Alma Mater Studiorum
Università di Bologna

Dottorato di Ricerca in
MATEMATICA

Ciclo XXVI

Settore Concorsuale di afferenza: 01/A5
Settore Scientifico disciplinare: MAT/08

Spectral estimates and preconditioning for saddle point systems arising from optimization problems

Presentata da: Mattia Tani

Coordinatore Dottorato:
Chiar.mo Prof.
Giovanna Citti

Relatore:
Chiar.mo Prof.
Valeria Simoncini

Esame Finale anno 2015

Contents

Notation

Sets and vector spaces

$\mathbb{R}$	Set of real numbers
$\mathbb{R}^+$	Set of positive real numbers
$\mathbb{R}^n$	Space of real vectors of dimension n
$\mathbb{R}^{m\times n}$	Space of $m \times n$ real matrices
$\mathbb{P}_k$	Space of polynomials of degree at most k

Matrices and vectors

I_n	Identity matrix of order n
$\text{blkdiag}(A, B, \ldots)$	Block diagonal matrix with blocks A, B,...
$\text{diag}(A)$	Diagonal matrix derived from the diagonal of A
$\text{diag}(v)$	Diagonal matrix with vector v on the diagonal
$\text{nnz}(A)$	Number of nonzero entries of A
$\text{rank}(A)$	Rank of matrix A
$\kappa(A)$	Condition number of matrix A
A_{ij}	Entry of matrix A in the $i-$th row and $j-$th column
$v_{\max}$	Maximum entry of vector v
$v_{\min}$	Minimum entry of vector v
v_i	$i-$th entry of vector v

Eigenvalues and singular values

$\text{spec}(A)$	spectrum of A
$\lambda_{\max}(A)$	Maximum eigenvalue of matrix A with real spectrum
$\lambda_{\min}(A)$	Minimum eigenvalue of matrix A with real spectrum
$\lambda_i(A)$	$i-$th eigenvalue of a matrix A with real spectrum, sorted in ascending order
$\sigma_{\max}(B)$	Maximum singular value of matrix B
$\sigma_{\min}(B)$	Minimum singular of matrix B
$\sigma_i(B)$	$i-$th singular value of matrix B, in descending order

Subspaces

$\ker(A)$	Null space of matrix A
$\text{range}(A)$	Range of matrix A
$\text{span}\{v_1, v_2, \ldots\}$	Linear space of vectors v_1, v_2,...

Given column vectors x and y, we write (x, y) for the column vector given by their concatenation, instead of using $[x^T, y^T]^T$. Given A and B two square symmetric matrices, we write $A \succeq B$ to intend that $A-B$ is positive semidefinite, and $A \succ B$ to intend that $A - B$ is positive definite. The Euclidean norm of a vector $v \in \mathbb{R}^n$ is defined as $\|v\| := \sqrt{v^T v}$. The matrix norm induced by the Euclidean norm is defined as $\|B\| := \max_{v \in \mathbb{R}^n, v \neq 0} \frac{\|Bv\|}{\|v\|}$, where $B \in \mathbb{R}^{m \times n}$. Note that we use the same symbol, $\|\cdot\|$, to denote both the vector and the matrix norm, as there is no ambiguity between them. Given a symmetric and positive definite matrix $\mathcal{D} \in \mathbb{R}^{n \times n}$, we define the associated vector norm as $\|v\|_{\mathcal{D}} = \sqrt{v^T \mathcal{D} v}$ for $v \in \mathbb{R}^n$, and the matrix norm induced by it as $\|A\|_{\mathcal{D}} := \max_{v \in \mathbb{R}^n, v \neq 0} \frac{\|Av\|_{\mathcal{D}}}{\|v\|_{\mathcal{D}}}$ for $A \in \mathbb{R}^{n \times n}$

Chapter 1

Introduction

We are interested in the numerical solution of the linear system

$$\mathcal{A}v = b$$

where $\mathcal{A} \in \mathbb{R}^{N\times N}$ is nonsingular and $b \in \mathbb{R}^N$. This problem arises very frequently in all areas of scientific computing. Indeed, numerical methods often reduce the solution of more complex problems, like the solution of nonlinear equations or differential equations, to the solution of one or more linear systems. As a consequence, the solution (possibly many) linear systems often constitutes the main computational effort when numerical methods are applied on real-world problems, and the importance of having fast algorithms dealing with such problems cannot be overestimated.

Despite the seeming simplicity of linear systems, the actual computation of their solution can be extremely hard when the dimension of the problem N is large. Nowadays applications often lead to very large linear system, but these are typically sparse and/or have some special structure which can be exploited by modern algorithms to gain in efficiency. Of course, the properties and the structure of a linear systems depend on the nature of the original problem.

In this thesis, we focus on linear systems stemming from optimization problems of the form

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad & \tfrac{1}{2}x^T Ax - c^T x \\ \text{s. t.} \quad & Bx = d \\ & Dx \geq f \end{aligned} \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite, $B \in \mathbb{R}^{m\times n}$ with $m \leq n$, $D \in \mathbb{R}^{l\times n}$, $c \in \mathbb{R}^n$, $d \in \mathbb{R}^m$, $f \in \mathbb{R}^l$, and the inequality between

vectors is intended componentwise. Such problems are referred to as convex Quadratic Programming (QP) problems (cf. [95, Chapter 16]).

The numerical solution of (1.1) often requires the solution of (one or more) symmetric indefinite linear systems having a peculiar block structure. As an example, consider the simplified case when $D = 0$ and $f = 0$, i.e. when there are no inequality constraints. If we introduce the vector $p \in \mathbb{R}^m$ of Lagrange multipliers, the first order optimality conditions for (1.1) read

$$\begin{aligned} Ax + B^T p &= c, \\ Bx &= d, \end{aligned}$$

or, written in matrix form

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ p \end{bmatrix} = \begin{bmatrix} c \\ d \end{bmatrix}. \tag{1.2}$$

Thus, the solution of the optimization problem is equivalent to the solution of the above linear system. A system whose coefficient matrix has the structure given in (1.2) is referred to as a *saddle point system*. Such systems form a very important family, whose relevance is not limited to optimization problems. Indeed, saddle point systems arise naturally in many areas of computational science and engineering, such as fluid dynamics, linear elasticity, electromagnetism, and many others. As a consequence, the literature on this topic is vast and often focused on particular applications. In the next chapter, we will discuss the main properties of saddle point systems and review some well-known approaches to solve them. For the moment, it is enough to mention that solving saddle point systems is often a challenging task.

When solving a linear system, one has to choose between a direct or an iterative method. Direct methods, which include the LU and Cholesky factorization, are characterized by the property of delivering the exact solution (if rounding errors are not considered) in a finite number of steps. Direct methods are considered robust and predictable, and they typically are the method of choice when $\mathcal{A}$ is dense and has small dimension.

On the other hand, when direct methods are applied to large and sparse systems, a complication arising is that some entries which are zero in $\mathcal{A}$ may become nonzero in the factors. This phenomenon, known as *fill-in*, may greatly increase the memory storage and the computational cost. Fill-in can be reduced using appropriate strategies, giving rise to the so-called sparse direct methods.

Nevertheless, when $\mathcal{A}$ is large and sparse, iterative methods are often preferred. These methods compute a sequence of approximate solutions $(v_k)_{k\in\mathbb{N}}$ which should converge to the exact solution.

Krylov subspace methods, or, in short, Krylov methods, are a very popular family of iterative methods. They take their name from the so-called Krylov subspaces, which are defined as

$$\mathcal{K}_k(\mathcal{A}, r_0) = \text{span}\left\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{k-1}r_0\right\}$$

where $r_0 = b - \mathcal{A}v_0$ is the initial residual. In Krylov methods, at each iteration the approximate solution v_k is sought in the $k-$dimensional affine subspace $v_0 + \mathcal{K}_k(\mathcal{A}, r_0)$.

The actual expression of v_k depends on the particular Krylov method employed. In many important cases, v_k satisfies some optimality property; for example, in MINRES [99] and GMRES [115], v_k is the vector of $v_0 + \mathcal{K}_k(\mathcal{A}, r_0)$ which minimizes the Euclidean norm of the residual $r_k = b - \mathcal{A}v_k$.

The eigenvalue distribution of the system matrix often plays a crucial role in Krylov methods. This is especially true when $\mathcal{A}$ is symmetric, as in (1.2). Indeed, the rate of convergence of Krylov methods for symmetric systems can be bounded by a term that depends only on the extreme eigenvalues of $\mathcal{A}$. As an example, consider MINRES, a Krylov method often employed in the context of saddle point systems. In MINRES, if the spectrum of $\mathcal{A}$ is contained in the set $[-d, -c] \cup [c, d]$, where $d > c > 0$, then the following can be shown (see e.g. [64, (3.15)]):

$$\frac{\|r_k\|}{\|r_0\|} \leq 2\left(\frac{d/c - 1}{d/c + 1}\right)^{[k/2]}, \qquad k = 1, 2, \ldots \tag{1.3}$$

where $[\cdot]$ denotes the integer part. It is apparent that, if $d/c \approx 1$ convergence will be fast; if, on the other hand, $d/c \gg 1$ then a slow convergence might be observed. Generally speaking, in the symmetric case, the more clustered the eigenvalues of $\mathcal{A}$ the faster the convergence.

The rate of convergence of Krylov methods can be improved (sometimes drastically) using a preconditioning strategy. By preconditioning, it is meant that the original problem is replaced with another one, which is equivalent but easier to solve. More precisely, instead of solving $\mathcal{A}v = b$, we solve for example the equivalent system

$$\mathcal{P}^{-1}\mathcal{A}v = \mathcal{P}^{-1}b \tag{1.4}$$

where $\mathcal{P} \in \mathbb{R}^{N \times N}$ is an invertible matrix called preconditioner. Two are the features of a good preconditioner. First, since at each iteration of a Krylov method we need to compute a matrix-vector product involving $\mathcal{P}^{-1}$ (or, equivalently, we need to solve a linear system with coefficient matrix $\mathcal{P}$), this operation must be sufficiently cheap. Second, the system $\mathcal{P}^{-1}\mathcal{A}$ must have better spectral properties than the original system $\mathcal{A}$.

System (1.4) is referred to as left preconditioning. Other reformulations for the system $\mathcal{A}v = b$ using the same $\mathcal{P}$ are possible, and will be discussed in the next chapter. For the moment we just note that, even if the original system matrix $\mathcal{A}$ is symmetric, such property may not hold for the preconditioned system.

When a Krylov method is applied to a nonsymmetric system, the role of the eigenvalues of $\mathcal{A}$ is less clear and no simple bound like (1.3) can be found. In some cases, the eigenvalue distribution may even be misleading. Consider for example the GMRES method. It has been proved in [65] that any nonincreasing convergence curve is possible for this method, regardless of the eigenvalues of the system matrix $\mathcal{A}$. Though this result might seem catastrophic at first, it is been observed that for many practical problems well-clustered eigenvalues (away from 0) still result in fast convergence, even in the nonsymmetric case.

As a consequence of the relation between the eigenvalues and the rate of convergence, it is extremely important to have meaningful spectral estimates for the matrices involved. What we mean here is that the estimates should accurately reflect the eigenvalue distributions of such matrices. Indeed, meaningful spectral estimates for $\mathcal{P}^{-1}\mathcal{A}$ give a precise idea of the worst-case scenario one can get when attempting to solve the linear system, and hence they can be considered a measure of the preconditioner quality. This is true even in the nonsymmetric case, although here information on the spectrum has to be coupled with some other information.

We emphasize that even spectral estimates for the unpreconditioned system may be useful, for a general reason not directly related with Krylov methods. If $\mathcal{A}$ is symmetric, as the matrices that will be discussed in this thesis, then spectral bounds for the unpreconditioned system give also an upper bound for its condition number, which in turn can be used to estimate the stability of numerical methods employed for its solution.

1.1 Aims and outline

The optimization problems that will be discussed in this thesis can be divided into two classes. First, we consider Quadratic Programming problems of the form (1.1) when $D = I_n$ and $f = 0$, i.e. when the inequality constraints are just $x \geq 0$. These are known as QP problems *in standard form*.

The second class of problems considered in this thesis stems from the discretization of certain PDE-constrained optimal control problems, with possible further inequality constraints on the control and/or on the state. In these cases, the variable x in (1.1) can be splitted in two parts, namely

$x = (u, y) \in \mathbb{R}^{2s}$, and the matrices A, B, D are chosen so that the whole problem (1.1) takes the form

$$\begin{aligned} \underset{y,u\in\mathbb{R}^s}{\text{minimize}} \quad & \tfrac{1}{2}(y-y_d)^T M(y-y_d) + \tfrac{\nu}{2}u^T M u \\ \text{s. t.} \quad & Lx = Mu - d \\ & a \leq \alpha_u u + \alpha_y y \leq b \end{aligned} \tag{1.5}$$

with $M, L \in \mathbb{R}^{s\times s}$, $a, b, d, y_d \in \mathbb{R}^s$, $\nu > 0$ and $\alpha_u, \alpha_y \geq 0$. Although problems of the form (1.5) are technically a subgroup of Quadratic Programming problems, because of their peculiar structure and other special features they deserve a separate treatment. More details about optimal control problems and their discrete version will be given in Part II.

The aim of this thesis is also twofold. First, we explore new preconditioning strategies for saddle point systems arising from the two classes of optimization problems just mentioned. Second, we derive novel spectral bounds for some of the matrices involved, both in the preconditioned and unpreconditioned case. The specific objectives are discussed below, in the description of each chapter.

The purpose of Part I is to present the main theoretical tools that will be used throughout the thesis. In Chapter 2, we review part of the theory about saddle point systems and their spectral properties, iterative Krylov methods and some well-known preconditioners for saddle point systems. In Chapter 3, we focus on the so-called augmented block diagonal preconditioner. After a review of the known results, we show spectral estimates for the preconditioned system when the (2,2) block of the original system is nonzero and discuss how the building blocks of the preconditioner should be chosen. This analysis first appeared in [91].

In Part II, we consider saddle point systems stemming from optimal control problems constrained by PDEs. In Chapter 4, we review an approach proposed in [117], based on a preconditioned Conjugate Gradient method implemented in a non-standard inner product. We present new and more accurate spectral estimates for the preconditioned system, and show how these estimates help understand the rate of convergence of the method. Most of these results have been published in [128]. In Chapter 5, we add control and state constraints to the optimal control problem, and consider the sequence of saddle point systems obtained when an active-set Newton method is applied to the discrete problem. We present two new preconditioners based on a full block matrix factorization of the Schur complement of the Jacobian matrix. We also derive spectral estimates for the preconditioned system and discuss the robustness of our approach with respect to the problem parameters, both theoretically and experimentally. We finally validate our approach

with numerical experiments on 3D problems, comparing our strategy with one of the state-of-the-art approaches. These results first appeared in [107].

Part III is devoted to sequences of saddle point systems obtained by applying an Interior Point (IP) method to a convex Quadratic Programming problem in standard form. Such systems have a natural 3×3 block structure, but a common approach is to deal with these systems in reduced form, by taking one or more steps of block Gaussian elimination. However, the reduced formulation becomes increasingly ill-conditioned as the IP iterates approach the exact solution, and some authors raised the question about which formulation should be preferred. Our contribution here aims at further exploring the features of the two approaches. In Chapter 6 we present new spectral estimates which confirm that, under suitable assumptions, the condition number of the unreduced formulation is bounded even when the iterates are close to the exact solution. The spectral analysis presented, which first appeared in [91], improves and completes the one recently given in [66]. The sharpness of the new estimates is illustrated by numerical experiments. In Chapter 7, we elaborate further on the comparison between reduced and unreduced formulation and discuss how their relation is affected by preconditioning. We consider some well-known preconditioners for saddle point systems (namely, constraint and augmented) and prove a spectral equivalence result which suggests that the unreduced formulation does not necessarily have better spectral properties than the reduced one, when preconditioning is considered. These results were first presented in [92]. Finally, we carry out an experimental analysis, to assess which of the two formulations should be preferred when solving large scale problems.

Part I

Numerical linear algebra preliminaries

Chapter 2

Saddle point systems and Krylov methods background

In this chapter we introduce the main subjects of this thesis, namely saddle point systems, discuss their main properties and present some approaches commonly employed in their solution. In Section 2.1 we clarify what a saddle point system is, give conditions for its nonsingularity and analyze its spectral properties. In Section 2.2, we briefly discuss Krylov methods, in particular CG, MINRES and GMRES and their main features. Finally, in Section 2.3 we present some well-known preconditioners for saddle point systems, and give details on their applicability and spectral properties of the preconditioned systems.

We remark that all the results presented here are taken from other sources. In particular, the whole chapter was inspired by the survey paper written by Benzi et al. [11]. We hence refer to that work and references therein for a more detailed discussion on saddle point systems and their numerical solution.

2.1 Saddle point systems

Consider a symmetric matrix having the following block structure

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix}, \tag{2.1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite, $B \in \mathbb{R}^{m\times n}$ with $n \geq m$, and $C \in \mathbb{R}^{m\times m}$ symmetric and positive semidefinite. An important

special case occurs when the C block is zero, i.e. $\mathcal{A}$ has the form

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}. \tag{2.2}$$

As we saw in the Introduction, this is the coefficient matrix of the linear system arising from the optimization problem

$$\begin{aligned} \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad & \tfrac{1}{2} x^T A x - c^T x, \\ \text{s. t.} \quad & Bx = d. \end{aligned}$$

The solution (x^*, p^*) of the system

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ p \end{bmatrix} = \begin{bmatrix} c \\ d \end{bmatrix}. \tag{2.3}$$

where p is the vector of Lagrange multipliers, is hence a *saddle point* for the Lagrangian function $\mathcal{L}(x, p)$ associated to (2.3), that is

$$\mathcal{L}(x^*, p) \leq \mathcal{L}(x^*, p^*) \leq \mathcal{L}(x, p^*) \qquad \forall\, x \in \mathbb{R}^n,\ p \in \mathbb{R}^m.$$

Due to this interpretation, system (2.3) is referred to as *saddle point system*, and the matrix (2.3) is sometimes called *saddle point matrix*.

Linear systems with coefficient matrix of the form (2.1) appear in different contexts, for example in the solution of mixed finite elements when a stabilization term is added to the problem (see e.g. [20, Chapter 3, Section 4], [42, Chapter 3]), or in Interior Point methods with regularization (see e.g. [1, 53] and Chapter 7 of this thesis). Probably due to these interpretations, matrices of the form (2.1) have been referred to as "stabilized" [61] or "regularized" saddle point matrices[32]. To keep the nomenclature as simple as possible, in this thesis we will use the term "saddle point matrix" to refer to a matrix of the more general form (2.1), whether $C = 0$ or not.

We observe that the block structure expressed by (2.1)-(2.2) is fairly general, and saddle point systems arise very frequently not only in constrained optimization or mixed finite elements, but also in many different areas of computational science and engineering. In [11, Section 2], the authors present a list of 16 different fields where saddle point systems naturally arise. In Part II and Part III of this thesis we will discuss two of these applications, and describe how they give rise to saddle point systems. In particular, the saddle point systems considered in Part II, which stem from optimal control problems, will be of the form (2.2). Instead, the saddle point systems considered in Part III, which stem from QP problems in standard form solved with an

Interior Point method, will be of the more general form (2.1), with a nonzero C.

We mention that in the literature matrices with a block structure similar to (2.1)-(2.2), but nonsymmetric, have often been referred to as generalized saddle point matrices [25]. These, however, will not be discussed here.

2.1.1 Nonsingularity

The first questions that arise when one has to solve the linear system $\mathcal{A}x = b$ are whether this system has a solution and whether this solution is unique. We therefore recall the conditions that ensure nonsingularity of $\mathcal{A}$.

We start by supposing that A, the (1,1) block of $\mathcal{A}$, is positive definite. Then $\mathcal{A}$ can be written as

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ BA^{-1} & I_m \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & -S \end{bmatrix} \begin{bmatrix} I_n & A^{-1}B^T \\ 0 & I_m \end{bmatrix}, \tag{2.4}$$

where $S = C + BA^{-1}B^T$ is the (negative) *Schur complement*[1] of A in $\mathcal{A}$. This important factorization shows in particular that $\mathcal{A}$ is nonsingular if and only if S is positive definite, that is, if and only if $\ker(C) \cap \ker(B^T) = 0$. In particular, $\mathcal{A}$ is nonsingular if $\ker(B^T) = 0$

If C is positive definite, then there exists a factorization analogous to (2.4), showing that $\mathcal{A}$ is nonsingular if and only if $A+B^TC^{-1}B$, i.e. the Schur complement of C in $\mathcal{A}$, is positive definite, that is if and only if $\ker(B) \cap \ker(A) = 0$.

If both A and C are singular, then none of the above conditions can be used to assess if $\mathcal{A}$ is nonsingular. We hence consider the following proposition.

Proposition 2.1.1. [5, Lemma 2.1] *Let $\mathcal{A}$ be a saddle point matrix of the form* (2.1) *with A and C symmetric and positive semidefinite. Then $\mathcal{A}$ is nonsingular if and only if all the following conditions are satisfied*

1. $\ker(A) \cap \ker(B) = 0$
2. $\ker(B^T) \cap \ker(C) = 0$
3. $\text{range} \begin{bmatrix} A \\ B \end{bmatrix} \cap \text{range} \begin{bmatrix} B^T \\ -C \end{bmatrix} = 0$

[1] For the rest of the thesis, we will drop the adjective "negative" and refer to S as the Schur complement of A in $\mathcal{A}$.

The above proposition is very general, as neither A nor C needs to be nonsingular, but its usefulness in practice is limited. An example of nonsingular saddle point system with A and C singular and $\ker(B^T) \neq 0$ will be discussed in Chapter 6.

In the special case when $\ker(B^T) = 0$ and $C = 0$, the conditions of the above proposition become simpler, and we have the following important result.

Proposition 2.1.2. (see e.g. [11, Theorem 3.2]) *Let $\mathcal{A}$ be a saddle point matrix of the form (2.2), with A symmetric and positive semidefinite and suppose* $\ker(B^T) = 0$. *Then $\mathcal{A}$ is nonsingular if and only if* $\ker(A) \cap \ker(B) = 0$. *In particular, if A is positive definite then $\mathcal{A}$ is nonsingular.*

2.1.2 Spectral properties

We now focus on the spectral properties of $\mathcal{A}$, and we assume for simplicity that $\mathcal{A}$ is nonsingular. Saddle point matrices are indefinite, i.e. they have both positive and negative eigenvalues. Indeed, if A is positive definite, it follows from the factorization (2.4) and from Sylvester's Law of Inertia that $\mathcal{A}$ has the same signature as the block diagonal matrix $\begin{bmatrix} A & 0 \\ 0 & -S \end{bmatrix}$. Since $\mathcal{A}$ is nonsingular, then S must be positive definite, implying that $\mathcal{A}$ has exactly n positive eigenvalues and m negative ones. Using a simple continuity argument it can be shown that this property holds even when A is singular. We summarize this result in the next proposition.

Proposition 2.1.3. see e.g. [11, pag. 21] *Suppose that $\mathcal{A}$ given by (2.1) is nonsingular. Then it has exactly n positive eigenvalues and m negative ones.*

Saddle point systems are sometimes said to be *highly indefinite* (see e.g. [11]), which means that $\mathcal{A}$ has many eigenvalues of both signs (unless of course m is very small compared to n).

In the literature, it is possible to find different results describing the eigenvalue distribution of saddle point matrices, depending on the assumptions on $\mathcal{A}$ that are considered. In 1992, Rusten and Winther [113], provided spectral estimates for $\mathcal{A}$ when $C = 0$, A is positive definite and B has full rank. In 1994, Silvester and Wathen [120] extended their result to the case when $C \neq 0$. In 2009, Gould and Simoncini [61] were able to relax the assumption of positive definiteness on A. Their result holds in the more general case when $\ker(A) \cap \ker(B) = 0$, although a condition on the norm of C is required. All these results express bounds for the eigenvalues of $\mathcal{A}$ that depend on parameters relative to the single blocks A, B and C, such as their

minimum and maximum eigenvalue or singular value. We also mention the works of Bai, Ng and Wang [6], and of Bai [5], where the authors derive spectral estimates for $\mathcal{A}$ in a general setting, that depend on the eigenvalues of A, of its Schur complement and of the matrix $BA^{-2}B^T$.

Below we recall the original result given by Rusten and Winther.

Theorem 2.1.4. [113, Lemma 2.1] *Let $\mathcal{A}$ be as in (2.2), with A symmetric and positive definite and $\ker(B^T) = 0$. Let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimum and maximum eigenvalues of A, and let $\sigma_{\min}$ and $\sigma_{\max}$ be the minimum and maximum singular value of B. Then it holds*

$$\operatorname{spec}(\mathcal{A}) \subseteq I^- \cup I^+,$$

where

$$I^- = \left[\frac{1}{2}\left(\lambda_{\min} - \sqrt{\lambda_{\min}^2 + 4\sigma_{\max}^2}\right),\ \frac{1}{2}\left(\lambda_{\max} - \sqrt{\lambda_{\max}^2 + 4\sigma_{\min}^2}\right)\right],$$

and

$$I^+ = \left[\lambda_{\min},\ \frac{1}{2}\left(\lambda_{\max} + \sqrt{\lambda_{\max}^2 + 4\sigma_{\max}^2}\right)\right].$$

We now make a few comments about why the numerical solution of saddle point systems is often considered a challenging problem. First, although saddle point systems are symmetric, their indefiniteness is considered an unfavorable property (see the next section). Second, one can use the results of Theorem 2.1.4 to gain insight into the condition number of $\mathcal{A}$, which can be written as

$$\kappa(\mathcal{A}) = \|\mathcal{A}\| \left\|\mathcal{A}^{-1}\right\| = \frac{\max\limits_{\lambda \in \operatorname{spec}(\mathcal{A})} |\lambda|}{\min\limits_{\lambda \in \operatorname{spec}(\mathcal{A})} |\lambda|}.$$

Unfortunately, in many applications the conditioning of $\mathcal{A}$ can be very high, and it often depends on some parameters of the problem. For example, in the optimal control problems discussed in Part II of this thesis, the blocks of $\mathcal{A}$ stem from the discretization of differential operators. As a consequence, both $\sigma_{\min}$ and $\lambda_{\min}$ go to zero as h, the mesh size parameter, goes to zero; this means that the condition number of $\mathcal{A}$ (and the rate of convergence of Krylov methods) worsens as the problem size increases. Moreover, the choice of other parameters (such as the regularization parameter) may affect the conditioning of $\mathcal{A}$.

A different cause for the ill-conditioning of $\mathcal{A}$ can be observed when Interior Point methods are used to solve QP problems, as will be discussed in Part III. In this case, a popular formulation for the linear system that has to

be solved at each iteration of the method is a saddle point system where $\lambda_{\max}$ goes to infinity (and possibly $\lambda_{\min}$ goes to zero) as the method approaches the exact solution of the QP problem.

Before the end of this section, we present an additional proposition which fully characterizes the eigenvalues of $\mathcal{A}$ when, in addition to the assumptions of Theorem 2.1.4, it holds $A = I_n$. We include this result as it will be used in Chapter 5.

Proposition 2.1.5. [44, Lemma 2.1] *Let $\mathcal{A}$ be as in (2.2), with $A = I_n$. Then $\mathcal{A}$ has the eigenvalue 1 with multiplicity $n - m$. Moreover, if σ_i, $i = 1, \ldots, m$ denote the singular values of B, then the remaining $2m$ eigenvalues are given by:*

$$\frac{1}{2}\left(1 \pm \sqrt{4\sigma_i^2 + 1}\right) \qquad i = 1, \ldots, n.$$

2.2 Krylov methods

In this section we summarize some results about iterative Krylov methods for the solution of linear systems, and in particular CG, MINRES, GMRES. Our goal is to mention the principal features of the particular methods that will be used throughout the thesis. For more details on the subject of Krylov methods, we refer to the monographs [64, 89, 114, 133, 85].

Given the linear system $\mathcal{A}v = b$, with $\mathcal{A} \in \mathbb{R}^{N\times N}$ and $b \in \mathbb{R}^N$, all Krylov methods produce a sequence of approximate solutions v_k such that

$$v_k \in v_0 + \mathcal{K}_k(\mathcal{A}, r_0), \qquad k = 1, 2, \ldots, \tag{2.5}$$

where v_0 is the initial guess for the solution and

$$\mathcal{K}_k(\mathcal{A}, r_0) = \text{span}\left\{r_0, \mathcal{A}r_0, \mathcal{A}^2 r_0, \ldots, \mathcal{A}^{k-1} r_0\right\} \tag{2.6}$$

is the k-th *Krylov subspace* generated by $\mathcal{A}$ and r_0. For ease of notation, when there is no ambiguity we will simply write $\mathcal{K}_k$. Since $\mathcal{K}_k$ is the linear space of k vectors, its dimension d_k is at most k. We suppose for the moment that $d_k = k$. In some Krylov methods, the $k-$th iterate v_k is uniquely determined by the requirement that the new residual $r_k := b - \mathcal{A}v_k$ is orthogonal to another subspace of dimension k, which is sometimes called the *constraint space*. The choice of this space plays a crucial role in Krylov methods. Two popular orthogonality conditions are:

(i) $b - \mathcal{A}v_k \perp \mathcal{K}_k$,

(ii) $b - \mathcal{A}v_k \perp \mathcal{A}\mathcal{K}_k$.

It is always possible to construct a Krylov method whose iterates satisfy property (ii), provided that $\mathcal{A}$ is nonsingular. On the other hand, it turns out that a vector v_k of the form (2.5) with property (i) does not necessarily exist. An assumption that guarantees the existence of v_k is such case is that $\mathcal{A}$ is symmetric and positive definite (SPD) [114, Proposition 5.1].

An orthogonality requirement on the residual r_k may correspond to an optimality property for the approximate solution v_k. This is indeed the case for (i) and (ii), as stated by the next theorem. Here and throughout, given an SPD matrix $\mathcal{A}$ we define the associated $\mathcal{A}$-norm as $\|v\|_{\mathcal{A}} := \sqrt{v^T \mathcal{A} v}$ for $v \in \mathbb{R}^N$.

Theorem 2.2.1. [114, Proposition 5.2 and Proposition 5.3] *Let $v_k \in v_0 + \mathcal{K}_k$. Then, the following statements hold.*

(i) *If $r_k \perp \mathcal{K}_k$ with $\mathcal{A}$ symmetric and positive definite, and v^* denotes the exact solution of $\mathcal{A}v = b$, then*

$$\|e_k\|_{\mathcal{A}} := \|v^* - v_k\|_{\mathcal{A}} = \min_{v \in v_0 + \mathcal{K}_k(\mathcal{A}, r_0)} \|v^* - v\|_{\mathcal{A}},$$

i.e. v_k minimizes the $\mathcal{A}-$norm of the error over $v_0 + \mathcal{K}_k$.

(ii) *If $r_k \perp \mathcal{A}\mathcal{K}_k$ with $\mathcal{A}$ nonsingular, then*

$$\|r_k\| = \|b - \mathcal{A}v_k\| = \min_{v \in v_0 + \mathcal{K}_k(\mathcal{A}, r_0)} \|b - \mathcal{A}v\|,$$

i.e. v_k minimizes the Euclidean norm of the residual over $v_0 + \mathcal{K}_k$.

The Conjugate Gradient (CG) method [70], is characterized by the property (i), and thus to be applied it requires $\mathcal{A}$ to be SPD. Of course, saddle point systems are not positive definite, however we will show in Chapter 4 that a variant of the CG method can be applied to preconditioned saddle point systems, by using an inner product different from the Euclidean one.

If the dimension of $\mathcal{K}_k$ is strictly smaller than k, than it can be shown that the exact solution v^* belongs to $\mathcal{K}_{k-1}$ [114, Proposition 5.6], and hence any approximate solution v_{k-1} satisfying (i) or (ii) is necessarily the exact solution. This means in particular that, in exact precision arithmetic, the corresponding Krylov methods always converge in a finite number of steps. This feature is known as *finite termination property*. More precisely, if $d_{\mathcal{A}}$ denotes the degree of the minimum polynomial of $\mathcal{A}$ (which is the polynomial of least degree such that $p(\mathcal{A}) = 0$), then it can be shown that the dimension of $\mathcal{K}_k$ cannot exceed $d_{\mathcal{A}}$, meaning that $v_{d_{\mathcal{A}}} = v^*$.

Nevertheless, since $d_{\mathcal{A}}$ is typically very large, it would be unpractical to compute iterates until the exact solution is found. Moreover, the user might not even need the exact solution, but rather be satisfied with a good enough approximation. Thus, it is extremely important to estimate how fast a Krylov methods converges, in order to predict how many steps are required to ensure that a suitable measure of the error is below the given tolerance.

To continue our analysis and also discuss other features of Krylov methods, we have to distinguish the case when $\mathcal{A}$ is symmetric from the case when $\mathcal{A}$ is nonsymmetric. We remark that, although this thesis is devoted to the solution of symmetric systems, the nonsymmetric case is still very important, since by preconditioning the symmetry of the original problem might be lost (see Section 2.3).

2.2.1 The symmetric case

We first note that $\mathcal{K}_k$ can be written as

$$\mathcal{K}_k = \{p(\mathcal{A})r_0 \,|\, p \in \mathbb{P}_{k-1}\} = \{\mathcal{A}p(\mathcal{A})e_0 \,|\, p \in \mathbb{P}_{k-1}\}\,,$$

where $\mathbb{P}_{k-1}$ is the space of polynomials of degree at most $k-1$.

If $\mathcal{A}$ is symmetric, than we can write $\mathcal{A} = V\Lambda V^T$ with Λ diagonal and V orthogonal. We first suppose that $\mathcal{A}$ is also positive definite and let e_k be the error at step k of the CG method. We observe that, if $v_k = v_0 + \mathcal{A}q(\mathcal{A})e_0$ with $q \in \mathbb{P}_{k-1}$, then we have

$$e_k = v^* - v_k = v^* - v_0 + \mathcal{A}q(\mathcal{A})e_0 = (I_N + \mathcal{A}q(\mathcal{A}))\, e_0 = p(\mathcal{A})$$

where $p \in \mathbb{P}_k$ such that $p(0) = 1$. By varying $q \in \mathbb{P}_{k-1}$, any polynomial $p \in \mathbb{P}_k$ with this restriction can be obtained. Then it holds

$$\begin{aligned} \|e_k\|_{\mathcal{A}} &= \min_{v \in v_0 + \mathcal{K}_k} \|v^* - v\|_{\mathcal{A}} = \min_{p \in \mathbb{P}_k, p(0)=1} \|p(\mathcal{A})e_0\|_{\mathcal{A}} = \\ &= \min_{p \in \mathbb{P}_k, p(0)=1} \left\|p(\mathcal{A})\mathcal{A}^{1/2}e_0\right\| \leq \min_{p \in \mathbb{P}_k, p(0)=1} \|p(\mathcal{A})\|\, \|e_0\|_{\mathcal{A}} = \\ &= \min_{p \in \mathbb{P}_k, p(0)=1} \|p(\Lambda)\|\, \|e_0\|_{\mathcal{A}} = \min_{p \in \mathbb{P}_k, p(0)=1} \max_{\lambda \in \mathrm{spec}(\mathcal{A})} |p(\lambda)|\, \|e_0\|_{\mathcal{A}}\,. \end{aligned} \tag{2.7}$$

It can be shown that bound (2.7) on the $\mathcal{A}-$norm of e_k is sharp, in the sense that for any given matrix $\mathcal{A}$ and any given index k strictly less than the number of distinct eigenvalues of $\mathcal{A}$, there exists an initial error such that the above relation holds with equality [63]. Note that if k is greater than or equal to the number of distinct eigenvalues of $\mathcal{A}$, and hence to the degree

of its minimum polynomial, then necessarily $e_k = 0$ as a consequence of the finite termination property.

We can derive a similar relation for a Krylov method with property (ii), when $\mathcal{A}$ is symmetric. This is the case of the Minimum Residual (MINRES) method [99]. If r_k is the residual at step k of MINRES, we have

$$\begin{aligned} \|r_k\| &= \min_{v \in v_0 + \mathcal{K}_k} \|b - \mathcal{A}v\| = \min_{p \in \mathbb{P}_k, p(0)=1} \|p(\mathcal{A}) r_0\| \leq \\ &\leq \min_{p \in \mathbb{P}_k, p(0)=1} \|p(\mathcal{A})\| \, \|r_0\| = \min_{p \in \mathbb{P}_k, p(0)=1} \|p(\Lambda)\| \, \|r_0\| = \\ &= \min_{p \in \mathbb{P}_k, p(0)=1} \max_{\lambda \in \mathrm{spec}(\mathcal{A})} |p(\lambda)| \, \|r_0\| , \end{aligned} \tag{2.8}$$

and this bound is sharp in the same sense as (2.7). We remark that the sharpness of inequalities (2.7) and (2.8) shows that in the symmetric case the worst-case scenario for the convergence of Krylov method is completely determined by the eigenvalues of $\mathcal{A}$.

We now report a more explicit expression for the bounds on the error and on the residual. Unfortunately, although the polynomial that solves the $\min - \max$ problems (2.7)-(2.8) is known [63], the value of the $\min - \max$ itself cannot be easily expressed. Thus, a common approach is to replace the maximum over the set of the eigenvalues of $\mathcal{A}$ with the maximum over a larger set. In the case of CG this set is the convex hull of $\mathrm{spec}(\mathcal{A})$, which we denote with $[\lambda_{\min}, \lambda_{\max}]$ (recall that in the CG method $\mathcal{A}$ has to be SPD). More precisely,

$$\frac{\|e_k\|_{\mathcal{A}}}{\|e_0\|_{\mathcal{A}}} \leq \min_{p \in \mathbb{P}_k, p(0)=1} \max_{\lambda \in \mathrm{spec}(\mathcal{A})} |p(\lambda)| \leq \min_{p \in \mathbb{P}_k, p(0)=1} \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |p(\lambda)| . \tag{2.9}$$

The new $\min - \max$ problem is solved by a scaled and shifted Chebyshev polynomial of the first kind (see e.g. [114, Theorem 6.25]) and the corresponding bound on the relative error reads

$$\frac{\|e_k\|_{\mathcal{A}}}{\|e_0\|_{\mathcal{A}}} \leq 2 \left(\frac{\sqrt{\kappa(\mathcal{A})} - 1}{\sqrt{\kappa(\mathcal{A})} + 1} \right)^k , \qquad \text{where} \quad \kappa(\mathcal{A}) = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{2.10}$$

In the $\min - \max$ problem (2.8), $\mathcal{A}$ is not restricted to be positive definite. Thus, we have to distinguish between the case when the eigenvalues of $\mathcal{A}$ are all positive and the case when $\mathcal{A}$ is indefinite. In the first case, a bound analogous to (2.10) can be derived for the Euclidean norm of the residual. In the second case, suppose $\mathrm{spec}(\mathcal{A}) \subseteq [a, b] \cup [c, d]$, with $a, b < 0$ and $c, d > 0$.

It holds

$$\frac{\|r_k\|}{\|r_0\|} \leq \min_{p\in\mathbb{P}_k, p(0)=1} \max_{\lambda\in\mathrm{spec}(\mathcal{A})} |p(\lambda)| \leq \min_{p\in\mathbb{P}_k, p(0)=1} \max_{\lambda\in[a,b]\cup[c,d]} |p(\lambda)|\,.$$

An explicit bound for $\dfrac{\|r_k\|}{\|r_0\|}$ can be obtained by assuming that $b-a=d-c$, i.e., that the two spectral intervals have the same length. In this case, the solution of the above problem is again given by a scaled and shifted Chebyshev polynomial of the first kind. The obtained bound is (see e.g. [64, (3.14)])

$$\frac{\|r_k\|}{\|r_0\|} \leq 2\left(\frac{\sqrt{|ad|}-\sqrt{|bc|}}{\sqrt{|ad|}+\sqrt{|bc|}}\right)^{[k/2]}, \tag{2.11}$$

where $[\cdot]$ denotes the integer part. In particular, if $a=-d$ and $b=-c$, i.e. if the spectral intervals are symmetric with respect to the origin, the above bound reduces to

$$\frac{\|r_k\|}{\|r_0\|} \leq 2\left(\frac{d/c-1}{d/c+1}\right)^{[k/2]}. \tag{2.12}$$

If we look at both (2.10) and (2.12) (or (2.11)), it is apparent that a *sufficient* (but not necessary) condition to observe fast convergence is that the eigenvalues of $\mathcal{A}$ are clustered. Indeed, if this is case, in both formulas the right-hand side becomes small very quickly as k increases. However, a comparison of the two bounds suggests that indefinite problems may pose a more serious challenge. Indeed, as pointed out in [64, Section 3.1], (2.12) is the bound one would obtain at step $[k/2]$ for a positive definite system with condition number of d^2/c^2, which is the *square* of the actual condition number.

Before turning to the nonsymmetric case, we discuss another important feature of Krylov methods for symmetric systems. As it generally happens in iterative methods, the approximate solution at step $k+1$ can be written as

$$v_{k+1} = v_k + \alpha_k p_k, \tag{2.13}$$

where $\alpha_k \in \mathbb{R}^+$ and $p_k \in \mathcal{K}_{k+1}(\mathcal{A}, r_0)\setminus\mathcal{K}_k(\mathcal{A}, r_0)$. It is then apparent that the set of vectors $\{p_0,\ldots,p_k\}$ form a basis for the Krylov subspace $\mathcal{K}_{k+1}(\mathcal{A}, r_0)$.

In the CG method, if $w \in \mathcal{K}_{k+1}$ from the orthogonality condition (i) we infer

$$0 = w^T r_{k+1} = w^T\left(b-\mathcal{A}\left(v_k+\alpha_k p_k\right)\right) = w^T\left(r_k-\alpha_k\mathcal{A}p_k\right) = w^T r_k - \alpha_k w^T\mathcal{A}p_k.$$

In particular, if we take $w \in \mathcal{K}_k$ then we have $w^T r_k = 0$ from the previous step, and hence the above formula reduces to $\alpha_k w^T\mathcal{A}p_k = 0$. Since $\alpha_k \neq 0$,

we have that p_k must be $\mathcal{A}$-orthogonal to $\mathcal{K}_k$, or equivalently to all its basis vectors $p_0, \ldots, p_{k-1}$. Note that α_k can be set to ensure that the orthogonality condition is satisfied even for $w \in \mathcal{K}_{k+1}(\mathcal{A}, r_0) \setminus \mathcal{K}_k(\mathcal{A}, r_0)$.

A possible way to construct p_k is to compute the matrix-vector product $\mathcal{A}p_{k-1}$ and then orthogonalize the resulting vector against all the previous p_i, $i = 1, \ldots, k-1$. This operation seemingly requires the storage of all the previous basis vectors and a computational effort which increases with k. However, the symmetry of the matrix $\mathcal{A}$ leads to the interesting property that $\mathcal{A}p_{k-1}$ is already $\mathcal{A}-$orthogonal to $p_1, \ldots, p_{k-3}$, and hence only the orthogonality with respect to p_{k-1} and p_{k-2} has to be imposed. As a consequence, the basis vectors of the Krylov subspace can be generated using a *short-term recurrence*, allowing a significant saving of memory and computational cost. This very important property holds also for the MINRES method, and in general for any Krylov method for symmetric systems.

It should be pointed out that the above discussion holds if exact arithmetic is assumed. Indeed, in finite precision arithmetic, algebraic errors may cause the basis vectors p_i to lose their $\mathcal{A}$-orthogonality quite soon. This was already noticed in the original paper of 1952 by Hestenes and Stiefel [70, Section 8], where the CG method was first presented. This loss of orthogonality in turn may cause a *delay* in the convergence of the method. We refer the reader to [64, Chapter 4], [85, Section 5.9] for a discussion on the behavior of CG and MINRES in finite precision arithmetic.

2.2.2 The nonsymmetric case

When $\mathcal{A}$ is nonsymmetric, or more precisely nonnormal (that is $\mathcal{A}^T\mathcal{A} \neq \mathcal{A}\mathcal{A}^T$), the convergence analysis of Krylov methods is harder than in the symmetric case. Here we consider the case of a Krylov method satisfying property (ii), such as the Generalized Minimum Residual (GMRES) method [115]. For simplicity we suppose that $\mathcal{A}$ is diagonalizable, i.e., $\mathcal{A} = X\Lambda X^{-1}$. Then we have

$$\|r_k\| = \min_{p\in\mathbb{P}_k, p(0)=1} \|p(\mathcal{A})r_0\| \leq \min_{p\in\mathbb{P}_k, p(0)=1} \|p(\mathcal{A})\| \, \|r_0\| =$$

$$= \min_{p\in\mathbb{P}_k, p(0)=1} \left\|Xp(\Lambda)X^{-1}\right\| \|r_0\| \leq \kappa(X) \min_{p\in\mathbb{P}_k, p(0)=1} \max_{\lambda\in\mathrm{spec}(\mathcal{A})} |p(\lambda)| \, \|r_0\| \,. \tag{2.14}$$

We first assume that $\mathcal{A}$ is normal, that is $\mathcal{A}^T\mathcal{A} = \mathcal{A}\mathcal{A}^T$. Then X is unitary, implying that $\kappa(X) = 1$, and the above bound is sharp in the same sense as (2.7)-(2.8), and convergence can be analyzed exactly as in the symmetric case. If, on the other hand, $\mathcal{A}$ is nonnormal, the bound is not guaranteed to

be sharp. In particular, if $\kappa(X) \gg 1$ then (2.14) might be a large overestimate of the true residual norm, even in the worst-case scenario.

As a consequence, in the nonnormal case information on the spectrum of $\mathcal{A}$ is not enough to predict the (worst-case) behavior of a Krylov method. Even worse, in some cases such information might be misleading. Indeed, as mentioned in the Introduction, it has been proved that for GMRES, any nonincreasing curve of relative residual norms is possible regardless of the eigenvalues of $\mathcal{A}$ [65].

We mention that (2.14) is not the only possible bound for the convergence of GMRES. Indeed, other bounds have been shown and analyzed in the literature, based not on the conditioning of the eigenvector matrix, but rather on the field of values or on the pseudospectrum of $\mathcal{A}$ [39, 94, 37, 38, 7]. An analysis of the convergence of GMRES that takes into account the initial residual can be found, e.g., in [82, 83, 84, 81, 129].

Another complication arising when a Krylov method is applied to a nonsymmetric system is that in general it is not possible to generate iterates with property (ii) using just a short-term recurrence for the update p_k in (2.13), as in the symmetric case. Indeed, in general all the basis vectors p_j, $j = 1, \ldots, k$, need to be stored and used to construct an optimal approximate solution v_{k+1}, as it happens in the standard implementation of GMRES. This fact may substantially increase the memory storage and the computational cost of the method, in particular if the iteration step k is high. Thus, the standard GMRES method is feasible only if we expect that only a moderate number of iterations is required to achieve convergence (which typically means that we have a good preconditioner at hand). If the iteration step k gets too high, a common approach is to *restart* GMRES, using the last approximate solution as the new initial guess.

It is still possible to devise a Krylov method for nonsymmetric systems which relies on a short-term recurrence, but then the optimality properties (i) or (ii) cannot be achieved. Krylov methods belonging to this class include the Biconjugate Gradient method (BiCG) [45], its stabilized version (BiCGStab) [132] and the Induced Dimnsion Reduction (IDR) method [135, 123], where the basis vectors of the Krylov subspace still satisfy an orthogonality condition [45, 122]. Other Krylov methods with short-term recurrence but no optimality property include the Quasi-Minimal Residual method (QMR) [52], and its transpose-free variant (TFQMR) [51]. None of these methods, however, will be further discussed in this thesis.

We finally mention that, in some cases, a possible approach to work around the limitations of nonsymmetry is to employ a method for symmetric systems implemented in a nonstandard inner product. See Section 4.2 for more details on this approach.

2.3 Preconditioners for saddle point systems

In practical computations, Krylov methods are very rarely applied without any preconditioning strategy. This is especially true in the case of saddle point systems, because the indefiniteness and the (typically high) condition number of $\mathcal{A}$ would make the convergence of Krylov methods unacceptably slow. As a consequence, a great deal of research is devoted to the development of efficient and reliable preconditioners.

2.3.1 Right, left and split preconditioners

Given the system $\mathcal{A}v = b$, we consider the left-preconditioned system

$$\mathcal{P}^{-1}\mathcal{A}v = \mathcal{P}^{-1}b \tag{2.15}$$

where $\mathcal{P}$ is a nonsingular matrix, known as *preconditioner*. It is apparent that system (2.15) has the same solution of the original system. Note that if we attempt to solve (2.15) with an iterative method, the preconditioned system does not need to be actually formed, as the coefficient matrix is needed only to compute matrix-vector products. Thus, at each iteration of the method, one matrix-vector product involving $\mathcal{A}$ and one involving $\mathcal{P}^{-1}$ (or, equivalently, one solve with $\mathcal{P}$) have to be computed.

If $\mathcal{P}^{-1}\mathcal{A}$ has more favorable spectral properties (e.g., more clustered eigenvalues) than $\mathcal{A}$, and the computational cost of applying $\mathcal{P}^{-1}$ to a vector is affordable, then it might be advantageous apply a Krylov method on (2.15) rather than on the original system.

Instead of (2.15), we can also consider the right-preconditioned system

$$\mathcal{A}\mathcal{P}^{-1}u = b \qquad \text{with} \quad u = \mathcal{P}v \tag{2.16}$$

Note that in this case, from the iterates u_k corresponding to the new system, we can recover the iterates v_k we are interested in with one additional solve with $\mathcal{P}$ (this is typically done only once at the end of the method).

If $\mathcal{P}$ is positive definite, then using the Cholesky factorization $\mathcal{P} = \mathcal{L}\mathcal{L}^T$ we can also consider the split-preconditioned system

$$\mathcal{L}^{-1}\mathcal{A}\mathcal{L}^{-T}u = \mathcal{L}^{-1}b, \qquad \text{with} \quad u = \mathcal{L}^T v. \tag{2.17}$$

Thus, once a preconditioner has been selected, one has to choose how to apply it to the system. A first observation shows that all the preconditioned matrices $\mathcal{P}^{-1}\mathcal{A}$, $\mathcal{A}\mathcal{P}^{-1}$ and $\mathcal{L}^{-1}\mathcal{A}\mathcal{L}^{-T}$ are similar, and hence they have the same eigenvalues. Still, there are differences that can make one approach

preferable over the others, depending on the properties of $\mathcal{A}$ and $\mathcal{P}$ and on the Krylov method employed.

Suppose that $\mathcal{A}$ is symmetric. Then if $\mathcal{P}$ is positive definite, the split-preconditioned (2.17) system is still symmetric. This fact allows the use of Krylov method for symmetric systems, like MINRES or CG, whose iterates can be generated using a short-term recurrence and whose convergence can be bounded as in (2.10) or (2.11). A very important feature of this approach is that there the Cholesky decomposition of $\mathcal{A}$ does not need to be actually computed, nor linear systems involving $\mathcal{L}$ have to be actually solved. Instead, the generation of the Krylov subspace associated with (2.17) can be done implicitly, and only matrix-vector products involving $\mathcal{P}^{-1}$ or $\mathcal{A}$ have to be computed. We refer to [64, Chapter 8] for the implementation details for CG and MINRES.

If the chosen preconditioner $\mathcal{P}$ is not SPD then (2.17) cannot be used and the preconditioned system is in general nonsymmetric. As a consequence, an iterative method for nonsymmetric systems, like GMRES, has to be employed. When GMRES is considered, it is still important to distinguish between the different formulations (2.15) and (2.16). To show the difference between the two approaches, let v_k^L and u_k^R be the approximate solutions at step k generated respectively by left-preconditioned and right-preconditioned GMRES. We suppose for simplicity $v_0 = 0$ (equivalently, $r_0 = b$) and observe that

$$v_k^L \in \mathcal{K}_k(\mathcal{P}^{-1}\mathcal{A}, \mathcal{P}^{-1}b) = \text{span}\left\{\mathcal{P}^{-1}b, \mathcal{P}^{-1}\mathcal{A}\mathcal{P}^{-1}b, \ldots, \left(\mathcal{P}^{-1}\mathcal{A}\right)^{k-1}\mathcal{P}^{-1}b\right\},$$

and

$$u_k^R \in \mathcal{K}_k(\mathcal{A}\mathcal{P}^{-1}, b) = \text{span}\left\{b, \mathcal{A}\mathcal{P}^{-1}b, \ldots, \left(\mathcal{A}\mathcal{P}^{-1}\right)^{k-1} b\right\}.$$

Writing $v_k^R = \mathcal{P}^{-1}u_k^R$, we have

$$v_k^R \in \mathcal{P}^{-1}\mathcal{K}_k(\mathcal{A}\mathcal{P}^{-1}, b) = \text{span}\left\{\mathcal{P}^{-1}b, \mathcal{P}^{-1}\mathcal{A}\mathcal{P}^{-1}b, \ldots, \mathcal{P}^{-1}\left(\mathcal{A}\mathcal{P}^{-1}\right)^{k-1} b\right\} =$$

$$= \text{span}\left\{\mathcal{P}^{-1}b, \mathcal{P}^{-1}\mathcal{A}\mathcal{P}^{-1}b, \ldots, \left(\mathcal{P}^{-1}\mathcal{A}\right)^{k-1}\mathcal{P}^{-1}b\right\}.$$

It is then apparent that v_k^R and v_k^L belong to the same subspace. We recall from the previous section that the GMRES iterates minimize the residual norm over the appropriate Krylov subspace. However, the residuals of the two systems (2.15) and (2.16) are different, and thus v_k^R and v_k^L minimize different quantities. More precisely, v_k^R is the vector of $\mathcal{K}_k(\mathcal{P}^{-1}\mathcal{A}, \mathcal{P}^{-1}b)$ which minimizes the Euclidean norm of $b - \mathcal{A}\mathcal{P}^{-1}u = b - \mathcal{A}v$, i.e., the residual of the original system. On the other hand, v_k^L is the vector of the same subspace which minimizes the norm of $\mathcal{P}^{-1}(b - \mathcal{A}v)$, which is the residual of an "artificial" system. As a consequence, when using GMRES the right-preconditioned system (2.16) is typically preferred.

2.3.2 Some well-known preconditioners

We now report three well-known classes of preconditioners, very popular in the context of saddle point systems, that will be used in the rest of the thesis. We also mention the spectral properties of their ideal versions.

Consider a saddle point matrix of the form (2.1). For the rest of the chapter, we will assume that the (1,1) block, A, is positive definite (see Chapter 3 for a preconditioning strategy that can also be applied when A is singular). The block diagonal preconditioner for $\mathcal{A}$ is defined as:

$$\mathcal{P}_D = \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \tag{2.18}$$

where, as before, $S = C + BA^{-1}B^T$ is the Schur complement of A. We recall that S is positive definite if and only if $\mathcal{A}$ is nonsingular. Under this assumption, we note that $\mathcal{P}_D$ is symmetric and positive definite, thus it has the important property of preserving the symmetry of $\mathcal{A}$. The following theorem characterizes the spectral properties of the preconditioned system in the special case when $C = 0$.

Proposition 2.3.1. [93, Proposition 1] *Let $\mathcal{A}$ be a saddle point matrix with A positive definite, $\ker(B^T) = 0$ and $C = 0$, and let $\mathcal{P}_D$ as in (2.18). Then, if $\mathcal{M} = \mathcal{P}_D^{-1}\mathcal{A}$, it holds*

$$(\mathcal{M} - I)\left(\mathcal{M} - \frac{1+\sqrt{5}}{2}I\right)\left(\mathcal{M} - \frac{1-\sqrt{5}}{2}I\right) = 0.$$

In particular, $\mathcal{M}$ is diagonalizable and $\operatorname{spec}(\mathcal{M}) = \left\{1, \frac{1 \pm \sqrt{5}}{2}\right\}$.

The above Theorem shows that MINRES (or GMRES) applied to the preconditioned system $\mathcal{P}_D^{-1}\mathcal{A}v = \mathcal{P}_D^{-1}b$ converges in at most three iterations.

If, on the other hand, $C \neq 0$, then the spectral properties of $\mathcal{P}_D^{-1}\mathcal{A}$ are not so favorable. Nevertheless, the following theorem can be proved.

Proposition 2.3.2. [61, Proposition 4.2] *Let $\mathcal{A}$ be a nonsingular saddle point matrix with A positive definite, and let $\mathcal{P}_D$ be as in (2.18). Moreover, let Θ be the set of finite eigenvalues of the matrix pencil $\left(C + BA^{-1}B^T, C\right)$. Then*

$$\operatorname{spec}(\mathcal{P}_D^{-1}\mathcal{A}) = \left\{1, \frac{1 \pm \sqrt{5}}{2}\right\} \cup \left\{\frac{1}{2\theta}\left(\theta - 1 \pm \sqrt{(1-\theta)^2 + 4\theta^2}\right) \mid \theta \in \Theta\right\}.$$

We emphasize that (2.18) represents an *ideal* preconditioner. Indeed, the application of the block diagonal preconditioner as it is in (2.18) would be almost as expensive as the actual computation of the inverse of $\mathcal{A}$ (see e.g. [11, Section 10.1.1]).

Thus, in practical computations the matrices A and S in (2.18) have to be replaced with easy-to-invert approximations, which we denote with $\widehat{A}$ and $\widehat{S}$. This necessity holds for all the preconditioners presented in this section; indeed, the main difficulty associated with their use in practice is finding good enough approximations for A and S which are at the same time computationally cheap. This task requires a deep knowledge on the origin and properties of the blocks of $\mathcal{A}$, and as a consequence the choice of $\widehat{A}$ and $\widehat{S}$ is strongly problem-dependent. Algebraic preconditioners, such as incomplete factorizations, sparse approximate inverses and multigrid methods might be helpful in some important cases (see Chapters 5 and 7 for two examples). We refer to [11, Section 10.1.3] and references therein for a review of some techniques employed in different contexts. Here we just mention that finding a suitable $\widehat{S}$ is typically a much harder problem than finding a suitable $\widehat{A}$.

Next, we consider the (ideal) *block triangular* preconditioner

$$\mathcal{P}_T = \begin{bmatrix} A & B^T \\ 0 & -S \end{bmatrix}. \tag{2.19}$$

To apply the inverse of $\mathcal{P}_T$ to a vector we can use the following factorization

$$\mathcal{P}_T^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & I_m \end{bmatrix} \begin{bmatrix} I_n & B^T \\ 0 & -I_m \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & S^{-1} \end{bmatrix},$$

which shows that the main computational effort required to apply $\mathcal{P}_D^{-1}$ is one solve with A and one with S, with the understanding that, as before, A and S have to be replaced with suitable approximations. Thus, the computational cost associated with $\mathcal{P}_T$ is very similar to the one associated with $\mathcal{P}_D$.

We now give a proposition which characterizes the eigenvalues of $\mathcal{P}_T^{-1}\mathcal{A}$.

Proposition 2.3.3. [73, Proposition 2] *Let $\mathcal{A}$ be a nonsingular saddle point matrix with A positive definite, and let $\mathcal{P}_T$ be as in* (2.19)*. Then it holds*

$$\left(\mathcal{P}_T^{-1}\mathcal{A} - I\right)^2 = 0.$$

In particular, $\operatorname{spec}(\mathcal{P}_T^{-1}\mathcal{A}) = \{1\}$.

This result shows that GMRES applied to the system $\mathcal{P}_T^{-1}\mathcal{A}v = \mathcal{P}_T^{-1}b$ converges in at most two iterations.

Unlike $\mathcal{P}_D$, $\mathcal{P}_T$ is not SPD and thus the preconditioned system is nonsymmetric. The use of a nonsymmetric solver, like GMRES, is then required. This can be regarded as a disadvantage, but in fact if the method converges in few iterations the extra computational cost required for GMRES is negligeable.

Finally, if $\widehat{A}$ denotes an approximation of A, which for simplicity we assume to be symmetric and positive definite, we define the constraint preconditioner as

$$\mathcal{P}_C = \begin{bmatrix} \widehat{A} & B^T \\ B & -C \end{bmatrix} \tag{2.20}$$

This preconditioner takes its name by the fact that if v^* is a solution for the system $\mathcal{P}_C v = b$, then v^* satisfies also the second equation of the original saddle point system, which in many applications represents some sort of constraint.

Even more so than in the previous cases, in $\mathcal{P}_C$ linear systems involving $\widehat{A}$ must be inexpensive to solve. This a consequence of the factorization

$$\mathcal{P}_C^{-1} = \begin{bmatrix} I_n & -\widehat{A}^{-1}B^T \\ 0 & I_m \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & -\left(C + B\widehat{A}^{-1}B^T\right)^{-1} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ -B & I_m \end{bmatrix} \begin{bmatrix} \widehat{A}^{-1} & 0 \\ 0 & I_m \end{bmatrix}$$

which shows that two solves with $\widehat{A}$ are needed at each application of the preconditioner. Thus, $\widehat{A}$ is often taken as a diagonal matrix. One solve with $C + B\widehat{A}^{-1}B^T$ has to be carried out as well, and this matrix is typically replaced with some suitable approximation.

To analyze the spectral properties $\mathcal{P}_C^{-1}\mathcal{A}$, we assume $C = 0$. We also assume $m < n$ (the case $n = m$ is trivial). Then the following proposition hold.

Theorem 2.3.4. [77, Theorem 2.1] *Let $\mathcal{A}$ be a saddle point matrix with A positive definite, $\ker(B^T) = 0$, $C = 0$ and $m < n$. Let $\mathcal{P}_C$ be as in (2.20), and suppose it is nonsingular. Moreover, let $Z \in \mathbb{R}^{n\times(n-m)}$ be a matrix whose columns form a basis for $\ker(B)$. Then the matrix $\mathcal{P}_C^{-1}\mathcal{A}$ has the eigenvalue 1 with multiplicity $2m$. Then the remaining $n-m$ eigenvalues are the solution of the generalized eigenvalue problem*

$$Z^T A Z v = \lambda Z^T \widehat{A} Z v \qquad v \in \mathbb{R}^{n-m}$$

In particular, $\mathcal{P}_C^{-1}\mathcal{A}$ has real and positive eigenvalues.

When $C \neq 0$ the eigenvalues of $\mathcal{P}_C^{-1}\mathcal{A}$ are more difficult to analyze (unless C is positive definite, see [32, Theorem 3.1]). As a consequence, spectral estimates for this case are more complicated, and we do not show them here for the sake of simplicity. We instead refer the reader to [6, Theorem 3.1 and Theorem 4.1].

Chapter 3

The augmented block diagonal preconditioner[1]

The development of solvers for saddle point systems is a very active area of research, and as a consequence a variety of preconditioning techniques have been proposed in the literature. Some of them, such as the block diagonal, block triangular and constraint preconditioners were discussed in the previous chapter. In this chapter we introduce another kind of preconditioner, the augmented block diagonal.

This preconditioner is based on the "augmentation" of the (1,1) block of the original system, which allows it to cope with the possible high singularity of both diagonal blocks. Moreover, it is symmetric and positive definite, so that it preserves the symmetry of the problem. This preconditioner was proposed in 2006 by Greif and Schötzau [68], although back to 2003 the same preconditioner (for a particular choice of W) was presented by Powell and Silvester [108, Section 2], in the context of second-order elliptic problems discretized using Raviart-Thomas mixed formulation. This approach was further studied in [110, 67]. In [24], Cao extended it to the nonsymmetric case.

To our knowledge, however, only the case $C = 0$ has been considered in the literature (i.e. the saddle point system has a zero (2,2) block). The main purpose of this chapter is to fill this gap, by providing spectral estimates for the preconditioned system in the case $C \neq 0$. In Section 3.1 we introduce the preconditioner and review the known spectral results. In section 3.2 we generalize these estimates and discuss some the optimal theoretical choices for the augmentation. Finally, in Section 3.3 we show some numerical experiments which illustrate the quality of our estimates, using the KKT systems

[1]The results presented in this chapter are taken from [91].

discussed in Chapters 6 and 7 as test matrices.

3.1 The preconditioner

Consider a nonsingular saddle point system of the form

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{3.1}$$

where, as usual, $A \in \mathbb{R}^{n\times n}$ symmetric and positive semidefinite, $B \in \mathbb{R}^{m\times n}$ with $n \geq m$, and $C \in \mathbb{R}^{m\times m}$ symmetric and positive semidefinite. We are mainly interested in the case when the (1,1) of the matrix is singular or very ill-conditioned. None of the preconditioners discussed in the previous chapter (namely the block diagonal, block triangular and constraint preconditioner) can be applied in this case, as they all require A to be nonsingular.

A possible way to work around this limitation is to use augmented Lagrangian techniques, that is, to replace the original system with an equivalent one where the (1,1) block is positive definite. Namely, if $C = 0$, then the system $\mathcal{A}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}$ is replaced with the "augmented" system

$$\begin{bmatrix} A + B^T W^{-1} B & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f + B^T W^{-1} g \\ g \end{bmatrix},$$

where $W \in \mathbb{R}^{m\times m}$, symmetric and positive definite. Then any of the preconditioners discussed in Chapter 2 can be applied on the augmented system (of course taking into account the new form of the (1,1) block), see [54, 55, 13, 14] and references therein. We give a special mention to [14], where a similar technique is applied when $C \neq 0$.

A related but different approach is to let the original system unchanged and still consider the augmentation in the preconditioner, that is

$$\mathcal{P}_{AD} = \begin{bmatrix} A + B^T W^{-1} B & 0 \\ 0 & W \end{bmatrix}. \tag{3.2}$$

Note that the condition $\ker(A) \cap \ker(B) = 0$ implies that $\mathcal{P}_{AD}$ is nonsingular.

It is apparent that $\mathcal{P}_{AD}$ is symmetric and positive definite. This important property allows the preservation of the symmetry of the original problem, as discussed in Chapter 2. Thus, a short-term recurrence like MINRES can be applied on the preconditioned system and its rate of convergence can be bounded by a term which depends only on the eigenvalues of the coefficient matrix.

At each application of the preconditioner, a linear system with coefficient matrix $A + B^T W^{-1} B$ has to be solved. This may be extremly costly, also because $A + B^T W^{-1} B$ could be much denser than each of its terms [110]. In fact $\mathcal{P}_{AD}$ in (3.2) is just an "ideal" preconditioner; in practice, some form (symmetric) approximation for the inverse of $A + B^T W^{-1} B$ has to be considered, which takes into account the specific application data. In any case, the matrix W has to be inexpensive to invert, e.g. diagonal, or a multiple of the identity matrix.

The spectral properties of this preconditioner have been studied in the case when $C = 0$ [68, 67]. Note that in this case we have to assume that B^T has full column rank to ensure the nonsingularity of $\mathcal{A}$. We report the following result regarding the eigenvalues of the preconditioned system. Recall that the nullity of a matrix if defined as the dimension of its null space.

Theorem 3.1.1. [68, Theorem 2.2] *Suppose* $\ker(B^T) = 0$ *and that* A *is positive semidefinite with nullity* r*. Then* $\mathcal{P}_{AD}^{-1}\mathcal{A}$ *has the eigenvalue* $\theta = 1$ *with algebraic multiplicity* n *and the eigenvalue* $\theta = -1$ *with algebraic multiplicity* r*. The remaining* $m - r$ *eigenvalues belong to the open interval (-1,0) and satisfy*

$$\theta_i = -\frac{\phi_i}{\phi_i + 1}, \qquad i = 1, \ldots, m - r \tag{3.3}$$

where the ϕ_i *are the* $m - r$ *positive solutions of the generalized eigenvalue problem*

$$\phi A v = B^T W^{-1} B v. \tag{3.4}$$

According to this theorem, the greater the nullity of A, the greater the multiplicity of the eigenvalue -1 for the preconditioned system. If the nullity of A is equal to m (the maximum possible if $\mathcal{A}$ is nonsingular) then $\mathcal{P}_{AD}^{-1}\mathcal{A}$ has only two distinct eigenvalues.

If A is nonsingular, then the m positive ϕ that satisfy (3.4) are the generalized eigenvalues of

$$\phi W u = B A^{-1} B^T u.$$

If, moreover, $W = \gamma I$ with $\gamma > 0$, then the eigenvalues of $\mathcal{P}_{AD}^{-1}\mathcal{A}$ different from ± 1 are (cf. [67, Theorem 2.4])

$$\theta_i = -\frac{\phi_i}{\phi_i + \gamma}, \qquad i = 1, \ldots, m,$$

where the ϕ_i are the eigenvalues of $B A^{-1} B^T$. The above relation allows us to make some comments on the choice of γ. As argued in [67], a small value of γ leads to a clustering of the negative eigenvalues in a neighborhood of -1.

However, if a too small value of γ is chosen, the preconditioner (3.2) may become very ill-conditioned. Thus, one has to chose a compromise value; see [68, 67, 110] for some empirical choices of γ. We also mention that a thorough algebraic analysis on the choice of W can be found in [54, 55], where however the focus was not the use of W within the context of (3.2).

We conclude this section, we mention that the possibility to augment the (1,1) block of the preconditioner is not restricted to the block diagonal case. Indeed, an augmented block triangular preconditioner has been introduced in [110]. In [119], Shen et al. studied its spectral properties for $C \neq 0$. See also [24, 28, 72] for some related preconditioning strategies. We will not discuss further this class of preconditioners in this chapter, although we will use it later in Chapter 7.

3.2 New spectral estimates for the case $C \neq 0$

We first give spectral bounds for the preconditioned matrix, which appear to be new for augmentation type preconditioned matrices with nonzero (2,2) block. In light of this analysis, we then discuss the choice of the matrix W.

The spectral estimates derived in this section hold for a general saddle point matrix like the one appearing in (3.1). However, differently from the convention, we do not require that $n \geq m$. We recall that, when $C \neq 0$, the condition $\ker(B)^T = 0$ is no longer necessary to ensure the nonsingularity of $\mathcal{A}$. Thus, we assume instead that the conditions of Proposition are satisfied.

3.2.1 Spectral estimates for B^T of full column rank

We first consider the case when $\ker(B^T) = \{0\}$.

Theorem 3.2.1. *Let $\mathcal{A}$ be a nonsingular saddle point system (3.1) with $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times m}$ be symmetric and positive semidefinite, $B \in \mathbb{R}^{m\times n}$ and suppose that $\ker(B^T) = \{0\}$. Let $W \in \mathbb{R}^{n\times n}$ be symmetric and positive definite.*

Let $c_0 = \lambda_{\min}(W^{-1}C)$, $c_1 = \lambda_{\max}(W^{-1}C)$, $b_0 = \lambda_{\min}\left(W^{-1}BB^T\right)$ and $b_1 = \lambda_{\max}\left(W^{-1}BB^T\right)$. Then, given $\mathcal{P}_{AD}$ in (3.2), it holds

$$\operatorname{spec}(\mathcal{P}_{AD}^{-1}\mathcal{A}) \subseteq I^- \cup I^+ = [\xi_1,\ \xi_2] \cup [\xi_3,\ 1]\,,$$

where

$$\begin{aligned}
\xi_1 &= \frac{(1-c_1)s - c_1 - \sqrt{((1-c_1)s-c_1)^2 + 4(1+c_1 s)(1+s)}}{2(1+s)},\\
\xi_2 &= \frac{1 - tc_0 - \sqrt{(1-tc_0)^2 + 4t\,(t-1+c_0)}}{2t},\\
\xi_3 &= \frac{(1-c_1)s - c_1 + \sqrt{((1-c_1)s-c_1)^2 + 4(1+c_1 s)(1+s)}}{2(1+s)},
\end{aligned}$$

with $s = \dfrac{\lambda_{\min}(A)}{b_1}$ *and* $t = 1 + \dfrac{b_0}{\|A\|}$.

Moreover, if ℓ denotes the nullity of B, then $\mathcal{P}_{AD}^{-1}\mathcal{A}$ has the eigenvalue 1 with multiplicity ℓ.

Proof. We first observe that the hypothesis on B yields $b_0 > 0$.

Consider the generalized eigenvalue problem $\mathcal{A}\begin{bmatrix} u \\ v \end{bmatrix} = \theta \mathcal{P}_{AD}\begin{bmatrix} u \\ v \end{bmatrix}$, i.e.

$$\begin{aligned}
Au + B^T v &= \theta\left(A + B^T W^{-1} B\right) u, &(3.5)\\
Bu - Cv &= \theta W v &(3.6)
\end{aligned}$$

and first suppose that $\theta > 0$. Since any vector $(u, 0)$ with $u \in \ker(B)$ satisfies (3.5) and (3.6) with $\theta = 1$, $\mathcal{P}_{AD}^{-1}\mathcal{A}$ has the eigenvalue one with multiplicity ℓ.

Suppose now $u \notin \ker(B)$. Since $\theta W + C$ is positive definite, we eliminate v from (3.6) and substitute it into (3.5). Further, we premultiply the resulting equation by u^T and obtain

$$u^T(A + B^T\left(\theta W + C\right)^{-1} B)u = \theta u^T\left(A + B^T W^{-1} B\right) u. \tag{3.7}$$

By using the inequality $(\theta W + C)^{-1} \preceq \dfrac{1}{\theta} W^{-1}$, and after some rearrangement we get

$$(1-\theta) u^T\left(\theta A + (1+\theta) B^T W^{-1} B\right) u \geq 0.$$

Noting that $\left(\theta A + (1+\theta) B^T W^{-1} B\right)$ is positive definite, it follows $\theta \leq 1$.

To show the lower bound for the positive eigenvalues, we reformulate (3.7) as

$$u^T B^T W^{-1/2}\left(\theta I_m - \left(\theta I_m + \widetilde{C}\right)^{-1}\right) W^{-\frac{1}{2}} B u = (1-\theta) u^T A u,$$

where $\widetilde{C} = W^{-\frac{1}{2}} C W^{-\frac{1}{2}}$ and note that

$$A \succeq \lambda_{\min}(A) I_n, \quad \theta I_m - \left(\theta I_m + \widetilde{C}\right)^{-1} \preceq \left(\theta - \frac{1}{\theta + c_1}\right) I_m, \quad B^T W^{-1} B \preceq b_1 I_n,$$

where the last inequality follows from the fact that $B^T W^{-1} B$ and $W^{-1} B B^T$ admit the same maximum eigenvalue. Using these inequalities and dividing by $\|u\|^2$, we find that θ satisfies

$$\left(\theta - \frac{1}{\theta + c_1}\right) b_1 \geq (1-\theta)\lambda_{\min}(A),$$

which is equivalent to

$$\theta^2 (1+s) + \theta ((c_1 - 1)s + c_1) - (1 + c_1 s) \geq 0,$$

with $s = \lambda_{\min}(A)/b_1$. The expression of the left extreme of I^+ readily follows.

Let us now consider $\theta < 0$. Equation (3.5) can be rewritten as

$$\left((1-\theta)A - \theta B^T W^{-1} B\right) u = -B^T v,$$

and the matrix on the left-hand side is now positive definite. This implies that $v \neq 0$ otherwise we would also have $u = 0$. If we eliminate u from (3.5), substitute into (3.6) and premultiply the resulting equation by $W^{-\frac{1}{2}}$ we get

$$\widetilde{B}\left((1-\theta)A - \theta \widetilde{B}^T \widetilde{B}\right)^{-1} \widetilde{B}^T w + \widetilde{C} w = -\theta w, \tag{3.8}$$

where $v = W^{-\frac{1}{2}} w$, $\widetilde{C} = W^{-\frac{1}{2}} C W^{-\frac{1}{2}}$, $\widetilde{B} = W^{-\frac{1}{2}} B$. It holds

$$\begin{aligned} \widetilde{B}\left((1-\theta)A - \theta \widetilde{B}^T \widetilde{B}\right)^{-1} \widetilde{B}^T &\preceq \widetilde{B}\left((1-\theta)\lambda_{\min}(A) I_n - \theta \widetilde{B}^T \widetilde{B}\right)^{-1} \widetilde{B}^T \\ &\preceq \frac{b_1}{(1-\theta)\lambda_{\min}(B) - \theta b_1} I_m, \end{aligned}$$

since the eigenvalues of $\widetilde{B}\left(\|A\|(1-\theta) I_n - \theta \widetilde{B}^T \widetilde{B}\right)^{-1} \widetilde{B}^T$ are of the form $\sigma^2/((1-\theta)\|A\| - \theta\sigma^2)$, with σ being a singular value of $\widetilde{B}^T$. We now multiply (3.8) by w^T and use the above inequality as well as $\widetilde{C} \preceq c_1 I_m$. Then, after some rearrangements we obtain

$$\theta^2 (1+s) + \theta ((c_1 - 1)s + c_1) - (1 + c_1 s) \leq 0,$$

with s as above, from which the lower bound for the negative eigenvalues follows.

To conclude, we prove the upper bound for the negative eigenvalues. We have $\widetilde{C} \succeq c_0 I_m$, as well as

$$\begin{aligned} \widetilde{B}\left((1-\theta)A - \theta \widetilde{B}^T \widetilde{B}\right)^{-1} \widetilde{B}^T &\succeq \widetilde{B}\left(\|A\|(1-\theta) I_n - \theta \widetilde{B}^T \widetilde{B}\right)^{-1} \widetilde{B}^T \\ &\succeq \frac{b_0}{(1-\theta)\|A\| - \theta b_0} I_m, \end{aligned}$$

using again the form of the eigenvalues of $\widetilde{B}\left(\|A\|(1-\theta)I_n - \theta\widetilde{B}^T\widetilde{B}\right)^{-1}\widetilde{B}^T$. We now multiply (3.8) from the left by w^T. Using the above inequalities and dividing by $\|w\|^2$, we obtain

$$\frac{b_0}{(1-\theta)\|A\| - \theta b_0} + c_0 \leq -\theta,$$

which implies

$$t\theta^2 - (1 - tc_0)\,\theta - t + 1 - c_0 \geq 0,$$

with $t = 1 + \dfrac{b_0}{\|A\|}$, and the stated expression for the right extreme of I^-. □

We now show that our estimates are consistent with the ones given by Theorem 3.1.1 when $C = 0$. In this special case, the bounds of Theorem 3.2.1 reduce to

$$I^- = \left[-\frac{1}{1+s}, -\frac{t-1}{t}\right], \qquad I^+ = \{1\}.$$

The interval I^+ and the left extreme of I^- coincide with the corresponding bounds in Theorem 3.1.1, when A is singular, i.e. $s = 0$.

The comparison between the right extreme of I^- and the expression for the negative eigenvalues provided by equations (3.3) and (3.4) requires a longer discussion. We first show that $\lambda \geq b_0/\|A\|$. Indeed, if in (3.4) we write $u = u_1 + u_2$, with $u_1 \in \ker(B)$ and $u_2 \in \ker(B)^\perp$ (note that if (3.4) holds, then $u_2 \neq 0$) and premultiply it by u_1^T, we obtain $\phi u_1^T A u_1 + \phi u_1^T A u_2 = 0$, which shows that $u_1^T A u_2$ is nonpositive. Premultiplying (3.4) by u_2^T we obtain

$$\phi u_2^T A(u_1 + u_2) = u_2^T B^T W^{-1} B u_2.$$

By using the inequalities

$$u_2^T A u_1 \leq 0, \qquad u_2^T A u_2 \leq \|A\|\,\|u_2\|^2 \quad \text{and} \quad u_2^T B^T W^{-1} B u_2 \geq b_0\,\|u_2\|^2,$$

we can conclude that $\phi \geq b_0/\|A\|$. Finally, we observe that the function $-\phi/(1+\phi)$ is monotonically decreasing, so that we get

$$\theta \leq -\frac{\dfrac{b_0}{\|A\|}}{1 + \dfrac{b_0}{\|A\|}}. \tag{3.9}$$

This bound coincides with the right extreme of I^-. If C is nonzero but singular, the upper bound for the negative eigenvalues is again given by (3.9). Analogously, if A is nonsingular, so that $\lambda_{\min}(A) > 0$, we get $\phi \leq b_1/\lambda_{\min}(A) = 1/s$, and the monotonicity of $-\lambda/(1+\lambda)$ ensures that $\theta \geq -1/(1+s)$, which is the left extreme of I^- above.

3.2.2 Spectral estimates for B^T column-rank deficient

In Theorem 3.2.1 we supposed $b_0 > 0$ to ensure that the left extreme of I^- is meaningful. Indeed, if b_0 were zero the upper bound for the negative eigenvalues would reduce to $\theta \leq -c_0$, which vanishes if C is singular. Even in case C is positive definite, c_0 may be very small in practice, making the bound little representative of the largest negative eigenvalue of $\mathcal{P}_{AD}^{-1}\mathcal{A}$. Thus, when C is singular the preceding upper bound for the negative eigenvalues has to be refined.

Theorem 3.2.2. *Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{m\times m}$ be symmetric and positive semidefinite, $B \in \mathbb{R}^{m\times n}$, $\mathcal{A}$ in (3.1) nonsingular. Let $W \in \mathbb{R}^{m\times m}$ be symmetric and positive definite and $c_1 = \lambda_{\max}(W^{-1}C)$. Suppose that B^T has a nontrivial null space, define*

$$\min_{0\neq x\in\ker(B^T)} \frac{x^TCx}{x^TWx} = c^* > 0,$$

and let b_+ be the minimum positive eigenvalue of $W^{-1}BB^T$. For $\mathcal{P}_{AD}$ the matrix in (3.2), it holds

$$\operatorname{spec}(\mathcal{P}_{AD}^{-1}\mathcal{A}) \subseteq I^- \cup I^+ = [\xi_1,\ \min\{\eta, \xi_2\}] \cup [\xi_3,\ 1],$$

where ξ_1, ξ_2 and ξ_3 are given in Theorem 3.2.1, and $\eta \geq -c^*$ is the largest negative root of the cubic polynomial

$$q(\theta) = t_+\theta^3 + \theta^2\left((c_1 + c^*)t_+ - 1\right) - \theta\left(c_1 + c^* - 1 + t_+\right) - (t_+ - 1)\,c^*, \quad (3.10)$$

with $t_+ = 1 + \dfrac{b_+}{\|A\|}$.

Proof. We only need to prove the upper bound for the negative eigenvalues. From the proof of Theorem 3.2.1 we infer that if θ is a negative eigenvalue of $\mathcal{P}_{AD}^{-1}\mathcal{A}$, then $\theta \leq \xi_2 = -c_0$.

Consider equation (3.8) and suppose that $w \in \ker(\widetilde{B}^T)$. Hence, we have $w^T\widetilde{C}w = -\theta\|w\|^2$, which implies $\theta \leq -c^*$.

We now suppose $\theta > -c^*$, (hence, $w \notin \ker(\widetilde{B}^T)$), and write $w = w_0 + w_1$, with $w_0 \in \ker(\widetilde{B}^T)$ and $w_1 \in \ker(\widetilde{B}^T)^\perp$. We premultiply equation (3.8) by w_0^T to get

$$w_0^T\widetilde{C}w_1 = -w_0^T\widetilde{C}w_0 - \theta\|w_0\|^2 \leq -\left(1 + \frac{\theta}{c^*}\right) w_0^T\widetilde{C}w_0,$$

from which, using the inequality $w_0^T \widetilde{C} w_1 \geq -\left(w_0^T \widetilde{C} w_0\right)^{1/2} \left(w_1^T \widetilde{C} w_1\right)^{1/2}$ we infer

$$-\left(w_0^T \widetilde{C} w_0\right)^{1/2} \geq -\frac{c^*}{c^* + \theta} \left(w_1^T \widetilde{C} w_1\right)^{1/2}.$$

Note that this inequality holds also when $\widetilde{C} w_0 = 0$. Thus,

$$w_0^T \widetilde{C} w_1 \geq -\left(w_0^T \widetilde{C} w_0\right)^{1/2} \left(w_1^T \widetilde{C} w_1\right)^{1/2} \geq -\frac{c^*}{c^* + \theta} w_1^T \widetilde{C} w_1. \tag{3.11}$$

We then premultiply equation (3.8) by w_1^T, and we bound the leftmost term as follows:

$$\begin{aligned} w_1^T \widetilde{B} \left((1-\theta)A - \theta \widetilde{B}^T \widetilde{B}\right)^{-1} \widetilde{B}^T w_1 &\geq w_1^T \widetilde{B} \left(\|A\| (1-\theta) I_n - \theta \widetilde{B}^T \widetilde{B}\right)^{-1} \widetilde{B}^T w_1 \\ &\geq \frac{b_+}{(1-\theta)\|A\| - \theta b_+} \|w_1\|^2, \end{aligned}$$

where the last inequality is justified by the fact that w_1 is orthogonal to the null space of $\widetilde{B} \left(\|A\| (1-\theta) I_n - \theta \widetilde{B}^T \widetilde{B}\right)^{-1} \widetilde{B}^T$, and that the eigenvalues of such matrix are of the form $\dfrac{\sigma^2}{(1-\theta)\|A\| - \theta\sigma^2}$, where σ is a singular value of $\widetilde{B}^T$. Therefore, we obtain

$$\frac{b_+}{(1-\theta)\|A\| - \theta b_+} \|w_1\|^2 + w_1^T \widetilde{C}(w_0 + w_1) \leq -\theta \|w_1\|^2. \tag{3.12}$$

According to the inequality (3.11), it holds

$$w_1^T \widetilde{C}(w_0 + w_1) \geq \left(1 - \frac{c^*}{c^* + \theta}\right) w_1^T \widetilde{C} w_1 \geq \frac{c_1 \theta}{c^* + \theta} \|w_1\|^2.$$

Thus, after dividing (3.12) by $\|w_1\|^2$, we obtain

$$\frac{b_+}{(1-\theta)\|A\| - \theta b_+} + \frac{c_1 \theta}{c^* + \theta} \leq -\theta,$$

and, after some rearrangements,

$$\theta^3 t_+ + \theta^2 \left((c_1 + c^*) t_+ - 1\right) - \theta \left(c_1 + c^* + t_+ - 1\right) - (t_+ - 1)\, c^* \geq 0.$$

where $t_+ = 1 + \dfrac{b_+}{\|A\|}$. If we call $q(\theta)$ the above cubic polynomial and η its largest negative root, then it holds that $\theta \leq \eta$. Since $q(-c^*) \geq 0$, then $-c^* \leq \eta$. □

3.2.3 An improvement of the upper bound on the positive eigenvalues

It turns out that if $\ker(B) = 0$ and C is positive definite, then the upper bound on the positive eigenvalues shown in Theorems 3.2.1 and 3.2.2 can be improved, as stated by the following result.

Theorem 3.2.3. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive semidefinite, $C, W \in \mathbb{R}^{m\times m}$ be symmetric and positive definite. Let $B \in \mathbb{R}^{m\times n}$, and assume that $\ker(B) = 0$ and that $\mathcal{A}$ in (3.1) nonsingular.*

Let $c_0 = \lambda_{\min}(W^{-1}C) > 0$ and $\mathfrak{b}_0 = \lambda_{\min}(B^T W^{-1} B) > 0$. Then, if θ is a positive eigenvalue of $\mathcal{P}_{AD}^{-1}\mathcal{A}$, with $\mathcal{P}_{AD}$ as in (3.2), it holds

$$\theta \leq \frac{1 - c_0\mathfrak{t} + \sqrt{(1-c_0\mathfrak{t})^2 - 4\mathfrak{t}(1-\mathfrak{t}-c_0)}}{2\mathfrak{t}},$$

with $\mathfrak{t} = 1 + \dfrac{\mathfrak{b}_0}{\|A\|}$.

Proof. We consider again equation (3.7) and reformulate it to obtain

$$(1-\theta)u^T A u + u^T \widetilde{B}^T \left(\left(\theta I_m + \widetilde{C}\right)^{-1} - \theta I_m \right) \widetilde{B} u = 0$$

where, as before, $\widetilde{C} = W^{-1}C$ and $\widetilde{B} = W^{-1/2}B$. Using $\left(\theta I_m + \widetilde{C}\right)^{-1} \preceq \dfrac{1}{\theta + c_0} I_m$, and then multiplying by $\theta + c_0$, we obtain

$$(1-\theta)(\theta + c_0) u^T A u + \left(-\theta^2 - c_0\theta + 1\right) \left\|\widetilde{B}u\right\|^2 \geq 0. \tag{3.13}$$

Now let $\nu_1 = \left(-\dfrac{c_0}{2} + \sqrt{\dfrac{c_0^2}{4} + 1} \right) < 1$ be the positive root of $p_1(\theta) := -\theta^2 - c_0\theta + 1$. Assume $\nu_1 < \theta \leq 1$ (otherwise $\theta \leq \nu_1$), which implies that $p_1(\theta) < 0$. Using $u^T A u \leq \|A\| \|u\|^2$, $\left\|\widetilde{B}u\right\| \geq \mathfrak{b}_0 \|u\|^2$ in (3.13), dividing by $\|u\|^2$ and rearranging we find that θ satisfies

$$\mathfrak{t}\theta^2 - (1 - c_0\mathfrak{t})\,\theta + 1 - \mathfrak{t} - c_0 \leq 0$$

with $\mathfrak{t} = \dfrac{\mathfrak{b}_0}{\|A\|}$. We denote with $p(\theta)$ the above quadratic polynomial and with ν its positive root. Then $\theta \leq \nu$.

So far we showed that either $\theta \leq \nu_1$ or $\theta \leq \nu$, and hence $\theta \leq \max\{\nu_1, \nu\}$. We now show that $\nu \geq \nu_1$. It holds

$$p(\nu_1) = -\mathfrak{t}p_1(\nu_1) + 1 - c_0 - \nu_1 = 1 - c_0 - \nu_1 = 1 - \frac{c_0}{2} - \sqrt{\frac{c_0^2}{4} + 1} < 0.$$

Thus, we can conclude that the positive root of p is greater than ν_1 and the proof is complete.

□

3.2.4 On the choice of W

The previous spectral analysis on the preconditioned matrix allows us to make some comments on the choice of W. We are interested in identifying the choices of W which are "optimal", in the sense that the eigenvalues of the preconditioned matrix are bounded by constants that do not depend on any of the parameters of the problem (e.g., the spectral properties of the blocks), as opposed to the spectral intervals of the original (unpreconditioned) problem.

If C is positive definite, then the choice $W = C$ leads to the familiar block diagonal preconditioner, where the Schur complement is computed with respect to the (2,2) block instead of the (1,1) one. In this case, we have $c_0 = c_1 = c^* = 1$, and the expression for I^- and I^+ reduces to:

$$\begin{aligned}
\xi_1 &= -\frac{1 + \sqrt{1 + 4(1+s)^2}}{2(1+s)} \geq -\frac{1}{2}\left(1 + \sqrt{5}\right), \\
\xi_2 &= \frac{(1-t) - \sqrt{(1-t)^2 + 4t^2}}{2t} \leq -1, \\
\xi_3 &= \frac{-1 + \sqrt{1 + 4(1+s)^2}}{2(1+s)} \geq \frac{1}{2}\left(-1 + \sqrt{5}\right),
\end{aligned}$$

regardless of the nullity of B^T, yielding an optimal W. We emphasize that the above bounds are consistent with the known results for the block diagonal preconditioner (see Proposition 2.3.2) when $C \neq 0$. However, this choice is feasible only if C is positive definite. If, on the other hand, C is singular or very ill-conditioned, we consider instead

$$W = C + \frac{1}{\|A\|} BB^T,$$

and show that this choice of W is also optimal, with similar spectral interval bounds.

Theorem 3.2.4. *Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{m\times m}$ be symmetric and positive semidefinite, $B \in \mathbb{R}^{m\times n}$, $\mathcal{A}$ in (3.1) nonsingular. Given $W = C + \dfrac{1}{\|A\|}BB^T$ and $\mathcal{P}_{AD}$ in (3.2), it holds that* $\operatorname{spec}(\mathcal{P}_{AD}^{-1}\mathcal{A}) \subseteq I^- \cup I^+$ *with*

$$\begin{aligned} I^- &= \left[-\frac{1+\sqrt{1+4(1+a)^2}}{2(1+a)},\ -\frac{1}{2}\right] \subseteq \left[-\frac{1}{2}\left(1+\sqrt{5}\right),\ -\frac{1}{2}\right], \\ I^+ &= \left[\frac{-1+\sqrt{1+4(1+a)^2}}{2(1+a)},\ 1\right] \subseteq \left[\frac{1}{2}\left(-1+\sqrt{5}\right),\ 1\right] \end{aligned}$$

and $a = \dfrac{\lambda_{\min}(A)}{\|A\|}$.

Moreover, if ℓ denotes the nullity of B^T, then $\mathcal{P}_{AD}^{-1}\mathcal{A}$ has the eigenvalue -1 with multiplicity ℓ.

Proof. Direct calculation shows that any vector of the form $(0, v)$, with $v \in \ker(B^T)$ is an eigenvector of $\mathcal{P}_{AD}^{-1}\mathcal{A}$ of eigenvalue -1.

Let $c_1 = \lambda_{\max}\left(W^{-1}C\right)$. Since $W \succeq C$, it follows $c_1 \leq 1$. Moreover,

$$c_1 \geq c^* = \min_{0\neq x\in\ker(B^T)} \frac{x^T C x}{x^T\left(C + \frac{1}{\|A\|}BB^T\right)x} = 1.$$

and as a consequence $c_1 = c^* = 1$.

From the relation $W \succeq \dfrac{1}{\|A\|}BB^T$, we obtain $B^T W^{-1} B \preceq \|A\|\, I_n$ and thus $s = \lambda_{\min}(A)/b_1 \geq \lambda_{\min}(A)/\|A\|$. Then all bounds except the left extreme of I^- follow from Theorem 3.2.1.

Regarding the upper bound for the negative eigenvalues, we start again from equations (3.5) and (3.6) and suppose $v \notin \ker(B^T)$ (if $v \in \ker(B^T)$ we immediately have that $\theta = -1$). If we eliminate u from (3.5) and substitute into (3.6), we obtain

$$B\left(\left((1-\theta)A - \theta B^T W^{-1} B\right)^{-1} + \frac{\theta}{\|A\|}I_n\right)B^T v + (1+\theta)Cv = 0.$$

Let us now multiply the above equation by v^T from the left. Assuming that $\theta \geq -1$, it holds that $(1+\theta)v^T C v \geq 0$ and we get

$$v^T B\left(\left((1-\theta)A - \theta B^T W^{-1} B\right)^{-1} + \frac{\theta}{\|A\|}I_n\right)B^T v \leq 0.$$

Since $v \notin \ker(B^T)$, the minimum eigenvalue of $\left((1-\theta)A - \theta B^T W^{-1} B\right)^{-1} + \dfrac{\theta}{\|A\|} I_n$ must be nonpositive. Thus, again by $B^T W^{-1} B \preceq \|A\| I_n$ and $A \preceq \|A\| I_n$, it follows

$$0 \geq \lambda_{\min}\left(\left((1-\theta)A - \theta B^T W^{-1} B\right)^{-1} + \frac{\theta}{\|A\|} I_n\right) \geq \frac{1}{(1-2\theta)\|A\|} + \frac{\theta}{\|A\|}.$$

Rearranging the above equation we obtain $-2\theta^2 + \theta + 1 \leq 0$ from which it follows $\theta \leq -\dfrac{1}{2}$. □

As already discussed in Section 3.1, a major consideration in the choice of W is that both W and $A + B^T W^{-1} B$ should be cheap to invert, and the choices discussed above, in particular $W = C + \dfrac{1}{\|A\|} BB^T$, are likely to be infeasible in practical computations. However, we argue that this optimal choice should be regarded as an *ideal* one, and should drive the choices of more economical approximations. Our spectral estimates no longer apply when approximations to $\mathcal{P}_{AD}$ are performed. However, since the preconditioned matrix remains symmetric, we expect to obtain small perturbations of these estimates as the approximate preconditioner slightly deviates from the ideal one.

We refer to Chapter 7 for an experimental discussion on the choice of a diagonal W in the context of systems arising from Quadratic Programming problems.

3.3 Numerical illustration

In this section we illustrate the quality of the spectral estimates derived for the proposed class of preconditioners. To construct the saddle point system $\mathcal{A}$, we consider a regularized Linear Programming problem of the form

$$\min_{x,r} \; c^T x + \frac{1}{2}\rho\|x\|^2 + \frac{1}{2}\|r\|^2 \quad \text{subject to} \quad Jx + \sqrt{\delta} r = b, \; x \geq 0, \qquad (3.14)$$

where $J \in \mathbb{R}^{m \times n}$, $n \geq m$, ρ and δ are nonnegative regularization parameters, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$. We refer to Chapter 6 and 7 for more details about such problems and their numerical solution via Interior Point (IP) methods. Here we just mention that the numerical solution of (3.14) with an IP method requires the solution of a sequence of linear systems, whose coefficient matrices

can be expressed in the form

$$K_{3,\mathrm{reg}} = \begin{bmatrix} \rho I_n & J^T & -Z^{1/2} \\ J & -\delta I_m & 0 \\ -Z^{1/2} & 0 & -X \end{bmatrix}.$$

Here $X = \mathrm{diag}(x)$, where x is the current approximation for the exact solution $\hat{x}$ of (3.14), and $Z = \mathrm{diag}(z)$ where z is the dual variable associated with the inequality constraints. The properties of IP methods ensure that x and z have strictly positive components, and hence X and Z are positive definite.

Note that for $K_{3,\mathrm{reg}}$ the role of A, B and C is not fixed. In the first set of experiments, we considered the block reordering

$$A = \begin{bmatrix} \delta I_m & 0 \\ 0 & X \end{bmatrix}, \quad C = \rho I_n, \quad B = \begin{bmatrix} -J & Z^{1/2} \end{bmatrix}, \tag{3.15}$$

while in the second one,

$$A = \rho I_m, \quad C = \begin{bmatrix} \delta I_m & 0 \\ 0 & X \end{bmatrix}, \quad B = \begin{bmatrix} J \\ -Z^{1/2} \end{bmatrix}, \tag{3.16}$$

Note that in (3.15), we have $n > m$ and $\ker(B^T) = \{0\}$, whereas in (3.16) we have $n < m$ and $\ker(B^T) \neq \{0\}$.

Various choices for the matrix W were considered, namely,

$$W = C, \quad W = C + \frac{1}{\|A\|} BB^T, \quad W = I_m, \quad W = \mathcal{R}^T\mathcal{R},$$

where $\mathcal{R}$ is an $m \times m$ matrix with normally distributed random entries (MATLAB function `randn`, [88]). In the setting (3.16), we also considered the choice $W = C + C_0$, where

$$C_0 = \begin{bmatrix} \gamma_1 I_m & 0 \\ 0 & \gamma_2 I_n \end{bmatrix} \quad \text{with} \quad \gamma_1 = \min\left\{1, \frac{1}{\rho}\right\}, \ \gamma_2 = \gamma_1 \cdot \min\left\{1, \mathrm{mean}(Z)\right\},$$

and $\mathrm{mean}(Z)$ denotes the algebraic mean of the diagonal elements of Z. We motivate this selection by pointing out that we will use an analogous selection in the numerical experiments of Chapter 7.

As a model problem, we used as J the matrix in LPnetlib/lp_scagr7 [131] with full row rank, $n = 185$ and $m = 129$. The actual matrix blocks were taken from the fifth iteration of the IP method. In all the experiments, we set $\rho = 10^{-6}$, so that $\gamma_1 = 10^6$. We also set $\delta = 0$, except for the case (3.16)

W	I^-	$[\lambda_1(\mathcal{P}_{AD}^{-1}\mathcal{A}),\ \lambda_m(\mathcal{P}_{AD}^{-1}\mathcal{A})]$
C	$[-1.61803,\ -1.61595]$	$[-1.61803,\ -1.61803]$
$C + \frac{1}{\|A\|}BB^T$	$[-1.00144,\ -0.5]$	$[-1.00000,\ -0.52722]$
I_m	$[-1.0000,\ -3.47771 \cdot 10^{-4}]$	$[-1.0000,\ -3.89544 \cdot 10^{-4}]$
$\mathcal{R}^T\mathcal{R}$	$[-1.00011, -2.07795 \cdot 10^{-6}\ ]$	$[-1.00011,\ -2.34485 \cdot 10^{-6}]$
W	I^+	$[\lambda_{m+1}(\mathcal{P}_{AD}^{-1}\mathcal{A}),\ \lambda_{n+m}(\mathcal{P}_{AD}^{-1}\mathcal{A})]$
C	$[0.61803,\ 1]$	$[0.61803,\ 1.00000]$
$C + \frac{1}{\|A\|}BB^T$	$[0.99856,\ 1]$	$[0.99901,\ 1.00000]$
I_m	$[1.00000,\ 1]$	$[1.00000,\ 1.00000]$
$\mathcal{R}^T\mathcal{R}$	$[0.99989,\ 1]$	$[0.99989,\ 1.00000]$

Table 3.1: Setting (3.15). Comparison between the intervals I^- and I^+ and the true extremal eigenvalues of $\mathcal{P}_{AD}^{-1}\mathcal{A}$.

with $W = C$, because then $C = \delta I_m$ is required to be positive definite; for this latter case we set $\delta = 10^{-6}$.

Tables 3.1 and 3.2 show the comparison between the intervals I^- and I^+ provided by Theorem 3.2.1, Theorem 3.2.2 and Theorem 3.2.4 and the true extremal eigenvalues of the matrix $\mathcal{P}_{AD}^{-1}\mathcal{A}$.

In general the tested matrices W provided a very favorable setting for our estimates, yielding rather sharp bounds for all interval extremes. Note that in the case $W = C \succ 0$ of setting (3.16) (Table 3.2) the assumptions of Theorem 3.2.3 are satisfied, and hence the bound provided by that result was reported. The less satisfactory estimates occurred on the negative eigenvalues for setting (3.16) and W chosen randomly or such that $W = C + 1/\|A\|BB^T$. It is also clear that both the random choice and the identity matrix give spectral intervals that would be too large for MINRES to achieve fast convergence.

We also refer to Chapter 7, where the performance of the augmented preconditioner is tested in the solution of saddle point systems (with $C \neq 0$) stemming from Quadratic Programming problems.

3.4 Conclusions

In this chapter, we have addressed the class of augmented block diagonal preconditioners. Such preconditioners cope with the (possibly high) singularity of the diagonal blocks of a saddle point system and also preserve the

W	I^-	$[\lambda_1(\mathcal{P}_{AD}^{-1}\mathcal{A}),\ \lambda_m(\mathcal{P}_{AD}^{-1}\mathcal{A})]$
C	$[-1.61803,\ -1.00000]$	$[-1.61803,\ -1.00000]$
$C + \frac{1}{\|A\|}BB^T$	$[-1.28078,\ -0.5]$	$[-1.00000,\ -0.50000]$
I_m	$[-3.04766,\ -0.01264]$	$[-2.82702,\ -0.01310]$
$\mathcal{R}^T\mathcal{R}$	$[-99.6126, -5.32932 \cdot 10^{-6}\ ]$	$[-99.6126,\ -4.29594 \cdot 10^{-5}]$
$C + C_0$	$[-1.58842,\ -0.02797]$	$[-1.58334, -0.02316\ ]$
W	I^+	$[\lambda_{m+1}(\mathcal{P}_{AD}^{-1}\mathcal{A}),\ \lambda_{n+m}(\mathcal{P}_{AD}^{-1}\mathcal{A})]$
C	$[0.61803,\ 0.61874]$	$[0.61803,\ 0.61874]$
$C + \frac{1}{\|A\|}BB^T$	$[0.78078,\ 1]$	$[0.99915,\ 1.00000]$
I_m	$[0.32812,\ 1]$	$[0.35692,\ 1.00000]$
$\mathcal{R}^T\mathcal{R}$	$[0.01004,\ 1]$	$[0.01178,\ 1.00000]$
$C + C_0$	$[0.62956, 1\ ]$	$[0.63150, 0.99999\ ]$

Table 3.2: Setting (3.16). Comparison between the intervals I^- and I^+ and the true extremal eigenvalues of $\mathcal{P}_{AD}^{-1}\mathcal{A}$.

symmetry of the problem. General spectral bounds for the preconditioned matrix were derived. These new bounds significantly extend results in the literature, as we are unaware of spectral estimates that cover the case of both nonzero diagonal blocks. We also derived theoretical choices for the matrix W for which the eigenvalues of the preconditioned system are bounded independently of any problem parameter. Finally, we presented numerical experiments on saddle point systems stemming from regularized LP problems, which showed the quality of our spectral estimates.

Part II

Saddle point systems arising in PDE-constrained optimization

Chapter 4

Refined spectral estimates for a preconditioned system in a nonstandard inner product[1]

When a symmetric system, for example a saddle point system, is solved using a symmetric and positive definite (SPD) preconditioner, then the symmetry of the problem is preserved. As a consequence, a short term iterative system solver like MINRES can be used. On the other hand, it has been observed that *indefinite* preconditioners may lead to very efficient solution methods. Various strategies have been proposed to cope with the resulting nonsymmetry, that aim to exploit the still rich algebraic structure [11, 86, 77, 111]. Since the paper by Bramble and Pasciak in 1988 [21], attention has also been given to strategies that allow one to use an iterative solver for positive definite matrices with short-term recurrences, by using a non-standard inner product during the iterative procedure, see [35, 69, 104, 117, 125, 126, 101, 102, 105, 79]. These approaches rely on elegant theoretical properties of Krylov subspace methods that allow the simplification of the general recurrence, whenever some symmetry relations can be exploited [75, 43, 2].

In this chapter, we consider saddle point system stemming from a family of PDE-constrained optimal control problems, and we concentrate on the strategy developed for these systems by Schöberl and Zulehner [117], where a CG method in a non-standard inner product is employed. Our main goal here is to refine the spectral analysis provided in that work. Moreover, we will experimentally show how this analysis may provide new insight in the understanding the performance of the linear system solver.

[1]The results presented in this chapter are taken from [128].

The chapter is organized as follows: in Section 4.1 we introduce the optimal control problem. In Section 4.2 we briefly review the theory of Conjugate Gradient methods in a nonstandard inner product, and in particular the approach proposed in [117]. Section 4.3 is the core of this chapter, and it is devoted to the derivation of the new spectral bounds. In Section 4.4 we show some numerical experiments and finally we draw the conclusions.

4.1 An optimal control problem with PDE constraints

We consider the following optimization problem on a Hilbert space. Find $\mathrm{y} \in H^1(\Omega)$ and $\mathrm{u} \in L^2(\Omega)$ that solve the minimum problem

$$\begin{aligned} &\min_{\mathrm{y,u}} \frac{1}{2}\|\mathrm{y}-\mathrm{y}_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|\mathrm{u}\|^2_{L^2(\Omega)} \\ &\text{s.t.} \begin{cases} -\Delta \mathrm{y} + \mathrm{y} = \mathrm{u} & \text{in } \Omega \\ \dfrac{\partial \mathrm{y}}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases} \end{aligned} \tag{4.1}$$

where $\nu \in \mathbb{R}^+$ is a regularization parameter, $\mathrm{y}_d \in H^1(\Omega)$ is a given function representing the desired state, and Ω is a domain in $\mathbb{R}^d$ with $d = 2, 3$. The function y, called *state*, and the function u, called *control*, are constrained to satisfy an elliptic partial differential equation with Neumann boundary conditions. For a detailed treatment of optimal control problems, we refer to [130].

The weak formulation of the differential equation in (4.1) reads:

$$\int_\Omega \nabla \mathrm{y}(x)\cdot\nabla \mathrm{v}(x)dx + \int_\Omega \mathrm{y}(x)\mathrm{v}(x)dx = \int_\Omega \mathrm{u}(x)\mathrm{v}(x)dx \qquad \forall\, \mathrm{v} \in H^1(\Omega)$$

Note that there is no integral over $\partial\Omega$, because of the condition $\dfrac{\partial \mathrm{y}}{\partial n} = 0$.

To solve (4.1), we follow a *discretize-then-optimize* approach. In other words, we first transform the original continuous problem into a standard Quadratic Programming (QP) problem by finite element discretization, and then we numerically solve the first-order conditions of the fully discretized optimization problem. Here we will not discuss in detail the finite element method; instead, we refer to the vast literature on the subject (see e.g. the monographs [20],[22],[98]).

Let V_h and U_h be finite dimension subspaces of $H^1(\Omega)$ and $L^2(\Omega)$, respectively. In this chapter, we take $V_h = U_h$ the space of continuous and piecewise

linear functions on a simplicial subdivision of Ω. Here h is the discretization parameter, which is a measure of the refinement of the mesh and determines the dimension of the subspaces.

Let $\{\phi_1, \ldots, \phi_n\}$ be a basis for V_h, we define the mass matrix $M \in \mathbb{R}^{n\times n}$, the stiffness matrix $K \in \mathbb{R}^{n\times n}$ and the vector $y_d \in \mathbb{R}^n$ as

$$M_{ij} = \int_\Omega \phi_i(x)\phi_j(x)dx \qquad K_{ij} = \int_\Omega \nabla\phi_i(x)\cdot\nabla\phi_j(x)dx \qquad i,j = 1,\ldots,n$$

$$(y_d)_i = \int_\Omega \mathrm{y}_d(x)\phi_i(x)dx \qquad i = 1,\ldots,n.$$

Then the discretization of problem (4.1) reads [42, Section 5.2]:

$$\begin{aligned} \min_{y,u\in\mathbb{R}^n} Q(u,y) &= \frac{1}{2}(y-y_d)^T M (y-y_d) + \frac{\nu}{2}u^T M u \\ \text{s.t.} \quad Ly &= Mu \end{aligned} \tag{4.2}$$

where $L = K + M$.

We note that (4.2) is a QP problem with no inequality constraints. As such, it is possible to solve it using the method of Lagrange multipliers. We introduce the Lagrangian function

$$\mathcal{L}(u,y,p) = \frac{1}{2}(y-y_d)^T M (y-y_d) + \frac{\nu}{2}u^T M u - p^T(Ly - Mu),$$

where $p \in \mathbb{R}^n$ is the vector of Lagrange multipliers. Since the function to be minimized in (4.2) is convex, we simply look for a triple $(u,y,p) \in \mathbb{R}^{3n}$ that satisfies the first-order optimality condition $\nabla\mathcal{L}(u,y,p) = 0$, i.e.

$$\begin{aligned} L^T u + M(y-y_d) &= 0 \\ \nu M u - Mp &= 0 \\ Ly - Mu &= 0 \end{aligned}$$

or, written in matrix form,

$$\begin{bmatrix} M & 0 & L^T \\ 0 & \nu M & -M \\ L & -M & 0 \end{bmatrix} \begin{bmatrix} y \\ u \\ p \end{bmatrix} = \begin{bmatrix} My_d \\ 0 \\ 0 \end{bmatrix}. \tag{4.3}$$

Note that, in our problem, both L and M are symmetric, so the transposition in the (1,3) block is purely conventional. When a different differential equation is considered (e.g. convection-diffusion, see the next chapter), L might become nonsymmetric. The system matrix appearing in (4.3) is apparently a saddle point system with blocks:

$$A = \begin{bmatrix} M & 0 \\ 0 & \nu M \end{bmatrix} \qquad B^T = \begin{bmatrix} L^T \\ -M \end{bmatrix} \qquad C = 0.$$

4.2 Conjugate Gradient in a nonstandard inner product

Consider the linear system

$$\mathcal{A}x = b \tag{4.4}$$

with $\mathcal{A} \in \mathbb{R}^{N\times N}$ nonsingular and $b \in \mathbb{R}^N$. As it is well-known (see Chapter 2), if one wants to apply the standard Conjugate Gradient (CG) method to solve (4.4), then $\mathcal{A}$ has to be symmetric and positive definite.

The standard definitions of symmetry and positive definiteness, however, are strongly tied to the choice of the standard Euclidean inner product, and different choices are possible. Let $\mathcal{B} \in \mathbb{R}^{N\times N}$ be an SPD matrix. The inner product associated to $\mathcal{B}$ is defined as

$$\langle v, w\rangle_{\mathcal{B}} := v^T\mathcal{B}w \qquad \forall\, v, w \in \mathbb{R}^N$$

We say that $\mathcal{A}$ is symmetric and positive definite in the inner product defined by $\mathcal{B}$ if

$$\mathcal{B}-\text{symmetry:} \qquad \langle x, \mathcal{A}y\rangle_{\mathcal{B}} = \langle \mathcal{A}x, y\rangle_{\mathcal{B}} \quad \forall x, y \in \mathbb{R}^N$$

$$\mathcal{B}-\text{positive definiteness:} \qquad \langle x, \mathcal{A}x\rangle_{\mathcal{B}} > 0 \qquad \forall\, x \in \mathbb{R}^N$$

Note that by taking $\mathcal{B} = I_N$ we recover the standard definitions of symmetry and positive definiteness. It can be shown that if a matrix $\mathcal{A}$ is symmetric and positive definite with respect to $\mathcal{B}$, then it is diagonalizable and its eigenvalues are real and positive. Moreover, the matrix X of eigenvectors for $\mathcal{A}$ can be chosen to be $\mathcal{B}$-orthogonal, i.e., $X^T\mathcal{B}X = I_N$.

It is possible to define a variant of the standard CG method, implemented in the $\mathcal{B}$-inner product. Such method can be applied to solve the system (4.4), provided that $\mathcal{A}$ is symmetric and positive definite with respect to $\mathcal{B}$. This idea, in fact, can be tracked back to the original paper by Hestenes and Stiefel [70, Section 10] where the CG method was first presented.

We refer to [2] for a full taxonomy of CG methods in nonstandard inner products. One such method, relative to the inner product defined by $\mathcal{B}$, is an iterative Krylov method (that is, the k-th iterate x_k lies in $x_k + \mathcal{K}_k(\mathcal{A}, r_0)$, where r_0 is the initial residual), which minimizes at each iteration the $\mathcal{B}$-norm of the error

$$\|e_k\|_{\mathcal{B}} := \|x^* - x_k\|_{\mathcal{B}} = \min_{x\in x_0+\mathcal{K}_k} \|x^* - x\|_{\mathcal{B}}\,;$$

As in the case of standard CG, by exploiting some orthogonality properties the approximate solution at the k-th is obtained from the previous iterates.

This leads to a short-term recurrence, and only a small number of vectors need to be stored in memory.

Necessary and sufficient conditions on $\mathcal{B}$ and $\mathcal{A}$ for a CG method to be computable were discussed by Faber and Manteuffel [43]. In our context, these conditions are met if $\mathcal{B} = \mathcal{D}\mathcal{A}$, with $\mathcal{D}$ SPD and if $\mathcal{D}s_i$ can be efficiently computed at every step of the algorithm, where s_i is the ith residual.

In nonstandard CG, the following a-priori estimate on the $\mathcal{B}$-norm of the error CG, analogous to (2.10), holds:

$$\frac{\|e_k\|_{\mathcal{B}}}{\|e_0\|_{\mathcal{B}}} \leq 2\left(\frac{\sqrt{\kappa(\mathcal{A})}-1}{\sqrt{\kappa(\mathcal{A})}+1}\right)^k \tag{4.5}$$

where $\kappa_{\mathcal{B}}(\mathcal{A}) = \|\mathcal{A}\|_{\mathcal{B}} \cdot \|\mathcal{A}^{-1}\|_{\mathcal{B}} = \lambda_{\max}(\mathcal{A})/\lambda_{\min}(\mathcal{A})$ is the $\mathcal{B}$-condition number of $\mathcal{A}$. Note that $\kappa_{\mathcal{B}}(\mathcal{A})$ is well defined, as the eigenvalues of $\mathcal{A}$ are real and positive.

The estimate in (4.5) shows that the error $\mathcal{B}$-norm is bounded by a quantity that only depends on the eigenvalues $\mathcal{A}$, and the use of the $\mathcal{B}-$norm is key for this happen. Recall that for a nonsymmetric matrix $\mathcal{A}$ convergence bounds for "standard" Krylov methods cannot be given by using eigenvalues only. Indeed, such bounds relies also on different information, for example on the eigenvectors of $\mathcal{A}$ (see (2.14)).

As already discussed, when attempting to solve a linear system with iterative methods, a preconditioner if often needed to improve the spectral properties of the system matrix. This is the case also when dealing with nonstandard CG methods. Thus, if we want to apply one such method to solve (4.4), we need to find a preconditioner $\mathcal{P}$ and an SPD matrix $\mathcal{B}$, such that the preconditioned system $\mathcal{P}^{-1}\mathcal{A}$ is symmetric and positive definite in the $\mathcal{B}-$inner product. Only then a CG method can be applied to solve $\mathcal{P}^{-1}\mathcal{A}x = \mathcal{P}^{-1}b$.

To solve the saddle point system (4.3), Schöberl and Zulehner in [117] considered the following symmetric and indefinite preconditioner:

$$\mathcal{P}_{SZ} = \begin{bmatrix} \widehat{A} & B^T \\ B & B\widehat{A}^{-1}B^T - \widehat{S} \end{bmatrix},$$

where $\widehat{A}$ and $\widehat{S}$ are SPD matrices which approximate A and $B\widehat{A}^{-1}B^T$, respectively, and satisfy

$$A < \widehat{A} \quad \text{and} \quad \alpha x^T \widehat{A} x \leq x^T A x \quad \forall\, x \in \ker B, \quad \alpha < 1, \tag{4.6}$$

$$\hat{S} < B\widehat{A}^{-1}B^T \leq \beta\widehat{S}, \quad \beta > 1. \tag{4.7}$$

The actual values of α and β are problem and method dependent. Estimates of these quantities can be derived, for instance, $\widehat{A}$ is obtained by an algebraic multigrid when A is the Laplace operator [19]. Ruge and Stuben [112] provides more details on algebraic multigrids. In [117], the building blocks of the preconditioner are taken as approximations of the matrices which represent an inner product in the underlying function spaces. The parameters α and β are estimated accordingly.

We recall the following result from [117].

Theorem 4.2.1. *Let* (4.6) *and* (4.7) *hold. Then* $\mathcal{D} := \mathcal{P}_{SZ} - \mathcal{A}$ *is* SPD *and* $\mathcal{D}\mathcal{P}_{SZ}^{-1}\mathcal{A}$ *is* SPD. *Moreover,*

$$\lambda_{\max}(\mathcal{P}_{SZ}^{-1}\mathcal{A}) \leq \beta(1+\sqrt{1-1/\beta}) \tag{4.8}$$

$$\lambda_{\min}(\mathcal{P}_{SZ}^{-1}\mathcal{A}) \;\geq\; \frac{1}{2}\left(2+\alpha-1/\beta-\sqrt{(2+\alpha-1/\beta)^2-4\alpha}\right). \tag{4.9}$$

Theorem 4.2.1 allows one to use CG to solve the system $\mathcal{P}_{SZ}^{-1}\mathcal{A}x = \mathcal{P}_{SZ}^{-1}b$, which, at every step, minimizes the error in the norm defined by $\mathcal{B} = \mathcal{D}\mathcal{P}_{SZ}^{-1}\mathcal{A}$. The same result can be employed to give an estimate of the convergence rate, according to (4.5).

4.3 Refined spectral estimates

We next give a refined result, where we do not restrict ourselves to the saddle point structure.

Proposition 4.3.1. *Let* $\widehat{\mathcal{A}}, \mathcal{A} \in \mathbb{R}^{N\times N}$ *be nonsingular symmetric matrices, such that* $\mathcal{D} = \widehat{\mathcal{A}} - \mathcal{A} \succeq 0$. *We suppose that both* $\widehat{\mathcal{A}}$ *and* $\mathcal{A}$ *have* n *positive eigenvalues and* $m = N - n$ *negative ones. Then* $\widehat{\mathcal{A}}^{-1}\mathcal{A}$ *has real and positive eigenvalues. Moreover, if* $\mathcal{D}$ *is positive definite,* $\widehat{\mathcal{A}}^{-1}\mathcal{A}$ *is diagonalizable and has* n *eigenvalues strictly smaller than 1 and* m *eigenvalues strictly greater than 1. If, on the other hand,* $\mathcal{D}$ *has the eigenvalue 0 with multiplicity* ℓ*, then* $\widehat{\mathcal{A}}^{-1}\mathcal{A}$ *has* ℓ *eigenvectors associated with the eigenvalue 1.*

Proof. We first assume that $\mathcal{D}$ is positive definite. Then $\mathcal{D}$ defines an inner product on $\mathbb{R}^N$. Since $\mathcal{D}\widehat{\mathcal{A}}^{-1}\mathcal{A} = (\widehat{\mathcal{A}}-\mathcal{A})\widehat{\mathcal{A}}^{-1}\mathcal{A} = \mathcal{A}-\mathcal{A}\widehat{\mathcal{A}}^{-1}\mathcal{A}$ is symmetric, there exists a $\mathcal{D}$-orthogonal matrix X of eigenvectors for $\widehat{\mathcal{A}}^{-1}\mathcal{A}$. Therefore

$$I_N = X^T\mathcal{D}X = X^T(\widehat{\mathcal{A}}-\mathcal{A})X = X^T\widehat{\mathcal{A}}(I_N - \widehat{\mathcal{A}}^{-1}\mathcal{A})X = X^T\widehat{\mathcal{A}}X(I_N-\Lambda),$$

and hence $X^T\widehat{\mathcal{A}}X = (I_N-\Lambda)^{-1}$ and is thus diagonal. Since $\widehat{\mathcal{A}}$ has m negative and n positive eigenvalues, the Sylvester Law of Inertia ensures that

$X^T\widehat{\mathcal{A}}X$ has m negative and n positive diagonal entries. Then Λ must have m eigenvalues greater than 1, and n smaller than 1. Similarly,

$$I_N = X^T(\widehat{\mathcal{A}} - \mathcal{A})X = X^T\mathcal{A}(\mathcal{A}^{-1}\widehat{\mathcal{A}} - I_N)X = X^T\mathcal{A}X(\Lambda^{-1} - I_N),$$

from which we deduce $X^T\mathcal{A}X = (\Lambda^{-1} - I_N)^{-1} = \Lambda(I_N - \Lambda)^{-1}$, and thus $X^T\mathcal{A}X$ is also diagonal. Moreover, this equation shows that Λ must have n eigenvalues lying in the interval $]0, 1[$ and m eigenvalues lying outside $[0, 1]$. Adding these conditions to the previous ones, we conclude that Λ has n eigenvalues lying in $]0, 1[$ and m eigenvalues lying in $]1, +\infty[$.

We now consider the case where $\mathcal{D}$ is positive semidefinite. We define $\widehat{\mathcal{A}}_\epsilon = \widehat{\mathcal{A}} + \epsilon I_N$ and $\mathcal{D}_\epsilon = \widehat{\mathcal{A}}_\epsilon - \mathcal{A}$ for $\epsilon > 0$. Since $\mathcal{D}_\epsilon$ is symmetric and positive definite, from the first part of the proof we deduce that $\widehat{\mathcal{A}}_\epsilon^{-1}\mathcal{A}$ has real and positive eigenvalues. Since $\widehat{\mathcal{A}}_\epsilon^{-1}\mathcal{A} \overset{\epsilon\to 0^+}{\longrightarrow} \widehat{\mathcal{A}}^{-1}\mathcal{A}$, for the continuity of the eigenvalues we conclude that $\widehat{\mathcal{A}}^{-1}\mathcal{A}$ (which is nonsingular) has real and positive eigenvalues. Finally, from the relation $\mathcal{D} = \widehat{\mathcal{A}}(I - \widehat{\mathcal{A}}^{-1}\mathcal{A})$ one deduces that $\mathcal{D}v = 0$ if and only if $\widehat{\mathcal{A}}^{-1}\mathcal{A}v = v$. □

A saddle point matrix of the form (4.3) has n positive and m negative eigenvalues; the same holds for $\mathcal{P}_{SZ}$. Thus, if we consider Theorem 4.3.1 taking $\widehat{\mathcal{A}} = \mathcal{P}_{SZ}$ we infer

$$\operatorname{spec}(\mathcal{P}_{SZ}^{-1}\mathcal{A}) \subseteq [\lambda_1, \lambda_n] \cup [\lambda_{n+1}, \lambda_{n+m}] \tag{4.10}$$

with $0 < \lambda_1 \le \lambda_n < 1 < \lambda_{n+1} \le \lambda_{n+m}$.

The result above shows that the spectral interval used in the convergence rate estimate is actually given by the union of two intervals, which do not include the value 1. We are interested in better understanding how far these intervals lie from 1, and whether this distance may influence convergence. In the following we provide new bounds for λ_n and λ_{n+1}, and also a new lower bound for λ_1. We first need to define two new quantities:

$$a = \lambda_{\max}\left(\widehat{A}^{-1}A\right), \quad s = \lambda_{\max}\left((B\widehat{A}^{-1}B^T)^{-1}\widehat{S}\right), \tag{4.11}$$

with $\alpha \le a < 1$ and $1/\beta < s < 1$ from (4.6) and (4.7). Since $\mathcal{D} = \mathcal{P}_{SZ} - \mathcal{A}$,

$$\mathcal{A}z = \lambda\mathcal{P}_{SZ}z \quad \text{is equivalent to} \quad \mathcal{A}z = \mu\mathcal{D}z, \text{ with } \mu = \frac{\lambda}{1-\lambda}. \tag{4.12}$$

We have that $\lambda < 1$ if and only if $\mu > 0$, and $\lambda > 1$ if and only if $\mu < -1$.

Lemma 4.3.2. *Let a, s be as in (4.11). Let μ be an eigenvalue of $\mathcal{A}z = \mu\mathcal{D}z$. Then either $\mu_- \leq \mu < -1$ or $0 < \mu \leq \mu_+$, with*

$$\mu_\pm = \frac{1}{2}\left(\frac{a}{1-a} \pm \sqrt{\left(\frac{a}{1-a}\right)^2 + \frac{4}{(1-a)(1-s)}}\right).$$

Proof. Let $z = [x; y]$ be an eigenvector associated with μ. Then

$$Ax + B^T y = \mu(\widehat{A} - A)x \tag{4.13}$$

$$Bx = \mu E y, \tag{4.14}$$

with $E = B\widehat{A}^{-1}B^T - \widehat{S}$. Note that $x \neq 0$, otherwise equation (4.13) would give $B^T y = 0$, and since B^T is full column rank, this would imply $y = 0$. Equation (4.14) is used to find y, which is then substituted into equation (4.13) to obtain

$$Ax + \frac{1}{\mu}B^T E^{-1}Bx = \mu(\widehat{A} - A)x.$$

Reordering the terms and premultiplying by μx^T we obtain

$$\mu^2 x^T \widehat{A}x - (\mu^2 + \mu)x^T Ax - x^T B^T E^{-1}Bx = 0.$$

Since $\mu \in]-\infty, -1[\cup]0, +\infty[$, $\mu^2 + \mu > 0$. Moreover, $x^T Ax \leq a x^T \widehat{A}x$. Thus,

$$(1-a)\mu^2 x^T \widehat{A}x - a\mu x^T \widehat{A}x - x^T B^T E^{-1}Bx \leq 0. \tag{4.15}$$

It holds that

$$x^T B^T E^{-1}Bx \leq \frac{1}{(1-s)}x^T B^T (B\widehat{A}^{-1}B^T)^{-1}Bx \leq \frac{1}{(1-s)}x^T \widehat{A}x.$$

Using this inequality in (4.15) and dividing by $(1-a)x^T\widehat{A}x$, we find:

$$\mu^2 - \mu\frac{a}{1-a} - \frac{1}{(1-a)(1-s)} \leq 0,$$

from which both extremes μ_- and μ_+ are derived. □

We emphasize that the bounds of Lemma 4.3.2 are sharp. Indeed, let us consider the case $n = 2$, $m = 1$, with

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1-\epsilon_A \end{bmatrix}, \quad B^T = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \widehat{A} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad \widehat{S} = 1 - \epsilon_S,$$

with $\epsilon_A < \frac{1}{2}$ and $\epsilon_S < 1$. Clearly, $a = \lambda_{\max}(\widehat{A}^{-1}A) = 1 - \epsilon_A$ and $s = \lambda_{\max}((B\widehat{A}^{-1}B^T)^{-1}\widehat{S}) = 1 - \epsilon_S$. The eigenvalues of the matrix

$$\mathcal{D}^{-1/2}\mathcal{A}\mathcal{D}^{-1/2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1-\epsilon_A}{\epsilon_A} & (\epsilon_S\epsilon_A)^{-1/2} \\ 0 & (\epsilon_S\epsilon_A)^{-1/2} & 0 \end{bmatrix}$$

satisfy the characteristic equation

$$(\mu - 1)\left(\mu^2 - \mu\frac{1-\epsilon_A}{\epsilon_A} - \frac{1}{\epsilon_A\epsilon_S}\right) = 0,$$

whose solutions are $\mu = 1$ and both bounds $\mu = \mu_-$, $\mu = \mu_+$.

Proposition 4.3.3. *Let λ_n and λ_{n+1} as in* (4.10)*. Then*

$$\begin{aligned} \lambda_n &\leq 1 - \frac{2(1-a)\sqrt{1-s}}{(2-a)\sqrt{1-s} + \sqrt{a^2(1-s) + 4(1-a)}} \\ &\leq 1 - \frac{(1-a)\sqrt{1-s}}{\sqrt{1-s} + \sqrt{1-a}} \end{aligned} \tag{4.16}$$

and

$$\begin{aligned} \lambda_{n+1} &\geq 1 + \frac{(2-a)(1-s) + \sqrt{a^2(1-s)^2 + 4(1-a)(1-s)}}{2s} \\ &\geq 1 + \frac{(1-s)\left(2a\sqrt{(1-s)(1-a)} + 2 - a\right)}{2s} \geq 1 + \frac{1-s}{2s}. \end{aligned} \tag{4.17}$$

Proof. Using Lemma 4.3.2 we find that

$$\lambda_n \leq \frac{\mu_+}{1+\mu_+} = 1 - \frac{1}{1+\mu_+}, \qquad \lambda_{n+1} \geq \frac{\mu_-}{1+\mu_-} = 1 - \frac{1}{1+\mu_-}.$$

Bounds (4.16) and (4.17) follow from simple, though tedious, calculations. □

Proposition 4.3.3 shows that the distance of λ_{n+1} from 1 depends linearly on s, the eigenvalue of $(B\widehat{A}^{-1}B^T)^{-1}\widehat{S}$ closest to 1, while the distance of λ_n from 1 depends nonlinearly on s and a. While it can be shown that the upper bound (4.8) is sharp, the lower bound (4.9) will be improved. The approach we follow deviates from that originally proposed in [117].

Proposition 4.3.4. *Let* (4.6) *and* (4.7) *hold. Let* λ *be an eigenvalue of* $\mathcal{P}_{SZ}^{-1}\mathcal{A}$. *Then*

$$\lambda \geq \min\left\{\alpha, \bar{\lambda}\right\} \text{ where } \bar{\lambda} = \frac{1}{2}\left(2\beta + \alpha - 1 - \sqrt{(2\beta+\alpha-1)^2 - 4\alpha\beta}\right). \tag{4.18}$$

Proof. We consider the generalized eigenvalue problem $\mathcal{A}\begin{bmatrix} x \\ y \end{bmatrix} = \lambda \mathcal{P}_{SZ}\begin{bmatrix} x \\ y \end{bmatrix}$, i.e.,

$$Ax + B^T y = \lambda\left(\widehat{A}x + B^T y\right) \tag{4.19}$$

$$By = \lambda\left(Bx + Ey\right) \tag{4.20}$$

with $E = B\widehat{A}^{-1}B^T - \widehat{S} \succ 0$. We observe that $x \neq 0$, otherwise it would follow $\lambda = 0$. We find y from the equation (4.20) and we substitute it into equation (4.19), giving

$$\left(\lambda\widehat{A} - A\right)x = \frac{(1-\lambda)^2}{\lambda}B^T E^{-1}Bx \tag{4.21}$$

We first consider the case $x \in \ker B$. Premultiplying the last equation by x^T we have

$$0 = x^T\left(\lambda\widehat{A} - A\right)x \leq \left(\frac{\lambda}{\alpha} - 1\right)x^T Ax,$$

and then $\lambda \geq \alpha$. In the general case, we write $x = x_1 + x_2$, with $x_1 \in \ker B$ and $0 \neq x_2 \in (\ker B)^{\perp_{\widehat{A}}} := \left\{u \in \mathbb{R}^n \mid u^T\widehat{A}v = 0\ \forall v \in \ker B\right\}$, which is well-defined since $\widehat{A}$ induces a scalar product in $\mathbb{R}^N$. Note that $x_2 = \widehat{A}^{-1}B^T w$ for some $w \in \mathbb{R}^m$.

We premultiply equation (4.21) by x_1^T and by x_2^T, and obtain (note that $x_1^T\widehat{A}x_2 = 0$)

$$x_1^T\left(\lambda\widehat{A} - A\right)x_1 - x_1^T Ax_2 = 0, \tag{4.22}$$

$$x_2^T\left(\lambda\widehat{A} - A\right)x_2 - x_2^T Ax_1 = \frac{(1-\lambda)^2}{\lambda}x_2^T B^T E^{-1}Bx_2. \tag{4.23}$$

We first consider the right-hand side of equation (4.23). Using (4.7), we write $E \leq (\beta-1)/\beta B^T\widehat{A}^{-1}B$. Hence, $x_2^T BE^{-1}B^T x_2 \geq c_\beta x_2^T B^T(B\widehat{A}^{-1}B^T)^{-1}Bx_2$, where $c_\beta = \beta/(\beta-1)$. Moreover, we have

$$x_2^T B^T(B\widehat{A}^{-1}B^T)^{-1}Bx_2 = x_2^T\widehat{A}\left(\widehat{A}^{-1}B^T(B\widehat{A}^{-1}B^T)^{-1}B\right)x_2 = x_2^T\widehat{A}x_2. \tag{4.24}$$

We now turn to the left-hand side of equation (4.23). We consider

$$\begin{aligned} -x_2^T A x_1 &= x_2^T(\widehat{A}-A)x_1 \leq \left(x_1^T(\widehat{A}-A)x_1\right)^{1/2}\left(x_2^T(\widehat{A}-A)x_2\right)^{1/2} \\ &\leq \sqrt{1-\alpha}\left(x_1^T\widehat{A}x_1\right)^{1/2}\left(x_2^T\widehat{A}x_2\right)^{1/2}. \end{aligned} \tag{4.25}$$

From (4.22) and condition (4.6) we deduce that $-x_2^T A x_1 \geq (\alpha-\lambda)\, x_1^T\widehat{A}x_1$. We suppose $\lambda < \alpha$ (if not, α is the sought after extreme). The last inequality, added to (4.25), shows that

$$\left(x_1^T\widehat{A}x_1\right)^{1/2} \leq \frac{\sqrt{1-\alpha}}{\alpha-\lambda}\left(x_2^T\widehat{A}x_2\right)^{1/2}.$$

Note that this inequality also holds for $x_1 = 0$. Returning to inequality (4.25) we now conclude that $-x_2^T A x_1 \leq (1-\alpha)/(\alpha-\lambda)x_2^T\widehat{A}x_2$, and thus

$$\begin{aligned} x_2^T\left(\lambda\widehat{A}-A\right)x_2 - x_2^T A x_1 &\leq \left(\lambda + \frac{1-\alpha}{\alpha-\lambda}\right)x_2^T\widehat{A}x_2 \\ &= \frac{(1-\lambda)(\lambda-\alpha+1)}{\alpha-\lambda}x_2^T\widehat{A}x_2. \end{aligned} \tag{4.26}$$

Collecting inequalities (4.24) and (4.26) we find that λ satisfies

$$\frac{\lambda-\alpha+1}{\alpha-\lambda} \geq \frac{(1-\lambda)}{\lambda}c_\beta,$$

or, after some algebra, $\lambda^2-(2\beta+\alpha-1)\lambda+\alpha\beta \leq 0$. We denote this polynomial by $p(\lambda)$. Since $p(0) = \alpha\beta > 0$ then the smallest positive root of $p(\lambda)$, which is precisely $\bar{\lambda}$, is a lower bound for λ, when $\bar{\lambda} < \alpha$. □

We next analyze the quality of $\bar{\lambda}$ by comparing it with the lower bound in Theorem 4.2.1, which will be denoted by $\bar{\lambda}_{SZ}$ in the following. We note that $\bar{\lambda}_{SZ}$ is the smallest positive root of a second degree polynomial, i.e., $p_{SZ}(\lambda) = \lambda^2 - (2+\alpha-1/\beta)\lambda + \alpha$. We observe that $p(\lambda) - p_{SZ}(\lambda) = (\beta-1)\,[(1/\beta-2)\lambda+\alpha]$. Therefore, $p(\lambda) > p_{SZ}(\lambda)$ if and only if $\lambda < \alpha/(2-1/\beta)$. If we show that $\bar{\lambda}_{SZ} < \alpha/(2-1/\beta)$, then necessarily $\bar{\lambda}_{SZ} < \bar{\lambda}$, and thus $\bar{\lambda}$ is a sharper lower bound for the eigenvalues of $\mathcal{P}_{SZ}^{-1}\mathcal{A}$. Let $\rho = 2-1/\beta$. Our condition reads

$$\frac{1}{2}\left(\rho+\alpha-\sqrt{(\rho+\alpha)^2-4\alpha}\right) < \frac{\alpha}{\rho},$$

which is equivalent to

$$\left(\rho+\alpha-\frac{2\alpha}{\rho}\right) - \sqrt{\left(\rho+\alpha-\frac{2\alpha}{\rho}\right)^2 + 4\frac{\alpha^2}{\rho}\left(1-\frac{1}{\rho}\right)} < 0,$$

which holds since $\rho > 1$, so that $(1-1/\rho) > 0$.

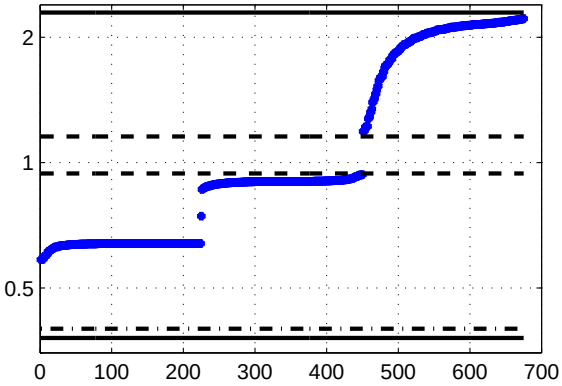

Figure 4.1: Blue dots are eigenvalues of $\mathcal{P}_{SZ}^{-1}\mathcal{A}$. The solid lines are the upper and lower bounds and the dashed lines are the interior bounds. The dash-dot lines is the improved lower bound.

4.4 Numerical experiments

In this section we report on some of our numerical experiments to illustrate our theoretical results. All computations were performed using MATLAB [88].

We considered the PDE-constrained optimal control problem described in Section 4.1, which is the same considered in [117]. The data we used to construct K, M and y_d were taken from [127, Target 1 – 2D]. We first consider the second level of discretization, that is, the dimension of $\mathcal{A}$ is 675×675.

Following Schöberl and Zulehner, we set

$$\widehat{A} = \frac{1}{\sigma}\begin{bmatrix} \widehat{Y} & 0 \\ 0 & \widehat{M} \end{bmatrix}, \qquad \widehat{S} = \frac{\sigma}{\tau\nu}\widehat{Y}. \tag{4.27}$$

where $\widehat{M}$ and $\widehat{Y}$ are SPD preconditioners for M and $Y := \sqrt{\nu}L + M$ respectively, while σ and τ are positive parameters, whose choice is crucial to obtain good values of α and β, and also to ensure that $\mathcal{D}$ is positive definite.

The above choice for $\widehat{A}$ and $\widehat{S}$ was guided by the properties of suitable inner products in the function spaces in which the state, control and Lagrange multiplier are sought; we refer again to [117] for a discussion. To actually construct $\widehat{M}$ we used three Gauss-Seidel iterations, while for $\widehat{Y}$ we used algebraic multigrid [19].

We set $\sigma = 0.9$, $\tau = 1.1 \cdot \frac{4}{3}$, $\nu = 10^{-4}$ [117, page 770, middle example, and page 768]; the value of ν does not seem to affect their analysis [117, Table 6.2]. Figure 4.1 shows the eigenvalues of $\mathcal{P}_{SZ}^{-1}\mathcal{A}$, together with the upper bound (4.8) and both interior bounds (4.16) and (4.17). We see that the estimates give a very realistic idea of the location of the true eigenvalues. For this example, we also observe that the bound (4.9) (lower solid line) is not sharp. Bound (4.18), represented by the dash-dotted line, slightly improves it.

The two parameters a and s, which are quality measures of the preconditioners $\widehat{A}$ and $\widehat{S}$ (and thus of $\mathcal{P}_{SZ}$), affect the distance of the eigenvalues of $\mathcal{P}_{SZ}^{-1}\mathcal{A}$ from 1, according to Proposition 4.3.3. More precisely, if a and s are close to 1, that is $\mathcal{P}_{SZ}$ is a good enough preconditioner for $\mathcal{A}$, the two spectral intervals $[\lambda_1, \lambda_n]$ and $[\lambda_{n+1}, \lambda_{n+m}]$ will be close to each other. Otherwise, if a and s are away from 1, the two intervals will be more distant. We also remark that, when $\widehat{A}$ and $\widehat{S}$ are constructed as proposed in [117], a is proportional to σ, and s is proportional to $1/\tau$. Numerical experiments show that parameter ν also has an influence on the distance between the two intervals; indeed, in our setting the distance is greater when $\nu \approx 1$.

The well-known bound (4.5) for the CG method considers only the extreme eigenvalues of a matrix $\mathcal{A}$. As such, it may overly pessimistic (even in the worst-case scenario), if the eigenvalues $\mathcal{A}$ are far from being uniformly distributed in the interval $[\lambda_{\min}, \lambda_{\max}]$. This is the case, for example, when $\mathcal{A}$ has a few eigenvalues far away from the others. In such cases, the real convergence is typically much faster than the one predicted by (4.5). Nevertheless, it is often possible to obtain more descriptive bounds by exploiting the peculiar spectral properties of the matrix considered, in case such information is available.

In 1977, Axelsson [3] showed a bound for CG that takes into account the presence of two separated spectral intervals of the same length. We report it here in the case of a non-standard inner product.

Theorem 4.4.1. [3, Section 4.1] *Let $\mathcal{M}$ be symmetric and positive definite with respect to the inner product defined by $\mathcal{D}$, and suppose* $\operatorname{spec}(\mathcal{M}) \subseteq [a, b] \cup [c, d]$, *with $0 < a < b < c < d$. We further suppose that $b - a = d - c$. Then a* CG *method in the $\mathcal{D}$-inner product satisfies*

$$\frac{\|e_{2k}\|_{\mathcal{B}}}{\|e_0\|_{\mathcal{D}}} \leq 2\left(\frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{k}} + 1}\right)^k \qquad \text{with} \quad \tilde{\kappa} = \frac{bc}{ad}, \quad \mathcal{B} = \mathcal{D}\mathcal{A}. \tag{4.28}$$

Note that if the spectral intervals of $\mathcal{A}$ are not of the same length, it is still possible to apply this result by simply "expanding" the smaller interval to match the larger one, although this may cause the bound to be less tight.

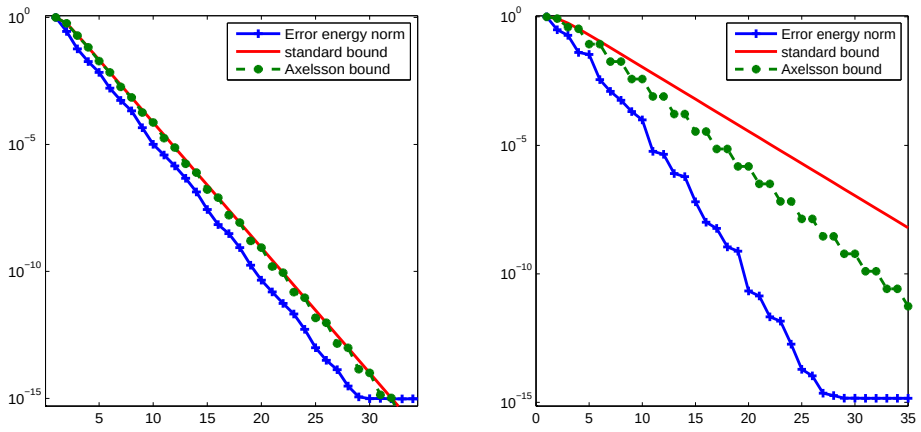

Figure 4.2: Convergence history and theoretical bound for $\sigma = 0.9$, $\tau = 1.1 \cdot \frac{4}{3}$, $\nu = 10^{-4}$ (left); and $\sigma = 0.5$, $\tau = 2 \cdot \frac{4}{3}$, $\nu = 1$ (right).

Figure 4.2 displays the convergence history of the method, in terms of the relative error $\mathcal{B}$-norm, namely $\|e_k\|_{\mathcal{B}}/\|e_0\|_{\mathcal{B}}$, along with the upper bounds (4.5) and (4.28). We used the same model previously considered in this Section but with a finer discretization, yielding $\mathcal{A}$ of size 11907. We used $x^* = \texttt{randn}(N, 1)$ as the exact solution, and $x_0 = 0$ as the initial guess.

We emphasize that, since (4.28) is defined only for even iterations, we considered the bound

$$\frac{\|e_{2k+1}\|_{\mathcal{B}}}{\|e_0\|_{\mathcal{B}}} \leq \min \left\{ 2 \left(\frac{\sqrt{\tilde{\kappa}} - 1}{\sqrt{\tilde{k}} + 1} \right)^k, \quad 2 \left(\frac{\sqrt{\kappa(\mathcal{A})} - 1}{\sqrt{\kappa(\mathcal{A})} + 1} \right)^{2k+1} \right\}$$

for odd iterations.

The left plot of Figure 4.2 considers the first choice of values for σ, τ and ν. We see that the observed behavior is in good agreement with both theoretical bounds. For this problem, the spectral intervals are $\text{spec}(\mathcal{P}_{SZ}^{-1}\mathcal{A}) \subseteq [0.5821, 0.9396] \cup [1.1553, 2.2891]$, and hence they are very close. The right plot refers instead to realistic values of the parameters that somewhat deviate from the ideal ones presented in [117]: $\sigma = 0.5$, $\tau = 2 \cdot \frac{4}{3}$, $\nu = 1$. In this case we see that the standard bound (4.5) fails to predict the rate of convergence of the method. On the other hand, bound (4.28), while not completely accurate, is a better approximation of the real rate of convergence. For this problem, the spectral intervals are $\text{spec}(\mathcal{P}_{SZ}^{-1}\mathcal{A}) \subseteq [0.4576, 0.6037] \cup [3.7977, 4.7748]$. They are quite separated, and this fact explain why (4.28) is more representative. Their length, however, is quite different, and this fact probably

motivates the distance between bound (4.28) and the real convergence curve.

4.5 Conclusions

We derived new sharper bounds for the spectrum of the preconditioned coefficient matrix of a saddle point linear system, that are used to analyze the convergence of CG in a non-standard inner product. In particular, we emphasized the presence of the union of *two* intervals containing the spectrum. Our results indicate that the standard theoretical estimates for the error energy norm at each iteration may not be representative of the actual convergence rate when the distance between these two intervals is sizable, and the bound (4.28) for two spectral intervals is a somewhat better approximation.

The recent surge of interest about CG methods in nonstandard inner products, in the context of saddle point systems, could motivate further work on this topic. In particular, one question that naturally arises is whether the use of the $\mathcal{B}-$norm might be beneficial when one is interested in reducing some different measure of the error, such as its Euclidean norm. Indeed, unlike the standard CG method [70, Theorem 6.5], the Euclidean norm of the error is not restricted to decrease at every iteration when a different inner product is considered. Moreover, even though the bound (4.5) does not depend on the conditioning of the (nonsymmetric) system matrix, the norm in which this convergence is achieved *does*, and may actually be very ill-conditioned. All these issues could be investigated in future work.

Chapter 5

New preconditioning strategies for optimal control problems with inequality constraints[1]

In the previous chapter, we considered a class of PDE-constrained optimal control problems, and we saw that the optimality conditions for the discretized problems lead to the solution of a single linear system of saddle point type. In this chapter we consider a similar class of problems, but this time box inequality constraints are imposed on the state and on the control. As a result of these additional constraints, the optimality conditions for the discretized problem now form a system of nonlinear equations. We emphasize that its dimension may be very large as soon as the desired accuracy requires a fine discretization of the partial differential equation.

If a Newton-type approach is applied to the nonlinear system, the method generates a sequence of sparse saddle point systems that have to be solved. It is well-known that a computationally effective solution of the linear algebra phase is crucial for the practical implementation of such approach and, if a Krylov method is employed, it is widely recognized that preconditioning is a critical ingredient of the inner solver.

Existing preconditioners for constrained optimal control problems have been tailored for specific elements of the family (5.1) and are generally suitable for problems where the operator characterizing the PDE is self-adjoint. Moreover, implementations based on preconditioned CG methods in a nonstandard inner product, analogous to the one described in the previous chapter, have often been preferred, in spite of possible strong limitations [69, 100, 126]. The works [100, 126] are for problems governed by symmet-

[1]The results presented in this chapter are taken from [107].

ric PDEs with box constraints on the control variables, while [69] considers problems with constraints (5.2)-(5.4) but mostly focuses on $\beta = 0$. In particular, our contribution was motivated by the problematic numerical behavior of the preconditioners proposed in [69] for $\beta \neq 0$.

In this chapter we present two new preconditioners, an indefinite preconditioner and a symmetric positive definite block diagonal preconditioner. Both strategies rely on a general factorized approximation of the Schur complement, and embed newly formed information of the nonlinear iteration, so that they dynamically change as the nonlinear iteration proceeds. The proposed preconditioners are very versatile, as they allow to handle mixed constraints as well as the corresponding limit cases, that is control and state constraints. In particular, we derive optimality and robustness theoretical properties for the spectrum of the preconditioned matrices, which hold for a relevant class of problem parameters; numerical experiments support this optimality also in terms of CPU time. A broad range of numerical experiments on three test problems is reported, for a large selection of the four problem parameters, indicating only a mild sensitivity of the preconditioner with respect to these values, especially when compared with existing approaches (for the parameters for which these latter strategies are defined). In addition, in most cases the indefinite preconditioner outperforms by at least 50% the block diagonal preconditioner, for the same Schur complement approximation.

The outline of this chapter is as follows. Sections 5.1 describes the problem and its formal numerical solution by an active-set Newton method. Section 5.2 reviews the preconditioning strategies that have been devised to solve (5.1) for some choices of the selected parameters. In Section 5.3 a new general approximation to the Schur complement is introduced and theoretically analyzed, while its impact on the new global preconditioners is investigated in Section 5.4. Section 5.5 is devoted to a wide range of numerical results. In particular, in Section 5.5.1 we discuss some algorithmic details, while in Section 5.5.2 we report on our numerical experiments on three model problems. Section 5.6 summarizes our conclusions.

5.1 The problem

The optimal control problems considered in this chapter have the form

$$\min_{\mathrm{y},u} \frac{1}{2}\|\mathrm{y}-\mathrm{y}_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|\mathrm{u}\|^2_{L^2(\Omega)}$$
$$\text{s.t.} \begin{cases} -\Delta \mathrm{y} - \beta\cdot\nabla \mathrm{y} = \mathrm{u} & \text{in } \Omega \\ \mathrm{y} = \bar{\mathrm{y}} & \text{on } \partial\Omega \\ \mathrm{a} \le \alpha_u \mathrm{u} + \alpha_y \mathrm{y} \le \mathrm{b} & \text{a.e. in } \Omega, \end{cases} \tag{5.1}$$

As in the previous chapter, $\nu \in \mathbb{R}^+$ is a regularization parameter, y_d is the desired state, and Ω is a domain in $\mathbb{R}^d$ with $d = 2, 3$. The state y and the control u are linked via an elliptic convection-diffusion equation with convection direction $\beta \in \mathbb{R}^d$. Dirichlet boundary conditions are assumed.

The main difference between (4.1) and (5.1) is the presence of box constraints of the form $\mathrm{a} \le \alpha_u \mathrm{u} + \alpha_y \mathrm{y} \le \mathrm{b}$ a.e. in Ω, where we assume $\mathrm{a}(x) < \mathrm{b}(x)$ a.e. in Ω and α_u, α_y nonnegative scalars such that $\max\{\alpha_y, \alpha_u\} > 0$. By varying the parameters α_u, α_y, we obtain optimal control problems with different inequality constraints. In particular, we will consider three specific choices which yield well studied problems. The first is $(\alpha_u, \alpha_y) = (1, 0)$, that is

$$\mathrm{a} \le \mathrm{u} \le \mathrm{b} \quad \text{a.e. in } \Omega, \tag{5.2}$$

which is refereed to as the optimal control problem with *Control Constraints* (CC). The second one is $(\alpha_u, \alpha_y) = (\epsilon, 1)$ yielding an optimal control problem with *Mixed Constraints* (MC) of the form

$$a \le \epsilon \mathrm{u} + \mathrm{y} \le \mathrm{b} \quad \text{a.e. in } \Omega. \tag{5.3}$$

The third choice is $(\alpha_u, \alpha_y) = (0, 1)$ yielding optimal control problems with *State Constraints* (SC)

$$\mathrm{a} \le \mathrm{y} \le \mathrm{b} \quad \text{a.e. in } \Omega. \tag{5.4}$$

Mixed Constraints (5.3) are commonly employed as a form of regularization of the state-constrained problem, where $\epsilon > 0$ represents the regularization parameter [90]. Indeed, pure state constrained problems are more complicated than control constrained ones, as in general the Lagrange multiplier associated with state constraints is only a measure, and therefore regularized versions with better regularity properties are needed [26, 74]; in the following we shall see the purely state-constrained problem as the limit case of the mixed-constrained one. As such, it may provide helpful information as a computational reference for the mixed-constrained problem when ϵ is very small.

5.1.1 The discrete optimization problem

We again follow a *discretize-then-optimize* approach to transform problem (5.1) into a discrete optimization problem. Issues related to the commutativity between the discretize-then-optimize and the optimize-then-discretize approach for convection diffusion control equations have been addressed in [102].

Let M represent the lumped (diagonal) mass matrix in an appropriate finite element space, and let L be the discretization of the differential operator $\mathcal{L}(\mathrm{y}) = -\Delta\mathrm{y} + \beta \cdot \nabla\mathrm{y}$; in particular, L is a nonsymmetric matrix of the form $L = K + C$, where K is the symmetric and positive definite discretization of the negative Laplacian operator and C is the "convection" matrix. In the following we shall assume that $L+L^T \succeq 0$ [2] Moreover, let n_h be the dimension of the discretized space depending on the mesh size h and let $y, u, a, b \in \mathbb{R}^{n_h}$ be the coefficients of y, u, a, b in the chosen finite element space basis. Then, the discretization of problem (5.1) is given by

$$\begin{aligned} &\min_{y,u} Q(u,y) = \frac{1}{2}(y-y_d)^T M(y-y_d) + \frac{\nu}{2}u^T M u \\ &\text{s.t.} \begin{cases} Ly = Mu - d \\ a \le \alpha_u u + \alpha_y y \le b \end{cases} \end{aligned} \tag{5.5}$$

where d represents the boundary data.

This is a convex Quadratic Programming problem with both linear equality and inequality constraints on variables. We now recall a well-known theorem which gives optimality conditions that characterize the solution of such problems. We state the theorem in a more general setting than the one described by (5.5), in order to include also the QP problems in standard form that will be discussed in Part III of this manuscript.

Theorem 5.1.1. (see e.g. [138, Theorems A.1 and A.2]) *Consider the following optimization problem*

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \tfrac{1}{2}x^T A x - c^T x \\ s.\ t. \quad & Bx = d \\ & Dx \ge f \end{aligned} \tag{5.6}$$

[2] This requirement is satisfied when, for instance, upwind finite differences over a regular grid, or upwind-type finite elements are used, with Dirichlet boundary conditions; see, e.g., [42, Chapter 6], [56].

where $A \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite, $B \in \mathbb{R}^{m\times n}$ with $m \leq n$, $D \in \mathbb{R}^{l\times n}$, $c \in \mathbb{R}^n$, $d \in \mathbb{R}^m$, $f \in \mathbb{R}^l$. Moreover, let

$$\mathcal{L}(x,p,q) = \frac{1}{2}x^T Ax - c^T x - p^T(Bx-d) - q^T(Dx-f),$$

be Lagrangian function associated with (5.6)*, where $p \in \mathbb{R}^m$ and $q \in \mathbb{R}^l$ are referred to as Lagrange multipliers or dual variables.*

Then, $\hat{x} \in \mathbb{R}^n$ is a solution for (5.6) *if and only if there exist Lagrange multiplier vectors $\hat{p} \in \mathbb{R}^m$ and $\hat{q} \in \mathbb{R}^l$ such that the following conditions hold:*

$$\begin{aligned}
\nabla_x \mathcal{L} = A\hat{x} - B^T\hat{p} - D^T\hat{q} - c &= 0, && \text{(5.7a)}\\
B\hat{x} &= d, && \text{(5.7b)}\\
D\hat{x} &\geq f, && \text{(5.7c)}\\
\hat{q} &\geq 0, && \text{(5.7d)}\\
\hat{q}^T(D\hat{x} - f) &= 0. && \text{(5.7e)}
\end{aligned}$$

Conditions (5.7a)-(5.7e) are known as the *Karush-Kuhn-Tucker conditions*, or KKT conditions for short. In particular, (5.7e) is known as *complementarity condition*; it states that either constraint i is active (i.e. the $i-th$ inequality of the system $D\hat{x} \geq f$ holds with equality) or $\hat{q}_i = 0$ (possibly both).

Note that, in the setting of problem (5.5), we have $x = (u,y) \in \mathbb{R}^{2n_h}$. Moreover, the Lagrange multiplier q can be splitted in two parts as well, namely $q = (\mu_a, \mu_b)$, where μ_a is associated with the lower constraint $\alpha_u u + \alpha_y y \geq a$, whereas μ_b is associated with the upper constraint $\alpha_u u + \alpha_y y \leq b$.

The Lagrangian function for problem (5.5) reads

$$\mathcal{L}(u,y,p,q) = Q(u,y) + (Ly - Mu + d)^T p + (\alpha_u u + \alpha_y y - b)^T \mu_b + (\alpha_u u + \alpha_y y - a)^T \mu_a.$$

The corresponding KKT conditions are

$$\begin{aligned}
&\nabla_y \mathcal{L} = M(y - y_d) + L^T p + \alpha_y(\mu_b + \mu_a) = 0\\
&\nabla_u \mathcal{L} = \nu Mu - Mp + \alpha_u(\mu_b + \mu_a) = 0\\
&Ly - Mu + d = 0\\
&\mu_b \geq 0, \quad \alpha_u u + \alpha_y y \leq b, \quad \mu_b^T(\alpha_u u + \alpha_y y - b) = 0\\
&\mu_a \leq 0, \quad a \leq \alpha_u u + \alpha_y y, \quad \mu_a^T(a - \alpha_u u - \alpha_y y) = 0.
\end{aligned} \tag{5.8}$$

5.1.2 The active-set Newton method

If we set $\mu = (\mu_b + \mu_a)$, the complementarity condition in (5.8) can be equivalently stated as the following nonlinear system

$$C(u,y,\mu) = 0$$

with C the following complementary function

$$C(u,y,\mu) = \mu - \max\{0, \mu + c(\alpha_u u + \alpha_y y - b)\} - \min\{0, \mu + c(\alpha_u u + \alpha_y y - a)\}, \tag{5.9}$$

with $c > 0$. Therefore, the KKT system (5.8) can be reformulated as the following nonlinear system

$$F(y,u,p,\mu) = \begin{bmatrix} M(y - y_d) + L^T p + \alpha_y \mu \\ \nu M u - M p + \alpha_u \mu \\ L y - M u + d \\ C(u,y,\mu) \end{bmatrix} = 0 \tag{5.10}$$

with $F : \mathbb{R}^{4n_h} \to \mathbb{R}^{4n_h}$, $y, u, p, \mu \in \mathbb{R}^{n_h}$.

In the following we recall a possible derivation of an active-set Newton type method for the solution of the KKT nonlinear system (5.10) following the description made in [71] where nonsmooth analysis was used.

Let us define the sets of active and inactive indices at the (discrete) optimal solution $(\hat{u}, \hat{y})$

$$\mathcal{A}_* = \mathcal{A}_*^b \cup \mathcal{A}_*^a \quad \text{and} \quad \mathcal{I}_* = \{1, \ldots, n_h\} \setminus \mathcal{A}_*, \tag{5.11}$$

where $\mathcal{A}_*^b, \mathcal{A}_*^a$ are the sets

$$\begin{aligned} \mathcal{A}_*^b &= \{i \mid \hat{\mu}_i + c(\alpha_u \hat{u}_i + \alpha_y \hat{y}_i - b_i) > 0\}, \\ \mathcal{A}_*^a &= \{i \mid \hat{\mu}_i + c(\alpha_u \hat{u}_i + \alpha_y \hat{y}_i - a_i) < 0\}. \end{aligned}$$

The nonlinearity and nonsmoothness of the function F in (5.10) are clearly gathered in the last block containing the complementarity function $C(u,y,\mu)$ defined in (5.9). Hintermüller et al. showed in [71] that the functions $v \to \min\{0, v\}$ and $v \to \max\{0, v\}$ from $\mathbb{R}^n \to \mathbb{R}^n$ are slantly differentiable with slanting functions given by the diagonal matrices $G_{\min}(v)$ and $G_{\max}(v)$ with diagonal elements

$$G_{\min}(v)_{ii} = \begin{cases} 1 & \text{if } v_i < 0 \\ 0 & \text{else} \end{cases}, \quad G_{\max}(v)_{ii} = \begin{cases} 1 & \text{if } v_i > 0 \\ 0 & \text{else} \end{cases}.$$

The above choice of $G_{\min}$ and $G_{\max}$ suggests to use the following element $F'(\hat{y}, \hat{u}, \hat{p}, \hat{\mu}) \in \mathbb{R}^{4n_h \times 4n_h}$ of the generalized Jacobian $\partial F(\hat{y}, \hat{u}, \hat{p}, \hat{\mu})$ (see [29])

$$F'(\hat{y}, \hat{u}, \hat{p}, \hat{\mu}) = \begin{bmatrix} M & 0 & L^T & \alpha_y I \\ 0 & \nu M & -M & \alpha_u I \\ L & -M & 0 & 0 \\ c\alpha_y \Pi_{\mathcal{A}_*} & c\alpha_u \Pi_{\mathcal{A}_*} & 0 & \Pi_{\mathcal{I}_*} \end{bmatrix}, \tag{5.12}$$

to construct a "semismooth" Newton scheme. Here $\Pi_{\mathcal{C}}$ denotes a diagonal binary matrix with nonzero entries in $\mathcal{C}$, and the sets $\mathcal{A}_*, \mathcal{I}_*$ are given in (5.11).

Given the kth iterate (y_k, u_k, p_k, μ_k), let $\mathcal{A}_k$ and $\mathcal{I}_k$ be the current active and inactive sets where

$$\mathcal{A}_k = \mathcal{A}_k^b \cup \mathcal{A}_k^a, \quad \mathcal{I}_k = \{1, \ldots, n_h\} \setminus \mathcal{A}_k \tag{5.13a}$$

$$\mathcal{A}_k^b = \{i \mid (\mu_k)_i + c(\alpha_u(u_k)_i + \alpha_y(y_k)_i - b_i) > 0\} \tag{5.13b}$$

$$\mathcal{A}_k^a = \{i \mid (\mu_k)_i + c(\alpha_u(u_k)_i + \alpha_y(y_k)_i - a_i) < 0\} \tag{5.13c}$$

and let $n_{\mathcal{A}_k} = card(\mathcal{A}_k)$ be the current number of active constraints. Using the Jacobian F' in (5.12), the semismooth Newton iteration [71] applied to system (5.10) is the following:

$$\begin{bmatrix} M & 0 & L^T & \alpha_y I \\ 0 & \nu M & -M & \alpha_u I \\ L & -M & 0 & 0 \\ c\alpha_y \Pi_{\mathcal{A}_k} & c\alpha_u \Pi_{\mathcal{A}_k} & 0 & \Pi_{\mathcal{I}_k} \end{bmatrix} \begin{bmatrix} y_{k+1} \\ u_{k+1} \\ p_{k+1} \\ \mu_{k+1} \end{bmatrix} = \begin{bmatrix} My_d \\ 0 \\ d \\ c(\Pi_{\mathcal{A}_k^b} b + \Pi_{\mathcal{A}_k^a} a) \end{bmatrix}.$$

Setting $(\mu_{k+1})_{\mathcal{I}_k} = 0$ (the multiplier associated with the inactive inequality constraints) and eliminating this variable, we obtain the sequence of Newton structured equations

$$J_k x_{k+1} = f_k, \quad k = 1, 2, \ldots \tag{5.14}$$

where $x_{k+1} = (y_{k+1}, u_{k+1}, p_{k+1}, (\mu_{k+1})_{\mathcal{A}_k}) \in \mathbb{R}^{3n_h + n_{\mathcal{A}_k}}$,

$$f_k = \begin{bmatrix} My_d \\ 0 \\ d \\ P_{\mathcal{A}_k^b} b + P_{\mathcal{A}_k^a} a \end{bmatrix}, \quad J_k = \begin{bmatrix} M & 0 & L^T & \alpha_y P_{\mathcal{A}_k}^T \\ 0 & \nu M & -M & \alpha_u P_{\mathcal{A}_k}^T \\ L & -M & 0 & 0 \\ \alpha_y P_{\mathcal{A}_k} & \alpha_u P_{\mathcal{A}_k} & 0 & 0 \end{bmatrix}, \tag{5.15}$$

where $P_{\mathcal{C}}$ is a rectangular matrix consisting of those rows of $\Pi_{\mathcal{C}}$ which belong to the indices in $\mathcal{C}$; with this notation, $\Pi_{\mathcal{C}} = P_{\mathcal{C}}^T P_{\mathcal{C}}$. We remark that the value of c has no influence on the solution of the Newton equation (5.14) but affects the updating of the active sets $\mathcal{A}_k$ in (5.13).

The above semismooth Newton scheme was proved to be equivalent to the Primal-Dual active-set method for solving constrained optimal control problems in [71] and this equivalence allowed to establish superlinear local and also global convergence results [71, 90, 80]. In fact, the active-set strategy works as a prediction technique in the sense that it is proved that if

$(u_k, y_k, p_k, \mu_k) \to (\hat{u}, \hat{y}, \hat{p}, \hat{\mu})$, then there exists an index $\bar{k}$ such that $\mathcal{A}_{\bar{k}} = \mathcal{A}_*$ and $\mathcal{I}_{\bar{k}} = \mathcal{I}_*$ [71, Remark 3.4].

Given x_k, the next iterate x_{k+1} is commonly computed by applying an iterative solver (in our case a preconditioned Krylov subspace method) to the Newton equation (5.14), and then generating a sequence of (inner) iterations $\{x_{k+1}^j\}_{j\geq 0}$. The inner iteration is started with $x_{k+1}^0 = x_k$ and stopped for $j_* > 0$ such that

$$\|J_k x_{k+1}^{j_*} - f_k\| \leq \eta_k \|J_k x_{k+1}^0 - f_k\| \tag{5.16}$$

and the next iterate x_{k+1} is set equal to $x_{k+1}^{j_*}$. The scalar $\eta_k > 0$ controls the accuracy in the solution of the unpreconditioned linear system. The choice $\eta_k = \eta_k^E$ with

$$\eta_k^E = \tau_1, \tag{5.17}$$

$k \geq 1$, with a small τ_1 (e.g. $\tau_1 = 10^{-10}$) allows us to compare various preconditioning techniques in solving the linear system (5.14), while the nonlinear iteration remains substantially unaffected by the use of each different inner strategy. This stopping criterion was used in all our numerical experiments of Sections 5.5.2 and 5.5.2.

Occasionally, for some choice of problem parameters we have experienced slow convergence of the Newton method in the solution of CC problems. This prompted us to also consider the adaptive choice $\eta_k = \eta_k^I$

$$\eta_0^I = \tau_2, \quad \eta_k^I = \min\{\eta_{k-1}^I, \tau_3 \|F(u_k, y_k, p_k, \mu_k)\|^2\}, \tag{5.18}$$

$k \geq 1$ (e.g. with $\tau_2 = 10^{-4}, \tau_3 = 10^{-2}$), which gives rise to the "inexact" solution of the Newton system [40, 76, 106]. In particular, (5.18) is intended to give the desirably fast local convergence near a solution and, at the same time, to minimize the occurrence of problem oversolving. We remark that the global convergence of the active set Newton method is no longer guaranteed if inexact steps are computed, but it is anyway expected for small values of the initial forcing term η_0^I [76]. Numerical tests with (5.18) are reported in Section 5.5.2.

The key step in the overall process is the efficient iterative solution of the linear systems (5.14), for which preconditioning is mandatory. The rest of the chapter is thus devoted to the analysis of effective preconditioning strategies.

5.2 Overview of the current approaches

In this section we review some of the preconditioning strategies that have been explored in the literature for the solution of CC, MC and SC problems.

In particular we consider the proposals [69, 126] both based on the use of the Preconditioned Conjugate Gradient method with a nonstandard inner product for the solution of the saddle point linear systems arising in the active-set Newton method for solving (5.10). This approach (from now on named BPCG), which is discussed in more detail in Chapter 4, was originally used in the context of saddle point systems by Bramble and Pasciak in [21], and then subsequently used in different settings where similar linear systems arise; see Chapter 4 of this thesis and references therein.

Herzog and Sachs in [69] consider the solution of CC, MC and SC problems by partitioning the Jacobian matrix J_k as follows

$$J_k = \left[\begin{array}{cc|cc} M & 0 & L^T & \alpha_y P_{\mathcal{A}_k}^T \\ 0 & \nu M & -M & \alpha_u P_{\mathcal{A}_k}^T \\ \hline L & -M & 0 & 0 \\ \alpha_y P_{\mathcal{A}_k} & \alpha_u P_{\mathcal{A}_k} & 0 & 0 \end{array}\right] = \begin{bmatrix} A & B_k^T \\ B_k & 0 \end{bmatrix}, \tag{5.19}$$

and therefore considering $(\alpha_y, \alpha_u) = (0, 1)$ in the CC case, $(\alpha_y, \alpha_u) = (1, \epsilon)$ in the MC case and $(\alpha_y, \alpha_u) = (1, 0)$ in the SC case. Following the approach presented in [117] and discussed in the previous chapter, Herzog and Sachs proposed the indefinite preconditioner

$$\mathcal{P}_k^{HS} = \begin{bmatrix} \widehat{A} & B_k^T \\ B_k & -\widehat{S}_k \end{bmatrix}, \tag{5.20}$$

where $\widehat{A}$ is an approximation of the (1,1) block A and $\widehat{S}_k$ is an approximation of $B_k \widehat{A}^{-1} B_k^T$. A feature of this approach is that $\widehat{A}$ and $\widehat{S}_k$ are block diagonal and their blocks can be chosen as (approximations of) the inner product matrices of the spaces where the continuous unknowns (y, u) and (p, μ) are sought. In particular (cf. (4.27)),

$$\widehat{A} = \widehat{A}(\sigma) = \frac{1}{\sigma}\begin{bmatrix} K & 0 \\ 0 & M \end{bmatrix} \quad \text{and} \quad \widehat{S}_k = \widehat{S}_k(\sigma, \tau) = \frac{\sigma}{\tau}\begin{bmatrix} K & 0 \\ 0 & P_{\mathcal{A}_k} M^{-1} P_{\mathcal{A}_k}^T \end{bmatrix},$$

where, as before, K represents the discretization of the (negative) Laplacian. Here σ and τ are positive scalars, whose choice is crucial for the method. Indeed, these parameters have to ensure that $\widehat{A} > A$ and $B_k \widehat{A}^{-1} B_k^T > \widehat{S}_k$, so that the preconditioned matrix $(\mathcal{P}_k^{HS})^{-1} J_k$ is positive definite with respect to the inner product defined by

$$\mathcal{D}_k = \mathcal{P}_k^{HS} - J_k = \begin{bmatrix} \widehat{A} - A & 0 \\ 0 & B_k \widehat{A}^{-1} B_k^T - \widehat{S}_k \end{bmatrix}.$$

Under these conditions, the CG method in this non-Euclidean inner product can be used. We mention that the choice of σ and τ can sometimes be estimated analytically, by considering the bilinear forms that underlies the problem. However, such estimates are not always available, and the numerical estimation of σ and τ might be computationally costly or cause a lack of stability in the BPCG approach.

The spectral analysis provided in [69, Corollary 2.3] for L symmetric ($\beta = 0$) shows that the eigenvalues of $(\mathcal{P}_k^{HS})^{-1} J_k$ are bounded independently of h, while they depend on ν in such a way that the condition number of the preconditioned matrix is proportional to $1/\nu$; also verified experimentally. Moreover, the authors show that the (preconditioned) condition number in MC problems scales like ϵ^{-2} for small ϵ, making the use of the proposed preconditioner prohibitive for values of ϵ smaller than 10^{-3}. Regarding the analysis for problems with $\beta = (\beta_1, 0, 0)$, $\beta_1 > 0$, a deterioration of the convergence behavior for large values of β_1 was theoretically analyzed for CC problems and confirmed in the few reported experiments. The difficulties in solving these problems are illustrated in the plots of Figure 5.1 which are in complete agreement with [69, Figure 4], and were obtained with the same codes[3], though on a different machine. In particular, we emphasize the strong dependence on β and h of the preconditioned strategy.

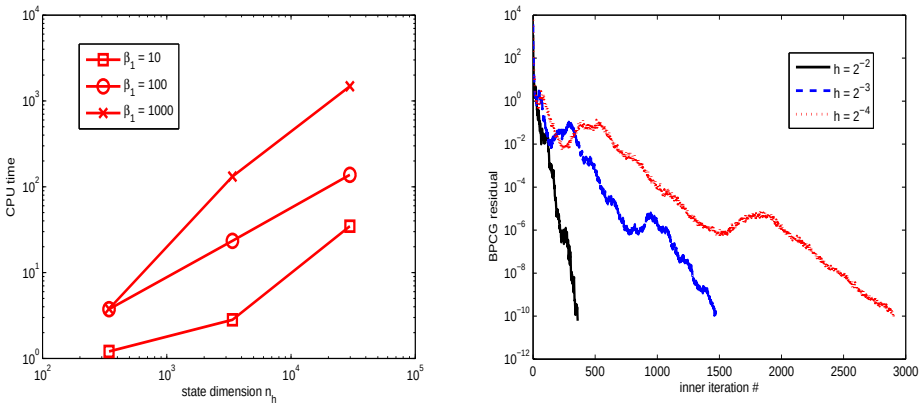

Figure 5.1: Unconstrained problem with convection (problem `CC-pb1` described in Section 5.5). Left: CPU time for a single Newton step vs. the discretized state space dimension, for $\beta = (\beta_1, 0, 0)$ with $\beta_1 = 10, 100, 1000$. Right: BPCG residual convergence history for various grid levels ($\beta_1 = 1000$).

In [126] CC problems with a self-adjoint and positive definite elliptic

[3]We thank Roland Herzog for providing us with all MATLAB codes used in [69].

operator as constraint is considered. Differently from [69], at each nonlinear iteration a saddle point system is obtained by eliminating $(\mu_{k+1})_{\mathcal{A}_k}$ from the system (5.14) and therefore solving a system of reduced dimensions with the following coefficient matrix

$$J_{k,red} = \begin{bmatrix} M & 0 & -L^T \\ 0 & \nu M_{\mathcal{A}_k,\mathcal{A}_k} & M_{\mathcal{A}_k,:} \\ -L & M_{:,\mathcal{A}_k} & 0 \end{bmatrix},$$

where $M_{\mathcal{C}_r,\mathcal{C}_c}$ is the submatrix of M obtained by taking the rows whose indices belong to the set $\mathcal{C}_r$ and the columns whose indices belong to the set $\mathcal{C}_c$. Here, ':' denotes the set of all indices $1,\dots,n_h$. However, the authors of [126] preferred to work with the full 3×3 block system,

$$J_F := \begin{bmatrix} M & 0 & -L^T \\ 0 & \nu M & M \\ -L & M & 0 \end{bmatrix}, \tag{5.21}$$

which they considered to be more practical to handle within the semismooth Newton method, than a system whose full dimension depends on the number of indices in the active sets. To solve these complete systems, the following block triangular preconditioner $\mathcal{P}^{BT}$ and inner product matrix $\mathcal{H}$ are introduced in [126]:

$$\mathcal{P}^{BT} = \begin{bmatrix} A_0 & 0 & \cdot \\ 0 & A_1 & 0 \\ -L & M & -S_0 \end{bmatrix}, \qquad \mathcal{H} = \begin{bmatrix} M-A_0 & 0 & 0 \\ 0 & \nu M - A_1 & 0 \\ 0 & 0 & S_0 \end{bmatrix}, \tag{5.22}$$

where A_0 and A_1 are appropriate approximations of M and νM, respectively, so that the matrix $\mathcal{H}$ is positive definite; moreover, $S_0 = LM^{-1}L$ approximates the following true Schur complement of J_F:

$$S_F = LM^{-1}L + \nu^{-1}M. \tag{5.23}$$

Note that the preconditioner $\mathcal{P}^{BT}$ does not depend on the nonlinear iteration k, and therefore on the current active set.

As in the previous approach, the preconditioned system $\left(\mathcal{P}^{BT}\right)^{-1} J_F$ is symmetric and positive definite with respect to the inner product associated with $\mathcal{H}$ and a CG method can be applied. In Section 5.5 we will report on the performance of the preconditioners $\mathcal{P}^{BT}$, compared with our new preconditioners.

In our analysis, we found the work [101] particularly inspiring, although a simplified setting was considered: a positive definite self-adjoint elliptic

operator in the equality constraint, and no bound-constraints. Under these hypotheses, the KKT conditions give a saddle point system with the coefficient matrix J_F in (5.21). In [101] the following factorized approximation to the Schur complement S_F in (5.23) is introduced (the scaling factor $\frac{1}{\nu}$ is omitted):

$$\widehat{S}_F = (\sqrt{\nu}L + M)M^{-1}(\sqrt{\nu}L + M), \tag{5.24}$$

which appears to possess nice independence properties with respect to the problem parameters: the eigenvalues of $\widehat{S}_F^{-1}S_F$ lie in the interval $\left[\frac{1}{2}, 1\right]$ independently of the values of h and ν [101, Theorem 4]. In the following we shall broadly generalize this idea so as to cover our more complete framework. Optimality results will also be discussed.

The Schur complement approximation (5.24) was also used in [102] in the solution of convection-diffusion (equality constrained) control problems where the authors generalized the above mentioned spectral properties of $\widehat{S}_F^{-1}S_F$ to the case where L is nonsymmetric.

5.3 A new approximation to the active-set Schur complement

In agreement with other commonly employed preconditioning strategies, the preconditioners we are going to present in Section 5.4 strongly rely on the quality of the used approximation to the Schur complement of the coefficient matrix J_k. In this section we introduce this approximation and analyze its spectral properties. In the following we shall make great use of the fact that M is a lumped mass matrix, and thus diagonal. This way, M and $\Pi_{\mathcal{A}_k}$ can commute and formulas simplify considerably. For ease of notation, we shall use the short-hand notation $\Pi_k = \Pi_{\mathcal{A}_k}$ and drop the subscript that identifies the dimension of the identity matrix.

Using the same partitioning as in (5.19), the *active-set Schur complement* associated with J_k, and its block factorization are given by

$$\begin{aligned} S_k = B_k A^{-1} B_k^T &= \frac{1}{\nu}\begin{bmatrix} \nu L M^{-1} L^T + M & (\alpha_y \nu L M^{-1} - \alpha_u I) P_{\mathcal{A}_k}^T \\ P_{\mathcal{A}_k}(\alpha_y \nu M^{-1} L^T - \alpha_u I) & (\alpha_y^2 \nu + \alpha_u^2) P_{\mathcal{A}_k} M^{-1} P_{\mathcal{A}_k}^T \end{bmatrix} \\ &= \frac{1}{\nu} R_k \begin{bmatrix} \mathbb{S}_k & 0 \\ 0 & (\alpha_y^2 \nu + \alpha_u^2) P_{\mathcal{A}_k} M^{-1} P_{\mathcal{A}_k}^T \end{bmatrix} R_k^T, \end{aligned}$$

with

$$R_k = \begin{bmatrix} I & \frac{1}{\alpha_y^2 \nu + \alpha_u^2}(\alpha_y \nu L M^{-1} - \alpha_u I)\Pi_k M P_{\mathcal{A}_k}^T \\ 0 & I \end{bmatrix}, \tag{5.25}$$

and

$$\mathbb{S}_k = \nu L M^{-1} L^T + M - \frac{1}{\alpha_y^2 \nu + \alpha_u^2} (\alpha_y \nu L M^{-1} - \alpha_u I) \Pi_k M \Pi_k (\alpha_y \nu L M^{-1} - \alpha_u I)^T.$$

We define the following *factorized approximation* of $\mathbb{S}_k$:

$$\widehat{\mathbb{S}}_k := L_1 M^{-1} L_1^T, \tag{5.26}$$

where

$$L_1 = \sqrt{\nu} L \, (I - \gamma_1 \Pi_k)^{\frac{1}{2}} + (I - \gamma_2 \Pi_k)^{\frac{1}{2}} \, M, \tag{5.27}$$

and

$$\gamma_1 = \frac{\alpha_y^2 \nu}{\alpha_y^2 \nu + \alpha_u^2}, \qquad \gamma_2 = \frac{\alpha_u^2}{\alpha_y^2 \nu + \alpha_u^2}. \tag{5.28}$$

Note that $\gamma_1 + \gamma_2 = 1$, which implies

$$(I - \gamma_1 \Pi_k)^{\frac{1}{2}} (I - \gamma_2 \Pi_k)^{\frac{1}{2}} = \sqrt{\gamma_1 \gamma_2} \; \Pi_k + (I - \Pi_k), \tag{5.29}$$

a property that will be used in the sequel. Moreover both (diagonal) matrices under square root have strictly positive diagonal elements for $\gamma_1, \gamma_2 \neq 1$, i.e., for $\alpha_u \neq 0$ and $\alpha_y \neq 0$, respectively. If $\gamma_1 = 1$ (or $\gamma_2 = 1$), then $(I - \gamma_1 \Pi_k)^{\frac{1}{2}}$ (or $(I - \gamma_2 \Pi_k)^{\frac{1}{2}}$) reduces to $(I - \Pi_k)$.

Remark 1. Our approach uses the fact that M is diagonal, both from a computational and a theoretical point of view. If the employed discretization is such that M is no longer diagonal, then we could define the preconditioner with $\mathrm{diag}(M)$ in place of M. As an alternative, we could keep M in the preconditioner, and solve systems with $P_{\mathcal{A}_k} M^{-1} P_{\mathcal{A}_k}^T$ as discussed in [69, (3.10)], and possibly also approximate the action of M^{-1} by a Chebyshev polynomial [134]. For the sake of simplicity we refrain from further exploring these possibilities.

We proceed with an analysis of the quality of the proposed Schur complement preconditioner.

Proposition 5.3.1. *Let $\mathbb{S}_k$ and $\widehat{\mathbb{S}}_k$ be as defined above. Then*

$$\widehat{\mathbb{S}}_k = \mathbb{S}_k + \sqrt{\nu}(L(I - \Pi_k) + (I - \Pi_k)L^T).$$

Proof. The result follows from

$$\begin{aligned} \mathbb{S}_k &= \nu L M^{-1} L^T + M - \frac{1}{\alpha_y^2 \nu + \alpha_u^2} (\alpha_u^2 \Pi_k M + \alpha_y^2 \nu^2 L \Pi_k M^{-1} L^T - \alpha_y \alpha_u \nu (\Pi_k L^T + L \Pi_k)) \\ &= \nu L (I - \gamma_1 \Pi_k) M^{-1} L^T + (I - \gamma_2 \Pi_k) M + \sqrt{\nu} (L \sqrt{\gamma_1 \gamma_2} \; \Pi_k + \sqrt{\gamma_1 \gamma_2} \; \Pi_k L^T), \end{aligned}$$

and

$$\begin{aligned}\widehat{\mathbb{S}}_k &= (\sqrt{\nu}L\,(I-\gamma_1\Pi_k)^{\frac{1}{2}} + (I-\gamma_2\Pi_k)^{\frac{1}{2}}\,M)M^{-1}(\sqrt{\nu}L\,(I-\gamma_1\Pi_k)^{\frac{1}{2}} + (I-\gamma_2\Pi_k)^{\frac{1}{2}}\,M)^T \\ &= \nu L(I-\gamma_1\Pi_k)M^{-1}L^T + (I-\gamma_2\Pi_k)M + \\ &\quad \sqrt{\nu}L\,(I-\gamma_1\Pi_k)^{\frac{1}{2}}\,(I-\gamma_2\Pi_k)^{\frac{1}{2}} + \sqrt{\nu}\,(I-\gamma_1\Pi_k)^{\frac{1}{2}}\,(I-\gamma_2\Pi_k)^{\frac{1}{2}}\,L^T \\ &= \nu L(I-\gamma_1\Pi_k)M^{-1}L^T + (I-\gamma_2\Pi_k)M + \\ &\quad +\sqrt{\nu}L\,(\sqrt{\gamma_1\gamma_2}\,\Pi_k + (I-\Pi_k)) + \sqrt{\nu}\,(\sqrt{\gamma_1\gamma_2}\,\Pi_k + (I-\Pi_k))\,L^T,\end{aligned}$$

where (5.29) was used. □

Note that the difference between the true and the approximate Schur complement does not depend on the γ's. The following special case of Proposition 5.3.1 occurs when all indices are active, so that $\Pi_k = I$.

Corollary 5.3.2. *If* $\mathcal{A}_k = \{1, \ldots, n_h\}$, *then* $\widehat{\mathbb{S}}_k = \mathbb{S}_k$.

The Schur complement approximation specializes when particular choices of α_u and α_v are made. In the CC case, that is for $(\alpha_u, \alpha_y) = (1, 0)$, we obtain

$$L_1 = \sqrt{\nu}L + (I - \Pi_k)M.$$

In the case of L symmetric and no bound constraints, that is for $\mathcal{A}_k = \emptyset$, we obtain $L_1 = \sqrt{\nu}L + M$, which corresponds to the factor in (5.24), as introduced in [101]. In the Mixed Constraints case, that is for $(\alpha_u, \alpha_y) = (\epsilon, 1)$, we obtain

$$L_1 = \sqrt{\nu}L\left(I - \frac{1}{1+\gamma}\Pi_k\right)^{\frac{1}{2}} + \left(I - \frac{\gamma}{1+\gamma}\Pi_k\right)^{\frac{1}{2}} M,$$

with $\gamma = \epsilon^2/\nu$. Note that both (diagonal) matrices under square root have strictly positive diagonal elements for $\gamma > 0$. Finally, in the pure State Constraints case, i.e. for $(\alpha_u, \alpha_y) = (0, 1)$, we obtain

$$L_1 = \sqrt{\nu}L\,(I - \Pi_k) + M.$$

In the next proposition we derive general estimates for the inclusion interval for the eigenvalues of the pencil $(\mathbb{S}_k, \widehat{\mathbb{S}}_k)$, whose extremes depend on the spectral properties of the nonsymmetric matrix L and on M, for general $\mathcal{A}_k$. Special cases will then be singled out.

Proposition 5.3.3. *Assume that $\widehat{\mathbb{S}}_k$ is nonsingular. Let*

$$G_k := F(I - \Pi_k) + (I - \Pi_k)F^T, \tag{5.30}$$

where $F = \sqrt{\nu}M^{-\frac{1}{2}}LM^{-\frac{1}{2}}$, with F nonsingular, and

$$H_k := F(I - \gamma_1\Pi_k)F^T + (I - \gamma_2\Pi_k) + \sqrt{\gamma_1\gamma_2}(F\Pi_k + \Pi_k F^T), \tag{5.31}$$

with γ_1, γ_2 as defined in (5.28). Then

$$\alpha_{\min} := \min_{z\neq 0} \frac{z^T G_k z}{z^T H_k z} > -1, \tag{5.32}$$

and the eigenvalues λ of the pencil $(\mathbb{S}_k, \widehat{\mathbb{S}}_k)$ satisfy $\lambda \in \left[\frac{1}{2}, \frac{1}{1+\alpha_{\min}}\right]$.

Proof. For the sake of readability, we omit the subscript k within this proof. The matrix H in (5.31) satisfies $H = M^{-\frac{1}{2}}\mathbb{S}M^{-\frac{1}{2}}$. Let

$$\widehat{H} = M^{-\frac{1}{2}}\widehat{\mathbb{S}}M^{-\frac{1}{2}}. \tag{5.33}$$

Then by Proposition 5.3.3 we have that G, H in (5.30) and (5.31) satisfy $\widehat{H} = H + G$. Therefore the problem $\mathbb{S}x = \lambda\widehat{\mathbb{S}}x$ can be written as $Hz = \lambda(H+G)z$, with $z = M^{\frac{1}{2}}x$, and for $z \neq 0$ we can write

$$\lambda = \frac{1}{1 + \frac{z^T G z}{z^T H z}}.$$

For $z \neq 0$ we have $\frac{z^T G z}{z^T H z} > -1$ if and only if $z^T(G + H)z > 0$. The latter inequality is satisfied since $G + H = M^{-\frac{1}{2}}\widehat{\mathbb{S}}M^{-\frac{1}{2}}$, and $\widehat{\mathbb{S}}$ is positive definite. This proves the upper bound for λ.

To prove the lower bound, we first consider the case $\gamma_2 \neq 1$. We define $W := (I - \gamma_2\Pi)^{-\frac{1}{2}}F(I - \gamma_1\Pi)^{\frac{1}{2}}$ and notice that

$$(I - \gamma_2\Pi)^{-\frac{1}{2}}\widehat{H}(I - \gamma_2\Pi)^{-\frac{1}{2}} = (W + I)(W + I)^T,$$

while

$$\begin{aligned}
& (I - \gamma_2\Pi)^{-\frac{1}{2}}H(I - \gamma_2\Pi)^{-\frac{1}{2}} \\
= \;& WW^T + I + \sqrt{\gamma_1\gamma_2}\left(W\Pi(I - \gamma_1\Pi)^{-\frac{1}{2}}(I - \gamma_2\Pi)^{-\frac{1}{2}} + (I - \gamma_2\Pi)^{-\frac{1}{2}}(I - \gamma_1\Pi)^{-\frac{1}{2}}\Pi W^T\right) \\
= \;& WW^T + I + (W\Pi + \Pi W^T),
\end{aligned}$$

where the relation (5.29) was used. For $x \neq 0$ we can thus write

$$\lambda = \frac{x^T\mathbb{S}x}{x^T\widehat{\mathbb{S}}x} = \frac{y^T(WW^T + I + (W\Pi + \Pi W^T))y}{y^T(W + I)(W + I)^T y},$$

where $y = (I - \gamma_2\Pi)^{\frac{1}{2}} M^{\frac{1}{2}} x$. Therefore, $\lambda \geq \frac{1}{2}$ if and only if

$$\frac{y^T(WW^T + I + (W\Pi + \Pi W^T))y}{y^T(W+I)(W+I)^T y} \geq \frac{1}{2}$$

which is equivalent to

$$\frac{1}{2} y^T(WW^T + I + W(2\Pi - I) + (2\Pi - I)W^T)y \geq 0.$$

Noticing that $I = (2\Pi - I)(2\Pi - I)$, it holds

$$WW^T + I + W(2\Pi - I) + (2\Pi - I)W^T = (W + (2\Pi - I))(W + (2\Pi - I))^T \succeq 0.$$

Therefore, the last inequality is always verified, proving the lower bound for λ.

Consider now the case $\gamma_2 = 1$ (which implies $\gamma_1 = 0$). We define $W := F^{-1}(I - \Pi)$. For $x \neq 0$ we can write

$$\lambda = \frac{x^T \mathbb{S} x}{x^T \widehat{\mathbb{S}} x} = \frac{z^T(FF^T + (I-\Pi))z}{z^T(F + (I-\Pi))(F + (I-\Pi))^T z} = \frac{y^T(WW^T + I)y}{y^T(W+I)(W+I)^T y},$$

where $z = M^{\frac{1}{2}} x$ and $y = F^T z$. As above, $\lambda \geq \frac{1}{2}$ if and only if

$$\frac{y^T(WW^T + I)y}{y^T(W+I)(W+I)^T y} \geq \frac{1}{2}$$

which holds since $2(WW^T + I) - (W+I)(W+I)^T = (W-I)(W-I)^T \succeq 0$. □

Proposition 5.3.3 reformulates the eigenvalue problem with the preconditioned Schur complement in terms of the eigenvalue problem with a different Rayleigh quotient, which seems to be easier to interpret. Numerical experiments confirm the sharpness of the lower extreme (see below); for the upper bound more insightful estimates can be given under additional hypotheses, and these are explored in the following.

Corollary 5.3.4. [102, Theorem 4.1] *Assume $L + L^T \succeq 0$ and let $\mathcal{A}_k = \emptyset$. Then the eigenvalues λ of the pencil $(\mathbb{S}_k, \widehat{\mathbb{S}}_k)$ satisfy $\lambda \in \left[\frac{1}{2}, 1\right]$.*

The result of Corollary 5.3.4 generalizes the result of [101] to nonsymmetric and positive semidefinite L, showing the optimality and robustness of the approximation with respect to the problem parameters.

To be able to analyze another interesting special case, we first need an auxiliary lemma whose proof is postponed to Appendix A.

Lemma 5.3.5. *Let $F \in \mathbb{R}^{n\times n}$ be such that $F + F^T \succeq 0$. Then*
i) $\|(F+I)^{-1}(F-I)\| \leq 1$;
ii) $\|(F+I)^{-1}(F+F^T)(F+I)^{-T}\| \leq \frac{1}{2}$.

We can now estimate the eigenvalues of $(\mathbb{S}_k, \widehat{\mathbb{S}}_k)$ for a particular choice of γ_1, γ_2.

Proposition 5.3.6. *Assume $L + L^T \succeq 0$ and let $\gamma_1 = \gamma_2 = \frac{1}{2}$. Then the eigenvalues λ of the pencil $(\mathbb{S}_k, \widehat{\mathbb{S}}_k)$ satisfy $\lambda \in \left[\frac{1}{2}, 3\right]$.*

Proof. For the sake of readability, we omit the subscript k within this proof. We only have to prove the upper bound. Let $F = \sqrt{\nu}M^{-\frac{1}{2}}LM^{-\frac{1}{2}}$, so that $F+F^T \succeq 0$. Proceeding as in the proof of Proposition 5.3.3, the eigenproblem $\mathbb{S}x = \lambda\widehat{\mathbb{S}}x$ can be transformed into

$$Hy = \lambda(H+G)y, \tag{5.34}$$

with $y = M^{\frac{1}{2}}x$, where H and G are given in (5.31) and (5.30), respectively. For $\gamma_1 = \gamma_2 = \frac{1}{2}$, we have $H + G = (F+I)\left(I - \frac{1}{2}\Pi\right)(F+I)^T$, while $H = (F-I)\left(I - \frac{1}{2}\Pi\right)(F-I)^T + F + F^T$, which can be readily verified. Therefore, problem (5.34) can be written as

$$\left((F-I)\left(I-\frac{1}{2}\Pi\right)(F-I)^T + F + F^T\right)y = \lambda(F+I)\left(I-\frac{1}{2}\Pi\right)(F+I)^T y,$$

or equivalently, with $u = (F+I)^T y$, as

$$\begin{aligned}(F+I)^{-1}\left((F-I)\left(I-\tfrac{1}{2}\Pi\right)(F-I)^T + F + F^T\right)(F+I)^{-T}u \\ = \lambda\left(I-\tfrac{1}{2}\Pi\right)u.\end{aligned} \tag{5.35}$$

We then multiply (5.35) from the left by $u^T \neq 0$,

$$\begin{aligned}u^T(F+I)^{-1}\left((F-I)\left(I-\tfrac{1}{2}\Pi\right)(F-I)^T + F + F^T\right)(F+I)^{-T}u \\ = \lambda u^T\left(I-\tfrac{1}{2}\Pi\right)u,\end{aligned} \tag{5.36}$$

and we note that $u^T\left(I - \frac{1}{2}\Pi\right)u \geq \frac{1}{2}\|u\|^2$. Moreover, using Lemma 5.3.5

$$\begin{aligned}u^T(F+I)^{-1}(F-I)(I-\frac{1}{2}\Pi)(F-I)^T(F+I)^{-T}u \\ \leq \|\left(I-\frac{1}{2}\Pi\right)\|\|(F-I)^T(F+I)^{-T}\|^2\|u\|^2 \leq \|u\|^2,\end{aligned}$$

and $u^T(F+I)^{-1}(F+F^T)(F+I)^{-T}u \leq \frac{1}{2}\|u\|^2$. Therefore, using these last bounds in (5.36) we obtain $\|u\|^2 + \frac{1}{2}\|u\|^2 \geq \lambda\frac{1}{2}\|u\|^2$, with $\|u\| \neq 0$, from which the upper estimate follows. □

In the notation of Proposition 5.3.3, for $\gamma_1 = \gamma_2$ we bounded $\alpha_{\min}$ by $-\frac{2}{3}$.

Remark 2. The case $\gamma_1 = \gamma_2$ comprises MC problems where $\alpha_u = \epsilon$ and $\alpha_y = 1$, so that the equality $\nu = \epsilon^2$ holds. Therefore, for $\nu = \epsilon^2$ Proposition 5.3.6 ensures a clustered spectrum of the preconditioned Schur complement, and this also strongly influences the spectrum of the overall preconditioned matrix - see Section 5.4 - predicting fast convergence of the iterative methods. From an application perspective, these experiments show that if $\nu \approx \epsilon^2$ in the given model, then a good performance of the solver is expected.

The good behavior for $\nu = \epsilon^2$ discussed in the remark above is confirmed by our numerical experiments (see Example 5.3), where problems with MC constraints (5.3) are tested for all combinations of values of ν and ϵ: the best performance is indeed obtained for $\nu = \epsilon^2$. It is also interesting to observe that our findings are in agreement with similar experimental observations reported in [17], where the case $\nu \approx \epsilon^2$ ensured the best performance of a multigrid solver for the MC problem.

Tables 5.1-5.3 display the spectral intervals for $\widehat{\mathbb{S}}_k^{-1}\mathbb{S}_k$ for the three considered model problems (see Table 5.4). In all tables, the minimum and maximum eigenvalues are reported for the kth iteration for which $\lambda_{\max}(\widehat{\mathbb{S}}_k^{-1}\mathbb{S}_k)$ is maximum. The CC case shows the largest, though still extremely modest, dependence of $\lambda_{\max}$ on the problem parameters, and this dependence quickly fades as β_1 increases. On the other hand, $\lambda_{\min}$ remains largely insensitive to parameter variations, with a small benign increase from the bound $\frac{1}{2}$ for $\nu = 10^{-2}$ as β_1 grows. In the mixed case and $\nu = \epsilon^2$, $\lambda_{\max}$ remains well below the upper estimate 3, for a variety of mesh parameter values.

		$\nu = 10^{-2}$				$\nu = 10^{-6}$			
β_1	h	k	$\lvert\mathcal{I}_k\rvert$	$\lambda_{\min}$	$\lambda_{\max}$	k	$\lvert\mathcal{I}_k\rvert$	$\lambda_{\min}$	$\lambda_{\max}$
0	2^{-2}	1	98	0.51	1.24	3	25	0.55	4.7
	2^{-3}	3	895	0.51	1.27	17	24	0.5	13.14
10	2^{-2}	1	73	0.64	1.18	5	57	0.54	5.32
	2^{-3}	1	891	0.61	1.24	6	44	0.50	10.72
100	2^{-2}	1*	0	1	1	4	49	0.51	4.82
	2^{-3}	1	120	0.95	1.01	6	201	0.5	6.64
1000	2^{-2}	1*	0	1	1	2	49	0.6	1.39
	2^{-3}	1*	0	1	1	2	675	0.58	1.63

* Newton terminates in 2 steps.

Table 5.1: Control-Constraints: Extreme eigenvalues of $\widehat{\mathbb{S}}_k^{-1}\mathbb{S}_k$, Newton iteration k, and dimension of the Inactive set, $|\mathcal{I}_k|$, as the mesh size h, the regularization parameter ν and the convection parameter $\beta = (\beta_1, 0, 0)$ vary.

		$\nu = 10^{-2}$					$\nu = 10^{-6}$				
β_1	h	ϵ	k	$\lvert\mathcal{I}_k\rvert$	$\lambda_{\min}$	$\lambda_{\max}$	ϵ	k	$\lvert\mathcal{I}_k\rvert$	$\lambda_{\min}$	$\lambda_{\max}$
0	2^{-2}	10^{-1}	1	196	0.53	1.10	10^{-1}	2	165	0.64	1.97
		10^{-2}	1	294	0.51	1.51	10^{-2}	1	196	0.75	1.10
		10^{-3}	2	303	0.50	1.93	10^{-3}	1	196	0.75	1.01
	2^{-3}	10^{-1}	2	2242	0.52	1.16	10^{-1}	3	1212	0.51	2.63
		10^{-2}	3	2782	0.51	1.97	10^{-2}	1	1800	0.51	1.29
		10^{-3}	3	3030	0.51	3.37	10^{-3}	1	1800	0.51	1.03
10	2^{-2}	10^{-1}	1	315	0.53	1.05	10^{-1}	2	147	0.57	3.28
		10^{-2}	0*	343	0.53	0.93	10^{-2}	1	196	0.69	1.37
		10^{-3}	0*	343	0.53	0.93	10^{-3}	1	196	0.69	1.03
	2^{-3}	10^{-1}	1	2549	0.56	1.18	10^{-1}	3	1406	0.50	5.20
		10^{-2}	1	3135	0.53	1.55	10^{-2}	2	1631	0.50	1.71
		10^{-3}	1	3303	0.53	2.45	10^{-3}	1	1800	0.50	1.08
100	2^{-2}	10^{-1}	0*	343	0.84	0.98	10^{-1}	2	196	0.51	4.40
		10^{-2}	0*	343	0.84	0.98	10^{-2}	2	196	0.51	2.49
		10^{-3}	0*	343	0.84	0.98	10^{-3}	1	147	0.51	1.24
	2^{-3}	10^{-1}	1	3299	0.84	1.01	10^{-1}	2	1575	0.50	5.9
		10^{-2}	1	3367	0.84	1.12	10^{-2}	2	1519	0.50	3.25
		10^{-3}	0*	3375	0.84	0.99	10^{-3}	2	1800	0.50	1.42
1000	2^{-2}	10^{-1}	0*	343	0.98	0.99	10^{-1}	1	294	0.51	1.22
		10^{-2}	0*	343	0.98	0.99	10^{-2}	1	294	0.51	1.22
		10^{-3}	0*	343	0.98	0.99	10^{-3}	1	294	0.52	1.24
	2^{-3}	10^{-1}	0*	3375	0.98	0.99	10^{-1}	2	2475	0.52	1.40
		10^{-2}	0*	3375	0.98	0.99	10^{-2}	2	2644	0.52	1.34
		10^{-3}	0*	3375	0.98	0.99	10^{-3}	2	2925	0.51	1.31

* Newton terminates in 1 step.

Table 5.2: Mixed-Constraints: Extreme eigenvalues of $\widehat{\mathbb{S}}_k^{-1}\mathbb{S}_k$, Newton iteration k, and dimension of the Inactive set, $|\mathcal{I}_k|$, as the mesh size h, the regularization parameters ν, ϵ and the convection parameter $\beta = (\beta_1, 0, 0)$ vary.

The dependence of $\lambda_{\max}$ on the parameters in the CC and SC cases can be analyzed by using the following result, whose proof is postponed to Appendix A.

Proposition 5.3.7. *Let λ be an eigenvalue of $\widehat{\mathbb{S}}_k^{-1}\mathbb{S}_k$. Then in the CC and SC case it holds*

$$\lambda \leq \zeta^2 + (1+\zeta)^2,$$

with
i) If $(\alpha_u, \alpha_y) = (1, 0)$ (CC case), then

$$\zeta = \|M^{\frac{1}{2}} \left(\sqrt{\nu}L + M(I-\Pi)\right)^{-1} \sqrt{\nu}LM^{-\frac{1}{2}}\|;$$

		$\nu = 10^{-2}$				$\nu = 10^{-6}$			
β_1	h	k	$\lvert\mathcal{I}_k\rvert$	$\lambda_{\min}$	$\lambda_{\max}$	k	$\lvert\mathcal{I}_k\rvert$	$\lambda_{\min}$	$\lambda_{\max}$
0	2^{-2}	2	303	0.50	2.01	1	196	0.75	1.00
	2^{-3}	3	3030	0.51	3.65	1	1800	0.51	1.02
10	2^{-2}	0*	343	0.53	0.93	1	196	0.69	1.02
	2^{-3}	1	3319	0.52	2.94	1	1800	0.50	1.06
100	2^{-2}	0*	343	0.84	0.98	1	196	0.50	1.23
	2^{-3}	0*	3375	0.84	0.99	2	2250	0.50	1.56
1000	2^{-2}	0*	343	0.98	0.99	0*	343	0.51	0.84
	2^{-3}	0*	3375	0.98	0.99	0*	3375	0.51	0.91

* Newton terminates in 1 step.

Table 5.3: State-Constraints: Extreme eigenvalues of $\widehat{\mathbb{S}}_k^{-1}\mathbb{S}_k$, Newton iteration k, and dimension of the Inactive set, $|\mathcal{I}_k|$, as the mesh size h, the regularization parameter ν and the convection parameter $\beta = (\beta_1, 0, 0)$ vary.

Moreover, if $L + L^T \succ 0$, *then for* $\nu \to 0$, ζ *is bounded by a constant independent of* ν;
ii) If $(\alpha_u, \alpha_y) = (0, 1)$ *(SC case), then*

$$\zeta = \|(I + \sqrt{\nu}M^{-\frac{1}{2}}LM^{-\frac{1}{2}}(I - \Pi_k))^{-1}\|;$$

Moreover, $\zeta \to 1$ *for* $\nu \to 0$.

The boundedness of ζ as $\nu \to 0$ in both the CC and SC cases justifies the good behavior of the eigenvalues shown in Tables 5.1 and 5.3.

5.4 New preconditioners for the active-set Newton method

In this section we propose two classes of preconditioners, which can be used throughout the nonlinear iterations, and automatically modified as the system dimensions dynamically change due to the different number of active indices. More precisely, for the problem partitioned as in (5.19) we consider the following block diagonal preconditioner $\mathcal{P}_k^{BDF}$, and indefinite preconditioner $\mathcal{P}_k^{CPF}$:

$$\mathcal{P}_k^{BDF} = \begin{bmatrix} A & 0 \\ 0 & \widehat{S}_k \end{bmatrix}, \tag{5.37}$$

and

$$\mathcal{P}_k^{CPF} = \begin{bmatrix} I & 0 \\ B_k A^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & -\widehat{S}_k \end{bmatrix} \begin{bmatrix} I & A^{-1}B_k^T \\ 0 & I \end{bmatrix}, \tag{5.38}$$

where in both cases, the matrix $\widehat{S}_k$ is factorized as

$$\widehat{S}_k = \frac{1}{\nu} R_k \begin{bmatrix} \widehat{\mathbb{S}}_k & 0 \\ 0 & (\alpha_y^2 \nu + \alpha_u^2) P_{\mathcal{A}_k} M^{-1} P_{\mathcal{A}_k}^T \end{bmatrix} R_k^T,$$

with $\widehat{\mathbb{S}}_k = L_1 M^{-1} L_1^T$, and R_k and L_1 given in (5.25) and (5.27), respectively. The following result can be readily proved from Proposition 5.3.3.

Proposition 5.4.1. *Assume that $\widehat{\mathbb{S}}_k$ is nonsingular and let $\alpha_{\min}$ be as defined in (5.32). Then the eigenvalues λ of the pencil $(J_k, \mathcal{P}_k^{BDF})$ satisfy*

$$\lambda\left(J_k,\ \mathcal{P}_k^{BDF}\right) \in \left\{1,\ \frac{1 \pm \sqrt{5}}{2}\right\} \cup I^- \cup I^+,$$

where

$$I^- = \left[\frac{1}{2}\left(1 - \sqrt{1 + \frac{4}{(1+\alpha_{\min})^2}}\right),\ \frac{1-\sqrt{2}}{2}\right],$$
$$I^+ = \left[\frac{1+\sqrt{2}}{2},\ \frac{1}{2}\left(1 + \sqrt{1 + \frac{4}{(1+\alpha_{\min})^2}}\right)\right].$$

The eigenvalues λ of the pencil $(J_k, \mathcal{P}_k^{CPF})$ satisfy

$$\lambda(J_k, \mathcal{P}_k^{CPF}) \in \{1\} \cup \left[\frac{1}{2}, \frac{1}{1+\alpha_{\min}}\right].$$

Proof. We observe that the pencil $\left(J_k,\ \mathcal{P}_k^{BDF}\right)$ has the same eigenvalues as:

$$\left(\mathcal{P}_k^{BDF}\right)^{-1/2} J_k \left(\mathcal{P}_k^{BDF}\right)^{-1/2} = \begin{bmatrix} I & A^{-1/2} B_k^T \widehat{S}_k^{-1/2} \\ \widehat{S}_k^{-1/2} B_k A^{-1/2} & 0 \end{bmatrix}.$$

Using 2.1.5, the eigenvalues of the pencil $(J_k,\ \mathcal{P}_k^{BDF})$ are either 1 or have the form $\frac{1}{2}\left(1 \pm \sqrt{1+4\sigma^2}\right)$, where σ is a singular value of $\widehat{S}_k^{-1/2} B_k A^{-1/2}$, that is, σ^2 is an eigenvalue of $\widehat{S}_k^{-1} S_k$. Considering that $\operatorname{spec}\left(\widehat{S}_k^{-1} S_k\right) = \{1\} \cup \operatorname{spec}\left(\widehat{\mathbb{S}}_k^{-1} \mathbb{S}_k\right)$, we have

$$\lambda\left(J_k,\ \mathcal{P}_k^{BDF}\right) \in \left\{1, \frac{1 \pm \sqrt{5}}{2}\right\} \cup \left\{\frac{1}{2}\left(1 \pm \sqrt{1+4\sigma^2}\right) \mid \sigma^2 \in \operatorname{spec}\left(\widehat{\mathbb{S}}_k^{-1} \mathbb{S}_k\right)\right\}.$$

The claim thus follows from Proposition 5.3.3.

As for the pencil $\left(J_k,\ \mathcal{P}_k^{IPF}\right)$, we have the factorization

$$\left(\mathcal{P}_k^{IPF}\right)^{-1} J_k = \begin{bmatrix} I & -A^{-1}B_k \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \widehat{S}_k^{-1}S_k \end{bmatrix} \begin{bmatrix} I & A^{-1}B_k \\ 0 & I \end{bmatrix}. \tag{5.39}$$

Again, the result follows from Proposition 5.3.3. □

Under the stated hypotheses, refined bounds for the eigenvalues of the indefinitely preconditioned problem can be derived using the bounds for the eigenvalues of $\widehat{\mathbb{S}}_k^{-1}\mathbb{S}_k$ obtained in Corollary 5.3.4, Proposition 5.3.6 and Proposition 5.3.7.

When the indefinite preconditioner is considered, the preconditioned matrix $\left(\mathcal{P}_k^{IPF}\right)^{-1} J_k$ has real spectrum (see Theorem 2.3.4), however it is no longer symmetric so that in general, a nonsymmetric solver needs to be applied. In our numerical experiments we used GMRES, for which it is known that the eigenvalues alone may not be sufficient to predict convergence, but that also eigenvectors play a role. In addition, indefinite preconditioners are often plagued by the presence of Jordan blocks, whose sensitivity may influence the use of inexact strategies; see, e.g., [118] for a detailed discussion. Fortunately, since the (1,1) block of J_k is reproduced exactly in the preconditioner, in our setting the spectral structure is considerably simplified, and in particular, Jordan blocks do not occur. The following proposition determines the complete eigenvector decomposition of the preconditioned matrix.

Proposition 5.4.2. *Let $\widehat{S}_k^{-1}S_kX = X\Lambda$ be the eigendecomposition of $\widehat{S}_k^{-1}S_k$, with $X = [X_1, X_2]$ and $\Lambda = \ blkdiag(I, \Lambda_2)$ partitioned so that X_1 contains the eigenvectors corresponding to the unit eigenvalue. Then the preconditioned matrix $\left(\mathcal{P}_k^{IPF}\right)^{-1} J_k$ admits the following eigenvalue decomposition*

$$\left(\mathcal{P}_k^{IPF}\right)^{-1} J_k = Q \begin{bmatrix} I & & \\ & I & \\ & & \Lambda_2 \end{bmatrix} Q^{-1},$$

with

$$Q = \left[\begin{array}{c|c|c} I & 0 & -A^{-1}B_kX_2 \\ 0 & X_1 & X_2 \end{array}\right], \qquad Q^{-1} = \left[\begin{array}{c|c} I & A^{-1}B_kX_2X_2^T\widehat{S}_k \\ 0 & X_1^T\widehat{S}_k \\ 0 & X_2^T\widehat{S}_k \end{array}\right].$$

Proof. Writing

$$\left(\mathcal{P}_k^{IPF}\right)^{-1} J_k = \begin{bmatrix} I & A^{-1}B_k(I - \widehat{S}_k^{-1}S_k) \\ 0 & \widehat{S}_k^{-1}S_k \end{bmatrix},$$

the decomposition can be explicitly verified upon substitution. The nonsingularity of Q follows from that of $X = [X_1, X_2]$. The inverse of Q can be derived by observing that X can be chosen so that $X^T \widehat{S}_k X = I$. □

The explicit form of Proposition 5.4.2 allows one to use standard results to bound the GMRES residual norm, by providing bounds for the norm of Q and its inverse Q^{-1}, and exploiting the fact that the spectrum of the preconditioned matrix is real (recall (2.14)).

5.5 Numerical experiments

In this section we provide a detailed performance analysis of the proposed preconditioners $\mathcal{P}_k^{CPF}$ in (5.38) and $\mathcal{P}_k^{BDF}$ in (5.37) for the active-set Newton method and use problems with constraints in (5.2)-(5.4) as prototypical problems. In particular, the analysis of the pure State Constraints case (5.4) will be analyzed as the limit case of the MC constraints (5.3) for $\epsilon \to 0$.

label	Ω	a	b	y_d
`CC-Pb1`	$(-1,1)^3$	0	2.5	1 for $\lvert x_1\rvert \leq \frac{1}{2}$, -2 otherwise
`CC-Pb2`	$(0,1)^3$	$\frac{1}{10}\exp(-\lVert x\rVert^2)$	$\frac{1}{2}$	$\exp(-64\lVert x - \frac{1}{2}\rVert^2)$
`MC-Pb1`	$(-1,1)^3$	$-\infty$	0	1 for $\lvert x_1\rvert \leq \frac{1}{2}$, -2 otherwise

Table 5.4: Problem data for the numerical experiments. Here $x = (x_1, x_2, x_3) \in \Omega$.

In all our examples, we use the three-dimensional data for the discretized problem generated by the codes in [69]. The matrices stem from the discretization by upwind finite differences on a uniform three-dimensional grid (so that $L + L^T \succ 0$). Zero Dirichlet boundary conditions, that is $\bar{y} = 0$ in (5.1), were used throughout. In Table 5.4 information on the data used in our numerical experiments can be found, for two test cases with control constraints, and one test case for mixed and state constraints; here $x = (x_1, x_2, x_3)$ is an element of Ω. The mesh parameter in each direction was taken as $h \in \{2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}\}$ which corresponds to a dimension for the state or control vectors $n_h \in \{343, 3375, 29791, 250047\}$. The total linear system dimension is thus between $3n_h$ and $4n_h$, depending on the number of indices in the active set at each Newton iteration.

5.5.1 Algorithmic considerations

Throughout this section we consider the implementation of the active-set Newton method with the following solvers and preconditioning strategies:

AS-GMRES-IPF	Active-set Newton method with linear solver GMRES preconditioned with $\mathcal{P}_k^{CPF}$;
AS-MINRES-BDF	Active-set Newton method with linear solver MINRES preconditioned with $\mathcal{P}_k^{BDF}$;
AS-BPCG-BT	Variant of active-set Newton method as proposed in [126], with BPCG preconditioned with $\mathcal{P}^{BT}$ defined in (5.22).

The application of the Schur complement approximation $\widehat{\mathbb{S}}_k$ requires solving with L_1 and its transpose in (5.27). These solves were replaced by the use of an algebraic multigrid operator (HSL-MI20, [19]), which needs to be recomputed at each Newton iteration. HSL-MI20 is used with all default parameters except for the value `control.st_parameter`=10^{-4}. Moreover, we set the number of pre/post smoothing steps equal to 5 for all the experiments with the MC problems, while with CC problems only for the finest mesh $h = 2^{-5}$. Although in most cases satisfactory results were obtained with this software, we did experience some anomalous behavior when strong convection was used. In these cases, ad-hoc algebraic multigrid strategies should be adopted. We also recall that both Π_k and M are diagonal, therefore L_1 is obtained from the convection-diffusion matrix by scaling, and then modifying its diagonal.

According to [134], we used $A_0 = 0.9M$ and $A_1 = 0.9(\nu\, M)$ for the parameterized preconditioners in (5.22) within the BPCG iteration. Systems with L to apply S_0 in (5.22) are approximately solved with the aforementioned HSL-MI20 code.

We set a limit of 80 GMRES iterations and 1000 MINRES and BPCG iterations. If a solver reaches the maximum number of iterations, the last computed iterate is used as the next Newton iterate.

As for the nonlinear iteration, in all tests we set the parameter c in the definition of the active-set strategy (5.13) equal to one, and we use a null starting guess x_0 in the Newton iteration, which by (5.13) implies that $\mathcal{A}_0 = \emptyset$ in all settings. As already mentioned, we used the stopping criterion (5.16) with $\eta_k = \eta_k^E$ in (5.17) where we further included the safeguard $\tau_s = 10^{-10}$ as follows

$$\|J_k x_{k+1}^{j_*} - f_k\| = \max\{\tau_s, \eta_k^E \|J_k x_{k+1}^0 - f_k\|\}, \tag{5.40}$$

$k \geq 1$, with the tight tolerance $\tau_1 = 10^{-10}$ in (5.17) [40]. While the residual 2-norm in (5.40) can be cheaply evaluated for GMRES when using right preconditioning, in the case of MINRES we explicitly computed the (unpreconditioned) residual vector at each iteration, and then computed its norm; for MINRES we thus slightly modified the code available in [41].

In the numerical tests in Section 5.5.2 we also experimented with the adaptive choice $\eta_k = \eta_k^I$ in (5.18), with $\tau_2 = 10^{-4}, \tau_3 = 10^{-2}$, together with the above safeguard threshold τ_s. We experimentally verified that this choice of tolerances preserved the global convergence of the active set Newton procedure.

Concerning the outer iteration, we followed [69] and we declare convergence when the nonlinear residual is sufficiently small, i.e.

$$\|F(u_k, y_k, p_k, \mu_k)\| \leq \tau_f, \qquad \text{with} \quad \tau_f = 10^{-8}.$$

We verified that this criterion was equivalent to terminating the iteration as soon as the active sets stay unchanged in two consecutive steps as proved in [16, 90]. On the contrary, any run performing more than 200 nonlinear iterations is considered a failure and will be denoted with the symbol '-' in the forthcoming tables.

All numerical experiments were performed on a 4xAMD Opteron 850, 2.4GHz, 16GB of RAM using MATLAB R2012a [88].

5.5.2 Numerical results

The presentation of the numerical results is organized as follows. Section 5.5.2 is devoted to the comparison of AS-GMRES-IPF and AS-MINRES-BDF with AS-BPCG-BT (see Section 5.2 and (5.22)) on symmetric CC problems. Section 5.5.2 collects the numerical results of the new proposals AS-GMRES-IPF and AS-MINRES-BDF on symmetric and nonsymmetric problems for a variety of problem parameters. Finally, in Section 5.5.2 an *inexact* active set approach is considered in the solution of nonsymmetric CC problems.

In some cases, a comparative computational analysis is carried out by using performance profiles for a given set of test problems and a given selection of algorithms [31]. For a problem P in our testing set and an algorithm A, we let $\texttt{ti}_{P,A}$ denote the total CPU time employed to solve problem P using algorithm A and $\texttt{ti}_P$ be the total CPU time employed by the *fastest* algorithm to solve problem P. As stated in [31], the CPU time performance profile is defined for algorithm A as

$$\pi_A(\tau) = \frac{\text{number of problems s.t. } \texttt{ti}_{P,A} \leq \tau\, \texttt{ti}_P}{\text{number of problems}}, \quad \tau \geq 1,$$

that is the probability[4] for solver A that a performance ratio $\texttt{ti}_{P,A}/\texttt{ti}_P$ is within a factor τ of the best possible ratio. The function $\pi_A(\tau)$ is the (cumulative) distribution function for the performance ratio.

[4]Or, more precisely, the frequency.

In the upcoming tables of results the following data will be reported: the average number of linear inner iterations (LI), the number of nonlinear outer iterations (NLI in brackets), the average elapsed CPU time of the inner solver (CPU), and the total elapsed CPU time (TCPU).

Finally, to be able to evaluate the effectiveness of the preconditioned linear solvers, we take as reference the computational cost of solving the whole system with a sparse direct solver ("backslash" in MATLAB). For the finest mesh, corresponding to $h = 2^{-5}$, the *compiled* direct solver takes 611 seconds to solve a single linear system with $\mathcal{A}_k = \emptyset$ for some k ($\nu = 10^{-2}, \beta = 0$). We note that this corresponds to the cost of the first iteration when the active set Newton algorithm is applied to every problem of the family (5.1). For comparison purposes, multiplying by the number of nonlinear iterations, the total cost of the process when the inner system is solved with a sparse direct method can be derived.

		AS-GMRES-IPF			AS-MINRES-BDF			AS-BPCG-BT		
ν	h	LI (NLI)	CPU	TCPU	LI (NLI)	CPU	TCPU	LI (NLI)	CPU	TCPU
10^{-2}	2^{-2}	9.6(3)	0.1	0.2	20(3)	0.1	0.2	11.3(3)	0.1	0.2
	2^{-3}	9.5(4)	0.8	3.2	19.5(4)	1.1	4.2	10.7(4)	0.7	2.7
	2^{-4}	8.5(4)	1.5	8.5	18.7(4)	2.5	9.9	10.0(4)	6.7	26.8
	2^{-5}	8.0(4)	12.1	48.2	19.2(4)	36.1	144.4	9.5(4)	17.5	69.9
10^{-4}	2^{-2}	6.5(7)	0.1	0.11	13.8(7)	0.1	0.2	17.5(7)	0.2	1.3
	2^{-3}	11.2(11)	0.7	8.1	23.8(11)	1.3	14.4	21.1(11)	1.3	14.7
	2^{-4}	10.7(17)	1.8	30.1	23.5(17)	3.1	51.6	18.0(17)	4.7	80.1
	2^{-5}	10.3(15)	16.1	241.4	24.3(15)	31.6	474.3	18.2(15)	30.9	463.3
10^{-6}	2^{-2}	10.3(9)	0.1	0.2	22.7(9)	0.1	0.35	41.1(9)	0.1	0.7
	2^{-3}	16.0(19)	1.1	21.5	34.6(19)	1.9	35.8	99.0(19)	6.1	115.6
	2^{-4}	17.6(54)	2.9	160.7	44.9(54)	5.7	289.8	93.5(54)	13.6	735.6
	2^{-5}	22.0(68)	38.4	2608.4	56.3(89)	63.2	5627.2	102.1(68)	136.7	9293.6
10^{-8}	2^{-2}	11.1(9)	0.1	0.2	25.4(9)	0.1	0.4	58.6(9)	0.1	1.0
	2^{-3}	18.3(27)	0.7	20.2	40.1(27)	2.1	57.6	133.2(27)	8.3	224.1
	2^{-4}	30.3(74)	7.3	540.5	72.1(66)	9.2	513.4	385.0(66)	60.1	3962.8
	2^{-5}	-	-	-	-	-	-	-	-	-

Table 5.5: Comparison among AS-GMRES-IPF, AS-MINRES-BDF and AS-BPCG-BT. Test problem `CC-Pb1` for a variety of h and ν (L symmetric, i.e., $\beta = 0$).

Comparison with the BPCG approach

In order to make comparisons with AS-BPCG-BT in the setting used in [126], we restrict our testing set to symmetric CC problems `CC-Pb1` and `CC-Pb2` with $\beta = 0$. Numerical results are reported in Tables 5.5 and 5.6. The number of nonlinear iterations remains quite low for most choices of the parameters, except for the finest grid and the limit case $\nu = 10^{-8}$. All methods seem to show some ν-dependence both in the (inner) linear solver,

		AS-GMRES-IPF			AS-MINRES-BDF			AS-BPCG-BT		
ν	h	LI (NLI)	CPU	TCPU	LI (NLI)	CPU	TCPU	LI (NLI)	CPU	TCPU
10^{-2}	2^{-2}	8.75(4)	0.1	0.2	18(4)	0.1	0.2	10.0(4)	0.03	0.10
	2^{-3}	8.0(5)	0.2	0.8	16.8(5)	0.9	4.7	9.0(5)	0.2	0.99
	2^{-4}	7.4(5)	1.3	6.5	16.2(5)	2.2	11.1	9.2(5)	1.8	8.77
	2^{-5}	7.4(5)	11.3	56.4	16.6(5)	19.4	96.7	8.4(5)	15.2	76.1
10^{-4}	2^{-2}	11.1(9)	0.1	0.2	23.2(9)	0.1	0.4	28.4(7)	0.1	0.6
	2^{-3}	12.9(13)	0.3	3.8	27.7(13)	1.5	19.8	24.4(13)	0.5	6.4
	2^{-4}	13.0(14)	2.1	29.5	28.7(14)	3.7	52.1	20.0(14)	3.6	50.1
	2^{-5}	11.7(13)	18.5	240.8	27.5(13)	32.0	416.1	20.5(13)	28.2	367.5
10^{-6}	2^{-2}	12.2(12)	0.1	0.3	26.6(12)	0.1	0.5	52.2(12)	0.1	1.3
	2^{-3}	16.8(22)	0.4	8.8	36.8(22)	2.0	44.6	115.5(22)	2.1	46.7
	2^{-4}	18.2(35)	3.1	106.9	43.5(36)	5.7	204.8	118.5(35)	19.3	675.2
	2^{-5}	20.0(41)	34.6	1416.9	52.5(53)	59.5	3151.9	84.3(40)	109.2	4367.1
10^{-8}	2^{-2}	10.4(11)	0.1	0.2	23.2(11)	0.1	0.4	64.3(11)	0.1	1.6
	2^{-3}	15.7(19)	0.4	8.2	35.5(19)	1.9	37.1	195.1(19)	3.8	71.2
	2^{-4}	27.6(55)	5.3	289.1	69.0(63)	9.1	572.0	360.5(54)	55.9	3021.7
	2^{-5}	41.0(156)	90.7	14156.0	-	-	-	343.3(131)	438.2	57406.9

Table 5.6: Comparison among AS-GMRES-IPF, AS-MINRES-BDF and AS-BPCG-BT. Test with `CC-Pb2` for a variety of h and ν (L symmetric, i.e., $\beta = 0$).

and in the (outer) nonlinear iteration; however, while in both problems for AS-GMRES-IPF and AS-MINRES-BDF such dependence is rather mild, this is significantly more evident for AS-BPCG-BT. Large values of LI for AS-BPCG-BT in the tables correspond to runs where the maximum number of inner iterations is reached. This shortcoming makes AS-BPCG-BT not competitive in almost all parameter combinations, with timings that differ significantly from the other methods, up to at most one order of magnitude. Finally, we recall that at each iteration AS-GMRES-IPF and AS-MINRES-BDF solve linear systems of dimension $3n_h + n_{\mathcal{A}_k}$, whereas AS-BPCG-BT solves systems of fixed dimension $3n_h$. The numbers in Tables 5.5 and 5.6 show that an appropriate explicit treatment of the active-set information within the preconditioner is capable of making up for the larger problem size, yielding an overall significant gain in CPU time.

Dependence on the problem parameters

We tested the new preconditioners on CC, MC and SC problems by analyzing their dependence on the parameters of the discretized problem, i.e. the regularization parameter ν, the convection coefficient β, the mesh size h and, for the MC case, the regularization parameter ϵ.

Example 5.1. For the CC problems, we varied $h \in \{2^{-2}, 2^{-3}, 2^{-4}, 10^{-5}\}$, $\nu \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}\}$, and we set $\beta = (\beta_1, 0, 0)$ with $\beta_1 \in \{0, 10, 100, 1000\}$. We remark that $\nu = 10^{-8}$ was included for completeness, however it will be

		$\nu = 10^{-2}$		$\nu = 10^{-4}$		$\nu = 10^{-6}$		$\nu = 10^{-8}$	
β_1	h	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU
0	2^{-2}	9.6(3)	0.2	6.5(7)	0.1	10.3(9)	0.2	11.1(9)	0.2
	2^{-3}	9.5(4)	3.2	11.2(11)	8.0	16.0(19)	21.5	18.3(27)	20.2
	2^{-4}	8.5(4)	8.5	10.7(17)	30.1	17.6(54)	160.7	30.3(74)	540.5
	2^{-5}	8.0(4)	48.2	10.3(15)	241.4	22.0(68)	2608.4	-	-
10	2^{-2}	9.0(3)	0.1	8.3(10)	0.2	10.4(10)	0.3	11.3(10)	0.3
	2^{-3}	8.5(4)	0.7	10.5(13)	3.0	15.4(18)	6.8	19.8(19)	10.7
	2^{-4}	8.5(4)	6.1	10.8(13)	25.3	18.6(41)	135.9	23.8(109)	509.1
	2^{-5}	8.0(4)	53.6	11.0(15)	277.6	20.9(47)	1810.7*	36.9(164)	13203.7*
100	2^{-2}	5.0(3)	0.1	7.0(4)	0.1	10.0(6)	0.1	13.7(8)	0.3
	2^{-3}	6.0(3)	0.4	9.6(5)	0.9	12.3(12)	3.5	23.7(19)	12.9
	2^{-4}	5.3(3)	2.9	8.8(6)	8.9	15.1(14)	41.4	34.3(46)	337.9
	2^{-5}	7.3(3)	40.3	10.0(6)	108.2	14.4(19)	690.5*	40.0(81)	8383.7*
1000	2^{-2}	3.0(2)	0.1	4.5(2)	0.1	6.0(4)	0.1	8.8(6)	0.2
	2^{-3}	4.0(2)	0.2	5.0(2)	0.2	5.8(6)	0.8	16.3(18)	7.6
	2^{-4}	4.5(2)	1.7	6.5(2)	2.4	8.1(6)	9.1	18.0(14)	53.2
	2^{-5}	4.5(2)	29.3	5.6(3)	53.5	7.2(7)	156.5	25.0(26)	2517.2*

		$\nu = 10^{-2}$		$\nu = 10^{-4}$		$\nu = 10^{-6}$		$\nu = 10^{-8}$	
β_1	h	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU
0	2^{-2}	8.7(4)	0.2	11.1(9)	0.2	12.1(12)	0.4	10.4(11)	0.3
	2^{-3}	8.0(5)	0.8	12.9(13)	3.8	16.8(22)	8.8	15.7(19)	8.2
	2^{-4}	7.4(5)	6.5	13.0(14)	29.5	18.2(35)	106.9	27.6(55)	289.1
	2^{-5}	7.4(5)	56.4	11.7(13)	240.8	20.0(41)	1416.9	41.0(156)	14156.0
10	2^{-2}	8.0(4)	0.1	10.6(10)	0.3	13.8(15)	0.5	15.1(15)	0.6
	2^{-3}	8.0(4)	0.7	13.1(12)	3.5	19.3(31)	14.7	24.6(30)	21.5
	2^{-4}	6.4(5)	6.7	13.5(12)	28.7	20.6(46)	167.6	34.1(67)	449.5
	2^{-5}	6.6(5)	59.1	12.3(13)	273.1	21.7(58)	2291.8	41.4(162)	15118.4*
100	2^{-2}	4.5(2)	0.1	9.8(6)	0.2	12.3(10)	0.3	15.5(12)	0.5
	2^{-3}	4.3(2)	0.3	10.3(6)	1.1	15.7(16)	5.2	27.2(24)	18.5
	2^{-4}	4.6(3)	2.7	9.6(6)	9.7	18.4(20)	61.5	33.2(47)	321.4
	2^{-5}	5.6(3)	69.3	8.5(7)	196.7	17.6(23)	907.8	34.3(82)	7428.8
1000	2^{-2}	3.0(2)	0.1	5.0(3)	0.1	10.1(7)	0.2	13.6(9)	0.3
	2^{-3}	2.5(2)	0.1	5.0(3)	0.3	10.1(7)	1.4	22.3(12)	6.8
	2^{-4}	2.5(2)	0.2	4.5(4)	3.7	9.5(7)	12.1	21.1(15)	58.9
	2^{-5}	3.0(2)	25.9	4.2(4)	66.4	8.2(8)	237.4	19.3(17)	2277.6*

Table 5.7: AS-GMRES-IPF for a variety of values for h, ν and β. The symbol '*' denotes runs where an HSL-MI20 warning occurred; in some of these cases, much larger timings were observed. Top: `CC-Pb1`. Bottom: `CC-Pb2`.

		$\nu = 10^{-2}$		$\nu = 10^{-4}$		$\nu = 10^{-6}$		$\nu = 10^{-8}$	
β_1	h	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU
0	2^{-2}	20.0(3)	0.2	13.8(7)	0.2	22.7(9)	0.4	25.4(9)	0.4
	2^{-3}	19.5(4)	4.2	23.8(11)	14.4	34.6(19)	35.8	40.1(27)	57.6
	2^{-4}	18.7(4)	9.9	23.5(17)	51.6	44.9(54)	289.8	72.1(66)	513.4
	2^{-5}	19.2(4)	144.4	24.3(15)	474.3	56.3(89)	5627.2	-	-
10	2^{-2}	18.3(3)	0.1	18.3(10)	0.3	25.3(10)	0.5	29.0(10)	0.5
	2^{-3}	17.7(4)	4.2	24.6(13)	19.0	37.7(18)	39.7	50.6(19)	56.7
	2^{-4}	17.7(4)	10.8	26.5(13)	40.0	53.7(33)	245.7	87.7(42)	500.7
	2^{-5}	19.2(4)	98.4	29.5(15)	550.6	63.1(74)	5572.9	174.1(146)†	$> 5h$*†
100	2^{-2}	10.5(2)	0.1	14.0(4)	0.1	20.5(6)	0.2	31.5(8)	0.4
	2^{-3}	11.6(3)	1.8	20.2(5)	5.2	27.1(12)	17.4	53.5(19)	54.3
	2^{-4}	11.6(3)	5.3	20.5(6)	17.8	37.0(14)	72.8	93.5(40)	546.5
	2^{-5}	98.3(3)	737.3	28.5(6)	382.7	74.8(19)†	2988.0*†	180.2(80)†	$> 5h$*†
1000	2^{-2}	6.5(2)	0.1	8.5(2)	0.1	11.5(4)	0.1	17.5(6)	0.2
	2^{-3}	7.5(2)	0.9	10.5(2)	1.1	11.8(6)	4.0	29.0(8)	13.5
	2^{-4}	9.5(2)	3.1	13.5(2)	4.3	16.8(6)	15.6	41.1(10)	61.7
	2^{-5}	9.5(2)	79.4	11.6(3)	146.3	19.0(7)	549.3	42.1(16)	3036.1*

		$\nu = 10^{-2}$		$\nu = 10^{-4}$		$\nu = 10^{-6}$		$\nu = 10^{-8}$	
β_1	h	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU
0	2^{-2}	18.0(4)	0.2	23.2(9)	0.4	26.5(12)	0.6	23.1(11)	0.5
	2^{-3}	16.8(5)	4.7	27.7(13)	19.8	36.8(22)	44.6	35.5(19)	37.1
	2^{-4}	16.2(5)	11.1	28.7(14)	52.1	43.5(36)	204.8	69.0(63)	572.0
	2^{-5}	16.6(5)	96.7	27.5(13)	416.1	52.5(53)	3152.0	123.2(133)†	$> 5h$†
10	2^{-2}	16.7(4)	0.1	23.0(10)	0.4	32.4(15)	0.9	37.4(15)	1.0
	2^{-3}	16.5(4)	4.0	30.6(12)	21.7	52.2(30)	90.8	70.3(30)	122.1
	2^{-4}	14.2(5)	10.8	32.0(12)	54.6	63.7(47)	409.8	108.2(79)	1151.1
	2^{-5}	14.8(5)	98.8	33.0(13)	533.3	78.4(100)	$> 5h$	194.2(159)†	$> 5h$*†
100	2^{-2}	9.5(3)	0.1	20.6(6)	0.2	27.1(10)	0.4	35.6(12)	0.7
	2^{-3}	9.0(3)	1.5	22.0(6)	7.0	37.3(16)	31.5	66.2(25)	86.9
	2^{-4}	9.6(3)	4.7	22.1(6)	19.1	47.1(20)	138.3	90.9(56)	744.5
	2^{-5}	226.0(3)	1206.2	112.5(7)	1411.8	109.2(23)†	4762.4†	-	-
1000	2^{-2}	5.5(2)	0.1	10.3(3)	0.1	21.0(7)	0.3	28.4(9)	0.5
	2^{-3}	5.5(2)	0.6	10.3(3)	1.8	20.4(7)	7.7	49.6(11)	31.3
	2^{-4}	5.5(2)	2.3	10.0(4)	6.6	19.5(7)	21.2	63.8(15)†	699.3†
	2^{-5}	6.5(2)	71.1	8.0(5)	8.3	25.5(8)†	828.3†	67.7(26)†	7897.7†

Table 5.8: AS-MINRES-BDF for a variety of values for h, ν and β. The symbol '*' denotes runs where an MI20 warning occurred; in some of these cases, much larger timings were observed ($> 5h$ means that the CPU is larger than 5 hours). Top: `CC-Pb1`. Bottom: `CC-Pb2`.

considered as a limit case because it is rather small. Analogously, $\beta_1 = 1000$ makes the operator very convection-dominated, providing anomalous behaviors in some exceptional cases; we did not explore whether for this extreme value of β_1 the upwind discretization was sufficient in these cases to damp the well-known numerical instabilities arising in the discretization phase. In fact, the value $\beta_1 = 1000$ was only considered for consistency with respect to the experiments carried out in [69]. In the same lines, we prefer to limit our speculations on the dependence with respect to β to the empirical level, as a deeper analysis would require a thorough discussion of both the discretization strategy and the employed convection; this is clearly beyond the scope of this paper.

We collect the results obtained with AS-GMRES-IPF and AS-MINRES-BDF for the problems `CC-Pb1` and `CC-Pb2` in Tables 5.7-5.8 and the corresponding total CPU time performance profile is displayed in Figure 5.2 (left plot) varying all the parameters for a total of 128 runs. The average number of inner iterations is quite homogeneous with respect to h and slightly dependent on ν and β. A comparison of Tables 5.7-5.7 and 5.8-5.8 shows that the number of nonlinear iterations is quite different between AS-GMRES-IPF and AS-MINRES-BDF when h is small and $\nu \in \{10^{-6}, 10^{-8}\}$. For these values the preconditioner in AS-MINRES-BDF is rather ill-conditioned and its performance deteriorates. In this case, the Newton steps computed with the two preconditioned solvers, using the stopping criterion (5.40), might differ so greatly that different convergence histories take place. Unfortunately, this resulted in the AS-MINRES-BDF failure in 10 instances. We recovered 9 over 10 failures by imposing the stricter tolerances $\tau_s = \tau_1 = 10^{-12}$ in (5.40) (this runs are marked with the symbol '†' in Table 5.8). A few unexpected large values of LI can still be observed in Table 5.8 for $\beta = 100$ and $h = 2^{-5}$, which can be presumably ascribed to an inaccuracy of the multigrid operator.

The superiority of AS-GMRES-IPF is also evident in the left plot of Figure 5.2, which reveals that AS-GMRES-IPF is much more efficient than AS-MINRES-BDF in terms of total CPU time and that in the 55% of the runs, the CPU time employed by AS-MINRES-BDF is within a factor 2 of the time employed by AS-GMRES-IPF.

Finally, for the sake of completeness, we also carried out experiments on `CC-pb1` using the AGMG algebraic multigrid operator [96, 97] in place of the HSL-MI20 in the solution of systems with L_1. The implementation of AGMG requires the use of the "flexible" variant of the linear system solver since the application of multigrid preconditioner is the result of an iterative process and therefore it changes step by step [114]. Table 5.9 shows the results obtained using Flexible GMRES (FGMRES) in combination with HSL-MI20 (first two columns) and AGMG (last two columns) in the application of

the $\mathcal{P}_k^{CPF}$ preconditioner. We only report experiments with $\nu \in \{10^{-6}, 10^{-8}\}$, as for larger values the performance with the two multigrid preconditioners is very similar. For $\nu = 10^{-6}$ the overall performance in terms of CPU time is still somewhat comparable, whereas it is clearly in favor of AGMG in the extreme case $\nu = 10^{-8}$. On the other hand, the average number of iterations (LI) with HSL-MI20 is in general lower, showing that the latter preconditioner is more effective in terms of approximation properties, but more expensive to apply.

		AS-FGMRES-IPF with HSL-MI20				AS-FGMRES-IPF with AGMG			
		$\nu = 10^{-6}$		$\nu = 10^{-8}$		$\nu = 10^{-6}$		$\nu = 10^{-8}$	
β_1	h	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU
0	2^{-2}	10.3(9)	0.3	11.1(9)	0.4	17.0(9)	0.3	18.7(9)	0.3
	2^{-3}	16.0(19)	21.3	18.3(27)	36.4	23.7(19)	12.3	28.5(27)	27.4
	2^{-4}	17.6(54)	223.6	28.5(57)	351.6	28.5(54)	157.4	41.7(57)	286.7
	2^{-5}	21.9(68)	2794.9	38.7(190)	17457.3	44.8(68)	3766.9	53.3(189)	14250.8
10	2^{-2}	10.4(10)	0.2	11.3(10)	0.4	19.0(10)	0.2	20.7(10)	0.3
	2^{-3}	15.4(18)	20.9	19.7(19)	31.1	25.0(18)	11.5	31.7(19)	23.1
	2^{-4}	18.6(41)	132.6	26.4(42)	239.4	25.0(41)	79.1	35.4(42)	272.3
	2^{-5}	20.9(47)	1938.1	37.9(138)*	12582.4*	39.3(47)	1951.3	50.2(147)	10105.1
100	2^{-2}	10.0(6)	0.1	13.7(8)	0.5	16.0(6)	0.1	20.3(8)	0.3
	2^{-3}	12.3(12)	10.5	23.7(19)	36.5	20.5(12)	7.5	31.5(19)	24.4
	2^{-4}	15.1(14)	39.8	36.8(34)	321.0	28.7(14)	37.6	39.4(34)	153.3
	2^{-5}	14.4(19)*	776.0*	38.8(82)	9172.9*	39.2(19)	755.6	50.2(81)	5658.0
1000	2^{-2}	6.0(4)	0.1	8.8(6)	0.37	8.2(4)	0.1	12.3(6)	0.2
	2^{-3}	5.8(6)	2.1	14.1(8)	9.3	9.8(6)	1.2	18.7(8)	5.1
	2^{-4}	8.1(6)	8.5	20.1(10)	43.2	12.0(6)	5.1	26.3(10)	31.6
	2^{-5}	7.2(7)	150.1	19.6(16)*	1660.9*	13.1(7)	43.9	26.6(16)	600.6

Table 5.9: AS-FGMRES-IPF (flexible variant) using HSL-MI20 (left) and AGMG (right) for a variety of values of h and β, and small values of ν. The symbol '*' denotes runs where an HSL-MI20 warning occurred; Test problem `CC-Pb1`.

Example 5.2. We further investigate the reliability of our proposals considering problem `CC-pb2` with the following nonconstant convection parameter

$$\beta(x, y, z) = \begin{pmatrix} -2x(1-x)(2y-1)z \\ (2x-1)y(1-y) \\ (2x-1)(2y-1)z(1-z) \end{pmatrix}; \tag{5.41}$$

see example 3D1 in [96]. The performance of AS-GMRES-IPF and AS-MINRES-BDF is analogous to that showed in Tables 5.7-5.8 for the constant and unidirectional $\beta = (\beta_1, 0, 0)$; a sample of this behavior for AS-GMRES-IPF is reported in Table 5.10 as ν and h vary.

	AS-GMRES-IPF on CC-Pb2 with convection (5.41)							
	$\nu = 10^{-2}$		$\nu = 10^{-4}$		$\nu = 10^{-6}$		$\nu = 10^{-8}$	
h	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU
2^{-2}	5.0(3)	0.3	9.0(8)	0.4	12.4(19)	0.5	14.5(19)	0.7
2^{-3}	5.0(3)	2.2	9.4(8)	5.4	15.1(32)	34.2	20.1(41)	66.9
2^{-4}	4.7(3)	5.9	8.3(9)	14.6	15.4(39)	133.2	26.5(92)	528.8
2^{-5}	5.0(3)	42.2	7.8(9)	139.9	14.9(36)	1089.0	28.6(137)	9643.7

Table 5.10: AS-GMRES-IPF on problem CC-Pb2 with convection β given in (5.41).

Example 5.3. For the MC and SC problems, we considered $h \in \{2^{-2}, 2^{-3}, 2^{-4}\}$, $\nu \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}\}$, $\beta = (\beta_1, 0, 0)$ with $\beta_1 \in \{0, 10, 100, 1000\}$, and $\epsilon \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-8}, 0\}$, where the values $\epsilon \in \{10^{-8}, 0\}$ are included to comprise the SC problems. We thus obtained a set of 288 runs. The numerical results for these problems do not significant differ from those of the CC problem, at least for the larger values of ϵ in the set. Therefore, to avoid proliferation of tables, we prefer not to include them, and report instead the overall performance profile in the right plot of Figure 5.2. For all considered runs, the profile clearly shows that AS-GMRES-IPF is the fastest in the 96% of the runs and that AS-MINRES-BDF is within a factor 2 of AS-GMRES-IPF for the majority (93%) of the runs.

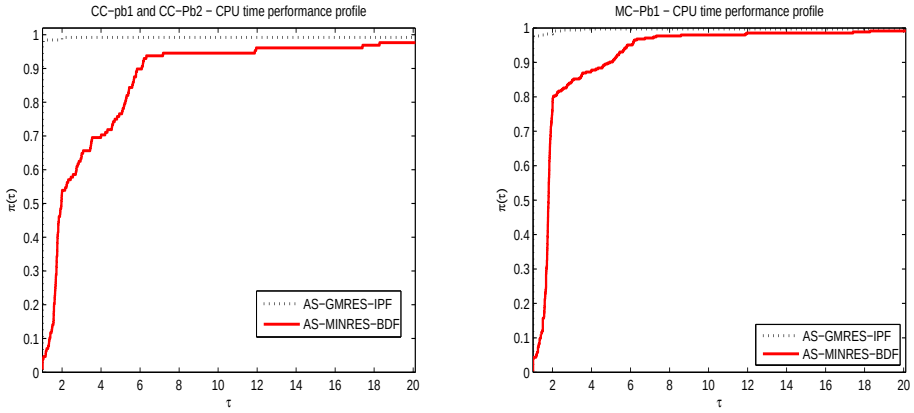

Figure 5.2: Total CPU time performance profile for AS-GMRES-IPF and AS-MINRES-BDF. Left: CC-Pb1 and CC-Pb2. Right: MC-Pb1.

A deeper exploration deserves the dependence on ϵ, and the mutual influence of ϵ and ν. In Tables 5.11 and 5.12 we report the average number of inner iterations for $h = 2^{-4}$ obtained with AS-GMRES-IPF and AS-MINRES-BDF, resp., as ν and ϵ vary.

$\beta_1 = 10$

ϵ / ν	-1	-2	-3	-4	-8	$-\infty$
-2	10.3	14.3	35.3	32.5	34.1	34.1
-4	13.5	13.3	16.6	20.2	21.0	21.3
-6	19.5	16.0	14	13.5	13.5	13.5
-8	25.8	18.4	12.0	10.5	10.5	10.5

$\beta_1 = 100$

ϵ / ν	-1	-2	-3	-4	-8	$-\infty$
-2	6.0	7.7	8.7	9.0	9.7	9.0
-4	12.3	13.3	16.3	22.3	26.6	26.4
-6	21.6	19.8	14.7	17.7	16.7	16.7
-8	40.4	34.2	18.0	14.0	13.5	13.5

Table 5.11: Mixed Constraints `MC-Pb1`: Average number of GMRES iterations using AS-GMRES-IPF with $h = 2^{-4}$ and varying ν and ϵ ($\log_{10}$ values of ν, ϵ).

$\beta_1 = 10$

ϵ / ν	-1	-2	-3	-4	-8	$-\infty$
-2	22.0	32.1	60.7	86.2	91.6*	93.8*
-4	27.8	27.0	34.4	42.2	44.1	44.8
-6	45.3	33.2	27.5	27.5	21.5	27.5
-8	65.7	39.8	24.5	21.5	21.5	21.5

$\beta_1 = 100$

ϵ / ν	-1	-2	-3	-4	-8	$-\infty$
-2	14.0	17.3	19.2	20.0	22.6	21.3
-4	27.0	28.0	33.7	44.2	56.0	55.2
-6	49.4	41.4	29.7	36.0	34.7	34.7
-8	93.6	71.9	36.3	28.5	27.5	27.5

* 6 pre/post smoothing steps set in HSL-MI20

Table 5.12: Mixed Constraints `MC-Pb1`: Average number of MINRES iterations using AS-MINRES-BDF with $h = 2^{-4}$ and varying ν and ϵ ($\log_{10}$ values of ν, ϵ).

We observe that for $\beta = 10$ the average number of inner iterations becomes large when ϵ is small and ν is large (top right corner) whereas for $\beta = 100$ the increase in iteration number is more evident in the opposite setting (bottom left corner). Overall, the variation of the reported values is quite modest and smallest values are located on the diagonal of the table (shaded cells), i.e. when $\nu = \epsilon^2$. We recall that $\nu = \epsilon^2$ corresponds to $\gamma_1 = \gamma_2 = \frac{1}{2}$ in the block L_1 of the Schur approximation (5.27), so that Proposition 5.3.6 holds (see Remark 2). We also notice that the variation in the number of iterations is significantly less pronounced for the indefinite preconditioner than for the block diagonal preconditioner. In particular, for a fixed ν, the average number of iterations for AS-GMRES-IPF varies very mildly. More significant variations for fixed ν are visible for AS-MINRES-BDF, see Table 5.12. Moreover, we observe that the behavior of the proposed preconditioner does not deteriorate for $\epsilon \to 0$ and, in particular, fully satisfying results are obtained for $\epsilon = 0$, i.e. in the solution of State Constrained problems.

We point out that similar digits were observed when using a direct solver (not reported here) in place of HSL-MI20 within the preconditioners. Therefore, the different performance as the parameters deviate from $\nu = \epsilon^2$ is not

due to the preconditioner inexactness, but rather, to the different quality of the (exact) preconditioner itself. The only exception is given by the two runs marked with the symbol '*' in Table 5.12, for which a lower average number of MINRES iterations was observed when using a direct solver in place of HSL-MI20.

The inexact active-set Newton method for CC problems

Performing the experiments on problems with CC constraints (5.2), we observed different trends in the nonlinear iteration progress varying the parameters ν and β, see e.g. the values of NLI in Table 5.7. To clarify this issue, we plot in Figure 5.3 the convergence history of AS-GMRES-IPF on `CC-Pb1` with mesh size $h = 2^{-4}$ varying $\beta_1 \in \{0, 10, 100, 1000\}$ and setting $\nu = 10^{-2}$ in the left plot and $\nu = 10^{-6}$ in the right plot.

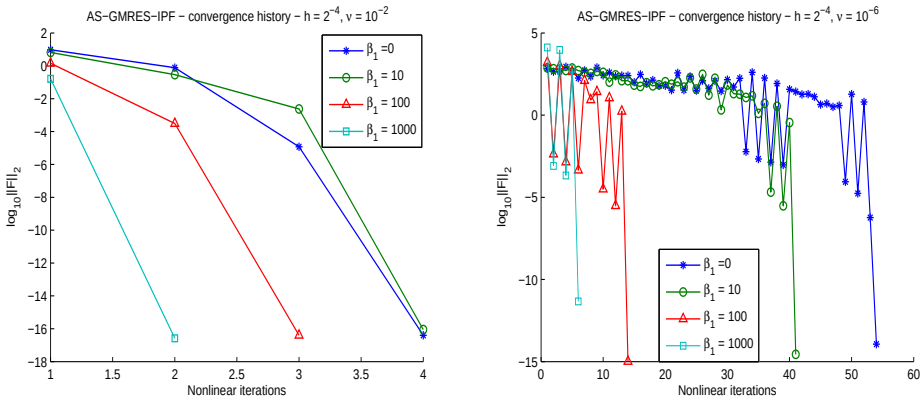

Figure 5.3: Convergence history of AS-GMRES-IPF for the `CC-pb1` with $h = 2^{-4}$. Left: $\nu = 10^{-2}$. Right: $\nu = 10^{-6}$.

Looking at each plot we note that the number of nonlinear iterations decreases as β becomes larger; moreover, comparing the two plots, we observe an increase of Newton steps for a smaller ν. More interestingly, the right plot in Figure 5.3 shows a long stagnation phase in the nonlinear process before reaching the local area of fast Newton convergence. In this first phase, away from a solution, choosing an η_k too small (as in (5.40)) can lead to oversolving the Newton equation (5.14): the corresponding step may result in little or no progress toward a solution, while involving pointless expense.

We therefore combined the active-set method with the inexact adaptive choice (5.18). We report in Table 5.13 the results of AS-GMRES-IPF using the adaptive value $\eta_k = \eta_k^I$ in (5.18) on problem `CC-Pb1` with $h \in \{2^{-4}, 2^{-5}\}$, $\beta_1 \in \{0, 10, 100, 1000\}$ and $\nu \in \{10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}\}$.

		$\nu = 10^{-2}$		$\nu = 10^{-4}$		$\nu = 10^{-6}$		$\nu = 10^{-8}$	
β_1	h	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU	LI(NLI)	TCPU
0	2^{-4}	3.2(5)	3.2	4.5(18)	15.9	7.0(60)	77.1	13.1(131)	281.9
	2^{-5}	4.0(6)	59.1	2.7(16)	102.0	3.5(92)	646.4	-	-
10	2^{-4}	3.5(4)	3.2	3.2(14)	11.1	5.2(60)	65.9	112.7(101)	233.9
	2^{-5}	4.0(5)	39.6	2.9(15)	100.7	3.0(53)	377.7*	-	-
100	2^{-4}	3.0(3)	1.8	3.3(6)	4.1	11.6(14)	30.4	10.6(51)	100.5
	2^{-5}	4.3(3)	25.4	3.1(6)	42.1	2.9(27)	332.4*	-	-
1000	2^{-4}	2.5(2)	1.1	3.3(3)	2.1	7.5(6)	8.5	20.4(10)	45.1
	2^{-5}	2.5(2)	18.1	3.0(3)	31.6	2.5(7)	67.3	8.1(18)	924.1

Table 5.13: AS-GMRES-IPF on `CC-Pb1` for $h \in \{2^{-4}, 2^{-5}\}$ and a variety of values for ν and β. The symbol '*' denotes runs where an HSL-MI20 warning occurred.

Let us compare values in Table 5.13 with the corresponding values in Table 5.7 (top table) obtained with η_k^E. The average number of linear iterations is smaller in Table 5.13 than in Table 5.7 while the number of nonlinear iterations is larger in 15 over 29 successful runs. Overall, the saving in number of inner iterations of AS-GMRES-IPF with η_k^I makes it faster than AS-GMRES-IPF with η_k^E in all runs. Two extra failures occur when η_k^I is used in the limit case $\nu = 10^{-8}$.

Summarizing, the inexact strategy is both cheaper and more effective in solving problem (5.10), especially for $\nu \in \{10^{-4}, 10^{-6}\}$ and $\beta_1 \leq 10$, that is values for which the stagnation phase is longer. Note that in particular, a less stringent inner accuracy allows a fast solution also in the limit case $\beta_1 = 1000$.

5.6 Conclusions

We have proposed two classes of preconditioners (a positive definite one and an indefinite one) for efficiently solving problem (5.1) by means of an active-set Newton method. Both acceleration strategies rely on a new effective approximation to the Schur complement of the Jacobian matrix, for which spectral estimates are provided.

A large set of numerical experiments shows the great potential of these preconditioners for a large range of all problem parameters. As opposed to the current literature, we cope with the indefiniteness of the problem by appropriately choosing the structured preconditioner, and we include active set information explicitly in the preconditioning blocks to exploit this information at later stages. Therefore, the preconditioner adapts dynamically with the modification of the active sets. This procedure allowed us to devise a general and simple to implement acceleration strategy, that can be employed

either within MINRES (in the block diagonal form) or within GMRES (in the indefinite factorized form). The latter formulation outperforms MINRES in all test cases, and shows significantly lower sensitivity to the extreme values of the parameters. In general, memory requirements of GMRES remain modest, as the number of iterations stays quite small throughout the nonlinear process. For the smallest values of ν, however, the number of GMRES iterations may make its memory requirements undesirably high. In this case, a short-term recurrence such as the symmetric version of QMR could be considered as an alternative; see, e.g., [103] for a discussion and related numerical experiments. We also mention that a dimension reduction could be employed in the original system (5.14). This strategy is discussed in [121] in the case when no bound constraints are imposed, and it could be naturally generalized to our setting.

Although some of the preconditioner blocks need to be recomputed at each Newton iteration, this cost does not seem to penalize the overall performance of the preconditioned solver. Numerical comparisons with state-of-the-art methods available in the literature support these claims.

Finally, we mention that more general regularization terms could be considered for the cost functionals, for instance, by enforcing sparsity constraints, see, e.g., [124]. We aim to address this important aspect in future research.

Part III

Saddle point systems arising in the solution of Quadratic Programming problems

Chapter 6

Spectral estimates for unreduced symmetric systems[1]

This chapter is devoted to the study of saddle point systems stemming from convex Quadratic Programming (QP) problems in standard form, i.e.

$$\min_x \quad c^T x + \frac{1}{2} x^T H x \qquad \text{subject to} \quad Jx = b,\ x \geq 0, \tag{6.1}$$

where $J \in \mathbb{R}^{m \times n}$ has full row rank $m \leq n$, $H \in \mathbb{R}^{n \times n}$ is symmetric and positive semidefinite, x, z, $c \in \mathbb{R}^n$, y, $b \in \mathbb{R}^m$.

Interior Point (IP) methods are effective iterative procedures for solving such problems, possibly of very large dimension, see [8, 12, 59, 48, 86, 138] and references therein. Since they are second-order methods, a linear algebra phase constitutes their computational core and its practical implementation is crucial for the efficiency of the overall optimization procedure. Therefore, linear algebra of IP methods has been extensively studied in all algorithmic issues, including formulation of the systems arising at each iteration, employment of direct and iterative solvers, preconditioning, inertia control.

The application of a primal-dual IP method gives rise, at each iteration, to a nonsymmetric 3×3 block system of dimension $2n+m$, sometimes referred to as KKT system [46, 49, 30]. Such system allows for alternative formulations differing for dimension, conditioning and definiteness [48, 49, 136].

The nonsymmetric 3×3 block matrix can be easily symmetrized without increasing the conditioning of the system [46], and here we will refer to the resulting symmetric matrix as the *unreduced* matrix. On the other hand, by exploiting the structure of the nonsymmetric 3×3 block matrix and block elimination, it is common to use a linear system of dimension $n + m$ with a *reduced* (or *augmented*) symmetric 2×2 block saddle point matrix.

[1] The results presented in this chapter are taken from [91].

The focus of this chapter is the theoretical and numerical study of the unreduced 3×3 formulation and, in some respects, a comparative analysis with the reduced 2×2 formulation.

Unlike the reduced 2×2 matrix, under suitable conditions on both the problem (6.1) and the solution, the unreduced matrix has condition number asymptotically uniformly bounded, and typically remains well-conditioned as the solution is approached [46, 48]. Motivated by this feature, in a very recent paper Greif et al. [66] presented a spectral analysis for the 3×3 matrix and claimed that this formulation can be preferable to the reduced one in terms of eigenvalues and conditioning, The study in [66] covers also regularized variants of KKT matrices arising from regularizations of the optimization problem.

The study conducted in [66] has renewed the interest in the unreduced formulation but leaves some issues open, that need to be addressed before any thorough comparison with the reduced formulation can be started. Specifically, some eigenvalue bounds presented in [66] may be overly pessimistic and not tight for the unregularized 3×3 matrix; in fact they may not reflect the nonsingularity of the matrix. In this chapter we aim at filling these gaps and offering new spectral bounds which improve results in [66].

We will not discuss preconditioning here, but this will be the main topic of the next chapter, where we will offer a comprehensive study of the impact of the system formulation on preconditioning techniques. Numerical experiments that compare the two formulations are also postponed to the next chapter.

The reminder of this chapter is organized as follows. In Section 6.1 we introduce the problem and briefly detail how the application of an IP method leads to the solution of the KKT system. In section 6.2 we introduce the different formulations of the system and report the main results presented in [66]. In Section 6.3 we give new estimates on the bounds of the unreduced KKT matrix and perform the analysis for the early and middle stage of the IP method, and for the late stage of the IP method, separately. In Section 6.4 we show numerical validation of the bounds obtained. Final conclusions are drawn in Section 6.5.

Notation. For any $x \in \mathbb{R}^n$ and set of indices $\mathcal{C} \subset \{1, 2, \ldots, n\}$, we write $x_{\mathcal{C}}$ for the subvector of x having components x_i with $i \in \mathcal{C}$. Further, if B is a matrix we write $B_{\mathcal{C}}$ for the submatrix of the columns of B with indices in $\mathcal{C}$.

6.1 Interior Point methods

We start this section by briefly reviewing the theory on the numerical solution of QP via Interior Point methods. For a more detailed treatment we refer to the monographs [138, 18, 95].

Similarly as done in Chapter 5, we introduce the Lagrangian function associated with the QP problem (6.1), that is

$$\mathcal{L}(x, y, z) = \frac{1}{2} x^T H x + c^T x - y^T (Jx - b) - z^T x.$$

Here, y is the Lagrange multiplier (or dual variable) associated with the equality constraints, while z is the Lagrange multiplier associated with the inequality constraints. Note that we use different names for the dual variables with respect to Theorem 5.1.1, to be consistent with the standard notation found in the optimization literature. The resulting KKT conditions, which characterize the primal-dual solution $(\hat{x}, \hat{y}, \hat{z})$ of (6.1) (cf. (5.7)) read:

$$\begin{aligned}
\nabla_x \mathcal{L} = H\hat{x} - J^T\hat{y} - \hat{z} + c &= 0, && \text{(6.2a)}\\
J\hat{x} &= b, && \text{(6.2b)}\\
\hat{x} &\geq 0, && \text{(6.2c)}\\
\hat{z} &\geq 0, && \text{(6.2d)}\\
\hat{x}^T\hat{z} &= 0. && \text{(6.2e)}
\end{aligned}$$

Interior Point methods generate a sequence of approximate solution (x, y, z) for the above equations for which the inequality constraints (6.2c) and (6.2d) are strictly satisfied, i.e. $(x, z) > 0$. The name "Interior Point" actually comes from this fundamental property.

We now report a possible derivation of these methods. To this end, we first introduce the notion of Central Path. Given a positive scalar τ, known as *barrier parameter*, we consider the following equations:

$$\begin{aligned}
Hx - J^T y - z + c &= 0, && \text{(6.3a)}\\
Jx &= b, && \text{(6.3b)}\\
x &> 0, && \text{(6.3c)}\\
z &> 0, && \text{(6.3d)}\\
x^T z &= \tau. && \text{(6.3e)}
\end{aligned}$$

These conditions are the same as the KKT conditions, except that now x and z have strictly positive components and the complementarity condition has

been replaced with $x^T z = \tau$. If we assume that there exists a triple (x, y, z) that satisfies the first four conditions (6.3a)-(6.3d), then equations (6.3) have a unique solution (x_τ, y_τ, z_τ) for any given value of τ. The Central Path is then defined as

$$\mathcal{C} = \{(x_\tau, y_\tau, z_\tau) \mid (x_\tau, y_\tau, z_\tau) \text{ satisfies (6.3)}, \ \tau > 0\}.$$

It is apparent that all the points of the Central Path satisfy $(x, z) > 0$, i.e., they lie in the interior of the set $(x, z) \geq 0$. Moreover, if we assume that the solutions of (6.3) converge as τ goes to 0, then they necessarily converge to an exact solution of (6.3).

The idea of Interior Point methods is to find approximate solutions which "follow" the Central Path towards the exact solution of the problem. To this end, at each iteration we take a Newton step towards the point (x_τ, y_τ, z_τ). The barrier parameter τ is gradually reduced to 0 as the method progresses, to ensure the convergence of the method to the solution of (6.2).

More precisely, if we introduce the nonlinear function

$$F_\tau(x, y, z) = \begin{bmatrix} Hx - J^T y - z + c \\ Jx - b \\ XZe - \tau e \end{bmatrix}$$

where $\widehat{X} = \text{diag}(x)$ and $Z = \text{diag}(z)$, then at each iteration of the method we have to solve the linear system

$$\mathcal{J}(x, y, z) \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} = -F_{\tau_i}(x, y, z),$$

where (x, y, z) is the current iterate and $\mathcal{J}$ denotes the Jacobian matrix of F_{τ_i}, i.e.,

$$\begin{bmatrix} H & J^T & -I \\ J & 0 & 0 \\ -Z & 0 & -X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} = \begin{bmatrix} -Hx + J^T y + z - c \\ -Jx + b \\ -XZe - \tau e \end{bmatrix}. \tag{6.4}$$

The next iterate is then computed as

$$(x^+, y^+, z^+) = (x, y, z) + \alpha(\Delta x, \Delta y, \Delta z)$$

where $\alpha > 0$ is the step length, chosen to ensure $(x^+, z^+) > 0$ and possibly other conditions.

Typically, $\tau = \sigma\mu$, where $\sigma \in [0, 1]$ is known as *centering parameter* and $\mu = x^T z/n$ is the *duality measure* of the current iterate. The duality

measure tells us how much the complementarity condition (6.2e) is violated in the approximate solutions, The centering parameter, whose actual value depends on the specific Interior Point method considered, plays a crucial role since it governs the reduction of the duality measure and the distance from the Central Path of the subsequent iteration. Indeed, we observe that we are taking a Newton step towards the Central Path point $(x_{\sigma\mu}, y_{\sigma\mu}, z_{\sigma\mu})$. If $\sigma = 0$, we are actually moving in the direction of the exact solution $(\hat{x}, \hat{y}, \hat{z})$. However, since we are not following the Central Path, we are often forced to take a very small step in that direction ($\alpha \ll 1$) to ensure that the condition $(x^+, z^+) > 0$ holds. On the other hand, if we take $\sigma = 1$, we are moving towards the point (x_μ, y_μ, z_μ) which lies in a much "central" zone, but it is characterized by having the same duality measure as the current iterate. Most Interior Point methods seek a compromise between these two extremes.

6.2 The KKT system

We now focus on the linear system (1.2), which in the literature is sometimes referred to as the *KKT system* (see e.g. [46, 49, 30]). Unsurprisingly, the numerical solution of KKT systems constitutes the computational core of Interior Point methods. As stated in the introduction of this chapter, different formulations for this system are possible. Indeed, the coefficient matrix that appears in (6.4), say $K_{3,\text{uns}}$, is symmetrizable by setting

$$K_3 = R^{-1} K_{3,\text{uns}} R, \quad \text{where} \quad R = \begin{bmatrix} I_n & 0 & 0 \\ 0 & I_m & 0 \\ 0 & 0 & Z^{\frac{1}{2}} \end{bmatrix}, \tag{6.5}$$

see [46]. Thus, we can consider the system equivalent to (6.4) with matrix[2]

$$K_3 = \begin{bmatrix} H & J^T & -Z^{\frac{1}{2}} \\ J & 0 & 0 \\ -Z^{\frac{1}{2}} & 0 & -X \end{bmatrix}, \tag{6.6}$$

Due to the presence of zero and diagonal blocks in (6.4), it is very common to eliminate Δz from the third equation and to obtain a KKT system of dimension $n + m$ with matrix

$$K_2 = \begin{bmatrix} H + X^{-1}Z & J^T \\ J & 0 \end{bmatrix}. \tag{6.7}$$

[2]There are other ways to symmetrize K_3; the symmetrization considered does not suffer inevitable ill-conditioning as the solution is approached [46, 48].

One further block elimination step yields the normal equation with matrix $K_1 = J(H+X^{-1}Z)^{-1}J^T$. In the rest of the paper, we focus on the symmetric matrices K_2 and K_3 and start by noticing that both matrices have a saddle point structure. Under the assumption that H is positive semidefinite and J has full rank, and as long as X and Z are diagonal with positive entries, K_2 and K_3 are nonsingular; K_2 has n positive and m negative eigenvalues, while K_3 has n positive and $n+m$ negative eigenvalues, see e.g. [12, Lemma 4.1], [66, Lemma 3.5, 3.8].

We concentrate on the use of the 3×3 formulation and investigate its spectral properties and preconditioning issues, as compared with the 2×2 formulation. A key issue is the behavior of the matrices in the limit of the IP procedure. We denote with $\mathcal{A}_*$ and $\mathcal{I}_*$ respectively the active and inactive sets at the exact solution $\hat{x}$, i.e. the sets

$$\mathcal{A}_* := \{i = 1, \ldots, n \,|\, \hat{x}_i = 0\}\,, \qquad \mathcal{I}_* := \{1, \ldots, n\} \setminus \mathcal{A}_*. \tag{6.8}$$

As stated by (6.2e), $\hat{x}$ and $\hat{z}$ are complementary, that is $\hat{x}_i\hat{z}_i = 0$ for every $i = 1, \ldots, n$. We also say that vectors $\hat{x}$, $\hat{z}$ are *strictly complementary* if $\hat{z}_i > 0$, for all $i \in \mathcal{A}_*$, that is, for every index $i = 1, \ldots, n$ it holds either $\hat{x}_i > 0$ or $\hat{z}_i > 0$.

As a consequence of complementarity, when the IP iterates approach a solution, some entries of $X^{-1}Z$ tend to zero while others tend to infinity and the eigenvalues of the $(1,1)$ block in K_2 may spread from zero to infinity. The effect of this feature on the conditioning of K_2 can be formally described in the situation where

$$\min_{1\le i\le n} \frac{z_i}{x_i} = \mathcal{O}(\mu), \qquad \text{and} \qquad \max_{1\le i\le n} \frac{z_i}{x_i} = \mathcal{O}(\mu^{-1}),$$

and μ is the duality measure, which as already discussed goes to 0 as the IP iterates approach the exact solution. These asymptotic estimates hold when strict complementary is in place, $\mathcal{A}_* \neq \emptyset$, $\mathcal{I}_* \neq \emptyset$, and the iterates are restricted to a suitable neighborhood of the central path, see, e.g., [66, 59]. As a consequence of these assumptions, the asymptotic condition number of K_2 may get as large as $\mathcal{O}(\mu^{-2})$, [66, Corollary 5.2], [59, Lemma 2.2]. Remedies to this occurrence may consist either in scalings of K_2 [50] or in regularization strategies [59, 53, 116]. For sake the of completeness, we recall here that ill-conditioning of the matrix is usually not harmful in case direct methods are applied [47, 138].

Under suitable conditions stated below, the unreduced matrix K_3 can be well conditioned eventually and nonsingular in the limit although the diagonal scaling (6.5) used for forming the right-hand side of the system and unscaling the variables remains *benignly* ill-conditioned [46, 48]. Therefore, a

spectral analysis of the original K_3 may give insight into both its conditioning and the necessity of regularization strategies.

Let q be the cardinality of the set $\mathcal{A}_*$. Without loss of generality, suppose that the zero components of $\hat{x}$ are its first q elements. Hence, if $\hat{x}$ and $\hat{z}$ are strictly complementary by (6.8) we have

$$\hat{x} = (0, \hat{x}_{\mathcal{I}_*}), \quad \hat{z} = (\hat{z}_{\mathcal{A}_*}, 0), \quad \hat{x}_{\mathcal{I}_*} > 0, \hat{z}_{\mathcal{A}_*} > 0, \tag{6.9}$$

where $\hat{x}_{\mathcal{I}_*} \in \mathbb{R}^{n-q}$, $\hat{z}_{\mathcal{A}_*} \in \mathbb{R}^q$. We now give a definition that will be used in the following.

Definition 6.1. The Linear Independence Constraint Qualification (LICQ) is satisfied at $\hat{x}$ if the matrix $\begin{bmatrix} J^T & -I_{\mathcal{A}_*} \end{bmatrix}$ has full column rank.

Note that a necessary condition for the LICQ condition to be satisfied at any point is that J has full row rank.

It is useful to make some comments on the matrices $K_{3,\text{uns}}$ and K_3 evaluated at $x = \hat{x}$, $z = \hat{z}$. To this end, we let

$$\widehat{K}_{3,\text{uns}} = \begin{bmatrix} H & J^T & -I_n \\ J & 0 & 0 \\ -\widehat{Z} & 0 & -\widehat{X} \end{bmatrix}, \quad \widehat{K}_3 = \begin{bmatrix} H & J^T & -\widehat{Z}^{\frac{1}{2}} \\ J & 0 & 0 \\ -\widehat{Z}^{\frac{1}{2}} & 0 & -\widehat{X} \end{bmatrix}, \tag{6.10}$$

where $\widehat{X} = \text{diag}(\hat{x})$, $\widehat{Z} = \text{diag}(\hat{z})$. Throughout the paper, $\widehat{K}_3$ and K_3 will denote the coefficient matrices at the QP solution and during the iterations, respectively.

The systems involving matrices $K_{3,\text{uns}}$ and K_3 are formally equivalent also at the exact solution $(\hat{x}, \hat{z})$, at least after the natural elimination of some equations. Indeed, let us assume for simplicity that $\hat{x}$ and $\hat{z}$ are partitioned as in (6.9) and strictly complementary. Then, if in equation (6.4) we substitute $x = \hat{x}$ and $z = \hat{z}$, $\widehat{K}_{3,\text{uns}}$ is block upper triangular and upon reduction of the components of Δz with indices in $\mathcal{I}_*$, we get a system with matrix

$$\begin{bmatrix} & H & & J_{\mathcal{A}_*}^T & -I_q \\ & & & J_{\mathcal{I}_*}^T & 0 \\ J_{\mathcal{A}_*} & & J_{\mathcal{I}_*} & 0 & 0 \\ -Z_{\mathcal{A}_*} & & 0 & 0 & 0 \end{bmatrix}$$

and $J_{\mathcal{A}_*} \in \mathbb{R}^{m \times q}$, $J_{\mathcal{I}_*} \in \mathbb{R}^{m \times (n-q)}$ and $Z_{\mathcal{A}_*} = \text{diag}(\hat{z}_{\mathcal{A}_*}) \in \mathbb{R}^{q \times q}$. By using the similarity transformation with a matrix of the form (6.5), namely

$$\begin{bmatrix} I_n & 0 & 0 \\ 0 & I_m & 0 \\ 0 & 0 & Z_{\mathcal{A}_*}^{\frac{1}{2}} \end{bmatrix},$$

the resulting system is symmetric with matrix obtained by removing in $\widehat{K}_3$ the last block row and column associated to the set $\mathcal{I}_*$.

We also observe that $\widehat{K}_3$ and $\widehat{K}_{3,\text{uns}}$ have the same eigenvalues. Indeed, by (6.5) K_3 and $K_{3,\text{uns}}$ have the same eigenvalues for every strictly positive x and z. A continuity argument shows that they also coincide when taking the limit as $x \to \hat{x}$ and $z \to \hat{z}$. The following theorem states conditions under which $\widehat{K}_3$ is nonsingular, see, e.g., [66, Theorem 3.10].

Theorem 6.2.1. *Suppose H is symmetric and positive semidefinite, $\widehat{X}$ and $\widehat{Z}$ are diagonal with nonnegative entries. Then $\widehat{K}_3$ in (6.10) is nonsingular if and only if*

1. *$\hat{x}$ and $\hat{z}$ are strictly complementary,*
2. *the LICQ is satisfied at $\hat{x}$,*
3. *the null spaces of matrices H, J, $\widehat{Z}$ satisfy*

$$\ker(H) \cap \ker(J) \cap \ker(\widehat{Z}) = \{0\}. \tag{6.11}$$

In the next theorem we summarize the bounds for the eigenvalues of K_3 given in [66, Corollary 5.3 and Corollary 5.4]. To simplify the notation we let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimum and maximum eigenvalues of H and $\sigma_{\min}$ and $\sigma_{\max}$ be the minimum and maximum singular values of J.

Theorem 6.2.2 ([66])**.** *Suppose H is symmetric and positive semidefinite and let K_3 be as in (6.6).*

i) If $\theta^- I_n + X$ is nonsingular for all the negative eigenvalues θ^- of K_3, then $\theta^- \in [\zeta, 0)$, where

$$\zeta = \min\left\{\frac{1}{2}\left(\lambda_{\min} - \sqrt{\lambda_{\min}^2 + 4\sigma_{\max}^2}\right), \min_{\{j | x_j + \theta^- < 0\}} \theta_j^*\right\}, \tag{6.12}$$

and θ_j^* is the smallest negative root of the cubic polynomial

$$p_j(\theta) = \theta^3 + (x_j - \lambda_{\min})\theta^2 - (\sigma_{\max}^2 + z_j + x_j\lambda_{\min})\theta - x_j\sigma_{\max}^2. \tag{6.13}$$

ii) If J has full rank and X and Z are diagonal with positive entries, then the positive eigenvalues θ^+ of K_3 satisfy $\theta^+ \in [\theta_3, \theta_4]$, where

$$\theta_3 = \min_{1 \le j \le n} \frac{1}{2}\left(\lambda_{\min} - x_j + \sqrt{(\lambda_{\min} + x_j)^2 + 4z_j}\right), \tag{6.14}$$

$$\theta_4 = \frac{1}{2}\left(\lambda_{\max} + \sqrt{\lambda_{\max}^2 + 4(\sigma_{\max}^2 + z_{\max})}\right). \tag{6.15}$$

Note that the definition of ζ in (6.12) depends on the eigenvalue θ^- considered, and as such it cannot be computed without some knowledge on the eigenvalue itself. To obtain a computable lower bound on the negative eigenvalues of K_3, following [66] we define

$$\theta_1 := \min\left\{\frac{1}{2}\left(\lambda_{\min} - \sqrt{\lambda_{\min}^2 + 4\sigma_{\max}^2}\right), \min_j \theta_j^*\right\},$$

with θ_j^* defined as in (6.13).

In [66, Section 5.2] Greif et al. observe that the lower bound in Theorem 6.2.2(i) is established excluding that some eigenvalue θ^- of K_3 belong to the spectrum of $-X$ but this assumption may fail both in the course of iterations and in the limit if there are inactive bounds. They also note that the zero upper bound in Theorem 6.2.2(i) is not particularly meaningful either in the case where $(x, z) > 0$ or in the limit. Finally, they point out that the lower bound θ_3 in Theorem 6.2.2(ii) is strictly positive as long as $(x, z) > 0$ but in the limit it reduces to $\lambda_{\min}$ and may be overly pessimistic if $\lambda_{\min} = 0$. In particular, the nonsingularity of K_3 stated in Theorem 6.2.1 is not reflected by this spectral analysis. In the next section we find new bounds for the spectrum of K_3 which clearly improve upon the existing results.

6.3 Spectral estimates

In this section we give new bounds for the eigenvalues of K_3 and distinguish between the matrix arising at a generic IP iteration and the matrix arising asymptotically or in the limit of the IP method. Therefore, first we only assume strict positivity of x and z. Then, we suppose that the assumptions in Theorem 6.2.1 hold and that (x, z) is either a positive vector approaching $(\hat{x}, \hat{z})$ or that it coincides with $(\hat{x}, \hat{z})$.

General IP iterations. For positive x and z we fill the incomplete analysis on the negative eigenvalues of K_3 given in [66]. If the leading block of K_3 is positive definite, an upper bound for the negative eigenvalues can be found in [120, Lemma 2.2], but this analysis does not apply to our case where H is only positive semidefinite. In [61, Proposition 3.2, Proposition 3.3], the authors derive eigenvalue bounds for general saddle point systems with possibly indefinite and also singular (1,1) block. Their results require that the (1,2) block of the saddle point matrix have full rank; this can be satisfied in our setting by a simple reordering of the blocks. However, they also require strong assumptions on the norm of the (2,2) block, which in our numerical experiments (see section 6.4) do not hold except during the very first few

iterations. We begin by investigating the lower bound, and work without assuming the nonsingularity of $X + \theta I_n$.

Theorem 6.3.1. *Suppose that H is symmetric and positive semidefinite, J has full rank and X and Z are diagonal with positive entries. Then the negative eigenvalues θ^- of matrix K_3 given in (6.6) satisfy*

$$\theta^- \geq \theta_1 := \min_{1\leq j\leq n} \theta_j^*, \tag{6.16}$$

where θ_j^* is the smallest negative root of the cubic polynomial (6.13).

Proof. Consider the eigenvalue problem for K_3.

$$\begin{bmatrix} H & J^T & -Z^{\frac{1}{2}} \\ J & 0 & 0 \\ -Z^{\frac{1}{2}} & 0 & -X \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \theta \begin{bmatrix} u \\ v \\ w \end{bmatrix},$$

and the block equations therein,

$$\begin{aligned} Hu + J^T v - Z^{\frac{1}{2}} w &= \theta u, & (6.17)\\ Ju &= \theta v, & (6.18)\\ -Z^{\frac{1}{2}} u - Xw &= \theta w. & (6.19) \end{aligned}$$

Necessarily $u \neq 0$; otherwise (6.18) gives $v = 0$, and by the positive definiteness of Z, (6.17) yields $w = 0$, which is a contradiction. Similarly, w must be nonzero since otherwise, (6.19) implies $u = 0$.

Since $\theta < 0$, the matrix $H - \theta I_n$ is positive definite, thus by eliminating u from (6.17) and substituting in (6.18) and (6.19) we obtain

$$\begin{aligned} -SS^T v + SR\xi &= \theta v, & (6.20)\\ Z^{\frac{1}{2}} RS^T v - Z^{\frac{1}{2}} R^2 \xi &= (\theta I_n + X) Z^{-\frac{1}{2}} \xi, & (6.21) \end{aligned}$$

where $R = (H - \theta I_n)^{-\frac{1}{2}}$, $S = JR$, $w = Z^{-\frac{1}{2}}\xi$. Let us now suppose

$$\theta < \gamma_0 := \frac{1}{2}\left(\lambda_{\min} - \sqrt{\lambda_{\min}^2 + 4\sigma_{\max}^2}\right),$$

otherwise, γ_0 provides the sought after lower bound for θ. Under this assumption, the matrix $SS^T + \theta I_m$ is negative definite and (6.20) gives $v = (SS^T + \theta I_m)^{-1} SR\xi$. Then, substituting v in (6.21) and premultiplying by $\xi^T Z^{-\frac{1}{2}}$ we obtain

$$\xi^T R \left[I_n - S^T(\theta I_m + SS^T)^{-1} S\right] R\xi + \xi^T Z^{-\frac{1}{2}} (\theta I_n + X) Z^{-\frac{1}{2}} \xi = 0. \tag{6.22}$$

In order to bound the left-hand side of the above equation, first note that

$$\xi^T Z^{-\frac{1}{2}}(X+\theta I_n)Z^{-\frac{1}{2}}\xi = \sum_{i=1}^{n} \frac{x_i+\theta}{z_i}\,\xi_i^2 \tag{6.23}$$

where $\ell = \arg(\max_j \frac{x_j+\theta}{z_j})$. We then note that the eigenvalues of $SS^T + \theta I_m$ are of the form $\dfrac{\sigma_i^2(S)}{\sigma_i^2(S)+\theta}$ and by the negative definiteness of the matrix, it follows that $\sigma_i^2 + \theta < 0,\ i = 1, \ldots, n$, and

$$S^T(\theta I_m + SS^T)^{-1}S \geq \frac{\sigma_{\max}^2(S)}{\sigma_{\max}^2(S)+\theta}. \tag{6.24}$$

Using inequalities (6.23) and (6.24), as well as $\xi^T R^2 \xi \leq \lambda_{\max}^2(R)\,\|\xi\|^2$ and $\|\xi\|^2 > 0$, equation (6.22) becomes

$$\frac{\theta\,\lambda_{\max}^2(R)}{\sigma_{\max}^2(S)+\theta} + \frac{x_\ell+\theta}{z_\ell} \geq 0. \tag{6.25}$$

By $\sigma_{\max}^2(S) \leq \dfrac{\sigma_{\max}^2}{\lambda_{\min}-\theta}$ and $\lambda_{\max}^2(R) = \dfrac{1}{\lambda_{\min}-\theta}$,

$$\frac{\theta}{\left(\dfrac{\sigma_{\max}^2}{\lambda_{\min}-\theta}+\theta\right)(\lambda_{\min}-\theta)} + \frac{x_\ell+\theta}{z_\ell} \geq 0,$$

which is equivalent to $p_\ell(\theta) \geq 0$, where $p_\ell(\theta)$ is the polynomial (6.13) for $j = \ell$. The polynomial $p_\ell(\theta)$ has two negative roots since $p_\ell(0) = -x_\ell\sigma_{\max}^2 < 0$ and $p_\ell(-x_\ell) = x_\ell z_\ell > 0$, and the product of the roots is negative. Hence, letting $\theta_\ell^* < -x_\ell$ be the smallest negative root of $p_\ell(\theta)$ it follows $\theta \geq \min\{\theta_\ell^*, \gamma_0\}$. Repeating the same arguments and letting θ_j^* be the smallest negative root of $p_j(\theta)$, $1 \leq j \leq n$, we can conclude that $\theta \geq \min\{\theta_\ell^*, \gamma_0\} \geq \min\left\{\min_j \theta_j^*, \gamma_0\right\}$ and

$$\min_j \theta_j^* < -x_{\max}. \tag{6.26}$$

To complete the proof, it remains to show that $\theta_j^* \leq \gamma_0$ for every $j = 1, \ldots, n$. To this end, for any j we write $p_j(\theta)$ as

$$p_j(\theta) = (\theta + x_j)(\theta^2 - \lambda_{\min}\theta - \sigma_{\max}^2) - z_j\theta,$$

and note that $p_j(\gamma_0) = -z_j\gamma_0 > 0$, i.e. the smallest root of $p_j(\theta)$ is smaller than γ_0. □

In (6.16) it is possible to remove the dependence on j by noting that if $\theta < -x_{\max}$ then (6.26) implies

$$\frac{x_j + \theta}{z_j} \leq \frac{x_{\max} + \theta}{z_{\max}}.$$

Using this inequality in (6.25) and repeating the same arguments as in Theorem 6.3.1, we find that the negative eigenvalues of K_3 are bounded from below by the smallest negative root θ^* of the cubic equation

$$\theta^3 + (x_{\max} - \lambda_{\min})\theta^2 - (x_{\max}\lambda_{\min} + \sigma_{\max}^2 + z_{\max})\theta - \sigma_{\max}^2 x_{\max} = 0.$$

However, this bound may be not as sharp as (6.16) as it is unlikely that $x_j = x_{\max}$ and $z_j = z_{\max}$ for the same index j.

We now turn on the problem of deriving an upper bound for the negative eigenvalues. If the leading block of K_3 is positive definite, one such bound can be found in [120, Lemma 2.2], but this analysis does not apply to our case where H is only positive semidefinite. In [61, Proposition 3.2, Proposition 3.3], the authors derive eigenvalue bounds for general saddle point systems with possibly indefinite and also singular (1,1) block. Their results require that the (1,2) block of the saddle point matrix have full rank; this can be satisfied in our setting by a simple reordering of the blocks. However, they also require strong assumptions on the norm of the (2,2) block, which in our numerical experiments (see section 6.4) do not hold except during the very first few iterations. In the following theorem we determine an upper bound for the negative eigenvalues of K_3 under weaker hypotheses, by exploiting the structure of the blocks.

Theorem 6.3.2. *Suppose that H is symmetric and positive semidefinite, J has full rank and X and Z have positive diagonal entries. Then the negative eigenvalues θ^- of K_3 given in (6.6) satisfy*

$$\theta^- \leq \theta_2 = \gamma, \tag{6.27}$$

where γ is the largest negative root of the cubic polynomial

$$p(\theta) = \theta^3 + (x_{\min} - \lambda_{\max})\theta^2 - (x_{\min}\lambda_{\max} + \sigma_{\min}^2 + z_{\max})\theta - \sigma_{\min}^2 x_{\min}, \tag{6.28}$$

with $\gamma > -x_{\min}$.

Proof. Consider equations (6.17), (6.18), (6.19), with $\theta < 0$. As before, we have $u, w \neq 0$. We first assume that $u \in \ker(J)$. From (6.18) we infer that $v = 0$. The first and second equations now read

$$Hu - Z^{\frac{1}{2}}w = \theta u, \qquad -Z^{\frac{1}{2}}u - Xw = \theta w.$$

If we determine u from the first equation above, substitute it in the second one, and multiply the resulting equation from the left by w^T, we obtain

$$w^T Z^{\frac{1}{2}}(H-\theta I_n)^{-1}Z^{\frac{1}{2}}w + w^T X w + \theta \|w\|^2 = 0.$$

Thus, using Rayleigh quotient arguments, we obtain $z_{\min}/(\lambda_{\max}-\theta)+x_{\min}+\theta \leq 0$, and

$$\theta \leq \gamma_1 := \frac{1}{2}\left(\lambda_{\max} - x_{\min} - \sqrt{(\lambda_{\max}+x_{\min})^2 + 4z_{\min}}\right). \tag{6.29}$$

We now suppose $u \notin \ker(J)$, and write $u = u_1 + u_2$, with $u_1 \in \ker(J)$ and $0 \neq u_2 \in \ker(J)^{\perp}$. Moreover, we suppose $\theta > -x_{\min}$ (otherwise, $-x_{\min}$ is the sought after upper bound), so that the matrix $X + \theta I_n$ is also positive definite. From (6.18) and (6.19) we respectively obtain

$$v = \frac{1}{\theta}Ju, \qquad w = -(X+\theta I_n)^{-1}Z^{\frac{1}{2}}u.$$

If we substitute in (6.17) and premultiply it by u_1^T and u_2^T, we respectively obtain:

$$\begin{aligned} u_1^T H(u_1+u_2) + u_1^T Z^{\frac{1}{2}}\left(X+\theta I_n\right)^{-1} Z^{\frac{1}{2}}(u_1+u_2) - \theta\|u_1\|^2 &= 0, \\ u_2^T H(u_1+u_2) + \frac{1}{\theta}\|Ju_2\|^2 + u_2^T Z^{\frac{1}{2}}\left(X+\theta I_n\right)^{-1} Z^{\frac{1}{2}}(u_1+u_2) - \theta\|u_2\|^2 &= 0. \end{aligned}$$

Subtracting the two equations,

$$\begin{aligned} u_2^T H u_2 - u_1 H u_1 + \frac{1}{\theta}\|Ju_2\|^2 + u_2^T Z^{\frac{1}{2}}\left(X+\theta I_n\right)^{-1} Z^{\frac{1}{2}}u_2 + \\ -u_1^T Z^{\frac{1}{2}}\left(X+\theta I_n\right)^{-1} Z^{\frac{1}{2}}u_1 - \theta\|u_2\|^2 + \theta\|u_1\|^2 &= 0. \end{aligned}$$

Since $-u_1^T H u_1$, $-u_1^T Z^{\frac{1}{2}}\left(X+\theta I_n\right)^{-1} Z^{\frac{1}{2}}u_1$ and $\theta\|u_1\|^2$ are nonpositive, it holds

$$u_2^T\left(H + \frac{1}{\theta}J^T J + Z^{\frac{1}{2}}\left(X+\theta I_n\right)^{-1} Z^{\frac{1}{2}} - \theta I_n\right)u_2 \geq 0,$$

from which we obtain

$$\left(\lambda_{\max} + \frac{\sigma_{\min}^2}{\theta} + \frac{z_{\max}}{x_{\min}+\theta} - \theta\right)\|u_2\|^2 \geq 0.$$

Dividing by $\|u_2\|^2$ and multiplying by $-\theta(\theta+x_{\min})$, we find that θ satisfies $p(\theta) \geq 0$ where $p(\theta)$ is the cubic polynomial in (6.28). Noting that $p(0) = -\sigma_{\min}^2 x_{\min} < 0$ and $p(-x_{\min}) = z_{\max}x_{\min} > 0$, it follows that $\theta \leq \gamma$, where γ is the largest negative root of $p(\theta)$, and $\gamma > -x_{\min}$. By (6.29) and $\gamma_1 < -x_{\min} < \gamma$, we can conclude that $\theta \leq \max\{\gamma_1, \gamma\} = \gamma$. □

Combining the above results with the bounds for the positive eigenvalues given in Theorem 6.2.2, we obtain

$$\operatorname{spec}(K_3) \subseteq I^- \cup I^+ = [\theta_1,\ \theta_2] \cup [\theta_3,\ \theta_4], \tag{6.30}$$

where θ_1 is as in (6.16), and θ_2 is as in (6.27). The new estimate in Theorem 6.3.2 provides a significant upper bound for the negative eigenvalues, as compared with the analysis given in [66]; see also Theorem 6.2.2.

The estimate θ_2 in Theorem 6.3.2 goes to zero with $x_{\min}$. This fact can also be appreciated by writing the polynomial $p(\theta)$ in the theorem statement as

$$p(\theta) = (\theta + x_{\min})(\theta^2 - \lambda_{\max}\theta - \sigma_{\min}^2 - z_{\max}) + z_{\max}x_{\min},$$

so that $p(\theta)$ differs by $z_{\max}x_{\min}$ from a polynomial having $-x_{\min}$ as one of its roots. Such a property shows the inadequacy of this technique to derive spectral estimates at later stages of the IP iterations, when the coefficient matrix remains fairly well conditioned.

Asymptotic IP iterations and limit point. The bounds in (6.30) are meaningful as long as (x, z) are either early or middle stage iterates of the IP method, or (x, z) are late stage iterates and K_3 tends to singularity. However, if (x, z) approaches a solution $(\hat{x}, \hat{z})$ satisfying the conditions in Theorem 6.2.1, then the bounds are unsatisfactory. Indeed, $\hat{K}_3$ is nonsingular whereas the upper negative eigenvalue θ_2 tends to 0 as $x_{\min}$ tends to 0, and so does the lower bound θ_3 on the positive eigenvalues if $\lambda_{\min} = 0$. We thus make a further step and focus on the case when $\hat{K}_3$ is nonsingular. It is therefore useful to analyze the assumptions made in Theorem 6.2.1. Considering the partitioning in (6.9) we can write

$$J = \begin{bmatrix} J_{\mathcal{A}_*} & J_{\mathcal{I}_*} \end{bmatrix}, \qquad \begin{bmatrix} J^T & -I_{\mathcal{A}_*} \end{bmatrix} = \begin{bmatrix} J_{\mathcal{A}_*}^T & -I_q \\ J_{\mathcal{I}_*}^T & 0 \end{bmatrix},$$

with $J_{\mathcal{A}_*} \in \mathbb{R}^{m\times q}$ and $J_{\mathcal{I}_*} \in \mathbb{R}^{m\times(n-q)}$. The LICQ condition is satisfied at $\hat{x}$ if and only if $J_{\mathcal{I}_*}^T$ has full column rank. This fact implies that $J_{\mathcal{I}_*}$ is a large or square matrix, i.e. $q \le n - m$, and that $\sigma_{\min}(J_{\mathcal{I}_*}) > 0$.

Concerning condition (6.11), we have that

$$\ker(\widehat{Z}) = \left\{(0, y) \in \mathbb{R}^n \mid y \in \mathbb{R}^{(n-q)}\right\},$$

and the vectors of $\ker(J) \cap \ker(\widehat{Z})$ are of the form $(0, y)$ with $y \in \ker(J_{\mathcal{I}_*})$. If $q = n - m$ then $J_{\mathcal{I}_*}$ is square and $\ker(J_{\mathcal{I}_*}) = \{0\}$; thus $\ker(J) \cap \ker(\widehat{Z}) = \{0\}$ and (6.11) is met. Otherwise, if $q < n-m$, then $\ker(J) \cap \ker(\widehat{Z})$ is a nontrivial

subspace and condition (6.11) is equivalent to

$$\min_{0\neq x\in\ker(J)\cap\ker(\widehat{Z})} \frac{x^T H x}{x^T x} = \lambda^* > 0. \tag{6.31}$$

Using the above properties, we prove nontrivial and sharp bounds for K_3 in the late stage of the IP method and for $\hat{K}_3$. To this end, the following technical lemma is needed. It provides bounds for the singular values of B, which will be used for later estimates; its proof is postponed to Appendix A.

Lemma 6.3.3. *Suppose that $\hat{x}$ and $\hat{z}$ are strictly complementary, and $\mathcal{A}_*$ and $\mathcal{I}_*$ are the index sets of active and inactive bounds at $\hat{x}$ defined in (6.8). Further, suppose that $\hat{x}$ and $\hat{z}$ are partitioned as in (6.9), the LICQ condition is satisfied at $\hat{x}$, and (6.11) holds. Let $Z_{\mathcal{A}_*} \in \mathbb{R}^{q\times q}$ be a diagonal positive definite matrix and*

$$B = \begin{bmatrix} J_{\mathcal{A}_*} & J_{\mathcal{I}_*} \\ -Z_{\mathcal{A}_*}^{\frac{1}{2}} & 0 \end{bmatrix}. \tag{6.32}$$

Then

$$\begin{aligned} \sigma^2_{\min}(B) &\geq \frac{1}{2}\left(\chi - \sqrt{\chi^2 - 4\sigma^2_{\min}(J_{\mathcal{I}_*})(z_{\mathcal{A}_*})_{\min}}\right), \\ \sigma^2_{\max}(B) &\leq \frac{1}{2}\left((z_{\mathcal{A}_*})_{\max} + \sigma^2_{\max} + \sqrt{((z_{\mathcal{A}_*})_{\max} - \sigma^2_{\max})^2 + 4(z_{\mathcal{A}_*})_{\max}\sigma^2_{\max}(J_{\mathcal{A}_*})}\right) \\ &\leq \sigma^2_{\max} + (z_{\mathcal{A}_*})_{\max}, \end{aligned}$$

with $\chi = \sigma^2_{\max}(J_{\mathcal{A}_*}) + \sigma^2_{\min}(J_{\mathcal{I}_*}) + (z_{\mathcal{A}_*})_{\min}$.

The following theorem provides bounds for all eigenvalues of K_3 under the stated assumptions; these bounds are based on perturbation theory results for symmetric matrices and on estimates in [61, 78].

Theorem 6.3.4. *Let H be symmetric and positive semidefinite with nontrivial null space, $\hat{x}$ and $\hat{z}$ strictly complementary, $\mathcal{A}_*$ and $\mathcal{I}_*$ be the index sets of active and inactive bounds at $\hat{x}$ defined in (6.8). Further, suppose that the cardinality of $\mathcal{A}_*$ is equal to q, $\hat{x}$ and $\hat{z}$ are partitioned as in (6.9), the LICQ condition is satisfied at $\hat{x}$, and condition (6.11) holds. Let x and z be sufficiently close to $\hat{x}$ and $\hat{z}$ and be such that $x = (x_{\mathcal{A}_*},\ x_{\mathcal{I}_*})$, $z = (z_{\mathcal{A}_*},\ z_{\mathcal{I}_*})$ with $x_{\mathcal{A}_*} \geq 0$, $x_{\mathcal{I}_*} > 0$, $z_{\mathcal{A}_*} > 0$, $z_{\mathcal{I}_*} \geq 0$. Then*

$$\operatorname{spec}(K_3) \subseteq [\mu_1,\ \mu_2] \cup [\mu_3,\ \mu_4],$$

where $\mu_1, \mu_2 < 0$ *and* μ_3, $\mu_4 > 0$ *are given by*

$$\begin{aligned}
\mu_1 &= \min\left\{-(x_{\mathcal{I}_*})_{\max},\ \frac{1}{2}\left(\lambda_{\min} - \sqrt{\lambda_{\min}^2 + 4\sigma_{\max}^2(B)}\right)\right\} \\
&\quad - \max\left\{(x_{\mathcal{A}_*})_{\max},\ \sqrt{(z_{\mathcal{I}_*})_{\max}}\right\}, \\
\mu_2 &= \max\left\{-(x_{\mathcal{I}_*})_{\min},\ \frac{1}{2}\left(\lambda_{\max} - \sqrt{\lambda_{\max}^2 + 4\sigma_{\min}^2(B)}\right)\right\} + \sqrt{(z_{\mathcal{I}_*})_{\max}}, \\
\mu_3 &= \mu_3^* - (x_{\mathcal{A}_*})_{\max}, \\
\mu_4 &= \frac{1}{2}\left(\lambda_{\max} + \sqrt{\lambda_{\max}^2 + 4\sigma_{\max}^2(B)}\right) + \sqrt{(z_{\mathcal{I}_*})_{\max}}.
\end{aligned}$$

If $q < n - m$, *the scalar* μ_3^* *is the smallest positive root of the cubic equation*

$$\mu^3 - \lambda_{\max}\mu^2 - \sigma_{\min}^2(B)\mu + \lambda^*\sigma_{\min}^2(B) = 0,$$

where λ^* *is defined as in* (6.31). *If* $q = n - m$ *we have instead*

$$\mu_3^* = \frac{1}{2}\left(\lambda_{\min} + \sqrt{\lambda_{\min}^2 + 4\sigma_{\min}^2(B)}\right).$$

Proof. We write K_3 in extended form

$$K_3 = \begin{bmatrix} & H & & J_{\mathcal{A}_*}^T & -Z_{\mathcal{A}_*}^{\frac{1}{2}} & 0 \\ & & & J_{\mathcal{I}_*}^T & 0 & -Z_{\mathcal{I}_*}^{\frac{1}{2}} \\ J_{\mathcal{A}_*} & & J_{\mathcal{I}_*} & 0 & 0 & 0 \\ -Z_{\mathcal{A}_*}^{\frac{1}{2}} & & 0 & 0 & -X_{\mathcal{A}_*} & 0 \\ 0 & & -Z_{\mathcal{I}_*}^{\frac{1}{2}} & 0 & 0 & -X_{\mathcal{I}_*} \end{bmatrix},$$

with $X_{\mathcal{A}_*} = \operatorname{diag}(x_{\mathcal{A}_*}) \in \mathbb{R}^{q\times q}$, $X_{\mathcal{I}_*} = \operatorname{diag}(x_{\mathcal{I}_*}) \in \mathbb{R}^{n-q\times n-q}$, $Z_{\mathcal{A}_*} = \operatorname{diag}(z_{\mathcal{A}_*}) \in \mathbb{R}^{q\times q}$, $Z_{\mathcal{I}_*} = \operatorname{diag}(z_{\mathcal{I}_*}) \in \mathbb{R}^{n-q\times n-q}$, and observe that

$$\begin{aligned}
K_3 &= \widetilde{K}_3 + \Delta_K \qquad (6.33)\\
&= \begin{bmatrix} & H & & J_{\mathcal{A}_*}^T & -Z_{\mathcal{A}_*}^{\frac{1}{2}} & 0 \\ & & & J_{\mathcal{I}_*}^T & 0 & 0 \\ J_{\mathcal{A}_*} & & J_{\mathcal{I}_*} & 0 & 0 & 0 \\ -Z_{\mathcal{A}_*}^{\frac{1}{2}} & & 0 & 0 & 0 & 0 \\ 0 & & 0 & 0 & 0 & -X_{\mathcal{I}_*} \end{bmatrix} + \begin{bmatrix} & 0 & & 0 & 0 & 0 \\ & & & 0 & 0 & -Z_{\mathcal{I}_*}^{\frac{1}{2}} \\ 0 & & 0 & 0 & 0 & 0 \\ 0 & & 0 & 0 & -X_{\mathcal{A}_*} & 0 \\ 0 & & -Z_{\mathcal{I}_*}^{\frac{1}{2}} & 0 & 0 & 0 \end{bmatrix}.
\end{aligned}$$

Standard perturbation arguments for symmetric matrices ensure that (see, e.g., [57, Theorem 8.1.5])

$$\lambda_i(\widetilde{K}_3) + \lambda_{\min}(\Delta_K) \le \theta \le \lambda_i(\widetilde{K}_3) + \lambda_{\max}(\Delta_K), \quad i = 1, \dots 2n + m. \quad (6.34)$$

Thus, estimates for θ can be derived from spectral information on $\widetilde{K}_3$ and Δ_K, where

$$\lambda_{\min}(\Delta_K) = -\max\left\{(x_{\mathcal{A}_*})_{\max},\ \sqrt{(z_{\mathcal{I}_*})_{\max}}\right\}, \quad \lambda_{\max}(\Delta_K) = \sqrt{(z_{\mathcal{I}_*})_{\max}}.$$

As for $\widetilde{K}_3$, we have that $\sigma\left(\widetilde{K}_3\right) = \sigma\left(-X_{\mathcal{I}_*}\right) \cup \sigma\left(\check{K}\right)$, where $\check{K}$ is the saddle point matrix

$$\check{K} = \begin{bmatrix} H & B^T \\ B & 0 \end{bmatrix},$$

with B given in (6.32). By Lemma 6.3.3 we know that B^T has full column rank. Moreover, $\ker(B) = \ker(J) \cap \ker(\widehat{Z})$, and by (6.11) either $\ker(B) = \{0\}$ (if B is square, i.e. $q = n - m$) or H is positive definite on $\ker(B)$. Thus, $\check{K}$ satisfies the hypothesis of [61, Proposition 2.2], and the expressions of $\mu_1,\ \mu_2$ and μ_4 are a direct consequence of that result.

A slightly different approach is needed to obtain μ_3. We consider the principal submatrix $\bar{K}$ of K_3 obtained by taking its first $n + m + q$ rows and columns, i.e.

$$\bar{K} = \begin{bmatrix} & H & & J_{\mathcal{A}_*}^T & -Z_{\mathcal{A}_*}^{\frac{1}{2}} \\ & & & J_{\mathcal{I}_*}^T & 0 \\ J_{\mathcal{A}_*} & & J_{\mathcal{I}_*} & 0 & 0 \\ -Z_{\mathcal{A}_*}^{\frac{1}{2}} & & 0 & 0 & -X_{\mathcal{A}_*} \end{bmatrix}.$$

It holds that K_3 has n positive and $n + m$ negative eigenvalues, and $\bar{K}$ has n positive eigenvalues and $m + q$ negative ones [66, Lemma 3.8]. Using interlacing properties of the eigenvalues and again the standard perturbation bounds for symmetric matrices, we infer

$$\lambda_{\min}^+(K_3) \geq \lambda_{\min}^+(\bar{K}) \geq \lambda_{\min}^+(\check{K}) - (x_{\mathcal{A}_*})_{\max},$$

where the symbol $\lambda_{\min}^+(\cdot)$ indicates the smallest positive eigenvalue of a matrix. If $q = n - m$ we can use again [61, Proposition 2.2] to obtain the expression of μ_3^*. If $q < n - m$, since we supposed H singular we can instead use the lower bound of the positive eigenvalues of $\check{K}$ given in [78, Theorem 2] to obtain the final result. □

It is interesting to observe that, whenever x and z are sufficiently close to $\hat{x}$ and $\hat{z}$, then $(x_{\mathcal{A}_*})_{\max}$ and $(z_{\mathcal{I}_*})_{\max}$ are small enough to guarantee that the intervals $[\mu_1,\ \mu_2]$ and $[\mu_3,\ \mu_4]$ are nontrivial, i.e., μ_2 is strictly negative and μ_3 is strictly positive.

Theorem 6.3.4 covers both the case where (x, z) is strictly positive and close enough to $(\hat{x}, \hat{z})$, and the case where $(x, z) = (\hat{x}, \hat{z})$. Thus, these bounds

are valid for the matrices K_3 occurring in the late stage of the IP method, and also for $\hat{K}_3$. The proof of this theorem relies on the perturbation theory for symmetric eigenvalue problems, and involves $\widetilde{K}_3$ and the scalars $(x_{\mathcal{A}_*})_{\max}$ and $(z_{\mathcal{I}_*})_{\max}$ which approach zero when (x, z) tends to $(\hat{x}, \hat{z})$. Hence, the smaller $(x_{\mathcal{A}_*})_{\max}$ and $(z_{\mathcal{I}_*})_{\max}$, the closer $\mu_1,\ \mu_2,\ \mu_3,\ \mu_4$ are to the spectral bounds for $\widetilde{K}_3$, for which bounds are available [61, 78].

Remark 3. In Theorem 6.3.4, for the case $q < n - m$, the value of μ_3^* relies on results from [78] and it holds for H singular. If H is nonsingular and $q < n - m$ then it holds that $\mu_3^* = \max\{\lambda_{\min}, \gamma\}$, where γ is the smallest positive root of the cubic polynomial $p_3(\mu) = \mu^3 - (\lambda_{\min} + \lambda^*)\,\mu^2 + (\lambda_{\min}\lambda^* - \lambda_{\max}^2 - \sigma_{\min}(B)^2)\,\mu + \lambda^*\sigma_{\min}^2(B)$. This follows from applying [61, Proposition 2.2] and [113, Lemma 2.1] to the matrix $\check{K}$ in the proof of Theorem 6.3.4. We also emphasize that all other bounds given by Theorem 6.3.4 still hold when H is positive definite.

6.4 Validation of the spectral bounds for K_3

We analyze the quality of our bounds by first using two examples with small matrices. In the first case, H is positive definite, while in the second H is only positive semidefinite.

Example 6.1. Given positive scalars $\lambda,\ \sigma,\ \rho$, let

$$H = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \quad J^T = \begin{bmatrix} 0 \\ \sigma \end{bmatrix},\ x = \begin{bmatrix} 0 \\ \rho \end{bmatrix},\ z = \begin{bmatrix} \sigma \\ 0 \end{bmatrix},\ \text{so that } B = \begin{bmatrix} 0 & \sigma \\ -\sigma & 0 \end{bmatrix}.$$

The characteristic polynomial of K_3 is given by $\pi(\theta) = (\theta+\rho)\,(\theta^2 - \lambda\theta - \sigma^2)^2$. The eigenvalues of K_3 are $-\rho$, $\frac{1}{2}\left(\lambda - \sqrt{\lambda^2 + 4\sigma^2}\right)$, $\frac{1}{2}\left(\lambda + \sqrt{\lambda^2 + 4\sigma^2}\right)$, and the bounds in Theorem 6.3.4 are sharp (note that $q = n - m$).

In this second example, we have $q < n - m$ and show that the estimates μ_1 and μ_3 can be sharp.

Example 6.2. Given positive scalars $\lambda_{\max},\ \lambda^*,\ \sigma,\ x_{\min}\ x_{\max},\ \lambda_{\max} > \lambda^*$, let

$$H = \begin{bmatrix} \lambda_{\max} - \lambda^* & \sqrt{\lambda^*(\lambda_{\max} - \lambda^*)} & 0 \\ \sqrt{\lambda^*(\lambda_{\max} - \lambda^*)} & \lambda^* & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad J^T = \begin{bmatrix} 0 \\ 0 \\ \sigma \end{bmatrix},$$

$$z = \begin{bmatrix} \sigma \\ 0 \\ 0 \end{bmatrix}, \quad x = \begin{bmatrix} 0 \\ x_{\min} \\ x_{\max} \end{bmatrix},\ \text{so that } B = \begin{bmatrix} 0 & 0 & \sigma \\ -\sigma & 0 & 0 \end{bmatrix}.$$

The characteristic polynomial of K_3 is

$$\pi(\theta) = (\theta + x_{\min})(\theta + x_{\max})(\sigma^2 - \theta^2)(\theta^3 - \lambda_{\max}\theta^2 - \sigma^2\theta + \lambda^*\sigma^2).$$

Since $\lambda_{\min} = 0$ and $q < n - m$, the bounds μ_1 and μ_3 are sharp.

6.4.1 Numerical validation: regularized LP problems

We then proceed by analyzing the quality of our spectral estimates on benchmark Linear Programming (LP) problems with primal regularization:

$$\min_{x\in\mathbb{R}^n} c^T x + \rho \|x\|^2, \qquad \text{subject to} \quad Jx = b, \quad x \geq 0,$$

where $n = 185$, $m = 129$, $J \in \mathbb{R}^{m\times n}$ is the matrix in LPnetlib/lp_scagr7 [131] with full row rank, b and c are fixed so that the $\hat{x} = (0, 1_{n-q})$ and $\hat{z} = (1_q, 0)$ are exact primal and dual solutions, $\rho \geq 0$ is a regularization parameter. The value of q is varied and it may affect the fulfillment of the LICQ condition at $\hat{x}$. It is apparent that regularized LP problems are just QP problem with $H = \rho I_n$. Note, moreover, that this is the same setting already considered for the numerical experiments presented in Chapter 3.

The problems were solved with the PDCO solver [109] and sequence of iterates approaching $\hat{x}$ and $\hat{z}$ were computed and stored. Then, for each iterate we formed matrix K_3, letting $\rho = 10^{-6}$. The eigenvalues of the resulting matrices were computed and compared with the bounds given in (6.30) and in Theorem 6.3.4 with μ_3^* as in Remark 3. Regarding the actual computation of the bounds, singular values $\sigma_{\min}(J)$, $\sigma_{\max}(J)$, $\sigma_{\min}(B)$ and $\sigma_{\max}(B)$ were computed while $\lambda_{\min} = \lambda_{\max} = \lambda^* = \rho$.

In our numerical experiments, bounds θ_1 and θ_4 are very similar to μ_1 and μ_4, therefore we do show them in the plots. On the other hand, μ_3, μ_2 were plotted only when meaningful, that is only when appearing with positive and negative sign, respectively.

We start by reporting on the accuracy of θ_2 in Theorem 6.3.2. For this purpose we set $q = n - m$, which makes the matrix $J_{\mathcal{I}_*}$, and thus B, rank deficient. The absolute value of the largest negative eigenvalue $\lambda_{n+m}(K_3)$ (solid line), and its bound θ_2 are displayed in Figure 6.1, showing that the estimate matches quite well the true eigenvalue of K_3.

Let $\mathrm{fix}(\cdot)$ be the function that rounds its argument to the nearest integer towards zero. We then set $q = \mathrm{fix}((n - m)/2)$ so that the assumptions of Theorem 6.3.4 hold; indeed, with this choice the matrix $J_{\mathcal{I}_*}$, and thus B, have full rank. In the left plot of Figure 6.2, for each iterate on the x-axis, the minimum positive eigenvalue $\lambda_{n+m+1}(K_3)$, and its bounds θ_3 and μ_3 are

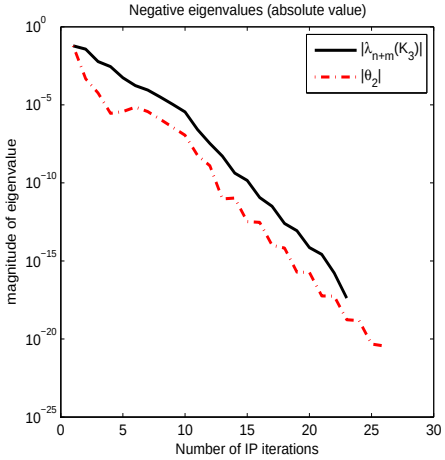

Figure 6.1: Negative eigenvalue of K_3 closest to zero (solid line) and its bound at every iteration, $q = n - m$.

displayed; θ_3 is a good lower bound and μ_3 is sharp as well during the later ones. Similarly, in the right plot of Figure 6.2 we report the absolute value of the negative eigenvalue $\lambda_{n+m}(K_3)$ closest to zero, along with the bounds θ_2 and μ_2. As expected, μ_2 is sharp during the final iterations, unlike θ_2.

It is worth testing the validity of μ_3 when H is semidefinite. For this reason, let us consider a QP problem where J is the same matrix as before and let the orthonormal columns of V span $\ker(J)$. Then by taking H in the form

$$H = \begin{bmatrix} V & Q \end{bmatrix} \begin{bmatrix} \rho I_{n-m} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V^T \\ Q^T \end{bmatrix}, \tag{6.35}$$

where ρ is positive and $[V\, Q]$ is an orthogonal matrix, we ensure that $\lambda^* = \lambda_{\max}$. It holds that $\lambda_{\min} = 0$ and $\lambda_{\max} = \rho$. Finally, the QP problem is built setting $\rho = 1$, so that $\hat{x} = (0, 1_{n-q})$ and $\hat{z} = (1_q, 0)$ are exact primal and dual solutions with $q = \text{fix}((n-m)/2)$.

The QP problem was solved with PDCO and a sequence of iterates approaching $\hat{x}, \hat{z}$ was formed. Figure 6.3 displays the positive eigenvalue $\lambda_{n+m+1}(K_3)$ of the matrices with the bounds θ_3, and μ_3, as the iterations proceed. Since $\lambda^* = \lambda_{\max}$ and $\mu_3^* = \min\{\sigma_{\min}(B), \rho\}$, during the last iterations of the interior point method, μ_3 gets close to μ_3^* and is sharp whereas eventually θ_3 is not representative of the minimum eigenvalue.

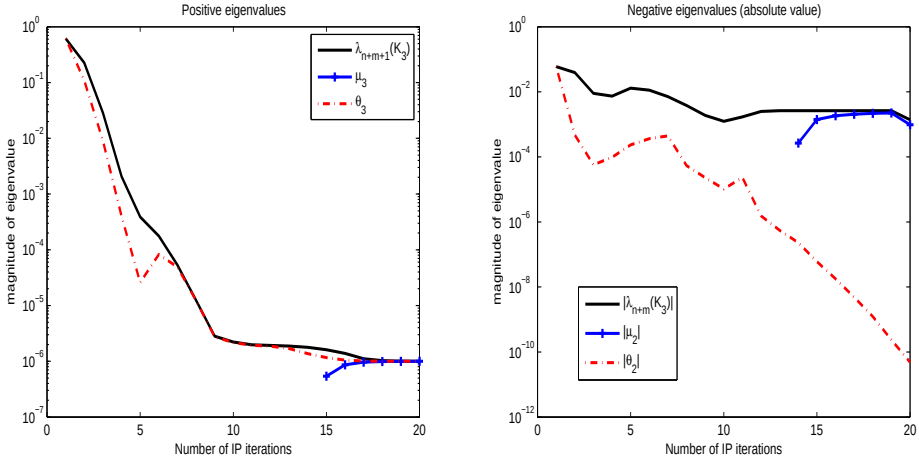

Figure 6.2: Eigenvalues of K_3 closest to zero (solid line) and their bounds at every iteration, $q = \text{fix}((n-m)/2)$ and H nonsingular. Left: positive eigenvalues. Right: negative eigenvalues.

6.4.2 Numerical validation: QP problems

We now extend the numerical validation of our estimates to more realistic QP problems. Here we consider problems taken from the Maros and Meszaros collection [87], which we report below together with their dimensions

(1) STCQP2 $n = 4097$ $m = 2052$
(2) AU3D $n = 3873$ $m = 1000$
(3) CONT-050 $n = 2597$ $m = 2401$

Note that all these problems have larger dimension than the ones considered in Section 6.4.1. In all the chosen problem, the matrix B is nonsingular, in late iterations.

A crucial issue when applying the results of Theorem 6.3.4 to practical computations is the choice of the two sets $\mathcal{A}_*$ and $\mathcal{I}_*$; indeed, the values of $x_{\mathcal{A}_*}$ should be well-separated from the values of $x_{\mathcal{I}_*}$, and this should hold also for $z_{\mathcal{A}_*}$ and $z_{\mathcal{I}_*}$. Motivated by this, we choose $\mathcal{A}_*$ and $\mathcal{I}_*$ so that the distance between the two sets of values $x_{\mathcal{A}_*}$ and $x_{\mathcal{I}_*}$ is maximized.

The three problems where were chosen so that, for the above of $\mathcal{A}_*$ and $\mathcal{I}_*$, the matrix B is nonsingular in late iterations.

We also found that scaling the problem has some effect on the quality of our bounds. More precisely, by considering the scaled variables $\tilde{x} = \dfrac{x}{s_x}$,

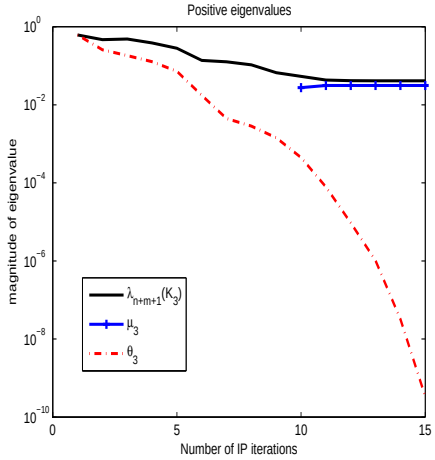

Figure 6.3: Minimum positive eigenvalue of K_3 and its bounds for H singular as in (6.35), $\rho = 1$, $q = \text{fix}((n-m)/2)$.

$\tilde{z} = \dfrac{z}{s_z}$, where s_x and s_z are positive parameters, the linear system

$$\begin{bmatrix} H & J^T & -I \\ J & 0 & 0 \\ -Z & 0 & -X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} = \begin{bmatrix} -Hx + J^T y + z + c \\ -Jx + b \\ -XZe - \tau e \end{bmatrix} \tag{6.36}$$

is replaced with the equivalent one

$$\begin{bmatrix} \widetilde{H} & J^T & -I \\ J & 0 & 0 \\ -\widetilde{Z} & 0 & -\widetilde{X} \end{bmatrix} \begin{bmatrix} \Delta\tilde{x} \\ \Delta\tilde{y} \\ \Delta\tilde{z} \end{bmatrix} = \begin{bmatrix} -\widetilde{H}\tilde{x} + J^T\tilde{y} + \tilde{z} + \tilde{c} \\ -J\tilde{x} + \tilde{b} \\ -\widetilde{X}\widetilde{Z}e - \tilde{\tau}e \end{bmatrix}, \tag{6.37}$$

where $\widetilde{H} = \dfrac{s_x}{s_z}H$, $\tilde{c} = \dfrac{c}{s_z}$, $\tilde{b} = \dfrac{b}{s_x}$, $\widetilde{X} = \text{diag}(\tilde{x})$, $\widetilde{Z} = \text{diag}(\tilde{z})$. We mention that if the unreduced formulation is considered, a scaling of the variables may have a great impact on the conditioning of the system matrix; indeed, s_z and s_x are often chosen to reduce the condition number of the (1,1) block of K_2. This, however, does not apply here, since we are working with the unreduced formulation.

To explain why scaling may effect the bounds consider the expression of μ_2 from Theorem 6.3.4. This bound will be meaningful only if

$$\sqrt{(z_{\mathcal{I}_*})_{\max}} \leq \min\left\{(x_{\mathcal{I}_*})_{\min},\ \frac{1}{2}\left(\sqrt{\lambda_{\max}^2 + 4\sigma_{\min}^2(B)} - \lambda_{\max}\right)\right\} \tag{6.38}$$

, which in particular means that $\sqrt{(z_{\mathcal{I}_*})_{\max}} \leq (x_{\mathcal{I}_*})_{\min}$. But if the variables are scaled, this relation is substituted by:

$$\sqrt{(\tilde{z}_{\mathcal{I}_*})_{\max}} \leq (\tilde{x}_{\mathcal{I}_*})_{\min} \quad \Longleftrightarrow \quad \sqrt{(z_{\mathcal{I}_*})_{\max}} \leq \frac{\sqrt{s_z}}{s_x}(x_{\mathcal{I}_*})_{\min}.$$

It is apparent that the greater the ratio $\frac{\sqrt{s_z}}{s_x}$, the more likely this relation will be satisfied. A similar observation applies also to the other term appearing in (6.38). Indeed, $\sqrt{\lambda_{\max}^2 + 4\sigma_{\min}^2(B)} - \lambda_{\max}$ is a monotonic decreasing function of $\lambda_{\max}$ and a monotonic increasing function of $\sigma_{\max}(B)$, and it holds

$$\lambda_{\max}(\widetilde{H}) = \frac{s_x}{s_z}\lambda_{\max}(H) < \lambda_{\max}(H),$$

if $\frac{s_z}{s_x} > 1$, and

$$\sigma_{\min}\left(\begin{bmatrix} J_{\mathcal{A}_*} & J_{\mathcal{I}_*} \\ \widetilde{Z}_{\mathcal{A}_*} & 0 \end{bmatrix}\right) = \sigma_{\min}\left(\begin{bmatrix} J_{\mathcal{A}_*} & J_{\mathcal{I}_*} \\ s_Z Z_{\mathcal{A}_*} & 0 \end{bmatrix}\right) \geq \sigma_{\min}\left(\begin{bmatrix} J_{\mathcal{A}_*} & J_{\mathcal{I}_*} \\ Z_{\mathcal{A}_*} & 0 \end{bmatrix}\right),$$

if $s_z \geq 1$.

On the other hand, a too small value of s_x may cause the lower bound on the positive eigenvalues to become meaningless, as the value of μ^* is perturbed by the quantity $(\tilde{x}_{\mathcal{A}_*})_{\max} = \frac{1}{s_x}(x_{\mathcal{A}_*})_{\max}$. However, we recall that μ_3 is useful only when θ_3 is not representative of $\lambda_{n+m+1}(K_3)$.

Below we report the scaling we used for each problem, along with the tolerances on complementarity `OptTol` (indeed, a necessary condition for PDCO to stop is that $\max_i x_i z_i \leq$ `OptTol`) that were set. We emphasize that these tolerances apply to the original variables, and not to the scaled ones.

(1) STCQP2	$s_x = 10^{-1}$	$s_z = 10^3$	$\text{tol} = 10^{-6}$
(2) AUG3D	$s_x = 10^{-2}$	$s_z = 10^2$	$\text{tol} = 10^{-6}$
(3) CONT-050	$s_x = 1$	$s_z = 1$	$\text{tol} = 10^{-10}$

The results are displayed in Figures 6.4-6.6. As before, we show only the eigenvalues K_3 closest to 0, namely $\lambda_{m+n}(K_3)$ and $\lambda_{m+n+1}(K_3)$, together with their bounds θ_2, θ_3, μ_2 and μ_3. The results are similar to the ones considered for the LP problems: θ_2 and θ_3 are good approximation for the true eigenvalues in early and middle iterations, and also in late iterations in the left plot of Figure 6.5, where eigenvalues approach 0 in the limit, and in the left plot of Figure 6.4. On the other hand, in late iterations μ_2 and μ_3 are often more representative. A particularly interesting case occurs for the positive

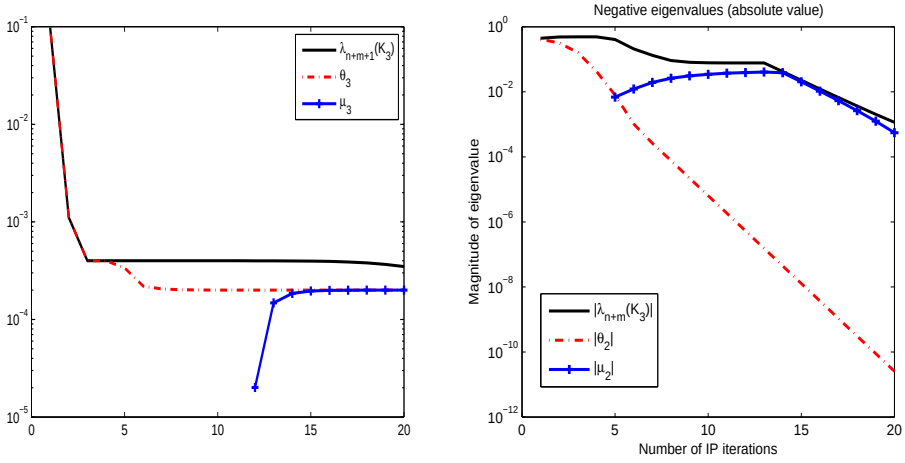

Figure 6.4: Problem STCQP2: eigenvalues of K_3 closest to zero (solid line) and their bounds at every iteration. Left: positive eigenvalues. Right: negative eigenvalues.

eigenvalues of problem CONT-050, shown in the left plot of Figure 6.6. Here H is positive definite, and hence in the last iterations $\theta_2 \approx \lambda_{\min}(H) > 0$. However, the real minimum positive eigenvalue K_3 appears to be several orders of magnitude greater than $\lambda_{\min}(H)$, and hence θ_2 cannot describe it. On the other hand, μ_2 is a very good approximation for $\lambda_{n+m}(K_3)$, in the last iterations. We emphasize that, only for this problem, we observed $q = n-m$, and so the second expression of μ_3^* from Theorem 6.3.4 was used.

6.5 Conclusions

In this chapter we have studied symmetric unreduced KKT systems, as they arise in the solution of convex quadratic programming problems solved by IP methods, and we have characterized the spectrum of the corresponding matrices.

In the unpreconditioned case, we distinguished between two stages of the IP method: generic iterations, and late or final stage. A spectral analysis should be able to reflect the peculiarities of each of these two phases, and in particular to capture the potential nonsingularity of the matrices at the limit. For the generic iteration, we were able to measure the distance from singularity for the negative eigenvalues. By appropriately partitioning the coefficient matrix, we were also able to characterize the spectral properties in the late stage of the IP iterations and at the solution, giving novel estimates

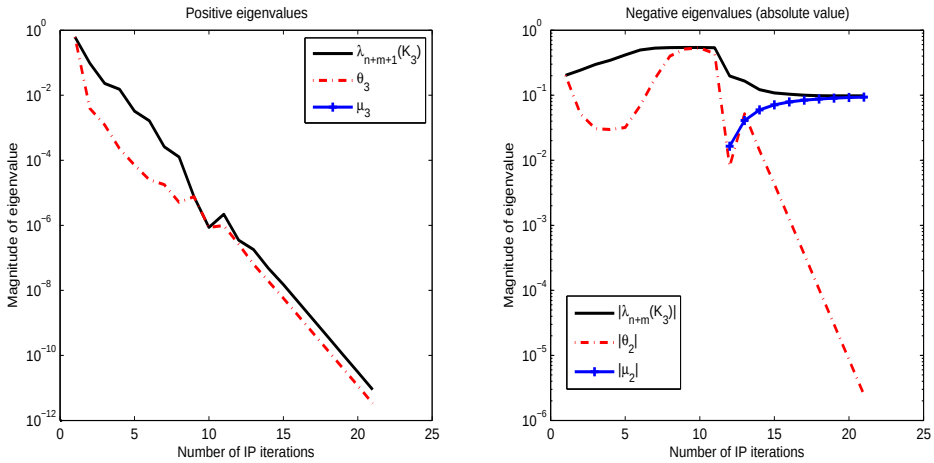

Figure 6.5: Problem AUG3D: eigenvalues of K_3 closest to zero (solid line) and their bounds at every iteration. Left: positive eigenvalues. Right: negative eigenvalues.

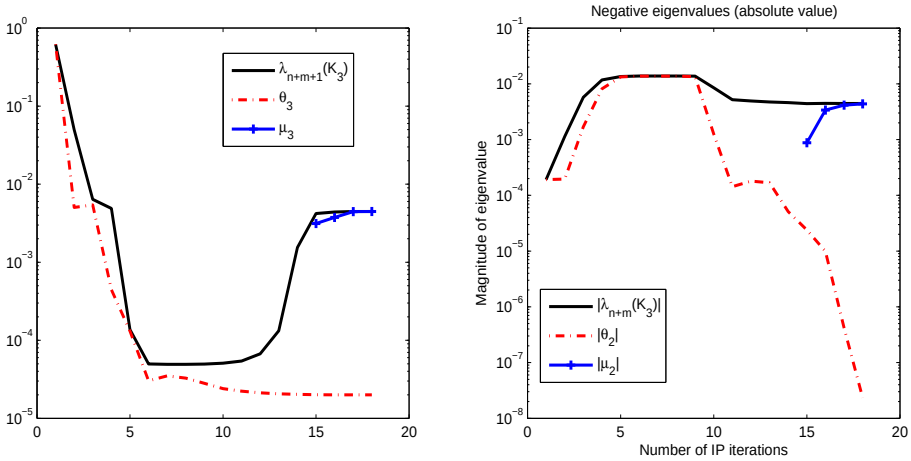

Figure 6.6: Problem CONT-050: eigenvalues of K_3 closest to zero (solid line) and their bounds at every iteration. Left: positive eigenvalues. Right: negative eigenvalues.

in this delicate case.

Chapter 7

Preconditioning for the reduced and unreduced formulation[1]

In the previous chapter, discussed Interior Point (IP) methods and their application to large-scale Quadratic Programming problems of the form ((6.1)). As already mentioned, the efficiency of such methods heavily depends on the per-iteration cost and this is mainly constituted by the solution of a structured algebraic linear system, namely the KKT system (6.4). Therefore, much effort has been devoted to developing properly tailored preconditioned iterative solvers, whose computational cost and memory requirements may be lower than those of direct solvers. The analysis and development of these resulting *inexact* IP methods have covered several different aspects such as the level of accuracy in the solution of linear systems, the design of suitable iterative methods and preconditioners, and the convergence analysis of the inexact IP solver, including worst-case iteration complexity, see, e.g., [15, 23, 30, 33, 36, 49, 59, 60, 110].

Two different and well-established formulations for the KKT system have been introduced in the previous chapter, which we called reduced and unreduced formulation. The aim of this chapter is to study the effect of preconditioning strategies on their relation. This analysis will be helpful in assessing which of the two formulations should be preferred when solving large scale problems. Relevant alternative formulations, possibly definite, such as condensed systems and doubly augmented systems (see, e.g., [49]) are not considered in this work.

The reduced systems become increasingly ill-conditioned in the progress of the IP iterations, though such ill-conditioning is benign if suitable direct methods are used [47, 137, 139]. On the other hand, as discussed in the

[1]The results presented in this chapter are taken from [91].

previous chapter, the unreduced systems may be well-conditioned throughout the IP iteration and nonsingular even in the limit; this distinguishing feature, observed in [46, 48] and supported by spectral analysis presented in [66] and in Chapter 6 of this thesis, motivates our interest in this possibly less exercised formulation.

A large number of papers have developed and analyzed preconditioning of indefinite systems in optimization see, e.g., the surveys [11, 30, 58]; however, a theoretical and experimental comparison of the reduced and unreduced formulations in the preconditioned regime has not been performed. We aim to assess whether the use of unreduced systems may still offer some advantage with respect to the reduced ones in terms of eigenvalues and conditioning as occurs in the unpreconditioned case.

Our study adopts a general form of the systems which includes the case where regularizations are applied to the QP problem, and investigates the use of constraint and augmented preconditioners. The analysis conducted shows that, for some frequently employed preconditioners, the two formulations remain strictly related, both in terms of spectra of the preconditioned matrices, and of the preconditioned systems. In particular, for typical constraint preconditioners, spectral invariance holds between the two formulations considered along with equality of equations in the preconditioned systems, making the use of the unreduced formulation questionable when using this class of preconditioners. On the other hand, these relations are no longer valid for augmented (diagonal or triangular) preconditioners; for these, an experimental comparison is performed in the context of the IP solver PDCO [109], so as to better assess the merits of each of the two formulations. Ad-hoc augmentation matrices are proposed for the reduced and unreduced forms. Numerical experiments confirm that the conditioning of the unreduced systems varies slowly with the IP iterations and may remain considerably smaller than the conditioning of the reduced formulation; this feature is shared by the corresponding augmented preconditioners. However, these better spectral properties do not seem to play a role as long as the systems are numerically nonsingular and the IP implementations with the unreduced and reduced formulations are both successful. In this case, the preconditioned linear solvers and the IP solvers behave similarly; thus the smallest dimension of the reduced systems makes such formulation preferable in terms of computational time. On the other hand, a potential benefit of the unreduced formulation may be observed in terms of robustness of the IP solver when the reduced systems become severely ill-conditioned.

The chapter is organized as follows. In Section 7.1 we introduce the regularized systems under study and state which classes of preconditioners we will consider. In Section 7.2 we show that the unreduced and reduced

formulations are closely related when some of these classes are used, and that only certain types of preconditioners allow to preserve differences between the two formulations. In Section 7.3 we experimentally compare the two formulations and their iterative solution in an IP solver and finally in Section 7.4 we draw our conclusions.

7.1 Preliminaries

We have seen in the previous chapter that when an Interior Point method is applied to the QP problem

$$\min_x \quad c^T x + \frac{1}{2} x^T H x \qquad \text{subject to} \quad Jx = b,\ x \geq 0, \tag{7.1}$$

then each iteration requires the solution of the KKT system, which we report here for the convenience of the reader:

$$\begin{bmatrix} H & J^T & -I \\ J & 0 & 0 \\ -Z & 0 & -X \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} = \begin{bmatrix} -Hx + J^T y + z + c \\ -Jx + b \\ -XZe - \tau e \end{bmatrix}. \tag{7.2}$$

In order to provide a comprehensive analysis of the symmetric 3×3 systems, it is useful to consider the case where regularizations are applied to the optimization problem (7.1). Several regularization techniques have been proposed in order to improve the numerical properties of the KKT systems and for details we refer to [1, 27, 53, 109, 116]. Here we focus on primal-dual regularizations such that system (7.2) becomes

$$\begin{bmatrix} H + \rho I_n & J^T & -I_n \\ J & -\delta I_m & 0 \\ -Z & 0 & -X \end{bmatrix} \begin{bmatrix} \Delta x \\ -\Delta y \\ \Delta z \end{bmatrix} = \begin{bmatrix} f_x \\ f_y \\ f_z \end{bmatrix}, \tag{7.3}$$

where $\delta, \rho \geq 0$ and the right-hand side vectors f_x, $f_z \in \mathbb{R}^n$ and $f_y \in \mathbb{R}^m$ are appropriately computed, see e.g. [53, 66, 109]. Using again the block diagonal matrix R in (6.5), the corresponding symmetric formulation is given by

$$K_{3,reg}\, \Delta_3 = f_3, \tag{7.4}$$

where

$$K_{3,reg} = \begin{bmatrix} H + \rho I_n & J^T & -Z^{\frac{1}{2}} \\ J & -\delta I_m & 0 \\ -Z^{\frac{1}{2}} & 0 & -X \end{bmatrix}, \tag{7.5}$$

and

$$\Delta_3 = \begin{bmatrix} \Delta x \\ -\Delta y \\ Z^{-\frac{1}{2}}\Delta z \end{bmatrix}, \qquad f_3 = \begin{bmatrix} f_x \\ f_y \\ Z^{-\frac{1}{2}} f_z \end{bmatrix}.$$

Further, upon reduction of Δz, system (7.2) becomes

$$K_{2,reg}\,\Delta_2 = f_2, \tag{7.6}$$

with

$$K_{2,\mathrm{reg}} = \begin{bmatrix} H + \rho I_n + X^{-1}Z & J^T \\ J & -\delta I_m \end{bmatrix},\ \Delta_2 = \begin{bmatrix} \Delta x \\ -\Delta y \end{bmatrix},\ f_2 = \begin{bmatrix} f_x - X^{-1} f_z \\ f_y \end{bmatrix}. \tag{7.7}$$

For $\delta, \rho \geq 0$, throughout the chapter we will refer to $K_{2,reg}$ and $K_{3,reg}$ as the *reduced* and *unreduced* matrices and recover K_2 and K_3 by setting $\delta = \rho = 0$ (analogously the systems (7.6) and (7.4) will be denoted reduced and unreduced systems, respectively). Clearly $K_{3,reg}$ can be cast into KKT form by proper block reordering. For later convenience, we observe that as long as X is nonsingular, these matrices are mathematically related by means of a congruence transformation. Indeed, setting

$$L = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ X^{-1}Z^{\frac{1}{2}} & 0 & I \end{bmatrix}, \tag{7.8}$$

it holds (see, e.g., [66])

$$\begin{aligned} K_{3,reg} &= \begin{bmatrix} I & 0 & Z^{\frac{1}{2}}X^{-1} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} K_{2,\mathrm{reg}} & 0 \\ 0 \quad 0 & -X \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ X^{-1}Z^{\frac{1}{2}} & 0 & I \end{bmatrix} \\ &= L^T \begin{bmatrix} K_{2,\mathrm{reg}} & 0 \\ 0 \quad 0 & -X \end{bmatrix} L. \end{aligned} \tag{7.9}$$

The following theorem characterizes the nonsingularity of $K_{2,\mathrm{reg}}$ and $K_{3,\mathrm{reg}}$.

Theorem 7.1.1. *Suppose that H is symmetric and positive semidefinite, X and Z are diagonal with positive entries. Let $K_{2,\mathrm{reg}}$ and $K_{3,\mathrm{reg}}$ be given in (7.7) and (7.5). Then $K_{2,\mathrm{reg}}$ and $K_{3,\mathrm{reg}}$ are nonsingular if and only if either $\delta > 0$, or $\delta = 0$ and J has full row rank.*

Proof. See e.g. [66, Corollary 3.4, Corollary 3.6, Theorem 3.9]. □

The use of the 2×2 formulation is supported by its reduced dimension and by the variety of direct solvers, iterative solvers and preconditioners available for its numerical solution [11]. However, the presence of matrix $X^{-1}Z$ may cause ill-conditioning of $K_{2,reg}$ while a solution is approached, and it represents the key difference between $K_{2,reg}$ and $K_{3,reg}$.

As we have seen in the previous chapter, under specific assumptions the unreduced matrix is nonsingular throughout the IP iterations and remains well-conditioned; at this regard, it is important to recall that ill-conditioning occurs in the matrix R defined in (6.5), see [46, 48], but the square root in the $(3,3)$ block has a damping effect on ill-conditioning at the final stage of the IP process. These properties may favor the use of unreduced systems and have indeed motivated the study of spectral estimates for $K_{3,reg}$ [66]. On the other hand, when the system (7.4) is solved iteratively and preconditioning is required, assessing the advantages of the unreduced formulation over the reduced one is still an open issue. Therefore, in this paper we study, both theoretically and computationally, systems (7.6) and (7.4) preconditioned by preconditioners in a same class and attempt to establish their distinguishing features.

We conclude this section by listing a few suitable preconditioners for our systems that will be analyzed in the following: constraint preconditioners and preconditioners based on augmentation of the $(1,1)$ block.

In order to handle systems (7.4) and (7.6) simultaneously, we consider the general formulation for a saddle point matrix

$$\mathcal{M} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix},$$

with $A \in \mathbb{R}^{n_1 \times n_1}$ and $C \in \mathbb{R}^{n_2 \times n_2}$ both positive semidefinite, $B \in \mathbb{R}^{n_2 \times n_1}$. We suppose that $\mathcal{M}$ is nonsingular and allow any relation between n_1 and n_2.

A suitable definite preconditioner is the augmented block diagonal and positive definite matrix

$$\mathcal{P}_{AD} = \begin{bmatrix} A + B^T W^{-1} B & 0 \\ 0 & W \end{bmatrix}, \tag{7.10}$$

where $W \in \mathbb{R}^{n_2 \times n_2}$ is a symmetric positive definite matrix. This preconditioner and its features have been discussed in Chapter 3 of this thesis. We recall that the use of $\mathcal{P}_{AD}$ may be advantageous over preconditioners based on the Schur complement of A, when the (1,1) block is singular or ill-conditioned. Interestingly, this may be the case for both the reduced and unreduced matrices, as the (1,1) block may be ill-conditioned in $K_{2,reg}$ at the late stage of the IP method and singular in K_3 at any IP iteration.

Constraint preconditioners for $\mathcal{M}$ are commonly used in optimization, see, e.g., [15, 23, 32, 33, 34, 36, 49, 77, 86]. We already introduced this class of preconditioners in Chapter 2; they take the form

$$\mathcal{P}_C = \begin{bmatrix} \widehat{A} & B^T \\ B & -C \end{bmatrix}, \tag{7.11}$$

where $\widehat{A}$ is a symmetric approximation of A.

Alternatively, block triangular preconditioners based on the augmentation of the $(1,1)$ block can be defined; see, e.g., [4, 24, 110, 119], where however most of the results are for a zero (2,2) block. In this nonsymmetric framework, one option is to set

$$\mathcal{P}_{AT} = \begin{bmatrix} A + B^T W^{-1} B & \kappa B^T \\ 0 & -W \end{bmatrix}, \tag{7.12}$$

where κ is a scalar and $W \in \mathbb{R}^{n_2 \times n_2}$ is a symmetric positive definite matrix. Clearly, setting $\kappa = 0$, $\mathcal{P}_{AT}$ is block diagonal and indefinite. Another possibility is to consider

$$\mathcal{T}_W = \begin{bmatrix} A + B^T W^{-1} B & \kappa B^T \\ 0 & -(W + C) \end{bmatrix}, \tag{7.13}$$

where κ is a scalar and $W \in \mathbb{R}^{n_2 \times n_2}$ is symmetric positive definite. This preconditioner was introduced in [119] with $\kappa = 2$.

7.2 Standard preconditioners for the unreduced systems

In this section we show that, for some of the preconditioners of type (7.10), (7.11) and (7.12), the unreduced formulation does not offer any advantage over the 2×2 reduced one. In particular, we establish two different results on the relationship between preconditioned 2×2 and 3×3 formulations. The first result (Theorems 7.2.1 and 7.2.2) holds for specific constraint preconditioners and augmented triangular preconditioners; invariance of the spectra of the preconditioned matrices and correspondence of block equations in the preconditioned systems are shown. The second result (Theorem 7.2.3) is valid for specific augmented diagonal and triangular preconditioners and indicates that, by applying the same elimination of Δz as in (7.4), the system $\mathcal{P}_{3,AT}^{-1} K_{3,reg} \Delta_3 = \mathcal{P}_{3,AT}^{-1} f_3$ reduces to $\mathcal{P}_{2,AT}^{-1} K_{2,reg} \Delta_2 = \mathcal{P}_{2,AT}^{-1} f_2$.

In view of such close relationships between the preconditioned systems (7.6) and (7.4) and the larger dimension of the latter system, we conclude

that there is no motivation for using the unreduced formulation. On the other hand, these results do not exclude the existence of effective preconditioners which preserve differences between the two formulations and may make the unreduced system preferable to the reduced one; this topic will be explored in Section 7.3.

7.2.1 Equivalence properties of preconditioned 2×2 and 3×3 formulations

In this section we study the application of specific occurrences of preconditioners (7.11) and (7.12) to the systems (7.6) and (7.4), establishing spectral correspondences between the preconditioned matrices, and equivalences between the associated linear systems.

The constraint preconditioners analyzed for $K_{3,reg}$ and $K_{2,reg}$ are of the form

$$\mathcal{P}_{2,C} = \begin{bmatrix} \text{diag}(H + \rho I_n + X^{-1}Z) & J^T \\ J & -\delta I_m \end{bmatrix}, \tag{7.14}$$

$$\mathcal{P}_{3,C} = \begin{bmatrix} \text{diag}(H + \rho I_n) & J^T & -Z^{\frac{1}{2}} \\ J & -\delta I_m & 0 \\ -Z^{\frac{1}{2}} & 0 & -X \end{bmatrix} \tag{7.15}$$

$$= L^T \begin{bmatrix} \mathcal{P}_{2,C} & 0 \\ 0 \quad 0 & -X \end{bmatrix} L, \tag{7.16}$$

where the last factorization is analogous to that for $K_{3,reg}$ in (7.9) and follows from the trivial equality $\text{diag}(H + \rho I_n + X^{-1}Z) = \text{diag}(H + \rho I_n) + X^{-1}Z$. Constraint preconditioners where the $(1,1)$ block is approximated by retaining its main diagonal are widely used, see, e.g., [15, 77, 86].

Augmentation in (7.12) is performed on the $(1,1)$ of $K_{2,reg}$ and $K_{3,reg}$ as follows. Let $R_d \in \mathbb{R}^{m \times m}$ be positive definite and such that

$$R_d = \delta I_m, \quad \text{if } \delta > 0.$$

Moreover, let $W = \begin{bmatrix} R_d & 0 \\ 0 & X \end{bmatrix}$ for system (7.4), and $W = R_d$ for system (7.6). Then, the augmented block triangular preconditioners $\mathcal{P}_{2,AT}$, $\mathcal{P}_{3,AT}$ are of

the form

$$\mathcal{P}_{2,AT} = \begin{bmatrix} H+\rho I_n + X^{-1}Z + J^T R_d^{-1} J & \kappa J^T \\ 0 & -R_d \end{bmatrix}, \tag{7.17}$$

$$\mathcal{P}_{3,AT} = \begin{bmatrix} H+\rho I_n + X^{-1}Z + J^T R_d^{-1} J & \kappa J^T & -\kappa Z^{\frac{1}{2}} \\ 0 & -R_d & 0 \\ 0 & 0 & -X \end{bmatrix} \tag{7.18}$$

$$= \begin{bmatrix} \mathcal{P}_{2,AT} & \begin{matrix} -kZ^{\frac{1}{2}} \\ 0 \end{matrix} \\ \begin{matrix} 0 & 0 \end{matrix} & -X \end{bmatrix}, \tag{7.19}$$

In the next two theorems we analyze the systems (7.6) and (7.4) preconditioned by the constraint preconditioners (7.14), (7.15) and the triangular preconditioners (7.17), (7.19) with $\kappa = 1$ and prove strong results which are straightforward consequences of the congruence between the reduced and unreduced formulations. Specifically, first we prove that, apart from the multiplicity of the unit eigenvalue, the eigenvalues of preconditioned $K_{3,reg}$ and $K_{2,reg}$ coincide. Second we show that the first two block equations of $\mathcal{P}_{3,C}^{-1}K_{3,reg}\Delta_3 = \mathcal{P}_{3,C}^{-1}f_3$ ($\mathcal{P}_{3,AT}^{-1}K_{3,reg}\Delta_3 = \mathcal{P}_{3,AT}^{-1}f_3$) are equal to $\mathcal{P}_{2,C}^{-1}K_{2,\mathrm{reg}}\Delta_2 = \mathcal{P}_{2,C}^{-1}f_2$ ($\mathcal{P}_{2,AT}^{-1}K_{2,\mathrm{reg}}\Delta_2 = \mathcal{P}_{2,AT}^{-1}f_2$) and the third equation is equivalent to the third equation in (7.4).

Theorem 7.2.1. *Suppose that H is symmetric and positive semidefinite, X and Z are diagonal with positive entries, $K_{2,\mathrm{reg}}$ and $K_{3,\mathrm{reg}}$ given in (7.7) and (7.5) are nonsingular. Let $\mathcal{P}_{2,C}$ and $\mathcal{P}_{3,C}$ be the preconditioners (7.14) and (7.15) respectively, and consider the systems (7.4) and (7.6). Then*

i) $\theta \in \mathrm{spec}(\mathcal{P}_{3,C}^{-1}K_{3,\mathrm{reg}})$ if and only if either $\theta = 1$ or $\theta \in \mathrm{spec}(\mathcal{P}_{2,C}^{-1}K_{2,\mathrm{reg}})$.

ii) Solving $\mathcal{P}_{3,C}^{-1}K_{3,\mathrm{reg}}\Delta_3 = \mathcal{P}_{3,C}^{-1}f_3$ reduces to solving $\mathcal{P}_{2,C}^{-1}K_{2,\mathrm{reg}}\Delta_2 = \mathcal{P}_{2,C}^{-1}f_2$ and recovering Δz from the third equation in (7.4).

Proof. To show *i)*, consider the generalized eigenvalue problem $K_{3,reg}u = \theta\mathcal{P}_{3,C}u$ with $u \in \mathbb{R}^{2n+m}$. Using (7.9) and (7.16), we have $K_{3,reg}u = \theta\mathcal{P}_{3,C}u$ if and only if

$$L^T \begin{bmatrix} K_{2,\mathrm{reg}} & \begin{matrix} 0 \\ 0 \end{matrix} \\ \begin{matrix} 0 & 0 \end{matrix} & -X \end{bmatrix} Lu = \theta L^T \begin{bmatrix} \mathcal{P}_{2,C} & \begin{matrix} 0 \\ 0 \end{matrix} \\ \begin{matrix} 0 & 0 \end{matrix} & -X \end{bmatrix} Lu,$$

and the result readily follows from the nonsingularity of L.

Concerning *ii)* we use again (7.9) and (7.16) and note that $\mathcal{P}_{3,C}^{-1}K_{3,\mathrm{reg}}\Delta_3 = \mathcal{P}_{3,C}^{-1}f_3$ if and only if

$$\begin{bmatrix} \mathcal{P}_{2,C} & 0 \\ 0 & 0 & -X \end{bmatrix}^{-1} \begin{bmatrix} K_{2,reg} & 0 \\ 0 & 0 & -X \end{bmatrix} (L\Delta_3) = \begin{bmatrix} \mathcal{P}_{2,C} & 0 \\ 0 & 0 & -X \end{bmatrix}^{-1} L^{-T} f_3.$$

Hence, the result follows from explicitly writing

$$L\Delta_3 = (\Delta x, -\Delta y, X^{-1}Z^{\frac{1}{2}}\Delta x + Z^{-\frac{1}{2}}\Delta z) = (\Delta_2, X^{-1}Z^{\frac{1}{2}}\Delta x + Z^{-\frac{1}{2}}\Delta z), \tag{7.20}$$

$$L^{-T}f_3 = (f_x - X^{-1}f_z, f_y, Z^{-\frac{1}{2}}f_z) = (f_2, Z^{-\frac{1}{2}}f_z), \tag{7.21}$$

with Δ_2 and f_2 given in (7.7). □

We observe that the spectral properties of $\mathcal{P}_{2,C}^{-1}K_{2,\mathrm{reg}}$ have been studied in a variety of papers, e.g. [15, 32, 33, 34, 77, 86], and in light of Theorem 7.2.1, these results apply to $\mathcal{P}_{3,C}^{-1}K_{3,\mathrm{reg}}$. In terms of performance, the result of Theorem 7.2.1 shows that little can be gained by the unreduced formulation when the popular constraint preconditioner is employed. Additional comments in this direction are postponed to the end of this section.

The second proof uses similar arguments.

Theorem 7.2.2. *Suppose that H is symmetric and positive semidefinite, X and Z are diagonal with positive entries, $K_{2,\mathrm{reg}}$ and $K_{3,\mathrm{reg}}$ given in (7.7) and (7.5) are nonsingular. Let $\mathcal{P}_{2,AT}$ and $\mathcal{P}_{3,AT}$ be the preconditioners (7.17) and (7.19) respectively with $\kappa = 1$, and consider the systems (7.4) and (7.6). Then*

i) $\theta \in \mathrm{spec}(\mathcal{P}_{3,AT}^{-1}K_{3,\mathrm{reg}})$ if and only if either $\theta = 1$ or $\theta \in \mathrm{spec}(\mathcal{P}_{2,AT}^{-1}K_{2,\mathrm{reg}})$.

ii) Solving $\mathcal{P}_{3,AT}^{-1}K_{3,\mathrm{reg}}\Delta_3 = \mathcal{P}_{3,AT}^{-1}f_3$ reduces to solving $\mathcal{P}_{2,AT}^{-1}K_{2,\mathrm{reg}}\Delta_2 = \mathcal{P}_{2,AT}^{-1}f_2$ and recovering Δz from the third equation in (7.4).

Proof. To characterize the spectrum of $\mathcal{P}_{3,AT}^{-1}K_{3,reg}$, we first observe that for L given in (7.8),

$$L^{-T}\mathcal{P}_{3,AT}L^{-1} = \begin{bmatrix} \mathcal{P}_{2,AT} & 0 \\ Z^{\frac{1}{2}} & 0 & -X \end{bmatrix}, \tag{7.22}$$

and that the eigenvalue problem $K_{3,\mathrm{reg}}u = \theta\mathcal{P}_{3,AT}u$, $u \in \mathbb{R}^{2n+m}$ can be written as

$$L^T \begin{bmatrix} K_{2,\mathrm{reg}} & 0 \\ 0 & 0 & -X \end{bmatrix} Lu = \theta L^T L^{-T}\mathcal{P}_{3,AT}L^{-1}Lu.$$

Thus, by using (7.22)

$$\begin{bmatrix} K_{2,reg} & 0 & \\ 0 & 0 & -X \end{bmatrix} Lu = \theta \begin{bmatrix} \mathcal{P}_{2,AT} & 0 & \\ Z^{\frac{1}{2}} & 0 & -X \end{bmatrix} Lu. \tag{7.23}$$

Setting $\hat{u} = Lu = (\hat{u}_1, \hat{u}_2)$ with $\hat{u}_1 \in \mathbb{R}^{n+m}$, $\hat{u}_2 \in \mathbb{R}^n$, the first block row gives the eigenproblem $K_{2,reg}\hat{u}_1 = \theta \mathcal{P}_{2,AT}\hat{u}_1$ while the second block row gives $-X\hat{u}_2 = \theta([Z^{\frac{1}{2}}, 0]\hat{u}_1 - X\hat{u}_2)$. Therefore, all eigenvalues of $(K_{2,reg}, \mathcal{P}_{2,AT})$ are also eigenvalues of the unreduced problem with $\hat{u}_2$ given by the second block row. The remaining eigenvalues are obtained for $\hat{u}_1 = 0$, which gives $\theta = 1$. Note that if $\theta = 1$ is also an eigenvalue of $(K_{2,reg}, \mathcal{P}_{2,AT})$, then the matrix pencil in (7.23) is not diagonalizable.

We now prove item *ii)*. By multiplying from the left by L in (7.8) and using (7.9), the preconditioned system $\mathcal{P}_{3,AT}^{-1}K_{3,reg}\Delta_3 = \mathcal{P}_{3,AT}^{-1}f_3$ can be written as

$$L\mathcal{P}_{3,AT}^{-1}L^T \begin{bmatrix} K_{2,reg} & 0 & \\ 0 & 0 & -X \end{bmatrix} L\Delta_3 = L\mathcal{P}_{3,AT}^{-1}L^T L^{-T} f_3.$$

Since $L\mathcal{P}_{3,AT}^{-1}L^T = (L^{-T}\mathcal{P}_{3,AT}L^{-1})^{-1}$ and $L^{-T}f_3 = (f_2, Z^{-\frac{1}{2}}f_z)$, from (7.22) it follows that the system can be rewritten as

$$\begin{bmatrix} \mathcal{P}_{2,AT} & 0 & \\ Z^{\frac{1}{2}} & 0 & -X \end{bmatrix}^{-1} \begin{bmatrix} K_{2,reg} & 0 & \\ 0 & 0 & -X \end{bmatrix} L\Delta_3 = \begin{bmatrix} \mathcal{P}_{2,AT} & 0 & \\ Z^{\frac{1}{2}} & 0 & -X \end{bmatrix}^{-1} \begin{bmatrix} f_2 \\ \\ Z^{-\frac{1}{2}}f_z \end{bmatrix}.$$

By (7.20) the first block equation coincides with the reduced preconditioned system while the third block equation is equivalent to the third equation in (7.4). □

Characterizations of the eigenvalues for $\mathcal{P}_{2,AT}^{-1}K_{2,reg}$ can be found in [24, 61, 110]. The above theorem is valid as long as $\kappa = 1$ and the matrix W is the one employed so far. It does not further hold if the $(3,3)$ block is different from X or if augmentation is performed using only the Schur complement $J^T R_d^{-1} J$.

We conclude this part by making some comments on the given correspondence results. They show that for the types of preconditioners considered, the spectral properties of the preconditioned 2×2 and 3×3 matrices are the same. As a consequence, the computational performance of an iterative solver, at least in terms of number of iterations, is expected to be the same for the reduced and unreduced systems. Taking into account that the unreduced

formulation requires larger memory allocations and computational costs per iteration, we claim that there does not seem to be any advantage in using it.

In looking for effective preconditioners for which the use of the unreduced formulation will pay off, more sophisticated preconditioning strategies should be explored. Such an experimental analysis is performed in Section 7.3.

7.2.2 Further relations between the 2×2 and 3×3 preconditioned systems

Further relationships between the systems (7.4) and (7.6) preconditioned by augmented diagonal and triangular preconditioners are shown in this section. The triangular preconditioners are as in (7.17), (7.19), and here we are interested in the case $\kappa \neq 1$. The diagonal preconditioners given in (7.10) are defined by using the same augmentation, thus,

$$\mathcal{P}_{2,AD} = \begin{bmatrix} H + \rho I_n + X^{-1}Z + J^T R_d^{-1} J & 0 \\ 0 & R_d \end{bmatrix}, \tag{7.24}$$

$$\mathcal{P}_{3,AD} = \begin{bmatrix} H + \rho I_n + X^{-1}Z + J^T R_d^{-1} J & 0 & 0 \\ 0 & R_d & 0 \\ 0 & 0 & X \end{bmatrix} = \begin{bmatrix} \mathcal{P}_{2,AD} & 0 \\ 0 \quad 0 & -X \end{bmatrix}. \tag{7.25}$$

We remark that in the regularized case, W is equal to the $(2,2)$ block of the matrices and this is an "optimal" choice for the diagonal preconditioner in terms of spectral distribution (see Section 3.2.4).

We now show that, upon reduction of Δz, the 3×3 preconditioned system reduces to the 2×2 preconditioned system.

Theorem 7.2.3. *Suppose that H is symmetric and positive semidefinite, X and Z are diagonal with positive entries, $K_{2,reg}$ and $K_{3,\mathrm{reg}}$ given in (7.7) and (7.5) are nonsingular. Let $\mathcal{P}_{2,AT}$ and $\mathcal{P}_{3,AT}$ be the preconditioners in (7.17) and (7.19) respectively, with $\kappa \in \mathbb{R}$.*

Then, by applying the same elimination of Δz as in $K_{3,\mathrm{reg}}\Delta_3 = f_3$, the system $\mathcal{P}_{3,AT}^{-1}K_{3,\mathrm{reg}}\Delta_3 = \mathcal{P}_{3,AT}^{-1}f_3$ reduces to $\mathcal{P}_{2,AT}^{-1}K_{2,\mathrm{reg}}\Delta_2 = \mathcal{P}_{2,AT}^{-1}f_2$.

The same feature holds with $\mathcal{P}_{2,AD}$, $\mathcal{P}_{3,AD}$ in (7.24), (7.25).

Proof. We first observe that

$$\mathcal{P}_{3,AT}^{-1} = \begin{bmatrix} \mathcal{P}_{2,AT}^{-1} & -\kappa\mathcal{P}_{2,AT}^{-1}\begin{bmatrix} X^{-1}Z^{\frac{1}{2}} \\ 0 \end{bmatrix} \\ 0 \quad 0 & -X^{-1} \end{bmatrix}.$$

Then, by using (7.9) the preconditioned system $\mathcal{P}_{3,AT}^{-1}K_{3,\text{reg}}\Delta_3 = \mathcal{P}_{3,AT}^{-1}f_3$ can be written as

$$\mathcal{P}_{3,AT}^{-1}L^T\begin{bmatrix}K_{2,reg} & 0\\ 0 & -X\end{bmatrix}L\Delta_3 = \mathcal{P}_{3,AT}^{-1}L^TL^{-T}f_3,$$

and takes the form

$$\begin{bmatrix}\mathcal{P}_{2,AT}^{-1}K_{2,reg} & -(1-\kappa)\mathcal{P}_{2,AT}^{-1}\begin{bmatrix}Z^{\frac{1}{2}}\\ 0\end{bmatrix}\\ 0 \quad 0 & I_m\end{bmatrix}L\Delta_3 = \begin{bmatrix}\mathcal{P}_{2,AT}^{-1} & (1-\kappa)\mathcal{P}_{2,AT}^{-1}\begin{bmatrix}X^{-1}Z^{\frac{1}{2}}\\ 0\end{bmatrix}\\ 0 \quad 0 & -X^{-1}\end{bmatrix}L^{-T}f_3.$$

Finally, by (7.20) and (7.21) it readily follows that by back substitution of Δz the first block equation coincides with the reduced preconditioned system.

The claim for the diagonal preconditioners $\mathcal{P}_{2,AD}$, $\mathcal{P}_{3,AD}$ in (7.24), (7.25) can be proved by repeating the above arguments. □

7.3 Qualitatively different preconditioning strategies

Our previous results show that the reduced and unreduced formulations are closely related, also when a large class of preconditioning strategies is used. Therefore, to be able to investigate the true potential of the unreduced formulation we need to select the free parameters of these acceleration strategies in a way that is peculiar to the 3×3 problem. To this end, we experimentally compare the use of both the reduced and unreduced formulations in an IP method for problem (7.1). The aim of our numerical experiments is twofold. First, we wish to compare the performance of the augmented block diagonal and block triangular preconditioners of the form (7.10) and (7.13), respectively, on the unreduced systems; second we wish to assess whether the unreduced formulation can be advantageous in terms of conditioning and execution time with respect to the reduced one.

The numerical experiments were conducted using MATLAB R2012a on a 4xAMD Opteron 850, 2.4GHz, 16GB of RAM processor. Elapsed times were measured by the `tic` and `toc` MATLAB commands.

Our numerical experiments were based on six convex QP problems from the CUTEr collection [62], whose matrix information is summarized in Table 7.1. All these datasets are characterized by a sufficiently large number of nonzeros in H so as to justify the use of an iterative solver.

The QP problems were solved by the MATLAB code PDCO (Primal Dual interior method for optimization with Convex Objectives) developed by Michael

Problem	n	m	nnz(H)	nnz(J)
CVXQP1	10000	5000	69968	14998
CVXQP2	10000	7500	69968	7499
CVXQP3	10000	7500	69968	22497
STCQP1_a	8193	4095	106473	28865
STCQP1_b	16385	8190	229325	61425
STCQP2	16385	8190	229325	61425

Table 7.1: Test problems: values of n, m, nonzeros in H and in J

Saunders, and available in [109]. For stability purposes, problem (7.1) is regularized as

$$\min_{x,r} \; c^T x + \frac{1}{2} x^T H x + \frac{1}{2}\|D_1 x\|^2 + \frac{1}{2}\|r\|^2 \quad \text{subject to} \quad Jx + D_2 r = b, \; x \geq 0,$$

where D_1 and D_2 are positive definite diagonal matrices specified by the user. PDCO implements an IP method, and it follows the general structure described in Section 6.1. Therefore, letting $D_1 = \sqrt{\rho} I_n$ and $D_2 = \sqrt{\delta} I_m$ for positive ρ and δ, the unreduced and reduced coefficient matrices of the systems to be solved take the form $K_{2,reg}$ and $K_{3,reg}$ respectively. In all our runs, we set $\delta = \rho = 10^{-6}$ and solved the problems with *no* variable scaling.

PDCO allows one to work with two alternative linear system formulations: one employs a reduced form with matrix $K_{2,reg}$, the other one is a condensed formulation where the coefficient matrix is a Schur complement. Symmetric indefinite systems are solved by a direct solver whereas for condensed systems both direct and iterative solvers are available. PDCO was modified so that the unreduced regularized formulation with matrix $K_{3,reg}$ could also be explicitly formed. In the following we report statistics on the results obtained and put emphasis on the solution of the sequence of linear systems generated with each formulation.

The systems in the sequence were solved by using either MATLAB sparse direct solver (function "\"), MINRES coupled with a diagonal augmented preconditioner $\mathcal{P}_{AD}$ of the form (7.10), or GMRES coupled with a triangular augmented preconditioner $\mathcal{T}_W$ of the form (7.13), where in the unreduced case we have

$$A = H + \rho I_n, \quad B = \begin{bmatrix} J \\ -Z^{\frac{1}{2}} \end{bmatrix}, \quad C = \begin{bmatrix} \delta I_m & 0 \\ 0 & X \end{bmatrix},$$

while in the reduced case we have

$$A = H + \rho I_n + X^{-1} Z, \quad B = J, \quad C = \delta I_m.$$

Following [119], we set $\kappa = 2$ in (7.13) and consequently

$$\mathcal{T}_W = \begin{bmatrix} A + B^T W^{-1} B & 2B^T \\ 0 & -(W + C) \end{bmatrix}. \tag{7.26}$$

For both MINRES and GMRES, the stopping criterion was based on the relative norm of the unpreconditioned system residual, with stopping tolerance equal to 10^{-6}. The ideal versions of $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ were replaced by practical ones, where $A + B^T W^{-1} B$ was replaced by its incomplete Cholesky factors (MATLAB function `ichol`), with truncation threshold 10^{-4}.

From a computational point of view, it is convenient to choose the matrix W to be diagonal so that it is inexpensive to invert. For the effectiveness of $\mathcal{P}_{AD}$ and $\mathcal{T}_W$, the choice of W is crucial and optimal choices in terms of eigenvalue distribution of the preconditioned matrices have been discussed in [119] and in Section 3.2.4 of this thesis. In the presence of regularization in the $(2, 2)$ block, one of such choices is $W = C$. In fact, this setting has provided fast convergence in our numerical experiments but revealed to be unsatisfactory in terms of conditioning, particularly when solving the unreduced systems. Indeed, as the IP iterates approach the exact solution, some components of x, and hence some entries of C, tend to become very small. As a consequence, the condition number of $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ tends to become much larger than the condition number of $K_{3,reg}$. Clearly, this occurrence is especially undesired in the unreduced formulation since it is expected to be better conditioned than the reduced one.

Alternative approximations such as $W = \gamma I_{n+m}$ with γ equal to either the arithmetic mean, geometric mean or median of $\text{diag}(C)$ actually led to slower convergence of preconditioned MINRES and GMRES.

For the unreduced system $K_{3,reg}$, an effective choice of W was obtained by setting

$$W = C + \Gamma = C + \begin{bmatrix} \gamma_1 I_m & 0 \\ 0 & \gamma_2 I_n \end{bmatrix} = \begin{bmatrix} \delta I_m + \gamma_1 I_m & 0 \\ 0 & X + \gamma_2 I_n \end{bmatrix}, \tag{7.27}$$

where γ_1 and γ_2 are given by

$$\gamma_1 = \min \left\{ 1, \frac{1}{\|H + \rho I_n\|_F} \right\}, \qquad \gamma_2 = \gamma_1 \cdot \text{mean}(z),$$

and $\|\cdot\|_F$ is the Frobenius norm. This choice aims at exploiting the partitioning of the original matrix, by giving different weights at the two diagonal blocks.

For the reduced system $K_{2,reg}$, we instead set

$$W = C + \gamma I_m = (\delta + \gamma) I_m, \qquad \gamma = \frac{1}{\|H + \rho I_n + X^{-1} Z\|_F}. \tag{7.28}$$

We point out that, due to the magnitude of the elements in $X^{-1}Z$ for all problems we worked with, this choice of γ always gave a value less than one. For different data, the minimum between this value and one could be considered, as for $K_{3,reg}$.

A first set of experiments was performed running PDCO with accuracy on feasibility (`FeaTol`) equal to 10^{-6}; the complementary tolerance (`OptTol`) was set to 10^{-3} for problems in the STCQP group, and to 1 for problems CVXQP1-CVXQP3 [2]. We used different values of `OptTol` for the two data sets to generate sequences of systems with conditioning at most comparable with the reciprocal of the machine precision. The chosen values depended on the fact that problems STCQP1_a, STCQP1_b and STCQP2 are well-scaled, whereas in CVXQP1-CVXQP3 the largest value of x and z at the solution is $\mathcal{O}(1)$ and $\mathcal{O}(10^4)$, respectively.

Table 7.2 displays statistics for the solution of the sequences of unreduced systems by preconditioned MINRES and GMRES. We report the minimum, maximum and average (min/max/avg) number of linear iterations performed through the IP iterations and the minimum and maximum condition number of $K_{3,reg}$ and of the preconditioner during the IP iterations. Note that although different when using different inner solvers, the condition number of $K_{3,reg}$ did not vary significantly, therefore we report only the values obtained when using MATLAB's direct solver. All condition numbers were estimated[3] by the MATLAB function `condest` and only the first significant digit is reported. In Table 7.3 we compare the number of IP (outer) iterations (i.e., the number of linear systems to be solved in each sequence) and the execution times in seconds for solving the systems with either the direct solver, preconditioned GMRES or preconditioned MINRES. Execution times in the iterative solvers include the time needed to form the preconditioner, that is to compute the incomplete Cholesky factorization of the preconditioner (1,1) block. Tables 7.4 and 7.5 show analogous statistics for the reduced systems.

The reported results show that the condition number of $K_{3,reg}$ varies in a small range during the IP iterations and the maximum value attained is

[2] PDCO stops when all the following conditions, based on the KKT conditions (6.2), are satisfied:

$$\begin{aligned} \left\| H\hat{x} - J^T\hat{y} - \hat{z} - c \right\|_\infty &\leq \texttt{FeaTol} \\ \left\| Jx - b \right\|_\infty &\leq \texttt{FeaTol} \\ \max_i \; x_i z_i &\leq \texttt{OptTol} \end{aligned}$$

[3] The estimation is based on the original preconditioning matrix, before the incomplete Cholesky factor is computed.

		$K_{3,reg}$, $\mathcal{P}_{AD}$-MINRES		$K_{3,reg}$, $T_{\mathcal{W}}$-GMRES	
Problem	condest($K_{3,reg}$) min/max	Inner It min/max/avg	condest($P_{\mathcal{D}}$) min/max	Inner It min/max/avg	condest($T_{\mathcal{W}}$) min/max
CVXQP1	$5 \cdot 10^{11}/7 \cdot 10^{11}$	21/42/35.6	$4 \cdot 10^{13}/4 \cdot 10^{13}$	10/17/13.9	$2 \cdot 10^{13}/2 \cdot 10^{13}$
CVXQP2	$1 \cdot 10^{11}/1 \cdot 10^{11}$	20/41/30.6	$3 \cdot 10^{13}/3 \cdot 10^{13}$	9/19/13.3	$2 \cdot 10^{13}/2 \cdot 10^{13}$
CVXQP3	$4 \cdot 10^{11}/7 \cdot 10^{11}$	23/47/37.4	$5 \cdot 10^{13}/5 \cdot 10^{13}$	10/18/15.4	$3 \cdot 10^{13}/3 \cdot 10^{13}$
STCQP1_a	$1 \cdot 10^{10}/2 \cdot 10^{10}$	12/25/19.8	$2 \cdot 10^{10}/2 \cdot 10^{10}$	9/13/11.0	$2 \cdot 10^{10}/2 \cdot 10^{10}$
STCQP1_b	$2 \cdot 10^{10}/3 \cdot 10^{10}$	14/26/21.1	$6 \cdot 10^{10}/6 \cdot 10^{10}$	11/14/12.1	$6 \cdot 10^{10}/6 \cdot 10^{10}$
STCQP2	$6 \cdot 10^{6}$ /$7 \cdot 10^{6}$	11/25/20.6	$7 \cdot 10^{9}/7 \cdot 10^{9}$	8/12/10.0	$7 \cdot 10^{9}/7 \cdot 10^{9}$

Table 7.2: Results from the iterative solution of the unreduced systems, with preconditioners $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ for $K_{3,reg}$. Here $W = C + \Gamma$ (see (7.27)). Number of iterations and conditioning.

	$K_{3,reg}$, $\mathcal{P}_{AD}$-MINRES		$K_{3,reg}$, $\mathcal{T}_W$-GMRES		$K_{3,reg}$, Backslash	
Problem	Outer It	Time min/max/avg	Outer It	Time min/max/avg	Outer It	Time min/max/avg
CVXQP1	14	0.27/0.58/0.46	14	0.20/0.34/0.28	13	0.85/1.10/0.97
CVXQP2	15	0.13/0.22/0.18	15	0.10/0.18/0.14	13	0.56/0.81/0.64
CVXQP3	14	0.42/0.77/0.67	14	0.30/0.46/0.42	13	0.91/1.18/1.08
STCQP1_a	24	0.13/0.19/0.17	24	0.13/0.19/0.15	24	0.11/0.12/0.12
STCQP1_b	28	0.41/0.60/0.52	28	0.41/0.47/0.44	28	0.35/0.37/0.36
STCQP2	33	0.16/0.48/0.31	33	0.16/0.37/0.25	33	0.70/1.21/0.94

Table 7.3: Results from the iterative solution of the unreduced systems, with preconditioners $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ for $K_{3,reg}$. Here $W = C + \Gamma$ (see (7.27)). Number of linear systems and CPU times in seconds for their solution.

of several orders of magnitude smaller than that of $K_{2,reg}$. The conditioning of $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ for $K_{3,reg}$ shows little changes too, and remains considerably smaller than the conditioning of the preconditioners for $K_{2,reg}$ for the STCQP dataset. As expected, $K_{2,reg}$ is increasingly ill-conditioned, together with the corresponding matrices $\mathcal{P}_{AD}$ and $\mathcal{T}_W$. In both formulations, the preconditioners $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ are very effective in solving the systems. Preconditioner $\mathcal{T}_W$ is more efficient than $\mathcal{P}_{AD}$ in terms of linear iterations and provides some savings in the computational time when the unreduced formulation is used. It is also interesting to note that the iterative solution of the systems does not deteriorate the overall performance of the IP method, as shown by the comparison of the outer iterations in Tables 7.3 and 7.5. Moreover, in most runs, the use of the reduced formulation slightly increases the number of outer iterations. This might be related to the fact that in the reduced case, Z is not estimated explicitly by the inner solver, but recovered from the reduction step.

The sparse direct solver is compiled in MATLAB whereas the implemented iterative solvers are based on interpreted, thus slower, commands. Nonethe-

		$K_{2,reg}$, $\mathcal{P}_{AD}$-MINRES		$K_{2,reg}$, $\mathcal{T}_W$-GMRES	
Problem	condest($K_{2,reg}$) min/max	Inner It min/max/avg	condest($\mathcal{P}_{AD}$) min/max	Inner It min/max/avg	condest($\mathcal{T}_W$) min/max
CVXQP1	$6 \cdot 10^{11}/2 \cdot 10^{16}$	17/26/21.6	$5 \cdot 10^{13}/7 \cdot 10^{13}$	6/10/8.7	$3 \cdot 10^{13}/3 \cdot 10^{15}$
CVXQP2	$1 \cdot 10^{11}/2 \cdot 10^{15}$	12/19/15.1	$3 \cdot 10^{13}/1 \cdot 10^{16}$	6/ 9/7.7	$2 \cdot 10^{13}/6 \cdot 10^{15}$
CVXQP3	$7 \cdot 10^{11}/5 \cdot 10^{15}$	20/28/23.3	$6 \cdot 10^{13}/5 \cdot 10^{15}$	8/11/9.6	$4 \cdot 10^{13}/2 \cdot 10^{15}$
STCQP1_a	$1 \cdot 10^{10}/7 \cdot 10^{15}$	4/11/8.6	$2 \cdot 10^{10}/6 \cdot 10^{15}$	3/ 9/6.8	$2 \cdot 10^{10}/3 \cdot 10^{15}$
STCQP1_b	$2 \cdot 10^{10}/2 \cdot 10^{16}$	4/13/9.9	$6 \cdot 10^{10}/9 \cdot 10^{15}$	3/10/8.0	$6 \cdot 10^{10}/4 \cdot 10^{15}$
STCQP2	$6 \cdot 10^{6}$ /$4 \cdot 10^{12}$	4/11/7.5	$7 \cdot 10^{9}/1 \cdot 10^{15}$	3/ 8/5.4	$7 \cdot 10^{9}/7 \cdot 10^{14}$

Table 7.4: Results from the iterative solution of the reduced systems, with preconditioners $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ for $K_{2,reg}$. Here $W = (\delta + \gamma)I_m$ (see (7.28)). Number of iterations and conditioning.

	$K_{2,reg}$, $\mathcal{P}_{AD}$-MINRES		$K_{2,reg}$, $\mathcal{T}_W$-GMRES		$K_{2,reg}$, Backslash	
Problem	Outer It	Time min/max/avg	Outer It	Time min/max/avg	Outer It	Time min/max/avg
CVXQP1	17	0.22/0.32/0.27	17	0.16/0.23/0.19	13	0.76/0.96/0.87
CVXQP2	17	0.08/0.12/0.09	17	0.06/0.15/0.07	13	0.58/0.71/0.64
CVXQP3	17	0.36/0.56/0.44	17	0.26/0.39/0.30	13	0.84/1.02/0.95
STCQP1_a	26	0.08/0.13/0.10	26	0.08/0.16/0.11	24	0.10/0.22/0.21
STCQP1_b	30	0.27/0.41/0.35	30	0.27/0.40/0.35	28	0.30/0.80/0.77
STCQP2	33	0.09/0.35/0.18	33	0.09/0.33/0.18	33	0.74/0.83/0.77

Table 7.5: Results from the iterative solution of the unreduced systems, with preconditioners $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ for $K_{2,reg}$. Here $W = (\delta + \gamma)I_m$ (see (7.28)). Number of linear systems and CPU times in seconds for their solution.

less, the iterative methods are able to largely overcome this disadvantage, leading to an overall much better performance than with the sparse direct solver: Table 7.6 displays the total execution time for solving the sequences of systems either by preconditioned GMRES, or by the sparse direct solver (MATLAB's backslash operator). Except for STCQP1_a and STCQP1_b with $K_{3,reg}$, the use of preconditioned iterative solvers speeds the solution of the linear algebra phase. Moreover, we notice that the fastest runs are those performed with the reduced systems. This last fact is not surprising, as preconditioned GMRES behaves very similarly in the two formulations in terms of number of iterations, while the reduced formulation deals with vectors of smaller dimensions.

On the other hand, the unreduced formulation seems to offer some advantages when more stringent complementary tolerances are used. Indeed, we performed an additional set of experiments with problems STCQP1_a, STCQP1_b, and STCQP2 using a complementary tolerance equal to 10^{-6}, and with problems CVXQP1-CVXQP3 using a tolerance equal to 10^{-4}. The considered smaller tolerances are still within the suggested complementary toler-

	Total execution times			
	$K_{3,reg}$		$K_{2,reg}$	
Problem	$\mathcal{T}_W$-GMRES	Backslash	$\mathcal{T}_W$-GMRES	Backslash
CVXQP1	3.89	12.62	3.15	11.31
CVXQP2	2.05	8.37	1.26	8.31
CVXQP3	5.91	13.98	5.02	12.40
STCQP1_a	3.63	2.81	2.82	5.10
STCQP1_b	12.21	10.02	10.53	21.42
STCQP2	8.31	30.96	5.90	25.48

Table 7.6: Total execution times for the sequence of systems generated during the IP method. Preconditioned GMRES and MATLAB sparse direct solver for the unreduced (left) and reduced (right) systems.

ance intervals suggested in PDCO. PDCO implemented with $K_{2,reg}$ and solved by preconditioned GMRES failed to solve CVXQP1-CVXQP3, whereas the other approaches succeeded. The results in Table 7.9 show the conditioning of the matrices and the solution statistics; as before, the condition numbers reported refer to the implementation with MATLAB's direct solver. Failures are indicated with the symbol '*'. We omit statistics on the number of linear iterations performed since the behavior of the iterative solver is very similar to that of the previous set of experiments. We observe that the conditioning of $K_{3,reg}$ remains bounded and PDCO implemented with $K_{3,reg}$ is as robust as PDCO with direct solver, though faster. On the contrary, $K_{2,reg}$ tends to become numerically singular and the implementation of PDCO with preconditioned GMRES fails in three runs out of six since the maximum number of linesearch backtracks is reached. These failures occurred at the late stage of the IP iterations and seem to indicate lack of precision in the computation of the steps with the current tolerance.

Since the cost of the application of $\mathcal{P}_{AD}$ and $\mathcal{P}_{AT}$ is strongly related to the sparsity level of the (1,1) block $A + B^T W^{-1} B$, and, more directly, of its incomplete Cholesky factor L_{IC}. For this reason in Table 7.3 we report minimum and maximum number of nonzero entries of L_{IC}, throughout the IP iteration, for $K_{3,reg}$ amd $K_{2,reg}$. We also show the number of nonzero entries of $A+B^T W^{-1} B$, which is reported only once for each problem as it is independent of the IP iteration and of whether we are using the reduced or unreduced formulation. This can be seen by exploiting the actual structure of A, B and W in the two cases. The number of nonzero entries of the (complete) Cholesky factor of $A+B^T W^{-1} B$ is also reported for a comparison. For each problem, the variation of this number with respect to iteration numeber and formulation is negligeable, and the number reported refers to

the first iteration in the unreduced formulation. For problems CVXQP1 and CVXQP3 the number of nonzero entries in the incomplete factor is quite large, while in other cases it remains within a factor of two from the number of nonzero entries in the original matrix.

It is worthwile to compare the performance of our strategies with the one observed using the constraint preconditioner $\mathcal{P}_C$ coupled with GMRES, as this preconditioner is very popular in the context of KKT systems. We do this for the unreduced formulation, and to approximate the ideal preconditioner $\mathcal{P}_C$ defined in (7.15) we again used an Incomplete Cholesky factorization to solve the system with the Schur complement. For problems of the STCQP family, we set 10^{-4} as dropping tolerance for `ichol`, while for problems CVXQP1-CVXQP3 the stricter tolerance 10^{-6} was needed to avoid failures of `ichol`. The results are shown in Table 7.7. In one problem, we observed that the maximum number of GMRES iteration, set to 50, was reached.

	$K_{3,reg}$, $\mathcal{P}_C$-GMRES		
Problem	Outer It	Inner It min/max/avg	Total Time
CVXQP1	14	9/29/20.4	2.29
CVXQP2	14	11/27/21.0	1.94
CVXQP3	14	9/31/22.1	3.93
STCQP1_a	24	14/44/29.7	11.93
STCQP1_b	28	28/50/36.7	39.65
STCQP2	33	3/25/14.2	6.89

Table 7.7: Results from the iterative solution of the unreduced system using the constraint preconditioner $\mathcal{P}_C$ and GMRES

Since the cost of the application of the preconditioner is strongly related to the sparsity level of the (1,1) block $A + B^T W^{-1} B$, and, more directly, of its incomplete Cholesky factor L. For this reason in Table 7.3 we report minimum and maximum number of nonzero entries of L, throughout the IP iteration, for $K_{3,reg}$ amd $K_{2,reg}$. We also show the number of nonzero entries of $A + B^T W^{-1} B$, which is repoted only once for each proble,m as it is independent of the IP iteration and of whether we are using the reduced or unreduced formulation. This can be seen by exploiting the actual structure of A, B and W in the two cases. The number of nonzero entries of the (complete) Cholesky factor of $A + B^T W^{-1} B$ is also reported for a comparison. For each problem, the variation of this number with respect to iteration numeber and formulation is negligeable, and the number reported refers to the first iteration in the unreduced formulation. For problems CVXQP1 and CVXQP3 the number of nonzero entries in the incomplete factor is quite

large, while in other cases it remains within a factor of two from the number of nonzero entries in the original matrix.

Problem	nnz($A + B^T W^{-1} B$)	nnz(L_C)	nnz(L_{IC})			
			Unreduced		Reduced	
			min	max	min	max
CVXQP1	99946	233440	282173	355051	264721	346446
CVXQP2	84960	749178	104399	132270	75771	128860
CVXQP3	114924	1444660	426783	539124	411303	530397
STCQP1_a	106629	3662714	154535	160891	143434	160728
STCQP1_b	229507	2559272	412669	432796	394330	432795
STCQP2	330033	3984326	112657	371027	63259	370949

Table 7.8: Sparisty of the preconditioner (1,1) block $A + B^T W^{-1} B$ and its complete and incomplete Choelsky factor, respectively L_C and L_{IC}, throughout the IP iteration.

7.3.1 Updating technique for the (1,1) block of $\mathcal{P}_{AD}$ and $\mathcal{T}_W$

In this section we explore a different approach for approximating the (1,1) block of the preconditioners $\mathcal{P}_{AD}$ and $\mathcal{T}_W$. Instead of computing the Incomplete Cholesky factorization of the (1,1) block of each preconditioner from scratch, we exploit an updating preconditioner technique proposed in [9].

Updating techniques aim at building preconditioners for a sequence of matrices at a low computational cost. Let $\{A_k\} = \{A_0 + \Sigma_k\}$, $k \geq 1$, be a sequence where A_0 is a symmetric positive definite *seed* matrix, available in factorized LDL^T form, and Σ_k is a diagonal positive semidefinite matrix. Then, the LDL^T factorization of matrices A_k can be formed by updating the factorization of the seed matrix as proposed in [9]. Given the factorization $A_0 = L_0 D_0 L_0^T$, the factorization $L_k D_k L_k^T$ of A_k is obtained by updating that of A_0, at a cost which is linear in the number of nonzero entries of Σ_k for building D_k, and linear in the number of nonzero entries of L_0 for building L_k. The appealing feature of such an approach is that the Incomplete Cholesky factorization is computed only for the seed matrix; successively, its update is expected to be cheaper than the computation of a new incomplete factorization from scratch. Thus, as long as the updated approximate factorization remains sufficiently accurate, computational time savings can be obtained.

		$K_{3,reg}$, $\mathcal{T}_W$-GMRES		$K_{3,reg}$, Backslash	
Problem	condest($K_{3,reg}$) min/max	Outer It	Time min/max/avg	Outer It	Time min/max/avg
CVXQP1	$4 \cdot 10^{11}/7 \cdot 10^{11}$	20	0.21/0.46/0.31	20	0.81/1.40/1.08
CVXQP2	$1 \cdot 10^{11}/1 \cdot 10^{11}$	21	0.10/0.36/0.17	21	0.54/0.90/0.72
CVXQP3	$4 \cdot 10^{11}/7 \cdot 10^{11}$	20	0.31/0.97/0.58	20	1.09/1.69/1.24
STCQP1_a	$1 \cdot 10^{10}/2 \cdot 10^{10}$	30	0.13/0.17/0.15	30	0.11/0.12/0.11
STCQP1_b	$2 \cdot 10^{10}/3 \cdot 10^{10}$	34	0.39/0.48/0.44	34	0.35/0.37/0.36
STCQP2	$6 \cdot 10^{6}$ /$7 \cdot 10^{6}$	39	0.15/0.39/0.25	39	0.70/1.22/0.98

		$K_{2,reg}$, $\mathcal{T}_W$-GMRES		$K_{2,reg}$, Backslash	
Problem	condest($K_{2,reg}$) min/max	Outer It	Time min/max/avg	Outer It	Time min/max/avg
CVXQP1	$5 \cdot 10^{11}/7 \cdot 10^{19}$	*	*	20	0.76/1.02/0.91
CVXQP2	$1 \cdot 10^{11}/4 \cdot 10^{19}$	*	*	21	0.55/0.74/0.67
CVXQP3	$6 \cdot 10^{11}/4 \cdot 10^{19}$	*	*	36	0.84/1.11/0.98
STCQP1_a	$1 \cdot 10^{10}/1 \cdot 10^{19}$	30	0.07/0.13/0.10	30	0.10/0.22/0.21
STCQP1_b	$2 \cdot 10^{10}/2 \cdot 10^{19}$	34	0.24/0.38/0.33	34	0.31/0.82/0.78
STCQP2	$6 \cdot 10^{6}$ /$7 \cdot 10^{15}$	39	0.09/0.33/0.16	39	0.74/0.81/0.78

Table 7.9: Results from the iterative and direct solution of the unreduced systems (upper table) and reduced ones (lower table) obtained with more stringent complementarity tolerances. Conditioning of the matrices, number of linear systems and CPU times in seconds. Failures of PDCO are indicated by '*'.

The sequence of the (1,1) blocks of both $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ may fit into the sequence of matrices described above. In the unreduced setting, the (1,1) block of both $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ reads

$$H + \rho I_n + \frac{1}{\delta + \gamma_1} J^T J + Z \left(X + \gamma_2 I_n\right)^{-1}.$$

Since we set γ_1 constant throughout the outer IP iterations, matrix $H + \rho I_n + \frac{1}{\delta + \gamma_1} J^T J$ is fixed and may play the role of the seed matrix A_0, while $Z \left(X + \gamma_2 I_n\right)^{-1}$ is a diagonal and positive definite matrix which changes at every IP iteration and may represent the diagonal modification Σ_k.

If we consider the reduced formulation, the (1,1) block of $\mathcal{P}_{AD}$ and $\mathcal{T}_W$ reads

$$H + \rho I_n + \frac{1}{\delta + \gamma} J^T J + X^{-1} Z,$$

and in the previous experiments γ changed at every IP iteration. In order to apply the updating strategy we reconsidered the choice of this parameter and fixed $\gamma = \frac{1}{\left\| H + \rho I_n + X_0^{-1} Z_0 \right\|_F}$, where X_0 and Z_0 are the diagonal matrices

	$K_{3,reg}$			
	$\mathcal{P}_{AD}$-MINRES		$\mathcal{T}_W$-GMRES	
Problem	Inner It min/max/avg	Total execution time	Inner It min/max/avg	Total execution time
STCQP1_a	12/27/21.7	3.10	9/13/11.7	2.45
STCQP1_b	14/30/23.2	10.49	11/15/12.8	7.87
STCQP2	11/27/22.6	10.26	8/13/10.8	7.06

	$K_{2,reg}$			
	$\mathcal{P}_{AD}$-MINRES		$\mathcal{T}_W$-GMRES	
Problem	Inner It min/max/avg	Total execution time	Inner It min/max/avg	Total execution time
STCQP1_a	3/16/10.6	1.49	3/10/7.9	1.50
STCQP1_b	3/18/12.4	5.80	3/12/9.5	5.50
STCQP2	1/16/11.0	5.23	2/10/7.7	4.78

Table 7.10: Numerical results for the unreduced systems (upper table) and reduced ones (lower table) obtained using the updating technique to approximate the (1,1) block of $\mathcal{P}_{AD}$ and $\mathcal{T}_W$

associated with the initial guesses x_0 and z_0. This way, $H + \rho I_n + \dfrac{1}{\delta + \gamma} J^T J$ can be used as the seed matrix and $X^{-1}Z$ is the positive definite diagonal modification Σ_k.

Experiments were performed with the first practical update from the paradigm given in [9, Section 3]. Table 7.10 displays the results obtained on the sequences of systems arising from problems STCQP1_a, STCQP1_b and STCQP2 with `OptTol` equal to 10^{-3}, i.e. the same sequences reported in Tables 7.2–7.6. The minimum, maximum and average (min/max/avg) number of iterations performed and the total execution times for both $K_{3,reg}$ and $K_{2,reg}$ are reported. Comparing these results with those in Tables 7.2–7.6, we note that the number of linear iterations is only slightly affected by the use of the updating strategy, while the computational time is reduced.

On the contrary, the updating strategy was not beneficial in the solution of problems from the CVXQP family. Loss of accuracy in the approximated factorization with respect to the Incomplete Cholesky factorization was observed and this deteriorated the performance of the preconditioned iterative solvers. Combining the updating strategy with low-rank matrix corrections may enhance the efficiency of the procedure [10], but this issue is beyond the scope of this chapter.

7.4 Conclusions

Classically, the high sparsity structure of KKT systems arising in the numerical solution of quadratic programming problems by means of interior point method encourages the use of reduction strategies before solving the systems. For stability purposes, however, the unreduced formulation may be appealing. Following previous analysis in [46, 49, 66], the aim of Chapter 6 and Chapter 7 of this work was to explore the actual advantages of the unreduced strategy.

In this chapter we have shown that when preconditioning is employed, much care must be put in even devising mathematically different problems. Moreover, once specifically designed preconditioners are used, and as long as implementations with the unreduced and reduced formulations are successful, the performance seems in favor of the latter in view of the smaller dimensions. The use of recently developed updating techniques provides similar saving for both formulations, thus not changing their computational comparison. On the other hand, the unreduced formulation may maintain better spectral properties than the reduced one and its use may enhance the robustness of the interior point solver when the reduced systems become severely ill-conditioned, or even numerically singular.

Numerical results on the considered datasets show that the upper triangular preconditioner is somewhat more efficient than the block diagonal one, although it requires additional memory allocations if the optimal solver GMRES is employed. In both preconditioners, our selection of generic augmentation matrix W seems to be well suited for all test cases.

Appendix A

Technical proofs

In this appendix we prove the most technical results used throughout this thesis.

Proof of Lemma 6.3.3. Matrix B has dimension $(m+q)\times n$ where q is the cardinality of the active set $\mathcal{A}_*$ at $\hat{x}$. By the LICQ condition, $q \leq n-m$ and $J_{\mathcal{I}_*}^T$ has full column rank. Consequently, B^T has full column rank.

We provide estimates for $\sigma_{\max}(B)$ and $\sigma_{\min}(B)$ by using the relations $\sigma_{\max}^2(B) = \lambda_{\max}(BB^T)$ and $\sigma_{\min}^2(B) = \lambda_{\min}(BB^T)$ and considering the eigenvalue problem for BB^T, that is

$$\begin{bmatrix} J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + J_{\mathcal{I}_*}J_{\mathcal{I}_*}^T & -J_{\mathcal{A}_*}Z_{\mathcal{A}_*}^{\frac{1}{2}} \\ -Z_{\mathcal{A}_*}^{\frac{1}{2}}J_{\mathcal{A}_*}^T & Z_{\mathcal{A}_*} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \lambda \begin{bmatrix} u \\ v \end{bmatrix}. \tag{A.1}$$

If $v = 0$, then from the first equation we find $\sigma_{\min}^2(J_{\mathcal{I}_*}) \leq \lambda \leq \sigma_{\max}^2(J_{\mathcal{I}_*})$. Then, we first focus on $\sigma_{\min}(B)$ and consider the case where $v \neq 0$ and $\lambda < \sigma_{\min}^2(J_{\mathcal{I}_*})$, otherwise $\sigma_{\min}^2(J_{\mathcal{I}_*})$ is the requested bound. By the first block equation in (A.1)

$$u = \left(J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + J_{\mathcal{I}_*}J_{\mathcal{I}_*}^T - \lambda I_m\right)^{-1} J_{\mathcal{A}_*}Z_{\mathcal{A}_*}^{\frac{1}{2}}v.$$

Then, the second block equation of (A.1) becomes

$$-Z_{\mathcal{A}_*}^{\frac{1}{2}}J_{\mathcal{A}_*}^T\left(J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + J_{\mathcal{I}_*}J_{\mathcal{I}_*}^T - \lambda I_m\right)^{-1} J_{\mathcal{A}_*}Z_{\mathcal{A}_*}^{\frac{1}{2}}v + Z_{\mathcal{A}_*}v - \lambda v = 0.$$

and premultiplying it by v^T we get

$$v^T Z_{\mathcal{A}_*}^{\frac{1}{2}}\left[I_q - J_{\mathcal{A}_*}^T\left(J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + J_{\mathcal{I}_*}J_{\mathcal{I}_*}^T - \lambda I_m\right)^{-1} J_{\mathcal{A}_*}\right] Z_{\mathcal{A}_*}^{\frac{1}{2}}v - \lambda\|v\|^2 = 0, \tag{A.2}$$

Now we observe that

$$\left(J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + J_{\mathcal{I}_*}J_{\mathcal{I}_*}^T - \lambda I_m\right) \succeq \left(J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + (\sigma_{\min}^2(J_{\mathcal{I}_*}) - \lambda)I_m\right),$$

and

$$J_{\mathcal{A}_*}^T\left(J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + J_{\mathcal{I}_*}J_{\mathcal{I}_*}^T - \lambda I_m\right)^{-1}J_{\mathcal{A}_*} \preceq J_{\mathcal{A}_*}^T\left(J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + (\sigma_{\min}^2(J_{\mathcal{I}_*}) - \lambda)I_m\right)^{-1}J_{\mathcal{A}_*}.$$

Then

$$\begin{aligned} w^T J_{\mathcal{A}_*}^T\left(J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + (\sigma_{\min}^2(J_{\mathcal{I}_*}) - \lambda)I_m\right)^{-1}J_{\mathcal{A}_*}w &\leq \max_i \frac{\sigma_i^2(J_{\mathcal{A}_*})}{\sigma_i^2(J_{\mathcal{A}_*}) + \sigma_{\min}^2(J_{\mathcal{I}_*}) - \lambda}\|w\|^2 \\ &= \frac{\sigma_{\max}^2(J_{\mathcal{A}_*})}{\sigma_{\max}^2(J_{\mathcal{A}_*}) + \sigma_{\min}^2(J_{\mathcal{I}_*}) - \lambda}\|w\|^2 \end{aligned}$$

where $w \in \mathbb{R}^q$ and (A.2) gives

$$(z_{\mathcal{A}_*})_{\min} - \frac{(z_{\mathcal{A}_*})_{\min}\sigma_{\max}^2(J_{\mathcal{A}_*})}{\sigma_{\max}^2(J_{\mathcal{A}_*}) + \sigma_{\min}^2(J_{\mathcal{I}_*}) - \lambda} - \lambda \leq 0,$$

which is equivalent to

$$r(\lambda) = \lambda^2 - (\sigma_{\max}^2(J_{\mathcal{A}_*}) + \sigma_{\min}^2(J_{\mathcal{I}_*}) + (z_{\mathcal{A}_*})_{\min})\lambda + (z_{\mathcal{A}_*})_{\min}\sigma_{\min}^2(J_{\mathcal{I}_*}) \leq 0.$$

Since $r(0) > 0$ and $r\left(\sigma_{\min}^2(J_{\mathcal{I}_*})\right) < 0$, then $\sigma_{\min}^2(J_{\mathcal{I}_*})$ is greater than the smallest root of $r(\lambda)$ and the stated bound on $\sigma_{\min}(B)$ follows.

Finally, if $w \in \mathbb{R}^q$, $\hat{w} \in \mathbb{R}^{n-q}$ we have

$$\left\|B\begin{bmatrix} w \\ \hat{w}\end{bmatrix}\right\|^2 = \left\|J\begin{bmatrix} w \\ \hat{w}\end{bmatrix}\right\|^2 + \left\|Z_{\mathcal{A}_*}^{\frac{1}{2}}w\right\|^2 \leq \left(\|J\|^2 + \|Z_{\mathcal{A}_*}\|\right)\left\|\begin{bmatrix} w \\ \hat{w}\end{bmatrix}\right\|^2,$$

from which the looser bound for $\sigma_{\max}(B)$ follows.

The sharper bound for $\sigma_{\max}(B)$, although more complicated, can be derived as follows. We start again from (A.1) and suppose $\lambda > \sigma_{\max}^2$, which in particular implies $v \neq 0$. As before, we find u from the first equation (note that $J_{\mathcal{A}_*}J_{\mathcal{A}_*}^T + J_{\mathcal{I}_*}J_{\mathcal{I}_*}^T = JJ^T$) and substitute into the second one. Premultiplying for v^T we obtain again equation (A.2). This time we use

$$J_{\mathcal{A}_*}^T\left(\lambda I_m - JJ^T\right)^{-1}J_{\mathcal{A}_*} \preceq \frac{\sigma_{\max}^2(J_{\mathcal{A}_*})}{\lambda - \sigma_{\max}^2}I_q.$$

Proceeding as above, we derive the inequality

$$(z_{\mathcal{A}_*})_{\max} + \frac{(z_{\mathcal{A}_*})_{\max}\sigma_{\max}^2(J_{\mathcal{A}_*})}{\lambda - \sigma_{\max}^2} - \lambda \geq 0,$$

which is equivalent to

$$q(\lambda) := \lambda^2 - \left((z_{\mathcal{A}_*})_{\max} + \sigma^2_{\max}\right)\lambda + (z_{\mathcal{A}_*})_{\max}\left(\sigma^2_{\max} - \sigma^2_{\max}(J_{\mathcal{A}_*})\right) \leq 0.$$

Since $q\left(\sigma^2_{\max}\right) < 0$, then $\sigma^2_{\max}$ is smaller than the largest root of $q(\lambda)$. We thus obtain

$$\sigma^2_{\max}(B) \leq \frac{1}{2}\left((z_{\mathcal{A}_*})_{\max} + \sigma^2_{\max} + \sqrt{\left((z_{\mathcal{A}_*})_{\max} - \sigma^2_{\max}\right)^2 + 4(z_{\mathcal{A}_*})_{\max}\sigma^2_{\max}(J_{\mathcal{A}_*})}\right).$$

Since $\sigma^2_{\max} - \sigma^2_{\max}(J_{\mathcal{A}_*}) \geq 0$, this bound is sharper than the simpler one.

Proof of Lemma 5.3.5 From $F + F^T \succeq 0$ it also follows that $F + I$ is nonsingular.

i) We consider the eigenvalue problem $(F+I)^{-1}(F-I)(F-I)^T(F+I)^{-T}x = \theta x$ with $\theta \geq 0$, or, equivalently, $(F-I)(F-I)^T y = \theta(F+I)(F+I)^T y$ with $y = (F+I)^{-T}x$. The largest eigenvalue coincides with $\|(F+I)^{-1}(F-I)\|^2$. We have $(F-I)(F-I)^T = FF^T + I - F - F^T$ and $(F+I)(F+I)^T = FF^T + I + F + F^T$. Substituting and rearranging terms gives

$$(1-\theta)(FF^T + I)y = (\theta+1)(F+F^T)y.$$

We multiply from the left by y^T. Since $FF^T + I \succ 0$, $F + F^T \succeq 0$ and $\theta + 1 > 0$, it must be that $1 - \theta \geq 0$, that is $\theta \leq 1$.

ii) We proceed in a similar way. Let us now consider $(F+I)^{-1}(F+F^T)(F+I)^{-T}x = \theta x$, with $\theta > 0$, which is equivalent to $(F+F^T)y = \theta(F+I)(F+I)^T y$, with $y = (F+I)^{-T}x$. Therefore, $(1-\theta)(F+F^T)y = \theta(FF^T+I)y$. We premultiply by y^T and rearrange to obtain

$$\frac{1-\theta}{\theta} = \frac{y^T(FF^T+I)y}{y^T(F+F^T)y}.$$

From the relation $(F-I)(F-I)^T \succeq 0$ it follows that $\dfrac{y^T(FF^T+I)y}{y^T(F+F^T)y} \geq 1$. Thus, $\dfrac{1-\theta}{\theta} \geq 1$ which implies $\theta \leq \dfrac{1}{2}$.

Proof of Proposition 5.3.7. Let $F = \sqrt{\nu}M^{-\frac{1}{2}}LM^{-\frac{1}{2}}$.

i) For $\alpha_u = 1$ and $\alpha_y = 0$ we have $\gamma_1 = 0$ and $\gamma_2 = 1$, so that $M^{-\frac{1}{2}}\mathbb{S}M^{-\frac{1}{2}} = FF^T + (I - \Pi)$, and

$$M^{-\frac{1}{2}}\widehat{\mathbb{S}}M^{-\frac{1}{2}} = (F + (I-\Pi))(F + (I-\Pi))^T,$$

where we used the fact that $(I-\Pi)^{\frac{1}{2}} = (I-\Pi)$. From $M^{-\frac{1}{2}}\mathbb{S}M^{-\frac{1}{2}}x = \lambda M^{-\frac{1}{2}}\widehat{\mathbb{S}}M^{-\frac{1}{2}}x$ we obtain for $y = M^{\frac{1}{2}}x$

$$(F + (I-\Pi))^{-1}(FF^T + (I-\Pi))(F + (I-\Pi))^{-T}y = \lambda y. \qquad \text{(A.3)}$$

Since F is nonsingular, we have

$$\begin{aligned}
&(F+(I-\Pi))^{-1}(FF^T+(I-\Pi))(F+(I-\Pi))^{-T}\\
&=(F+(I-\Pi))^{-1}F(I+F^{-1}(I-\Pi)(I-\Pi)F^{-T})F^T(F+(I-\Pi))^{-T}\\
&=(I+F^{-1}(I-\Pi))^{-1}(I+F^{-1}(I-\Pi)(I-\Pi)F^{-T})(I+F^{-1}(I-\Pi))^{-T}\\
&=:(I+Z)^{-1}(I+ZZ^T)(I+Z)^{-T},
\end{aligned}$$

with $Z=F^{-1}(I-\Pi)$. Therefore, from (A.3) it follows

$$\begin{aligned}
\lambda &\leq \|(I+Z)^{-1}(I+ZZ^T)(I+Z)^{-T}\| \leq \|(I+Z)^{-1}\|^2+\|(I+Z)^{-1}Z\|^2\\
&= \|(I+Z)^{-1}\|^2+\|I-(I+Z)^{-1}\|^2\\
&\leq \|(I+Z)^{-1}\|^2+(1+\|(I+Z)^{-1}\|)^2. \qquad \text{(A.4)}
\end{aligned}$$

We then recall that $Z=F^{-1}(I-\Pi)=\frac{1}{\sqrt{\nu}}M^{\frac{1}{2}}L^{-1}M^{\frac{1}{2}}(I-\Pi)$, so that

$$\begin{aligned}
\|(I+Z)^{-1}\| &= \|(I+\frac{1}{\sqrt{\nu}}M^{\frac{1}{2}}L^{-1}M^{\frac{1}{2}}(I-\Pi))^{-1}\|\\
&= \|M^{\frac{1}{2}}\left(\sqrt{\nu}L+M(I-\Pi)\right)^{-1}\sqrt{\nu}LM^{-\frac{1}{2}}\|.
\end{aligned}$$

To analyze the behavior for $\nu\to 0$, let us suppose that $L+L^T\succ 0$, and write $Z=\frac{1}{\sqrt{\nu}}\widetilde{F}^{-1}(I-\Pi)$; without loss of generality also assume that $I-\Pi=\mathrm{blkdiag}(I_\ell,0)$. The eigendecomposition of $\widetilde{F}^{-1}(I-\Pi)$ is given by[1] $\widetilde{F}^{-1}(I-\Pi)=X\Lambda X^{-1}$ where $\Lambda=\mathrm{diag}(\lambda_i)$ and $\lambda_i\in\mathrm{spec}((\widetilde{F}^{-1})_{11})\cup\{0\}$. Here $(\widetilde{F}^{-1})_{11}$ is the top left $\ell\times\ell$ block of $\widetilde{F}^{-1}$. Note that all eigenvalues of $(\widetilde{F}^{-1})_{11}$ have strictly positive real part, thanks to the condition $L+L^T\succ 0$. Therefore

$$\|(I+Z)^{-1}\| = \|X(I+\frac{1}{\sqrt{\nu}}\Lambda)^{-1}X^{-1}\|\leq\kappa(X)\max\left\{\frac{1}{\min\limits_{\lambda\in\mathrm{spec}((\widetilde{F}^{-1})_{11})}|1+\lambda/\sqrt{\nu}|},1\right\}.$$

We thus have

$$\max\left\{\frac{1}{\min\limits_{\lambda\in\mathrm{spec}((\widetilde{F}^{-1})_{11})}|1+\lambda/\sqrt{\nu}|},1\right\}\to 1 \qquad \textit{for} \qquad \nu\to 0,$$

so that $\|(I+Z)^{-1}\|\leq\eta\,\kappa(X)$ with $\eta\to 1$ for $\nu\to 0$.

[1] In the unlikely case of a Jordan decomposition, the proof proceeds with the maximum over norms of Jordan blocks inverses, which leads to the same final result.

ii) For $\alpha_u = 0$ and $\alpha_y = 1$ we have $\gamma_1 = 1$ and $\gamma_2 = 0$, so that $M^{-\frac{1}{2}}\mathbb{S}M^{-\frac{1}{2}} = F(I-\Pi)F^T + I$, and

$$M^{-\frac{1}{2}}\widehat{\mathbb{S}}M^{-\frac{1}{2}} = (F(I-\Pi)+I)(F(I-\Pi)+I)^T.$$

As before, setting this time $Z = F(I-\Pi)$ we obtain the bounds (A.4) for λ with $\|(I+Z)^{-1}\| = \|(I+\sqrt{\nu}M^{-\frac{1}{2}}LM^{-\frac{1}{2}}(I-\Pi))^{-1}\|$. Finally, it is apparent from the above expression that $\|(I+Z)^{-1}\| \to 1$ as $\nu \to 0$.

Bibliography

[1] Anna Altman and Jacek Gondzio. Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization. *Optim. Methods Softw.*, 11(1-4):275–302, 1999.

[2] Steven F. Ashby, Thomas A. Manteuffel, and Paul E. Saylor. A taxonomy for conjugate gradient methods. *SIAM J. Numer. Anal.*, 27(6):1542–1568, 1990.

[3] Owe Axelsson. Solution of linear systems of equations: iterative methods. In *Sparse matrix techniques*, pages 1–51. Springer, 1977.

[4] Owe Axelsson. Preconditioners for regularized saddle point matrices. *J. Numer. Math.*, 19(2):91–112, 2011.

[5] Zhong-Zhi Bai. Eigenvalue estimates for saddle point matrices of Hermitian and indefinite leading blocks. *J. Comput. Appl. Math.*, 237(1):295–306, 2013.

[6] Zhong-Zhi Bai, Michael K. Ng, and Zeng-Qi Wang. Constraint preconditioners for symmetric indefinite matrices. *SIAM J. Matrix Anal. Appl.*, 31(2):410–433, 2009.

[7] Bernhard Beckermann, Sergei A. Goreinov, and Eugene E. Tyrtyshnikov. Some remarks on the Elman estimate for GMRES. *SIAM J. Matrix Anal. Appl.*, 27(3):772–778, 2006.

[8] Stefania Bellavia. Inexact interior-point method. *J. Optim. Theory and Appl.*, 96(1):109–121, 1998.

[9] Stefania Bellavia, Valentina De Simone, Daniela Di Serafino, and Benedetta Morini. A preconditioning framework for sequences of diagonally modified linear systems arising in optimization. *SIAM J. Numer. Anal.*, 50(6):3280–3302, 2012.

[10] Stefania Bellavia, Valentina De Simone, Daniela Di Serafino, and Benedetta Morini. On the update of constraint preconditioners for regularized KKT systems. *Optimization Online, http://www.optimizationonline.org/DB HTML/2014/03/4283.html*, 2014.

[11] Michele Benzi, Gene H. Golub, and Jörg Liesen. Numerical solution of saddle point problems. *Acta Numer.*, 14:1–137, 2005.

[12] Michele Benzi, Eldad Haber, and Lauren Taralli. Multilevel algorithms for large-scale interior point methods. *SIAM J. Sci. Comput.*, 31(6):4152–4175, 2009.

[13] Michele Benzi and Maxim A. Olshanskii. An augmented Lagrangian-based approach to the Oseen problem. *SIAM J. Sci. Comput.*, 28(6):2095–2113, 2006.

[14] Michele Benzi, Maxim A. Olshanskii, and Zhen Wang. Modified augmented Lagrangian preconditioners for the incompressible Navier–Stokes equations. *Internat. J. Numer. Methods Fluids*, 66(4):486–508, 2011.

[15] Luca Bergamaschi, Jacek Gondzio, and Giovanni Zilli. Preconditioning indefinite systems in interior point methods for optimization. *Comput. Optim. Appl.*, 28(2):149–171, 2004.

[16] Maitine Bergounioux, Kazufumi Ito, and Karl Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM J. Control Optimi.*, 37(4):1176–1194, 1999.

[17] Alfio Borzì. Smoothers for control and state-constrained optimal control problems. *Comput. Vis. Sci.*, 11(1):59–66, 2008.

[18] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[19] Jonathan Boyle, Milan Mihajlović, and Jennifer Scott. HSL_MI20: an efficient AMG preconditioner for finite element problems in 3D. *Internat. J. Numer. Methods Engrg.*, 82(1):64–98, 2010.

[20] Dietrich Braess. *Finite elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, 2001.

[21] James H. Bramble and Joseph E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comp.*, 50(181):1–17, 1988.

[22] Susanne C. Brenner and Ridgway Scott. *The mathematical theory of finite element methods*. Springer, 2008.

[23] Sonia Cafieri, Marco D'Apuzzo, Valentina De Simone, and Daniela Di Serafino. On the iterative solution of KKT systems in potential reduction software for large-scale quadratic problems. *Comput. Optim. Appl.*, 38(1):27–45, 2007.

[24] Zhi-Hao Cao. Augmentation block preconditioners for saddle point-type matrices with singular (1, 1) blocks. *Numer. Linear Algebra Appl.*, 15(6):515–533, 2008.

[25] Zhi-Hao Cao. A note on spectrum distribution of constraint preconditioned generalized saddle point matrices. *Numer. Linear Algebra Appl.*, 16(6):503–516, 2009.

[26] Eduardo Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control Optim.*, 24(6):1309–1318, 1986.

[27] Jordi Castro and Jordi Cuesta. Quadratic regularizations in an interior-point method for primal block-angular problems. *Math. Program.*, 130(2):415–445, 2011.

[28] Guang-Hui Cheng, Ting-Zhu Huang, and Shu-Qian Shen. Block triangular preconditioners for the discretized time-harmonic Maxwell equations in mixed form. *Comput. Phys. Comm.*, 180(2):192–196, 2009.

[29] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley New York, 1983.

[30] Marco D'Apuzzo, Valentina De Simone, and Daniela di Serafino. On mutual impact of numerical linear algebra and large-scale optimization with focus on interior point methods. *Comput. Optim. Appl.*, 45(2):283–310, 2010.

[31] Elizabeth D. Dolan and Jorge J. Moré. Benchmarking optimization software with performance profiles. *Math. Program.*, 91(2):201–213, 2002.

[32] H. Sue Dollar. Constraint-style preconditioners for regularized saddle point problems. *SIAM J. Matrix Anal. Appl.*, 29(2):672–684, 2007.

[33] H. Sue Dollar, Nicholas I. M. Gould, Wil H. A. Schilders, and Andrew J. Wathen. Implicit-factorization preconditioning and iterative solvers for regularized saddle-point systems. *SIAM J. Matrix Anal. Appl.*, 28(1):170–189, 2006.

[34] H. Sue Dollar, Nicholas I. M. Gould, Wil H. A. Schilders, and Andrew J. Wathen. Using constraint preconditioners with regularized saddle-point problems. *Comput. Optim. Appl.*, 36(2-3):249–270, 2007.

[35] H. Sue Dollar, Nicholas I. M. Gould, Martin Stoll, and Andrew J. Wathen. Preconditioning saddle-point systems with applications in optimization. *SIAM J. Sci. Comput.*, 32(1):249–270, 2010.

[36] Carla Durazzi and Valeria Ruggiero. Indefinitely preconditioned conjugate gradient method for large sparse equality and inequality constrained quadratic problems. *Numer. Linear Algebra Appl.*, 10(8):673–688, 2003.

[37] Michael Eiermann. Fields of values and iterative methods. *Linear Algebra Appl.*, 180:167–197, 1993.

[38] Michael Eiermann and Oliver G. Ernst. Geometric aspects of the theory of Krylov subspace methods. *Acta Numer.*, 10:251–312, 2001.

[39] Stanley C. Eisenstat, Howard C. Elman, and Martin H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 20(2):345–357, 1983.

[40] Stanley C. Eisenstat and Homer F. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM J. Sci. Comput.*, 17(1):16–32, 1996.

[41] Howard C. Elman, Alison Ramage, and David J. Silvester. IFISS, a Matlab toolbox for modelling incompressible flow. *ACM Trans. Math. Software*, 33(2):14, 2007.

[42] Howard C. Elman, David J. Silvester, and Andrew J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Oxford University Press, second edition, 2014.

[43] Vance Faber and Thomas Manteuffel. Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. Numer. Anal.*, 21(2):352–362, 1984.

[44] Bernd Fischer, Alison Ramage, David J. Silvester, and Andrew J. Wathen. Minimum residual methods for augmented systems. *BIT*, 38(3):527–543, 1998.

[45] Roger Fletcher. Conjugate gradient methods for indefinite systems. In *Numerical Analysis*, pages 73–89. Springer, 1976.

[46] Anders Forsgren. Inertia-controlling factorizations for optimization algorithms. *Appl. Numer. Math*, 43(1):91–107, 2002.

[47] Anders Forsgren, Philip E. Gill, and Joseph R. Shinnerl. Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization. *SIAM J. Matrix Anal. Appl.*, 17(1):187–211, 1996.

[48] Anders Forsgren, Philip E. Gill, and Margaret H. Wright. Interior methods for nonlinear optimization. *SIAM Rev.*, 44(4):525–597, 2002.

[49] Anders Forsgren, Alison. Ramage, and Joshua D. Griffin. Iterative solution of augmented systems arising in interior methods. *SIAM J. Optim.*, 18(2):666–690, 2007.

[50] Robert Fourer and Sanjay Mehrotra. Solving symmetric indefinite systems in an interior-point method for linear programming. *Math. Program.*, 62(1-3):15–39, 1993.

[51] Roland W. Freund. A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems. *SIAM J. Sci. Comput.*, 14(2):470–482, 1993.

[52] Roland W Freund and Noël M Nachtigal. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numer. Math.*, 60(1):315–339, 1991.

[53] Michael P. Friedlander and Dominique Orban. A primal–dual regularized interior-point method for convex quadratic programs. *Math. Program. Comput.*, 4(1):71–107, 2012.

[54] Gene H. Golub and Chen Greif. On solving block-structured indefinite linear systems. *SIAM J. Sci. Comput.*, 24(6):2076–2092 (electronic), 2003.

[55] Gene H. Golub, Chen Greif, and James M. Varah. An algebraic analysis of a block diagonal preconditioner for saddle point systems. *SIAM J. Matrix Anal. Appl.*, 27(3):779–792, 2005.

[56] Gene H. Golub, David J. Silvester, and Andrew J. Wathen. Diagonal dominance and positive definiteness of upwind approximations for advection diffusion problems. In *Numerical analysis*, pages 125–131. World Sci. Publ., River Edge, NJ, 1996.

[57] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. JHU Press, 2012.

[58] Jacek Gondzio. Interior point methods 25 years later. *European J. Oper. Res.*, 218(3):587–601, 2012.

[59] Jacek Gondzio. Matrix-free interior point method. *Comput. Optim. Appl.*, 51(2):457–480, 2012.

[60] Jacek Gondzio. Convergence analysis of an inexact feasible interior point method for convex quadratic programming. *SIAM J. Optim.*, 23(3):1510–1527, 2013.

[61] Nicholas I. M. Gould and Valeria Simoncini. Spectral analysis of saddle point matrices with indefinite leading blocks. *SIAM J. Matrix Anal. Appl.*, 31(3):1152–1171, 2009.

[62] Nicholas I.M. Gould, Dominique Orban, and Philippe L. Toint. CUTEr and SifDec: A constrained and unconstrained testing environment, revisited. *ACM Trans. Math. Software*, 29(4):373–394, 2003.

[63] Anne Greenbaum. Comparison of splittings used with the conjugate gradient algorithm. *Numer. Math.*, 33(2):181–193, 1979.

[64] Anne Greenbaum. *Iterative methods for solving linear systems*. SIAM, 1997.

[65] Anne Greenbaum, Vlastimil Pták, and Zdeněk Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM J. Matrix Anal. Appl.*, 17(3):465–469, 1996.

[66] Chen Greif, Erin Moulding, and Dominique Orban. Bounds on eigenvalues of matrices arising from interior-point methods. *SIAM J. Optim.*, 24(1):49–83, 2014.

[67] Chen Greif and Michael L. Overton. An analysis of low-rank modifications of preconditioners for saddle point systems. *Electron. Trans. Numer. Anal.*, 37:307–320, 2010.

[68] Chen Greif and Dominik Schötzau. Preconditioners for saddle point linear systems with highly singular (1, 1) blocks. *ETNA, Special Volume on Saddle Point Problems*, 22:114–121, 2006.

[69] Roland Herzog and Ekkehard Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints. *SIAM J. Matrix Anal. Appl.*, 31(5):2291–2317, 2010.

[70] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952.

[71] Michael Hintermüller, Kazufumi Ito, and Karl Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3):865–888 (2003), 2002.

[72] Ting-Zhu Huang, Li-Tao Zhang, Tong-Xiang Gu, and Xian-Yu Zuo. New block triangular preconditioner for linear systems arising from the discretized time-harmonic Maxwell equations. *Comput. Phys. Comm.*, 180(10):1853–1859, 2009.

[73] Ilse C. F. Ipsen. A note on preconditioning nonsymmetric matrices. *SIAM J. Sci. Comput.*, 23(3):1050–1051, 2001.

[74] Kazufumi Ito and Karl Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems Control Lett.*, 50(3):221–228, 2003.

[75] Wayne D. Joubert and David M. Young. Necessary and sufficient conditions for the simplification of generalized conjugate-gradient algorithms. *Linear Algebra Appl.*, 88/89:449–485, 1987.

[76] Christian Kanzow. Inexact semismooth Newton methods for large-scale complementarity problems. *Optim. Methods Softw.*, 19(3-4):309–325, June 2004.

[77] Carsten Keller, Nicholas I. M. Gould, and Andrew J. Wathen. Constraint preconditioning for indefinite linear systems. *SIAM J. Matrix Anal. Appl.*, 21(4):1300–1317, 2000.

[78] Wolfgang Krendl, Valeria Simoncini, and Walter Zulehner. Stability estimates and structural spectral properties of saddle point problems. *Numer. Math.*, 124(1):183–213, 2013.

[79] Piotr Krzyżanowski. On block preconditioners for saddle point problems with singular or indefinite (1, 1) block. *Numer. Linear Algebra Appl.*, 18(1):123–140, 2011.

[80] Karl Kunisch and Arnd Rösch. Primal-dual active set strategy for a general class of constrained optimal control problems. *SIAM J. Optim.*, 13(2):321–334, 2002.

[81] Ren-Cang Li and Wei Zhang. The rate of convergence of GMRES on a tridiagonal Toeplitz linear system. *Numer Math.*, 112(2):267–293, 2009.

[82] Jörg Liesen. Computable convergence bounds for GMRES. *SIAM J. Matrix Anal. Appl.*, 21(3):882–903, 2000.

[83] Jörg Liesen and Zdeněk Strakoš. Convergence of GMRES for tridiagonal Toeplitz matrices. *SIAM J. Matrix Anal. Appl.*, 26(1):233–251, 2004.

[84] Jörg Liesen and Zdeněk Strakoš. GMRES convergence analysis for a convection-diffusion model problem. *SIAM J. Sci. Comput.*, 26(6):1989–2009, 2005.

[85] Jörg Liesen and Zdeněk Strakoš. *Krylov subspace methods: principles and analysis*. Oxford University Press, 2012.

[86] Ladislav Lukšan and Jan Vlček. Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems. *Numer. Linear Algebra Appl.*, 5(3):219–247, 1998.

[87] Istvan Maros and Csaba Mészáros. A repository of convex quadratic programming problems. *Optim. Methods Softw.*, 11(1-4):671–681, 1999.

[88] The MathWorks, Inc. *MATLAB 7*, R2013b edition, 2013.

[89] Gérard A Meurant. *Computer solution of sparse linear systems*. Elsevier, 1999.

[90] Christian Meyer, Uwe Prüfert, and Fredi Tröltzsch. On two numerical methods for state-constrained elliptic control problems. *Optim. Methods Softw.*, 22(6):871–899, 2007.

[91] Benedetta Morini, Valeria Simoncini, and Mattia Tani. Unreduced symmetric KKT systems arising from Interior Point methods. Part I: spectral estimates. *Optimization Online, http://www.optimization-online.org/DB_HTML/2014/05/4356.html*, 2014.

[92] Benedetta Morini, Valeria Simoncini, and Mattia Tani. Unreduced symmetric KKT systems arising from Interior Point methods. part II: preconditioning. *Optimization-Online, http://www.optimization-online.org/DB_HTML/2014/07/4418.html*, 2014.

[93] Malcolm F. Murphy, Gene H. Golub, and Andrew J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21(6):1969–1972, 2000.

[94] Noël M. Nachtigal, Satish C. Reddy, and Lloyd N. Trefethen. How fast are nonsymmetric matrix iterations? *SIAM J. Matrix Anal. Appl.*, 13(3):778–795, 1992.

[95] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer, second edition, 2006.

[96] Yvan Notay. Aggregation-based algebraic multigrid for convection-diffusion equations. *SIAM J. Sci. Comput.*, 34(4):A2288–A2316, 2012.

[97] Yvan Notay. AGMG software and documentation. *http://homepages.ulb.ac.be/∼ynotay/AGMG*, 2014.

[98] John T. Oden and Junuthula N. Reddy. *An introduction to the mathematical theory of finite elements*. Courier Dover Publications, 2012.

[99] Yvan. Paige, ChNotay and Michael A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975.

[100] John W. Pearson, Martin Stoll, and Andrew J. Wathen. Preconditioners for state-constrained optimal control problems with Moreau–Yosida penalty function. *Numer. Linear Algebra Appl.*, 21(1):81–97, 2014.

[101] John W. Pearson and Andrew J. Wathen. A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numer. Linear Algebra Appl.*, 19(5):816–829, 2012.

[102] John W. Pearson and Andrew J. Wathen. Fast iterative solvers for convection-diffusion control problems. *Electron. Trans. on Numer. Anal.*, 40:294–310, 2013.

[103] Ilaria Perugia and Valeria Simoncini. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. *Numer. Linear Algebra Appl.*, 7(7-8):585–616, 2000.

[104] J. Pestana and A. J. Wathen. Combination preconditioning of saddle point systems for positive definiteness. *Numer. Linear Algebra Appl.*, 20(5):785–808, 2013.

[105] Jennifer Pestana. *Nonstandard inner products and preconditioned iterative methods*. PhD thesis, Oxford University, 2011.

[106] Margherita Porcelli. On the convergence of an inexact Gauss–Newton trust-region method for nonlinear least-squares problems with simple bounds. *Optim. Lett.*, 7(3):447–465, 2013.

[107] Margherita Porcelli, Valeria Simoncini, and Mattia Tani. Preconditioning of active-set Newton methods for PDE-constrained optimal control problems. *arXiv preprint, arXiv:1407.1144*, 2014.

[108] Catherine Elizabeth Powell and David Silvester. Optimal preconditioning for Raviart–Thomas mixed formulation of second-order elliptic problems. *SIAM J. Matrix Anal. Appl.*, 25(3):718–738, 2003.

[109] PDCO: Primal-Dual interior method for Convex Objectives. http://convexoptimization.com/sol-software/pdco.html.

[110] Tim Rees and Chen Greif. A preconditioner for linear systems arising from interior point optimization methods. *SIAM J. Sci. Comput.*, 29(5):1992–2007, 2007.

[111] Miroslav Rozložník and Valeria Simoncini. Krylov subspace methods for saddle point problems with indefinite preconditioning. *SIAM J. Matrix Anal. Appl.*, 24(2):368–391, 2002.

[112] John W. Ruge and Klaus Stüben. Algebraic multigrid. In *Multigrid methods*, volume 3 of *Frontiers Appl. Math.*, pages 73–130. SIAM, Philadelphia, PA, 1987.

[113] Torgeir Rusten and Ragnar Winther. A preconditioned iterative method for saddlepoint problems. *SIAM J. Matrix Anal. Appl.*, 13(3):887–904, 1992.

[114] Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, second edition, 2003.

[115] Yousef Saad and Martin H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. and Stat. Comput.*, 7(3):856–869, 1986.

[116] Michael A. Saunders et al. Cholesky-based methods for sparse least squares: The benefits of regularization. In *Linear and Nonlinear Conjugate Gradient-Related Methods*, pages 92–100. SIAM, 1996.

[117] Joachim Schöberl and Walter Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM J. Matrix Anal. Appl.*, 29(3):752–773 (electronic), 2007.

[118] Debora Sesana and Valeria Simoncini. Spectral analysis of inexact constraint preconditioning for symmetric saddle point matrices. *Linear Algebra Appl.*, 438(6):2683–2700, 2013.

[119] Shu-Qian Shen, Ting-Zhu Huang, and Jian-Song Zhang. Augmentation block triangular preconditioners for regularized saddle point problems. *SIAM J. Matrix Anal. Appl.*, 33(3):721–741, 2012.

[120] David J. Silvester and Andrew J. Wathen. Fast iterative solution of stabilised Stokes systems. II. Using general block preconditioners. *SIAM J. Numer. Anal.*, 31(5):1352–1367, 1994.

[121] Valeria Simoncini. Reduced order solution of structured linear systems arising in certain PDE-constrained optimization problems. *Comput. Optim. Appl.*, 53(2):591–617, 2012.

[122] Valeria Simoncini and Daniel B Szyld. Interpreting IDR as a Petrov-Galerkin method. *SIAM J. Sci. Comput.*, 32(4):1898–1912, 2010.

[123] Peter Sonneveld and Martin B. van Gijzen. IDR (s): A family of simple and fast algorithms for solving large nonsymmetric systems of linear equations. *SIAM Journal on Scientific Computing*, 31(2):1035–1062, 2008.

[124] Georg Stadler. Elliptic optimal control problems with L 1-control cost and applications for the placement of control devices. *Comput. Optim. Appl.*, 44(2):159–181, 2009.

[125] Martin Stoll and Andy Wathen. The Bramble-Pasciak+ preconditioner for saddle point problems. *SIAM J. Matrix Anal. Appl.*, 30(2):582–608, 2008.

[126] Martin Stoll and Andy Wathen. Preconditioning for partial differential equation constrained optimization with control constraints. *Numer. Linear Algebra Appl.*, 19(1):53–71, 2012.

[127] H. Sue Thorne. Distributed control and constraint preconditioners. *Comput. & Fluids*, 46(1):461–466, 2011.

[128] Mattia Tani and Valeria Simoncini. Refined spectral estimates for preconditioned saddle point linear systems in a non-standard inner product. *ANZIAM J.*, 54:C291–C308, 2013.

[129] David Titley-Peloquin, Jennifer Pestana, and Andrew J Wathen. GMRES convergence bounds that depend on the right-hand-side vector. *IMA J. Numer. Anal.*, 34(2):462–479, 2014.

[130] Fredi Tröltzsch. *Optimal control of partial differential equations*. American Mathematical Society, 2010.

[131] The University of Florida Sparse Matrix Collection. http://www.cise.ufl.edu/research/sparse/matrices/.

[132] Henk A. Van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 13(2):631–644, 1992.

[133] Henk A. Van der Vorst. *Iterative Krylov methods for large linear systems*. Cambridge University Press, 2003.

[134] Andrew J. Wathen and Tyrone Rees. Chebyshev semi-iteration in preconditioning for problems including the mass matrix. *Electron. Trans. Numer. Anal.*, 34:125–135, 2009.

[135] Piet Wesseling and Peter Sonneveld. Numerical experiments with a multiple grid and a preconditioned Lanczos type method. In *Approximation methods for Navier-Stokes problems*, pages 543–562. Springer, 1980.

[136] Margaret H. Wright. Ill-conditioning and computational error in interior methods for nonlinear programming. *SIAM J. Optim.*, 9(1):84–111, 1998.

[137] Stephen J. Wright. Stability of linear equations solvers in interior-point methods. *SIAM J. Matrix Anal. Appl.*, 16(4):1287–1307, 1995.

[138] Stephen J. Wright. *Primal-dual interior-point methods*. SIAM, 1997.

[139] Stephen J. Wright. Effects of finite-precision arithmetic on interior-point methods for nonlinear programming. *SIAM J. Optim.*, 12(1):36–78, 2001.

Acknowledgements

First and foremost, I would like to thank my supervisor, prof. Valeria Simoncini, who introduced me to the rich field of numerical linear algebra and guided me with authority in these years of study. I would like to thank her in particular for always being present whenever I needed advice, both on a mathematical topic or on some different matter. She never saved myself the efforts that were needed, but always encouraged me to do my best.

My gratitude goes also to the other coauthors of the results presented in this thesis. In particular, I would like to thank prof. Benedetta Morini for her kindness and for being a great source of knowledge on optimization issues, and dott. Margherita Porcelli for her friendship and tireless support on everything related to software.

I would like also to thank Zdenek Strakos for the kind hospitality at the Charles University of Prague, during the first half of 2014, where I had the chance to meet some of the highest-profile researchers in my field. It was a pleasure to study and carry out my research in such a stimulating environment.

A special thanks goes to the fellow Ph.D. students of the Mathematics Department for sharing with me the everyday life of the research work; in particular, to Giulio Tralli, for the long discussions on math topics, and to Agnese Baldisserri, for her invaluable friendship.

Last but not least, I would like to thank my family and all the friends who supported me during these years.