.3.4 immediately implies that

$$\begin{aligned}((\widehat{Q}_i - \widehat{Q}_{i-1})\, v,\, (\widehat{Q}_i - \widehat{Q}_{i-1})\, v) &= ((\widehat{Q}_i - \widehat{Q}_{i-1})\, v,\, v), \quad \text{for} \quad i = 1, \ldots, J \\ (\widehat{Q}_0\, v,\, \widehat{Q}_0\, v) &= (\widehat{Q}_0\, v,\, v) \\ ((\widehat{Q}_i - \widehat{Q}_{i-1})\, v,\, (\widehat{Q}_j - \widehat{Q}_{j-1})\, v) &= 0, \quad \text{for} \quad i, j = 1, \ldots, J, \quad i \neq j \\ \text{and} \quad ((\widehat{Q}_i - \widehat{Q}_{i-1})\, v,\, \widehat{Q}_0\, v) &= 0, \quad \text{for} \quad i = 1, \ldots, J.\end{aligned}$$

These characteristics follow immediately as the operators are all orthogonal projections with respect to the standard inner product.

The operators $A_j \in \mathbb{R}^{n_j \times n_j},\ j = 0, \ldots, J-1$ as defined in (2.7) represent the operator A on the subspaces $V_j,\ j = 0, \ldots, J-1$. So as we want to use the operators A_j to solve for an given operator $A \in \mathbb{R}^{n \times n}$ and a given $f \in \mathbb{R}^n$ the equation $A\, u = f$ we need that the operators A_j are non singular.

Lemma: 2.3.5. *1. Let $A \in \mathbb{R}^{n \times n}$ be a non singular matrix. Then it follows that A_j is non singular if and only if there is no $v_j \in V_j$ with $A\, v_j \in V_j^{\perp}$.*

2. If A is s.p.d then is A_j for all $j = 0, \ldots, J-1$ s.p.d.
In particular this implies that A_j is non singular.

3. If A is real positive then is A_j for all $j = 0, \ldots, J-1$ real positive.
In particular this implies that A_j is non singular.

proof. 1. For an arbitrary $j = 0, \ldots, J-1$ is $\widehat{Q}_j = P_j \, \widehat{S}_j \, R_j$ the orthogonal projection $V \to V_j$. And as it is $\widehat{Q}_j \, v = 0$ if and only if it is $R_j \, v = 0$ we obtain $R_j \, v = 0$ if and only if it is $v \in V_j^{\perp}$. As we have $rk(P_j) = n_j$ for $P_j \in \mathbb{R}^{n_J \times n_j}$ we obtain that there is an $\widetilde{v}_j \in \widetilde{V}_j$ with

$$A_j \widetilde{v}_j = R_j \, A \, P_j \, \widetilde{v}_j = 0$$

if and only if there is an $v_j = P_j \, \widetilde{v}_j \in V_j$ that holds $R_j \, A \, v_j = 0$. And this is equivalent to $A \, v_j \in V_j^{\perp}$.

2. If A is s.p.d. then it follows

$$A_j^T = (R_j \, A \, P_j)^T = P_j^T \, A^T \, R_j^T = R_j \, A \, P_j = A_j.$$

and for an arbitrary $\widetilde{v}_j \in \widetilde{V}_j$ with $\widetilde{v}_j \neq 0$

$$\widetilde{v}_j^T \, A_j \, \widetilde{v}_j = \widetilde{v}_j^T \, R_j \, A \, P_j \, \widetilde{v}_j = (P_j \, \widetilde{v}_j)^T \, A \, \underbrace{P_j \, \widetilde{v}_j}_{\neq 0} > 0.$$

This implies that A_j is s.p.d. and therewith non singular.

3. If A is real positiv then it follows $a_{i,k}^j \in \mathbb{R}$ for all elements $a_{i,k}^j$ of $A_j = R_j \, A \, P_j$. Again for an arbitrary $\widetilde{v}_j \in \widetilde{V}_j$ with $\widetilde{v}_j \neq 0$ we obtain

$$\widetilde{v}_j^T \, A_j \, \widetilde{v}_j = \widetilde{v}_j^T \, R_j \, A \, P_j \, \widetilde{v}_j = (P_j \, \widetilde{v}_j)^T \, A \, \underbrace{P_j \, \widetilde{v}_j}_{\neq 0} > 0.$$

This implies that A_j is real positive and therewith non singular.

□

Corollary: 2.3.6. *Let $A \in \mathbb{R}^{n \times n}$ be a non singular matrix. Then it follows that A_j is non singular if and only if there is no $w_{j+1} \in V_j^{\perp}$ that holds $A^{-1} w_{j+1} \in V_j$.*

proof. The proposition is equivalent to the first proposition of Lemma 2.3.5. □

The following technical aspect we will use frequently:

Lemma: 2.3.7. *For $j = 1, \ldots, J-1$ it is*

$$R_{j-1}^j (I_j - Q_{j-1}) = 0.$$

proof. Based on the definition of S_{j-1} we obtain

$$\begin{aligned} R^j_{j-1}(I_j - Q_{j-1}) &= R^j_{j-1}(I_j - P_j^{j-1}\, S_{j-1}\, R^j_{j-1}) \\ &= R^j_{j-1} - R^j_{j-1}\, P_j^{j-1}\, S_{j-1}\, R^j_{j-1} = R^j_{j-1} - R^j_{j-1}. \end{aligned}$$

□

Now we will give an alternative representation for $\widehat{Q}_j$, that applies under a certain condition.

Lemma: 2.3.8. *By the definitions of $P_j^{j-1}, S_{j-1}, \widehat{S}_j$ and $\widehat{S}_{j-1}$ we have for $j = 1, \ldots, J-1$*

$$\widehat{Q}_{j-1} = P_j\, \widehat{S}_j\, Q_{j-1}\, R_j$$

if and only if we have

$$\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}.$$

proof. By the definition of $\widehat{Q}_{j-1}$ by $\widehat{Q}_{j-1} = P_{j-1}\, \widehat{S}_{j-1}\, R_{j-1}$ and because of $rk(P_j) = rk(R_j) = n_j$ we obtain

$$\begin{aligned} \widehat{Q}_{j-1} = P_j\, \widehat{S}_j\, Q_{j-1}\, R_j = P_j\, \widehat{S}_j\, P_j^{j-1}\, S_{j-1}\, R^j_{j-1}\, R_j \\ \Leftrightarrow \quad P_{j-1}\, \widehat{S}_{j-1}\, R_{j-1} = P_j\, \widehat{S}_j\, P_j^{j-1}\, S_{j-1}\, R_{j-1} \\ \Leftrightarrow \quad P_j^{j-1}\, \widehat{S}_{j-1} = \widehat{S}_j\, P_j^{j-1}\, S_{j-1}. \end{aligned}$$

This is the proposition. □

The meaning of this lemma is that the operators $\widehat{S}_j, P_j^{j-1}$ commutate and $\widehat{S}_{j-1} = \widehat{S}_j \circ S_{j-1}$ holds. If we consider the matrix representations of the operators, the equation $\widehat{S}_{j-1} = \widehat{S}_j \circ S_{j-1}$ is not well posed just by the dimensions of the matrices. But even if we use the definitions of the operators just by their effect, the equation $\widehat{S}_{j-1} = \widehat{S}_j \circ S_{j-1}$ does not hold in all situations. Furthermore it is obvious that the equations

$$\widehat{S}_{j-1} = \widehat{S}_j \circ S_{j-1} \quad \text{and} \quad \widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$$

are the same if we interpret P_j^{j-1} as the identity. Remember that we have done this as we have associated the spaces $\mathbb{R}^n, \mathbb{R}^{n_i}$ with the spaces of linear or bilinear functions.

Lemma 2.3.8 immediately implies that if the assumption of this lemma is fulfilled

$$\begin{aligned} P_j\,\widehat{S}_j\,(I_j - Q_{j-1})\,R_j &= P_j\,\widehat{S}_j\,(I_j - P_j^{j-1}\,S_{j-1}\,R_{j-1}^j)\,R_j \\ &= \widehat{Q}_j - P_j\,\widehat{S}_j\,P_j^{j-1}\,S_{j-1}\,R_{j-1}^j\,R_j \\ &= \widehat{Q}_j - \widehat{Q}_{j-1} \quad \text{follows.} \end{aligned}$$

Furthermore, it follows in this case that the matrices $\widehat{S}_j$, $j = 0, \ldots, J-1$ are not needed to get matrix representations for orthogonal projections $\widehat{Q}_j$. This results from the following corollary.

Corollary: 2.3.9. *If $\widehat{S}_{j+1}\,P_{j+1}^j\,S_j = P_{j+1}^j\,\widehat{S}_j$ holds for all $j = 0, \ldots, J-1$, it follows for all $v \in V$*

$$P_j\,\widehat{S}_j = P_J^{J-1}\,S_{J-1}\,P_{J-1}^{J-2}\,S_{J-2}\ldots P_{j+1}^j\,S_j$$

$$\textit{and} \quad \widehat{Q}_j = P_J^{J-1}\,S_{J-1}\,P_{J-1}^{J-2}\,S_{J-2}\ldots P_{j+1}^j\,S_j\,R_j.$$

proof. By the representation $\widehat{Q}_j = P_j\,\widehat{S}_j\,P_j$ for $j = 0, \ldots, J-1$, the second proposition follows immediately from the first one. The first proposition obviously holds for $j = J-1$ according to the assumption. Assume now that the equation holds for $k > j$. Then we obtain

$$\begin{aligned} P_j\,\widehat{S}_j &= P_J^{j+1}\,P_{j+1}^j\,\widehat{S}_j \\ &= P_J^{j+1}\,\widehat{S}_{j+1}\,P_{j+1}^j\,S_j \\ &= P_J^{J-1}\,S_{J-1}\,P_{J-1}^{J-2}\,S_{J-2}\ldots P_{j+1}^j\,S_j. \end{aligned}$$

□

Further Lemma 2.3.8 directly implies the following representation of the norm:

Corollary: 2.3.10. *If $\widehat{S}_i\,P_i^{i-1}\,S_{i-1} = P_i^{i-1}\,\widehat{S}_{i-1}$ holds, then it follows for all $v \in V$ that*

$$\|(\widehat{Q}_i - \widehat{Q}_{i-1})v\|^2 = (\widehat{S}_i\,(I_i - Q_{i-1})\,R_i\,v,\,(I_i - Q_{i-1})\,R_i\,v)$$

proof. By the assumption we obtain for an arbitrary $v \in V$

$$\begin{aligned} &(\widehat{S}_i\,(I_i - Q_{i-1})\,R_i\,v, (I_i - Q_{i-1})\,R_i\,v) \\ &\qquad = (\widehat{S}_i\,(I_i - Q_{i-1})\,R_i\,v,\,R_i\,P_i\,\widehat{S}_i\,(I_i - Q_{i-1})\,R_i\,v) \\ &\qquad = (P_i\,\widehat{S}_i\,(I_i - Q_{i-1})\,R_i\,v,\,P_i\,\widehat{S}_i\,(I_i - Q_{i-1})\,R_i\,v) \\ &\qquad = ((\widehat{Q}_i - \widehat{Q}_{i-1})v,\,(\widehat{Q}_i - \widehat{Q}_{i-1})v). \end{aligned}$$

□

2.4 The aggregation method

2.4.1 The general setting

Now we will introduce the aggregation method. Hence we set for the nodes $\mathcal{N}_i^J = \{\mathcal{N}_i\}$, for $i = 1, \ldots, n$. If $\{\mathcal{N}_i^j\}_{i=1,\ldots,n_j}$ and $\varphi_{j,i}$ are defined for $j \in \{1, \ldots, J\}$, then we define some kind of node-like sets

$$\mathcal{N}_i^{j-1} \subset \{\mathcal{N}_1^j, \ldots, \mathcal{N}_{n_j}^j\}, \quad i = 1, \ldots, n_{j-1}$$

in such a way that we have

$$\bigcup_{i=1}^{n_{j-1}} \mathcal{N}_i^{j-1} = \{\mathcal{N}_1^j, \ldots, \mathcal{N}_{n_j}^j\} \quad \text{and} \quad \mathcal{N}_i^{j-1} \cap \mathcal{N}_k^{j-1} = \emptyset \quad \text{for} \quad k \neq i.$$

Furthermore, we define the following sets of indices

$$I_i^{j-1,j} := \{l \in \{1, \ldots, n_j\} \,|\, \mathcal{N}_l^j \subset \mathcal{N}_i^{j-1}\}.$$

So the aggregation method is defined as we set for $i = 1, \ldots, n_{j-1}$

$$\varphi_{j-1,i}(x) := \sum_{k \in I_i^{j-1,j}} \varphi_{j,k}(x).$$

V_{j-1} is defined as

$$V_{j-1} =< \varphi_{j-1,1}, \ldots, \varphi_{j-1,n_{j-1}} >$$

and we obtain $dim(V_{j-1}) = n_{j-1}$. As $v \in V = V_J$ has a unique representation $v \in \mathbb{R}^n$, each $v_j \in V_j$, $j = 0, \ldots, J$ has a unique representation $v_j \in \mathbb{R}^{n_j}$. We get these representations the same way. For $j = 0, \ldots, J$ and $i = 1, \ldots, n_j$, we set the unit vector $e_i^j \in \mathbb{R}^{n_j}$ for $\varphi_{j,i}$. We obtain

$$v_j(x) = \sum_{i=1}^{n_i} v_j(\mathcal{N}_i^j)\varphi_{j,i}(x) \equiv \left(v(\mathcal{N}_1^j), \ldots, v(\mathcal{N}_n^j)\right).$$

With the definition of the spaces $V_0, \ldots, V_J$, we can also define the sets

$$I_i^{j-k,j} := \{l \in \{1, \ldots, n_j\} \,|\, \mathcal{N}_l^j \subset \mathcal{N}_i^{j-k}\}$$

and in particular the sets

$$I_i^k := \{l \in \{1, \ldots, n_J\} \,|\, \mathcal{N}_l^J \subset \mathcal{N}_i^k\}.$$

Now we can define the following expressions: We say that two points (or sets) $\mathcal{N}_i^j, \mathcal{N}_k^j$, $i \neq k$ of level j are *aggregated* if and only if $i, k \in I_t^{j-1,j}$ holds for a $t = 1, \ldots, n_{j-1}$. Further, we say that $\mathcal{N}_i^j$ is *isolated* if and only if $i \in I_s^{j-1,j}$ and $|I_s^{j-1,j}| = 1$ hold.
More generally, we say for $1 \leq j \leq J$ and $1 \leq k \leq j$ that the points $\mathcal{N}_{i(1)}^j, \ldots, \mathcal{N}_{i(l)}^j$ are aggregated to $\mathcal{N}_t^{j-k}$ if $i(1), \ldots, i(l) \in I_t^{j-k,j}$ holds. Further, we say in this case $\mathcal{N}_{i(1)}^j, \ldots, \mathcal{N}_{i(l)}^j \subset \mathcal{N}_t^{j-k}$.

Now for $j = 1, \ldots, J$ the linear restriction operators $R_{j-1}^j : V_j \to V_{j-1}$ which imply these subspaces are given as follows:

$$R_{j-1}^j \varphi_{j,i} = \varphi_{j-1,k}, \quad \text{with} \quad i \in I_k^{j-1,j}.$$

For the linear prolongation operators $P_j^{j-1} : V_{j-1} \to V_j$ we want to ensure that $P_j^{j-1} \equiv id_{V_{j-1}}$ holds. If we use $\mathbb{R}^{n_{j-1}}$ for V_{j-1}, we want to have $P_j^{j-1} = (R_{j-1}^j)^T$. For that we define $P_j^{j-1} : V_{j-1} \to V_j$ as

$$P_j^{j-1} \varphi_{j-1,i} = \sum_{k \in I_i^{j-1,j}} \varphi_{j,k} = \varphi_{j-1,i}.$$

In Figure 2.1 we have illustrated the set of nodes $\mathcal{N}^2, \mathcal{N}^1$ and $\mathcal{N}^0$ that describe the decrease of the system's dimension. For the one-dimensional case we have illustrate the effect of R_1^2 and $R_0^2 = R_0^1 R_1^2$ on a function v that is given as the sum of two basis functions in Figure 2.2.

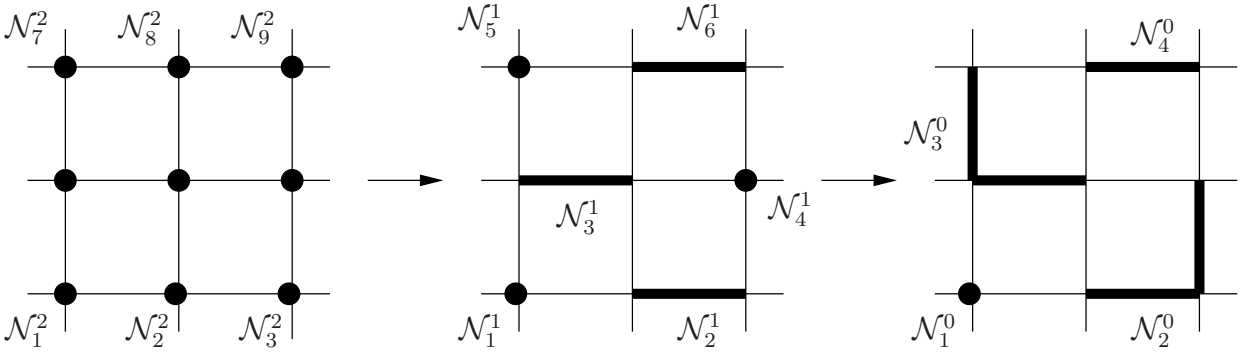


Figure 2.1: Coarsing of the grids

So far the restriction and prolongation operators are defined for function spaces V_j. The same way we will introduce the operators $S_j, \widehat{S}_j$. Hence we define the linear operator

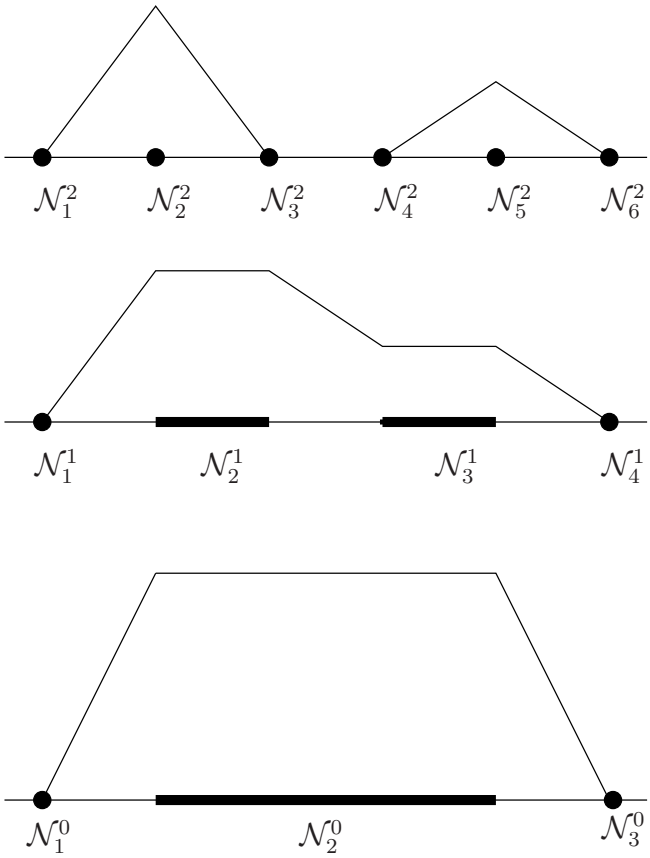


Figure 2.2: Effect of R_1^2 and R_0^1 on $v = \varphi_2^2 + \varphi_5^2$

$S_j : V_{j+1} \to V_{j+1}$ for $j = 0, \ldots, J-1$ by

$$S_j\, \varphi_{j+1,i} := \frac{1}{|I_k^{j,j+1}|} \varphi_{j+1,i}, \quad \text{with} \quad i \in I_k^{j,j+1}. \tag{2.10}$$

Analogously, we define the linear operator $\widehat{S}_j : V_j \to V_j$ for $j = 0, \ldots J-1$ by

$$\widehat{S}_j\, \varphi_{J,i} := \frac{1}{|I_k^{j,J}|} \varphi_{J,i}, \quad \text{with} \quad i \in I_k^j \tag{2.11}$$

$$\text{and} \quad \widehat{S}_J := Id. \tag{2.12}$$

The following lemma will show that one of the characteristics of these operators is that $S_j\, R_j^{j+1}$ and $\widehat{S}_j\, R_j$ respectively are the identity on certain subspaces of $\widetilde{V}_{j+1}$ and V respectively.

Lemma: 2.4.1. *For $j = 1, \ldots, J$*

1. *the operator S_j as defined in (2.10) fulfils*

$$S_j\, R_j^{j+1} \varphi_{j,i} = \varphi_{j,i} = R_j^{j+1}\, S_j \varphi_{j,i}, \quad \forall \varphi_{j,i} \in V_j.$$

2. *the operator $\widehat{S}_j$ as defind in (2.11) fulfils*

$$\widehat{S}_j\, R_j \varphi_{j,i} = \varphi_{j,i} = R_j\, \widehat{S}_j \varphi_{j,i}, \quad \forall \varphi_{j,i} \in V_j.$$

proof. 1. Let $\varphi_{j,i} \in V_j$ be an arbitrary base function with

$$\varphi_{j,i} = \sum_{k \in I_i^{j,j+1}} \varphi_{j+1,k}.$$

Then we have

$$S_j\, R_j^{j+1} \varphi_{j,i} = S_j\, R_j^{j+1} \sum_{k \in I_i^{j,j+1}} \varphi_{j+1,k} = S_j \sum_{k \in I_i^{j,j+1}} R_j^{j+1}\, \varphi_{j+1,k} = S_j \sum_{k \in I_i^{j,j+1}} \varphi_{j,i}$$

$$= |I_i^{j,j+1}| \cdot S_j\, \varphi_{j,i} = |I_i^{j,j+1}| \cdot \frac{1}{|I_i^{j,j+1}|} \varphi_{j,i} = \varphi_{j,i}$$

and

$$R_j^{j+1}\, S_j\, \varphi_{j,i} = R_j^{j+1}\, S_j \sum_{k \in I_i^{j,j+1}} \varphi_{j+1,k} = R_j^{j+1} \sum_{k \in I_i^{j,j+1}} S_j\, \varphi_{j+1,k}$$

$$= R_j^{j+1} \sum_{k \in I_i^{j,j+1}} \frac{1}{|I_i^{j,j+1}|}\, \varphi_{j+1,k} = \frac{1}{|I_i^{j,j+1}|} \sum_{k \in I_i^{j,j+1}} R_j^{j+1}\, \varphi_{j+1,k}$$

$$= \frac{1}{|I_i^{j,j+1}|}\, |I_i^{j,j+1}|\, \varphi_{j,i} = \varphi_{j,i}.$$

2. The proposition for $\widehat{S}_j$ follows by exactly the same arguments.

□

In Section 2.3.2 we have seen in Lemma 2.3.4 that by the definition of P_{j+1}^j as $P_{j+1}^j = (R_j^{j+1})^T$ and S_j as $(R_j^{j+1}\, P_{j+1}^j)^{-1}$ we have a representation of the orthogonal projection Q_j by $Q_j = P_{j+1}^j\, S_j\, R_j^{j+1}$. Analogously this also holds for $\widehat{Q}_j$. Now we will see that by the given definitions for the operators $P_{j+1}^j, R_j^{j+1}, S_j$ and $\widehat{S}_j$ and their characteristics, which we have shown above, we get the same representations for the orthogonal projections Q_j and $\widehat{Q}_j$ respectively.

Lemma: 2.4.2. *With the operators $S_j, \widehat{S}_j$ as defined in (2.10), (2.11), it holds for Q_j and $\widehat{Q}_j$ that*

$$Q_j = P_{j+1}^j\, S_j\, R_j^{j+1} \quad \textit{and} \quad \widehat{Q}_j = P_j\, \widehat{S}_j\, R_j.$$

proof. Because of the definition of $Q_j, \widehat{Q}_j$ as orthogonal projections with respect to the inner product $(.,.)$ and the uniqueness of these operators we have to prove that the operators $P_{j+1}^j\, S_j\, R_j^{j+1}, P_j\, \widehat{S}_j\, R_j$ are orthogonal projections too. This will only be shown for the operator Q_j because of the the proof follows by the same arguments for $\widehat{Q}_j$.
First we show that Q_j is a projector, i.e. $Q_j = (Q_j)^2$. For all $v_{j+1} \in V_{j+1}$ we obtain

$$P_{j+1}^j\, S_j\, R_j^{j+1}\, v_{j+1} \in V_j.$$

Further, we obtain

$$S_j\, R_j^{j+1} v_j = v_j \quad \forall v_j \in V_j$$

and P_{j+1}^j is the identity. So we get

$$P_{j+1}^j\, S_j\, R_j^{j+1} v_j = v_j \quad \forall v_j \in V_j.$$

Considering these conlusions, we have for all $v_{j+1} \in V_{j+1}$

$$(P_{j+1}^j\, S_j\, R_j^{j+1})\, \underbrace{P_{j+1}^j\, S_j\, R_j^{j+1}\, v_{j+1}}_{\in V_j} = P_{j+1}^j\, S_j\, R_j^{j+1}\, v_{j+1}.$$

This means that $(Q_j)^2 = Q_j$.
Further, we need to prove that for the operator Q_j, that for all $v_j \in V_j$ and all $v_{j+1} \in V_{j+1}$, we obtain

$$(Q_j\, v_{j+1},\, v_j) = (v_{j+1},\, v_j).$$

By the linearity of the operator, it is sufficient to prove this for the basis functions of the spaces. Let $\varphi_{j+1,i} \in V_{j+1}$ and $\varphi_{j,k} \in V_j$ be two arbitrary base functions. Then there exist unique $n, m \in \{1, \ldots, n_j\}$ with

$$\begin{aligned} P_{j+1}^j S_j R_j^{j+1} \varphi_{j+1,i} = P_{j+1}^j S_j \varphi_{j,n} = P_{j+1}^j \frac{1}{|I_n^{j,j+1}|} \varphi_{j,n} \\ = \frac{1}{|I_n^{j,j+1}|} \sum_{t \in I_n^{j,j+1}} \varphi_{j+1,t} \\ \text{and} \quad \varphi_{j,k} = \sum_{s \in I_m^{j,j+1}} \varphi_{j+1,s}. \end{aligned}$$

Now we remember that by the definition of the sets $I_n^{j,j+1}$, we obtain $n = m$ or $I_n^{j,j+1} \cap I_m^{j,j+1} = \emptyset$. Then $n = m$ is equivalent to $i \in I_m^{j,j+1}$. Hence we get

$$\begin{aligned} (Q_j \varphi_{j+1,i}, \varphi_{j,k}) &= \left(\sum_{t \in I_n^{j,j+1}} \varphi_{j+1,t}, \sum_{s \in I_m^{j,j+1}} \varphi_{j+1,s} \right) \\ &= \begin{cases} \sum_{t \in I_m^{j,j+1}} (\varphi_{j+1,t}, \varphi_{j+1,t}) & \text{if} \quad n = m \\ 0 & \text{else} \end{cases} \\ \text{and} \quad (\varphi_{j+1,i}, \varphi_{j,k}) &= \left(\varphi_{j+1,i}, \sum_{s \in I_m^{j,j+1}} \varphi_{j+1,s} \right) \\ &= \begin{cases} \sum_{t \in I_m^{j,j+1}} (\varphi_{j+1,t}, \varphi_{j+1,t}) & \text{if} \quad i \in I_m^{j,j+1} \\ 0 & \text{else} \end{cases} \end{aligned}$$

This completes the poof of Lemma 2.4.2 □

2.4.2 Matrix representations

In this section we will give matrix representations of the linear operators $R_{j-k}^j, P_j^{j-k}, S_j$ and $\widehat{S}_j$ used for the aggregation method so far. With the unit-vectors $e_t^j \in \mathbb{R}^{n_j}$, we define the matrix $R_{M,j-1}^j \in \mathbb{R}^{n_{j-1} \times n_j}$ by its rows $(R_{M,j-1}^j)_{i,.}$, $i = 1, \ldots, n_{j-1}$ with

$$(R_{M,j-1}^j)_{i,.} = \sum_{t \in I_i^{j-1,j}} (e_t^j)^T.$$

Furthermore, we define the following matrices for $j = 1, \ldots, J$ and $1 \leq k < j$:

$$R^j_{M,j-k} \in \mathbb{R}^{n_{j-k} \times n_j} \quad \text{by} \quad R^j_{M,j-k} := R^{j-k+1}_{M,j-k} \cdot \cdots \cdot R^j_{M,j-1}$$

$$P^j_{M,j+1} \in \mathbb{R}^{n_{j+1} \times n_j} \quad \text{by} \quad P^j_{M,j+1} = (R^{j+1}_{M,j})^T$$

$$P^{j-k}_{M,j} \in \mathbb{R}^{n_{j-k} \times n_j} \quad \text{by} \quad P^{j-k}_{M,j} := P^{j-1}_{M,j} \cdot \cdots \cdot P^{j-k}_{M,j-k+1}.$$

Lemma: 2.4.3. *By the definitions of R^j_{j-k}, P^{j-k}_j and $R^j_{M,j-k}, P^{j-k}_{M,j}$ and the definition of $e^j_i \in \mathbb{R}^{n_j}$ as a representation of $\varphi_{j,i} \in V_j \equiv \mathbb{R}^{n_j}$, we have for all $j = 1, \ldots, J$ and all $1 \leq k < j$ that*

1. *$R^j_{M,j-k}$ is a matrix representation of R^j_{j-k}.*
2. *$P^{j-k}_{M,j}$ is a matrix representation of P^{j-k}_j.*

proof. It is sufficient to prove both propositions for $k = 1$. The rest follows by the iterative definition of the operators for $k > 1$.

1. For an arbitrary $j \in 1, \ldots, J$ and an arbitrary base function $\varphi_{j,i} \in V_j$ with $\varphi_{j,i} \equiv e^j_i \in R^{n_j}$, there is a unique $k \in 1, \ldots, n_{j-1}$ with $i \in I^{j-1,j}_k$. By the definition of R^j_{j-1}, we obtain

$$R^{j-1}_j \, \varphi_{j,i} = \varphi_{j-1,k} \equiv e^{n_{j-1}}_k \in R^{n_{j-1}}.$$

   However, we have

$$\begin{aligned} R^j_{M,j-1} \, e^j_i &= \left((R^j_{M,j-1})_{1,.} \, e^j_i, \ldots, (R^j_{M,j-1})_{n_{j-1},.} \, e^j_i \right) \\ &= \left(\sum_{t \in I^{j-1,j}_1} (e^j_t)^T e^j_i, \ldots, \sum_{t \in I^{j-1,j}_{n_{j-1}}} (e^j_t)^T e^j_i \right) \\ &= e^{j-1}_k. \end{aligned}$$

   The last equation follows from the uniqueness of k with $i \in I^{j-1,j}_k$. Therefore the proof for R^j_{j-1} follows by the linearity of this operator.

2. Similary, for an arbitrary $j \in 1, \ldots, J$ and an arbitrary base function $\varphi_{j-1,i} \in V_{j-1}$ with $\varphi_{j-1,i} \equiv e^{j-1}_i \in R^{n_{j-1}}$ follows that there is a unique $k \in 1, \ldots n_j$ with $k \in I^{j-1,j}_i$. By the definition of P^{j-1}_j we now have

$$P^{j-1}_j \varphi_{j-1,i} = \sum_{t \in I^{j-1,j}_i} \varphi_{j,t} \equiv \sum_{t \in I^{j-1,j}_i} e^j_t \quad \text{with} \quad e^j_t \in \mathbb{R}^{n_j}.$$

Moreover we obtain $(R^j_{M,j-1})^T_{s,t} = 1$ if and only if $\mathcal{N}^j_s \subset \mathcal{N}^{j-1}_t$ holds. Hence it follows that

$$\begin{aligned} P^{j-1}_{M,j}\, e^{j-1}_i &= (R^j_{M,j-1})^T e^{j-1}_i \\ &= \left(((R^j_{M,j-1})_{.,1})^T e^{j-1}_i, \ldots, ((R^j_{M,j-1})_{.,n_j})^T e^{j-1}_i\right) \\ &= \left((e^{j-1}_{l(1)})^T e^{j-1}_i, \ldots, (e^{j-1}_{l(n_j)})^T e^{j-1}_i\right) \\ &= \sum_{t \in I^{j-1,j}_i} e^j_t \quad \text{with} \quad e^j_t \in \mathbb{R}^{n_j}. \end{aligned}$$

In this calculation, $l(x) \in \mathbb{N}$ is the index with $x \in I^{j-1,j}_{l(x)}$. Again the equivalence of the operators follows from the linearity.

As already mentioned iteration proves the proposition for $k > 1$. □

Since it is always obvious whether or not we use a matrix, we will write $P^{j-k}_j, R^j_{j-k}, \ldots$ in both cases and drop the denotation with the index M.

Lemma: 2.4.4. *By the definitions of R^j_{j-k}, P^{j-k}_j, for the matrix representations of the operators we have*

1. $(R^j_{j-k})_{i,.} = \sum_{t \in I^{j-k,j}_i} (e^j_t)^T$, *for all* $j = 1, \ldots, J$ *and* $1 \le k \le j$.
2. $R^j_{j-k}\, P^{j-k}_j = diag(|I^{j-k,j}_1|, \ldots, |I^{j-k,j}_{n_{j-k}}|)$, *for all* $j = 1, \ldots, J$ *and* $1 \le k \le j$.
3. $(S_{j-1})^{-1} = diag(|I^{j-1,j}_1|^{-1}, \ldots, |I^{j-1,j}_{n_{j-1}}|^{-1})$, *for all* $j = 1, \ldots, J$.
4. $(\widehat{S}_j)^{-1} = diag(|I^{j,J}_1|^{-1}, \ldots, |I^{j,J}_{n_j}|^{-1})$, *for all* $j = 0, \ldots, J$.

proof. 1. For an arbitrary $j = 1, \ldots, J$ and $k = 1$, the proposition holds by the definition of the matrix R^j_{j-1}. Assume that the proposition holds for a $k - 1 \ge 1$. Then we consider $R^j_{j-k} = R^{j-k+1}_{j-k}\, R^j_{j-k+1}$. By the assumption, we obtain for the i-th row of R^{j-k+1}_{j-k}

$$(R^{j-k+1}_{j-k})_{i,.} = \sum_{t \in I^{j-k,j-k+1}_i} e^{j-k+1}_t,$$

with

$$I^{j-k,j-k+1}_i = \{t \in \{1, \ldots, n_{j-k+1}\} : \mathcal{N}^{j-k+1}_t \subset \mathcal{N}^{j-k}_i\}.$$

Analogously, it follows for the z-th column of R^j_{j-k+1} that in the row s the entry is one if and only if it is

$$\mathcal{N}^j_z \subset \mathcal{N}^{j-k+1}_s$$

and zero otherwise. This implies for the element $(R^j_{j-k})_{i,s}$ that

$$\begin{aligned}(R^j_{j-k})_{i,s} &= (R^{j-k+1}_{j-k})_{i,\cdot}\,(R^j_{j-k+1})_{\cdot,s} \\ &= \begin{cases} 1 & \text{if} \quad \mathcal{N}^j_s \subset \mathcal{N}^{j-k+1}_t \quad \wedge \quad \mathcal{N}^{j-k+1}_t \subset \mathcal{N}^{j-k}_i \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

With $I^{j-k,j}_z = \{l \in \{1,\ldots,n_j\} : \mathcal{N}^j_l \subset \mathcal{N}^{j-k}_z\}$.

2. From the first result we obtain

$$\begin{aligned}(R^j_{j-k})_{s,\cdot} \cdot (P^{j-k}_j)_{\cdot,t} &= (R^j_{j-k})_{s,\cdot} \cdot ((R^j_{j-k})_{t,\cdot})^T \\ &= \left(\sum_{x \in I^{j-k,j}_s} (e^j_x)^T\right) \cdot \left(\sum_{y \in I^{j-k,j}_t} e^j_y\right) \qquad = \begin{cases} |I^{j-k,j}_s| & \text{if} \quad s = t \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

The zero is given as we have $I^{j-k,j}_s \cap I^{j-k,j}_t = \emptyset$ for $s \neq t$.

3. This proposition follows from the definition of $S_{j-1} := (R^j_{j-1}\,P^{j-1}_j)^{-1}$ and from the second proposition of this lemma.

4. This proposition follows from the definition of $\widehat{S}_j := (R_j\,P_j)^{-1}$ and also from the result of this lemma.

□

The Lemma 2.4.4 gives a matrix representation of the operators $S_j, \widehat{S}_j$. As for the operators $R^j_{j-1}, \ldots$, we use the same symbol for the operator and its matrix representation. Later on, we will need some characteristics of the cardinal number of the sets $I^{j-k,j}_i$. Therefore, we will take a look at this now.

Lemma: 2.4.5. *For all $j \leq J$ and $0 \leq k \leq j$ and all $i \in \{1,\ldots,n_{j-k}\}$,*

$$|I^{j-k,j}_i| = \sum_{l_{k-1} \in I^{j-k,j-k+1}_i} \left| \sum_{l_{k-2} \in I^{j-k+1,j-k+2}_{l_{k-1}}} \right| \cdots \sum_{l_1 \in I^{j-2,j-1}_{l_2}} |I^{j-1,j}_{l_1}| \Bigg| \ldots \Bigg|.$$

In particular, if for all $z \in \{j-k+1, \ldots, j\}$, *we have*

$$|I_t^{z-1,z}| = s_z, \quad \textit{for all } t \textit{ with } \quad \mathcal{N}_t^{z-1} \subset \mathcal{N}_i^z,$$

we also have

$$|I_i^{j-k,j}| = s_j \cdot \cdots \cdot s_{j-k+1}.$$

proof. By the definition of the set $I_i^{j-k,k}$ we obtain

$$\begin{aligned}
I_i^{j-k,j} &:= \left\{ l \in \{1, \ldots, n_j\} \,:\, \mathcal{N}_l^j \subset \mathcal{N}_i^{j-k} \right\} \\
&= \left\{ l \in \{1, \ldots, n_j\} \,:\, \mathcal{N}_l^j \subset \mathcal{N}_{l_1}^{j-1} \wedge \mathcal{N}_{l_1}^{j-1} \subset \mathcal{N}_i^{j-k} \right\} \\
&\vdots \\
&= \Big\{ l \in \{1, \ldots, n_j\} \,:\, \mathcal{N}_l^j \subset \mathcal{N}_{l_1}^{j-1} \wedge \mathcal{N}_{l_1}^{j-1} \subset \mathcal{N}_{l_2}^{j-2} \wedge \ldots \\
&\qquad\qquad \ldots \wedge \mathcal{N}_{l_{k-1}}^{j-k+1} \subset \mathcal{N}_i^{j-k} \Big\}.
\end{aligned}$$

Hence it follows for the cardinal number of the set $I_i^{j-k,j}$ that

$$\begin{aligned}
|I_i^{j-k,j}| &= \sum_{l_{k-1} \in I_i^{j-k,j-k+1}} |I_{l_{k-1}}^{j-k+1,j}| \\
&= \sum_{l_{k-1} \in I_i^{j-k,j-k+1}} \left| \sum_{l_{k-2} \in I_{l_{k-1}}^{j-k+1,j-k+2}} |I_{l_{k-2}}^{j-k+2,j}| \right| \\
&\vdots \\
&= \sum_{l_{k-1} \in I_i^{j-k,j-k+1}} \left| \sum_{l_{k-2} \in I_{l_{k-1}}^{j-k+1,j-k+2}} \left| \cdots \sum_{l_2 \subset I_{l_1}^{j-2,j-1}} |I_{l_1}^{j-1,j}| \right| \ldots \right|. \qquad (2.13)
\end{aligned}$$

This completes the proof of this lemma. If we additionally have

$$|I_t^{z-1,z}| = s_z \quad \text{for all } t \text{ with } \mathcal{N}_t^{z-1} \subset \mathcal{N}_i^z,$$

it follows for all sets $I_t^{z-1,z}$ which are used in (2.13) that their cardinal number only depends on the index z. Thus, they are given by $I_t^{z-1,z} = s_z$. This implies that

$$|I_i^{j-k,j}| = s_j \cdot \cdots \cdot s_{j-k+1}.$$

□

2.4.3 The condition $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$:

In Lemma 2.3.8, we have assumed that $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ and as a result we have arrived at a representation of the orthogonal projection $\widehat{Q}_{j-1}$ by $\widehat{Q}_{j-1} = P_j\, \widehat{S}_j\, Q_{j-1}\, R_j$. Furthermore, we have seen that by an iterative use of this condition we can drop the operators $\widehat{S}_j$. If we consider the assumption in the function space, P_j^{j-1} is the identity. Hence, the assumption can be interpreted as $\widehat{S}_j \circ S_{j-1} = \widehat{S}_{j-1}$. By the matrix representations used in the previous section it is obvious from the dimensions of the matrices, that this term is not well-posed. That is why we consider the equation as given in Lemma 2.3.8.

It will be the primary aim of this section to show an equivalent characterisation for the assumption $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$, that will depend on the structure of the sets $I_i^{j-1,j}$. Further, we will show additional characteristics that hold by this assumption.

We define the following condition for the restriction operators:
If we have $\mathcal{N}_x^J, \mathcal{N}_y^J \subset \mathcal{N}_k^j$ for an $k \in \{1, \dots, n_j\}$ and

$$\mathcal{N}_x^J \subset \mathcal{N}_{i_{J-1}(x)}^{J-1}, \dots, \mathcal{N}_x^J \subset \mathcal{N}_{i_{j+1}(x)}^{j+1}$$

$$\text{as well as} \quad \mathcal{N}_y^J \subset \mathcal{N}_{i_{J-1}(y)}^{J-1}, \dots, \mathcal{N}_y^J \subset \mathcal{N}_{i_{j+1}(y)}^{j+1}$$

then it follows that

$$|I_{i_k(x)}^{k,k+1}| = |I_{i_k(y)}^{k,k+1}|, \quad \text{for all} \quad k = j, \dots, J-1. \tag{2.14}$$

In short, we denote this with condition (2.14).

The condition means that two grid points (or rows of A) $\mathcal{N}_x^J, \mathcal{N}_y^J$, that are at least in the $(J-j)$-th step aggregated, have in all previous aggregation steps the same number of grid points $\mathcal{N}^J$ (or rows of A) that are aggregated with them to one new grid point $\mathcal{N}^k$ (or new row of A_k) for $k \geq j$.

Now we show some technical lemmata that are coherent with this condition. The first one will give us an easy sufficient condition for the condition (2.14). The other lemmata will show characteristics of the condition.

Lemma: 2.4.6. *If we assume for all $j = 0, \dots, J$ that $S_j \in \mathbb{R}^{n_j \times n_j}$ fulfils*

$$S_j = s_j I_j \quad \textit{with} \quad s_j \in \mathbb{R},$$

then the condition (2.14) is fulfiled.

proof. Based on the assumption that

$$S_j = s_j I_j \quad \text{with} \quad s_j \in \mathbb{R}$$

for all $j = 0, \ldots, J-1$, we obtain for all $j = 0, \ldots, J-1$

$$|I_i^{j,j+1}| = |I_t^{j,j+1}| = s_j \quad \text{for all} \quad i, t \in \{1, \ldots, n_j\}.$$

□

The situation given by the assumption of Lemma 2.4.6 is that in a single aggregation step the number of aggregated points is always the same. Of course, this also holds for the aggregation of two arbitrary points $\mathcal{N}_i^j, \mathcal{N}_k^j$: In all previous steps, they are aggregated with the equal number of points to a new grid point.

Lemma: 2.4.7. *Assume that $\mathcal{N}_x^J, \mathcal{N}_y^J \subset \mathcal{N}_t^j$. Then the following three statements are equivalent:*

1. *condition (2.14) holds for $\mathcal{N}_x^J, \mathcal{N}_y^J \subset \mathcal{N}_t^j$.*

2. *the equation $|I_{i_j(a)}^{k,k+1}| = |I_{i_j(x)}^{k,k+1}|$ holds for all $\mathcal{N}_a^J \subset \mathcal{N}_t^j$ with $\mathcal{N}_a^J \subset \mathcal{N}_{i_{J-1}(a)}^{J-1}, \ldots, \mathcal{N}_a^J \subset \mathcal{N}_{i_j(a)}^j$ and all $k = j, \ldots, J-1$.*

3. *the equation $|I_{i(x)}^{p,q}| = |I_{i(y)}^{p,q}|$ holds for all $j \leq p \leq q \leq J$ for $|I_{i(x)}^{p,q}|$ and $|I_{i(y)}^{p,q}|$ with $\mathcal{N}_{i(x)}^p \subset \mathcal{N}_x^j, \mathcal{N}_{i(y)}^p \subset \mathcal{N}_y^j$.*

proof. We prove the proposition by three implications:

$1 \Rightarrow 2$: If condition (2.14) holds, we get for an arbitrary $\mathcal{N}_a^J \subset \mathcal{N}_t^{j-1}$ and $\mathcal{N}_x^J$ that $\mathcal{N}_x^J, \mathcal{N}_a^J$ are at least aggregated in step $j-1$. So the assumption of condition (2.14) is fulfiled. Hence it follows for all $k = j, \ldots, J-1$ that

$$|I_{i_j(a)}^{k,k+1}| = |I_{i_j(x)}^{k,k+1}|.$$

$2 \Rightarrow 3$: By Lemma 2.4.5 the cardinal numbers of $I^{p,q}_{i(x)}, I^{p,q}_{i(y)}$ are given by

$$|I^{p,q}_{i(x)}| = \sum_{l_p \in I^{p,p+1}_{i(x)}} \left| \sum_{l_{p+1} \in I^{p+1,p+2}_{l_p}} \right| \cdots \sum_{l_{q-1} \in I^{q-2,q-1}_{q-2}} |I^{q-1,q}_{q-1}| \Bigg| \cdots \Bigg|$$

$$|I^{p,q}_{i(y)}| = \sum_{l_p \in I^{p,p+1}_{i(y)}} \left| \sum_{l_{p+1} \in I^{p+1,p+2}_{l_p}} \right| \cdots \sum_{l_{q-1} \in I^{q-2,q-1}_{q-2}} |I^{q-1,q}_{q-1}| \Bigg| \cdots \Bigg|$$

As all used sets $I^{r,r+1}_s$ represent a node $\mathcal{N}^r_s$ that is aggregated to $\mathcal{N}^{j-1}_t$, all sets have the same cardinal number by the assumption of the second characteristic. Therefore, both sums have the same value. This proves the proposition.

$3 \Rightarrow 1$: This is obvious if we set $k = 1$.

□

On the next lemma, we will see in particular that the equation $\widehat{S}_j \, P^{j-1}_j \, S_{j-1} = P^{j-1}_j \, \widehat{S}_{j-1}$ holds if and only if condition (2.14) is fulfilled. We can see this as the central characteristic of this condition.

Lemma: 2.4.8. *The condition (2.14) is equivalent to the following three statements:*

1. *for all* $t \in \{1, \ldots, n_{j-1}\}$ $(\widehat{S}_j)_{k,k}$ *is the same number for all* $k \in I^{j-1,j}_t$.
2. $\widehat{S}_j \, Q_{j-1} = Q_{j-1} \, \widehat{S}_j$
3. $\widehat{S}_j \, P^{j-1}_j \, S_{j-1} = P^{j-1}_j \, \widehat{S}_{j-1}$ *holds.*

proof. 1. As it is $(\widehat{S}_j)_{k,k} = \frac{1}{|I^{j,J}_k|}$ we obtain from Lemma 2.4.7 that this is equal for all $k \in I^{j-1,j}_t$, if and only if condition (2.14) holds.

2. First we assume that condition (2.14) holds. Let $e^j_i \in \mathbb{R}^{n_j}$ be a unit-vector. Then it follows, that $R^j_{j-1} e^j_i = e^{j-1}_t \in \mathbb{R}^{n_{j-1}}$ and that

$$P^{j-1}_j \, S_{j-1} e^{j-1}_t = P^{j-1}_j \, \frac{1}{|I^{j-1,j}_t|} \, e^{j-1}_t = \frac{1}{|I^{j-1,j}_t|} \sum_{k \in I^{j-1,j}_t} e^j_k.$$

Assuming that condition (2.14) holds, it follows for all $k \in I_t^{j-1,j}$ that $(\widehat{S}_j)_{k,k}$ is the same number s. This implies that

$$\widehat{S}_j\, Q_{j-1}\, e_i^j = s\, \frac{1}{|I_t^{j-1,j}|} \sum_{k \in I_t^{j-1,j}} e_k^j$$

$$\text{and} \quad Q_{j-1}\, \widehat{S}_j\, e_i^j = s\, Q_{j-1}\, e_i^j = s\, \frac{1}{|I_t^{j-1,j}|} \sum_{k \in I_t^{j-1,j}} e_k^j.$$

Assume now that condition (2.14) does not hold. Then there are two points $\mathcal{N}_x^j, \mathcal{N}_y^j \subset \mathcal{N}_t^{j-1}$ with $\mathcal{N}_x^j \subset \mathcal{N}_x^{j,J}$, $\mathcal{N}_y^j \subset \mathcal{N}_y^{j,J}$ and

$$(\widehat{S}_j)_{x,x}^{-1} = |I_x^{j,J}| = n_x \neq n_y = |I_y^{j,J}| = (\widehat{S}_j)_{y,y}^{-1}.$$

It follows that

$$\begin{aligned} Q_{j-1}\, \widehat{S}_j\, (e_x^j + e_y^j) &= Q_{j-1} \left(\frac{1}{n_x} e_x^j + \frac{1}{n_y} e_y^j \right) = P_j^{j-1}\, S_{j-1} \left(\frac{1}{n_x} + \frac{1}{n_y} \right) e_t^{j-1} \\ &= P_j^{j-1}\, \frac{1}{|I_t^{j-1,j}|} \left(\frac{1}{n_x} + \frac{1}{n_y} \right) e_t^{j-1} \\ &= \frac{1}{|I_t^{j-1,j}|} \left(\frac{1}{n_x} + \frac{1}{n_y} \right) \sum_{i \in I_t^{j-1,j}} e_i^j \quad \in P_j^{j-1}(\widetilde{V}_{j-1}) \end{aligned}$$

$$\begin{aligned} \text{and} \quad \widehat{S}_j\, Q_{j-1}\, (e_x^j + e_y^j) &= \widehat{S}_j\, \frac{2}{|I_t^{j-1,j}|} \sum_{i \in I_t^{j-1,j}} e_i^j \\ &= \frac{2}{|I_t^{j-1,j}|} \sum_{i \in I_t^{j-1,j}} \frac{1}{|I_i^{j,J}|} e_i^j \quad \notin P_j^{j-1}(\widetilde{V}_{j-1}). \end{aligned}$$

So it is obvious that the elements can not be the same.

3. As shown in the proof of Lemma 2.4.5, we can generally say that

$$|I_i^{j-1,J}| = \sum_{t \in I_i^{j-1,j}} |I_t^{j,J}|. \tag{2.15}$$

If and only if condition (2.14) holds, it follows that

$$\sum_{t \in I_i^{j-1,j}} |I_t^{j,J}| = |I_i^{j-1,j}| \cdot |I_t^{j,J}|, \quad \text{with } t \in I_i^{j-1,j}. \tag{2.16}$$

First we assume that condition (2.14) holds. Then we obtain for an arbitrary unit-vector $e_i^{j-1} \in \mathbb{R}^{n_{j-1}}$ that

$$\widehat{S}_j\, P_j^{j-1}\, S_{j-1}\, e_i^{j-1} = \frac{1}{|I_i^{j-1,j}|}\, \widehat{S}_j\, P_j^{j-1}\, e_i^{j-1} = \frac{1}{|I_i^{j-1,j}|}\, \widehat{S}_j \sum_{t \in I_i^{j-1,j}} e_t^j$$

$$= \frac{1}{|I_i^{j-1,j}|} \sum_{t \in I_i^{j-1,j}} \frac{1}{|I_t^{j,J}|}\, e_t^j.$$

According to condition (2.14), $\frac{1}{|I_t^{j,J}|}$ is constant over $t \in I_i^{j-1,j}$ for given i, j. With equation (2.16), this implies for an arbitrary $t \in I_i^{j-1,j}$ the equation

$$\frac{1}{|I_i^{j-1,J}|} = \frac{1}{|I_i^{j-1,j}|}\, \frac{1}{|I_t^{j,J}|} \quad \Rightarrow \quad \widehat{S}_j\, P_j^{j-1}\, S_{j-1}\, e_i^{j-1} = \frac{1}{|I_i^{j-1,J}|} \sum_{t \in I_i^{j-1,j}} e_t^j.$$

On the other side we have

$$P_j^{j-1}\, \widehat{S}_{j-1}\, e_i^{j-1} = \frac{1}{|I_i^{j-1,J}|}\, P_j^{j-1}\, e_i^{j-1} = \frac{1}{|I_i^{j-1,J}|} \sum_{t \in I_i^{j-1,j}} e_t^j.$$

This proves the proposition. Now we assume that condition (2.14) does not hold. Then there is a $t \in \{1, \ldots, n_{j-1}\}$ and $\mathcal{N}_x^j, \mathcal{N}_y^j$ with $x, y \in I_t^{j-1,j}$. By the first proposition of this lemma we can assume that $(\widehat{S}_j)_{x,x}^{-1} = |I_x^{j,J}| \neq |I_y^{j,J}| = (\widehat{S}_j)_{y,y}^{-1}$. Hence it follows that

$$\widehat{S}_j\, P_j^{j-1}\, S_{j-1}\, (e_t^{j-1}) = \widehat{S}_j\, P_j^{j-1}\, \frac{1}{|I_t^{j-1,j}|} e_t^{j-1} = \frac{1}{|I_t^{j-1,j}|}\, \widehat{S}_j \sum_{i \in I_t^{j-1,j}} e_i^j$$

$$= \frac{1}{|I_t^{j-1,j}|} \sum_{i \in I_t^{j-1,j}} \frac{1}{|I_i^{j,J}|}\, e_i^j$$

$$\text{and} \quad P_j^{j-1}\, \widehat{S}_{j-1}\, e_t^{j-1} = P_j^{j-1}\, \frac{1}{|I_t^{j-1,J}|}\, e_t^{j-1} = \frac{1}{|I_t^{j-1,J}|} \sum_{i \in I_t^{j-1,j}} e_i^j.$$

As in this case $|I_t^{j-1,j}| \cdot |I_x^{j,J}| \neq |I_t^{j-1,J}|$, the proposition obviously follows if we multiply both expressions with $(e_x^j)^T$ ($(e_y^j)^T$). It follows that

$$(e_x^j)^T\, \widehat{S}_j\, P_j^{j-1}\, S_{j-1}\, (e_t^{j-1}) = \frac{1}{|I_t^{j-1,j}|}\, \frac{1}{|I_x^{j,J}|}\, (e_x^j)^T\, e_i^j = \frac{1}{|I_t^{j-1,j}|}\, \frac{1}{|I_x^{j,J}|}$$

$$\text{and} \quad (e_x^j)^T\, P_j^{j-1}\, \widehat{S}_{j-1}\, e_t^{j-1} = \frac{1}{|I_t^{j-1,J}|} (e_x^j)^T\, e_i^j = \frac{1}{|I_t^{j-1,J}|}.$$

□

At the least, we get a result for the kernel of the operator R_{j-1}^{j}, which we will use as a condition in estimations on the condition and proofs of the non singularity of operators.

Lemma: 2.4.9. *Assume that condition (2.14) holds. Then* $ker(R_{j-1}^{j}) = ker(R_{j-1}^{j} R_j P_j)$ *holds as well.*

proof. Consider an arbitrary $\widetilde{v}_j \in \widetilde{V}_j$ with $\widetilde{v}_j \in ker(R_{j-1}^{j})$. By Lemma 2.3.4, this is equivalent to $\widetilde{v}_j \in ker(Q_{j-1})$. From the second proposition of Lemma 2.4.8 we obtain

$$\begin{aligned} 0 = Q_{j-1}\widetilde{v}_j = Q_{j-1}\,\widehat{S}_j\,R_j\,P_j\,\widetilde{v}_j = \widehat{S}_j\,Q_{j-1}\,R_j\,P_j\,\widetilde{v}_j \\ = \widehat{S}_j\,P_j^{j-1}\,S_{j-1}\,R_{j-1}^{j}\,R_j\,P_j\,\widetilde{v}_j. \end{aligned}$$

As we have $\widehat{S}_j\,P_j^{j-1}\,S_{j-1} \in \mathbb{R}^{n_j \times n_{j-1}}$ with $rk(\widehat{S}_j\,P_j^{j-1}\,S_{j-1}) = n_{j-1}$ this is equivalent to $R_j^{j-1}\,R_j\,P_j\,\widetilde{v}_j = 0$. This completes the proof. □

2.4.4 The black box method

As we also want to treat the preconditioning operators as a black box preconditioner for linear equation systems that has nothing to do with partial differential equations, we will introduce them accordingly. So we set

$$V := V_J \equiv \mathbb{R}^{n_J} \quad \text{and} \quad \widetilde{V}_j \equiv \mathbb{R}^{n_j}, \quad \text{for} \quad j = 0, \ldots, J-1.$$

Then we choose arbitrary matrices $P_j^{j-1} \in \mathbb{R}^{n_j \times n_{j-1}}$ with $rk(P_j^{j-1}) = n_{j-1}$ for $j = 0, \ldots, J-1$. We define R_{j-1}^{j} by $R_{j-1}^{j} = (P_j^{j-1})^T$ and define the matrices $R_{j-k}^{j}, P_j^{j-k}, P_j, R_j$ for $j = 1, \ldots, J$ and $1 \le k \le j$ as done in section 2.3.1.

Further, we define the spaces V_j for $j = 0, \ldots, J-1$ by

$$V_j := \left\langle P_j\,e_1^j, \ldots, P_j\,e_{n_j}^j \right\rangle$$

with the unit-vectors $e_1^j, \ldots, e_{n_j}^j \in \mathbb{R}^{n_j} \equiv \widetilde{V}_j$. Similar to section 2.3.1 we define the operators

$$\begin{aligned} S_{i-1} = (R_{i-1}^{i}\,P_i^{i-1})^{-1}, \quad \widehat{S}_i = (R_i\,P_i)^{-1} \\ Q_{i-1} = P_i^{i-1}\,S_{i-1}\,R_{i-1}^{i} \quad \text{and} \quad \widehat{Q}_i = P_i\,\widehat{S}_i\,R_i. \end{aligned}$$

So they are just defined by the matrices. As shown in this section, this setting is sufficient to obtain that $Q_{i-1} : \mathbb{R}^{n_i} \to P_i^{i-1}(\mathbb{R}^{n_{i-1}})$ and $\widehat{Q}_i : \mathbb{R}^{n} \to P_i(\mathbb{R}^{n_i})$ are the

orthogonal projections with respect to the inner product $(.,.)$. Analogously, we can define the spaces $W_j^t, \widetilde{W}_j$ and W_j by

$$W_j^t := \left\langle (\widehat{Q}_j - \widehat{Q}_{j-1})\varphi_{J,1}, \ldots, (\widehat{Q}_j - \widehat{Q}_{j-1})\varphi_{J,n} \right\rangle$$

$$\widetilde{W}_j := \langle (I_j - Q_{j-1})R_j\, \varphi_{J,1}, \ldots, (I_j - Q_{j-1})R_j\, \varphi_{J,n} \rangle$$

$$W_j := \langle P_j(I_j - Q_{j-1})R_j\, \varphi_{J,1}, \ldots, P_j(I_j - Q_{j-1})R_j\, \varphi_{J,n} \rangle\,.$$

This leads to the same setting as the discussion of the finite elements. But here we just choose arbitrary prolongations P_{j+1}^j. They give the hole structure of the subspaces.

We get the structure of the aggregation method if we set $P_j^{j-1} := (R_{j-1}^j)^T$. We define R_{j-1}^j by its rows $(R_{j-1}^j)_{i,.}$, if we set

$$(R_{j-1}^j)_{i,.} = \sum_{t \in I_i^{j-1,j}} (e_t^j)^T$$

with sets $I_k^{j-1,j}$ that fulfil

$$\bigcup_{k=1}^{n_{j-1}} I_k^{j-1,j} = \{1, \ldots, n_j\} \quad \text{and} \quad I_i \cap I_j = \emptyset \quad \text{for} \quad i \neq j.$$

In conjunction with the iterative definition of A_j as $A_j := R_j^{j+1}\, A_{j+1}\, P_{j+1}^j$, we can interpret $\mathcal{N}_i^j$ as the i-th row or column of A_j. $I_k^{j,j+1}$ is the set of rows (or colums) that will be added in A_{j+1} to get the k-th row (or column) of A_j. Hence, $|I_k^{j,j+1}|$ is the number of row or columns that are added.

2.5 The standard geometrical method (no aggregation)

For numerical results we will also use the standard geometrical method. Since we will make some modifications to this method, we will give a short introduction to it. As this method is mainly defined for grids with characteristic step widths, we introduce it accordingly. Further we introduce the method for the two grid case. And as usual, the multigrid situation follows if we use the setting iteratively.
Let $V = V_h \supset V_H$ be spaces of bilinear functions over the grids $\mathcal{T}_h, \mathcal{T}_H$. Assume that they have the grid points $\mathcal{N}_i^h, \mathcal{N}_j^H$ with $i = 1, \ldots, n_h$ and $j = 1, \ldots, n_H$
Assume that the situation is given as shown in Figure 2.3. Then the prolongation $P_h^H : V_H \to V_h$ is given as follows. We distinguish three types of grid points.

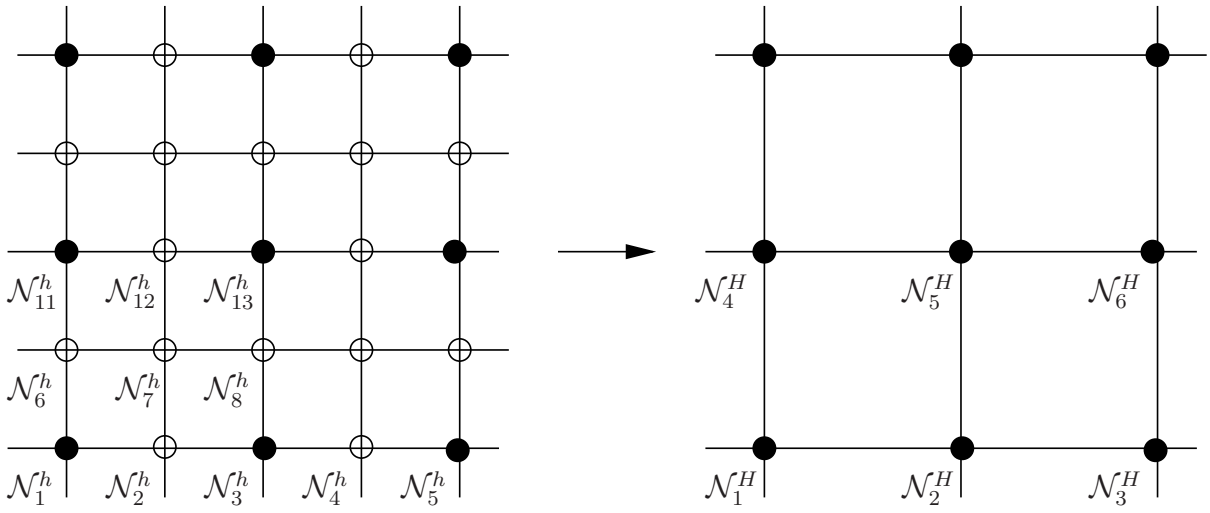


Figure 2.3: Geometrical coarsing of the grid

1. For points $\mathcal{N}_i^h = \mathcal{N}_j^H$ that also belong to the coarser grid (as for example $\mathcal{N}_1^h$), we define

$$(P_h^H v_H)(\mathcal{N}_i^h) = v_H(\mathcal{N}_j^H).$$

2. For points $\mathcal{N}_i^h$ whose right and left neighbours $\mathcal{N}_{i-1}^h = \mathcal{N}_j^H, \mathcal{N}_{i+1}^h = \mathcal{N}_{j+1}^H$ belong to the coarser grid (as for example $\mathcal{N}_2^h$), we define

$$(P_h^H v_H)(\mathcal{N}_i^h) = (v_H(\mathcal{N}_j^H) + v_H(\mathcal{N}_{j+1}^H))/2.$$

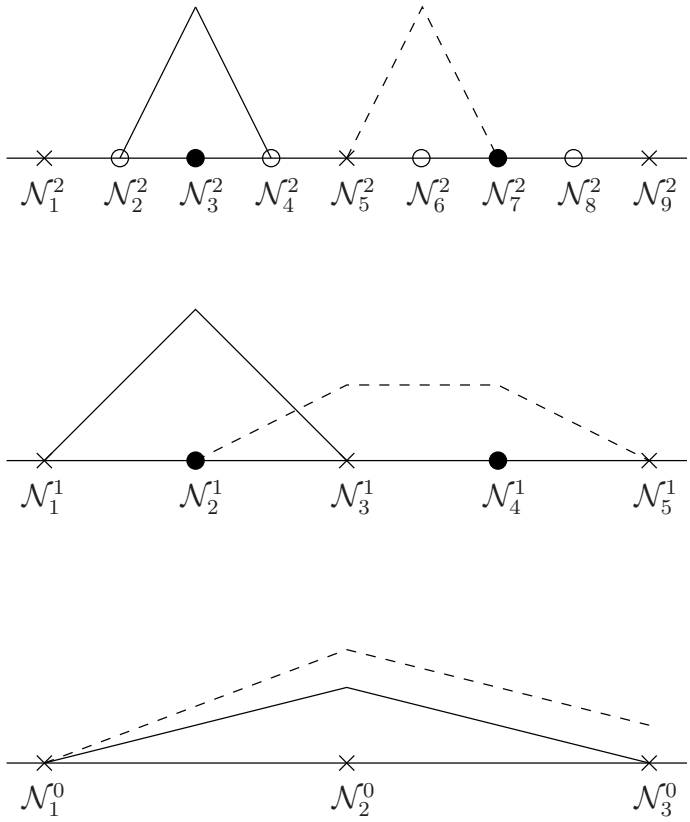
   The same definition holds for points $\mathcal{N}_i^h$ whose upper and lower neighbours are also points of the coarser grid (as for example $\mathcal{N}_6^h$).

3. For points $\mathcal{N}_i^h$ whose right and left neighbours and upper and lower neighbours are no grid points of the coarser grid (as for example $\mathcal{N}_7^h$), we define

$$(P_h^H v_H)(\mathcal{N}_i^h) = \frac{(v_H(\mathcal{N}_j^H) + v_H(\mathcal{N}_{j+1}^H) + v_H(\mathcal{N}_k^H) + v_H(\mathcal{N}_{k+1}^H)}{4}.$$

   Hence, $\mathcal{N}_j^H, \mathcal{N}_{j+1}^H$ are the lower left and lower right neighbours of $\mathcal{N}_i^h$ and $\mathcal{N}_k^H, \mathcal{N}_{k+1}^H$ are the upper left and upper right neighbours.

This defines the prolongation P_H^h and we define the restriction R_H^h by $R_H^h = (P_h^H)^T$. For the one-dimensional situation, this defines basis functions as shown in Figure 2.4 on page 53 for the three grids $\mathcal{T}_2, \mathcal{T}_1$ and $\mathcal{T}_0$.



——— φ_3, $R_1^2\,\varphi_3$ and $R_0^2\,\varphi_3$ respectively.

- - - - φ_6, $R_1^2\,\varphi_6$ and $R_0^2\,\varphi_6$ respectively.

Figure 2.4: Restriction of two Basis functions

2.6 Decompositions and representations

At last, we will give some decompositions and representation of elements $v \in \mathbb{R}^n$ and the inner products $(v,\, v)$ respectively, that we want to use. For an arbitrary $v \in V$, we have

(2.17)
$$v = \sum_{j=1}^{J} (\widehat{Q}_j - \widehat{Q}_{j-1})\, v + \widehat{Q}_0\, v \quad = \sum_{j=1}^{J} P_j\, \widehat{S}_j (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j}) R_j\, v + P_0\, \widehat{S}_0\, R_0\, v.$$

For this representation, no assumption on $\widehat{Q}_i$ for $i < J$ is needed. The only assumption on $\widehat{Q}_i$ is given by $\widehat{Q}_J = I$. If we additionally have $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$, as for example given for the aggregation method by condition (2.14), this implies that

(2.18)
$$v = \sum_{j=1}^{J} P_j\, \widehat{S}_j (I_j - Q_{j-1}) R_j\, v + P_0\, \widehat{S}_0\, R_0\, v.$$

So if we consider the two representations of v as given by (2.17) and (2.18), we get a first idea of the meaning of condition (2.14).

By the calculations of section 2.3.2 we obtain for an arbitrary $v \in V$

$$\begin{aligned}
(v,\, v) &= \left(\sum_{j=1}^{J} (\widehat{Q}_j - \widehat{Q}_{j-1}) v + \widehat{Q}_0 v,\, \sum_{j=1}^{J} (\widehat{Q}_j - \widehat{Q}_{j-1}) v + \widehat{Q}_0 v \right) \\
&= \sum_{j=1}^{J} \Big((\widehat{Q}_j - \widehat{Q}_{j-1}) v,\, (\widehat{Q}_j - \widehat{Q}_{j-1}) v \Big) + (\widehat{Q}_0\, v,\, \widehat{Q}_0\, v) \\
&= \sum_{j=1}^{J} \Big(v,\, (\widehat{Q}_j - \widehat{Q}_{j-1}) v \Big) + (v,\, \widehat{Q}_0\, v) \\
&= \sum_{j=1}^{J} (v,\, P_j\, \widehat{S}_j\, (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j})\, R_j\, v) + (\widehat{S}_0\, R_0\, v,\, R_0\, v). \qquad (2.19)
\end{aligned}$$

If additionally $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ holds this can be represented as

(2.20)
$$(v,\, v) = \sum_{j=1}^{J} (v,\, P_j\, \widehat{S}_j\, (I_j - Q_{j-1})\, R_j\, v) + (\widehat{S}_0\, R_0\, v,\, R_0\, v).$$

Further, we will use the following representation and estimation in the context of the BPX method:

$$\left(\sum_{j=0}^{J} \widehat{Q}_j v,\, v\right) = \sum_{j=0}^{J} (\widehat{Q}_j\, v,\, v) = \sum_{j=0}^{J} (\widehat{S}_i\, R_i\, v,\, R_i\, v).$$

As $\widehat{Q}_j$ is the orthogonal projection with respect to the dot product $(.,.)$, it follows for all $\widehat{Q}_j$ that

$$(v,\, v) = (\widehat{Q}_j\, v,\, v) + ((I - \widehat{Q}_j)\, v,\, v) \geq (\widehat{Q}_j\, v,\, v).$$

This obviously implies for all $v \in V$ that

$$(v,\, v) \leq \sum_{j=0}^{J} (\widehat{Q}_j\, v,\, v) \leq (J+1)(v,\, v).$$

3 Introduction of the preconditioners

In this chapter we will introduce three different preconditioners to solve the linear system of equations

$$Au = f$$

for a non singular $A \in \mathbb{R}^{n\times n}$ and $f \in \mathbb{R}^n = V$. As we only consider the linear system of equations we set $V = \mathbb{R}^n$.
All preconditioners are additive methods. We will introduce the preconditioners as two grid methods by using the vector spaces $V, V_0 \subset V$ and $W = V_0^{\perp}$. Our main interest is the characteristic of the operators which follows from the subspaces that are used. So we do not care about the quality of the solution on the different spaces. Hence we use the exact inverse of the operators A and A_0.

If we use the preconditioners in the context of partial differential equations and grids, the methods are defined by using two grids. But in general we will introduce them as a kind of black box method, that means the subspace V_0 is defined by $V_0 = Im(P_0(\widetilde{V}_0))$, with $\widetilde{V}_0 = \mathbb{R}^{n_0}$ (cf. section 2.4.4).

3.1 Common Setting

As we will introduce the preconditioners as two-grid methods, we can drop some of the indices we have used in Chapter 2 to simplify the notation. We set

$$P = P_1^0 = P_0, \quad R = R_0^1 = R_0$$

$$S = S_0 = \widehat{S}_0 \quad \text{and} \quad Q_0 = \widehat{Q}_0.$$

The equations $S_{J-1} = \widehat{S}_{J-1}$ and $Q_{J-1} = \widehat{Q}_{J-1}$ hold independently of J.
Furthermore we remember that $A_0 \in \mathbb{R}^{n_0\times n_0}$ is defined as

$$A_0 = R\,A\,P$$

and we still assume $P = R^T$.
Furthermore, we define two common constants c_1, d_1, to use for some estimations. We define c_1 by

$$c_1 := \max_{v_0 \in V_0 \setminus \{0\}} \frac{\|A\, v_0\|^2}{\|Q_0\, A\, v_0\|^2} \tag{3.1}$$

and d_1 by

$$d_1 := \min_{v_0 \in V_0 \setminus \{0\}} \frac{\|A\, v_0\|^2}{\|Q_0\, A\, v_0\|^2}. \tag{3.2}$$

The constants c_1, d_1 hence depend on the structure of the matrix A and the structure of the subspace V_0. The last dependency is more obvious if we transform the equation (3.1) (and (3.2) respectively) into

$$c_1 = \max_{\widetilde{v}_0 \in \widetilde{V}_0 \setminus \{0\}} \frac{\|A\, P\, \widetilde{v}_0\|^2}{\|Q_0\, A\, P\, \widetilde{v}_0\|^2}.$$

As Q_0 is the orthogonal projection concerning the dot product $(.,.)$, we have

$$\|A\, v_0\|^2 \geq \|Q_0\, A\, v_0\|^2$$

for all $v_0 \in V_0$. Consequently, one is a lower bound for d_1. For c_1, it is not as easy to get an upper bound. For further estimations, we want to show the following simple but useful results of these constants.

Lemma: 3.1.1. *Let $A \in \mathbb{R}^{n \times n}, P \in \mathbb{R}^{n \times n_0}$ be given matrices. Assume that A_0 is non singular. Let $\widetilde{c}_1, \widetilde{d}_1$ be two constants that fulfil for all $v \in V$*

$$\widetilde{d}_1 (Q_0\, v,\, v) \leq (A\, P\, A_0^{-1}\, R\, v,\, A\, P\, A_0^{-1}\, R\, v) \leq \widetilde{c}_1 (Q_0\, v,\, v). \tag{3.3}$$

Then it follows that

$$c_1 \leq \widetilde{c}_1 \quad \textit{and} \quad d_1 \geq \widetilde{d}_1.$$

If the inequality (3.3) for $\widetilde{c}_1$ ($\widetilde{d}_1$) is also true by equality for an $v^* \in V$, then it follows that

$$c_1 = \widetilde{c}_1 \quad (d_1 = \widetilde{d}_1).$$

proof. We start with the proposition for c_1. As we have that $R : \mathbb{R}^n \to \mathbb{R}^{n_0}$ is surjective and $A_0^{-1} \in \mathbb{R}^{n_0 \times n_0}$ is non singular, it follows that

$$\begin{aligned}
&\widetilde{c}_1(P\,S\,R\,v,\,v) \geq (A\,P\,A_0^{-1}\,R\,v,\,A\,P\,A_0^{-1}\,R\,v), \quad \forall v \in \mathbb{R}^n \\
\Leftrightarrow \quad &\widetilde{c}_1(S\,\widetilde{v}_0,\,\widetilde{v}_0) \geq (A\,P\,A_0^{-1}\,\widetilde{v}_0,\,A\,P\,A_0^{-1}\,\widetilde{v}_0), \quad \forall \widetilde{v}_0 \in \mathbb{R}^{n_0} \\
\Leftrightarrow \quad &\widetilde{c}_1(S\,A_0\,A_0^{-1}\,\widetilde{v}_0,\,A_0\,A_0^{-1}\,\widetilde{v}_0) \geq (A\,P\,A_0^{-1}\,\widetilde{v}_0,\,A\,P\,A_0^{-1}\,\widetilde{v}_0), \quad \forall \widetilde{v}_0 \in \mathbb{R}^{n_0} \\
\Leftrightarrow \quad &\widetilde{c}_1(S\,A_0\,\widetilde{v}_0,\,A_0\,\widetilde{v}_0) \geq (A\,P\,\widetilde{v}_0,\,A\,P\,\widetilde{v}_0), \quad \forall \widetilde{v}_0 \in \mathbb{R}^{n_0} \\
\Leftrightarrow \quad &\widetilde{c}_1(S\,R\,A\,P\,\widetilde{v}_0,\,R\,A\,P\widetilde{v}_0) \geq (A\,P\,\widetilde{v}_0,\,A\,P\,\widetilde{v}_0), \quad \forall \widetilde{v}_0 \in \mathbb{R}^{n_0} \\
\Leftrightarrow \quad &\widetilde{c}_1(P\,S\,R\,A\,P\,\widetilde{v}_0,\,A\,P\widetilde{v}_0) \geq (A\,P\,\widetilde{v}_0,\,A\,P\,\widetilde{v}_0), \quad \forall \widetilde{v}_0 \in \mathbb{R}^{n_0} \\
\Leftrightarrow \quad &\widetilde{c}_1(Q_0\,A\,\widetilde{v}_0,\,Q_0\,A\,\widetilde{v}_0) \geq (A\,P\,\widetilde{v}_0,\,A\,P\,\widetilde{v}_0), \quad \forall \widetilde{v}_0 \in \mathbb{R}^{n_0}.
\end{aligned}$$

As this inequality is true for all $v \in V$ this implies

$$\widetilde{c}_1 \geq \max_{\widetilde{v}_0 \in \widetilde{V}_0 \setminus \{0\}} \frac{\|A\,P\,\widetilde{v}_0\|^2}{\|Q_0\,A\,P\,\widetilde{v}_0\|^2} = \max_{v_0 \in V_0 \setminus \{0\}} \frac{\|A\,v_0\|^2}{\|Q_0\,A\,v_0\|^2} = c_1.$$

This shows the inequality for c_1. The proposition for d_1 follows by the same arguments. Further, the proof shows that if there is a $v^* \in V$ that fulfils the inequality (3.3) for $\widetilde{c}_1$ by equality, and we define

$$v_0^* = A_0^{-1}\,R\,v^* \quad \text{and} \quad \widetilde{v}_0^* = P v_0^*$$

then the equality holds for $\widetilde{v^*}_0$ with

$$\widetilde{c}_1 = \frac{\|A\,\widetilde{v}_0^*\|^2}{\|Q_0\,A\,\widetilde{v}_0^*\|^2} = c_1.$$

This shows the additional proposition for c_1. The proposition for d_1 follows again by the same arguments. □

We highlight that for explicit calculations respectively estimations of c_1, d_1 we use a representation without a use of the inverse of A or A_0. The form as given in (3.3) is the form we want to use in estimations for the condition of the preconditioned systems. Further we remember that it is $d_1 \geq 1$. Hence we can estimate $d_1 = 1$.

Next, we will illustrate some characteristics of the operators R, Q_0 and the spaces W, V_0, that are shown in Lemma 2.3.4, for the multigrid situation.

Remark: 3.1.2. *For the operators R, Q_0 and the spaces W_1, V_0 as defined in chapter 2 the following characteristics hold in a two grid context:*

1. *For an arbitray $v \in V$, there are unique $v_0 \in V_0$, $w \in W$ so that we have*

$$v = Q_0\, v + (I - Q_0)\, v = v_0 + w.$$

   Further, it follows that $(v_0,\, w) = 0$.

2. *For all $w \in W$, it follows that $R\, w = 0$. For $v \in V$ we have $R\, v = 0$ if and only if we have $v \in W = (I - Q_0)(V)$.*

proof. These propositions follow immediately from Lemma 2.3.4. □

3.2 Introduction of C_{BPX}^{-1}

Now we will introduce into this quite general setting a preconditioner C_{BPX}^{-1} for the equation $Au = f$, which is in more special cases well-known as the BPX method. As already mentioned we use the exact inverse of A and A_0 respectively as we do not consider the quality of the approximation for these operators. We only consider the relation between the spaces and neglect the solutions in subspaces. Hence it is sufficient for our results to use the inverse operators. As we assume that A is non singular, the existence of the operator A_0^{-1} is discussed in Lemma 2.3.5. In this section, its existence is an assumption. For a non singular $A \in \mathbb{R}^{n\times n}$ and a non singular $A_0 \in \mathbb{R}^{n_0\times n_0}$ we define $C_{BPX}^{-1} \in \mathbb{R}^{n\times n}$ by

$$C_{BPX}^{-1} := A^{-1} + P\, A_0^{-1}\, R. \tag{3.4}$$

Our aim is to determine constants $c_{BPX}, d_{BPX} > 0$ that fulfil for all $v \in V$ the inequalities

$$c_{BPX}(A\, C_{BPX}^{-1}\, v,\, A\, C_{BPX}^{-1}\, v) \le (v,\, v) \le d_{BPX}(A\, C_{BPX}^{-1}\, v,\, A\, C_{BPX}^{-1}\, v). \tag{3.5}$$

More precisely, we will show on which characteristics the constants c_{BPX}, d_{BPX} depend.

As the space V is finite-dimensional, the existence of a constant $c_{BPX} > 0$ is always given. The existence of the constant $d_{BPX} > 0$ is equivalent to the non singularity of the operator $A\, C_{BPX}^{-1}$. Therefore, we will first show the existence of $d_{BPX} > 0$ and then give an estimation for it.

Lemma: 3.2.1. *Let $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$ be non singular. Then the matrix*

$$A\, C_{BPX}^{-1}$$

is also non singular.

proof. Suppose that $A\, C_{BPX}^{-1}$ is singular. Then there must be a $v \in V\backslash\{0\}$ with

$$\begin{aligned} & 0 = A\, C_{BPX}^{-1} v \\ \Leftrightarrow \quad & 0 = v + A\, P\, A_0^{-1}\, R\, v \\ \Leftrightarrow \quad & -v = A\, P\, A_0^{-1}\, R\, v \\ \Rightarrow \quad & -R\, v = \underbrace{R\, A\, P}_{=A_0}\, A_0^{-1}\, R\, v \\ \Leftrightarrow \quad & -R\, v = R\, v. \end{aligned}$$

For the given $v \in V$, we obtain $R\, v = 0$. However, in the case of $R\, v = 0$, we get

$$0 = AC_{BPX}^{-1} v = v + A\, P\, A_0^{-1}\, R\, v = v.$$

And hence, this is in contradiction to the assumption. □

To determine the constants c_{BPX} and d_{BPX}, we further need the angle between the two addends of $A\, C_{BPX}^{-1} v$. We define $\gamma_{BPX}^{+}, \gamma_{BPX}^{-}$ as

$$\gamma_{BPX}^{+} = \min\{t \in \mathbb{R}_+ \,:\, (A\, P\, A_0^{-1}\, R\, v,\, v) \le t\, \|A\, P\, A_0^{-1}\, R\, v\|\, \|v\|,\ \forall v \in V\} \tag{3.6}$$

(3.7)

$$\text{and} \quad \gamma_{BPX}^{-} = \min\{t \in \mathbb{R}_+ \,:\, (A\, P\, A_0^{-1}\, R\, v,\, v) \ge -t\, \|A\, P\, A_0^{-1}\, R\, v\|\, \|v\|,\ \forall v \in V\}.$$

So we get the following proposition:

Proposition: 3.2.2. *For non singular matrices $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$ and a given $R \in \mathbb{R}^{n_0\times n}$ the inequalities (3.5) hold with*

$$c_{BPX} = \frac{1}{1 + 2\gamma_{BPX}^{+}\sqrt{c_1} + c_1} \quad \textit{and} \quad d_{BPX} = \frac{1}{1 - (\gamma_{BPX}^{-})^2}.$$

proof. According to the definitions of this chapter, the inequality of Young A.0.3 with $\varepsilon = \sqrt{c_1}$ and Lemma 3.1.1 we have

$$\begin{aligned}
(A\,C_{BPX}^{-1}\,v,\,A\,C_{BPX}^{-1}\,v) &= (A\,A^{-1}\,v,\,A\,A^{-1}\,v) + 2(A\,A^{-1}\,v,\,A\,P\,A_0^{-1}\,R\,v) \\
&\quad + (A\,P\,A_0^{-1}\,R\,v,\,A\,P\,A_0^{-1}\,R\,v) \\
&\leq (v,\,v) + (A\,P\,A_0^{-1}\,R\,v,\,A\,P\,A_0^{-1}\,R\,v) \\
&\quad + 2\gamma_{BPX}^{+}\|v\|\,\|A\,P\,A_0^{-1}\,R\,v\| \\
&\leq (1+\gamma_{BPX}^{+}\varepsilon)(v,\,v) \\
&\quad + \left(1+\frac{\gamma_{BPX}^{+}}{\varepsilon}\right)(A\,P\,A_0^{-1}\,R\,v,\,A\,P\,A_0^{-1}\,R\,v) \\
&\leq (1+\gamma_{BPX}^{+}\varepsilon)(v,\,v) + \left(1+\frac{\gamma_{BPX}^{+}}{\varepsilon}\right)c_1(Q_0\,v,\,v) \\
&= (1+\gamma_{BPX}^{+}\sqrt{c_1})(v,\,v) + \left(1+\frac{\gamma_{BPX}^{+}}{\sqrt{c_1}}\right)c_1(Q_0\,v,\,v) \\
&\leq (1+\gamma_{BPX}^{+}\sqrt{c_1})(v,\,v) + \left(1+\frac{\gamma_{BPX}^{+}}{\sqrt{c_1}}\right)c_1(v,\,v) \\
&= (1+2\gamma_{BPX}^{+}\sqrt{c_1}+c_1)(v,\,v).
\end{aligned}$$

This proves the proposition for c_{BPX}. For d_{BPX}, it follows from the same arguments according to the inequality of Young with $\varepsilon = \gamma_{BPX}^{-}$

$$\begin{aligned}
(A\,C_{BPX}^{-1}\,v,\,A\,C_{BPX}^{-1}\,v) &\geq (v,\,v) + (A\,P\,A_0^{-1}\,R\,v,\,A\,P\,A_0^{-1}\,R\,v) \\
&\quad - 2\gamma_{BPX}^{-}\|v\|\,\|A\,P\,A_0^{-1}\,R\,v\| \\
&\geq (1-\gamma_{BPX}^{-}\varepsilon)(v,\,v) \\
&\quad + (1-\frac{\gamma_{BPX}^{-}}{\varepsilon})(A\,P\,A_0^{-1}\,R\,v,\,A\,P\,A_0^{-1}\,R\,v) \\
&\geq (1-\gamma_{BPX}^{-}\gamma_{BPX}^{-})(v,\,v) + \left(1-\frac{\gamma_{BPX}^{-}}{\gamma_{BPX}^{-}}\right)d_1(Q_0\,v,\,v) \\
&\geq (1-(\gamma_{BPX}^{-})^2)(v,\,v).
\end{aligned}$$

□

Lastly we will show a simple restriction for γ_{BPX}^{-} that will be useful later.

Corollary: 3.2.3. *Let $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$ be non singular. Then it follows that $\gamma_{BPX}^{-} < 1$.*

proof. Assume that $\gamma_{BPX}^{-} = 1$ holds. Then we have a $v \in V\backslash\{0\}$ that fulfils

$$\begin{aligned}(A\,P\,A_0^{-1}\,R\,v,\,v) &= -\|A\,P\,A_0^{-1}\,R\,v\|\,\|v\| \\ \Rightarrow \quad v &= -A\,P\,A_0^{-1}\,R\,v \\ \Rightarrow \quad 0 &= v + A\,P\,A_0^{-1}\,R\,v = A(A^{-1}\,v + P\,A_0^{-1}\,R\,v) \\ &= A\,C_{BPX}^{-1}\,v.\end{aligned}$$

Hence $A\,C_{BPX}^{-1}$ is singular and that contradicts Lemma 3.2.1. □

3.3 Introduction of C_{DT}^{-1}

In this section we will introduce the preconditioner C_{DT}^{-1} for the system of linear equations $Au = f$ the same setting. By the same arguments as in section 3.2, we also use the inverse A^{-1}, A_0^{-1} for the definition of the preconditioner. For non singular matrices $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$ we define C_{DT}^{-1} by

$$C_{DT}^{-1} := A^{-1}(I - Q_0) + P\,A_0^{-1}\,R. \tag{3.8}$$

Of course, our aim is again to determine constants $c_{DT}, d_{DT} > 0$ that fulfil the inequalities

$$c_{DT}(A\,C_{DT}^{-1}\,v,\,A\,C_{DT}^{-1}\,v) \le (v,\,v) \le d_{DT}(A\,C_{DT}^{-1}\,v,\,A\,C_{DT}^{-1}\,v) \tag{3.9}$$

for all $v \in V$. We will do the same steps as in section 3.2 and start with a proof of the existence of $d_{DT} > 0$.

Lemma: 3.3.1. *Let $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$ be non singular. Then the matrix*

$$A\,C_{DT}^{-1}$$

is also non singular.

proof. Assume that $A\,C_{DT}^{-1}$ is singular. Then there must be a $v \in V\backslash\{0\}$ with

$$\begin{aligned} & 0 = AC_{DT}^{-1}v \\ \Leftrightarrow \quad & 0 = (I - Q_0)\,v + A\,P\,A_0^{-1}\,R\,v \\ \Leftrightarrow \quad & -(I - P\,S\,R)\,v = A\,P\,A_0^{-1}\,R\,v \\ \Rightarrow \quad & -\underbrace{R\,(I - P\,S\,R)}_{=0}\,v = \underbrace{R\,A\,P}_{=A_0}\,A_0^{-1}\,R\,v \\ \Leftrightarrow \quad & 0 = R\,v. \end{aligned}$$

So it follows that $R\,v = 0$. But in this case we obtain

$$0 = AC_{DT}^{-1}v = (I - P\,S\,R)\,v + A\,P\,A_0^{-1}\,\underbrace{R\,v}_{=0} = v - P\,S\,R\,v = v.$$

And hence, this is in contradiction to the assumption. □

In analogy to the last section we need the angles between the addends of $A\,C_{DT}^{-1}$. Therefore we define

$$\gamma_{DT}^{+} := \min\Big\{t \in \mathbb{R}_+ \,:\, (A\,P\,A_0^{-1}\,R\,v,\,(I - Q_0)v) \leq t\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I - Q_0)v\|,\ \forall v \in V\Big\} \tag{3.10}$$

$$\text{and} \quad \gamma_{DT}^{-} := \min\Big\{t \in \mathbb{R}_+ \,:\, (A\,P\,A_0^{-1}\,R\,v,\,(I - P\,S\,R)v) \geq -t\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I - Q_0)v\|,\ \forall v \in V\Big\}. \tag{3.11}$$

Then we get the following estimations for the constants c_{DT} and d_{DT} :

Proposition: 3.3.2. *For non singular matrices $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$ and a given $R \in \mathbb{R}^{n_0\times n}$ the inequalities (3.9) hold with*

$$c_{DT} = \frac{2}{1 + c_1 + \sqrt{(c_1 - 1)^2 + 4c_1(\gamma_{DT}^{+})^2}} \quad \textit{and} \quad d_{DT} \;= \frac{2}{1 + d_1 - \sqrt{(1 - d_1)^2 + 4d_1(\gamma_{DT}^{-})^2}}.$$

In particular, we can also estimate that

$$d_{DT} = \frac{1}{1 - \gamma_{DT}^{-}}.$$

proof. According to the definition of C_{DT}^{-1} and the inequality of Young A.0.3 with

$$\varepsilon = \frac{c_1 - 1 + \sqrt{(1-c_1)^2 + 4c_1(\gamma_{DT}^+)^2}}{2\gamma_{DT}^+}$$

we obtain for all $v \in V$

$$\begin{aligned}
(A\,C_{DT}^{-1}\,v,\, A\,C_{DT}^{-1}\,v) &= ((I-Q_0)\,v, (I-Q_0)\,v) + (A\,P\,A_0^{-1}\,R\,v,\, A\,P\,A_0^{-1}\,R\,v) \\
&\quad + 2((I-Q_0)\,v,\, A\,P\,A_0^{-1}\,R\,v) \\
&\le ((I-Q_0)\,v, (I-Q_0)\,v) + (A\,P\,A_0^{-1}\,R\,v,\, A\,P\,A_0^{-1}\,R\,v) \\
&\quad + 2\gamma_{DT}^+ \|(I-Q_0)\,v\|\,\|A\,P\,A_0^{-1}\,R\,v\| \\
&\le ((I-Q_0)\,v, (I-Q_0)\,v)\,(1+\gamma_{DT}^+\varepsilon) \\
&\quad + (A\,P\,A_0^{-1}\,R\,v,\, A\,P\,A_0^{-1}\,R\,v)\left(1 + \frac{\gamma_{DT}^+}{\varepsilon}\right) \\
&\le ((I-Q_0)\,v,\, v)(1+\gamma_{DT}^+\varepsilon) + (Q_0\,v,\, v)\left(1 + \frac{\gamma_{DT}^+}{\varepsilon}\right) c_1 \\
&= ((I-Q_0)\,v,\, v)\left(1 + \gamma_{DT}^+ \frac{c_1 - 1 + \sqrt{(1-c_1)^2 + 4c_1(\gamma_{DT}^+)^2}}{2\gamma_{DT}^+}\right) \\
&\quad + (Q_0\,v,\, v)\left(1 + \frac{\gamma_{DT}^+}{\frac{c_1-1+\sqrt{(1-c_1)^2+4c_1(\gamma_{DT}^+)^2}}{2\gamma_{DT}^+}}\right) c_1 \\
&= ((I-Q_0)\,v,\, v)\left(\frac{c_1 + 1 + \sqrt{(1-c_1)^2 + 4c_1(\gamma_{DT}^+)^2}}{2}\right) \\
&\quad + (Q_0\,v,\, v)\left(1 + \frac{2\gamma_{DT}^+}{c_1 - 1 + \sqrt{(1-c_1)^2 + 4c_1(\gamma_{DT}^+)^2}}\right) c_1 \\
&= \left(\frac{c_1 + 1 + \sqrt{(1-c_1)^2 + 4c_1(\gamma_{DT}^+)^2}}{2}\right)\Big(((I-Q_0)\,v,\, v) + (Q_0\,v,\, v)\Big) \qquad (3.12) \\
&= \left(\frac{c_1 + 1 + \sqrt{(1-c_1)^2 + 4c_1(\gamma_{DT}^+)^2}}{2}\right)(v,\, v).
\end{aligned}$$

The equation (3.12) follows from the calculation:

$$\frac{c_1+1+\sqrt{(c_1-1)^2+4c_1(\gamma_{DT}^+)^2}}{2} = \left(1+\frac{2(\gamma_{DT}^+)^2}{c_1-1+\sqrt{(1-c_1)^2+4c_1(\gamma_{DT}^+)^2}}\right)c_1$$

$$\Leftrightarrow \quad c_1+1+\sqrt{(c_1-1)^2+4c_1(\gamma_{DT}^+)^2} = \left(\frac{c_1-1+\sqrt{(1-c_1)^2+4c_1(\gamma_{DT}^+)^2}+2(\gamma_{DT}^+)^2}{c_1-1+\sqrt{(1-c_1)^2+4c_1(\gamma_{DT}^+)^2}}\right)2c_1$$

$$\begin{aligned}\Leftrightarrow \quad & (c_1+1+\sqrt{(c_1-1)^2+4c_1(\gamma_{DT}^+)^2})(c_1-1+\sqrt{(1-c_1)^2+4c_1(\gamma_{DT}^+)^2}) \\ = \; & (c_1-1+\sqrt{(1-c_1)^2+4c_1(\gamma_{DT}^+)^2}+2(\gamma_{DT}^+)^2)2c_1\end{aligned}$$

$$\begin{aligned}\Leftrightarrow \quad & 2c_1^2-2c_1+2c_1\sqrt{(1-c_1)^2+4c_1(\gamma_{DT}^+)^2}+4c_1(\gamma_{DT}^+)^2 \\ = \; & 2c_1^2-2c_1+2c_1\sqrt{(1-c_1)^2+4c_1(\gamma_{DT}^+)^2}+4c_1(\gamma_{DT}^+)^2.\end{aligned}$$

This completes the proof for c_{DT}. For d_{DT} we get according to the inequality of Young with

$$\varepsilon = \frac{1-d_1+\sqrt{(1-d_1)^2+4d_1(\gamma_{DT}^-)^2}}{2\gamma_{DT}^-}$$

by the same arguments

$$\begin{aligned}(A\,C_{DT}^{-1}\,v,\,A\,C_{DT}^{-1}\,v) \geq\; & ((I-Q_0)\,v,(I-Q_0)\,v)+(A\,P\,A_0^{-1}\,R\,v,\,A\,P\,A_0^{-1}\,R\,v) \\ & -2\gamma_{DT}^-\|(I-Q_0)\,v\|\,\|A\,P\,A_0^{-1}\,R\,v\| \\ \geq\; & ((I-Q_0)\,v,\,v)\,(1-\gamma_{DT}^-\varepsilon)+(Q_0\,v,\,v)\left(1-\frac{\gamma_{DT}^-}{\varepsilon}\right)d_1 \\ =\; & ((I-Q_0)\,v,\,v)\left(1-\gamma_{DT}^-\frac{1-d_1+\sqrt{(1-d_1)^2+4d_1(\gamma_{DT}^-)^2}}{2\gamma_{DT}^-}\right) \\ & +(Q_0\,v,\,v)\left(1-\frac{\gamma_{DT}^-}{\frac{1-d_1+\sqrt{(1-d_1)^2+4c_1(\gamma_{DT}^-)^2}}{2\gamma_{DT}^-}}\right)d_1 \\ =\; & \frac{1}{d_{DT}}(v,\,v).\end{aligned}$$

The more simple expression for d_{DT} follows if we estimate $d_1 = 1$. This implies

$$\sup_{d_1 \geq 1} \frac{2}{1 + d_1 - \sqrt{(1-d_1)^2 + 4d_1(\gamma_{DT}^-)^2}} = \frac{2}{1 + d_1 - \sqrt{(1-d_1)^2 + 4d_1(\gamma_{DT}^-)^2}}\Bigg|_{d_1=1}$$
$$= \frac{2}{2 - \sqrt{4(\gamma_{DT}^-)^2}} = \frac{1}{1 - \gamma_{DT}}.$$

This completes the proof of the proposition. □

Again, we will conclude this section with a simple restriction for γ_{DT}^- that we will use later.

Corollary: 3.3.3. *Let $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$ be non singular. Then it follows that $\gamma_{DT}^- < 1$.*

proof. Assume that

$$(A\,P\,A_0^{-1}\,R\,v,\,(I - Q_0)\,v) = -\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I - Q_0)v\|$$

for a $v \in V\backslash\{0\}$. Then it follows that

$$\begin{aligned}(I - Q_0)v &= -A\,P\,A_0^{-1}\,R\,v \\ \Rightarrow \quad 0 &= (I - Q_0)v + A\,P\,A_0^{-1}\,R\,v = A(A^{-1}\,(I - Q_0)v + P\,A_0^{-1}\,R\,v) \\ &= A\,C_{DT}^{-1}\,v.\end{aligned}$$

Hence $A\,C_{BPX}^{-1}$ is singular and that contradicts Lemma 3.3.1. □

3.4 Relations between the constants

In this section we will show the relations between the constants we have defined in the last sections. We will see that we can reduce them to one constant.

Lemma: 3.4.1. *For a non singular $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$, the operator*

$$A\,P\,A_0^{-1}\,R : V \to \left\langle A\,P\,A_0^{-1}\,Re_1, \ldots, A\,P\,A_0^{-1}\,R\,e_n \right\rangle$$

is a projection and it follows

$$Q_0\,A\,P\,A_0^{-1}\,R\,v = Q_0\,v \quad \textit{for all} \quad v \in V. \tag{3.13}$$

proof. The calculation

$$(A\,P\,A_0^{-1}\,\underbrace{R)\,(A\,P}_{A_0}\,A_0^{-1}\,R) = A\,P\,A_0^{-1}\,A_0\,A_0^{-1}\,R$$

$$= A\,P\,A_0^{-1}\,R$$

shows that $A\,P\,A_0^{-1}\,R$ is a projection. The equation (3.13) follows from

$$Q_0\,A\,P\,A_0^{-1}\,R\,v = P\,S\,\underbrace{R\,A\,P}_{A_0}\,A_0^{-1}\,R\,v = P\,S\,R\,v = Q_0\,v.$$

□

From Lemma 3.4.1 we can conclude that the direction of the projection $A\,P\,A_0^{-1}\,R$ is orthogonal to V_0. That means that for all $v_0 \in V_0$ a $w \in W$ exists so that for all $w_1 \in W$ follows that

$$A\,P\,A_0^{-1}\,R\,(v_0 + w_1) = A\,P\,A_0^{-1}\,R\,v_0 = v_0 + w. \tag{3.14}$$

This points out that in general it is not a projection in the space V_0. In other words:

$$V_0 \neq (A\,P\,A_0^{-1}\,R)(V) = \left\langle A\,P\,A_0^{-1}\,Re_1, \ldots, A\,P\,A_0^{-1}\,R\,e_n \right\rangle.$$

This would only be the case if we additionally had $w = 0$ in (3.14). Furthermore, it is obvious that if

$$A\,P\,A_0^{-1}\,R\,v_0^* = v_0^* + w, \tag{3.15}$$

holds for an $v_0^* \in V_0$ then we obtain

$$Q_0\,A\,P\,A_0^{-1}\,R\,v_0^* = v_0^* \tag{3.16}$$

$$\text{and} \quad (I - Q_0)\,A\,P\,A_0^{-1}\,R\,v_0^* = w \tag{3.17}$$

This is illustrated in Figure 3.1 on page 69.
Now we can give a result for the constants $\gamma_{DT}^+, \gamma_{DT}^-$. The main aspect of this lemma is given by the fact that the elments

$$R\,v = v_0 \quad \text{and} \quad (I - Q_0)v = w$$

are elements of orthogonal subspaces. So for an $v \in V$ with $v = v_0 + w$, $v_0 \in V_0$, $w \in W$, the addends v_0, w can each be modified without a modification of the other one.

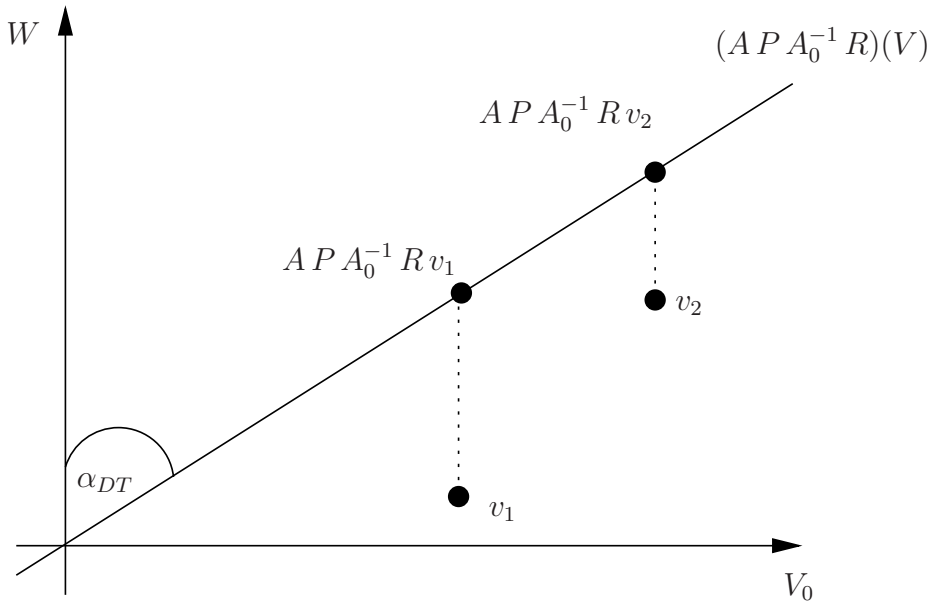


Figure 3.1: Direction of the projection $A\,P\,A_0^{-1}\,R$

Lemma: 3.4.2. *According to the definitions (3.10), (3.11) for* $\gamma_{DT}^+, \gamma_{DT}^-$

$$\gamma_{DT}^+ = \gamma_{DT}^- \quad \textit{holds.}$$

proof. We show that, for an arbitrary $v \in V$ with

$$(A\,P\,A_0^{-1}\,R\,v,\,(I-Q_0)v) = t\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I-Q_0)v\|, \quad t \geq 0$$

there is a $v_1 \in V$ that fulfils

$$(A\,P\,A_0^{-1}\,R\,v_1,\,(I-Q_0)v_1) = -t\,\|A\,P\,A_0^{-1}\,R\,v_1\|\,\|(I-Q_0)v_1\|.$$

Hence it follows that $\gamma_{DT}^+ \leq \gamma_{DT}^-$.
We consider an arbitrary $v \in V$. We can decompose this into $v = v_0 + w$, $v_0 \in V_0$, $w \in W$. Then we have

$$(A\,P\,A_0^{-1}\,R\,v,\,(I-Q_0)v) = t\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I-Q_0)v\|$$

with $t \in [0, \gamma_{DT}^+]$. According to Lemma 3.4.1 there is a $w_1 \in W$ so that it follows

$$A\,P\,A_0^{-1}\,R\,v = A\,P\,A_0^{-1}\,R\,v_0 = v_0 + w_1.$$

This implies that

$$\begin{aligned}(A\,P\,A_0^{-1}\,R\,v,\,(I-Q_0)v) &= t\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I-Q_0)v\| \\ \Leftrightarrow \quad (v_0+w_1,\,w) &= t\,\|v_0+w_1\|\,\|w\| \\ \Leftrightarrow \quad (w_1,\,w) &= t\,\|v_0+w_1\|\,\|w\|.\end{aligned}$$

It follows for

$$v_1 := v - 2w$$

that

$$\begin{aligned}-(w_1,\,w) &= -t\,\|v_0+w_1\|\,\|w\| \\ \Leftrightarrow \quad (v_0+w_1,\,-w) &= -t\,\|v_0+w_1\|\,\|-w\| \\ \Leftrightarrow \quad (A\,P\,A_0^{-1}\,R\,v_1,\,(I-Q_0)v_1) &= -t\,\|A\,P\,A_0^{-1}\,R\,v_1\|\,\|(I-Q_0)v_1\|.\end{aligned}$$

This shows the inequality $\gamma_{DT}^{+} \leq \gamma_{DT}^{-}$.
Based on the same arguments, it follows for an arbitrary $\widetilde{v} \in V$ with

$$(A\,P\,A_0^{-1}\,R\,\widetilde{v},\,(I-Q_0)\widetilde{v}) = -t\,\|A\,P\,A_0^{-1}\,R\,\widetilde{v}\|\,\|(I-Q_0)\widetilde{v}\|$$

for $\widetilde{v}_1 := \widetilde{v} - 2(I-Q_0)\widetilde{v}$ the equality

$$(A\,P\,A_0^{-1}\,R\,\widetilde{v}_1,\,(I-Q_0)\widetilde{v}_1) = t\,\|A\,P\,A_0^{-1}\,R\,\widetilde{v}_1\|\,\|(I-Q_0)\widetilde{v}_1\|.$$

This shows $\gamma_{DT}^{-} \leq \gamma_{DT}^{+}$. □

From the result of Lemma 3.4.2 we drop the constants $\gamma_{DT}^{+}, \gamma_{DT}^{-}$ and only use $\gamma_{DT} = \gamma_{DT}^{+} = \gamma_{DT}^{-}$ in the following.
Next we will prove a technical estimation for the relation of $(I-Q_0)\,A\,P\,A_0^{-1}\,R\,v_0$ and $Q_0\,A\,P\,A_0^{-1}\,R\,v_0$ that follows immediately from the angle γ_{DT}.

Lemma: 3.4.3. *By the constant γ_{DT} it holds for all $v \in V$ the inequality*

$$\|(I-Q_0)\,A\,P\,A_0^{-1}\,R\,v\|^2 \leq \frac{\gamma_{DT}^2}{1-\gamma_{DT}^2}\|Q_0\,A\,P\,A_0^{-1}\,R\,v\|^2.$$

proof. By Remark 3.1.2, it is sufficient to prove the inequality for all $v_0 \in V_0$. Therefore, we consider an arbitrary $v_0 \in V_0$. Then we obtain again by Lemma 3.4.1 that

$$A\,P\,A_0^{-1}\,R\,v_0 = v_0 + w$$

with $w \in W$. Furthermore, it follows that

$$v_0 = Q_0\, A\, P\, A_0^{-1}\, R\, v_0$$
$$w = (I - Q_0)\, A\, P\, A_0^{-1}\, R\, v_0.$$

From the definition of γ_{DT}, it follows for $v = v_0 + w$ that

$$\begin{aligned}
(A\, P\, A_0^{-1}\, R\, v,\, (I - Q_0)\, v) &\leq \gamma_{DT} \|A\, P\, A_0^{-1}\, R\, v\|\, \|(I - Q_0)\, v\| \\
\Leftrightarrow \quad (v_0 + w,\, w) &\leq \gamma_{DT} \|v_0 + w\|\, \|w\| \\
\Leftrightarrow \quad \|w\|^2 &\leq \gamma_{DT} \|v_0 + w\|\, \|w\| \\
\Leftrightarrow \quad \|w\| &\leq \gamma_{DT} \|v_0 + w\| \\
\Leftrightarrow \quad \|w\|^2 &\leq \gamma_{DT}^2 \|v_0 + w\|^2 = \gamma_{DT}^2 (\|v_0\|^2 + \|w\|^2) \\
\Leftrightarrow \quad \|w\|^2 (1 - \gamma_{DT}^2) &\leq \gamma_{DT}^2 \|v_0\|^2 \\
\Leftrightarrow \quad \|w\|^2 &\leq \frac{\gamma_{DT}^2}{(1 - \gamma_{DT}^2)} \|v_0\|^2 \\
\Leftrightarrow \quad \|(I - Q_0)\, A\, P\, A_0^{-1}\, R\, v_0\|^2 &\leq \frac{\gamma_{DT}^2}{(1 - \gamma_{DT}^2)} \|Q_0\, A\, P\, A_0^{-1}\, R\, v_0\|^2.
\end{aligned}$$

This shows the proposition. □

The dependency between $\|(I - Q_0)\, A\, P\, A_0^{-1}\, R\, v\|$ and $\|Q_0\, A\, P\, A_0^{-1}\, R\, v\|$ is illustrated in Figure 3.2. With $\cos(\alpha_{DT}) = \gamma_{DT}$, the figure also illustrates the angle between the spaces $(A\, P\, A_0^{-1}\, R)(V), W$.

Remark: 3.4.4. *If the inequality*

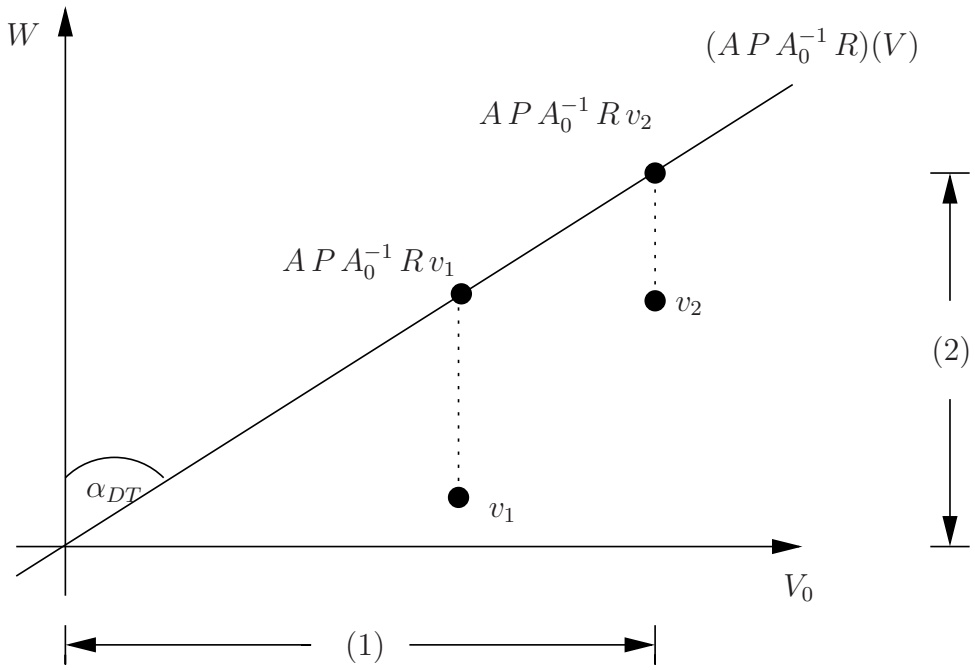
$$(A\, P\, A_0^{-1}\, R\, v,\, (I - Q_0)\, v) \leq \gamma_{DT} \|A\, P\, A_0^{-1}\, R\, v\|\, \|(I - Q_0)\, v\|$$

is also true by equality for a v^* then Lemma 3.4.3 also holds for this v^* by equality.

We go on and consider the constants c_1, d_1. As we have already noticed, we have $d_1 \geq 1$. Hence we can estimate d_1 by its lower bound and set $d_1 = 1$. For c_1, we can give an estimation that depends only on γ_{DT}.

Lemma: 3.4.5. *For the constants γ_{DT} and c_1 as defined in (3.1), we have*

$$c_1 \leq \frac{1}{1 - \gamma_{DT}^2}.$$



$$(1) = \|Q_0\,A\,P\,A_0^{-1}\,R\,v_2\| \qquad (2) = \|(I - Q_0)\,A\,P\,A_0^{-1}\,R\,v_2\|$$

Figure 3.2: Illustration of the quotient $\frac{\|Q_0\,A\,P\,A_0^{-1}\,R\,v_2\|}{\|(I-Q_0)\,A\,P\,A_0^{-1}\,R\,v_2\|}$

proof. From the Lemmata 3.4.1 and 3.4.3 we obtain for an arbitrary $v \in V$ with $v = v_0 + w_1$, $v_0 \in V_0$, $w_1 \in W$

$$A\,P\,A_0^{-1}\,R\;v = v_0 + w$$

$$\text{with} \quad w \in W, \quad \text{and} \quad \|w\|^2 \leq \frac{\gamma_{DT}^2}{1-\gamma_{DT}^2}\|v_0\|^2.$$

Hence it follows that

$$\begin{aligned}
\|A\,P\,A_0^{-1}\,R\,v\|^2 &= \|A\,P\,A_0^{-1}\,R\,v_0\|^2 \\
&= \|Q_0\,A\,P\,A_0^{-1}\,R\,v_0 + (I-Q_0)\,A\,P\,A_0^{-1}\,R\,v_0\|^2 \\
&= \|v_0+w\|^2 = \|v_0\|^2 + \|w\|^2 \\
&\leq \|v_0\|^2\left(1+\frac{\gamma_{DT}^2}{1-\gamma_{DT}^2}\right) = \|Q_0\,v\|^2\frac{1}{1-\gamma_{DT}^2}.
\end{aligned}$$

This shows the proposition. □

Remark: 3.4.6. *In particular, Lemma 3.4.5 shows the implication*

$$\gamma_{DT} = 0 \quad \Rightarrow \quad c_1 = 1.$$

As we consider finite-dimensional linear spaces, it is well posed to define γ_{DT} as the minimum of the set

$$\Big\{t \in \mathbb{R}_+ \,:\, (A\,P\,A_0^{-1}\,R\,v,\,(I-Q_0)v) \leq t\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I-Q_0)v\|,\;\forall v \in V\Big\}.$$

Hence there is a v^* for which it follows that

$$A\,P\,A_0^{-1}\,R\;v_0 = v_0 + w$$

$$\text{with} \quad (A\,P\,A_0^{-1}\,R\;v_0,\,w) = \gamma_{DT}\|A\,P\,A_0^{-1}\,R\;v_0\|\,\|w\|.$$

By Remark 3.4.4, this is the best estimation for c_1.
At last we will consider the constants $\gamma_{BPX}^{\pm}$. The next lemma shows that there is no case in which we have $\gamma_{BPX}^{+} < 1$ and that $\gamma_{BPX}^{-}, \gamma_{DT}$ are given by each other.

Lemma: 3.4.7. *For the constants $\gamma_{BPX}^{+}, \gamma_{BPX}^{-}$ as defined in (3.6), (3.7) we have*

$$\gamma_{BPX}^{+} = 1 \quad \textit{and} \quad \gamma_{BPX}^{-} = \gamma_{DT}.$$

proof. We start with the assertion for γ^{+}_{BPX}. By Lemma 3.4.1, we obtain for an arbitrary, but fixed $v = v_0 + w$ with $v_0 \in V_0$, $w \in W$, that there is a $w_1 \in W$ with

$$A\,P\,A_0^{-1}\,R\,v = A\,P\,A_0^{-1}\,R\,(v_0 + w) = v_0 + w_1.$$

Then we set

$$v_1 = v_0 + w_1$$

and hence it follows that

$$\begin{aligned}(A\,P\,A_0^{-1}\,R\,v_1,\, v_1) &= (v_0 + w_1,\, v_0 + w_1)\\ &= \|v_0 + w_1\|\,\|v_0 + w_1\|\\ &= \|A\,P\,A_0^{-1}\,R\,v_1\|\,\|v_1\|.\end{aligned}$$

This shows the proposition for γ^{+}_{BPX}.
For the proposition concerning γ^{-}_{BPX} and γ_{DT} we will show two inequalities.

$\gamma^{-}_{BPX} \le \gamma_{DT}$: We assume that there is a $\gamma_{DT} \le 1$ with

$$(A\,P\,A_0^{-1}\,R\,v,\,(I - Q_0)\,v) \ge -\gamma_{DT}\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I - Q_0)\,v\|$$

for all $v \in V$. We prove that this γ_{DT} fulfils also

$$(A\,P\,A_0^{-1}\,R\,v,\,v) \ge -\gamma_{DT}\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|v\| \quad \forall v \in V$$

$$(3.18) \quad \Leftrightarrow \quad (v_0 + w_1,\, v_0 + w) \ge -\gamma_{DT}\,\|v_0 + w_1\|\,\|v_0 + w\| \quad \forall v_0 \in V_0,\, \forall w \in W.$$

This implies $\gamma^{-}_{BPX} \le \gamma_{DT}$.
We consider an arbitray $v_0 \in V_0$ with $A\,P\,A_0^{-1}\,R\,v_0 = v_0 + w_1$. Then for all $w \in W$ with $\|w\| = \|w_1\|$ the left side of (3.18) is minimized if we set $w = -w_1$ and the right side of (3.18) is constant. Hence it is sufficient to consider $w = -\lambda w_1$ with $\lambda \in \mathbb{R}_+$. We obtain that (3.18) is equivalent to

$$\begin{aligned}(3.19) \qquad (v_0 + w_1,\, v_0 - \lambda w_1) &\ge -\gamma_{DT}\,\|v_0 + w_1\|\,\|v_0 - \lambda w_1\|\\ &= -\gamma_{DT}\,\sqrt{\|v_0\|^2 + \|w_1\|^2}\,\sqrt{\|v_0\| + \lambda^2\|w_1\|}\end{aligned}$$

With the shortcut $\|w_1\|^2 = b$ and the scaling $\|v_0\| = 1$ (for $v_0 = 0$, both sides of all inequalities are zero) this is equivalent to

$$(3.20) \qquad 1 - \lambda b \ge -\gamma_{DT}\sqrt{1 + b}\sqrt{1 + \lambda^2 b}.$$

If it is $\lambda b \leq 1$ then the inequality holds independent of γ_{DT}. Hence it is sufficient to consider the situation $\lambda b > 1$. (3.20) is fulfiled if we have

$$(1-\lambda b)^2 \leq \gamma_{DT}(1+b+\lambda^2 b+\lambda^2 b^2)$$

$$g := \frac{(1-\lambda b)^2}{(1+b+\lambda^2 b+\lambda^2 b^2)} \leq \gamma_{DT}.$$

If we differentiate g with respect to λ we obtain

$$\frac{\partial\, g}{\partial \lambda} = 2b(1+b)\frac{\lambda^2 b+\lambda b-\lambda-1}{\left(1+\lambda^2 b^2+(1+\lambda^2)b\right)^2}.$$

From the assumption $\lambda\, b > 1$ follows

$$\frac{\partial g}{\partial \lambda} = 2b(1+b)\frac{(\lambda+1)(\lambda\, b-1)}{\left(1+\lambda^2 b^2+(1+\lambda^2)b\right)^2} \geq 0.$$

So it is sufficient if the inequalities (3.20) and (3.19) respectively hold for the limit $\lambda \to \infty$. If we consider in (3.19) the limit with respect to λ we obtain

$$\begin{aligned}
&\lim_{\lambda\to\infty}\left((v_0+w_1,\, v_0-\lambda w_1) \geq -\gamma_{DT}\,\|v_0+w_1\|\,\|\lambda w_1\|\right)\\
\Leftrightarrow\quad &\lim_{\lambda\to\infty}\left(\|v_0\|^2-\lambda\,\|w_1\|^2 \geq -\gamma_{DT}\,\|v_0+w_1\|\,\|\lambda w_1\|\right)\\
\Leftrightarrow\quad &\lim_{\lambda\to\infty}\left(-\lambda\,\|w_1\|^2 \geq -\gamma_{DT}\,\|v_0+w_1\|\,\|\lambda w_1\|\right)\\
\Leftrightarrow\quad &-\|w_1\|^2 \geq -\gamma_{DT}\,\|v_0+w_1\|\,\|w_1\|)\\
\Leftrightarrow\quad &(v_0+w_1,\, -w_1) \geq -\gamma_{DT}\,\|v_0+w_1\|\,\|w_1\|\\
\Leftrightarrow\quad &(A\,P\,A_0^{-1}\,R\,v,\,(I-Q_0)v) \geq\\
&\qquad -\gamma_{DT}\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I-Q_0)v\|.
\end{aligned}$$

This inequality holds based on the assumptions.

$\gamma_{DT} \leq \gamma^-_{BPX}$: Now we assume that it holds

$$(A\,P\,A_0^{-1}\,R\,v,\,v) \geq -\gamma^-_{BPX}\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|v\| \quad \forall v \in V. \tag{3.21}$$

We prove that it follows

$$(A\,P\,A_0^{-1}\,R\,v,\,(I-Q_0)\,v) \geq -\gamma^-_{BPX}\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I-Q_0)\,v\| \quad \forall v \in V. \tag{3.22}$$

This implies $\gamma_{DT} \leq \gamma_{BPX}^{-}$.
We consider the inequality (3.22) for an arbitrary but fix $v \in V$ with $v = v_0 + w$ and $A\,P\,A_0^{-1}\,R\,v = v_0 + w_1$. If it is $v_0 = 0$ then it follows $w_1 = 0$ and the inequality (3.22) holds for all $\gamma_{BPX}^{-} \in [0, 1]$. Hence it is sufficient to consider $v_0 \neq 0$. Then the inequality (3.21) holds also for

$$v_\lambda = \lambda v_0 + w, \quad \lambda > 0.$$

This implies $A\,P\,A_0^{-1}\,R\,v_\lambda = \lambda(v_0 + w_1)$ and we obtain

$$(\lambda v_0 + \lambda w_1,\, \lambda v_0 + w) \geq -\gamma_{BPX}^{-}\,\|\lambda(v_0 + w_1)\|\,\|\lambda v_0 + w\|. \tag{3.23}$$

Based on $\lambda > 0$, this is equivalent to

$$(v_0 + w_1,\, \lambda v_0 + w) \geq -\gamma_{BPX}^{-}\,\|v_0 + w_1\|\,\|\lambda v_0 + w\|. \tag{3.24}$$

Since the inequality (3.24) holds for all $\lambda > 0$, this is also true for the limit $\lambda \to 0$. We obtain

$$\begin{aligned} & (v_0 + w_1,\, w) \geq -\gamma_{BPX}^{-}\,\|v_0 + w_1\|\,\|w\| \\ \Leftrightarrow \quad & (A\,P\,A_0^{-1}\,R\,v,\, v) \geq -\gamma_{BPX}^{-}\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|v\| \end{aligned}$$

for $v = v_0 + w$. This proves the second inequality.

□

3.5 Introduction of C_{2P}^{-1}

In this section we will introduce the preconditioner C_{2P}^{-1} into the same setting as $C_{DT}^{-1}, C_{BPX}^{-1}$ as a third possibility of a preconditioning. This preconditioner is motivated by the idea that for a symmetric matrix $A\,C_{BPX}^{-1}$ is just the symmetric alternative to C_{DT}^{-1}. So there is also a second possibility to modify C_{DT}^{-1} to a symmetric preconditioner. We use the same elements and define the preconditioner C_{2P}^{-1} for a non singular $A \in \mathbb{R}^{n\times n}$ and a non singular A_0 by

$$C_{2P}^{-1} := (I - Q_0)\,A^{-1}\,(I - Q_0) + P\,A_0^{-1}\,R. \tag{3.25}$$

Of course our aim is again to determine constants $c_{2P}, d_{2P} > 0$ that fulfil the inequalities

$$c_{2P}\|A\,C_{2P}^{-1}\,v\|^2 \leq \|v\|^2 \leq d_{2P}\|A\,C_{2P}^{-1}\,v\|^2 \tag{3.26}$$

for all $v \in V$. We will do the same steps as for the BPX and the DT method. Thus we will start with a proposition about the existence of d_{2P}. Therefore we get the same condition as we have in the DT-method and the BPX-method.

Lemma: 3.5.1. *Let $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$ be a non singular matrices. Then the matrix*

$$A\, C_{2P}^{-1}$$

is non singular.

proof. As A is non singular by the assumptions, the operator $A\, C_{2P}^{-1}$ is singular if and only if C_{2P}^{-1} is singular. This is given if and only if there is an $v \in V\backslash\{0\}$ that fulfils $C_{2P}^{-1}v = 0$. As we have for an arbitrary $v \in V$

$$\begin{aligned}((I - Q_0)\, A^{-1}\, (I - Q_0)\, v,\ P\, A_0^{-1}\, R\, v)&\\ = (R\, (I - Q_0)\, A^{-1}\, (I - Q_0)\, v,\ A_0^{-1}\, R\, v) = 0&\end{aligned}$$

the two addends are orthogonal to each other with respect to the inner procuct $(.,.)$. So for $C_{2P}^{-1}v = 0$ its a necessary condition that both addends are equal zero.
From the assumptions follows $P\, A_0^{-1} \in \mathbb{R}^{n\times n_0}$ and $rk(P\, A_0^{-1}) = n_0$. Thus we have

$$P\, A_0^{-1}\, R\, v = 0$$

if and only if we have $R\, v = 0$. This is equivalent to $v \in W = V_0^{\perp}$. If we assume $v \in W = V_0^{\perp}$ then it holds for the other addend

$$(I - Q_0)\, A^{-1}\, (I - Q_0)\, v = (I - Q_0)A^{-1}v$$

This is zero if and only if $A^{-1}\, v \in W^{\perp} = V_0$ holds. By Corollary 2.3.6 this contradicts the assumtion that A_0 is non singular. □

In this section we define the angles $\gamma_{2P}^{0,1}, \gamma_{2P}^{1,0}$ and γ_{2P} by

(3.27)
$$\gamma_{2P}^{0,1} = \min\left\{t \in \mathbb{R} \,:\, (A\,P\,A_0^{-1}R\,\widetilde{v}_0,\, w) \le t\|A\,P\,A_0^{-1}R\,\widetilde{v}_0\|\,\|w\|,\ \forall w \in W,\, v_0 \in V_0\right\}$$

(3.28)
$$\gamma_{2P}^{1,0} = \min\left\{t \in \mathbb{R} \,:\, (A\,(I-Q)\,A^{-1}\,(I-Q)\,w,\, v_0) \le t\|A\,(I-Q)\,A^{-1}\,(I-Q)\,w\|\,\|v_0\|,\ \forall w \in W,\, v_0 \in V_0\right\}$$

(3.29)
$$\gamma_{2P} = \min\left\{t \in \mathbb{R} \,:\, (A\,(I-Q)\,A^{-1}\,(I-Q)\,v,\, A\,P\,A_0^{-1}R\,v) \le t\,\|A\,(I-Q)\,A^{-1}\,(I-Q)\,v\|\,\|A\,P\,A_0^{-1}R\,v\|,\ \forall v \in V\right\}.$$

Based on these definitions, it is obvious that $\gamma_{2P}^{0,1}$ is the same constant as γ_{DT}. Furthermore, for a given matrix the constants $\gamma_{2P}^{0,1}$ and $\gamma_{2P}^{1,0}$ are easier to determine than γ_{2P}. However, the constant we will use for estimations for the c_{2P}, d_{2P} is γ_{2P}. We will do that as in sections 3.3 and 3.2. Therefore, we will prove a relation between these constants in the next lemma. This result is similar to the result of Lemma 3.4.3.

Lemma: 3.5.2. *Let $\gamma_{2P}^{0,1}, \gamma_{2P}^{1,0}$ be as defined in (3.27), (3.28). If we assume that $\gamma_{2P}^{0,1}, \gamma_{2P}^{1,0} < 1$ holds then we have*

$$\|(I-Q_0)\,A\,P\,A_0^{-1}R\,v_0\| \le \frac{\gamma_{2P}^{0,1}}{\sqrt{1-(\gamma_{2P}^{0,1})^2}}\|Q_0\,A\,P\,A_0^{-1}R\,v_0\| \quad \textit{for all} \quad v_0 \in V_0 \quad \textit{and}$$

$$\|Q_0\,A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,w\| \le \frac{\gamma_{2P}^{1,0}}{\sqrt{1-(\gamma_{2P}^{1,0})^2}}\|(I-Q_0)\,A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,w\| \quad \textit{for all} \quad w \in W.$$

proof. As mentioned, the constant $\gamma_{2P}^{0,1}$ is the same as γ_{DT}. For this constant we have proved the propostion in section 3.4. For $\gamma_{2P}^{1,0}$ the proof follows by the same arguments: for an arbitrary $w \in W$ we obtain

$$A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,w = w_1 + v_0 \quad \text{with} \quad w_1 \in W,\, v_0 \in V_0.$$

Hence it follows that

$$(I-Q_0)\,A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,w = w_1 \quad \text{and} \quad Q_0\,A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,w = v_0$$

and by the definition of $\gamma_{2P}^{1,0}$ we obtain for the selected w and for $v_0 = A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,w$ that

$$\begin{aligned}
(A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,w,\,v_0) &\le \gamma_{2P}^{1,0}\|A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,w\|\,\|v_0\|\ \forall v_0 \in V_0\\
\Rightarrow\quad (v_0+w_1,\,v_0) &\le \gamma_{2P}^{1,0}\|v_0+w_1\|\,\|v_0\|\\
\Leftrightarrow\quad \|v_0\|^2 &\le \gamma_{2P}^{1,0}\|v_0+w_1\|\,\|v_0\|\\
\Leftrightarrow\quad \|v_0\| &\le \gamma_{2P}^{1,0}\|v_0+w_1\|\\
\Leftrightarrow\quad \|v_0\|^2 &\le (\gamma_{2P}^{1,0})^2\|v_0+w_1\|^2 = (\gamma_{2P}^{1,0})^2\|v_0\|^2 + (\gamma_{2P}^{1,0})^2\|w_1\|^2\\
\Leftrightarrow\quad \|v_0\| &\le \frac{\gamma_{2P}^{1,0}}{\sqrt{1-(\gamma_{2P}^{1,0})^2}}\|w_1\|.
\end{aligned}$$

This proves the proposition for $\gamma_{2P}^{1,0}$. □

By the result of Lemma 3.5.2 we can represent the result of $A\,P\,A_0^{-1}R\,v$ and $(I-Q_0)\,A^{-1}\,(I-Q_0)\,v$ respectively for an arbitrary $v \in V$ by

$$\begin{aligned}
A\,P\,A_0^{-1}R\,v &= v_0+w_0, \quad \text{with} \quad v_0 \in V_0,\, w_0 \in W\\
\text{and}\quad A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,v &= v_1+w_1 \quad \text{with} \quad v_1 \in V_0,\, w_1 \in W.
\end{aligned}$$

Then we can represent the dot products (v_1, v_0) and (w_0, w_1) as follows:

$$(v_0,\,v_1) = \mu_1\,\|v_0\|\,\|w_1\|, \quad \text{with} \quad \mu_1 \le \frac{\gamma_{2P}^{1,0}}{\sqrt{1-(\gamma_{2P}^{1,0})^2}} \tag{3.30}$$

$$\text{and} \quad (w_0,\,w_1) = \mu_0\,\|w_1\|\,\|v_0\| \quad \text{with} \quad \mu_0 \le \frac{\gamma_{2P}^{0,1}}{\sqrt{1-(\gamma_{2P}^{0,1})^2}}. \tag{3.31}$$

Now we will prove an estimation for γ_{2P}.

Lemma: 3.5.3. *Let $\gamma_{2P}^{0,1}, \gamma_{2P}^{1,0}, \gamma_{2P}$ be as defined in (3.27), (3.28) and (3.29). If we assume that we have $\mu_0, \mu_1 < 1$ then*

$$\gamma_{2P} \le \gamma_{2P}^{1,0}\sqrt{1-(\gamma_{2P}^{0,1})^2} + \gamma_{2P}^{0,1}\sqrt{1-(\gamma_{2P}^{1,0})^2}$$

holds.

proof. Based on the definition of γ_{2P} we have to prove that the inequality

$$(A\,(I-Q)\,A^{-1}\,(I-Q)\,v, A\,P\,A_0^{-1}R\,v) \leq \gamma_{2P}\|A\,(I-Q)\,A^{-1}\,(I-Q)\,v\|\,\|A\,, P\,A_0^{-1}R\,v\| \tag{3.32}$$

holds for an arbitrary $v \in V$. For an arbitrary $v \in V$ we obtain

$$\begin{aligned} A\,P\,A_0^{-1}R\,v &= v_0 + w_0, \quad \text{with} \quad v_0 \in V_0,\, w_0 \in W \\ A\,(I-Q)\,A^{-1}\,(I-Q)\,v &= v_1 + w_1, \quad \text{with} \quad v_1 \in V_0,\, w_1 \in W. \end{aligned}$$

Hence the Proposition (3.32) is equivalent to

$$\begin{aligned} & (v_0 + w_0,\, v_1 + w_1) \leq \gamma_{2P}\|v_0 + w_0\|\,\|v_1 + w_1\| \\ \Leftrightarrow \quad & (v_0,\, v_1) + (w_0,\, w_1) \leq \gamma_{2P}\sqrt{\|v_0\|^2 + \|w_0\|^2}\,\sqrt{\|v_1\|^2 + \|w_1\|^2}. \end{aligned} \tag{3.33}$$

By the result of Lemma 3.5.2 we get

$$\begin{aligned} \|w_0\| = \mu_0\,\|v_0\| \quad \text{with} \quad \mu_0 \leq \frac{\gamma_{2P}^{0,1}}{\sqrt{1-(\gamma_{2P}^{0,1})^2}} \\ \|v_1\| = \mu_1\,\|w_1\| \quad \text{with} \quad \mu_1 \leq \frac{\gamma_{2P}^{1,0}}{\sqrt{1-(\gamma_{2P}^{1,0})^2}}. \end{aligned}$$

Hence the inequality (3.33) follows if we have

$$\begin{aligned} & \|v_0\|\,\|v_1\| + \|w_0\|\,\|w_1\| \leq \gamma_{2P}\sqrt{\|v_0\|^2 + \|w_0\|^2}\,\sqrt{\|v_1\|^2 + \|w_1\|^2} \\ \Leftarrow \quad & \|v_0\|\,\|w_1\|(\mu_0 + \mu_1) \leq \gamma_{2P}\sqrt{\|v_0\|^2(1+\lambda_0^2)}\,\sqrt{\|w_1\|^2(1+\lambda_1^2)} \\ \Leftrightarrow \quad & \frac{\mu_0 + \mu_1}{\sqrt{1+\lambda_0^2}\sqrt{1+\lambda_1^2}} \leq \gamma_{2P}. \end{aligned} \tag{3.34}$$

If we differentiate the left side with respect to μ_i, $i = 0, 1$, we obtain

$$\begin{aligned} \frac{d}{d\mu_0}\frac{(\mu_0+\mu_1)^2}{(1+\mu_0^2)(1+\mu_1^2)} &= \frac{(1-\mu_0\mu_1)}{(1+\mu_0^2)^2(1+\mu_1^2)^2} \\ \frac{d}{d\mu_1}\frac{(\mu_0+\mu_1)^2}{(1+\mu_0^2)(1+\mu_1^2)} &= \frac{(1-\mu_0\mu_1)}{(1+\mu_0^2)^2(1+\mu_1^2)^2}. \end{aligned}$$

By the assumption of $\mu_0\mu_1 < 1$ for all $v \in V$ is (3.34) increasing in μ_1, μ_0. Hence we can estimate them by their upper bound. This leads to the estimation

$$\begin{aligned}\frac{\mu_0+\mu_1}{\sqrt{1+\mu_0^2}\sqrt{1+\mu_1^2}} &\leq \frac{\frac{\gamma_{2P}^{0,1}}{\sqrt{1-(\gamma_{2P}^{0,1})^2}}+\frac{\gamma_{2P}^{1,0}}{\sqrt{1-(\gamma_{2P}^{1,0})^2}}}{\sqrt{1+\frac{(\gamma_{2P}^{0,1})^2}{(1-(\gamma_{2P}^{0,1})^2)}}\sqrt{1+\frac{(\gamma_{2P}^{1,0})^2}{(1-(\gamma_{2P}^{1,0})^2)}}} = \frac{\frac{\gamma_{2P}^{0,1}}{\sqrt{1-(\gamma_{2P}^{0,1})^2}}+\frac{\gamma_{2P}^{1,0}}{\sqrt{1-(\gamma_{2P}^{1,0})^2}}}{\sqrt{\frac{1}{1-(\gamma_{2P}^{0,1})^2}}\sqrt{\frac{1}{1-(\gamma_{2P}^{1,0})^2}}} \\ &= \sqrt{1-(\gamma_{2P}^{1,0})^2}\sqrt{1-(\gamma_{2P}^{0,1})^2}\left(\frac{\gamma_{2P}^{0,1}}{\sqrt{1-(\gamma_{2P}^{0,1})^2}}+\frac{\gamma_{2P}^{1,0}}{\sqrt{1-(\gamma_{2P}^{1,0})^2}}\right) \\ &= \gamma_{2P}^{1,0}\sqrt{1-(\gamma_{2P}^{0,1})^2}+\gamma_{2P}^{0,1}\sqrt{1-(\gamma_{2P}^{1,0})^2}.\end{aligned}$$

This shows the proposition. □

Thus Lemma 3.5.3 gives us an estimation for the angle γ_{2P} if $\gamma_{2P}^{0,1}$ and $\gamma_{2P}^{1,0}$ are small enough. We should mention that the special cases of $\gamma_{2P}^{0,1} = 0$ or $\gamma_{2P}^{1,0} = 0$ are included in the estimation above. Furthermore, we get the following corollary:

Corollary: 3.5.4. *Assume that $\gamma_{2P}^{0,1} = 0$ ($\gamma_{2P}^{1,0} = 0$) holds. Then we have*

$$\gamma_{2P} \leq \gamma_{2P}^{1,0} \quad (\gamma_{2P} \leq \gamma_{2P}^{0,1}).$$

proof. As we have $\gamma_{2P}^{0,1} = 0$ it follows from Lemma 3.5.2 $\mu_0 = 0$. Hence the proof follows immediately from Lemma 3.5.3. □

Next we will give an estimation for the condition of $A\,C_{2P}^{-1}$. For this estimation we have to quantify the condition that A_0 is non singular. This characteristic can be expressed by the introduction of the following constant: There exist constants $d_2, c_2 > 0$ which fulfil

$$\begin{aligned}(3.35)\quad & d_2\|(I-Q_0)\,v\|^2 \leq \|A\,(I-Q_0)\,A^{-1}\,(I-Q)\,v\|^2 \leq c_2\|(I-Q_0)\,v\|^2, \quad \forall v \in V \\ \Leftrightarrow\quad & d_2\|w\|^2 \leq \|A\,(I-Q_0)\,A^{-1}\,w\|^2 \leq c_2\|w\|^2, \quad \forall w \in W.\end{aligned}$$

The existence of c_2 is always given as the operators are finite dimensional. The constant d_2 exists if and only if there is no $v \in W$ that fulfils $A^{-1}\,v \in V_0$. And in Corollary 2.3.6 we have seen that this is equivalent to the non singularity of A_0. Further, we remember that $\gamma_{2P}^{0,1} = \gamma_{DT}$ and thus we obtain from the results of sections 3.4 and 3.1 that

$$d_1\|Q_0\,v\|^2 \leq \|A\,P\,A_0^{-1}\,R\,v\|^2 \leq c_1\|Q_0\,v\|^2 \quad \text{with} \quad d_1 = 1,\, c_1 = \frac{1}{1-\gamma_{DT}}$$

holds for all $v \in V$. By these assumptions we can give the following estimation for the condition of $A\,C_{2P}^{-1}$.

Proposition: 3.5.5. *Let $A \in \mathbb{R}^{n\times n}$, $A_0 \in \mathbb{R}^{n_0\times n_0}$ be non singular matrices. Then*

$$c_{2P}\|A\,C_{2P}^{-1}\,v\|^2 \le \|v\|^2 \le d_{2P}\|A\,C_{2P}^{-1}\,v\|^2$$

holds for all $v \in V$ with

$$c_{2P} = \frac{2}{c_2 + c_1 + \sqrt{(c_1-c_2)^2 + 4c_1c_2\gamma_{2P}^2}} \quad \textit{and} \quad d_{2P} = \frac{2}{d_1 + d_2 - \sqrt{(d_1-d_2)^2 + 4\gamma_{2P}^2 d_1 d_2}}$$

In particular, if V_0 is invariant with respect to A, the inequality holds with

$$c_{2P} = \frac{2}{c_2 + 1 + \sqrt{(c_2-1)^2 + 4c_2(\gamma_{2P}^{1,0})^2}} \quad \textit{and} \quad d_{2P} = \frac{2}{1 + d_2 - \sqrt{(1-d_2)^2 + 4(\gamma_{2P}^{1,0})^2}}.$$

proof. For d_{2P} we obtain with the constants d_1, d_2 and γ_{2P} from the inequality of Young with

$$\varepsilon = \frac{d_1 - d_2 + \sqrt{(d_1-d_2)^2 + 4d_2d_1\gamma_{2P}^2}}{2\gamma_{2P}}$$

for an arbitrary $v \in V$ that

$$\begin{aligned}
\|A\,C_{2P}^{-1}\|^2 &\ge \|A\,P\,A_0^{-1}\,R\,v\|^2 + \|A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,v\|^2 \\
&\quad - 2\gamma_{2P}\|A\,P\,A_0^{-1}\,R\,v\|\,\|A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,v\| \\
&\ge \|A\,P\,A_0^{-1}\,R\,v\|^2(1-\gamma_{2P}\varepsilon) + \|A\,(I-Q_0)\,A^{-1}\,(I-Q_0)\,v\|^2\left(1-\frac{\gamma_{2P}}{\varepsilon}\right) \\
&\ge \|Q_0\,v\|^2(1-\gamma_{2P}\varepsilon)d_1 + \|(I-Q_0)\,v\|^2(1-\frac{\gamma_{2P}}{\varepsilon})d_2 \\
&= \frac{d_1 + d_2 - \sqrt{(d_1-d_2)^2 + 4d_2d_1\gamma_{2P}^2}}{2}\|Q_0v\|^2.
\end{aligned}$$

This proves the proposition for d_{2P}. The proposition for c_{2P} follows similary with

$$\varepsilon = \frac{c_2 - c_1 + \sqrt{(c_1-c_2)^2 + 4\gamma_{2P}^2c_2c_1}}{2\gamma_{2P}c_1}.$$

If V_0 is invariant with respect to A then we obtain $\gamma_{2P}^{0,1} = \gamma_{DT} = 0$. Hence we have $c_1 \le \frac{1}{1-\gamma_{DT}^2} = 1$ and the result of Corollary 3.5.4 for γ_{2P}. This implies the additional assertion. □

3.6 Estimations by angles

In the sections 3.2 and 3.3 we have introduced the two preconditioners C_{DT}^{-1} and C_{BPX}^{-1} for the linear system $A\,u = f$ and we have given estimations for the condition of $A\,C_{DT}^{-1}$ and $A\,C_{BPX}^{-1}$ in the Euclidean norm. In section 3.4 we have shown relations between the constants that determine the estimations for the condition. Now we can reduce those constants that are given by the stiffness matrix and the structure of the subspace to γ_{DT}. First we will show some additional resutlts for the angle γ_{DT}. With that it will also be possible to get the best possible estimation for the condition in the Euclidean norm. Hence we can compare the methods and analyse the behaviour with respect to this characteristic. As a short cut we set

$$\mu_{\gamma_{DT}} := \frac{\gamma_{DT}}{\sqrt{1-\gamma_{DT}^2}} \quad \text{for} \quad \gamma_{DT} \in [0,1). \tag{3.36}$$

The operators C_{DT}^{-1}, C_{BPX} are well-posed if A, A_0 are non singular. By Corollary 3.3.3 we have in this case $\gamma_{DT} < 1$ and hence $\mu_{\gamma_{DT}}$ is also well posed for the operators of our interest .
Further, we define for an $v \in V$ the Operator $Q_v : V \to < v >$ as the orthogonal projection with respect to the dot product $(.,.)$.

3.6.1 Basics for angles

As we will need some basic results for the angles between spaces we will present them first. In this process, we will also take a look at the situation that is given for the operators defined in previous sections of this chapter.

Lemma: 3.6.1. *Let V be a vector space, W be a vector subspace and $B : V \to V$, a linear operator. Then we have for an arbitrary but fix $v \in V$ with $B\,v \neq 0$*

$$t_0^v := \inf\{t \in \mathbb{R} \,:\, (B\,v,\, w) \leq t\,\|B\,v\|\,\|w\|,\ \forall w \in W\} = \sup_{w\in W,\, w\neq 0} \frac{(B\,v,\ w)}{\|B\,v\|\,\|w\|} =: \widetilde{t}_0^v.$$

proof. For an arbitrary $v \in V$ we obtain from the definition of $\widetilde{t}_0^v$

$$\frac{(B\,v,\ w)}{\|B\,v\|\,\|w\|} \leq \widetilde{t}_0^v,\ \forall w \in W\backslash\{0\}$$

$$\Rightarrow \quad (B\,v,\ w) \leq \widetilde{t}_0^v\,\|B\,v\|\,\|w\|, \quad \forall w \in W.$$

This implies

$$\widetilde{t}_0^v \geq t_0^v.$$

The other inequality follows by the same argument. This completes the proof. □

Lemma: 3.6.2. *Let V be a finite dimensional vector space and V_0, W two vector subspaces that fulfil $(v_0,\, w) = 0$ for all $v_0 \in V_0, w \in W$. Assume that $I - Q : V \to W$ is the orthogonal projection with respect to the inner product $(.,.)$. Assume further that $B : V \to V$ is a linear operator with $ker(B) = W$ and $B\,v \notin W$ for all $v \in V$. Assume that, for an $v^* \in V$,*

$$t_0 = \min\{t \in \mathbb{R} \,:\, (B\,v^*,\, w) \leq t\,\|B\,v^*\|\,\|w\|,\ \forall w \in W\} \tag{3.37}$$

holds with a $t_0 \in [0, 1)$. Then we can draw the following conclusions:

1. *The inequality (3.37) holds if and only if we have*

$$(B\,v^*,\, (I - Q)\,B\,v^*) = t_0\,\|B\,v^*\|\,\|(I - Q)\,B\,v^*\|. \tag{3.38}$$

2. *In the case of $v^* \notin W$, we have*

$$\frac{\|(I-Q)Bv^*\|^2}{\|Q\,Bv^*\|^2} = \frac{t_0^2}{1-t_0^2} \quad \textit{and} \quad \frac{\|B\,v^*\|^2}{\|Q\,B\,v^*\|^2} = \frac{1}{1-t_0^2}.$$

3. *If we have $v^* \notin W$ and $\lambda \in \mathbb{R}_+$ is given then it follows for all $w \in W$ with $\|w\| = 1$ that*

$$(B\,v^*,\, \lambda w) \leq \left(B\,v^*,\, \lambda\frac{(I-Q)Bv^*}{\|(I-Q)Bv^*\|}\right) = \|Q\,B\,v^*\|\left(\lambda\frac{t_0}{\sqrt{1-t_0^2}}\right)$$

$$\textit{and} \quad (B\,v^*,\, \lambda w) \geq \left(B\,v^*,\, -\lambda\frac{(I-Q)Bv^*}{\|(I-Q)Bv^*\|}\right) = -\|Q\,B\,v^*\|\left(\lambda\frac{t_0}{\sqrt{1-t_0^2}}\right).$$

proof. 1. First we consider the case that the condition (3.37) holds with $t_0 = 0$. As we have $(I - Q)B\,v^* \in W$, we obtain that the inequality

$$(B\,v^*,\, w) \leq t\|B\,v^*\|\,\|w\|$$

must also hold for $w = (I - Q)B\,v^*$. This implies

$$0 \leq \|(I - Q)B\,v^*\|^2 = (B\,v^*, (I - Q)B\,v^*) \leq 0 \cdot \|B\,v^*\|\,\|(I - Q)B\,v^*\| = 0.$$

Hence it, follows that $B\,v^* = 0$ or $(I - Q)B\,v^* = 0$. In both cases, the equality (3.38) holds with $t_0 = 0$. The other direction follows by the same argument.
Now we assume that we have $t_0 > 0$. Then we obtain $B\,v^* \neq 0$. First, in this, we assume that t_0 follows from (3.37). Based on Lemma 3.6.1 the definition of t_0 is equivalent to

$$t_0 = \sup_{w \in W,\, w \neq 0} \frac{(B\,v^*,\ w)}{\|B\,v^*\|\,\|w\|}.$$

As we have $(I - Q)\,B\,v^* \in W$, it follows that

$$\frac{(B\,v^*,\ (I-Q)\,B\,v^*)}{\|B\,v^*\|\,\|(I-Q)\,B\,v^*\|} \leq \sup_{w \in W,\, w \neq 0} \frac{(B\,v^*,\ w)}{\|B\,v^*\|\,\|w\|} = t_0.$$

However for a $w \in W$, $w \neq (I - Q)\,B\,v$ we obtain that there are $\widetilde{w} \in W$ and $\lambda \in \mathbb{R}$ that fulfil

$$w = \widetilde{w} + \lambda\,(I - Q)\,B\,v^*$$

$$\text{and} \quad 0 = (\widetilde{w},\ (I - Q)\,B\,v^*).$$

It follows that

$$\frac{(B\,v^*,\ w)}{\|B\,v^*\|\,\|w\|} = \frac{(B\,v^*,\ \lambda(I-Q)\,B\,v^* + \widetilde{w})}{\|B\,v^*\|\,\|\lambda(I-Q)\,B\,v^* + \widetilde{w}\|} = \frac{(B\,v^*,\ \lambda(I-Q)\,B\,v^*)}{\|B\,v^*\|\,\sqrt{\|\lambda(I-Q)\,B\,v^*\|^2 + \|\widetilde{w}\|^2}}$$

$$\leq \frac{(B\,v^*,\ \lambda(I-Q)\,B\,v^*)}{\|B\,v^*\|\,\|\lambda(I-Q)\,B\,v^*\|} = \frac{(B\,v^*,\ (I-Q)\,B\,v^*)}{\|B\,v^*\|\,\|(I-Q)\,B\,v^*\|}.$$

This implies

$$\sup_{w \in W,\, w \neq 0} \frac{(B\,v^*,\ w)}{\|B\,v^*\|\,\|w\|} = \frac{(B\,v^*,\ (I-Q)\,B\,v^*)}{\|B\,v^*\|\,\|(I-Q)\,B\,v^*\|}.$$

Based on the same calculation we obtain (3.37) if we define t_0 by (3.38).

2. From the first result of this lemma it follows that

$$(B\,v^*,\ (I-Q)\,Bv^*) = t_0\,\|B\,v^*\|\,\|(I-Q)\,B\,v^*\|$$

$$\Leftrightarrow \quad \|(I-Q)\,Bv^*\|^2 = t_0\,\|B\,v^*\|\,\|(I-Q)\,B\,v^*\|$$

$$\Leftrightarrow \quad \|(I-Q)\,Bv^*\| = t_0\,\|B\,v^*\|$$

$$\Leftrightarrow \quad \|(I-Q)\,Bv^*\|^2 = t_0^2\,(\|Q\,B\,v^*\|^2 + \|(I-Q)\,B\,v^*\|^2)$$

$$\Leftrightarrow \quad \|(I-Q)\,Bv^*\|^2(1-t_0^2) = t_0^2\,\|Q\,B\,v^*\|^2$$

$$\Leftrightarrow \quad \frac{\|(I-Q)Bv^*\|^2}{\|Q\,Bv^*\|^2} = \frac{t_0^2}{1-t_0^2}$$

and also that

$$\frac{\|Bv^*\|^2}{\|Q\,Bv^*\|^2} = \frac{\|Q\,B\,v^*\|^2 + \|(I-Q)Bv^*\|^2}{\|Q\,Bv^*\|^2} = 1 + \frac{t_0^2}{1-t_0^2} = \frac{1}{1-t_0^2}.$$

3. From the result of Lemma 3.6.1 and the first result of this lemma, as well as from the definition of t_0, it follows for all $w \in W\backslash\{0\}$ that

$$\frac{(B\,v^*,\,w)}{\|Bv^*\|\,\|w\|} \leq \frac{(B\,v^*,\,(I-Q)B\,v^*)}{\|Bv^*\|\,\|(I-Q)B\,v^*\|}.$$

Hence it follows for all $w \in W$ with $\|w\| = 1$ that

$$(B\,v^*,\,w) \leq \left(B\,v^*,\,\frac{(I-Q)Bv^*}{\|(I-Q)Bv^*\|}\right).$$

And from the second result of this lemma we obtain

$$\begin{aligned}\left(B\,v^*,\,\lambda\frac{(I-Q)Bv^*}{\|(I-Q)Bv^*\|}\right) &= \left((I-Q)\,B\,v^*,\,\lambda\frac{(I-Q)Bv^*}{\|(I-Q)Bv^*\|}\right)\\ &= \lambda\frac{\|(I-Q)Bv^*\|^2}{\|(I-Q)Bv^*\|} = \lambda\|Q\,B\,v^*\|\frac{t_0}{\sqrt{1-t_0^2}}.\end{aligned}$$

However, by the same arguments we have

$$(B\,v^*,\,\lambda w) \geq -\lambda\left(B\,v^*,\,\frac{(I-Q)Bv^*}{\|(I-Q)Bv^*\|}\right) = -\|Q\,B\,v^*\|\lambda\frac{t_0}{\sqrt{1-t_0^2}}$$

for all $w \in W$ with $\|w\| = 1$.

□

Remark: 3.6.3. *By Lemma 3.6.1 we obtain for $v \in V$ with $B\,v \neq 0$ that*

$$\sup_{w\in W,\,w\neq 0}\frac{(B\,v,\,w)}{\|Bv\|\,\|w\|} = \frac{(B\,v,\,(I-Q)Bv)}{\|Bv\|\,\|(I-Q)Bv\|}.$$

As we have $ker(B) = W$, we get for the case of $B = A\,P\,A_0^{-1}\,R$ by the definition of γ_{DT} that

$$\begin{aligned}\gamma_{DT} &= \sup_{v\in V, w\in W;\,Bv,w\neq 0}\frac{(B\,v,\,w)}{\|Bv\|\,\|w\|} = \sup_{v\in V;\,Bv,(I-Q)Bv\neq 0,}\frac{(B\,v,\,(I-Q)Bv)}{\|Bv\|\,\|(I-Q)Bv\|}\\ &= \sup_{v_0\in V_0;\,Bv_0,(I-Q)Bv\neq 0}\frac{(B\,v_0,\,(I-Q)Bv_0)}{\|Bv_0\|\,\|(I-Q)Bv_0\|}.\end{aligned}$$

As we consider finite dimensional spaces V, there is a $v^* \in V$ *with*

$$\frac{(B\,v^*;\,(I-Q)Bv^*)}{\|Bv^*\|\,\|(I-Q)Bv^*\|} = \gamma_{DT}.$$

Otherwise we could only conclude that there is a sequence $(v^{*,k})_{k\in\mathbb{N}}$, *for which*

$$\lim_{k\to\infty} \frac{(B\,v^{*,k},\,(I-Q)Bv^{*,k})}{\|Bv^{*,k}\|\,\|(I-Q)Bv^{*,k}\|} = \gamma_{DT}$$

holds.

3.6.2 Estimations for the preconditioners

In this section we will show estimations for the conditions of $A\,C_{DT}^{-1}$ and $A\,C_{BPX}^{-1}$ in the Euclidean norm. As already mentioned, we will also show that these estimations are the best possible estimations. Further, we will compare the methods with each other and analyse the behaviour of the condition if the constant γ_{DT} increases or decreases.

Before we can start with the estimations we have to highlight two simple propositions for real numbers μ:

Remark: 3.6.4. *For all* $\mu \in \mathbb{R}_+$ *we have*

$$\frac{2+\mu^2-\mu\sqrt{4+\mu^2}}{2} \le 1 \le \frac{2+\mu^2+\mu\sqrt{4+\mu^2}}{2}.$$

Furthermore,

$$\frac{2+\mu^2-\mu\sqrt{4+\mu^2}}{2} = 1 = \frac{2+\mu^2+\mu\sqrt{4+\mu^2}}{2}$$

holds if and only if $\mu = 0$ *holds as well.*

proof. Based on $\mu \in \mathbb{R}_+$ it is obvious that

$$1 \le \frac{2+\mu^2+\mu\sqrt{4+\mu^2}}{2}$$

$$\text{and} \quad 1 = \frac{2+\mu^2+\mu\sqrt{4+\mu^2}}{2} \quad \Leftrightarrow \quad \mu = 0.$$

The other inequality follows from

$$1 \geq \frac{2+\mu^2-\mu\sqrt{4+\mu^2}}{2} \quad \Leftrightarrow \quad \mu^2 \leq \mu\sqrt{4+\mu^2}$$

$$\Leftrightarrow \quad \mu^2 \leq 4+\mu^2$$

$$\text{and} \quad 1 = \frac{2+\mu^2-\mu\sqrt{4+\mu^2}}{2} \quad \Leftrightarrow \quad \mu^2 = \mu\sqrt{4+\mu^2}.$$

The last equality holds if and only if $\mu = 0$ holds. □

Theorem: 3.6.5. *Let $A \in \mathbb{R}^{n\times n}$, $A_0 \in \mathbb{R}^{n_0\times n_0}$ be non singular and C_{DT}^{-1} as defined in (3.8). Then the inequalities*

$$c_{DT}\|A\,C_{DT}^{-1}\,v\|^2 \leq \|v\|^2 \leq d_{DT}\|A\,C_{DT}^{-1}\,v\|^2 \tag{3.39}$$

hold for all $v \in V$ with

$$c_{DT} := \frac{2+\mu_{\gamma_{DT}}^2-\mu_{\gamma_{DT}}\sqrt{4+\mu_{\gamma_{DT}}^2}}{2} \quad \textit{and} \quad d_{DT} := \frac{2+\mu_{\gamma_{DT}}^2+\mu_{\gamma_{DT}}\sqrt{4+\mu_{\gamma_{DT}}^2}}{2}.$$

proof. We consider an arbitrary $v \in V$. We can decompose this into $v = v_0 + w$, with $v_0 \in V_0$ and $w \in W$. If we have $v_0 = 0$, it follows that

$$\begin{aligned}\|A\,C_{DT}^{-1}\,v\|^2 &= \|(I-Q)v\|^2 + \|A\,P\,A_0^{-1}\,R\,v\|^2 \\ &\quad + 2(A\,P\,A_0^{-1}\,R\,v,\,(I-Q)\,v) \\ &= \|(I-Q)v\|^2 = \|w\|^2 = \|v\|^2.\end{aligned}$$

Hence the inequalities (3.39) hold with $c_{DT} = d_{DT} = 1$.

Now we assume that we have $v_0 \neq 0$. By Remark 3.6.4 we obtain for the given terms, $c_{DT} \leq 1, d_{DT} \geq 1$ and $1 = c_{DT} = d_{DT}$ if and only if we have $\mu_{\gamma_{DT}} = 0$. First we consider the inequality concerning c_{DT}. We can scale the inequality so that we can assume w.l.o.g. $\|v_0\| = \|w\| = 1$ and $v = v_0 + w\lambda$ with $\lambda \in \mathbb{R}$. For the given v_0 we obtain by Corollary 3.3.3

$$\left|(A\,P\,A_0^{-1}\,R\,v_0, (I-Q)\,A\,P\,A_0^{-1}\,R\,v_0)\right| = t\,\|A\,P\,A_0^{-1}\,R\,v_0\|\,\|(I-Q)\,A\,P\,A_0^{-1}\,R\,v_0\|$$

with a $t \in [0, \gamma_{DT}], \gamma_{DT} < 1$. So the setting $\mu_t^2 = t^2/(1-t^2)$ is well posed. Further we remember that

$$Q_0\,A\,P\,A_0^{-1}\,R\,v = Q_0\,v = v_0.$$

It follows that

$$
\begin{aligned}
&-c_{DT}\|A\,C_{DT}^{-1}\,v\|^2+\|v\|^2\\
&\quad=-c_{DT}\Big(\|(I-Q)v\|^2+\|A\,P\,A_0^{-1}\,R\,v\|^2+2(A\,P\,A_0^{-1}\,R\,v,\,(I-Q)\,v)\Big)\\
&\qquad+\|v_0\|^2+\|w\|^2\\
&\quad=-c_{DT}\left(\|\lambda w\|^2+\|A\,P\,A_0^{-1}\,R\,v\|^2+2(A\,P\,A_0^{-1}\,R\,v,\,\lambda\,w,)\right)\\
&\qquad+\|\lambda w\|^2+\|v_0\|^2\\
&\quad=-c_{DT}\Big(\lambda^2+\|Q_0\,A\,P\,A_0^{-1}\,R\,v\|^2+\|(I-Q_0)\,A\,P\,A_0^{-1}\,R\,v\|^2\\
&\qquad+2(\lambda\,w,\,(I-Q_0)\,A\,P\,A_0^{-1}\,R\,v)\Big)+\lambda^2+1\\
&\quad=-c_{DT}\left(\lambda^2+\|v_0\|^2+\mu_t^2\|v_0\|^2+2(\lambda\,w,\,(I-Q_0)\,A\,P\,A_0^{-1}\,R\,v)\right)+\lambda^2+1\\
&\quad\geq-c_{DT}\left(\lambda^2+1+\mu_t^2+2\lambda\mu_t\right)+\lambda^2+1 \qquad (3.40)\\
&\quad\geq-c_{DT}\left(\lambda^2+1+\mu_{\gamma_{DT}}^2+2\lambda\mu_{\gamma_{DT}}\right)+\lambda^2+1. \qquad (3.41)
\end{aligned}
$$

From the calculation above and Lemma 3.6.2 with $B = A\,P\,A_0^{-1}\,R$ we obtain the inequality (3.40). By the algebraic signs it is sufficient to consider $\lambda \in \mathbb{R}_+$. Hence we obtain the inequality (3.41) from the monotonie $t \leq \gamma_{DT} \Rightarrow \mu_t \leq \mu_{\gamma_{DT}}$ (cf. (A.0.6)). Further, we see that the inequality

$$0 \leq -c_{DT}\left(\lambda^2+1+\mu_{\gamma_{DT}}^2+2\lambda\mu_{\gamma_{DT}}\right)+\lambda^2+1 \qquad (3.42)$$

holds in the case of $\mu_{\gamma_{DT}} = 0$ with $c_{DT} = 1$. Now we can go on and consider the case of $\mu_{\gamma_{DT}} > 0$. Hence it follows $c_{DT} < 1$ from remark 3.6.4. If we differentiate (3.42) with respect to λ, we get

$$\frac{d}{d\lambda}\left[-c_{DT}\left(\lambda^2+1+\mu_{\gamma_{DT}}^2+2\lambda\mu_{\gamma_{DT}}\right)+\lambda^2+1\right]=2\lambda(1-c_{DT})-2c_{DT}\mu_{\gamma_{DT}}.$$

From the assumption of $c_{DT} < 1$ we obtain that (3.41) is minimized by $\lambda = \frac{c_{DT}\mu_{\gamma_{DT}}}{1-c_{DT}}$. From the minimizing value, it follows that

$$-c_{DT}\left(\lambda^2+1+\mu_{\gamma_{DT}}^2+2\lambda\mu_{\gamma_{DT}}\right)+\lambda^2+1=\frac{c_{DT}^2-c_{DT}(2+\mu_{\gamma_{DT}}^2)+1}{1-c_{DT}}.$$

As the denominator is positive, we just consider the nominator. The proposition for c_{DT} follows as the roots of the nominator with respect to c_{DT} are given by

$$c_{DT}=\frac{2+\mu_{\gamma_{DT}}^2\pm\mu_{\gamma_{DT}}\sqrt{4+\mu_{\gamma_{DT}}^2}}{2}.$$

We have to set the negative algebraic sign for c_{DT}, then we obtain for all $v \in V$ with $v = v_0 + \lambda w$, $v_0 \in V_0, w \in W$, $\lambda \in \mathbb{R}$ and $\|v_0\| = \|w\| = 1$ that

$$0 \leq \frac{c_{DT}^2 - c_{DT}(2 + \mu_{\gamma_{DT}}^2) + 1}{1 - c_{DT}} \leq -c_{DT}\|A\, C_{DT}^{-1}\, v\|^2 + \|v\|^2.$$

Next, we consider the estimation concerning d_{DT}. One more we decompose an arbitrary $v \in V$ into $v = v_0 + \lambda w$, with $v_0 \in V_0, w \in W$, $\|v_0\| = \|w\| = 1$ and $\lambda \in \mathbb{R}$. So we obtain

$$\left|(A\, P\, A_0^{-1}\, R\, v_0, (I - Q)\, A\, P\, A_0^{-1}\, R\, v_0)\right| = t\, \|A\, P\, A_0^{-1}\, R\, v_0\|\, \|(I - Q)\, A\, P\, A_0^{-1}\, R\, v_0\|$$

with a $t \in [0, \gamma_{DT}]$. Similary to the calculation done for as c_{DT} and Lemma 3.6.2 respectively we have

$$\begin{aligned} & d_{DT}\|A\, C_{DT}^{-1}\, v\|^2 - \|v\|^2 \\ & \quad = d_{DT}\left(\|\lambda w\|^2 + \|A\, P\, A_0^{-1}\, R\, v\|^2 + 2(\lambda\, w,\ A\, P\, A_0^{-1}\, R\, v)\right) \\ & \qquad - (\|\lambda w\|^2 + \|v_0\|^2) \\ & \quad \geq d_{DT}\left(\lambda^2 + 1 + \mu_t^2 - 2\lambda\mu_t\right) - (\lambda^2 + 1). \end{aligned} \tag{3.43}$$

The inequality (3.43) follows as it is again sufficient to consider $\lambda \in \mathbb{R}_+$. Again, for $\mu_t = 0$ we have

$$0 \leq d_{DT}\left(\lambda^2 + 1 + \mu_t^2 - 2\lambda\mu_t\right) - (\lambda^2 + 1)$$

with $d_{DT} = 1$. So we can further assume that we have $\mu_t > 0$ and $d_{DT} > 1$ for the proposed d_{DT}. We differentiate the term (3.43) with respect to λ. Then we get

$$\frac{d}{d\lambda}\left[d_{DT}\left(\lambda^2 + 1 + \mu_t^2 - 2\lambda\mu_t\right) - (\lambda^2 + 1)\right] = 2\lambda(d_{DT} - 1) - 2d_{DT}\mu_t.$$

Hence, (3.43) is minimized by $\lambda = \frac{d_{DT}\mu_t}{d_{DT}-1}$. It follows for $d_{DT} > 1$ that

$$\begin{aligned} d_{DT}\left(\lambda^2 + 1 + \mu_t^2 - 2\lambda\mu_t\right) - (\lambda^2 + 1) &= \lambda^2(d_{DT} - 1) - 2\lambda d_{DT}\mu_t + (\mu_t^2 + 1)d_{DT} - 1 \\ & \geq \frac{\mu_t^2 d_{DT}^2}{(d_{DT} - 1)} - \frac{2d_{DT}^2\mu_t^2}{d_{DT} - 1} + \frac{((\mu_t^2 + 1)d_{DT} - 1)(d_{DT} - 1)}{d_{DT} - 1} \\ & = \frac{d_{DT}^2 - d_{DT}(2 + \mu_t^2) + 1}{d_{DT} - 1} \\ & \geq \frac{d_{DT}^2 - d_{DT}(2 + \mu_{\gamma_{DT}}^2) + 1}{d_{DT} - 1}. \end{aligned} \tag{3.44}$$

The last inequality follows as the term is decreasing in μ_t, and from $t \in [0, \gamma_{DT}]$ follows $\mu_{\gamma_{DT}} \geq \mu_t$. As the denominator is positive for $d_{DT} > 1$, we just consider the nominator. We obtain that the roots of $d_{DT}^2 - d_{DT}(2 + \mu_{\gamma_{DT}}^2) + 1$ are given by

$$d_{DT} = \frac{2 + \mu_{\gamma_{DT}}^2 \pm \mu_{\gamma_{DT}} \sqrt{4 + \mu_{\gamma_{DT}}^2}}{2}. \tag{3.45}$$

If we set the positive algebraic sign for d_{DT} it follows for all $v \in V$ with $v = v_0 + \lambda w$, $v_0 \in V_0, w \in W$, $\lambda \in \mathbb{R}$ and $\|v_0\| = \|w\| = 1$ that

$$0 \leq \frac{d_{DT}^2 - d_{DT}(2 + \mu_{\gamma_{DT}}^2) + 1}{d_{DT} - 1} \leq d_{DT} \|A\, C_{DT}^{-1}\, v\|^2 - \|v\|^2.$$

This completes the proof. □

We can see immediately from the proof of Theorem 3.6.5 that the constants c_{DT}, d_{DT} are best possible. We only have to construct the minimizing elements given in the theorem.

Corollary: 3.6.6. *With the constants c_{DT}, d_{DT} as defined in Theorem 3.6.5, there is no $c^* > c_{DT}$ and no $d^* < d_{DT}$ so that the inequalities (3.39) hold with c^* and d^* respectively for all $v \in V$.*

proof. As the space V is finite-dimensional, there is a $v_0^* \in V_0$ with $\|v_0^*\| = 1$ so that we have

$$\begin{aligned}
&(A\, P\, A_0^{-1}\, R\, v_0^*, (I - Q)\, A\, P\, A_0^{-1}\, R\, v_0^*) \\
&\qquad = \gamma_{DT}\, \|A\, P\, A_0^{-1}\, R\, v_0^*\|\, \|(I - Q)\, A\, P\, A_0^{-1}\, R\, v_0^*\| \\
\Rightarrow \quad &(A\, P\, A_0^{-1}\, R\, v_0^*, \lambda(I - Q)\, A\, P\, A_0^{-1}\, R\, v_0^*) = \lambda \frac{\gamma_{DT}}{\sqrt{1 - \gamma_{DT}^2}} \|Q_0 v_0^*\|^2 \\
\text{and} \quad &\|A\, P\, A_0^{-1}\, R\, v_0^*\|^2 = \|Q_0 v_0^*\|^2 \frac{1}{1 - \gamma_{DT}^2}.
\end{aligned}$$

If we set

$$v_{\lambda,c}^* = v_0^* + \frac{(I - Q)\, A\, P\, A_0^{-1}\, R\, v_0^*}{\|(I - Q)\, A\, P\, A_0^{-1}\, R\, v_0^*\|} \underbrace{\frac{c_{DT} \mu_{\gamma_{DT}}}{1 - c_{DT}}}_{\lambda},$$

we obtain that the inequality

$$c_{DT} \|A\, C_{DT}^{-1} v_{\lambda,c}^*\|^2 \leq \|v_{\lambda,c}^*\|^2$$

holds by equality. So there is no bigger $c \in R$ that fulfils the inequality for all $v \in V$. The same follows for d_{DT} if we set

$$v^*_{\lambda,d} = v^*_0 - \frac{(I-Q)\,A\,P\,A_0^{-1}\,R\,v^*_0}{\|(I-Q)\,A\,P\,A_0^{-1}\,R\,v^*_0\|}\frac{d_{DT}\mu_{\gamma_{DT}}}{d_{DT}-1}.$$

□

Next, we will show that according to the Theorem, 3.6.5 the constants c_{DT}, d_{DT} are given by a simple one-dimensional optimization problem or a two dimensional restricted optimization problem respectively.

Corollary: 3.6.7. *The constants c_{DT}, d_{DT} of Theorem 3.6.5 are equivalent to*

$$c_{DT} = \min_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+1+\mu^2_{\gamma_{DT}}+2\lambda\mu_{\gamma_{DT}}} = \min_{\lambda\in\mathbb{R},\,\mu\in[0,\mu_{\gamma_{DT}}]} \frac{\lambda^2+1}{\lambda^2+1+\mu^2+2\lambda\mu}$$

$$d_{DT} = \max_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+1+\mu^2_{\gamma_{DT}}-2\lambda\mu_{\gamma_{DT}}} = \max_{\lambda\in\mathbb{R},\,\mu\in[0,\mu_{\gamma_{DT}}]} \frac{\lambda^2+1}{\lambda^2+1+\mu^2-2\lambda\mu}.$$

proof. By inequality (3.41), we obtain c_{DT} by

$$\begin{aligned} & 0 \le -c_{DT}\left(\lambda^2+1+\mu^2_{\gamma_{DT}}+2\lambda\mu_{\gamma_{DT}}\right)+1+\lambda^2,\ \forall\lambda\in\mathbb{R} \\ \Leftrightarrow \quad & c_{DT} \le \frac{\lambda^2+1}{\lambda^2+1+\mu^2_{\gamma_{DT}}+2\lambda\mu_{\gamma_{DT}}},\ \forall\lambda\in\mathbb{R}. \end{aligned}$$

The proof of Theorem 3.6.5 shows that the minimum of the right side exists. From Corollary 3.6.6 we obtain that the given constant is the biggest one, so it is given by the minimum of the right side. From the inequality (3.40), the proposition for the two-dimensional restricted system by the same arguments.
The proposition for d_{DT} follows by the same arguments. □

At last, we will consider the behaviour of the constants with respect to γ_{DT}. The results are quite easy to see.

Corollary: 3.6.8. *For the constants c_{DT}, d_{DT} of Theorem 3.6.5 we have*

$$\frac{d}{d\,\gamma_{DT}}[c_{DT}] < 0 \quad \textit{and} \quad \frac{d}{d\,\gamma_{DT}}[d_{DT}] > 0.$$

proof. By Lemma A.0.6, we have $\frac{d}{d\gamma_{DT}}[\mu_{DT}] > 0$. Hence it follows that

$$\begin{aligned}\frac{d}{d\,\gamma_{DT}}[d_{DT}] &= \frac{d}{d\,\mu_{DT}}[d_{DT}] \cdot \frac{d}{d\,\gamma_{DT}}[\mu_{DT}] \\ &= \frac{2\mu_{DT} + \sqrt{4+\mu_{DT}^2} + \frac{\mu_{DT}}{\sqrt{4+\mu_{DT}^2}}}{2} \cdot \underbrace{\frac{d}{d\,\gamma_{DT}}[\mu_{DT}]}_{>0} > 0\end{aligned}$$

and

$$\begin{aligned}\frac{d}{d\,\gamma_{DT}}[c_{DT}] &= \frac{d}{d\,\mu_{DT}}[c_{DT}] \cdot \frac{d}{d\,\gamma_{DT}}[\mu_{DT}] \\ &= \frac{2\mu_{DT} - \sqrt{4+\mu_{DT}^2} - \frac{\mu_{DT}^2}{\sqrt{4+\mu_{DT}^2}}}{2} \cdot \underbrace{\frac{d}{d\,\gamma_{DT}}[\mu_{DT}]}_{>0} < 0.\end{aligned}$$

The last inequality thus follows by

$$\begin{aligned}0 &> 2\mu_{DT} - \sqrt{4+\mu_{DT}^2} - \frac{\mu_{DT}^2}{\sqrt{4+\mu_{DT}^2}} \\ &= \frac{2\mu_{DT}\sqrt{4+\mu_{DT}^2} - (4+\mu_{DT}^2) - \mu_{DT}^2}{\sqrt{4+\mu_{DT}^2}} \\ &= -\frac{(\sqrt{4+\mu_{DT}^2} - \mu_{DT})^2}{\sqrt{4+\mu_{DT}^2}}.\end{aligned}$$

□

If we consider the two dimensional restricted system in Corollay 3.6.7, then the proposition for the behaviour of c_{DT}, d_{DT} follows quite simple as an bigger γ_{DT} implies a bigger $\mu_{\gamma_{DT}}$, and thus c_{DT} (or d_{DT}) is given as the minimum (maximum) for the same function on a bigger set.

Next, we will consider the BPX-method and we will get similar results. So at first we will again highlight a basic proposition on real numbers that will give the constants for the BPX-method.

Remark: 3.6.9. *For all $\mu \in \mathbb{R}_+$ we have*

$$\frac{5+\mu^2 - \sqrt{9+10\mu^2+\mu^4}}{8} < 1 \leq \frac{5+\mu^2 + \sqrt{9+10\mu^2+\mu^4}}{8}$$

$$\textit{and} \quad 1 = \frac{5+\mu^2 + \sqrt{9+10\mu^2+\mu^4}}{8} \quad \Leftrightarrow \quad \mu = 0.$$

proof. For

$$\frac{5+\mu^2+\sqrt{9+10\mu^2+\mu^4}}{8}$$

the proposition is obvious from $\mu \in \mathbb{R}_+$. We obtain the other inequality by

$$\begin{aligned} 1 &\geq \frac{5+\mu^2-\sqrt{9+10\mu^2+\mu^4}}{8} \\ \Leftrightarrow \quad \mu^2 - 3 &\leq \sqrt{9+10\mu^2+\mu^4} \\ \Leftrightarrow \quad \mu^4 - 6\mu^2 + 9 &\leq 9 + 10\mu^2 + \mu^4. \end{aligned}$$

If we insert $\mu = 0$, this implies $\frac{5+\mu^2-\sqrt{9+10\mu^2+\mu^4}}{8} < 1$. □

Later we will see that we can conclude from $c_{BPX}|_{\mu_{\gamma_{DT}}=0} = 1/2$ that $1/2$ is an upper bound for c_{BPX} (cf. Corollary 3.6.13).
Now we will proceed to the central result for the BPX-method.

Theorem: 3.6.10. *Let $A \in \mathbb{R}^{n\times n}$, $A_0 \in \mathbb{R}^{n_0\times n_0}$ be non singular and C_{BPX}^{-1} be as defined in (3.4). Then the inequalities*

$$c_{BPX}\|A\,C_{BPX}^{-1}\,v\|^2 \leq \|v\|^2 \leq d_{BPX}\|A\,C_{BPX}^{-1}\,v\|^2 \tag{3.46}$$

hold for all $v \in V$ with

$$c_{BPX} := \frac{5+\mu_{\gamma_{DT}}^2-\sqrt{9+10\mu_{\gamma_{DT}}^2+\mu_{\gamma_{DT}}^4}}{8} \tag{3.47}$$

$$\text{and} \quad d_{BPX} := \frac{5+\mu_{\gamma_{DT}}^2+\sqrt{9+10\mu_{\gamma_{DT}}^2+\mu_{\gamma_{DT}}^4}}{8}.$$

proof. We consider an arbitrary $v \in V$. First we assume that we have $v = w$, with $w \in W$. As $W = ker(R)$, it follows that

$$\|A\,C_{BPX}^{-1}\|^2 = \|v\|^2 + \|A\,P\,A_0^{-1}\,R\,v\|^2 + 2(A\,P\,A_0^{-1}\,R\,v,\,v) = \|v^2\|.$$

Hence, the inequalities (3.46) hold with $c_{BPX} = d_{BPX} = 1$ and by Remark 3.6.9 this is fulfil by the given constants.
Now we assume that we have $v = v_0 + w$, with $v_0 \in V_0$, $w \in W$ and $v_0 \neq 0$. Further, we highlight that we have $c_{BXP} < 1$ and $d_{BPX} \geq 1$. By the assumption of $v_0 \neq 0$, we can scale the inequality that way that we have $v = v_0 + \lambda w$, with $\lambda \in \mathbb{R}_+$ and $\|v_0\| = \|w\| = 1$. Further, we obtain from Corollary 3.2.3 for the given v_0 that

$$\left|(A\,P\,A_0^{-1}\,R\,v_0, (I-Q)\,A\,P\,A_0^{-1}\,R\,v_0)\right| = t\,\|A\,P\,A_0^{-1}\,R\,v_0\|\,\|(I-Q)\,A\,P\,A_0^{-1}\,R\,v_0\|$$

with a $t \in [0, \gamma_{DT}], \gamma_{DT} < 1$. Thus the setting $\mu_t^2 = t^2/(1-t^2)$ is well posed. As we have

$$Q_0\, A\, P\, A_0^{-1}\, R\, v = Q_0\, v = v_0.$$

it follows for c_{BPX} that

$$\begin{aligned}
(3.48)\quad & c_{BPX}\|A\, C_{BPX}^{-1}\|^2 \\
&= c_{BPX}\left(\|v\|^2 + \|A\, P\, A_0^{-1}\, R\, v\|^2 + 2(A\, P\, A_0^{-1}\, R\, v,\ v)\right) \\
&= c_{BPX}\Big(\|\lambda w\|^2 + \|v_0\|^2 + \|Q_0\, A\, P\, A_0^{-1}\, R\, v\|^2 + \|(I-Q_0)\, A\, P\, A_0^{-1}\, R\, v\|^2 \\
&\quad + 2(Q_0\, A\, P\, A_0^{-1}\, R\, v,\ v) + 2((I-Q_0)\, A\, P\, A_0^{-1}\, R\, v,\ v)\Big) \\
&= c_{BPX}\Big(\|\lambda w\|^2 + \|v_0\|^2 + \|v_0\|^2 + \mu_t^2\, \|v_0\|^2 \\
&\quad + 2(Q_0\, v,\ v) + 2((I-Q_0)\, A\, P\, A_0^{-1}\, R\, v,\ \lambda\, w)\Big) \\
&= c_{BPX}\Big(\lambda^2 + 2 + \mu_t^2 + 2(Q_0\, v,\ v) + 2((I-Q_0)\, A\, P\, A_0^{-1}\, R\, v,\ \lambda\, w)\Big) \\
(3.49)\quad & \leq c_{BPX}\Big(\lambda^2 + 4 + \mu_t^2 + 2\lambda\mu_t\Big) \\
& \leq c_{BPX}\Big(\lambda^2 + 4 + \mu_{\gamma_{DT}}^2 + 2\lambda\mu_{\gamma_{DT}}\Big).
\end{aligned}$$

Based on the algebraic signs it is sufficient to consider $\lambda \in \mathbb{R}_+$. This implies the last inequality above. As we have $\|v\|^2 = \|v_0\|^2 + \|\lambda w\|^2 = 1 + \lambda^2$, by the scaling of v_0, w we obtain

$$(3.50)\qquad -c_{BPX}\|A\, C_{BPX}\|^2 + \|v\|^2 \geq \lambda^2 + 1 - c_{BPX}\Big(\lambda^2 + 4 + \mu_{\gamma_{DT}}^2 + 2\lambda\mu_{\gamma_{DT}}\Big).$$

We differentiate (3.50) with respect to λ, we obtain

$$\begin{aligned}
&\frac{d}{d\lambda}\left(\lambda^2 + 1 - c_{BPX}\Big(\lambda^2 + 4 + \mu_{\gamma_{DT}}^2 + 2\lambda\mu_{\gamma_{DT}}\Big)\right) \\
&\qquad = 2\lambda - c_{BPX}(2\lambda + 2\mu_{\gamma_{DT}}).
\end{aligned}$$

Hence (3.50) is minimized by $\lambda = \frac{\mu_{\gamma_{DT}} c_{BPX}}{1-c_{BPX}}$ and we get

$$\lambda^2 + 1 - c_{BPX}\Big(\lambda^2 + 4 + \mu_{\gamma_{DT}}^2 + 2\lambda\mu_{\gamma_{DT}}\Big) \geq \frac{c_{BPX}^2 - c_{BPX}\frac{5+\mu_{\gamma_{DT}}^2}{4} + \frac{1}{4}}{1 - c_{BPX}}.$$

As the denominator is positive, we only take a closer look at the nominator. The roots of the nominator with respect to c_{BPX} are given by

$$c_{BPX} = \frac{5 + \mu_{\gamma_{DT}^2} \pm \sqrt{9 + 10\mu_{\gamma_{DT}}^2 + \mu_{\gamma_{DT}}^4}}{8}.$$

If we take the negative algebraic sign, it follows for all $v \in V$ with $v = v_0 + \lambda w$, $v_0 \in V_0, w \in W$, $\lambda \in \mathbb{R}$ and $\|v_0\| = \|w\| = 1$ that

$$0 \leq \frac{5 + \mu_{\gamma_{DT}^2} - \sqrt{9 + 10\mu_{\gamma_{DT}}^2 + \mu_{\gamma_{DT}}^4}}{8(1 - c_{BPX})} \leq \|v\|^2 - c_{BPX}\|A\, C_{BPX}^{-1}\, v\|^2.$$

This proves the propostion for c_{BPX}.
Next, we consider the proposition for d_{BPX}. By the calculation above we can again consider a $v \in V$ that fulfils $v = v_0 + \lambda w$ with $\|v_0\| = \|w\| = 1$ and $\lambda \in \mathbb{R}$. Hence it follows that

$$\begin{aligned}
&d_{BPX}\|A\, C_{BPX}^{-1}\|^2 \\
&\quad = d_{BPX}\Big(\|\lambda w\|^2 + \|v_0\|^2 + \|Q_0\, A\, P\, A_0^{-1}\, R\, v\|^2 + \|(I - Q_0)\, A\, P\, A_0^{-1}\, R\, v\|^2 \\
&\qquad + 2(Q_0\, A\, P\, A_0^{-1}\, R\, v,\ v) + 2((I - Q_0)\, A\, P\, A_0^{-1}\, R\, v,\ v)\Big) \\
&\quad = d_{BPX}\Big(\lambda^2 + 1 + \|Q_0\, v\|^2 + \mu_t^2\|Q_0\, v\|^2 \\
&\qquad + 2(Q_0\, v,\ v) + 2((I - Q_0)\, A\, P\, A_0^{-1}\, R\, v,\ v)\Big) \\
&\quad = d_{BPX}\Big(\lambda^2 + 2 + \mu_t^2 + 2(Q_0\, v,\ v) + 2((I - Q_0)\, A\, P\, A_0^{-1}\, R\, v,\ v)\Big) \\
&\quad \geq d_{BPX}\Big(\lambda^2 + 4 + \mu_t^2 - 2\lambda\mu_t\Big).
\end{aligned}$$

So we obtain

$$d_{BPX}\|A\, C_{BPX}^{-1}\|^2 - \|v\|^2 \geq d_{BPX}\Big(\lambda^2 + 4 + \mu_t^2 - 2\lambda\mu_t\Big) - (1 + \lambda^2). \tag{3.51}$$

In the case of $\mu_t = 0$, this holds by $d_{BPX} = 1$. This proofs the proposition for $\mu_t = 0$. Now we can assume that we have $\mu_t > 0$ and $d_{BPX} > 1$ for the proposed d_{BPX}. If we differentiate (3.51) with respect to λ, we get

$$\frac{d}{d\lambda}\Big[d_{BPX}\Big(\lambda^2 + 4 + \mu_t^2 - 2\lambda\mu_t\Big) - (1 + \lambda^2)\Big] = 2\lambda(d_{BPX} - 1) - 2\mu_t d_{BPX}.$$

Hence, (3.51) is minimized with respect to λ if we set $\lambda = \frac{\mu_t d_{BPX}}{d_{BPX}-1}$. From this value we obtain

$$d_{BPX}\Big(\lambda^2 + 4 + \mu_t^2 - 2\lambda\mu_t\Big) - (1+\lambda^2) = \frac{d_{BPX}^2 - d_{BPX}\frac{5+\mu_t^2}{4} + \frac{1}{4}}{d_{BPX}-1}$$
$$\geq \frac{d_{BPX}^2 - d_{BPX}\frac{5+\mu_{\gamma_{DT}}^2}{4} + \frac{1}{4}}{d_{BPX}-1}.$$

The denominator is positive, so we consider only the nominator. For this one, the roots with respect to d_{BPX} are given by

$$d_{BPX} = \frac{5 + \mu_{\gamma_{DT}}^2 \pm \sqrt{9 + 10\mu_{\gamma_{DT}}^2 + \mu_{\gamma_{DT}}^4}}{8}.$$

As we take the positive algebraic sign, it follows for all $v \in V$ with $v = v_0 + \lambda w$, $v_0 \in V_0, w \in W$, $\lambda \in \mathbb{R}$ and $\|v_0\| = \|w\| = 1$ that

$$0 \leq \frac{5 + \mu_{\gamma_{DT}^2} + \sqrt{9 + 10\mu_{\gamma_{DT}}^2 + \mu_{\gamma_{DT}}^4}}{8(d_{BPX}-1)} \leq d_{BPX}\|A\,C_{BPX}^{-1}\,v\|^2 - \|v\|^2.$$

□

As the constants c_{DT}, d_{DT} defined in Theorem 3.6.5 are best possible estimations, this is also the case for the constants c_{BPX}, d_{BPX}. In more formal words:

Corollary: 3.6.11. *With the constants c_{BPX}, d_{BPX} defined in Theorem 3.6.10, there is no $c^* > c_{BPX}$ and no $d^* < d_{BPX}$ so that the inequalieties (3.46) hold with c^* and d^* respectively for all $v \in V$.*

proof. As the space V is finite-dimensional, there is a $v_0^* \in V_0$ with $\|v_0^*\| = 1$ for which we obtain

$$(A\,P\,A_0^{-1}\,R\,v_0^*, (I-Q)\,A\,P\,A_0^{-1}\,R\,v_0^*)$$
$$= \gamma_{DT}\,\|A\,P\,A_0^{-1}\,R\,v_0^*\|\,\|(I-Q)\,A\,P\,A_0^{-1}\,R\,v_0^*\|$$
$$\Rightarrow \quad (A\,P\,A_0^{-1}\,R\,v_0^*, \lambda(I-Q)\,A\,P\,A_0^{-1}\,R\,v_0^*) = \lambda\frac{\gamma_{DT}}{\sqrt{1-\gamma_{DT}^2}}\|Q_0 v_0^*\|^2$$

and $\quad \|A\,P\,A_0^{-1}\,R\,v_0^*\|^2 = \|Q_0 v_0^*\|^2\frac{1}{1-\gamma_{DT}^2}.$

If we set

$$v_{\lambda,c}^* = v_0^* + \frac{(I-Q)\,A\,P\,A_0^{-1}\,R\,v_0^*}{\|(I-Q)\,A\,P\,A_0^{-1}\,R\,v_0^*\|}\underbrace{\frac{c_{BPX}\mu_{\gamma_{DT}}}{1-c_{BPX}}}_{\lambda},$$

the inequality

$$c_{BPX}\|A\,C_{BPX}^{-1}v_{\lambda,c}^*\|^2 \leq \|v_{\lambda,c}^*\|^2$$

is true with equality. So there is no bigger $c \in R$ that fulfils the inequality for all $v \in V$. The same follows for d_{BPX} if we set

$$v_{\lambda,d}^* = v_0^* - \frac{(I-Q)\,A\,P\,A_0^{-1}\,R\,v_0^*}{\|(I-Q)\,A\,P\,A_0^{-1}\,R\,v_0^*\|}\frac{d_{BPX}\mu_{\gamma_{DT}}}{d_{BPX}-1}.$$

□

Like we did for the DT-method, we also want to give a characterisation for c_{BPX}, d_{BPX} by a one-dimensional optimization problem (and a two-dimensional restricted optimization problem, respectively).

Corollary: 3.6.12. *The constants c_{BPX}, d_{BPX} defined in Theorem 3.6.10 are equivalent defined as*

$$c_{BPX} = \min_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+4+\mu_{\gamma_{DT}}^2+2\lambda\mu_{\gamma_{DT}}} = \min_{\lambda\in\mathbb{R},\mu\in[0,\mu_{\gamma_{DT}}]} \frac{\lambda^2+1}{\lambda^2+4+\mu^2+2\lambda\mu}$$

$$d_{BPX} = \max_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+4+\mu_{\gamma_{DT}}^2-2\lambda\mu_{\gamma_{DT}}} = \max_{\lambda\in\mathbb{R},\mu\in[0,\mu_{\gamma_{DT}}]} \frac{\lambda^2+1}{\lambda^2+4+\mu^2-2\lambda\mu}.$$

proof. From the inequality (3.50), we obtain c_{BPX} by

$$0 \leq -c_{BPX}\left(\lambda^2+4+\mu_{\gamma_{DT}}^2+2\lambda\mu_{\gamma_{DT}}\right)+1+\lambda^2, \quad \forall\lambda\in\mathbb{R}$$

$$\Leftrightarrow \quad c_{BPX} \leq \frac{\lambda^2+1}{\lambda^2+4+\mu_{\gamma_{DT}}^2+2\lambda\mu_{\gamma_{DT}}}, \quad \forall\lambda\in\mathbb{R}.$$

The proof of Theorem 3.6.10 shows the existence of the minimum of the right side. Hence, the minimum is the best estimation and by Corollary 3.6.11 we obtain that c_{BPX} is given as the best estimation. The representation by a restricted problem follows by the same arguments from (3.49).

$$0 \leq -c_{BPX}\left(\lambda^2+4+\mu_t^2+2\lambda\mu_t\right)+1+\lambda^2, \quad \forall\lambda\in\mathbb{R}, \forall\mu_t \leq \mu_{\gamma_{DT}}.$$

The assertion for d_{BPX} follows by the same arguments. □

At last we will consider the behaviour of the constants with respect to γ_{DT}. The results are again quite easy to see.

Corollary: 3.6.13. *For the constants c_{BPX}, d_{BPX} of Theorem 3.6.10, we have*

$$\frac{d}{d\,\gamma_{DT}}[c_{BPX}] < 0 \quad \textit{and} \quad \frac{d}{d\,\gamma_{DT}}[d_{BPX}] > 0.$$

proof. From Lemma A.0.6, it follows that $\frac{d}{d\gamma_{DT}}[\mu_{DT}] > 0$. Hence we obtain that

$$\begin{aligned}\frac{d}{d\,\gamma_{DT}}[d_{BPX}] &= \frac{d}{d\,\mu_{\gamma_{DT}}}[d_{BPX}] \cdot \frac{d}{d\,\gamma_{DT}}[\mu_{\gamma_{DT}}] \\ &= \frac{2\mu_{\gamma_{DT}} + \frac{10\mu_{\gamma_{DT}}+2\mu^3_{\gamma_{DT}}}{\sqrt{9+10\mu^2_{\gamma_{DT}}+\mu^4_{\gamma_{DT}}}}}{8} \cdot \underbrace{\frac{d}{d\,\gamma_{DT}}[\mu_{\gamma_{DT}}]}_{>0} > 0\end{aligned}$$

and

$$\begin{aligned}\frac{d}{d\,\gamma_{DT}}[c_{DT}] &= \frac{d}{d\,\mu_{DT}}[c_{BPX}] \cdot \frac{d}{d\,\gamma_{DT}}[\mu_{DT}] \\ &= \frac{2\mu_{DT} - \frac{10\mu_{\gamma_{DT}}+2\mu^3_{\gamma_{DT}}}{\sqrt{9+10\mu^2_{\gamma_{DT}}+\mu^4_{\gamma_{DT}}}}}{8} \cdot \underbrace{\frac{d}{d\,\gamma_{DT}}[\mu_{DT}]}_{>0} < 0.\end{aligned}$$

The last inequality follows from the consideration below:

$$\begin{aligned} & 0 > 2\mu_{\gamma_{DT}} - \frac{10\mu_{\gamma_{DT}} + 2\mu^3_{\gamma_{DT}}}{\sqrt{9 + 10\mu^2_{\gamma_{DT}} + \mu^4_{\gamma_{DT}}}} \\ \Leftrightarrow \quad & 4\mu^2_{\gamma_{DT}}(9 + 10\mu^2_{\gamma_{DT}} + \mu^4_{\gamma_{DT}}) < (10\mu_{\gamma_{DT}} + 2\mu^3_{\gamma_{DT}})^2 \\ \Leftrightarrow \quad & 36\mu^2_{\gamma_{DT}} < 100\mu^2_{\gamma_{DT}}.\end{aligned}$$

□

Again, as for the *DT*-method, we could also conclude from the restricted optimization problem as given in Corollary 3.6.12, that a smaller angle γ_{DT} gives a lower constant d_{BPX} and a bigger constant c_{BPX}.

Remark: 3.6.14. *In Corollaries 3.6.6 and 3.6.11, the space V and thus $V_0 \subset V$ in particular is finite-dimensional in each case. As we consider the case $V = \mathbb{R}^n$ this is sufficient. However, the assumption is not necessary as the proposition for the constants would also follow for sequences $(v_k^*)_{k\in\mathbb{N}}$ so that*

$$\lim_{k\to\infty} \frac{A\,P\,A_0^{-1}\,R\,v_k^*,\,(I-Q)\,A\,P\,A_0^{-1}\,R\,v_k^*}{\|A\,P\,A_0^{-1}\,R\,v_k^*\|\,\|(I-Q)\,A\,P\,A_0^{-1}\,R\,v_k^*\|} = \gamma_{DT}.$$

Now we have given estimations for the condition of $A\,C_{BPX}^{-1}$ and $A\,C_{DT}^{-1}$ in the Euclidean norm. As the estimations all depend only on the constant γ_{DT} and μ_{DT} respectively, and as we have shown in the Corollaries 3.6.6 and 3.6.11 that these are the best possible estimations, we can now compare the two methods. More exactly: As we have exactly calculated (rather than just estimated) the condition with respect to the Euclidean norm we can compare the methods with respect to this characteristic. As the condition is given by $\frac{d_i}{c_i}$, $i = DT, BPX$, we will take a look at the a relations between d_{DT} and d_{BPX}, and between $\frac{1}{c_{DT}}$ and $\frac{1}{c_{BPX}}$. At last, we will calculate a relation for the quotient.

Theorem: 3.6.15. *For the constants c_{DT}, d_{DT}, c_{BPX} and d_{BPX} as defined in this section, it follows that*

$$c_{BPX} \leq c_{DT}, \quad d_{BPX} \leq d_{DT}$$
$$\textit{and} \quad \frac{d_{DT}}{c_{DT}} \leq \frac{d_{BPX}}{c_{BPX}} \quad \Leftrightarrow \quad \gamma_{DT} \leq \sqrt{1/2}.$$

proof. From the Corollaries 3.6.12 and 3.6.7 we obtain

$$d_{BPX} = \max_{\lambda \in \mathbb{R}} \frac{\lambda^2 + 1}{\lambda^2 + 4 + \mu_{\gamma_{DT}}^2 - 2\lambda\mu_{\gamma_{DT}}}$$
$$\text{and} \quad d_{DT} = \max_{\lambda \in \mathbb{R}} \frac{\lambda^2 + 1}{\lambda^2 + 1 + \mu_{\gamma_{DT}}^2 - 2\lambda\mu_{\gamma_{DT}}}.$$

As for a given $\mu_{\gamma_{DT}} \in \mathbb{R}_+$ and for all $\lambda \in \mathbb{R}$, we have

$$\frac{\lambda^2 + 1}{(\lambda^2 - \mu_{\gamma_{DT}})^2 + 4} < \frac{\lambda^2 + 1}{(\lambda^2 - \mu_{\gamma_{DT}})^2 + 1}$$
$$\Leftrightarrow \quad \frac{\lambda^2 + 1}{\lambda^2 + 4 + \mu_{\gamma_{DT}}^2 - 2\lambda\mu_{\gamma_{DT}}} < \frac{\lambda^2 + 1}{\lambda^2 + 1 + \mu_{\gamma_{DT}}^2 - 2\lambda\mu_{\gamma_{DT}}}.$$

It follows that

$$d_{BPX} = \max_{\lambda \in \mathbb{R}} \frac{\lambda^2 + 1}{\lambda^2 + 4 + \mu_{\gamma_{DT}}^2 - 2\lambda\mu_{\gamma_{DT}}} = \frac{(\lambda^*)^2 + 1}{(\lambda^*)^2 + 4 + \mu_{\gamma_{DT}}^2 - 2\lambda^*\mu_{\gamma_{DT}}}$$
$$< \frac{(\lambda^*)^2 + 1}{(\lambda^*)^2 + 1 + \mu_{\gamma_{DT}}^2 - 2\lambda^*\mu_{\gamma_{DT}}} \leq \max_{\lambda \in \mathbb{R}} \frac{\lambda^2 + 1}{\lambda^2 + 1 + \mu_{\gamma_{DT}}^2 - 2\lambda\mu_{\gamma_{DT}}} = d_{DT}.$$

This implies $d_{BPX} \leq d_{DT}$. By the same arguments, we obtain for c_{DT}, c_{BPX} that

$$\begin{aligned} c_{DT} &= \min_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+1+\mu_{\gamma_{DT}}^2-2\lambda\mu_{\gamma_{DT}}} = \frac{(\lambda^*)^2+1}{(\lambda^*)^2+1+\mu_{\gamma_{DT}}^2-2\lambda^*\mu_{\gamma_{DT}}} \\ &> \frac{(\lambda^*)^2+1}{(\lambda^*)^2+4+\mu_{\gamma_{DT}}^2-2\lambda^*\mu_{\gamma_{DT}}} \geq \min_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+4+\mu_{\gamma_{DT}}^2-2\lambda\mu_{\gamma_{DT}}} = c_{BPX}. \end{aligned}$$

From the definition of $\mu_{\gamma_{DT}}$, we obtain

$$\gamma_{DT} \leq \sqrt{1/2} \quad \Leftrightarrow \quad \mu_{\gamma_{DT}}^2 \leq 1.$$

Hence we get the last proposition by

$$\begin{aligned} &\frac{d_{DT}}{c_{DT}} \leq \frac{d_{BPX}}{c_{BPX}} \\ \Leftrightarrow \quad &\frac{2+\mu_{\gamma_{DT}}^2+\mu_{\gamma_{DT}}\sqrt{4+\mu_{\gamma_{DT}}^2}}{2+\mu_{\gamma_{DT}}^2-\mu_{\gamma_{DT}}\sqrt{4+\mu_{\gamma_{DT}}^2}} \leq \frac{5+\mu_{\gamma_{DT}}^2+\sqrt{9+10\mu_{\gamma_{DT}}^2+\mu_{\gamma_{DT}}^4}}{5+\mu_{\gamma_{DT}}^2-\sqrt{9+10\mu_{\gamma_{DT}}^2+\mu_{\gamma_{DT}}^4}} \\ \Leftrightarrow \quad &(5+\mu_{\gamma_{DT}}^2)\mu_{\gamma_{DT}}\sqrt{4+\mu_{\gamma_{DT}}^2} \leq (2+\mu_{\gamma_{DT}}^2)\sqrt{9+10\mu_{\gamma_{DT}}^2+\mu_{\gamma_{DT}}^4} \\ \Leftrightarrow \quad &(5+\mu_{\gamma_{DT}}^2)^2\mu_{\gamma_{DT}}^2(4+\mu_{\gamma_{DT}}^2) \leq (2+\mu_{\gamma_{DT}}^2)^2(9+10\mu_{\gamma_{DT}}^2+\mu_{\gamma_{DT}}^4) \\ &\Leftrightarrow \quad 0 \leq \mu_{\gamma_{DT}}^4+2\mu_{\gamma_{DT}}^2-3 \\ &\Rightarrow \quad \mu_{\gamma_{DT}}^2 = 1. \end{aligned}$$

The proof is completed by the fact that we have $\mu_{\gamma_{DT}} \in \mathbb{R}_+$ for $\gamma_{DT} \in [0,1)$. □

3.7 First Summary

We will close this section by a first summary of our results. We have introduced the preconditioners $C_{BPX}^{-1}, C_{DT}^{-1}$ and C_{2P}^{-1} by using the inverse of A, A_0. In this introduction we have seen that the three operators are all well posed if A, A_0 are non singular. For all preconditioners, we have given an easy basic estimation for the condition of $A\,C^{-1}$ in the Euclidean norm. For the preconditioners C_{BPX}^{-1} and C_{DT}^{-1}, we have seen that we can give much better estimations. In particular, they depend only on the angle γ_{DT} and the value $\mu_{\gamma_{DT}}$ respectively.

Further, we have seen in the Corollaries 3.6.8 and 3.6.13 that the constants d_{BPX}, d_{DT} are increasing in γ_{DT} and c_{BPX}, c_{DT} are decreasing in γ_{DT}. This gives us a first idea to

modify the system that way that we have $\gamma_{DT} = 0$. As the spaces V_0, W are orthogonal to each other, it is easy to see that the condition of $\gamma_{DT} = 0$ is equivalent to the condition that V_0 is invariant with respect to the operator A. We will take a closer look at this aspect in the next chapter. So far we will just consider what would happen in the case of $\gamma_{DT} = 0$.

Corollary: 3.7.1. *Assume thatwe have $\gamma_{DT} = 0$. Then it follows for the constants c_{DT}, d_{DT} and c_{BPX}, d_{BPX} of section 3.6.2 that*

$$\frac{d_{DT}}{c_{DT}} = 1 \quad \textit{and} \quad \frac{d_{BPX}}{c_{BPX}} = 4.$$

proof. The proof follows as $\gamma_{DT} = 0$ implies $\mu_{\gamma_{DT}} = 0$. Hence the proof follows from the results of Theorems 3.6.5 and 3.6.10. □

As the exact calculation for the conditions only depends on the constant γ_{DT}, we can compare the two methods. This is done in Theorem 3.6.15. By the quotients d/c, we see that the BPX-method is better if the angle is bigger. More exactly this is the case if we have $\gamma_{DT} \geq \sqrt{1/2}$. These are the more serious problems. If the angle is small, the solution is quite exactly given by the addition of the solutions of the subspaces. In this case, the DT-method has the lower condition. By the relations for d_{DT}, d_{BPX} and c_{DT}, c_{BPX} as given in Theorem 3.6.15, we can interpret this as follows:

Interpretation of d_{DT}, d_{BPX} : As already mentioned, the constants d_{DT}, d_{BPX} are the more serious problem. These constants exist if and only if the operators $A\,C_{DT}^{-1}$ and $A\,C_{BPX}^{-1}$ respectively are non singular. So for a robust preconditioner it is important that the constant d has an upper bound which is as small as possible. Hence, as we have $d_{BPX} \leq d_{DT}$, we can conclude that the BPX-method is more robust.

From the representation

$$\frac{\|v\|^2}{\|A\,C_i^{-1}\,v\|^2} \leq d_i \quad \text{for all} \quad v \in V$$

$$\text{follows} \quad \lambda_{min}(A\,C_i^{-1}) \geq \frac{1}{d_i}, \qquad i = DT, BPX$$

with

$$\lambda_{min}(A\,C_i^{-1}) := \min\{|\lambda| \in \mathbb{R}_+ \,:\, A\,C_i^{-1}\,v = \lambda\,v \text{ for an } v \in \mathbb{R}^n \backslash \{0\}\}.$$

Now we can see that $\frac{1}{d_i}$ is a lower bound for absolute value of the eigenvalues of $A\,C_i^{-1}$. Thus, a bigger d_i means a lower bound for the eigenvalue with the smallest absolute value.

Interpretation of c_{DT}, c_{BPX} : The constants c_{DT}, c_{BPX} determine how exact the solution can be. In Corollary 3.7.1, we have seen that by the condition $\gamma_{DT} = 0$, the BPX-method is not exact. We obtain this as we have $c_{BPX} = \frac{1}{4}$ in this case. By comparison, we get $c_{DT} = 1$ in this case. Hence, the constant c_i, $i = DT, BPX$ can be seen as a measure for how exact a method can be.

As done for the constants d from the representation

$$\frac{\|v\|^2}{\|A\,C_i^{-1}\,v\|^2} \geq c_i \quad \text{for all} \quad v \in V$$

$$\text{follows} \quad \lambda_{max}(A\,C_i^{-1}) \leq \frac{1}{c_i}\; i = DT, BPX$$

with

$$\lambda_{max}(A\,C_i^{-1}) := \max\{|\lambda| \in \mathbb{R}_+ \,:\, A\,C_i^{-1}\,v = \lambda\,v \text{ for an } v \in \mathbb{R}^n \backslash \{0\}\}.$$

We can see $\frac{1}{c_i}$ as an upper bound for the eigenvalues of $A\,C_i^{-1}$. Thus, a smaller c_i means a lower bound for the biggest eigenvalue.

4 Modification of the BPX and DT Method

In chapter 3 we have introduced the preconditioners $C_{BPX}^{-1}, C_{DT}^{-1}$ and C_{2P}^{-1}. Then we have given quite simple estimations for the condition of $C_i^{-1} A$, $i = BPX, DT, 2P$ in the Eulidian norm. In section 3.6 we have proved a better estimation for the BPX and the DT-method with respect to the same norm. In particular we have seen that we can estimate the condition just by one constant, the given estimations are best and the behaviour of the condition with respect to the constant is quite easy to see.
So in this chapter we will modify the preconditioner. First, only by modifying the prolongation (one sided), then by modifying the prolongation and the restriction (two sided). As it is obvious that the constant γ_{DT} as defined in chapter 3 is zero if and only if the subspace V_0 is invariant with respect to the operator A, the aim will be to modify the prolongation in a way, that holds this invariance. Furthermore we highlight that the restriction and prolongation must not be given by an aggregation method.

As in the last chapter, we will introduce the modification for the two grid methods. Hence we drop the same indices as in the last chapter.

4.1 A one sided modification

First we will try to modify the DT and the BPX-method by a one sided modification. A modification matrix $X \in \mathbb{R}^{n \times n}$ should have the property

$$rk(X\,P) = n_0.$$

We define a modified prolongation $P_X \in \mathbb{R}^{n \times n_0}$ by

$$P_X := X\,P.$$

To get a consistence on the subspace V_0 we define a modified lower dimensional operator $A_{0,X}$ by

$$A_{0,X} := R\,A\,P_X.$$

For the non singularity of $A_{0,X}$ we get by the analogy to Lemma 2.3.5 and Corollary 2.3.6 respectively the following result:

Lemma: 4.1.1. *Let $A \in \mathbb{R}^{n\times n}$ be a non singular matrix. Then it follows that $A_{0,X}$ is non singular if and only if there is no $v_0 \in V_0$ that holds $A\,X\,v_0 \in W = V_0^{\perp\cdot}$*

proof. The proof follows from the same arguments as in Lemma 2.3.5. □

Analogue to γ_{DT} we define the angles $\gamma_{DT,X}$ by

$$\gamma_{DT,X} := \min\Big\{t \in \mathbb{R}_+ \,:\, (A\,P_X\,A_{0,X}^{-1}\,R\,v,\,(I-Q_0)v) \leq t\,\|A\,P_X\,A_{0,X}^{-1}\,R\,v\|\,\|(I-Q_0)v\|,\ \forall v \in V\Big\}. \tag{4.1}$$

So we get a first simple result for the operator $A\,P_X\,A_{0,X}^{-1}\,R\,v$ that we will use for both methods. The result is the simple generalization of Lemma 3.4.1.

Lemma: 4.1.2. *Let $A \in \mathbb{R}^{n\times n}, R \in \mathbb{R}^{n_0\times n}$ be given matrics. Assume that $A_{0,X}$ is non singular. Then the operator*

$$A\,P_X\,A_{0,X}^{-1}\,R : V \to \left\langle A\,P_X\,A_{0,X}^{-1}\,Re_1, \ldots, A\,P_X\,A_{0,X}^{-1}\,R\,e_n\right\rangle$$

is a projection and

$$Q_0\,A\,P_X\,A_{0,X}^{-1}\,R\,v = Q_0\,v \tag{4.2}$$

holds for all $v \in V$.

proof. The calculation

$$(A\,P_X\,A_{0,X}^{-1}\,\underbrace{R)\,(A\,P_X}_{A_{0,X}}\,A_{0,X}^{-1}\,R) = A\,P_X\,A_{0,X}^{-1}\,A_{0,X}\,A_{0,X}^{-1}\,R$$

$$= A\,P_X\,A_{0,X}^{-1}\,R$$

shows that $A\,P_X\,A_{0,X}^{-1}\,R$ is a projection. The equation (4.2) follows from

$$Q_0\,A\,P_X\,A_{0,X}^{-1}\,R\,v = P\,S\,\underbrace{R\,A\,P_X}_{A_{0,X}}\,A_{0,X}^{-1}\,R\,v = P\,S\,R\,v = Q_0\,v.$$

□

By Lemma 4.1.2 we can conclude that the direction of the projection $A\,P_X\,A_{0,X}^{-1}\,R$ is orthogonal to V_0. That means that for an arbitrary $v \in V$ with $v = v_0 + w,\ w \in W,\ v_0 \in V_0$ it is

$$Q_0\,v = v_0, \quad (I - Q_0)\,v = w$$

$$\text{and} \quad A\,P_X\,A_{0,X}^{-1}\,R\,v = A\,P_X\,A_{0,X}^{-1}\,R\,v_0 = v_0 + w_1, \quad w_1 \in W.$$

But the image space of $A\,P_X\,A_{0,X}^{-1}\,R$ is also in this case in general not given by V_0. We will take a look at a condition for this. It is:

Lemma: 4.1.3. *For given non singular $A \in \mathbb{R}^{n\times n}$, $A_{0,X} \in \mathbb{R}^{n_0\times n_0}$ the following three statements are equivalent:*

1. *It holds $\gamma_{DT,X} = 0$.*
2. *V_0 is invariant with respect to $A\,X$.*
3. *It holds $(A\,P_X\,A_{0,X}^{-1}\,R)\,(V) = V_0$.*

proof. To prove the equivalences we will show three implications:

$1 \Rightarrow 2:$ For an arbitrarily given $v_0 \in V_0$ there is an $w_1 \in W$ that holds for all $w \in W$

$$A\,P_X\,A_{0,X}^{-1}\,R\,(v_0 + w) = v_0 + w_1.$$

Hence we obtain for $v = v_0 + w_1$ by $\gamma_{DT,X} = 0$

$$\begin{aligned}(A\,P_X\,A_{0,X}^{-1}\,R\,v,\ (I - Q)\,v) &\le \gamma_{DT,X}\,\|A\,P_X\,A_{0,X}^{-1}\,R\,v\|\,\|(I - Q)\,v\| \\ \Leftrightarrow \quad (v_0 + w_1,\ w_1) &\le 0 \\ \Leftrightarrow \quad (w_1,\ w_1) &\le 0 \\ \Rightarrow \quad w_1 &= 0.\end{aligned}$$

This implies $A\,P_X\,A_{0,X}^{-1}\,R\,(v_0 + w) = v_0 \in V_0$. As we have

$$V_0 = Im((P\,A_{0,X}^{-1}\,R)(V_0))$$

it follows that V_0 is invariant with respect to $A\,X$.

$2 \Rightarrow 3$: As it holds $P\, A_{0,X}^{-1}\, R\, v \in V_0$ for all $v \in V$ we obtain by the invariance of V_0 with respect to $A\, X$ the inclusion $A\, P_X\, A_{0,X}^{-1}\, R\,(V) \subset V_0$. As it is

$$R\, A\, P_X\, A_{0,X}^{-1}\, R = R$$

we obtain

$$rk(A\, P_X\, A_{0,X}^{-1}\, R) \geq rk(R\, A\, P_X\, A_{0,X}^{-1}\, R) = rk(R) = n_0$$

This implies that $A\, P_X\, A_{0,X}^{-1}\, R : V \to V_0$ is surjective.

$3 \Rightarrow 1$: As we obtain $A\, P_X\, A_{0,X}^{-1}\, R \in V_0 = W^{\perp}$ for all $v \in V$. It follows

$$(A\, P_X\, A_{0,X}^{-1}\, R\, v,\, w) = 0 \quad \forall v \in V,\, \forall w \in W.$$

This implies the proposition.

□

The information given by the last two lemmata is similar to the conlusions we get in section 3.4 for the unmodified method. In particular we have seen in the Lemmata 3.4.1 and 4.1.2 analogoues propositions for the operators $A\, P\, A_0^{-1}\, R$ and $A\, P_X\, A_{0,X}^{-1}\, R$. This result will be the main aspect to get similar results for the modified method. The effect of the modification is illustrated in Figure 4.1. Based on the mentioned analogy the Figure 4.1 is the modification of Figure 3.1 at page 69.

4.1.1 The DT-method

For a non singular matrix $A \in \mathbb{R}^{n \times n}$ and a non singular $A_{0,X} \in \mathbb{R}^{n_0 \times n_0}$ we define the modified preconditioner $C_{DT,X}^{-1}$ by

$$C_{DT,X}^{-1} := A^{-1}(I - Q_0) + P_X\, A_{0,X}^{-1}\, R. \tag{4.3}$$

First we will show that the operator $C_{DT,X}^{-1}$ is non singular. This follows in the next lemma based on the same assumption and arguments as used in Lemma 3.3.1 for the unmodified operator.

Lemma: 4.1.4. *Let $A \in \mathbb{R}^{n \times n}, A_0 \in \mathbb{R}^{n_0 \times n_0}$ be non singular. Then the matrix*

$$A\, C_{DT,X}^{-1}$$

is non singular.

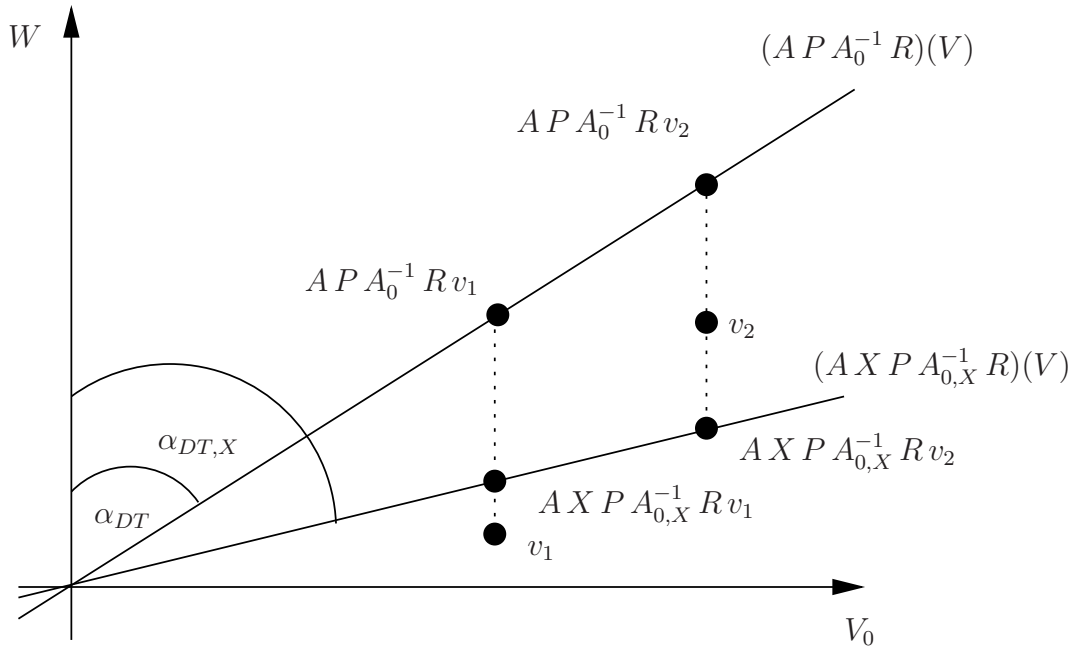


Figure 4.1: Effect of the modified projection $A\,X\,P\,A_{0,X}^{-1}\,R$

proof. Suppose that $A\,C_{DT,X}^{-1}$ is singular. Then it must exist an $v \in V\backslash\{0\}$ with

$$\begin{aligned} & 0 = A\,C_{DT,X}^{-1} v \\ \Leftrightarrow \quad & 0 = (I - Q_0)\,v + A\,P_X\,A_{0,X}^{-1}\,R\,v \\ \Leftrightarrow \quad & -(I - Q_0)\,v = A\,P_X\,A_{0,X}^{-1}\,R\,v \\ \Rightarrow \quad & -R\,(I - Q_0)\,v = \underbrace{R\,A\,P_X}_{=A_{0,X}}\,A_{0,X}^{-1}\,R\,v \\ \Leftrightarrow \quad & 0 = R\,v. \end{aligned}$$

So the for the given $v \in V$ we obtain $R\,v = 0$. But in the case of $R\,v = 0$ we obtain

$$0 = A\,C_{DT,X}^{-1} v = (I - Q_0)\,v + A\,P_X\,A_{0,X}^{-1}\,R\,v = v.$$

This is in contradiction to the assumption. □

As shown in Corollary 3.3.3 in the unmodified situation the proof of the non singularity of $A\,C_{DT,X}^{-1}$ implies $\gamma_{DT,X} < 1$.
To get estimations for the condition of $A\,C_{DT,X}^{-1}$ that can be compared with the estimations of section 3.6 we need similar results. So for the modified method we will give

an estimation just depending on the constant $\gamma_{DT,X}$. So we define for this chapter

$$\mu_{\gamma_{DT,X}} = \frac{\gamma_{DT,X}}{\sqrt{1-\gamma_{DT,X}^2}}$$

Theorem: 4.1.5. *Let $A \in \mathbb{R}^{n\times n}$, $A_{0,X} \in \mathbb{R}^{n_0\times n_0}$ be non singular and $C_{DT,X}^{-1}$ as defined in (4.3). Then the inequalities*

$$c_{DT,X}\|A\,C_{DT,X}^{-1}\,v\|^2 \le \|v\|^2 \le d_{DT,X}\|A\,C_{DT,X}^{-1}\,v\|^2 \tag{4.4}$$

hold for all $v \in V$ with

$$c_{DT,X} := \frac{2+\mu_{\gamma_{DT,X}}^2 - \mu_{\gamma_{DT,X}}\sqrt{4+\mu_{\gamma_{DT,X}}^2}}{2} \tag{4.5}$$

$$\text{and} \quad d_{DT,X} := \frac{2+\mu_{\gamma_{DT,X}}^2 + \mu_{\gamma_{DT,X}}\sqrt{4+\mu_{\gamma_{DT,X}}^2}}{2}.$$

proof. The proof follows exactly the same arguments as the proof of Theorem 3.6.5. We use again Lemma 3.6.2. This time we set

$$B = A\,P_X\,A_{0,X}^{-1}\,R \quad \text{instead of} \quad B = A\,P\,A_0^{-1}\,R$$

as done in the proof of Theorem 3.6.5. Then the proof follows as

$$Q_0\,A\,P_X\,A_{0,X}^{-1}\,R\,v = Q_0 v$$

holds as $Q_0\,A\,P\,A_0^{-1}\,R\,v = Q_0\,v$ in the proof of Theorem 3.6.5. □

As the constants $c_{DT,X}, d_{DT,X}$ that determine the condition of $A\,C_{DT,X}^{-1}$ have the same structure as c_{DT}, d_{DT} in Theorem 3.6.5 it is obvious that we obtain for the constants $c_{DT,X}, d_{DT,X}$ the same characteristics as for c_{DT}, d_{DT}. These are summarized in the next proposition.

Proposition: 4.1.6. *Let $c_{DT,X}, d_{DT,X}$ be as given in Theorem 4.1.5 then it follows:*

1. *$c_{DT,X} \le 1 \le d_{DT,X}$ and it is $c_{DT,X} = 1 = d_{DT,X}$ if and only if it is $\gamma_{DT,X} = 0$.*
2. *It is*

$$\frac{d}{d\gamma_{DT,X}}[c_{DT,X}] < 0 \quad \text{and} \quad \frac{d}{d\gamma_{DT,X}}[d_{DT,X}] > 0.$$

3. *There is no $c^* > c_{DT,X}$ and no $d^* < d_{DT,X}$ that hold for all $v \in V$*

$$c^* \|C_{DT,X}^{-1} A\, v\|^2 \leq \|v\|^2 \leq d^* \|C_{DT,X}^{-1} A\, v\|^2.$$

4. *The constants $c_{DT,X}, d_{DT,X}$ are given by*

$$c_{DT,X} = \min_{\lambda \in \mathbb{R}} \frac{\lambda^2 + 1}{\lambda^2 + 1 + \mu_{\gamma_{DT,X}}^2 + 2\lambda\mu_{\gamma_{DT,X}}} = \min_{\lambda \in \mathbb{R},\, \mu \in [0, \mu_{\gamma_{DT,X}}]} \frac{\lambda^2 + 1}{\lambda^2 + 1 + \mu^2 + 2\lambda\mu}$$

$$d_{DT,X} = \max_{\lambda \in \mathbb{R}} \frac{\lambda^2 + 1}{\lambda^2 + 1 + \mu_{\gamma_{DT,X}}^2 - 2\lambda\mu_{\gamma_{DT,X}}} = \max_{\lambda \in \mathbb{R},\, \mu \in [0, \mu_{\gamma_{DT,X}}]} \frac{\lambda^2 + 1}{\lambda^2 + 1 + \mu^2 - 2\lambda\mu}.$$

proof. The proof follows the same arguments as the proofs of Remark 3.6.4 and the Corollaries 3.6.6, 3.6.7 and 3.6.8. □

Using these results we can compare the modified method with respect to different modification matrices X_1, X_2 with each other.

Proposition: 4.1.7. *Let $A \in \mathbb{R}^{n \times n}$ be non singular and $X_1, X_2 \in \mathbb{R}^{n \times n}$ two modifications so that $A_{0,X_1}, A_{0,X_2} \in \mathbb{R}^{n_0 \times n_0}$ are non singular. Assume further that it is*

$$\gamma_{DT,X_1} := \min \Big\{ t \in \mathbb{R}_+ \,:\, (A\, P_{X_1}\, A_{0,X_1}^{-1}\, R\, v,\, (I - Q_0)v) \leq t\, \|A\, P_{X_1}\, A_{0,X_1}^{-1}\, R\, v\|\, \|(I - Q_0)v\|,\ \forall v \in V \Big\}$$

$$\gamma_{DT,X_2} := \min \Big\{ t \in \mathbb{R}_+ \,:\, (A\, P_{X_2}\, A_{0,X_2}^{-1}\, R\, v,\, (I - Q_0)v) \leq t\, \|A\, P_{X_2}\, A_{0,X_2}^{-1}\, R\, v\|\, \|(I - Q_0)v\|,\ \forall v \in V \Big\}$$

$$\textit{and} \quad \gamma_{DT,X_1} < \gamma_{DT,X_2}$$

then it holds

$$c_{DT,X_1} > c_{DT,X_2} \quad \textit{and} \quad d_{DT,X_1} < d_{DT,X_2}.$$

proof. The proposition is immediately followed by the second proposition of 4.1.6. □

Looking at the proposition 4.1.7 it is obvious that the aim of a modification should be a low angle $\gamma_{DT,X}$. At the same time there are two restrictions for practical causes:

1. As we will in general use iterative methods instead of to determine $A_{0,X}^{-1}$ the matrix X should induce for $A_{0,X}$ good characteristics for common iterative methods (cf. chapter 9).

2. The effort to determine X, P_X and to calculate $X\,v_0$ for, a $v_0 \in V_0$ or $P_X\widetilde{v}_0$ for a $\widetilde{v}_0 \in \widetilde{V}_0$ should be limited.

We have assumed for the modification matrix only $X \in \mathbb{R}^{n\times n}$ with $rk(X\,P) = n_0$. In particular the *DT*-method do only use $X\,v_0$ for $v_0 \in V_0$. So it is obvious that the modification matrix is not unique because for an arbitrary $w \in W$ the result of $X\,w$ does not matter. We will see this in an explicit example (cf section 5.1.1). We will conclude this section with the example $X = A^{-1}$. In this case we obtain:

$$\begin{aligned}(A\,X\,P\,A_{0,X}^{-1}\,R\,v,\,(I-Q)\,v) &= (P\,A_{0,X}^{-1}\,R\,v,\,(I-Q)\,v)\\ &= (A_{0,X}^{-1}\,R\,v,\,\underbrace{R\,(I-Q)}_{=0}\,v) = 0.\end{aligned}$$

Hence it is $\gamma_{DT,X} = 0$ in this case. Furthermore it is quite obvious that V_0 is invariant with respect to $A\,X = id$. We have shown above that this implies the invariance, too. This implies $c_{DT,X} = d_{DT,X} = 1$. If we take a closer look at this example we get

$$A_{0,X} = R\,A\,X\,P = R\,P \quad \Rightarrow \quad A_{0,X}^{-1} = S.$$

This implies

$$\begin{aligned}A\,C_{DT,X}^{-1} &= A\left(A^{-1}\,(I-Q) + A^{-1}\,P\,A_{0,X}^{-1}\,R\right)\\ &= (I-Q) + P\,S\,R = I.\end{aligned}$$

So the preconditioner is in this case the exact inverse of A.

4.1.2 The BPX-method

Now we will consider the effect on the BPX-method if we modify this method in the same way as we have done it for the DT-method. For a non singular $A \in \mathbb{R}^{n\times n}$ and a non singular $A_{0,X} \in \mathbb{R}^{n_0\times n_0}$ we define the one sided modified BPX-preconditioner $C_{BPX,X}^{-1}$ by

$$C_{BPX,X}^{-1} = A^{-1} + P_X\,A_{0,X}^{-1}\,R. \tag{4.6}$$

Again we will first show that the operator $C_{BPX,X}^{-1}$ is non singular. The next lemma is the modified version of Lemma 3.2.1.

Lemma: 4.1.8. *Let $A \in \mathbb{R}^{n\times n}, A_0 \in \mathbb{R}^{n_0\times n_0}$ be non singular. Then the matrix*

$$A\, C_{BPX,X}^{-1}$$

is non singular.

proof. Suppose that $A\, C_{BPX,X}^{-1}$ is singular. Then there is an $v \in V\backslash\{0\}$ with

$$\begin{aligned} & 0 = A\, C_{BPX,X}^{-1} v \\ \Leftrightarrow \quad & 0 = v + A\, P_X\, A_{0,X}^{-1}\, R\, v \\ \Leftrightarrow \quad & -v = A\, P_X\, A_{0,X}^{-1}\, R\, v \\ \Rightarrow \quad & -R\, v = \underbrace{R\, A\, P_X}_{=A_{0,X}}\, A_{0,X}^{-1}\, R\, v \\ \Leftrightarrow \quad & -R\, v = R\, v. \end{aligned}$$

So the for the given $v \in V$ we obtain $R\, v = 0$. But in the case of $R\, v = 0$ we obtain

$$0 = A\, C_{BPX,X}^{-1} v = v + A\, P_X\, A_{0,X}^{-1}\, R\, v = v.$$

This gives the contradiction. □

So we obtain also for the modified BPX-method a central result that estimates the condition of $A\, C_{BPX,X}^{-1}$ in the Euclidean norm just by the angle $\gamma_{DT,X}$. So the result is the generalization of Theorem 3.6.10.

Theorem: 4.1.9. *Let $A \in \mathbb{R}^{n\times n}$, $A_{0,X} \in \mathbb{R}^{n_0\times n_0}$ be non singular and $C_{BPX,X}^{-1}$ as defined in (4.6). Then the inequalities*

$$c_{BPX,X}\|A\, C_{BPX,X}^{-1}\, v\|^2 \le \|v\|^2 \le d_{BPX,X}\|A\, C_{BPX,X}^{-1}\, v\|^2 \tag{4.7}$$

hold for all $v \in V$ with

$$c_{BPX,X} := \frac{5 + \mu_{\gamma_{DT,X}}^2 - \sqrt{9 + 10\mu_{\gamma_{DT,X}}^2 + \mu_{\gamma_{DT,X}}^4}}{8} \tag{4.8}$$

$$\text{and} \quad d_{BPX,X} := \frac{5 + \mu_{\gamma_{DT,X}}^2 + \sqrt{9 + 10\mu_{\gamma_{DT,X}}^2 + \mu_{\gamma_{DT,X}}^4}}{8}.$$

proof. The proof follows exactly the same arguments as the proof of Theorem 3.6.10. We use again Lemma 3.6.2 and set

$$B = A\,P_X\,A_{0,X}^{-1}\,R \quad \text{instead of} \quad B = A\,P\,A_0^{-1}\,R$$

as done in the proof of Theorem 3.6.10. □

As for the modification of the DT-method, the constants $c_{BPX,X}, d_{BPX,X}$ that determine the condition of $A\,C_{BPX,X}^{-1}$ follow the same structure as c_{BPX}, d_{BPX} in Theorem 3.6.10. So it is obvious that we obtain for the constants $c_{BPX,X}, d_{BPX,X}$ the same properties as for c_{BPX}, d_{BPX}. These are summarized in the next proposition.

Proposition: 4.1.10. *Let $c_{BPX,X}, d_{BPX,X}$ be as given in Theorem 4.1.9 then it follows:*

1. *$c_{BPX,X} < 1 \leq d_{BPX,X}$ and it is $d_{BPX,X} = 1$ if and only if it is $\gamma_{DT,X} = 0$.*

2. *It is*

$$\frac{d}{d\gamma_{DT,X}}[c_{BPX,X}] < 0 \quad \textit{and} \quad \frac{d}{d\gamma_{DT,X}}[d_{BPX,X}] > 0.$$

3. *There is no $c^* > c_{BPX,X}$ and no $d^* < d_{BPX,X}$ that hold for all $v \in V$*

$$c^*\|C_{BPX,X}^{-1}\,A\,v\|^2 \leq \|v\|^2 \leq d^*\|C_{BPX,X}^{-1}\,A\,v\|^2.$$

4. *The constants $c_{BPX,X}, d_{BPX,X}$ are given by*

$$c_{BPX,X} = \min_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+4+\mu_{\gamma_{DT,X}}^2+2\lambda\mu_{\gamma_{DT,X}}} = \min_{\lambda\in\mathbb{R},\,\mu\in[0,\mu_{\gamma_{DT,X}}]} \frac{\lambda^2+1}{\lambda^2+4+\mu^2+2\lambda\mu}$$

$$d_{BPX,X} = \max_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+4+\mu_{\gamma_{DT,X}}^2-2\lambda\mu_{\gamma_{DT,X}}} = \max_{\lambda\in\mathbb{R},\,\mu\in[0,\mu_{\gamma_{DT,X}}]} \frac{\lambda^2+1}{\lambda^2+4+\mu^2+2\lambda\mu}.$$

proof. The proof follows the same arguments as the proofs of Remark 3.6.9 and the Corollaries 3.6.11, 3.6.12 and 3.6.13. □

So as for the DT-method we can also compare the modified BPX-method for two different modifications X_1, X_2 with each other.

Proposition: 4.1.11. *Let $A \in \mathbb{R}^{n\times n}$ be non singular and $X_1, X_2 \in \mathbb{R}^{n\times n}$ two modifications so that $A_{0,X_1}, A_{0,X_2} \in \mathbb{R}^{n_0\times n_0}$ are non singular. Assume further that it is*

$$\gamma_{DT,X_1} := \min\Big\{t \in \mathbb{R}_+ \,:\, (A\,P_{X_1}\,A_{0,X_1}^{-1}\,R\,v,\,(I-Q_0)v) \\ \leq t\,\|A\,P_{X_1}\,A_{0,X_1}^{-1}\,R\,v\|\,\|(I-Q_0)v\|,\ \forall v \in V\Big\}$$

$$\gamma_{DT,X_2} := \min\Big\{t \in \mathbb{R}_+ \,:\, (A\,P_{X_2}\,A_{0,X_2}^{-1}\,R\,v,\,(I-Q_0)v) \\ \leq t\,\|A\,P_{X_2}\,A_{0,X_2}^{-1}\,R\,v\|\,\|(I-Q_0)v\|,\ \forall v \in V\Big\}$$

$$\textit{and} \quad \gamma_{DT,X_1} < \gamma_{DT,X_2}$$

then we have

$$c_{BPX,X_1} > c_{BPX,X_2} \quad \textit{and} \quad d_{BPX,X_1} < d_{BPX,X_2}.$$

proof. The proposition is immediately followed by the second proposition of 4.1.10. □

So as for the DT-method we can conclude that a lower angle $\gamma_{DT,X}$ implies a lower condition of $A\,C_{BPX,X}^{-1}$. And as we have done for the DT-method we will conclude this section with the example $X = A^{-1}$. We have seen in section 4.1.1 that we obtain in this case

$$\gamma_{DT,X} = 0 \quad \Rightarrow \quad \mu_{\gamma_{DT,X}} = 0 \quad \text{and} \quad A_{0,X} = S^{-1}.$$

So this implies for the modified BPX-method:

$$A\,C_{BPX,X}^{-1} = A\left(A^{-1} + A^{-1}\,P\,A_{0,X}^{-1}\,R\right) = I + P\,S\,R = I + Q_0.$$

It is therefore obvious that we get in this case

$$(A\,C_{BPX,X}^{-1}\,v,\ A\,C_{BPX,X}^{-1}\,v) = ((I+Q_0)\,v,\,v)$$
$$\Rightarrow \quad (v,\,v) \leq ((I+Q_0)\,v,\,v) \leq 2(v,\,v).$$

Hence the BPX-method is not exact in this case. But the smallest eigenvalue is given by $\lambda_{min} = 1$ in this case.

4.1.3 Summary

So we can summarize the results of section 4.1 as follows: As its possible to set $X = I_n$ for the modification matrix, it is obvious that the modified preconditioners are a generalization of the non modified ones. Furthermore we have seen that based on the fact that

$$A\,P_X\,A_{0,X}^{-1}\,R$$

is a projection and $Q_0\,A\,P_X\,A_{0,X}^{-1}\,R\,v = Q_0\,v$ holds for all $v \in V$ we get the same results as for the unmodified method. The methods differ with regard to the assumption. For a given matrix A and a given prolongation P (and therewith the structure of the subspace V_0) the matrices A_0^{-1} and $A_{0,X}^{-1}$ respectively are well-posed if and only if the operators A_0 and $A_{0,X}$ respectively are non singular. And in both cases this implies that the preconditioners are well posed too. We have pointed out that the conditions for this are that there is no $v_0 \in V_0$ that holds $A\,v_0 \in W$, and $A\,X\,v_0 \in W$ respectively. So it depends on the modification whether the preconditioners are well posed.

If we compare the modified preconditioners $C_{DT,X}^{-1}$ and $C_{BPX,X}^{-1}$ then we can do this in the same way as we have done in the unmodified situation in Theorem 3.6.15. So it is obvious that we obtain

$$\frac{d_{DT,X}}{c_{DT,X}} \leq \frac{d_{BPX,X}}{c_{BPX,X}} \quad \Leftrightarrow \quad \gamma_{DT,X} \leq \sqrt{1/2}.$$

If we compare the modified methods with the unmodified methods, then the result is illustrated in Figure 4.1 at page 109. As already mentioned we can interpret the unmodified method as a modification with $X = I$. Hence the modified methods can be compared with the unmodified if we use the results of Propositions 4.1.7 and 4.1.11. This points out that the modification makes the method better if and only if we have

$$\gamma_{DT,X} < \gamma_{DT}.$$

However, there are some problems for the modification, too. First, as already mentioned, the problem of the effort concerns the preconditioners. For practical issues we will not determine X but P_X. Hence the number of multiplication in a iterative solution method remains the same. But there can be the problem of a fill in for $P_X, A_{0,X}$. And of course this raises the effort per multiplication.

Another problem is that if A is a symmetric matrix then this is also true for the coarse grid operator A_0. In general this is not true for the modified operator $A_{0,X}$. If we want to use the preconditioner as a symmetric one this is a problem. To solve this problem we will modify the preconditioner in a symmetric way. We will show this in the next section.

4.2 A two sided modification

As mentioned before, the modification of the preconditioner by an operator X has the side effect that for a symmetric operator A the operator $A_{0,X}$ is in general not symmetric. In particular the hole operator $C^{-1}_{BPX,X}$ is no more symmetric. The aim of this section is therefore to keep for symmetric operators A the coarse grid operators and the modified operator C^{-1}_{BPX} symmetric. So we will concentrate on symmetric operators A.

Similarly to the one sided modification, we define for a non singular operator $A \in \mathbb{R}^{n\times n}$ and a modification $X \in \mathbb{R}^{n\times n}$ with $rk(X\,P) = n_0$ the modified prolongation $P_X \in \mathbb{R}^{n\times n_0}$ and restriction $R_X \in \mathbb{R}^{n_0\times n}$ as follows

$$P_X := X\,P \quad \text{and} \quad R_X := (P_X)^T.$$

Furthermore, we define the coarse grid operator $A_{0,XX} \in \mathbb{R}^{n\times n}$ as follows

$$A_{0,XX} := R_X\,A\,P_X.$$

Then we define the operators $Q_{0,X} \in \mathbb{R}^{n\times n}$ and $S_X \in \mathbb{R}^{n_0\times n_0}$ as follows

$$S_X := (R_X\,P_X)^{-1} \quad \text{and} \quad Q_{0,X} := P_X\,S_X\,R_X.$$

Based on this definition for S_X we can highlight two important characteristics that we have also used for S in the unmodified, respectively one sided modified situation.

Remark: 4.2.1. *Based on the definitions as given above it follows that S_X is symmetric and positive definite.*

proof. As it is

$$(S_X^{-1})^T = (R_X\,P_X)^T = P_X^T\,R_X^T = R_X\,P_X = S_X^{-1}$$

it holds that S_X^{-1} is symmetric. As we have for an arbitrary $\widetilde{v}_0 \in \mathbb{R}^{n_0}$

$$(\widetilde{v}_0,\, S_X^{-1}\,\widetilde{v}_0) = (P_X\,\widetilde{v}_0,\, P_X\,\widetilde{v}_0) = \|P_X\,\widetilde{v}_0\|^2 \geq 0.$$

Based on the condition $rk(P_X) = n_0$ it follows $\|P_X\,\widetilde{v}_0\|^2 = 0$ if and only if it is $\widetilde{v}_0 = 0$. So it is S_X^{-1} s.p.d. and hence also S_X. □

Furthermore we define in analogy to the spaces V_0, W the vector spaces $V_{0,X}$ and W_X as follows

$$V_{0,X} := Im(Q_{0,X}(\mathbb{R}^n))$$

$$W_X := Im((I - Q_{0,X})(\mathbb{R}^n)).$$

Based on these definitions we get the following basics for the operators and vector spaces:

Lemma: 4.2.2. *Based on the definitions of this section it holds:*

1. *$Q_{0,X} : V \to V_{0,X}$ and $(I - Q_{0,X}) : V \to W_X$ are orthogonal projections with respect to the inner product $(.,.)$.*
2. *For a given $v \in V$ the following three characteristics are equivalent:*
   a) It is $Q_{0,X}\,v = 0$.
   b) It is $R_X\,v = 0$.
   c) It is $v \in W_X$.
3. *For a non singular matrix $A \in \mathbb{R}^{n\times n}$ the matrix $A_{0,XX} \in \mathbb{R}^{n_0\times n_0}$ is non singular if and only if there is no $v_{0,X} \in V_{0,X}$ with $A\,v_{0,X} \in W_X$.*
4. *If A is s.p.d. then this holds also for $A_{0,XX}$. In particular $A_{0,XX}$ is in this case non singular.*
5. *If A is real positive then $A_{0,XX}$ is also real positive.*

proof. 1. Based on the calculation

$$Q_{0,X}^2 = (P_X\,S_X\,\underbrace{R_X)(P_X\,S_X}_{=I}\,R_X) = P_X\,S_X\,R_X = Q_{0,X}$$

it follows that $Q_{0,X}$ is a projection. Based on the symmetry

$$Q_{0,X}^T = (P_X\,S_X\,R_X)^T = R_X^T\,S_X\,P_X = P_X\,S_X\,R_X$$

it follows that the projection is orthogonal with respect to $(.,.)$. The proposition of the image space follows the definition of $V_{0,X}$.
The proposition for $I - Q_{0,X}$ follows the same arguments.

2. We prove three implications:

a) $\Rightarrow$ b) Based on the definition of $Q_{0,X}$ as $Q_{0,X} = P_X\, S_X\, R_X$ it follows from the non singularity of S_X and the assumption $rk(P_X) = n_0$ that $(P_X\, S_X) \in \mathbb{R}^{n \times n_0}$ has rank n_0. Therewith $Q_{0,X}\, v = 0$ implies $R_X\, v = 0$.

b) $\Rightarrow$ c) If it is $R_X\, v = 0$ then it follows

$$(I - Q_{0,X})\, v = v - P_X\, S_X\, R_X\, v = v.$$

This implies $v \in W_X$.

c) $\Rightarrow$ a) As $(I - Q_{0,X}) : V \to W_X$ is a projection, it follows for $v \in W_X$

$$(I - Q_{0,X})\, v = v \quad \Rightarrow \quad Q_{0,X}\, v = 0.$$

3. We obtain that $A_{0,XX}$ is singular if and only if there is a $\widetilde{v}_0^* \in \mathbb{R}^{n_0} \backslash \{0\}$ with

$$A_{0,XX}\, \widetilde{v}_0^* = R_X\, A\, P_X\, \widetilde{v}_0^* = 0.$$

Based on the definition of $V_{0,X}$ and the assumption $rk(P_X) = n_0$ we get $P_X\, \widetilde{v}_0 \neq 0$ for all $\widetilde{v}_0 \in \mathbb{R}^{n_0} \backslash \{0\}$. This implies $A\, P_X\, \widetilde{v}_0 \neq 0$ for all $\widetilde{v}_0 \in \mathbb{R}^{n_0} \backslash \{0\}$. Furthermore, it is $P_X\, \widetilde{v}_0 \in V_{0,X}$ based on the definition of $V_{0,X}$. As we have $ker(R_X) = W_X$ it follows

$$A_{0,XX}\, \widetilde{v}_0^* = 0 \quad \Leftrightarrow \quad R_X\, A\, (P_X\, \widetilde{v}_0^*) = 0 \quad \Leftrightarrow \quad A\, (P_X\, \widetilde{v}_0^*) \in W_X.$$

This proves the proposition.

4. If A is s.p.d. then we obtain that $A_{0,XX}$ is symmetric based on

$$A_{0,XX}^T = (R_X\, A\, P_X)^T = P_X^T\, A^T\, R_X^T = R_X\, A\, P_X = A_{0,XX}.$$

And we obtain that $A_{0,XX}$ is positive definite as follows

$$(A_{0,XX}\, \widetilde{v}_0,\, \widetilde{v}_0) = (A\, P_X\, \widetilde{v}_0,\, P_X\, \widetilde{v}_0) = \|P_X\, \widetilde{v}_0\|_A^2 \geq 0.$$

From the assumption $rk(P_X) = n_0$ follows $P_X\, \widetilde{v}_0 \neq 0$ for $\widetilde{v}_0 \neq 0$ and hence $\|P_X\, \widetilde{v}_0\|_A^2 > 0$ for $\widetilde{v}_0 \neq 0$.

5. The fifth assertion of this Lemma follows immediately from the proof of the fourth assertion.

□

So we define an angle $\gamma_{DT,XX} \in R_+$ as follows

$$\gamma_{DT,XX} := \min\Big\{t \in \mathbb{R}_+ \,:\, (A\,P_X\,A_{0,XX}^{-1}\,R_X\,v,\,(I-Q_{0,X})v) \\ \leq t\,\|A\,P_X\,A_{0,XX}^{-1}\,R_X\,v\|\,\|(I-Q_{0,X})v\|,\ \forall v \in V\Big\}. \tag{4.9}$$

Therewith we define for $\gamma_{DT,XX} < 1$ the constant $\mu_{\gamma_{DT,XX}}$ as

$$\mu_{\gamma_{DT,XX}} := \frac{\gamma_{DT,XX}}{\sqrt{1-\gamma_{DT,XX}^2}}.$$

Hence we obtain two results that are similar to the Lemmata 4.1.2 and 4.1.3.

Lemma: 4.2.3. *For a non singular $A \in \mathbb{R}^{n\times n}$ the operator*

$$A\,P_X\,A_{0,XX}^{-1}\,R_X : V \to \left\langle A\,P_X\,A_{0,XX}^{-1}\,R_X e_1^1, \ldots, A\,P_X\,A_{0,X}^{-1}\,R_X\,e_n^1\right\rangle \\ = \left\langle A\,P_X\,e_1^0, \ldots, A\,P_X\,e_{n_0}^0\right\rangle$$

is a projection and it holds for all $v \in V$

$$Q_{0,X}\,A\,P_X\,A_{0,XX}^{-1}\,R_X\,v = Q_{0,X}\,v \tag{4.10}$$

proof. The calculation

$$(A\,P_X\,A_{0,XX}^{-1}\,\underbrace{R_X)\,(A\,P_X}_{=A_{0,XX}}\,A_{0,XX}^{-1}\,R_X) = A\,P_X\,A_{0,XX}^{-1}\,A_{0,XX}\,A_{0,XX}^{-1}\,R_X \\ = A\,P_X\,A_{0,XX}^{-1}\,R_X$$

shows that $A\,P_X\,A_{0,XX}^{-1}\,R_X$ is a projection. The equality of the two spaces is a result of the non singularity of $A_{0,XX}$. Hence the matrix $A_{0,XX}^{-1}\,R_X \in \mathbb{R}^{n_0\times n}$ has rank n_0. Therewith $\{e_0^1, \ldots, e_{n_0}^0\}$ is a basis of $Im(A_{0,XX}\,R_X\,(\mathbb{R}^n))$. The equation (4.10) follows from

$$Q_{0,X}\,A\,P_X\,A_{0,XX}^{-1}\,R_X\,v = P_X\,S_X\,\underbrace{R_X\,A\,P_X}_{A_{0,XX}}\,A_{0,XX}^{-1}\,R_X\,v = P_X\,S_X\,R_X\,v = Q_{0,X}\,v.$$

□

Lemma: 4.2.4. *For non singular $A \in \mathbb{R}^{n\times n}$, $A_{0,XX} \in \mathbb{R}^{n_0\times n_0}$ the following three characteristics are equivalent:*

1. *It holds* $\gamma_{DT,XX} = 0$.

2. $V_{0,X}$ *is invariant with respect to* A.

3. *It holds* $(A\,P_X\,A_{0,XX}^{-1}\,R_X)\,(V) = V_{0,X}$.

proof. We show three implications:

$1 \Rightarrow 2$: For an arbitrarily given $v_{0,X} \in V_{0,X}$ and all $w_X \in W_X$ it is

$$A\,P_X\,A_{0,XX}^{-1}\,R_X\,(v_{0,X} + w_X) = v_{0,X} + w_{1,X}$$

with $w_{1,X} \in W_X$. Hence we obtain for $v = v_{0,X} + w_{1,X}$ by $\gamma_{DT,XX} = 0$

$$\begin{aligned}(A\,P_X\,A_{0,XX}^{-1}\,R_X\,v,\,(I - Q_{0,X})\,v) &\leq \gamma_{DT,XX}\,\|A\,P_X\,A_{0,XX}^{-1}\,R\,v\|\,\|(I - Q_{0,X})\,v\| \\ \Leftrightarrow \quad (v_{0,X} + w_{1,X},\,w_{1,X}) &\leq 0 \\ \Leftrightarrow \quad (w_{1,X},\,w_{1,X}) &\leq 0 \\ \Rightarrow \quad w_{1,X} &= 0.\end{aligned}$$

This implies that $V_{0,X}$ is invariant with respect to A.

$2 \Rightarrow 3$: As it holds $P_X\,A_{0,XX}^{-1}\,R_X\,v \in V_{0,X}$ for all $v \in V$ it follows based on the invariance of $V_{0,X}$ with respect to A the inclusion $A\,P_X\,A_{0,XX}^{-1}\,R_X\,(V) \subset V_{0,X}$. As it is

$$R_X\,A\,P_X\,A_{0,XX}^{-1}\,R_X = R_X$$

we obtain

$$rk(A\,P_X\,A_{0,XX}^{-1}\,R_X) \geq rk(R_X\,A\,P_X\,A_{0,XX}^{-1}\,R_X) = rk(R_X) = n_0.$$

This implies that $A\,P_X\,A_{0,XX}^{-1}\,R_X : V \to V_{0,X}$ is surjective. Hence we have $(A\,P_X\,A_{0,XX}^{-1}\,R_X)\,(V) = V_{0,X}$.

$3 \Rightarrow 1$: As we obtain $A\,P_X\,A_{0,XX}^{-1}\,R_X \in V_{0,X} = W_X^{\perp}$ for all $v \in V$. It follows

$$(A\,P_X\,A_{0,XX}^{-1}\,R_X\,v,\,w_X) = 0 \quad \forall v \in V,\,\forall w_X \in W_X.$$

This implies the proposition.

□

So like the similar lemmata for the one sided modification the two lemmata above show the structure of the modification as well as the aim of the modifications. The major stake of the difference is that for the one sided modification we still consider the subspace V_0 like in the unmodified situation. Then we try to create the situation in which it holds that V_0 is invariant with respect to $A\,X$. So we modify the operator. As the lemmata above suggested in the two sided modification, we will consider $V_{0,X}, W_X$. We will try to create the modification in a way that the modified space $V_{0,X}$ is invariant with respect to the operator A. This situation is illustrated in Figure 4.2 (cf. Figure 3.1 at page 69 and Figure 4.1 at page 109).

Before we take a look at the two sided modified preconditioners we will consider the condition that $V_{0,X}$ is invariant with respect to A with respect to the aggregation method. Hence we assume that we have two sets I_1, I_2 with

$$I_1 := \big\{ i \in \{1,\dots,n\} \,:\, \mathcal{N}_i^1 \text{ is an isolated point} \big\}$$
$$I_2 := \big\{ (i,j) \in \{1,\dots,n\} \times \{1,\dots,n\} \,:\, \mathcal{N}_i^1, \mathcal{N}_j^1 \text{ are aggregated.} \big\}$$

According to the definition of V_0 this implies that

$$\{e_i^1 \,:\, i \in I_1\} \cup \{e_i^1 + e_j^1 \,:\, (i,j) \in I_2\}$$

is a basis of V_0. Therewith

$$\{X_{\cdot,i} \,:\, i \in I_1\} \cup \{X_{\cdot,i} + X_{\cdot,j} \,:\, (i,j) \in I_2\} \tag{4.11}$$

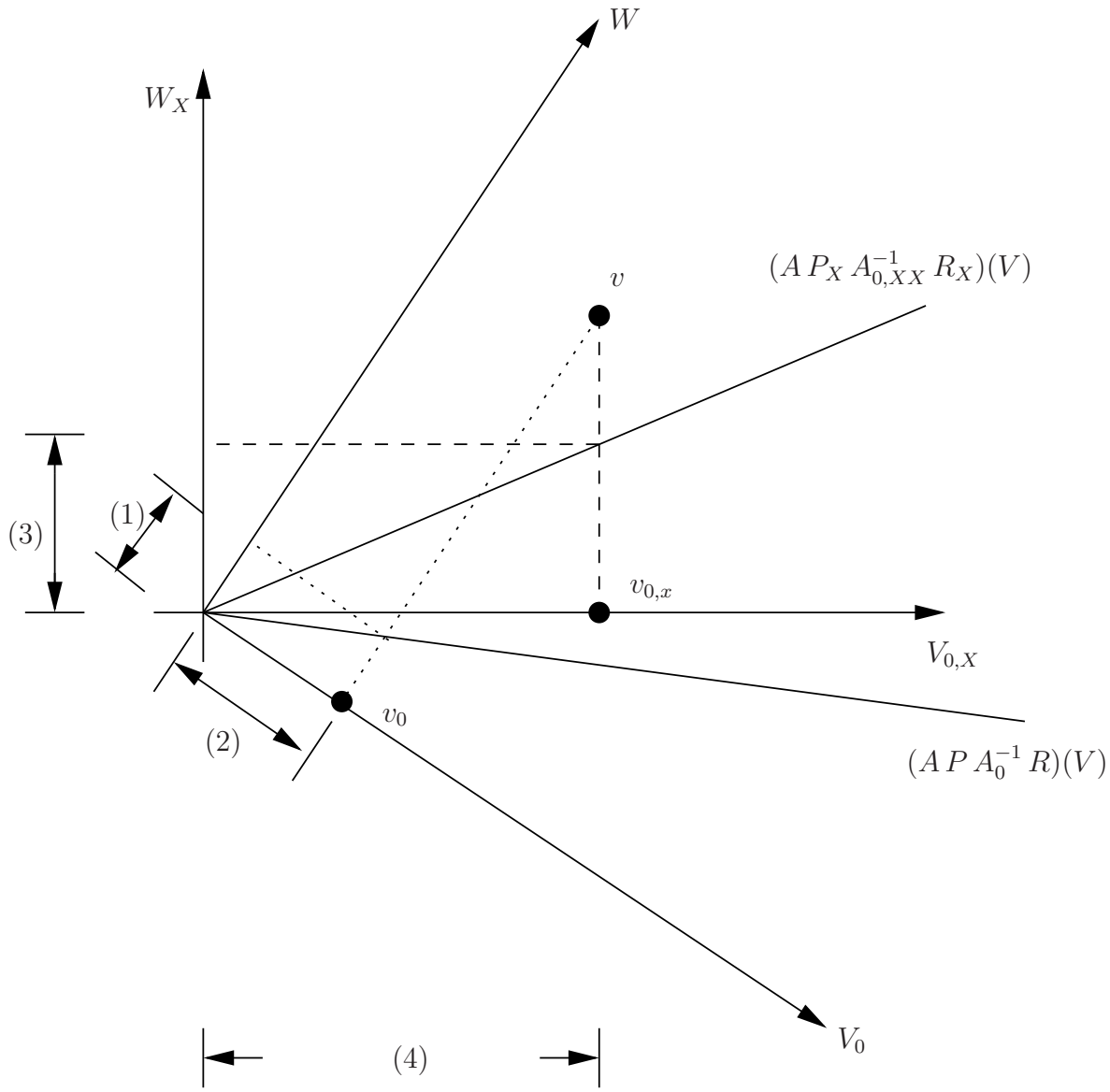
is a basis of $V_{0,X}$ if we use the aggregation method to construct P and $V_{0,X}$, respectively. We obtain the following result concerning the invariance of $V_{0,X}$ with respect to A.

Proposition: 4.2.5. *Let $A \in \mathbb{R}^{n\times n}$ be s.p.d. Then $V_{0,X}$ is invariant with respect to A if and only if there are $z_1,\dots,z_{n_0}$ with*

$$V_{0,X} = \Big\langle z_1,\dots,z_{n_0} \Big\rangle$$

and $A\,z_i = \lambda_i\,z_i$ for $i = 1,\dots,n_0$.

proof. We prove two implications. First we assume that $\{z_1,\dots,z_{n_0}\}$ is a basis of $V_{0,X}$ with $A\,z_i = \lambda_i\,z_i$ for $i = 1,\dots,n_0$. Then it follows obviously $A\,z_i \in V_{0,X}$ for $i = 1,\dots,n_0$. Hence it follows that $V_{0,X}$ is invariant with respect to A based on the linearity of the



$(1) = \|(I - Q_0)\,A\,P\,A_0^{-1}\,R\,v\|$ $\qquad$ $(2) = \|Q_0\,A\,P\,A_0^{-1}\,R\,v\| = \|v_0\|$

$(3) = \|(I - Q_{0,X})\,A\,P_X\,A_{0,XX}^{-1}\,R_X\,v\|$ $\qquad$ $(4) = \|Q_{0,X}\,A\,P_X\,A_{0,XX}^{-1}\,R_X\,v\| = \|v_{0,X}\|$

Figure 4.2: Modification of spaces $V_{0,X}$

operator. Second we assume that $V_{0,X}$ is not given by n_0 eigenvectors of A. Then we have that a basis of $V_{0,X}$ is given by

$$(z_1 + \cdots + z_k) \cup \{b_2, \ldots, b_{n_0}\}$$

with $A\, z_i = \lambda_i\, z_i$ for $i = 1, \ldots, k,\ k \geq 2$ and there are i, j with $\lambda_i \neq \lambda_j$. Furthermore, we have

$$z_i^T\, b_j = 0 \quad \text{for} \quad i = 1, \ldots, k, \text{ and } j = 2, \ldots, n_0.$$

Hence we obtain

$$A\,(z_1 + \cdots + z_k) = \lambda_1\, z_1 + \cdots + \lambda_k\, z_k.$$

Based on

$$(\lambda_1\, z_1 + \cdots + \lambda_k\, z_k)^T\, b_j = 0 \quad \text{for} \quad j = 2, \ldots, n_0$$

and $\lambda_1\, z_1 + \cdots + \lambda_k\, z_k \notin < z_1 + \cdots + z_k >$ we obtain

$$\lambda_1\, z_1 + \cdots + \lambda_k\, z_k \notin V_{0,X}.$$

This proves the second implication. □

Based on the Proposition 4.2.5 it follows from the representation (4.11) for a basis of $V_{0,X}$ that $V_{0,X}$ is invariant with respect to A if and only if the columns and the sum of columns of X, respectively are given by n_0 linear independent eigenvectors of A. Our ideas to modify this system will be based on this characteristic. We will carry out this modification at the end of the section.

Based on the results above we can give estimations for the DT-method and the BPX-method in the two sided modified situation that are analogue to the unmodified methods. The major stake will be that we use the angles $\gamma_{DT,XX}$ instead of γ_{DT} and respectively the spaces $V_{0,X}, W_X$ instead of V_0, W.

4.2.1 The DT-method

We will start by the definition of a two sided modified preconditioner $C^{-1}_{DT,XX}$. For a non singular A and a non singular $A_{0,XX}$ we define $C^{-1}_{DT,XX}$ as follows

$$C^{-1}_{DT,XX} := A^{-1}(I - Q_{0,X}) + P_X\, A^{-1}_{0,XX}\, R_X. \tag{4.12}$$

As usual we will start by proving the non singularity of $A\, C^{-1}_{DT,XX}$. This is based on the same arguments as Lemma 4.1.4.

Lemma: 4.2.6. *Let $A \in \mathbb{R}^{n\times n}, A_{0,XX} \in \mathbb{R}^{n_0\times n_0}$ be non singular. Then the matrix*

$$A\, C_{DT,XX}^{-1}$$

is non singular.

proof. Suppose that $A\, C_{DT,XX}^{-1}$ is singular. Then it must exist a $v \in V\backslash\{0\}$ with

$$\begin{aligned} & 0 = A\, C_{DT,XX}^{-1} v \\ \Leftrightarrow \quad & 0 = (I - Q_{0,X})\, v + A\, P_X\, A_{0,XX}^{-1}\, R_X\, v \\ \Rightarrow \quad & -R_X\,(I - Q_{0,X})\, v = \underbrace{R_X\, A\, P_X}_{=A_{0,XX}}\, A_{0,XX}^{-1}\, R_X\, v \\ \Leftrightarrow \quad & 0 = R_X\, v. \end{aligned}$$

So the for the given $v \in V$ we obtain $R_X\, v = 0$. But in the case of $R_X\, v = 0$ it follows

$$0 = A\, C_{DT,XX}^{-1} v = (I - Q_{0,X})\, v + A\, P_X\, A_{0,XX}^{-1}\, R_X\, v = v.$$

This is in contradiction to the assumption. □

Again as mentioned after the proof of the non singularity of $A\, C_{DT,X}^{-1}$ the non singulariy of $A\, C_{DT,XX}^{-1}$ immediately implies $\gamma_{DT,XX} < 1$ for all non singular operators A, and all prolongations P and modifications X that fulfil that $A_{0,XX}$ is non singular.

So we can prove for the two sided modified preconditioner the same characteristics as for the one sided and the unmodified preconditioner.

Theorem: 4.2.7. *Let $A \in \mathbb{R}^{n\times n}$, $A_{0,XX} \in \mathbb{R}^{n_0\times n_0}$ be non singular and $C_{DT,XX}^{-1}$ as defined in (4.12). Then the inequalities*

$$c_{DT,XX}\|A\, C_{DT,XX}^{-1}\, v\|^2 \le \|v\|^2 \le d_{DT,XX}\|A\, C_{DT,XX}^{-1}\, v\|^2 \tag{4.13}$$

holds for all $v \in V$ with

$$c_{DT,XX} := \frac{2 + \mu_{\gamma_{DT,XX}}^2 - \mu_{\gamma_{DT,XX}}\sqrt{4 + \mu_{\gamma_{DT,XX}}^2}}{2}$$

$$\text{and} \quad d_{DT,XX} := \frac{2 + \mu_{\gamma_{DT,XX}}^2 + \mu_{\gamma_{DT,XX}}\sqrt{4 + \mu_{\gamma_{DT,XX}}^2}}{2}.$$

proof. We can set for the vector spaces V_0, W of Lemma 3.6.2 the spaces $V_{0,X}, W_X$. Moreover, we can set in this lemma $B = A\,P_X\,A_{0,XX}^{-1}\,R_X$. Furthermore, we have shown for all $v \in V$ the equality

$$Q_{0,X}\,v = Q_{0,X}\,A\,P_X\,A_{0,XX}^{-1}\,R_X\,v.$$

Hence we obtain the propositions based on the same arguments as the propositions of Theorem 3.6.5. □

Again the constants $c_{DT,XX}, d_{DT,XX}$ that determine the condition of $A\,C_{DT,XX}^{-1}$ have the same structure as c_{DT}, d_{DT} in Theorem 3.6.5 and $c_{DT,X}, d_{DT,X}$ in Theorem 4.1.5, respectively. It is obvious that we obtain for the constants the same characteristics as before. These are summarised in the following proposition.

Proposition: 4.2.8. *Let $c_{DT,XX}, d_{DT,XX}$ be as given in Theorem 4.2.7 then it follows:*

1. *$c_{DT,XX} \leq 1 \leq d_{DT,XX}$ and it is $c_{DT,XX} = 1 = d_{DT,XX}$ if and only if it is $\gamma_{DT,XX} = 0$.*

2. *It is*

$$\frac{d}{d\gamma_{DT,XX}}[c_{DT,XX}] < 0 \quad \textit{and} \quad \frac{d}{d\gamma_{DT,XX}}[d_{DT,XX}] > 0.$$

3. *There is no $c^* > c_{DT,XX}$ and no $d^* < d_{DT,XX}$ that hold for all $v \in V$*

$$c^*\|C_{DT,XX}^{-1}\,A\,v\|^2 \leq \|v\|^2 \leq d^*\|C_{DT,XX}^{-1}\,A\,v\|^2.$$

4. *The constants $c_{DT,XX}, d_{DT,XX}$ are given by*

$$c_{DT,XX} = \min_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+1+\mu_{\gamma_{DT,XX}}^2+2\lambda\mu_{\gamma_{DT,XX}}} = \min_{\lambda\in\mathbb{R},\,\mu\in[0,\mu_{\gamma_{DT,XX}}]} \frac{\lambda^2+1}{\lambda^2+1+\mu^2+2\lambda\mu}$$

$$d_{DT,XX} = \max_{\lambda\in\mathbb{R}} \frac{\lambda^2+1}{\lambda^2+1+\mu_{\gamma_{DT,XX}}^2-2\lambda\mu_{\gamma_{DT,XX}}} = \max_{\lambda\in\mathbb{R},\,\mu\in[0,\mu_{\gamma_{DT,XX}}]} \frac{\lambda^2+1}{\lambda^2+1+\mu^2-2\lambda\mu}.$$

proof. As the constants $c_{DT,XX}, d_{DT,XX}$ have the same structure as c_{DT}, d_{DT} in Theorem 3.6.5 the proof follows again the same arguments as the proofs of Remark 3.6.4 and the Corollaries 3.6.6, 3.6.7 and 3.6.8. □

To conclude the section we will take a look at some examples for modification matrices. These will be motivated for symmetric matrices. We therefore take into consideration the result concerning the invariance of $V_{0,X}$ given in Proposition 4.2.5. If A is symmetric then there is an orthogonal matrix $O \in \mathbb{R}^{n\times n}$ that holds

$$A = O\, D_A\, O^T,$$

with a diagonal matrix D_A. Then we set

$$A^s = O\, D_A^s\, O^T \quad \text{with} \quad D_A^s = diag\big((D_A)_{1,1}^s, \ldots, (D_A)_{n,n}^s\big)$$

for $s \in \mathbb{R}$. For the modification we set $X = A^{-1/2}$. Therewith X is in this case also given as symmetric. This implies

$$A_{0,XX} = R\, X^T\, A\, X\, P = R\, P = S^{-1}$$

$$S_X = (R\, X^T\, X\, P)^{-1} = (R\, A^{-1}\, P)^{-1}.$$

Therewith it follows

$$\begin{aligned} A\, C_{DT,XX}^{-1} &= (I - A^{-1/2}\, P\, (R\, A^{-1}\, P)^{-1}\, R\, A^{-1/2}) + A\, A^{-1/2}\, P\, S\, R\, A^{-1/2} \\ &= A^{1/2}\, (I - A^{-1}\, P\, (R\, A^{-1}\, P)^{-1}\, R)\, A^{-1/2} + A^{1/2}\, Q_0\, A^{-1/2} \\ &= A^{1/2}\, (I + Q_0 - A^{-1}\, P\, (R\, A^{-1}\, P)^{-1}\, R)\, A^{-1/2}. \end{aligned}$$

This is the exact inverse if and only if the term in brackets is the identity. This is equivalent to

$$Q_0\, v = A^{-1}\, P\, (R\, A^{-1}\, P)^{-1}\, R\, v \quad \forall v \in V.$$

Based on this characteristic we see that such a modification is senseless. For the unmodified system we have the problem that V_0 is not invariant with respect to A. For the two sided modification with $X = A^{-1/2}$ we obtain the problem that V_0 is not invariant with respect to A^{-1}. Furthermore we want to highlight that we obtain the same problem if we use A^{-1} to carry out the modification.

As mentioned above we will consider a modification that is based only on the eigenvectors of A. We highlight that we modify the symmetric operator with an unsymmetric X. We still assume that $A = O\, D_A\, O^T$. Then it follows for $X = O$

$$A_{0,XX} = R\, X^T\, A\, X\, P = R\, O^T\, O\, D_A\, O^T\, O\, P = R\, D_A\, P$$

$$S_X = (R\, X^T\, X\, P)^{-1} = (R\, P)^{-1} = S.$$

This implies

$$\begin{aligned} A\, C_{DT,XX}^{-1} &= (I - X\, P\, S_X\, R\, X^T) + O\, D_A\, O^T\, X\, P\, A_{0,XX}^{-1}\, R\, X^T \\ &= (I - O\, P\, S\, R\, O^T) + O\, D_A\, O^T\, O\, P\, (R\, D_A\, P)^{-1}\, R\, O^T \\ &= O\, (I - Q_0)\, O^T + O\, D_A\, P\, (R\, D_A\, P)^{-1}\, R\, O^T \\ &= O\, (I - Q_0 + D_A\, P\, (R\, D_A\, P)^{-1}\, R)\, O^T. \end{aligned}$$

And again this is exact if the term in brackets is the identity. Again this is in general not fulfilled as this is equivalent to

$$Q_0\, v = D_A\, P\, (R\, D_A\, P)^{-1}\, R\, v \quad \forall v \in V.$$

As it is $Q_0\, v \in V_0$ and if $\mathcal{N}_1^1, \mathcal{N}_2^1$ are aggregated we obtain for

$$P\, (R\, D_A\, P)^{-1}\, R\, v = (1, 1, 0, \ldots, , 0)^T$$

$$D_A\, P\, (R\, D_A\, P)^{-1}\, R\, v = (\lambda_1, \lambda_2, 0, \ldots, 0) \notin V_0.$$

Hence for this modification we have the problem that V_0 is not invariant with respect to D_A. We take a closer look at the space $V_{0,X}$ follows from the use of $X = O$. Let $z_i = X_{.,i}$ be the eigenvectors of A. With the sets I_1, I_2 as used for the representation (4.11) it follows that

$$\{z_i \,:\, i \in I_1\} \cup \{z_i + z_j \,:\, (i,j) \in I_2\}$$

is a basis of $V_{0,X}$. Hence the assumptions of Proposition 4.2.5 are not fulfilled. To construct a modification that fulfils the assumptions of Proposition 4.2.5 we define $\widetilde{I} \in \mathbb{R}^{n \times n}$ as follows

$$\widetilde{I} = diag(\widetilde{i}_{1,1}, \ldots, \widetilde{i}_{n,n})$$

$$\widetilde{i}_{i,i} = \begin{cases} 1 & \text{if } \mathcal{N}_i^1 \text{ is an isolated point} \\ & \text{or } \mathcal{N}_i^1, \mathcal{N}_j^1 \text{ are aggregated and it is } i < j \\ 0 & \text{otherwise.} \end{cases}$$

Based on the definition of $\widetilde{I}$ it follows for an arbitrary $T \in \mathbb{R}^{n \times n}$

$$(T\, \widetilde{I})_{.,i} = \begin{cases} T_{.,i} & \text{if } \mathcal{N}_i^1 \text{ is an isolated point} \\ & \text{or } \mathcal{N}_i^1, \mathcal{N}_j^1 \text{ are aggregated and it is } i < j \\ (0, \ldots, 0)^T & \text{otherwise.} \end{cases}$$

So the matrix $\widetilde{I}$ selects n_0 columns of T. Based on the same argument $\widetilde{I}\,T$ selects the same n_0 rows of T. Then we define the matrix $X = \widetilde{O} = O\,\widetilde{I}$. Based on this modification we obtain

$$\begin{aligned} S_X &= (R\,\widetilde{I}\,O^T\,O\widetilde{I}\,P)^{-1} = I_0. \\ Q_{0,X} &= P_X\,S_X\,R_X = O\,\widetilde{I}\,P\,R\,\widetilde{I}\,O^T = O\,\widetilde{I}\,O^T \\ A_{0,XX} &= R_X\,A\,P_X = R\,\widetilde{I}\,O^T\,A\,O\,\widetilde{I}\,P = R\,\widetilde{I}\,O^T\,O\,D_A\,O^T\,O\,\widetilde{I}\,P = R\,\widetilde{I}\,D_A\,\widetilde{I}\,P =: \widetilde{D}_A \end{aligned}$$

with $\widetilde{D}_A \in \mathbb{R}^{n_0 \times n_0}$

$$\widetilde{D}_A = diag((\widetilde{D}_A)_{1,1}, \ldots, (\widetilde{D}_A)_{n_0,n_0}).$$

As it is $D_A = diag(\lambda_1, \ldots, \lambda_n)$ with the eigenvalues λ_i, $i = 1, \ldots, n$ of A it follows that $(\widetilde{D}_A)_{i,i} = \lambda_j$ with $j \in \{1, \ldots, n\}$. Moreover, we obtain from the definiton of $Q_{0,X}$ that this operator is the projection $V \to V_{0,X}$ that is orthogonal with respect to the Euclidean norm.

From the calculations above we obtain

$$\begin{aligned} A\,C_{DT,XX}^{-1} &= A\,A^{-1}\,(I - Q_{0,X}) + A\,P_X\,A_{0,XX}^{-1}\,R_X \\ &= I - O\,\widetilde{I}\,P\,R\,\widetilde{I}\,O^T + O\,D_A\,O^T\,O\,\widetilde{I}\,P\,(\widetilde{D}_A)^{-1}\,R\,\widetilde{I}\,O^T \\ &= I - O\,\widetilde{I}\,P\,R\,\widetilde{I}\,O^T + O\,D_A\,\widetilde{I}\,P\,(\widetilde{D}_A)^{-1}\,R\,\widetilde{I}\,O^T. \end{aligned}$$

Therewith $A\,C_{DT,XX}^{-1}$ is the identity if and only if it is

$$O\,\widetilde{I}\,P\,R\,\widetilde{I}\,O^T = O\,D_A\,\widetilde{I}\,P\,(\widetilde{D}_A)^{-1}\,R\,\widetilde{I}\,O^T. \tag{4.14}$$

We have the equality above if

$$\widetilde{I}\,P = D_A\,\widetilde{I}\,P\,(\widetilde{D}_A)^{-1}$$

holds. For the left side of this equation if e_i^1 is the k-th column of P we obtain that $(\widetilde{I}\,P)_{.,k} = e_i^1$. And if $e_i^1 + e_j^1$ with $i < j$ is the k-column of P that $(\widetilde{I}\,P)_{.,k} = e_i^1$. For the right side we obtain if the e_i^1 is the k-column of P that

$$\begin{aligned} (P\,(\widetilde{D}_A)^{-1})_{.,k} &= \frac{1}{\lambda_i}\,e_i^1 \\ \Rightarrow \quad (\widetilde{I}\,P\,(\widetilde{D}_A)^{-1})_{.,k} &= \frac{1}{\lambda_i}\,e_i^1 \\ \Rightarrow \quad (D_A\,\widetilde{I}\,P\,(\widetilde{D}_A)^{-1})_{.,k} &= e_i^1. \end{aligned}$$

And if $e_i^1 + e_j^1$ with $i < j$ is the k-column of P that

$$(P\,(\widetilde{D}_A)^{-1})_{.,k} = \frac{1}{\lambda_i}\,(e_i^1 + e_j^1)$$

$$\Rightarrow \quad (\widetilde{I}\,P\,(\widetilde{D}_A)^{-1})_{.,k} = \frac{1}{\lambda_i}\,e_i^1$$

$$\Rightarrow \quad (D_A\,\widetilde{I}\,P\,(\widetilde{D}_A)^{-1})_{.,k} = e_i^1.$$

Therewith the equation (4.14) holds and we obtain $A\,C_{DT,XX}^{-1} = I$.

To conclude this example we want to highlight two characteristics of the modification:

1. For practical issues there is so far no rule which n_0 eigenvectors of A should be chosen.

2. We obtain the same result if we modify with $X = O\,D_A^{-1/2}\,\widetilde{I}$. In this case we scale the eigenvectors with the associated eigenvalue. We get $A_{0,XX} = I_0$ in this case and $A\,C_{DT,XX}^{-1} = I$ follows from a similar calculation.

4.2.2 The BPX-method

Similarly to the two sided modified DT-method we define the two sided modified BPX preconditioner. For a non singular A and a non singular $A_{0,XX}$ we define $C_{BPX,XX}^{-1}$ as follows

$$C_{BPX,XX}^{-1} := A^{-1} + P_X\,A_{0,XX}^{-1}\,R_X. \tag{4.15}$$

As already mentioned we obtain for the two sided modification that $A_{0,XX}$ is symmetric if this holds for A. But in contrast to the DT-method we obtain for the BPX-method the symmetry of the operator $C_{BPX,XX}^{-1}$. The result is

$$\begin{aligned}(C_{BPX,XX}^{-1})^T &= (A^{-1} + P_X\,A_{0,XX}^{-1}\,R_X)^T = (A^{-1})^T + (R_X)^T\,(A_{0,XX}^{-1})^T\,(P_X)^T \\ &= A^{-1} + P_X\,A_{0,XX}^{-1}\,R_X = C_{BPX,XX}^{-1}.\end{aligned}$$

Then we will show that $A\,C_{BPX,XX}^{-1}$ is also non singular based on the same condition as used all the time.

Lemma: 4.2.9. *Let $A \in \mathbb{R}^{n\times n}$, $A_{0,XX} \in \mathbb{R}^{n_0\times n_0}$ be non singular. Then the matrix*

$$A\,C_{BPX,XX}^{-1}$$

is non singular.

proof. Suppose that $A\,C^{-1}_{BPX,XX}$ is singular. Then there is a $v \in V\backslash\{0\}$ with

$$\begin{aligned} & 0 = A\,C^{-1}_{BPX,XX} v \\ \Leftrightarrow \quad & 0 = v + A\,P_X\,A^{-1}_{0,XX}\,R_X\,v \\ \Rightarrow \quad & -R_X\,v = \underbrace{R_X\,A\,P_X}_{=A_{0,XX}}\,A^{-1}_{0,XX}\,R_X\,v \\ \Leftrightarrow \quad & -R_X\,v = R_X\,v. \end{aligned}$$

So for the given $v \in V$ we obtain $R_X\,v = 0$. But in the case of $R_X\,v = 0$ it follows

$$0 = A\,C^{-1}_{BPX,XX} v = v + A\,P_X\,A^{-1}_{0,XX}\,R_X\,v = v.$$

Hence this is in contradiction to the assumption. □

We have shown for the DT-method how we have to modify the assumptions and the steps to get the same result as for the unmodified method, we obtain this also for the modified BPX-method. This is obvious as we used the same arguments in section 3.6 for both methods.

Theorem: 4.2.10. *Let $A \in \mathbb{R}^{n\times n}$, $A_{0,XX} \in \mathbb{R}^{n_0\times n_0}$ be non singular and $C^{-1}_{BPX,XX}$ as defined in (4.15). Then the inequalities*

$$c_{BPX,XX}\|A\,C^{-1}_{BPX,XX}\,v\|^2 \le \|v\|^2 \le d_{BPX,XX}\|A\,C^{-1}_{BPX,XX}\,v\|^2 \tag{4.16}$$

holds for all $v \in V$ with

$$c_{BPX,XX} := \frac{5 + \mu^2_{\gamma_{DT,XX}} - \sqrt{9 + 10\mu^2_{\gamma_{DT,XX}} + \mu^4_{\gamma_{DT,XX}}}}{8} \tag{4.17}$$

$$\text{and} \quad d_{BPX,XX} := \frac{5 + \mu^2_{\gamma_{DT,XX}} + \sqrt{9 + 10\mu^2_{\gamma_{DT,XX}} + \mu^4_{\gamma_{DT,XX}}}}{8}. \tag{4.18}$$

proof. As mentioned above the proposition follows the same arguments as the propositions of Theorem 3.6.10 if we modify the spaces for that and we use the arguments as explained in the proof of Theorem 4.2.7. □

So it is obvious that we get the same characteristics as before. To sum up it is:

Proposition: 4.2.11. *Let $c_{BPX,XX}, d_{BPX,XX}$ be as given in Theorem 4.2.10 then it follows:*

1. *$c_{BPX,XX} < 1 \leq d_{BPX,XX}$ and it is $d_{BPX,XX} = 1$ if and only if it is $\gamma_{DT,XX} = 0$.*

2. *It is*
$$\frac{d}{d\gamma_{DT,XX}}[c_{BPX,XX}] < 0 \quad \textit{and} \quad \frac{d}{d\gamma_{DT,XX}}[d_{BPX,XX}] > 0.$$

3. *There is no $c^* > c_{BPX,XX}$ and no $d^* < d_{BPX,XX}$ that hold for all $v \in V$*
$$c^* \|C_{BPX,XX}^{-1} A\, v\|^2 \leq \|v\|^2 \leq d^* \|C_{BPX,XX}^{-1} A\, v\|^2.$$

4. *The constants $c_{BPX,XX}, d_{BPX,XX}$ are given by*
$$c_{BPX,XX} = \min_{\lambda \in \mathbb{R}} \frac{\lambda^2 + 1}{\lambda^2 + 4 + \mu_{\gamma_{DT,XX}}^2 + 2\lambda\mu_{\gamma_{DT,XX}}} = \min_{\lambda \in \mathbb{R},\, \mu \in [0, \mu_{\gamma_{DT,XX}}]} \frac{\lambda^2 + 1}{\lambda^2 + 4 + \mu^2 + 2\lambda\mu}$$
$$d_{BPX,XX} = \max_{\lambda \in \mathbb{R}} \frac{\lambda^2 + 1}{\lambda^2 + 4 + \mu_{\gamma_{DT,XX}}^2 - 2\lambda\mu_{\gamma_{DT,XX}}} = \max_{\lambda \in \mathbb{R},\, \mu \in [0, \mu_{\gamma_{DT,XX}}]} \frac{\lambda^2 + 1}{\lambda^2 + 4 + \mu^2 - 2\lambda\mu}.$$

proof. The proof follows again based on the same structure as for the unmodified or one sided modified BPX-method. □

To conclude this section we will also consider for the two sided BPX-method the same modifications as done for the DT-method. For $X = A^{-1/2}$ we obtain
$$\begin{aligned} A\, C_{BPX,XX}^{-1} &= I + A\, A^{-1/2}\, P\, S\, R\, A^{-1/2} \\ &= A^{1/2}\, (I + Q_0)\, A^{-1/2}. \end{aligned}$$
Hence this illustrates again that the BPX-method can not be exact. Furthermore we see that this is as far from the identity as the unmodified method. As unsymmetric example we consider for $A = O\, D_A\, O^T$ again $X = O$. Then it follows
$$\begin{aligned} A\, C_{BPX,XX}^{-1} &= I + O\, D_A\, O^T\, O\, P\, (R\, D_A\, P)^{-1}\, R\, O^T \\ &= O\, (I + D_A\, P\, (R\, D_A\, P)^{-1}\, R)\, O^T. \end{aligned}$$
This is more senseful. In particular if it is $D_A\, P\, (R\, D_A\, P)^{-1}\, R = Q_0$. However, the problems concerning this characteristic are explained for the DT-method. Finally we will carry out the modification with $X = \widetilde{O} = O\, \widetilde{I}$. In this case we obtain from the same calculation as done for the DT-method
$$\begin{aligned} A\, C_{BPX,XX}^{-1} &= I + O\, D_A\, O^T\, O\, \widetilde{I}\, P\, (\widetilde{D}_A)^{-1}\, R\, \widetilde{I}\, O^T \\ &= O\, (I + \widetilde{I})\, O^T = I + Q_{V_{0,X}}. \end{aligned}$$
Since $Q_{V_{0,X}}$ is the projection $V \to V_{0,X}$ that is orthogonal with respect to the Euclidean norm we obtain the biggest (smallest) eigenvalue of $A\, C_{BPX,XX}^{-1}$: two (one).

5 Examples for modifications

In the last chapter we have seen that we can modify the preconditioninig methods as we modify the prolongation, or the prolongation and the restriction respectively. Now we will consider the problems we have introduced as model problems in section 2.2. We will motivate modifications by meaningful results for quite simple special cases of these problems. Afterwards we will considere more general cases of the modell problems to get an idea of what happens in these cases in relation to our modification. Of course we will get in the more general cases not the meaningful results which we have for the simple problems. Further we will use for all the examples the aggregation method to get the coarser grids.

5.1 Convection diffusion equation

5.1.1 One dimensional convection

We will start with our unsymmetric model problem and consider the convection diffusion equation defined in (2.5). As a more simple version we will consider the stiffness matrices we get in the case of the one dimensional system with $\varepsilon = 0$. The equation we consider is given by

$$b(x)\, D_x\, u(x) = f(x) \quad \forall x \in \Omega \subset \mathbb{R}$$

$$u(x) = c(x) \quad \forall x \in \partial\Omega.$$

Furthermore, we assume that it is $b(x) > 0$, for all $x \in \overline{\Omega}$. We use finite differences for the discritisation and by appliying the upwind method we get in $\mathcal{N}_i$ the stencil

$$[-b_i, b_i, 0]$$

with $b_i > 0$. To set $\varepsilon = 0$ can be seen as the limit $\varepsilon \to 0$. As the diffusion is often small compared to the convection (that means $\varepsilon << b(x)$) a discritisation and a solution method should at least confirm that the equation we get from the limit $\varepsilon \to 0$ is as

exact as possible. This can be seen as motivation for this problem. Furthermore, the one dimensional situation can be seen as the situation of a one dimensional convection in a two or three dimensional system.

For the most aspects it is sufficient to consider a small system given by the four grid points $\mathcal{N}_1^1, \ldots, \mathcal{N}_4^1$. Then we assume that we aggregate the grid points $\mathcal{N}_2^1, \mathcal{N}_3^1$ to the new point $\mathcal{N}_2^0$ (cf. Figure 5.1 at page 135). So we obtain that the restriction R and the prolongation P are give by

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad R = P^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5.1}$$

Therewith also follows that

$$R\,P = diag(1,\, 2,\, 1), \quad S = diag(1,\, 1/2,\, 1),$$

$$P\,S\,R = Q_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad (I - Q_0) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -1/2 & 1/2 & 0 \\ 0 & 1/2 & -1/2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

So this also gives the structure of the subspaces V_0, W. It is

$$\{(1,0,0,0),\ (0,1,1,0),\ (0,0,0,1)\}$$

a basis of V_0 and

$$\{(0,-1,1,0)\}$$

a basis of W and we obtain the stiffness matrix A as

$$A = \begin{pmatrix} b_1 & 0 & 0 & 0 \\ -b_2 & b_2 & 0 & 0 \\ 0 & -b_3 & b_3 & 0 \\ 0 & 0 & -b_4 & b_4 \end{pmatrix}. \tag{5.2}$$

We highlight, that this implies that the coarse gird operator A_0 is follows as

$$A_0 = R\,A\,P = \begin{pmatrix} b_1 & 0 & 0 \\ -b_2 & b_2 & 0 \\ 0 & -b_4 & b_4 \end{pmatrix}. \tag{5.3}$$

Therewith the coefficient b_3 that represents the conjunction between the grid points $\mathcal{N}_2, \mathcal{N}_3$ has no effect for the coarser operator. The hole system is illustrated in Figure 5.1.

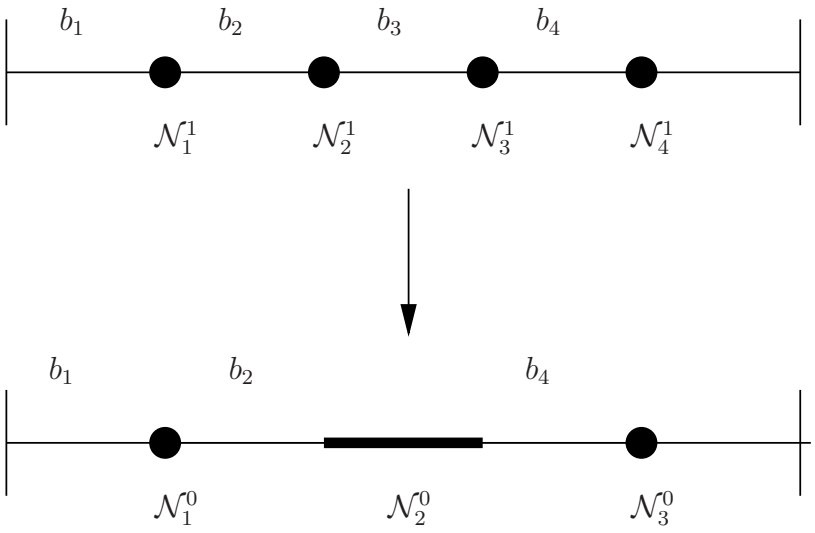


Figure 5.1: Coarsing of the four point system

Unmodified method

First we will consider the result for the unmodified method. In particular we will consider the angle γ_{DT} that determines the condition of $A\,C_{BPX}^{-1}$ and $A\,C_{DT}^{-1}$ in the Euclidean norm. Based on the definition of γ_{DT} and the results of the previous chapters the following equivalence is obvious:

$$\begin{aligned} \gamma_{DT} := \min \Big\{ t \in \mathbb{R}_+ \,:\, & (A\,P\,A_0^{-1}\,R\,v,\,(I-Q_0)v) \\ & \leq t\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I-Q_0)v\|,\ \forall v \in V \Big\} \\ \Leftrightarrow \quad \gamma_{DT} := \min \Big\{ t \in \mathbb{R}_+ \,:\, & (A\,v_0,\,w) \leq t\,\|A\,v_0\|\,\|w\|,\ \forall v_0 \in V_0,\,\forall w \in W \Big\}. \end{aligned}$$

To consider the inequality

$$(A\,v_0,\,w) \leq \gamma_{DT}\,\|A\,v_0\|\,\|w\| \quad \forall v_0 \in V_0,\, w \in W$$

is more simple for explicit calculations. Thus we obtain in the simple situation the following result:

Proposition: 5.1.1. *Let A be given as in (5.2) and P, R as given in (5.1). Then*

$$(A\,v_0,\,w) \leq \gamma\|A\,v_0\|\,\|w\|, \quad \forall v_0 \in V_0,\, w \in W$$

holds with $\gamma = \sqrt{1/2}$. Furthermore this is the best possible estimation.

proof. As an arbitrary $v_0 \in V_0$ and an arbitrary $w \in W$ is given by

$$\begin{aligned} v_0 &= (f, u, u, g), \quad f, u, g \in \mathbb{R} \\ w &= (0, s, -s, 0), \quad s \in \mathbb{R} \end{aligned}$$

we obtain

$$A\,v_0 = (b_1 f,\, b_2(u-f),\, 0, b_4(g-u)).$$

And therewith follows

$$\begin{aligned} (A\,v_0,\,w) &= b_2\,s\,(u-f) \\ \|A\,v_0\|^2 &= b_1^2 f^2 + b_2^2(u-f)^2 + b_4^2(g-u)^2 \\ \|w\|^2 &= 2s^2. \end{aligned}$$

This implies

$$\begin{aligned} (A\,v_0,\,w)^2 = b_2^2 s^2 (u-f)^2 &= \frac{1}{2}\,(2\,s^2)\,(b_2^2(u-f)^2) \\ &\leq \frac{1}{2}\,(2\,s^2)\,(b_2^2(u-f)^2 + b_1^2 f^2 + b_4^2(g-u)^2) \\ &= \frac{1}{2}\,\|w\|^2\,\|A\,v_0\|^2. \end{aligned}$$

It is obvious that this is the best possible estimation if we consider the case of $f = 0$ and $g = u$. □

Furthermore, it is obvious that this result does not depend on the low dimension of the problem. For a matrix of the same structure and an arbitrary dimension we get the same result.

We consider the situation of n grid points. Then we assume that the stencil in $\mathcal{N}_i^1$ is given by $[-b_i, b_i, 0]$. By the choosen numeration of the grid points follows that the stiffness matrix $A \in \mathbb{R}^{n\times n}$ follows as

$$a_{i,j} = \begin{cases} b_i & \text{for} \quad j = i \\ -b_i & \text{for} \quad j = i-1 \\ 0 & \text{else.} \end{cases} \tag{5.4}$$

For an illustration see Figure 5.2 at page 142. Furthermore we still assume that $R \in \mathbb{R}^{n_0\times n}$ follows from the aggregation method. Hence we have

$$R_{j,\cdot} = \begin{cases} (e_i^1)^T & \text{if } \mathcal{N}_i^1 \subset \mathcal{N}_j^0 \text{ is an isolated point} \\ (e_i^1)^T + (e_{i+1}^1)^T & \text{if } \mathcal{N}_i^1, \mathcal{N}_{i+1}^1 \text{ are aggregated to } \mathcal{N}_j^0. \end{cases} \tag{5.5}$$

Then the result of Proposition 5.1.1 can be generalized as follows:

Proposition: 5.1.2. *Let A, R be given as in (5.4), (5.5) and $P = R^T$. Then*

$$(A\, v_0,\, w) \le \gamma \, \|A\, v_0\| \, \|w\|, \quad \forall v_0 \in V_0,\, w \in W.$$

holds with $\gamma = \sqrt{1/2}$. Furthermore this estimation is best possible.

proof. We distinguish two different situations for the grid points. First we consider a point $\mathcal{N}_i^1$ that is isolated. Then it is $w(i) = 0$ and we have

$$\begin{aligned} (A\, v_0)(i)\, w(i) &\le 1/2\, ((A\, v_0)(i))^2\, (w(i))^2 \\ \Leftrightarrow \quad 0 &\le 0. \end{aligned}$$

And of course this inequality is fulfilled.
Now we consider a point $\mathcal{N}_i^1$ that is aggregated with $\mathcal{N}_{i+1}^1$ to $\mathcal{N}_j^0$. In this case we have

$$\begin{aligned} & w(i) = -w(i+1) \\ \text{and} \quad & v_0(i) = v_0(i+1) \quad \Rightarrow \quad (A\, v_0)(i+1) = 0, \quad \forall v_0 \in V_0. \end{aligned}$$

Therewith follows that

$$\Big([\,(A\,v_0)(i),\,(A\,v_0)(i+1)\,],\,[w(i),w(i+1)] \Big) = (A\,v_0)(i)w(i)$$

$$\big\| \big[\,(A\,v_0)(i),\,(A\,v_0)(i+1)\,\big] \big\|^2 = ((A\,v_0)(i))^2$$

$$\big\| \big[\,w(i),w(i+1)\,\big] \big\|^2 = 2w(i)^2.$$

So the inequality holds with $\gamma_{DT} = \sqrt{1/2}$ for all one or two dimensional subsystem. The proposition follows from Lemma A.0.4. That this is the best possible estimation follows immediately from Proposition 5.1.1. □

An exact modification

Now we will construct for the simple system a modification $X = (x_{i,j})$ that realises $\gamma_{DT,X} = 0$. We will do this for the low dimensional system. Then we will show that we can generalise this to an arbitrary big system of the given structure. The main idea of this approch is that we invert the flux that is described by A for aggregated points.

So our aim is to construct X such that V_0 is invariant with respect to $A\,X$. On our four point system this is equivalent to

$$(A\,X\,v_0)(2) = (A\,X\,v_0)(3) \quad \text{holds for all} \quad v_0 \in V_0 \equiv \mathbb{R}^4.$$

Based on the basis of V_0 as shown above the equality must hold for all basis vectors. Hence we obtain that this is equivalent to

$$(A\,X)_{2,1} = (A\,X)_{3,1}, \quad (A\,X)_{2,4} = (A\,X)_{3,4} \tag{5.6}$$

$$\text{and} \quad (A\,X)_{2,2} + (A\,X)_{2,3} = (A\,X)_{3,2} + (A\,X)_{3,3}.$$

As $\mathcal{N}_1^1, \mathcal{N}_4^1$ are isolated points (i.e. as e_1^1, e_4^1 are basis elements of V_0) the values of the first and the fourth row of $A\,X$ does not matter. As we will modify few elements this motivates to set for the first and the fourth row of X the first and the fourth unit vector of $\mathbb{R}^4$. That means

$$X_{1,\cdot} = (e_1^1)^T \quad \text{and} \quad X_{4,\cdot} = (e_4^1)^T.$$

Next we will consider the three equations given in (5.6). It is

$$(A\,X)_{2,4} = (A\,X)_{3,4}$$

$$\Leftrightarrow \quad a_{2,1}x_{1,4} + a_{2,2}x_{2,4} + a_{2,3}x_{3,4} + a_{2,4}x_{4,4} = a_{3,1}x_{1,4} + a_{3,2}x_{2,4} + a_{3,3}x_{3,4} + a_{3,4}x_{4,4}$$

$$\Leftrightarrow \quad a_{2,2}x_{2,4} = a_{3,2}x_{2,4} + a_{3,3}x_{3,4}$$

This is fulfilled if we set $x_{2,4} = x_{3,4} = 0$. Furthermore, we obtain

$$(A\,X)_{2,1} = (A\,X)_{3,1}$$

$$\Leftrightarrow \quad a_{2,1}x_{1,1} + a_{2,2}x_{2,1} + a_{2,3}x_{3,1} + a_{2,4}x_{4,1} = a_{3,1}x_{1,1} + a_{3,2}x_{2,1} + a_{3,3}x_{3,1} + a_{3,4}x_{4,1}$$

$$\Leftrightarrow \quad a_{2,1} + a_{2,2}x_{2,1} = a_{3,2}x_{2,1} + a_{3,3}x_{3,1}$$

We set in this equation $x_{3,1} = 0$. Therewith the last equality is equivalent to

$$x_{2,1} = \frac{-a_{2,1}}{a_{2,2} - a_{3,2}} = \frac{b_2}{b_2 + b_3}.$$

So this equation is also fulfilled if we set $x_{2,1}$ as given above. At least we consider

$$(A\,X)_{2,2} + (A\,X)_{2,3} = (A\,X)_{3,2} + (A\,X)_{3,3}$$

$$\Leftrightarrow \quad a_{2,2}x_{2,2} + a_{2,2}x_{2,3} = a_{3,2}x_{2,2} + a_{3,3}x_{3,2} + a_{3,2}x_{2,3} + a_{3,3}x_{3,3}$$

We set for the consistence $x_{2,2} = 1 = x_{3,3}$ and $x_{3,2} = 0$. Therewith the equality is equivalent to

$$a_{2,2} + a_{2,2}x_{2,3} = a_{3,2} + a_{3,2}x_{2,3} + a_{3,3}$$

$$\Leftrightarrow \quad x_{2,3} = \frac{a_{3,3} - a_{2,2} + a_{3,2}}{a_{2,2} - a_{3,2}}.$$

For the matrix A this implies

$$x_{2,3} = \frac{-b_2}{b_2 + b_3} = -x_{2,1}.$$

Alltogether we get the matrices X, P_X as follows:

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{b_2}{b_2+b_3} & 1 & -\frac{b_2}{b_2+b_3} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad P_X = \begin{pmatrix} 1 & 0 & 0 \\ \frac{b_2}{b_2+b_3} & \frac{b_3}{b_2+b_3} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.7}$$

Thus it it follows $rk(P_X) = n_0 = 3$ as the rows $1, 3, 4$ of P_X are linearly independent. This property for a modification we have always assumed in the chapter 4.

And we can summarize the main properties of the matrix X as follows:

Proposition: 5.1.3. *Let A, P be as defined in (5.2), (5.1) and X as defined in (5.7). Then it follows that V_0 is invariant with respect to $A\,X$.*

proof. The proof follows the calculation as done above in this section. □

Moreover, we will highlight some interesting characteristics of this modification that we will prove later in a more general system. First it is obvious as the first row of X is given by e_1^1 that it follows $(A\,X\,v)(1) = (A\,v)(1)$ for all $v \in V$. As $(X)_{3,.}$ and $(X)_{4,.}$ are also given by the unit vectors e_3^1, e_4^1 we obtain analogue $(A\,X\,v)(4) = (A\,v)(4)$ for all $v \in V$.

At least we highlight that it follows

$$A_{0,X} = \begin{pmatrix} b_1 & 0 & 0 \\ -\frac{2\,b_2\,b_3}{b_2+b_3} & \frac{2\,b_2\,b_3}{b_2+b_3} & 0 \\ 0 & -b_4 & b_4 \end{pmatrix}$$

So if we compare the matrices $A_0, A_{0,X}$ (cf. (5.3)) we see that the modification maintains a lot of useful characteristics.

1. We have $(A_0)_{i,j} \neq 0 \Leftrightarrow (A_{0,X})_{i,j} \neq 0$. So there is no fill in if we use the modification. That means that the effort for the lower dimension grids does not increase if we we compare the modified system with the non modified system.

2. Like A, A_0 the matrix $A_{0,X}$ fulfils

$$a_{i,i} > 0, \quad a_{i,j} \leq 0 \quad \text{for } i \neq j \text{ and } \quad \sum_{j=1,\, j\neq i}^{n} |a_{i,j}| \leq a_{i,i}.$$

   So the matrix $A_{0,X}$ is also an M-matrix. A more detailed analysis of this aspect is done in section 9.1.

3. The link between $\mathcal{N}_1^0$ and $\mathcal{N}_2^0$ is in the modified system given by $\frac{2\,b_2\,b_3}{b_2+b_3}$. In the unmodified system this is just given by b_3, so the link is modified by the factor $\frac{2b_2}{b_2+b_3}$. This makes sense as a small b_2 should imply that the value on $\mathcal{N}_4^1$ and $\mathcal{N}_3^0$ respectively does not depend so strong on the value on $\mathcal{N}_1^1$ and $\mathcal{N}_1^0$, respectively. In the unmodified system this is not realized. For an illustration see again Figure 5.1 at page 135.

4. The modification holds

$$\sum_{j=1}^{4} x_{i,j} = 1, \quad \text{for all} \quad i = 1, \ldots, 4.$$

Because of this characteristic it follows $X\,v = v$ for all constant vectors v. It is easy to see that we obtain $A\,v \in V_0$ for constant $v \in V$. So in this case no modification is necessary. Therewith it is a kind of consistence that the choosen modification has no effect on such vectors.

Because of the inverse of A as defined in (5.2) is

$$A^{-1} = \begin{pmatrix} \frac{1}{b_1} & 0 & 0 & 0 \\ \frac{1}{b_1} & \frac{1}{b_2} & 0 & 0 \\ \frac{1}{b_1} & \frac{1}{b_2} & \frac{1}{b_3} & 0 \\ \frac{1}{b_1} & \frac{1}{b_2} & \frac{1}{b_3} & \frac{1}{b_4}. \end{pmatrix}$$

it is $X \neq A^{-1}$ for X given in (5.7). Together with the modification as given in (5.7) this illustrates that for a modification it is not necessary to determine the inverse of A to get the invariance of V_0 with respect to $A\,X$.

Exact modification for one dimensional convection systems of arbitrary size

In the last section we have seen that we can give a perfect modification for the matrix A we get for the one dimensional convection on the small system given by four grid points. Now we will show that we can generalize this to an arbitrary number of grid points and an arbitray structure of grid points that are aggregated pairwise to new grid points. This will explain some of the choices for the matrix X we have done in the last section and they seem to be arbitrary.

We consider the situation as illustrated in Figure 5.2. That means the stencil in $\mathcal{N}_i^1$ is given by $[-b_i, b_i, 0]$. Based on the chosen numeration of the grid points follows that the stiffness matrix $A \in \mathbb{R}^{n \times n}$ is given as

$$a_{i,j} = \begin{cases} b_i & \text{for} \quad j = i \\ -b_i & \text{for} \quad j = i - 1 \\ 0 & \text{else} \end{cases} \tag{5.8}$$

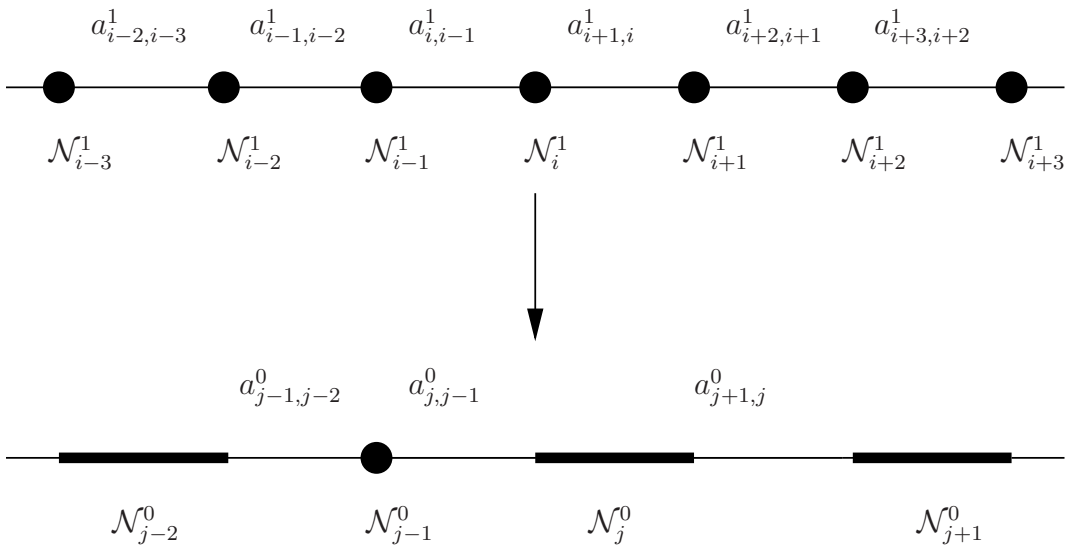


Figure 5.2: Coarsing of an arbitrary one dimensional system

with $b_i > 0$, $i = 1, \ldots, n$. For the restriction we define $R \in \mathbb{R}^{n_0 \times n}$ of the form

$$R_{j,\cdot} = \begin{cases} (e_i^1)^T & \text{if } \mathcal{N}_i^1 \subset \mathcal{N}_j^0 \text{ is an isolated point} \\ (e_i^1)^T + (e_{i+1}^1)^T & \text{if } \mathcal{N}_i^1, \mathcal{N}_{i+1}^1 \text{ are aggregated to } \mathcal{N}_j^0 \end{cases} \tag{5.9}$$

Then we highlight that by the structure of R given above

$$\Big\{ \{e_i^1 \,:\, \mathcal{N}_i^1 \text{ is isolated}\} \cup \{e_i^1 + e_j^1 \,:\, \mathcal{N}_i^1, \mathcal{N}_j^1 \text{ are aggregated}\} \Big\} \tag{5.10}$$

is a basis of V_0.

Then for a restriction $R \in \mathbb{R}^{n_0 \times n}$ of the structure defined in (5.9), we define the

modification matrix $X \in \mathbb{R}^{n\times n}$ based on its rows $X_{i,\cdot}$ as

(5.11)

$$X_{i,\cdot} := (e_i^1)^T, \quad \text{if } \mathcal{N}_i^1 \text{ is an isolated point.}$$

$$X_{i,\cdot} := (e_i^1)^T, \quad \text{if } \mathcal{N}_i^1, \mathcal{N}_{i-1}^1 \text{ are aggregated to } \mathcal{N}_j^0 \text{ for an } j \in \{1,\dots,n_0\}.$$

$$X_{i,\cdot} := (e_i^1)^T + \frac{b_i}{b_i + b_{i+1}}((e_{i-1}^1)^T - (e_{i+1}^1)^T), \quad \text{for } i > 1 \text{ if } \mathcal{N}_i^1, \mathcal{N}_{i+1}^1 \text{ are aggregated to } \mathcal{N}_j^0 \text{ for an } j \in \{1,\dots,n_0\}.$$

$$X_{1,\cdot} := (e_1^1)^T - \frac{b_1}{b_1 + b_2}(e_2^1)^T, \quad \text{if } \mathcal{N}_1^1, \mathcal{N}_2^1 \text{ are aggregated to } \mathcal{N}_j^0 \text{ for an } j \in \{1,\dots,n_0\}.$$

Therewith this modification matrix is a generalization of the matrix defined in (5.7). We get the same meaningful result as in the situation of the small system with four grid points:

Proposition: 5.1.4. *Let A be as defined in (5.8). Let R be a restriction operator as defined in (5.9), $P = R^T$ and X be the modification defined in (5.11). Then it follows that V_0 is invariant with respect to $A\,X$.*

proof. To prove that it is $A\,X\,v_0 \in V_0$ it is sufficient to prove that for two aggregated points $\mathcal{N}_i^1, \mathcal{N}_{i+1}^1$ we obtain

$$(A\,X\,v_0)(i) = (A\,X\,v_0)(i+1) \quad \forall v_0 \in V_0.$$

Because of the definition of A as given in (5.8) we can represent A through its rows $A_{i,\cdot}$ as

$$A_{i,\cdot} = \begin{cases} b_1(e_1^1)^T & \text{for } i = 1 \\ b_i(e_i^1 - e_{i-1}^1)^T & \text{for } i \neq 1. \end{cases}$$

First we assume that it is $i > 1$. Then it follows from the definitions

(5.12)

$$X_{i-1,\cdot} = (e_{i-1}^1)^T, \quad X_{i,\cdot} = (e_i^1)^T + \frac{b_i}{b_i + b_{i+1}}((e_{i-1}^1)^T - (e_{i+1}^1)^T), \quad X_{i+1,\cdot} = (e_{i+1}^1)^T$$

$$A_{i,\cdot} = b_i((e_i^1)^T - (e_{i-1}^1)^T), \quad \text{and} \quad A_{i+1,\cdot} = b_i((e_{i+1}^1)^T - (e_i^1)^T).$$

Based on $a_{i,j} = a_{i+1,j} = 0$ for $j \notin \{i-1, i, i+1\}$ follows from the rows $i-1, i, i+1$ of X as given above

$$(A\,X)_{i,j} = A_{i,.} \cdot X_{.,j} = 0 = A_{i+1,.} \cdot X_{.,j} = (A\,X)_{i+1,j} \quad \text{for } j \notin \{i-1, i, i+1\}.$$

Hence it is sufficient to prove

$$\begin{aligned} (A\,X)_{i,i-1} &= (A\,X)_{i+1,i-1} \\ \text{and} \quad (A\,X)_{i,i} + (A\,X)_{i,i+1} &= (A\,X)_{i+1,i} + (A\,X)_{i+1,i+1}. \end{aligned}$$

Based on the definition of X the columns $i-1, i, i+1$ of X follow as

$$\begin{aligned} X_{.,i} &= e_i, \quad X_{.,i-1} = e_{i-1} + \frac{b_i}{b_i + b_{i+1}}\, e_i + x_{i-2,i-1}\, e_{i-2} \\ X_{.,i+1} &= e_{i+1} - \frac{b_i}{b_i + b_{i+1}}\, e_i + x_{i+2,i+1}\, e_{i+2}. \end{aligned}$$

In the equation above is $x_{i-2,i} = 0$ if $\mathcal{N}^1_{i-2}$ is not aggregated with $\mathcal{N}^1_{i-1}$ and $x_{i+2,i+1} = 0$ if $\mathcal{N}^1_{i+2}$ is not aggreageted with $\mathcal{N}^1_{i+3}$. Independent of the values of $x_{i-2,i}, x_{i+2,i+1}$ follows for the rows of $(A\,X)$ that it is

$$\begin{aligned} (A\,X)_{i,.} &= A_{i,.}(X_{.,i-1}\,(e^1_{i-1})^T + X_{.,i}\,(e^1_i)^T + X_{.,i+1}\,(e^1_{i+1})^T) \\ &= \left(\frac{b_i^2}{b_i + b_{i+1}} - b_i\right)(e^1_{i-1})^T + b_i (e^1_i)^T - \frac{b_i^2}{b_i + b_{i+1}} (e^1_{i+1})^T \\ &= -\frac{b_i b_{i+1}}{b_i + b_{i+1}} (e^1_{i-1})^T + b_i (e^1_i)^T - \frac{b_i^2}{b_i + b_{i+1}} (e^1_{i+1})^T \\ \text{and} \quad (A\,X)_{i+1,.} &= A_{i+1,.}(X_{.,i-1}\,(e^1_{i-1})^T + X_{.,i}\,(e^1_i)^T + X_{.,i+1}\,(e^1_{i+1})^T) \\ &= -\frac{b_i b_{i+1}}{b_i + b_{i+1}} (e^1_{i-1})^T - b_{i+1} (e^1_i)^T + \left(\frac{b_i b_{i+1}}{b_i + b_{i+1}} + b_{i+1}\right)(e^1_{i+1})^T. \end{aligned}$$

Since for all $v_0 \in V_0$ holds $v_0(i) = v_0(i+1)$ it follows for an arbitrary $v_0 \in V_0$

$$\begin{aligned}(A\,X\,v_0)(i) &= (A\,X)_{i,.}\,v_0 \\ &= -\frac{b_i b_{i+1}}{b_i + b_{i+1}} v_0(i-1) + \left(b_i - \frac{b_i^2}{b_i + b_{i+1}}\right) v_0(i) \\ &= -\frac{b_i b_{i+1}}{b_i + b_{i+1}} v_0(i-1) + \frac{b_i b_{i+1}}{b_i + b_{i+1}} v_0(i)\end{aligned}$$

$$\begin{aligned}\text{and} \quad (A\,X\,v_0)(i+1) &= (A\,X)_{i+1,.}\,v_0 \\ &= -\frac{b_i b_{i+1}}{b_i + b_{i+1}}\, v_0(i-1) + \left(-b_{i+1} + \frac{b_i b_{i+1}}{b_i + b_{i+1}} + b_{i+1}\right) v_0(i) \\ &= -\frac{b_i b_{i+1}}{b_i + b_{i+1}}\, v_0(i-1) + \frac{b_i b_{i+1}}{b_i + b_{i+1}}\, v_0(i).\end{aligned}$$

This completes the proof for $i > 1$. The case of $i = 1$ we only have to consider if $\mathcal{N}_1^1, \mathcal{N}_2^1$ are aggregated. In this case follows, based on the same arguments as for $i > 1$, that

$$\begin{aligned}(A\,X)_{1,.} &= b_1 (e_1^1)^T - \frac{b_1^2}{b_1 + b_2} (e_2^1)^T \\ \text{and} \quad (A\,X)_{2,.} &= -b_2 (e_1^1)^T + \left(\frac{b_1 b_2}{b_1 + b_2} + b_2\right) (e_2^1)^T.\end{aligned}$$

Based on $v_0(1) = v_0(2)$ for all $v_0 \in V_0$ we obtain for an arbitrary $v_0 \in V_0$

$$\begin{aligned}(A\,X\,v_0)(1) &= (A\,X)_{1,.}\,v_0 = \left(b_1 - \frac{b_1^2}{b_1 + b_2}\right) v_0(1) = \frac{b_1 b_2}{b_1 + b_2} v_0(1) \\ \text{and} \quad (A\,X\,v_0)(2) &= (A\,X)_{2,.}\,v_0 = \left(-b_2 + \frac{b_1 b_2}{b_1 + b_2} + b_2\right) v_0(1) = \frac{b_1 b_2}{b_1 + b_2} v_0(1).\end{aligned}$$

□

From the Proposition 5.1.4 follows that with the matrix X as given in (5.11) V_0 is invariant with respect to $A\,X$. From Lemma 4.1.3 follows therewith that the angle $\gamma_{DT,X}$ is zero. We know from section 4.1 that this is the best possible result.
To conclude the discussion of this modification we will show that the structure of the coarse grid operators $A_{0,X}$ is an operator of the same structure as A_0 also for the arbitrary big system. To show this we will show the structure of A_0 and $A_{0,X}$. W.l.o.g. we assume that we have for the prolongation matrix P the following order condition:

$$p_{i,j} \neq 0 \quad \Rightarrow \quad p_{s,t} = 0, \quad \forall\, s \geq i,\, \forall\, t < j. \tag{5.13}$$

To illustrate this assumption we assume that the structure of P is given for example as follows:

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad P \neq \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

This is always possible to reach by means of a permutation matrix Π we use. This transforms P to $\Pi\, P\, \Pi^T = \widehat{P}$ with the assumed structure. So the permutation matrix just represents another numeration on the coarse grid points. Furthermore, it is obvious based on the structure of A that

$$(e_i^1)^T\, A = b_i (e_i^1 - e_{i-1}^1)^T \quad \text{holds for all } i \geq 2$$

$$\text{and} \quad (e_1^1)^T\, A = b_1\, (e_1^1)^T.$$

We will use this property to prove the propositions concerning the structure of $A_0, A_{0,X}$.

Lemma: 5.1.5. *Let A be a matrix as defined in (5.8), R a restriction operator as defined in (5.9) so that $P = R^T$ fulfils the condition (5.13). Let X be the modification matrix as given in (5.11). Then it follows:*

1. *For $P\, e_k^0 = e_i^1$ or $P\, e_k^0 = e_i^1 + e_{i+1}^1$ we obtain for $i > 1$*

$$a_{k,t}^0 = \begin{cases} b_i & \text{if} \quad t = k \\ -b_i & \text{if} \quad t+1 = k \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad a_{1,t}^0 = \begin{cases} b_1 & \text{if} \quad t = 1 \\ 0 & \text{otherwise} \end{cases}$$

   for $i = 1$.

2. *For $P\, e_k^0 = e_i^1$ we obtain for $i > 1$*

$$a_{k,t}^{0,X} = \begin{cases} b_i & \text{if} \quad t = k \\ -b_i & \text{if} \quad t+1 = k \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad a_{1,t}^{0,X} = \begin{cases} b_1 & \text{if} \quad t = 1 \\ 0 & \text{otherwise} \end{cases}$$

   for $i = 1$.

3. *For* $P\,e_k^0 = e_i^1 + e_{i+1}^1$ *we obtain for* $i > 1$

$$a_{k,t}^{0,X} = \begin{cases} \frac{2b_i\,b_{i+1}}{b_i+b_{i+1}} & \text{if} \quad t = k \\ -\frac{2b_i\,b_{i+1}}{b_i+b_{i+1}} & \text{if} \quad t+1 = k \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad a_{1,t}^{0,X} = \begin{cases} \frac{b_1 b_2}{b_1+b_2} & \text{if} \quad t = 1 \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1$.

proof. For all propositions we only prove the proposition for the case of $P\,e_k^0 = e_i^1$ or $P\,e_k^0 = e_i^1 + e_{i+1}^1$ with $i > 1$. For $i = 1$ it follows $k = 1$ from the condition (5.13). Then the prove always follows the same arguments as for $i > 1$.

1. First we assume that we have $P\,e_k^0 = e_i^1$. As it is generally

$$\begin{aligned} a_{k,t}^0 &= (e_k^0)^T\,A_0\,e_t^0 = (e_k^0)^T\,R\,A\,P\,e_t^0 = (e_k^0)^T\,P^T\,A\,P\,e_t^0 \\ &= (P\,e_k^0)^T\,A\,(P\,e_t^0) = (e_i^1)^T\,A\,(P\,e_t^0) \\ &= b_i(e_i^1 - e_{i-1}^1)^T\,(P\,e_t^0) \end{aligned}$$

we obtain for $t = k$

$$a_{k,k}^0 = b_i(e_i^1 - e_{i-1}^1)^T\,e_i^1 = b_i.$$

For $t = k-1$ follows by the assumption (5.13) on the structure of P

$$\begin{aligned} P\,e_t^0 = P\,e_{k-1}^0 = e_{i-1}^1 \quad &\text{if } \mathcal{N}_{i-1} \text{ is isolated, or} \\ P\,e_t^0 = P\,e_{k-1}^0 = e_{i-1}^1 + e_{i-2}^1 \quad &\text{if } \mathcal{N}_{i-1}, \mathcal{N}_{i-2} \text{ are aggregated.} \end{aligned}$$

In both cases follows the propositon for $a_{k,k-1}^0$ that is

$$\begin{aligned} a_{k,k-1}^0 &= b_i(e_i^1 - e_{i-1}^1)^T\,(P\,e_{k-1}^0) = b_i(e_i^1 - e_{i-1}^1)^T\,e_{i-1}^1 = -b_i \\ \text{or} \quad a_{k,k-1}^0 &= b_i(e_i^1 - e_{i-1}^1)^T\,(P\,e_{k-1}^0) = b_i(e_i^1 - e_{i-1}^1)^T\,(e_{i-1}^1 + e_{i-2}^1) = -b_i. \end{aligned}$$

For $t < k-1$ it follows from the condition (5.13)

$$P\,e_t^0 = e_j^1 \quad \text{or} \quad P\,e_t^0 = e_j^1 + e_{j-1}^1 \quad \text{with} \quad j \le i-2.$$

This implies

$$a_{k,t}^0 = b_i(e_i^1 - e_{i-1}^1)^T\,e_j^1 = b_i(e_i^1 - e_{i-1}^1)^T\,(e_j^1 + e_{j-1}^1) = 0 \quad \text{for} \quad t < k-1.$$

Similary it follows for $t > k$ from the condition (5.13)

$$P\,e_t^0 = e_j^1 \quad \text{or} \quad P\,e_t^0 = e_j^1 + e_{j+1}^1 \quad \text{with} \quad j \geq i+1.$$

And this implies

$$a_{k,t}^0 = b_i(e_i^1 - e_{i-1}^1)^T\,e_j^1 = b_i(e_i^1 - e_{i-1}^1)^T\,(e_j^1 + e_{j+1}^1) = 0 \quad \text{for} \quad t \geq k+1.$$

Then we assume that we have $P\,e_k^0 = e_i^1 + e_{i+1}^1$. Based on the same arguments as above we obtain

$$a_{k,t}^0 = (e_k^0)^T\,A_0\,e_t^0 = \left(b_i(e_i^1 - e_{i-1}^1)^T + b_{i+1}(e_{i+1}^1 - e_i^1)^T\right)(P\,e_t^0)$$

we obtain for $t = k$

$$a_{k,k}^0 = \left(b_i(e_i^1 - e_{i-1}^1)^T + b_{i+1}(e_{i+1}^1 - e_i^1)^T\right)(e_i^1 + e_{i+1}) = b_i.$$

For $t = k-1$ follows again by the assumption (5.13) on the structure of P

$$P\,e_t^0 = P\,e_{k-1}^0 = e_{i-1}^1 \quad \text{if } \mathcal{N}_{i-1} \text{ is isolated, or}$$

$$P\,e_t^0 = P\,e_{k-1}^0 = e_{i-1}^1 + e_{i-2}^1 \quad \text{if } \mathcal{N}_{i-1}, \mathcal{N}_{i-2} \text{ are aggregated.}$$

In both cases follows the propositon for $a_{k,k-1}^0$ that is

$$\begin{aligned} a_{k,k-1}^0 &= \left(b_i(e_i^1 - e_{i-1}^1)^T + b_{i+1}(e_{i+1}^1 - e_i^1)^T\right)(P\,e_{k-1}^0) \\ &= \left(b_i(e_i^1 - e_{i-1}^1)^T + b_{i+1}(e_{i+1}^1 - e_i^1)^T\right)\,e_{i-1}^1 = -b_i \\ \text{or} \quad a_{k,k-1}^0 &= \left(b_i(e_i^1 - e_{i-1}^1)^T + b_{i+1}(e_{i+1}^1 - e_i^1)^T\right)(P\,e_{k-1}^0) \\ &= \left(b_i(e_i^1 - e_{i-1}^1)^T + b_{i+1}(e_{i+1}^1 - e_i^1)^T\right)(e_{i-1}^1 + e_{i-2}^1) = -b_i. \end{aligned}$$

For $t < k-1$ and $t > k$ the assertion follows as in the case of $P\,e_k^0 = e_i^1$. This shows the proposition about the structure of A_0.

2. We assume again that we have $P\,e_k^0 = e_i^1$ with $i > 1$. Again we will consider the elements of the k-th row of $A_{0,X}$. We obtain

$$\begin{aligned} a_{k,t}^{0,X} &= (e_k^0)^T\,A_{0,X}\,e_t^0 = (P\,e_k^0)^T\,A\,X\,(P\,e_t^0) = (e_i^1)^T\,A\,X\,(P\,e_t^0) \\ &= b_i(e_i^1 - e_{i-1}^1)^T\,X\,(P\,e_t^0). \end{aligned}$$

Based on the assumption that $\mathcal{N}_i^1$ is an isolated point we obtain for the i-th and the $(i-1)$-th row of X

$$X_{i,\cdot} = (e_i^1)^T \quad \text{and} \quad X_{i-1,\cdot} = (e_{i-1}^1)^T \quad \text{and therewith}$$
$$(e_i^1 - e_{i-1}^1)^T X = (e_i^1 - e_{i-1}^1)^T.$$

So it follows

$$a_{k,t}^{0,X} = b_i (e_i^1 - e_{i-1}^1)^T (P\, e_t^0)$$

and we have shown in the first part of the proof that this is b_i for $t = k$, $-b_i$ for $t = k-1$ and zero otherwise.

3. Now we assume that we have $P\, e_k^0 = e_i^1 + e_{i+1}^1$ with $i > 1$. Then it follows for $A_{0,X}$

$$\begin{aligned} a_{k,t}^{0,X} &= (e_k^0)^T A_{0,X}\, e_t^0 = (P\, e_k^0)^T A\, X\, (P\, e_t^0) = (e_i^1 + e_{i+1}^1)^T A\, X\, (P\, e_t^0) \\ &= b_i\, (e_i^1 - e_{i-1}^1)^T X\, (P\, e_t^0) + b_{i+1}\, (e_{i+1}^1 - e_i^1)^T X\, (P\, e_t^0) \\ &= (b_i - b_{i+1})\, (e_i^1)^T X\, (P\, e_t^0) + (b_{i+1}\, (e_{i+1}^1)^T - b_i\, (e_{i-1}^1)^T) 0\, X\, (P\, e_t^0). \end{aligned}$$

As $\mathcal{N}_i^1, \mathcal{N}_{i+1}^1$ are aggregated it follows for the rows of X

$$X_{i-1,\cdot} = (e_{i-1}^1)^T, \quad X_{i+1,\cdot} = (e_{i+1}^1)^T$$
$$\text{and} \quad X_{i,\cdot} = (e_i^1)^T + \frac{b_i}{b_i + b_{i+1}} (e_{i-1}^1 - e_{i+1}^1)^T.$$

Hence we obtain

$$(e_{i-1}^1)^T X = X_{i-1,\cdot} = (e_{i-1}^1)^T, \quad (e_{i+1}^1)^T X = X_{i+1,\cdot} = (e_{i+1}^1)^T$$
$$(e_i^1)^T X = X_{i,\cdot} = (e_i^1)^T + \frac{b_i}{b_i + b_{i+1}} (e_{i-1}^1 - e_{i+1}^1)^T.$$

Therewith we get for $a_{k,t}^{0,X}$

$$\begin{aligned} a_{k,t}^{0,X} &= (b_i - b_{i+1}) \left((e_i^1)^T + \frac{b_i}{b_i + b_{i+1}} (e_{i-1}^1 - e_{i+1}^1)^T \right) (P\, e_t^0) \\ &\quad + \left(b_{i+1}\, (e_{i+1}^1)^T - b_i\, (e_{i-1}^1)^T \right) (P\, e_t^0) \\ &= \frac{b_i^2 - b_{i+1}^2}{b_i + b_{i+1}} (e_i^1)^T P\, e_t^0 - \frac{2 b_i\, b_{i+1}}{b_i + b_{i+1}} (e_{i-1}^1)^T P\, e_t^0 \\ &\quad + \frac{-b_i^2 + b_{i+1}^2 + 2 b_i b_{i+1}}{b_i + b_{i+1}} (e_{i+1}^1)^T P\, e_t^0. \end{aligned}$$

In the case of $t = k$ it follows $P\, e_t^0 = e_i^1 + e_{i+1}^1$ and therewith that

$$\begin{aligned} a_{k,k}^{0,X} &= \frac{b_i^2 - b_{i+1}^2}{b_i + b_{i+1}} (e_i^1)^T \, (e_i^1 + e_{i+1}^1) - \frac{2b_i\, b_{i+1}}{b_i + b_{i+1}} (e_{i-1}^1)^T \, (e_i^1 + e_{i+1}^1) \\ &\quad + \frac{-b_i^2 + b_{i+1}^2 + 2b_i b_{i+1}}{b_i + b_{i+1}} (e_{i+1}^1)^T \, (e_i^1 + e_{i+1}^1) \\ &= \frac{b_i^2 - b_{i+1}^2}{b_i + b_{i+1}} + 0 + \frac{-b_i^2 + b_{i+1}^2 + 2b_i b_{i+1}}{b_i + b_{i+1}} = \frac{2b_i b_{i+1}}{b_i + b_{i+1}}. \end{aligned}$$

In the case of $t = k - 1$ we get

$$P\, e_{k-1}^0 = e_{i-1}^1 \quad \text{or} \quad P\, e_{k-1}^0 = e_{i-1}^1 + e_{i-2}^1.$$

In both cases it follows

$$\begin{aligned} a_{k,k-1}^{0,X} &= \frac{b_i^2 - b_{i+1}^2}{b_i + b_{i+1}} (e_i^1)^T \, e_{i-1}^1 - \frac{2b_i\, b_{i+1}}{b_i + b_{i+1}} (e_{i-1}^1)^T \, e_{i-1}^1 \\ &\quad + \frac{-b_i^2 + b_{i+1}^2 + 2b_i b_{i+1}}{b_i + b_{i+1}} (e_{i+1}^1)^T \, e_{i-1}^1 \\ &= -\frac{2b_i b_{i+1}}{b_i + b_{i+1}}. \end{aligned}$$

For $t \leq k - 2$ it follows again

$$P\, e_t^0 = e_j^1 \quad \text{or} \quad P\, e_t^0 = e_j^1 + e_{j-1}^1 \quad \text{with} \quad j \leq i - 2$$

and for $t \geq k + 1$

$$P\, e_t^0 = e_j^1 \quad \text{or} \quad P\, e_t^0 = e_j^1 + e_{j+1}^1 \quad \text{with} \quad j \geq i + 2.$$

This proves again $a_{k,t}^{0,X} = 0$ for $t \neq k, k - 1$.

□

To illustrate the assertion of Lemma 5.1.5 we give the following example: Assume that the grids are structured as in Figure 5.3. Then it results $A_0, A_{0,X}$ as follows:

$$A_0 = \begin{pmatrix} b_1 & & & \\ -b_3 & b_3 & & \\ & -b_4 & b_4 & \\ & & -b_6 & b_6 \end{pmatrix} \quad A_{0,X} = \begin{pmatrix} \frac{2\,b_1 b_2}{b_1 + b_2} & & & \\ -b_3 & b_3 & & \\ & -\frac{2\,b_4 b_5}{b_4 + b_5} & \frac{2\,b_4 b_5}{b_4 + b_5} & \\ & & -\frac{2\,b_6 b_7}{b_6 + b_7} & \frac{2\,b_6 b_7}{b_6 + b_7} \end{pmatrix}$$

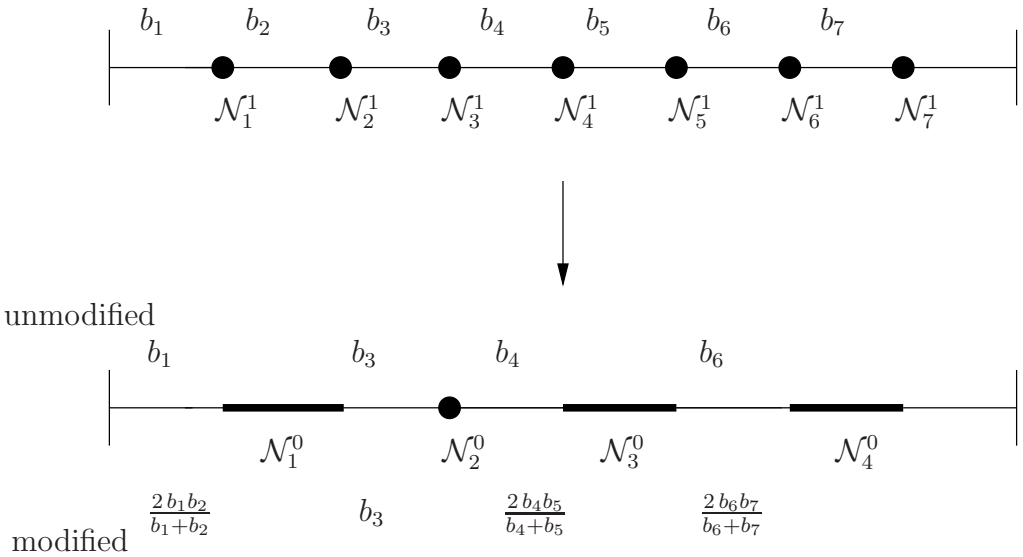


Figure 5.3: Unmodified and modified links in the coarse grid operator

To conclude this section we highlight that the modification again fulfils

$$rk(P_X) = n_0.$$

This is obtained as it follows

$$(P_X)_{j,\cdot} = (e_t^0)^T$$

if $\mathcal{N}_j^1 \subset \mathcal{N}_t^0$ is an isolated point or $\mathcal{N}_i^1, \mathcal{N}_j^1$ are aggregated to $\mathcal{N}_t^0$. This implies that n_0 rows of P_X are given by the n_0 unit basis vectors of $\mathbb{R}^{n_0}$.

Modification based on the inverse of blocks

Next we will consider the idea to consider the aggregated points as blocks which are independent of the rest of the system. So we are back in the situation as given in section 5.1.1 and consider the system of four points. The stiffness matrix of our intrest is still

$$A = \begin{pmatrix} b_1 & 0 & 0 & 0 \\ -b_2 & b_2 & 0 & 0 \\ 0 & -b_3 & b_3 & 0 \\ 0 & 0 & -b_4 & b_4 \end{pmatrix}. \tag{5.14}$$

Based on the structure of R as defined in (5.9) we consider the blocks

$$B_1 = (b_1), \quad B_2 = \begin{pmatrix} b_2 & 0 \\ -b_3 & b_3 \end{pmatrix}, \quad B_3 = (b_4).$$

Then we will set X as the inverse of blocks given by

$$X = \begin{pmatrix} B_1^{-1} & & \\ & B_2^{-1} & \\ & & B_3^{-1} \end{pmatrix}. \tag{5.15}$$

As it is

$$B_1^{-1} = \begin{pmatrix} \frac{1}{b_1} \end{pmatrix}, \quad B_2^{-1} = \begin{pmatrix} 1/b_2 & 0 \\ 1/b_2 & 1/b_3 \end{pmatrix}, \quad B_3^{-1} = \begin{pmatrix} \frac{1}{b_4} \end{pmatrix}$$

we obtain

$$A\,X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -b_2/b_1 & 1 & 0 & 0 \\ 0 & -b_3/b_2 & 1 & 0 \\ 0 & 0 & -b_4/b_3 & 0 \end{pmatrix} \quad \text{and} \quad A_{0,X} = \begin{pmatrix} 1 & 0 & 0 \\ -b_2/b_1 & 2 - b_3/b_2 & 0 \\ 0 & -b_3/b_4 & 1 \end{pmatrix}.$$

First we highlight that for this modification we obtain $rk(X) = n$. This implies $rk(P_X) = n_0$.

Compared with the modification we have done bevor its obvious that this idea is simpler to implement in a numerical algorithm. But the problems of this modification are quite obvious.

1. The matrix $A_{0,X}$ can be singular. This is for example the case if it is $b_3 = 2b_2$.
2. For $b_3 > 2b_2$ the matrix $A_{0,X}$ is non singular but it is obvious that $A_{0,X}$ is no M-matrix in this case. (For a closer look at the characteristics of $A_{0,X}$ concerning M-matrices, cf. chapter 9.)
3. At last there is no local estimation for $\gamma_{DT,X} < 1$ that fulfils

$$(A\,X\,P\,A_{0,X}^{-1}\,R\,v,\,(I - Q_0)\,v) \leq \gamma_{DT,X}\,\|A\,X\,P\,A_{0,X}^{-1}\,R\,v\|\,\|(I - Q_0)\,v\|$$

for all $v \in V$. This can be seen as follows: As already mentioned is $R\,v = R\,Q_0\,v$. Hence the inequality above is equivalent to

$$(A\,X\,v_0,\, w) \leq \gamma_{DT,X}\,\|A\,X\,v_0\|\,\|w\|$$

for all $v_0 \in V_0$ and all $w \in W$. Since we only want to consider the local situation of the two aggregated points this is equivalent to

$$\begin{aligned}&((A\,X\,v_0)(2)\,w(2) + (A\,X\,v_0)(3)\,w(3))^2\\ &\qquad \leq \gamma_{DT,X}\,\left([(A\,X\,v_0)(2)]^2 + [(A\,X\,v_0)(3)]^2\right)(w(2)^2 + w(3)^2).\end{aligned}$$

For $v_0 = (f, u, u, g),\ f, u, g \in \mathbb{R}$ and $w = (0, s, -s, 0),\ s \in \mathbb{R}$ this is equivalent to

(5.16)
$$\left[\left(u - f\frac{b_2}{b_1}\right) - u\left(1 - \frac{b_3}{b_2}\right)\right]^2 s^2 \leq \gamma_{DT,X}^2(2s^2)\left[\left(u - f\frac{b_2}{b_1}\right)^2 + u^2\left(1 - \frac{b_3}{b_2}\right)^2\right]$$

Then we see that for

$$f = \left(2 - \frac{b_3}{b_2}\right)\frac{b_1}{b_2}\,u$$

the inequality (5.16) is only fulfilled for $\gamma_{DT,X} = 1$.

Of course there are some reasons for this modification. First of all the modification is quite simple and we can use it for many systems. The only assumption we need is that the blocks we get are not singular and that the modified coarser operator $A_{0,X}$ is not singular. The second one is that the effort of this modification is quite small since we modify for aggregated points $\mathcal{N}_i^1, \mathcal{N}_j^1$ only the values $(P\,v_0)(i)$ and $(P\,v_0)(j)$. Hence there is no additional effort to search other neighbours and modify the values for them. At least there is the idea that the matrix A is mainly given by blocks and other links are weak. Then B is a good approximation for A. We will see that this idea better suits for the symmetric problems. In particular for the two sided modification for symmetric problems. The problem of the modification for the convection system is obvious as in the example above the value on $\mathcal{N}_2^1$ is mainly given by the the value on $\mathcal{N}_1^1$. The interpretation as blocks implicates that the value on $\mathcal{N}_2^1$ mainly depends on the value on $\mathcal{N}_3^1$.

5.1.2 Modifications for a two dimensional convection

Now we will consider a two dimensional convection system. As the block inversion offers problems for the convection system already for the one dimensional case, we only want to consider for this example the modification we have introduced as an exact modification in section 5.1.1.
So we will consider a problem given as follows: The matrix $A \in \mathbb{R}^{n\times n}$ is given as

$$\begin{aligned} A = b_{i,j} \quad \text{with} \quad b_{i,i} > 0, \quad b_{i,j} \leq 0, \quad \text{for} \quad i \neq j \\ b_{i,i} \geq \sum_{j\neq i} |b_{i,j}| \quad \text{for} \quad i = 1, \ldots, n \\ \text{and} \quad b_{i,j} \neq 0 \quad \Rightarrow \quad b_{j,i} = 0 \quad \text{for} \quad i \neq j. \end{aligned} \tag{5.17}$$

The matrix in (5.17) represents a convection system of two dimensions (or higher). This motivates the condition $b_{i,j} \neq 0 \Rightarrow b_{j,i} = 0$. So there is neither a convection from $\mathcal{N}_i^1$ to $\mathcal{N}_j^1$ nor vice versa, but both directions in one system are meaningless.

To define the modification X, we first define the set M_0 of indices as follows

$$M_0(i) := \big\{ t \in \{1, \ldots, n\} \backslash \{i\} \, : \, b_{i,t} \neq 0. \big\}.$$

Based on the interpretation as a convection system, $M_0(i)$ is the set of the indices of the grid points $\mathcal{N}_t^1$ that have an influence on $\mathcal{N}_i^1$.

Then we define our modification matrix $X \in \mathbb{R}^{n\times n}$ also by its rows as

$$X_{i,\cdot} = \begin{cases} (e_i^1)^T & \text{if } \mathcal{N}_i^1 \text{ is isolated or } \mathcal{N}_i^1 \text{ is aggregated} \\ & \text{with } \mathcal{N}_j^1 \text{ and it is } b_{i,j} \neq 0. \\ (e_i^1)^T + x_{i,j}(e_j^1)^T + x_{i,k}(e_k^1)^T & \text{if } \mathcal{N}_i^1, \mathcal{N}_j^1 \text{ are aggregated, it is } b_{i,j} = 0 \\ & \text{and it is } k \in M_0(i). \end{cases} \tag{5.18}$$

If we aggregate two points $\mathcal{N}_i^1, \mathcal{N}_j^1$ with $b_{i,j} \neq 0$ and it is $M_0(i) = \emptyset$ then we set $x_{i,k} = 0$. If it is $|M_0(i)| > 1$ then we choose just one of the indices. Based on the idea to reduce the influence of grid points which only influence one of the two points $\mathcal{N}_i^1, \mathcal{N}_j^1$, it is a feasible heuristic to choose an index $k \in M_0(i)$ that holds

$$|b_{i,k}| \geq |b_{i,s}| \quad \forall s \in M_0(i).$$

We want to consider examples for $x_{i,j}, x_{i,k}$ that will give results similar to those we had for the one dimensional situation and we want to show to what extent this is a generalization of the modification given in this context. First we define for an arbitrary grid point $\mathcal{N}_k^1$ or row $A_{k,.}$ of A respectively the set $M(k)$ as follows

(5.19)
$$M(k) = \big\{ t \in \{1,\ldots,n\}\backslash\{k\} \,:\, \mathcal{N}_t^1, \mathcal{N}_s^1,\ s \neq k, \text{ are aggregated and it is } x_{t,k} \neq 0. \big\}.$$

Based on this definition follows that it is $t \in M(k)$ if it is $x_{t,k} \neq 0$. That implies $\mathcal{N}_t^1$ is aggregated with another grid point $\mathcal{N}_s^1$ and it is $b_{t,k} \neq 0$. Further is this $\mathcal{N}_k^1$ used to modify the situation for $\mathcal{N}_t^1, \mathcal{N}_s^1$. Based on the definitions of $M(k), X, A$ we obtain

$$t \in M(k) \quad \Rightarrow \quad a_{k,t} = 0. \tag{5.20}$$

This implication holds as we have $x_{t,k} \neq 0$ if it is $t \in M(k)$ from the definition of $M(k)$. The definition of X implies $b_{t,k} \neq 0$ and therewith follows $b_{k,t} = 0$ from the definition of A.

We obtain the following result:

Proposition: 5.1.6. *Let A, X be matrices as defined in (5.17),(5.18). Let $\mathcal{N}_i^1, \mathcal{N}_j^1$ be two aggregated points with $b_{j,i} \neq 0$. Let $\mathcal{N}_k^1$ be a grid point with $b_{i,k} \neq 0$.*

1. *If it is $b_{j,k} = 0$ and it is $b_{i,t} = b_{j,t} = 0$ for all $t \in M(k)\backslash\{i\}$ and we define $x_{i,k} = \frac{|b_{i,k}|}{b_{i,i}+|b_{j,i}|}$ then*

$$(A\,X)_{i,k} = (A\,X)_{j,k} = \frac{b_{j,i}|b_{i,k}|}{b_{i,i} + |b_{j,i}|} \quad \textit{holds.}$$

2. *If it is $b_{j,t} = 0$ for all t with $t \in M(i)$, $b_{i,t} = 0$ for all t with $t \in M(j)$ and we define $x_{i,j} = \frac{b_{j,j}-b_{i,i}+b_{j,i}}{b_{i,i}+|b_{j,i}|}$ then*

$$(A\,X)_{i,i} + (A\,X)_{i,j} = (A\,X)_{j,j} + (A\,X)_{j,i} = \frac{b_{i,i}b_{j,j}}{b_{i,i} + |b_{j,i}|} \quad \textit{holds.}$$

proof. Based on the assumption of $b_{j,i} \neq 0$ and the definition of X it follows that we have $X_{j,.} = (e_j^1)^T$ and $X_{i,.} = (e_i^1)^T + x_{i,k}(e_k^1)^T + x_{i,j}(e_j^1)^T$.

1. We obtain that the $k-$th column of X follows as

$$X_{.,k} = e_k^1 + \sum_{t=M(k)}^{n} x_{t,k} e_t^1 = e_k^1 + x_{i,k} e_i^1 + \sum_{t=M(k)\backslash\{i\}}^{n} x_{t,k} e_t^1$$

Hence we have based on the assumption $b_{i,t} = b_{j,t} = 0$ for all $t \in M(k)\backslash\{i\}$

$$(A\,X)_{i,k} = A_{i,.}\,X_{.,k} = b_{i,i}x_{i,k} + b_{i,k}$$

$$(A\,X)_{j,k} = A_{j,.}\,X_{.,k} = b_{j,i}x_{i,k}.$$

The second equation thereby results from the condition $b_{j,k} = 0$. This implies that they are equal if it is

$$x_{i,k} = \frac{|b_{i,k}|}{b_{i,i} + |b_{j,i}|}.$$

And we obtain

$$(A\,X)_{j,k} = \frac{b_{j,i}|b_{i,k}|}{b_{i,i} + |b_{j,i}|} = (A\,X)_{i,k}.$$

2. Based on the assumption that $\mathcal{N}_i^1, \mathcal{N}_j^1$ are aggregated with $b_{j,i} \neq 0$ it follows for the columns i, j of X

$$X_{.,i} = e_i^1 + \sum_{t \in M(i)} x_{t,i}e_t^1$$

$$\text{and} \quad X_{.,j} = e_j^1 + x_{i,j}e_i^1 + \sum_{t \in M(j)} x_{t,j}e_t^1.$$

Furthermore we obtain from the assumption $b_{j,t} = 0$ for all t with $t \in M(i)$, $b_{i,t} = 0$ for all t with $t \in M(j)$ and the implication (5.20) that we have $b_{i,t} = b_{j,t} = 0$ for all t with $t \in M(i) \cup M(j)$. This implies

$$\begin{aligned}(A\,X)_{i,i} + (A\,X)_{i,j} &= A_{i,.}X_{.,i} + A_{i,.}X_{.,j} \\ &= b_{i,i} + b_{i,i}x_{i,j} = (1 + x_{i,j})b_{i,i} \\ (A\,X)_{j,i} + (A\,X)_{j,j} &= A_{j,.}X_{.,i} + A_{j,.}X_{.,j} \\ &= b_{j,i} + x_{i,j}b_{j,i} + b_{j,j} = (1 + x_{i,j})b_{j,i} + b_{j,j}.\end{aligned}$$

So this is equal if we have

$$x_{i,j} = \frac{b_{j,j} - b_{i,i} + b_{j,i}}{b_{i,i} + |b_{j,i}|}.$$

And we obtain in this case

$$\begin{aligned}(A\,X)_{i,i} + (A\,X)_{i,j} &= (1 + x_{i,j})b_{i,i} \\ &= \frac{b_{i,i}\big(b_{j,j} - b_{i,i} + b_{j,i} + b_{i,i} + |b_{j,i}|\big)}{b_{i,i} + |b_{j,i}|} = \frac{b_{i,i}b_{j,j}}{b_{i,i} + |b_{j,i}|}.\end{aligned}$$

□

First we will take a closer look on the assertion of the Proposition 5.1.6 then we will consider the assumptions. To ensure that V_0 is invariant with respect to $A\,X$ we have to prove for two aggregated points $\mathcal{N}_i^1, \mathcal{N}_j^1$ that

$$(A\,X\,v_0)(i) = (A\,X\,v_0)(j) \quad \forall v_0 \in V_0.$$

The result of Proposition 5.1.6 is that we have

$$(A\,X\,v_0)(i) = (A\,X\,v_0)(j)$$
$$\text{for} \quad v_0 = (e_i^1 + e_j^1) \quad \text{and} \quad v_0 = e_k^1.$$

Therewith we do not have the invariance of V_0 with respect to $A\,X$, but we are a little bit closer to this as in the unmodified method.

Further we obtain that the values for

$$(A\,X)_{i,i}, (A\,X)_{i,j}, (A\,X)_{i,k} \quad \text{and} \quad (A\,X)_{j,i}, (A\,X)_{j,j}, (A\,X)_{j,k},$$

respectively are as given in the one dimensional system. Hence we obtain a good modification if the system is mainly a one dimensional system.

So we will take a look at the assumptions we have in the last proposition and what kind of convection system can be described by them. If we have a two dimensional convection system then we assume the stencil in $\mathcal{N}_i^1$ given as

$$\begin{pmatrix} 0 & 0 & 0 \\ -b_{i,x} & b_{i,x} + b_{i,y} & 0 \\ 0 & -b_{i,y} & 0 \end{pmatrix} \quad \text{with} \quad b_{i,x}, b_{i,y} \geq 0.$$

This means that the convection locally has two directions, and this does not change its direction. So we assume that after a permutation of rows and columns it is $k = i - 1$ and $j = i + 1$. Then the rows k, i, j of A are given as

$$A = \begin{pmatrix} \dots & -b_{k,y} & \dots & & -b_{k,x} & b_{k,x} + b_{k,y} & \dots & & \\ \dots & & -b_{i,y} & \dots & & -b_{i,x} & b_{i,x} + b_{i,y} & \dots & \\ \dots & & & -b_{j,y} & \dots & & -b_{j,x} & b_{j,x} + b_{j,y} & \dots \end{pmatrix}.$$

So if we aggregate the points $\mathcal{N}_i^1$ and $\mathcal{N}_j^1$ and all other points are isolated it is obvious that the assumtions of Proposition 5.1.6 are fulfilled. This is illustrated in Figure 5.4

(a). Moreover, there is no restriction for the direction of the links, so the situation as given in Figure 5.4 (b) also fulfils the condition. Furhtermore, the assumptions also hold if $\mathcal{N}_k^1$ is aggregated with its left neighbour $\mathcal{N}_{k-1}^1$ and we use $\mathcal{N}_l^1$ to modify this aggregation. Then the assumptions of the Proposition 5.1.6 are fulfilled. This follows as we have $b_{k,k-1} \neq 0$ and so only the $(k-1)$-th row of X is modified and it is $b_{i,k-1} = b_{j,k-1} = 0$ (Figure 5.4 (c)). Based on the same arguments the assumtions hold in a situation as shown in Figure 5.4 (d). The situation illustrated in (e) also does not infringe the assumptions if $\mathcal{N}_l^1$ is used to modifiy the aggregation between $\mathcal{N}_s^1, \mathcal{N}_t^1$. It is $b_{i,t} \neq 0$ and $b_{j,s} \neq 0$, but the aggregation of $\mathcal{N}_s^1, \mathcal{N}_t^1$ implies a modification of the t-th row of X. There are only the entries $x_{t,l}$ and $x_{t,s}$ that are modified. This changes values in the i-th and the j-th row of $(A\,X)$. But the entries $(A\,X)_{i,i}, (A\,X)_{i,j}, (A\,X)_{i,k}$ and $(A\,X)_{j,i}, (A\,X)_{j,j}, (A\,X)_{j,k}$ do not depend on this. In (f) $\mathcal{N}_s^1$ is aggregated with $\mathcal{N}_t^1$ and $b_{s,t} \neq 0$. So from the definition of X follows that the t-th row of X is modified. In the situation of (f) it is $x_{t,k}$ modified and $t \in M(k)$. This infringes the assumptions for the first assertion of Proposition 5.1.6. Obviously the situation (g) infringes the assumption $b_{j,k} = 0$ and hence also a condition for the first assertion of the proposition. At last we will consider an example that infringes the assumptions of the second assertion, but this is not possible based on the given situation of a locally unique direction of the convection. So we have to construct the example as shown in (h) (The direction of the arrows give the direction of the convection in this case). $\mathcal{N}_i^1$ is used to modifiy the aggregation between $\mathcal{N}_t^1, \mathcal{N}_s^1$. This implies $x_{t,s}, x_{t,i} \neq 0$. Hence it follows from the definition $t \in M(i)$. As it is $b_{j,i} \neq 0$ this infringes the assumptions.

We can summarize this as follows: The assumptions are weaker than they seem at first sight. Especially for the second assertion they are always fulfilled if the directions of the convection are locally unique. If the convection is only one dimensional in a two (or three) dimensional system then it follows for example $b_{i,y} = 0$ for all $i = 1, \ldots, n$. Also the assumptions for the first assertion of Proposition 5.1.6 are therefore always fulfilled. Furthermore, we can see that in the one dimensional convection system as given in section 5.1.1 with the modification as given in (5.11) has the same structure as a modification given by the conditions of Proposition 5.1.6. We will prove this in the next lemma.

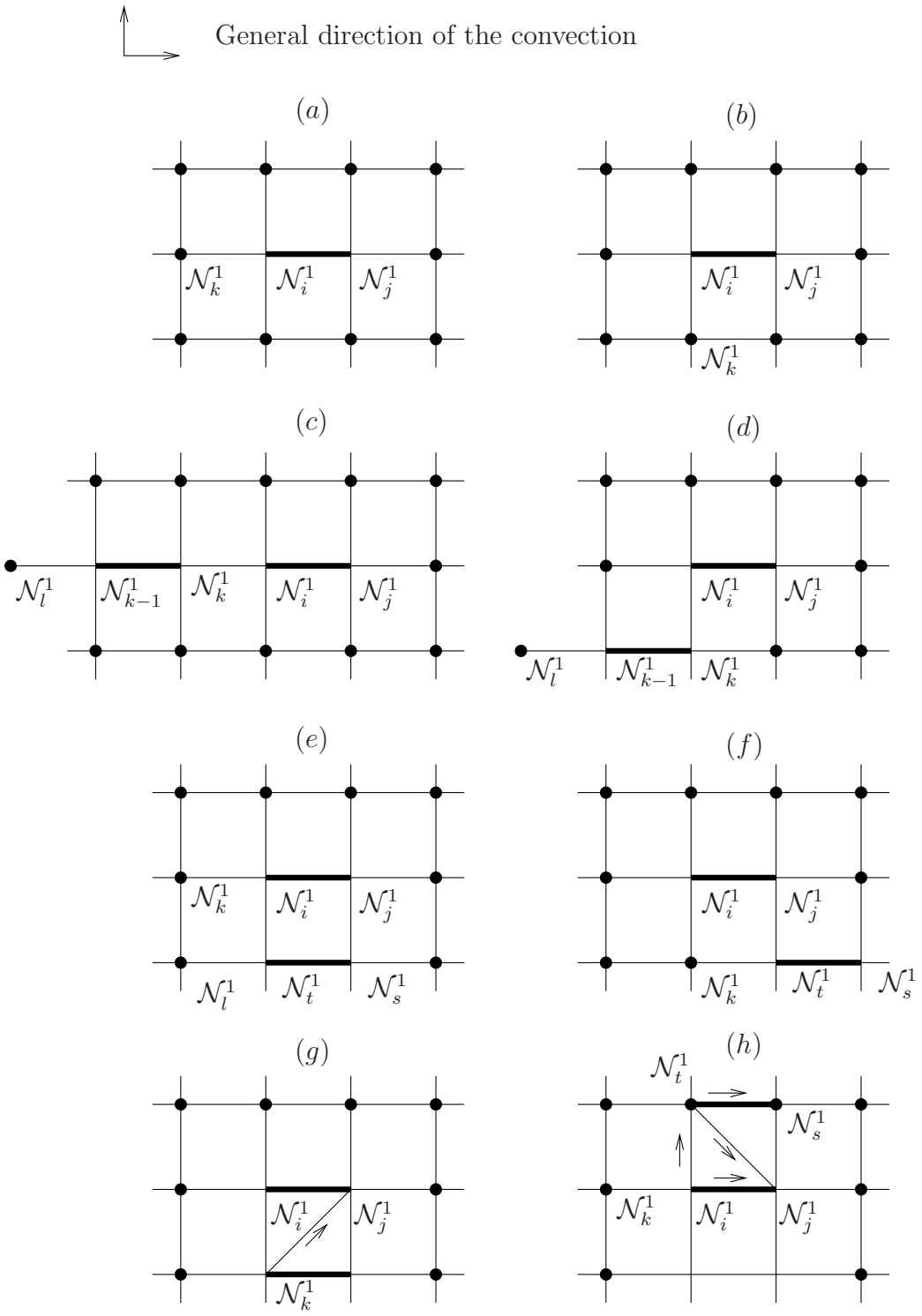


Figure 5.4: Illustration of the assumptions of Proposition 5.1.6

Lemma: 5.1.7. *Let A, X be as given in (5.8), (5.11). Then for two aggregated points $\mathcal{N}_i^1, \mathcal{N}_{i+1}^1$ the modification holds the structure as given in (5.18) with*

$$x_{i,i-1} = \frac{|b_{i,i-1}|}{b_{i,i} + |b_{i+1,i}|} \quad \textit{and} \quad x_{i,i+1} = \frac{b_{i+1,i+1} - b_{i,i} + b_{i+1,i}}{b_{i,i} + |b_{i+1,i}|}.$$

proof. The modification as given in (5.11) obviously has the same structure as the modification given in (5.18). Furthermore, it is in the one dimensional system

$$b_{i,i} = -b_{i,i-1} = b_i \quad \text{and} \quad b_{i+1,i+1} = -b_{i+1,i} = b_{i+1}.$$

Hence follows for $x_{i,i-1}$ and $x_{i,i+1}$

$$x_{i,i-1} = \frac{|b_{i,i-1}|}{b_{i,i} + |b_{i+1,i}|} = \frac{-b_i}{b_i + b_{i+1}}$$

$$\text{and} \quad x_{i,i+1} = \frac{b_{i+1} - b_i - b_{i+1}}{b_i + b_{i+1}} = \frac{b_i}{b_i + b_{i+1}}.$$

This is the structure we have given in (5.11) for the modification. □

In the case of the one dimensional convection the modfication fulfils additionally $x_{i,i-1} + x_{i,i+1} = 0$. For functions and vectors respectively this implies that the image of constant function is a constant function. For the modified coarse grid operator $A_{0,X}$ this implies that it is also a M-matrix. This we will discuss more detailed in chapter 9.

To conclude this section we will present two propositions which are similar to Proposition 5.1.6. The first one has the same result as the first result of Proposition 5.1.6 for a slightly more general situation. The second porposition gives a perfect result in a quite theoretical situation.

Proposition: 5.1.8. *Let A, X be matrices as defined in (5.17),(5.18). Let $\mathcal{N}_i^1, \mathcal{N}_j^1$ be two aggregated points with $b_{j,i} \neq 0$. Let $\mathcal{N}_k^1$ be a grid point with $b_{i,k} \neq 0$. If $b_{i,t} = b_{j,t} = 0$ for all $t \in M(k) \backslash \{i\}$ and we define $x_{i,k} = \frac{|b_{i,k}| - |b_{j,k}|}{b_{i,i} + |b_{j,i}|}$, then*

$$(A\,X)_{i,k} = (A\,X)_{j,k} = \frac{b_{j,i}|b_{i,k}| - b_{i,i}|b_{j,k}|}{b_{i,i} + |b_{j,i}|} \quad \textit{holds.}$$

proof. Similarly to the proof of Proposition 5.1.6, we obtain that the $k-$th column of X follows as

$$X_{.,k} = e_k^1 + \sum_{t=M(k)}^{n} x_{t,k} e_t^1 = e_k^1 + x_{i,k} e_i^1 + \sum_{t=M(k)\backslash\{i\}}^{n} x_{t,k} e_t^1.$$

Hence, based on the assumption $b_{i,t} = b_{j,t} = 0$ for all $t \in M(k)\backslash\{i\}$, we have

$$(A\,X)_{i,k} = A_{i,.}\,X_{.,k} = b_{i,i}x_{i,k} + b_{i,k}$$

$$(A\,X)_{j,k} = A_{j,.}\,X_{.,k} = b_{j,i}x_{i,k} + b_{j,k}.$$

Thus we have $(A\,X)_{i,k} = (A\,X)_{j,k}$ if we have

$$x_{i,k} = \frac{|b_{i,k}| - |b_{j,k}|}{b_{i,i} + |b_{j,i}|}.$$

And we obtain

$$(A\,X)_{j,k} = \frac{b_{j,i}|b_{i,k}| - b_{i,i}|b_{j,k}|}{b_{i,i} + |b_{j,i}|} = (A\,X)_{i,k}.$$

□

If we compare the assumptions of the Propositions 5.1.6 and 5.1.8 then we see that in Proposition 5.1.8 we drop the assumption of $b_{j,k} = 0$. If we consider the cases as illustrated in Figure 5.4 then we obtain that the situation as presented in (g) does not infringe the assumptions of Proposition 5.1.8. But we have to determine one more element of A for the modification. This implies a higher effort for the construction of P_X. The assertion of the two propositions is more or less the same.

Proposition: 5.1.9. *Let A, X be matrices as defined in (5.17),(5.18). Let $\mathcal{N}_i^1, \mathcal{N}_j^1$ be two aggregated points with $b_{j,i} \neq 0$. Assume that we have $x_{k,i} = x_{k,j} = 0$ for all $k \neq i, j$ with $b_{i,k} \neq 0$ or $b_{j,k} \neq 0$. If we define*

$$x_{i,k} = \begin{cases} 1 & for \quad k = i \\ \frac{b_{j,j} - b_{i,i} + b_{j,i}}{b_{i,i} + |b_{j,i}|} & for \quad k = j \\ \frac{|b_{i,k}| - |b_{j,k}|}{b_{i,i} + |b_{j,i}|} & for \quad k \neq i, j. \end{cases}$$

then we have

$$(A\,X)_{i,k} = (A\,X)_{j,k} \quad for \quad k \neq i, j$$

$$and \quad (A\,X)_{i,i} + (A\,X)_{i,j} = (A\,X)_{j,j} + (A\,X)_{j,i}.$$

proof. The second equality follows immediately from the second assertion of Proposition 5.1.6. The first equality follows like the first assertion of Propositon 5.1.6. The k-th column of X is

$$X_{.,k} = e_k^1 + \sum_{t=M(k)}^{n} x_{t,k}e_t^1 = e_k^1 + x_{i,k}e_i^1 + \sum_{t=M(k)\backslash\{i\}}^{n} x_{t,k}e_t^1.$$

Hence we have, based on the assumption $x_{k,i} = x_{k,j} = 0$ for $k \neq i, j$ with $b_{i,k} \neq 0$ or $b_{j,k} \neq 0$,

$$(A\,X)_{i,k} = A_{i,.}\,X_{.,k} = b_{i,i}x_{i,k} + b_{i,k}$$
$$(A\,X)_{j,k} = A_{j,.}\,X_{.,k} = b_{j,i}x_{i,k} + b_{j,k}.$$

Thus we have $(A\,X)_{i,k} = (A\,X)_{j,k}$ if we have

$$x_{i,k} = \frac{|b_{i,k}| - |b_{j,k}|}{b_{i,i} + |b_{j,i}|}.$$

□

As already mentioned, the assumptions of Proposition 5.1.9 are quite restrictive. They are for example fulfilled if only $\mathcal{N}_i^1, \mathcal{N}_j^1$ are aggregated and all other points are isolated points.

5.1.3 Modifications for a convection diffusion system

Finally we want to consider a convection diffusion system. To show the effect that occurs compared with the convection system it is sufficient to consider the small system given by four points. So we will do this first. Then we will show that the results of Proposition 5.1.6 hold in a weaker sense.

So we will start by a system given by four grid points as illustrated in Figure 5.5.

The stencil in $\mathcal{N}_i^1$ is given by

$$[-\varepsilon_{i-1} - b_i,\ b_i + \varepsilon_{i-1} + \varepsilon_i,\ -\varepsilon_i].$$

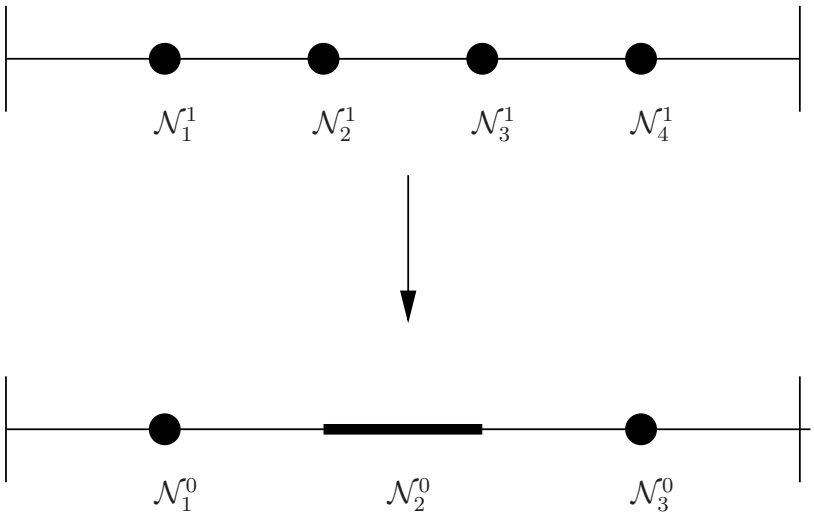


Figure 5.5: Coarsing of the four point system

So the stiffness matrix for the small system is given by

(5.21)

$$A = \begin{pmatrix} b_1+\varepsilon_0+\varepsilon_1 & -\varepsilon_1 & 0 & 0 \\ -b_2-\varepsilon_1 & b_2+\varepsilon_1+\varepsilon_2 & -\varepsilon_2 & 0 \\ 0 & -b_3-\varepsilon_2 & b_3+\varepsilon_2+\varepsilon_3 & -\varepsilon_3 \\ 0 & 0 & -b_4-\varepsilon_3 & b_4+\varepsilon_3+\varepsilon_4 \end{pmatrix}$$

$$= \begin{pmatrix} b_1 & 0 & 0 & 0 \\ -b_2 & b_2 & 0 & 0 \\ 0 & -b_3 & b_3 & 0 \\ 0 & 0 & -b_4 & b_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_0+\varepsilon_1 & -\varepsilon_1 & 0 & 0 \\ -\varepsilon_1 & \varepsilon_1+\varepsilon_2 & -\varepsilon_2 & 0 \\ 0 & -\varepsilon_2 & \varepsilon_2+\varepsilon_3 & -\varepsilon_3 \\ 0 & 0 & -\varepsilon_3 & \varepsilon_3+\varepsilon_4 \end{pmatrix} = B + E.$$

To have an idea of what happens in this system if we use no modification we will first consider the second and the third row of A. We remember that a basis of V_0 is given as

$$\{(1,0,0,0),(0,1,1,0),(0,0,0,1)\}.$$

Hence V_0 is invariant with respect to A if and only if

$$a_{2,1} = a_{3,1}, \quad a_{2,4} = a_{3,4} \quad \text{and} \quad a_{2,2} + a_{2,3} = a_{3,3} + a_{3,2}.$$

As shown in section 5.1.1 we have for an arbitrary $v_0 \in V_0$ the representation $v_0 = (f, u, u, g)^T$ with $f, u, g \in \mathbb{R}$ and therewith

$$\begin{aligned} A\,v_0 &= B\,v_0 + E\,v_0 \\ &= (b_1\,f,\, b_2(u-f),\, 0,\, b_4(u-g)) \\ &+ (\varepsilon_0 f + \varepsilon_1(f-u),\, \varepsilon_1(u-f),\, \varepsilon_2(u-g),\, \varepsilon_3(g-u) + \varepsilon_4 g). \end{aligned}$$

So we can split the problem of the invariance into two subproblems. The first one with respect to the matrix B and the other one with respect to E. So we will use the modification as figured out in the last section and section 5.1.1 respectively and show that the bias is only given by the the matrix E. The idea is that the main influence for the system is given by B as this represents the convection.

So we set

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{|a_{2,1}|}{a_{2,2}+|a_{3,2}|} & 1 & \frac{a_{3,3}-a_{2,2}+a_{3,2}}{a_{2,2}+|a_{3,2}|} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.22}$$

We therefore have the same structure of the modification. In particular this modification is easily given by the elements of A since we do not differ for the calculation of X between the symmetric and the antisymmetric part of the operator. Then we obtain

$$\begin{aligned} (A\,X)_{2,1} &= \frac{|a_{2,1}|a_{2,2}}{a_{2,2}+|a_{3,2}|} + a_{2,1} = \frac{|a_{3,2}|a_{2,1}}{a_{2,2}+|a_{3,2}|} \\ (A\,X)_{2,2} + (A\,X)_{2,3} &= a_{2,2} + a_{2,3} + \frac{a_{2,2}(a_{3,3}-a_{2,2}+a_{3,2})}{a_{2,2}+|a_{3,2}|} = \frac{a_{2,2}a_{3,3}}{a_{2,2}+|a_{3,2}|} + a_{2,3} \\ (A\,X)_{2,4} &= 0 \\ \text{and} \quad (A\,X)_{3,1} &= \frac{a_{3,2}|a_{2,1}|}{a_{2,2}+|a_{3,2}|} \\ (A\,X)_{3,2} + (A\,X)_{3,3} &= a_{3,2} + a_{3,3} + \frac{a_{3,2}(a_{3,3}-a_{2,2}+a_{3,2})}{a_{2,2}+|a_{3,2}|} = \frac{a_{3,3}a_{2,2}}{a_{2,2}+|a_{3,2}|} \\ (A\,X)_{3,4} &= a_{3,4}. \end{aligned}$$

With these calculations we can summarize the results we get for this small system as follows:

Lemma: 5.1.10. *Let A, X be matrices as given in (5.21), (5.22) then*

1. $(A\,X)_{2,1} = (A\,X)_{3,1}$ *holds.*
2. $(A\,X\,v_0)(2) - (A\,X\,v_0)(3) = -\varepsilon_2 v_0(3) + \varepsilon_3 v_0(4)$. *In particular this difference is independent of B for all $v_0 \in V_0$.*
3. *follows if $\varepsilon_i = \varepsilon$ for all $i = 0, \ldots, 4$ then we have $A\,X\,v_0 \in V_0$ for all constant vectors v_0.*

proof. 1. The first proposition follows immediately form the calculation above the lemma.

2. For $v_0 \in V_0$ with $v_0 = (f, u, u, g)^T$ it follows

$$\begin{aligned}(A\,X\,v_0)(2) - (A\,X\,v_0)(3)&\\ &= [(A\,X)_{2,1} - (A\,X)_{3,1}]f + [(A\,X)_{2,4} - (A\,X)_{3,4}]g\\ &\quad + [(A\,X)_{2,2} + (A\,X)_{2,3} - (A\,X)_{3,2} - (A\,X)_{3,3}]u\\ &= a_{2,3}u - a_{3,4}g = -\varepsilon_2 u + \varepsilon_3 g.\end{aligned}$$

The last equation follows thereby again from the calculations above the lemma.

3. If we have $\varepsilon_i = \varepsilon$ for $i = 0, \ldots, 4$ and it is v_0 constant then it follows from the calculation done for the second proposition

$$(A\,X\,v_0)(2) - (A\,X\,v_0)(3) = -\varepsilon_2 u + \varepsilon_3 g = \varepsilon(g - u) = 0.$$

□

From the calculation above and the results of the Lemma 5.1.10 respectively it is obvious that we can not transform the result as easily to a more general result as done for the convection. We will see that if we try to do this we always get a dependency on the elements of E. Let $A = B + E \in \mathbb{R}^{n\times n}$ be matrices which fulfil

$$B = (b_{i,j}), \quad \text{with} \quad b_{i,i} > 0, \quad b_{i,j} \leq 0, \quad \text{for} \quad i \neq j \tag{5.23}$$

$$b_{i,i} \geq \sum_{j\neq i} |b_{i,j}| \quad \text{and} \quad b_{i,j} \neq 0 \quad \Rightarrow \quad b_{j,i} = 0$$

$$\text{and} \quad E = (\varepsilon_{i,j}), \quad \text{with} \quad \varepsilon_{i,i} > 0, \quad \varepsilon_{i,j} \leq 0, \quad \text{for} \quad i \neq j \tag{5.24}$$

$$\varepsilon_{i,i} \geq \sum_{j\neq i} |\varepsilon_{i,j}| \quad \text{and} \quad \varepsilon_{i,j} = \varepsilon_{j,i}.$$

To define the modification X we remember the definition of the set $M_0(i)$ and define the set $M_1(i)$ as

$$M_1(i) := \big\{t \in \{1,\dots,n\}\backslash\{i\} \,:\, |a_{i,t}| > |a_{t,i}| \neq 0.\big\}.$$

Based on the interpretation as a convection system, $M_1(i)$ is the set of the indices of the grid points $\mathcal{N}_t^1$ that have an influence by the convection on $\mathcal{N}_i^1$.

Then we define X as done in (5.18) as

(5.25)
$$X_{i,\cdot} = \begin{cases} (e_i^1)^T & \text{if } \mathcal{N}_i^1 \text{ is isolated or } \mathcal{N}_i^1 \text{ is aggregated} \\ & \text{with } \mathcal{N}_j^1 \text{ and it is } |a_{i,j}| > |a_{j,i}|. \\ (e_i^1)^T + x_{i,j}(e_j^1)^T + x_{i,k}(e_k^1)^T & \text{if } \mathcal{N}_i^1, \mathcal{N}_j^1 \text{ are aggregated, it is } |a_{i,j}| < |a_{j,i}| \\ & \text{and it is } k \in M_1(i). \end{cases}$$

(Based on the modification it is implicit that we only aggregate $\mathcal{N}_i^1, \mathcal{N}_j^1$ if it is $a_{i,j} \neq a_{j,i}$.) As in the definition (5.18), we choose one of the indices if it is $|M_1(i)| > 1$. And again it is a feasible heuristic to choose the index k with $|b_{i,k}| \geq |b_{i,s}|$ for all $s \in M_1(i)$. But if it is $M_1(i) = \emptyset$, it can be also useful to choose an index k with $a_{i,k} = a_{k,i}$.

Then we get a result that can be seen as a generalization of Proposition 5.1.6.

Proposition: 5.1.11. *Let $A = B + E$ and X be matrices as defined in (5.23),(5.24) and (5.25). Let $\mathcal{N}_i^1, \mathcal{N}_j^1$ be two aggregated points with $|a_{j,i}| > |a_{i,j}|$. Let further $\mathcal{N}_k^1$ be a grid point with $b_{i,k} \neq 0$.*

1. *If it is $b_{j,k} = 0$ and it is $b_{i,t} = b_{j,t} = 0$ for all $t \in M(k)\backslash\{i\}$ and we define $x_{i,k} = \frac{|a_{i,k}|}{a_{i,i}+|a_{j,i}|}$ then*

$$(A\,X)_{i,k} - \sum_{t \in M(k)\backslash\{i\}} \varepsilon_{i,t} x_{t,k} = (A\,X)_{j,k} - \sum_{t \in M(k)} \varepsilon_{j,t} x_{t,k} = \frac{a_{i,k}|a_{j,i}|}{a_{i,i} + |a_{j,i}|}$$

   holds.

2. *If it is $b_{i,t} = b_{j,t} = 0$ for all t with $t \in M(i) \cup M(j)$ and we define $x_{i,j} = \frac{a_{j,j} - a_{i,i} + a_{j,i}}{a_{i,i}+|a_{j,i}|}$*

then

$$(A\,X)_{i,i} + (A\,X)_{i,j} - \varepsilon_{i,j} - \sum_{t\in M(i)} \varepsilon_{i,t} x_{t,i} - \sum_{t\in M(j)} \varepsilon_{i,t} x_{t,j}$$

$$= (A\,X)_{j,j} + (A\,X)_{j,i} - \sum_{t\in M(i)} \varepsilon_{j,t} x_{t,i} - \sum_{t\in M(j)} \varepsilon_{j,t} x_{t,j} = \frac{a_{i,i} a_{j,j}}{a_{i,i} + |a_{j,i}|}$$

holds.

proof. Based on the assumption of $|a_{j,i}| > |a_{i,j}|$ and the definition of X it follows that we have $X_{j,.} = (e_j^1)^T$ and $X_{i,.} = (e_i^1)^T + x_{i,k}(e_k^1)^T + x_{i,j}(e_j^1)^T$.

1. We obtain that the $k-$th column of X is given as

$$X_{.,k} = e_k^1 + \sum_{t=M(k)}^{n} x_{t,k} e_t^1 = e_k^1 + x_{i,k} e_i^1 + \sum_{t=M(k)\setminus\{i\}}^{n} x_{t,k} e_t^1$$

Hence we have based on the assumption $b_{i,t} = b_{j,t} = 0$ for all $t \in M(k)\setminus\{i\}$

$$(A\,X)_{i,k} = A_{i,.}\,X_{.,k} = a_{i,i} x_{i,k} + a_{i,k} + \sum_{t=M(k)\setminus\{i\}}^{n} x_{t,k} \varepsilon_{i,t}$$

$$(A\,X)_{j,k} = A_{j,.}\,X_{.,k} = a_{j,i} x_{i,k} + \sum_{t=M(k)\setminus\{i\}}^{n} x_{t,k} \varepsilon_{j,t}.$$

The second equation results thereby from the condition $b_{j,k} = 0$. So we obtain for $x_{i,k} = \frac{|a_{i,k}|}{a_{i,i}+|a_{j,i}|}$ the following

$$(A\,X)_{i,k} = \frac{a_{i,k}|a_{j,i}|}{a_{i,i} + |a_{j,i}|} + \sum_{t=M(k)\setminus\{i\}}^{n} x_{t,k} \varepsilon_{i,t}$$

$$\text{and} \quad (A\,X)_{j,k} = \frac{a_{i,k}|a_{j,i}|}{a_{i,i} + |a_{j,i}|} + \sum_{t=M(k)\setminus\{i\}}^{n} x_{t,k} \varepsilon_{j,t}$$

2. Based on the assumption that $\mathcal{N}_i^1, \mathcal{N}_j^1$ are aggregated with $b_{j,i} \neq 0$ it follows for the columns i, j of X

$$X_{.,i} = e_i^1 + \sum_{t\in M(i)} x_{t,i} e_t^1$$

$$\text{and} \quad X_{.,j} = e_j^1 + x_{i,j} e_i^1 + \sum_{t\in M(j)} x_{t,j} e_t^1.$$

As it is $b_{i,t} = b_{j,t} = 0$ for all t with $t \in M(i) \cup M(j)$ this implies

$$\begin{aligned}
(A\,X)_{i,i} + (A\,X)_{i,j} &= A_{i,.}X_{.,i} + A_{i,.}X_{.,j} \\
&= a_{i,i} + a_{i,i}x_{i,j} + a_{i,j} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{i,t}x_{t,i} + \sum_{t \in M(j)} \varepsilon_{i,t}x_{t,j} \\
&= (1 + x_{i,j})a_{i,i} + a_{i,j} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{i,t}x_{t,i} + \sum_{t \in M(j)} \varepsilon_{i,t}x_{t,j} \\
(A\,X)_{j,i} + (A\,X)_{j,j} &= A_{j,.}X_{.,i} + A_{j,.}X_{.,j} \\
&= a_{j,i} + x_{i,j}a_{j,i} + a_{j,j} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{j,t}x_{t,i} + \sum_{t \in M(j)} \varepsilon_{j,t}x_{t,j} \\
&= (1 + x_{i,j})a_{j,i} + a_{j,j} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{j,t}x_{t,i} + \sum_{t \in M(j)} \varepsilon_{j,t}x_{t,j}.
\end{aligned}$$

If we have $x_{i,j} = \frac{a_{j,j} - a_{i,i} + a_{j,i}}{a_{i,i} + |a_{j,i}|}$ then we obtain

$$\begin{aligned}
&(A\,X)_{i,i} + (A\,X)_{i,j} \\
&\quad= (1 + x_{i,j})a_{i,i} + a_{i,j} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{i,t}x_{t,i} + \sum_{t \in M(j)} \varepsilon_{i,t}x_{t,j} \\
&\quad= \frac{a_{i,i}\big(a_{i,i} + |a_{j,i}| + a_{j,j} - a_{i,i} + a_{j,i}\big)}{a_{i,i} + |a_{j,i}|} + a_{i,j} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{i,t}x_{t,i} + \sum_{t \in M(j)} \varepsilon_{i,t}x_{t,j} \\
&\quad= \frac{a_{i,i}a_{j,j}}{a_{i,i} + |a_{j,i}|} + a_{i,j} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{i,t}x_{t,i} + \sum_{t \in M(j)} \varepsilon_{i,t}x_{t,j}
\end{aligned}$$

and

$$
\begin{aligned}
&(A\,X)_{j,i} + (A\,X)_{j,j} \\
&\quad = (1 + x_{i,j})a_{j,i} + a_{j,j} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{j,t} x_{t,i} + \sum_{t \in M(j)} \varepsilon_{j,t} x_{t,j} \\
&\quad = \frac{a_{j,i}\big(a_{j,j} - a_{i,i} + a_{j,i} + a_{i,i} + |a_{j,i}|\big)}{a_{i,i} + |a_{j,i}|} + \frac{a_{j,j}a_{i,i} + a_{j,j}|a_{j,i}|}{a_{i,i} + |a_{j,i}|} \\
&\qquad + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{j,t} x_{t,i} + \sum_{t \in M(j)} \varepsilon_{j,t} x_{t,j} \\
&\quad = \frac{a_{j,i}a_{j,j}}{a_{i,i} + |a_{j,i}|} + \frac{a_{j,j}a_{i,i} + a_{j,j}|a_{j,i}|}{a_{i,i} + |a_{j,i}|} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{j,t} x_{t,i} + \sum_{t \in M(j)} \varepsilon_{j,t} x_{t,j} \\
&\quad = \frac{a_{j,j}a_{i,i}}{a_{i,i} + |a_{j,i}|} + \sum_{t \in M(i)\setminus\{k\}} \varepsilon_{j,t} x_{t,i} + \sum_{t \in M(j)} \varepsilon_{j,t} x_{t,j}.
\end{aligned}
$$

This proves the proposition.

□

The main aspect of the Proposition 5.1.11 is that if we have a convection diffusion system and we do the same modification as for a convection system then we get the same result with a bias that only depends on the elements of E and X. Based on the given structure the idea is that the elements of B are much bigger than the elements of E. If we assume that the elements of B have the size b and the elements of E the size ε, then based on the definition of X the elements $x_{i,j}$ which are used for the modification also have the size $\frac{b}{b} = 1$. Therefore the bias is given by the size of ε.

5.2 Modifications for the symmetric model problem (one sided)

Now we will consider modifications for the symmetric problem as introduced in section 2.2. The continious problem is given by the equation

$$
\begin{aligned}
-div(\alpha(x)\, grad\, u(x)) &= f(x), \quad \forall x \in \Omega \\
u(x) &= g(x), \quad \forall x \in \partial\Omega.
\end{aligned}
$$

with a symmetric $\alpha(x) \in \mathbb{R}^{2\times 2}$. So we obtain the stencils as

$$\begin{pmatrix} -\delta_{nw} & -\varepsilon_n & -\delta_{ne} \\ -\varepsilon_w & m & -\varepsilon_e \\ -\delta_{sw} & -\varepsilon_s & -\delta_{se} \end{pmatrix}$$

$$\text{with} \quad m = \varepsilon_n + \varepsilon_e + \varepsilon_s + \varepsilon_w + \delta_{ne} + \delta_{se} + \delta_{sw} + \delta_{nw}$$

$$\text{and} \quad \varepsilon_i > 0, \quad \text{for} \quad i = n, e, s, w$$

$$\delta_i \geq 0, \quad \text{for} \quad i = ne, se, sw, nw.$$

As for the convection we will start from a one dimensional system. That means $\Omega \subset \mathbb{R}$ and $\alpha(x) \in \mathbb{R}_+$. The stencil follows in $\mathcal{N}_i^1$ as

$$[-a_{i-1}, a_{i-1} + a_i, -a_i], \quad \text{with} \quad a_{i-1}, a_i > 0.$$

So again we will first consider the small system of four grid points that is given by $\mathcal{N}_1^1, \ldots, \mathcal{N}_4^1$. Then we assume that we aggregate the grid points $\mathcal{N}_2^1, \mathcal{N}_3^1$ to the new point $\mathcal{N}_2^0$. We obtain that the restriction R and the prolongation P are

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad R = P^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5.26}$$

The links and the system are illustrated in Figure 5.6.

And therewith A, A_0 follow as

$$A = \begin{pmatrix} \varepsilon + \varepsilon_0 & -\varepsilon & 0 & 0 \\ -\varepsilon & a + \varepsilon & -a & 0 \\ 0 & -a & a + \delta & -\delta \\ 0 & 0 & -\delta & \delta + \delta_0 \end{pmatrix} \quad \text{and} \quad A_0 = \begin{pmatrix} \varepsilon + \varepsilon_0 & -\varepsilon & 0 \\ -\varepsilon & \varepsilon + \delta & -\delta \\ 0 & -\delta & \delta + \delta_0 \end{pmatrix}. \tag{5.27}$$

A short discussion for an estimation of γ_{DT} of the unmodified method will be presented in the section 8.1. With regard to this we will only consider modifications and the modified systems, respectively.

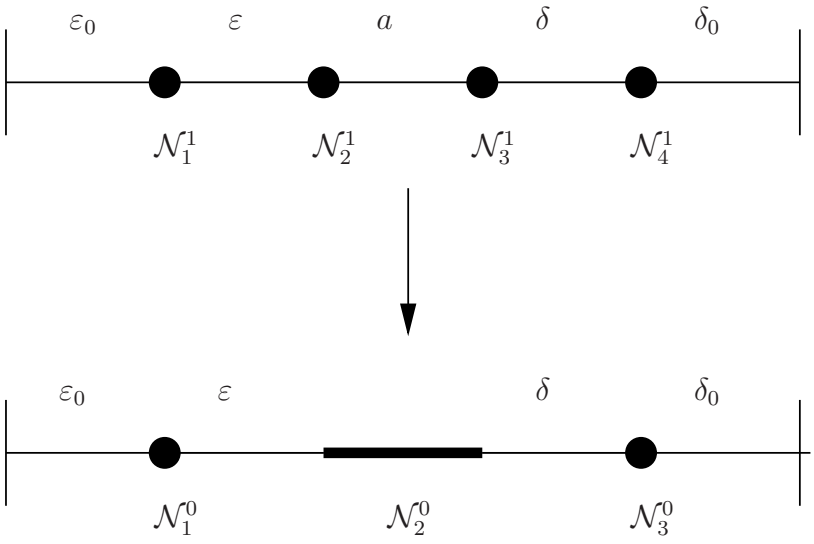


Figure 5.6: Coarsing of the symmetric four point system

5.2.1 An exact modification

First we will construct for the given small system of A, P, R a modification X that holds

$$A\,X\,v_0 \in V_0 \quad \text{for all} \quad v_0 \in V_0. \tag{5.28}$$

As for the unsymmetric example of the small system shown in (5.6) based on the basis $\{e_1^1, e_2^1 + e_3^1, e_4^1\}$ of V_0 the condition (5.28) is equivalent to

$$(A\,X)_{2,1} = (A\,X)_{3,1}, \quad (A\,X)_{2,4} = (A\,X)_{3,4} \tag{5.29}$$

$$\text{and} \quad (A\,X)_{2,2} + (A\,X)_{2,3} = (A\,X)_{3,2} + (A\,X)_{3,3}. \tag{5.30}$$

Again to do as few modifications as possible we set the first and the fourth row of X as follows

$$X_{1,\cdot} = (e_1^1)^T \quad \text{and} \quad X_{4,\cdot} = (e_4^1)^T.$$

Next we will consider the three conditions given in (5.29), (5.30) seperated. For

$(A\,X)_{2,1} = (A\,X)_{3,1}$ we obtain

$$\begin{aligned} & (A\,X)_{2,1} = (A\,X)_{3,1} \\ \Leftrightarrow \quad & -ax_{3,1} + (a+\varepsilon)x_{2,1} - \varepsilon = (a+\delta)x_{3,1} - ax_{2,1} \\ \Leftrightarrow \quad & x_{2,1} = \frac{2a+\delta}{2a+\varepsilon}x_{3,1} + \frac{\varepsilon}{2a+\varepsilon}. \end{aligned}$$

Again as explained above we set $x_{3,1} = 0$ and the equation above implies $x_{2,1} = \frac{\varepsilon}{2a+\varepsilon}$. For the equation $(A\,X)_{2,4} = (A\,X)_{3,4}$ we obtain based on the same arguments

$$\begin{aligned} & (A\,X)_{2,4} = (A\,X)_{3,4} \\ \Leftrightarrow \quad & x_{3,4} = \frac{2a+\varepsilon}{2a+\delta}x_{2,4} + \frac{\delta}{2a+\delta}. \end{aligned}$$

We set again $x_{2,4} = 0$ and we obtain $x_{3,4} = \frac{\delta}{2a+\delta}$. So far we should remark that the values for the modification are easy to calculate based on the elements of the matrix A. It is

$$x_{2,1} = \frac{a_{2,2} + a_{2,3}}{a_{2,2} - a_{2,3}} \quad \text{and} \quad x_{3,4} = \frac{a_{3,3} + a_{3,2}}{a_{3,3} - a_{3,2}}.$$

So we consider the condition (5.30). It follows

$$\begin{aligned} & (A\,X)_{2,2} + (A\,X)_{2,3} = (A\,X)_{3,2} + (A\,X)_{3,3} \\ \Leftrightarrow \quad & (2a+\delta)(x_{3,2} + x_{3,3}) = (2a+\varepsilon)(x_{2,2} + x_{2,3}). \end{aligned}$$

This is fulfilled if we set

$$x_{2,3} = 0 = x_{3,2}, \quad x_{2,2} = \frac{1}{2a+\varepsilon} \quad \text{and} \quad x_{3,3} = \frac{1}{2a+\delta}.$$

Altogether this gives the modification matrix X in the form

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{\varepsilon}{2a+\varepsilon} & \frac{1}{2a+\varepsilon} & 0 & 0 \\ 0 & 0 & \frac{1}{2a+\delta} & \frac{\delta}{2a+\delta} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.31}$$

and we can summarize the result for this system and the so defined modification as follows:

Proposition: 5.2.1. *Let A, P, X be as defined in (5.27), (5.26) and (5.31). Then V_0 is invariant with respect to $A\,X$.*

proof. For the proof see the calculation above in this section. □

So far it seems that we could modify the symmetric problem as well as the problem given by the one dimensional convection. In the next two sections we will first consider a more general one dimensional problem that is not possible to modify that way. Then we will consider a special case that can be solved in higher dimensions, too. But because of the structure of the coarser operators we will see that this is more or less a theoretical result.

5.2.2 Problems for exact modifications

The problem of the modification is obvious if we consider a system of the same structure that belongs to six grid points $\mathcal{N}_1^1, \ldots, \mathcal{N}_6^1$ and we assume that the points $\mathcal{N}_2^1, \mathcal{N}_3^1$ and $\mathcal{N}_4^1, \mathcal{N}_5^1$ are aggregated to $\mathcal{N}_2^0$ and $\mathcal{N}_3^0$, respectively. This is illustrated in Figure 5.7.

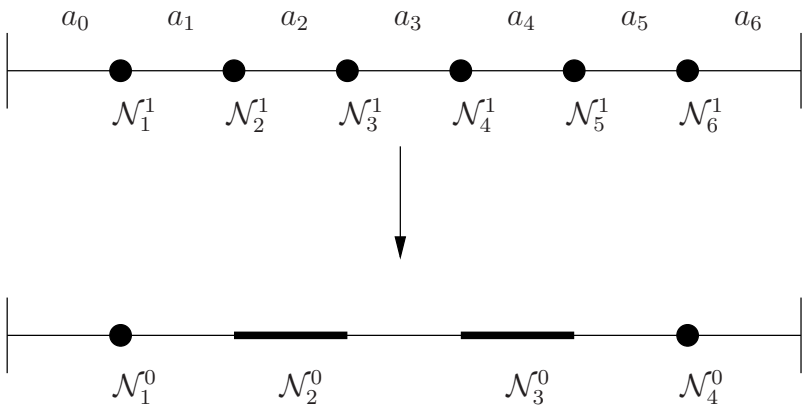


Figure 5.7: Coarsing of the symmetric six point system

We have the matrices as follows:

(5.32)
$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} a_0 + a_1 & -a_1 & 0 & 0 & 0 & 0 \\ -a_1 & a_1 + a_2 & -a_2 & 0 & 0 & 0 \\ 0 & -a_2 & a_2 + a_3 & -a_3 & 0 & 0 \\ 0 & 0 & -a_3 & a_3 + a_4 & -a_4 & 0 \\ 0 & 0 & 0 & -a_4 & a_4 + a_5 & -a_5 \\ 0 & 0 & 0 & 0 & -a_5 & a_5 + a_6 \end{pmatrix}$$

Then we generalize the modification as given in (5.31) and define $X \in \mathbb{R}^{6\times 6}$ as follows

(5.33)
$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{a_1}{2a_2+a_1} & \frac{1}{2a_2+a_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2a_2+a_3} & \frac{a_3}{2a_2+a_3} & 0 & 0 \\ 0 & 0 & \frac{a_3}{2a_4+a_3} & \frac{1}{2a_4+a_3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2a_4+a_5} & \frac{a_5}{2a_4+a_5} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

A basis of V_0 is in this case given as follows

$$\big\{(1,0,0,0,0,0),(0,1,1,0,0,0),(0,0,0,1,1,0),(0,0,0,0,0,1)\big\}.$$

Hence to keep that result that V_0 is invariant with respect to $A\,X$ it must hold

$$A\,X\,v_0 \in V_0$$

Hence it is necessary that the second and the third row of $A\,X$ fulfil the following equations:

$$(A\,X)_{2,1} = (A\,X)_{3,1}, \quad (A\,X)_{2,6} = (A\,X)_{3,6}$$

$$(A\,X)_{2,2} + (A\,X)_{2,3} = (A\,X)_{3,2} + (A\,X)_{3,3}$$

$$(A\,X)_{2,4} + (A\,X)_{2,5} = (A\,X)_{3,4} + (A\,X)_{3,5}.$$

If we calculate these two rows we obtain

$$(A\,X)_{2,\cdot} = \left(\frac{-a_1 a_2}{a_1 + 2a_2}, \frac{a_1 + a_2}{a_1 + 2a_2}, \frac{-a_2}{a_3 + 2a_2}, \frac{-a_2 a_3}{2a_2 + a_3}, 0, 0 \right)$$

$$(A\,X)_{3,\cdot} = \left(\frac{-a_1 a_2}{a_1 + 2a_2}, \frac{-a_2}{a_1 + 2a_2}, \frac{a_2 + a_3}{a_3 + 2a_2} - \frac{a_3^2}{a_3 + 2a_4}, \frac{(a_2 + a_3)a_3}{2a_2 + a_3} - \frac{a_3}{a_3 + 2a_4}, 0, 0 \right).$$

So it is obvious that the meaningful result of the grid given by four points does not hold in this situation. A closer look shows quite simply that there is no local estimation in this case.

In the next section we will show that this problem results from the situation in which there are neighbours of aggregated points that are not isolated points.

5.2.3 A solvable situation in arbitrary dimensions

Now we will show that we can generalize the modification to a quite general situation having only the restriction that the neigbours of aggregated points are isolated points. This assumption we will also consider in chapter 8. Based on this strict assumption we obtain a meaningful result. This gives us the motivation for more general systems.

So the situation should be given as follows: Let $A \in \mathbb{R}^{n \times n}$ be a s.p.d. matrix that fulfils

$$\begin{aligned} & a_{i,i} > 0, \quad \forall i = 1, \ldots, n \\ & a_{i,j} \leq 0, \quad \forall i, j = 1, \ldots, n, \quad i \neq j \\ & a_{i,i} \geq \sum_{j=1,\, i \neq j}^{n} |a_{i,j}|. \end{aligned} \tag{5.34}$$

Let $\mathcal{N}_i^1, \mathcal{N}_j^1$ be two points that will be aggregated to $\mathcal{N}_t^0$ for an $t \in \{1, \ldots, n_0\}$ and all points $\mathcal{N}_k^1,\ k \neq i, j$ with $a_{i,k} \neq 0$ or $a_{j,k} \neq 0$ are isolated points. This is illustrated in Figure 5.8 at page 176 .Then we define the modification $X \in \mathbb{R}^{n \times n}$ by its rows $X_{i,\cdot}$ with

$$X_{i,\cdot} = (e_i^1)^T \quad \text{if } \mathcal{N}_i^1 \text{ is an isolated point.} \tag{5.35}$$

$$X_{i,\cdot} = \frac{1}{a_{i,i} + |a_{i,j}|} \left((e_i^1)^T + \sum_{k=1,\, k \neq i,j}^{n} |a_{i,k}| (e_k^1)^T \right) \quad \text{if } \mathcal{N}_i^1, \mathcal{N}_j^1 \text{ are aggregated to } \mathcal{N}_t^0 \text{ for an } t \in \{1, \ldots, n_0\}.$$

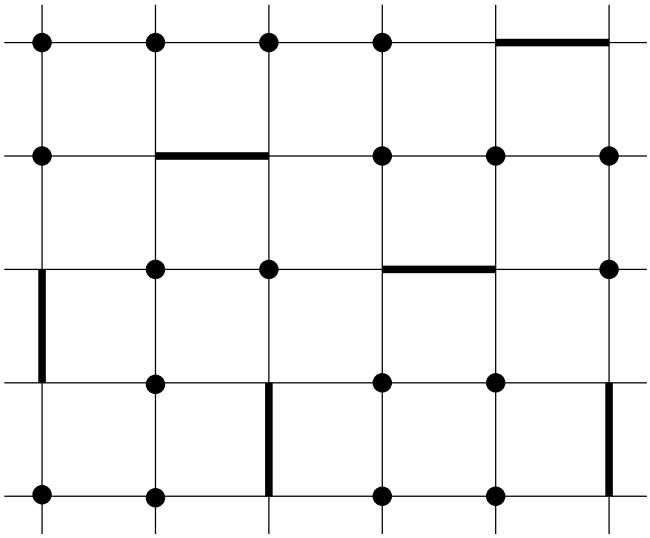


Figure 5.8: Coarsing of a symmetric system for an exact modification

Proposition: 5.2.2. *Assume the situation as given above in this section. Then for two aggregated points $\mathcal{N}_1^1, \mathcal{N}_2^1$ it holds*

$$(A\,X)_{1,1} = \frac{a_{1,1}}{a_{1,1} + |a_{1,2}|}, \quad (A\,X)_{1,2} = -\frac{|a_{1,2}|}{a_{2,2} + |a_{1,2}|},$$

$$(A\,X)_{1,k} = -\frac{|a_{1,2}\,a_{1,k}|}{a_{1,1} + |a_{1,2}|} - \frac{|a_{1,2}\,a_{2,k}|}{a_{2,2} + |a_{1,2}|} \quad \textit{for} \quad k \neq 1, 2.$$

$$(A\,X)_{2,2} = \frac{a_{2,2}}{a_{2,2} + |a_{2,1}|}, \quad (A\,X)_{2,1} = -\frac{|a_{2,1}|}{a_{1,1} + |a_{2,1}|}$$

$$(A\,X)_{2,k} = -\frac{|a_{2,1}\,a_{2,k}|}{a_{2,2} + |a_{2,1}|} - \frac{|a_{2,1}\,a_{1,k}|}{a_{1,1} + |a_{2,1}|} \quad \textit{for} \quad k \neq 1, 2.$$

proof. Based on the symmetry of the proposition it is sufficient to prove the propositions for the entries in the first row of $A\,X$. We define the set

$$M := \{t \in \{1, \ldots, n\} \,:\, X_{t,.} \neq e_t^1\}.$$

Based on the definition of X it is $t \in M$ if and only if $\mathcal{N}_t^1$ is not an isolated point. Based on the assumptions we have

$$t \in M \backslash \{1, 2\} \quad \Rightarrow \quad a_{1,t} = a_{2,t} = 0.$$

We start by the proposition for $(A\,X)_{1,1}$. Based on the definition above we obtain

$$X_{.,1} = \frac{1}{a_{1,1} + |a_{1,2}|} e_1^1 + \sum_{t \in M \setminus \{1,2\}} x_{t,1} e_t^1$$

$$\Rightarrow \quad (A\,X)_{1,1} = A_{1,.} \cdot X_{.,1} = a_{1,1} \frac{1}{a_{1,1} + |a_{1,2}|}.$$

Based on the same argument we obtain

$$X_{.,2} = \frac{1}{a_{2,2} + |a_{2,1}|} e_2^1 + \sum_{t \in M \setminus \{1,2\}} x_{t,2} e_t^1$$

$$\Rightarrow \quad (A\,X)_{1,2} = A_{1,.} \cdot X_{.,2} = -\frac{|a_{1,2}|}{a_{2,2} + |a_{2,1}|}.$$

At last we will consider an arbitrary $k \neq 1, 2$. Then we will distinguish two situations. First we will assume that $\mathcal{N}_k^1$ is a isolated point. Then it is

$$X_{.,k} = e_k^1 + \sum_{t \in M} x_{t,k} e_t^1 = e_k^1 + e_1^1 x_{t,1} + e_2^1 x_{t,2} + \sum_{t \in M \setminus \{1,2\}} x_{t,k} e_t^1.$$

As we have $a_{1,t} = a_{2,t} = 0$ for $t \in M \setminus \{1, 2\}$ we obtain

$$(A\,X)_{1,k} = A_{1,.} \cdot X_{.,k} = A_{1,.} \cdot \left(e_k^1 + e_1^1 x_{1,k} + e_2^1 x_{2,k} + \sum_{t \in M \setminus \{1,2\}} x_{t,k} e_t^1 \right)$$

$$= a_{1,k} + x_{1,k} a_{1,1} + x_{2,k} a_{1,2}$$

$$= a_{1,k} + \frac{|a_{1,k}|}{a_{1,1} + |a_{1,2}|} a_{1,1} + \frac{|a_{2,k}|}{a_{2,2} + |a_{2,1}|} a_{1,2}$$

$$= \frac{a_{1,k}|a_{1,2}|}{a_{1,1} + |a_{1,2}|} + \frac{|a_{2,k}|a_{1,2}}{a_{2,2} + |a_{2,1}|} = -\frac{|a_{1,k}a_{1,2}|}{a_{1,1} + |a_{1,2}|} - \frac{|a_{2,k}a_{1,2}|}{a_{2,2} + |a_{2,1}|}.$$

Secondly we will assume that $\mathcal{N}_k^1$ is aggregated with $\mathcal{N}_l^1$ to $\mathcal{N}_s^0$. As this implies $a_{1,k} = a_{2,k} = 0$ we obtain from the definition of X that the k-th column is

$$X_{.,k} = \sum_{t=1}^{n} x_{t,k} e_t^1 = \sum_{t=3}^{n} x_{t,k} e_t^1.$$

This implies

$$(A\,X)_{1,k} = A_{1,.} \cdot X_{.,k} = 0 = -\frac{|a_{1,k}a_{1,2}|}{a_{1,1} + |a_{1,2}|} - \frac{|a_{2,k}a_{1,2}|}{a_{2,2} + |a_{2,1}|}.$$

This proves the assertion. □

Theorem: 5.2.3. *Assume the situation as given above in this section. Then it follows that V_0 is invariant with respect to $A\,X$.*

proof. From the results of proposition 5.2.2 we obtain

$$\begin{aligned}
&(A\,X)_{1,1} + (A\,X)_{1,2} = (A\,X)_{2,2} + (A\,X)_{2,1} \\
\Leftrightarrow \quad &\frac{a_{1,1}}{a_{1,1}+|a_{1,2}|} - \frac{|a_{1,2}|}{a_{2,2}+|a_{1,2}|} = \frac{a_{2,2}}{a_{2,2}+|a_{2,1}|} - \frac{|a_{2,1}|}{a_{1,1}+|a_{1,2}|} \\
\Leftrightarrow \quad &\frac{a_{1,1}\,a_{2,2} + a_{1,1}|a_{1,2}| - (a_{1,1}|a_{1,2}| + a_{1,2}^2)}{(a_{1,1}+|a_{1,2}|)(a_{1,2}+|a_{1,2}|)} = \frac{a_{1,1}\,a_{2,2} + a_{2,2}|a_{1,2}| - (a_{2,2}|a_{2,1}| + a_{2,1}^2)}{(a_{2,2}+|a_{2,1}|)(a_{1,1}+|a_{1,2}|)}.
\end{aligned}$$

This equation holds since it is based on the symmetry of A we have $a_{1,2} = a_{2,1}$. For an arbitrary $k \neq 1, 2$ we obtain

$$\begin{aligned}
&(A\,X)_{1,k} = (A\,X)_{2,k} \\
&-\frac{|a_{1,2}\,a_{1,k}|}{a_{1,1}+|a_{1,2}|} - \frac{|a_{1,2}\,a_{2,k}|}{a_{2,2}+|a_{1,2}|} = -\frac{|a_{2,1}\,a_{2,k}|}{a_{2,2}+|a_{2,1}|} - \frac{|a_{2,1}\,a_{1,k}|}{a_{1,1}+|a_{2,1}|}.
\end{aligned}$$

Again this holds considering the characteristic $a_{1,2} = a_{2,1}$. Therewith we have $(A\,X\,v_0)(1) = (A\,X\,v_0)(2)$ for all $v_0 \in V_0$. □

We therefore obtain a result that seems to be the same as for the one dimensional convection. The difference in the assumptions is given as follows. For the convection we have only considered a one dimensional system but for the symmetric problem we have the condition that the neighbours of aggregated points are isolated points. The problem related to this condition is that the dimension of the matrix hardly changes if the coarsing holds this condition. This can be seen in Figure 5.8 at page 176. Additionally, in a numerical algorithm it takes a huge effort to control that the assumptions are fulfilled in each step.

5.2.4 Modification by the inverse of blocks

Again, as for the convection system it can be an idea to modify the system with the inverse of small blocks. Therefore we consider again the system given by four grid points as described at the beginning of this section. We remember that the matrix A

is given as

$$A = \begin{pmatrix} \varepsilon+\varepsilon_0 & -\varepsilon & 0 & 0 \\ -\varepsilon & a+\varepsilon & -a & 0 \\ 0 & -a & a+\delta & -\delta \\ 0 & 0 & -\delta & \delta+\delta_0 \end{pmatrix} \tag{5.36}$$

As the second and the third row and column are aggregated we define the blocks

$$A_1 = (\varepsilon+\varepsilon_0), \quad A_2 = \begin{pmatrix} a+\varepsilon & -a \\ -a & a+\delta \end{pmatrix} \quad \text{and} \quad A_3 = (\delta+\delta_0).$$

From these definitions we obtain

$$A_1^{-1} = (\varepsilon+\varepsilon_0)^{-1}, \quad A_2^{-1} = \frac{1}{(a+\varepsilon)(a+\delta)-a^2} \begin{pmatrix} a+\delta & a \\ a & a+\varepsilon \end{pmatrix} \quad \text{and} \quad A_3^{-1} = (\delta+\delta_0)^{-1}.$$

and with these blocks we define the modification X by

$$X = \begin{pmatrix} A_1^{-1} & & \\ & A_2^{-1} & \\ & & A_3^{-1} \end{pmatrix}. \tag{5.37}$$

We obtain

$$A\,X = \begin{pmatrix} 1 & -\frac{\varepsilon(a+\delta)}{N} & -\frac{\varepsilon a}{N} & 0 \\ -\frac{\varepsilon}{\varepsilon+\varepsilon_0} & 1 & 0 & 0 \\ 0 & 0 & 1 & -\frac{\delta}{\delta+\delta_0} \\ 0 & -\frac{\delta a}{N} & -\frac{\delta(a+\varepsilon)}{N} & 1 \end{pmatrix} \quad \text{and} \quad A_{0,X} = \begin{pmatrix} 1 & -\frac{\varepsilon(2a+\delta)}{N} & 0 \\ -\frac{\varepsilon}{\varepsilon+\varepsilon_0} & 2 & -\frac{\delta}{\delta+\delta_0} \\ 0 & -\frac{\delta(2a+\delta)}{N} & 1 \end{pmatrix}$$

with $\quad N = \dfrac{1}{(a+\delta)(a+\varepsilon)-a^2}.$

So we see for the block inversion in the small symmetric system the following characteristics:

1. The modification in general does not fulfil that V_0 is invariant with respect to $A\,X$.

2. If the links to the outside of $\mathcal{N}_2^1, \mathcal{N}_3^1$ are all equal (that means $\varepsilon = \delta = \varepsilon_0 = \delta_0$) then we have $A\,X\,v_0 \in V_0$ for constant vectors $v_0 \in V_0$.

3. If we take a look at the second row of $A_{0,X}$ and compare this with the second row of A_0 ($(A_0)_{2,\cdot} = (-\varepsilon, \varepsilon+\delta, -\delta)$) we can see that in the modified system the diagonal element is bigger than the other elements of this row. This characteristic is not so strong in the unmodified systems. More formally we obtain for $\varepsilon_0, \delta_0 > 0$:

$$\frac{a^0_{2,2}}{|a^0_{2,1}| + |a^0_{2,3}|} = \frac{\varepsilon+\delta}{\varepsilon+\delta} = 1 < \frac{2}{\frac{\varepsilon}{\varepsilon+\varepsilon_0} + \frac{\delta}{\delta+\delta_0}} = \frac{a^{0,X}_{2,2}}{|a^{0,X}_{2,1}| + |a^{0,X}_{2,3}|}.$$

In particular we obtain in the case of $\varepsilon = \varepsilon_0$ and $\delta = \delta_0$

$$\frac{a^{0,X}_{2,2}}{|a^{0,X}_{2,1}| + |a^{0,X}_{2,3}|} = 2.$$

This motivates the idea that the system of linear equations

$$A_{0,X} u_0 = R\, f$$

is more simple to solve using an iterative method than the unmodified system

$$A_{0,X} u_0 = f.$$

4. Furhtermore, in the case of $\varepsilon = \delta$ the modified coarser system also has the following structure

$$a_{i,i} > 0, \quad a_{i,j} \leq 0, \text{ for } i \neq j$$

$$\text{and} \quad a_{i,i} \geq \sum_{j=1,\, j\neq 1}^{4} |a_{i,j}|.$$

Unfortunately this characteristic does not hold in the case of $\varepsilon \neq \delta$ for the first or the fourth row.

5.3 Modifications for the symmetric model problem (two sided)

In this section we will consider the idea of a two sided modification. Again we will consider the system that is given on the four grid points $\mathcal{N}^1_1, \ldots, \mathcal{N}^1_4$. So the situation

is the same as the one at the beginning of section 5.2. The stiffness matrix is given by

$$A = \begin{pmatrix} \varepsilon + \varepsilon_0 & -\varepsilon & 0 & 0 \\ -\varepsilon & a + \varepsilon & -a & 0 \\ 0 & -a & a + \delta & -\delta \\ 0 & 0 & -\delta & \delta + \delta_0 \end{pmatrix} \tag{5.38}$$

and for the coarser grid we aggregate the points $\mathcal{N}_2^1, \mathcal{N}_3^1$ to a new one. As shown in section 4.2 for the condition of $A\, C_{DT,XX}^{-1}$ and $A\, C_{BPX,XX}^{-1}$ in the Euclidean norm the relevant constant $\gamma_{DT,XX}$ is given as

$$\gamma_{DT,XX} := \min \Big\{ t \in \mathbb{R}_+ \,:\, (A\, P_X\, A_{0,XX}^{-1}\, R_X\, v,\, (I - Q_{0,X})v) \\ \leq t\, \|A\, P_X\, A_{0,XX}^{-1}\, R_X\, v\|\, \|(I - Q_{0,X})v\|,\ \forall v \in V \Big\}.$$

As we have concluded in section 4.2 the aim is to minimise $\gamma_{DT,XX}$. The optimal constant $\gamma_{DT,XX} = 0$ is given if and only if $V_{0,X}$ is invariant with respect to A. For the given model problem

$$\{(1,0,0,0)^T, (0,1,1,0)^T, (0,0,0,1)^T\}$$

is a basis of V_0. As we have the assumption $rk(X\, P) = n_0 = 3$

$$\{X\,(1,0,0,0)^T, X\,(0,1,1,0)^T, X\,(0,0,0,1)^T\} = \{X_{.,1},\, (X_{.,2} + X_{.,3}),\, X_{.,4}\}$$

is a basis of $V_{0,X}$.

5.3.1 Exact modification

We have seen in section 4.2 that an optimal two sided modification depends on the knowledge and the use of the eigenvectors of the operator A. First of all if we know them there are easier possibilities to solve $A\, u = f$ than to use the presented preconditioner. Furthermore, it will result a modification X with $x_{i,j} \neq 0$ for almost all i, j. Hence the effort to use this modification is for big systems much higher than for the iterative method itself.

Summary: The exact modification is theoretically well known but it is not interesting for practical issues.

5.3.2 Approximations

As done for the one sided modification with the inverse of blocks we will approximate the operator A by three blocks which are motivated by the given restriction. For the two sided modification the idea will belong to the eigenvectors of the system. We set

$$B_1 = (\varepsilon + \varepsilon_0), \quad B_2 = \begin{pmatrix} a+\varepsilon & -a \\ -a & a+\delta \end{pmatrix} \quad \text{and} \quad B_3 = (\delta + \delta_0).$$

Then we determine the eigenvectors of

$$B = \begin{pmatrix} B_1 & & \\ & B_2 & \\ & & B_3 \end{pmatrix}.$$

Based on the block structure of B it is obvious that we can determine the eigenvectors for the separated blocks.

The blocks B_1, B_4 imply the eigenvectors $(e_1^1)^T, (e_4^1)^T$. For B_2 the eigenvectors are follow as

$$v_1 = \left(\frac{\delta - \varepsilon + \sqrt{4a^2 + (\delta - \varepsilon)^2}}{2a},\ 1 \right)^T \tag{5.39}$$

$$v_2 = \left(\frac{\delta - \varepsilon - \sqrt{4a^2 + (\delta - \varepsilon)^2}}{2a},\ 1 \right)^T$$

with the eigenvalue

$$\lambda_1 = \frac{2a + \delta + \varepsilon - \sqrt{4a^2 + (\delta - \varepsilon)^2}}{2}$$

$$\lambda_2 = \frac{2a + \delta + \varepsilon + \sqrt{4a^2 + (\delta - \varepsilon)^2}}{2}.$$

To take a closer look at this part we will consider the basis vector $X(e_2^1 + e_3^1)^T$ of $V_{0,X}$.

We obtain the equation system

$$\begin{pmatrix} (A\,X\,(e_2^1+e_3^1))(2) \\ (A\,X\,(e_2^1+e_3^1))(3) \end{pmatrix} = \lambda_{2,3} \begin{pmatrix} (X\,(e_2^1+e_3^1))(2) \\ (X\,(e_2^1+e_3^1))(3) \end{pmatrix}$$

$$\Leftrightarrow \quad \begin{pmatrix} (a+\varepsilon)(x_{2,2}+x_{2,3}) - a(x_{3,2}+x_{3,3}) \\ (a+\delta)(x_{3,3}+x_{3,2}) - a(x_{2,2}+x_{2,3}) \end{pmatrix} = \lambda_{2,3} \begin{pmatrix} x_{2,2}+x_{2,3} \\ x_{3,3}+x_{3,2} \end{pmatrix}$$

With the shortcuts $y_2 = x_{2,2} + x_{2,3}$ and $y_3 = x_{3,3} + x_{3,2}$ we obtain the linear system of equations

$$\begin{pmatrix} (a+\varepsilon)y_2 - ay_3 \\ (a+\delta)y_3 - ay_2) \end{pmatrix} = \lambda_{2,3} \begin{pmatrix} y_2 \\ y_3 \end{pmatrix} \quad \Leftrightarrow \quad \begin{pmatrix} (a+\varepsilon) & -a \\ -a & (a+\delta) \end{pmatrix} \begin{pmatrix} y_2 \\ y_3 \end{pmatrix} = \lambda_{2,3} \begin{pmatrix} y_2 \\ y_3 \end{pmatrix}$$

Hence we obtain that $(y_2, y_3)^T$ is an eigenvector of

$$\begin{pmatrix} (a+\varepsilon) & -a \\ -a & (a+\delta) \end{pmatrix}.$$

As there is no further condition on $x_{2,2}, x_{2,3}, x_{3,2}$ and $x_{3,3}$ we set

$$x_{2,3} = 0 = x_{3,2}$$

and $(x_{2,2}, x_{3,3})^T$ is an eigenvector of B_2. Therewith

$$\{(1,0,0,0)^T,\ (0, x_{2,2}, x_{3,3}, 0),\ (0,0,0,1)^T\}$$

is a basis of $V_{0,X}$. Since there are two eigenvectors of B_2 we have to choose one of them. Therefore we consider the situation of $\varepsilon = \delta$. In that case we obtain for the eigenvectors and eigenvalues

$$v_1 = (1,\ 1)^T, \quad \text{with} \quad \lambda_1 = \varepsilon \tag{5.40}$$

$$v_2 = (-1,\ 1)^T, \quad \text{with} \quad \lambda_2 = 2a + \varepsilon.$$

Based on these vectors it is obvious that to choose v_1 and $(x_{2,2}, x_{3,3}) = (1,1)$, respectively it means that the system will not be modified. This implies two characteristics:

1. The use of v_2 is more influential than v_1.

2. In the case of $\varepsilon = \delta$ the unmodified system is equal to the system modified with v_1.

This modification is based on the general example we presented in section 4.2.1. In this section we have mentioned that we do not know which eigenvectors we should choose. By the block structure this problem is partly solved. The question is still open which eigenvector we use for the block B_2. Based on the example of section 4.2.1 it is obvious that we have to choose one of them.

So far we have discussed only the structure of the subspace. For practical issues also the scaling of the vectors we use as columns in X or P_X play a role. To conclude this section we will define this idea of modification for an arbitrary big system. Afterwards we will discuss the aspect of the scaling of the columns of X. This discussion will lead us to additional modifications.

For a given $A \in \mathbb{R}^{n\times n}$, s.p.d. we set $X = diag(x_{1,1}, \ldots, x_{n,n})$. Furthermore, we set $x_{i,i} = 1$ if $\mathcal{N}_i^1$ that is an isolated point. If $\mathcal{N}_i^1, \mathcal{N}_j^1$ are aggregated then we define $A^{(i,j)} \in \mathbb{R}^{2\times 2}$ by

$$A^{(i,j)} := \begin{pmatrix} a_{i,i} & a_{i,j} \\ a_{j,i} & a_{j,j} \end{pmatrix}. \tag{5.41}$$

Then we set

$$\begin{pmatrix} x_{i,i} \\ x_{j,j} \end{pmatrix} = v_{i,j}$$

where $v_{i,j}$ is an eigenvector of $A^{(i,j)}$.

Now we will consider the problem of the scaling. This was already mentioned in section 4.2.1 and above, respectively. So far, for two aggregated points $\mathcal{N}_i^1, \mathcal{N}_j^1$ we have just set $(x_{i,i}, x_{j,j}) = v_{i,j}$ with an eigenvector $v_{i,j}$. To keep a consistence to $x_{k,k}$ with an isolated point $\mathcal{N}_k^1$ and to fulfil $\widetilde{O}\,\widetilde{O}^T = \widetilde{I}$ it seems reasonable to set

$$(x_{i,i}, x_{j,j}) = \frac{v_{i,j}}{\|v_{i,j}\|}.$$

As already presented in section 4.2.1 this implies $S_{0,X} = I_0$. In general we have $S_{0,X} = (R\,X^T\,X\,P)^{-1}$. By the given block structure for X it is obvious that this implies

$$S_{0,X} = diag(s_{1,1}^{0,X}, \ldots, s_{n_0,n_0}^{0,X})$$

with $s_{k,k}^{0,X} = 1$ if $\mathcal{N}_t^1$ is an isolated point with $\{k\} = I_t^{1,0}$ and $s_{k,k}^{0,X} = (x_{i,i}, x_{j,j})\,(x_{i,i}, x_{j,j})^T$ if $\mathcal{N}_i^1, \mathcal{N}_j^1$ are aggregated to $\mathcal{N}_k^0$ ($\{i,j\} = I_k^{1,0}$). Hence the scaling

$$(x_{i,i}, x_{j,j}) = \frac{v_{i,j}}{\|v_{i,j}\|}$$

implies $s_{k,k}^{0,X} = 1$ for all $k = 1, \ldots, n_0$. We obtain that $S_{0,X}$ is easy to calculate independent of the scaling. We will see in the section 7.2.2 that an advantage is given by this scaling. However there is a numerical problem in this case. We consider the situation of the matrix A given in (5.38). Additionally we assume $\varepsilon = \delta$. As we have already highlighted $(1,1)$ is in this case an eigenvector of B_2. If we use the scaling presented above we obtain

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \sqrt{1/2} & 0 & 0 \\ 0 & 0 & \sqrt{1/2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad A_{0,XX} = \begin{pmatrix} \varepsilon + \varepsilon_0 & -\frac{\varepsilon}{\sqrt{2}} & 0 \\ -\frac{\varepsilon}{\sqrt{2}} & \frac{\varepsilon+\delta}{\sqrt{4}} & -\frac{\delta}{\sqrt{2}} \\ 0 & -\frac{\delta}{\sqrt{2}} & \delta + \delta_0 \end{pmatrix}.$$

Therewith it is obvious that $A_{0,XX}$ does not hold

$$a_{i,i}^{0,XX} \geq \sum_{j=1,\, j\neq i}^{n} |a_{i,j}^{0,XX}|.$$

Since we use iterative methods to solve

$$A_{0,XX} u_0 = R_X\, f$$

we lose a useful characteristic for the methods (cf. chapter 9). This motivates to set for two aggregated points $\mathcal{N}_i^1, \mathcal{N}_j^1$ with the eigenvector $v_{i,j}$ of $A^{(i,j)}$

$$(x_{i,i}, x_{j,j}) = \sqrt{2}\frac{v_{i,j}}{\|v_{i,j}\|}.$$

For $\varepsilon = \delta$ this implies in our example that the modified method does not differ from the unmodified one. However we will see in the multigrid setting that this will again imply a problem. This one will be solved by the assumption that a condition is fulfilled.

An additional idea is to set again $X = diag(x_{1,1}, \ldots, x_{n,n})$. Then we set

$$x_{i,i} = \frac{1}{\sqrt{a_{i,i}}}$$

if $\mathcal{N}_i^1$ is an isolated point. If $\mathcal{N}_i^1, \mathcal{N}_j^1$ are aggregated we set

$$\begin{pmatrix} x_{i,i} \\ \\ x_{j,j} \end{pmatrix} = \frac{1}{\sqrt{\lambda_{i,j}}} \frac{v_{i,j}}{\|v_{i,j}\|}$$

where $v_{i,j}$ is an eigenvector of $A^{(i,j)}$ and $\lambda_{i,j}$ the associated eigenvalue. This is motivated by the idea to set $X = D_A^{-1/2} O \widetilde{I}$ as we have presented in section 4.2. However the problem of such a setting is discussed above.

5.4 Summary

To conclude this chapter we will summarise the results we have shown based on some simple characteristics. As in the last chapter we will mainly consider the one sided modifications.

For the one sided modificatios we distinguish two kinds of modifications. The exact modifications and the modification based on blocks. First it is quite obvious that if $\mathcal{N}_i^1, \mathcal{N}_j^1$ are aggregated then using $x_{i,i}, x_{i,j}, x_{j,i}$ and $x_{j,j}$ is not sufficient in any situation for an exact modification. This results as the subsystem $\mathcal{N}_i^1, \mathcal{N}_j^1$ is always influenced from other points. So we always need the influence of other points for an exact modification. As we do not use any geometrical structure we have to determine the other points by the entries of A. This can imply a much higher effort than modifications which use no other points. However we have seen that such modifications can be simple and have useful characteristics if the influence for two aggregated points is only given by one point. See therefore the example of the one dimensional convection. In the case of the two (or higher) dimensional convection we have seen that if the influence is given by different points we can control the bias based on one direction with a quite simple modification. Moreover we have seen for the convection diffusion system that if there is a main influence that has the structure of the convection than we get the same result with a little bias based on the diffusion. But the terms which are used for the modification are still easy to calculate.

For the modification based on blocks we have also for quite simple problems not so meaningful results concernig the invarianc of subspaces. However, this modification is easy to define. In particular this holds if the structure of the system is complex. Additionally the effort in an algorithm is low. This results as we need no $\mathcal{N}_k^1, k \neq i, j$ to modify the aggregation between $\mathcal{N}_i^1, \mathcal{N}_j^1$.
At least in the symmetric example we have seen that in such a situation we need a high effort and strict assumptions to obtain similar results as for the exact modifications in a convection system. The main aspect of this characteristic is that for aggregated points $\mathcal{N}_i^1, \mathcal{N}_j^1$ it has an influence whether the a neighbour is aggregated or not. In addition, the modification of the neighbours also has an influence on the values for $\mathcal{N}_i^1, \mathcal{N}_j^1$. This is not the case in the convection system. Afterwards we have seen that the modification by the inverse of blocks seems a good and simple idea for the symmetric system. However, as for a symmetric matrix A the one sided modified operator $A_{0,X}$ is not symmetric we can use this only as a two grid method. Hence, this is more or less only a theoretical result.

For the two sided modification of a symmetric system we know that the eigenvectors of the operator play a role. Afterwards we have presented approximations which are based on using eigenvectors of smaller subsystems. This idea based on the characteristics that these eigenvectors are simpler to determine and as the subsystems are given by a block matrix, the most entries of the eigenvectors are zero. Furthermore we have seen that in particular we have a diagonal matrix for the modification matrix. As for other methods based on blocks it follows that the effort is quite low.

There is an additional (positive) effect obtained by the modification. For example if we use the modification by the inverse of blocks then it is possible that the modified coarse grid operator $A_{0,X}$ has better numerical characteristics than the operator A_0. As we mainly consider the angles between the solutions and always assume the exact solution in all subspaces this is a minor effect in this work. However, we should highlight that for numerical experiments this could be the main effect.

6 Multigrid aspects for the preconditioners

So far we have introduced the preconditioning operators $C_{DT}^{-1}, C_{BPX}^{-1}$ and C_{2P}^{-1} as two grid algorithms. Now we want to generalize them to multigrid algorithems. Thereby we generalize also the approximation that is used for the inverse on the different grids. So far we have just set them as A^{-1}, A_0^{-1}. That means that we have used the exact inverse on the differnt grids. Therewith we have obtained the estimations of our intrest. Now we are more interested in the existence of these preconditioners. As in the chapter 3 it is in this chapter not necessary that the restriction is given by the aggregation method. However there will some situations in which we use assumptions that are for the aggregation method given by the condition (2.14).

6.1 Multigrid aspects for C_{BPX}^{-1}

We will define the BPX preconditioner in the setting of $J+1$ grids in a way that for $J=1$ it is the same as the two grid algorithm defined in section 3.2. So for non singular $B^{(j)} \in \mathbb{R}^{n_j \times n_j}$, $j = 0, \ldots, J$ we define the operator C_{BPX}^{-1} as follows

$$C_{BPX}^{-1}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=0}^{J} P_j \, (B^{(j)})^{-1} \, R_j. \tag{6.1}$$

First we will show a sufficient condition for the operators $B^{(j)}$, $j = 0, \ldots, J$ so that the operator $C_{BPX}^{-1}(B^{(0)}, \ldots, B^{(J)})$ is non singular. We will proof this the same way as done in the two grid case.

Lemma: 6.1.1. *Assume that there is a matrix $\widetilde{B} \in \mathbb{R}^{n \times n}$ with*

$$(R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1})(\widetilde{v}_k) = \tau_{(\widetilde{v}_k)}^k \widetilde{v}_k, \quad \tau_{(\widetilde{v}_k)}^k > 0 \tag{6.2}$$

for all $\widetilde{v}_k \in \widetilde{V}_k$ and all $k = 0, \ldots, J$. Then $C_{BPX}^{-1}(B^{(0)}, \ldots, B^{(J)})$ is non singular.

proof. We will show that based on the assumptions there is no $v \in V\backslash\{0\}$ that fulfils $C_{BPX}^{-1}\, v = 0$. Assume that such an $v \in V\backslash\{0\}$ exists. For an arbitrary $j \leq J$ we have $R_0 = R_0^j\, R_j$ and therewith follows

$$0 = C_{BPX}^{-1}\, v = \sum_{j=0}^{J} P_j\, (B^{(j)})^{-1}\, R_j\, v$$

$$\Rightarrow \quad 0 = R_0\, \widetilde{B} \left(\sum_{j=0}^{J} P_j\, (B^{(j)})^{-1}\, R_j\, v \right) = \sum_{j=0}^{J} R_0^j\, (R_j\, \widetilde{B}\, P_j\, (B^{(j)})^{-1})\, \underbrace{(R_j\, v)}_{\in \widetilde{V}_j}$$

$$= \sum_{j=0}^{J} R_0^j\, \tau^j_{(R_j\, v)} (R_j\, v) = \sum_{j=0}^{J} \tau^j_{(R_j\, v)} R_0\, v = R_0\, v \left(\sum_{j=1}^{J} \tau^j_{(R_j\, v)} \right).$$

Hence we obtain $R_0\, v = 0$. Assume now for an $k \leq J$ it is $R_i\, v = 0$ for all $i \leq k-1$. Then it follows

$$0 = \sum_{j=0}^{J} P_j\, (B^{(j)})^{-1}\, R_j\, v = \sum_{j=k}^{J} P_j\, (B^{(j)})^{-1}\, R_j\, v.$$

And this implies

$$0 = \sum_{j=k}^{J} R_k\, \widetilde{B}\, P_j\, (B^{(j)})^{-1}\, R_j\, v = \sum_{j=k}^{J} R_k^j\, (R_j\, \widetilde{B}\, P_j\, (B^{(i)})^{-1})\, \underbrace{(R_j\, v)}_{\in \widetilde{V}_j}$$

$$= \sum_{j=k}^{J} R_k^j\, \tau^j_{(R_j\, v)} (R_j\, v) = \sum_{j=k}^{J} \tau^j_{(R_j\, v)} (R_k\, v) = R_k\, v \left(\sum_{j=k}^{J} \tau^j_{(R_j\, v)} \right).$$

Hence it is $R_k\, v = 0$. Based on the argument of the induction this implies $R_J\, v = v = 0$ and this is in contradiction to the assumptions. □

The assumption of the existence of an operator $\widetilde{B}$ that fulfils the equation (6.2) seems to be quite strong. The following lemma will show that this condition can be fulfilled rather simply.

Lemma: 6.1.2. *Let $B \in \mathbb{R}^{n\times n}$ be a non singular matrix so that the matrices $B^{(j)} \in \mathbb{R}^{n_j \times n_j}$ defined as follows*

$$B^{(j)} := \frac{1}{\sigma_j}\, R_j\, B\, P_j, \quad j = 0, \ldots, J, \quad \sigma_j > 0 \tag{6.3}$$

are non singular. If we set $\widetilde{B} = B$ then the equation (6.2) holds for all $j = 0, \ldots, J$ and all $\widetilde{v}_j \in \widetilde{V}_j$ with $\tau^j_{(\widetilde{v}_j)} = \sigma_j$ for all $\widetilde{v}_j \in \widetilde{V}_j$.

proof. For an arbitrary $j \in \{0, \ldots, J\}$ we obtain from the assumptions for an arbitrary $\widetilde{v}_j \in \widetilde{V}_j$

$$(R_j\, B\, P_j\, (B^{(j)})^{-1})(\widetilde{v}_j) = \sigma_j\, B_j\, (B^{(j)})^{-1}\, (\widetilde{v}_j) = \sigma_j\, \widetilde{v}_j.$$

□

With $B = A$ and $\sigma_j = 1$ for all $j = 0, \ldots, J$ this is the situation we have considered for the two grid algorithms. This leads to a definition for the multigrid case that is, in the case of $J = 1$ the same as considered in the chapter for the two grid method. We define for a non singular $B \in \mathbb{R}^{n\times n}$ the operator C_{BPX}^{-1} as follows

$$C_{BPX}^{-1}(B) := \sum_{j=0}^{J} P_j\, B_j^{-1}\, R_j \quad \text{with} \quad B_j = R_j\, B\, P_j. \tag{6.4}$$

From the results above C_{BPX}^{-1} is non singular if this holds for the matrices B_j, $j = 0, \ldots, J$. The non singularity of the matrices $B^{(j)}$ are discussed in Lemma 2.3.5.

If we use the BPX-method as a preconditioner to solve $A\, u = f$ then the lemmata above give us the idea that the quality of the preconditioner depends on some aspects. The first one is that the coarse grid operators $B^{(j)}$ should be good approximations of A in certain subspaces. So it seems to be a good idea to set $B^{(j)}$ as done in (6.3) with $B = A$. If we do this the next aspect is that we need a good scaling operator σ_i. At last we have to consider the structure of the used subspaces. As in the two grid situation it is obvious that the angle between them has an influence on the quality of the preconditioner. For $v_k \in V_k$ and $B \in \mathbb{R}^{n\times n}$ we can decompose $B\, v_k$ as

$$B\, v_k = \widehat{Q}_k(B\, v_k) + (I - \widehat{Q}_k)(B\, v_k) = \underbrace{P_k\, \widehat{S}_k\, R_k\, (B\, v_k)}_{\in V_k} + \underbrace{(I - P_k\, \widehat{S}_k\, R_k)(B\, v_k)}_{\in V_k^{\perp}}.$$

The idea we got from the two grid method that the quality depends on the bias of B as given by

$$\frac{\|(I - \widehat{Q}_k)(B\, v_k)\|}{\|\widehat{Q}_k(B\, v_k)\|}$$

and that it is optimal if the spaces V_k are invariant with respect to B.

To specify this idea we consider a representation of $A\,C_{BPX}^{-1}(B^{(0)},\dots B^{(J)})$. We obtain

(6.5)
$$\begin{aligned} A\,C_{BPX}^{-1}(B^{(0)},\dots B^{(J)}) &= \sum_{j=0}^{J} A\,P_j\,(B^{(j)})^{-1}\,R_j \\ &= \sum_{j=0}^{J} \widehat{Q}_j\,A\,P_j\,(B^{(j)})^{-1}\,R_j + \sum_{j=0}^{J}(I-\widehat{Q}_j)\,A\,P_j\,(B^{(j)})^{-1}\,R_j \\ &= \sum_{j=0}^{J} P_j\,\widehat{S}_j\,A_j\,(B^{(j)})^{-1}\,R_j + \sum_{j=0}^{J}(I-\widehat{Q}_j)\,A\,P_j\,(B^{(j)})^{-1}\,R_j. \end{aligned}$$

If we consider the special case that $B^{(j)} = A_j$ for $j = 0,\dots,J$ then this is equivalent to

(6.6)
$$\begin{aligned} A\,C_{BPX}^{-1}(A) &= \sum_{j=0}^{J} P_j\,\widehat{S}_j\,R_j + \sum_{j=0}^{J}(I-\widehat{Q}_j)\,A\,P_j\,A_j^{-1}\,R_j \\ &= \sum_{j=0}^{J} \widehat{Q}_j + \sum_{j=0}^{J}(I-\widehat{Q}_j)\,A\,P_j\,A_j^{-1}\,R_j \end{aligned}$$

Therewith we furthermore see that we obtain

$$A\,C_{BPX}^{-1}(A) = \sum_{j=0}^{J} \widehat{Q}_j$$

if V_i is invariant with respect to A.

6.2 Multigrid aspects for C_{DT}^{-1}

In this section we will introduce the DT-method in the context of $J+1$ grids. Similary to for the BPX-method it should be done in a way that in the case $J = 1$ we get the preconditioner as given in in section 3.3. However, for the DT-method there are two possible generalisations that we will present.

6.2.1 Version 1

Again for non singular $B^{(j)} \in \mathbb{R}^{n_j \times n_j}$, $j = 0,\dots,J$ we define the operator $C_{DT,1}^{-1}$

(6.7)
$$C_{DT,1}^{-1}(B^{(0)},\dots,B^{(J)}) := \sum_{j=1}^{J} P_j\,(B^{(j)})^{-1}\,(I_j - Q_{j-1})\,R_j + P_0\,(B^{(0)})^{-1}\,R_0.$$

As we have done for the BPX-method we will show a sufficient condition for the non singularity of $C_{DT,1}^{-1}$. Similary to the proof for the BPX-method this will depend on an assumption concerning the operators $B^{(j)}$, $j = 0, \ldots, J$.

Lemma: 6.2.1. *Assume that there is a matrix $\widetilde{B} \in \mathbb{R}^{n\times n}$ implying for all $k = 1, \ldots, J$ that*

$$\widetilde{W_k} \quad \textit{is invariant with respect to} \quad (R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1}) \tag{6.8}$$

and $rk(R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1}) = n_k$ for $k = 0, \ldots, J$. Then $C_{DT,1}^{-1}(B^{(0)}, \ldots, B^{(J)})$ is non singular.

proof. We will prove that based on the assumptions there is no $v \in V\backslash\{0\}$ that fulfils $C_{DT}^{-1}(B^{(0)}, \ldots, B^{(J)})\,v = 0$. Assume that such an $v \in V\backslash\{0\}$ exists. For an arbitrary $j \leq J$ it is $R_0 = R_0^j\,R_j$. Furthermore, we remember that we can represent an arbitrary $\widetilde{w}_k \in \widetilde{W_k}$ as

$$\widetilde{w}_k = (I_k - Q_{k-1})\,R_k\,v^{(k)} = (I_k - P_k^{k-1}\,S_{k-1}\,R_{k-1}^k)\,R_k\,v^{(k)} \quad \text{with an } v^{(k)} \in V.$$

Therewith follows

$$\begin{aligned}
0 = C_{DT,1}^{-1}\,v &= \sum_{j=1}^{J} P_j\,(B^{(j)})^{-1}\,(I_j - Q_{j-1})\,R_j\,v + P_0\,(B^{(0)})^{-1}\,R_0\,v \\
\Rightarrow \quad 0 &= R_0\,\widetilde{B}\left(\sum_{j=1}^{J} P_j\,(B^{(i)})^{-1}\,(I_j - Q_{j-1})\,R_j\,v + P_0\,(B^{(0)})^{-1}\,R_0\,v\right) \\
&= \sum_{j=1}^{J} R_0^j\,(R_j\,\widetilde{B}\,P_j\,(B^{(j)})^{-1})\,\underbrace{(I_j - Q_{j-1})\,R_j\,v}_{\in\widetilde{W}_j} + (R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1})\,\underbrace{R_0\,v}_{\in\widetilde{V}_0} \\
&= \sum_{j=1}^{J} R_0^{j-1}\,\underbrace{R_{j-1}^j\,(I_j - Q_{j-1})}_{=0}\,R_j\,v^{(j)} + (R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1})\,R_0\,v \\
&= (R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1})\,R_0\,v.
\end{aligned}$$

Based on the assumption $rk(R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1}) = n_0$ we have $R_0\,v = 0$. Assume now

that for an $k \leq J$ we have $R_j v = 0$ for all $j < k$. Then we obtain

$$\begin{aligned} 0 &= \sum_{j=1}^{J} P_j \, (B^{(j)})^{-1} \, (I_j - Q_{j-1}) \, R_j \, v + P_0 \, (B^{(0)})^{-1} \, R_0 \, v \\ &= \sum_{j=k}^{J} P_j \, (B^{(j)})^{-1} \, (I_j - Q_{j-1}) \, R_j \, v. \end{aligned}$$

And this implies

$$\begin{aligned} 0 &= \sum_{j=k}^{J} R_k \, \widetilde{B} \, P_j \, (B^{(i)})^{-1} \, (I_j - Q_{j-1}) \, R_j \, v \\ &= \sum_{j=k}^{J} R_k^j \, (R_j \, \widetilde{B} \, P_j \, (B^{(i)})^{-1}) \, \underbrace{(I_j - Q_{j-1}) \, R_j \, v}_{\in \widetilde{W}_j} \\ &= (R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1}) \, (I_k - Q_{k-1}) \, R_k \, v + \sum_{j=k+1}^{J} R_k^j \, (I_j - Q_{j-1}) \, R_j \, v^{(j)} \\ &= (R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1}) \, (I_k - Q_{k-1}) \, R_k \, v + \sum_{j=k+1}^{J} R_k^{j-1} \, \underbrace{R_{j-1}^j \, (I_j - Q_{j-1})}_{=0} \, R_j \, v^{(j)} \\ &= (R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1}) \, (I_k - Q_{k-1}) \, R_k \, v. \end{aligned}$$

Based on the assumption $rk(R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1}) = n_k$ this implies $(I_k - Q_{k-1}) \, R_k \, v = 0$. Hence we obtain from the definition of Q_{k-1} and the assumption $R_j \, v = 0$ for $j < k$

$$0 = (I_k - Q_{k-1}) \, R_k \, v = R_k \, v - P_k^{k-1} \, S_{k-1} \, R_{k-1} \, v = R_k \, v.$$

Iteratively we obtain $R_J \, v = 0$. Hence the contradiction follows from $R_J \, v = v$. □

To compare the two proofs of the non singularity for $C_{DT,1}^{-1}$ and C_{BPX}^{-1} we will start by comparing the sufficient conditions for the proofs.

Lemma: 6.2.2. *If there is a non singular $\widetilde{B} \in \mathbb{R}^{n \times n}$ which fulfils*

$$(R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1})(\widetilde{v}_k) = \tau_{(\widetilde{v}_k)}^k \widetilde{v}_k, \quad \tau_{(\widetilde{v}_k)}^k > 0$$

for all $k = 0, \ldots, J$ and all $\widetilde{v}_k \in \widetilde{V}_k$ then for the matrix $\widetilde{B}$ also holds

$$\widetilde{W}_k \quad \textit{is invariant with respect to} \quad (R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1})$$

for all $k = 1, \ldots, J$ and $rk(R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1}) = n_k$ for $k = 0, \ldots, J$.

proof. If it is for a $\widetilde{B}$

$$(R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1})(\widetilde{v}_k) = \tau^k_{(\widetilde{v}_k)}\widetilde{v}_k, \quad \tau^k_{(\widetilde{v}_k)} > 0$$

then $\widetilde{W}_k \subset \widetilde{V}_k$ is invariant with respect to this operator. Furthermore, the assumption obviously implies the non singularity of $R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1}$. □

Remark: 6.2.3. *Lemma 6.2.2 implies together with Lemma 6.1.2 that the technical condition about the invariance of subspaces is easily fulfilled if there is an non singular $B \in \mathbb{R}^{n\times n}$ for that we have*

$$B^{(j)} := \frac{1}{\sigma_j} R_j\,B\,P_j, \quad \textit{with} \quad \sigma_j > 0, j = 0,\ldots,J.$$

Additionally the assumption

$$rk(R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1}) = n_k$$

is equivalent to

$$rk(R_k\,\widetilde{B}\,P_k) = n_k.$$

For a matrix $\widetilde{B} \in \mathbb{R}^{n\times n}$ this is discussed in Lemma 2.3.5.
So at the first view it seems that the sufficient condition for the DT-method is weaker than the condition for BPX-method. But in the proof for $C^{-1}_{DT,1}$ we additionally use

$$P_k^{k-1}\,S_{k-1}\,R_{k-1}^k : \widetilde{V}_k \to P_k^{k-1}(\widetilde{V}_{k-1})$$

is the orthogonal projection concerning the inner product $(.,.)$. This is equivalent to the condition $S_{k-1} = (R^k_{k-1}\,P_k^{k-1})^{-1}$. That means that we have to calculate for each $j = 0,\ldots,J-1$ an inverse matrix. For the aggregation method this is possible without an additional big effort as $(R^k_{k-1}\,P_k^{k-1})$ is a diagonal matrix. But this is not so easy if we want to use other restriction operators as for example the standard geometrical restriction. This is not necessary for the BPX-method.

As we have done for the BPX-method we have the analogy to the two grid method also for the DT-method. We define $C^{-1}_{DT,1}(B)$ by

$$C^{-1}_{DT,1}(B) := \sum_{j=0}^{J} P_j\,B_j^{-1}\,(I_j - Q_{j-1})\,R_j + P_0\,B_0^{-1}\,R_0 \tag{6.9}$$

$$\text{with} \quad B_j = R_j\,B\,P_j, \quad j = 0,\ldots,J. \tag{6.10}$$

Based on the results above this is a non singular matrix if the matrices B_j, $j = 0, \ldots, J$ are non singular. This is the result if we set $\widetilde{B} = B$ in Lemma 6.2.1. Then B_j^{-1} is well posed and we have

$$(R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1}) = (R_k \, B \, P_k \, B_k^{-1}) = I_k.$$

Hence $\widetilde{W_k}$ is invariant with respect to $(R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1})$ and we obtain

$$rk(R_k \, \widetilde{B} \, P_k \, (B^{(k)})^{-1}) = n_k$$

Thus the quality of the DT-method as a preconditioner for $A \, u = f$ depends on the aspects that have been important for the BPX-method. The first is that the matrices $B^{(j)}$ should be good approximations for A_j. The second is again that if we decompose an arbitrary $v_k \in V_k$

$$A \, v_k = \widehat{Q}_k(A \, v_k) + (I - \widehat{Q}_k)(A \, v_k) = \underbrace{P_k \, \widehat{S}_k \, R_k \, (A \, v_k)}_{\in V_k} + \underbrace{(I - P_k \, \widehat{S}_k \, R_k)(A \, v_k)}_{\in V_k^{\perp}}.$$

Then the quality of the preconditioner depends on the bias given by

$$\frac{\|(I - \widehat{Q}_k)(A \, v_k)\|}{\|\widehat{Q}_k(A \, v_k)\|}.$$

This is obvious if we take a look at the following representations of $A \, C_{DT,1}^{-1}(B^{(0)}, \ldots B^{(J)})$ that will conclude this section and motivate modifications as

done in the two grid case. We have

(6.11)

$$\begin{aligned}
&A\, C_{DT,1}^{-1}(B^{(0)},\dots B^{(J)}) \\
&\qquad = \sum_{j=1}^{J} A\, P_j\, (B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j + A\, P_0\, (B^{(0)})^{-1}\, R_0 \\
&\qquad = \sum_{j=1}^{J} \widehat{Q}_j\, A\, P_j\, (B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j + \widehat{Q}_0\, A\, P_0\, (B^{(0)})^{-1}\, R_0 \\
&\qquad\quad + \sum_{j=1}^{J} (I - \widehat{Q}_j)\, A\, P_j\, (B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j + (I - \widehat{Q}_0)\, A\, P_0\, (B^{(0)})^{-1}\, R_0 \\
&\qquad = \sum_{j=1}^{J} P_j\, \widehat{S}_j\, A_j\, (B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j + P_0\, \widehat{S}_0\, A_0\, (B^{(0)})^{-1}\, R_0 \\
&\qquad\quad + \sum_{j=0}^{J} (I - \widehat{Q}_j)\, A\, P_j\, (B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j + (I - \widehat{Q}_j)\, A\, P_0\, (B^{(0)})^{-1}\, R_0.
\end{aligned}$$

If we consider the special case that $B^{(j)} = A_j$ for $j = 0, \dots, J$ then this is equivalent to

$$\begin{aligned}
A\, C_{DT,1}^{-1}(A) &= \sum_{j=1}^{J} P_j\, \widehat{S}_j\, (I_j - Q_{j-1})\, R_j + P_0\, \widehat{S}_0\, R_0 \\
&\quad + \sum_{j=0}^{J} (I - \widehat{Q}_j)\, A\, P_j\, A_j^{-1}\, (I_j - Q_{j-1})\, R_j + (I - \widehat{Q}_0)\, A\, P_0\, A_0^{-1}\, R_0.
\end{aligned} \tag{6.12}$$

If we have $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ for $j = 1, \dots, J$ we get based on Lemma 2.3.8 that this is the same as

$$\begin{aligned}
A\, C_{DT,1}^{-1}(A) &= \sum_{j=1}^{J} (\widehat{Q}_j - \widehat{Q}_{j-1}) + Q_0 \\
&\quad + \sum_{j=0}^{J} (I - \widehat{Q}_j)\, A\, P_j\, (B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j + (I - \widehat{Q}_0)\, A\, P_0\, (B^{(0)})^{-1}\, R_0.
\end{aligned}$$

Therewith we obtain that if V_i is invariant with respect to A then

$$A\, C_{DT,1}^{-1}(A) = \sum_{j=1}^{J} (\widehat{Q}_j - \widehat{Q}_{j-1}) + \widehat{Q}_0 = I.$$

To conclude this we want to highlight that if the equation $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ does not hold we have another kind of bias because even if V_i is invariant with respect to A we only obtain

$$A\, C_{DT,1}^{-1}(A) = \sum_{j=1}^{J} P_j\, \widehat{S}_j\, (I_j - Q_{j-1})\, R_j + P_0\, \widehat{S}_0\, R_0$$

$$= \sum_{j=1}^{J} (\widehat{Q}_j - P_j\, \widehat{S}_j\, Q_{j-1}\, R_j) + \widehat{Q}_0.$$

This leads to the idea of another kind of modification.

Remark: 6.2.4. *Based on Lemma 2.4.8 we obtain that for the aggreagation method the equation $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ holds if and only if the condition (2.14) is fulfilled.*

6.2.2 Version 2

For the DT-methods we have seen that in the context of two grids it holds $A\, C_{DT}^{-1} = I$ if V_0 is A-invariant. In the last section we have seen that for the multigrid situation we need an additional condition to obtain this property. Hence we will propose a second generalisation for the two grid method. This one should hold that we only need the assumption that V_i is A-invariant to obtain $A\, C_{DT,2}^{-1} = I$.
In other words we want to get a certain consistence without using the additional assumption that $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ holds or the condition (2.14) is fulfilled, respectively.
For non singular $B^{(j)} \in \mathbb{R}^{n_j \times n_j}$, $j = 0, \ldots, J$ we define the operator $C_{DT,2}^{-1}$ as follows

(6.13)

$$C_{DT,2}^{-1}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=1}^{J} P_j\, (B^{(j)})^{-1}\, (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1} R_{j-1}^{j})\, R_j + P_0\, (B^{(0)})^{-1}\, R_0.$$

Then we have a similar result for the non singularity of $C_{DT,2}^{-1}(B^{(0)}, \ldots, B^{(J)})$ as in the last section for $C_{DT,1}^{-1}(B^{(0)}, \ldots, B^{(J)})$.

Lemma: 6.2.5. *Assume that there is a matrix $\widetilde{B} \in \mathbb{R}^{n \times n}$ that implies*

(6.14)

$$Im(((I_k - \widehat{S}_k^{-1}\, P_k^{k-1}\, \widehat{S}_{k-1}\, R_{k-1}^{k}) R_k)(V)) \quad \textit{is invariant with respect to} \quad (R_k\, \widetilde{B}\, P_k\, (B^{(k)})^{-1})$$

for $k = 1, \ldots, J$ and $rk(R_k\, \widetilde{B}\, P_k\, (B^{(k)})^{-1}) = n_k$ for $k = 0, \ldots, J$. Then $C_{DT,2}^{-1}(B^{(0)}, \ldots, B^{(J)})$ is non singular.

proof. We will show that based on the assumption there is no $v \in V\backslash\{0\}$ with $C_{DT,2}^{-1}(B^{(0)},\dots,B^{(J)})\,v = 0$. Assume that such an $v \in V\backslash\{0\}$ exists. From the assumption (6.14) on the invarianve follows that

$$(R_j\,\widetilde{B}\,P_j\,(B^{(j)})^{-1})(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j)\,R_j\,v = (I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j)\,R_j\,v^j$$

for all $j = 1,\dots j$ with an $v^j \in V$. Hence we obtain

$$\begin{aligned}
0 &= \sum_{j=1}^{J} P_j\,(B^{(j)})^{-1}(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j)\,R_j\,v + P_0\,(B^{(0)})^{-1}\,R_0\,v \\
\Rightarrow\quad 0 &= \sum_{j=1}^{J} \widehat{Q}_0\,\widetilde{B}\,P_j\,(B^{(j)})^{-1}(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j)\,R_j\,v + \widehat{Q}_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1}\,R_0\,v \\
&= \sum_{j=1}^{J} P_0\,\widehat{S}_0\,R_0^j\,(R_j\,\widetilde{B}\,P_j\,(B^{(j)})^{-1})(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j)\,R_j\,v \\
&\quad + P_0\,\widehat{S}_0\,(R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1})\,R_0\,v \\
&= \sum_{j=1}^{J} P_0\,\widehat{S}_0\,R_0^j\,(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j)\,R_j\,v^j + P_0\,\widehat{S}_0\,(R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1})\,R_0\,v \\
&= \sum_{j=1}^{J} \Big(\widehat{Q}_0 - P_0\,\widehat{S}_0\,R_0^j\,\widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j\Big)v^j + P_0\,\widehat{S}_0\,(R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1})\,R_0\,v \\
&= \sum_{j=1}^{J} \Big(\widehat{Q}_0 - P_0\,\widehat{S}_0\,R_0^j\,R_j\,P_j\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j\Big)v^j + P_0\,\widehat{S}_0\,(R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1})\,R_0\,v \\
&= \sum_{j=1}^{J} \Big(\widehat{Q}_0 - \widehat{Q}_0\widehat{Q}_{j-1}\Big)v^j + P_0\,\widehat{S}_0\,(R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1})\,R_0\,v \\
&= P_0\,\widehat{S}_0\,(R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1})\,R_0\,v.
\end{aligned}$$

As we have

$$rk(\widehat{S}_0) = rk(R_0\,\widetilde{B}\,P_0\,(B^{(0)})^{-1}) = rk(P_0) = n_0$$

this implies $R_0\,v = 0$. Assume that we have $R_j\,v = 0$ for an $k \le J$ and all $j < k$. Then

it follows

$$
\begin{aligned}
0 &= \sum_{j=1}^{J} P_j\,(B^{(j)})^{-1}(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\,R_j\,v + P_0\,(B^{(0)})^{-1}\,R_0\,v \\
&= \sum_{j=k}^{J} P_j\,(B^{(j)})^{-1}(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\,R_j\,v \\
\Rightarrow \quad 0 &= \sum_{j=k}^{J} \widehat{Q}_k\,\widetilde{B}\,P_j\,(B^{(j)})^{-1}(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\,R_j\,v \\
&= \sum_{j=k+1}^{J} \widehat{Q}_k\,\widetilde{B}\,P_j\,(B^{(j)})^{-1}(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\,R_j\,v \\
&\quad + \widehat{Q}_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1}(R_k - \widehat{S}_k^{-1}\,P_k^{k-1}\,\widehat{S}_{k-1}\,R_{k-1})\,v \\
&= \sum_{j=k+1}^{J} P_k\,\widehat{S}_k\,R_k^{j}\,(R_j\,\widetilde{B}\,P_j\,(B^{(j)})^{-1})(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\,R_j\,v \\
&\quad + P_k\,\widehat{S}_k\,(R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1})\,R_k\,v \\
&= \sum_{j=k+1}^{J} P_k\,\widehat{S}_k\,R_k^{j}\,(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\,R_j\,v^j \\
&\quad + P_k\,\widehat{S}_k\,(R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1})\,R_k\,v \\
&= \sum_{j=k+1}^{J} \left(P_k\,\widehat{S}_k\,R_k - \underbrace{P_k\widehat{S}_k\,R_k^{j}\,R_j}_{\widehat{Q}_k}\,\underbrace{P_j\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j}\,R_j}_{\widehat{Q}_{j-1}} \right) v^j \\
&\quad + P_k\,\widehat{S}_k\,(R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1})\,R_k\,v \\
&= \sum_{j=k+1}^{J} (\widehat{Q}_k - \widehat{Q}_k\,\widehat{Q}_{j-1})\,v^j + P_k\,\widehat{S}_k\,(R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1})\,R_k\,v \\
&= P_k\,\widehat{S}_k\,(R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1})\,R_k\,v.
\end{aligned}
$$

From the assumption

$$rk(\widehat{S}_k) = rk(R_k\,\widetilde{B}\,P_k\,(B^{(k)})^{-1}) = rk(P_k) = n_k$$

follows $R_k\,v = 0$. Hence $R_J\,v = v = 0$ follows based on the argument of induction. □

Of course it is again a good idea to set $B^{(j)} = A_j = R_j\, A\, P_j$. We obtain in this situation

$$\begin{aligned}
A\, P_j\, A_j^{-1}(I_j - &\widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j)\, R_j \\
&= \widehat{Q}_j\, A\, P_j\, A_j^{-1}(I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j)\, R_j \\
&\quad + (I - \widehat{Q}_j)\, A\, P_j\, A_j^{-1}(I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j)\, R_j \\
&= P_j\, \widehat{S}_j\, (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j)\, R_j \\
&\quad + (I - \widehat{Q}_j)\, A\, P_j\, A_j^{-1}(I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j)\, R_j \\
&= (\widehat{Q}_j - P_j\, \widehat{S}_j\, R_j\, P_j\, P_j^{j-1}\, \widehat{S}_{j-1}\, Rj - 1) \\
&\quad + (I - \widehat{Q}_j)\, A\, P_j\, A_j^{-1}(I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j)\, R_j \\
&= (\widehat{Q}_j - \widehat{Q}_{j-1}) \\
&\quad + (I - \widehat{Q}_j)\, A\, P_j\, A_j^{-1}(I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j)\, R_j.
\end{aligned}$$

This implies

$$\begin{aligned}
A\, C_{DT,2}^{-1}(A) = &\sum_{j=1}^{J}(\widehat{Q}_j - \widehat{Q}_{j-1}) + \widehat{Q}_0 \\
&+ \sum_{j=1}^{J}(I - \widehat{Q}_j)\, A\, P_j\, A_j^{-1}(I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j)\, R_j + (I - \widehat{Q}_0) A\, P_0\, A_0^{-1}\, R_0.
\end{aligned}$$

And if we additionally have V_i that is invariant with respect to A this implies

$$A\, C_{DT,2}^{-1}(A) = \sum_{j=1}^{J}(\widehat{Q}_j - \widehat{Q}_{j-1}) + \widehat{Q}_0 = I.$$

That means that if we take the operator $C_{DT,2}^{-1}$ instead of $C_{DT,1}^{-1}$ then we can drop the assumption that the equation $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ holds or the condition (2.14) is fulfilled to get an easy representation of $A\, C_{DT,2}^{-1}\, v$ as the sum of the identity and a bias. The bias is for this operator again induced by the non invariance of V_i with respect to A. However using $C_{DT,2}^{-1}$ instead of $C_{DT,1}^{-1}$ implicates a higher effort. The matrices $\widehat{S}_i$ have to be calculated and saved. For the aggregation method the effort is not so high as the matrices $\widehat{S}_j,\ j = 0, \ldots, J$ are diagonal matrices too.

We will conclude the section with the proof that if the equation $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ holds the two methods $C_{DT,1}^{-1}$ and $C_{DT,2}^{-1}$ are the same.

Lemma: 6.2.6. *Let $C_{DT,1}^{-1}, C_{DT,2}^{-1}$ be as defined in (6.7),(6.7). If we assume that the equation $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ holds then we have*

$$C_{DT,1}^{-1} = C_{DT,2}^{-1}.$$

proof. Based on the assumptions we obtain

$$\begin{aligned} Q_{j-1} &= P_j^{j-1}\, S_{j-1}\, R_{j-1}^j = \widehat{S}_j^{-1}\, \widehat{S}_j\, P_j^{j-1}\, S_{j-1}\, R_{j-1}^j \\ &= \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j. \end{aligned}$$

This proves the assertion. □

6.3 Multigrid aspects for C_{2P}^{-1}

Similarly to the DT-method we will present two modifications for the generalisation of the $2P$-method. They will distinguish by the assumption of the equation $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ holds or the condition (2.14) is fulfilled respectively. If we consider the aggregation method then we can sum this up to the assumption that the condition (2.14) is fulfilled. Moreover, for the first version of this preconditioner, we will need this condition for the non singularity of the preconditioner and not only for a good estimation in a theoretical situation.

6.3.1 Version 1

We will introduce a generalisation of the the $2P$-method in the context of $J+1$ grids that is similar to the generalisation done for $C_{DT,1}^{-1}$. So again for non singular $B^{(j)} \in \mathbb{R}^{n_j \times n_j}$, $j = 0, \ldots, J$ we define the operator $C_{2P,1}^{-1}$ by

$$C_{2P,1}^{-1}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=1}^{J} P_j\, (I_j - Q_{j-1})(B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j + P_0\, (B^{(0)})^{-1}\, R_0.$$

First we have that for symmetric $B^{(j)},\ j = 0, \ldots, J$ the operator $C_{2P,1}^{-1}$ is symmetric, too. This follows from

$$\begin{aligned} &\Big(P_j\, (I_j - Q_{j-1})(B^{(j)})^{-1}\, (I_j - Q_j - 1)\, R_j\Big)^T \\ &\qquad = \Big((I_j - Q_{j-1})\, R_j\Big)^T \Big((B^{(j)})^{-1}\Big)^T \Big(P_j\, (I_j - Q_{j-1})\Big)^T \\ &\qquad = P_j\, (I_j - Q_{j-1})(B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j. \end{aligned}$$

If we want to consider the non singularity of this preconditioner we need stronger assumptions than for the DT or the $BPX-$method. First we need a simple result for the decomposition of an $v \in V$ by the operators

$$(I_j - P_j^{j-1} S_{j-1} R_{j-1}^j) R_j v, \quad R_0 v.$$

Lemma: 6.3.1. *Based on the definition of R_j^k there is no $v \in V\backslash\{0\}$ with*

$$(I_j - P_j^{j-1} S_{j-1} R_{j-1}^j) R_j v = 0 \quad \textit{for all} \quad j = 1, \ldots, J$$
$$\textit{and} \quad R_0 v = 0.$$

proof. Assume that such an $v \in V\backslash\{0\}$ exists. Then it is in particular $R_0 v = 0$. Assume now that for an $k \leq J$ we have $R_j v = 0$ for all $j < k$. We obtain from the assumption

$$0 = (I_k - P_k^{k-1} S_{k-1} R_{k-1}^k) R_k v = R_k v - P_k^{k-1} S_{k-1} R_{k-1} v = R_k v.$$

Based on the argument of the induction it follows $R_J v = v = 0$. This gives the contradiction to the assumption. □

We should highlight two aspects of Lemma 6.3.1. The first is that this lemma does not need any assumption concerning S_{k-1}. The second is that the lemma does not give any information whether

$$\sum_{j=1}^{J} P_j (I_j - P_j^{j-1} S_{j-1} R_{j-1}^j) R_j + P_0 R_0$$

is singular or not. However, now we can give a proof for the non singularity of the operator $C_{2P,1}^{-1}$.

Lemma: 6.3.2. *Let $B^{(j)} \in \mathbb{R}^{n_j \times n_j}$ be non singular, with*

(6.15)
$$(I_j - Q_{j-1}) (B^{(j)})^{-1} (I_j - Q_{j-1}) \widetilde{v}_j = 0 \;\Rightarrow\; (I_j - Q_{j-1}) \widetilde{v}_j = 0 \quad \textit{for} \quad j = 1, \ldots, J.$$

If we assume that

(6.16)
$$ker(R_{i-1}^i) = ker(R_{i-1}^i R_i P_i)$$

holds, then the operator $C_{2P,1}^{-1}$ is non singular.

proof. Assume that there is an $v \in V \backslash \{0\}$ with $C_{2P,1}^{-1}\, v = 0$. Then it follows by Lemma 6.3.1 that

$$(I_j - P_j^{j-1}\, S_{j-1}\, R_{j-1}^{j})\, R_j\, v \quad \text{for } j = 1, \ldots, J \quad \text{and} \quad R_0\, v$$

are not all zero. Based on the assumption on the non singularity of $(B^{(j)})^{-1}$, $j = 0, \ldots J$ it follows that

$$(B^{(j)})^{-1}\, (I_j - P_j^{j-1}\, S_{j-1}\, R_{j-1}^{j})\, R_j\, v \quad j = 1, \ldots, J \quad \text{and} \quad (B^{(0)})^{-1}\, R_0\, v$$

does not all vanish. Briefly we write

$$\begin{aligned} &(B^{(j)})^{-1}\, (I_j - P_j^{j-1}\, S_{j-1}\, R_{j-1}^{j})\, R_j\, v = u_j \in \mathbb{R}^{n_j} \quad \text{for } j = 1, \ldots, J \\ &\quad \text{and} \quad (B^{(0)})^{-1}\, R_0\, v = u_0 \in \mathbb{R}^{n_0}. \end{aligned}$$

This implies

$$\begin{aligned} 0 = C_{2P,1}^{-1}\, v &= \sum_{j=1}^{J} P_j\, (I_j - Q_{j-1}) u_j + P_0\, u_0 \\ \Rightarrow \quad 0 &= R_0 \left(\sum_{j=1}^{J} P_j\, (I_j - Q_{j-1}) u_j + P_0\, u_0 \right) \\ &= \sum_{j=1}^{J} R_0^{j-1}\, R_{j-1}^{j}\, R_j\, P_j\, (I_j - Q_{j-1}) u_j + R_0\, P_0\, u_0 \end{aligned} \tag{6.17}$$

Based on the definition of S_{j-1} as $S_{j-1} = R_{j-1}^{j}\, P_j^{j-1}$ we obtain $R_{j-1}^{j}\, (I_j - Q_{j-1}) u_j = 0$ for all $u_j \in \mathbb{R}^{n_j}$. From the assumption (6.16) follows that

$$R_{j-1}^{j}\, R_j\, P_j\, (I_j - Q_{j-1}) u_j = 0$$

holds for all $j = 1, \ldots, J$. Therewith it follows from (6.17) $R_0\, P_0\, u_0 = \widehat{S}_0^{-1}\, u_0 = 0$. Hence we have $u_0 = 0$. Assume now that it is $u_j = 0$ for an $k \leq J$ and all $j < k$. Then it

follows

$$\begin{aligned}
0 &= R_k \left(\sum_{j=1}^{J} P_j \left(I_j - Q_{j-1} \right) u_j + P_0 \, u_0 \right) \\
&= \sum_{j=k}^{J} R_k \, P_j \left(I_j - Q_{j-1} \right) u_j \\
&= R_k \, P_k \left(I_k - Q_{k-1} \right) u_k + \sum_{j=k+1}^{J} R_k^{j-1} \, R_{j-1}^{j} \, R_j \, P_j \left(I_j - Q_{j-1} \right) u_j \\
&= R_k \, P_k \left(I_k - Q_{k-1} \right) u_k.
\end{aligned}$$

Thereby follows the last equation again from the assumption (6.16) and $R_{j-1}^{j} \left(I_j - Q_{j-1} \right) u_j = 0$ for all $j = k+1, \ldots, J$. By the definition of u_k and $rk(R_k \, P_k) = n_k$ this is equivalent to

$$\left(I_k - Q_{k-1} \right) \left(B^{(k)} \right)^{-1} \left(I_k - Q_{k-1} \right) R_k \, v = 0.$$

Based on the assumption (6.15) this implies $\left(I_k - Q_{k-1} \right) R_k \, v = 0$. Hence we have also

$$u_k = \left(B^{(k)} \right)^{-1} \left(I_k - Q_{k-1} \right) R_k \, v = 0.$$

From the argument of the induction it follows $u_j = 0$, $j = 0, \ldots, J$. This is in contradiction to the assumption of $v \neq 0$. □

We want to highlight that the condition (6.16) follows for the aggregation method from the condition (2.14). This is proved in Lemma 2.4.9.

To conclude this section we will take a look at the condition used in the proof of Lemma 6.3.2. As above mentioned the second condition is for the main aspect of this work equivalent to the assumption that condition (2.14) is fulfilled. This condition is well-known from section 2.4.3. So we consider here only the other assumption given by (6.15). If we set $B^{(i)} = A_i = R_i \, A \, P_i$ then the condition is equivalent to the assumption that A_i is non singular. This we will show in the next lemma. Moreover we have shown that if A is s.p.d. then we obtain that A_i is also s.p.d. In particular this implies that A_i is not singular.

Lemma: 6.3.3. *If we set $B^{(j)} = A_j = R_j\, A\, P_j$ and A_j is non singular for $j = 0, \ldots, J$ then the condition*

(6.18)
$$(I_j - Q_{j-1})\,(B^{(j)})^{-1}\,(I_j - Q_{j-1})\widetilde{v}_j = 0 \;\Rightarrow\; (I_j - Q_{j-1})\widetilde{v}_j = 0 \quad \textit{for} \quad j = 1, \ldots, J.$$

holds. In particular (6.18) holds if we set $B^{(j)} = A_j = R_j\, A\, P_j$ and A is symmetric positive definite.

proof. If there is an $\widetilde{v}_j \in \widetilde{V}_j$ with $\widetilde{w}_j = (I_j - Q_{j-1})\widetilde{v}_j \neq 0$ and

$$(I_j - Q_{j-1})\, A_j^{-1}\,(I_j - Q_{j-1})\,\widetilde{v}_j = 0 \tag{6.19}$$

then it follows from the non singularity of A_j

$$A_j^{-1}\,(I_j - Q_{j-1})\,\widetilde{v}_j = \widetilde{v}_j^* \neq 0.$$

It follows from (6.19)

$$(I_j - Q_{j-1})\,\widetilde{v}_j^* = 0 \quad \Leftrightarrow \quad \widetilde{v}_j^* = Q_{j-1}\,\widetilde{v}_j^*.$$

We highlight that $S_{j-1}\, R_{j-1}^j \widetilde{v}_j^* \neq 0$ follows from this equation. Hence we obtain

$$\begin{aligned} A_{j-1}\, S_{j-1}\, R_{j-1}^j \widetilde{v}_j^* &= R_{j-1}^j\, A_j\, P_j^{j-1}\, S_{j-1}\, R_{j-1}^j \widetilde{v}_j^* \\ &= R_{j-1}^j\, A_j\, Q_{j-1}\, \widetilde{v}_j^* = R_{j-1}^j\, A_j\, \widetilde{v}_j^* \\ &= R_{j-1}^j\, A_j\, A_j^{-1}\,(I_j - Q_{j-1})\widetilde{v}_j \\ &= R_{j-1}^j\,(I_j - Q_{j-1})\widetilde{v}_j = 0. \end{aligned}$$

This is in contradiction to the non singularity of A_{j-1}.
The additional result follows because we obtain from Lemma 2.3.5 that A_j is non singular in this case. □

So in the case of a symmetric A the condition for the non singularity is given by condition (2.14). To conclude this section we want to summarize the result concerning the non singularity of $C_{2P,1}^{-1}$ for the aggregation method and a symmetric positive definite operator A.

Theorem: 6.3.4. *Let $A \in \mathbb{R}^{n\times n}$ be s.p.d. Assume that the aggregation method is used to construct $V_{J-1}, \ldots, V_0$ and the condition (2.14) is fulfilled. Then $C_{2P,1}^{-1}$ with $B^{(j)} = A_j^{-1}$ is non singular.*

proof. The proof follows from the arguments presented in this section. □

6.3.2 Version 2

Again as done for the DT-method we will introduce a second version that is independent of the condition (2.14). Hence we define for non singular $B^{(i)} \in \mathbb{R}^{n_i \times n_i}$, $i = 0, \ldots, J$ the operator $C_{2P,2}^{-1}$ by

$$C_{2P,2}^{-1}(B^{(0)}, \ldots, B^{(J)}) :=$$

$$\sum_{j=1}^{J} P_j \, (I_j - P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j \, \widehat{S}_j^{-1})(B^{(j)})^{-1} \, (I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j) \, R_j$$

$$+ P_0 \, (B^{(0)})^{-1} \, R_0.$$

As a first result it is obvious that for symmetric $B^{(j)}$, $j = 0, \ldots, J$ the operator $C_{2P,2}^{-1}$ is symmetric, too. This follows from

$$\begin{aligned}
&\Big(P_j \, (I_j - P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j \, \widehat{S}_j^{-1})(B^{(j)})^{-1} \, (I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j) \, R_j\Big)^T \\
&= \Big((I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j) \, R_j\Big)^T \Big((B^{(j)})^{-1}\Big)^T \Big(P_j \, (I_j - P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j \, \widehat{S}_j^{-1})\Big)^T \\
&= P_j \, (I_j - P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j \, \widehat{S}_j^{-1})(B^{(j)})^{-1} \, (I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j) \, R_j.
\end{aligned}$$

For a result concerning the non singularity of $C_{2P,2}^{-1}$ we need a technical result similar to Lemma 6.3.1.

Lemma: 6.3.5. *Based on the definition of $R_J = I$ and $R_j = R_j^{j+1} \, R_{j+1}$ there is no $v \in V \backslash \{0\}$ with*

$$(I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j) \, R_j \, v = 0 \quad \textit{for all} \quad j = 1, \ldots, J$$

$$\textit{and} \quad R_0 \, v = 0.$$

proof. The proof holds based on the same arguments as the proof of Lemma 6.3.1. □

Therewith, we obtain the result of our interest:

Lemma: 6.3.6. *Let $B^{(j)} \in \mathbb{R}^{n_j \times n_j}$ be non singular, with*

$$(I_j - P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j \, \widehat{S}_j^{-1}) \, (B^{(j)})^{-1} \, (I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j) \widetilde{v}_j = 0 \tag{6.20}$$

$$\Rightarrow \quad (I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j) \widetilde{v}_j = 0 \quad \textit{for} \quad j = 1, \ldots, J.$$

Then the operator $C_{2P,2}^{-1}$ is non singular.

proof. Assume that there is an $v \in V\backslash\{0\}$ with $C_{2P,2}^{-1} v = 0$. Then it follows from Lemma 6.3.5 that

$$(I_j - \widehat{S}_j^{-1} P_j^{j-1} \widehat{S}_{j-1} R_{j-1}^j) R_j v \quad j = 1, \ldots, J \quad \text{and} \quad R_0 v$$

are not all zero. From the assumption on the non singularity of $(B^{(j)})^{-1}$, $j = 0, \ldots J$ it follows that

$$(B^{(j)})^{-1} (I_j - \widehat{S}_j^{-1} P_j^{j-1} \widehat{S}_{j-1} R_{j-1}^j) R_j v \quad \text{for } j = 1, \ldots, J \quad \text{and} \quad (B^{(0)})^{-1} R_0 v$$

does not all vanish. Briefly we write

$$\begin{aligned}(B^{(j)})^{-1} (I_j - \widehat{S}_j^{-1} P_j^{j-1} \widehat{S}_{j-1} R_{j-1}^j) R_j v = u_j \in \mathbb{R}^{n_j} \quad \text{for } j = 1, \ldots, J \\ \text{and} \quad (B^{(0)})^{-1} R_0 v = u_0 \in \mathbb{R}^{n_0}.\end{aligned}$$

This implies

$$\begin{aligned}
0 = C_{2P,2}^{-1} v &= \sum_{j=1}^{J} P_j (I_j - P_j^{j-1} \widehat{S}_{j-1} R_{j-1}^j \widehat{S}_j^{-1}) u_j + P_0 u_0 \\
\Rightarrow \quad 0 &= R_0 \left(\sum_{j=1}^{J} P_j (I_j - P_j^{j-1} \widehat{S}_{j-1} R_{j-1}^j \widehat{S}_j^{-1}) u_j + P_0 u_0 \right) \\
&= \sum_{j=1}^{J} R_0^{j-1} R_{j-1}^j R_j P_j (I_j - P_j^{j-1} \widehat{S}_{j-1} R_{j-1}^j \widehat{S}_j^{-1}) u_j + R_0 P_0 u_0 \\
&= \sum_{j=1}^{J} R_0^{j-1} (R_{j-1}^j \widehat{S}_j^{-1} - R_{j-1}^j R_j P_j P_j^{j-1} \widehat{S}_{j-1} R_{j-1}^j \widehat{S}_j^{-1}) u_j + R_0 P_0 u_0 \\
&= \sum_{j=1}^{J} R_0^{j-1} (R_{j-1}^j \widehat{S}_j^{-1} - R_{j-1}^j \widehat{S}_j^{-1}) u_j + R_0 P_0 u_0 \\
&= R_0 P_0 u_0.
\end{aligned}$$

This implies $u_0 = 0$. The rest of the proof follows again the argument of induction where the same calculation is used in each step. □

Next we will take a look at the conditions of Lemma 6.3.6. We will show that we get for this generalisation a similar result as in Lemma 6.3.3.

Lemma: 6.3.7. *If we set $B^{(j)} = A_j = R_j\, A\, P_j$ and A_j is non singular for $j = 0, \ldots, J$ then the condition*

$$(I_j - P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j}\,\widehat{S}_j^{-1})\,(B^{(j)})^{-1}\,(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\widetilde{v}_j = 0 \tag{6.21}$$
$$\Rightarrow \quad (I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\widetilde{v}_j = 0 \quad \textit{for} \quad j = 1, \ldots, J.$$

holds. In particular (6.21) holds if we set $B^{(j)} = A_j = R_j\, A\, P_j$ and A is symmetric positive definite.

proof. If there is an $\widetilde{v}_j \in \widetilde{V}_j$ with $\widetilde{w}_j = (I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\widetilde{v}_j \neq 0$ and

$$(I_j - P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j}\,\widehat{S}_j^{-1})\,A_j^{-1}\,(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\widetilde{v}_j = 0$$

then we obtain from the non singularity of A_j

$$A_j^{-1}\,(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\widetilde{v}_j = \widetilde{v}_j^{*} \neq 0.$$

As we have $(I_j - P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j}\,\widehat{S}_j^{-1})\widetilde{v}_j^{*} = 0$ we obtain

$$\widetilde{v}_j^{*} = P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j}\,\widehat{S}_j^{-1}\,\widetilde{v}_j^{*}.$$

We highlight that $\widehat{S}_{j-1}\,R_{j-1}^{j}\,\widehat{S}_j^{-1}\,\widetilde{v}_j^{*} \neq 0$ follows. Hence we have

$$\begin{aligned}
A_{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j}\,\widehat{S}_j^{-1}\,\widetilde{v}_j^{*} &= R_{j-1}^{j}\,A_j\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j}\,\widehat{S}_j^{-1}\,\widetilde{v}_j^{*} \\
&= R_{j-1}^{j}\,A_j\,\widetilde{v}_j^{*} \\
&= R_{j-1}^{j}\,A_j\,A_j^{-1}\,(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\widetilde{v}_j \\
&= R_{j-1}^{j}\,(I_j - \widehat{S}_j^{-1}\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\widetilde{v}_j \\
&= (R_{j-1}^{j} - R_{j-1}^{j}\,(R_j\,P_j)\,P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\widetilde{v}_j \\
&= (R_{j-1}^{j} - \widehat{S}_{j-1}^{-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\widetilde{v}_j = 0
\end{aligned}$$

This is therefore in contradiction to the non singularity of A_{j-1}.
The additional result follows as we obtain from Lemma 2.3.5 that A_j is non singular in this case. □

We will conclude the section by highlighting that we have $C_{2P,1}^{-1} = C_{2P,2}^{-1}$ if we use the aggregation method and the condition (2.14) holds. The proof for this follows the same argument as the proof of Lemma 6.2.6 which gives the same result for the two generalisations of C_{DT}^{-1}.

7 Multigrid aspects for the modified preconditioners

As we have modified the two grid preconditioners we will also try to do this for the multigrid preconditioners. In the context of two grids the aim was that V_0 is invariant with respect to $A\,X$ (one sided) and $V_{0,X}$ is invariant with respect to A (two sided), respectively. So in this section we will formulate this in the context of $J+1$ grids. Hence we will use J different modifications.

To motivate this we will first consider a version of the one sided modified DT-method and the BPX-method in the context of $J+1$ grids and J modifications, respectively. These modifications will all have all the dimension $n \times n$ so it is obvious that this implies a huge effort. Hence this is not based on practical issues but rather on motivation. Afterwards we will present an idea to define the modifications iteratively by matrices which have a reduced dimension. For the aggregation method we will see that the condition (2.14) plays an important role, again. Furthermore, we will see that for the two sided modified preconditioner this is not as easy as for the one sided modified one.

7.1 Full modifications: A motivation for modifications on $J+1$ grids.

To generalise the results of chapter 4 the aim is to have modifications $\widehat{X}_i$, with

$$\begin{aligned} & A\,\widehat{X}_i\,P_i\,\widetilde{v}_i \in V_i \quad \text{for all} \quad \widetilde{v}_i \in \widetilde{V}_i \\ \Leftrightarrow \qquad & V_i \text{ is invariant with respect to } A\widehat{X}_i. \end{aligned}$$

for all $i = 0, \ldots, J-1$.

For operators $\widehat{X}_i$, $i = 0, \ldots, J-1$ we define modified prolongations $P_{i,\widehat{X}}$ as

$$P_{i,\widehat{X}} := \widehat{X}_i \, P_i \quad \text{for} \quad i = 0, \ldots, J-1$$

$$\text{and} \quad P_{J,\widehat{X}} := I.$$

For $A \in \mathbb{R}^{n \times n}$ we define the one sided modified coarse grid operators $A_{i,\widehat{X}} \in \mathbb{R}^{n_i \times n_i}$ as follows

$$A_{i,\widehat{X}} := R_i \, A \, \widehat{X}_i \, P_i \quad \text{for} \quad i = 0, \ldots, J. \tag{7.1}$$

Lemma: 7.1.1. *Let $A \in \mathbb{R}^{n \times n}$ be non singular, let $A_{j,X} \in \mathbb{R}^{n_j \times n_j}$ be non singular for $j = 0, \ldots, J$ and $\widehat{X}_i \in \mathbb{R}^{n \times n}$, $i = 0, \ldots, J-1$. Then it follows for all $v \in V$*

1. *We have*

$$\Big(A \, \widehat{X}_i \, P_i \, A_{i,X}^{-1} \, R_i \, v, \; A \, \widehat{X}_j \, P_j \, A_{j,X}^{-1} \, R_j \, v \Big)$$
$$= (\widehat{Q}_i \, v, \; v) + \Big((I - \widehat{Q}_i) \, A \, \widehat{X}_i \, P_i \, A_{i,X}^{-1} \, R_i \, v, \; A \, \widehat{X}_j \, P_j \, A_{j,X}^{-1} \, R_j \, v \Big).$$

   for $i, j = 0, \ldots, J, \; i \leq j$.

2. *We have*

$$\Big(A \, \widehat{X}_i \, P_i \, A_{i,X}^{-1} \, (I_i - Q_{i-1}) \, R_i \, v, \; A \, \widehat{X}_j \, P_j \, A_{j,X}^{-1} \, (I_j - Q_{j-1}) \, R_j \, v \Big)$$
$$= \delta_{i,j} \Big(\widehat{S}_i \, (I_i - Q_{i-1}) \, R_i \, v, (I_i - Q_{i-1}) \, R_i \, v \, v \Big)$$
$$+ \Big((I - \widehat{Q_i}) \, A \, \widehat{X}_i \, P_i \, A_{j,X}^{-1} \, R_i \, v, \; A \, \widehat{X}_j \, P_j \, A_{j,X}^{-1} \, R_j \, v \Big).$$

   for all $i, j = 1, \ldots, J$ with $i \leq j$ and

$$\Big(A \, \widehat{X}_0 \, P_0 \, A_{0,X}^{-1} \, R_0 \, v, \; A \, \widehat{X}_j \, P_j \, A_{j,X}^{-1} \, (I_j - Q_{j-1}) \, R_j \, v \Big)$$
$$= \Big((I - \widehat{Q}_0) \, A \, \widehat{X}_0 \, P_0 \, A_{0,X}^{-1} \, R_0 \, v, \; A \, \widehat{X}_j \, P_j \, A_{j,X}^{-1} \, (I_j - Q_{j-1}) \, R_j \, v \Big)$$

   for all $j = 1, \ldots, J$.

3. *If we additionally have $\widehat{S}_i \, P_i^{i-1} \, S_{i-1} = P_i^{i-1} \, \widehat{S}_{i-1}$ for all $i = 1, \ldots, J-1$ then it follows*

$$\Big(A \, \widehat{X}_i \, P_i \, A_{i,X}^{-1} \, (I_i - Q_{i-1}) \, R_i \, v, \; A \, \widehat{X}_j \, P_j \, A_{j,X}^{-1} \, (I_j - Q_{j-1}) \, R_j \, v \Big)$$
$$= \delta_{i,j} ((\widehat{Q}_i - \widehat{Q}_{i-1}) \, v, \; v) + \Big((I - \widehat{Q_i}) \, A \, \widehat{X}_i \, P_i \, A_{j,X}^{-1} \, R_i \, v, \; A \, \widehat{X}_j \, P_j \, A_{j,X}^{-1} \, R_j \, v \Big).$$

for all $i,j = 1,\dots,J$ *with* $i \le j$ *and*

$$\Big(A\,\widehat{X}_0\,P_0\,A_{0,X}^{-1}\,R_0\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,(I_j - Q_{j-1})\,R_j\,v\Big)$$
$$= \Big((I-\widehat{Q}_0)\,A\,\widehat{X}_0\,P_0\,A_{0,X}^{-1}\,R_0\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,(I_j - Q_{j-1})\,R_j\,v\Big)$$

for all $j = 1,\dots,J$.

proof. 1. Based on the definition of $A_{i,X}$ we obtain

$$\begin{aligned}
&(A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,R_j\,v)\\
&\quad= (\widehat{Q}_i\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,R_j\,v)\\
&\qquad+ ((I-\widehat{Q}_i)\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,R_j\,v)\\
&\quad= (\widehat{S}_i\,R_i\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,R_i\,v, R_i^j\,R_j\,A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,R_j\,v)\\
&\qquad+ ((I-\widehat{Q}_i)\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,R_j\,v)\\
&\quad= (\widehat{S}_i\,R_i\,v, R_i^j\,R_j\,v)\\
&\qquad+ ((I-\widehat{Q}_i)\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,R_j\,v)\\
&\quad= (\widehat{Q}_i\,v,\; v) + ((I-\widehat{Q}_i)\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,R_j\,v).
\end{aligned}$$

2. For $i \le j$ we obtain with $\widehat{Q}_j = P_j\,\widehat{S}_j\,R_j$ for all $v \in V$

$$\begin{aligned}
&(A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,(I_i - Q_{i-1})\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,(I_j - Q_{j-1})\,R_j\,v)\\
&= (P_i\,\widehat{S}_i\,R_i\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,(I_i - Q_{i-1})\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,(I_j - Q_{j-1})\,R_j\,v)\\
&\quad+ ((I-\widehat{Q}_i)\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,(I_i - Q_{i-1})\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,(I_j - Q_{j-1})\,R_j\,v)\\
&= (\widehat{S}_i\,(I_i - Q_{i-1})\,R_i\,v,\; R_i^j\,(I_j - Q_{j-1})\,R_j\,v)\\
&\quad+ ((I-\widehat{Q}_i)\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,(I_i - Q_{i-1})\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,(I_j - Q_{j-1})\,R_j\,v)\\
&= \delta_{i,j}\,(\widehat{S}_i\,(I_i - Q_{i-1})\,R_i\,v,\; (I_i - Q_{i-1})\,R_i\,v)\\
&\quad+ ((I-\widehat{Q}_i)\,A\,\widehat{X}_i\,P_i\,A_{i,X}^{-1}\,(I_i - Q_{i-1})\,R_i\,v,\; A\,\widehat{X}_j\,P_j\,A_{j,X}^{-1}\,(I_j - Q_{j-1})\,R_j\,v)
\end{aligned}$$

The assertion for $i = 0$ follows the same arguments.

3. Based on the assumption $\widehat{S}_i\, P_i^{i-1}\, S_{i-1} = P_i^{i-1}\, \widehat{S}_{i-1}$ for all $i = 1, \ldots, J-1$ we obtain from Corollary 2.3.10

$$\|(\widehat{Q}_i - \widehat{Q}_{i-1})v\|^2 = (\widehat{S}_i\,(I_i - Q_{i-1})\,R_i\,v,\,(I_i - Q_{i-1})\,R_i\,v).$$

Hence the equation follows from the second assertion of this Lemma. The assertion for $i = 0$ follows the same arguments.

□

Considering the results of Lemma 7.1.1 we highlight that if V_i is invariant with respect to $A\,\widehat{X}_i$ then it follows $(I - \widehat{Q}_i)\,A\,\widehat{X}_i\,P_i\,\widetilde{v}_i = 0$ for all $\widetilde{v}_i \in \widetilde{V}_i$. We will use this assumption later to motivate the modifications in the multigrid setting.

Based on the matrices $\widehat{X}_i$, $i = 0, \ldots, J$ we can define one sided modified DT-method and BPX-method on $J+1$ grids. Let $B^{(j)}$, $j = 0, \ldots J$ be non singular matrices, then we define $C^{-1}_{DT,\widehat{X}}(B^{(0)}, \ldots, B^{(J)}), C^{-1}_{BPX,\widehat{X}}(B^{(0)}, \ldots, B^{(J)})$ as follows

$$C^{-1}_{DT,\widehat{X}}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=1}^{J} P_{j,\widehat{X}}\,(B^{(j)})^{-1}\,(I_j - Q_{j-1})\,R_j + P_{0,\widehat{X}}\,(B^{(0)})^{-1}\,R_0.$$

$$C^{-1}_{BPX,\widehat{X}}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=1}^{J} P_{j,\widehat{X}}\,(B^{(j)})^{-1}\,R_j + P_{0,\widehat{X}}\,(B^{(0)})^{-1}\,R_0.$$

Therewith we get the following proposition for the existence of both preconditioners:

Lemma: 7.1.2. *Let $A \in \mathbb{R}^{n\times n}$, $B^{(j)} \in \mathbb{R}^{n_j\times n_j}$, $j = 0, \ldots J$ be non singular and $\widehat{X}_j \in \mathbb{R}^{n\times n}$ modifications with $rk(\widehat{X}_j\,P_j) = n_j$. If there is a matrix $\widetilde{B} \in \mathbb{R}^{n\times n}$ with*

$$(R_j\,\widetilde{B}\,P_{j,\widehat{X}}\,(B^{(j)})^{-1})(\widetilde{v}_j) = \tau^j_{(\widetilde{v}_j)}\widetilde{v}_j, \quad \tau^j_{(\widetilde{v}_j)} \in \mathbb{R},\, \tau^j_{(\widetilde{v}_j)} > 0 \tag{7.2}$$

for all $j = 1, \ldots, J$ then $C^{-1}_{BPX,\widehat{X}}(B^{(0)}, \ldots, B^{(J)})$ is non singular.
If there is a matrix $\widetilde{B} \in \mathbb{R}^{n\times n}$ with

$$\widetilde{W_j} \quad \text{is invariant with respect to} \quad (R_j\,\widetilde{B}\,P_{j,\widehat{X}}\,(B^{(j)})^{-1}) \tag{7.3}$$

and $rk(R_j\,\widetilde{B}\,P_{j,\widehat{X}}\,(B^{(j)})^{-1}) = n_j$ for $j = 0, \ldots, J$. Then $C^{-1}_{DT,\widehat{X}}(B^{(0)}, \ldots, B^{(J)})$ is non singular.

proof. The proof follows by the same arguments as used in the proofs of the Lemmata 6.1.1 and 6.2.1 respectively. □

We highlight that we can again define a second version of the modified preconditioner $C^{-1}_{DT,\widehat{X}}$ if we set

$$C^{-1}_{DT,\widehat{X},2}(B^{(0)},\ldots,B^{(J)}) := \sum_{j=1}^{J} P_{j,\widehat{X}}\,(B^{(j)})^{-1}\,(I_j - \widehat{S}_j^{-1}P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^{j})\,R_j + P_{0,\widehat{X}}\,(B^{(0)})^{-1}\,R_0.$$

We obtain the analogue result concerning the non singularity of $C^{-1}_{DT,\widehat{X},2}(B^{(0)},\ldots,B^{(J)})$ as given in Lemma 6.2.5 for the unmodified version. Furthermore we want to highlight that as in the unmodified situation the assumptions on the inivariance are fulfilled if we consider for $j=0,\ldots,J$ the situation

$$A_{j,\widehat{X}} = B^{(j)}.$$

Similarly to the unmodified situation we can give representations for $A\,C^{-1}_{BPX,\widehat{X}}$ and $A\,C^{-1}_{DT,\widehat{X}}$, respectively. We will see that based on the strong assumption that we have matrices $\widehat{X}_j$, $j=1,\ldots,J$ which fulfil that V_j is invariant with respect to $A\,\widehat{X}_j$ we obtain a meaningful result for both preconditioners.

For the BPX-method we obtain from the assumption that V_i is invariant with respect to $A\,\widehat{X}_i$ for an arbitrary $v\in V$

$$\begin{aligned} A\,C^{-1}_{BPX,\widehat{X}}\,v &= \sum_{j=0}^{J} A\,\widehat{X}_j\,P_j\,(B^{(j)})^{-1}\,R_j\,v = \sum_{j=0}^{J}\widehat{Q}_j\,A\,\widehat{X}_j\,P_j\,(B^{(j)})^{-1}\,R_j\,v \\ &= \sum_{j=0}^{J} P_j\,\widehat{S}_j\,A_{j,\widehat{X}}\,(B^{(j)})^{-1}\,R_j\,v. \end{aligned}$$

In the case of $A_{j,\widehat{X}} = B^{(j)}$ it follows

$$A\,C^{-1}_{BPX,\widehat{X}}\,v = \sum_{j=0}^{J}\widehat{Q}_j\,v.$$

Based on the same characteristic for $\widehat{X}_j$ it follows for the DT-method

$$\begin{aligned} A\, C^{-1}_{DT,\widehat{X}}\, v &= \sum_{j=1}^{J} A\, \widehat{X}_j\, P_j\, (B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j\, v + A\, \widehat{X}_0\, P_0\, (B^{(0)})^{-1}\, R_0\, v \\ &= \sum_{j=1}^{J} \widehat{Q}_j\, A\, \widehat{X}_j\, P_j\, (B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j\, v + \widehat{Q}_0\, A\, \widehat{X}_0\, P_0\, (B^{(0)})^{-1}\, R_0\, v \\ &= \sum_{j=1}^{J} P_j\, \widehat{S}_j\, A_{j,\widehat{X}}\, (B^{(j)})^{-1}\, (I_j - Q_{j-1})\, R_j\, v + P_0\, \widehat{S}_0\, A_{0,\widehat{X}}\, (B^{(0)})^{-1}\, R_0\, v. \end{aligned}$$

Again in the case of $A_{j,\widehat{X}} = B^{(j)}$ it follows

$$A\, C^{-1}_{DT,\widehat{X}}\, v = \sum_{j=1}^{J} P_j\, \widehat{S}_j\, (I_j - Q_{j-1})\, R_j\, v + P_0\, \widehat{S}_0\ R_0\, v$$

If additionally $\widehat{S}_j\, P_j^{j-1}\, S_{j-1} = P_j^{j-1}\, \widehat{S}_{j-1}$ holds then it follows based on Corollary 2.3.9

$$A\, C^{-1}_{DT,\widehat{X}}\, v = \sum_{j=1}^{J} (\widehat{Q}_j - \widehat{Q}_{j-1})\, v + \widehat{Q}_0\, v = v.$$

Of course we have again that for the aggregation method the condition (2.14) is equivalent to the characteristic that the equation $\widehat{S}_j\, P_{j-1}^{j}\, S_{j-1} = P_{j-1}^{j}\, \widehat{S}_{j-1}$ holds. And if the equation does not hold it is obvious that we can take the operator $C^{-1}_{DT,\widehat{X},2}$.

The calculations above motivate the aim that V_j is invariant with respect to $A\, \widehat{X}_j$ for the multigrid situation. We can also motivate this based on the results of Lemma 7.1.1.

For an arbitrary $v \in V$ this implies for $B^{(k)} = A_{k,\widehat{X}}$

$$
\begin{aligned}
&\|A\, C^{-1}_{DT,\widehat{X}}\, v\|^2 \\
&= \sum_{j=1}^{n} (A\, P_{j,X}\, A^{-1}_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v,\ A\, P_{j,X}\, A^{-1}_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v) \\
&\quad + (A\, P_{0,X}\, A^{-1}_{0,X}\, R_0\, v,\ A\, P_{0,X}\, A^{-1}_{0,X}\, R_0\, v) \\
&\quad + 2 \sum_{j=1}^{n} (A\, P_{j,X}\, A_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v,\ A\, P_{0,X}\, A^{-1}_{0,X}\, R_0\, v) \\
&\quad 2 \sum_{j=1}^{n} \sum_{i=j+1}^{J} (A\, P_{j,X}\, A^{-1}_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v,\ A\, P_{i,X}\, A^{-1}_{i,\widehat{X}} (I_i - Q_{i-1}) R_i\, v) \\
&= \sum_{j=1}^{n} (\widehat{Q}_j\, A\, P_{j,X}\, A^{-1}_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v,\ A\, P_{j,X}\, A^{-1}_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v) \\
&\quad + \sum_{j=1}^{n} ((I - \widehat{Q}_j)\, A\, P_{j,X}\, A^{-1}_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v,\ A\, P_{j,X}\, A^{-1}_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v) \\
&\quad + (\widehat{Q}_0\, A\, P_{0,X}\, A^{-1}_{0,X}\, R_0\, v,\ A\, P_{0,X}\, A^{-1}_{0,X}\, R_0\, v) \\
&\quad + ((I - \widehat{Q}_0)\, A\, P_{0,X}\, A^{-1}_{0,X}\, R_0\, v,\ A\, P_{0,X}\, A^{-1}_{0,X}\, R_0\, v) \\
&\quad + 2 \sum_{j=1}^{n} (\widehat{Q}_0\, A\, P_{j,X}\, A_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v,\ A\, P_{0,X}\, A^{-1}_{0,X}\, R_0\, v) \\
&\quad + 2 \sum_{j=1}^{n} ((I - \widehat{Q}_0)\, A\, P_{j,X}\, A_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v,\ A\, P_{0,X}\, A^{-1}_{0,X}\, R_0\, v) \\
&\quad + 2 \sum_{j=1}^{n} \sum_{i=j+1}^{J} (\widehat{Q}_j\, A\, P_{j,X}\, A^{-1}_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v,\ A\, P_{i,X}\, A^{-1}_{i,\widehat{X}} (I_i - Q_{i-1}) R_i\, v) \\
&\quad + 2 \sum_{j=1}^{n} \sum_{i=j+1}^{J} ((I - \widehat{Q}_j)\, A\, P_{j,X}\, A^{-1}_{j,\widehat{X}} (I_j - Q_{j-1}) R_j\, v,\ A\, P_{i,X}\, A^{-1}_{i,\widehat{X}} (I_i - Q_{i-1}) R_i\, v)
\end{aligned}
$$

If we assume that $\widehat{S}_j\, P^j_{j-1}\, S_{j-1} = P^j_{j-1}\, \widehat{S}_{j-1}$ holds then it follows from the calculation

above and the results of Lemma 7.1.1

$$
\begin{aligned}
&\|A\,C^{-1}_{DT,\widehat{X}}\,v\|^2 \\
&= \sum_{j=1}^{n}((\widehat{Q}_j - \widehat{Q}_{j-1})\,v,\,v) + (\widehat{Q}_0\,v,\,v) \\
&\quad + \sum_{i,j=1}^{n}((I-\widehat{Q}_j)\,A\,P_{j,X}\,A^{-1}_{j,\widehat{X}}(I_j - Q_{j-1})R_j\,v,\; A\,P_{j,X}\,A^{-1}_{j,\widehat{X}}(I_j - Q_{j-1})R_j\,v) \\
&\quad + ((I-\widehat{Q}_0)\,A\,P_{0,X}\,A^{-1}_{0,X}\,R_0\,v,\; A\,P_{0,X}\,A^{-1}_{0,X}\,R_0\,v) \\
&\quad + 2\sum_{j=1}^{n}((I-\widehat{Q}_0)\,A\,P_{j,X}\,A_{j,\widehat{X}}(I_j - Q_{j-1})R_j\,v,\; A\,P_{0,X}\,A^{-1}_{0,X}\,R_0\,v) \\
&\quad 2\sum_{j=1}^{n}\sum_{i=j+1}^{J}((I-\widehat{Q}_j)\,A\,P_{j,X}\,A^{-1}_{j,\widehat{X}}(I_j - Q_{j-1})R_j\,v,\; A\,P_{i,X}\,A^{-1}_{i,\widehat{X}}(I_i - Q_{i-1})R_i\,v)
\end{aligned}
$$

The last equation in the calculation above follows from Lemma 7.1.1. If we have additionally that V_j is invariant with respect to $A\widehat{X}_j$ for $j = 0, \ldots, J-1$ then it follows from the calculation above

$$\|A\,C^{-1}_{DT,\widehat{X}}\,v\|^2 = \|v\|^2.$$

All the term

$$
\begin{aligned}
&((I-\widehat{Q}_j)\,A\,P_{j,X}\,A^{-1}_{j,\widehat{X}}(I_j - Q_{j-1})R_j\,v,\; A\,P_{i,X}\,A^{-1}_{i,\widehat{X}}(I_i - Q_{i-1})R_i\,v) \quad i,j = 1,\ldots,J \\
&((I-\widehat{Q}_0)\,A\,P_{0,X}\,A^{-1}_{0,X}\,R_0\,v,\; A\,P_{0,X}\,A^{-1}_{0,X}\,R_0\,v) \\
&((I-\widehat{Q}_0)\,A\,P_{j,X}\,A_{j,\widehat{X}}(I_j - Q_{j-1})R_j\,v,\; A\,P_{0,X}\,A^{-1}_{0,X}\,R_0\,v) \quad j = 1,\ldots,J
\end{aligned}
$$

represents a kind of bias that vanishes if the condition of the invariance holds.

Although the results above are meaningful, it is not a good idea to determine matrices $\widehat{X}_j \in \mathbb{R}^{n_j \times n_j}$, $j = 0, \ldots, J$ with the assumed characteristics. As they all have the dimension $(n \times n)$ there is a huge effort to determine them, and also to apply them. We should highlight that in an algorithm we have to use them twice. First for the calculation $\widehat{X}_j\,P_j\big((B^{(-1)})\,R_j\,v\big)$. The second time is for the approximation of $A_{j,\widehat{X}}$.

So we use the basic idea of multigrid algorithms to determine modifications. This means that we will construct modifications $X_j \in \mathbb{R}^{n_j \times n_j}$ for which we use only the operator $A_{j,X}$ and the space $\widetilde{V}_{j-1}$ and $Im(P_j^{j-1}(\widetilde{V}_{j-1}))$, respectively. Hence they can be constructed iteratively.

7.2 Modification of reduced systems

As motivated at the end of the last section we will consider modifications $X_j \in \mathbb{R}^{n_j \times n_j}$ for $j = 1, \ldots, J$. Similarly to the two grid situation we require for $P_j^{j-1} \in \mathbb{R}^{n_j \times n_{j-1}}$ that $rk(X_j\, P_j^{j-1}) = n_{j-1}$. Then we define

$$
\begin{aligned}
(7.4) \qquad & P_{j,X}^{j-1} := X_j\, P_j^{j-1} \quad \text{for} \quad j = 1, \ldots, J \\
& P_{j,X}^{i} := P_{j,X}^{j-1} \circ \cdots \circ P_{i+1,X}^{i} \quad \text{for} \quad i, j = 1, \ldots, J,\ j \geq i \\
& P_{j,X} := P_{J,X}^{j} \quad \text{for } j = 0, \ldots, J-1 \text{ and} \quad P_{J,X} := I.
\end{aligned}
$$

For a two sided modification we define the operators

$$
\begin{aligned}
(7.5) \qquad & R_{j,X}^{i} = (P_{i,X}^{j})^T \quad \text{for} \quad i, j = 1, \ldots, J,\ j \leq i \\
& R_{j,X} = (P_{j,X})^T \quad \text{for} \quad j = 0, \ldots, J \\
& S_{j,X} = (P_{j+1,X}^{j}\, R_{j,X}^{j+1})^{-1}, \quad \widehat{S}_{j,X} = (P_{j,X}\, R_{j,X})^{-1} \quad \text{for} \quad j = 0, \ldots, J-1 \\
& Q_{j,X} = P_{j+1,X}^{j}\, S_{j,X}\, R_{j,X}^{j+1}, \quad \widehat{Q}_j = P_{j,X}\, \widehat{S}_{j,X}\, R_{j,X} \quad \text{for} \quad j = 0, \ldots, J-1.
\end{aligned}
$$

Then we define the spaces

$$
\begin{aligned}
(7.6) \qquad & V_{j,X} := P_{j,X}(\widetilde{V}_j) \equiv P_{j,X}(\mathbb{R}^{n_j}) \quad \text{for} \quad j = 0, \ldots, J \\
& W_{j,X} := (I_j - Q_{j-1,X})(\widetilde{V}_j) \quad \text{for} \quad j = 1, \ldots, J.
\end{aligned}
$$

Furthermore, we set iteratively

$$
(7.7) \qquad A_{j,X} := R_j^{j+1}\, A_{j+1,X}\, P_{j+1,X}^{j} \quad \text{and} \quad A_{j,XX} := R_{j,X}^{j+1}\, A_{j+1,XX}\, P_{j+1,X}^{j}
$$

for $j = 0, \ldots, J-1$ and $A_{J,X} = A = A_{J,XX}$. Based on the definition of the operators it follows immediately

$$
(7.8) \qquad A_{j,X} := R_j\, A\, P_{j,X} \quad \text{and} \quad A_{j,XX} := R_{j,X}\, A\, P_{j,X}
$$

for $j = 0, \ldots, J-1$. Based on these definitions we define for non singular $B^{(j)} \in \mathbb{R}^{n_j \times n_j}, j = 0, \ldots, J$ the one or two sided precondition operators as follows

(7.9)

$$C^{-1}_{BPX,X}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=0}^{J} P_{j,X} \, (B^{(j)})^{-1} \, R_j$$

$$C^{-1}_{DT,1,X}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=1}^{J} P_{j,X} \, (B^{(j)})^{-1} \, (I_j - Q_{j-1}) \, R_j + P_{0,X} \, (B^{(0)})^{-1} \, R_0$$

$$C^{-1}_{DT,2,X}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=1}^{J} P_{j,X} \, (B^{(j)})^{-1} \, (I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^{j}) \, R_j$$

$$+ P_{0,X} \, (B^{(0)})^{-1} \, R_0$$

$$C^{-1}_{BPX,XX}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=0}^{J} P_{j,X} \, (B^{(j)})^{-1} \, R_{j,X}$$

$$C^{-1}_{DT,1,XX}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=1}^{J} P_{j,X} \, (B^{(j)})^{-1} \, (I_j - Q_{j-1,X}) \, R_{j,X} + P_{0,X} \, (B^{(0)})^{-1} \, R_{0,X}$$

$$C^{-1}_{DT,2,XX}(B^{(0)}, \ldots, B^{(J)}) := \sum_{j=1}^{J} P_{j,X} \, (B^{(j)})^{-1} \, (I_j - \widehat{S}_{j,X}^{-1} \, P_{j,X}^{j-1} \, \widehat{S}_{j-1,X} \, R_{j-1,X}^{j}) \, R_{j,X}$$

$$+ P_{0,X} \, (B^{(0)})^{-1} \, R_{0,X}$$

7.2.1 One sided modification

In this section we will first briefly consider a short assertion concerning the non singularity of the one sided modified precondition operators. Afterwards we will take a look at the properties we get from the iterative modifiation as given from the iterative definition of $P_{j,X}$. The main result will be that if we use the aggregation method, the condition (2.14) is fulfilled and we have that $Im(P_j^{j-1}(\widetilde{V}_{j-1}))$ is invariant with respect to $A_{j,X} \, X_j$ then we obtain

$$A \, P_{j,X} = \widehat{Q}_j A \, P_{j,X}.$$

For the two grid situation this is equivalent to the characteristic that V_0 is invariant with respect to $A \, X$. Also for the multigrid situation we will see that this has the same characteristic.

Lemma: 7.2.1. *Let $A \in \mathbb{R}^{n\times n}$, $B^{(j)} \in \mathbb{R}^{n_j\times n_j}$, $j = 0,\dots J$ be non singular and $X_j \in \mathbb{R}^{n_j\times n_j}$, $j = 1,\dots,J$ modifications that hold $rk(X_j\, P_{j-1}^j) = n_j$. If there is a matrix $\widetilde{B} \in \mathbb{R}^{n\times n}$ with*

$$(R_j\, \widetilde{B}\, P_{j,X}\, (B^{(j)})^{-1})(\widetilde{v}_j) = \tau^j_{(\widetilde{v}_j)}\widetilde{v}_j, \quad \tau^j_{(\widetilde{v}_j)} > 0 \tag{7.10}$$

for all $j = 1,\dots,J$ then $C^{-1}_{BPX,X}(B^{(0)},\dots,B^{(J)})$ is non singular.
If there is a matrix $\widetilde{B} \in \mathbb{R}^{n\times n}$ with

$$\widetilde{W_j} \quad \textit{is invariant with respect to} \quad (R_j\, \widetilde{B}\, P_{j,X}\, (B^{(j)})^{-1}) \tag{7.11}$$

$$\Big(Im(((I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j)R_j)(V)) \quad \textit{is invariant with respect to} \tag{7.12}$$

$$(R_j\, \widetilde{B}\, P_{j,X}\, (B^{(j)})^{-1})\Big)$$

and $rk(R_j\, \widetilde{B}\, P_{j,X}\, (B^{(j)})^{-1}) = n_j$ for $j = 0,\dots,J$. Then $C^{-1}_{DT,1,X}(B^{(0)},\dots,B^{(J)})$ $\Big(C^{-1}_{DT,2,X}(B^{(0)},\dots,B^{(J)})\Big)$ is non singular.

proof. The proof follows the same arguments as used in the proofs of Lemmata 6.1.1, 6.2.1 and 6.2.5. □

Moreover, we highlight that like in the unmodified situation the conditions (7.10), (7.11) and (7.12) are fulfilled if we set $B^{(j)} = A_{j,X}$. Now we consider the main aspect of this section. Based on the characteristics for the two grid situation we obtain in the following characteristic for the modifications on the reduced systems:

Lemma: 7.2.2. *Let $A \in \mathbb{R}^{n\times n}$ be non singular. Let $P_{j,X}$ be as defined in (7.4). If we assume that $Im(P_{j+1}^j(\widetilde{V}_j))$ is invariant with respect to $A_{j+1,X}\, X_{j+1}$ for $j = 0,\dots,J-1$ then we have*

$$A\, P_{j,X}\, \widetilde{v}_j \in V_j \quad \textit{for all} \quad \widetilde{v}_j \in \widetilde{V}_j \tag{7.13}$$

if and only if we have

$$P_J^{J-1}\, S_{J-1}\, P_{J-1}^{J-2}\, S_{J-2}\, \dots, P_{j+1}^j\, \widetilde{v}_j \in V_j \quad \textit{for all} \quad \widetilde{v}_j \in \widetilde{V}_j. \tag{7.14}$$

proof. Based on the invariance of $Im(P_{j+1}^j(\widetilde{V}_j))$ with respect to $A_{j+1,X}\, X_{j+1}$ we obtain

$$A_{j+1,X}\, X_{j+1}\, P_{j+1}^j = Q_j\, A_{j+1,X}\, X_{j+1}\, P_{j+1}^j = P_{j+1}^j\, S_j\, R_j^{j+1}\, A_{j+1,X}\, P_{j+1}^j = P_{j+1}^j\, S_j\, A_{j,X}$$

for $j = 1, \ldots, J$. If we use this characteristic iteratively we obtain

$$\begin{aligned} A\,P_{j,X} &= A\,X_J\,P_J^{J-1}\,P_{J-1,X}^{j} \\ &= Q_{J-1}\,A\,X_J\,P_J^{J-1}\,P_{J-1,X}^{j} \\ &= P_J^{J-1}\,S_{J-1}\,R_{J-1}^{J}\,A_{J,X}\,X_J\,P_J^{J-1}\,P_{J-1,X}^{j} \\ &= P_J^{J-1}\,S_{J-1}\,A_{J-1,X}\,P_{J-1,X}^{j} \\ &= P_J^{J-1}\,S_{J-1}\,Q_{J-1}\,A_{J-1,X}\,P_{J-1,X}^{j} \\ &= P_J^{J-1}\,S_{J-1}\,P_{J-1}^{J-2}\,S_{J-2}\,A_{J-2,X}\,P_{J-2,X}^{j} \\ &\vdots \\ &= P_J^{J-1}\,S_{J-1}\,P_{J-1}^{J-2}\,S_{J-2}\,\ldots,P_{j+1}^{j}\,S_j\,A_{j,X}. \end{aligned}$$

This proves the proposition. □

It is obvious that in the multigrid situation the equation (7.13) is the generalisation of the characteristic that V_0 is invariant with respect to $A\,X$ in the two grid situation. Based on Lemma 7.2.2 we obtain that if the invariance is given for the two grid situations between the grids j and $j-1$ for $j = 1, \ldots, J$ then the equation (7.13) is equivalent to the condition (7.14). This condition does not depends on the operators $A_{j,X}$ but only on the structure of the aggregation and the spaces V_j, respectively. Hence for the aggregation method we can sum this up as follows:

Corollary: 7.2.3. *Let $A \in \mathbb{R}^{n\times n}$ be non singular and $P_{j,X}$ be as defined in (7.4). Assume that the aggregation method is used to construct the coarser grids and the condition (2.14) holds. If $Im(P_k^{k-1}(\widetilde{V}_{k-1}))$ is invariant with respect to $A_{k,X}\,X_k$ for $k = j+1, \ldots, J$ then we have*

$$A\,P_{j,X}\,\widetilde{v}_j \in V_j \quad \textit{for all} \quad \widetilde{v}_j \in \widetilde{V}_j.$$

proof. Based on Lemma 2.4.8 we have that for the aggregation method the condition (2.14) is equivalent to the equation $\widehat{S}_k\,P_k^{k-1}\,S_{k-1} = P_k^{k-1}\,\widehat{S}_{k-1}$ for $k = j+1, \ldots, J$. Based on Corollary 2.3.9 we obtain that this implies

$$P_J^{J-1}\,S_{J-1}\,P_{J-1}^{J-2}\,S_{J-2}\,\ldots,P_{j+1}^{j}\,S_j = \widehat{Q}_j.$$

Hence we have

$$P_J^{J-1}\,S_{J-1}\,P_{J-1}^{J-2}\,S_{J-2}\,\ldots,P_{j+1}^{j}\,\widetilde{v}_j \in V_j \quad \text{for all} \quad \widetilde{v}_j \in V_j$$

and the assertion follows from Lemma 7.2.2. □

Based on the results above and the calculations in chapter 6 it is obvious that if the assumptions of Lemma 7.2.1 and Corollary 7.2.3, respectively are fulfilled then we obtain

$$A\,P_{j,X} = \widehat{Q}_j\,A\,P_{j,X}$$

for $j = 0, \ldots, J-1$. From this result it follows

$$\begin{aligned} A\,C^{-1}_{DT,1,X}(A) &= \sum_{j=1}^{J} A\,P_{j,X}\,A^{-1}_{j,X}\,(I_j - Q_{j-1})\,R_j + A\,P_{0,X}\,A^{-1}_{0,X}\,R_0 \\ &= \sum_{j=1}^{J} \widehat{Q}_j\,A\,P_{j,X}\,A^{-1}_{j,X}\,(I_j - Q_{j-1})\,R_j + \widehat{Q}_0 A\,P_{0,X}\,A^{-1}_{0,X}\,R_0 \\ &= \sum_{j=1}^{J} (\widehat{Q}_j - \widehat{Q}_{j-1}) + \widehat{Q}_0 = I. \end{aligned}$$

Based on the same assumptions we obtain for the BPX-method

$$A\,C^{-1}_{BPX,X}(A) = \sum_{j=0}^{J} A\,P_{j,X}\,A^{-1}_{j,X}\,R_j = \sum_{j=0}^{J} \widehat{Q}_j\,A\,P_{j,X}\,A^{-1}_{j,X}\,R_j = \sum_{j=0}^{J} \widehat{Q}_j.$$

7.2.2 Two sided modification

As done for the one sided modification in the multigrid setting at the beginning of this section we will present a sufficient condition for the non singularity of the two sided modifications. Afterwards we will consider the characteristics of the modifications on the reduced systems. We will see that if we can fulfil the condition that $Im(P^{j-1}_{j,X}(\widetilde{V}_{j-1}))$ is invariant with respect to $A_{j,XX}$ then there are no additional problems for the multigrid situation. As already discussed in section 5.3 this assumption is only theoretically interesting. If we consider the block matrices as done in section 5.3.2 we will see that again the condition (2.14) plays a role in the multigrid situation.

Lemma: 7.2.4. *Let $A \in \mathbb{R}^{n\times n}$, $B^{(j)} \in \mathbb{R}^{n_j\times n_j}$, $j = 0, \ldots J$ be non singular and $X_j \in \mathbb{R}^{n_j\times n_j}$, $j = 1, \ldots, J$ modifications that hold $rk(X_j\,P_j^{j-1}) = n_j$. If there is a matrix $\widetilde{B} \in \mathbb{R}^{n\times n}$ with*

$$(R_{j,X}\,\widetilde{B}\,P_{j,X}\,(B^{(j)})^{-1})(\widetilde{v}_j) = \tau^j_{(\widetilde{v}_j)}\widetilde{v}_j, \quad \tau^j_{(\widetilde{v}_j)} > 0 \tag{7.15}$$

for all $j = 1, \ldots, J$ *then* $C^{-1}_{BPX,XX}(B^{(0)}, \ldots, B^{(J)})$ *is non singular.*
If there is a matrix $\widetilde{B} \in \mathbb{R}^{n \times n}$ *with*

(7.16) $\quad \widetilde{W}_{j,X}$ *is invariant with respect to* $(R_{j,X}\, \widetilde{B}\, P_{j,X}\, (B^{(j)})^{-1})$

(7.17) $\quad \Big(Im(((I_j - \widehat{S}^{-1}_{j,X}\, P^{j-1}_{j,X}\, \widehat{S}_{j-1,X}\, R^j_{j-1,X}) R_{j,X})(V))$ *is invariant with respect to*
$(R_{j,X}\, \widetilde{B}\, P_{j,X}\, (B^{(j)})^{-1}) \Big)$

and $rk(R_{j,X}\, \widetilde{B}\, P_{j,X}\, (B^{(j)})^{-1}) = n_j$ *for* $j = 0, \ldots, J$. *Then* $C^{-1}_{DT,1,XX}(B^{(0)}, \ldots, B^{(J)})$ $\Big(C^{-1}_{DT,2,XX}(B^{(0)}, \ldots, B^{(J)}) \Big)$ *is non singular.*

proof. As for the one sided modification the proof follows by the same arguments as used in the proofs of Lemmata 6.1.1, 6.2.1 and 6.2.5, respectively. □

Furthermore, we highlight as in the unmodified and one sided modified situation, respectively that the conditions (7.15), (7.16) and (7.17) are fulfilled if we set $B^{(j)} = A_{j,XX}$.

Now we consider the main aspect of this section. For the one sided modification the aim is to use only a two grid situation to determine one modification. This is given as

$$\begin{aligned} & A_{j,XX}\, P^{j-1}_{j,X}\, \widetilde{v}_{j-1} \in Im(P^{j-1}_{j,X}(\widetilde{V}_{j-1})) \quad \text{for all} \quad \widetilde{v}_{j-1} \in \widetilde{V}_{j-1} \\ \Leftrightarrow \quad & Im(P^{j-1}_{j,X}(\widetilde{V}_{j-1})) \text{ is invariant with respect to } A_{j,XX} \end{aligned}$$

for $j = 1, \ldots, J$. It is obvious that for $J = 1$ this condition is equivalent to the characteristic that $V_{0,X}$ is invariant with respect to A. Hence we can give a first result that is the generalisation of Proposition 4.2.5.

Proposition: 7.2.5. *Let* $A_{j,XX} \in \mathbb{R}^{n_j \times n_j}$ *be s.p.d. Then* $Im(P^{j-1}_{j,X}(\widetilde{V}_{j-1}))$ *is invariant with respect to* $A_{j,XX}$ *if and only if there are* $z_1, \ldots, z_{n_{j-1}}$ *with*

$$Im(P^{j-1}_{j,X}(\widetilde{V}_{j-1})) = \left\langle z_1, \ldots, z_{n_{j-1}} \right\rangle$$

and $A_{j,XX}\, z_i = \lambda_i\, z_i$ *for* $i = 1, \ldots, n_{j-1}$.

proof. The proof follows the same arguments as the proof of Proposition 4.2.5. □

Based on the Propositon 7.2.5 we have a characteristic for the modifications X_j that imply for all two gird situations $(\widetilde{V}_{j-1}, Im(P^{j-1}_{j,X}(\widetilde{V}_{j-1}))$ for $j = 1, \ldots, J)$ the same result

as we have in the situation given by only two grids. Hence we consider the result we obtain for $A\,P_{k,X}$. Based on the charactersitics mentioned so far the aim is

$$A\,P_{j,X}\,\widetilde{v}_j \in V_{j,X} \quad \text{for all } \widetilde{v}_j \in \widetilde{V}_j \tag{7.18}$$

for all $j = 0, \ldots, J-1$.
Similarly to Lemma 7.2.2 for the one sided modification we obtain the following result.

Lemma: 7.2.6. *Let $A \in \mathbb{R}^{n\times n}$ be non singular. Let $P_{j,X}$ be as defined in (7.4). If we assume that $Im(P^j_{j+1,X}(\widetilde{V}_j))$ is invariant with respect to $A_{j+1,XX}$ for $j = 0, \ldots, J-1$ then we have*

$$A\,P_{j,X}\,\widetilde{v}_j \in V_{j,X} \quad \textit{for all} \quad \widetilde{v}_j \in \widetilde{V}_j \tag{7.19}$$

if and only if we have

$$P^{J-1}_{J,X}\,S_{J-1,X}\,P^{J-2}_{J-1,X}\,S_{J-2,X}\,\ldots, P^j_{j+1,X}\,\widetilde{v}_j \in V_{j,X} \quad \textit{for all} \quad \widetilde{v}_j \in \widetilde{V}_j. \tag{7.20}$$

proof. Based on the invariance of $Im(P^j_{j+1,X}(\widetilde{V}_j))$ with respect to $A_{j+1,XX}$ we obtain

$$\begin{aligned} A_{j+1,XX}\,P^j_{j+1,X} &= Q_{j,X}\,A_{j+1,XX}\,X_{j+1}\,P^j_{j+1} = P^j_{j+1,X}\,S_{j,X}\,R^{j+1}_{j,X}\,A_{j+1,XX}\,P^j_{j+1,X} \\ &= P^j_{j+1}\,S_{j,X}\,A_{j,X}. \end{aligned}$$

If we use this characteristic iteratively we obtain

$$\begin{aligned} A\,P_{j,X} &= A\,X_J\,P^{J-1}_J\,P^j_{J-1,X} \\ &= P^{J-1}_{J,X}\,S_{J-1,X}\,A_{J-1,XX}\,P^j_{J-1,X} \\ &= P^{J-1}_{J,X}\,S_{J-1,X}\,Q_{J-1,X}\,A_{J-1,XX}\,P^j_{J-1,X} \\ &= P^{J-1}_{J,X}\,S_{J-1,X}\,P^{J-2}_{J-1,X}\,S_{J-2,X}\,A_{J-2,XX}\,P^j_{J-2,X} \\ &\vdots \\ &= P^{J-1}_{J,X}\,S_{J-1,X}\,P^{J-2}_{J-1,X}\,S_{J-2,X}\,\ldots, P^j_{j+1,X}\,S_{j,X}\,A_{j,XX}. \end{aligned}$$

This proves the proposition. □

Based on Lemma 7.2.6 we need to fulfil the condition (7.20). Based on the Propositon 7.2.5 it is obvious that to fulfil $Im(P^{j-1}_{j,X}(\widetilde{V}_{j-1}))$ invariant with respect to $A_{j,XX}$ it is necessary that $V_{j-1,X}$ is given by n_{j-1} eigenvectors of $A_{j,XX}$. Since we have done it for

the one sided modification we will concentrate on the aggregation method. Theirfore we can given the following sufficient condition for the invariance given in (7.20).

As a generalisation of the setting used in section 4.2 for the two grid situation we set

$$I_1^j := \big\{ i \in \{1,\dots,n\} \,:\, \mathcal{N}_i^j \text{ is an isolated point} \big\}$$
$$I_2^j := \big\{ (i,k) \in \{1,\dots,n\} \times \{1,\dots,n\} \,:\, \mathcal{N}_i^j, \mathcal{N}_k^j \text{ are aggregated to } \mathcal{N}_t^{j-1}. \big\}$$

We mark again with $X_{.,i}^j$ the i-th colum of X_j. Hence

$$\{X_{.,i}^j \,:\, i \in I_1\} \cup \{X_{.,i}^j + X_{.,k} \,:\, (i,k) \in I_2\}$$

is a basis of $Im(P_{j,X}^{j-1}(\widetilde{V}_{j-1}))$ if we use the aggreagation method.

Lemma: 7.2.7. *Let P_j^{j-1} be given by the aggregation method. Assume that $X_j \in \mathbb{R}^{n_j \times n_j}$ is given based on its columns as*

$$X_{.,k}^j = z_t \quad \textit{for} \quad k \in I_1^j \quad \textit{and} \quad X_{.,k}^j + X_{.,i}^j = z_t \quad \textit{for} \quad (i,k) \in I_2^j$$
$$\textit{with} \quad A\, z_t = \lambda_t z_t$$

Assume further

$$\big(X_{.,k}^j\big)^T X_{.,k}^j = 1 \quad \textit{for} \quad k \in I_1^j$$
$$\big(X_{.,k}^j + X_{i,.}^j\big)^T \big(X_{.,k}^j + X_{.,i}^j\big) = 1 \quad \textit{for} \quad (i,k) \in I_2^j$$

and

$$\big(X_{.,k_1}^j\big)^T X_{.,k_2}^j = 0 \quad \textit{for} \quad k_1, k_2 \in I_1^j,\ k_1 \neq k_2$$
$$\big(X_{.,k_1}^j + X_{.,i_1}^j\big)^T \big(X_{.,k_2}^j + X_{.,i_2}^j\big) = 0 \quad \textit{for} \quad (k_1,i_1),(k_2,i_2) \in I_2^j,\ (k_1,i_1) \neq (k_2,i_2)$$
$$\big(X_{.,k}^j + X_{.,i}^j\big)^T X_t^j = 0 \quad \textit{for} \quad (k,i) \in I_2^j,\ t \in I_1^j$$

then it is

$$S_{j-1,X} = I_{j-1}.$$

proof. Based on the definitions the columns of $P_{j,X}^{j-1}$ are given from n_{j-1} orthonormal eigenvectors. This implies the assertion. □

Corollary: 7.2.8. *Assume that the assumptions of Lemma 7.2.7 are fulfilled for* $j = 0, \ldots, J-1$*, then it follows that*

$$P_{J,X}^{J-1} S_{J-1,X} P_{J-1,X}^{J-2} S_{J-2,X} \ldots, P_{j+1,X}^{j} \widetilde{v}_j \in V_{j,X} \quad \textit{for all} \quad \widetilde{v}_j \in \widetilde{V}_j.$$

proof. Based on the assumptions of Lemma 7.2.7 we obtain $S_{j,X} = I_j$ for $j = 0, \ldots, J-1$. This implies for an arbitrary $\widetilde{v}_j \in \widetilde{V}_j$

$$\begin{aligned} P_{J,X}^{J-1} S_{J-1,X} P_{J-1,X}^{J-2} S_{J-2,X} \ldots, P_{j+1,X}^{j} \widetilde{v}_j & \\ = P_{J,X}^{J-1} P_{J-1,X}^{J-2} \ldots, P_{j+1,X}^{j} \widetilde{v}_j = P_{j,X} \widetilde{v}_j \in V_{j,X}. & \end{aligned}$$

□

To sum up the results above we can maintain that if we have the modifications X_j that hold $A_{j,XX}(Im(P_{j,X}^{j-1}(\widetilde{V}_{j-1}))) \in Im(P_{j,X}^{j-1}(\widetilde{V}_{j-1}))$ then there is no problem concerning the multigrid aspects. As already discussed in section 5.3 this is only a theoretical result even for the most simple problems. Also as discussed in section 5.3 we consider that we use the eigenvectors of a block matrix that should approximate A. Then we will consider two situations presented in section 5.3:

1. If we set for two aggregated points $\mathcal{N}_i^j, \mathcal{N}_k^j$ with the eigenvector $v_{i,k}^j$ of $A_j^{(i.k)}$ the modification

$$(x_{i,i}^j, x_{k,k}^j) = \frac{v_{i,k}^j}{\|v_{i,k}^j\|}$$

then it is $\|(x_{i,i}^j, x_{k,k}^j)\| = 1$. Based on the same arguments used in Lemma 7.2.7 this implies $S_{j-1,X} = I_{j-1}$. Hence we have in this case also

$$\begin{aligned} P_{J,X}^{J-1} S_{J-1,X} P_{J-1,X}^{J-2} S_{J-2,X} \ldots, P_{j,X}^{j-1} \widetilde{v}_j & \\ = P_{J,X}^{J-1} P_{J-1,X}^{J-2} \ldots, P_{j,X}^{j-1} \widetilde{v}_{j-1} \in V_{j-1,X}. & \end{aligned}$$

However the problem is mentioned in section 5.3. Based on this modification X_j we obtain that $A_{j-1,XX}$ is in general no M-matrix because it is possible to lose the property

$$a_{i,i}^{j-1,XX} \geq \sum_{k=1,\, k \neq i}^{n} |a_{i,k}^{j-1,XX}|.$$

2. If we set for two aggregated points $\mathcal{N}_i^j, \mathcal{N}_k^j$ with the eigenvector $v_{i,k}^j$ of $A_j^{(i,k)}$ the modification

$$(x_{i,i}^j, x_{k,k}^j) = \sqrt{2}\,\frac{v_{i,k}^j}{\|v_{i,k}^j\|}$$

then it is $\|(x_{i,i}^j, x_{k,k}^j)\| = \sqrt{2}$. As already mentioned, this is the same situation as in the unmodified situation. Hence we obtain the same characteristic. It is

$$S_{j,X} = diag(s_{1,1}^{j,X}, \ldots, s_{n_j,n_j}^{j,X})$$

with $s_{k,k}^{j,X} = 1$ if $|I_k^{j,j+1}| = 1$ and $s_{k,k}^{j,X} = (x_{i,i}, x_{j,j})\,(x_{i,i}, x_{j,j})^T = 2$ if $I_k^{j,j+1} = \{\mathcal{N}_i^1, \mathcal{N}_j^1\}$. Therewith the

$$P_{J,X}^{J-1}\, S_{J-1,X}\, P_{J-1,X}^{J-2}\, S_{J-2,X}\, \ldots, P_{j,X}^{j-1}\, \widetilde{v}_j \in V_j \quad \text{for all} \quad \widetilde{v}_j \in \widetilde{V}_j \tag{7.21}$$

does not hold in general. Based on the arguments above we need the property

$$P_J^{J-1}\, S_{J-1}\, P_{J-1}^{J-2}\, S_{J-2}\, \ldots, P_j^{j-1}\, \widetilde{v}_j \in V_j \quad \text{for all} \quad \widetilde{v}_j \in \widetilde{V}_j.$$

It is obvious that we also obtain for the modified situation that (7.21) holds if the condition (2.14) is fulfilled.

8 Symmetric Problems

In this chapter we will consider the preconditioners C_{DT}^{-1} and C_{BPX}^{-1} for a symmetric stiffness matrix A. As already mentioned in chapter 2 the problem will be motivated by the partial differential equation

$$div(a(x)\, grad\, u(x)) = f(x) \quad \forall\, x \in \Omega \tag{8.1}$$
$$u(x) = 0 \quad \forall\, x \in \partial\Omega$$

as discussed in the section 2. We only take the general structure of these matrices and so we do not distinguish whether the matrix is the result of a finite element method or by finite differences. So it will be our aim in this chapter to give properties of the preconditioners $C_{BPX}^{-1}, C_{DT}^{-1}$ for s.p.d. matrices A that fulfil additional

$$a_{i,i} > 0, \quad \text{for all } i = 1, \ldots, n \tag{8.2}$$
$$a_{i,j} \leq 0, \quad \text{for all } i, j = 1, \ldots, n \text{ with } i \neq j$$
$$\sum_{j=1,\, j\neq i}^{n} |a_{i,j}| \leq |a_{i,i}|, \quad \text{forall } i = 1, \ldots, n.$$

We will take a look at the angle γ_{DT} that determines our estimations for the condition as done in chapter 3. We will determine the constant γ_{DT} for two special situations. Our main aspect will be to consider the condition of the preconditioner in the norm that is induced by A. We will take a closer (quite technical) view of the constants that appear by this estimations. As the motivation is given by the problem (8.1) this is also the situation in which we will give estimations for the constants.

As we have done before, we will first consider the two grid situation. In this case we will drop the indices on the prolongation and restriction operators. Furthermore, if we want to determine a constant we assume that the aggregation method is used to construct $V_{J-1}, \ldots, V_0$.

8.1 Introduction and problem

In this section we will consider a matrix A that results from the discritisation of (8.1) on a small sector $\Omega_s \subset \mathbb{R}$. In this sector we will see the structure of the stiffness matrices we consider, the structure of the coarser operators and the problems which arise. The problem will be that if we only consider the local situation it is in general not possible to determine a constant γ_{DT} that fulfils for all $v \in V$ the inequality

$$(8.3) \qquad ((I - Q_0)v,\, A\,P\,A_0^{-1}\,R\,v) \leq \gamma_{DT}\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I - Q_0)v\|$$
$$\Leftrightarrow \quad (w,\, A\,v_0) \leq \gamma_{DT}\,\|A\,v_0\|\,\|w\| \quad \forall v_0 \in V_0, \forall w \in W.$$

Hence we will consider two special situations in which we can give estimations for the constant. From the results of chapter 3 we know that by assuming that A is *s.p.d.* it follows that A_0 is non singular. Hence there must be a $\gamma_{DT} < 1$ because otherwise $A\,C_{DT}^{-1}$ would be singular. But this is a contradiction to Lemma 3.3.1. As a sector of the hole system we consider the situation as given in Figure 8.1.

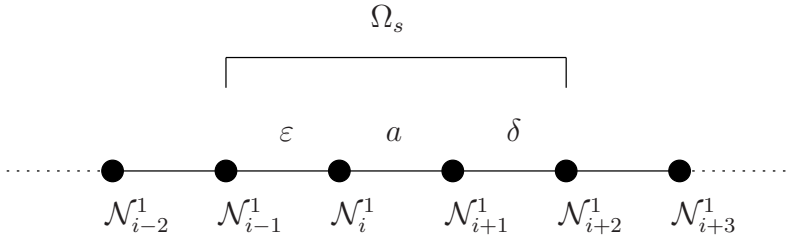


Figure 8.1: Small sector $\Omega_s \subset \Omega$

We will consider the inequality (8.3) for the sector Ω_s as shown in the Figure 8.1. We assume that for the coarser grid the points $\mathcal{N}_i^1, \mathcal{N}_{i+1}^1$ are aggregated and $\varepsilon, a, \delta > 0$ are constants that are given by the used material. So we will consider the stiffness matrix $A \in \mathbb{R}^{4\times 4}$ given by

$$(8.4) \qquad A = \begin{pmatrix} \varepsilon + * & -\varepsilon & 0 & 0 \\ -\varepsilon & a + \varepsilon & -a & 0 \\ 0 & -a & a + \delta & -\delta \\ 0 & 0 & -\delta & \delta + * \end{pmatrix}$$

As there is no conjunction to the boundary, the matrix A as given above is singular in the case of $* = 0$. It is in this case $ker(A) = (1, 1, 1, 1)^T$. That means that we can

switch the boundary to an arbitrary value. As we will do the estimations independet of the link that is given to points which does not belong to Ω_s we set $* = 0$ (We could consider the case of a sequence that converges to zero). Then we obtain for an arbitrary $v \in V$

$$P\,A_0^{-1}\,R\,v = (u_L, u, u, u_R)^T, \quad \text{with } u_L, u, u_R \in \mathbb{R}$$

$$(I - Q_0)v = (s, w, -w, t) \quad \text{with } s, w, t \in \mathbb{R}.$$

We obtain

$$s = 0, \quad (t = 0)$$

if $\mathcal{N}_{i-1}^1$, $(\mathcal{N}_{i+2}^1)$ is an isolated point. Otherwise s, t depend on values given by $\mathcal{N}_{i-1}^1$, $(\mathcal{N}_{i+2}^1)$ and left (right) neighbours $\mathcal{N}_{i-2}^1$, $(\mathcal{N}_{i+3}^1)$. So if we do not assume that $\mathcal{N}_{i-1}^1$ and $\mathcal{N}_{i+2}^1$ are isolated points we can consider the situation as follows:

$$\varepsilon = \delta = a = w = t = u_R = 1$$

$$u_L = s = -1 \quad \text{and} \quad u = 0$$

In this case follows

$$((I - Q_0)v,\, A\,P\,A_0^{-1}\,R\,v) = (-1, 1, -1, 1) \begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \end{pmatrix} = 4$$

$$\|(I - Q_0)v\| = \|\,A\,P\,A_0^{-1}\,R\,v\| = \|(-1, 1, -1, 1)\| = 2.$$

So we can not determine a $\gamma_{DT} < 1$ for the local situation. We have to consider in this case a bigger sector of Ω to get an estimation for γ_{DT}. Otherwise this implies that we can not use the estimations of chapter 3 to obtain an estimation for the condition of $A\,C_{DT}^{-1}$ and $A\,C_{BPX}^{-1}$ respectively.

In the next two sections we will consider special cases in which we can give estimation for γ_{DT} based on the small sector.

8.1.1 Both neighbours are isolated points

First we assume that both neighbours of the points we aggregate are isolated points. Therewith it is in the sector $\Omega_s \subset \Omega$

$$A\,P\,A_0^{-1}\,R\,v = (\varepsilon(u_L - u), \varepsilon(u - u_L), \delta(u - u_R), \delta(u_R - u))^T$$

$$(I - Q_0)v = (0, w, -w, 0).$$

And we obtain

$$\begin{aligned}
\|A\,P\,A_0^{-1}\,R\,v\|^2 &= 2\varepsilon^2(u_L - u)^2 + 2\delta^2(u_R - u)^2 \\
\|(I - Q_0)v\|^2 &= 2w^2 \\
\left[(A\,P\,A_0^{-1}\,R\,v,\ (I - Q_0)v)\right]^2 &= \left[w\varepsilon(u - u_L) - w\delta(u - u_R)\right]^2 \\
&\leq 2w^2(\varepsilon^2(u - u_L)^2 + \delta^2(u - u_R)^2).
\end{aligned}$$

The estimation above is based on the inequality of Young (A.0.3). So it is obvious that

$$(A\,P\,A_0^{-1}\,R\,v,\ (I - Q_0)v) \leq \gamma_{DT}\,\|A\,P\,A_0^{-1}\,R\,v\|\,\|(I - Q_0)v\|$$

holds with $\gamma_{DT} = \sqrt{1/2}$. So the estimation is independent of the elements of the matrix A. However, the assumption is quite restrictive.

8.1.2 One neighbour is an isolated point

Now we assume that one of the neighbours is an isolated point. W.l.o.g. we consider the case that $\mathcal{N}_{i-1}^1$ is isolated. As already mentioned, this implies $s = 0$. Furthermore, we assume that $\varepsilon, \delta \leq 1$, so we avoid having to distinguish many cases. As we have $\gamma_{DT} = \sqrt{1/2}$ in the situation that both neighbours are isolated points, it is obvious that $\gamma_{DT} = \sqrt{1/2}$ is a lower bound for the case that only one neighbour is an isolated point. We obtain

$$\begin{aligned}
\|A\,P\,A_0^{-1}\,R\,v\|^2 &= 2\varepsilon^2(u_L - u)^2 + 2\delta^2(u_R - u)^2 \\
\|(I - Q_0)v\|^2 &= 2w^2 + t^2 \\
\left[(A\,P\,A_0^{-1}\,R\,v,\ (I - Q_0)v)\right]^2 &= \left[w\varepsilon(u - u_L) - w\delta(u - u_R) + t\delta(u_R - u)\right]^2.
\end{aligned}$$

We transform the variables in $x := u_L - u$ and $y := u_R - u$. Therewith we have to determine an $\gamma_{DT} < 1$ that holds

$$g := \gamma_{DT}^2\,(2\varepsilon^2x^2 + 2\delta^2y^2) \cdot (2w^2 + t^2) - [w\varepsilon x - w\delta y + t\delta y]^2 \geq 0.$$

To minimize g with respect to t we differentiate g with respect to t. We obtain that the minimizing t is given by

$$t = \frac{w(\varepsilon x - \delta y)\delta y}{\gamma_{DT}^2(2\varepsilon x^2 + 2\delta y^2) - \delta^2 y^2}.$$

We highlight, that the denominator is positive if it is $2\gamma_{DT}^2 \geq \delta$. By assuming that $\delta \leq 1$ this is always fulfiled if it is $\gamma_{DT}^2 \geq 1/2$. If we insert the value for t in g it follows

$$g = \frac{2\gamma_{DT}^2 w^2(\varepsilon^2 x^2 + \delta^2 y^2)(\varepsilon^2(-1 + 4\gamma_{DT}^2)x^2 + 2\delta\varepsilon xy + \delta^2(-3 + 4\gamma_{DT}^2)y^2)}{(2\gamma_{DT}^2(\varepsilon^2 x^2 + \delta^2 y^2) - \delta^2 y^2}$$

As the denominator is positive it is sufficient to consider the numerator. We obtain that this is positive if it is

$$\begin{aligned} g_0 &:= (\varepsilon^2(-1 + 4\gamma_{DT}^2)x^2 + 2\delta\varepsilon xy + \delta^2(-3 + 4\gamma_{DT}^2)y^2) \\ &= 4\gamma_{DT}^2(x^2\varepsilon^2 + y^2\delta^2) - 2\delta^2 y^2 - (\varepsilon x - \delta y)^2 \geq 0. \end{aligned}$$

Then is g_0 minimized with respect to x if it is

$$x = -\frac{\delta y}{\varepsilon(-1 + 4\gamma_{DT}^2)}.$$

And again by assuming that $\varepsilon \leq 1$ we get for $\gamma_{DT}^2 \geq \sqrt{1/4}$ that the denominator is positive. It follows by this value for x

$$g_0 = \frac{2\delta^2(1 - 8\gamma_{DT}^2 + 8\gamma_{DT}^4)y^2}{-1 + 4\gamma_{DT}^2}$$

As the denominator is positive for $\gamma_{DT}^2 \geq \sqrt{1/4}$ we consider again only the numerator. This is positive if it is

$$(1 - 8\gamma_{DT}^2 + 8\gamma_{DT}^4 \geq 0 \quad \Leftrightarrow \quad \gamma_{DT}^2 \geq \frac{1}{2} + \frac{\sqrt{2}}{4}.$$

8.2 Two grid estimations for $C_{DT}^{-1} A$ in the A-norm

For other estimations of the eigenvalues of $C_{DT}^{-1} A$ we consider the condition of $C_{DT}^{-1} A$ that is given by the norm induced by A. That means for $A \in \mathbb{R}^{n \times n}$ s.p.d. we consider the inner product and the induced norm follow as:

$$a(u, v) := u^T A\, v \quad \text{and} \quad \|u\|_A = \sqrt{a(u, u)}. \tag{8.5}$$

For $A \in \mathbb{R}^{n\times n}$ and $f \in \mathbb{R}^n$ we define $u^* \in \mathbb{R}^n$ by

$$A\,u^* := f \quad \text{and} \quad u^* := A^{-1}\,f, \quad \text{respectively.} \tag{8.6}$$

For the same A, f we define $u_1, u_0 \in \mathbb{R}^n$ by

$$u_1 := A^{-1}\,(I - Q_0)\,f \tag{8.7}$$

$$u_0 := P\,A_0^{-1}\,R\,f. \tag{8.8}$$

Based on these definitions it is obvious that we obtain from the definition (3.9) that

$$u_0 + u_1 = C_{DT}^{-1}\,f.$$

We therefore get a relation between the solution u^* and the vectors u_0, u_1 we get by using the preconditioner C_{DT}^{-1}. Since the solution u^* of $A\,u = f$ is given by

$$a(u^*,\, v) = (f,\, v) \quad \forall v \in V$$

we will see in the next lemma that the vectors u_0, u_1 can be interpreted as solutions in a subspace of $V \equiv \mathbb{R}^n$. For u_1 this is quite obvious. For u_0 we take it as a result.

Lemma: 8.2.1. *Let $A \in \mathbb{R}^{n\times n}$ be s.p.d. For u_1, u_0 as defined in (8.7), (8.8) it holds*

$$a(u_1,\, v) = a(u^*,\, v - Q_0\,v) = (f,\, v - Q_0\,v), \quad \forall v \in V \tag{8.9}$$

$$a(u_0,\, v_0) = a(u^*,\, v_0) = (f,\, v_0) \quad \forall v_0 \in V_0. \tag{8.10}$$

proof. As $Q_0, (I - Q_0)$ are orthogonal projection with respect to the inner product $(.,.)$ they hold $Q_0^T = Q_0$ and $(I - Q_0)^T = (I - Q_0)$. Hence we get for the equation (8.9) that it is for all $v \in V$

$$\begin{aligned} a(u_1, v) &= a(A^{-1}(I - Q_0)f,\, v) = ((I - Q_0)\,f,\, v) \\ &= (f,\, (I - Q_0)\,v) = (A\,A^{-1}\,f,\, (I - Q_0)\,v) = a(u^*,\, (I - Q_0)v). \end{aligned}$$

The second equation of this lemma is obtained as it is $Q_0\,v_0 = P\,S\,R\,v_0 = v_0$ for all $v_0 \in V_0$ and $P = R^T$. Therewith we obtain for all $v_0 \in V_0$

$$\begin{aligned} a(u_0,\, v_0) &= a(u_0,\, Q_0\,v_0) = a(P\,A_0^{-1}\,R\,f,\, P\,S\,R\,v_0) = (A\,P\,A_0^{-1}\,R\,f,\, P\,S\,R\,v_0) \\ &= (\underbrace{R\,A\,P}_{A_0}\,A_0^{-1}\,R\,f,\, S\,R\,v_0) = (A_0\,A_0^{-1}\,R\,f,\, S\,R\,v_0) \\ &= (R\,f,\, S\,R\,v_0) = (f,\, Q_0\,v_0) = a(u^*,\, v_0). \end{aligned}$$

□

As we will use for estimations the norm $\|\,.\,\|_A$ for the single additional terms u_1, u_0 we get by using the preconditioning operator C_{DT}^{-1}, we will need a relation between

$$\|u_1\|_A,\ \|u_0\|_A \quad \text{and} \quad \|u_1 + u_0\|_A.$$

We will provide this in the next lemma.

Lemma: 8.2.2. *Let A be a s.p.d. matrix. For u_1, u_0 as defined in (8.7), (8.8) it is*

$$a(u_0, u_1) = 0.$$

proof. By using the definitions of u_0, u_1 we get

$$\begin{aligned} a(u_0,\, u_1) &= a(A^{-1}\,(I - Q_0)\,f,\, P\,A_0^{-1}\,R\,f) \\ &= ((I - Q_0)\,f,\, P\,A_0^{-1}\,R\,f) = (R(I - P\,S\,R)\,f,\, A_0^{-1}\,R\,f) \\ &= ((R - R)\,f,\, A_0^{-1}\,R\,f) = 0. \end{aligned}$$

The last equality follows as S is defined by $S = (R\,P)^{-1}$. □

To get the estimation between u_1, u_0 and u^* we define the constant c_a by

$$c_a := \sup\left\{ \frac{\|Q_0\,v\|_A}{\|v\|_A} \ :\ v \in V\backslash\{0\} \right\}. \tag{8.11}$$

Based on the definiton of c_a it is obvious that with the constant c_a the inequality

$$\|Q_0\,v\|_A \le c_a\,\|v\|_A \tag{8.12}$$

holds for all $v \in V$. Moreover, it is obvious that c_a depends on the structure of the matrix A. As $Q_0 : V \to V_0$ is the orthogonal projection the definition (8.11) is equivalent to

$$c_a = \sup\left\{ \frac{\|v_0\|_A}{\|v_0 + w\|_A} \ :\ v_0 \in V_0, w \in W,\, v_0 + w \neq 0. \right\}.$$

Therewith it is obvious that c_a depends on the structure of the matrix A and the subspaces $V_0, W \subset V$.

As the inequality (8.12) holds for $v \in V_0$, it follows $c_a \ge 1$ from $Q_0\,v_0 = v_0$ for all $v_0 \in V_0$. If we take the constant c_a as given then this implies the following result:

Lemma: 8.2.3. *Let $A \in \mathbb{R}^{n\times n}$ s.p.d. For u_1, u_0 and u^* as defined in (8.7), (8.8) and (8.6) then*

$$\|u_0\|_A \leq \|u^*\|_A \tag{8.13}$$

$$\|u_1\|_A \leq (1 + c_a)\|u^*\|_A \tag{8.14}$$

$$\|u^*\|_A \leq c_a(\|u_0\|_A + \|u_1\|_A). \tag{8.15}$$

holds.

proof. For u_0 based on the equation (8.10) of Lemma 8.2.1 and the inequality of Cauchy-Schwarz (A.0.1)

$$\|u_0\|_A^2 = a(u_0,\, u_0) = a(u^*, u_0) \leq \|u^*\|_A \, \|u_0\|_A.$$

This implies

$$\|u_0\|_A \leq \|u^*\|_A.$$

Based on the equation (8.9) of Lemma 8.2.1 and the inequality of Cauchy-Schwarz (A.0.1) we get for u_1

$$\begin{aligned}\|u_1\|_A^2 = a(u_1, u_1) = a(u^*,\, u_1 - Q_0\, u_1) = a(u^*,\, u_1) - a(u^*,\, Q_0\, u_1)\\ \leq \|u^*\|_A(\|u_1\|_A + \|Q_0\, u_1\|_A) \leq (1 + c_a)\|u^*\|_A \, \|u_1\|_A.\end{aligned}$$

This implies

$$\|u_1\|_A \leq (1 + c_a)\|u^*\|_A.$$

The inequality (8.15) is also obtained by using the results of Lemma 8.2.1 and the inequality of Cauch-Schwarz (A.0.1). We obtain

$$\begin{aligned}\|u^*\|_A^2 &= a(u^*,\, u^*) = a(u^*,\, u^* - Q_0 u^*) + a(u^*,\, Q_0 u^*)\\ &= a(u_1,\, u^*) + a(u_0,\, Q_0 u^*) \leq \|u_1\|_A \, \|u^*\|_A + \|u_0\|_A \, \|Q_0 u^*\|_A\\ &\leq \|u^*\|_A \, \|u_1\|_A + c_a \, \|u_0\|_A \, \|u^*\|_A \leq c_a \, \|u^*\|_A \, (\|u_1\|_A + \|u_0\|_A).\end{aligned}$$

In the last inequality we use $c_a \geq 1$. □

From these results we can give an estimation for the condition of $C_{DT}^{-1} A$ that only depends on the constant c_a.

Theorem: 8.2.4. *Let A be a non singular s.p.d. matrix. With c_a as defined in (8.11) then*

$$c_{DT}\|v\|_A \leq \|C_{DT}^{-1} A\, v\|_A \leq d_{DT}\|v\|_A$$

holds for all $v \in V$ with

$$c_{DT} = \frac{1}{c_a\sqrt{2}} \quad \textit{and} \quad d_{DT} = 2 + c_a.$$

proof. To prove

$$c_{DT}\|v\|_A \leq \|C_{DT}^{-1} A\, v\|_A \leq d_{DT}\|v\|_A$$

for all $v \in V$ it is equivalent to set $v = A^{-1}f$ and prove

$$\begin{aligned} & c_{DT}\|A^{-1} f\|_A \leq \|C_{DT}^{-1} f\|_A \leq d_{DT}\|A^{-1} f\|_A \\ \Leftrightarrow \quad & c_{DT}\|u^*\|_A \leq \|u_0 + u_1\|_A \leq d_{DT}\|u^*\|_A. \end{aligned}$$

The second equivalence follows the defintion of u^*, u_0 and u_1. From the equations (8.13) and (8.14) of Lemma 8.2.3 it follows

$$\|u_0 + u_1\|_A \leq \|u_0\|_A + \|u_1\|_A \leq (2 + c_a)\|u^*\|_A.$$

This proves the proposition for d_{DT}. As u_0, u_1 are orthogonal with respect to the inner product $a(.,.)$ we obtain from equation 8.15 and Lemma A.0.5

$$\|u^*\|_A \leq c_a\,(\|u_0\|_A + \|u_1\|_A) \leq \sqrt{2}c_a\|u_0 + u_1\|_A.$$

This complets the proof for c_{DT}. □

8.3 Technical view of the constant c_a. (Neigbours are isolated points)

In section 8.2 we have proved an estimation for the condition of the operator $C_{DT}^{-1} A$ in the norm $\|\,.\,\|_A$. The estimation depends on a constant c_a that fulfils the inequality

$$\|Q_0\, v\|_A \leq c_a\, \|v\|_A.$$

We have already highlighted that this constant depends on the elements of the matrix A and the structure of the subspaces. In particular we have seen that it is

$$\frac{1}{\sqrt{2}c_a}\|v\|_A \leq \|C_{DT}^{-1}\,A\,v\|_A.$$

Therewith $\frac{1}{\sqrt{2}c_a}$ is our lower bound for the absolute value of the eigenvalue λ of $C_{DT}^{-1}\,A$. So if we can estimate an upper bound for c_a that only depends on the elements of the matrix A we get a lower bound for the constant c_{DT} with the same dependency. That is what we will do in this section for problems that result from the discretisation of the problem (8.1) and for matrices that have the structure as given in (8.2) respectively. The restriction we assume in this section is that we only aggregate two points into a new one and the neighbours of the aggregated points are all isolated points. That means the sets $I_t^{0,1}$ all have the cardinal number one or two. And if $\mathcal{N}_1^1, \mathcal{N}_2^1$ are aggregated to $\mathcal{N}_t^0$ and it is $a_{1,k} \neq 0$ or $a_{2,k} \neq 0$ for an $k \in \{3, \ldots, n\}$ then $\mathcal{N}_k^1$ is an isolated point. So this is a strict assumption concerning the aggregations that are done. These assumtions are also used in section 5.2.1 for an exact one sided modification for this problem. For it we will have no restriction concerning the geometrical structure of the grid points or the dimension of the system the stiffness matrix is based on.

We start with a simple one dimensional situation and consider the local situation as illustrated in Figure 8.2. Furthermore we assume that the two grid points $\mathcal{N}_i^1, \mathcal{N}_{i+1}^1$ will be aggregated

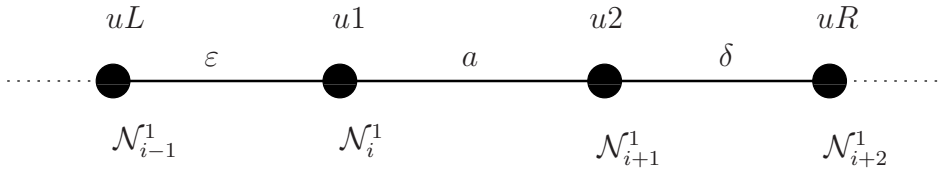


Figure 8.2: One dimensional situation

If we only consider the sector $\Omega_s = [\mathcal{N}_{i-1}, \mathcal{N}_{i+2}]$ then we locally get the stiffness matrix A as

$$A = \begin{pmatrix} \varepsilon & -\varepsilon & & \\ -\varepsilon & a+\varepsilon & -a & \\ & -a & a+\delta & -\delta \\ & & -\delta & \delta \end{pmatrix}.$$

As a result of arguments from section 8.1 the matrix A in the sector is singular. Therewith in this sector for a function $u \in V$

$$u = (u_L, u_1, u_2, u_R)$$

we obtain the values

$$\|u\|_A^2 = \varepsilon(u_L - u_1)^2 + a(u_1 - u_2)^2 + \delta(u_2 - u_R)^2$$

$$\|Q_0\, u\|_A^2 = \varepsilon\left(\frac{u_1 + u_2}{2} - u_L\right)^2 + \delta\left(\frac{u_1 + u_2}{2} - u_R\right)^2.$$

Therefore, we can give an estimation for c_a that only depends on the elements of the matrix.

Lemma: 8.3.1. *Assume that the situation is given as above and $\mathcal{N}_{i-1}^1, \mathcal{N}_{i+2}^1$ are isolated points then the inequality*

$$\|Q_0\, u\|_A \leq c_A \|u\|_A$$

holds for all $u \in V$ with

$$c_a = \sqrt{1 + \frac{\varepsilon + \delta}{4a}}.$$

proof. We just have to prove this for the restricted area as mentioned in the setting. So we can set $u = (u_L, u_1, u_2, u_R) \in \mathbb{R}^4$ and show the inequality $g \geq 0$ with

$$\begin{aligned} g &= c_a^2 \|u\|_a^2 - \|Q_0\, u\|_a^2 \\ &= c_a^2 \left(\varepsilon(u_L - u_1)^2 + a(u_1 - u_2)^2 + \delta(u_2 - u_R)^2\right) \\ &\quad - \left(\varepsilon\left(\frac{u_1 + u_2}{2} - u_L\right)^2 + \delta\left(\frac{u_1 + u_2}{2} - u_R\right)^2\right) \end{aligned}$$

To minimize g with respect to u_L, u_R we differentiatethe function g with respect to u_L and u_R. It follows

$$\frac{dg}{du_L} = 2c_a^2 \varepsilon(u_L - u_1) - 2\varepsilon\left(u_L - \frac{u_1 + u_2}{2}\right) \tag{8.16}$$

$$\frac{dg}{du_R} = 2c_a^2 \delta(u_R - u_2) - 2\delta\left(u_R - \frac{u_1 + u_2}{2}\right) \tag{8.17}$$

and we obtain by the first order condition for a minimum

$$u_L = \frac{c_a^2 u_1 - (u_1+u_2)/2}{c_a^2 - 1} \tag{8.18}$$

$$u_R = \frac{c_a^2 u_2 - (u_1+u_2)/2}{c_a^2 - 1} \tag{8.19}$$

With the g minimizing values for u_L, u_R we get for the function g :

$$\begin{aligned} g &\geq c_a^2 \left[\frac{\varepsilon}{(c_a^2-1)^2} \left(\frac{u_1-u_2}{2} \right)^2 + \frac{\delta}{(c_a^2-1)^2} \left(\frac{u_1-u_2}{2} \right)^2 + a(u_1-u_2)^2 \right] \\ &\quad - \left[\frac{\varepsilon}{(c_a^2-1)^2} \left(c_a^2 \frac{u_1-u_2}{2} \right)^2 + \frac{\delta}{(c_a^2-1)^2} \left(c_a^2 \frac{u_1-u_2}{2} \right)^2 \right] \\ &= -\frac{(c_a^2-1)c_a^2}{(c_a^2-1)^2}(\varepsilon+\delta)\frac{(u_1-u_2)^2}{4} + \frac{c_a^2}{(c_a^2-1)^2} a (c_a^2-1)^2 (u_1-u_2)^2 \\ &= \frac{(c_a^2-1)\, c_a^2\, (u_1-u_2)^2}{(c_a^2-1)^2} \left[-\frac{\varepsilon+\delta}{4} + (c_a^2-1)a \right] \end{aligned}$$

Since we have $c_a \geq 1$ last term is non negative if and only if it holds

$$(c_a^2-1)a \geq \frac{\varepsilon+\delta}{4} \quad \Leftrightarrow \quad c_a^2 \geq \sqrt{1+\frac{\varepsilon+\delta}{4a}}.$$

This shows the proposition. □

Moreover, the proof of the last lemma implies that the chosen constant c_a is the smallest possible constant.

Remark: 8.3.2. *Assume the situation is given as in Lemma 8.3.1. Then there is no* $\widetilde{c}_a < \sqrt{1+\frac{\varepsilon+\delta}{4a}}$ *that holds the inequality* $\|Q_0\, u\|_A \leq c_A \|u\|_A$ *for all* $u \in V$.

proof. We take the system of Figure 8.2 and set

$$u_1 = 1 \quad \text{and} \quad u_2 = c_a^2.$$

The equalities (8.18) and (8.19) motivate to set

$$u_L = \frac{c_a^2 - (1+c_a^2)/2}{c_a^2-1} = 1/2 \quad \text{and} \quad u_R = \frac{c_a^2 \cdot c_a^2 - (1+c_a^2)/2}{c_a^2-1}.$$

In this case we get

$$(c_a^2-1)\big(c_a^2\|u\|_A^2-\|Q_a u\|_A^2\big)=c_a^8 a+c_a^6\left(-3a-\frac{|\varepsilon+\delta|}{4}\right)$$
$$+c_a^4\left(3a+\frac{|\varepsilon+\delta|}{2}\right)+c_a^2\left(-a-\frac{|\varepsilon+\delta|}{4}\right).$$

The expression on the right side is zero if we set $c_a^2=1+(\varepsilon+\delta)/4a$. □

Moreover, we highlight that the minimizing values for u_L, u_R as given in (8.18) and (8.19) do not depend on the couplings represented by ε, δ and a. This characteristic will be usefull for generalizations.

We generalise the considered system in a two dimensional grid. The matrix A and the two points we will aggregate respectively fulfil the following condition:
Let $\mathcal{N}_1^1, \mathcal{N}_2^1$ two grid points that will be aggregated and the stiffness matrix A satisfies $a_{1,2} \neq 0$. Moreover, $\mathcal{N}_k^1$ is an isolated point if it is $k \in \{3,\dots,n\}$ and $a_{1,k} \neq 0$ or $a_{2,k} \neq 0$. And there is no $k \in \{3,\dots,n\}$ with

$$a_{1,k} \neq 0 \quad \text{and} \quad a_{2,k} \neq 0.$$

We call such a situation an open system. The structure is as shown in Figure 8.3.
Furthermore, we assume that the grid points $\mathcal{N}_1^1, \mathcal{N}_2^1$ are interior points of Ω. That means that it holds for $i = 1, 2$

$$a_{i,i} = \sum_{k=1,\, k\neq i} |a_{i,k}|.$$

The following definitions are for an easier notation:

$$a^L := \sum_{k=3}^{n} |a_{1,k}| \quad \text{and} \quad a^R := \sum_{k=3}^{n} |a_{2,k}|$$

$$I := \{1,2\} \cup I_1 \cup I_2$$

$$\text{with} \quad k \in I_1 :\Leftrightarrow a_{1,k} \neq 0 \wedge a_{2,k} = 0$$

$$\text{and} \quad k \in I_2 :\Leftrightarrow a_{2,k} \neq 0 \wedge a_{1,k} = 0.$$

Based on this setting we can generalize the result of Lemma 8.3.1 in the following way:

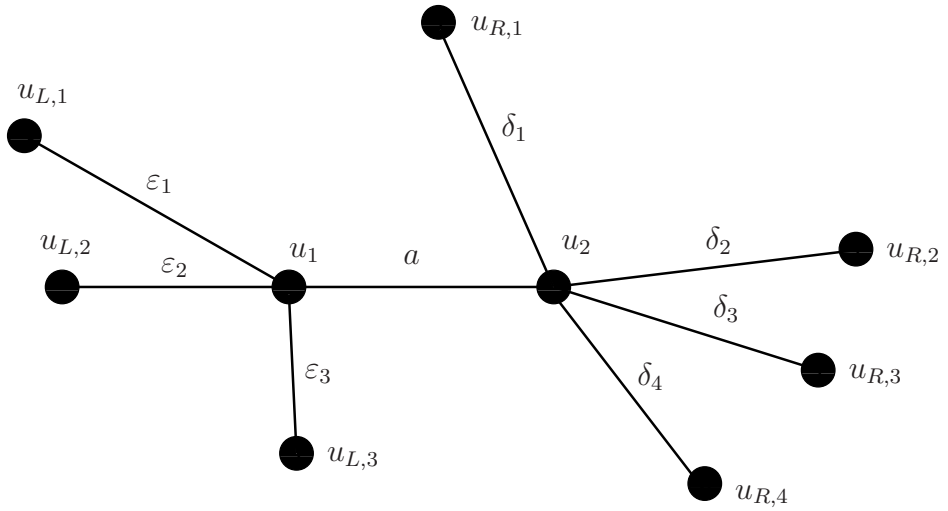


Figure 8.3: Open system

Lemma: 8.3.3. *Let $A \in \mathbb{R}^{n\times n}$ s.p.d. be a matrix as defined in (8.2). Let $\mathcal{N}_1^1, \mathcal{N}_2^1$ be two aggregatet interior points with $a_{1,2} \neq 0$. If we assume that all neighbours of $\mathcal{N}_1^1, \mathcal{N}_2^1$ are isolated points and the setting as given above then the inequality*

$$c_a \|u\|_A \geq \|Q_0\, u\|_A \tag{8.20}$$

holds for all $u \in V$ with $c_a = \sqrt{1 + (a^L + a^R)/4|a_{1,2}|}$

proof. As in the proof of Lemma 8.3.1 it is sufficient to prove the inequality of the restricted area that is connected to u_1, u_2. Similary to the proof of Lemma 8.3.1 we prove that it is

$$g := c_a^2 \|u\|_A^2 - \|Q_0\, u\|_A^2 \geq 0$$

for all $u \in \mathbb{R}^n$. For $u = (u_1, \ldots, u_n)$ it follows

$$\|u\|_A^2 = |a_{1,2}|(u_1 - u_2)^2 + \sum_{k\in I_1} |a_{1,k}|(u_1 - u_k)^2 + \sum_{k\in I_2} |a_{2,k}|(u_2 - u_k)^2 + \left(\sum_{i,j\in I_1\cup I_2} |a_{i,j}|(u_i - u_j)^2 \right) \tag{8.21}$$

$$\|Q_0\, u\|_A^2 = \sum_{k\in I_1} |a_{1,k}| \left(\frac{u_1 + u_2}{2} - u_k \right)^2 + \sum_{k\in I_2} |a_{2,k}| \left(\frac{u_1 + u_2}{2} - u_k \right)^2 . + \left(\sum_{i,j\in I_1\cup I_2} |a_{i,j}|(u_i - u_j)^2 \right) . \tag{8.22}$$

So we consider again the weighted difference by c_a of these expressions. As the proposed constant c_a holds $c_a \geq 1$ we can estimate as follows:

$$\begin{aligned} g &= c_a^2 \left(|a_{1,2}|(u_1-u_2)^2 + \sum_{k\in I_1} |a_{1,k}|(u_1-u_k)^2 + \sum_{k\in I_2} |a_{2,k}|(u_2-u_k)^2 \right) \\ &\quad + (c_a^2-1) \left(\sum_{i,j\in I_1\cup I_2} |a_{i,j}|(u_i-u_j)^2 \right) \\ &\quad - \left(\sum_{k\in I_1} |a_{1,k}| \left(\frac{u_1+u_2}{2} - u_k \right)^2 + \sum_{k\in I_2} |a_{2,k}| \left(\frac{u_1+u_2}{2} - u_k \right)^2 \right) \\ &\geq c_a^2 \left(|a_{1,2}|(u_1-u_2)^2 + \sum_{k\in I_1} |a_{1,k}|(u_1-u_k)^2 + \sum_{k\in I_2} |a_{2,k}|(u_2-u_k)^2 \right) \\ &\quad - \left(\sum_{k\in I_1} |a_{1,k}| \left(\frac{u_1+u_2}{2} - u_k \right)^2 + \sum_{k\in I_2} |a_{2,k}| \left(\frac{u_1+u_2}{2} - u_k \right)^2 \right) \\ &=: g_0. \end{aligned}$$

Now we minimize the function g_0 in the variables $u_k,\ k \in I_1 \cup I_2$. We get

$$\begin{aligned} \frac{dg_0}{du_k} &= c_a^2 2u_1|a_{1,k}| - (u_1+u_2)|a_{1,k}| - 2u_k(c_a^2-1)|a_{1,k}| \quad \text{for} \quad k \in I_1 \\ \text{and} \quad \frac{dg_0}{du_k} &= c_a^2 2u_2|a_{2,k}| - (u_1+u_2)|a_{2,k}| - 2u_k(c_a^2-1)|a_{2,k}| \quad \text{for} \quad k \in I_2. \end{aligned}$$

And we get as the first order condition for the g_0 minimizing values

$$u_k = \frac{c_a^2 u_1 - (u_1+u_2)/2}{c_a^2-1} \quad \text{for} \quad k \in I_1 \tag{8.23}$$

$$\text{and} \quad u_k = \frac{c_a^2 u_2 - (u_1+u_2)/2}{c_a^2-1} \quad \text{for} \quad k \in I_2. \tag{8.24}$$

As in the proof of Lemma 8.3.1 we use the minimizing values for $u_k,\ k \in I_1 \cup I_2$ then

we obtain

$$\begin{aligned} g_0 \geq c_a^2 &\left[|a_{1,2}|(u_1-u_2)^2 + \sum_{k\in I_1} \frac{|a_{1,k}|}{(c_a^2-1)^2} \left(\frac{u_1-u_2}{2}\right)^2 + \sum_{k\in I_2} \frac{|a_{2,k}|}{(c_a^2-1)^2} \left(\frac{u_1-u_2}{2}\right)^2 \right] \\ &- \left[\sum_{k\in I_1} \frac{|a_{1,k}|}{(c_a^2-1)^2} \left(c_a^2\frac{u_1-u_2}{2}\right)^2 + \sum_{k\in I_2} \frac{|a_{2,k}|}{(c_a^2-1)^2} \left(c_a^2\frac{u_1-u_2}{2}\right)^2 \right] \\ = &\frac{(1-c_a^2)c_a^2}{(c_a^2-1)^2} \sum_{k\in I_1\cup I_2} (|a_{1,k}|+|a_{2,k}|) \left(\frac{u_1-u_2}{2}\right)^2 + \frac{c_a^2(c_a^2-1)^2}{(c_a^2-1)^2}|a_{1,2}|(u_1-u_2)^2 \\ = &\frac{(1-c_a^2)c_a^2}{(c_a^2-1)^2}(a^L+a^R) \left(\frac{u_1-u_2}{2}\right)^2 + \frac{c_a^2(c_a^2-1)^2}{(c_a^2-1)^2}|a_{1,2}|(u_1-u_2)^2 \end{aligned}$$

As we have $(u_1-u_2)^2 \geq 0$ and $(c_a^2-1) \geq 0$ it is $g_0 \geq 0$ if and only if it is

$$0 \leq (c_a^2-1)|a_{1,2}| - \frac{a^L+a^R}{4}$$

As in the proof of Lemma 8.3.1 we obtain

$$c_a \geq \sqrt{1+\frac{a^L+a^R}{4|a_{1,2}|}}$$

as sufficient condition for $g \geq g_0 \geq 0$. This shows the proposition. □

Remark: 8.3.4. *If we take a look at the two conditions (8.23) and (8.24) then we see that the minimizing situation for the neighbours of u_1, u_2 does not depend on the number of neighbours or a structure of the grid. This property implies the same structure for the constant c_a independently of the structure of the grid or the number of neighbours. The constant only depends on the relation of the link between the points $\mathcal{N}_1^1, \mathcal{N}_2^1$ compared with the sum of the links to other points, no matter how the sum $a^L + a^R$ of links is partitioned among neighbours.*

The Remark 8.3.2 shows for the simple one dimensional problem that the constant c_a can not be estimated in a better way than in Lemma 8.3.1. As the Lemma 8.3.3 is a generalisation of this, we get the same result for the constant c_a defined in Lemma 8.3.3.

This holds based on the property that the constant c_a in Lemma 8.3.3 is the same as in Lemma 8.3.1, if we only have the one dimensional situation. We have seen that the worst values for the neighbours of the aggregated points do not depend on the dimension or the structure of the links.

Now we will estimate the constant c_a for the aggregation of two arbitrary interior points. The generelisation in this step is that there may be indices $k \in I_1 \cap I_2$. That means the sets I_1, I_2 are now defined as

$$I_1 := \{k \in \{3, \dots, n\} \,:\, a_{1,k} \neq 0\} \quad \text{and} \quad I_2 := \{k \in \{3, \dots, n\} \,:\, a_{2,k} \neq 0\}.$$

Moreover, it still holds the definitions of a^L, a^R as

$$a^L := \sum_{k \in I_1} |a_{1,k}| \quad \text{and} \quad a^R := \sum_{k \in I_2} |a_{2,k}|.$$

We assume $a_{1,2} \neq 0$ and if $a_{1,k} \neq 0$ or $a_{2,k} \neq 0$ for an $k \in \{3, \dots, n\}$ then $\mathcal{N}_k^1$ is an isolated point. Then we define the sets

$$I^* := I_1 \cap I_2, \quad I_1^* := I_1 \backslash I^* \quad \text{and} \quad I_2^* := I_2 \backslash I^*.$$

Lemma: 8.3.5. *Let A s.p.d. be a matrix as defined in (8.2). Let $\mathcal{N}_1^1, \mathcal{N}_2^1$ be two aggregated interior points with $a_{1,2} \neq 0$. If we assume the setting given above then the inequality*

$$c_a \, \|u\|_A \geq \|Q_0 \, u\|_A$$

holds for all $u \in V$ with $c_a = \sqrt{1 + (a^L + a^R)/4|a_{1,2}|}$

proof. We consider the following expressions:

$$\begin{aligned} \|\widetilde{u}\|_A^2 = {} & |a_{1,2}|(u_1 - u_2)^2 + \sum_{k \in I_1^*} |a_{1,k}|(u_1 - u_k)^2 + \sum_{k \in I^*} |a_{1,k}|(u_1 - u_{k,1})^2 \\ & + \sum_{k \in I_2^*} |a_{2,k}|(u_2 - u_k)^2 + \sum_{k \in I^*} |a_{2,k}|(u_2 - u_{k,2})^2 \\ & + \left(\sum_{i,j \in I_1 \cup I_2} |a_{i,j}|(u_i - u_j)^2 \right) \end{aligned} \tag{8.25}$$

$$\begin{aligned} \|\widetilde{Q_0 \, u}\|_A^2 = {} & \sum_{k \in I_1^*} |a_{1,k}| \left(\frac{u_1 + u_2}{2} - u_k \right)^2 + \sum_{k \in I^*} |a_{1,k}| \left(\frac{u_1 + u_2}{2} - u_{k,1} \right)^2 \\ & + \sum_{k \in I_2^*} |a_{2,k}| \left(\frac{u_1 + u_2}{2} - u_k \right)^2 + \sum_{k \in I^*} |a_{2,k}| \left(\frac{u_1 + u_2}{2} - u_{k,2} \right)^2 \\ & + \left(\sum_{i,j \in I_1 \cup I_2} |a_{i,j}|(u_i - u_j)^2 \right). \end{aligned} \tag{8.26}$$

Then we obtain in the case of $u_{k,1} = u_{k,2}$ for all $k \in I^*$

$$\|\widetilde{u}\|_A^2 = \|u\|_A^2 \quad \text{and} \quad \|\widetilde{Q_0\, u}\|_A^2 = \|Q_0\, u\|_A^2.$$

From the result of Lemma 8.3.3 and the given constant c_a we obtain

$$\sup_{\widetilde{u}\neq 0} \frac{\|\widetilde{Q_0\, u}\|_A^2}{\|\widetilde{u}\|_A^2} \leq c_a.$$

Thus the proposition follows from

$$\sup_{u\neq 0} \frac{\|Q_0\, u\|_A^2}{\|u\|_A^2} = \sup_{u\neq 0,\, u_{k,1}=u_{k,2}\,\forall k\in I^*} \frac{\|\widetilde{Q_0\, u}\|_A^2}{\|\widetilde{u}\|_A^2} \leq \sup_{u\neq 0} \frac{\|\widetilde{Q_0\, u}\|_A^2}{\|\widetilde{u}\|_A^2} \leq c_a.$$

□

The central point of the proof of Lemma 8.3.5 is that the open system we have considered in Lemma 8.3.3 is more general than the system we consider in Lemma 8.3.5. Therewith the system in Lemma 8.3.5 is a special case of Lemma 8.3.3. The system we have considered in this lemma and the structure of the proof are illustrated in Figure 8.4 at page 247. We start from a general system and then we cut it open to get the situation as considered before in Lemma 8.3.3 Figure 8.3 on page 242. The sets of indices as used in the proof are

$$I_1 = \{(L,2),(L,3),(G,1),(G,2)\} \quad I_2 = \{(R,2),(R,3),(R,4),(G,1),(G,2)\}$$

$$I^* = \{(G,1),(G,2)\} \quad I_1^* = \{(L,2),(L,3)\}$$

and $\quad I_2^* = \{(R,2),(R,3),(R,4)\}.$

The last generalisation is to consider points that can be coupled to the boundary, too. We consider s.p.d. matrices $A \in \mathbb{R}^{n\times n}$ for their elements holding

$$a_{i,i} > 0, \quad \text{for all} \quad i = 1,\ldots n, \quad a_{i,j} \leq 0, \quad \text{for all} \quad i \neq j$$

$$\text{and} \quad \sum_{j=1,\, j\neq i}^{n} |a_{i,j}| \leq |a_{i,i}| \quad \text{for all} \quad i = 1,\ldots,n.$$

If it is

$$0 < r_i = a_{i,i} - \sum_{j=1,\, j\neq i}^{n} |a_{i,j}|$$

then r_i is the couple of the boundary. We still assume that if $a_{1,k} \neq 0$ or $a_{2,k} \neq 0$ for an $k \in \{3,\ldots,n\}$ then $\mathcal{N}_k^1$ is an isolated point. Hence we get the same structure as in the estimation above.

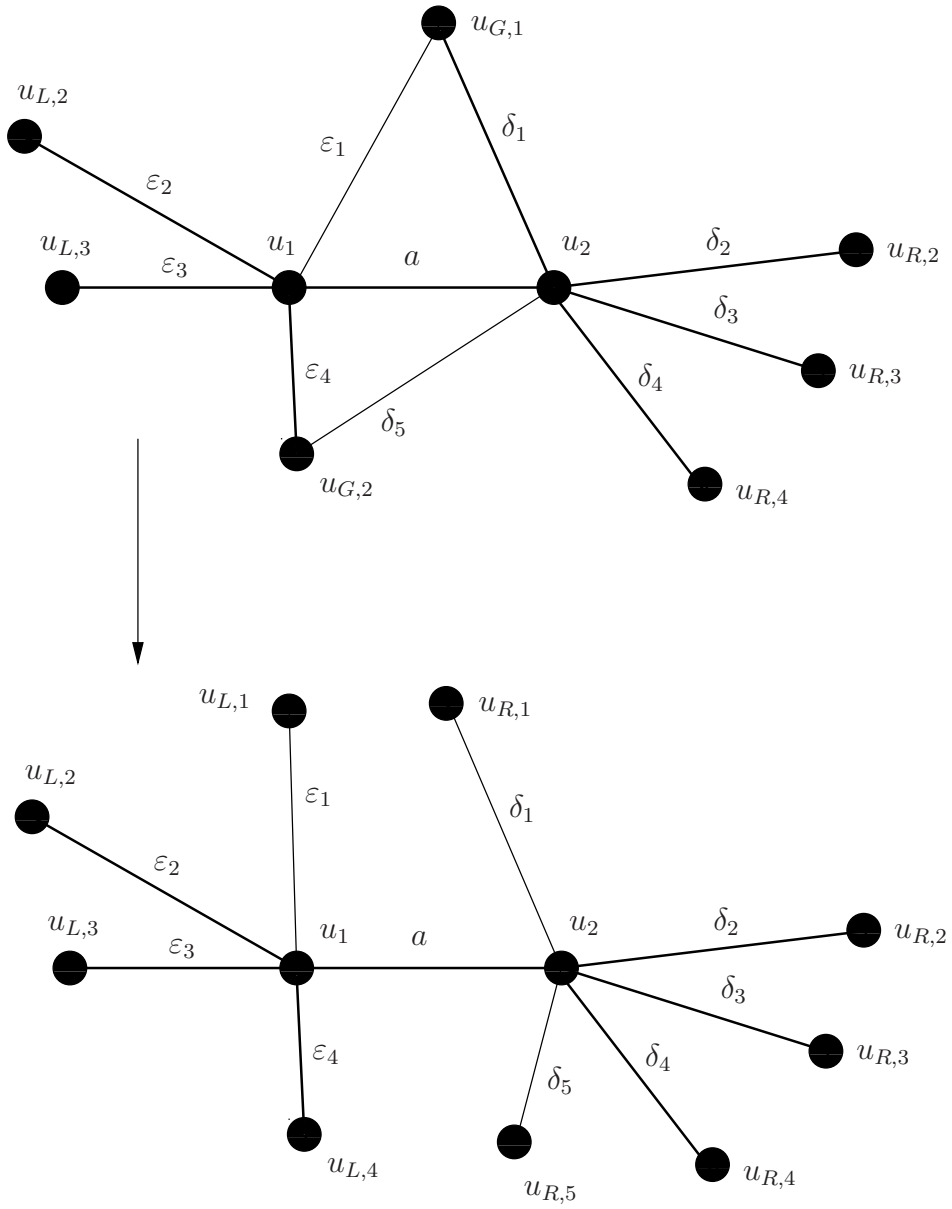


Figure 8.4: General system and the system to that it is cut open.

Lemma: 8.3.6. *Let A s.p.d. be a matrix as defined above. Let $\mathcal{N}_1^1, \mathcal{N}_2^1$ be two aggregated points with $a_{1,2} \neq 0$. If we assume that all neighbours are isolated points and the setting is as above then the inequality*

$$c_a \, \|u\|_A \geq \|Q_0 \, u\|_A$$

holds for all $u \in V$ with $c_a = \sqrt{1 + \frac{a_{1,1}+a_{2,2}-2|a_{1,2}|}{4|a_{1,2}|}}$

proof. For the situation based on the above mentioned sets we define

$$\|\widetilde{u}\|_A^2 = |a_{1,2}|(u_1 - u_2)^2 + \sum_{k \in I_1} |a_{1,k}|(u_1 - u_k)^2 + \sum_{k \in I_2} |a_{2,k}|(u_2 - u_k)^2 \tag{8.27}$$

$$+ r_1(u_1 - x)^2 + r_2(u_2 - x) + \left(\sum_{i,j \in I_1 \cup I_2} |a_{i,j}|(u_i - u_j)^2 \right)$$

$$\|\widetilde{Q_0 \, u}\|_A^2 = \sum_{k \in I_1} |a_{1,k}| \left(\frac{u_1 + u_2}{2} - u_k \right)^2 + \sum_{k \in I_2} |a_{2,k}| \left(\frac{u_1 + u_2}{2} - u_k \right)^2 \tag{8.28}$$

$$(r_1 + r_2) \left(\frac{u_1 + u_2}{2} - x \right)^2 + \left(\sum_{i,j \in I_1 \cup I_2} |a_{i,j}|(u_i - u_j)^2 \right).$$

with

$$r_i := a_{i,i} - \sum_{j=1, j \neq i}^{n} |a_{i,j}| \quad i = 1, 2.$$

As we have for $x = 0$

$$\|\widetilde{u}\|_A = \|u\|_A \quad \text{and} \quad \|\widetilde{Q_0 \, u}\|_A = \|Q_0 \, u\|_A$$

it follows

$$\sup_{u \neq 0} \frac{\|Q_0 u\|_A}{\|u\|_A} = \sup_{u \neq 0, \, x = 0} \frac{\|\widetilde{Q_0 u}\|_A}{\|\widetilde{u}\|_A} \leq \sup_{u \neq 0} \frac{\|\widetilde{Q_0 u}\|_A}{\|\widetilde{u}\|_A} \leq c_a.$$

The last equation is obtained by the proposition of Lemma 8.3.5 as for an interior point it is by the symmetry of A

$$a^L = a_{1,1} - |a_{1,2}| \quad \text{and} \quad a^R = a_{2,2} - |a_{1,2}|$$

and for the points of this lemma we obtain

$$a^L + r_1 = a_{1,1} - |a_{1,2}| \quad \text{and} \quad a^R + r_2 = a_{2,2} - |a_{1,2}|.$$

□

In the proof of Lemma 8.3.5 we have the system cut open and so induced new points and therewith degrees of freedom. In the proof of Lemma 8.3.6 we introduce a free point (and therewith one more degree of freedom) as we drop the condition $u = 0$ for points that belong to the boundary.

So we can summarize the results of this section in one central and global theorem. For this we assume that there is an arbitrary number of grid points that are aggregated in the coarser grid. We set for each pair $\mathcal{N}_i^1, \mathcal{N}_j^1$ of aggregated points

$$a^{i,j} := a_{i,i} - |a_{i,j}| \quad \text{and} \quad c_a^{i,j} := \sqrt{1 + \frac{a^{i,j} + a^{j,i}}{4|a_{i,j}|}}.$$

And for a given restriction operator R we set

$$Ind = \{(i,j) \,:\, \mathcal{N}_i^1, \mathcal{N}_j^1 \quad \text{are aggregated}\}.$$

Then Ind is the set of the aggregated points. So we can summarize the results as follows:

Theorem: 8.3.7. *Let A s.p.d. be a matrix as defined in 8.2 and the given grid. We assume that it is $a_{i,j} \neq 0$ for all $(i,j) \in Ind$ and all the neighbours of aggregated points are isolated points. Then*

$$c_a \, \|u\|_A \geq \|Q_0 \, u\|_A$$

holds for all $u \in V$ with $c_a = \max\{c_a^{i,j} \,:\, (i,j) \in Ind\}$.

proof. Let $\mathcal{N}_i^1, \mathcal{N}_j^1$ be two aggregated points. Then the inequality

$$c_a \, \|u\|_A \geq \|Q_0 \, u\|_A$$

holds locally for $\mathcal{N}_i^1, \mathcal{N}_j^1$ with $c_a \geq c_a^{i,j}$. The proof is completed by the fact that $\mathcal{N}_i^1, \mathcal{N}_j^1$ are arbitrary points. □

At least we mention that a coupling between two aggregated points is necessary to obtain a locall estimation for c_a. That means the condition $a_{i,j} \neq 0$ for two aggregated points is necessary. This is obvious if we set $a = 0$ in the most simple considered system of Lemma 8.3.1. Then it follows in the considered small sector

$$\|u\|_A^2 = \varepsilon (f - u_1)^2 + \delta (g - u_2)^2$$

$$\|Q_0 u\|_A^2 = \varepsilon \left(f - \frac{u_1 + u_2}{2} \right)^2 + \delta \left(g - \frac{u_1 + u_2}{2} \right)^2$$

If we set $u_1 = f = 1$ and $u_2 = g = -1$ then for $\varepsilon, \delta > 0$ is no estimation for c_a possible. In a two dimensional system that is non irreducible a proof could be obtained by using a link that is constructed with some points between the two aggregated points. Nevertheless the condition is in general not necessary therefore the preconditioner is well posed (non singular). This holds as we have proved the non singularity of the operator C_{DT}^{-1} in chapter 3 only by the condition that A, A_0 are non singular.

8.4 Technical view on the constant c_a. (One Dimension)

In section 8.3 we have given estimations for the constant c_a for a quite general geometrical situation. But therefore we had a strict restriction for the aggregation and for the neighbours of aggregated points, respectively. Now we will drop this assumption. For the sake of simplicity, we consider a one dimensional system.

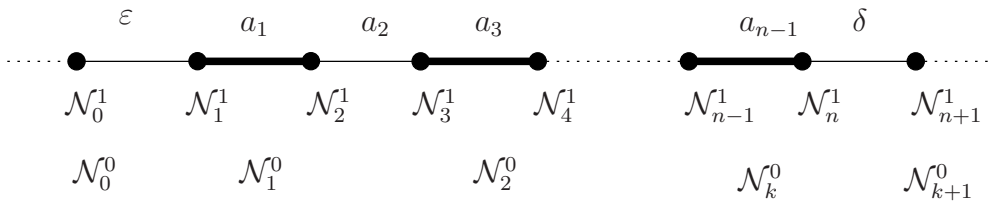


Figure 8.5: One dimensional system with an arbitrary situation for neighbours

We assume that the situation is given as in Figure 8.5. That means we have $n = 2k$ with $k \in \mathbb{N}$ points $\mathcal{N}_1^1, \ldots, \mathcal{N}_n^1$ that are pairwise aggregated to k new points $\mathcal{N}_1^0, \ldots, \mathcal{N}_k^0$. Moreover, these n points have a left and a right neighbour $\mathcal{N}_0^1, \mathcal{N}_{n+1}^1$ or $\mathcal{N}_0^0, \mathcal{N}_{k+1}^0$ respectively, that are isolated or belong to the boundary of Ω. Furthermore, the values of u are given by

$$u(\mathcal{N}_0^1) = u_L, \quad u(\mathcal{N}_{n+1}^1) = u_R \quad \text{and} \quad u(\mathcal{N}_i^1) = u_i, \quad \text{for} \quad i = 1, \ldots, n.$$

Then we obtain

$$Q_0 u = \left(u_L, \ \frac{u_1 + u_2}{2}, \ \frac{u_1 + u_2}{2} \ \frac{u_3 + u_4}{2}, \ \ldots, \ \frac{u_{n-1} + u_n}{2}, \ u_R \right).$$

Furhermore, for the sector of Ω the stiffness matrix is

$$A = \begin{pmatrix} \varepsilon & -\varepsilon & & & & & \\ -\varepsilon & \varepsilon + a_1 & -a_1 & & & & \\ & -a_1 & a_1 + a_2 & -a_2 & & & \\ & & \ddots & & & & \\ & & & & -a_{n-1} & a_{n-1} + \delta & -\delta \\ & & & & & -\delta & \delta \end{pmatrix}$$

with $\varepsilon, \delta, a_i > 0$. Therewith the expressions of our interest in the sector of Ω follow as

$$\|u\|_A^2 = \varepsilon(u_L - u_1)^2 + \sum_{i=1}^{n-1}(u_i - u_{i+1})^2 a_i + \delta(u_R - u_n)^2$$

$$\|Q_0 u\|_A^2 = \varepsilon \left(u_L - \frac{u_1 + u_2}{2}\right)^2 + \delta \left(u_R - \frac{u_{n-1} + u_n}{2}\right)^2 + \sum_{i=1}^{k-1} \left(\frac{u_{2i-1} + u_{2i}}{2} - \frac{u_{2i+1} + u_{2i+2}}{2}\right)^2 a_{2i}.$$

As we have done in the previous section we define a function $g := c_a^2 \|u\|_A^2 - \|Q_0 u\|_A^2$. We have to determine a c_a that fulfils $g \geq 0$ for all $u \in V$. We begin by minimizing the function $g(u_L, u_1, \ldots, u_n, u_R)$ with respect to the values u_L, u_R given on the isolated points. As already known it follows that the g minimizing values for u_L, u_R are given by

$$u_L = \frac{c_a^2 u_1 - (u_1 + u_2)/2}{c_a^2 - 1} \quad \text{and} \quad u_R = \frac{c_a^2 u_n - (u_{n-1} + u_n)/2}{c_a^2 - 1}.$$

If we insert these values in g then we obtain

$$g = -\frac{c_a^2}{(c_a^2 - 1)} \frac{\varepsilon}{4} (u_1 - u_2)^2 - \frac{c_a^2}{(c_a^2 - 1)} \frac{\delta}{4} (u_n - u_{n-1})^2 + c_a^2 \sum_{i=1}^{n-1}(u_i - u_{i+1})^2 a_i - \sum_{i=1}^{k-1} \left(\frac{u_{2i-1} + u_{2i}}{2} - \frac{u_{2i+1} + u_{2i+2}}{2}\right)^2 a_{2i}. \tag{8.29}$$

For an easier notation we transform the variables u_i in y_i with

$$y_i := (u_i - u_{i+1}), \quad \text{for} \quad i = 1, \ldots, n-1$$

$$y_n := u_n$$

(As the values only depends on the differences the setting $y_n = u_n$ is just for completeness. This will be obvious in some steps). Then we obtain

$$\begin{aligned}(u_{2i-1} + u_{2i} - u_{2i+1} - u_{2i+2})^2 &= ((u_{2i-1} - u_{2i+1}) + (u_{2i} - u_{2i+2}))^2 \\ &= ((y_{2i-1} + y_{2i}) + (y_{2i} + y_{2i+1}))^2 \\ &\leq 4y_{2i-1}^2 + 8y_{2i}^2 + 4y_{2i+1}^2. \end{aligned} \tag{8.30}$$

With g as given in (8.29) and the representation and estimation as given in (8.30)

$$\begin{aligned} g \geq &-\frac{c_a^2}{(c_a^2-1)}\frac{\varepsilon}{4}y_1^2 - \frac{c_a^2}{(c_a^2-1)}\frac{\delta}{4}y_{n-1}^2 - \left(a_2 y_1^2 + a_{n-2}y_{n-1}^2\right) + c_a^2\sum_{i=1}^{n-1} y_i^2 a_i \\ &- 2\sum_{i=1}^{k-1} a_{2i}y_{2i}^2 - \sum_{i=2}^{k-2} y_{2i-1}^2(a_{2i} + a_{2i-2}) \end{aligned} \tag{8.31}$$

holds. This inequality must hold for all $y_i,\ i = 1, \ldots, n-1$. We consider the variables $y_i,\ i = 1, \ldots, n-1$ separated:

1. y_1 : For y_1 we obatin

$$\begin{aligned} c_a^2 a_1 &\geq \frac{c_a^2}{(c_a^2-1)}\frac{\delta}{4} + a_2 \\ \Leftarrow \quad c_a^2 &\geq \frac{a_1 + a_2 + \varepsilon/4 + \sqrt{(a_1 + a_2 + \varepsilon/4)^2 - 4a_1a_2}}{2a_1}. \end{aligned}$$

2. y_{n-1} : Similary to y_1 it is sufficient

$$\begin{aligned} c_a^2 a_{n-1} &\geq \frac{c_a^2}{(c_a^2-1)}\frac{\delta}{4} + a_{n-2} \\ \Leftarrow \quad c_a^2 &\geq \frac{a_{n-1} + a_{n-2} + \delta/4 + \sqrt{(a_{n-1} + a_{n-2} + \delta/4)^2 - 4a_{n-1}a_{n-2}}}{2a_{n-1}}. \end{aligned}$$

3. y_i, $2 \leq i \leq n-2$, and i even: Then we obtain

$$c_a^2\, a_i \geq 2a_i \quad \Leftarrow c_a^2 \geq 2.$$

Hence

$$c_a^2 \geq 2$$

is sufficient for $g \geq 0$ for all y_i, $2 \leq i \leq n-2$, and i even.

4. y_i, $2 \leq i \leq n-2$, and i odd: For these constants we have

$$c_a^2\, a_i \geq a_{i-1} + a_{i+1} \quad \Leftarrow c_a^2 \geq \frac{a_{i-1} + a_{i+1}}{a_i}.$$

Hence

$$c_a^2 \geq \frac{a_{i-1} + a_{i+1}}{a_i}.$$

is sufficient for $g \geq 0$ for all y_i, $2 \leq i \leq n-2$, and i odd.

Therewith we can summarize the result for this situation in the following theorem:

Theorem: 8.4.1. *Let A be the matrix given by the structure of (8.4). We assume that there are n points $\mathcal{N}_1^1, \ldots, \mathcal{N}_n^1$ that are pairwise aggregated in V_0 and that the left and the right neighbours $\mathcal{N}_0^1, \mathcal{N}_{n+1}^1$ of this system are isolated points or belong to the boundary of Ω. The links are given as described above. Then the inequality*

$$\|Q_0\, u\|_A \leq c_a\, \|u\|_A$$

holds with

$$c_a^2 = \max\Bigg\{ \max\{(a_{i-1} + a_{i+1})/a_i \,:\, 2 \leq i \leq n-1, i \text{ odd}\}, 2,$$
$$\frac{a_1 + a_2 + \varepsilon/4 + \sqrt{(a_1 + a_2 + \varepsilon/4)^2 - 4a_1a_2}}{2a_1},$$
$$\frac{a_{n-1} + a_{n-2} + \delta/4 + \sqrt{(a_{n-1} + a_{n-2} + \delta/4)^2 - 4a_{n-1}a_{n-2}}}{2a_{n-1}} \Bigg\}.$$

proof. See the calculation above in this section. □

So as in the situation in section 8.3 the estimation depends on the ratio of the links between the points that are aggregated to the links they have with their neighbours. The easiest way to see this, is the restriction $c_a^2 \geq \frac{a_{i-1}+a_{i+1}}{a_i}$. In this equation a_i is the link between the points $\mathcal{N}_i^1, \mathcal{N}_{i+1}^1$ and they are aggregated. a_{i-1}, a_{i+1} are the links to their neighbours $\mathcal{N}_{i-1}^1, \mathcal{N}_{i+2}^1$.
At least we should highlight that the estimation given in Theorem 8.4.1 does not depend on the number k of aggregated pairs.

8.5 Two grid estimation for $C_{BPX}^{-1} A$ in the A-norm

We also want to estimate the condition of $C_{BPX}^{-1} A$ in the norm induced by A. So we remember the definition of C_{BPX}^{-1} by

$$C_{BPX}^{-1} f = A^{-1} f + P A_0^{-1} R f.$$

By the definition of u^* and u_0 as

$$u^* = A^{-1} f \quad \text{and} \quad u_0 = P A_0^{-1} R f$$

defined in (8.6), (8.8) we can write this for a given $f \in V$ as

$$C_{BPX}^{-1} f = u^* + u_0. \tag{8.32}$$

We also remember that it still holds

$$\|u_0\|_A \leq \|u^*\|_A$$

as proved in Lemma 8.2.3. For further estimations we need an estimation for $a(u^*, u_0)$. This is given in the next lemma.

Lemma: 8.5.1. *Let A be a s.p.d. matrix. For u^*, u_0 as defined in (8.6), (8.8) then*

$$a(u^*, u_0) = \|u_0\|_A^2 \quad \textit{holds.}$$

proof. By using the definitions of u_0, u^* we get

$$a(u^*, u_0) = (A A^{-1} f,\, P A_0^{-1} R f) = (R f,\, A_0^{-1} R f) = \|R f\|_{A_0^{-1}}$$

$$\begin{aligned} \text{and} \quad \|u_0\|_A &= (A P A_0^{-1} R f,\, P A_0^{-1} R f) \\ &= (A_0^{-1} R A P A_0^{-1} R f,\, R f) \\ &= (A_0^{-1} R f,\, R f) = \|R f\|_{A_0^{-1}}. \end{aligned}$$

This proves the proposition. □

This result is sufficient to prove a strong proposition for the condition of $C_{BPX}^{-1} A$ if we consider the operator in the norm induced by A.

Theorem: 8.5.2. *Let A be a non singular s.p.d. matrix. Then*

$$c_{BPX} \|v\|_A \leq \|C_{BPX}^{-1} A v\|_A \leq d_{BPX} \|v\|_A$$

holds for all $v \in V$ with

$$c_{BPX} = 1 \quad \textit{and} \quad d_{BPX} = 2.$$

proof. To prove

$$c_{BPX}\|v\|_A \leq \|C_{BPX}^{-1} A\, v\|_A \leq d_{BPX}\|v\|_A$$

for all $v \in V$ it is equivalent to set $v = A^{-1} f$ and prove

$$\begin{aligned} & c_{BPX}\|A^{-1} f\|_A \leq \|C_{BPX}^{-1} f\|_A \leq d_{BPX}\|A^{-1} f\|_A \\ \Leftrightarrow \quad & c_{BPX}\|u^*\|_A \leq \|u^* + u_0\|_A \leq d_{BPX}\|u^*\|_A. \end{aligned}$$

The second equivalence follows from the definition of u^*, u_0 and the representation (8.32). From the result of Lemma 8.5.1 follows

$$\begin{aligned} \|u^*\|_A &\leq \sqrt{\|u^*\|_A^2 + 3\|u_0\|_A^2} = \sqrt{\|u^*\|_A^2 + 2a(u_0,\, u^*) + \|u_0\|_A^2} \\ &= \sqrt{\|u^* + u_0\|_A^2} = \|u^* + u_0\|_A. \end{aligned}$$

This proves the proposition for c_{BPX}. On the other side, follows the assertion for d_{BPX} as we obtain from the same arguments

$$\|u^* + u_0\|_A = \sqrt{\|u^*\|_A^2 + 3\|u_0\|_A^2} \leq \sqrt{4\|u^*\|_A^2} = 2\|u^*\|_A.$$

□

So we see by using the $\|\,.\,\|_A$ norm that we can give for the BPX-preconditioner a strong estimation for the eigenvalues.

8.6 Multigrid estimation for $C_{BPX}^{-1} A$ in the A-norm

We will give an estimation for the condition of $C_{BPX}^{-1} A$ in the multigrid case. So we consider again the condition concerning the norm induced by A. We remember the multigrid definition of C_{BPX}^{-1} by

$$C_{BPX}^{-1} f = \sum_{j=0}^{J} P_j\, A_j^{-1}\, R_j\, f.$$

We keep the definition of u^* as $u^* = A^{-1} f$ and define for the same $f \in \mathbb{R}^n$ the vectors u_j by

$$u_j := P_j\, A_j^{-1}\, R_j\, f \quad \text{for} \quad j = 0, \ldots, J. \tag{8.33}$$

Therewith we obtain

$$C_{BPX}^{-1} f = \sum_{j=0}^{J} u_j$$

and further follows $u_J = u^*$. For further estimations we need an estimation for $a(u_i, u_j)$, for $i, j = 0, \ldots, J$. These are given in the next lemma.

Lemma: 8.6.1. *Let A be a s.p.d. matrix. For u_j as defined in (8.33) we have*

1. $a(u_j,\, v_j) = a(u^*,\, v_j), \quad \forall j = 0, \ldots, J,\ \forall v_j \in V_j$
2. $\|u_j\|_A \leq \|u^*\|_A, \quad \textit{for} \quad j = 0, \ldots, J.$
3. $a(u_i,\, u_j) = a(u_i,\, u_i) = \|u_i\|_A^2, \quad \textit{for} \quad i < j.$

proof. 1. For an arbitrary $j \in \{0, \ldots, J\}$ and an arbitrary $v_j \in V_i$ we obtain $\widehat{Q}_j\, v_j = P_j\, \widehat{S}_j\, R_j\, v_j = v_j$. So it follows for an arbitrary $v_j \in V_j$

$$\begin{aligned} a(u_j,\, v_j) &= a(u_j,\, P_j\, \widehat{S}_j\, R_j\, v_j) = a(P_j\, A_j^{-1}\, R_j\, f,\, P_j\, \widehat{S}_j\, R_j\, v_j) \\ &= (R_j\, A\, P_j\, A_j^{-1}\, R_j\, f,\, \widehat{S}_j\, R_j\, v_j) = (R_j\, f,\, \widehat{S}_j\, R_j\, v_j) \\ &= (f,\, \widehat{Q}_j\, v_j) = a(u^*,\, v_j). \end{aligned}$$

This shows the first proposition.

2. The second assertion follows the first one and the inequalitiy of Cauchy-Schwarz

$$\begin{aligned} & \|u_j\|_A^2 = a(u_j,\, u_j) = a(u^*,\, u_j) \leq \|u^*\|_A\, \|u_j\| \\ \Rightarrow \quad & \|u_j\|_A \leq \|u^*\|_A. \end{aligned}$$

3. The third proposition is obtained by the following two representations:

$$\begin{aligned} a(u_i,\, u_j) &= (A\, P_i\, A_i^{-1}\, R_i\, f,\, P_j\, A_j^{-1}\, R_j\, f) = (A_i^{-1}\, R_i\, f,\, R_i\, A\, P_j\, A_j^{-1}\, R_j\, f) \\ &= (A_i^{-1}\, R_i\, f,\, R_i^j\, R_j\, A\, P_j\, A_j^{-1}\, R_j\, f) = (A_i^{-1}\, R_i\, f, R_i^j\, R_j\, f) \\ &= (A_i^{-1}\, R_i\, f,\, R_i\, f) \\ \text{and} \quad a(u_i,\, u_i) &= (A\, P_i\, A_i^{-1}\, R_i\, f,\, P_i\, A_i^{-1}\, R_i\, f) = (A_i^{-1}\, R_i\, f, R_i\, A\, P_i\, A_i^{-1}\, R_i\, f) \\ &= (A_i^{-1}\, R_i\, f,\, R_i\, f). \end{aligned}$$

□

Therewith we can give the central result for the condition concerning the norm induced by A of the preconditiond system in the multilevel situation.

Theorem: 8.6.2. *Let $A \in \mathbb{R}^{n\times n}$ be s.p.d. Then*

$$c_{BPX}\|v\|_A \leq \|C_{BPX}^{-1} A\, v\|_A \leq d_{BPX}\|v\|_A$$

holds for all $v \in V$ with

$$c_{BPX} = 1 \quad \textit{and} \quad d_{BPX} = J + 1.$$

proof. To prove

$$c_{BPX}\|v\|_A \leq \|C_{BPX}^{-1} A\, v\|_A \leq d_{BPX}\|v\|_A$$

for all $v \in V$ it is equivalent to set $v = A^{-1}f$ and prove

$$c_{BPX}\|A^{-1} f\|_A \leq \|C_{BPX}^{-1} f\|_A \leq d_{BPX}\|A^{-1} f\|_A$$

$$\Leftrightarrow \quad c_{BPX}\|u^*\|_A \leq \left\|\sum_{j=0}^{J} u_j\right\|_A \leq d_{BPX}\|u^*\|_A.$$

From the results of Lemma 8.6.1 follows

$$\|u^*\|_A^2 = \|u_J\|_A^2 \leq \sum_{j=0}^{J} \|u_j\|_A^2 \leq \sum_{j=0}^{J}(2j+1)\|u_j\|_A^2 = \left\|\sum_{j=0}^{J} u_j\right\|_A^2.$$

This proves the proposition for c_{BPX}. On the other side, we obtain the proposition for d_{BPX} based on the same arguments and the estimation $\|u_i\|_A \leq \|u^*\|_A$ for all $i = 0, \ldots, J$. This implies

$$\|u_J + u_{J-1} + \cdots + u_0\|_A \leq \sum_{j=0}^{J} \|u_j\|_A \leq \sum_{j=0}^{J} \|u^*\|_A = (J+1)\|u^*\|_A.$$

□

Therefore, by using the $\|\,.\,\|_A$ norm we see that we can give for the BPX-preconditioner a strong estimation for the eigenvalues of $C_{BPX}^{-1} A$. In particular the estimations are independent of the elements of the matrix A. We should highlight that these results are as strong as the assumption included to use the inverse of A.

To understand better the structure of the BPX-method that is used in the Theorem 8.6.2 we can show the following calculation:

$$\begin{aligned}\|u^*\|_A &\leq \sqrt{\|u^*\|_A^2 + 3\|u_{J-1}\|_A^2 + 5\|u_{J-2}\|_A^2 + \cdots + (2J+1)\|u_0\|_A^2} \\ &= \Bigg[\|u_J\|_A^2 + \Big(\|u_{J-1}\|_A^2 + 2a(u_J,\, u_{J-1})\Big) + \Big(\|u_{J-2}\|_A^2 + 2a(u_J,\, u_{J-2}) + 2a(u_{J-1},\, u_{J-2})\Big) \\ &\quad + \cdots + \Big(\|u_0\|_A^2 + 2a(u_0,\, u_J) + \cdots + 2a(u_0,\, u_1)\Big)\Bigg]^{1/2} \\ &= \sqrt{\|u_J + u_{J-1} + \cdots + u_0\|_A^2} = \|u_J + u_{J-1} + \cdots + u_0\|_A.\end{aligned}$$

We can do the same for the other estimation if we use additional $\|u_i\|_A \leq \|u^*\|_A$. This shows that the estimations of Theorem 8.6.2 are exact if

$$u^* = u_J = \cdots = u_0$$

holds. This is more or less the problem for the BPX-method.

However to conclude this section we highlight that the result for the condition of $C_{BPX}^{-1} A$ in the norm $\|\,.\,\|_A$ is in general the same as in the optimal situation if we use the Euclidean norm.

8.7 Multigrid estimations for $C_{DT}^{-1} A$ in the A-norm

We will consider the multigrid situation for the estimation of the condition of $C_{DT}^{-1} A$ concerning the norm induced by A. In chapter 6 we have already seen that there are two possible multigrid preconditioners which are generalisations of C_{DT}^{-1} as defined for two grids. We will see that both have their own problems. We will present them both and point out the problems. Then we will show that if we use the aggregation method and we assume that the condition (2.14) holds the generalisations are equal and the additional problems we get for the multigrid situation are solved. As the definition of the constants is simpler in this case we will present the preconditioner we have introduced in chapter 6 as second version in this chapter first.

8.7.1 Generalisation of C_{DT}^{-1}. Version 2.

For a given $f \in \mathbb{R}^n$ we define $u^* \in \mathbb{R}^n$ by

$$A\, u^* := f, \quad \text{respectively} \quad u^* := A^{-1}\, f. \tag{8.34}$$

For the same f we define $u_{2,j} \in \mathbb{R}^n$ by

$$u_{2,j} := P_j\, A_j^{-1} (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j) R_j\, f, \quad \text{for} \quad j = 1, \ldots, J \tag{8.35}$$

$$\text{and} \quad u_{2,0} := P_0\, A_0^{-1}\, R_0\, f. \tag{8.36}$$

As it always holds $\widehat{S}_J = I_n$ and $\widehat{S}_{J-1} = S_{J-1}$ we obtain by these definitions that in the case of $J = 1$ this is the operator as used in section 8.2 as two grid operator. So we write

$$\begin{aligned} C_{DT,2}^{-1}\, f &= \sum_{j=0}^{J} u_{2,j} \\ &= \sum_{j=1}^{J} P_j\, A_j^{-1} (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j) R_j + P_0\, A_0^{-1}\, R_0. \end{aligned}$$

With these definitions we can show some properties of the elements $u_{2,j}$, $j = 0, \ldots, J$ as done in the two gird situation. Similary to the two grid situation, the elements $u_{2,j}$ can by interpreted as solutions of $A\, u = f$ in subspaces.

Lemma: 8.7.1. *Let $A \in \mathbb{R}^{n\times n}$ be s.p.d. For $u_{2,j}$, $j = 0, \ldots, J$ as defined in (8.35), (8.36) we have*

$$a(u_{2,j},\, v_j) = a(u^*,\, (\widehat{Q}_j - \widehat{Q}_{j-1})\, v_j) = (f,\, (\widehat{Q}_j - \widehat{Q}_{j-1})\, v_j) \tag{8.37}$$
$$\forall v_j \in V_j \quad \textit{and all} \quad \forall\, j = 1, \ldots, J$$

$$\textit{and} \quad a(u_{2,0},\, v_0) = a(u^*,\, v_0) = (f,\, v_0), \quad \forall v_0 \in V_0. \tag{8.38}$$

proof. As it holds $\widehat{Q}_j v_j = v_j$ for all $v_j \in V_j$ it follows from the definitions for an arbitrary $v_j \in V_j$

$$\begin{aligned}
a(u_{2,j},\, v_j) &= a(u_{2,j},\, P_j\, \widehat{S}_j\, R_j\, v_j) \\
&= (A\, P_j\, A_j^{-1}\, (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j}) R_j\, f,\ P_j\, \widehat{S}_j\, R_j\, v_j) \\
&= (R_j\, A\, P_j\, A_j^{-1}\, (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j}) R_j\, f,\ \widehat{S}_j\, R_j\, v_j) \\
&= ((I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j}) R_j\, f,\ \widehat{S}_j\, R_j\, v_j) \\
&= (f,\ P_j (I_j - P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j} \widehat{S}_j^{-1}) \widehat{S}_j\, R_j\, v_j) \\
&= (f,\ (P_j\, \widehat{S}_j\, R_j - P_{j-1}\, \widehat{S}_{j-1}\, R_{j-1})\, v_j) \\
&= (f,\ (\widehat{Q}_j - \widehat{Q}_{j-1})\, v) = a(u^*,\, (\widehat{Q}_j - \widehat{Q}_{j-1})\, v).
\end{aligned}$$

This shows the first assertion. For the second we go through the same steps and we obtain

$$\begin{aligned}
a(u_{2,0},\, v_0) &= (A\, P_0\, A_0^{-1}\, R_0\, f,\ P_0\, \widehat{S}_0\, R_0\, v_0) \\
&= (R_0\, f,\ \widehat{S}_0\, R_0\, v_0) \\
&= (f,\, \widehat{Q}_0\, v_0) = (f,\, v_0) = a(u^*,\, v_0).
\end{aligned}$$

□

Remark: 8.7.2. *Unfortunately*

$$a(u_{2,i},\, u_{2,j}) = 0 \quad \textit{for } i \neq j$$

does not hold in this setting. This can be seen if we consider for $i < j$ the following calculation:

$$\begin{aligned}
a(u_{2,j},\, u_{2,i}) &= \big(A\, P_j\, A_j^{-1}\, (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j}) R_j\, f,\ P_i\, A_i^{-1}\, (I_i - \widehat{S}_i^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j}) R_i\, f\big) \\
&= \big(R_j^i\, (I_j - \widehat{S}_j^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j}) R_j\, f,\ A_i^{-1}\, (I_i - \widehat{S}_i^{-1}\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^{j}) R_i\, f\big)
\end{aligned}$$

As it is not necessarily $(I_j - \widehat{S}_j^{-1} P_j^{j-1} \widehat{S}_{j-1} R_{j-1}^j) R_j f \in \widetilde{W}_j = ker(R_{j-1}^j)$. *So the inner product above is in general unequal zero.*

For further estimation between $u_{2,0} + \cdots + u_{2,J}$ and u^* we define the constants $c_{a,2,j}, c_{G,j}$ as follows

$$c_{a,2,j} := \sup \left\{ \frac{\|\widehat{Q}_{j-1} v_j\|_A}{\|v_j\|_A} \ : \ v_j \in V_j \backslash \{0\} \right\}, \quad \text{for} \quad j = 1, \ldots, J \tag{8.39}$$

$$c_{G,j} := \sup \left\{ \frac{\|\widehat{Q}_j v\|_A}{\|v\|_A} \ : \ v \in V \backslash \{0\} \right\}, \quad \text{for} \quad j = 0, \ldots, J. \tag{8.40}$$

By the definiton of $c_{a,2,j}$ it is obvious that these constants hold the inequality

$$\|\widehat{Q}_{j-1} v_j\|_A \leq c_{a,2,j} \, \|v_j\|_A$$

for all $v_j \in V_j$. Moreover, we can represent the expressions as follows:

$$\begin{aligned} \|\widehat{Q}_{j-1} v_j\|_A^2 &= (A \, P_{j-1} \widehat{S}_{j-1} R_{j-1} v_j, \ P_{j-1} \widehat{S}_{j-1} R_{j-1} v_j) \\ &= (A_j \, P_j^{j-1} \widehat{S}_{j-1} R_{j-1} v_j, \ P_j^{j-1} \widehat{S}_{j-1} R_{j-1} v_j) \\ \|v_j\|_A^2 = \|\widehat{Q}_j v_j\|_A^2 &= (A \, P_j \widehat{S}_j R_j v_j, \ P_j \widehat{S}_j R_j v_j) \\ &= (A_j \widehat{S}_j R_j v_j, \ \widehat{S}_j R_j v_j). \end{aligned} \tag{8.41}$$

Similary to the constant c_a in the two grid situation, the constant $c_{a,2,j}$ depends on the matrix A_j and the relation of the subspaces $P_j^{j-1} \widehat{S}_{j-1} R_{j-1}^j(\widetilde{V}_j), \widetilde{V}_j$ to each other. Based on these constants we get the following estimations:

Lemma: 8.7.3. *Let* $A \in \mathbb{R}^{n \times n}$ *be s.p.d. For* $u_{2,j}$, $j = 0, \ldots, J$ *and* u^* *we have*

$$\|u_{2,0}\|_A \leq \|u^*\|_A \tag{8.42}$$

$$\|u_{2,j}\|_A \leq (c_{a,2,j} + 1) \|u^*\|_A, \quad \textit{for } j = 1, \ldots, J. \tag{8.43}$$

$$\|u^*\|_A \leq \sum_{j=0}^{J} c_{G,j} \|u_{2,j}\|_A. \tag{8.44}$$

proof. As it is $u_{2,0} \in V_0$ it follows from Lemma 8.7.1

$$\begin{aligned} &\|u_{2,0}\|_A^2 = a(u_{2,0}, \, u_{2,0}) = a(u^*, \, u_{2,0}) \leq \|u^*\|_A \, \|u_{2,0}\|_A \\ \Rightarrow \quad &\|u_{2,0}\|_A \leq \|u^*\|_A. \end{aligned}$$

The second proposition also results from Lemma 8.7.1. As we have $u_{2,j} \in V_j$ it follows with the definition of $c_{a,2,j}$

$$\begin{aligned}
\|u_{2,j}\|_A^2 = a(u_{2,j},\, u_{2,j}) &= a(u^*,\, (\widehat{Q}_j - \widehat{Q}_{j-1})u_{2,j}) \\
&\leq \|u^*\|_A \, \|(\widehat{Q}_j - \widehat{Q}_{j-1})u_{2,j}\|_A \\
&\leq \|u^*\|_A (\|\widehat{Q}_j u_{2,j}\|_A + \|\widehat{Q}_{j-1}u_{2,j}\|_A) \\
&\leq \|u^*\|_A \, \|u_{2,j}\|_A (1 + c_{a,2,j}) \\
\Rightarrow \quad \|u_{2,j}\|_A &\leq (c_{a,2,j} + 1)\|u^*\|_A.
\end{aligned}$$

This shows the second proposition. For the third we decompose u^* as

$$u^* = \sum_{j=1}^{J} (\widehat{Q}_j - \widehat{Q}_{j-1})u^* + \widehat{Q}_0 u^*.$$

Then we obtain

$$\begin{aligned}
(\widehat{Q}_j - \widehat{Q}_{j-1})v = (\widehat{Q}_j - \widehat{Q}_{j-1})\widehat{Q}_j\, v, \quad \forall v \in V \\
\text{and} \quad \widehat{Q}_j\, v \in V_j
\end{aligned}$$

for all $v \in V$. This implies with Lemma 8.7.1

$$\begin{aligned}
\|u^*\|_A^2 = a(u^*,\, u^*) &= a\left(u^*,\, \sum_{j=1}^{J} (\widehat{Q}_j - \widehat{Q}_{j-1})u^* + \widehat{Q}_0 u^*\right) \\
&= \sum_{j=1}^{J} a(u^*,\, (\widehat{Q}_j - \widehat{Q}_{j-1})u^*) + a(u^*,\, \widehat{Q}_0 u^*) \\
&= \sum_{j=1}^{J} a(u^*,\, (\widehat{Q}_j - \widehat{Q}_{j-1})\, \widehat{Q}_j\, u^*) + a(u^*,\, \widehat{Q}_0 u^*) \\
&= \sum_{j=0}^{J} a(u_{2,j},\, \widehat{Q}_j u^*) \leq \sum_{j=0}^{J} \|u_{2,j}\|_A \, \|\widehat{Q}_j\, u^*\|_A \leq \sum_{j=0}^{J} c_{G,j}\|u_{2,j}\|_A \, \|u^*\|_A \\
\Rightarrow \quad \|u^*\|_A &\leq \sum_{j=0}^{J} c_{G,j}\|u_{2,j}\|_A
\end{aligned}$$

□

Furthermore we define a constant K_2 that holds

$$\sum_{j=0}^{J} \|u_{2,j}\|_A \leq K_2 \left\| \sum_{j=0}^{J} u_{2,j} \right\|_A . \tag{8.45}$$

The problem for further estimations is that until there is no knowlege about the angles $\gamma_{i,j} < 1$ that holds

$$a(u_{2,i},\, u_{2,j}) \leq \gamma_{i,j} \, \|u_{2,i}\|_A \, \|u_{2,j}\|_A$$

we can not give any estimation for K_2.
If such an K_2 existed, then we could give an estimation for the condition of $C_{DT,2}^{-1} A$ that only depends on the constants $c_{a,2,j}, c_{G,j}$ and K_2.

Proposition: 8.7.4. *Let A be a s.p.d. matrix and assume that the inequality (8.45) holds with K_2. With $c_{a,2,j}, c_{G,j}$ as defined in (8.39) and (8.40)*

$$c_{DT,2}\|v\|_A \leq \|C_{DT,2}^{-1} A\, v\|_A \leq d_{DT,2}\|v\|_A$$

holds for all $v \in V$ with

$$c_{DT,2} = \frac{1}{\max_{j=0,\ldots,J} c_{G,j} K_2} \quad \textit{and} \quad d_{DT,2} = (J+1) + \sum_{j=1}^{J} c_{a,2,j}.$$

proof. To prove

$$c_{DT,2}\|v\|_A \leq \|C_{DT,2}^{-1} A\, v\|_A \leq d_{DT,2}\|v\|_A$$

for all $v \in V$ it is equivalent to set $v = A^{-1} f$ and prove

$$c_{DT,2}\|A^{-1} f\|_A \leq \|C_{DT,2}^{-1} f\|_A \leq d_{DT,2}\|A^{-1} f\|_A$$

$$\Leftrightarrow \quad c_{DT,2}\|u^*\|_A \leq \|u_{2,0} + \cdots + u_{2,J}\|_A \leq d_{DT,2}\|u^*\|_A.$$

From the estimations of Lemma 8.7.3 we obtain

$$\|u_{2,0} + \cdots + u_{2,J}\|_A \leq \|u_{2,0}\|_A + \cdots + \|u_{2,J}\|_A \leq \left((J+1) + \sum_{j=1}^{J} c_{a,2,j} \right) \|u^*\|_A.$$

This proves the propostion for $d_{DT,2}$. Again from Lemma 8.7.3 by assuming that there is a K_2 that fulfils the inequality (8.45) it follows

$$\|u^*\|_A \leq \sum_{j=0}^{J} c_{G,j}\|u_{2,j}\|_A \leq \max_{i=0,\ldots,J} c_{G,i} \left(\sum_{j=0}^{J} \|u_{2,j}\|_A \right)$$

$$\leq \max_{i=0,\ldots,J} c_{G,i} K_2 \|u_{2,0} + \cdots + u_{2,J}\|_A.$$

This proves the estimation for $c_{DT,2}$. □

A short discussion on the constants: In section 8.2 we have highlighted that the constant c_a does only depend on the elements of the matrix A and the structure of the subspace V_0. The calculation (8.41) shows that the constants $c_{a,2,j}$ only depend on the elements of A_j and the structure of $P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j(\widetilde{V}_j), \widetilde{V}_j$ in relation to each other. One of the problems is that in general we do not have

$$P_j^{j-1}\,\widehat{S}_{j-1}\,R_{j-1}^j\widetilde{v}_j \in P_j^{j-1}(\widetilde{V}_{j-1}).$$

If we use the aggregation method, then the inclusion above only holds if the condition (2.14) is fulfilled. The constant $c_{G,j}$ depends on the elements of A and the structure of V_j in relation to V. Hence $c_{G,j}$ is not as easy to handle as c_a.

Moreover, the definitions imply immediately

$$c_{a,2,J} = c_a = c_{G,J-1} \quad \text{and} \quad c_{G,J} = 1. \tag{8.46}$$

Now we can compare the result of Proposition 8.7.4 with the result of Theorem 8.2.4 that holds in the two grid situation:
In Proposition 8.7.4 we obtain from equation (8.46) in the case of $J = 1$

$$c_{DT,2} = \frac{1}{c_a\,K_2} \quad \text{and} \quad d_{DT,2} = 2 + c_a.$$

And in the two gird situation considered in Theorem 8.2.4 we obtain

$$c_{DT} = \frac{1}{c_a\,\sqrt{2}} \quad \text{and} \quad d_{DT} = 2 + c_a.$$

From Lemma A.0.5 follows $K_2 = \sqrt{2}$ if we add two orthogonal vectors. Thus the estimations for $d_{DT}, d_{DT,2}$ and $c_{DT}, c_{DT,2}$ are the same.
This shows that the results of this section are a generalisation of the results we have in the two grid situation.

8.7.2 Generalisation of C_{DT}^{-1}. Version 1.

For $f \in \mathbb{R}^n$ we take over the definition of $u^* \in \mathbb{R}^n$ by

$$A\,u^* := f \quad \text{and} \quad u^* := A^{-1}\,f \quad \text{respectively.} \tag{8.47}$$

For the same $f \in \mathbb{R}^n$ we define $u_{1,j} \in \mathbb{R}^n$ by

$$u_{1,j} := P_j\,A_j^{-1}(I_j - Q_{j-1})R_j\,f, \quad \text{for} \quad j = 1, \ldots, J \tag{8.48}$$

$$\text{and} \quad u_{1,0} := P_0\,A_0^{-1}\,R_0\,f. \tag{8.49}$$

Based on these definitions it is obvious that in the case of $J = 1$ this is also the operator as used in section 8.2. Furthermore, it is obvious that we obtain $u_{1,0} = u_{2,0}$. We write

$$\begin{aligned} C_{DT,1}^{-1} f &= \sum_{j=0}^{J} u_{1,j} \\ &= \sum_{j=1}^{J} P_j A_j^{-1} (I_j - Q_{j-1}) R_j + P_0 A_0^{-1} R_0. \end{aligned}$$

In this case we obtain the following characteristics that are similar to the properties proved in Lemma 8.7.1 for the situation of $C_{DT,2}^{-1}$.

Lemma: 8.7.5. *Let $A \in \mathbb{R}^{n \times n}$ be s.p.d. For $u_{1,j}$, $j = 0, \ldots, J$ as defined in (8.48), (8.49) we have:*

$$\begin{aligned} a(u_{1,j},\, v_j) &= a(u^*,\, P_j (I - Q_{j-1}) \widehat{S}_j\, R_j\, v_j) \\ &= (f,\, P_j (I - Q_{j-1}) \widehat{S}_j\, R_j\, v_j) \quad \forall v_j \in V_j, \quad j = 1, \ldots, J \end{aligned} \tag{8.50}$$

$$a(u_{1,0},\, v_0) = a(u^*,\, v_0) = (f,\, v_0), \quad \forall v_0 \in V_0 \tag{8.51}$$

$$\text{and} \quad a(u_{1,i},\, u_{1,j}) = 0, \quad \forall i \neq j. \tag{8.52}$$

proof. For an arbitrary $j \in \{1, \ldots, J\}$ we have $Q_{j-1} = Q_{j-1}^T$ and also $I_j - Q_{j-1} = (I_j - Q_{j-1})^T$. Furthermore, for an arbitrary $v_j \in V_j$ it is $\widehat{Q}_j\, v_j = P_j\, \widehat{S}_j\, R_j\, v_j = v_j$. The first assertion follows from

$$\begin{aligned} a(u_{1,j},\, v_j) &= a(u_{1,j},\, \widehat{Q}_j\, v_j) \\ &= (A\, P_j\, A_j^{-1}\, (I_j - Q_{j-1}) R_j\, f,\; P_j\, \widehat{S}_j\, R_j\, v_j) \\ &= (R_j\, A\, P_j\, A_j^{-1}\, (I_j - Q_{j-1}) R_j\, f,\; \widehat{S}_j\, R_j\, v_j) \\ &= ((I_j - Q_{j-1}) R_j\, f,\; \widehat{S}_j\, R_j\, v_j) \\ &= (f,\; P_j\, (I - Q_{j-1})\, \widehat{S}_j\, R_j\, v_j) \\ &= a(u^*,\; P_j\, (I_j - Q_{j-1})\, \widehat{S}_j\, R_j\, v_j). \end{aligned}$$

The second assertion follows as it holds $u_{1,0} = u_{2,0}$ and the is the same as the equation (8.38) in Lemma 8.7.1. For the third proposition we consider first $i, j \geq 1$ with w.l.o.g.

$i < j$. Then we obtain

$$\begin{aligned} a(u_{1,i},\, u_{1,j}) &= (P_i\, A_i^{-1}\,(I - Q_{j-1})R_i\, f,\; A\, P_j\, A_j^{-1}\,(I - Q_{j-1})R_j\, f) \\ &= (A_i^{-1}\,(I - Q_{j-1})R_i\, f,\; R_i^j\, R_j\, A\, P_j\, A_j^{-1}\,(I - P_j^{j-1}\, S_{j-1}\, R_{j-1}^j)R_j\, f) \\ &= (A_i^{-1}\,(I - Q_{j-1})R_i\, f,\; R_i^j (I - P_j^{j-1}\, S_{j-1}\, R_{j-1}^j)R_j\, f) \\ &= (A_i^{-1}\,(I - Q_{j-1})R_i\, f,\; R_i^{j-1}(R_{j-1}^j - R_{j-1}^j\, P_j^{j-1}\, S_{j-1}\, R_{j-1}^j)R_j\, f) \\ &= (A_i^{-1}\,(I - Q_{j-1})R_i\, f,\; R_i^{j-1}(R_{j-1}^j - R_{j-1}^j)R_j\, f) = 0. \end{aligned}$$

If it is $i < j$ and $i = 0$ then the proposition follows the same way by

$$\begin{aligned} a(u_{1,0},\, u_{1,j}) &= (P_0\, A_0^{-1}\, R_0\, f,\; A\, P_j\, A_j^{-1}\,(I - Q_{j-1})R_j\, f) \\ &= (A_0^{-1}\, R_0\, f,\; R_0\, A\, P_j\, A_j^{-1}\,(I - Q_{j-1})R_j\, f) \\ &= (A_0^{-1}\, R_0\, f,\; R_0^j\,(I - Q_{j-1})R_j\, f) = 0. \end{aligned}$$

This proves the third proposition. □

Remark: 8.7.6. *For the equation (8.50) we can also write*

$$\begin{aligned} a(u_{1,j},\, v_j) &= a(u^*,\; P_j(I_j - Q_{j-1})\widehat{S}_j\, R_j\, v_j) \\ &= a(u^*,\; \widehat{Q}_j - P_j\, Q_{j-1}\, \widehat{S}_j\, R_j\, v_j), \quad \forall v_j \in V_j \quad \textit{and all} \quad j = 1, \ldots, J. \end{aligned}$$

This representation clearly shows the problem of the operator. In general we have $P_j\, Q_{j-1}\, \widehat{S}_j\, R_j \neq \widehat{Q}_{j-1}$. *Therewith it is*

$$v \neq \sum_{j=1}^{J} P_j(I_j - Q_{j-1})\widehat{S}_j\, R_j\, v + P_0\, \widehat{S}_0\, R_0\, v$$

and we can not decompose u *as done in the two grid case or by using* $C_{DT,1}^{-1}$.

We define again some constants that are for this generalisation of the two grid case the generalisation of c_a. So we define for $j = 1, \ldots, J$ the constants $c_{a,1,j}, K_1$ by

$$c_{a,1,j} := \sup \left\{ \frac{\|Q_{j-1}\widetilde{v}_j\|_{A_j}}{\|\widetilde{v}_j\|_{A_j}} \;:\; \widetilde{v}_j \in \widetilde{V}_j \backslash \{0\} \right\} \tag{8.53}$$

$$K_1 := \sup \left\{ \frac{a(v,\, v)}{a\left(v,\; \sum_{j=1}^{J} P_j(I_j - Q_{j-1})\, \widehat{S}_j\, R_j\, v + P_0\, \widehat{S}_0\, R_0 v\right)} \;:\; v \in V \backslash \{0\} \right\}. \tag{8.54}$$

From the definiton of $c_{a,1,j}$ it is again given that these constants hold the inequality

$$\|Q_{j-1}\widetilde{v}_j\|_{A_j} \leq c_{a,1,j} \, \|\widetilde{v}_j\|_{A_j}$$

for all $\widetilde{v}_j \in \widetilde{V}_j$. Furthermore we obtain

$$\begin{aligned} \|P_j \, Q_{j-1} \, \widehat{S}_j \, R_j \, v_j\|_A^2 &= (A \, P_j \, Q_{j-1} \, \widehat{S}_j \, R_j \, v_j, \; P_j \, Q_{j-1} \, \widehat{S}_j \, R_j \, v_j) \\ &= (A_j \, Q_{j-1} \, \widehat{S}_j \, R_j \, v_j, \; Q_{j-1} \, \widehat{S}_j \, R_j \, v_j) \\ \|v_j\|_A^2 = \|\widehat{Q}_j \, v_j\|_A^2 &= (A \, P_j \, \widehat{S}_j \, R_j \, v_j, \; P_j \, \widehat{S}_j \, R_j \, v_j) \\ &= (A_j \, \widehat{S}_j \, R_j \, v_j, \; \widehat{S}_j \, R_j \, v_j). \end{aligned}$$

As $\widehat{S}_j \, R_j : V_j \to \widetilde{V}_j$ is bijective we obtain for a constant $c_{a,1,j}$

$$\begin{aligned} & \|Q_{j-1}\widetilde{v}_j\|_{A_j} \leq c_{a,1,j} \, \|\widetilde{v}_j\|_{A_j} \quad \forall \widetilde{v}_j \in \widetilde{V}_j \\ \Leftrightarrow \quad & \|P_j \, Q_{j-1} \, \widehat{S}_j \, R_j \, v_j\|_A^2 \leq c_{a,1,j} \|v_j\|_A \quad \forall v_j \in V_j. \end{aligned}$$

Furthermore we highlight that for $c_{a,1,j}$ the same situation as for c_a in the two grid situation remains. We will discuss this more in-depth in section 8.7.3.

By these characteristics we can show some estimations for isolated elements $u_{1,j}$:

Lemma: 8.7.7. *Let A be a s.p.d. matrix. For $u_{1,j}$, $j = 0, \ldots, J$ and u^* we have*

$$\|u_{1,0}\|_A \leq \|u^*\|_A \tag{8.55}$$

$$\|u_{1,j}\|_A \leq (1 + c_{a,1,j}) \|u^*\|_A \tag{8.56}$$

$$\|u^*\|_A \leq K_1 \max_{j=0,\ldots,J} c_{G,j} \sqrt{J+1} \|u_{1,0} + \cdots + u_{1,J}\|_A. \tag{8.57}$$

proof. As the first proposition is proved for $u_{2,0}$ in Lemma 8.7.3 the first propostion holds again by $u_{1,0} = u_{2,0}$. The second proposition follows by $u_{1,j} \in V_j$ and Lemma 8.7.5 with

$$\begin{aligned} \|u_{1,j}\|_A^2 = a(u_{1,j}, \, u_{1,j}) &= a(u^*, \, P_j(I - P_j^{j-1} \, S_{j-1} \, R_{j-1}^j)\widehat{S}_j \, R_j \, u_{1,j}) \\ &= a(u^*, \, (\widehat{Q}_j - P_{j-1} \, S_{j-1} \, R_{j-1}^j \, \widehat{S}_j \, R_j) \, u_{1,j}) \\ &\leq \|u^*\|_A (\|\widehat{Q}_j u_{1,j}\|_A + \|P_{j-1} \, S_{j-1} \, R_{j-1}^j \, \widehat{S}_j \, R_j \, u_{1,j}\|_A) \\ &\leq \|u^*\|_A (\|u_{1,j}\|_A + c_{a,1,j} \|u_{1,j}\|_A) \\ \Rightarrow \quad \|u_{1,j}\|_A &\leq (1 + c_{a,1,j}) \|u^*\|_A. \end{aligned}$$

For the third proposition we use the definition of K_1. The third proposition of this lemma follows from the calculation

$$\begin{aligned}
\|u^*\|_A^2 &= a(u^*,\, u^*) \\
&\leq K_1\, a\left(u^*,\, \sum_{j=1}^{J} P_j\,(I_j - Q_{j-1})\widehat{S}_j\, R_j\, u^* + P_0\,\widehat{S}_0\, R_0\, u^*\right) \\
&= K_1 \sum_{j=1} a(u^*,\; P_j\,(I_j - Q_{j-1})\widehat{S}_j\, R_j\, u^*) + a(u^*, P_0\,\widehat{S}_0\, R_0\, u^*) \\
&= K_1 \sum_{j=1} a(u^*,\; P_j\,(I_j - Q_{j-1})\widehat{S}_j\, \underbrace{R_j\, P_j\, \widehat{S}_j}_{=I_j}\, R_j\, u^*) + a(u^*, P_0\,\widehat{S}_0\, R_0\, u^*) \\
&= K_1 \sum_{j=1} a(u^*,\; P_j\,(I_j - Q_{j-1})\widehat{S}_j\, R_j\, \underbrace{P_j\, \widehat{S}_j\, R_j\, u^*}_{\in V_j}) + a(u^*, P_0\,\widehat{S}_0\, R_0\, u^*) \\
&= K_1 \sum_{j=0}^{J} a(u_{1,j}, \widehat{Q}_j\, u^*) \leq K_1 \sum_{j=0}^{J} \|u_{1,j}\|_A\, \|\widehat{Q}_j\, u^*\|_A \\
&\leq K_1 \sum_{j=0}^{J} c_{G,j} \|u_{1,j}\|_A\, \|u^*\|_A \\
\Rightarrow \quad \|u^*\|_A &\leq K_1 \sum_{j=0}^{J} c_{G,j}\, \|u_{1,j}\|_A.
\end{aligned}$$

From the orthogonality of $u_{1,i}, u_{1,j}$ for $i \neq j$ as shown in Lemma 8.7.5, we obtain with Lemma A.0.5

$$K_1 \sum_{j=0}^{J} c_{G,j}\, \|u_{1,j}\|_A \leq \sqrt{J+1}\, K_1 \max_{j=0,\ldots,J} c_{G,j} \|u_{1,0} + \cdots + u_{1,J}\|_A.$$

□

In this setting we can give an estimation for the condition of $C_{DT,1}^{-1} A$.

Proposition: 8.7.8. *Let $A \in \mathbb{R}^{n\times n}$ be s.p.d. With $c_{a,1,j}, c_{G,j}$ and K_1 as defined in (8.53),(8.40) and (8.54) then*

$$c_{DT,1}\|v\|_A \leq \|C_{DT,1}^{-1}\, A\, v\|_A \leq d_{DT,1}\|v\|_A$$

holds for all $v \in V$ with

$$c_{DT,1} = \frac{1}{\sqrt{J+1}\, K_1 \, \max_{j=0,\dots,J} c_{G,j}} \quad \textit{and} \quad d_{DT,1} = (J+1) + \sum_{j=1}^{J} c_{a,1,j}.$$

proof. We prove again that it is

$$c_{DT,1}\|u^*\|_A \leq \|u_{1,0} + \cdots + u_{1,J}\|_A \leq d_{DT,1}\|u^*\|_A.$$

From the estimations of Lemma 8.7.7 we obtain

$$\|u_{1,0} + \cdots + u_{1,J}\|_A \leq \|u_{1,0}\|_A + \cdots + \|u_{1,J}\|_A \leq \left((J+1) + \sum_{j=1}^{J} c_{a,1,j} \right) \|u^*\|_A.$$

This proves the assertion for $d_{DT,1}$. The proposition for $c_{DT,1}$ is still proved by the proof of (8.57) in Lemma 8.7.7. □

A short discussion on the constants: As already mentioned, the constant c_a of section 8.2 only depends on the elements of A and the structure of V_0. As shown in section 8.7.1 the constants $c_{a,2,j}, c_{G,j}$ do also depend on the elements of A_j and on a relation of spaces that is not as easy to handle as the relation of V_0 to V. For $c_{a,1,j}$ it is obvious that this constant depends on the elemets of A_j and the structure of $\widetilde{V}_j, P_j^{j-1}(\widetilde{V}_{j-1})$. Hence we have for $c_{a,1,j}$ the same relation of spaces as for c_a.

Furthermore, we also obtain

$$c_{a,1,J} = c_a = c_{G,J-1} \quad \text{and} \quad c_{G,J} = 1.$$

As the equation $Q_{J-1} = \widehat{Q}_{J-1}$ holds independently of the number J of levels, it follows that

$$K_1 = 1$$

for the two grid situation. Now we can again compare the result of Proposition 8.7.8 with the result of Theorem 8.2.4 that holds in the two grid situation:

For $J = 1$, we obtain $c_{a,1,1} = c_a$ from the relations above. Thus it follows that $d_{DT} = d_{DT,1}$.
For $J = 1$, the estimation for c_{DT} and $c_{DT,1}$, is also the same. Therefore, the estimations we gave for $C_{DT}^{-1} A$ are the same as in the two grid situation.

Furthermore, the characteristics $\widehat{S}_J = I_J$ and $Q_{J-1} = \widehat{Q}_{J-1}$ imply that both multigrid versions are the same in the two grid situation. So they are both generalisations of the same two grid structure.

In addition we should highlight that in general it is not possible to replace the global constant K_1 by constants $k_{1,j}$ that would fulfil

$$a(v, (\widehat{Q}_j - \widehat{Q}_{j-1})v) \leq k_{1,j} a(v, P_j(I_j - Q_{j-1})\,\widehat{S}_j\,R_j\,v). \tag{8.58}$$

This is impossible as in general we have

$$\begin{aligned} & ker(R_{j-1}) \neq ker(R^j_{j-1}\,\widehat{S}_j\,R_j) \\ \Rightarrow \quad & ker(\widehat{Q}_{j-1}) \neq ker(P_j\,Q_{j-1}\,\widehat{S}_j\,R_j) \\ \Rightarrow \quad & ker(\widehat{Q}_j - \widehat{Q}_{j-1}) \neq ker(P_j(I_j - Q_{j-1})\,\widehat{S}_j\,R_j). \end{aligned}$$

So the existence of a constant K_1 is not sufficient for the existence of constants $k_{1,j}$ for all $j = 1, \ldots, J$ that would fulfil the inequalities (8.58).

8.7.3 Technical view of the constants

We will now take a look at the constants that determine the condition of $C^{-1}_{DT,1}\,A$ and $C^{-1}_{DT,2}\,A$ in the A-norm. As we have considered the constant c_a for the two grid case quite in-depth we can now use this knowlegde. So we with regard to the estimations for the constants we will refer to the estimations we have done for c_a. Again we will only consider the case in which two points are aggregated to a new one.

The constant $c_{a,1,j}$: First we remember that the constant c_a for the two grid method in section 8.2 was given by the inequality

$$(A\,Q_0\,v,\;\; Q_0\,v) \leq c_a^2\,(A\,v,\;\, v), \quad \text{for all} \quad v \in V.$$

And so as already mentioned, we obtain for the multigrid setting

$$(A\,Q_{J-1}\,v_J,\;\; Q_{J-1}\,v_J) \leq c^2_{a,1,J}\,(A\,v_J,\;\, v_J), \quad \text{for all} \quad v_J \in V_J$$

with $c_{a,1,J} = c_a$. And by the Theorem 8.3.7 or local Lemma 8.3.6 we can estimate this for two aggregated points $\mathcal{N}^J_i, \mathcal{N}^J_k$ as follows

$$c_a = \sqrt{1 + \frac{a_{i,i} + a_{k,k} - 2|a_{i,k}|}{4|a_{i,k}|}}.$$

For $j = 1, \ldots, J$ the constant $c_{a,1,j}$ follows from the inequality

$$\|Q_{j-1}\,\widetilde{v}_j\|_{A_j}^2 \leq c_{a,1,j}^2\,\|\widetilde{v}_j\|_{A_j}^2 \quad \text{for all} \quad \widetilde{v}_j \in \widetilde{V}_j$$

Then we have for each $j = 1, \ldots, J-1$ the same situation as in the two grid situation if we use the entries of A_j instead of A. Therewith we can estimate $c_{a,1,j}$ locally for two points $\mathcal{N}_i^j, \mathcal{N}_k^j$ that are aggregated by

$$c_{a,1,j} = \sqrt{1 + \frac{a_{i,i}^j + a_{k,k}^j - 2|a_{i,k}^j|}{4|a_{i,k}^j|}}.$$

Thereby $a_{i,k}^j$ is the element (i,k) of A_j. So the constants $c_{a,1,j}$ are the generalisation of c_a.

So we can summarize the results for $c_{a,1,j}$ in a theorem that is the generalisation of the Theorem 8.3.7. We set for each pair $\mathcal{N}_i^j, \mathcal{N}_k^j$ of aggregated points

$$c_{a,1,j}^{i,k} := \sqrt{1 + \frac{a_{i,i}^j + a_{k,k}^j - 2|a_{i,k}^j|}{4|a_{i,k}^j|}}.$$

And for a Restriction R_{j-1}^j operator we set

$$Ind_j = \Big\{(i,k) \,:\, i,k \in I_t^{j-1,j} \quad \text{for an } t \in \{1, \ldots, n_{j-1}\}\Big\}.$$

Then Ind_j is the set of points that are aggregated from level j to $j-1$. Therefore we obtain the following result:

Theorem: 8.7.9. *Let A s.p.d. be a matrix as defined for Theorem 8.3.7. We assume that the neighbours of aggregated points are isolated points and that it is $a_{i,k}^j \neq 0$ for all $(i,k) \in Ind_j$. Then*

$$\|Q_{j-1}\widetilde{v}_j\|_{A_j} \leq c_{a,1,j}\,\|\widetilde{v}_j\|_{A_j}$$

holds for all $\widetilde{v}_j \in \widetilde{V}_j$ with $c_{a,1,j} = \max\{c_{a,1,j}^{i,k} \,:\, (i,k) \in Ind_j\}$.

proof. See the calculation above and the proof of Theorem 8.3.7. □

The constant $c_{a,2,j}$: Also for the constant $c_{a,2,j}$ we refer to the situation and estimation we had for c_a and the two grid case. Here we have with $\widehat{Q}_j\, v_j = v_j$ for all

$v_j \in V_j$

$$\begin{aligned}
& \|\widehat{Q}_{j-1}\, v_j\|_A^2 \le c_{a,2,j}^2\, \|v_j\|_A^2, \quad \text{for all} \quad v_j \in V_j \\
\Leftrightarrow \quad & (A\, P_j\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}\, v_j,\ P_j\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}\, v_j) \\
& \qquad \le c_{a,2,j}^2\, (A\, P_j \widehat{S}_j\, R_j\, v_j,\ P_j \widehat{S}_j\, R_j\, v_j), \quad \text{for all} \quad v_j \in V_j \\
\Leftrightarrow \quad & (A_j\, P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}\, v_j,\ P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}\, v_j) \\
& \qquad \le c_{a,2,j}^2\, (A_j\, \widehat{S}_j\, R_j\, v_j,\ \widehat{S}_j\, R_j\, v_j), \quad \text{for all} \quad v_j \in V_j \\
\Leftrightarrow \quad & \|P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j\, \widetilde{v}_j\|_{A_j}^2 \le c_{a,2,j}^2\, \|\widehat{S}_j\, \widetilde{v}_j\|_{A_j}^2, \quad \text{for all} \quad \widetilde{v}_j \in \widetilde{V}_j.
\end{aligned}$$

And again the last equivalence foollows as $R_j : V_j \to \widetilde{V}_j$ is bijective.
If we consider at level j the local situation as shown in Figure 8.6 we assume that the points $\mathcal{N}_i^j$ and $\mathcal{N}_{i+1}^j$ are aggregated to $\mathcal{N}_k^{j-1}$.

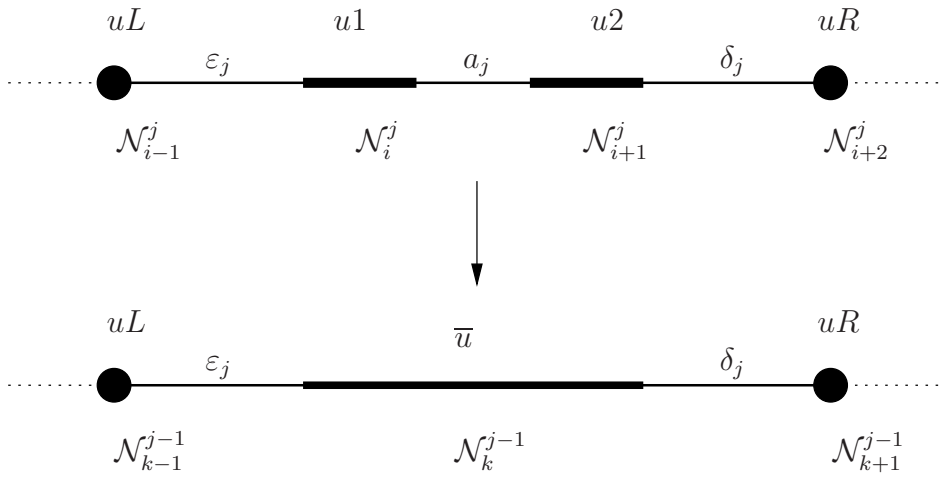


Figure 8.6: Coarsing between the j-th and the $(j-1)$-th grid

Based on the definition of $\widehat{S}_{j-1}$ and the result of Lemma 2.4.4 showing that the structure of $\widehat{S}_{j-1}$ is

$$(\widehat{S}_{j-1})^{-1} = diag(|I_1^{j-1,J}|, \ldots, |I_{n_{j-1}}^{j-1,J}|).$$

We obtain for $\widetilde{v}_j \in \widetilde{V}_j$ that is locally given by

$$\widetilde{v}_j = (u_L, u_1, u_2, u_R)$$

$$P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j\, \widetilde{v}_j = \left(u_L,\ \frac{n_1 u_1 + n_2 u_2}{n_1 + n_2},\ \frac{n_1 u_1 + n_2 u_2}{n_1 + n_2},\ u_R\right)$$

$$\text{with} \quad n_1 := |I_i^{j,J}| \quad \text{and} \quad n_2 := |I_{i+1}^{j,J}|.$$

And by the definitions of n_1, n_2 we obtain

$$(\widehat{S}_{j-1})_{k,k} = |I_k^{j-1,J}| = n_1 + n_2.$$

With

$$A_j = \begin{pmatrix} \varepsilon_j & -\varepsilon_j & 0 & 0 \\ -\varepsilon_j & a_j + \varepsilon_j & -a_j & 0 \\ 0 & -a_j & a_j + \delta_j & -\delta_j \\ 0 & 0 & -\delta_j & \delta_j \end{pmatrix}$$

and the shortcut $\overline{u} = \frac{n_1 u_1 + n_2 u_2}{n_1 + n_2}$ the inequality

$$0 \leq c_{a,1,j}^2 \|\widetilde{v}_j\|_A^2 - \|P_j^{j-1}\, \widehat{S}_{j-1}\, R_{j-1}^j\, \widetilde{v}_j\|_A^2$$

is locally equivalent to

$$\begin{aligned} 0 \leq c_{a,2,j}^2 \left(\varepsilon_j (u_L - u_1)^2 + a_j (u_1 - u_2)^2 + \delta_j (u_2 - u_R)^2 \right) \\ - \left(\varepsilon_j \left(\overline{u} - u_L\right)^2 + \delta_j \left(\overline{u} - u_R\right)^2 \right) =: g. \end{aligned}$$

We differentiate the function g with respect to u_L, u_R to minimize g concerning these variables. We get

$$\begin{aligned} \frac{\partial g}{\partial u_L} &= c_{a,2,j}^2 \varepsilon_j (u_L - u_1) - 2\varepsilon_j (u_L - \overline{u}) \\ \frac{\partial g}{\partial u_R} &= c_{a,2,j}^2 \delta_j (u_R - u_2) - 2\delta_j (u_R - \overline{u}). \end{aligned}$$

So g is minimized with respect to u_L, u_R if we set them

$$u_L = \frac{c_{a,2,j}^2 u_1 - \overline{u}}{c_{a,2,j}^2 - 1} \quad \text{and} \quad u_R = \frac{c_{a,2,j}^2 u_2 - \overline{u}}{c_{a,2,j}^2 - 1}. \tag{8.59}$$

If we put these values in the function g based on the same calculation as done in Lemma 8.3.1 it follows

$$\begin{aligned} g = \frac{c_{a,2,j}^2 (1 - c_{a,2,j}^2)}{(c_{a,2,j}^2 - 1)^2} \left(\varepsilon_j \left[\frac{n_2 (u_2 - u_1)}{n_2 + n_1} \right]^2 + \delta_j \left[\frac{n_1 (u_1 - u_2)}{n_2 + n_1} \right]^2 \right) \\ + \frac{c_{a,2,j}^2 (1 - c_{a,2,j}^2)^2}{(c_{a,2,j}^2 - 1)^2} a_j (u_1 - u_2)^2. \end{aligned}$$

So $g \geq 0$ holds for all $u_1, u_2 \in \mathbb{R}$ if it is

$$(c_{a,2,j}^2 - 1)a_j \geq \frac{\varepsilon_j n_2^2 + \delta_j n_1^2}{(n_1+n_2)^2} \quad \Leftrightarrow \quad c_{a,2,j} = \sqrt{1 + \frac{\varepsilon_j n_2^2 + \delta_j n_1^2}{a_j(n_1+n_2)^2}}.$$

We see that in the case of $n_1 = n_2$ this is the same situation as for c_a and $c_{a,1,j}$ respectively. This will be obvious in the next section as we will see that by assuming that the condition (2.14) holds both generalisations of the two grid method are the same and therewith $c_{a,1,j} = c_{a,2,j}$ obviously holds. And the condition $n_1 = n_2$ is locally for the given level the same assumtions as that the condition (2.14) holds. Generally we can not state wether the estimation for $c_{a,1,j}$ is smaller or bigger as the estimation for $c_{a,2,j}$. This happens because this depends on ε_j, δ_j. If we assume $\varepsilon_j = \delta_j$ it follows $c_{a,1,j} \leq c_{a,2,j}$ and the constants are equal if and only if it is $n_1 = n_2$. Furthermore, $c_{a,2,j} \geq c_{a,1,j}$ holds if we have $\varepsilon_j \geq \delta_j$ and $n_2 \geq n_1$ or $\delta_j \geq \varepsilon_j$ and $n_1 \geq n_2$.

As the minimizing values for u_L, u_R indicated in (8.59) do not depend on ε_j, δ_j we can go through all the generalisation steps we did for the constant c_a.
So we define

$$c_{a,2,j}^{i,k} := \sqrt{1 + \frac{(a_{i,i}^j - |a_{i,k}^j|)|I_k^{j,J}|^2 + (a_{k,k}^j - |a_{i,k}^j|)|I_i^{j,J}|^2|}{4|a_{i,k}^j|(|I_k^{j,J}| + |I_i^{j,J}|)^2}}.$$

and we can summarize the results for $c_{a,2,j}$ as follows:

Theorem: 8.7.10. *Let A s.p.d. be a matrix as defined in Theorem 8.3.7. We assume that the neighbours of aggregated points are isolated points and that it is $a_{i,k}^j \neq 0$ for all $(i,k) \in Ind_j$. Then*

$$c_{a,2,j} \, \|v_j\|_A \geq \|\widehat{Q}_{j-1} \, v_j\|_A$$

holds for all $v_j \in V_j$ with

$$c_{a,2,j} = \max\{c_{a,2,j}^{i,k} \, : \, (i,k) \in Ind_j\}.$$

proof. See the calculation above. □

The constant $c_{G,j}$: To give an estimation for the constant $c_{G,j}$ we have to determine an $c_{G,j}$ so that

$$\|\widehat{Q}_j \, v\|_A^2 \leq c_{G,j}^2 \|v\|_A^2,$$

holds for all $v \in V$. We therefore need a relation between an element $v \in V_J$ and an element $\widehat{Q}_j\, v \in V_j$. This is more difficult to control since the constants $c_{a,i,j}$ we regarded before. That is why we need quite strong assumptions to show quite a weaker result as for the constants $c_{a,i,j},\ i = 1, 2$.

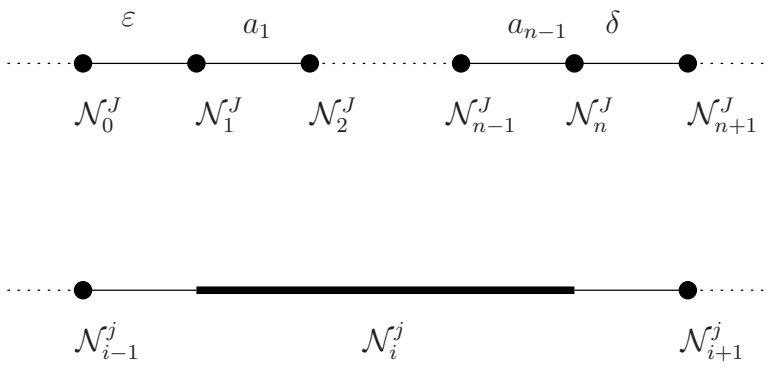


Figure 8.7: Coarsing between the grids J, j

We assume that the situation is given as shown in Figure 8.7. That means the stiffness matrix is given by a one dimensional problem. We have

$$|I_i^{j,J}| = n \quad \text{and} \quad |I_{i-1}^{j,J}| = |I_{i+1}^{j,J}| = 1.$$

So at j level two points $\mathcal{N}_0^J, \mathcal{N}_{n+1}^J$ are isolated points in all previous steps and there are n points $\mathcal{N}_1^J, \ldots, \mathcal{N}_n^J$ that are aggregated at j level to the point $\mathcal{N}_i^j$. Then for $u \in V$ the values for $\|u\|_A^2$ and $\|\widehat{Q}_j\, u\|_A^2$ with $\overline{u} := \frac{1}{n}\sum_{i=1}^n u_i$ are given as follows:

$$\|u\|_A^2 = \varepsilon(u_L - u_1)^2 + \sum_{i=1}^{n-1} a_i(u_{i+1} - u_i)^2 + \delta(u_R - u_n)^2$$

$$\|\widehat{Q}_j\, u\|_A^2 = \varepsilon(u_L - \overline{u})^2 + \delta(u_R - \overline{u})^2.$$

As done before we set $g := c_{G,j}^2 \|u\|_A^2 - \|Q_j\, u\|_A^2$ and differentiate g with respect to u_L, u_R. This implies as minimizing expressions

$$u_L = \frac{c_{G,j}^2 u_1 - \overline{u}}{c_{G,j}^2 - 1} \quad \text{and} \quad u_R = \frac{c_{G,j}^2 u_n - \overline{u}}{c_{G,j}^2 - 1}.$$

If we insert the minimizing values and go through the same calculation steps as done for $c_a, c_{a,1,j}$ or $c_{a,2,j}$ it follows that $g \geq 0$ is implied by

$$(c_{G,j}^2 - 1)\sum_{i=1}^{n-1} a_i(u_{i+1} - u_i)^2 \geq \varepsilon(u_1 - \overline{u})^2 + \delta(u_n - \overline{u})^2. \tag{8.60}$$

We transform the variables into

$$y_1 := (u_2 - u_1), \quad y_2 := (u_3 - u_2), \ldots, y_{n-1} := (u_n - u_{n-1}) \quad \text{and} \quad y_n := u_1$$

then we obtain based on a simple calculation

$$\overline{u} - u_1 = \frac{1}{n} \sum_{i=1}^{n-1} y_i \, (n-i) \quad \text{and} \quad \overline{u} - u_n = -\frac{1}{n} \sum_{i=1}^{n-1} y_i \, i$$

Therewith it is (8.60) equivalent to

$$(c_{G,j}^2 - 1) \sum_{i=1}^{n-1} a_i y_i^2 \geq \frac{\varepsilon}{n^2} \left(\sum_{i=1}^{n-1} y_i \, (n-i) \right)^2 + \frac{\delta}{n^2} \left(\sum_{i=1}^{n-1} y_i \, i \right)^2 . \tag{8.61}$$

Then we use on the right side pairwise the inequality of Young. So we obtain

$$2 \, (y_i \, i) \, (y_j \, j) \leq y_i^2 \, j^2 + y_j^2 \, i^2$$

$$\text{and} \quad 2 \, (y_i \, (n-i)) \, (y_j \, (n-j)) \leq y_i^2 \, (n-j)^2 + y_j^2 \, (n-i)^2.$$

Furthermore, we use the result

$$\sum_{i=1}^{n-1} i^2 = \frac{n(n-1)(2n-1)}{6}.$$

Therewith we get

$$\frac{\delta}{n^2} \left(\sum_{i=1}^{n-1} y_i i \right)^2 \leq \frac{\delta}{n^2} \sum_{i=1}^{n-1} \left(y_i^2 \sum_{k=1}^{n-1} k^2 \right) = \frac{\delta}{n^2} \sum_{i=1}^{n-1} y_i^2 \frac{n(n-1)(2n-1)}{6}$$

As we get the same result for $\frac{\varepsilon}{n^2} \left(\sum_{i=1}^{n-1} y_i (n-i) \right)^2$ the inequality (8.61) is fulfilled if

$$(c_{G,j}^2 - 1) \sum_{i=1}^{n-1} a_i y_i^2 \geq \frac{(\delta + \varepsilon)(n-1)(2n-1)}{6n} \sum_{i=1}^{n-1} y_i^2$$

holds. This is fulfiled if we have

$$c_{G,j} \geq \sqrt{1 + \frac{(\delta + \varepsilon)(n-1)(2n-1)}{6na_i}}, \quad \forall i = 1, \ldots, n-1.$$

Therefore, we can give an estimation for $c_{G,j}$. However, we have to limit the result as we assumed $|I_{i-1}^{j,J}| = |I_{i+1}^{j,J}| = 1$ and the estimation depends on the number of points that are aggregated.

8.7.4 Generalisation of C_{DT}^{-1}. Common Version.

In this section we will assume that we use the aggregation method to construct the spaces $V_{J-1}, \ldots, V_0$ and the condition (2.14) holds. We will see that in this case the two generalisations of C_{DT}^{-1} we have presented are the same. So we have two different representations for the same operator. Hence, we can use the good properties of both and prove a stronger proposition for the constants c_{DT}, d_{DT} as we have done in the sections 8.7.1 or 8.7.2. More precisely, we will see that based on this assumption we can drop the constant K_1, K_2.

Based on Lemma 2.4.9 the condition (2.14) is equivalent to the equation

$$\widehat{S}_j \, P_j^{j-1} \, S_{j-1} = P_j^{j-1} \, \widehat{S}_{j-1}.$$

Therewith we obtain for $j = 1, \ldots, J$

$$\begin{aligned}
u_{2,j} &= P_j \, A_j^{-1} \, (I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j) \, R_j \, f \\
&= P_j \, A_j^{-1} \, (I_j - \widehat{S}_j^{-1} \, \widehat{S}_j \, P_j^{j-1} \, S_{j-1} \, R_{j-1}^j) \, R_j \, f \\
&= P_j \, A_j^{-1} \, (I_j - \, P_j^{j-1} \, S_{j-1} \, R_{j-1}^j) \, R_j \, f \\
&= P_j \, A_j^{-1} \, (I_j - \, Q_{j-1}) \, R_j \, f \\
&= u_{1,j}.
\end{aligned}$$

Based on the definitions $u_{1,0} = u_{2,0}$ the both preconditioners are the same. We set $u_j := u_{1,j}$, for $j = 0, \ldots, J$. So assuming that the condition (2.14) holds we have two representations of the same preconditioner

$$\begin{aligned}
C_{DT}^{-1} \, f &= \sum_{j=1}^{J} P_j \, A_j^{-1} (I_j - \widehat{S}_j^{-1} \, P_j^{j-1} \, \widehat{S}_{j-1} \, R_{j-1}^j) \, R_j \, f + P_0 \, A_0^{-1} \, R_0, f \\
&= \sum_{j=1}^{J} P_j \, A_j^{-1} (I_j - Q_{j-1}) \, R_j \, f + P_0 \, A_0^{-1} \, R_0, f \\
&= \sum_{j=0}^{J} u_j.
\end{aligned}$$

So far it is obvious that we have two representations of the same operator. Hence we can always use the more useful representation. In one case, this means we use version

1 to obtain the orthogonality of u_i, u_j for $i \neq j$ with respect to the dotproduct $a(.,.)$, while in the other case we use version 2 to obtain a decomposition of f by orthogonal projections $\widehat{Q}_j$.

The result for K_1 we will prove as an example (this results from the decomposition of f). Therefore we remember that in the Lemmata 2.3.8, 2.4.8 we proved that if the condition (2.14) holds, it follows

$$\widehat{Q}_{j-1} = P_j\,\widehat{S}_j\,Q_{j-1}\,R_j \quad \text{and} \quad \widehat{S}_j\,Q_{j-1} = Q_{j-1}\,\widehat{S}_{j-1}.$$

Hence we get for the constants K_1

(8.62)

$$\begin{aligned}
K_1 &= \sup\left\{ \frac{a(v,\,v)}{a\left(v,\,\sum_{j=1} P_j(I_j - Q_{j-1})\,\widehat{S}_j\,R_j\,v + P_0\,\widehat{S}_0\,R_0\,v\right)} \;:\; v \in V\backslash\{0\} \right\} \\
&= \sup\left\{ \frac{a(v,\,v)}{a\left(v,\,\sum_{j=1}(P_j\,\widehat{S}_j\,R_j - P_j\,Q_{j-1}\,\widehat{S}_j\,R_j)\,v + P_0\,\widehat{S}_0\,R_0\,v\right)} \;:\; v \in V\backslash\{0\} \right\} \\
&= \sup\left\{ \frac{a(v,\,v)}{a\left(v,\,\sum_{j=1}(\widehat{Q}_j - \widehat{Q}_{j-1})\,v + \widehat{Q}_0\,v\right)} \;:\; v \in V\backslash\{0\} \right\} \\
&= \sup\left\{ \frac{a(v,\,v)}{a(v,\,v)} \;:\; v \in V\backslash\{0\} \right\} = 1.
\end{aligned}$$

So we can drop the constant K_1. On the other side, we can set $\sqrt{J+1}$ for the constant K_2 as used for the estimations concerning $C_{DT,2}^{-1}$ (this results form the orthogonality of u_i, u_j). As already mentioned, in this situation we obtain a result for the condition of $C_{DT}^{-1}\,A$ that combines the good characteristics of both versions. This is what we meant when we mentioned that we can drop these constants.

Theorem: 8.7.11. *Let $A \in \mathbb{R}^{n\times n}$ be s.p.d. and assume that it holds the condition (2.14). With $c_{a,1,j}, c_{G,j}$ as defined in (8.53) and (8.40)*

$$c_{DT}\|v\|_A \leq \|C_{DT}^{-1}\,A\,v\|_A \leq d_{DT}\|v\|_A$$

holds for all $v \in V$ with

$$c_{DT} = \frac{1}{\sqrt{J+1}\max_{j=0,\ldots,J} c_{G,j}} \quad \textit{and} \quad d_{DT,2} = (J+1) + \sum_{j=1}^{J} c_{a,1,j}.$$

proof. The proof follows from Proposition 8.7.8 and the characteristic that $K_1 = 1$ holds in the assumed situation as shown in (8.62). □

9 Numerical results

In this chapter we will first sum up properties of matrices which are useful for iterative methods. Then we will consider for $j = 0, \ldots, J$ the matrices $A_j, A_{j,X}$ of our model problems with respect to these characteristics. To conclude the chapter we will present some numerical results for the model problems, highlighting that the numerical results for the modification which is presented in section 5.1.3 and motivated as the exact modification in section 5.1.1 are correct. For other modifications we have to take a closer look at the individual situations.

9.1 Characteristics of matrices

In chapter 3 we have used the operators A^{-1}, A_0^{-1} to define the preconditioners $C_{BPX}^{-1}, C_{DT}^{-1}, C_{2P}^{-1}$ and the associated modified options. We have mentioned that we do not want to use the exact inverse. In the multigrid situation we have defined the same preconditioners with non singular matrices $B^{(j)}$ for $j = 0, \ldots, J$ and we have recognized $B^{(j)} = A_j$ only as an example. For the numerical experiments we will carry out in the next section we set for $(B^j)^{-1}$ some iterations of an iterative method. In paricular we will use the Jacobi method and the SSOR method.

9.1.1 Basics for iterative methods

In this section we want to introduce some splitting methods. Afterwards we will sum up some results for these methods presented in [GrR94], [Hac85]. As usual we define for a matrix A the spectrum $\sigma(A)$ and the spectral radius $\rho(A)$ as follows

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

$$\rho(A) := \max\{|\lambda| \,:\, \lambda \in \sigma(A)\}.$$

A linear iteration method to solve $Ax = b$ can be presented as

$$x^{k+1} := Mx^k + Nb \quad x_0 \in \mathbb{R}^n$$
$$\text{with} \quad M, N \in \mathbb{R}^{n\times n}.$$

The matrix M is said to be the *iteration matrix.* Moreover we define for a $A \in \mathbb{R}^{n\times n}$

$$A = D - L - R, \quad \text{with} \quad D = diag(a_{1,1}, \ldots, a_{n,n})$$

$$-L = \begin{pmatrix} 0 & & & 0 \\ a_{2,1} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n,1} & \cdots & a_{n,n-1} & 0 \end{pmatrix} \quad \text{and} \quad -U = \begin{pmatrix} 0 & a_{1,2} & \cdots & a_{1,n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ 0 & & & 0 \end{pmatrix}$$

Then the Jacobi method is defined as follows

$$\begin{aligned} x^{k+1} &:= D^{-1}(L+U)\, x^k + D^{-1}\, b \\ &= (I - D^{-1}\, A)\, x^k + D^{-1}\, b. \end{aligned}$$

The Gauss-Seidel method is defined as follows

$$x^{k+1} := (D-L)^{-1}\, R\, x^k + (D-L)^{-1}\, b.$$

Hence we have the iteration matrices

$$M_J := (I - D^{-1}\, A) \quad \text{and} \quad M_{GS} := (D-L)^{-1}\, R$$

The main aspect of these methods is given by the following result:

Proposition: 9.1.1. *A linear iteration method, with the iteration matrix M converges if and only if it is $\rho(M) < 1$.*

proof. Cf. [GrR94] Lemma 5.2 or [Hac85] Proposition 3.2.7. □

Furthermore we define the residual error e^k as follows

$$e^k := b - A\, x^k.$$

For a linar iteration method we obtain for an arbitrary matrix norm $\|.\|$ that the sequence of x^k converges if we have $\|M\| < 1$. Additionally we obtain in this case

$$\|e^k\| \leq \|M\|^k\, \|e^0\|.$$

9.1.2 Basics for matrices

In this section we will define some basic characteristics for matrices. Afterwards we will consider the relation to the property $\rho(M_J), \rho(M_{GS}) < 1$. For the characterisation of matrices we follow the definition given in [Hac85]. Hence we define for $A \in \mathbb{R}^{n\times n}$

$$A \geq 0 \quad (A > 0) \quad \text{if it is} \quad a_{i,j} \geq 0 \quad (a_{i,j} > 0) \quad \forall i, j.$$

Based on this notation we define the following matrices:

Definition: 9.1.2. *A matrix $A \in \mathbb{R}^{n\times n}$ is said to be*

1. *a L_0-matrix, if it is $a_{i,j} \leq 0$ for $i \neq j$.*
2. *a L-matrix, if it is $a_{i,j} \leq 0$ for $i \neq j$ and $a_{i,i} > 0$*
3. *a M-matrix, if A is a non singular L-matrix and it is $A^{-1} \geq 0$.*

Furthermore we define the graph $G(A)$ of a matrix $A \in \mathbb{R}^{n\times n}$ as

$$G(A) := \{(i, j) \,:\, a_{i,j} \neq 0\}.$$

The elements $(i, j) \in G(A)$ are also called *edges*, and the rows (or columns) of the matrix are also called *verteces* in this case. Then we say that i is *adjacent* to j if it it $(i, j) \in G(A)$. We say that i, j are *connected*, if there are $k_0, k_1, \ldots, k_t$ with

$$i = k_0, k_1, \ldots, k_{t-1}, k_t = j \quad \text{with} \quad (k_{s-1}, k_s) \in G(A) \quad \text{for all} \quad s = 1, \ldots, t.$$

Otherwise we say that i, j are *disconnected.* Furthermore we say that $G(A)$ is *connected* if all $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$ are connected. Otherwise we say that $G(A)$ is *disconnected.*

Definition: 9.1.3. *A matrix $A \in \mathbb{R}^{n\times n}$ is said to be*

1. weakly (strictly) diagonally dominant*, if it is*

$$|a_{i,i}| \geq \sum_{j=1,\, j\neq i}^{n} |a_{i,j}| \quad \left(|a_{i,i}| > \sum_{j=1,\, j\neq i}^{n} |a_{i,j}| \right)$$

   for all $i = 1, \ldots, n$.

2. irreducible *if $G(A)$ is connected. Otherwise A is said to be* reducible.

3. irreducible diagonal dominant*, if A is irreducible, weakly diagonally dominant and there is one $i_0 \in \{1, \ldots, n\}$ with*

$$|a_{i_0,i_0}| > \sum_{j=1,\, j\neq i_0}^{n} |a_{i_0,j}|.$$

Remark: 9.1.4. *A matrix $A \in \mathbb{R}^{n\times n}$ is reducible, if and only if there is a permutation matrix Π that holds*

$$\Pi\, A\, \Pi^T = \begin{pmatrix} A^{(1,1)} & A^{(1,2)} \\ 0 & A^{(2,2)} \end{pmatrix}$$

with $A^{(1,1)} \in \mathbb{R}^{k,k}$, $A^{(2,2)} \in \mathbb{R}^{(n-k)\times(n-k)}$ and $A^{(1,2)} \in \mathbb{R}^{k\times(n-k)}$ for a $k \in 1, \ldots, n-1$.

Basesd on these definitions we obtain the following results:

Proposition: 9.1.5. *Let $A \in \mathbb{R}^{n\times n}$ be a strict diagonal dominant or irreducible diagonal dominant, then we have*

$$\rho(M_J) = \rho(I - D_A^{-1}\, A) < 1.$$

proof. Cf. [Hac85] Proposition 6.4.10. □

Based on a similar condition we obtain that A is a M-matrix if we additionally use the condition of the algebraic signs of the elements of A :

Proposition: 9.1.6. *Let $A \in \mathbb{R}^{n\times n}$ be a L-matrix. If A is strict diagonal dominant or irreducible diagonal dominant then A is a M-matrix.*

proof. S. [GrR94] Proposition 1.6 or [Ost]. □

If $A \in \mathbb{R}^{n\times n}$ is s.p.d. we can give a simple sufficient condition to ensure that A is an M-matrix. It holds:

Proposition: 9.1.7. *Let $A \in \mathbb{R}^{n\times n}$ be s.p.d. If it is $a_{i,j} \leq 0$ for all $i \neq j$ then A is an M-matrix.*

proof. Cf. [Hac85] Proposition 6.4.18. □

The relation between the Propositions 9.1.5, 9.1.6 are obvious based on the following result:

Proposition: 9.1.8. *Let $A \in R^{n\times n}$ be a L_0-matrix.*

a) The following two assertions are equivalent:

1) A is non singular and it is $A^{-1} \geq 0$

2) It is $a_{i,i} > 0$ for $i = 1, \ldots, n$, $M_J \geq 0$ and $\rho(M_J) < 1$.

b) Additionally it follows that if the condition in a) is fulfilled then A is an M-matrix. Vice versa it holds $M_J \geq 0$ and $\rho(M_J) < 1$ if A is an M-matrix.

proof. Cf. [Hac85] Proposition 6.4.4. □

As mentioned before the definition follows the definition given in [Hac85]. The proposition 9.1.8 proves that we can drop the condition $a_{i,i} > 0$ in the definition of the M-matrix. This is the principle we followed in [GrR94]. It is less well-known that we can also drop the assumption $A^{-1} \geq 0$ if we assume, instead of this that A is weak diagonal dominant.

Proposition: 9.1.9. *Let $A \in R^{n \times n}$ be a non singular, irreducible diagonal dominant L-matrix. Then it follows $A^{-1} \geq 0$.*

proof. It is obvious that $A^{-1} \geq 0$ is equivalent to

$$Ax = b \quad \text{with} \quad b \geq 0 \quad \Rightarrow \quad x \geq 0.$$

Hence we assume that it is $Ax = b$ with $b \geq 0$ and $x_{i_0} < 0$. W.l.o.g. we assume $x_{i_0} \leq x_j$ for all $j = 1, \ldots, n$. Then we obtain

$$0 \leq b_{i_0} = \sum_{j=1}^{n} a_{i_0,j} x_j \quad \Leftrightarrow \quad x_{i_0} \geq \sum_{j=1, j \neq i_0}^{n} \frac{a_{i_0,j}}{a_{i_0,i_0}} x_j.$$

Based on $x_{i_0} \leq x_j$ it follows $a_{i_0,j} \neq 0 \Rightarrow x_j = x_{i_0}$ for all $j = 1, \ldots, n$. Hence we obtain that there are $m \geq 2$ elementes $x_j \in x$ with $x_j = x_{i_0}$. With a permutation matrix Π we obtain

$$A^* = \Pi \, A \, \Pi^T = \begin{pmatrix} A^{(1,1)} & 0 \\ A^{(2,1)} & A^{(2,2)} \end{pmatrix}$$

with $A^{(1,1)} \in \mathbb{R}^{m \times m}$. As A is non singular this also holds for A^* and $A^{(1,1)}$. We obtain by operations with the rows of A^* a non singular matrix

$$A^{**} = \begin{pmatrix} A^{(1,1)} & 0 \\ 0 & \widetilde{A}^{(2,2)} \end{pmatrix}$$

But for $v = e_1 + \cdots + e_m$ we obtain $A^{**} v = 0$. This is in contradiction to the assumption. □

Finally we will present a relation between M_J and M_{GS}.

Proposition: 9.1.10. *Let $A \in \mathbb{R}^{n\times n}$ be an M-matrix. Then it follows*

$$\rho(M_{GS}) \leq \rho(M_J) < 1.$$

proof. Cf. [Hac85] Proposition 6.6.3. □

Hence we can summarize the result for the matrices with a look at the splitting methods: If the matrices we use are M-matrices or irreducible (strict) diagonal dominants then the splitting methods converge.

9.1.3 Results for $A_j, A_{j,X}$

Based on the definition of our model problems in chapter 2, we obtain that A is a weak diagonal dominant L-matrix for these problems. The symmetric problem implies always an irreducible matrix A. For the convection system this is the case if there is a diffusion, too ($\varepsilon \neq 0$). Hence it is obvious that if we assume that the stiffness matrix is based on the discretisation of one of the partial differential equations presented in chapter 2, then the irreducibility depends on the diffusion. Based on proposition 9.1.6 we obtain that A is an M-matrix in this case. Now we discuss if the coarser matrices $A_j, A_{j,X}$ are M-matrices, too. Again, we only consider the situation of two grids and drop the indices for P, R. It is obvious that this contains all the information as the coarser operators are defined iteratively. Hence we can use all the arguments iteratively. We remember that if we use the operators P, R given by the aggregation method then

$$R_{k,\cdot} = (e_j^1)^T \quad \text{implies} \quad (R\,A)_{k,\cdot} = A_{j,\cdot}$$
$$\text{and} \quad R_{k,\cdot} = (e_i^1 + e_j^1)^T \quad \text{implies} \quad (R\,A)_{k,\cdot} = A_{j,\cdot} + A_{i,\cdot}.$$

Furthermore we obtain that

$$P_{\cdot,k} = e_j^1 \quad \text{implies} \quad (A\,P)_{\cdot,k} = A_{\cdot,j}$$
$$\text{and} \quad P_{\cdot,k} = e_i^1 + e_j^1 \quad \text{implies} \quad (A\,P)_{\cdot,k} = A_{\cdot,j} + A_{\cdot,i}.$$

Based on these characteristics we obtain the following result:

Proposition: 9.1.11. *Let $A \in \mathbb{R}^{n\times n}$ be an irreducible diagonal dominant L-matrix. If we assume that P, R are based on the aggregation method then A_0 is also a irreducible diagonal dominant L-matrix.*

proof. It is sufficient to consider the situation we obtain if we aggregate two points. W.l.o.g. we assume that we aggregate the points $\mathcal{N}_n^1, \mathcal{N}_{n-1}^1$ to $\mathcal{N}_{n-1}^0$. Then it follows

$$R = \begin{pmatrix} I_{n-2} & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

We consider two different kinds of rows of A_0.
For $k < n-1$ the rows of A_0 are follows as

$$(A_0)_{k,\cdot} = (a_{k,1}, \ldots, a_{k,n-2}, a_{k,n-1} + a_{k,n}).$$

Hence it holds for these rows

$$a_{k,j}^0 = \begin{cases} a_{k,j} \leq 0, & \text{for } j \neq k, n \\ a_{k,n-1} + a_{k,n} \leq 0, & \text{for } j = n-1 \\ a_{k,k} > 0 & \text{for } j = k. \end{cases}$$

Furthermore we obtain

$$a_{k,k}^0 = a_{k,k} \geq \sum_{j=1,\, j\neq k}^{n} |a_{k,j}| = \sum_{j=1,\, j\neq k}^{n-1} |a_{k,j}^0|.$$

Now we consider the $(n-1)$-th row of A_0. We obtain

$$(A_0)_{n-1,\cdot} = (a_{n-1,1} + a_{n,1}, \ldots, a_{n-1,n-2} + a_{n,n-2}, a_{n-1,n-1} + a_{n-1,n} + a_{n,n-1} + a_{n,n}).$$

Based on the characteristics for $a_{n-1,j}, a_{n,j}$ for $j = 1, \ldots, n$ we obtain

$$a_{n-1,k}^0 = a_{n-1,k} + a_{n,k} \leq 0 \quad \text{for} \quad k \neq n-1$$

$$a_{n-1,n-1}^0 = \underbrace{a_{n-1,n-1} + a_{n-1,n}}_{\geq 0} + \underbrace{a_{n,n-1} + a_{n,n}}_{\geq 0} \geq 0.$$

The last inequality above is follows from the weak diagonal dominance of A. To prove that A_0 is an L-matrix we have to prove $a_{n-1,n-1}^0 > 0$. Assume that we have $a_{n-1,n-1}^0 = 0$. From the weak diagonal dominance of A it follows in this case

$$a_{n,n} = -a_{n,n-1}, \quad a_{n-1,n-1} = -a_{n-1,n}$$

$$\text{and} \quad a_{n-1,k} = a_{n,k} = 0 \quad \text{for} \quad k < n-1.$$

This implies

$$A = \begin{pmatrix} A^{(1,1)} & A^{(1,2)} \\ 0 & A^{(2,2)} \end{pmatrix}$$

with $A^{(1,1)} \in \mathbb{R}^{(n-2)\times(n-2)}$ and $A^{(2,2)} \in \mathbb{R}^{2\times 2}$. This is in contradiction to the irreducibility of A. Hence we have $a^0_{n-1,n-1} > 0$.
Concerning the weak diagonal dominance of A_0 we obtain

$$\begin{aligned} a^0_{n-1,n-1} &= a_{n-1,n-1} + a_{n-1,n} + a_{n,n-1} + a_{n,n} \\ &\geq \sum_{j=1,\, j\neq n-1,n}^{n} (|a_{n-1,j}| + |a_{n,j}|) \\ \Leftrightarrow \quad a_{n-1,n-1} + a_{n,n} &\geq \sum_{j=1,\, j\neq n}^{n} |a_{n,j}| + \sum_{j=1,\, j\neq n-1}^{n} |a_{n-1,j}|) \\ \Leftarrow \quad a_{n-1,n-1} &\geq \sum_{j=1,\, j\neq n-1}^{n} |a_{n-1,j}| \quad \text{and} \quad a_{n,n} \geq \sum_{j=1,\, j\neq n}^{n} |a_{n,j}|. \end{aligned}$$

In the calculations above we have proved that A_0 is a weak diagonal dominant. So far we have proved that and A_0 is a weak diagonal dominant L-matrix.
The irreducibility of A_0 follows as it is

$$|a^0_{i,j}| \geq |a_{i,j}| \quad \text{for} \quad (i,j) \neq (n-1, n-1).$$

Based on the assumption that A is irreducible diagonal dominant, there is a $k \in \{1, \ldots, n\}$ with

$$\sum_{i=1,\, i\neq k}^{n} |a_{k,i}| < a_{k,k}.$$

Based on this property the calculations above prove that for $k \leq n-2$ it follows that

$$\sum_{i=1,\, i\neq k}^{n-1} |a^0_{k,i}| < a^0_{k,k}$$

and for $k = n-1$ or $k = n$ it follows that

$$\sum_{i=1}^{n-2} |a^0_{n-1,i}| < a^0_{n-1,n-1}.$$

Hence A_0 is also irreducible diagonal dominant. □

Based on the Proposition above for the aggregation method much of the structure of the matrix A is maintained if we consider the coarser operator A_0. In section 5.1.1 we have seen that also for the simple one dimensional convection this is not true for $A_{0,X}$ if we use the modification based on the inverse of blocks. In particular we have seen that the matrix $A_{0,X}$ can be singular in this case.
In section 5.1.1 we have also seen that if we use the exact modification for the one dimensional convection (without diffusion, i.e. $\varepsilon = 0$) then $A_{0,X}$ has the same structure as A_0 and A, respectively. In particular they are all weak diagonal dominant L-matrices. We have to highlight that for this example neither A nor $A_0, (A_{0,X})$ are irreducible.

We will take a closer look at the matrices $A_{0,X}$ we obtain from the modifications presented in Lemma 5.1.10 or more generally in proposition 5.1.11. We remember that these are the generalisations of the exact modification presented in section 5.1.1.

We consider the convection diffusion system composed of four grid points. We remember that the matrices A, R are defined as

(9.1)
$$A = \begin{pmatrix} b_1 + \varepsilon_0 + \varepsilon_1 & -\varepsilon_1 & 0 & 0 \\ -b_2 - \varepsilon_1 & b_2 + \varepsilon_1 + \varepsilon_2 & -\varepsilon_2 & 0 \\ 0 & -b_3 - \varepsilon_2 & b_3 + \varepsilon_2 + \varepsilon_3 & -\varepsilon_3 \\ 0 & 0 & -b_4 - \varepsilon_3 & b_4 + \varepsilon_3 + \varepsilon_4 \end{pmatrix} \quad R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(cf. (5.21)). As a generalisation of X defined in (5.22) we set

(9.2)
$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ p & 1 & -q & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{with} \quad p, q \in \mathbb{R}_+.$$

Then we obtain the following proposition for the modified operator $A_{0,X}$:

Lemma: 9.1.12. *Assume that A, R are as defined in (9.1) and it is $\varepsilon_i = \varepsilon$ for $i = 0, \ldots, 4$. If X is defined in (9.2) then we have*

1. $(A_{0,X})_{2,2} > 0$ *if it is* $q \leq 1$ *or* $b_3 \geq b_2 + \varepsilon$.
2. $(A_{0,X})_{2,j} \leq 0$ *for* $j = 1, 3$ *if it is* $p \leq 1$ *or* $b_3 \geq b_2 + \varepsilon$.

3. *that* $A_{0,X}$ *is a L-matrix if it is* $p, q \leq 1$.

4. $(a^{0,X})_{2,2} \geq |(a^{0,X})_{2,1} + (a^{0,X})_{2,3}|$ *if we have* $p = q \leq 1$ *or* $1 \geq q > p$ *and* $b_3 \geq b_2 + \varepsilon$.

5. $A_{0,X}$ *is a weak diagonal dominant if we have* $q = p \leq 1$.

proof. Based on the definitions for A, R, X we obtain

$$A_{0,X} = \begin{pmatrix} b_1 + \varepsilon(1-p) & -\varepsilon(1-q) & 0 \\ -b_2 - \varepsilon + (b_2 - b_3 + \varepsilon)p & b_2 + 2\varepsilon - (b_2 - b_3 + \varepsilon)q & -\varepsilon \\ 0 & -b_4 - \varepsilon & b_4 + 2\varepsilon \end{pmatrix}.$$

1. The first assertion follows from

$$(A_{0,X})_{2,2} = b_2 + 2\varepsilon - (b_2 + \varepsilon - b_3)q = b_2(1-q) + b_3 + \varepsilon(2-q).$$

2. Similar to the first one, the second assertion follows from

$$(A_{0,X})_{2,1} = -b_2 - \varepsilon + (b_2 - b_3 + \varepsilon)p = -b_2(1-p) - b_3 p - \varepsilon(1-p)$$
$$(A_{0,X})_{2,3} = -\varepsilon$$

3. If we consider the first row of $A_{0,X}$ then we obtain $(A_{0,X})_{1,1} > 0$ for $p \leq 1$ and $(A_{0,X})_{1,3} \leq 0$ for $q \leq 1$. Together with the two results above this implies that $A_{0,X}$ is an L-matrix in this case.

4. Based on the second row of $A_{0,X}$ we obtain

$$\begin{aligned} & (A_{0,X})_{2,2} \geq |(A_{0,X})_{2,1}| + |(A_{0,X})_{2,3}| \\ \Leftrightarrow \quad & b_2 + 2\varepsilon - (b_2 - b_3 + \varepsilon)q \geq b_2 + \varepsilon - (b_2 - b_3 + \varepsilon)p + \varepsilon \\ \Leftrightarrow \quad & (p-q)(b_2 + \varepsilon - b_3) \geq 0. \end{aligned}$$

5. For $p = q \leq 1$ the first row of $A_{0,X}$ holds $(A_{0,X})_{1,1} \geq |(A_{0,X})_{1,2}| + |(A_{0,X})_{1,3}|$. Hence the last assertion follows from previous assertion.

□

We can sum up the technical results before in a simple proposition that follows immediately:

Corollary: 9.1.13. *Assume that A, R are as defined in (9.1) and it is $\varepsilon_i = \varepsilon$ for $i = 0, \ldots, 4$. If X is defined in (9.2) with $p = q \leq 1$ then $A_{0,X}$ is a weak diagonal dominant L-matrix. If we have $p = q < 1$ then $A_{0,X}$ is irreducible diagonal dominant.*

proof. The first assertion follows immediately from Lemma 9.1.12. If we have additionally $p, q < 1$ then we obatin $(A_{0,X})_{1,2} < 0$. Hence $A_{0,X}$ is irreducible in this case. Furthermore, for the first row of $A_{0,X}$ we have

$$a_{1,1}^{0,X} > |a_{1,2}^{0,X}| + |a_{1,3}^{0,X}|.$$

Thus $A_{0,X}$ is irreducible diagonal dominant. □

Proposition: 9.1.14. *Assume that A, R are as defined in (9.1) and it is $\varepsilon_i = \varepsilon$ for $i = 0, \ldots, 4$. If X is defined in (9.2) with p, q as defined in (5.22) then $A_{0,X}$ is an irreducible diagonal dominant L-matrix.*

proof. Based on the Corollary 9.1.13 it is sufficient to prove $p = q < 1$. For the defined values we have

$$\begin{aligned} p &= \frac{|a_{2,1}|}{a_{2,2} + |a_{3,2}|} = \frac{b_2 + \varepsilon}{b_2 + b_3 + 3\varepsilon} < 1 \\ q &= -\frac{a_{3,3} - a_{2,2} + a_{3,2}}{a_{2,2} + |a_{3,2}|} = -\frac{b_3 + 2\varepsilon - b_2 - 2\varepsilon - b_3 - \varepsilon}{b_2 + b_3 + 3\varepsilon} \\ &= \frac{b_2 + \varepsilon}{b_2 + b_3 + 3\varepsilon} = p. \end{aligned}$$

□

Thus the presented exact modification has for the coarser operators characteristics which are useful from a numerical point of view. But as the invariance does not hold for more complex systems this is also true for these characteristics.

If we drop the condition of $\varepsilon_i = \varepsilon$ for $i = 0, \ldots, 4$ and consider the matrix X as presented in (5.22) then it follows

$$\begin{aligned} p &= \frac{|a_{2,1}|}{a_{2,2} + |a_{3,2}|} = \frac{b_2 + \varepsilon_1}{b_2 + \varepsilon_1 + 2\varepsilon_2 + b_3} \\ \text{and} \quad q &= -\frac{a_{3,3} - a_{2,2} + a_{3,2}}{a_{2,2} + |a_{3,2}|} = -\frac{b_3 + \varepsilon_2 + \varepsilon_3 - b_2 - \varepsilon_1 - \varepsilon_2 - b_3 - \varepsilon_2}{b_2 + \varepsilon_1 + 2\varepsilon_2 + b_3} \\ &= \frac{b_2 + \varepsilon_1 + (\varepsilon_2 - \varepsilon_3)}{b_2 + \varepsilon_1 + 2\varepsilon_2 + b_3}. \end{aligned}$$

Hence it follows obviously $p \neq q$ for $\varepsilon_2 \neq \varepsilon_3$. But if we assume that it is $b_i >> \varepsilon_j$ then it is $p \approx q$. A similar problem we obtain in the case of a two dimensional convection system also without any kind of diffusion. If the stencil in $\mathcal{N}_i^1$ is given as

$$\begin{pmatrix} 0 & 0 & 0 \\ -b_{i,x} & b_{i,x} + b_{i,y} & 0 \\ 0 & -b_{i,y} & 0 \end{pmatrix} \quad \text{with} \quad b_{i,x}, b_{i,y} \in \mathbb{R}_+$$

and we assume that the points $\mathcal{N}_i^1, \mathcal{N}_{i+1}^1$ fulfill $a_{i,i+1} = 0, a_{i+1,i} = b_{i+1,x}$. That means that there is a flow from $\mathcal{N}_i^1$ to $\mathcal{N}_{i+1}^1$ in the x-direction. Furthermore we assume that $\mathcal{N}_i^1, \mathcal{N}_{i+1}^1$ are aggregated and we follow the modification as presented in section 5.1.2. Then the entries of the modification matrix are

$$p = \frac{b_{i,x}}{b_{i,x} + b_{i,y} + b_{i+1,x}}$$

$$q = -\frac{b_{i+1,x} + b_{i+1,y} - b_{i,x} - b_{i,y} - b_{i+1,x}}{b_{i,x} + b_{i,y} + b_{i+1,x}} = \frac{b_{i,x} + (b_{i,y} - b_{i+1,y})}{b_{i,x} + b_{i,y} + b_{i+1,x}}.$$

Hence in this situation the convection in the y-direction has an influence that terminates the structure. As long as we have $b_{i,y} \approx b_{i+1,y}$ or more precise

$$\frac{b_{i,y} - b_{i+1,y}}{b_{i,x} + b_{i,y} + b_{i+1,x}} \ll \frac{b_{i,x}}{b_{i,x} + b_{i,y} + b_{i+1,x}}$$

we are still close to $p = q$.

The characteristics as mentioned above motivate two ideas:

1. If we consider the results of Lemma 9.1.12, then $b_3 \geq b_2$ seems a feasible heuristic. Additionally this implies the same rule for the aggregation as the results of section 8.3.

2. Another numerical idea is to determine only $p \in [0, 1]$ and to set $q = p$. It is obvious that if the modification must hold $p \neq q$ to fulfill the invariance, we lose this characteristic.

We want to look briefly at the structure of the matrices we get from the second idea mentioned above. W.l.o.g. we consider only the situation that two points are aggregated to a new one. More general results are obtained by the iterative use of the arguments.

We consider a matrix $A \in \mathbb{R}^{n\times n}$ that is an irreducible diagonal dominant L-matrix. Based on proposition 9.1.6 A is an M-matrix. We will show that if certain conditions are fulfilled then the assumed characteristics of the matrix are also true for $A_{0,X}$ and hence $A_{0,X}$ is an M-matrix based on the same proposition. It is $A_0 \in \mathbb{R}^{(n-1)\times(n-1)}$ and w.l.o.g. we assume that $\mathcal{N}^1_{n-1}, \mathcal{N}^1_n$ are aggregated to $\mathcal{N}^0_{n-1}$. Moreover, we assume that $\mathcal{N}^1_{n-2}$ is used to modify the prolongation. The structure of R, P_X follows as

(9.3)

$$R = \begin{pmatrix} I_{n-2} & & \\ & 1 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} I_{n-3} & & & \\ & 1 & & \\ & p & 1 & -p \\ & & & 1 \end{pmatrix} \quad \text{and} \quad P_X = \begin{pmatrix} I_{n-3} & & \\ & 1 & \\ & p & 1-p \\ & & 1 \end{pmatrix}.$$

To prove that $A_{0,X}$ has the same structure as A we consider three types of rows:

$j \le n-3$: For $j \le n-3$ the row j-th row of $A_{0,X}$ is

$$\begin{aligned}(A_{0,X})_{j,.} &= \left(a^{0,X}_{j,1}, \ldots, a^{0,X}_{j,n-3}, a^{0,X}_{j,n-2}, a^{0,X}_{j,n-1}\right) \\ &= \left(a_{j,1}, \ldots, a_{j,n-3}, p\, a_{j,n-1} + a_{j,n-2}, (1-p)a_{j,n-1} + a_{j,n}\right)\end{aligned}$$

It follows for $p \in [0,1]$

$$a^{0,X}_{j,j} = a_{j,j} > 0, \qquad a^{0,X}_{j,k} \le 0 \quad \text{for} \quad k \ne j$$

$$\begin{aligned}\text{and} \quad \sum_{i=1, i\ne j}^{n-1} |a^{0,X}_{j,i}| &= \left(\sum_{i=1, i\ne j}^{n-3} |a_{j,i}|\right) + |p\, a_{j,n-1} + a_{j,n-2}| + |(1-p)a_{j,n-1} + a_{j,n}| \\ &= \left(\sum_{i=1, i\ne j}^{n-3} |a_{j,i}|\right) + |p\, a_{j,n-1}| + |a_{j,n-2}| + (1-p)|a_{j,n-1}| + |a_{j,n}| \\ &= \sum_{i=1, i\ne j}^{n} |a_{j,i}| \le a_{j,j} = a^{0,X}_{j,j}.\end{aligned}$$

$j = n-2$: The $(n-2)$-th row of $A_{0,X}$ is

$$(A_{0,X})_{n-2,.} = \left(a_{n-2,1}, \ldots, a_{n-2,n-3}, p\, a_{n-2,n-1} + a_{n-2,n-2}, (1-p)a_{n-2,n-1} + a_{n-2,n}\right)$$

It follows obviously $a^{0,X}_{n-2,j} \le 0$ for $j \ne n-2$ and $p \in [0,1]$. Moreover we obtain from the weak diagonal dominance of A the inequality

$$p\,|a_{n-2,n-1}| \le a_{n-2,n-2}$$

for $p \in [0,1]$. This implies $a^{0,X}_{n-2,n-2} > 0$ for $p \in [0,1)$. Furthermore we obtain from the weak diagonal dominance of A for $p \in [0,1]$

$$\begin{aligned}\sum_{i=1,i\neq n-2}^{n-1} |a^{0,X}_{n-2,i}| &= \left(\sum_{i=1}^{n-3} |a_{n-2,i}|\right) + |(1-p)a_{n-2,n-1} + a_{n-2,n}| \\ &= \sum_{i=1,i\neq n-1}^{n} |a_{n-2,i}| - p|a_{n-2,n-1}| \\ &\leq a_{n-2,n-2} - p\,|a_{n-2,n-1}| = a^{0,X}_{n-2,n-2}.\end{aligned}$$

Hence the $(n-2)$-th row of $A_{0,X}$ holds the condition for the weak diagonal dominance.

$j = n-1$: The $(n-1)$-th row of $A_{0,X}$ is

$$\begin{aligned}(A_{0,X})_{n-1,.} = \Big(&a_{n-1,1} + a_{n,1}, \ldots, a_{n-1,n-3} + a_{n,n-3}, \\ &a_{n-1,n-2} + a_{n,n-2} + p\,(a_{n-1,n-1} + a_{n,n-1}), \\ &a_{n-1,n} + a_{n,n} + (1-p)(a_{n-1,n-1} + a_{n,n-1})\Big).\end{aligned}$$

The inequality

$$a^{0,X}_{n-1,j} \leq 0 \quad \text{for} \quad j \leq n-3$$

follows immediately from the representation above.

Furthermore we obtain

$$a^{0,X}_{n-1,n-2} \leq 0 \quad \Leftrightarrow \quad a_{n-1,n-2} + a_{n,n-2} + p(a_{n-1,n-1} + a_{n,n-1}) \leq 0. \tag{9.4}$$

As we have assumed that $\mathcal{N}^1_{n-2}$ is used to modify the prolongation it follows $a_{n-1,n-2} < 0$. Therewith the inequality (9.4) holds if p is small enough. As an example we consider $p = \frac{|a_{n-1,n-2}|}{a_{n-1,n-1}+|a_{n,n-1}|}$ as done for the invariance in chapter 5 then it follows

$$\begin{aligned}&0 \geq a_{n-1,n-2} + a_{n,n-2} + p(a_{n-1,n-1} + a_{n,n-1}) \\ \Leftrightarrow\quad &|a_{n-1,n-2}| + |a_{n,n-2}| \geq p(a_{n-1,n-1} + a_{n,n-1}) \\ \Leftrightarrow\quad &(|a_{n-1,n-2}| + |a_{n,n-2}|)(a_{n-1,n-1} + |a_{n,n-1}|) \geq |a_{n-1,n-2}|(a_{n-1,n-1} - |a_{n,n-1}|) \\ \Leftarrow\quad &|a_{n,n-2}|(a_{n-1,n-1} + |a_{n,n-1}|) \geq 0 \\ &\text{and} \quad |a_{n-1,n-2}|(a_{n-1,n-1} + |a_{n,n-1}|) \geq |a_{n-1,n-2}|(a_{n-1,n-1} - |a_{n,n-1}|).\end{aligned}$$

Hence we have $a^{0,X}_{n-1,n-2} \leq 0$ in this case.

For the diagonal element $a^{0,X}_{n-1,n-1}$ of $A_{0,X}$ we obtain

$$\begin{aligned} & 0 < a^{0,X}_{n-1,n-1} \\ \Leftrightarrow \quad & 0 < a_{n-1,n} + a_{n,n} + (1-p)(a_{n-1,n-1} + a_{n,n-1}). \end{aligned}$$

This inequality is fulfilled by several assumptions, too. If it is $A = A^T$ then we obtain

$$a_{n,n} \geq |a_{n,n-1}| = |a_{n-1,n}| \quad \text{and} \quad a_{n-1,n-1} \geq |a_{n-1,n}| = |a_{n,n-1}|.$$

And for $n \geq 3$ it follows from the irreducibility of A

$$a_{n,n} > |a_{n,n-1}| \quad \text{or} \quad a_{n-1,n-1} > |a_{n-1,n}|.$$

Hence the inequality $0 < a^{0,X}_{n-1,n-1}$ is true for $p \in [0,1]$.
For a matrix A which is not necessarily symmetric we consider the situation again for $p = \frac{|a_{n-1,n-2}|}{a_{n-1,n-1}+|a_{n,n-1}|}$. We obtain

$$\begin{aligned} & (a_{n-1,n-1} + |a_{n,n-1}|)a^{0,X}_{n-1,n-1} \\ & \quad = (a_{n-1,n-1} + |a_{n,n-1}|)(a_{n-1,n-1} - |a_{n,n-1}|) \\ & \qquad - (a_{n-1,n-1} + |a_{n,n-1}|)\frac{|a_{n-1,n-2}|}{a_{n-1,n-1} + |a_{n,n-1}|}(a_{n-1,n-1} - a_{n,n-1}) \\ & \qquad + (a_{n-1,n-1} + |a_{n,n-1}|)(a_{n-1,n-1} - a_{n,n-1}) \\ & \quad = (a_{n-1,n-1} + |a_{n,n-1}| - |a_{n-1,n-2}|)(a_{n-1,n-1} - |a_{n,n-1}|) \\ & \qquad + (a_{n,n} - |a_{n-1,n}|)(a_{n-1,n-1} + |a_{n,n-1}|) \\ & \quad = a_{n-1,n-1}\big(a_{n-1,n-1} - |a_{n-1,n-2}| - |a_{n-1,n}|\big) + |a_{n,n-1}|\big(a_{n,n} - |a_{n,n-1}|\big) \\ & \qquad + \big(a_{n,n}a_{n-1,n-1} - |a_{n-1,n}||a_{n,n-1}|\big). \end{aligned}$$

Based on the weak diagonal dominance the first and second bracket are non negative. For the third bracket we obtain

$$a_{n,n}a_{n-1,n-1} - |a_{n-1,n}||a_{n,n-1}| > 0$$

because

$$a_{n,n} = |a_{n,n-1}| \quad \text{and} \quad a_{n-1,n-1} = |a_{n-1,n}|$$

contradicts again the irreducibility of A.

For the condition concerning the weak diagonal dominance of $A_{0,X}$ we only consider the assumption $p = \frac{|a_{n-1,n-2}|}{a_{n-1,n-1}+|a_{n,n-1}|}$. In this case we have $a^{0,X}_{n-1,k} \leq 0$ for $k \leq n-2$. Thus we obtain

$$\begin{aligned}
\sum_{i=1}^{n-2} |a^{0,X}_{n-1,i}| &= \sum_{i=1}^{n-2} -a^{0,X}_{n-1,i} \\
&= \sum_{i=1}^{n-3} -(a_{n-1,i} + a_{n,i}) - (a_{n-1,n-2} + a_{n,n-2} + p\,(a_{n-1,n-1} + a_{n,n-1})) \\
&= \sum_{i=1}^{n-2} -(a_{n-1,i} + a_{n,i}) - p\,a_{n-1,n-1} - p\,a_{n,n-1} \\
&\leq a_{n-1,n-1} - |a_{n-1,n}| + a_{n,n} - |a_{n,n-1}| - p\,a_{n-1,n-1} - p\,a_{n,n-1} \\
&\leq (1-p)a_{n-1,n-1} + a_{n-1,n} + a_{n,n} + (1-p)a_{n,n-1} = a^{0,X}_{n-1,n-1}.
\end{aligned}$$

Finally we want to highlight that for $p = \frac{|a_{n-1,n-2}|}{a_{n-1,n-1}+|a_{n,yn-1}|}$ it follows that $A_{0,X}$ is irreducible diagonal dominant from the same arguments as in Proposition 9.1.11. We sum up this result in the following proposition.

Theorem: 9.1.15. *Assume that $A \in \mathbb{R}^{n\times n}$ is an irreducible diagonal dominant L-matrix. Assume that R, P_X are as defined in (9.3) and we set $p = \frac{|a_{n-1,n-2}|}{a_{n-1,n-1}+|a_{n,n-1}|}$. Then $A_{0,X}$ is also an irreducible diagonal dominant L-matrix.*

proof. See the calculations above the proposition. □

9.2 Numerical experiments

In this section we want to present and discuss some numerical results. We will compare the preconditioner $C^{-1}_{BPX}, C^{-1}_{DT}, C^{-1}_{2P}$ and the associated modified options. As in the analytical consideration we will concentrate on the operators $C^{-1}_{BPX}, C^{-1}_{DT}$. To compare the methods we will consider the number of iterations needed to solve the equation (Itter), the time for one iteration (t_{It}, msec.), the time for the setup (t_{set}, sec.) and the time for the total algorithm (t, sec.). In doing so we have to highlight the following: The main aspect of this paper is to make theoretical assertions on numerical methods and not to give the best possible implementation method. It is possible to get a good

idea of which modifiactions or algorithms imply more effort. Furthermore we only use the $GMRes$-method preconditioned with the different operators. That means that for symmetric stiffness matrics we drop the possibility to use the CG-method if the preconditioner is symmetric, too. We will discuss this more in-depth in the section for symmetic problems.
Furthermore we highlight that there are possibilities to parallelize some steps. We will discuss the possibilities therefore at the end of the section.

For all experiments we set $\Omega = [0,1] \times [0,1]$. h is the step width and $n_x = n_y = \frac{1}{h} - 1$ the number of grid points in the direction of x, y. We use $n_x = n_y = 2^8 - 1$. Thus we obtain $n = 255^2 = 65025$ grid points. Moreover we stop the iteration if the condition

$$\|A\, x^k - f\| \leq \delta = 10^{-8} \quad \text{is fulfilled.}$$

Furthermore we only aggregate only two grid points to a new one as done in all calculations. The rule concerning which points are aggregated follows the arguments of chapter 8. Hence we classify the possible aggregations by the relative size of the links to the neighbours. We set $J = 15$. Hence we use 16 different grids.

9.2.1 The unsymmetric model problem

In this section we will consider three different problems based on the convection diffusion equation

$$b_1(x,y)\frac{\partial u}{\partial x} + b_2(x,y)\frac{\partial u}{\partial y} - \varepsilon \triangle u(x,y) = f(x,y) \quad \forall (x,y) \in \Omega \tag{9.5}$$

$$u(x,y) = g(x,y) \quad \forall (x,y) \in \partial\Omega.$$

In all problems we set $\varepsilon = const, f \equiv 1$ and $g \equiv 0$. To obtain the stiffness matrix we use the method of the finite differences and the upwind method.

To obtain the different problems we define the functions b_1, b_2 on the grid points as follows:

PS1:

$$b_1(x,l) = \frac{-(n_y+1)/2 + l}{n_y}$$

$$\text{and} \quad b_2(k,y) = \frac{-k + (n_x+1)/2}{n_x}$$

for $l = 1, \ldots, n_y$ and $k = 1, \ldots, n_x$. (Hence the constant solutions are given by circles. Cf. Figure 9.1)

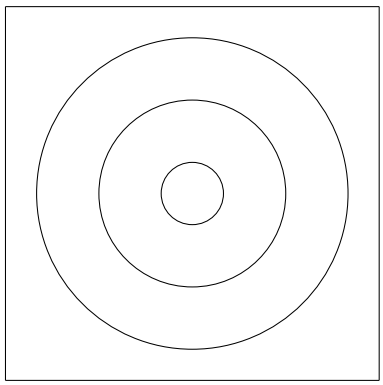


Figure 9.1:

PS2: $b_1 \equiv 1, \quad b_2 \equiv 0.$

PS3: $b_1(k,y) = \frac{n_x}{k}$, for $k = 1, \ldots n_x$ and $b_2 \equiv 0$.

For the modifications we set

$X = 0$: No modification. $X_j = I_j$ for $j = 1, \ldots, J$.

$X = 1$: $X_j = D_{A_j}^{-1}$ for $j = 1, \ldots, J$. (The main diagonal.)

$X = 2$: Modification by the inverse of blocks. We only invert the blocks of the dimension 2×2. (Cf. section 5.1.1)

$X = 3$: Modification by the inverse of blocks. We invert the blocks of the dimension 2×2 and of the dimension 1×1. (Cf. section 5.1.1)

$X = 4$: „exact“ modification as presented in Proposition 5.1.11 for the arbitrary situation. For $\varepsilon = 0$ and the one dimensional problem this is proved as a perfect choice concerning the angle (Cf. section 5.1.1.)

$X = 5$: Modification for symmetric matrices. The modification is introduced for symmetric matrices but it can be used for other matrices as well (Cf. section 5.2).

$X = 6$: Modification as motivated in section 9.1 based on M-matrix properties.

For the matrices $B^{(j)}$ that should approximate A_j we set ν iterations of the Jacobi-method ($meth. = Jac$) or the symmetric Gauss-Seidel-method ($meth. = SSOR$).

We start with the tables 9.1 and 9.2. In these tables we present the results for the problem $PS1$ for all the three preconditioners and the six different kinds of modification.

Problem: $PS1$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-8}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	708	290	120	843	435	244	1024	454	280	1082
1	807	321	132	970	454	204	1147	450	308	1183
2	808	302	124	953	397	200	1078	404	284	1119
3	809	306	120	958	400	204	1081	382	284	1092
4	935	152	148	984	130	228	985	127	352	996
5	1046	116	172	1081	103	256	1085	110	364	1100
6	935	137	144	977	130	240	987	136	304	935

Table 9.1:

First we consider the unmodified methods. It is obvious to see that for this complex problem the BPX-method is more effective than the other methods. This is is not suprising as we have proved in section 3.6 hat

$$\frac{d_{DT}}{c_{DT}} \leq \frac{d_{BPX}}{c_{BPX}} \quad \Leftrightarrow \quad \gamma_{DT} \leq \sqrt{1/2}. \tag{9.6}$$

Hence the BPX-method is more robust than the DT-method. Furthermore, we see that the effort for the $2P$-method is higher than for the DT-method and this one is higher than the effort for the BPX-method. This is based on the projections that are needed. Moreover we see that in this example the modifications $X = 1, 2, 3$ almost have no influence on the number of iterations. The modifications $X = 4, 5, 6$ work well on all preconditioners. This is obvious by the number of the iterations in both tables 9.1, 9.2. Unfortuantly we also see that the effort to determine the modification is to high to obtain a faster method than the unmodified one. Moreover, we highlight that the effect of the modifications $X = 4, 5, 6$ is much higher for the DT-method and the $2P$-method than for the BPX-method. This result suggests that the DT-method is more sensitive with respect to the angle. In the figures 9.2 and 9.3 we consider the derivation of d/c with respect to $\mu_{\gamma_{DT}}$ for the DT and the BPX-method. In the figures

Problem: $PS1$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = SSOR$, $\nu = 1, \delta = 10^{-8}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	706	137	156	749	180	240	787	406	332	1032
1	807	138	152	851	179	236	887	390	316	1111
2	808	142	152	853	206	236	907	375	316	1095
3	808	141	156	853	201	240	904	364	316	1083
4	935	96	192	965	84	276	966	119	352	993
5	1046	89	232	1076	68	316	1073	103	396	1104
6	934	96	188	963	84	276	966	126	352	998

Table 9.2:

9.4 and 9.5 we do the same for the deviation with respect to γ_{DT}. The result is

$$\frac{d}{d\mu_{\gamma_{DT}}}\left(\frac{d_{DT}}{c_{DT}}\right) > \frac{d}{d\mu_{\gamma_{DT}}}\left(\frac{d_{BPX}}{c_{BPX}}\right) \quad \forall \mu_{\gamma_{DT}} > 0$$

$$\Leftrightarrow \quad \frac{d}{d\gamma_{DT}}\left(\frac{d_{DT}}{c_{DT}}\right) > \frac{d}{d\gamma_{DT}}\left(\frac{d_{BPX}}{c_{BPX}}\right) \quad \forall \gamma_{DT} \in (0,1).$$

The equivalence follows from

$$\frac{d}{d\gamma_{DT}}\left(\frac{d_{DT}}{c_{DT}}\right) = \frac{d}{d\mu_{\gamma_{DT}}}\left(\frac{d_{DT}}{c_{DT}}\right) \underbrace{\left(\frac{d}{d\gamma_{DT}}\mu_{ga}\right)}_{>0}.$$

The result of this consideration is obvious. The DT method is more sensitive with respect to the angle than the BPX method. And the bigger γ_{DT} is, the bigger is the difference between the methods.

If we compare the three modifications $X = 4, 5, 6$ more in-depth then we see that $X = 5$ always has the lowest number of iterations, but also the highest effort. The higher effort results as we have to determine two directions in which a modification is done. And the lower number of iterations result as information of two directions are used to modify the system. The methods $X = 4$ and $X = 6$ are more or less equal for this example. At last we want to highlight that the relations are the same if we raise

ν.

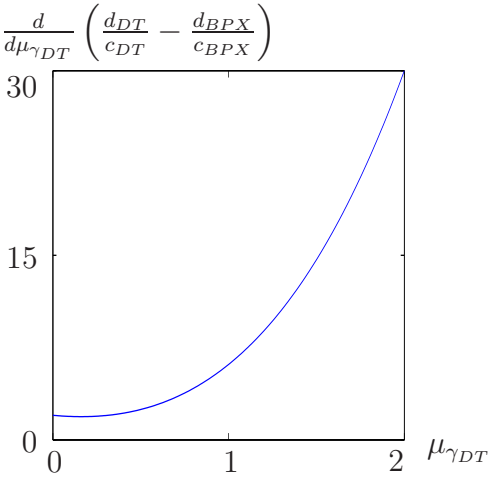


Figure 9.2:

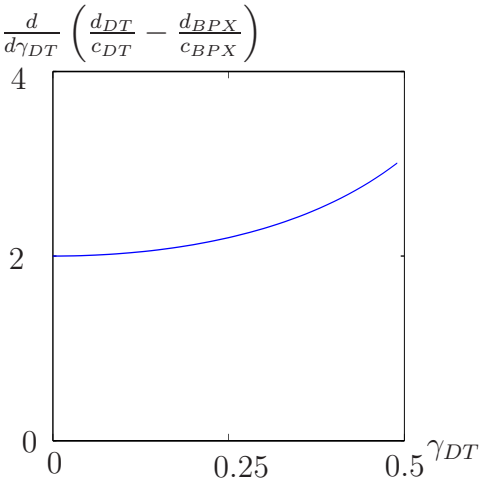


Figure 9.3:

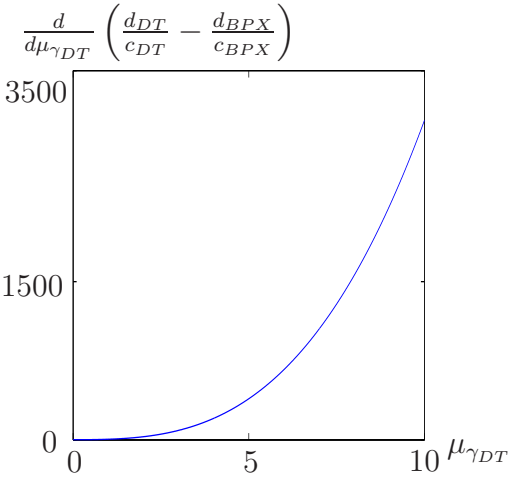


Figure 9.4:

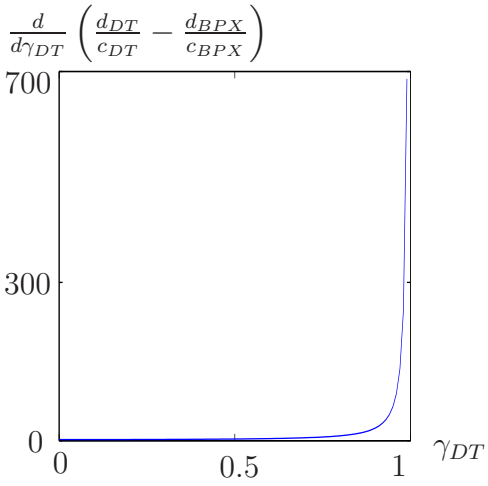


Figure 9.5:

Next we consider the results in the tables 9.3 and 9.4 for the problem $PS2$. It is again the case that the modifications $X = 1, 2, 3$ have no effect on the number of iterations. Again the methods $X = 4, 5, 6$ work in a way that the number of iterations reduces. We highlight that especially the modification $X = 4$ is more or less constructed for exactly this situation. If we now take a closer look at the modifications $X = 4, 5, 6$ then we see that the method $X = 5$ does not have the lowest number of iterations anymore. Probably this is based on the fact that $X = 5$ uses two directions for each

Problem: $PS2$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-8}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	719	78	128	737	101	204	752	114	276	766
1	819	108	128	819	109	200	855	113	320	869
2	819	118	120	850	111	208	858	112	316	868
3	818	118	120	849	112	200	856	117	273	868
4	950	45	144	959	33	220	956	40	296	961
5	1064	42	168	1073	38	248	1075	40	324	1079
6	947	43	140	956	45	224	960	59	304	969

Table 9.3:

Problem: $PS2$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = SSOR$, $\nu = 1, \delta = 10^{-8}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	718	36	152	725	44	264	732	89	308	755
1	819	46	168	830	43	228	831	93	308	858
2	819	39	152	827	35	232	829	110	348	869
3	819	41	152	827	32	228	827	118	336	877
4	948	31	184	955	20	276	953	41	348	964
5	1064	34	228	1073	19	308	1070	40	424	1083
6	948	28	188	954	35	312	960	58	344	927

Table 9.4:

grid point that is modified. But in this example the causality is mainly given by one direction (the second direction has only an influence based on the diffusion). Hence the second direction in the modification makes effort and has no important causality. Furthermore we see in particular in table 9.4 that $X = 6$ is no longer equal to $X = 4$. Thus for the DT-method $X = 4$ is a better choice. We remember that $X = 6$ is based on numerical ideas for the iterative methods used on subspaces. We can assume that in this example and in particular for $meth = SSOR$ the solutions are good. Also for $X = 4$. Hence the effect of the right angle is more important.

Now we consider in the table 9.5 and 9.6 results for $PS3$. We remember that $PS3$ represents is a one dimensional flux without the condition of a constant b.

Problem: $PS3$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-8}$										
		C^{-1}_{BPX}			$C^{-1}_{DT,1}$			$C^{-1}_{2P,1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	752	120	132	785	124	216	799	507	292	1212
1	861	125	136	896	108	212	898	116	292	911
2	858	104	132	884	90	212	887	372	320	1140
3	858	106	144	886	91	216	887	92	324	897
4	971	69	148	987	52	232	986	76	332	1003
5	1083	78	172	1103	60	256	1102	215	332	1212
6	972	70	148	988	61	232	990	484	308	1405

Table 9.5:

For the modifications $X = 4, 5, 6$ there is nothing new. But for this problem the modifications $X = 1, 2, 3$ also have a positive effect on the number of iterations. In particular if we consider the methods DT and $2P$ which are more sensitive with respect to the angle than the BPX method we see that the number of iterations reduces for $X = 3$. This is more obvious for $meth = SSOR$. This may result from the fact that the solution on the subspace is in this case better than for $meth = Jac$. Hence the effect of the angle is more important.

Now we are going to take a look at the methods $C^{-1}_{DT,2}, C^{-1}_{2P,2}$. We remember that these

Problem: $PS3$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = SSOR$, $\nu = 1, \delta = 10^{-8}$										
		C^{-1}_{BPX}			$C^{-1}_{DT,1}$			$C^{-1}_{2P,1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	752	44	168	762	64	256	774	417	336	1102
1	861	48	172	872	55	252	878	100	328	906
2	858	42	168	867	41	252	871	190	324	965
3	858	42	168	868	34	256	868	91	356	901
4	971	37	200	980	34	284	982	83	392	1012
5	1083	39	236	1094	27	320	1093	110	396	1141
6	971	36	200	980	41	284	986	447	400	1375

Table 9.6:

methods are based on the idea to obtain an orthogonality in the multigrid situation independent of the condition (2.14).
In the tables 9.7 and 9.8 we see that the results are always worse than for $C^{-1}_{DT,1}, C^{-1}_{2P,1}$. Worse means that we have a higher effort and a higher number of iterations. Perhaps this will be better if we use the matrices $\widehat{S}_k, \widehat{S}_k^{-1}$ also to construct the coarser operators A_k and not only for the projections (Cf. the scaled tentative prolongator in [GJV08]). This idea will not be outlined in the current work.

We will conclude this section with two remarks. First is that if the stop criterion becomes harder then the modifications $X = 4, 5, 6$ should become better if we consider the total time of the algorithm. This is because the number of iterations raises for all methods and the setting time is fixed independent of the number of iterations. We present this for the problem $PS3$. We set

$$\|A\,x^k - f\| \le \delta = 10^{-12} \quad \text{or} \quad \|A\,x^k - f\| \le \delta = 10^{-15}$$

$$\text{instead of} \quad \|A\,x^k - f\| \le \delta = 10^{-8}$$

as used befor. In the tables 9.9 and 9.11 we present the results for $meth = Jac$. Hence this tables should be compared with the table 9.5. In the tables 9.10 and 9.12 we present the results for $meth = SSOR$. Thus this tables should be compared with the table 9.6.

Problem: $PS1$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-8}$										
		C_{BPX}^{-1}			$C_{DT,2}^{-1}$			$C_{2P,2}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	705	290	120	840	436	228	1034	450	324	1091
1	805	321	128	970	462	212	1160	474	296	1223
2	806	302	124	952	402	236	1087	424	332	1158
3	806	306	124	956	401	224	1090	409	316	1136
4	984	152	164	932	133	260	987	133	328	998
5	1044	116	172	1080	104	304	1088	112	340	1097
6	932	137	152	975	130	260	986	133	316	996

Table 9.7:

Problem: $PS1$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = SSOR$, $\nu = 1, \delta = 10^{-8}$										
		C_{BPX}^{-1}			$C_{DT,2}^{-1}$			$C_{2P,2}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	704	137	168	749	180	268	792	405	348	1043
1	805	138	152	848	182	240	889	409	328	1138
2	806	142	152	851	218	280	922	396	392	1137
3	806	141	164	853	209	244	909	382	324	1104
4	932	96	188	961	87	332	969	124	392	1000
5	1044	89	256	1076	68	320	1071	106	436	1105
6	932	96	188	961	86	320	969	124	392	999

Table 9.8:

In both cases we see that $\delta = 10^{-12}$ has no large influence on the total time the algorithm needs. Hence there is no difference to $\delta = 10^{-8}$. For $\delta = 10^{-15}$ we see that a good modification makes the preconditioner more robust. (Additionally we should highlight that we stop all methods if we have done 1000 iterations.)

Problem: $PS3$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-12}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	752	173	132	810	178	248	841	726	316	1611
1	858	187	152	926	165	212	926	176	292	946
2	856	154	140	905	141	212	909	558	300	1407
3	856	155	144	906	142	212	909	144	316	926
4	968	104	160	998	81	252	996	115	336	1022
5	1080	121	184	1119	94	276	1117	335	352	1332
6	968	107	156	999	95	268	1005	685	332	1761

Table 9.9:

The second aspect is the behaviour of the modifications if the convection shrinks compared to the diffusion. For this aspect we have to differ between the problems.

In the tables 9.13, 9.14 we consider the problem $PS1$ for $meth = Jac$, $b = 1$ and $\varepsilon = 2^{-8}, 2^{-5}$. Hence these tables should be compared with the table 9.1. If we consider the unmodified operators we see that the problem becomes easier if ε grows. If we consider the modified methods with $X = 4, 5, 6$ then we see that this effect is weaker for these methods. This implies that the effect of the modification is smaller for a bigger ε. Hence the modified preconditioners are more robust concerning a small diffusion as the unmomdified methods. In particular for $X = 4, 6$ the number of iterations is almost constant. For $X = 5$ it occurs the effect that with the bigger diffusion there is more than one direction that has an influence on the behaviour of the system. Hence for a bigger ε we have for $X = 5$ again a lower number of iterations as needed for $X = 4, 6$.

In the tables 9.15, 9.16 we consider the problem $PS3$ for $meth = Jac$, $b = 1$ and $\varepsilon = 2^{-8}, 2^{-5}$. Hence this tables should be compared to the table 9.5. Considering the

Problem: $PS3$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = SSOR$, $\nu = 1, \delta = 10^{-12}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	752	67	184	770	99	272	790	591	356	1383
1	858	73	172	877	84	260	888	152	356	939
2	856	62	172	871	63	260	877	267	344	1032
3	856	61	172	871	52	260	872	141	340	927
4	968	56	200	983	49	288	985	124	372	1032
5	1080	60	248	1098	40	328	1095	159	408	1175
6	968	56	200	984	62	296	991	632	376	1684

Table 9.10:

Problem: $PS3$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-15}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	752	1000	132	2087	1000	236	2193	1000	288	2240
1	858	231	132	951	208	232	959	231	288	993
2	855	1000	128	2189	1000	252	2290	1000	324	2350
3	855	1000	132	2192	1000	212	2277	1000	288	2345
4	968	131	168	1009	102	232	1003	1000	308	2484
5	1080	158	176	1138	1000	272	2537	1000	332	2621
6	968	1000	148	2324	1000	232	2397	1000	304	2479

Table 9.11:

Problem: $PS3$ with $b = 1, \varepsilon = 2^{-15}$										
$meth = SSOR$, $\nu = 1, \delta = 10^{-15}$										
		C^{-1}_{BPX}			$C^{-1}_{DT,1}$			$C^{-1}_{2P,1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	752	1000	172	2125	1000	268	2224	1000	356	2310
1	858	93	168	884	105	260	899	193	344	968
2	855	1000	172	2225	1000	260	2320	1000	344	2399
3	855	1000	172	2226	1000	260	2318	1000	340	2397
4	967	71	200	987	61	288	990	1000	372	2539
5	1079	1000	236	2517	1000	328	2611	1000	408	2693
6	969	70	220	990	81	288	1000	1000	368	2545

Table 9.12:

Problem: $PS1$ with $b = 1, \varepsilon = 2^{-8}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-8}$										
		C^{-1}_{BPX}			$C^{-1}_{DT,1}$			$C^{-1}_{2P,1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	708	278	124	834	320	204	895	320	276	921
1	808	295	120	948	333	204	1008	332	280	1032
2	808	264	128	925	294	204	971	318	288	1018
3	809	264	128	920	289	200	967	290	280	990
4	934	140	144	978	129	224	983	129	304	993
5	1027	83	160	1048	69	248	1050	92	324	1067
6	935	144	144	980	141	224	989	157	304	1012

Table 9.13:

Problem: $PS1$ with $b = 1, \varepsilon = 2^{-5}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-8}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	707	196	128	777	204	204	798	211	280	818
1	807	198	128	880	202	204	897	202	284	913
2	808	201	120	882	204	204	899	220	288	929
3	808	206	120	884	210	204	903	209	280	920
4	936	114	144	967	112	224	976	116	304	987
5	1026	72	160	1044	63	244	1046	73	324	1056
6	936	122	144	971	129	232	985	137	312	999

Table 9.14:

unmodified methods we see that in this case the problem becomes more comlex if ε grows. But also in this case we see that this effect is smaller for the modified methods and in particular for the modifications $X = 4, 5, 6$. Based on the same arguments as for $PS1$ we observe the lowest effect for $X = 5$. This is probably again based on the causality of a second direction.

Remarks on parallelisation

As the three preconditioners $C_{BPX}^{-1}, C_{DT}^{-1}$ and C_{2P}^{-1} are all additive methods we can parallelize the multiplication

$$C_i^{-1}\, v$$

in that way that we calculate

$$(B^{(J)})^{-1}\, R_J\, v, \ldots, (B^{(1)})^{-1}\, R_1\, v, (B^{(0)})^{-1}\, R_0\, v$$

$$\text{or} \quad (B^{(J)})^{-1}\, (I_J - Q_{J-1})\, R_J\, v, \ldots, (B^{(1)})^{-1}\, (I_1 - Q_0)\, R_1, (B^{(0)})^{-1}\, R_0\, v$$

at the same time. It is obvious that this makes all preconditioners faster.

As mentioned at the beginning of this section we set for $B^{(j)}$ ν iterations of the Jacobi method or the $SSOR$ method. In doing so it is obvious that for the Jacobi method

Problem: $PS3$ with $b = 1, \varepsilon = 2^{-8}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-8}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	752	172	156	814	265	216	894	404	296	1066
1	859	174	160	923	226	212	968	228	292	988
2	856	152	140	904	175	212	930	437	292	1213
3	856	151	136	903	173	216	929	172	292	942
4	967	75	152	985	64	232	987	68	308	994
5	1075	67	176	1092	52	256	1091	128	328	1137
6	967	80	152	987	80	232	994	246	304	1115

Table 9.15:

Problem: $PS3$ with $b = 1, \varepsilon = 2^{-5}$										
$meth = Jac$, $\nu = 1, \delta = 10^{-8}$										
		C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	t_{set}	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	754	267	136	875	342	212	963	350	316	999
1	853	257	128	966	297	212	1022	297	328	1046
2	855	231	128	949	258	212	989	331	292	1082
3	853	214	132	935	235	212	969	235	292	987
4	968	121	148	1003	115	232	1006	124	312	1019
5	1065	82	168	1087	68	260	1085	84	328	1099
6	968	125	148	1005	133	232	1014	193	300	1066

Table 9.16:

we can parallize the calculation for the different grid points. If we have enough kernels then its also possible to use for each grid point its own kernel. For the $SSOR$-method this is not possible. Additionally as concerns it is difficult to check the relation between the numeration of the grid points on coarser grids and the geometrical structure. This would result in an additional effort.

9.2.2 The symmetric model problem

In this section we are going to consider the problem based on the equation

$$-div(\alpha(x,y)\, grad\, u(x,y)) = f(x,y), \quad \forall (x,y) \in \Omega \tag{9.7}$$
$$u(x,y) = g(x,y), \quad \forall (x,y) \in \partial\Omega$$

with

$$\alpha(x,y) = \begin{pmatrix} a(x,y) & 0 \\ 0 & b(x,y) \end{pmatrix}$$

We set $f \equiv 1$ and $g \equiv 0$. To obtain the associated stiffness matrix we use the method of the finite elements.

In our example the functions $a(x,y) = b(x,y)$ are set constant on the single elements. Furthermore the constants for the function $a(x,y)$ are increasing in the diagonal through the unit square. More exactly we set

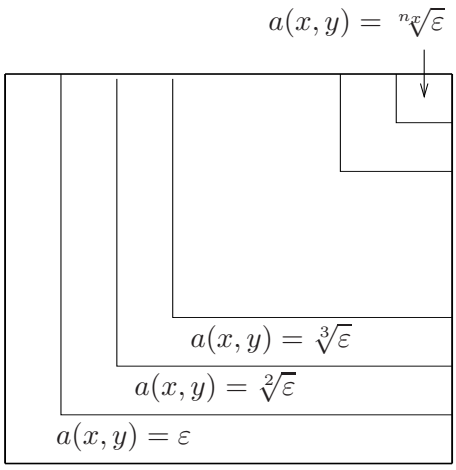


Figure 9.6:

$$a(x,y) = \sqrt[k]{\varepsilon} \quad \text{for} \quad x \le y \quad \text{and} \quad (k-1)\,h \le x < k\,h$$
$$\text{or} \quad y < x \quad \text{and} \quad (k-1)\,h \le y < k\,h \quad \text{(Cf. figure 9.6).}$$

For the modification we use the idea as presented in section 5.3. Hence we calculate the eigenvectors of the blocks of the dimension 2×2. In section 5.3 we have also seen that they have for $\varepsilon, \delta \to 0$ the limits

$$v_1 = (1,\ 1)^T \quad \text{and} \quad v_2 = (-1,\ 1)^T .$$

If we use v_1 in the modification then we say that the modification is based on the long waves. If we use v_2 then we say that the modification is based on the short waves. And for the modifications we set

$X = 0$: No modification. $X_j = I_j$ for $j = 1, \ldots, J$.

$X = 7$: modification with the long waves. $\|v_1\| = 1$.

$X = 8$: modification with the long waves. $\|v_1\| = \sqrt{2}$.

$X = 9$: modification with the long waves, scaled with the eigenvalues. $\|v_1\| = \sqrt{2/\lambda_1}$.

$X = 10$: modification with the short waves. $\|v_2\| = 1$.

$X = 11$: modification with the short waves. $\|v_2\| = \sqrt{2}$.

$X = 12$: modification with the short waves, scaled with the eigenvalues. $\|v_2\| = \sqrt{2/\lambda_2}$.

Unfortunately we see in the tables 9.17 and 9.18 that the modification cause an additional effort and raise the number of iterations. Only if the solutions on the subspaces are very good as in the case of $meth = SSOR$ and $\nu = 10$ then we see that the modifications $X = 10, 11$ have a positive influence on the number of iterations. But as the idea is to use the CG-method instead of the $GMRes$-method for symmetric problems this result is quite worse because for a symmetric preconditioner we need symmetric matrices $B^{(j)}$. Hence we may only do one iteration of the Jacobi - method or the Symmetric Gauss-Seidel method. We will sum this up to the following two aspects:

1. The idea to scale the modification with the eigenvalues not a good idea.

2. If we use the two sided modification then the modification must use an approximation on the waves with short frequences.

To conclude the consideration of a two sided modification, we highlight that there are methods for which we obtain better results for the two sided modification. For the stiffness matrix as defined above we consider the BPX-method with prolongation

$meth = Jac,\ \nu = 1, \delta = 10^{-8}, \varepsilon = 2^{-15}$									
	C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	198	124	791	211	264	819	264	372	902
7	328	124	1134	327	256	1186	315	448	1254
8	460	128	1277	397	224	1248	382	404	1325
9	442	128	1255	1000	212	2391	1000	400	2628
10	1000	128	2294	649	244	1631	641	376	1726
11	1000	128	2302	650	212	1612	641	332	1690
12	1000	128	2294	1000	216	2392	1000	308	2479

Table 9.17:

$meth = SSOR,\ \nu = 10, \delta = 10^{-8}, \varepsilon = 2^{-15}$									
	C_{BPX}^{-1}			$C_{DT,1}^{-1}$			$C_{2P,1}^{-1}$		
X	Itter	t_{It}	t	Itter	t_{It}	t	Itter	t_{It}	t
0	90	884	809	107	956	835	261	1032	1071
7	160	884	1137	146	980	1135	311	1072	1417
8	178	880	1161	154	956	1141	368	1036	1512
9	183	872	1167	1000	976	3143	1000	1052	3252
10	216	868	1208	92	992	1066	224	1096	1326
11	213	880	1205	91	960	1062	225	1136	1330
12	216	868	1208	1000	956	3129	1000	1480	3553

Table 9.18:

and restriction operators which are based on the geometrical method (cf section 2.5). Modifications for this are presented in [Tic06]. The results are presented in table 9.19. We see that for this example the modified preconditioners need less iterations to solve the system of linear equations. In contrast to the algebraic methods, the setup time does not rise if we use the modification. This results as for geometrical methods the modification follows mainly from the geometrical structure. Hence there is no further effort to determine neighbours for the modification from the elements of the matrix.

We emphasize that we do not consider the DT or the $2P$ method as it is not as easy to determine the operator $S_X = (R_X\, P_X)^{-1}$ for the geometrical methods. This would induce effects we do not want to discuss in this thesis.

$C_{BPX^{-1}}$ with $J = 7$							
$\nu = 1, \delta = 10^{-8}, \varepsilon = 2^{-15}$							
		$meth = Jac$			$meth = SSOR$		
X	t_{set}	Itter	t_{It}	t_{ges}	Itter	t_{It}	t_{ges}
unmodified	841	44	76	847	35	100	845
modified	805	26	76	808	22	96	808

Table 9.19:

A Basics

We will sum up some elemantary definitions and results which do not fit to any chapter. For this section we assume the following setting:
Let V be a linear vector space and U, W linear subspaces. Let $(.,.)$ be a dotproduct and $\|v\| := \sqrt{(v,v)}$ be the associated vectornorm.

Proposition: A.0.1. *For all $v_1, v_2 \in V$ it is*

$$|(v_1, v_2)| \leq \|v_1\| \, \|v_2\|.$$

Definition: A.0.2. *If there is a $\gamma \in [0,1)$ with*

$$(u, w) \leq \gamma \, \|u\| \, \|w\| \quad \forall u \in U,\, w \in W$$

then we say that U, W hold a strengthened Cauchy-Schwarz Inequality.

Proposition: A.0.3. *For $a, b \in \mathbb{R}$ and $\varepsilon \in \mathbb{R}_+$ we have*

$$a\, b \leq \frac{a^2}{2\varepsilon} + \frac{\varepsilon b^2}{2}.$$

Lemma: A.0.4. *For $j = 1, \ldots, m$ let a^j, b^j be in $\mathbb{R}^{n_j}$. Let further $\gamma_j \leq 1$ be constants that fulfil*

$$(a^j,\, b^j) \leq \gamma_j \, \|a^j\| \, \|b^j\|$$

then we obtain for $a := (a^1, a^2, \ldots, a^m) \in \mathbb{R}^{n_1+\ldots,+n_m}$ and $b := (b^1, b^2, \ldots, b^m) \in \mathbb{R}^{n_1+\ldots,+n_m}$ the inequality

$$(a,\, b) \leq \gamma \, \|a\| \, \|b\|$$

with $\gamma = \max_{j=1,\ldots,m}\{\gamma_j\}$.

proof. We show the proposition by induction over m. For $m = 1$ the proposition follows by the assumption. So we show the induction step. Assume that the assertion is fulfilled for $m-1$. With

$$\widetilde{a} := (a^1, \ldots, a^{m-1}) \quad \text{and} \quad \widetilde{b} := (b^1, \ldots, b^{m-1})$$

it follows

$$\begin{aligned}(a,\,b)^2 &= ((\widetilde{a},a^m),\,(\widetilde{b},b^m))^2 = (\widetilde{a},\,\widetilde{b})^2 + (a^m,\,b^m)^2 + 2(\widetilde{a},\,\widetilde{b})(a^m,\,b^m)\\ &\le \gamma^2\,\|\widetilde{a}\|^2\,\|\widetilde{b}\|^2 + \gamma^2\,\|a^m\|^2\,\|b^m\|^2 + 2\gamma^2\,\|\widetilde{a}\|\,\|\widetilde{b}\|\,\|a^m\|\,\|b^m\|\\ &\le \gamma^2\,\|\widetilde{a}\|^2\,\|\widetilde{b}\|^2 + \gamma^2\,\|a^m\|^2\,\|b^m\|^2 + \gamma^2\left(\|\widetilde{a}\|^2\|b^m\|^2 + \|\widetilde{b}\|^2\,\|a^m\|^2\right)\\ &= \gamma^2\,\|(\widetilde{a},a^m)\|^2\,\|(\widetilde{b},b^m)\|^2 = \gamma^2\,\|a\|^2\,\|b\|^2.\end{aligned}$$

This shows the proposition for m. □

Lemma: A.0.5. *Let $v_1,\ldots,v_n \in V$ be orthogonal by pairs. That means*

$$(v_i,\,v_j) = 0 \quad \textit{for} \quad i \neq 0.$$

Then we obtain

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\|v_i\|\right)^2 \le \left\|\sum_{i=1}^{n} v_i\right\|^2 \le \left(\sum_{i=1}^{n}\|v_i\|\right)^2.$$

proof. Based on the orthogonality of the elements we obtain

$$\sum_{i=1}^{n}\|v_i\|^2 = \left\|\sum_{i=1}^{n} v_i\right\|^2.$$

Thus it follows for the first inequality by using the inequality of Young (A.0.3)

$$\begin{aligned}\left(\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\|v_i\|\right)\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\|v_i\|^2 + \frac{2}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\|v_i\|\,\|v_j\|\\ &\le \frac{1}{n}\sum_{i=1}^{n}\|v_i\|^2 + \frac{2}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\frac{\|v_i\|^2+\|v_j\|^2}{2}\\ &= \frac{1}{n}\sum_{i=1}^{n} n\,\|v_i\|^2 = \sum_{i=1}^{n}\|v_i\|^2 = \left\|\sum_{i=1}^{n} v_i\right\|^2.\end{aligned}$$

This implies the first inequality. The second inquality is obtained as follows:

$$\left\|\sum_{i=1}^{n} v_i\right\|^2 = \sum_{i=1}^{n}\|v_i\|^2 \le \sum_{i=1}^{n}\|v_i\|^2 + \frac{2}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\|v_i\|\,\|v_j\| = \left(\sum_{i=1}^{n}\|v_i\|\right)^2.$$

□

Lemma: A.0.6. *For $\gamma \in (0,1)$ we obtain that*

$$\mu(\gamma) := \frac{\gamma}{\sqrt{1-\gamma^2}}$$

increases in γ.

proof. We differentiate μ with respect to γ. Hence we have

$$\frac{d\,\mu}{d\gamma} = \frac{1}{\sqrt{1-\gamma^2}} - \frac{\gamma}{2}\frac{(-2\gamma)}{(\sqrt{1-\gamma^2})^3} > 0.$$

□

Bibliography

[ASF09] D'Ambra, P., di Serafino, D. and Filippone, S.: MLD2P4: a Package of Parallel Multilevel Algebraic Domain Decomposition Preconditioners in Fortran 95, Consiglio Nazionale delle Ricerche Istituto di Calcolo e Reti ad Alte Prestazioni, 2009.

[AxB84] Axelson, O., Gustafson, I.: Preconditioning and two level-multigrid methods of arbitrary degree of approximation. Math. Comp. Vol. 40, 1983, p 219 - 242.

[Bak66] Bakhvalov, N.S.: On the convergence of a Relaxion Method with natural constraints on the elliptic operator, U.S.S.R. Computational Mathematics and Mathematical Physics 6, 1966, p 101 - 135.

[BaD81] Bank, R.E. Dupont, T.: Analysis of two level scheme for solving finite element equations. Report CNA-159, Center for Numerical Analysis, University of Texas at Austin 1980.

[BDY88] Bank, R.E., Dupont, T.F. and Yserentant, H.: The hierachical basis multigrid method, Numer. Math., Vol 52, 1988, p 427 - 458.

[BjH88] Bjoerstad,P.E. und Hvidsten, A.: Iterative Methods for substructured elasticity problems in structural analysis, in Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, 1988, p. 301 - 312.

[BjW84] Bjoerstad,P.E. und Wilund, O.B.: Solving elliptic problems on regions partitioned into substructures, in Elliptic Problem Solver II, G. Birkhoff and A. Schoenstadt, eds., Academic Press, New York, 1984, p 245 - 256.

[Brk60] Brackhage, H.: Über die numerische Behandlung von Integralgleichungen nach der Quadraturformelmethode, Numerische Mathematik Vol 2, 1960, p 183 - 196.

[Bra81] Braess, D.: The contrction number of a multigrid method for solving the Poisson equation. Numer. Math. Vol. 37, 1981, p 387 - 404.

[Bra95] Braess, D.: Towards algebraic multigrid for elliptic problems of second order, Computing Vol 55, 1995, p 379 - 393.

[Brd73] Brandt, A.: Multi-level adaptive technique (MLAT) for fast numerical solutions to boundary problems, in Proceedings of the 3rd International Conference on Numerical Methods in Fluid Mechanics, Paris, 1972, H. Cabannes and R. Temam, eds., Springer Verlag, Berlin, 1973, p 82 - 89.

[Brd77] Brandt, A.: Multi-level adaptive solutions to boundary value problems, Mathematics of Computation, Vol. 31, 1977, p 333 - 390.

[Brd82] Brandt, A.: A guide to multigrid development, in Multigrid Methods, W.Hackbusch and U. Trottenberg, eds. Springer Verlag, Berlin, 1982, p 220 - 312.

[Brd86] Brandt, A.: Algebraic multigrid theory: The symmetric case, Applied Matematics and Computation, 19, 1986, p 23 - 56.

[BMR82a] Brandt, A., McCormick, S.F., Ruge, J.: Algebraic multigrid (AMG) for automatic algorithm design and problem solution. A preliminary report. Report, Inst. Comp. Studies, Colorado State University, Ft Collins, Co, 1982.

[BMR82b] Brandt, A. McCormick, S.F., Ruge, J.: Algebraic multigrid (AMG) for automatic algorithm design and problem solution. A preliminary report. Report, Inst. Comp. Studies, Colorado State University, Ft Collins, Co, 1982.

[BMR84] Brandt, A. McCormick, S.F., Ruge, J.: Algebraic multigrid (AMG) for sparse matrix equations, in Sparsity and Its Applications. D.J. Evans, ed., Camebridge University Press, Camebridge, 1984, p 257 - 284.

[BoM08] Bondy, J.A., Murty, U.S.R.: Graph Theory, Springer Verlag 2008, ISBN 978-1-84628-969-9.

[BrP87] Brambel, J.H., Pasciak, J.E.: New Convergence Estimates for Multigrid Algorithems. Math. Comp. Vol 49, 1987, p 311 - 329.

[BPS86] Brambel, J.H., Pasciak, J.E., Schatz, A.H.: The construction preconditioners for elliptic Problems by Substructuring. I. Math. Comp. Vol 47, 1986, p 103 - 134.

[BPS87] Brambel, J.H., Pasciak, J.E., Schatz, A.H.: The construction preconditioners for elliptic Problems by Substructuring. II. Math. Comp. Vol 49, 1987, p 1 - 16.

[BPS88] Brambel, J.H., Pasciak, J.E., Schatz, A.H.: The construction preconditioners for elliptic Problems by Substructuring. III. Math. Comp. Vol 51, 1988, p 415 - 430.

[BPS89] Brambel, J.H., Pasciak, J.E., Schatz, A.H.: The construction preconditioners for elliptic Problems by Substructuring. IV. Math. Comp. Vol 53, 1989, p 1 - 24.

[BPWX90] Brambel, J.H., Pasciak, J.E., Wang, J., Xu, J.: Convegence estimates for multigrid algorithems without regularity assumptions. Math. Comp. Vol 57, 1991, p 23 - 45.

[BPX90] Brambel, J.H., Pasciak, J.E.,Xu, J.: Parallel Multilevel Preconditioners. Math. Comp. Vol 55, 1990, p 1 - 22.

[BFLMRC04] Brezina, M., Falgout, R., MacLachlan, S., Manteuffel,T., McCormick, S. and Ruge, J.: Adaptive smoothed aggregation (αSA) Multigrid, SIAM J. Sci. Comput. Vol 25. No 6, 2004. p 1896 - 1920.

[BrV90] Brezina, M. and Vanek, P.: A black box iterative solver based on a two level Schwarz Method. Math. Comp. Vol 63, 1999, p 233 - 263.

[ChS89] Chen, H.-C., Sameh, A.H. : A matrix decomposition method for orthotropic elasticity problems, SIAM Journal on Matrix Analysis and Applications, Vol. 10, 1989, p 39 - 64.

[ClH94] Clark, J., Holton, A.H.: Graphentheorie, Grundlagen und Anwendung, Spektrum, Akademischer Verlag 1994, ISBN 3-86025-331.

[Fed62] Fedorenko, R.P.: A relaxation method for solving elliptic difference equations, U.S.S.R. Computational Mathematics and Mathematical Physics Vol. 1, 1962, p 1092 - 1096.

[Fed64] Fedorenko, R.P.: The speed convergence of one iterative process, U.S.S.R. Computational Mathematics and Mathematical Physics Vol. 4, 1964, p 227 - 235.

[GJV08] Guillard, H., Janka, A., Vanek, P.: Analysis of an algebraic Petrov-Galerkin smoothed aggregation multigrid method, Applied Numerical Mathematics, Vol. 58, 2008, p 1861 - 1874.

[GrR94] Grossmann, C., Roos, H.-G.: Numerik Partieller Differentialgleichungen, Teubner Verlag 1994, ISBN 3-519-12089-5.

[Hac85] Hackbusch, W.: Multigrid Methods, Springer Verlag, Berlin, 1985, ISBN 3-540-12761-5.

[Hac85] Hackbusch, W.: Iterative Lösung großer schwachbesetzter Gleichungssysteme, Teubner Verlag 1991, ISBN 3-519-02372-5.

[KrV96] Krizkova, J. and Vanek, P.: Two level preconditioner with small coarse grid appropriate for unstructured mesh, Num. Lin. Alg. Appl. Vol 3, 1996, p 255 - 274.

[KrD72] Kronsjö, L., Dahlquist, G..: On the design of nested iterations for elliptic difference equation, BIT Vol 12, 1972, p 63 - 71.

[MMB90] Mandel, J., McCormick, S.F., and Bank, R.: Variational multigrid theory, Multigrid Methods (s. McCormick, ed.), SIAM Philadelphia, PA 1987, p 131 - 178.

[McC87] McCormick, S.F. eds.: Multigrid methods, SIAM Frontiers in Applied Mathematics, Philadelphia, 1987.

[McR83] McCormick, S.F., and Ruge, J.W.: Unigrid for multigrid simulation, Math. Comp., Vol. 41, 1983, p 43 - 62.

[Ost] Ostrowski, A.M.: Über die Determinanten mit überwiegender Hauptdiagonale, Commentari Mathematici Helvetici, Vol 10 (1937), p 69 - 96.

[PLH09] Prill, F., Lukacova-Medvidova, M. and Hartmann, R.: Smoothed aggregation multigrid for the discontinous Galerkin Method, SIAM J. Sci. Comput. Vol 31. No 5, 2009, p 3503 - 3528.

[Rou89] Roux, F.X: Acceleration of the outer conjugate gradient by reorthogonalization for a domain decomposition method with Lagrange multiplier, in Proceedings of the third International Symposium on Domain Decomposition Methods, Houston, March 1989, SIAM, Philadelphia, 1990, p. 314 - 321.

[Sad03] Saad, Y.: Iterative Methods for Sparse Linear Systems, SIAM (Society for Industrial and Applied Mathematics) 2003, ISBN 0-89871-534-2.

[Sch70] Schwarz, H.A.: Gesammelte Mathematische Abhandlung Vol. 2, Springer Verlag, Berlin 1890, p 133 - 143. First published in Viertelsjahresschrift der Naturforschenden Gesellschaft in Zürich, Vol. 15, 1870, p 272 - 286.

[Smi92] Smith, B.F. An optimal domain decomposition preconditioner for finite elemente solution of linear elasticity problems, SIAM Journal on Scientific and Statistical Computing, Vol. 13, 1992, p. 199 - 233.

[Stu83] Stüben, K.: Algebraic Multigrid (AMG): Experience and comparison, Appl. Math. Comput. Vol 13, 1983, p 419 - 452.

[TRV91] Tallac, P.L., Roeck, Y.-H. D., Vidrascu, M.: Domain-decomposition methods for large linear elliptic three dimensional problems, Journal of Computational and Applied Mathematics, Vol. 34, 1991.

[Tic06] Tichy, M.: BPX-Verfahren und verschärfte Cauchy Ungleichung. Diplomarbeit am Institut für Mathematik, Bayrische Julius Maximilians Universität Würzburg, 2006.

[Van92] Vanek, P.: Acceleration of concergence of a two-level algorithm by smoothing transfer operators, Appl. Math. Vol 37, 1992, p 265 - 274.

[Van95] Vanek, P.: Fast multigrid solver, Appl. Math. Vol 40, 1995, p. 1 - 20.

[VBM96] Vanek, P., Brezina, M., Mandel, J.: Algebraic multigrid based on smoothed aggregation for second and fourth order problems, Computing, Vol 56, 1996, p 179 - 196.

[VBM01] Vanek, P., Brezina, M., Mandel, J.: Convergence of algebraic multigrid based on smoothed aggregation, Numerische Mathematik, Vol 88, 2001, p 559 - 579.

[VBT99] Vanek, P., Brezina, M.,Tezauer, R.: Two-grid method for linear elasticity on unstructured meshes , SIAM J. Sci. Comp., Vol 21, 1999, p 900 - 923.

[VAR60] Varga, R.S.: Factorizations and normalized iterative methods, in Boundary Problems in Differential Equations, University of Wisconsin Press, Madison, WI, 1960, p 121 - 142.

[Yse83] Yserentant, H.: On the Multi-Level Splitting of Finite Element Spaces, Bericht Nr. 21, Inst. für Geometrie und Praktische Mathematik der RWTH, Achen, 1983.

[Yse85] Yserentant, H.: Hierachical Bases of Finite Element Spaces in the Discritisation on Nonsymmetric Elliptic Boundary Value Problems. Computing Vol. 35, 1985, p 39 - 49.

[Yse86] Yserentant, H.: On the Multi-Level Splitting of Finite Element Spaces, Numer. Math., Vol 49, 1986, p 379 - 412.

[Yse86a] Yserentant, H.: On the Multi-Level Splitting of finite Element Spaces for Indefinite Elliptic Boundary Value Problems. SIAM J. Numer. Anal., Vol 23, 1986, p 581 - 595.

[ZKGB82] Zienkiewicz, O.C., Kelly, D.W., Gago, J. and Babuska, I.: Hierchical finite element approches, error estimates and adaptive refinements, in the Mathematics of Finite Elements and Applications IV (J.R. Whiteman, Ed), Mafelab 1981, London, 1982.