Science & Technology Facilities Council

Technical Report

Properties of linear systems in PDE-constrained optimization. Part II: Neumann boundary control

H S Thorne

April 22, 2010

RAL-TR-2009-018

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

ISSN 1358-6254

RAL-TR-2009-018

Properties of linear systems in PDE-constrained optimization. Part II: Neumann boundary 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 Neumann boundary 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, boundary 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 of the saddle-point system. 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 Neumann boundary 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 $\Omega \subset \mathbb{R}^2$ or $\mathbb{R}^3$ (in this case, the Poisson equation):

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

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

$$\frac{\delta u}{\delta n} = g \ \text{ on } \ \partial\Omega. \tag{1.3}$$

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 boundary conditions, g, (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 boundary conditions of the PDE, g, rapidly vary along the boundary $\partial\Omega$. 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 of approaches: 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 mesh size h and 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) and the Schur complement (Section 5). In Section 4.4, we will also show that if $\hat{u}$ is defined over the whole of the domain Ω, 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. Finally, we draw our conclusions in Section 6.

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 discretise our problems with finite elements. Using the weak formulation of (1.2)-(1.3), we wish to find $u \in H^1(\Omega)$ and $g \in L_2(\partial\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x - \int_{\partial\Omega} gv \,\mathrm{d}s = \int_\Omega fv \,\mathrm{d}x \qquad \forall v \in H^1(\Omega).$$

We discretize the problem using the finite element method, using a triangulation where the total number of vertices is m_T and the number of vertices on the boundary is m_B. Assume that V^h is a m_T dimensional vector space of test functions with basis $\{\phi_1, \ldots, \phi_{m_T}\}$. If $u_h \in V^h \subset H^1(\Omega)$, it is uniquely determined by $\mathbf{u} = (U_1 \ldots U_{m_T})^T$ in

$$u_h = \sum_{j=1}^{m_T} U_j \phi_j.$$

Here the ϕ_i, $i = 1, \ldots, m_T$, define a set of shape functions: we shall assume that ϕ_i, $i = 1, \ldots, m_I$ are in the interior of the domain, the remaining ϕ_i are on the boundary. We also assume that this approximation is conforming, i.e. $V^h = \mathrm{span}\{\phi_1, \ldots, \phi_{m_T}\} \subset H^1(\Omega)$.

Similarly, let W^h be a m_B dimensional vector space of test functions with basis $\{\hat{\phi}_1, \ldots, \hat{\phi}_{m_B}\}$ g_h such that $W^h \subset L_2(\partial\Omega)$. Hence, $g_h \in W^h$ is uniquely determined by $\mathbf{g} = (G_1 \ldots G_{m_B})^T$ in

$$g_h = \sum_{j=1}^{m_B} G_j \hat{\phi}_j$$

and we can write the discrete analogue of the minimization problem as

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

$$\text{such that} \qquad \int_\Omega \nabla u_h \cdot \nabla v_h \,\mathrm{d}x - \int_{\partial\Omega} u_h g_h \,\mathrm{d}s = \int_\Omega f v_h \,\mathrm{d}x \qquad \forall v_h \in V^h. \tag{2.1}$$

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

$$\min_{u_h, f_h} \frac{1}{2} \|u_h - \hat{u}\|_2^2 + \beta \|g_h\|_2^2 = \min_{\mathbf{u},\mathbf{f}} \frac{1}{2} \mathbf{u}^T \bar{M} \mathbf{u} - \mathbf{u}^T \mathbf{b} + \alpha + \beta \mathbf{g}^T M_g \mathbf{g}, \tag{2.2}$$

where $\mathbf{u} = (U_1, \ldots, U_{m_T})^T$, $\mathbf{g} = (G_1, \ldots, G_{m_B})^T$, $\mathbf{b} = \{\int_\Omega \hat{u}\phi_i\}_{i=1\ldots m_T}$, $\alpha = \|\hat{u}\|_2^2$, $M = \{\int_\Omega \phi_i \phi_j\}_{i,j=1\ldots m_T}$ is a mass matrix, $\bar{M} = M$, and $M_g = \{\int_{\partial\Omega} \hat{\phi}_i \hat{\phi}_j\}_{i,j=1\ldots m_B}$ is the Neumann matrix. 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 (2.2) but α, b and $\bar{M}$ are defined by

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

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

We now turn our attention to the constraint: (2.1) is equivalent to finding $\mathbf{u}$ such that

$$\int_\Omega \nabla \left(\sum_{i=1}^{m_T} U_i \phi_i \right) \cdot \nabla \phi_j \, \mathrm{d}x - \int_{\partial\Omega} \left(\sum_{i=1}^{m_B} G_i \hat{\phi}_i \right) \phi_j \, \mathrm{d}s = \int_\Omega f \phi_j \, \mathrm{d}x, \qquad j = 1, \ldots, m_T,$$

which is

$$K\mathbf{u} - E\mathbf{g} = \mathbf{d},$$

where the matrix $K = \{\int_\Omega \nabla\phi_i \cdot \nabla\phi_j\}_{i,j=1\ldots m_T} \in \mathbb{R}^{m_T \times m_T}$ is the discrete Laplacian (the stiffness matrix), $E = \{\int_{\partial\Omega} \widehat{\phi_j} \ \phi_i\}_{i=1\ldots m_T, j=1\ldots m_B} \in \mathbb{R}^{m_T \times m_B}$ and $\mathbf{d} = \{\int_\Omega f\phi_i\}_{i=1\ldots m_T} \in \mathbb{R}^{m_T}$. We will assume that $\phi_{i+m_I} = \hat{\phi}_i$, $i = 1, \ldots, m_B$. Hence, $E = [0, M_g]^T$.

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

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

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

$$\begin{bmatrix} 2\beta M_g & 0 & -E^T \\ 0 & \bar{M} & K^T \\ -E & K & 0 \end{bmatrix} \begin{bmatrix} \mathbf{g} \\ \mathbf{u} \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{b} \\ \mathbf{d} \end{bmatrix}. \tag{2.3}$$

We will discuss the properties of this system in Section 4.

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

Using the fact that we have assumed ϕ_i, $i = 1, \ldots, m_I$ are in the interior of the domain, and the remaining ϕ_i are on the boundary, we can partition K and M as

$$K = \begin{bmatrix} K_{II} & K_{BI}^T \\ K_{BI} & B_{BB} \end{bmatrix}, \quad \text{and} \quad M = \begin{bmatrix} M_{II} & M_{BI}^T \\ M_{BI} & M_{BB} \end{bmatrix}, \tag{2.4}$$

where $K_{II} \in \mathbb{R}_{m_I \times m_I}$, $K_{BB} \in \mathbb{R}^{m_B \times m_B}$, $M_{II} \in \mathbb{R}_{m_I \times m_I}$ and $M_{BB} \in \mathbb{R}^{m_B \times m_B}$. In addition, we note that, with this partitioning,

$$E = \begin{bmatrix} 0 \\ M_g \end{bmatrix}. \tag{2.5}$$

Throughout this paper, we will assume that a shape regular, quasi-uniform division of the domain is used [8] with $\mathbf{P_m}$ or $\mathbf{Q_m}$ ($\mathbf{m} \geq 1$) finite element approximations. Using these assumptions, we have the following theorem [8]:

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

$$\begin{array}{lll} \lambda_{\min}(K) = 0, & \lambda^{+}_{\min}(K) = ch^p, & \lambda_{\max}(K) = Ch^{p-2}, \\ \lambda_{\min}(K_{II}) = c_{II}h^p, & \lambda_{\max}(K) = C_{II}h^{p-2}, & \\ \lambda_{\min}(M) = dh^p, & \lambda_{\max}(M) = Dh^p, & \\ \lambda_{\min}(M_g) = d_g h^{p-1}, & \lambda_{\max}(M_g) = D_g h^{p-1}, & \end{array}$$

where c, c_{II}, d, d_g, C, C_{II}, D and D_g are constants independent of the mesh size h but dependent on p.

In addition, we have the following less well-known result.

Theorem 2.2 Consider the p-dimensional problem with $p \in \{2, 3\}$. There exist constants c_{BB} and C_{BB} independent of the mesh size h such that

$$\lambda_{\min}(K_{BB}) = c_{BB}h^{p-2} \quad \text{and} \quad \lambda_{\max}(K_{BB}) = C_{BB}h^{p-2}.$$

Proof. From [9, Theorem 1], if $\{K_E\}$ is the set of individual element matrices that comprise K_{BB}, then

$$\max_E \left(\lambda_{\max}\left(K_E\right)\right) \leq \lambda_{\max}\left(K_{BB}\right) \leq t \max_E \left(\lambda_{\max}(K_E)\right)$$

and

$$\min_E \left(\lambda_{\min}\left(K_E\right)\right) \leq \lambda_{\min}\left(K_{BB}\right),$$

where t is a positive constant independent of the mesh size h.

For p-dimensional problems with $p \in \{2, 3\}$, each element matrix K_E takes the form $K_E = h^{p-2}\tilde{K}_E$, where the eigenvalues of $\tilde{K}_E$ are independent of the mesh size h and nonzero. Thus,

$$h^{p-2} \max_E \left(\lambda_{\max}\left(\tilde{K}_E\right)\right) \leq \lambda_{\max}\left(K_{BB}\right) \leq th^{p-2} \max_E \left(\lambda_{\max}(\tilde{K}_E)\right)$$

and

$$h^{p-2} \min_E \left(\lambda_{\min}\left(\tilde{K}_E\right)\right) \leq \lambda_{\min}\left(K_{BB}\right).$$

This completes the proof. □

In the following, we have used Cauchy's interlacing theorem [16]

Theorem 2.3 Suppose $T \in \mathbb{R}^{n\times n}$ is symmetric and

$$T = \begin{bmatrix} H & \star \\ \star & \star \end{bmatrix},$$

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

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

and the eigenvalue perturbation theorem [24, pp. 101-2]

Theorem 2.4 If $\mathcal{M}$ and $\mathcal{M}+\mathcal{E} \in \mathbb{R}^{N\times N}$ are symmetric matrices, then

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

to bound the eigenvalues of M_{BB} and M_{II}, and the singular values of $\begin{bmatrix} E & K \end{bmatrix}$.

Theorem 2.5 Consider the p-dimensional problem with $p \in \{2, 3\}$. Let E be defined by (2.5). Now

$$\begin{array}{ll} \lambda_{\min}(M_{BB}) = d_{BB}h^p, & \lambda_{\max}(M_{BB}) = D_{BB}h^p, \\ \lambda_{\min}(M_{II}) = d_{II}h^p, & \lambda_{\max}(M_{II}) = D_{II}h^p, \\ \sigma_{\min}(\begin{bmatrix} E & K \end{bmatrix}) = c_B h^{p-\alpha}, & \sigma_{\max}(\begin{bmatrix} E & K \end{bmatrix}) = C_B h^{p-2}, \end{array}$$

where α, d_{BB}, d_{II}, D_{BB} and D_{II} are constants independent of the mesh size h but dependent on p, and $0 \leq \alpha \leq 1$.

Let $S = K_{BB} - K_{BI}K_{II}^{-1}K_{BI}^T$. Since K is singular, the Schur complement S must be singular. Additionally, define

$$\tilde{Z}_1 = \begin{bmatrix} -K_{II}^{-1}K_{BI}^T \\ I \end{bmatrix} \quad \text{and} \quad \tilde{Z} = \begin{bmatrix} M_q^{-1}S \\ Z_1 \end{bmatrix}. \tag{2.6}$$

We will use the following assumptions that we have numerically justified.

Assumption 2.1 Consider the p-dimensional problem with $p \in \{2, 3\}$. Let $\tilde{Z}$ and $\tilde{Z}_1$ be defined by (2.6),

$$\hat{M}_g = \begin{bmatrix} 0 & 0 \\ 0 & M_g \end{bmatrix} \in \mathbb{R}^{m_T \times m_T}, \quad \text{and} \quad S = K_{BB} - K_{BI} K_{II}^{-1} K_{BI}^T.$$

We will assume that $\sigma_{\min}(M_{21}) = \sigma_{\min}(K_{21}) = 0$ and

$$\begin{aligned}
\sigma_{\min^+}(K_{BI}) &= c_{BI} h^{p-1}, & \sigma_{\max}(K_{BI}) &= C_{BI} h^{p-2}, \\
\sigma_{\min^+}(K_{BI} K_{II}^{-1}) &= \hat{c}_{BI} h, & \sigma_{\max}(K_{BI} K_{II}^{-1}) &= \hat{C}_{BI} h^{-\frac{1}{2}}, \\
\sigma_{\min}(\begin{bmatrix} E & K \end{bmatrix}) &= c_B h^{p-\alpha}, & \sigma_{\max}(\begin{bmatrix} E & K \end{bmatrix}) &= C_B h^{p-2}, \\
\sigma_{\min^+}(M_{BI}) &= d_{BI} h^{p-1}, & \sigma_{\max}(M_{BI}) &= D_{BI} h^{p-1}, \\
\lambda_{\min^+}(S) &= c_S h^{p-1}, & \lambda_{\max}(S) &= C_S h^{p-2}, \\
\lambda_{\min^+}(S M_g^{-1} S, \tilde{Z}^T \tilde{Z}) &= c_{S1} h^{p+1}, & \lambda_{\max}(S M_g^{-1} S, \tilde{Z}^T \tilde{Z}) &= C_{S1} h^{p-1}, \\
\lambda_{\min}(\tilde{Z}_1^T M \tilde{Z}_1, \tilde{Z}^T \tilde{Z}) &= c_Z h^{p+2}, & \lambda_{\max}(\tilde{Z}_1^T M \tilde{Z}_1, \tilde{Z}^T \tilde{Z}) &= C_Z h^p, \\
\lambda_{\min^+}(KMK, \hat{M}_g^2 + K^2) &= c_{K1} h^{p+2}, & \lambda_{\max}(KMK, \hat{M}_g^2 + K^2) &= C_{K1} h^p, \\
\lambda_{\min^+}(\hat{M}_g^3, \hat{M}_g^2 + K^2) &= d_{G1} h^{p+1}, & \lambda_{\max}(\hat{M}_g^3, \hat{M}_g^2 + K^2) &= D_{G1} h^{p-1}, \\
\hat{M}_g v_{\min^+}(KMK, \hat{M}_g^2 + K^2) &\neq 0, & K v_{\min^+}(\hat{M}_g^3, \hat{M}_g^2 + K^2) &\neq 0,
\end{aligned}$$

where c_{BI}, $\hat{c}_{BI}$, c_S, d_{BI}, C_{BI}, $\hat{C}_{BI}$, C_S and D_{BI} are constants independent of the mesh size h but dependent on p.

If the target $\hat{u}$ is defined over the whole of the domain, then $\bar{M} = M$. Suppose that the target $\hat{u}$ is only defined over a sub-domain of $\hat{\Omega} \subset \Omega$. If we partition $\bar{M}$ as we partition M in (2.4), then

$$\bar{M} = \begin{bmatrix} \bar{M}_{II} & \bar{M}_{BI}^T \\ \bar{M}_{BI} & \bar{M}_{BB} \end{bmatrix}, \tag{2.7}$$

where $\bar{M}_{II} \in \mathbb{R}_{m_I \times m_I}$ and $\bar{M}_{BB} \in \mathbb{R}^{m_B \times m_B}$. Noting that we can permute the rows and columns of $\bar{M}_{II}$ and M_{II} such that the assumptions of Theorem 2.3 hold (and similarly for $\bar{M}_{BB}$ and M_{BB}), combining Theorem 2.3 and Theorem 2.5 we obtain

Theorem 2.6 Consider the p-dimensional problem with $p \in \{2, 3\}$ and $\hat{\Omega} \subset \Omega$. Let Theorem 2.5 hold. Then

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

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

In addition, we will use the following assumption.

Assumption 2.2 Consider the p-dimensional problem with $p \in \{2,3\}$ and $\hat{\Omega} \subset \Omega$. Let

$$\hat{M}_g = \begin{bmatrix} 0 & 0 \\ 0 & M_g \end{bmatrix}.$$

We will assume that Assumption 2.1 holds, $\sigma_{\min}(\bar{M}_{21}) = 0$ and

$$\begin{aligned} \sigma_{\min^+}(\bar{M}_{BI}) &= \bar{d}_{BI} h^{p-1}, & \sigma_{\max}(\bar{M}_{BI}) &= \bar{D}_{BI} h^{p-1}, \\ \lambda_{\min}(\tilde{Z}_1^T \bar{M} \tilde{Z}_1, \tilde{Z}^T \tilde{Z}) &= \bar{c}_Z h^{p+2}, & \lambda_{\max}(\tilde{Z}_1^T \bar{M} \tilde{Z}_1, \tilde{Z}^T \tilde{Z}) &= \bar{C}_Z h^p, \\ \lambda_{\min^+}(K \bar{M} K, \hat{M}_g^2 + K^2) &= \bar{c}_{K1} h^{p+2}, & \lambda_{\max}(K \bar{M} K, \hat{M}_g^2 + K^2) &= \bar{C}_{K1} h^p, \end{aligned}$$

where $\bar{c}_{K1}$, $\bar{C}_{K1}$, $\bar{c}_Z$, $\bar{C}_Z$, $\bar{d}_{BI} \geq d_{BI}$ and $\bar{D}_{BI} \leq D_{BI}$ are constants independent of the mesh size h but dependent on p.

Throughout the rest of this paper, we shall assume that the matrices K, M and $\bar{M}$ are partitioned as in (2.4) and (2.7).

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 the control will rapidly vary from one extreme to another along the boundary [15] and would often be difficult to impose in real life. 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 may be used. In practice, a tolerance is 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. To use such a strategy, the coarse mesh must already be refined enough such that further refinements have minimal effect on the value of β, see [1, 7].

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 consider problems of the form (1.1)–(1.3) and discretise the problem using bilinear quadrilateral $\mathbf{Q_1}$ finite elements.

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. We will consider both continuous and discontinuous targets and 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 $\hat{\Omega}_2$.

4 Properties of the saddle-point matrices

We observe that the coefficient matrix in (2.3) can be written in the form

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}, \tag{4.1}$$

	$\hat{\Omega}$	$\hat{\Omega}_1$	$\hat{\Omega}_2$	$\hat{u}(x,y)\|_{\hat{\Omega}_1}$	$\hat{u}(x,y)\|_{\hat{\Omega}_2}$
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	$\hat{\Omega}_1 \cup \hat{\Omega}_2$	$\left\{(x,y) : x^2+y^2 \leq \frac{1}{4}\right\}$	$\Omega/\hat{\Omega}_1$	2	0
Target 3	$\hat{\Omega}_1$	$\left\{(x,y) : x^2+y^2 \leq \frac{1}{4}\right\}$	—	2	—
Target 4	$\hat{\Omega}_1$	$\delta\Omega$	—	$e^{x^2+y^{0.5}}$	—

Table 3.1: Target functions for 2D problems

	$\hat{\Omega}$	$\hat{\Omega}_1$	$\hat{\Omega}_2$	$\hat{u}(x,y)\|_{\hat{\Omega}_1}$	$\hat{u}(x,y)\|_{\hat{\Omega}_2}$
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	$\hat{\Omega}_1 \cup \hat{\Omega}_2$	$\left\{(x,y,z) : x^2+y^2+z^2 \leq \frac{1}{4}\right\}$	$\Omega/\hat{\Omega}_1$	2	0
Target 3	$\hat{\Omega}_1$	$\left\{(x,y,z) : x^2+y^2+z^2 \leq \frac{1}{4}\right\}$	—	2	—
Target 4	$\hat{\Omega}_1$	$\delta\Omega$	—	$e^{x^2+y^{0.5}+z^4}$	—

Table 3.2: Target functions for 3D problems

where

$$A = \begin{bmatrix} 2\beta M_g & 0 \\ 0 & \bar{M} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -E & K \end{bmatrix}.$$

Systems of the general form given in (4.1) are known as saddle-point matrices [2]. We will consider the case where the target $\hat{u}$ is defined over the whole domain Ω separately to the case where it is only defined on subdomains.

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 [18] 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, then we obtain the following result from [7, Corollary 4.3].

Corollary 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

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

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

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

$l_1 < \lambda_{\min}(Z_B^T A Z_B)$ 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

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

In particular,

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

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

4.2 Neumann boundary control problems with target $\hat{u}$ defined over whole domain Ω

Suppose that the target $\hat{u}$ is defined over the whole of the domain Ω. Observe that the matrix A is symmetric and positive definite whilst B has full rank, where A and B are defined by (4.1). We will consider the cases $\beta > \frac{h}{2}\max\{\frac{d}{d_g}, \frac{D}{D_g}\}$ and $\beta < \frac{h}{2}\min\{\frac{d}{d_g}, \frac{D}{D_g}\}$ separately. Initially, we will assume that $\beta > \frac{h}{2}\max\{\frac{d}{d_g}, \frac{D}{D_g}\}$.

Since A is symmetric and positive definite, we will apply Theorem 4.1 to bound the eigenvalues of $\mathcal{A}$. Applying Theorem 2.1 and Assumption 2.1 we obtain $\lambda_{\min}(A) = dh^p$, $\lambda_{\max}(A) = 2\beta D_g h^{p-1}$,

$\sigma_{\min}(B) = c_B h^{p-\alpha}$ and $\sigma_{\max}(B) = C_B h^{p-2}$, where $0 \leq \alpha \leq 1$ is a constant independent of the mesh size h. Substituting these into Theorem 4.1 and using the Taylor series expansion $\sqrt{1+x} = 1 + \frac{x}{2} - \frac{x^2}{8} + \mathcal{O}(x^3)$ for $0 \leq x < 1$, find that there exists a constant u_1 independent of the mesh size h and regularization parameter β such that $\lambda(\mathcal{A}) \in I^- \cup I^+$, where

$$\begin{aligned} I^- &= \left[\tfrac{1}{2}h^{p-2}\left(dh^2 - \sqrt{d^2h^4 + 4C_B^2}\right), \beta h^{p-1} D_g \left(1 - \sqrt{1 + \frac{c_B^2}{\beta^2 D_g^2} h^{2(1-\alpha)}}\right)\right] \\ &\subset \left[\tfrac{1}{2}h^{p-2}\left(dh^2 - \sqrt{d^2h^4 + 4C_B^2}\right), -\frac{u_1}{\beta} h^{p+1-2\alpha}\right], \\ I^+ &= \left[dh^p, h^{p-2}\left(\beta D_g h + \sqrt{\beta^2 D_g^2 h^2 + C_B^2}\right)\right]. \end{aligned} \tag{4.2}$$

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

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

Note that $\mathcal{M}$ is singular. Using Theorem 4.1, the positive eigenvalues of $\mathcal{M}$ lie in

$$\left[2\beta d_g h^{p-1}, 2\beta D_g h^{p-1}\right] \cup \left[dh^p, \tfrac{1}{2}h^{p-2}\left(Dh^2 + \sqrt{D^2h^4 + 4C^2}\right)\right].$$

Applying Theorem 2.4 to these bounds and incorporating the bounds given in (4.2), we obtain the following result.

Corollary 4.3 Consider the p-dimensional problem with $p \in \{2, 3\}$. Let

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

assume that Assumption 2.1 holds and $\beta > \frac{h}{2}\max\{\frac{d}{d_g}, \frac{D}{D_g}\}$. Let $\lambda(\mathcal{A})$ denote the spectrum of $\mathcal{A}$. There exists a constant u_1 independent of the mesh size h such that

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

where

$$\begin{aligned} I^- &= \left[\tfrac{1}{2}h^{p-2}\left(dh^2 - \sqrt{d^2h^4 + 4C_B^2}\right), -\frac{u_1}{\beta} h^{p+1-2\alpha}\right] \\ &\subset \left[\tfrac{1}{2}h^{p-2}\left(dh^2 - \sqrt{d^2h^4 + 4C_B}\right), -\frac{u_1}{\beta} h^{p+1-2\alpha}\right], \\ I_1^+ &= \left[dh^p, \tfrac{1}{2}h^{p-2}\left(Dh^2 + 2D_g h + \sqrt{D^2h^4 + 4C^2}\right)\right], \\ I_2^+ &= \left[\left(2\beta d_g - D_g\right)h^{p-1}, h^{p-2}\left(\beta D_g h + \sqrt{\beta^2 D_g^2 h^2 + C_B^2}\right)\right]. \end{aligned}$$

Consequently, if $\beta \gg \frac{h}{2}\max\{\frac{d}{d_g}, \frac{D}{D_g}\}$, then, for the p-dimensional problem with $p \in \{2, 3\}$, there exist constants u_1 and U_1 independent of h such that

$$\begin{aligned} \sigma_{\min}(\mathcal{A}) &\geq \frac{u_1}{\beta} h^{p+1-2\alpha}, \\ \sigma_{\max}(\mathcal{A}) &\leq \beta U_1 h^{p-1}, \end{aligned}$$

giving, $\kappa(\mathcal{A}) = \mathcal{O}(\beta^2 h^{2(\alpha-1)})$. Additionally, if $2\beta d_g h \gg C$, then the intervals I_1^+ and I_2^+ will be well separated. Each set will contain n_T eigenvalues.

Consider the case $\beta < \frac{h}{2}\min\{\frac{d}{d_g}, \frac{D}{D_g}\}$. As with the previous case, we may apply Theorem 4.1 to bound the eigenvalues of $\mathcal{A}$. Applying Theorem 2.1 and Assumption 2.1 we obtain $\lambda_{\min}(A) = 2\beta d_g h^{p-1}$, $\lambda_{\max}(A) = Dh^p$, $\sigma_{\min}(B) = c_B h^{p-\alpha}$ and $\sigma_{\max}(B) = C_B h^{p-2}$, where $0 \leq \alpha \leq 1$ is a constant independent of the mesh size h. Substituting these into Theorem 4.1 yields

$$\begin{aligned} I^- &= \left[h^{p-2}\left(\beta d_g h - \sqrt{\beta^2 d_g^2 h^2 + C_B^2}\right), \tfrac{1}{2}h^{p-\alpha}\left(Dh^\alpha - \sqrt{D^2h^{2\alpha} + 4c_B^2}\right)\right], \\ I^+ &= \left[2\beta d_g h^{p-1}, \tfrac{1}{2}h^{p-2}\left(Dh^2 + \sqrt{D^2h^4 + 4C_B^2}\right)\right]. \end{aligned} \tag{4.3}$$

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

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

Note that $\mathcal{M}$ is a saddle point system and if we let

$$A = \begin{bmatrix} 0 & 0 \\ 0 & M \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -E & K \end{bmatrix},$$

then $\lambda_{\min^+}(A) = dh^p$, $\lambda_{\max}(A) = Dh^p$, $\sigma_{\min}(B) = c_B h^{p-\alpha}$ and $\sigma_{\max}(B) = C_B h^{p-2}$. Since A is positive semidefinite, we can apply Corollary 4.2 to bound the eigenvalues of $\mathcal{M}$. Now

$$Y_A = \begin{bmatrix} 0 \\ I_{m_T} \end{bmatrix} \quad \text{and} \quad Z_A = \begin{bmatrix} I_{m_B} \\ 0 \end{bmatrix},$$

where I_k is the identity matrix with k rows and columns. Hence, $\sigma_{\min}(BY_A) = \sigma_{\min}(K) = 0$, $\|BY_A\| = \|K\| = Ch^{p-2}$ and $\sigma_{\min}(BZ_A) = \sigma_{\min}(-E) = d_g h^{p-1}$. Let $\hat{M}_g$ and S be as defined in Assumption 2.1, $\tilde{Z}$ and $\tilde{Z}_1$ be as defined in (2.6). If we set $\tilde{Y}_B = [-E, K]^T$, then

$$\lambda_{\max}(Y_B^T A Y_B) = \lambda_{\max}(\tilde{Y}_B^T A \tilde{Y}_B, \tilde{Y}_B^T \tilde{Y}_B) = \lambda_{\max}(KMK, \hat{D}_g{}^2 + K^2) = C_{K1} h^p$$

from Assumption 2.1. Additionally, using the same assumption and letting $\tilde{Z}_B = \tilde{Z}$,

$$\lambda_{\min}(Z_B^T A Z_B) = \lambda_{\min}(\tilde{Z}^T A \tilde{Z}, \tilde{Z}^T \tilde{Z}) = \lambda_{\min}(\tilde{Z}_1^T M \tilde{Z}_1, \tilde{Z}^T \tilde{Z}) = c_Z h^{p+2}$$

and

$$\lambda_{\max}(Z_B^T A Z_B) = \lambda_{\max}(\tilde{Z}_1^T M \tilde{Z}_1, \tilde{Z}^T \tilde{Z}) = C_Z h^p.$$

Substituting these values into Corollary 4.2 we find that the eigenvalues of $\mathcal{M}$ lie in $I_{\mathcal{M}}^- \cup I_{\mathcal{M}}^+$, where

$$\begin{aligned} I_{\mathcal{M}}^- &= \left[-C_B h^{p-2}, \tfrac{1}{2}h^{p-\alpha}\left(C_{K1}h^\alpha - \sqrt{C_{K1}^2 h^{2\alpha} + 4c_B^2}\right)\right], \\ I_{\mathcal{M}}^+ &= \left[l^+, \tfrac{1}{2}h^{p-2}\left(Dh^2 + \sqrt{D^2h^4 + 4C_B^2}\right)\right], \end{aligned}$$

where

$$
\begin{aligned}
l^+ &= \max(l_1, \min(l_2, l_3)),\\
l_1 &\geq -\frac{D^2h^{2\alpha}+c_B^2}{2c_Z}h^{p-2-2\alpha} + \sqrt{\frac{\left(D^2h^{2\alpha}+c_B^2\right)^2}{4c_Z^2}h^{2p-4-4\alpha} + c_B^2h^{2p-\alpha}}\\
&= \frac{D^2h^{2\alpha}+c_B^2}{2c_Z}h^{p-2-2\alpha}\left(-1+\sqrt{1+4\frac{c_B^2c_Z^2}{\left(D^2h^{2\alpha}+c_B^2\right)^2}h^{4+2\alpha}}\right)\\
&\geq \frac{c_B^2c_Z}{D^2h^{2\alpha}+c_B^2}h^{p+2} - \frac{c_B^4c_Z^3}{\left(D^2h^{2\alpha}+c_B^2\right)^3}h^{p+6+2\alpha}\\
&\geq c_1h^{p+2},\\
l_2 &= dh^p,\\
l_3 &\geq \tfrac{1}{2}\left(-\frac{d_g^2h^2+C^2}{d}h^{p-4} + \sqrt{\frac{\left(d_g^2h^2+C^2\right)^2}{d^2}h^{2p-8} + 4d_g^2h^{2p-2}}\right)\\
&= \tfrac{1}{2}\frac{d_g^2h^2+C^2}{d}h^{p-4}\left(-1+\sqrt{1+4\frac{d^2d_g^2}{\left(d_g^2h^2+C^2\right)^2}h^6}\right)\\
&\geq \tfrac{1}{2}\frac{d_g^2h^2+C^2}{d}h^{p-4}\left(-1+1+2\frac{d^2d_g^2}{\left(d_g^2h^2+C^2\right)^2}h^6 - 8\frac{d^4d_g^4}{\left(d_g^2h^2+C^2\right)^4}h^{12}\right)\\
&= \frac{dd_g^2}{d_g^2h^2+C^2}h^{p+2} - \frac{d^3d_g^4}{\left(d_g^2h^2+C^2\right)^3}h^{p+10}\\
&\geq c_3h^{p+2},
\end{aligned}
$$

and c_1, c_3 are constants independent of the mesh size h. Hence,

$$
I^+_{\mathcal{M}} \subset \left[\min(c_1,c_2)h^{p+2}, \tfrac{1}{2}h^{p-2}\left(Dh^2+\sqrt{D^2h^4+4C_B^2}\right)\right].
$$

Using the above results for $\mathcal{M}$ and applying Theorem 2.4, we find that the eigenvalues of $\mathcal{A}$ lie in

$$
\begin{aligned}
I^- &= \left[-C_Bh^{p-2}, \tfrac{1}{2}h^{p-\alpha}\left(C_{K1}h^\alpha + 4\beta D_gh^{\alpha-1} - \sqrt{C_{K1}^2h^{2\alpha}+4c_B^2}\right)\right],\\
I^+ &\subset \left[\min(c_1,c_3)h^{p+2}, \tfrac{1}{2}h^{p-2}\left(Dh^2+4\beta D_gh+\sqrt{D^2h^4+4C_B^2}\right)\right].
\end{aligned}
\tag{4.4}
$$

Combining equations (4.3)–(4.4), we obtain the following result.

Corollary 4.4 Consider the p-dimensional problem with $p \in \{2, 3\}$. Let

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

assume that Assumption 2.1 holds, and $\beta < \frac{h}{2}\min\{\frac{d}{d_g}, \frac{D}{D_g}\}$. Let $\lambda(\mathcal{A})$ denote the spectrum of $\mathcal{A}$. Then there exists a constant $c_1 \lesssim \min(c_Z, \frac{dd_g^2}{C^2})$ independent of the mesh size h and regularization parameter β such that

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

where

$$\begin{aligned} \hat{I}^- &= \left[-C_B h^{p-2}, , \tfrac{1}{2}h^{p-\alpha}\left(Dh^\alpha - \sqrt{D^2h^{2\alpha} + 4c_B^2}\right)\right], \\ \hat{I}^+ &= \left[\max(c_1 h^{p+2}, 2\beta d_g h^{p-1}), \tfrac{1}{2}h^{p-2}\left(Dh^2 + \sqrt{D^2h^4 + 4C_B^2}\right)\right]. \end{aligned}$$

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

In Figure 4.2, we plot the condition number of $\mathcal{A}$ with respect to β for the 2D Neumann boundary control with Target 1. Results are given for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$. We observe that, as expected, if $\beta \gg \frac{C}{d}h^{-1}$, then the condition number of $\mathcal{A}$ is proportional to β and inversely proportional to the mesh size h. For $\frac{c_Z d_I^2 h^3}{d_g\left(c_B^2 + D^2\right)} \leq \beta < \frac{dh}{2D_g}$, 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^4.

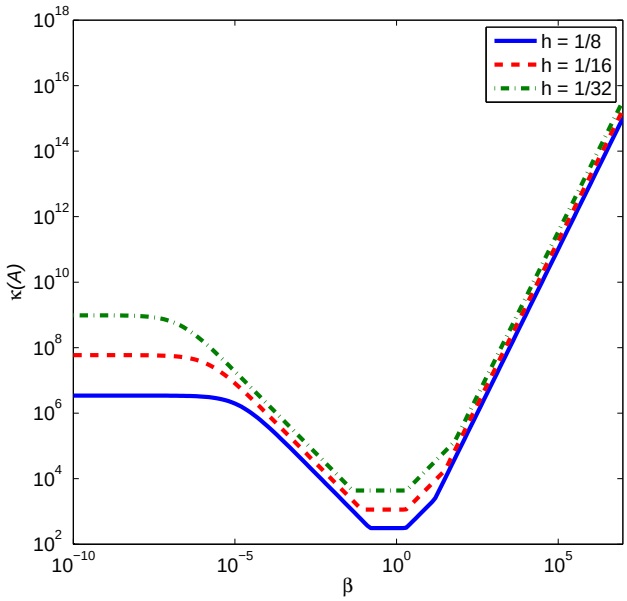

Figure 4.1: Condition number of $\mathcal{A}$ for the 2D Neumann boundary control with Target 1 and different values of β. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$.

4.3 Neumann boundary control problems with target $\hat{u}$ defined over a subdomain of Ω

For the case where the target $\hat{u}$ is not defined over all of the domain Ω, the matrix A defined in (4.1) will be positive semi-definite and singular. As for the case where $\hat{u}$ was defined over all of Ω, we will consider large values of the regularisation parameter, β, separately to small values.

Let $\beta > \max(\frac{h}{2}\max\{\frac{\bar{d}}{d_g}, \frac{\bar{D}}{D_g}\})$. We will apply Corollary 4.2 to bound the eigenvalues of $\mathcal{A}$. Clearly $\lambda_{\min^+}(A) = \bar{d}h^p$, $\lambda_{\max}(A) = 2\beta D_g h^{p-1}$ and, from Assumption 2.1, $\sigma_{\min}(B) = c_B h^{p-\alpha}$ and $\sigma_{\max}(B) = C_B h^{p-2}$, where α is a constant independent of the mesh size h and satisfies $0 \leq \alpha \leq 1$. Writing

$$Y_A = \begin{bmatrix} I_{m_B} & 0 \\ 0 & \Pi_Y \end{bmatrix} \quad \text{and} \quad Z_A = \begin{bmatrix} 0 \\ \Pi_Z \end{bmatrix},$$

where $[\Pi_Y, \Pi_Z]$ is a permutation matrix, we find that $BY_A = [-E, K\Pi_Y]$ and $BZ_A = K\Pi_Z$. Clearly we have two possibilities: $\sigma_{\min}(BY_A) = 0$ and $\sigma_{\min}(BY_A) > 0$. In the following, let $\hat{M}_g$ and S be as defined in Assumption 2.1, $\tilde{Z}$ and $\tilde{Z}_1$ be as defined in (2.6). Applying Theorem 2.4, if $\sigma_{\min}(BY_A) > 0$, then

$$\begin{aligned} \sigma^2_{\min}(BY_A) &= \lambda_{\min}(BY_A Y_A^T B^T) \\ &= \lambda_{\min}(\hat{M}_g^2 + K\Pi_Y \Pi_Y^T K) \\ &\geq \min(\lambda_{\min^+}(\hat{M}_g^2), \lambda_{\min^+}(K\Pi_Y\Pi_Y^T K)) \\ &= \min(d_g^2 h^{2(p-1)}, \bar{c}^2 h^{2(p-\alpha_1)}), \end{aligned}$$

where $\alpha_1 \leq 2$ is a constant independent of the mesh size h satisfying $\alpha \leq \alpha_1 \leq 2$. Also,

$$\begin{aligned} \sigma^2_{\min}(BY_A) &= \leq \min(\lambda_{\min}(\hat{M}_g^2) + \lambda_{\max}(K\Pi_Y\Pi_Y^T K), \lambda_{\max}(\hat{M}_g^2) + \lambda_{\min}(K\Pi_Y\Pi_Y^T K)) \\ &= \min(\bar{C}^2 h^{2(p-\alpha_2)}, D_g^2 h^{2(p-1)}), \end{aligned}$$

where α_2 is a constant independent of the meshsize h satisfying $\alpha_1 \leq \alpha_2 \leq 2$. Hence,

$$\min(d_g h^{p-1}, \bar{c}h^{p-\alpha_1}) \leq \sigma_{\min}(BY_A) \leq \min(\bar{C}h^{p-\alpha_2}, D_g h^{p-1}).$$

Therefore, if $\sigma_{\min}(BY_A) > 0$, we will assume that $\sigma_{\min}(BY_A) = c_\nu h^{p-\nu}$, where ν and c_ν are constants independent of the mesh size h, and $\nu \leq 1$.

Similarly, we can show that $\sigma_{\max}(BY_A) = c_\gamma h^{p-\gamma}$ and $\sigma_{\min}(BZ_A) = c_\delta h^{p-\delta}$, where γ, δ, c_δ and c_γ are constants independent of the mesh size h, and $\nu \leq \gamma \leq 1$ and $\delta \leq 2$.

Let $\tilde{Y}_B = [-E, K]^T$, then

$$\left\|Y_B^T A Y_B\right\| = \lambda_{\max}(\tilde{Y}_B^T A \tilde{Y}_B, \tilde{Y}_B^T \tilde{Y}_B) = \lambda_{\max}(2\beta\hat{M}_g^3 + K\bar{M}K, \hat{M}_g^2 + K^2).$$

Applying Theorem 2.4,

$$\begin{aligned} \left\|Y_B^T A Y_B\right\| &\leq 2\beta\lambda_{\max}(\hat{M}_g^3, \hat{M}_g^2 + K^2) + \lambda_{\max}(K\bar{M}K, \hat{M}_g^2 + K^2) \\ &= 2\beta D_{G1} h^{p-1} + \bar{C}_{K1} h^p, \end{aligned}$$

$$\begin{aligned} \left\|Y_B^T A Y_B\right\| &\geq 2\beta\lambda_{\min}(\hat{M}_g^3, \hat{M}_g^2 + K^2) + \lambda_{\max}(K\bar{M}K, \hat{M}_g^2 + K^2) \\ &= \bar{C}_{K1} h^p \end{aligned}$$

and

$$\begin{aligned} \left\|Y_B^T A Y_B\right\| &\geq 2\beta\lambda_{\max}(\hat{M}_g^3, \hat{M}_g^2 + K^2) + \lambda_{\min}(K\bar{M}K, \hat{M}_g^2 + K^2) \\ &= 2\beta D_{G1} h^{p-1}. \end{aligned}$$

Hence, $\left\|Y_B^T A Y_B\right\| \geq 2\beta D_{G1} h^{p-1}$ for $\beta \geq \frac{\bar{c}_{K1}}{2D_{G1}} h$. Therefore, we shall assume that $\left\|Y_B^T A Y_B\right\| = 2\beta D_{G1} h^{p-1} + c_\eta h^{p+\eta}$, where c_η is a constant (possibly zero) independent of the meshsize h and η is a constant independent of the meshsize h satisfying $0 \leq \eta \leq 2$. Using the same methodology, we can show that $\lambda_{\min}(Z_B^T A Z_B) = c_\mu h^{p+\mu}$, where μ is a constant independent of the meshsize h satisfying $0 \leq \mu \leq 2$.

Substituting the above results into Corollary 4.2, we find that the eigenvalues of $\mathcal{A}$ lie in $I^- \cup I^+$, where

$$
\begin{aligned}
I^- &= \left[-C_B h^{p-2}, \tfrac{1}{2}\left(2\beta D_{G1} h^{p-1} + c_\eta h^{p+\eta} - \sqrt{\left(2\beta D_{G1} h^{p-1} + c_\eta h^{p+\eta}\right)^2 + 4c_B^2 h^{2(p-\alpha)}}\right)\right] \\
&\subset \left[-C_B h^{p-2}, -\frac{c_B^2 h^{p+1-2\alpha}}{2\beta D_{G1} + c_\eta h^{1+\eta}} + \frac{c_B^4 h^{p+3-4\alpha}}{\left(2\beta D_{G1} + c_\eta h^{1+\eta}\right)^3}\right] \\
&\subset \left[-C_B h^{p-2}, -\bar{c}_1 h^{p+1-2\alpha}\beta^{-1}\right], \\
I^+ &= \left[\max\left(l_1, \min(l_2, l_3)\right), \left(\beta D_g h^{p-1} + \sqrt{\beta^2 D_g^2 h^{2(p-1)} + c_B^2 h^{2(p-\alpha)}}\right)\right], \\
&\subset \left[\max\left(l_1, \min(l_2, l_3)\right), 2\beta D_g h^{p-1} + \frac{c_B^2}{D_g} h^{p+1-2\alpha}\beta^{-1}\right], \\
l_1 &\geq \frac{4\beta^2 D_g^2 + c_B^2 h^{2(1-\alpha)}}{2c_\mu} h^{p-2-\mu}\left(-1 + \sqrt{1 + \frac{4c_\mu^2 c_B^2}{\left(4\beta^2 D_g^2 + c_B^2 h^{2(1-\alpha)}\right)} h^{2(2+\mu-\alpha)}}\right) \\
&\geq \frac{c_\mu c_B^2}{4\beta^2 D_g^2 + c_B^2 h^{2(1-\alpha)}} h^{p+2+\mu-2\alpha} - \frac{c_\mu^3 c_B^4}{\left(4\beta^2 D_g^2 + c_B^2 h^{2(1-\alpha)}\right)^3} h^{p+6+3\mu-4\alpha} \\
&\geq \bar{c}_2 h^{p+2+\mu-2\alpha}\beta^{-2} := \bar{l}_1, \\
l_2 &= \tfrac{1}{2}\left(\bar{d} h^p + \sqrt{\bar{d}^2 h^{2p} + 4\left(\sigma_{\min}(BY_A)\right)^2}\right), \\
l_3 &\geq \tfrac{1}{2}\left(-\frac{c_\delta^2 h^{-2\delta} + c_\gamma^2 h^{-2\gamma}}{\bar{d}} h^p + \sqrt{\frac{\left(c_\delta^2 h^{-2\delta} + c_\gamma^2 h^{-2\gamma}\right)^2}{\bar{d}^2} h^{2p} + 4c_\delta^2 h^{2(p-\delta)}}\right) \\
&\geq \frac{\bar{d} c_\delta^2}{c_\delta^2 h^{-2\delta} + c_\gamma^2 h^{-2\gamma}} h^{p-2\delta} - \frac{\bar{d}^3 c_\delta^4}{\left(c_\delta^2 h^{-2\delta} + c_\gamma^2 h^{-2\gamma}\right)^3} h^{p-4\delta} \\
&\geq \bar{c}_3 h^{p+2\max(0,\gamma-\delta)} := \bar{l}_3,
\end{aligned}
$$

where $\bar{c}_1$, $\bar{c}_2$ and $\bar{c}_3$ are constants independent of the mesh size h and regularization parameter β. Now $l^+ = \max(l_1, \min(l_2, l_3)) \geq \max(\bar{l}_1, \min(l_2, \bar{l}_3))$. If $\sigma_{\min}(BY_A) = 0$, then $l_2 = \bar{d} h^p$; otherwise $l_2 = \frac{1}{2} h^{p-\nu}\left(\bar{d} h^\nu + \sqrt{\bar{d}^2 h^{2\nu} + 4c_\nu^2}\right)$. In either case, $\max(\bar{l}_1, \min(l_2, \bar{l}_3)) = \max(\bar{l}_1, \bar{l}_3) = \bar{c}_4 h^{p+2\max(0,\gamma-\delta)}$ for some constant $\bar{c}_4$ and $\beta > \sqrt{\frac{\bar{c}_2}{\bar{c}_4}} h^{1-\frac{\mu}{2}-\alpha-\max(0,\gamma-\delta)}$ and

$$
I^+ = \left[\bar{c}_4 h^{p+2\max(0,\gamma-\delta)}, 2\beta D_g h^{p-1} + \frac{c_B^2}{D_g} h^{p+1-2\alpha}\beta^{-1}\right]. \tag{4.5}
$$

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

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

Note that $\mathcal{M}$ is singular. Using Corollary 4.2, the positive eigenvalues of $\mathcal{M}$ lie in

$$
\left[2\beta d_g h^{p-1}, 2\beta D_g h^{p-1}\right] \cup \left[l^+, \tfrac{1}{2} h^{p-2}\left(\bar{D} h^2 + \sqrt{\bar{D}^2 h^4 + 4C^2}\right)\right]
$$

for some positive l^+. Applying Theorem 2.4 to these bounds and incorporating the bounds given in (4.5), we obtain the following result.

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

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

assume that Assumptions 2.1 and 2.2 hold, $\sigma_{\min}(BZ_A) = c_\delta h^{p-\delta}$, $\sigma_{\max}(BY_A) = c_\gamma h^{p-\gamma}$ and $\lambda_{\min}(Z_B^T A Z_B) = c_\mu h^{p+\mu}$, where δ, γ, μ, c_δ, c_γ and c_μ are constants independent of the mesh size h and regularization parameter β subject to $\alpha \leq \delta \leq 2$ and $\alpha \leq \gamma \leq 1$. There exist constants c_0, $\bar{c}_1$ and $\bar{c}_2$ independent of the mesh size h and regularization parameter β such that if $\beta > c_0 h^{1-\frac{\mu}{2}-\alpha-\max(0,\gamma-\delta)}$, then

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

where

$$\begin{aligned} I^- &= \left[-C_B h^{p-2}, -\bar{c}_1 h^{p+1-2\alpha}\beta^{-1}\right], \\ I_1^+ &= \left[\bar{c}_2 h^{p+2\max(0,\gamma-\delta)}, \tfrac{1}{2}h^{p-2}\left(\bar{D}h^2 + 2D_g h + \sqrt{\bar{D}^2 h^4 + 4C^2}\right)\right], \\ I_2^+ &= \left[(2\beta d_g - D_g)h^{p-1}, 2\beta D_g h^{p-1} + \frac{c_B^2}{D_g} h^{p+1-2\alpha}\beta^{-1}\right]. \end{aligned}$$

Consequently, if $\beta \gg c_0 h^{1-\frac{\mu}{2}-\alpha-\max(0,\gamma-\delta)}$, then

$$\begin{aligned} \sigma_{\min}(\mathcal{A}) &\geq \bar{c}_1 h^{p+1-2\alpha}\beta^{-1}, \\ \sigma_{\max}(\mathcal{A}) &\leq \beta \bar{c}_3 h^{p-1}, \end{aligned}$$

where $\bar{c}_3$ is a constant independent of the mesh size h and regularization parameter β. The condition number of $\mathcal{A}$ bounded from above by a term that is proportional to $\beta^2 h^{2(\alpha-1)}$.

Consider the case $\beta < \min\left(\frac{\bar{d}h}{2d_g}, \frac{\bar{D}h}{2D_g}\right)$. If $\beta = 0$, then $\mathcal{A}_0 = \mathcal{A}$, where

$$\mathcal{A}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & -M_g \\ 0 & \bar{M}_{II} & \bar{M}_{IB} & K_{II} & K_{BI}^T \\ 0 & \bar{M}_{BI} & \bar{M}_{BB} & K_{BI} & K_{BB} \\ 0 & K_{II} & K_{BI}^T & 0 & 0 \\ -M_g & K_{BI} & K_{BB} & 0 & 0 \end{bmatrix}.$$

The matrix M_g is nonsingular and, hence, $\mathcal{A}_0$ is nonsingular if and only if

$$\mathcal{H} = \begin{bmatrix} \bar{M}_{II} & \bar{M}_{IB} & K_{II} \\ \bar{M}_{BI} & \bar{M}_{BB} & K_{BI} \\ K_{II} & K_{BI}^T & 0 \end{bmatrix}$$

is nonsingular. $\mathcal{H}$ is a saddle-point system and is nonsingular if and only if $Z_K^T \bar{M} Z_K$ is nonsingular, where the columns of Z_K span the nullspace of $\left[K_{II}, K_{BI}^T\right]$, see [2, Section 3]. Setting $Z_K = \tilde{Z}_1$, where $\tilde{Z}_1$ is defined in (2.6), we determine that $\mathcal{A}_0$ is nonsingular if and only if $K_{BI}K_{II}^{-1}\bar{M}_{II}K_{II}^{-1}K_{IB} - \bar{M}_{BI}K_{II}^{-1}K_{IB} - K_{BI}K_{II}^{-1}\bar{M}_{IB} + \bar{M}_{BB}$ is nonsingular. Note that if the target $\hat{u}$ is not defined on the boundary, then $\bar{M}_{BB} = 0$ and $\bar{M}_{BI} = 0$. Hence, $\tilde{Z}_1^T \bar{M} \tilde{Z}_1 = K_{BI}K_{II}^{-1}\bar{M}_{II}K_{II}^{-1}K_{IB}$ which is singular. Therefore, a necessary (but not sufficient) condition for $\mathcal{A}_0$ to be nonsingular is that $\hat{u}$ must be defined on part of the boundary. If $Z_K^T \bar{M} Z_K$ is nonsingular, we will expect the eigenvalues of $\mathcal{A}$ to be bounded away from zero as $\beta \to 0$; if $Z_K^T \bar{M} Z_K$ is singular, then some of the eigenvalues of $\mathcal{A}$ will converge to zero as $\beta \to 0$.

Let

$$A = \begin{bmatrix} 2\beta M_g & 0 \\ 0 & \bar{M} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -E & K \end{bmatrix},$$

where E is defined by (2.5). We can apply Corollary 4.2 to obtain bounds on the eigenvalues of $\mathcal{A}$. From Assumptions 2.1 and 2.2, and Theorem 2.6, we obtain $\sigma_{\min}^B = c_B h^{p-\alpha}$, $\sigma_{\max}^B = C_B h^{p-2}$ and $\mu_{\max}^A = \bar{D}h^p$. As for the case when β was large, we shall assume that $\sigma_{\min}(BY_A) = 0$ or $\sigma_{\min}(BY_A) = c_\nu h^{p-\nu}$, $\sigma_{\max}(BY_A) = c_\gamma h^{p-\gamma}$ and $\sigma_{\min}(BZ_A) = c_\delta h^{p-\delta}$, where ν, γ, δ, c_ν, c_δ and c_γ are constants independent of the mesh size h and regularization parameter β, and $\nu \leq 1$, $\nu \leq \gamma \leq 1$ and $\delta \leq 2$. Similarly, we can also show that $\left\|Y_B^T A Y_B\right\| = C_{K1}h^p + 2\beta c_\eta h^{p+\eta}$ for $\beta < \frac{C_{K1}}{2\bar{D}G1}h^3$ and $\lambda_{\min}(Z_B^T A Z_B) = c_3 h^{p+2} + 2\beta c_\mu h^{p+\mu}$, where η, μ, c_3, c_μ and c_η are constants independent of the mesh size h and regularization parameter β, and $-1 \leq \eta \leq 1$ and $-1 \leq \mu \leq 1$. If $Z_K^T \bar{M} Z_K$ is nonsingular, then $c_3 = \bar{c}_Z$; otherwise $c_3 = 0$. Applying Corollary 4.2, the eigenvalues of $\mathcal{A}$ lie in $I^- \cup I^+$, where

$$
\begin{aligned}
I^- &= \left[-C_B h^{p-2}, \tfrac{1}{2}h^{p-\alpha}\left(\bar{D}h^\alpha - \sqrt{\bar{D}^2h^{2\alpha} + c_B^2}\right)\right], \\
I^+ &= \left[l^+, \tfrac{1}{2}h^{p-2}\left(\bar{D}h^2 + \sqrt{\bar{D}^2h^4 + 4C_B^2}\right)\right], \\
l^+ &= \max(l_1, \min(l_2, l_3)), \\
l_1 &\geq -\frac{\bar{D}^2h^{2\alpha} + c_B^2}{2\left(c_3h^{2-\mu} + 2\beta c_\mu\right)}h^{p-\mu-2\alpha} + \sqrt{\frac{\left(\bar{D}^2h^{2\alpha} + c_B^2\right)^2}{4\left(c_3h^{2-\mu} + 2\beta c_\mu\right)^2}h^{2p-2\mu-4\alpha} + c_B^2 h^{2p-2\alpha}} \\
&\geq \frac{c_B^2\left(c_3h^{2-\mu} + 2\beta c_\mu\right)}{\bar{D}^2h^{2\alpha} + c_B^2}h^{p+\mu} - \frac{c_B^4\left(c_3h^{2-\mu} + 2\beta c_\mu\right)^3}{\left(\bar{D}^2h^{2\alpha} + c_B^2\right)^3}h^{p+\mu} := \bar{l}_1, \\
l_2 &= \beta d_g h^{p-1} + \sqrt{\beta^2 d_g^2 h^{2p-2} + \left(\sigma_{\min}(BY_A)\right)^2}, \\
l_3 &\geq \tfrac{1}{2}\left(-\frac{c_\delta^2h^{-2\delta} + c_\gamma^2h^{-2\gamma}}{2\beta d_g}h^{p+1}\right) + \sqrt{\frac{\left(c_\delta^2h^{-2\delta} + c_\gamma^2h^{-2\gamma}\right)^2}{4\beta^2 d_g^2}h^{2p+2} + 4c_\delta^2 h^{2p-3\delta}} \\
&\geq \frac{2c_\delta^2 d_g}{c_\delta^2h^{-2\delta} + c_\gamma^2h^{-2\gamma}}\beta h^{p-1-2\delta} - \frac{8c_\delta^4 d_g^3}{\left(c_\delta^2h^{-2\delta} + c_\gamma^2h^{-2\gamma}\right)^3}\beta^3 h^{p-3-4\delta} := \bar{l}_3.
\end{aligned}
$$

Now $l^+ \geq \max(\bar{l}_1, \min(l_2, \bar{l}_3))$. If $\sigma_{\min}(BY_A) = 0$, then $l_2 = 2\beta d_g h^{p-1}$; otherwise

$$l_2 = \beta d_g h^{p-1} + \sqrt{\beta^2 d_g^2 h^{2p-2} + c_\nu h^{2p-2\nu}}.$$

For either case, $\min(l_2, \bar{l}_3) = \bar{l}_3$. If $Z_K^T \bar{M} Z_K$ is nonsingular, then $l^+ \geq \bar{l}_1$; otherwise,

$$l^+ \geq u_3 \beta h^{p-\max(1-2\max(0,\gamma-\delta),\mu)},$$

where u_3 is a constant independent of h and β.

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

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

assume that Assumptions 2.1 and 2.2 hold, $\sigma_{\min}(BZ_A) = c_\delta h^{p-\delta}$, $\sigma_{\max}(BY_A) = c_\gamma h^{p-\gamma}$ and $\lambda_{\min} Z_B^T A Z_B = c_2 h^{p+2} + 2\beta c_\mu h^{p-\mu}$, where δ, γ, μ, c_δ, c_2, c_μ and c_γ are constants independent of the mesh size h and regularization parameter β subject to $\alpha \leq \delta \leq 2$ and $\alpha \leq \gamma \leq 1$. Additionally, if $\tilde{Z}_1^T \bar{M} \tilde{Z}_1$ is nonsingular, then $c_2 = \bar{c}_Z$; otherwise, $c_2 = 0$. There exist constants c_0, $\bar{c}_1$ and $\bar{c}_2$ independent of the mesh size h and regularization parameter β such that if $\beta < c_0 h^3$, then

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

where

$$\begin{aligned} I^- &= \left[-C_B h^{p-2}, \tfrac{1}{2} h^{p-\alpha} \left(\bar{D} h^\alpha - \sqrt{\bar{D}^2 h^{2\alpha} + c_B^2} \right) \right], \\ I^+ &= \left[l^+, \tfrac{1}{2} h^{p-2} \left(\bar{D} h^2 + \sqrt{\bar{D}^2 h^4 + 4 C_B^2} \right) \right], \\ l^+ &= \begin{cases} \bar{c}_1 \left(\bar{c}_Z h^{p+2} + 2\beta c_\mu h^{p-\mu} \right) & \text{if } \tilde{Z}_1^T \bar{M} \tilde{Z}_1 \text{ is nonsingular,} \\ \bar{c}_2 \beta h^{p - \max(1 - 2\max(0, \gamma - \delta), \mu)} & \text{otherwise.} \end{cases} \end{aligned}$$

Consequently, $\sigma_{\max}(\mathcal{A}) \leq \frac{h^{p-2}}{2} \left(\bar{D} h^2 - \sqrt{\bar{D} h^4 + 4 C_B^2} \right)$. If $\tilde{Z}_1^T \bar{M} \tilde{Z}_1$ is singular, then

$$\sigma_{\min}(\mathcal{A}) \geq \bar{c}_2 \beta h^{p - \max(1 - 2\max(0, \gamma - \delta), \mu)}$$

and

$$\kappa(\mathcal{A}) = \mathcal{O}(\beta^{-1} h^{\max(-1 - 2\max(0, \gamma - \delta), \mu - 2)}).$$

If $\tilde{Z}_1^T \bar{M} \tilde{Z}_1$ is nonsingular, then $\sigma_{\min}(\mathcal{A}) \geq \bar{c}_1 \left(\bar{c}_Z h^{p+2} + 2\beta c_\mu h^{p-\mu} \right)$. If, in addition, $\beta \gg \frac{\bar{c}_Z}{2 c_\mu} h^{2+\mu}$, then $\kappa(\mathcal{A}) = \mathcal{O}(\beta^{-1} h^{\mu - 2})$. If $\beta \ll \frac{\bar{c}_Z}{2 c_\mu} h^{2+\mu}$, then $\kappa(\mathcal{A}) = \mathcal{O}(h^{-4})$.

In Figure 4.3, we plot the condition number of $\mathcal{A}$ with respect to β for the 2D Neumann boundary control with Target 3. Results are given for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$. Now $\tilde{Z}_1^T \bar{M} \tilde{Z}_1$ is singular. Numerical experiments reveal that $\alpha \approx \frac{1}{4}$, $\delta = 0$, $\gamma = 2$ and $c_\nu = 0$. If β is large, then $\mu = 0$. From Corollary 4.5, we will expect there to be a constant c_0 such that if $\beta \gg c_0 h^{-\frac{5}{4}}$, then $\kappa(\mathcal{A})$ is bounded from above by a function that is proportional to $\beta^2 h^{-\frac{3}{2}}$. If β is small, then $\mu = 1$ and, from Corollary 4.6, we will expect there to be a constant c_0 such that if $\beta < c_0 h^3$, then $\kappa(\mathcal{A})$ to be bounded from above by a function that is proportional to $\beta^{-1} h^{-1}$. For both these cases of β, we find that these upper bounds accurately reflect the condition number of $\mathcal{A}$, Figure 4.3.

If we consider the 2D Neuman boundary control problem with Target 4, then $\tilde{Z}_1^T \bar{M} \tilde{Z}_1$ is nonsingular. Similarly to Target 3, numerical experiments reveal that $\alpha \approx \frac{1}{4}$, $\delta = 0$, $\gamma = 2$ and $c_\nu = 0$. If β is small, then $\mu = 0$; otherwise $\mu = 1$. Applying Corollary 4.5, we will expect the condition number of $\mathcal{A}$ to be bounded above by a function that is proportional to $\beta^2 h^{-\frac{3}{2}}$ for all $\beta \gg c_0 h^{-\frac{5}{4}}$, where c_0 is a constant independent of h and β. For small values of β, we will expect there to be constants c_0 and c_1 such that if $c_1 h^3 \beta \ll c_0 h^3$, then the condition number of $\mathcal{A}$ will be bounded from above by a function proportional to $\beta^{-1} h^{-1}$; if $\beta \ll c_0 h^3$, then $\kappa(\mathcal{A}) = \mathcal{O}(h^{-4})$. In Figure 4.3, we plot the condition number of $\mathcal{A}$ with respect to β for the 2D Neumann boundary control with Target 4. We observe that these upper bounds accurately reflect $\kappa(\mathcal{A})$.

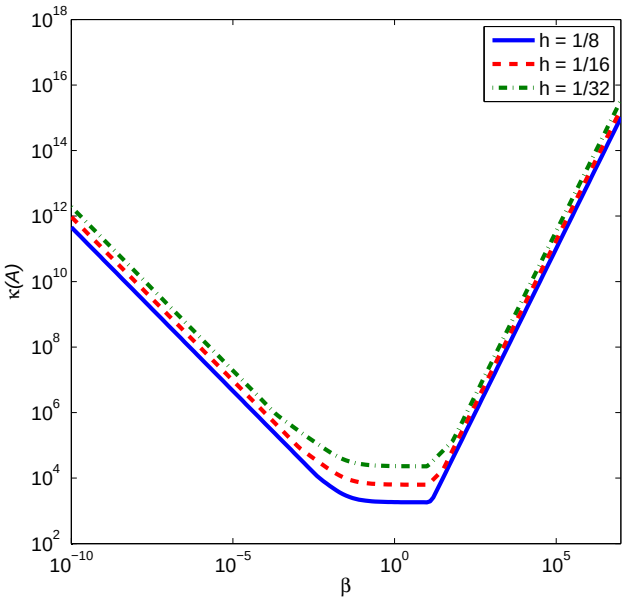

Figure 4.2: Condition number of $\mathcal{A}$ for the 2D Neumann boundary control with Target 3 and different values of β. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$.

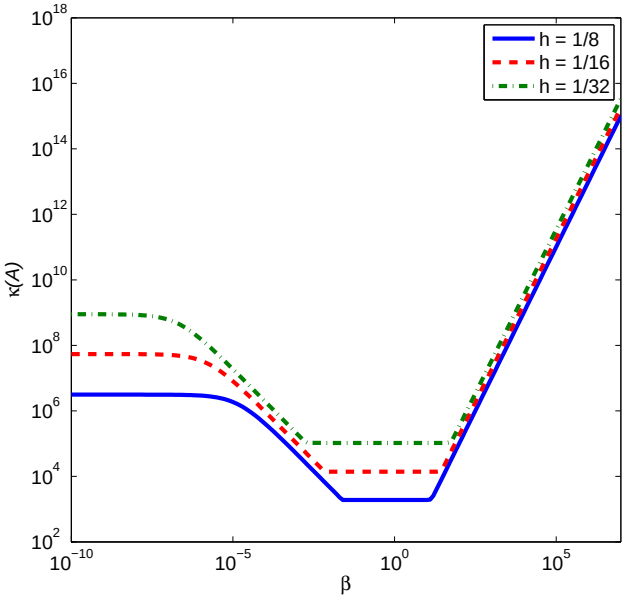

Figure 4.3: Condition number of $\mathcal{A}$ for the 2D Neumann boundary control with Target 4 and different values of β. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$.

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 $\Delta s \approx \tilde{s} - s$:

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

from which we obtain the bound

$$\|\Delta s\| \leq \left\|\mathcal{A}^{-1}\right\|_2 \|\Delta\mathcal{A}\|_2 \|s\|. \tag{4.2}$$

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

$$\|\Delta\mathcal{A}\| \leq \mathtt{u}\gamma_N \|\mathcal{A}\|, \tag{4.3}$$

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 [10] and [20], 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 (4.2) and (4.3) we obtain

$$\|\Delta s\| \leq \mathtt{u}\gamma_N\kappa(\mathcal{A}) \|s\|. \tag{4.4}$$

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 [25] 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 (4.4).

Let $\mathcal{A}$ be factorized as

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

where U 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, $\|\mathcal{A}\| = \|\Sigma_L\|$, $\left\|\mathcal{A}^{-1}\right\| = \left\|\Sigma_S^{-1}\right\|$ and $\kappa(\mathcal{A}) = \|\Sigma_L\| \left\|\Sigma_S^{-1}\right\|$. 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

$$\begin{aligned} b &= b_L + b_S = U_L\delta_L + U_S\delta_S, \\ s &= s_L + s_S = V_L\psi_L + V_S\psi_S, \end{aligned} \tag{4.5}$$

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

$$\Sigma\psi = \begin{bmatrix} \Sigma_L\psi_L \\ \Sigma_S\psi_S \end{bmatrix} = \begin{bmatrix} \delta_L \\ \delta_S \end{bmatrix} = \delta. \tag{4.6}$$

From (4.6) we obtain

$$\|b_L\| \le \|\Sigma_L\| \|s_L\|, \quad \|b_S\| \le \|\Sigma_S\| \|s_S\|,$$

and

$$\|s_L\| \le \left\|\Sigma_L^{-1}\right\| \|b_L\|, \quad \|s_S\| \le \left\|\Sigma_S^{-1}\right\| \|b_S\|.$$

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

$$G = \begin{bmatrix} G_L \\ G_S \end{bmatrix} = \begin{bmatrix} G_{L1} & G_{L2} \\ G_{S1} & G_{S2} \end{bmatrix}.$$

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

$$\begin{bmatrix} \Delta\psi_L \\ \Delta\psi_S \end{bmatrix} = -\begin{bmatrix} \Sigma_L^{-1}G_L\psi \\ \Sigma_S^{-1}G_S\psi \end{bmatrix}. \tag{4.7}$$

This implies that

$$\begin{aligned} \|\Delta s_L\| &\le \left\|\Sigma_L^{-1}\right\| \|G_L\| \|s\| \le \left\|\Sigma_L^{-1}\right\| \|\Delta\mathcal{A}\| \|s\|, \\ \|\Delta s_S\| &\le \left\|\Sigma_S^{-1}\right\| \|G_S\| \|s\| \le \left\|\Sigma_S^{-1}\right\| \|\Delta\mathcal{A}\| \|s\|. \end{aligned} \tag{4.8}$$

Since $\|\Sigma_L\| = \|\mathcal{A}\|$, (4.8) implies that

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

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.

From [25, Theorem 3.1], we have the following theorem.

Theorem 4.7 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 $d_1 = \text{sep}(L_1, L_2) - \|E_{11}\| - \|E_{22}\|$ and $v = \|E_{12}\|/d_1$, where $\text{sep}(L_1, L_2) = \min_{i,j} |\lambda_i(L_1) - \lambda_j(L_2)|$. If $d_1 > 0$ and $v < \frac{1}{2}$, then

(i) 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 $\left\|X_2 - \tilde{X}_2\right\| \leq 2v$;

(ii) for $i = 1, \ldots, l$, there is an eigenvalue $\tilde{\lambda}$ of $\tilde{\mathcal{M}}$ satisfying $\left|\tilde{\lambda} - \breve{\lambda}_i\right| \leq 3\|E_{12}\| v$, where $\{\breve{\lambda}\}$ are the eigenvalues of $X_1^T \tilde{\mathcal{M}} X_1$.

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

$$\mathcal{M} = \left[\begin{array}{c|cccc} 2\beta M_g & 0 & 0 & 0 & 0 \\ \hline 0 & M_{II} & M_{IB} & K_{II} & 0 \\ 0 & M_{BI} & M_{BB} & 0 & K_{BB} \\ 0 & K_{II} & 0 & 0 & 0 \\ 0 & 0 & K_{BB} & 0 & 0 \end{array}\right] \text{ and } \mathcal{E} = \left[\begin{array}{c|cccc} 0 & 0 & 0 & 0 & -M_g \\ \hline 0 & 0 & 0 & 0 & K_{IB} \\ 0 & 0 & 0 & K_{BI} & 0 \\ 0 & 0 & K_{IB} & 0 & 0 \\ -M_g & K_{BI} & 0 & 0 & 0 \end{array}\right].$$

If

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

then $[X_1, X_2]$ is orthogonal, and both range(X_1) and range(X_2) are simple invariant subspaces of $\mathcal{M}$. Applying Theorem 4.7, we have

$$L_1 = 2\beta M_g, \quad L_2 = \begin{bmatrix} M_{II} & M_{IB} & K_{II} & 0 \\ M_{BI} & M_{BB} & 0 & K_{BB} \\ K_{II} & 0 & 0 & 0 \\ 0 & K_{BB} & 0 & 0 \end{bmatrix}.$$

Correspondingly,

$$E_{11} = 0, \quad E_{22} = \begin{bmatrix} 0 & 0 & 0 & K_{IB} \\ 0 & 0 & K_{BI} & 0 \\ 0 & K_{IB} & 0 & 0 \\ K_{BI} & 0 & 0 & 0 \end{bmatrix}, \quad \text{and} \quad E_{12} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -M_g \end{bmatrix}^T.$$

Observe that L_2 is of saddle-point form (4.1), with $A = M$ and $B = \begin{bmatrix} K_{II} & 0 \\ 0 & K_{BB} \end{bmatrix}$. From Theorems 2.1,

2.2 and 4.1, the eigenvalues of L_2 lie in $I^- \cup I^+$, where

$$\begin{aligned} I^- &= \left[\tfrac{1}{2}h^{p-2}\left(dh^2 - \sqrt{minM^2h^4 + 4C_1^2}\right), \tfrac{1}{2}h^p\left(D - \sqrt{D^2 + 4c_{II}^2}\right)\right], \\ I^+ &= \left[dh^p, \tfrac{1}{2}h^{p-2}\left(Dh^2 + \sqrt{D^2h^4 + 4C_1^2}\right)\right], \\ C_1 &= \max\left(C_{BB}, C_{II}\right). \end{aligned}$$

Thus, $d_1 := \text{sep}(L_1, L_2) - \|E_{11}\| - \|E_{22}\| \geq 2\beta d_g h^{p-1} - \frac{1}{2}h^{p-2}\left(Dh^2 + \sqrt{D^2h^4 + 4C_1^2}\right) - C_{BI}h^{p-2} > 0$ for $\beta \geq \frac{1}{4d_g}h^{-1}\left(Dh^2 + \sqrt{D^2h^4 + 4C_1^2} + 2C_{BI}\right)$. Now,

$$v := \|E_{12}\|/d_1 = \frac{D_g h}{2\beta d_g h - \frac{1}{2}\left(Dh^2 + \sqrt{D^2h^4 + 4C_1^2}\right) - C_{BI}}.$$

Hence, $v \to 0$ as $\beta \to +\infty$. Clearly, $v < \frac{1}{2}$ if $\beta \geq \frac{D_g}{d_g} + \frac{1}{4d_g}h^{-1}\left(Dh^2 + \sqrt{D^2h^4 + 4C_1^2} + 2C_{BI}\right)$.

Suppose that $\beta \gg \frac{1}{4d_g}h^{-1}\left(Dh^2 + \sqrt{D^2h^4 + 4C^2} + 2D_g h\right)$, then Corollary 4.3 and its derivation tells us that if $M_g \in \mathbb{R}^{m_B \times m_B}$, $\mathcal{A}$ has m_B 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 4.3 and part(i) of Theorem 4.7, we find that there are orthonormal bases $\tilde{X}_1$ and $\tilde{X}_2$ such that

$$\tilde{X}_1 = \begin{bmatrix} I_{m_B} \\ 0 \end{bmatrix} + \tfrac{D_g}{\beta d_g}\begin{bmatrix} \Upsilon_{11} \\ \Upsilon_{12} \end{bmatrix} + \mathcal{O}(\beta^{-2}), \quad \tilde{X}_2 = \begin{bmatrix} 0 \\ I_{2m_T} \end{bmatrix} + \tfrac{D_g}{\beta d_g}\begin{bmatrix} \Upsilon_{21} \\ \Upsilon_{22} \end{bmatrix} + \mathcal{O}(\beta^{-2}), \tag{4.9}$$

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_B 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

$$U_L = \tilde{X}_1\tilde{U}_L, \quad U_S = \tilde{X}_2\tilde{U}_S, \quad V_L = \tilde{X}_1\tilde{U}_L, \quad \text{and} \quad V_S = \tilde{X}_2\tilde{V}_S. \tag{4.10}$$

When $\beta \gg \frac{1}{2d_g}h^{-1}\left(C_1 + C_{BI}\right)$, we find that the solution of (2.3) satisfies $\mathbf{g} = \mathcal{O}(\beta^{-1})$, $\lambda = \mathcal{O}(1)$ and $\mathbf{u} = \mathcal{O}(1)$, Section 5. Using (4.5) and defining $w = [\mathbf{u}^T, \lambda^T]^T$, we obtain

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

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 $|\Delta\mathcal{A}| \leq \mathtt{u}\gamma_N|\mathcal{A}|$, [13]. Hence, we may write $\Delta\mathcal{A}$ as

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

where the entries of E_{11}, E_{12}, E_{21} and E_{22} are $\mathcal{O}(1)$. Using (4.9), (4.10) 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 (4.7) we have

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

From the derivation of Corollary 4.3, we know that $\left\|\Sigma_L^{-1}\right\| = \frac{1}{2\beta D_g}+\mathcal{O}(\beta^{-2})$. Hence,

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

where c_2 is a constant independent of β. Hence, for $\beta \gg \frac{1}{2d_g}h^{-1}\left(C_1+C_{BI}\right)$, 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 4.2 gives an upper bound that is proportional to β, which is a gross overestimate. Finally, we observe that since $\tilde{U}_L\psi_L \rightarrow \mathbf{f}$ as $\beta\rightarrow\infty$ and $\mathbf{g}=\mathcal{O}(\beta^{-1})$, if $\beta \gg \frac{1}{2d_g}h^{-1}\left(C_1+C_{BI}\right)$, then $\|\Delta\mathbf{g}\|/\|\mathbf{g}\|$ will be $\mathcal{O}(\mathtt{u})$.

Let us now consider ψ_S. From (4.7),

$$\begin{aligned}\Delta\psi_S &= -\Sigma_S^{-1}\left[\begin{array}{cc} G_{S1} & G_{S2}\end{array}\right]\left[\begin{array}{c}\psi_L\\ \psi_S\end{array}\right]\\ &= -\Sigma_S^{-1}\left[\begin{array}{cc} \beta^{-1}G_{S1} & G_{S2}\end{array}\right]\left[\begin{array}{c}\rho_L\\ \psi_S\end{array}\right]\\ &= -\mathtt{u}\gamma_N\Sigma_S^{-1}\left[\begin{array}{cc} \beta^{-1}\hat{E}_{21}+\mathcal{O}(\beta^{-2}) & \hat{E}_{22}+\mathcal{O}(\beta^{-1})\end{array}\right]\left[\begin{array}{c}\rho_L\\ \psi_S\end{array}\right].\end{aligned}$$

Applying Corollary 4.3, we can bound $\left\|\Sigma_S^{-1}\right\|$ from above by $c_3\beta$, where c_3 is a constant independent of β. Hence,

$$\begin{aligned}\|\Delta s_S\| &= \|\Delta\psi_S\|\\ &\leq \mathtt{u}\gamma_N\left\|\Sigma_S^{-1}\right\|\left\|\left[\begin{array}{cc} \beta^{-1}\hat{E}_{21}+\mathcal{O}(\beta^{-2}) & \hat{E}_{22}+\mathcal{O}(\beta^{-1})\end{array}\right]\right\|\left\|\left[\begin{array}{c}\rho_L\\ \psi_S\end{array}\right]\right\|\\ &= \mathtt{u}\beta c_4+\mathcal{O}(1),\end{aligned}$$

where c_4 is a constant independent of β. Hence, we will expect the relative error to be proportional to β for large enough values of β. Now, the condition number of $\mathcal{A}$ is proportional to β^2 for such values of β and, hence, the illconditioning of $\mathcal{A}$ is not effecting us as much as we might have expected.

Similarly, if $\beta \gg \frac{1}{2d_g}h^{-1}\left(C_1+C_{BI}\right)$ and $\hat{\Omega}\neq\Omega$, we can show that $\|\Delta\mathbf{g}\|/\|\mathbf{g}\|$ will be $\mathcal{O}(\mathtt{u})$ and $\|\Delta\mathbf{u}\|/\|\mathbf{u}\|$ to be $\mathcal{O}(\mathtt{u}\beta)$. If β is small, then similar arguments do not hold for either $\hat{\Omega}=\Omega$ or $\hat{\Omega}\neq\Omega$. As a result, if h remains fixed, we will expect $\|\Delta\mathbf{g}\|/\|\mathbf{g}\|$ and $\|\Delta\mathbf{u}\|/\|\mathbf{u}\|$ to be proportional to the condition number of $\mathcal{A}$ as $\beta\rightarrow 0$.

In Figure 4.1, we plot $\|\Delta\mathbf{g}\|/\|\mathbf{g}\|$ and $\|\Delta\mathbf{u}\|/\|\mathbf{u}\|$ against β for the 2D Neuman boundary control problem with Target 3. 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 $\mathcal{O}(\mathtt{u})$.

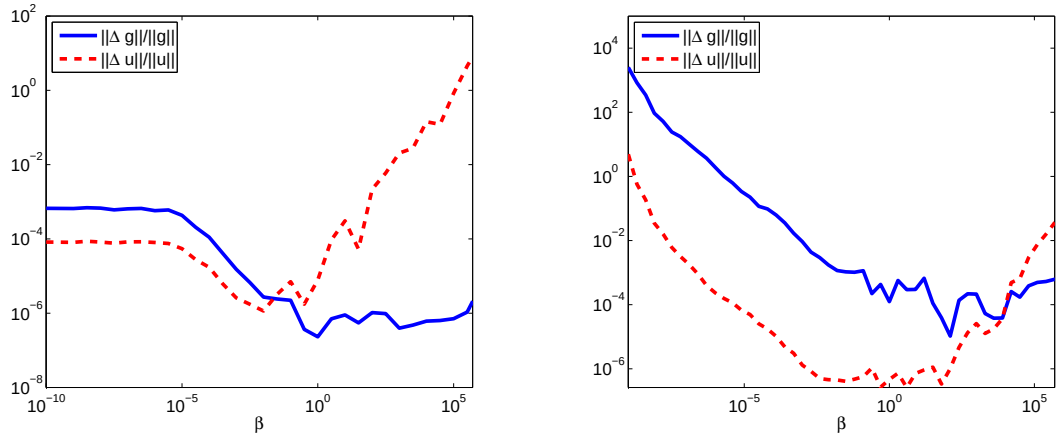

Figure 4.1: Plot of $\|\Delta \mathbf{g}\| / \|\mathbf{g}\|$ and $\|\Delta \mathbf{u}\| / \|\mathbf{u}\|$ with respect to β for the 2D Neuman boundary control problem with Target 1 (left) and Target 3 (right). Results are shown for $h = \frac{1}{8}$.

5 Schur complement method

A common method for solving systems of the form (2.3) is to reduce it to a series of smaller systems that need to be solved. For example, 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 (2.3), this method will only be applicable if the target $\hat{u}$ is defined over all of Ω. Applying the Schur complement method to (2.3) we obtain

$$\begin{aligned} \lambda &= -T^{-1}\left(\mathbf{d} - KM^{-1}\mathbf{b}\right), \\ \mathbf{g} &= \tfrac{1}{2\beta}\left[0 \;\; I_{m_B}\right]\lambda, \\ \mathbf{u} &= M^{-1}\left(\mathbf{b} - K\lambda\right), \end{aligned}$$

where

$$T = \tfrac{1}{2\beta}\hat{M}_g + KM^{-1}K$$

and $\hat{M}_g$ is defined in Assumption 2.1. Thus, it is necessary to be able to carry out solves with M_g, M and the Schur complement S. There are good methods available for solving systems of the form $Mx = b$ and $M_g x = b$, see [23], so we will concern ourselves with the Schur complement.

Suppose that the theorems and assumptions in Section 2.1 hold, then T is positive definite and

$$\begin{aligned}
\lambda_{\min}(T) &\geq \min\left(\tfrac{1}{2\beta}\lambda_{\min^+}\left(\hat{M}_g\right), \lambda_{\min^+}\left(KM^{-1}K\right)\right) \\
&\geq \min\left(\tfrac{d_g h^{p-1}}{2\beta}, \tfrac{c^2 h^p}{D}\right), \\
\lambda_{\min}(T) &\leq \min\left(\tfrac{1}{2\beta}\lambda_{\max}(\hat{M}_g), \lambda_{\max}(KM^{-1}K)\right) \\
&\leq \min\left(\tfrac{D_g h^{p-1}}{2\beta}, \tfrac{C^2 h^{p-4}}{d}\right), \\
\lambda_{\max}(T) &\geq \max\left(\tfrac{1}{2\beta}\lambda_{\max}(\hat{M}_g), \lambda_{\max}(KM^{-1}K)\right) \\
&\geq \max\left(\tfrac{D_g h^{p-1}}{2\beta}, \tfrac{C^2 h^{p-4}}{d}\right), \\
\lambda_{\max}(T) &\leq \tfrac{1}{2\beta}\lambda_{\max}\left(\hat{M}_g\right) + \lambda_{\max}\left(KM^{-1}K\right) \\
&\leq \tfrac{D_g h^{p-1}}{2\beta} + \tfrac{C^2 h^{p-4}}{d}.
\end{aligned}$$

From this we obtain the following result.

Corollary 5.1 Consider the p-dimensional problem with $p \in \{2, 3\}$. Let $T = \frac{1}{2\beta}\hat{M}_g + KM^{-1}K$ and assume that Assumption 2.1 holds.
If $\beta > \frac{dD_g}{2C^2}h^3$, then

$$\kappa(T) \geq \frac{2\beta C^2}{D_g h^3};$$

otherwise

$$\kappa(T) \geq \frac{D_g h^3}{2\beta C^2}.$$

If $\beta > \frac{d_g D}{2c^2 h}$, then

$$\kappa(T) \leq \frac{D_g}{d_g} + \frac{2C^2\beta}{d d_g h^3};$$

otherwise

$$\kappa(T) \leq \frac{D D_g}{2c^2\beta h} + \frac{C^2 D}{c^2 d h^4}.$$

Consequently, if $\beta > \frac{d_g D}{2c^2 h}$, we will expect $\kappa(T)$ to be proportional to β and inversely proportional to h^3. If $\beta \ll \frac{dD_g}{2C^2}h^3$, then there exist constants c_0 and c_1 such that $c_0\beta^{-1}h^3 \leq \kappa(T) \leq c_1\beta^{-1}h^{-1}$. Finally, if $\frac{dD_g}{2C^2}h^3 \gg \beta \leq \frac{d_g D}{2c^2 h}$, there exist constants c_2 and c_3 such that $c_2\beta^{-1}h^3 \geq \kappa(T) \leq c_3 h^{-4}$.

In Figure 5, we plot the condition number of the Schur complement T with respect to β for the 2D Neumann boundary control with Target 1. Results are given for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$. As expected, if $\beta > \frac{d_g D}{2c^2 h}$, then $\kappa(T)$ is proportional to βh^{-3}. For $\beta \ll \frac{dD_g}{2C^2}h^3$, we find that the condition number closely follows the upper bound $c_1\beta^{-1}h^{-1}$. Finally, for intermediate values of β, the condition number satisfies $\kappa(T) \approx c_4 h^{-4}$ for some constant $c_4 \lesssim c_3$.

6 Conclusions

We have presented results about the spectral properties of the discretized systems that arise in Neumann boundary control problems: the PDE in the constraints is assumed to be the Poisson problem. The Neumann boundary 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 β but inversely proportional to h^4 for β smaller

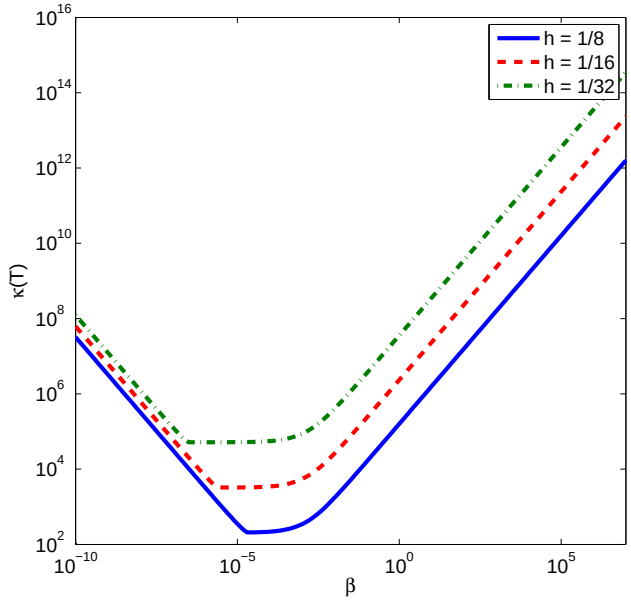

Figure 5.1: Condition number of the Schur complement T for the 2D Neumann boundary control with Target 1 and different values of β. Results are shown for $h = \frac{1}{8}$, $h = \frac{1}{16}$ and $h = \frac{1}{32}$.

than $c_1 h^3$, where h is the mesh size and c_1 is a constant independent of h and β; if $c_1 h^4 \ll \beta < c_2 h$, where c_2 is a constant independent of h and β, then the condition number will be bounded from above by a function that is inversely proportional to β and h^2; if $\beta \gg c_3 h^{-1}$, where c_3 is a constant independent of h and β, the condition number is bounded from above by a function that is proportional to β^2 but inversely proportional to h. Conversely, if $\hat{u}$ is only defined over a sub-domain of the overall problem, then we have two possibilities. If $\mathcal{A}$ is nonsingular when $\beta = 0$, then the spectral properties will resemble those for the case when the target $\hat{u}$ is defined over the whole of the domain; otherwise, the condition number is no longer bounded from above by a function that is independent of β when β is small and the behaviour is as for slightly larger values of β. 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 $\mathbf{g}$ and the relative error of $\mathbf{u}$ is proportional to β and not β^2 as would be expected using standard error bounds. However, if β is small, then the relative errors reflect the condition number of $\mathcal{A}$.

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 given constants c_4 and c_5, if $\beta \ll c_4 h^3$, then the condition number of the Schur complement is bounded from above by a function that is inversely proportional to β and h, and bounded from below by function that is inversely proportional to β and proportional tp h^3; if $\beta \gg c_5 h^{-1}$, then the condition number of the Schur complement is bounded from above by a function that is proportional to β and but inversely proportional to h^3. Hence, refining the mesh will result in a larger condition number.

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 [19, 22]. 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, [17, 21]. 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 Neumann boundary 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. S. Dollar, *Properties of linear systems in PDE-constrained optimization. part I: Distributed control*, Tech. Rep. RAL-TR-2009-017, Rutherford Appleton Laboratory, 2009.

[8] 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.

[9] I. Fried, *Bounds on the extremal eigenvalues of the finite element stiffness and mass matrices and their spectral condition number*, Journal of Sound and Vibration, 22 (1972), pp. 407–418.

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

[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] 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.

[16] 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.

[17] 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.

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

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

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

[21] 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.

[22] 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.

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

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

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