Science & Technology
Facilities Council

Technical Report

Properties of linear systems in PDE-constrained optimization. Part I: Distributed control

H S Thorne

August 25, 2009, revised April 22, 2010

RAL-TR-2009-017

Enquires about copyright, reproduction and requests for additional copies of this report should be addressed to:

Library and Information Services
SFTC Rutherford Appleton Laboratory
Harwell Science and Innovation Campus
Didcot
OX11 0QX
UK
Tel: +44 (0)1235 445384
Fax: +44(0)1235 446403
Email: library@rl.ac.uk

The STFC ePublication archive (epubs), recording the scientific output of the Chilbolton, Daresbury, and Rutherford Appleton Laboratories is available online at: http://epubs.cclrc.ac.uk/

ISSN 1358-6254

Neither the Council nor the Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigation

RAL-TR-2009-017

Properties of linear systems in PDE-constrained optimization. Part I: Distributed control[1]

H. Sue Thorne[2]

ABSTRACT

Optimization problems with constraints that contain a partial differential equation arise widely in many areas of science. In this paper, we consider distributed control problems in which the 2- and 3-dimensional Poisson problem is the PDE. If a discretize-then-optimization approach is used to solve the optimization problem, then a large dimensional, symmetric and indefinite linear system must be solved. In general, distributed control problems include a regularization term, the size of which is determined by a real value known as the regularization parameter. The spectral properties and, hence, the condition number of the linear system are highly dependent on the size of this regularization parameter. We derive intervals that contain the eigenvalues of the linear systems and, using these, we are able to show that if the regularization parameter is larger than a certain value, then backward-stable direct methods will compute solutions to the discretized optimization problem that have relative errors of the order of machine precision: changing the value of the regularization parameter within this interval will have negligible effect on the accuracy but the condition number of the system may have dramatically changed. We also analyse the spectral properties of the Schur complement and reduced systems derived via the nullspace method. Throughout the paper, we complement the theoretical results with numerical results.

[1] This work was supported by the EPSRC grant EP/E053351/1.

[2] Computational Science and Engineering Department, Rutherford Appleton Laboratory,
Chilton, Oxfordshire, OX11 0QX, England, EU.
Email: sue.thorne@stfc.ac.uk
Current reports available from "http://www.numerical.rl.ac.uk/reports/".

Computational Science and Engineering Department
Atlas Centre
Rutherford Appleton Laboratory
Oxfordshire OX11 0QX

April 22, 2010

1 Introduction

In this paper, we consider the linear algebraic properties of distributed control problems after their discretization. The problems considered consist of a cost functional to be minimized subject to a partial differential equation (PDE) posed on a domain in $\Omega \subset \mathrm{I\!R}^2$ or $\mathrm{I\!R}^3$ (in this case, the Poisson equation):

$$\min_{u,f} \frac{1}{2} \|u - \hat{u}\|^2_{L^2(\hat{\Omega})} + \beta \|f\|^2_{L^2(\Omega)} \tag{1.1}$$

$$\text{subject to} \quad -\nabla^2 u = f \;\text{ in }\; \Omega \tag{1.2}$$

$$\text{with} \quad u = g \quad \text{or} \quad \frac{\delta u}{\delta n} = g \;\text{ on }\; \delta\Omega. \tag{1.3}$$

Here, the function $\hat{u}$ (the 'desired state') is known and we want to find u that satisfies the PDE and is as close to $\hat{u}$ as possible in the L_2 norm sense over the domain $\hat{\Omega} \subseteq \Omega$ for which $\hat{u}$ is known. In order to do this, the right-hand side of the PDE, f, (also known as the 'control') can be varied. The second term in the cost functional (1.1), a Tikhonov regularization term, is added because the problem may be either ill-posed or the right-hand side of the PDE, f, rapidly varies across the domain Ω. In general, the Tikhonov parameter β needs to be determined, although it is often selected a priori – a value around $\beta = 10^{-2}$ is commonly used (see [6, 11, 14]).

In PDE-constrained optimization there is the choice as to whether to discretize-then-optimize or optimize-then-discretize, and there are differing opinions regarding which route to take (see Collis and Heinkenschloss [6] for a discussion). We have chosen to discretize-then-optimize, as then we are guaranteed symmetry in the resulting linear system. The underlying optimization problems are naturally self-adjoint and by this choice we avoid non-symmetry due to discretization that can arise with the optimize-then- discretize approach (as shown in, for example, Collis and Heinkenschloss [10]). We discuss the formulation and general structure of our discretized problem in Section 2.

In this paper, we will consider how the size of the regularization parameter effects the spectral properties of the linear systems associated with problems of the above form. In particular, we will consider the overall saddle-point system (Section 4), the Schur complement (Section 5) and the reduced system from the nullspace method (Section 6). In Section 4.4, we will also show that if the regularization parameter β is large, then solving the overall saddle-point system with a backward-stable direct method will result in the computed state and control variables being of much higher accuracy than standard bounds based on the condition number of the system would suggest. We draw our conclusions in Section 7.

1.1 Notation

All norms are two-norms; the eigenvalues $\{\lambda_i\}$ of a matrix (or generalised eigenvalue problem) are ordered such that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$; the singular values $\{\sigma_i\}$ of a matrix are ordered such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$. The condition number of a matrix A, $\kappa(A)$, is defined by $\kappa(A) := \|A\| \left\|A^{-1}\right\|$. We will use the following notation. We will use the notation $\lambda_{\min}(A)$, $\lambda_{\min^+}(A)$ and $\lambda_{\max}(A)$ ($\sigma_{\min}(A)$, $\sigma_{\min^+}(A)$ and $\sigma_{\max}(A)$) to denote the minimum, minimum positive and maximum eigenvalues (singular values), respectively, of a matrix A. Similarly, $\lambda_{\min}(A, B)$, $\lambda_{\min^+}(A, B)$ and $\lambda_{\max}(A, B)$ denotes the minimum, minimum positive and maximum eigenvalues of the A, respectively, of the generalised eigenvalue problem $Av = \lambda Bv$. For each eigenvalue $\lambda_i(A, B)$, we denote the corresponding eigenvector by $v_i(A, B)$. We define

$$\min_x^+ (f(x)) = \min\{f(x) : f(x) > 0\}.$$

Definition 1.1 (Order notation) Let ϕ be a scalar, vector, or matrix function of a positive variable α, let p be fixed, and let c_u and c_l denote constants.

- If there exists $c_u > 0$ such that $\|\phi\| \leq c_u\alpha^p$ for all sufficiently small/large α, we write $\phi = \mathcal{O}(\alpha^p)$.
- If there exists $c_l > 0$ and $c_u > 0$ such that $c_l\alpha^p \leq \|\phi\| \leq c_u\alpha^p$ for all sufficiently small/large α, we write $\phi = \Theta(\alpha^p)$.

2 Formulation and structure

We have chosen to discretize our problem with finite elements. In order to use these, we require weak formulations of (1.1)–(1.3). For definiteness and clarity we describe the formulation for the purely Dirichlet problem; the formulation for the mixed and purely Neumann problem is standard (see [7]). In the Dirichlet problem we wish to find $u \in H^1_g = \{u \,:\, u \in H^1(\Omega), u = g \text{ on } \delta\Omega\}$ such that

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega vf \quad \forall v \in H^1_0(\Omega). \tag{2.1}$$

We assume that $V^h_0 \subset H^1_0$ is an n-dimensional vector space of test functions with basis $\{\phi_1, \ldots, \phi_n\}$. Then, for the boundary condition to be satisfied, we extend the basis by defining functions $\phi_{n+1}, \ldots, \phi_{n+\delta n}$ and coefficients U_j so that $\sum_{j=n+1}^{n+\delta n} U_j\phi_j$ interpolates the boundary data. Then, if $u_h \in V^h_g \subset H^1_g(\Omega)$, it is uniquely determined by $\mathbf{u} = (U_1 \ldots U_n)^T$ in

$$u_h = \sum_{j=1}^{n} U_j\phi_j + \sum_{j=n+1}^{n+\partial n} U_j\phi_j.$$

Here the ϕ_i, $i = 1, \ldots, n$, define a set of shape functions. We also assume that this approximation is conforming, i.e. $V^h_g = \text{span}\{\phi_1, \ldots, \phi_{n+\partial n}\} \subset H^1_g(\Omega)$. Then we get the finite-dimensional analogue of (2.1): find $u_h \in V^h_g$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega v_h f \qquad \forall v_h \in V^h_0.$$

We also need a discretization of f, as this appears in (1.1). We discretize this using the same basis used for u, so

$$f_h = \sum_{j=1}^{n} F_j\phi_j$$

since it is well known that $f_h = 0$ on $\partial\Omega$. Thus we can write the discrete analogue of the minimization problem as

$$\min_{u_h, f_h} \frac{1}{2}||u_h - \hat{u}||_2^2 + \beta||f_h||_2^2 \tag{2.2}$$

$$\text{such that} \qquad \int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega v_h f \qquad \forall v_h \in V_0^h. \tag{2.3}$$

If $\hat{u}$ is defined over the whole of Ω, we can write the discrete cost functional as

$$\min_{u_h, f_h} \frac{1}{2}||u_h - \hat{u}||_2^2 + \beta||f_h||_2^2 = \min_{\mathbf{u},\mathbf{f}} \frac{1}{2}\mathbf{u}^T \bar{M}\mathbf{u} - \mathbf{u}^T\mathbf{b} + \alpha + \beta\mathbf{f}^T M\mathbf{f}, \tag{2.4}$$

where $\mathbf{u} = (U_1, \ldots, U_n)^T$, $\mathbf{f} = (F_1, \ldots, F_n)^T$, $\mathbf{b} = \{\int \hat{u}\phi_i\}_{i=1\ldots n}$, $\alpha = \|\hat{u}\|_2^2$, $M = \{\int \phi_i\phi_j\}_{i,j=1\ldots n}$ is a mass matrix, and $\bar{M} = M$. If $\hat{u}$ is only defined on part of the domain, defining

$$\tilde{u}_i = \begin{cases} \hat{u}_i & \text{if } \hat{u}_i \text{ defined} \\ 0 & \text{if } \hat{u}_i \text{ not defined,} \end{cases}$$

we obtain (2.4) but α, b and $\bar{M}$ are defined by

$$\begin{aligned} \alpha &= \|\tilde{u}\|_2^2, \\ \mathbf{b}_i &= \int \tilde{u}\phi_i, \\ \bar{M}_{ij} &= \begin{cases} M_{i,j} & \text{if } \hat{u} \text{ is defined atnodes } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

In this case $\bar{M}$ will be singular.

We now turn our attention to the constraint: (2.3) is equivalent to finding $\mathbf{u}$ such that

$$\int_\Omega \nabla\left(\sum_{i=1}^n U_i\phi_i\right) \cdot \nabla\phi_j + \int_\Omega \nabla\left(\sum_{i=n+1}^{n+\partial n} U_i\phi_i\right) \cdot \nabla\phi_j = \int_\Omega \left(\sum_{i=1}^n F_i\phi_i\right)\phi_j, \quad j = 1, \ldots, n$$

which is

$$\sum_{i=1}^n U_i \int_\Omega \nabla\phi_i \cdot \nabla\phi_j = \sum_{i=1}^n F_i \int_\Omega \phi_i\phi_j - \sum_{i=n+1}^{n+\partial n} U_i \int_\Omega \nabla\phi_i \cdot \nabla\phi_j, \quad j = 1, \ldots, n$$

or

$$K\mathbf{u} = M\mathbf{f} + \mathbf{d}, \tag{2.5}$$

where the matrix $K = \{\int \nabla\phi_i \cdot \nabla\phi_j\}_{i,j=1\ldots n}$ is the discrete Laplacian (the stiffness matrix) and $\mathbf{d}$ contains the terms coming from the boundary values of u_h. Thus (2.4) and (2.5) together are equivalent to (2.2) and (2.3).

One way to solve this minimization problem is by considering the Lagrangian

$$\mathcal{L} := \frac{1}{2}\mathbf{u}^T\bar{M}\mathbf{u} - \mathbf{u}^T\mathbf{b} + \alpha + \beta\mathbf{f}^T M\mathbf{f} + \lambda^T(K\mathbf{u} - M\mathbf{f} - \mathbf{d}),$$

where λ is a vector of Lagrange multipliers. Using the stationarity conditions of $\mathcal{L}$, we find that $\mathbf{f}$, $\mathbf{u}$ and λ are defined by the linear system

$$\begin{bmatrix} 2\beta M & 0 & -M \\ 0 & \bar{M} & K^T \\ -M & K & 0 \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{u} \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \\ \mathbf{d} \end{bmatrix}. \tag{2.6}$$

We will discuss the properties of this system in Section 4

2.1 Properties of K, M and $\bar{M}$

Throughout this paper, we will assume that a shape regular, quasi-uniform division of the domain is used [7] with $\mathbf{P_m}$ or $\mathbf{Q_m}$ ($\mathbf{m} \geq 1$) finite element approximations. Using these assumptions, we have the following theorem [7]:

Theorem 2.1 Consider the p-dimensional problem with $p \in \{2, 3\}$. Now

$$\begin{aligned} \lambda_{\min}(K) = ch^p, &\quad \lambda_{\max}(K) = Ch^{p-2}, \\ \lambda_{\min}(M) = dh^p, &\quad \lambda_{\max}(M) = Dh^p, \end{aligned}$$

where c, d, C and D are constants independent of the mesh size h but dependent on p. In addition, $D < \frac{2c}{3}$.

If the target $\hat{u}$ is defined over the whole domain Ω, then $\bar{M} = M$. Suppose that the target $\hat{u}$ is only defined on a sub-domain of Ω. We will use Cauchy's interlacing theorem [17]:

Theorem 2.2 Suppose $T \in \mathbb{R}^{n \times n}$ is symmetric and

$$T = \left[\begin{array}{cc} H & \star \\ \star & \star \end{array} \right],$$

where $H \in \mathbb{R}^{m \times m}$ with $m < n$. Label the eigenpairs of T and H as

$$\begin{aligned} Tz_i = \alpha_i z_i, &\quad i = 1, \ldots, n, \quad \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n, \\ Hy_i = \lambda_i y_i, &\quad i = 1, \ldots, m, \quad \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m. \end{aligned}$$

Then

$$\alpha_k \leq \lambda_k \leq \alpha_{k+n-m}, \quad k = 1, \ldots, m.$$

There exists a permutation matrix Π such that

$$\Pi^T \bar{M} \Pi = \left[\begin{array}{cc} \bar{M}_{11} & 0 \\ 0 & 0 \end{array} \right],$$

where $\bar{M}_{11} \in \mathbb{R}^{\bar{m} \times \bar{m}}$ is nonsingular. Applying Theorem 2.2, the eigenvalues of $\bar{M}_{11}$ lie in the interval $\left[\lambda_{\min}(\Pi^T M \Pi), \lambda_{\max}(\Pi^T M \Pi)\right]$. Hence, we have the following theorem.

Theorem 2.3 Consider the p-dimensional problem with $p \in \{2,3\}$. Assume that Theorem 2.1 holds and that the target $\hat{u}$ in (1.1) is only defined on a sub-domain of Ω. Then

$$\begin{aligned} &\lambda_{\min}(\bar{M}) = 0, \qquad &&\lambda_{\max}(\bar{M}) = \bar{D}h^p, \\ &\lambda_{\min^+}(\bar{M}) = \bar{d}h^p, && \end{aligned}$$

where $\bar{d} \geq d$ and $\bar{D} \leq D$ are constants independent of the mesh size h but dependent on p.

Let

$$\Pi^T K \Pi = \begin{bmatrix} K_{11} & K_{21}^T \\ K_{21} & K_{22} \end{bmatrix},$$

where Π is as defined above and $K_{11} \in \mathbb{R}^{\bar{m}\times\bar{m}}$. We will use the following assumptions.

Assumption 2.1 Consider the p-dimensional problem with $p \in \{2,3\}$. Assume that Theorem 2.1 holds and that the target $\hat{u}$ in (1.1) is only defined on a sub-domain of Ω. We will assume that

$$\lambda_{\min}(K_{22}) = \bar{c}h^p, \quad \lambda_{\max}(K_{22}) = \bar{C}h^{p-2},$$

where $\bar{c} \geq c$ and $\bar{C} \leq C$ are constants independent of the mesh size h but dependent on p.

2.2 The role of β in (1.1)

The second term in the cost functionals is added because, in general, the problem with be ill-posed or a *bang-bang* control state would be obtained. Bang-bang control states are states which rapidly vary from one extreme to another over the domain [16] and would often be difficult to impose in real life applications. By varying the value of the regularization parameter β, the balance between the two terms in the cost functionals will be altered. If it is extremely important for $\|u - \hat{u}\|$ to be very small but we are less concerned by the size of $\|f\|$, then a small value of β should be chosen. Conversely, if u does not need to closely match $\hat{u}$ but it is important that $\|f\|$ remains small, then a larger value of β would be used. In practice, a tolerance is often given that determines how small $\|u - \hat{u}\| / \|\hat{u}\|$ should be. A coarse grid is then used to *cheaply* determine the value of β that corresponds to this tolerance for this grid size: this value of β is then used to solve the problem on the refined mesh [1]. Of course, the coarse grid must be fine enough such that grid refinement is not expected to make a marked difference in terms of the regularization. As we will see in Section 3, there may be instances when the coarse grid has to have a very small mesh size for this to be the case.

3 Test problems

As we proceed through this paper, we will use several test examples to illustrate our results. For all of our tests, we discretize the problem with bilinear quadrilateral $\mathbf{Q_1}$ finite elements.

First, we will describe the target functions that we use. We will consider a continuous and a discontinuous target that are both described over the whole of Ω, and a target that is only defined on a sub-domain of Ω.

In Tables 3.1 and 3.2, we define the different targets used within this paper for 2D and 3D problems, respectively. For the 2D and 3D problems, we define $\Omega = [0,1]^2$ and $\Omega = [0,1]^3$, respectively. Additionally, let

$$\tilde{\Omega} = \left\{(x,y,z) : (x-\tfrac{5}{8})^2 + (y-\tfrac{3}{4})^2 + (z-\tfrac{7}{10})^2 \leq \tfrac{1}{16}\right\}.$$

We will define the domain $\hat{\Omega}$ over which the target is defined: for some cases it will be useful to split $\hat{\Omega}$ into two subregions $\hat{\Omega}_1$ and $\hat{\Omega}_2$.

	$\hat{\Omega}$	$\hat{\Omega}_1$	$\hat{\Omega}_2$	$\hat{u}(x,y)\vert_{\hat{\Omega}_1}$	$\hat{u}(x,y)\vert_{\hat{\Omega}_2}$
Target 1	$\hat{\Omega}_1 \cup \hat{\Omega}_2$	$[0,\frac{1}{2}]^2$	$\Omega/\hat{\Omega}_1$	$(2x-1)^2\,(2y-1)^2$	0
Target 2	$\hat{\Omega}_1 \cup \hat{\Omega}_2$	$\left\{(x,y) : (x-\frac{5}{8})^2 + (y-\frac{3}{4})^2 \leq \frac{1}{25}\right\}$	$\Omega/\hat{\Omega}_1$	2	0
Target 3	$\hat{\Omega}_1 \cup \hat{\Omega}_2$	$\left\{(x,y) : (x-\frac{5}{8})^2 + (y-\frac{3}{4})^2 \leq \frac{1}{25}\right\}$	$\partial\Omega$	2	0

Table 3.1: Target functions for 2D problems

	$\hat{\Omega}$	$\hat{\Omega}_1$	$\hat{\Omega}_2$	$\hat{u}(x,y)\vert_{\hat{\Omega}_1}$	$\hat{u}(x,y)\vert_{\hat{\Omega}_2}$
Target 1	$\hat{\Omega}_1 \cup \hat{\Omega}_2$	$[0,\frac{1}{2}]^3$	$\Omega/\hat{\Omega}_1$	$(2x-1)^2\,(2y-1)^2$	0
Target 2	$\hat{\Omega}_1 \cup \hat{\Omega}_2$	$\tilde{\Omega}$	$\Omega/\hat{\Omega}_1$	2	0
Target 3	$\hat{\Omega}_1 \cup \hat{\Omega}_2$	$\tilde{\Omega}$	$\partial\Omega$	2	0

Table 3.2: Target functions for 3D problems

We now describe the test examples with which we will use our targets $\hat{u}$. Our first example has Dirichlet boundary conditions.

Example 3.1 *Let $\Omega = [0,1]^2$ or $\Omega = [0,1]^3$, and $\hat{\Omega} \subseteq \Omega$ be the domain over which $\hat{u}$ is defined. Consider the problem*

$$\min_{u,f} \frac{1}{2}||u-\hat{u}||^2_{L_2(\hat{\Omega}))} + \beta||f||^2_{L_2(\Omega)}$$

$$\text{s.t.} \qquad -\nabla^2 u = f \text{ in } \Omega, \tag{3.1}$$

$$u = \begin{cases} \hat{u} & \text{on } \partial\Omega \cap \hat{\Omega} \\ 0 & \text{on } \partial\Omega/\hat{\Omega}. \end{cases} \tag{3.2}$$

Our second example has Neumann boundary conditions:

Example 3.2 *Let $\Omega = [0,1]^2$ and $\hat{\Omega} \subseteq \Omega$ be the domain over which $\hat{u}$ is defined. Consider the Neumann problem*

$$\min_{u,f} \frac{1}{2}||u-\hat{u}||^2_{L_2(\hat{\Omega})} + \beta||f||^2_{L_2(\Omega)}$$

$$\text{s.t.} \qquad -\nabla^2 u = f \qquad \text{in } \Omega, \tag{3.3}$$

$$\frac{\partial u}{\partial n} = 0 \qquad \text{on } \partial\Omega. \tag{3.4}$$

Our final problem has mixed boundary conditions:

Example 3.3 *Let* $\Omega = [0,1]^2$ *and* $\hat{\Omega} \subseteq \Omega$ *be the domain over which* $\hat{u}$ *is defined. Consider the problem*

$$\min_{u,f} \frac{1}{2}||u-\hat{u}||^2_{L_2(\hat{\Omega})} + \beta||f||^2_{L_2(\Omega)}$$

$$\begin{aligned} \text{s.t.} \qquad -\nabla^2 u &= f \text{in } \Omega, & (3.5)\\ u &= \begin{cases} \hat{u} & \text{on } \partial_1\Omega \cap \hat{\Omega} \\ 0 & \text{on } \partial\Omega_1/\hat{\Omega}, \end{cases} & (3.6)\\ \frac{\partial u}{\partial n} &= 0 \text{ on } \partial\Omega_2, & (3.7) \end{aligned}$$

where $\partial\Omega_1 = (0 \times [0,1)) \cup ((0,1] \times 0)$ *and* $\partial\Omega_2 = (1 \times (0,1]) \cup ([0,1) \times 1)$.

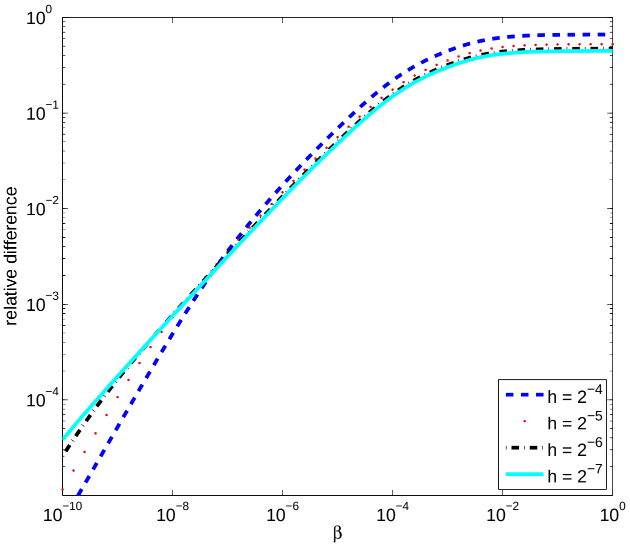

Figure 3.1: The relative difference $\frac{\|u-\hat{u}\|}{\|u\|}$ for Example 3.1 in 2D with Target 1 and different values of β. Results are shown for $h = \frac{1}{16}$, $h = \frac{1}{32}$, $h = \frac{1}{64}$ and $h = \frac{1}{128}$.

Let us consider the 2D version of Example 3.1 and the continuous target $\hat{u}$ defined by Target 1. In Figure 3.1, we plot the value of $\|u - \hat{u}\| / \|\hat{u}\|$ against the regularization parameter β. Results are given for different choices of mesh size h. We observe that, for the fixed $\beta > 10^{-7}$, the larger values of h produce values of $\|u - \hat{u}\| / \|\hat{u}\|$ that are of the same order of magnitude. For this problem, it is therefore possible to find a suitable value of β from a coarse discretization and then use this value of β with a fine discretization to compute the desired u.

If we use the Target 2, then we would not expect the same behaviour of $\|u - \hat{u}\| / \|\hat{u}\|$ because the discontinuity of the target will not be well approximated on coarse meshes. Indeed, in Figure 3.2, if we wanted $\|u - \hat{u}\| / \|\hat{u}\| \approx 0.1$, then the required value of β would drastically change as we refine h. Eventually, h will be small enough relative to the required tolerance for our discretization to be good enough give (almost) mesh independent results.

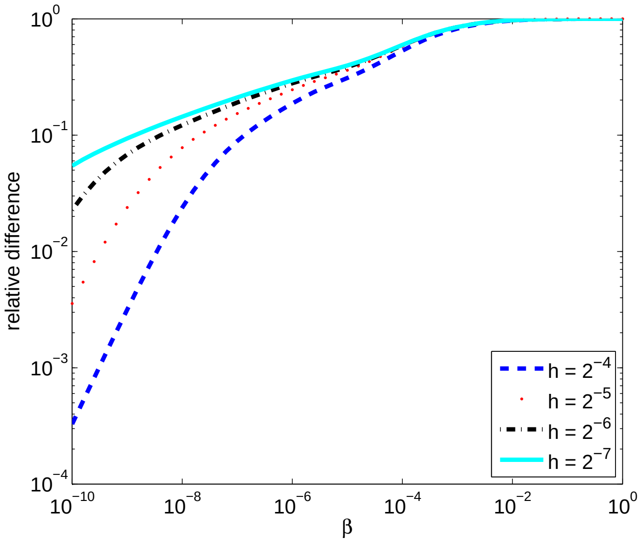

Figure 3.2: The relative difference $\frac{\|u-\hat{u}\|}{\|u\|}$ for Example 3.1 in 2D with Target 2 and different values of β. Results are shown for $h = \frac{1}{16}$, $h = \frac{1}{32}$, $h = \frac{1}{64}$ and $h = \frac{1}{128}$.

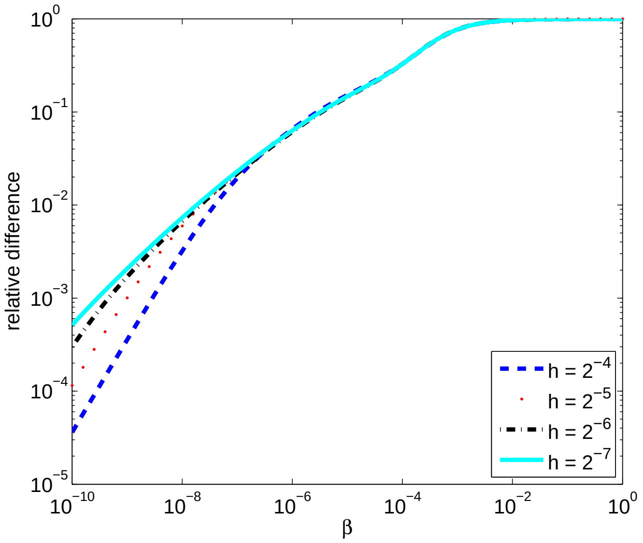

Figure 3.3: The relative difference $\frac{\|u-\hat{u}\|}{\|u\|}$ for Example 3.1 in 2D with Target 3 and different values of β. Results are shown for $h = \frac{1}{16}$, $h = \frac{1}{32}$, $h = \frac{1}{64}$ and $h = \frac{1}{128}$.

If we instead consider the Target 3, then, so long as $\|u - \hat{u}\| / \|\hat{u}\|$ is not too small, the coarse meshes will generate values of β that can be used with the finer meshes to generate solutions that provide the correct level of accuracy with respect the target $\hat{u}$. In Figure 3.3 we observe that the dependency on h is only an issue for small values of β. In fact, in this case, the mesh size it not small enough for the required tolerance to be relevant.

Therefore, as shown in [1], it is paramount that the value of β is chosen very carefully and according to the characteristics of the underlying problem. In literature, it is common to see the choice $\beta = 0.01$ or $\beta = 0.001$ used [6, 11, 14]. From Figures 3.1–3.3 we see that such values of β would produce a large value of $\|u - \hat{u}\| / \|\hat{u}\|$. In practice, we feel that u will be needed to differ from the target by at most 10% and, hence, for our examples, we should have $\beta \approx 10^{-5}$ or 10^{-6}. Because the choice of β is very dependent on the target $\hat{u}$, we will not restrict ourselves to any particular value of β in the following analysis.

4 Spectral properties of the saddle-point matrices

We observe that the system (2.6) can be written in the form

$$\underbrace{\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}}_{\mathcal{A}} \underbrace{\begin{bmatrix} x \\ y \end{bmatrix}}_{s} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}}_{b}, \tag{4.1}$$

where

$$A = \begin{bmatrix} 2\beta M & 0 \\ 0 & \bar{M} \end{bmatrix}, \quad B = \begin{bmatrix} -M & K \end{bmatrix}. \tag{4.2}$$

Systems of the general form given in (4.1) are known as saddle-point matrices [2]. We note that, within this application, the matrix B always has full row rank and if $\beta > 0$, then $\mathcal{A}$ is guaranteed to be nonsingular. In addition, if the target $\hat{u}$ is defined over the whole of the domain Ω, then A will be nonsingular for $\beta = 0$.

4.1 Eigenvalue intervals for saddle-point problems

If $A \in \mathbb{R}^{n\times n}$ is positive definite and $B \in \mathbb{R}^{m\times n}$ has full rank, then $\mathcal{A}$ defined by (4.1) has m negative eigenvalues and n positive eigenvalues [2] (similarly for A positive semidefinite and $\mathcal{A}$ nonsingular). The following result from [19] can be used to establish eigenvalue bounds for (4.1).

Theorem 4.1 Assume A is positive definite and B has full rank. Then

$$\lambda(\mathcal{A}) \subset I^- \cup I^+,$$

where $\mathcal{A}$ is defined by (4.1),

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

and

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

If A is positive semidefinite and we let $B\left[Y_B, Z_B\right] = \left[L, 0\right]$, where $\left[Y_B, Z_B\right]$ is orthogonal and L is nonsingular, it has been conjectured that the lower bound of I^+ is replaced by $\lambda_{\min}(Z_B^T A Z_B)$ and the other bounds are unchanged, [23]. We note that, in this case, $\mathcal{A}$ is nonsingular if and only if $Z_B^T A Z_B$ is positive definite [2]. Consider the problem

$$\mathcal{A} = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 100 & 1 & 1 \\ \hline 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{array}\right].$$

Now $Z_B = \frac{1}{\sqrt{3}}\left[\begin{array}{ccc} -1 & -1 & 1 \end{array}\right]^T$, $Z_B^T A Z_B = \frac{101}{3}$ and, if the conjecture was true, we would expect the positive eigenvalues of $\mathcal{A}$ to be greater than or equal to $\frac{101}{3}$. However, $\lambda_{\min^+}(\mathcal{A}) = 0.9950$.

Theorem 4.2 Assume $\mathcal{A}$ is nonsingular, A is positive semidefinite and possibly nonsingular, and let $A\left[Y_A, Z_A\right] = \left[L_A, 0\right]$, where $\left[Y_A, Z_A\right]$ is orthogonal and L_A has full column rank. Assume B has full rank and let $B\left[Y_B, Z_B\right] = \left[L_B, 0\right]$, where $\left[Y_B, Z_B\right]$ is orthogonal and L_B is nonsingular. Then

$$\lambda(\mathcal{A}) \subset I^- \cup I_1^+ \cup I_2^+ \cup I_3^+,$$

where $\mathcal{A}$ is defined by (4.1),

$$\begin{aligned}
I^- &= \left[-\sigma_{\max}(B), \tfrac{1}{2}\left(\left\|Y_B^T A Y_B\right\| - \sqrt{\left\|Y_B^T A Y_B\right\| + 4(\sigma_{\min}(B))^2}\right)\right], \\
I_1^+ &= \left[\lambda_{\min}(Z_B^T A Z_B), \lambda_{\max}(Z_B^T A Z_B)\right], \\
I_2^+ &= \left[l_2, u_2\right], \\
I_3^+ &= \left[l_3, \tfrac{1}{2}\left(\|A\| + \sqrt{\|A\| + 4\|B\|^2}\right)\right], \\
l_2 &= \tfrac{1}{2}\left(\lambda_{\min^+}(A) + \sqrt{\left(\lambda_{\min^+}(A)\right)^2 + 4\left(\sigma_{\min}(BY_A)\right)^2}\right) \\
&\geq \tfrac{1}{2}\left(\lambda_{\min^+}(A) + \sqrt{\left(\lambda_{\min^+}(A)\right)^2 + 4\left(\sigma_{\min}(B)\right)^2}\right), \\
u_2 &= \tfrac{1}{2}\left(\|A\| + \sqrt{\|A\| + 4\|BY_A\|^2}\right) \\
&\leq \tfrac{1}{2}\left(\|A\| + \sqrt{\|A\|^2 + 4\|B\|^2}\right),
\end{aligned}$$

and $l_3 < \sigma_{\min}(BZ_A)$ is the smallest positive root of the cubic equation

$$\mu^3 - \mu^2\lambda_{\min^+}(A) - \mu\left(\left(\sigma_{\min}(BZ_A)\right)^2 + \|BY_A\|^2\right) + \lambda_{\min^+}(A)\left(\sigma_{\min}(BZ_A)\right)^2 = 0.$$

In particular,

$$\begin{aligned}
l_3 &\geq \frac{1}{2}\left(-\bar{l}_3 + \sqrt{\bar{l}_3^2 + 4\left(\sigma_{\min}(BZ_A)\right)^2}\right), \\
\bar{l}_3 &= \frac{\left(\sigma_{\min}(BZ_A)\right)^2 + \|BY_A\|^2}{\lambda_{\min^+}(A)}.
\end{aligned}$$

If $m = n$, we have

$$\begin{aligned}
I_1^+ &= \emptyset, \\
I_3^+ &= \left[\sigma_{\min}(B), \tfrac{1}{2}\left(\|A\| + \sqrt{\|A\|^2 + 4\|B\|^2}\right)\right].
\end{aligned}$$

If $m < n$ and A is nonsingular, then $I_3^+ = \emptyset$.

Proof. Let $[x; y]$ be an eigenvector corresponding to an eigenvalue λ. Expanding out the eigenvalue problem we obtain

$$Ax + B^T y = \lambda x, \tag{4.3}$$

$$Bx = \lambda y. \tag{4.4}$$

Consider the case $y = 0$. From (4.4) we find that $x = Z_B z_B$ for some vector $z_B \neq 0$. Premultiplying (4.3) by x^T and substituting in $x = Z_B z_B$ we obtain the interval I_1^+. If $x = 0$, then (4.3) implies that $y = 0$. Since $[x; y]$ is an eigenvector, this case can not occur.

We will firstly calculate the extremes of I^-. Premultiplying (4.3) by x^T and using (4.4) to eliminate y we obtain

$$\lambda^2 \|x\|^2 - \lambda x^T Ax - \|Bx\|^2 = 0. \tag{4.5}$$

Since $\lambda < 0$, $-\lambda x^T Ax \geq 0$. Additionally $-\|Bx\|^2 \geq -\|B\|^2 \|x\|^2$. Hence

$$\lambda^2 - \|B\|^2 \leq 0$$

and we obtain the lower bound of I^-.

Writing $x = Z_B z_B + Y_B y_B$ and using (4.4) to eliminate y from (4.3) we find that

$$\lambda^2 Z_B z_B + \lambda^2 Y_B y_B - \lambda A Z_B z_B - \lambda A Y_B y_B - B^T B Y_B y_B. \tag{4.6}$$

Premultiplying (4.6) by $y_B^T Y_B^T$ we obtain

$$\lambda^2 \|y_B\|^2 - \lambda y_B^T Y_B^T A Z_B z_B - \lambda y_B^T Y_B^T A Y_B y_B - y_B^T Y_B^T B^T B Y_B y_B = 0. \tag{4.7}$$

Premultiplying (4.6) by $z_B^T Z_B^T$ we obtain

$$\lambda^2 \|z_B\|^2 - \lambda z_B^T Z_B^T A Z_B z_B - \lambda z_B^T Z_B^T A Y_B y_B = 0. \tag{4.8}$$

Subtracting (4.8) from (4.7) gives

$$\begin{aligned}
0 &= \lambda^2 \|y_B\|^2 - \lambda^2 \|z_B\|^2 - \lambda y_B^T Y_B^T A Y_B y_B + \lambda z_B^T Z_B^T A Z_B z_B - y_B^T Y_B^T B^T B Y_B y_B \\
&\leq \lambda^2 \|y_B\|^2 - \lambda y_B^T Y_B^T A Y_B y_B - y_B^T Y_B^T B^T B Y_B y_B \\
&\leq \left(\lambda^2 - \left\|Y_B^T A Y_B\right\| - (\sigma_{\min}(B))^2\right) \|y_B\|^2 \\
&\leq \left(\lambda^2 - \|A\| - (\sigma_{\min}(B))^2\right) \|y_B\|^2 .
\end{aligned}$$

Hence, we obtain the upper bounds of I^-.

Consider $\lambda > 0$. Let $y \neq 0$ and $Ax = 0$, then $x = Z_A z_A$ for some vector $z_A \neq 0$. Premultiplying (4.3) by x^T and using (4.4) to eliminate y we obtain

$$z_A^T Z_A^T B^T B Z_A z_A = \lambda^2 \|z_A\|^2 .$$

Note that BZ_A must have full column rank for $\mathcal{A}$ to be nonsingular. Hence,

$$\lambda \in \left[\sigma_{\min}(BZ_A), \|BZ_A\|\right]. \tag{4.9}$$

Let $y \neq 0$ and $Ax \neq 0$. Substituting $x = Z_A z_A + Y_A y_A$ into (4.3) and (4.4), and premultiplying (4.3) by Z_A^T and Y_A^T, respectively we obtain

$$\begin{aligned} Z_A^T B^T y &= \lambda z_A, && (4.10)\\ Y_A^T A Y_A y_A + Y_A^T B^T y &= \lambda y_A, && (4.11)\\ B Z_A z_A + B Y_A y_A &= \lambda y. && (4.12) \end{aligned}$$

Consider the case $z_A = 0$. Now $Ax \neq 0$ and, hence, $y_A \neq 0$. Using (4.12) to eliminate y from (4.11) and premultiplying the resultant equation by y_A^T, we obtain

$$\begin{aligned} 0 &= \lambda^2 \|y_A\|^2 - \lambda y_A^T Y_A^T A Y_A y_A - y_A^T Y_A^T B^T B Y_A y_A \\ &\leq \left(\lambda^2 - \lambda\lambda_{\min}(Y_A^T A Y_A) - (\sigma_{\min}(BY_A))^2\right) \|y_A\|^2 \\ &\leq \left(\lambda^2 - \lambda\lambda_{\min}(Y_A^T A Y_A) - (\sigma_{\min}(B))^2\right) \|y_A\|^2 . \end{aligned}$$

Hence,we obtain the lower bounds of I_2^+. Similarly, we find that

$$\begin{aligned} 0 &= \lambda^2 \|y_A\|^2 - \lambda y_A^T Y_A^T A Y_A y_A - y_A^T Y_A^T B^T B Y_A y_A \\ &\geq \left(\lambda^2 - \lambda \left\|Y_A^T A Y_A\right\| - \|BY_A\|^2\right) \|y_A\|^2 \\ &\geq \left(\lambda^2 - \lambda \left\|Y_A^T A Y_A\right\| - \|B\|^2\right) \|y_A\|^2 . \end{aligned}$$

This gives the upper bounds of I_2^+.

Consider the case $z_A \neq 0$. Then (4.10) implies that $Z_A^T B^T y \neq 0$. From (4.5) we obtain

$$\left(\lambda^2 - \lambda \|A\| - \|B\|^2\right) \|x\|^2 \leq 0.$$

From this we obtain the upper bound of I_3^+. If $m = n$, we also find that

$$\left(\lambda^2 - (\sigma_{\min}(B))^2\right) \|x\|^2 \geq 0$$

and, hence, obtain the required lower bound of I_3^+.

Assume that $m < n$ and $\lambda < \sigma_{\min}(BZ_A)$. Using (4.10) to eliminate z_A from (4.12), and premultiplying the resulting equation by y^T we obtain

$$\lambda y^T B Y_A y_A = \lambda^2 \|y\|^2 - y^T B Z_A Z_A^T y \leq \left(\lambda^2 - (\sigma_{\min}(BZ_A))^2\right) \|y\|^2 < 0.$$

Also, $y^T B Y_A \geq - \|BY_A\| \|y\| \|y_A\|$. Combining the two inequalities we find that

$$\|y\| \leq -\frac{\lambda \|BY_A\| \|y_A\|}{\lambda^2 - (\sigma_{\min}(BZ_A))^2},$$

which implies that

$$-y^T B Y_A y_A \leq -\frac{\lambda \|BY_A\|^2 \|y_A\|^2}{\lambda^2 - (\sigma_{\min}(BZ_A))^2}.$$

Premultiplying (4.11) by y_A^T we find that

$$\begin{aligned} 0 &= \lambda \|y_A\|^2 - y_A^T Y_A^T A Y_A y_A - y_A^T Y_A^T B^T y \\ &\leq \left(\lambda - \lambda_{\min}(Y_A^T A Y_A) - \frac{\lambda \|BY_A\|^2}{\lambda^2 - (\sigma_{\min}(BZ_A))^2} \right) \|y_A\|^2 \\ &\leq \left(\lambda \|y_A\|^2 - \lambda_{\min}(Y_A^T A Y_A) - \frac{\lambda \|BY_A\|^2}{\lambda^2 - (\sigma_{\min}(B))^2} \right) \|y_A\|^2 , \end{aligned}$$

which gives

$$\begin{aligned} 0 &\geq \lambda^3 - \lambda^2 \lambda_{\min}(Y_A^T A Y_A) - \lambda \left((\sigma_{\min}(BZ_A))^2 + \|BY_A\|^2 \right) + \lambda_{\min}(Y_A^T A Y_A) \, (\sigma_{\min}(BZ_A))^2 \\ &\geq -\lambda^2 \lambda_{\min}(Y_A^T A Y_A) - \lambda \left((\sigma_{\min}(BZ_A))^2 + \|BY_A\|^2 \right) + \lambda_{\min}(Y_A^T A Y_A) \, (\sigma_{\min}(BZ_A))^2 . \end{aligned}$$

This gives the lower bound for I_3^+. Finally, we observe that the interval given by (4.9) is contained within I_3^+. □

Combining Theorem 4.2 with Propositions 2.2 and 2.9 from [10] we obtain the following result.

Corollary 4.3 Assume $\mathcal{A}$ is nonsingular, A is positive semidefinite and possibly nonsingular, and let $A\,[Y_A, Z_A] = [L_A, 0]$, where $[Y_A, Z_A]$ is orthogonal and L_A has full column rank. Assume B has full rank and let $B\,[Y_B, Z_B] = [L_B, 0]$, where $[Y_B, Z_B]$ is orthogonal and L_B is nonsingular. Then

$$\lambda(\mathcal{A}) \subset I^- \cup I^+,$$

where $\mathcal{A}$ is defined by (4.1),

$$\begin{aligned}
I^- &= \left[-\sigma_{\max}(B), \tfrac{1}{2}\left(\left\|Y_B^T A Y_B\right\| - \sqrt{\left\|Y_B^T A Y_B\right\|^2 + 4(\sigma_{\min}(B))^2}\right)\right], \\
&\subset \left[-\sigma_{\max}(B), \tfrac{1}{2}\left(\|A\| - \sqrt{\|A\|^2 + 4(\sigma_{\min}(B))^2}\right)\right], \\
I^+ &= \left[l^+, \tfrac{1}{2}\left(\|A\| + \sqrt{\|A\|^2 + 4\,\|B\|^2}\right)\right], \\
l^+ &= \max\left(l_1, \min\left(l_2, l_3\right)\right), \\
l_2 &= \tfrac{1}{2}\left(\lambda_{\min^+}(A) + \sqrt{\left(\lambda_{\min^+}(A)\right)^2 + 4\left(\sigma_{\min}(BY_A)\right)^2}\right) \\
&\geq \tfrac{1}{2}\left(\lambda_{\min^+}(A) + \sqrt{\left(\lambda_{\min^+}(A)\right)^2 + 4\left(\sigma_{\min}(B)\right)^2}\right),
\end{aligned}$$

$l_1 < \lambda_{\min}(Z_B^T A Z_B)$ is the smallest positive root of the cubic equation

$$\mu^3 - \mu^2 \lambda_{\min}(Z_B^T A Z_B) - \mu\left(\|A\|^2 + (\sigma_{\min}(B))^2\right) + \lambda_{\min}(Z_B^T A Z_B)\left(\sigma_{\min}(B)\right)^2 = 0$$

and $l_3 < \sigma_{\min}(BZ_A)$ is the smallest positive root of the cubic equation

$$\mu^3 - \mu^2 \lambda_{\min^+}(A) - \mu\left(\left(\sigma_{\min}(BZ_A)\right)^2 + \|BY_A\|^2\right) + \lambda_{\min^+}(A)\left(\sigma_{\min}(BZ_A)\right)^2 = 0.$$

In particular,

$$\begin{aligned}
l_1 &\geq -\frac{\|A\|^2 + (\sigma_{\min}(B))^2}{2\lambda_{\min}(Z_B^T A Z_B)} + \sqrt{\frac{\left(\|A\|^2 + (\sigma_{\min}(B))^2\right)^2}{4\left(\lambda_{\min}(Z_B^T A Z_B)\right)^2} + (\sigma_{\min}(B))^2}, \\
l_3 &\geq \frac{1}{2}\left(-\bar{l}_3 + \sqrt{\bar{l}_3^2 + 4\left(\sigma_{\min}(BZ_A)\right)^2}\right), \\
\bar{l}_3 &= \frac{\left(\sigma_{\min}(BZ_A)\right)^2 + \left(\sigma_{\max}(BY_A)\right)^2}{\lambda_{\min^+}(A)}.
\end{aligned}$$

If $m = n$, we have $l^+ = \sigma_{\min}(B)$.
If $m < n$ and either A is nonsingular, then $l^+ = \max\left(l_1, l_2\right)$.

4.2 Target $\hat{u}$ defined over whole domain Ω

Suppose that the target $\hat{u}$ is defined over the whole of the domain Ω. We will consider the p-dimensional problem with $p \in \{2, 3\}$. It will be necessary to consider the cases $\beta \geq \frac{1}{2}$ and $\beta < \frac{1}{2}$ separately.

Consider the case $\beta \geq \frac{1}{2}$. We will start by bounding the positive eigenvalues of $\mathcal{A}$ by applying Theorem 4.1. Now $\lambda_{\min}(A) = dh^p$, $\|A\| = 2\beta Dh^p$ and there exist constants $\tilde{c} \geq c$ and $\tilde{C} \approx C$ such that $\sigma_{\min}(B) = \tilde{c}h^p$ and $\sigma_{\max}(B) = \tilde{C}h^{p-2}$. Hence, the positive eigenvalues of $\mathcal{A}$ lie in

$$\left[dh^p, h^{p-2}\left(\beta Dh^2 + \sqrt{\beta^2 D^2 h^4 + \tilde{C}^2}\right)\right].$$

From [26, pp. 101-2] we obtain

Theorem 4.4 If $\mathcal{M}$ and $\mathcal{M} + \mathcal{E} \in \mathbb{R}^{N \times N}$ are symmetric matrices, then

$$\lambda_k(\mathcal{M}) + \lambda_{\min}(\mathcal{E}) \leq \lambda_k(\mathcal{M} + \mathcal{E}) \leq \lambda_k(\mathcal{M}) + \lambda_{\max}(\mathcal{E}), \quad k = 1, \ldots, N.$$

Let $\mathcal{A} = \mathcal{M} + \mathcal{E}$, where

$$\mathcal{M} = \begin{bmatrix} 2\beta M & 0 & 0 \\ 0 & M & K^T \\ 0 & K & 0 \end{bmatrix} \quad \text{and} \quad \mathcal{E} = \begin{bmatrix} 0 & 0 & -M \\ 0 & 0 & 0 \\ -M & 0 & 0 \end{bmatrix}.$$

The matrix $\mathcal{M}$ is block diagonal with one of the block being of saddle-point form and the other equal to $2\beta M$: applying Theorem 2.1 and Theorem 4.1, and noting that $K \in \mathbb{R}^{n \times n}$, we find that the $\mathcal{M}$ has n negative eigenvalues that lie in

$$\left[\tfrac{1}{2}h^{p-2}\left(dh^2 - \sqrt{d^2h^4 + 4C^2}\right), \tfrac{1}{2}h^p\left(D - \sqrt{D^2 + 4c^2}\right)\right].$$

Applying Theorem 4.4, $\mathcal{A}$ will have n eigenvalues that lie in

$$\left[\tfrac{1}{2}h^{p-2}\left(dh^2 - 2Dh^2 - \sqrt{d^2h^4 + 4C^2}\right), \tfrac{1}{2}h^p\left(3D - \sqrt{D^2 + 4c^2}\right)\right].$$

Since $D < \frac{2c}{3}$, these n eigenvalues will all be negative and this accounts for all of the negative eigenvalues of $\mathcal{A}$.

Corollary 4.5 Consider the p-dimensional problem with $p \in \{2,3\}$. Let

$$\mathcal{A} = \begin{bmatrix} 2\beta M & 0 & -M \\ 0 & M & K^T \\ -M & K & 0 \end{bmatrix},$$

assume that Theorem 2.1 holds and $\beta > \frac{1}{2}$. Then there exists a constant $\tilde{C} \approx C$ such that

$$\lambda(\mathcal{A}) \subset I^- \cup I^+,$$

where

$$\begin{aligned} I^- &= \left[\tfrac{1}{2}h^{p-2}\left(dh^2 - 2Dh^2 - \sqrt{d^2h^4 + 4C^2}\right), \tfrac{1}{2}h^p\left(3D - \sqrt{D^2 + 4c^2}\right)\right], \\ I^+ &= \left[dh^p, \left(\beta D h^p + \sqrt{\beta^2 D^2 h^{2p} + \tilde{C}^2 h^{2p-4}}\right)\right]. \end{aligned}$$

If $\beta \gg \frac{\tilde{C}}{D}h^{-2}$, then

$$\begin{aligned} \sigma_{\min}(\mathcal{A}) &\geq \tfrac{1}{2}h^p\left(\sqrt{D^2 + 4c^2} - 3D\right), \\ \sigma_{\max}(\mathcal{A}) &\leq \left(\beta h^p + \sqrt{\beta^2 h^{2p} + \tilde{C}^2 h^{2p-4}}\right) \approx 2\beta h^p, \end{aligned}$$

giving, $\kappa(\mathcal{A}) = \mathcal{O}(\beta^{-1})$.

Consider the case $\beta \leq \frac{1}{2}$. Initially we will apply Theorem 4.1 to bound the eigenvalues of $\mathcal{A}$. Now, $\lambda_{\min}(A) = 2\beta d h^p$, $\|A\| = Dh^2$, and there exist constants $\tilde{c} \geq c$ and $\tilde{C} \approx C$ such that $\sigma_{\min}(B) = \tilde{c}h^p$ and $\sigma_{\max}(B) = \tilde{C}h^{p-2}$. Therefore, $\lambda(\mathcal{A}) \in I^- \cup I^+$, where

$$\begin{aligned} I^- &= \left[h^{p-2}\left(\beta dh^2 - \sqrt{\beta^2 d^2 h^4 + \tilde{C}^2}\right), \tfrac{h^p}{2}\left(D - \sqrt{D^2 + 4\tilde{c}^2}\right)\right], && (4.13) \\ I^+ &= \left[2\beta d h^p, \tfrac{1}{2}h^{p-2}\left(Dh^2 + \sqrt{D^2h^4 + 4\tilde{C}^2}\right)\right]. && (4.14) \end{aligned}$$

Alternatively, define

$$\mathcal{M} = \begin{bmatrix} 0 & 0 & -M \\ 0 & M & K \\ -M & K & 0 \end{bmatrix}, \mathcal{E} = \begin{bmatrix} 2\beta M & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

then $\mathcal{A} = \mathcal{M} + \mathcal{E}$.

We will use Theorem 4.2 to bound the eigenvalues of $\mathcal{M}$. Now, $Y_A^T A Y_A = M$, $BY_A = K$, $BZ_A = -M$ and there exist constants $\tilde{c} \geq c$ and $\tilde{C} \approx C$ such that $\sigma_{\min}(B) = \tilde{c}h^p$ and $\sigma_{\max}(B) = \tilde{C}h^{p-2}$. Let $\tilde{Z} = [-M^{-1}, K^{-1}]^T$, then

$$\begin{aligned} \mu_{\min}^Z &= \min_{x_z} \frac{x_z^T \tilde{Z}^T A \tilde{Z} x_z}{x_z^T \tilde{Z}^T \tilde{Z} x_z} = \min_{\tilde{x}_z} \frac{\tilde{x}_z^T K^{-1} M K^{-1} \tilde{x}_z}{\tilde{x}_z^T \left(M^{-2} + K^{-2}\right) \tilde{x}_z} \geq \frac{d^3 c^2 h^{p+4}}{C^2\left(d^2 + c^2\right)}, \\ \mu_{\max}^Z &= \max_{x_z} \frac{x_z^T \tilde{Z}^T A \tilde{Z} x_z}{x_z^T \tilde{Z}^T \tilde{Z} x_z} = \max_{\tilde{x}_z} \frac{\tilde{x}_z^T K^{-1} M K^{-1} \tilde{x}_z}{\tilde{x}_z^T \left(M^{-2} + K^{-2}\right) \tilde{x}_z} \leq \frac{D^3 C^2 h^p}{c^2\left(D^2 h^4 + C^2\right)} \end{aligned}$$

In fact, $\mu^Z_{\min} = c_1 h^{p+4}$ and $\mu^Z_{\max} = c_2 h^p$, where $c_1 \approx \frac{d^3 c^2}{C^2(d^2+c^2)}$ and $c_2 \approx \frac{D^3}{c^2}$ are constants independent of h.

Combining Corollary 4.3 and Theorem 4.4, and using the fact that $\lambda_{\min}(\mathcal{E}) = 0$ and $\lambda_{\max}(\mathcal{E}) = 2\beta d h^2$, we are able to find alternative bounds for the eigenvalues of $\mathcal{A} : \lambda(\mathcal{A}) \subset \tilde{I}^- \cup \tilde{I}^+$, where $\mathcal{A}$ is defined by (4.1),

$$\begin{aligned}
\tilde{I}^- &= \left[-\tilde{C} h^{p-2}, \tfrac{h^p}{2}\left(D + 4\beta D - \sqrt{D^2 + 4\tilde{c}^2}\right)\right], && (4.15)\\
\tilde{I}^+_1 &= \left[c_1 h^{p+4}, \left(c_2 + 2\beta D\right) h^p\right], && (4.16)\\
\tilde{I}^+_2 &= \left[\tfrac{1}{2} h^p \left(d + \sqrt{d^2 + 4c^2}\right), \tfrac{1}{2} h^{p-2}\left(\left(1 + 2\beta\right) D h^2 + \sqrt{D^2 h^4 + 4C}\right)\right], && (4.17)\\
\tilde{I}^+_3 &= \left[\max\left(l_1, \min\left(l_2, l_3\right)\right), \tfrac{1}{2} h^{p-2}\left(\left(1 + 2\beta\right) D h^2 + \sqrt{D^2 h^4 + 4\tilde{C}}\right)\right], && (4.18)
\end{aligned}$$

where

$$\begin{aligned}
l_1 &\geq -\frac{D^2 + \tilde{c}^2}{2c_1} h^{p-4} + \frac{D^2 + \tilde{c}^2}{2c_1} h^{p-4} \sqrt{1 + \frac{4c_1^2 \tilde{c}^2 h^8}{\left(D^2 + \tilde{c}^2\right)^2}}\\
&= \frac{c_1 \tilde{c}^2}{D^2 + \tilde{c}^2} h^{p+4} - \frac{c_1^3 \tilde{c}^4}{\left(D^2 + \tilde{c}^2\right)^3} h^{p+12} + \mathcal{O}(h^{p+20})\\
&\geq c_3 h^{p+4},\\
l_2 &= \tfrac{1}{2} h^p \left(d + \sqrt{d^2 + 4\tilde{c}^2}\right),\\
l_3 &\geq \tfrac{1}{2}\left(-\frac{\left(d^2 h^4 + C^2\right) h^{p-4}}{d}\right) + \sqrt{\frac{\left(d^2 h^4 + C^2\right)^2 h^{2p-8}}{d^2} + 4d^2 h^{2p}}\\
&= \tfrac{1}{2}\left(-\frac{\left(d^2 h^4 + C^2\right) h^{p-4}}{d}\right) + \frac{\left(d^2 h^4 + C^2\right) h^{p-4}}{d} \sqrt{1 + \frac{4d^4 h^8}{\left(d^2 h^4 + C^2\right)^2}}\\
&= \frac{d^3 h^{p+4}}{d^2 h^4 + C^2} - \frac{d^7 h^{p+12}}{\left(d^2 h^4 + C^2\right)^3} + \mathcal{O}(h^{p+20})\\
&\geq c_4 h^{p+4},
\end{aligned}$$

where c_3 and c_4 are constants independent of h and β. Hence $\lambda_{\min^+}(\mathcal{A}) \geq \max\left(c_3, c_4\right) h^{p+4}$.

Combining (4.13), (4.14), (4.15), (4.16), (4.17), and (4.18) we obtain the following result.

Corollary 4.6 Consider the p-dimensional problem with $p \in \{2,3\}$. Let

$$\mathcal{A} = \begin{bmatrix} 2\beta M & 0 & -M \\ 0 & M & K^T \\ -M & K & 0 \end{bmatrix},$$

assume that Theorem 2.1 holds and $\beta < \frac{1}{2}$. Let $\lambda(\mathcal{A})$ denote the spectrum of $\mathcal{A}$. Then there exist constants c_1 and $\tilde{C} \approx C$ such that

$$\lambda(\mathcal{A}) \subset I^- \cup I^+,$$

where,

$$\begin{aligned} I^- &= \left[h^{p-2}\left(\beta d h^2 - \sqrt{\beta^2 d^2 h^4 + \tilde{C}^2}\right), \tfrac{h^p}{2}\left(D - \sqrt{D^2 + 4\tilde{c}^2}\right)\right], \\ I^+ &= \left[\max\left(2\beta d h^p, c_1 h^{p+4}\right), \tfrac{1}{2}h^{p-2}\left(Dh^2 + \sqrt{D^2 h^4 + 4\tilde{C}^2}\right)\right]. \end{aligned}$$

For both the 2D and 3D cases, if $\frac{c_1 h^4}{2d} \leq \beta < \frac{1}{2}$, we will expect the condition number of $\mathcal{A}$ to be at most inversely proportional to both β and h^2. For $\beta < \frac{c_1 h^4}{2d}$, we will expect the condition number to be independent of β but at most inversely proportional to h^6.

In Figure 4.1, we plot the condition number of $\mathcal{A}$ with respect to β for Example 3.1 in 3D (left) and Example 3.2 in 2D (right) with a target $\hat{u}$ defined over the whole of Ω. Results are given for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and, for the 2D problem, $h = \frac{1}{32}$. We observe that, as expected, if $\beta \gg \frac{\tilde{C}}{D}h^{-2}$, then the condition number of $\mathcal{A}$ is proportional to β but (essentially) independent of the mesh size h. For $\frac{c_4 h^4}{2d} \leq \beta \leq \frac{1}{2}$, the condition number varies inversely proportionally with β. Additionally, the condition number is inversely proportional to h^2. Finally, for very small β, the condition number is independent of the regularization parameter but inversely proportional to h^6.

4.3 Distributed control problems with target $\hat{u}$ only defined over a subdomain

In the case where the target $\hat{u}$ is not defined over all of the domain Ω, the matrix A defined in (4.2) will be positive semi-definite and singular. As a result, we will use Corollary 4.3 to obtain bounds for the eigenvalues of $\mathcal{A}$.

Let $\mathcal{A} = \mathcal{M} + \mathcal{E}$, where

$$\mathcal{M} = \begin{bmatrix} 2\beta M & 0 & 0 \\ 0 & \bar{M} & K^T \\ 0 & K & 0 \end{bmatrix} \quad \text{and} \quad \mathcal{E} = \begin{bmatrix} 0 & 0 & -M \\ 0 & 0 & 0 \\ -M & 0 & 0 \end{bmatrix}.$$

The matrix $\mathcal{M}$ is block diagonal with one of the block being of saddle-point form and the other equal to $2\beta M$: applying Theorem 2.1, Theorem 2.3 and Corollary 4.3, and noting that $K \in \mathrm{I\!R}^{n\times n}$, we find that the $\mathcal{M}$ has n negative eigenvalues that lie in

$$\left[-Ch^{p-2}, \tfrac{1}{2}h^p\left(\bar{D} - \sqrt{\bar{D}^2 + 4c^2}\right)\right].$$

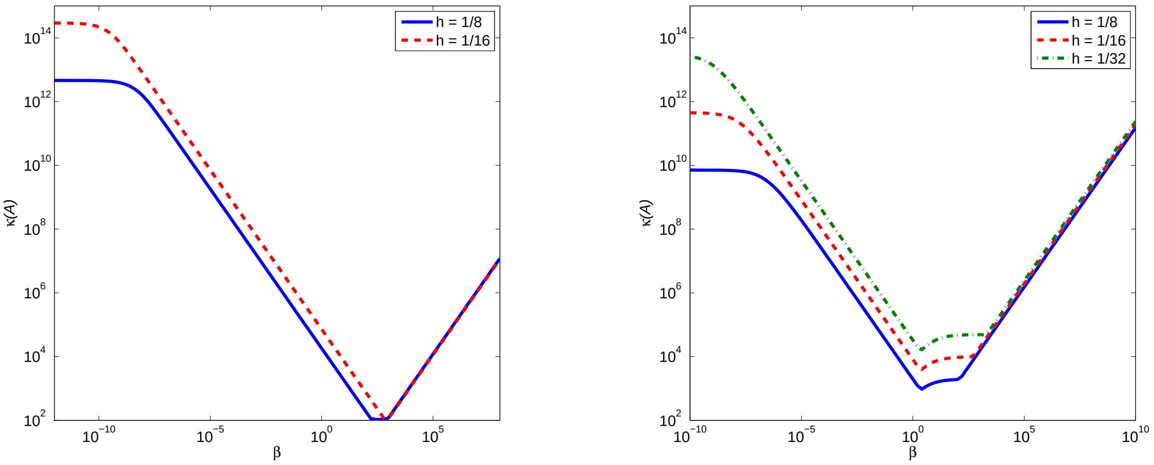

Figure 4.1: Condition number of $\mathcal{A}$ with respect to β for Example 3.1 in 3D (left) and Example 3.2 in 2D (right) with a target $\hat{u}$ defined over the whole of Ω. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and, for the 2D problem, $h = \frac{1}{32}$.

Applying Theorem 4.4, $\mathcal{A}$ will have n eigenvalues that lie in

$$\left[-\left(Dh^2 + C\right) h^{p-2}, \tfrac{1}{2}h^p \left(2D + \bar{D} - \sqrt{\bar{D}^2 + 4c^2}\right)\right].$$

Since $\bar{D} \leq D < \frac{2c}{3}$, these n eigenvalues will all be negative and this accounts for all of the negative eigenvalues of $\mathcal{A}$. Additionally, the positive eigenvalues of $\mathcal{M}$ will bounded from below by $\min\left(2\beta d h^p, \frac{1}{2}h^p \left(\bar{d} + \sqrt{\bar{d}^2 + 4\tilde{c}^2}\right)\right)$. Assuming that $\beta > 0.5d^{-1}\left(\bar{d} + \sqrt{\bar{d}^2 + 4\tilde{c}^2}\right)$, we obtain $\lambda_{\mathrm{min}^+}(\mathcal{A}) \geq \frac{1}{2}h^p \left(\bar{d} - 2D + \sqrt{\bar{d}^2 + 4\tilde{c}^2}\right) > 0$. Applying Corollary 4.3 to $\mathcal{A}$, we find that the positive eigenvalues are bounded from above by $h^{p-2}\left(\beta D h^2 + \sqrt{\beta^2 D^2 h^4 + \tilde{C}^2}\right)$.

Corollary 4.7 Let

$$\mathcal{A} = \begin{bmatrix} 2\beta M & 0 & -M \\ 0 & \bar{M} & K^T \\ -M & K & 0 \end{bmatrix},$$

assume that Theorem 2.3 holds and $\beta > 0.5d^{-1}\left(\bar{d} + \sqrt{\bar{d}^2 + 4\tilde{c}^2}\right)$. There exist constants $\tilde{c} \geq c$ and $\tilde{C} \approx C$ independent of β and h such that

$$\lambda(\mathcal{A}) \subset I^- \cup I^+,$$

where

$$\begin{aligned} I^- &= \left[-\left(Dh^2 + C\right) h^{p-2}, \tfrac{1}{2}h^p \left(2D + \bar{D} - \sqrt{\bar{D}^2 + 4c^2}\right)\right], \\ I^+ &= \left[\tfrac{1}{2}h^p \left(\bar{d} - 2D + \sqrt{\bar{d}^2 + 4\tilde{c}^2}\right), h^{p-2}\left(\beta D h^2 + \sqrt{\beta^2 D^2 h^4 + \tilde{C}^2}\right)\right]. \end{aligned}$$

If $\beta \gg \frac{\tilde{C}}{D}h^{-2}$, then there exist constants c_1 and C_1 such that

$$\sigma_{\min}(\mathcal{A}) \geq c_1 h^p \quad \text{and} \quad \sigma_{\max}(\mathcal{A}) \leq \beta C_1 h^p,$$

giving, $\kappa(\mathcal{A}) = \mathcal{O}(\beta^{-1})$.

Consider the case $\beta \ll \frac{\bar{D}}{2D}$. Similarly to the case where $\hat{u}$ is defined over the whole of Ω, applying Theorem 2.1, Theorem 2.3 and Corollary 4.3 we obtain the following result.

Corollary 4.8 Let

$$\mathcal{A} = \begin{bmatrix} 2\beta M & 0 & -M \\ 0 & \bar{M} & K^T \\ -M & K & 0 \end{bmatrix},$$

assume that Theorem 2.3 holds and $\beta \ll \frac{\bar{D}}{2D}$. Then there exist constants c_1, c_2 and $\tilde{C} \approx C$ independent of β and h such that

$$\lambda(\mathcal{A}) \subset I^- \cup I^+,$$

where,

$$\begin{aligned} I^- &= \left[-\tilde{C}h^{p-2}, \tfrac{1}{2}h^p\left(\bar{D} - \sqrt{\bar{D}^2 + 4c^2}\right)\right], \\ I^+ &= \left[\beta c_2 h^p, \tfrac{1}{2}h^{p-2}\left(\bar{D}h^2 + \sqrt{\bar{D}^2 h^4 + 4\tilde{C}^2}\right)\right]. \end{aligned}$$

Thus, for $\beta \ll \frac{\bar{D}}{2D}$, we will expect the condition number to grow at most inversely proportionally with β and h^2.

In Figure 4.2, we plot the condition number of $\mathcal{A}$ with respect to β for Example 3.1 in 3D (left) and Example 3.2 in 2D (right) with Target 3. Results are given for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and, for the 2D problem, $h = \frac{1}{32}$. We observe that, as expected, if $\beta \gg \frac{\tilde{C}}{D}h^{-2}$, then the condition number of $\mathcal{A}$ is proportional to β but (essentially) independent of the mesh size h. For $\beta \ll \frac{\bar{D}}{2D}$, the condition number varies inversely proportionally with β and h^2 : this is as we expected.

4.4 Effect of β on direct solvers applied to the saddle-point problem

Suppose that we wish to solve a system of the form $\mathcal{A}s = b$, where $\mathcal{A} \in \mathbb{R}^{N \times N}$ is symmetric, by using a backward-stable direct method. If $\mathcal{A}$ is nonsingular but ill-conditioned, the relative sensitivity of the solution is bounded by (and in the worse case equal to) the condition number of $\mathcal{A}$ multiplied by the relative perturbations in b or $\mathcal{A}$, [13]. In this paper, we will only consider relative perturbations in $\mathcal{A}$.

When the matrix $\mathcal{A}$ changes by $\Delta\mathcal{A}$, the exact solution $\tilde{s}$ of the perturbed system satisfies

$$(\mathcal{A} + \Delta\mathcal{A})\,\tilde{s} = \mathcal{A}s = b, \quad \text{or} \quad \tilde{s} - s = -\left(\mathcal{A} + \Delta\mathcal{A}\right)^{-1}\Delta\mathcal{A}s. \tag{4.1}$$

If $\kappa(\mathcal{A}) \approx \kappa(\mathcal{A} + \Delta\mathcal{A})$, then we may ignore second-order terms and an approximation to (4.1) is satisfied by $\Delta s \approx \tilde{s} - s$:

$$\mathcal{A}\Delta s = -\Delta\mathcal{A}s, \tag{4.2}$$

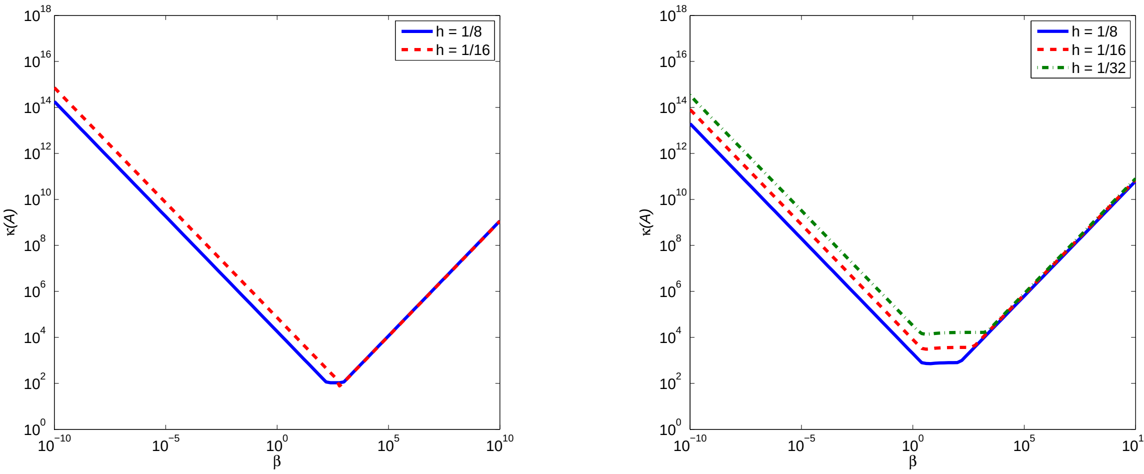

Figure 4.2: Condition number of $\mathcal{A}$ with respect to β for Example 3.1 in 3D (left) and Example 3.2 in 2D (right) with Target 3. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and, for the 2D problem, $h = \frac{1}{32}$.

from which we obtain the bound

$$\|\Delta s\| \leq \left\|\mathcal{A}^{-1}\right\|_2 \|\Delta\mathcal{A}\|_2 \|s\| \,. \tag{4.3}$$

Equality can hold in this relation, [13].

We may assume that $\Delta\mathcal{A} = (\Delta\mathcal{A})^T$ [4, Theorem 3]. For the most common backward-stable methods performed on a machine with unit roundoff $\mathtt{u}$, the perturbation $\Delta\mathcal{A}$ satisfies

$$\|\Delta\mathcal{A}\| \leq \mathtt{u}\gamma_N \|\mathcal{A}\| \,, \tag{4.4}$$

where γ_N is a function containing a low-order polynomial in N and charateristics of $\mathcal{A}$ such as the growth factor. Charaterizations of γ_N are known for various conditions:

- the Cholesky factorization when $\mathcal{A}$ is sufficiently positive definite, [12];
- the symmetric indefinite factorization with partial pivoting, [13];
- Gaussian elimination with partial pivoting, [13];
- the modified Cholesky factorizations of [8] and [21], see [5].

If extreme growth is not exhibited (as we expect the case to be), then γ_N is of reasonable size for all of these methods, i.e., $\mathtt{u}\gamma_N \ll 1$.

Combining (4.3) and (4.4) we obtain

$$\|\Delta s\| \leq \mathtt{u}\gamma_N \kappa(\mathcal{A}) \|s\| \,. \tag{4.5}$$

Thus, if condition number of $\mathcal{A}$ is small, then the error will be small. The converse is not true but it might be the case for some problems.

In interior-point methods, the singular values of the linear system split into two subgroups. Wright [27] was able to use the fact that these subgroups are well-behaved to show that the portion of the solution associated with one of these subgroups has an absolute error bound comparable to machine precision even though the overall system is extremely ill-conditioned. We will use similar arguments to show that backward-stable methods applied to some of our linear systems will achieve much better accuracy than we might expect from (4.5).

Let $\mathcal{A}$ be factorized as

$$\mathcal{A} = U\Sigma V^T = \left[\begin{array}{cc} U_L & U_S \end{array} \right] \left[\begin{array}{cc} \Sigma_L & 0 \\ 0 & \Sigma_S \end{array} \right] \left[\begin{array}{c} V_L^T \\ V_S^T \end{array} \right], \tag{4.6}$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix whose diagonal entries are all positive and ordered in decreasing order. Let Σ_L have dimension $\hat{N}$. Assume that $0 < \hat{N} < N$ and $\sigma_{\hat{N}} > \sigma_{\hat{N}+1}$.

Clearly, $\|\mathcal{A}\| = \|\Sigma_L\|$, $\|\mathcal{A}^{-1}\| = \|\Sigma_S^{-1}\|$ and $\kappa(\mathcal{A}) = \|\Sigma_L\| \|\Sigma_S^{-1}\|$. Suppose that Σ_L and Σ_S are individually much better conditioned than $\mathcal{A}$, i.e.,

$$\frac{\sigma_1}{\sigma_{\hat{N}}} \ll \frac{\sigma_1}{\sigma_N} \quad \text{and} \quad \frac{\sigma_{\hat{N}+1}}{\sigma_N} \ll \frac{\sigma_1}{\sigma_N}.$$

This can clearly be the case for the problems considered in this paper.

We wish to solve $\mathcal{A}s = b$. Writing

$$\begin{aligned} b &= b_L + b_S = U_L\delta_L + U_S\delta_S, \\ s &= s_L + s_S = V_L\psi_L + V_S\psi_S, \end{aligned} \tag{4.7}$$

and using the fact that U and V are orthogonal matrices we obtain

$$\|b\|^2 = \|\delta_L\|^2 + \|\delta_S\|^2, \quad \|b_L\| = \|\delta_L\| \quad \text{and} \quad \|b_S\| = \|\delta_S\|.$$

We can similarly relate s and ψ. Solving $\mathcal{A}s = b$ is equivalent to solving

$$\Sigma\psi = \left[\begin{array}{c} \Sigma_L\psi_L \\ \Sigma_S\psi_S \end{array} \right] = \left[\begin{array}{c} \delta_L \\ \delta_S \end{array} \right] = \delta. \tag{4.8}$$

From (4.8) we obtain

$$\|b_L\| \leq \|\Sigma_L\| \|s_L\|, \quad \|b_S\| \leq \|\Sigma_S\| \|s_S\|, \tag{4.9}$$

and

$$\|s_L\| \leq \|\Sigma_L^{-1}\| \|b_L\|, \quad \|s_S\| \leq \|\Sigma_S^{-1}\| \|b_S\|. \tag{4.10}$$

When the matrix $\mathcal{A}$ changes, we can use the first-order approximation (4.1), $\Delta s = -\mathcal{A}^{-1}\Delta\mathcal{A}s$. Let $\Delta\mathcal{A} = UGV^T$ for some matrix G, then $\|\Delta\mathcal{A}\| = \|G\|$. Now, $G = U^T\Delta\mathcal{A}V$ and we partition G as

$$G = \left[\begin{array}{c} G_L \\ G_S \end{array} \right] = \left[\begin{array}{cc} G_{L1} & G_{L2} \\ G_{S1} & G_{S2} \end{array} \right].$$

Suppose that we also express Δs as a linear combination of the columns of V, that is, $\Delta s = V\Delta\psi$, then we have

$$\left[\begin{array}{c} \Delta\psi_L \\ \Delta\psi_S \end{array} \right] = - \left[\begin{array}{c} \Sigma_L^{-1}G_L\psi \\ \Sigma_S^{-1}G_S\psi \end{array} \right]. \tag{4.11}$$

This implies that

$$\begin{aligned}\|\Delta s_L\| &\leq \left\|\Sigma_L^{-1}\right\| \|G_L\| \|s\| \leq \left\|\Sigma_L^{-1}\right\| \|\Delta \mathcal{A}\| \|s\|, &(4.12)\\ \|\Delta s_S\| &\leq \left\|\Sigma_S^{-1}\right\| \|G_S\| \|s\| \leq \left\|\Sigma_S^{-1}\right\| \|\Delta \mathcal{A}\| \|s\|. &(4.13)\\ & &(4.14)\end{aligned}$$

Since $\|\Sigma_L\| = \|\mathcal{A}\|$, (4.12) implies that

$$\frac{\|\Delta s_L\|}{\|s\|} \leq \left\|\Sigma_L^{-1}\right\| \|\Sigma_L\| \frac{\|\Delta \mathcal{A}\|}{\|\mathcal{A}\|},$$

so that the change in s_l relative to s compared to the relative perturbation in $\mathcal{A}$ can only be blown up by $\kappa(\Sigma_L)$ rather than $\kappa(\mathcal{A})$. In contrast, the perturbation in s_S relative to s can, in general, be blown up by $\kappa(\mathcal{A})$. We can use the structure of G to give better bounds for Δs_L and Δs_S.

4.5 Saddle-point formulation: target $\hat{u}$ defined over the whole of Ω

Initially, we will assume that β is large. From [27, Theorem 3.1], we have the following theorem.

Theorem 4.9 Let $\mathcal{M}$ denote a real symmetric matrix, and define the perturbed matrix $\tilde{\mathcal{M}}$ as $\mathcal{M} + \mathcal{E}$, where $\mathcal{E}$ is symmetric. Consider an orthogonal matrix $[X_1, X_2]$, where X_1 has l columns, such that the range(X_1) is a simple invariant subspace of $\mathcal{M}$, where

$$\begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} \mathcal{M} \begin{bmatrix} X_1 & X_2 \end{bmatrix} = \begin{bmatrix} L_1 & 0 \\ 0 & L_2 \end{bmatrix} \text{and} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix} \mathcal{E} \begin{bmatrix} X_1 & X_2 \end{bmatrix} = \begin{bmatrix} E_{11} & E_{12} \\ E_{12}^T & E_{22} \end{bmatrix}.$$

Let $d_1 = \operatorname{sep}(L_1, L_2) - \|E_{11}\| - \|E_{22}\|$ and $v = \|E_{12}\|/d_1$, where $\operatorname{sep}(L_1, L_2) = \min_{i,j} |\lambda_i(L_1) - \lambda_j(L_2)|$. If $d_1 > 0$ and $v < \frac{1}{2}$, then there are orthonormal bases $\tilde{X}_1$ and $\tilde{X}_2$ for simple invariant subspaces of the perturbed matrix $\tilde{\mathcal{M}}$ satisfying $\left\|X_1 - \tilde{X}_1\right\| \leq 2v$ and $\left\|X_2 - \tilde{X}_2\right\| \leq 2v$.

Suppose that we let $\mathcal{A} = \mathcal{M} + \mathcal{E}$, where

$$\mathcal{M} = \left[\begin{array}{c|cc} 2\beta M & 0 & 0 \\ \hline 0 & M & K^T \\ 0 & K & 0 \end{array}\right] \text{ and } \mathcal{E} = \left[\begin{array}{c|cc} 0 & 0 & -M \\ \hline 0 & 0 & 0 \\ -M & 0 & 0 \end{array}\right].$$

If

$$X_1 = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix} \text{ and } X_2 = \begin{bmatrix} 0 & 0 \\ I & 0 \\ 0 & I \end{bmatrix},$$

then $[X_1, X_2]$ is orthogonal, and both range(X_1) and range(X_2) are simple invariant subspaces of $\mathcal{M}$. From Theorem 4.9, we have

$$L_1 = 2\beta M, \quad L_2 = \begin{bmatrix} M & K \\ K & 0 \end{bmatrix}.$$

Correspondingly, $E_{11}=0$, $E_{22}=0$, and $E_{12}=[0,-M]$. Observe that L_2 is of saddle-point form (4.1), with $A=M$ and $B=K$. From Theorems 2.1 and 4.1, the eigenvalues of L_2 lie in $I^-\cup I^+$, where

$$I^- \quad = \quad \left[\tfrac{1}{2}h^{p-2}\left(dh^2-\sqrt{d^2h^4+4C^2}\right), \tfrac{1}{2}h^p\left(D-\sqrt{D^2+4c^2}\right)\right]$$

and

$$I^+ \quad = \quad \left[dh^p, \tfrac{1}{2}h^{p-2}\left(Dh^2+\sqrt{D^2h^4+4C^2}\right)\right].$$

Thus, $d_1 := \operatorname{sep}(L_1,L_2)-\|E_{11}\|-\|E_{22}\| \geq 2\beta dh^p - \frac{1}{2}h^{p-2}\left(Dh^2+\sqrt{D^2h^4+4C^2}\right) > 0$ for $\beta \gg \frac{1}{4d}h^{-2}\left(Dh^2+\sqrt{D^2h^4+4C^2}\right)$. Now,

$$v := \|E_{12}\|/d_1 = \frac{2Dh^2}{4\beta dh^2-\left(Dh^2+\sqrt{D^2h^4+4C^2}\right)}.$$

Hence, $v\to 0$ as $\beta\to+\infty$.

Suppose that $\beta \gg \frac{C}{d}h^{-2}$, then Corollary 4.5 and its derivation tells us that if $M\in\mathbb{R}^{n\times n}$, $\mathcal{A}$ has n singular values that are $\mathcal{O}(\beta)$; the remaining eigenvalues are $\mathcal{O}(1)$. We shall assume that the mesh size h remains fixed. From the derivation of Corollary 4.5 and Theorem 4.9, we find that there are orthonormal bases $\tilde{X}_1$ and $\tilde{X}_2$ such that

$$\tilde{X}_1 = \begin{bmatrix} I_m \\ 0 \end{bmatrix} + \tfrac{D}{\beta d}\begin{bmatrix} \Upsilon_{11} \\ \Upsilon_{12} \end{bmatrix} + \mathcal{O}(\beta^{-2}), \quad \tilde{X}_2 = \begin{bmatrix} 0 \\ I_{2m} \end{bmatrix} + \tfrac{D}{\beta d}\begin{bmatrix} \Upsilon_{21} \\ \Upsilon_{22} \end{bmatrix} + \mathcal{O}(\beta^{-2}), \tag{4.15}$$

and

$$\begin{bmatrix} \tilde{X}_1^T \\ \tilde{X}_2^T \end{bmatrix} \mathcal{A} \begin{bmatrix} \tilde{X}_1 & \tilde{X}_2 \end{bmatrix} = \begin{bmatrix} \tilde{L}_1 & 0 \\ 0 & \tilde{L}_2 \end{bmatrix},$$

where Υ_1 and Υ_2 are $\mathcal{O}(1)$, and $\tilde{L}_1$ has eigenvalues equal to the n eigenvalues of $\mathcal{A}$ that are $\mathcal{O}(\beta)$. If we write the singular value decompositions of $\tilde{L}_1$ and $\tilde{L}_2$ as $\tilde{L}_1 = \tilde{U}_L\Sigma_L\tilde{U}_L^T$ and $\tilde{L}_2 = \tilde{U}_S\Sigma_S\tilde{V}_S^T$, we may factorize $\mathcal{A}$ as

$$\mathcal{A} = \begin{bmatrix} U_L & U_S \end{bmatrix} \begin{bmatrix} \Sigma_L & 0 \\ 0 & \Sigma_S \end{bmatrix} \begin{bmatrix} V_L^T \\ V_S^T \end{bmatrix},$$

where

$$U_L = \tilde{X}_1\tilde{U}_L, \quad U_S = \tilde{X}_2\tilde{U}_S, \quad V_L = \tilde{X}_1\tilde{U}_L, \quad \text{and} \quad V_S = \tilde{X}_2\tilde{V}_S. \tag{4.16}$$

When $\beta \gg \frac{C}{d}h^{-2}$, we find that the solution of (2.6) satisfies $\mathbf{f}=\mathcal{O}(\beta^{-1})$, $\lambda=\mathcal{O}(1)$ and $\mathbf{u}=\mathcal{O}(1)$, Section 5. Using (4.7) and defining $g=[\mathbf{u}^T,\lambda^T]^T$, we obtain

$$\begin{bmatrix} \mathbf{f} \\ g \end{bmatrix} = \begin{bmatrix} \tilde{U}_L\psi_L \\ 0 \end{bmatrix} + \frac{D}{\beta d}\begin{bmatrix} \Upsilon_{11} \\ \Upsilon_{12} \end{bmatrix}\tilde{U}_L\psi_L + \begin{bmatrix} 0 \\ \tilde{U}_S\psi_S \end{bmatrix} + \frac{D}{\beta d}\begin{bmatrix} \Upsilon_{21} \\ \Upsilon_{22} \end{bmatrix}\tilde{V}_S\psi_S + \mathcal{O}(\beta^{-2}).$$

This implies that, for large β, ψ_S is $\mathcal{O}(1)$ and we can introduce a vector ρ_L with $\mathcal{O}(1)$ entries such that $\rho_L=\beta\psi_L$.

For common backward-stable methods, we can assume that $|\Delta\mathcal{A}| \leq \mathrm{u}\gamma_N|\mathcal{A}|$, [13]. Hence, we may write $\Delta\mathcal{A}$ as

$$\Delta A = \mathrm{u}\gamma_N \begin{bmatrix} \beta E_{11} & E_{21}^T \\ E_{21} & E_{22} \end{bmatrix}, \tag{4.17}$$

where the entries of E_{11}, E_{12}, E_{21} and E_{22} are $\mathcal{O}(1)$. Using (4.15), (4.16) and the fact that $G = U^T\Delta\mathcal{A}V$, we find that

$$G = \begin{bmatrix} G_{L1} & G_{L2} \\ G_{S1} & G_{S2} \end{bmatrix} = \mathtt{u}\gamma_N \begin{bmatrix} \beta\tilde{U}_L^T E_{11}\tilde{U}_L + \hat{E}_{11} & \hat{E}_{12} \\ \hat{E}_{21} & \hat{E}_{22} \end{bmatrix} + \mathtt{u}\gamma_N\mathcal{O}(\beta^{-1}),$$

where $\|\hat{E}_{11}\|$, $\|\hat{E}_{12}\|$, $\|\hat{E}_{21}\|$ and $\|\hat{E}_{22}\|$ are $\mathcal{O}(1)$.

From (4.11) we have

$$\begin{aligned} \Delta\psi_L &= -\Sigma_L^{-1}\begin{bmatrix} G_{L1} & G_{L2} \end{bmatrix}\begin{bmatrix} \psi_L \\ \psi_S \end{bmatrix} \\ &= -\Sigma_L^{-1}\begin{bmatrix} \beta^{-1}G_{L1} & G_{L2} \end{bmatrix}\begin{bmatrix} \rho_L \\ \psi_S \end{bmatrix} \\ &= -\mathtt{u}\gamma_N\Sigma_L^{-1}\begin{bmatrix} \tilde{U}_L^T E_{11}\tilde{U}_L + \beta^{-1}\hat{E}_{11} + \mathcal{O}(\beta^{-2}) & \hat{E}_{12} + \mathcal{O}(\beta^{-1}) \end{bmatrix}\begin{bmatrix} \rho_L \\ \psi_S \end{bmatrix}. \end{aligned}$$

From the derivation of Corollary 4.5, we know that $\left\|\Sigma_L^{-1}\right\| = \frac{1}{2\beta d} + \mathcal{O}(\beta^{-2})$. Hence,

$$\begin{aligned} \|\Delta s_L\| &= \|\Delta\psi_L\| \\ &\leq \mathtt{u}\gamma_N \left\|\Sigma_L^{-1}\right\| \left\|\begin{bmatrix} \tilde{U}_L^T E_{11}\tilde{U}_L + \beta^{-1}\hat{E}_{11} + \mathcal{O}(\beta^{-2}) & \hat{E}_{12} + \mathcal{O}(\beta^{-1}) \end{bmatrix}\right\| \left\|\begin{bmatrix} \rho_L \\ \psi_S \end{bmatrix}\right\| \\ &= \mathtt{u}\left(\frac{c_2}{\beta} + \mathcal{O}(\beta^{-2})\right), \end{aligned}$$

where c_2 is a constant independent of β. Hence, for $\beta \gg \frac{c}{d}h^{-2}$, we will expect the change in s_L relative to s to be at most inversely proportional to the regularization parameter β. Similarly, we can show that we can expect $\frac{\Delta s_S}{s}$ to be bounded above by a constant that is of the order $\mathtt{u}\gamma_N$. Equation 4.3 gives an upper bound that is proportional to β, which is a gross overestimate. Finally, we observe that since $\tilde{U}_L\psi_L \to \mathbf{f}$ as $\beta \to \infty$ and $\mathbf{f} = \mathcal{O}(\beta^{-1})$, if $\beta \gg \frac{c}{d}h^{-2}$, then $\|\Delta\mathbf{f}\| / \|\mathbf{f}\|$ will be $\mathcal{O}(\mathtt{u})$. Also, $\tilde{U}_S\psi_S \to g$ as $\beta \to \infty$ and $g = \mathcal{O}(1)$. Hence, we will expect $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ to be $\mathcal{O}(\mathtt{u})$.

In Figure 4.1, we plot $\|\Delta s_L\| / \|s\|$,$\|\Delta s_S\| / \|s\|$, $\|\Delta\mathbf{f}\| / \|\mathbf{f}\|$ and $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ against β for Example 3.1 with Target 1. The solution s is calculated with the backslash command in Matlab, whilst $\tilde{s}$ is calculated by applying Matlab's `ldl` function to factor a single precision version of $\mathcal{A}$ and this factorization is then used to solve the system. For large β, we observe that, as expected, the change in s_L relative to s is inversely proportional to β but the change in s_S relative to s remains (approximately) constant. Also, as predicted, both $\|\Delta\mathbf{f}\| / \|\mathbf{f}\|$ and $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ are $\mathcal{O}(\mathtt{u})$.

We will now consider the case where β is small. Let $\mathcal{A} = \mathcal{M} + \mathcal{E}$, where

$$\mathcal{M} = \left[\begin{array}{c|cc} 2\beta M & 0 & 0 \\ \hline 0 & M & K^T \\ 0 & K & 0 \end{array}\right] \text{ and } \mathcal{E} = \left[\begin{array}{c|cc} 0 & 0 & -M \\ \hline 0 & 0 & 0 \\ -M & 0 & 0 \end{array}\right].$$

If

$$X_1 = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix} \text{ and } X_2 = \begin{bmatrix} 0 & 0 \\ I & 0 \\ 0 & I \end{bmatrix},$$

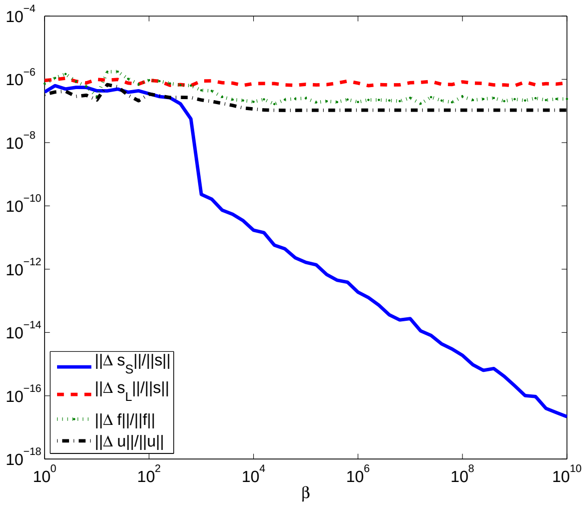

Figure 4.1: Plot of Δs_L and Δs_S with respect to β for Example 3.1 with Target 1. Results are shown for $h = \frac{1}{8}$.

then $[X_1, X_2]$ is orthogonal, and both range(X_1) and range(X_2) are simple invariant subspaces of $\mathcal{M}$. From Theorem 4.9, we obtain

$$L_1 = 2\beta M, \qquad L_2 = \begin{bmatrix} M & K \\ K & 0 \end{bmatrix}.$$

Correspondingly, $E_{11} = 0$, $E_{22} = 0$, and $E_{12} = [-M, K]^T$. Combining Theorem 4.1 and Corollary 4.3 we find that the eigenvalues of L_2 lie in $I^- \cup I^+$, where

$$\begin{aligned} I^- &= \left[\tfrac{1}{2}h^{p-2}\left(dh^2 - \sqrt{d^2h^4 + 4C^2}\right), \frac{1}{2}h^p\left(D - \sqrt{D^2 + 4c^2}\right)\right], \\ I^+ &= \left[ch^p, \frac{1}{2}h^{p-2}\left(Dh^2 + \sqrt{D^2h^4 + 4C^2}\right)\right]. \end{aligned}$$

Thus, $d_1 := \text{sep}(L_1, L_2) - \|E_{11}\| - \|E_{22}\| \geq ch^p - \beta Dh^p > 0$ for $\beta < \frac{c}{D}$. Now,

$$v := \|E_{12}\| / d_1 = \frac{D}{c - \beta D} < \tfrac{1}{2}.$$

From Theorem 4.9, we find that there exist orthonormal bases $\tilde{X}_1$ and $\tilde{X}_2$ such that

$$\tilde{X}_1 = \begin{bmatrix} I_m \\ 0 \end{bmatrix} + 2v \begin{bmatrix} \Upsilon_{11} \\ \Upsilon_{12} \end{bmatrix}, \quad \tilde{X}_2 = \begin{bmatrix} 0 \\ I_{2m} \end{bmatrix} + 2v \begin{bmatrix} \Upsilon_{21} \\ \Upsilon_{22} \end{bmatrix}, \tag{4.18}$$

and

$$\begin{bmatrix} \tilde{X}_1^T \\ \tilde{X}_2^T \end{bmatrix} \mathcal{A} \begin{bmatrix} \tilde{X}_1 & \tilde{X}_2 \end{bmatrix} = \begin{bmatrix} \tilde{L}_1 & 0 \\ 0 & \tilde{L}_2 \end{bmatrix},$$

where Υ_1 and Υ_2 are $\mathcal{O}(1)$, and $\tilde{L}_1$ has eigenvalues equal to the m eigenvalues of $\mathcal{A}$ that are $\mathcal{O}(\beta)$. If we write the singular value decompositions of $\tilde{L}_1$ and $\tilde{L}_2$ as $\tilde{L}_1 = \tilde{U}_S \Sigma_S \tilde{U}_S^T$ and $\tilde{L}_2 = \tilde{U}_L \Sigma_L \tilde{V}_L^T$, we may factorize $\mathcal{A}$ as

$$\mathcal{A} = \begin{bmatrix} U_L & U_S \end{bmatrix} \begin{bmatrix} \Sigma_L & 0 \\ 0 & \Sigma_S \end{bmatrix} \begin{bmatrix} V_L^T \\ V_S^T \end{bmatrix},$$

where

$$U_L = \tilde{X}_2 \tilde{U}_L, \quad U_S = \tilde{X}_1 \tilde{U}_S, \quad V_L = \tilde{X}_2 \tilde{V}_L, \quad \text{and} \quad V_S = \tilde{X}_1 \tilde{U}_S. \tag{4.19}$$

When $\beta \ll \frac{1}{2}$, we find that the solution of (2.6) satisfies $\mathbf{f} = \mathcal{O}(1)$, $\lambda = \mathcal{O}(\beta)$ and, in general, $\mathbf{u} = \mathcal{O}(1)$. Using (4.7) and defining $g = [\mathbf{u}^T, \lambda^T]^T$, we obtain

$$\begin{bmatrix} \mathbf{f} \\ g \end{bmatrix} = \begin{bmatrix} \tilde{U}_S \psi_S \\ 0 \end{bmatrix} + 2v \begin{bmatrix} \Upsilon_{11} \\ \Upsilon_{12} \end{bmatrix} \tilde{U}_S \psi_S + \begin{bmatrix} 0 \\ \tilde{U}_L \psi_L \end{bmatrix} + 2v \begin{bmatrix} \Upsilon_{21} \\ \Upsilon_{22} \end{bmatrix} \tilde{V}_L \psi_L.$$

This implies that ψ_S is $\mathcal{O}(1)$ and $\psi_L = \rho_1 + \beta \rho_2$, where ρ_1 and ρ_2 are vectors independent of β.

Using (4.17), (4.15), (4.16) and the fact that $G = U^T \Delta \mathcal{A} V$, we find that

$$G = \begin{bmatrix} G_{L1} & G_{L2} \\ G_{S1} & G_{S2} \end{bmatrix} = \mathtt{u}\gamma_N \begin{bmatrix} \tilde{U}_L^T E_{22} \tilde{V}_L + 2v\hat{E}_{11} & \tilde{U}_L^T E_{21} \tilde{V}_L + 2v\hat{E}_{12} \\ \tilde{U}_S^T E_{21}^T \tilde{V}_L + 2v\hat{E}_{21} & \tilde{U}_S^T E_{11} \tilde{U}_S + 2v\hat{E}_{22} \end{bmatrix} + \mathtt{u}\gamma_N v^2 \bar{G},$$

where E_{11}, $\hat{E}_{11}$, $\hat{E}_{12}$, $\hat{E}_{21}$, $\hat{E}_{22}$ and $\bar{G}$ are $\mathcal{O}(1)$.

From (4.11) we have

$$\begin{aligned} \Delta \psi_L &= -\Sigma_L^{-1} \begin{bmatrix} G_{L1} & G_{L2} \end{bmatrix} \begin{bmatrix} \psi_L \\ \psi_S \end{bmatrix} \\ &= -\mathtt{u}\gamma_N \Sigma_L^{-1} \begin{bmatrix} \tilde{U}_L^T E_{22} \tilde{V}_L + 2v\hat{E}_{11} & \tilde{U}_L^T E_{21} \tilde{V}_L + 2v\hat{E}_{12} \end{bmatrix} \begin{bmatrix} \rho_1 + \beta \rho_2 \\ \psi_S \end{bmatrix}. \end{aligned}$$

From Corollary 4.6, we know that $\left\| \Sigma_L^{-1} \right\| = \mathcal{O}(1)$ and, therefore,

$$\begin{aligned} \|\Delta s_L\| &= \|\Delta \psi_L\| \\ &\leq \mathtt{u}\gamma_N \left\|\Sigma_L^{-1}\right\| \left\| \begin{bmatrix} \tilde{U}_L^T E_{11} \tilde{U}_L + \frac{D}{c-\beta D} \hat{E}_{11} & \hat{E}_{12} \end{bmatrix} \right\| \left(\left\| \begin{bmatrix} \rho_1 \\ \psi_S \end{bmatrix} + \beta \begin{bmatrix} \rho_2 \\ 0 \end{bmatrix} \right\| \right) \\ &= \mathtt{u}\gamma_N \left(c_2 + c_3 \beta \right). \end{aligned}$$

where c_2 and c_3 are constants independent of β. Hence, for $\beta \ll \frac{1}{2}$, we will expect the change in s_L relative to s to initially decrease as β decreases and then to become constant. Similarly, we can show that we can expect $\frac{\Delta s_S}{s}$ to be proportional $\mathtt{u}\gamma_N \beta^{-1}$ for $\frac{c_5 h^2}{2d} \leq \beta \ll \frac{1}{2}$ and to become (approximately) constant but several orders of magnitude larger than $\mathtt{u}$ when $\beta \leq \frac{c_5 h^2}{2d}$, where $c_5 \approx \frac{c^2 d^3}{C^2 (c^2 + d^2 + D^2)}$.

Now, as $\beta \longrightarrow 0$, the value of v tends towards a positive constant and, therefore, $\tilde{U}_L \psi_L \not\rightarrow g$. Hence, we can only conclude that $\|\Delta g\|^2 \leq \|\Delta s_L\|^2 + \|\Delta s_S\|^2$. For small β, $\|\Delta s_L\| / \|s\|$ is expected to initially decrease and then become constant but the dominance of $\|\Delta s_S\|$ implies that $\|\Delta \mathrm{u}\| / \|\mathrm{u}\|$ will initially increase and then become constant but several orders of magnitude larger than $\mathtt{u}$.

In Figure 4.2, we plot $\|\Delta s_L\| / \|s\|$,$\|\Delta s_S\| / \|s\|$, $\|\Delta \mathbf{f}\| / \|\mathbf{f}\|$ and $\|\Delta \mathbf{u}\| / \|\mathbf{u}\|$ against β for Example 3.1 with Target 1. The solution s is calculated by using the backslash command in Matlab, whilst $\tilde{s}$ is calculated by applying Matlab's `ldl` function to factor a single precision version of $\mathcal{A}$ and this factorization is then used to solve the system. We observe that, as expected, the change in s_L relative to s initially decreases as β decreases and then becomes constant; the change in s_S relative to s initially increases as β decreases and then becomes constant. However, the nice properties of s_L are not reflected in the change in either $\mathbf{f}$ or $\mathbf{u}$.

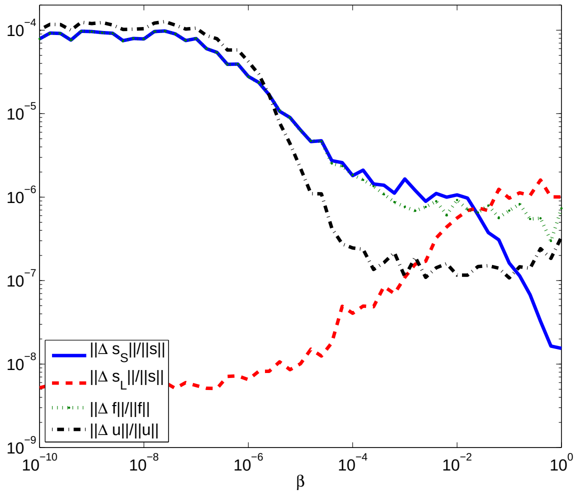

Figure 4.2: Plot of Δs_L and Δs_S with respect to β for Example 3.1 with Target 1. Results are shown for $h = \frac{1}{8}$.

4.6 Saddle-point formulation: target $\hat{u}$ defined over a subdomain of Ω

Let us consider the saddle-point system defined by (2.6) and the case where β is large. Following the methodology for the case when $\hat{u}$ is defined over all of the domain Ω, we write $\mathcal{A} = \mathcal{M} + \mathcal{E}$, where

$$\mathcal{M} = \left[\begin{array}{c|cc} 2\beta M & 0 & 0 \\ \hline 0 & \bar{M} & K^T \\ 0 & K & 0 \end{array}\right] \text{ and } \mathcal{E} = \left[\begin{array}{c|cc} 0 & 0 & -M \\ \hline 0 & 0 & 0 \\ -M & 0 & 0 \end{array}\right].$$

If

$$X_1 = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix} \text{ and } X_2 = \begin{bmatrix} 0 & 0 \\ I & 0 \\ 0 & I \end{bmatrix},$$

then $[X_1, X_2]$ is orthogonal, and both range(X_1) and range(X_2) are simple invariant subspaces of $\mathcal{M}$. From Theorem 4.9, we have

$$L_1 = 2\beta M, \quad L_2 = \begin{bmatrix} \bar{M} & K \\ K & 0 \end{bmatrix}.$$

Correspondingly, $E_{11} = 0$, $E_{22} = 0$, and $E_{12} = [0, -M]$.

The matrix L_2 is of saddle-point form (4.1), with $A = \bar{M}$ and $B = K$: Corollary 4.3 can be used to establish eigenvalue bounds for L_2. B is square and nonsingular so we can set $Y = I$. The eigenvalues of L_2 lie in $I^- \cup I^+$, where

$$I^- \; = \; \left[-Ch^{p-2}, \tfrac{1}{2}h^p\left(\bar{D} - \sqrt{\bar{D}^2 + 4c^2}\right)\right]$$

and

$$I^{+} = \left[ch^{p}, \tfrac{1}{2}h^{p-2}\left(\bar{D}h^{2}+\sqrt{\bar{D}^{2}h^{4}+4C^{2}}\right)\right].$$

Now, $d_1 := \operatorname{sep}(L_1, L_2) - \|E_{11}\| - \|E_{22}\| \geq 2\beta d h^{p} - \frac{1}{2}h^{p-2}\left(\bar{D}h^{2}+\sqrt{\bar{D}^{2}h^{4}+4C^{2}}\right) > 0$ for $\beta \gg \frac{1}{4d}h^{-2}\left(\bar{D}h^{2}+\sqrt{\bar{D}^{2}h^{4}+4C^{2}}\right)$, and

$$v := \|E_{12}\| / d_1 = \frac{2Dh^{2}}{4\beta d h^{2} - \left(\bar{D}h^{2}+\sqrt{\bar{D}^{2}h^{4}+4C^{2}}\right)}.$$

Hence, $v \to 0$ as $\beta \to +\infty$. From Theorem 4.9 and the above derivation, we deduce the following result.

If the regularization parameter $\beta \gg \frac{C}{d}h^{-2}$, then $\mathbf{f} = \mathcal{O}(\beta^{-1})$, $\mathbf{u} = \mathcal{O}(1)$ and $\lambda = \mathcal{O}(1)$, Section 6. Similarly to the case where the target $\hat{u}$ is defined over the whole of Ω, we can show that if $\beta \gg \frac{C}{d}h^{-2}$, we will expect the change in s_L relative to s to be at most inversely proportional to β and for the change in s_S relative to s to be roughly constant as β increases. As a result, we will expect $\|\Delta\mathbf{f}\| / \|\mathbf{f}\|$ and $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ to be $\mathcal{O}(\mathtt{u})$.

In Figure 4.3, we plot $\|\Delta s_L\| / \|s\|$,$\|\Delta s_S\| / \|s\|$, $\|\Delta\mathbf{f}\| / \|\mathbf{f}\|$ and $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ against β for Example 3.1 with Target 3 and $h = \frac{1}{8}$. The solution s is calculated by using the backslash command in Matlab, whilst $\tilde{s}$ is calculated by applying Matlab's `ldl` function to factor a single precision version of $\mathcal{A}$ and this factorization is then used to solve the system. For large β, we observe that, as expected, the change in s_L relative to s is inversely proportional to β but the change in s_S relative to s remains (approximately) constant. Also, both $\|\Delta\mathbf{f}\| / \|\mathbf{f}\|$ and $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ are (approximately) constant.

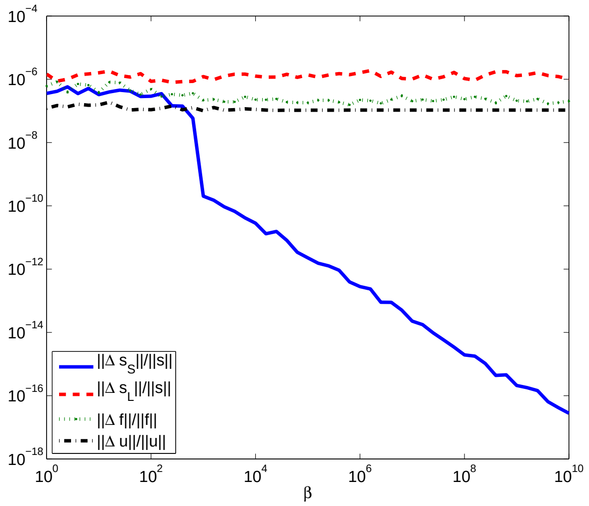

Figure 4.3: Plot of Δs_L and Δs_S with respect to β for Example 3.1 with Target 3. Results are shown for $h = \frac{1}{8}$.

For $\beta \ll \frac{\bar{D}}{2D}$, anticipating the nullspace method in Section 6, we find that $\mathbf{f}$ and $\mathbf{u}$ are $\mathcal{O}(1)$, whilst $\lambda = \mathcal{O}(\beta)$. In a similar manner to the case where $\hat{u}$ is defined over the whole of Ω, we can show that the change in s_L relative to s will initially decrease as β decreases and then remain (approximately) constant. However, we will expect the change in s_S relative to s to be inversely proportional to β as the regularization parameter decreases. As for the case when $\hat{u}$ is defined over the whole of Ω, this does not imply that $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ will initially decrease as β decreases and then remain (approximately) constant; in fact, we will expect $\|\Delta\mathbf{f}\| / \|\mathbf{f}\|$ and $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ to be inversely proportional to β for small enough β

In Figure 4.4, we plot $\|\Delta s_L\| / \|s\|$,$\|\Delta s_S\| / \|s\|$, $\|\Delta\mathbf{f}\| / \|\mathbf{f}\|$ and $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ against β for Example 3.1 with Target 3 and $h = \frac{1}{8}$. The solution s is calculated by using the backslash command in Matlab, whilst $\tilde{s}$ is calculated by applying Matlab's `ldl` function to factor a single precision version of $\mathcal{A}$ and this factorization is then used to solve the system. We observe that, as expected, the change in s_L relative to s initially decreases as β decreases and then becomes constant; the change in s_S relative to s increases as β decreases. However, $\|\Delta\mathbf{f}\| / \|\mathbf{f}\|$ and $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ are inversely proportional to β for small enough β.

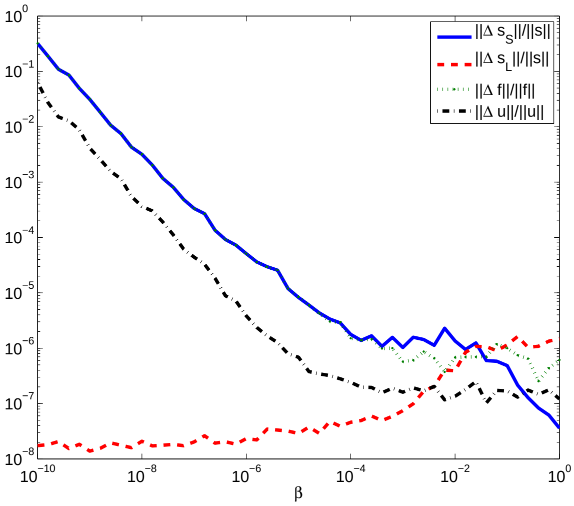

Figure 4.4: Plot of Δs_L and Δs_S with respect to β for Example 3.1 with Target 3. Results are shown for $h = \frac{1}{8}$.

5 Schur complement method

A common method for solving systems of the form (2.6) is to reduce it to a series of smaller systems that need to be solved. If A is nonsingular, then we can form the Schur complement factorization:

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

Thus, if

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

then

$$\begin{aligned} y &= -(BA^{-1}B^T)^{-1}\left(b_2 - BA^{-1}b_1\right), \\ x &= A^{-1}\left(b_1 - B^T y\right). \end{aligned}$$

In terms of (2.6), A is only non-singular if the target $\hat{u}$ is defined over all of Ω. Applying the Schur complement method to (2.6) we obtain

$$\begin{aligned} \lambda &= -S^{-1}\left(\mathbf{d} - KM^{-1}\mathbf{b}\right), \\ \mathbf{f} &= \tfrac{1}{2\beta}\lambda, \\ \mathbf{u} &= M^{-1}\left(\mathbf{b} - K\lambda\right), \end{aligned}$$

where

$$S = \tfrac{1}{2\beta}M + KM^{-1}K. \tag{5.1}$$

Thus, it is necessary to be able to carry out solves with M and the Schur complement S. Clearly, solves with M are independent of the regularization parameter β : there are good methods available for efficiently carrying out these solves [25].

Using the assumptions of Section 2.1, from Theorem 4.4 we obtain

$$\begin{aligned} \lambda_{\min}(S) &\geq \tfrac{1}{2\beta}\lambda_{\min}(M) + \lambda_{\min}\left(KM^{-1}K\right) \\ &= \tfrac{d}{2\beta}h^p + c_1 h^p, \\ \lambda_{\min}(S) &\leq \min\left(\tfrac{1}{2\beta}\lambda_{\min}(M) + \lambda_{\max}\left(KM^{-1}K\right), \tfrac{1}{2\beta}\lambda_{\max}(M) + \lambda_{\min}\left(KM^{-1}K\right)\right) \\ &= \min\left(\tfrac{d}{2\beta}h^p + C_1 h^{p-4}, \tfrac{D}{2\beta}h^p + c_1 h^p\right), \\ \lambda_{\max}(S) &\leq \tfrac{1}{2\beta}\lambda_{\max}(M) + \lambda_{\max}\left(KM^{-1}K\right) \\ &= \tfrac{D}{2\beta}h^p + C_1 h^{p-4}, \\ \lambda_{\max}(S) &\geq \max\left(\tfrac{1}{2\beta}\lambda_{\min}(M) + \lambda_{\max}\left(KM^{-1}K\right), \tfrac{1}{2\beta}\lambda_{\max}(M) + \lambda_{\min}\left(KM^{-1}K\right)\right) \\ &= \max\left(\tfrac{d}{2\beta}h^p + C_1 h^{p-4}, \tfrac{D}{2\beta}h^p + c_1 h^p\right), \end{aligned}$$

where $c_1 \approx \frac{c^2}{D}$ and $C_1 \approx \frac{C^2}{d}$ are constants independent of β and h. Hence,

$$\kappa(S) \leq \frac{D + 2\beta C_1 h^{-4}}{d + 2\beta c_1}$$

and if $\beta \geq \frac{D-d}{2(C_1 - c_1 h^4)}h^4$, then

$$\kappa(S) \geq \frac{d + 2\beta C_1 h^{-4}}{D + 2\beta c_1};$$

otherwise

$$\kappa(S) \geq \frac{D + 2\beta c_1}{d + 2\beta C_1 h^{-4}}.$$

For $\beta \gg \frac{d}{2c_1}$, we have

$$\kappa(S) \leq \frac{D + 2\beta C_1 h^{-4}}{d + 2\beta c_1} \approx \frac{C_1}{c_1}h^{-4}$$

and

$$\kappa(S) \geq \frac{d + 2\beta C_1 h^{-4}}{D + 2\beta c_1} \approx \frac{C_1}{c_1}h^{-4}.$$

Therefore, we will expect the condition number to be inversely proportional to h^4. For $\beta \ll \frac{D-d}{2(C_1-c_1h^4)}h^4 < \frac{D}{2C_1}h^4$, we find that

$$\kappa(S) \leq \frac{D+2\beta C_1 h^{-4}}{d+2\beta c_1} \approx \frac{D}{d}$$

and

$$\kappa(S) \geq \frac{D+2\beta c_1}{d+2\beta C_1 h^{-4}} \approx \frac{D}{d}.$$

Thus, the condition number will be bounded from above by a constant.

The upper bound for $\kappa(S)$ is monotonically increasing as β increases. If $\beta \geq \frac{D-d}{2(C_1-c_1h^4)}h^4$, then the lower bound is also monotonically increasing as β increases.

For intermediate values,

$$\kappa(S) \leq \frac{D+2\beta C_1 h^{-4}}{d+2\beta c_1} \leq C_2\beta h^{-4} + C_3,$$

where C_2 and C_3 are constants independent of β and h. We can also show that

$$\kappa(S) \geq \frac{d+2\beta C_1 h^{-4}}{D+2\beta c_1} \geq c_2\beta h^{-4} + c_3,$$

where c_2 and c_3 are constants independent of β and h. Thus, as β increases, it is reasonable to expect the condition number of S to increase at a rate proportional to β. For small mesh sizes, h, the condition number will be at most inversely proportional to h^4.

In Figure 5.1, we plot the condition number of $S = BA^{-1}B^T$ with respect to β for Example 3.3 with a target $\hat{u}$ defined over all of the domain Ω. Results are given for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$. We observe that, as expected, if $\beta \gg \frac{dD}{2c^2}$, then the condition number of Z^TAZ is inversely proportional to h^4 but (essentially) independent of the regularization parameter β. For $\beta \ll \frac{dDh^4}{2C^2}$, the condition number independent of both β and h : this is as we expected. Finally, for intermediate values, as β decreases, the condition number of S decreases at a rate proportional to β. Additionally, the condition number is inversely proportional to h^4.

Remark 5.1 *If the mesh size h remains fixed and $\beta \gg \frac{dD}{2c^2}$, then we will expect $\mathbf{u}$ and λ to be $\mathcal{O}(1)$, and $\mathbf{f}$ to be $\mathcal{O}(\beta^{-1})$. If $\beta \ll \frac{dD}{2c^2}$, we will expect λ to be $\mathcal{O}(\beta)$, whilst $\mathbf{f}$ and $\mathbf{u}$ will be $\mathcal{O}(1)$.*

6 Nullspace method and spectral properties

The nullspace method is another commonly used method that recasts the saddle-point system into systems of reduced order. Consider the solution of (4.1). Suppose that $B \in \mathbb{R}^{m\times n}$. Let the columns of $Z \in \mathbb{R}^{n\times(n-m)}$ span the nullspace of B and the columns of $Y \in \mathbb{R}^{n\times m}$ span the range space of B^T, then we can write $x = Yx_y + Zx_z$, where $x_y \in \mathbb{R}^m$ and $x_z \in \mathbb{R}^{n-m}$. Substituting this into (4.1) and premultiplying the resulting system by

$$\begin{bmatrix} Y^T & 0 \\ Z^T & 0 \\ 0 & I \end{bmatrix}$$

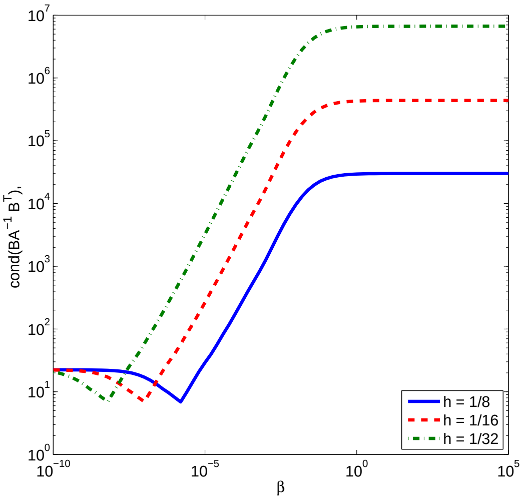

Figure 5.1: Condition number of $S = BA^{-1}B^T$ with respect to β for Example 3.3 with target $\hat{u}$ defined over the whole of Ω. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$.

we obtain

$$\begin{bmatrix} Y^TAY & Y^TAZ & Y^TB^T \\ Z^TAY & Z^TAZ & 0 \\ BY & 0 & 0 \end{bmatrix} \begin{bmatrix} x_y \\ x_z \\ y \end{bmatrix} = \begin{bmatrix} Y^Ts \\ Z^Ts \\ t \end{bmatrix}.$$

Thus, we can solve (4.1) by carrying out the following steps:

- Solve $BYx_y = b_2$;
- Solve $Z^TAZx_z = Z^T(b_1 - AYx_y)$;
- Set $x = Yx_y + Zx_z$;
- Solve $Y^TB^Ty = Y^T(b_1 - Ax)$.

There are many possibilities for the matrices Y and Z but we will focus on three standard choices for Z and two standard choices for Y. One possibility is to use the full QR factorization: let $QR = [Y\ Z]\left[R^T\ 0^T\right]^T$ be an orthogonal factorization of B^T, where $R \in \mathbb{R}^{m\times m}$ is upper triangular. For small choices of mesh size h, the QR factorization of B^T, where B is defined by (4.2), will be very expensive to form and, hence, we will not consider this method for forming Y and Z. Now, the simplest choice for Y is $Y = B^T$. Since

$$B = \begin{bmatrix} -M & K \end{bmatrix},$$

we find that solving $BYx_y = b_2$ is equivalent to solving

$$\left(M^2 + K^2\right) x_y = b_2.$$

From our analysis in Section 4, we know that there exist constants $\tilde{c} \geq c$ and $\tilde{C} \approx C$ such that $\lambda_{\min}(M^2 + K^2) = \tilde{c}^2h^{2p}$ and $\lambda_{\max}(M^2 + K^2) = \tilde{C}^2h^{2p-4}$. Thus, this system will become increasingly ill-conditioned as the mesh size h is refined. Alternatively, we can use a generalized form of the nullspace method:

- Find $\hat{x}$ such that $B\hat{x} = b_2$;
- Solve $Z^TAZx_z = Z^T(b_1 - A\hat{x})$;
- Set $x = \hat{x} + Zx_z$;
- Find y such that $B^Ty = b_1 - Ax$.

If we choose a symmetric matrix G such that Z^TGZ is nonsingular, then we can find $\hat{x}$ such that $B\hat{x} = d$ by solving

$$\underbrace{\begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix}}_{\mathcal{P}} \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} = \begin{bmatrix} 0 \\ b_2 \end{bmatrix}.$$

Similarly, once x has been calculated we can obtain y by solving

$$\begin{bmatrix} G & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \tilde{x} \\ y \end{bmatrix} = \begin{bmatrix} b_1 - Ax \\ 0 \end{bmatrix}.$$

Note that $\tilde{x} = 0$ when x is exact.

Considering our problem (2.6), if we set

$$G = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix},$$

then we obtain

$$\mathcal{P}_I = \begin{bmatrix} 0 & 0 & -M \\ 0 & I & K^T \\ -M & K & 0 \end{bmatrix}.$$

It is straightforward to see that solves with $\mathcal{P}_I$ will only require the solution of two systems with coefficient matrix M. As noted in Section 5, since M is a mass matrix, there are a number of efficient methods available to us to carry out these solves. We note that $\mathcal{P}$ is sometimes called a constraint preconditioner [15] for $\mathcal{A}$. If Z^TGZ is symmetric and positive definite, it may be used in combination with a projected preconditioned conjugate gradient method to solve saddle-point systems of the form (4.1) that have Z^TAZ symmetric and positive definite [9]. See [18] for a discussion on the use of constraint preconditioners for solving problems of the form (1.1)–(1.3).

Two further possibilities for defining Z are

$$Z = \begin{bmatrix} M^{-1}K \\ I \end{bmatrix} \quad \text{or} \quad Z = \begin{bmatrix} I \\ K^{-1}M \end{bmatrix}.$$

We will consider these two choices for Z and analyse how the condition number of Z^TAZ varies with the regularization parameter β. As a result, we hope to be able to choose the optimal value of Z for our problems.

Consider the case $Z = \left[KM^{-1},\ I\right]^T$. Given A as in (4.2), we obtain

$$Z^TAZ = \bar{M} + 2\beta KM^{-1}K.$$

If the target $\hat{u}$ is defined over all of the domain Ω, then $\bar{M} = M$ and $Z^T AZ = 2\beta S$, where S is the Schur complement defined by (5.1). Hence, the condition number of $Z^T AZ$ will be equal to the condition number of S.

Suppose that the target $\hat{u}$ is only defined over a sub-domain of Ω. Consider the Rayleigh quotient of $Z^T AZ$. Clearly, if $v^T \bar{M} v = 0$ and $\|v\| = 1$, then

$$v^T Z^T AZv = 2\beta v^T K M^{-1} K v,$$

and, assuming that Theorem 2.3 and Assumption 2.1 hold,

$$\frac{2\beta\bar{c}^2}{D} h^p \leq v^T Z^T AZv \leq \frac{2\beta\bar{C}^2}{d} h^{p-4}; \tag{6.1}$$

if $v^T \bar{M} v \neq 0$ and $\|v\| = 1$, then

$$\bar{d}h^p + \frac{2\beta c^2}{D} h^p \leq v^T Z^T AZv \leq \bar{D}h^p + \frac{2\beta C^2}{d} h^{p-4}. \tag{6.2}$$

Consider the case where β is small. Suppose that we compare the lower bounds in (6.1)–(6.2). If $\beta \ll \frac{\bar{d}dh^4}{2\bar{C}^2}$, then we can assume that the eigenvector $v_{\min}$ corresponding to $\lambda_{\min}(Z^T AZ)$ will be such that $v_{\min}^T \bar{M} v_{\min} = 0$ and, hence, we obtain $\lambda_{\min}(Z^T AZ) \geq \frac{2\beta\bar{c}^2}{D} h^p$. Clearly, $\lambda_{\max}(Z^T AZ) \leq \bar{D}h^p + \frac{2\beta C^2}{d} h^{p-4}$.

Consider the case where β is large. Comparing the lower bounds in (6.1)–(6.2), we observe that if $\beta \gg \frac{\bar{d}D}{2(\bar{c}^2 - c^2)}$, then $\lambda_{\min}(Z^T AZ) \geq \bar{d}h^p + \frac{2\beta c^2}{D} h^p$ and $\lambda_{\max}(Z^T AZ) \leq \bar{D}h^p + \frac{2\beta C^2}{d} h^{p-4}$. Hence, we obtain the following:

- if $\beta \ll \frac{d\bar{D}h^4}{2\bar{C}^2}$, then $\kappa(Z^T AZ) \leq \frac{D\left(d\bar{D}\beta^{-1} + 2C^2 h^{-4}\right)}{2\bar{c}^2 d} \approx \frac{D\bar{D}\beta^{-1}}{2\bar{c}^2}$;
- if $\beta \gg \frac{\bar{d}D}{2(\bar{c}^2 - c^2)}$, then $\kappa(Z^T AZ) \leq \frac{D\left(d\bar{D} + 2\beta C^2 h^{-4}\right)}{d\left(\bar{d}D + 2\beta c^2\right)}$.

If $\beta \gg \frac{\bar{d}D}{2(\bar{c}^2 - c^2)}$, then $\kappa(Z^T AZ)$ is monotonically increasing and, if $\beta \gg \frac{\bar{d}D}{2c^2}$, we have $\kappa(Z^T AZ) \approx \frac{C^2 D}{c^2 d} h^{-4}$.

In Figure 6.1, we plot the condition number of $Z^T AZ = \bar{M} + 2\beta K M^{-1} K$ with respect to β for Example 3.3 with Target 3. Results are given for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$. We observe that, as expected, if $\beta \gg \frac{\bar{d}D}{2c^2}$, then the condition number of $Z^T AZ$ is inversely proportional to h^4 but independent of the regularization parameter β; for $\beta \ll \frac{d\bar{D}h^4}{2\bar{C}^2}$, the condition number is independent of h but inversely proportional to β : this is as we expected. Finally, for intermediate values, as β increases, the condition number of $Z^T AZ$ increases.

Now consider the case $Z = \left[I,\ MK^{-1}\right]^T$. Given A as in (4.2), we obtain

$$Z^T AZ = 2\beta M + MK^{-1}\bar{M}K^{-1}M.$$

Let us firstly assume that the target $\hat{u}$ is defined over all of Ω, in which case, $\bar{M} = M$. If $\|v\| = 1$, then from Theorem 2.1

$$d\left(2\beta + \tfrac{d^2}{C^2} h^4\right) h^p \leq v^T Z^T AZv \leq Dh^p \left(2\beta + \tfrac{D^2}{c^2}\right),$$

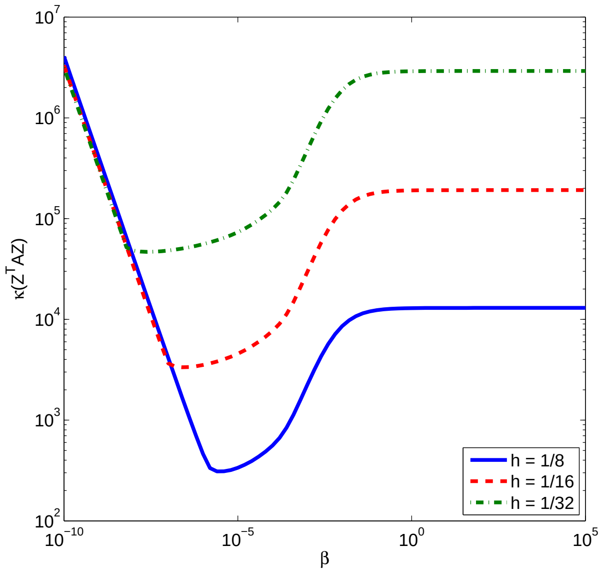

Figure 6.1: Condition number of $Z^T AZ = \bar{M} + 2\beta K M^{-1} K$ with respect to β for Example 3.3 with Target 2. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$.

and

$$\kappa(Z^T AZ) \leq \tfrac{C^2 D}{c^2 d} \left(\frac{2\beta c^2 + D^2}{2\beta C^2 + d^2 h^4} \right).$$

We note that this upper bound is monotonically decreasing as β increases and, hence, we will expect the condition number to improve as β increase. For $\beta \ll \frac{d^2}{2C^2} h^4$, the condition number of $Z^T AZ$ will be bounded above by a function that is approximately equal to $\frac{C^3 D^3}{c^2 d^3} h^{-4}$. If $\beta \gg \frac{D^2}{2c^2}$, then $\kappa(Z^T AZ)$ is bounded above by a function that is approximated by $\frac{D}{d}$.

In Figure 6.2, we plot the condition number of $Z^T AZ = 2\beta M + MK^{-1} MK^{-1} M$ with respect to β for Example 3.3 with a $\hat{u}$ defined over the whole of Ω. Results are given for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$. We observe that, as expected, if $\beta \gg \frac{D^3}{2c^2 d}$, then the condition number of $Z^T AZ$ is inversely proportional to h^4 but independent of the regularization parameter β; for $\beta \ll \frac{d^3 h^4}{2C^2 D}$, the condition number independent of h and β : this is as we expected. Finally, for intermediate values, as β decreases, the condition number of $Z^T AZ$ also decreases.

Now suppose that the target $\hat{u}$ is only defined on a sub-domain of Ω. If $\|v\| = 1$ and $\bar{M} K^{-1} M v \neq 0$, then

$$dh^p \left(2\beta + \tfrac{d\bar{d}}{C^2} h^4\right) \leq v^T Z^T AZ v \leq Dh^p \left(2\beta + \tfrac{D\bar{D}}{c^2}\right);$$

if $\|v\| = 1$ and $\bar{M} K^{-1} M v = 0$, then

$$2\beta dh^p \leq v^T Z^T AZ v \leq 2\beta D h^p.$$

Hence, $\lambda_{\min}(Z^T AZ) \geq 2\beta dh^p$, $\lambda_{\max}(Z^T AZ) \leq Dh^p \left(2\beta + \frac{D\bar{D}}{c^2}\right)$. Note that as β increases, this upper bound on $\kappa(Z^T AZ)$ will decrease and, hence, we will expect the condition number to decrease. If $\beta \gg \frac{D\bar{D}}{2c^2}$, then $\kappa(Z^T AZ) \lesssim \frac{D}{d}$; if $\beta \ll \frac{D\bar{D}}{2c^2}$, then $\kappa(Z^T AZ) \lesssim \frac{D^2 \bar{D}}{2c^2 d} \beta^{-1}$.

In Figure 6.3, we plot the condition number of $Z^T AZ = 2\beta M + MK^{-1} \bar{M} K^{-1} M$ with respect to β for Example 3.3 with Target 3. Results are given for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$. We observe

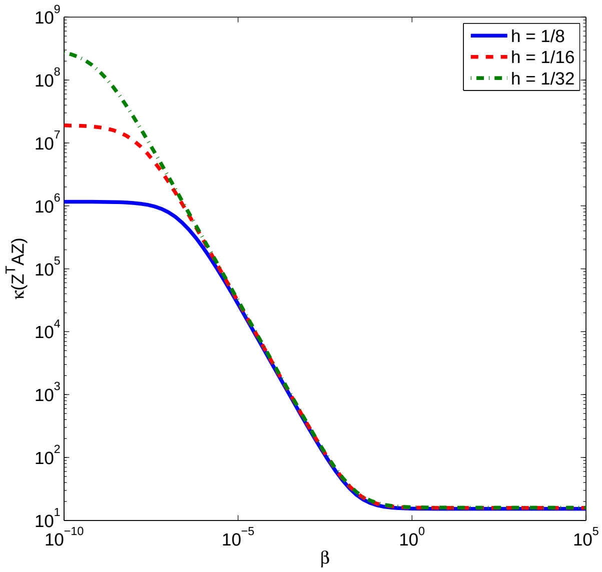

Figure 6.2: Condition number of $Z^TAZ = 2\beta M + MK^{-1}MK^{-1}M$ with respect to β for Example 3.3 with $\hat{u}$ defined over the whole of Ω. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$.

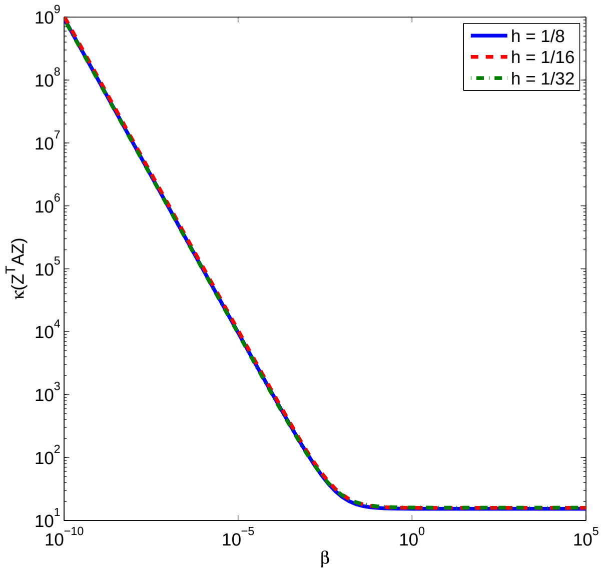

Figure 6.3: Condition number of $Z^TAZ = 2\beta M + MK^{-1}\bar{M}K^{-1}M$ with respect to β for Example 3.3 with Target 2. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$.

that, as expected, if $\beta \gg \frac{D\bar{D}}{2c^2}$, then the condition number of Z^TAZ is independent of the both the mesh size h and regularization parameter β; for $\beta \ll \frac{D\bar{D}}{2c^2}$, the condition number is independent of h but inversely proportional to β.

7 Conclusions

We have presented results about the spectral properties of the discretized systems that arise in distributed control problems: the PDE in the constraints is assumed to be the Poisson problem. The distributed control problems considered include a target $\hat{u}$. If $\hat{u}$ is defined over the whole of the domain, then we have shown that the condition number of the resulting saddle-point system will be bounded from above by a function that is independent of of the regularization parameter β but inversely proportional to h^6 for β smaller than $c_1 h^4$, where h is the mesh size and c_1 is a constant independent of h and β; if $c_1 h^4 \ll \beta < \frac{1}{2}$, then the condition number will be bounded from above by a function that is inversely proportional to β and h^2; if $\beta \gg c_2 h^{-2}$, where c_2 is a constant independent of h and β, the condition number is bounded from above by a function that is independent of h but proportional to β. Conversely, if $\hat{u}$ is only defined over a sub-domain of the overall problem, then the condition number is no longer bounded from above by a function that is independent of β when β is small: the upper bound is inversely proportional to β and h^2. In all of our numerical examples, we observed that the behaviour of the upper bound was well reflected in the calculated condition number. We were also able to show that if β is large and a backward-stable direct method is used to solve the saddle-point system, then the large condition number is not reflected in the relative error of $\mathbf{f}$ and $\mathbf{u}$: the relative error in these components is of order machine precision. However, this is not the case if β is small.

If the Schur complement method is used to solve the saddle-point system when $\hat{u}$ is defined over the whole of the domain, we were able to show that as $\beta \to 0$, the condition number of the Schur complement converges to the condition number of the mass matrix M. As $\beta \to +\infty$, the condition number of the Schur complement converges to $\kappa(M)\,(\kappa(K))^2$. Hence, refining the mesh will result in a larger condition number. We obtain more favourable condition numbers when the regularization parameter is small.

Alternatively, we could solve the saddle-point system by using a nullspace method. We have analyzed two different choices for the nullspace and were able to show that the spectral properties and, hence, the condition number, significantly altered when we changed which nullspace was used.

In practice, as the mesh is refined, the resulting linear systems will become too large for direct methods to be feasible and iterative methods will be required. The large condition numbers of the systems analyzed in this paper mean that popular iterative methods, for example, Krylov methods, may perform many iterations before reaching the desired level of accuracy [20, 24]. As a result, a preconditioner should be used such that the condition number of the preconditioned system is small. Only a handful of papers in the literature consider the saddle-point structure of the matrices when solving distributed control problems of the type considered in this paper, see, for example, [18, 22]. We hope that the analysis in this paper will be a building block for the derivation of preconditioners that will be effective for realistic values of the regularization parameter.

In this paper, we have concentrated on distributed control problems containing the Poisson problem. In many applications, this may be replaced by the Stokes or Navier-Stokes problem [3]. In these cases, the constraints will be degenerate but it is possible to deal with this degeneracy. Similar methods to those used in this paper can be applied to characterize the spectral properties of the resulting saddle-point systems, the Schur complement, and the reduced system from the nullspace method.

Acknowledgment

I would like to thank Tyrone Rees for providing me with numerical test examples that I was able to adapt for use within this paper. I would also like to thank Tyrone, Nick Gould and Andy Wathen for their helpful discussions and valuable suggestions during the process of this work.

References

[1] U. M. Ascher and E. Haber, *Grid refinement and scaling for distributed parameter estimation problems*, Inverse Problems, 17 (2001), pp. 571–590.

[2] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numerica, 14 (2005), pp. 1–137.

[3] G. Biros and O. Ghattas, *Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization. I. The Krylov-Schur solver*, SIAM J. Sci. Comput., 27 (2005), pp. 687–713 (electronic).

[4] J. R. Bunch, J. W. Demmel, and C. F. Van Loan, *The strong stability of algorithms for solving symmetric linear systems*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 494–499.

[5] S. H. Cheng and N. J. Higham, *A modified Cholesky algorithm based on a symmetric indefinite factorization*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 1097–1110 (electronic).

[6] S. S. Collis and M. Heinkenschloss, *Analysis of the streamline upwind/Petrov Galerkin method applied to the solution of optimal control problems*, Tech. Rep. TR02–01, Department of Computational and Applied Mathematics, Rice University, 2002.

[7] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford University Press, Oxford, 2005.

[8] P. E. Gill, W. Murray, and M. H. Wright, *Practical optimization*, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1981.

[9] N. I. M. Gould, M. E. Hribar, and J. Nocedal, *On the solution of equality constrained quadratic programming problems arising in optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 1376–1395.

[10] N. I. M. Gould and V. Simoncini, *Spectral analysis of saddle point matrices with indefinite leading blocks*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 1152–1171.

[11] E. Haber and U. M. Ascher, *Preconditioned all-at-once methods for large, sparse parameter estimation problems*, Inverse Problems, 17 (2001), pp. 1847–1864.

[12] N. J. Higham, *Analysis of the Cholesky decomposition of a semi-definite matrix*, in Reliable numerical computation, Oxford Sci. Publ., Oxford Univ. Press, New York, 1990, pp. 161–185.

[13] ——, *Accuracy and stability of numerical algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2002.

[14] K. Ito and K. Kunisch, *Augmented Lagrangian-SQP methods for nonlinear optimal control problems of tracking type*, SIAM J. Control Optim., 34 (1996), pp. 874–891.

[15] C. Keller, N. I. M. Gould, and A. J. Wathen, *Constraint preconditioning for indefinite linear systems*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1300–1317.

[16] H. Maurer and H. D. Mittelmann, *Optimization techniques for solving elliptic control problems with control and state constraints. I. Boundary control*, Comput. Optim. Appl., 16 (2000), pp. 29–55.

[17] B. N. Parlett, *The symmetric eigenvalue problem*, vol. 20 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[18] T. Rees, H. S. Dollar, and A. J. Wathen, *Optimal solvers for PDE-constrained optimization*, Tech. Rep. RAL-TR-2008-018, Rutherford Appleton Laboratory, 2008.

[19] T. Rusten and R. Winther, *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904.

[20] Y. Saad, *Iterative methods for sparse linear systems*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2003.

[21] R. B. Schnabel and E. Eskow, *A new modified Cholesky factorization*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 1136–1158.

[22] J. Schöberl and W. Zulehner, *Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems*, Tech. Rep. 2006-19, Johannes Kepler University, 2006.

[23] V. Simoncini, *Spectral properties of saddle point linear systems and relations to iterative solvers I.* London Mathematical Society Durham Symposium on Computational Linear Algebra for Partial Differential Equations, July 2008.

[24] H. A. van der Vorst, *Iterative Krylov Methods for Large Linear Systems*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, United Kingdom, 1 ed., 2003.

[25] A. J. Wathen and T. Rees, *Chebyshev semi-iteration in preconditioning*, Tech. Rep. NA-08/14, Oxford University Computing Laboratory, 2008.

[26] J. H. Wilkinson, *The algebraic eigenvalue problem*, Monographs on Numerical Analysis, The Clarendon Press Oxford University Press, New York, 1988. Oxford Science Publications.

[27] M. H. Wright, *Ill-conditioning and computational error in interior methods for nonlinear programming*, SIAM J. Optim., 9 (1999), pp. 84–111.