Operator preconditioning for a class of constrained optimal control problems

Anton Schiela*& Stefan Ulbrich †

May 15, 2012

Abstract

We propose and analyze two strategies for preconditioning linear operator equations that arise in PDE constrained optimal control in the framework of conjugate gradient methods. Our particular focus is on control or state constrained problems, where we consider the question of robustness with respect to critical parameters. We construct a preconditioner that yields favorable robustness properties with respect to critical parameters.

AMS MSC 2000: 65F08, 49M05, 65J10

Keywords: preconditioner, optimal control, control constraints, state constraints

1 Introduction

In this paper we are concerned with the solution of optimization problems, subject to partial differential equations and inequality constraints on the control and/or the state. Such problems can be considered as optimization problems in infinite dimensional function spaces, and in recent years, algorithms have been constructed which tackle these problems in function space. The common feature of these algorithms is that they can be formulated and analyzed in the infinite dimensional setting, and each step of such an algorithm requires the solution of an infinite dimensional problem. Taking, for example, Newton methods, this means that in each iteration a linear operator equation is solved. In general terms a perturbed saddle point problem of the form

$$\begin{pmatrix} M^*M & A^* \\ A & -CC^* \end{pmatrix} \begin{pmatrix} \delta y \\ \delta p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{1}$$

has to be solved in each Newton step (we will give a derivation and a precise functional analytic setting in the next sections). Of course, implementations have to deal with discretized versions of these subproblems, but have the conceptual advantage that the methods inherit much of the structure of the infinite dimensional problem.

In this paper we pursue this line of thought one step further and construct a preconditioned iterative solver for the linear systems that occur in PDE constrained optimization. In particular we consider two block preconditioners for the cg-method in function space,

*Supported by the DFG Research Center MATHEON "Mathematics for key technologies"

†Supported by the DFG within the Graduate School "Computational Engineering", the SPP 1253 "Optimization with PDE Constraints" and by SFB 666 and SFB 805

applied to this problem class. One of them is straightforward and already well known. The second is applicable under certain circumstances and yields increased robustness with respect to certain critical parameters, which may become small or large, if (1) comes from constrained optimal control problems.

Finally, we want to point out that our results are valid also in discretized settings, where the usually infinite dimensional spaces are replaced by finite dimensional (finite element) subspaces. Our analysis includes but does not require infinite dimensional spaces.

Preconditioning and multigrid for optimality systems in PDE constrained optimization is an active topic of research and there are several lines of research. Early attempts were made by Battermann et al [1, 2]. Borzi [3, 4] considers collective Gauss-Seidel preconditioners, while Zulehner et al [22, 27, 21] and Wathen et al [17, 7] propose and analyze block preconditioners for such systems. While cases without inequality constraints are well understood meanwhile, the case of control and/or state constraints is still mostly open. First approaches were taken by Herzog and Sachs [10] which observed lack of robustness of standard block preconditioners in particular for state constrained optimization problems and in [23]. In a very recent preprint [16] a preconditioner with favorable stability properties for state constrained problems was proposed, but the analysis does not provide useful estimates for the condition number. This preconditioner fits into our general framework, which works for control constraints and for state constraints.

2 Examples of optimal control problems

To clarify our abstract setting, let us consider a couple of examples for optimal control problems. For simplicity of presentation we consider linear quadratic problems here. Nonlinear problems can be solved iteratively via a Newton type algorithm, which requires the solution of a linear system of operator equations in each step.

Example 2.1 (Elliptic optimal control with control constraints)**.** Let us consider as an example the optimal control problem

$$\min \frac{1}{2}\|Ey - y_d\|^2_{L_2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L_2(\Omega)} \text{ s.t. } Ay - Bu = 0 \quad u \geq 0.$$

For elliptic optimal control on a bounded Lipschitz domain $\Omega \in \mathbb{R}^d$ with boundary Γ, we define A as follows

$$\begin{aligned} &A : H^1(\Omega) \to H^1(\Omega)^* \\ &\quad y \mapsto Ay \ : (Ay)(v) := a(y,v) := \int_\Omega \langle \kappa(x)\nabla y, \nabla v\rangle_{\mathbb{R}^d} + a_0(x)yv\,dx, \end{aligned} \tag{2}$$

where $\kappa(x) : \Omega \to \mathbb{R}^{d\times d}$ defines a symmetric bounded elliptic bilinear form, and $a_0 : \Omega \to \mathbb{R}$ is positive and bounded. The operator E_S is defined as the Sobolev embedding $H^1(\Omega) \hookrightarrow L_2(\Omega)$, while $B : L_2(\Omega) \to H^1(\Omega)^*$ is, in the case of distributed control defined via

$$(Bu)(v) = \int_\Omega u \cdot E_S v, dx.$$

More generally one could also consider a Sobolev space V such that $H^1_0(\Omega) \subset V \subset H^1(\Omega)$ to cover Dirichlet and mixed boundary conditions.

In the case of boundary control we have instead $B : L_2(\Gamma) \to H^1(\Omega)^*$ via

$$(Bu)(v) = \int_\Gamma u \cdot \tau(v), dx,$$

where τ is the trace operator.

It can be shown that the minimizer of this problem can be characterized by the following control reduced optimality system, where p is the so called *adjoint state* and the optimal control is given pointwise by $u = \max\{\alpha^{-1}B^*p, 0\}$:

$$\begin{aligned} 0 &= E^*(Ey - y_d) + A^*p \\ 0 &= Ay - B\max\{\alpha^{-1}B^*p, 0\}. \end{aligned} \tag{3}$$

This problem can be solved for example by a semi-smooth Newton method [11, 25], whose Jacobian matrix can be written as

$$\begin{pmatrix} E^*E & A^* \\ A & -B\alpha^{-1}\chi_{\mathcal{I}}(p)B^* \end{pmatrix}, \tag{4}$$

where $\chi_{\mathcal{I}}(p) = 0$ for $p > 0$ and $\chi_{\mathcal{I}}(p) = 1$ for $p \le 0$. In the case of bilateral constraints $\underline{u} \le u \le \overline{u}$ one obtains similar systems with $\max\{\alpha^{-1}p, 0\}$ replaced by $\mathrm{Proj}_{[\underline{u},\overline{u}]}(\alpha^{-1}p)$ and $\chi_{\mathcal{I}}(p) = 0$ for $\alpha^{-1}p \in]\underline{u}, \overline{u}[$ and $\chi_{\mathcal{I}}(p) = 1$ otherwise. This fits into our notational framework, if we set $M := E$ and $C := B\alpha^{-1/2}\chi_{\mathcal{I}}(p)$.

The reduced system formulation (3) has several advantages, compared to a classical KKT system, containing y, p, u and additional Lagrangian multipliers for the constraint $u \ge 0$. First of all, it is a system of two PDEs, and thus, the solutions of these system are contained in a smoother space than the corresponding right hand sides. This has fundamental consequences for the convergence theory of Newton's method, applied to this system [18]. Second, elimination of u makes an additional special treatment (such as barrier of penalty regularization) of the control constraints unnecessary, and the system can be solved directly. Third, only the smooth variables y and p have to be discretized, leading to optimal discretization schemes [13].

If we apply Galerkin's method to discretize (4), then the same definitions as above can be made on finite dimensional subspaces $Y_h \subset Y$ and $P_h \subset P$. Moreover, in the following derivations, no mesh-size parameter appears, so that the following results are automatically independent of the choice of the mesh.

Example 2.2 (Parabolic optimal control with control constraints)**.** Consider now the parabolic optimal control problem

$$\min \frac{1}{2}\|y - y_d\|^2_{L_2([0,T],L_2(\Omega))} + \frac{\alpha}{2}\|u\|^2_{L_2([0,T],L_2(\Omega))} \text{ s.t. } Ay - Bu = 0 \quad u \ge 0.$$

This case runs quite similarly to the elliptic case, with the only difference that

$$A : W([0,T]) \to L_2([0,T], H^1(\Omega))^*$$

is now a parabolic operator on

$$W([0,T]) = \{y \in L_2([0,T], H^1(\Omega)), y_t \in L_2([0,T], H^1(\Omega)^*)\},$$

defined via

$$(Ay)(v) = \int_{[0,T]} \left(y_t(v) + \int_\Omega \langle \nabla y, \kappa(t,x)\nabla v\rangle_{\mathbb{R}^d} + a_0(t,x)yv\,dx \right) dt.$$

Here $y_t(v)(t)$ is the application of $y_t \in H^1(\Omega)^*$ to $v \in H^1(\Omega)$, which yields an integrable function in time. For a more detailed description consider, e.g., [24].

In both of these examples the parameter $\alpha > 0$ appears as a finite (possibly small) but fixed value. The next two examples describe situations, where during algorithmic progress towards the solution such a parameter is adjusted, and makes the problem at hand more and more difficult as the solution is approached.

Example 2.3 (Regularized bang-bang control)**.** In some application so called bang-bang control is of interest:

$$\min \frac{1}{2}\|Ey - y_d\|^2_{L_2(\Omega)} \text{ s.t. } Ay - Bu = 0 \quad \underline{u} \leq u \leq \overline{u}.$$

Usually an optimal solution of such a problem almost everywhere takes either the value $\underline{u}$ or $\overline{u}$. A simple idea to solve this problem is to consider its regularized versions

$$\min \frac{1}{2}\|Ey - y_d\|^2_{L_2(\Omega)} + \frac{\alpha}{2}\|u\|_{L_2(\Omega)} \text{ s.t. } Ay - Bu = 0 \quad \underline{u} \leq u \leq \overline{u}.$$

and pass to the limit $\alpha \to 0$ (cf. e.g. [26]). Under certain assumptions on the adjoint state p it can be shown that the regularized solutions u_α tend to the solution of the original problem, and that the Lebesgue measure of the set $\mathcal{I} := \{x \in \Omega : \underline{u} < u_\alpha(x) < \overline{u}\}$ becomes smaller and smaller. In most cases it is observed that $\text{meas}(I) \leq C\alpha$.

In this setting it is desirable to obtain a preconditioner that is robust for $\alpha \to 0$.

Example 2.4 (Elliptic optimal control with state constraints)**.** In the case of state constraints, algorithms typically apply some kind of regularization technique [12, 19], usually based on classical approaches such as penalty or barrier methods. As an example, consider a classical penalty approach, also called "generalized Moreau-Yosida" regularization in this context, for the problem

$$\min \frac{1}{2}\|Ey - y_d\|^2_{L_2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L_2(\Omega)} \text{ s.t. } Ay - Bu = 0 \quad y \geq 0.$$

Penalization of the constraint $y \geq 0$ yields the problem

$$\min \frac{1}{2}\|Ey - y_d\|^2_{L_2(\Omega)} + \frac{\gamma}{2}\|\min\{y,0\}\|^2_{L_2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L_2(\Omega)} \text{ s.t. } Ay - Bu = 0,$$

which is often tackled by semi-smooth Newton methods. To approach the solution of the original problem, γ is driven towards $+\infty$ in a path-following method. In practice, the algorithm is terminated with values of γ in the range of 10^8 to 10^{12}. The presence of large γ affects the condition number of the problem severely, if no appropriate measures are taken.

Similarly, this problem can be tackled by barrier methods with a barrier functional $l(\cdot;\mu) :]0,\infty] \to \mathbb{R}$ parametrized by $\mu > 0$ such that $\lim_{t\to 0} l(t;\mu) = +\infty$. Path-following algorithms drive μ towards 0 to converge towards the original solution, and similar effects for the condition number occur.

In both cases computing a Newton step amounts in the solution of a linear system of the form

$$\begin{pmatrix} E^*b(x)E & A^* \\ A & -B\alpha^{-1}B^* \end{pmatrix} \begin{pmatrix} \delta y \\ \delta p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{5}$$

where $b(x)$ is either $1 + \gamma\chi_{y<0}(x)$ for the penalty method, or $1 + l''(y(x);\mu)$ for the barrier method.

Also this problem will fit into our theoretical framework, if we set $M := \sqrt{b}E$ and $C^* := \alpha^{-1/2}B^*$. Clearly, also combinations of control and state constraints can be treated.

3 Conjugate gradients in function space

In this paper we concentrate for simplicity on the solution of (1) via a Schur complement approach and a conjugate gradient method. Several alternatives have been proposed in the literature, such as preconditioned MINRES [17] and a Bramble-Pasciak cg-method [10]. We are certain, that our ideas apply also to these settings.

On a Hilbert space X consider a convex, quadratic functional

$$\psi(x) := b(x) + \frac{1}{2}\langle x, x\rangle_K,$$

with $b \in X^*$ and $\langle\cdot,\cdot\rangle_K$ a scalar product. Further, let a different scalar product $\langle\cdot,\cdot\rangle_Q$ (a preconditioner) be given. Denote by $\nabla_Q\psi(x)$ the Q-gradient of ψ, i.e. $\nabla_Q\psi(x)$ satisfies

$$\langle\nabla_Q\psi(x), w\rangle_Q = \psi'(x)(w) = b(w) + \langle x, w\rangle_K \quad \forall w \in X.$$

Then the method of conjugate gradients can be written as follows (cf. also [9]):

Algorithm 3.1. (preconditioned cg in function space)

```
$x_0$ given, $d_0 := -\nabla_Q\psi(x_0)$

$k = 0, 1, 2, \dots$

    $x_{k+1} = x_k - \frac{\psi'(x_k)d_k}{\langle d_k, d_k\rangle_K} d_k$   (exact linesearch along $d_k$)
    $g_{k+1} = -\nabla_Q\psi(x_{k+1})$   (direction of steepest descent w.r.t $\langle\cdot,\cdot\rangle_Q$)
    $d_{k+1} = g_{k+1} - \frac{\langle g_{k+1}, d_k\rangle_K}{\langle d_k, d_k\rangle_K} d_k$   (orthogonalization w.r.t. $\langle\cdot,\cdot\rangle_K$)
```

It is well known that speed of convergence of the cg-method depends on the condition number κ_Q of $\langle\cdot,\cdot\rangle_K$ with respect to $\langle\cdot,\cdot\rangle_Q$. It can be defined as follows. If the following (sharp) estimates hold,

$$m_Q\langle v, v\rangle_Q \le \langle v, v\rangle_K \le M_Q\langle v, v\rangle_Q, \tag{6}$$

then the condition number is given by $\kappa_Q := M_Q/m_Q$. Then, if x_* denotes the minimizer of ψ, we have the estimate

$$\|x_k - x_*\|_K \le 2\left(\frac{\sqrt{\kappa_Q}-1}{\sqrt{\kappa_Q}+1}\right)^k \|x_0 - x_*\|_K \tag{7}$$

and the number of iterations to reach a certain accuracy is proportional to $\sqrt{\kappa_Q}$ (cf. e.g. [6, Sec. 5.3.2]). Thus, it is crucial to find a good preconditioner Q that renders κ_Q small.

3.1 The saddle point system as a convex minimization problem

As already mentioned in the introduction, application of Newton's method to the reduced Kuhn-Tucker conditions requires the solution of a system of the form:

$$\begin{pmatrix} M^*M & A^* \\ A & -CC^* \end{pmatrix}\begin{pmatrix} \delta y \\ \delta p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}. \tag{8}$$

Let us fix the theoretical framework for this system. The following abstract and very basic assumptions will be used throughout the paper.

Assumption 3.2. (Basic Assumptions)

(i) Assume that Y and P are reflexive Banach spaces, and U and H are Hilbert spaces.

Further, as a matter of notation, we use $x^*(x)$ for dual pairings, while $\langle x_1, x_2\rangle_X$ denotes a Hilbert space scalar product.

(ii) Let $A : Y \to P^*$ be an isomorphism, which implies that $A^* : P \to Y^*$ is an isomorphism as well.

Let further $C : U \to P^*$ be a continuous operator, and $M : Y \to H$ a continuous operator with dense range.

The adjoint $C^* : P \to U$ is defined via

$$\langle C^* p, u\rangle_U = (Cu)(p) \quad \forall u \in U$$

and continuous as well. The adjoint $M^* : H \to Y^*$, defined via

$$(M^* h)(y) = \langle My, h\rangle_H \quad \forall y \in Y$$

is continuous and injective.

Thus, we use the definition of adjoints in Hilbert space, whenever applicable.

Then our system of equations is just another way of writing down the following weak form:

$$\begin{aligned} \langle M\delta y, Mv\rangle_H + (Av)(\delta p) &= f(v) \qquad \forall v \in Y \\ (A\delta y)(w) - \langle C^*\delta p, C^* w\rangle_U &= g(w) \qquad \forall w \in P. \end{aligned}$$

Density of $\operatorname{ran} M$ in H implies injectivity of $M^* : H \to Y^*$, so that $M^* : H \to \operatorname{ran} M^*$ can be considered as a bijective operator with inverse M^{-*}. Hence, if we define the new space

$$D_K := A^{-*}(\operatorname{ran} M^*) = \{p \in P : A^* p \in \operatorname{ran} M^*\} \subset P,$$

then $M^{-*}A^* p$ is well defined for all $p \in D_K$. Thus, on D_K the following bilinear form is well defined:

$$\langle v, w\rangle_K := \langle M^{-*}A^* v, M^{-*}A^* w\rangle_H + \langle C^* v, C^* w\rangle_U.$$

So, we can consider the following minimization problem for $b \in D_K^*$,

$$\min_{p \in D_K} \frac{1}{2}\langle p, p\rangle_K + b(p). \tag{9}$$

Remark 3.3. The invertibility requirement for M^* can equivalently be replaced by a similar condition on C. In this case the following analysis can be carried out completely with the roles of C and M^* switched.

Example 3.4. In Example 2.1 M is the Sobolev embedding $H^1(\Omega) \hookrightarrow L_2(\Omega)$, such that for an H^2-regular problem we get $D_K = A^{-*}(L_2(\Omega)) = H^2(\Omega)$.

If M is the trace operator $H^1(\Omega) \hookrightarrow L_2(\Gamma)$, which corresponds to boundary observation, then D_K can be computed as the set of all functions that correspond to solutions of the following problem in strong form (cf. (25)): $\mathcal{A}\varphi = 0$ on Ω with inhomogeneous Neumann boundary conditions $\partial_{\kappa\nu}\varphi = g$ on Γ for some $g \in L_2(\Gamma)$.

Lemma 3.5. *The minimization problem* (9) *has a unique solution in* D_K *for any* $b \in D_K^*$. *Moreover, for any right hand sides* $f \in Y^*$ *and* $g \in P^*$ *the system* (8) *has a solution* $(\delta y, \delta p) \in Y \times P$. *It can be computed from the solution* $\tilde{p}$ *of*

$$\min_{p \in D_K} \frac{1}{2}\langle p, p\rangle_K + g(p) + \langle C^* A^{-*} f, C^* w\rangle_U.$$

via

$$\begin{aligned} \delta p &= \tilde{p} + A^{-*} f \ \text{ in } Y^* \\ \delta y &= A^{-1}(g + CC^* \delta p) \ \text{ in } P^*. \end{aligned} \tag{10}$$

Proof. Since M^* is continuous, and A^* is an isomorphism, the norm $\|\cdot\|_{D_K}$, defined by $\|v\|_{D_K} := \|M^{-*}A^*v\|_H$ is a stronger norm than $\|\cdot\|_P$, and thus renders $D_K \subset P$ a Hilbert space. Hence, by continuity of C^* it is easy to see that our functional is continuous, strictly convex and coercive on the Hilbert space $(D_K, \|\cdot\|_{D_K})$. Thus, a unique minimizer of eqrefeq:min exists, and its first order optimality conditions read

$$\langle M^{-*}A^*\tilde{p}, M^{-*}A^*w\rangle_H + \langle C^*\tilde{p}, C^*w\rangle_U + b(w) = 0 \quad \forall w \in D_K.$$

Setting $b(p) := g(p) + \langle C^* A^{-*} f, C^* p\rangle_U$ this yields

$$\langle M^{-*}A^*\tilde{p}, M^{-*}A^*w\rangle_H + \langle C^*\tilde{p}, C^*w\rangle_U + g(w) + \langle C^*A^{-*}f, C^*w\rangle_U = 0 \quad \forall w \in D_K.$$

By definition, δy and δp solve the second row of (8), so it remains to show that they solve the first row. Inserting (10) we conclude via $A\delta y = CC^*\tilde{p} + g + CC^*A^{-*}f$ that

$$\langle M^{-*}(A^*\delta p - f), M^{-*}A^*w\rangle_H + (A\delta y)(w) = 0 \quad \forall w \in D_K. \tag{11}$$

For arbitrary $v \in Y$ define w as the solution of the equation $A^*w = M^*Mv$, or more explicitly

$$(A\eta)(w) = \langle M\eta, Mv\rangle_H \quad \forall \eta \in Y.$$

By definition, $A^*w \in \operatorname{ran} M^*$ and thus $w \in D_K$. Then in particular $(A\delta y)(w) = \langle M\delta y, Mv\rangle_H$, and $M^{-*}A^*w = Mv$, and we conclude from (11)

$$\langle M^{-*}(A^*\delta p - f), Mv\rangle_H + \langle M\delta y, Mv\rangle_H = 0 \quad \forall v \in Y$$

which yields

$$(A^*\delta p - f)(v) + (M^*M\delta y)(v) = 0 \quad \forall v \in Y$$

and thus, in short, the first row of (8). □

Hence, if we assume that our original system (8) is uniquely solvable, we can find its solution by solving (9) for $\tilde{p}$ and then computing δp and δy sequentially.

In the remainder of the paper we consider application of Algorithm 3.1 to (9) and construct and analyze bilinear forms $\langle\cdot,\cdot\rangle_Q$ on D_K. We will establish estimates of the form (6) such that κ_Q is small. In particular we want to avoid that κ_Q depends strongly on certain critical parameters, e.g. α in Example 2.3 or γ in Example 2.4, that arise in optimal control problems.

4 Two strategies for operator preconditioning

In the following we shall derive and justify some operator preconditioners for solving the minimization problem (9) with the conjugate gradient method.

4.1 Preconditioning via the pure differential operators

Our first preconditioner, which is similar to the ones proposed in [10], is defined via the first part of $\langle \cdot, \cdot \rangle_K$:

$$(Q_0 v)(w) := \langle v, w \rangle_{Q_0} := \langle M^{-*} A^* v, M^{-*} A^* w \rangle_H \quad \forall w \in D_K. \tag{12}$$

Its inverse can be applied to a residual vector r by computing

$$v := {Q_0}^{-1} r = A^{-*} M^* M A^{-1} r.$$

This involves one solve of the state equation and one solve of the adjoint equation. Clearly we have $A^* v \in \operatorname{ran} M^*$, hence $v \in D_K$.

Lemma 4.1. *Assume that*

$$\gamma_{Q_0} := \sup_v \frac{\langle C^* v, C^* v \rangle_U}{\langle v, v \rangle_{Q_0}} \tag{13}$$

is finite. Then

$$\langle v, v \rangle_{Q_0} \leq \langle v, v \rangle_K \leq (1 + \gamma_{Q_0}) \langle v, v \rangle_{Q_0}, \tag{14}$$

and thus the condition number κ of K, relative to Q_0 is bounded by

$$\kappa_{Q_0} \leq (1 + \gamma_{Q_0}).$$

Proof. The left part of (14) follows simply from the positive semi-definiteness of $\langle C^* \cdot, C^* \cdot \rangle_U$. Further, since $\langle C^* v, C^* v \rangle_U \leq \gamma_{Q_0} \langle v, v \rangle_{Q_0}$, we obtain the right part. □

4.2 A preconditioner with increased robustness

Our next preconditioner exploits the positive definiteness of CC^* to improve our condition number estimate. In order to render it well defined, we have to impose the following assumption, which restricts the class of problems to be tackled:

Assumption 4.2. (Compatibility Assumption) Assume that there is a continuous mapping

$$I : H \to U.$$

For simplicity, assume that $\|I\| \leq 1$.

For the Hilbert space adjoint $I^* : U \to H$, defined by

$$\langle I^* u, h \rangle_H = \langle u, Ih \rangle_U$$

we note that in general $\|I^*\| = \|I\| \leq 1$.

Example 4.3. Let Ω_H and Ω_U be two subsets of $\mathbb{R}^d$ of non-zero measure. Define $H = L_2(\Omega_H)$ and $U = L_2(\Omega_U)$. Then the mapping $I : H \to U$ can be defined by restriction of $h \in H$ to $\Omega_H \cap \Omega_U$, followed by extension by zero onto Ω_U. In turn, $I^* : U \to H$ is the restriction to $\Omega_H \cap \Omega_U$ and extension by zero to Ω_H, namely for all $h \in L_2(\Omega_H), u \in L_2(\Omega_U)$

$$\langle u, Ih \rangle_{L_2(\Omega_U)} = \int_{\Omega_U} u(Ih)\, dx = \int_{\Omega_H \cap \Omega_U} uh\, dx = \int_{\Omega_H} (I^* u) h\, dx = \langle I^* u, h \rangle_{L_2(\Omega_H)}.$$

The extreme cases are $\Omega_H \cap \Omega_U = \emptyset$, then $I \equiv 0$ and $\Omega_H = \Omega_U$, then $I = Id$.

By our assumptions the composition CIM is well defined, and thus also its adjoint $M^*I^*C^*$ via the relations

$$(CIMy)(v) = \langle IMy, C^*v\rangle_U = \langle My, I^*C^*v\rangle_H = (M^*I^*C^*v)(y).$$

Definition 4.4. Define the preconditioner $Q : D_K \to D_K^*$ by:

$$(Qv)(w) := \langle v, w\rangle_Q := \langle M^{-*}(A + CIM)^*v, M^{-*}(A + CIM)^*w\rangle_H \;\forall w \in D_K. \tag{15}$$

The application of the preconditioner $Q^{-1}r$ proceeds in three steps. For given r one has to solve the modified PDE

$$(A + CIM)x = r.$$

Then one has to compute

$$y := M^*Mx.$$

Finally, one has to solve

$$(A + CIM)^*v = y.$$

so that $Q^{-1}r := v$. Since $y \in \operatorname{ran} M^*$ and $M^*I^*C^*v \in \operatorname{ran} M^*$ we conclude that $A^*v \in \operatorname{ran} M^*$, hence $v \in D_K$.

An equivalent preconditioner has been proposed and analyzed recently for the unconstrained case of distributed control by [27, 21]. For the case of regularized state constraints an equivalent preconditioner has been proposed recently and independently in [16], but no useful estimates for the condition number were derived.

Remark 4.5. Already at this point we can predict the main features of this type of preconditioner. In contrast to Q_0 it also includes the operator C in its formulation. Hence, more information of the problem enters into the construction of the preconditioner. We will see that this leads to a significant improvement of condition numbers, in cases where Q can be applied.

However, we also observe the main limitations of our approach. The composition CIM has to be non-zero, otherwise $Q = Q_0$. Here our main focus is restricted to simple mappings I, as defined in Example 4.3, because CIM has to be simple enough to make the equation $(A + CIM)x = r$ solvable at low cost.

The following lemma plays a pivotal role in our analysis:

Lemma 4.6. *Assume that the following quantity is finite:*

$$\gamma_Q := \sup_v \frac{\langle C^*v, C^*v\rangle_U}{\langle v, v\rangle_Q}. \tag{16}$$

Then we have the following estimates:

$$\frac{1}{2}\langle v, v\rangle_Q \le \langle v, v\rangle_K \le (2 + 3\gamma_Q)\langle v, v\rangle_Q, \tag{17}$$

and thus the condition number κ of K, relative to Q is bounded by

$$\kappa_Q \le 4 + 6\gamma_Q.$$

Proof. For the proof we recall the parallelogram law in Hilbert spaces:

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2),$$

which implies that each summand on the left hand side can be estimated from above by the right hand side.

We use this estimate to compute

$$\begin{aligned}\langle v,v\rangle_Q &= \|M^{-*}A^*v + I^*C^*v\|_H^2 \le 2(\|M^{-*}A^*v\|_H^2 + \|I^*C^*v\|_H^2)\\ &\le 2(\|M^{-*}A^*v\|_H^2 + \|C^*v\|_U^2) = 2\langle v,v\rangle_K.\end{aligned}$$

Let us consider the opposite direction. For given $x \in D_K$ define

$$v := (M^{-*}A^* + I^*C^*)x$$

so that $\langle v,v\rangle_H = \langle x,x\rangle_Q$. Then, using the definition of γ_Q, we can compute:

$$\begin{aligned}\langle x,x\rangle_K &= \|M^{-*}A^*x\|_H^2 + \|C^*x\|_U^2 = \|v - I^*C^*x\|_H^2 + \|C^*x\|_U^2\\ &\le 2\|v\|_H^2 + 2\|I^*C^*x\|_H^2 + \|C^*x\|_U^2 \le 2\langle x,x\rangle_Q + 3\|C^*x\|_U^2 \qquad (18)\\ &\le (2+3\gamma_Q)\langle x,x\rangle_Q.\end{aligned}$$

□

With some additional effort one can refine this estimate slightly. However, for us, the asymptotics $\kappa_Q = O(\gamma_Q)$ for large γ_Q is the main point of interest. This relation cannot be improved substantially, as we will briefly explain.

Consider the left bound first. Since usually A^* is a differential operator, we may assume that there is a sequence v_k, such that $\|M^{-*}A^*v_k\|_H = 1$, while $\|C^*v_k\|_U \to 0$. Then the terms in K and Q containing C^* can be neglected, and we obtain $\langle v_k,v_k\rangle_Q/\langle v_k,v_k\rangle_K \to 1$.

As for the right bound in (17), assume that $(1-\varepsilon)\gamma_Q$ is attained by some x in (16). Then instead of (18) we can compute

$$\langle x,x\rangle_K = \|M^{-*}A^*x\|_H^2 + \|C^*x\|_U^2 \ge \|C^*x\|_U^2 \ge (1-\varepsilon)\gamma_Q\langle x,x\rangle_Q.$$

Hence, taking both estimates together we obtain a lower bound for the condition number, given by

$$\kappa_Q \ge \gamma_Q. \qquad (19)$$

In order to motivate this estimate further, consider the auxiliary quantity r, defined by

$$(A + CIM)^*x = r,$$

which means in turn that $x(r)$ is the solution of a partial differential equation with right hand side r. Then γ_Q can be written as follows:

$$\gamma_Q = \sup_r \frac{\langle C^*x(r), C^*x(r)\rangle_U}{\langle M^{-*}r, M^{-*}r\rangle_H}.$$

Hence, our task will be to establish estimates of the form

$$\|C^*x\|_U \le c(A, CIM)\|M^{-*}r\|_H$$

on the solution v of the above PDE in terms of r. This reduces the estimation of the condition number of our saddle-point system to an estimate for a PDE solution. We will use this technique in Sections 6 and 7 below, where particular structure of the PDE at hand is used. In the following and in Section 5, we keep arguing purely in terms of functional analytic estimates.

4.3 Application to PDE optimization problems

In this section we provide estimates for γ_Q in the setting of elliptic and parabolic problems. Of course other settings are conceivable, as well.

Elliptic problems. Consider the preconditioner Q from (15) for the elliptic case, where $A : H^1(\Omega) \to H^1(\Omega)^*$ may, for example be defined via an elliptic bilinear form $a(\cdot,\cdot)$ as in (2) from Example 2.1.

Lemma 4.7. *For the elliptic equation* (2) *we have the estimate*

$$\gamma_Q \le c \sup_v \frac{\langle C^*v, C^*v\rangle_U \langle Mv, Mv\rangle_H}{(a(v,v) + \langle C^*v, IMv\rangle_U)^2}$$

Proof. Let $v \in Y = P = H^1(\Omega)$, so that $(A^* + M^*I^*C^*)v \in Y^*$. We start with the Cauchy-Schwarz inequality:

$$\begin{aligned}((A^* + M^*I^*C^*)v)(v) &= \langle M^{-*}(A^* + M^*I^*C^*)v, Mv\rangle_H \\ &\le \|M^{-*}(A^* + M^*I^*C^*)v\|_H \|Mv\|_H \\ &= \sqrt{\langle v, v\rangle_Q \langle Mv, Mv\rangle_H}.\end{aligned}$$

Hence, we may estimate (using $(A^*v)(v) = a(v,v)$)

$$\begin{aligned}\gamma_Q = \sup_v \frac{\langle C^*v, C^*v\rangle_U \langle Mv, Mv\rangle_H}{\langle v, v\rangle_Q \langle Mv, Mv\rangle_H} &\le \sup_v \frac{\langle C^*v, C^*v\rangle_U \langle Mv, Mv\rangle_H}{\left(((A^* + M^*I^*C^*)v)(v)\right)^2} \\ = \sup_v \frac{\langle C^*v, C^*v\rangle_U \langle Mv, Mv\rangle_H}{(a(v,v) + \langle C^*v, IMv\rangle_U)^2}.\end{aligned}$$

□

Parabolic problems. To establish an estimate for γ_Q in the parabolic case, as in Example 2.2, we have to modify our proof slightly in order to cope with the non-symmetry of its differential operator, which reads

$$(Ay)(v) = \int_{[0,T]} \left(\langle y_t, v\rangle + \int_\Omega \langle \nabla y, \kappa(t,x)\nabla v\rangle_{\mathbb{R}^d} + a_0(t,x) y v \, dx \right) dt, \tag{20}$$

together with initial conditions $y(0) = 0$. We have to employ a special scalar and duality product. For $\omega > 0$ we set

$$\langle v, w\rangle_{e^{-\omega t}} := \int_{[0,T]} e^{-\omega t} \langle v(t), w(t)\rangle_{L_2(\Omega)} \, dt,$$

which induces an equivalent norm $e^{-\omega T}\|\cdot\|_{L_2([0,T]\times\Omega)} \le \|\cdot\|_{e^{-\omega t}} \le \|\cdot\|_{L_2([0,T]\times\Omega)}$. Similarly, for a Banach space V we write the duality product on $L_2([0,T], V^*) \times L_2([0,T], V)$

$$\langle v^*, v\rangle_{e^{-\omega t}} := \int_{[0,T]} e^{-\omega t} (v^*(t)(v(t))) \, dt,$$

Our motivation is that A^* is positive definite with respect to this scalar product, as long as ω is chosen sufficiently large.

Lemma 4.8. *Let A be defined as in* (20)*. Assume that $\langle M^*v, w\rangle_{e^{-\omega t}} = \langle v, Mw\rangle_{e^{-\omega t}}$. Then A^* is positive definite w.r.t. $\langle\cdot,\cdot\rangle_{e^{-\omega t}}$, and we obtain the following condition number:*

$$\gamma_Q \le c(T) \sup_v \frac{\langle C^*v, C^*v\rangle_U \langle Mv, Mv\rangle_{e^{-\omega t}}}{(\langle A^*v, v\rangle_{e^{-\omega t}} + \langle I^*C^*v, Mv\rangle_{e^{-\omega t}})^2}$$

Proof. First, we show that A^* is positive definite w.r.t. the scalar product $\langle\cdot,\cdot\rangle_{e^{-\omega t}}$. Inserting $w = e^{-\omega t}v$ into the formula of partial integration (cf. e.g. [8, Satz 1.17])

$$\langle v(T), w(T)\rangle - \langle v(0), w(0)\rangle = \int_{[0,T]} \langle v_t(t), w(t)\rangle + \langle w_t(t), v(t)\rangle \, dt$$

and taking into account our restriction $v(0) = w(0) = 0$ we infer after a short computation

$$\langle v_t, v\rangle_{e^{-\omega t}} = \frac{1}{2}\left(e^{-\omega T}\|v(T)\|^2 + \omega\langle v, v\rangle_{e^{-\omega t}}\right).$$

As for the remaining part of A^* we have

$$\int_{[0,T]}\int_\Omega \langle\nabla v, \kappa(t,x)\nabla v\rangle_{\mathbb{R}^d} + a(t,x)v^2 \, dx e^{-\omega t} \, dt \ge 0,$$

and hence

$$\langle A^*v, v\rangle_{e^{-\omega t}} \ge \frac{\omega}{2}\langle v, v\rangle_{e^{-\omega t}}$$

From this point our proof runs in parallel to the elliptic case. Similar as before, let $v \in Y = W([0,T])$. Then also $v \in P = L_2([0,T], H^1(\Omega))$, such that $(A^* + M^*I^*C^*)v \in Y^*$. Thus, we can use the Cauchy-Schwarz inequality:

$$\begin{aligned}
&\langle (A^* + M^*I^*C^*)v, v\rangle^2_{e^{-\omega t}} \\
&\quad \le \langle M^{-*}(A^* + M^*I^*C^*)v, M^{-*}(A^* + M^*I^*C^*)v\rangle_{e^{-\omega t}}\langle Mv, Mv\rangle_{e^{-\omega t}} \\
&\quad \le \langle v, v\rangle_Q \langle Mv, Mv\rangle_{e^{-\omega t}}.
\end{aligned}$$

Hence, we may estimate, also as before:

$$\begin{aligned}
\gamma_Q &= \sup_v \frac{\langle C^*v, C^*v\rangle_U \langle Mv, Mv\rangle_{e^{-\omega t}}}{\langle v, v\rangle_Q \langle Mv, Mv\rangle_{e^{-\omega t}}} \le \sup_v \frac{\langle C^*v, C^*v\rangle_U \langle Mv, Mv\rangle_{e^{-\omega t}}}{\langle (A^* + M^*I^*C^*)v, v\rangle^2_{e^{-\omega t}}} \\
&= \sup_v \frac{\langle C^*v, C^*v\rangle_U \langle Mv, Mv\rangle_{e^{-\omega t}}}{(\langle A^*v, v\rangle_{e^{-\omega t}} + \langle I^*C^*v, Mv\rangle_{e^{-\omega t}})^2}.
\end{aligned}$$

□

Our assumption $\langle M^*v, w\rangle_{e^{-\omega t}} = \langle v, Mw\rangle_{e^{-\omega t}}$ can easily be verified, if for example M can be written as $(Mv)(t) = M(t)v(t)$, where $M(t)$ depends on the "slice" $v(t)$ only.

The dependence of γ_Q on the interval length T can be worked out to be proportional to T by choosing ω optimally.

We conclude this section with the observation that both elliptic and parabolic problems admit the common estimate

$$\gamma_Q \le c\frac{\langle C^*v, C^*v\rangle_U \langle Mv, Mv\rangle}{(\langle v, v\rangle + \langle C^*v, IMv\rangle)^2}, \tag{21}$$

where the notation $\langle\cdot,\cdot\rangle$ stands for one of the scalar products $\langle\cdot,\cdot\rangle_H$ and $\langle\cdot,\cdot\rangle_{e^{-\omega t}}$ for the elliptic and parabolic case, respectively. The estimate holds, because A^* is positive definite in the corresponding scalar product, i.e.,

$$(A^*v)(v) \ge c_A\langle v, v\rangle.$$

5 Applications to concrete problems

In this section we will discuss a couple of examples for which our preconditioning strategy can (or cannot) be applied effectively. This should clarify the advantages, but also the limitations of the preconditioner Q from (15). Included are problems with control constraints, and also with state constraints.

The following bounds hold in a quite general setting. Under stronger assumptions they can be refined, as shown in Section 6 and Section 7.

5.1 Distributed control problems with control bounds

Let us consider as an example the optimal control problem

$$\min \frac{1}{2}\|y-y_d\|^2_{L_2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L_2(\Omega)} \text{ s.t. } Ay - Bu = 0 \quad u \geq 0$$

As explained in Example 2.1 this problem can be solved by a semi-smooth Newton method, which leads to systems of the form (4). Then, $Y = P = H^1(\Omega)$ and $U = H = L_2(\Omega)$. Further, $M : H^1(\Omega) \to L_2(\Omega)$ is the Sobolev embedding E_S, $B = E_S^* : L_2(\Omega) \to H^1(\Omega)^*$, and $C^* : H^1(\Omega) \to L_2(\Omega)$ is defined as $C^* = \alpha^{-1/2}\chi_{\mathcal{I}}(p)E$. Finally, the mapping $I : H \to U$ is just the identity in $L_2(\Omega)$.

As a parabolic example we consider

$$\min \frac{1}{2}\|y-y_d\|^2_{L_2([0,T],L_2(\Omega))} + \frac{\alpha}{2}\|u\|^2_{L_2([0,T],L_2(\Omega))} \text{ s.t. } Ay - Bu = 0 \quad u \geq 0.$$

Here, $Y = W([0,T])$, $P = L_2([0,T],H^1)$ and $U = H = L_2([0,T],L_2(\Omega))$. Similar to before, $M : Y \to H$ is the Sobolev embedding $W([0,T]) \hookrightarrow L_2([0,T],L_2(\Omega))$. Using the Sobolev embedding $E_S : L_2([0,T],H^1(\Omega)) \to L_2([0,T],L_2(\Omega))$ we can define $B = E_S^* : U \to P^*$, and $C^* : P \to U$ as $C^* = \alpha^{-1/2}\chi_{\mathcal{I}}(p)E_S$. Also here, the mapping $I : H \to U$ is just the identity on $L_2([0,T],L_2(\Omega))$.

With our analysis from the previous section we obtain the following results for our preconditioners:

Proposition 5.1. *Consider the preconditioner Q_0 from* (12) *applied to the block operator* (4)*. Then we obtain the following condition number:*

$$\kappa_{Q_0} \leq 1 + c\alpha^{-1}.$$

Proof. In view of the lemmas in the previous section, we have to provide estimates for γ_{Q_0}. For the numerator we can compute

$$\langle C^*v, C^*v\rangle \leq \|\max\{\alpha^{-1},0\}\|_\infty \|v\|^2_{L_2}$$

and the denominator yields $\langle A^*v, A^*v\rangle \geq c\|v\|^2_{L_2}$ by continuity of $(A^*)^{-1} : L_2(\Omega) \to L_2(\Omega)$. So $\gamma_{Q_0} \leq c\alpha^{-1}$, which yields the desired result for κ_{Q_0} via Lemma 4.1. □

Now we consider our preconditioner Q for the elliptic and parabolic case.

Proposition 5.2. *Consider the preconditioner Q from* (15) *applied to the block operator* (4)*. Consider the elliptic or the parabolic operator from the last section. Then*

$$\kappa_Q \leq c(1 + \alpha^{-1/2})$$

Proof. Let $\langle\cdot,\cdot\rangle$ be one of $\langle\cdot,\cdot\rangle_{L_2(\Omega)}$ (for the elliptic case) or $\langle\cdot,\cdot\rangle_{e^{-\omega t}}$ (for the parabolic case). Then, from (21) we estimate (taking into account $Mv = v$).

$$\gamma_Q \le c \sup_v \frac{\langle C^*v, C^*v\rangle_U \langle v, v\rangle}{(\langle v, v\rangle + \langle C^*v, v\rangle)^2} \le c \sup_v \frac{\langle C^*v, C^*v\rangle \langle v, v\rangle}{2\langle v, v\rangle \langle C^*v, v\rangle}.$$

The last inequality follows from the general relation $a^2 + b^2 \ge 2ab$. Moreover, in both cases, $\langle C^*v, C^*v\rangle_U \le c\langle C^*v, C^*v\rangle$. By definition of C we obtain

$$\langle C^*v, C^*v\rangle_U \le c\alpha^{-1/2}\langle C^*v, v\rangle.$$

Hence,

$$\gamma_Q \le c\alpha^{-1/2},$$

and thus by Lemma 4.6 $\kappa_Q \le c(1 + \alpha^{-1/2})$. □

Thus, we have obtained $\kappa_Q \sim \sqrt{\kappa_{Q_0}}$ which already yields a considerable gain of efficiency for a cg method via (7).

Corollary 5.3. *In the unconstrained case, we obtain the α independent bound*

$$\kappa_Q \le c$$

Proof. Just as before, we compute

$$\gamma_Q = \sup_v \frac{\langle C^*v, C^*v\rangle \langle v, v\rangle}{(c\langle v, v\rangle + \langle C^*v, v\rangle)^2} \le c \sup_v \frac{\alpha^{-1}\langle v, v\rangle^2}{\langle C^*v, v\rangle^2} = c \sup_v \frac{\alpha^{-1}\langle v, v\rangle^2}{(\alpha^{-1/2}\langle v, v\rangle)^2}.$$

Hence $\gamma_Q \le c$. □

5.2 Disjoint control and observation regions

Consider the problem (for simplicity, let A be the elliptic operator from (2))

$$\min \frac{1}{2}\|y - y_d\|^2_{L_2(\Omega_H)} + \frac{\alpha}{2}\|u\|^2_{L_2(\Omega_U)} \text{ s.t. } Ay - Bu = 0,$$

where $B : L_2(\Omega_U) \to P^*$ is a continuous mapping, and Ω_H and Ω_U are subsets of $\overline{\Omega}$, such that their intersection is a set of measure zero in both $H := L_2(\Omega_H)$ and $U := L_2(\Omega_U)$. An important special case is boundary control and observation in the domain, i.e., $\Omega_H = \Omega$ and $\Omega_U = \Gamma = \partial\Omega$. With the obvious choice $I : L_2(\Omega) \to L_2(\Gamma)$ via $I \equiv 0$ this problem fits into our theoretical framework, and we obtain $CIM \equiv 0$. In this case $Q \equiv Q_0$ and we have not gained anything in terms of condition numbers.

Nevertheless, our block elimination can be applied, even if, at first sight, M^* is not invertible. Since the cg iterates remain in D_K, which is here $\{v : A^*v \in L_2(\Omega_H)\}$, M^{-*} remains well defined in the context of the cg iteration.

5.3 Distributed control with state constraints

In the case of state constrained optimal control problems, algorithms often employ a path-following scheme, which leads to a block M that is very ill conditioned towards the end of the algorithm. Both our preconditioners can be applied in this case. Moreover, it can be shown that usually the preconditioner Q is significantly more robust than Q_0 with respect to the path-following parameters.

As elaborated in Example 2.4 in state constrained problems with distributed control linear systems of the form

$$\begin{pmatrix} E^*b(x)E & A^* \\ A & -B\alpha^{-1}B^* \end{pmatrix} \begin{pmatrix} \delta y \\ \delta p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{22}$$

have to be solved, where $E = B^*$ is the Sobolev embedding, in the case of distributed control, where $b(x) > 0$, and typically $\|b\|_\infty$ is very large, tending to infinity towards the end of the algorithm. Here we can define $H = U = L_2(\Omega)$ and $M = \sqrt{b(x)}E$ and $C^* = \alpha^{-1/2}B$ and again $I = Id$.

Proposition 5.4. *Assume that $p \in L_\infty$. Consider the preconditioner Q_0 from* (12) *applied to the operator* (22)*. Then we obtain the following condition number:*

$$\kappa_{Q_0} \leq 1 + c\|b\|_\infty \alpha^{-1}$$

Proof. In view of the lemmas in the previous section, we have to provide estimates for γ_{Q_0}. For the numerator we can compute

$$\langle C^*v, C^*v\rangle \leq \|\max\{\alpha^{-1}, 0\}\|_\infty \|v\|^2_{L_2}$$

and the denominator yields $\langle M^{-*}A^*v, M^{-*}A^*v\rangle \geq c\|b\|^{-1}_\infty \|v\|^2_{L_2(\Omega)}$ by continuity of $(A^*)^{-1} : L_2(\Omega) \to L_2(\Omega)$. □

The following result crucially depends on the assumption that C and M are defined as multiplication operators with functions that have the same support. This is true in particular for purely state constrained problems with distributed control.

Proposition 5.5. *Consider the preconditioner Q from* (15) *applied to the block operator* (22)*. Then the condition number is bounded by*

$$\kappa_Q \leq c(1 + \|b\|^{1/2}_\infty \alpha^{-1/2})$$

Proof. Lemma 4.7 yields

$$\gamma_Q \leq c \sup_v \frac{\langle C^*v, C^*v\rangle_{L_2}\langle Mv, Mv\rangle_{L_2}}{(\langle v, v\rangle_{L_2} + \langle C^*v, IMv\rangle_{L_2})^2} \leq c \sup_v \frac{\langle C^*v, C^*v\rangle_{L_2}\langle Mv, Mv\rangle_{L_2}}{2\langle v, v\rangle_{L_2}\langle C^*v, IMv\rangle_{L_2}}. \tag{23}$$

By definition of C and M we obtain

$$\begin{aligned} \langle C^*v, C^*v\rangle_{L_2} &= \alpha^{-1}\langle v, v\rangle_{L_2} \\ \langle Mv, Mv\rangle_{L_2} &\leq \alpha^{1/2}\|b\|^{1/2}_\infty \langle C^*v, IMv\rangle_{L_2}. \end{aligned}$$

Here we used that $M^*I^*C^*v = \alpha^{-1/2}(b(x))^{1/2}v$ is just a pointwise multiplication by positive functions. Hence, we finally compute

$$\gamma_Q \leq c\alpha^{-1/2}\|b\|^{1/2}_\infty.$$

□

A similar situation holds for boundary control problems if the state constraints are only imposed on the boundary. For parabolic problems, a similar result is obtained analogously, replacing Lemma 4.7 by Lemma 4.8.

The improved preconditioner Q is not effective for boundary control and state constraints, if these are imposed on the whole domain. In that case, we have again $CIM = 0$, and thus $Q_0 = Q$ just as described in Section 5.2. As a remedy, it is conceivable to introduce an artificial "virtual" control [15] on the domain equipped with a regularization parameter that is driven to ∞.

A similar situation occurs for additional control constraints, in case that the control is active, i.e. $\chi_{\mathcal{I}}(p)(x) = 0$ where $b(x)$ is large. It is, however, still feasible in this case to use Q as a preconditioner. If active control and state set are disjoint, then we conjecture that Q is still more efficient than Q_0. Otherwise, its efficiency may degrade to a level comparable to Q_0 if this assumption is not valid.

6 A splitting argument for refined estimates

In some situations our estimates for Q can still be refined, if we are willing to impose additional assumptions on CIM. In broad terms, we will assume that CIM is defined as multiplication operators via piecewise constant functions with smoothly bounded level sets. The applications below comprise Newton systems for control constrained problems and for penalty methods for state constrained problems as described in Example 2.1 and Example 2.4. For simplicity we concentrate on the elliptic case.

In the following, let $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ be an elliptic bilinear form, as defined, e.g., in (2):

$$a(y, v) = \int_\Omega \langle \kappa(x)\nabla y, \nabla v\rangle_{\mathbb{R}^d} + a_0(x)yv\,dx \tag{24}$$

Further, let us define the corresponding operator in strong form:

$$\mathcal{A}y := -\operatorname{div}(\kappa(x)\nabla y) + a_0(x)y, \tag{25}$$

which results from $a(\cdot,\cdot)$ via integration by parts, assuming that κ is sufficiently smooth.

Assumption 6.1. Assume that CIM is a multiplication operator, defined by a piecewise constant function ϕ, i.e., $(CIMv)(x) = \phi(x)v(x)$, which assumes two non-negative values ϕ_1 and ϕ_0, such that $\phi_1 > 0$ and $\phi_1 \geq \phi_0 \geq 0$, namely:

$$\phi(x) = \begin{cases} \phi_0 & : \quad x \in J_0 \\ \phi_1 & : \quad x \in J_1 := \Omega \setminus J_0. \end{cases} \tag{26}$$

Let us denote by ∂J_0 the boundary of J_0 relative to Ω, i.e., $\partial J_0 \subset \Omega$, so that $\partial J_0 = \partial J_1$.

Assume that J_0 and J_1 are Lipschitz domains and that the solution v_J of the problem

$$v_J \in H_0^1(J_0) : \quad a(v_J, w) = \int_{J_0} fw\,dx \quad \forall w \in H_0^1(J_0)$$

gives rise to the following trace estimate:

$$\|\partial_{\kappa\nu} v_J\|_{L_2(\partial J_0)} \leq c_{tr,1}\|f\|_{L_2(J_0)}. \tag{27}$$

Here $\partial_{\kappa\nu}$ stands for the derivative in direction of the outer normal ν at a point $x \in \partial J_0$ with respect to the scalar product induced by $\kappa(x)$. This normal has to be defined almost everywhere on ∂J_0.

The validity of (27) certainly depends on the smoothness of ∂J_0 and on the coefficients of $a(\cdot,\cdot)$, and is known to hold e.g., for $H^{3/2+\varepsilon}$-regular problems. In general one only has (27) for $\|\partial_{\kappa\nu} v_J\|_{H^{-1/2}(\partial J_0)}$. In the context of optimal control the smoothness of ∂J_0 can usually only be observed a-posteriori, for example, after the active set of the optimal control has been computed. From an intuitive point of view, smoothness of ∂J_0 results in a relatively weak coupling of J_0 and J_1. So it should have a certain influence on the condition number.

Lemma 6.2. *Let J be an open domain with Lipschitz boundary. Then for $v \in H^1(J)$ the following estimate holds for the trace operator $\tau : H^1(J) \to L_2(\partial J)$:*

$$\|\tau(v)\|_{L_2(\partial J)} \leq c_{tr,2}\sqrt{\|v\|_{H^1(J)}\|v\|_{L_2(J)}}.$$

Proof. Since $v \in H^1(J)$, $\tau(v)$ exists by the classical trace theorem. After localization and transformation of a part of the boundary to the first coordinate axis, we end up in showing (dividing the coordinates into $x = (x', t)$):

$$\|v(\cdot, 0)\|_{L_2}^2 \leq c\|v\|_{H^1(J)}\|v\|_{L_2(J)}.$$

This can be obtained by the formula of integration by parts, well known from parabolic problems:

$$2\int_0^T \langle \frac{d}{dt}v(t), v(t)\rangle\, dt = \|v(T,\cdot)\|_{L_2}^2 - \|v(0,\cdot)\|_{L_2}^2$$

Choosing T large enough, such that $v(T) \equiv 0$, we obtain

$$\|v(0,\cdot)\|_{L_2}^2 \leq 2\int_0^T \|\frac{d}{dt}v(t)\|_{L_2}\|v(t)\|_{L_2}\, dt \leq 2\|v\|_{H^1}\|v\|_{L_2}.$$

□

This estimate can *not* be acquired directly via interpolation theory of Sobolev spaces, because there exists no continuous trace operator $H^{1/2}(J) \to L_2(\partial J)$.

In the following lemma we will consider the problem

$$v \in H_0^1(\Omega): \quad a(v,w) + \int_\Omega \phi(x)v(x)w(x)\, dx = \int_\Omega fw\, dx \quad \forall w \in H_0^1(\Omega). \tag{28}$$

This can also be written in operator notation, if we define (E_S is the Sobolev embedding):

$$\begin{aligned} \Phi : H_0^1(\Omega) &\to H_0^1(\Omega)^* \\ (\Phi v)(x) &:= E_S^*\phi(x)(E_S v)(x) \end{aligned} \tag{29}$$

Then (28) reads

$$(A + \Phi)v = f.$$

Lemma 6.3. *Consider problem (28) such that ϕ has the properties, defined in Assumption 6.1, and assume that $f \in L_2(\Omega)$. Then*

$$\|\Phi v\|_{L_2(\Omega)} \leq c(\|f\|_{L_2(\Omega)} + \phi_1^{1/4}\|f\|_{L_2(J_0)}). \tag{30}$$

If additionally $\phi_0 > 0$, then

$$\|v\|_{L_2(\Omega)} \leq c(1 + \phi_1^{1/4})\|\Phi^{-1}f\|_{L_2(\Omega)}. \tag{31}$$

In all these estimates c depends on the regularity of J_0 and J_1, and on the ellipticity of $a(\cdot,\cdot)$ in (28).

Proof. The idea of the proof is to split v into two parts $v = v_0 + \tilde{v}$. To define v_0, we first consider v_J, the solution of the problem:

$$v_J \in H_0^1(J_0): \quad a(v_J, w) + \phi_0 \int_{J_0} v_J w \, dx = \int_{J_0} f w \, dx \quad \forall w \in H_0^1(J_0), \tag{32}$$

and extend this function by 0 to a function $v_0 \in H_0^1(\Omega)$, such that $v_0|_{J_0} = v_J$, and $v_0 = 0$ on J_1. Testing (32) with $\phi_0 v_J$ and dividing by $\|\phi_0 v_J\|_{L_2(J_0)}$ we obtain

$$\phi_0 \|v_J\|_{L_2(J_0)} \le \|f\|_{L_2(J_0)}.$$

Thus, v_J satisfies

$$a(v_J, w) = \int_{J_0} \underbrace{(f - \phi_0 v_J)}_{\tilde{f}} w \, dx \quad \forall w \in H_0^1(J_0),$$

with $\|\tilde{f}\|_{L_2} \le 2\|f\|_{L_2}$. By our trace assumption we conclude

$$\|\partial_{\kappa\nu} v_J\|_{L_2(\partial J_0)} \le c_{\mathrm{tr},1} 2 \|f\|_{L_2(J_0)}. \tag{33}$$

Integration by parts on J_0 yields

$$\begin{aligned} \mathcal{A} v_0 + \phi_0 v_0 = f \text{ on } J_0 \\ v_0 = 0 \text{ on } \partial J_0 \cup \partial\Omega \cup J_1. \end{aligned}$$

Testing this equation with $w \in H_0^1(\Omega)$ and separate integration by parts on J_0 and J_1 reveals that v_0 satisfies the weak form

$$v_0 \in H_0^1(\Omega): \quad a(v_0, w) + \phi_0 \int_{J_0} v_0 w \, dx + \int_{\partial J_0} \partial_{\kappa\nu} v_J w \, ds = \int_{J_0} f w \, dx \quad \forall w \in H_0^1(\Omega).$$

Hence, $\tilde{v} = v - v_0$ satisfies the equation

$$\tilde{v} \in H_0^1(\Omega): \quad a(\tilde{v}, w) + \int_\Omega \phi(x) \tilde{v} w \, dx = \int_{\partial J_0} \partial_{\kappa\nu} v_J w \, ds + \int_{J_1} f w \, dx \quad \forall w \in H_0^1(\Omega).$$

This follows from subtraction of the weak forms for v and v_0, taking into account that $v_0 = 0$ on J_1. Testing with $\tilde{v}$ we get

$$a(\tilde{v}, \tilde{v}) + \int_\Omega \phi(x) \tilde{v}^2 \, dx \le \|\partial_{\kappa\nu} v_J\|_{L_2(\partial J_0)} \|\tilde{v}\|_{L_2(\partial J_0)} + \|f\|_{L_2(J_1)} \|\tilde{v}\|_{L_2(J_1)} \tag{34}$$

By Lemma 6.2, we obtain:

$$\|\tilde{v}\|_{L_2(\partial J_0)} \le c_{\mathrm{tr},2} \sqrt{\|\tilde{v}\|_{L_2(J_1)} \|\tilde{v}\|_{H^1(J_1)}}.$$

Division of (34) by the square-root of its left-hand-side and taking into account $\phi_1 > 0$ we obtain due to (33)

$$\begin{aligned} &\|\phi^{1/2} \tilde{v}\|_{L_2(\Omega)} \le \sqrt{a(\tilde{v}, \tilde{v}) + \int_\Omega \phi(x) \tilde{v}^2 \, dx} \\ &\le \frac{\|\partial_{\kappa\nu} v_J\|_{L_2(\partial J_0)} c_{\mathrm{tr},2} \sqrt{\|\tilde{v}\|_{L_2(J_1)} \|\tilde{v}\|_{H^1(J_1)}} + \|f\|_{L_2(J_1)} \|\tilde{v}\|_{L_2(J_1)}}{\sqrt{c_a \|\tilde{v}\|^2_{H^1(\Omega)} + \phi_1 \|\tilde{v}\|^2_{L_2(J_1)}}} \\ &\le c \|\partial_{\kappa\nu} v_J\|_{L_2(\partial J_0)} \phi_1^{-1/4} + \|f\|_{L_2(J_1)} \phi_1^{-1/2} \le c \|f\|_{L_2(J_0)} \phi_1^{-1/4} + \|f\|_{L_2(J_1)} \phi_1^{-1/2}. \end{aligned}$$

Since $\phi_1 \geq \phi_0$ we can use the triangle inequality to obtain our first result:

$$\begin{aligned}\|\phi v\|_{L_2(\Omega)} &= \|\phi_0 v_0 + \phi\tilde{v}\|_{L_2(\Omega)} \leq \phi_0\|v_J\|_{L_2(J_0)} + \phi_1^{1/2}\|\phi^{1/2}\tilde{v}\|_{L_2(\Omega)} \\ &\leq \|f\|_{L_2(J_0)} + c\phi_1^{1/4}\|f\|_{L_2(J_0)} + \|f\|_{L_2(J_1)} \leq \sqrt{2}\|f\|_{L_2(\Omega)} + c\phi_1^{1/4}\|f\|_{L_2(J_0)}.\end{aligned}$$

Tracing back the constant c, we notice that it depends solely on $c_{\mathrm{tr},1}$, $c_{\mathrm{tr},2}$, and c_a.

For the second result we apply a duality technique. For given $w \in H_0^1(\Omega)$ define $z_w := (A+\Phi)^{-1}w$. Then we can compute

$$\begin{aligned}\|v\|_{L_2(\Omega)} &= \sup_{\|w\|_{L_2(\Omega)}=1} \langle v, w\rangle_{L_2(\Omega)} = \sup_{\|w\|_{L_2(\Omega)}=1} \langle (A^* + \Phi)^{-1} f, w\rangle \\ &= \sup_{\|w\|_{L_2(\Omega)}=1} \langle f, z_w\rangle = \sup_{\|w\|_{L_2(\Omega)}=1} \langle \Phi^{-1} f, \Phi z_w\rangle \\ &\leq \sup_{\|w\|_{L_2(\Omega)}=1} \|\Phi^{-1} f\|_{L_2(\Omega)}\|\Phi z_w\|_{L_2(\Omega)} \\ &\leq \sup_{\|w\|_{L_2(\Omega)}=1} \|\Phi^{-1} f\|_{L_2(\Omega)}(\sqrt{2} + c\phi_1^{1/4})\|w\|_{L_2(\Omega)},\end{aligned}$$

which implies our assertion, since $\|w\|_{L_2(\Omega)} = 1$. □

Remark 6.4. If the smoothness of ∂J_0 does not admit an L_2-estimate of the form (27) but only in a weaker norm (e.g. in $\|\partial_{\kappa\nu} v_J\|_{H^{-s}(\partial J_0)}$ for $s \in [0, 1/2]$), one can show a similar result, where $\phi_1^{1/4}$ is replaced by $\phi_1^{1/4+s/2}$.

Sharpness of Lemma 6.3. In the following we will briefly argue that the estimate (30) is sharp. Let $\Omega =]0,2[\subset \mathbb{R}$, and choose $\phi = \phi_0 = 0$ on $]1,2[$ and $\phi = \phi_1 = \mathit{const}$ on $]0,1[$. Moreover, set $f = 0$ on $]0,1[$ and $f = 2$ on $[1,2[$. Consider the problem

$$-v'' + \phi v = f \text{ on }]0,2[,\ v'(0) = 0,\ v(2) = 0,$$

which is by symmetry one half of a Dirichlet problem on $]-2,2[$. We can now proceed along the lines of our proof and split $v = \tilde{v} + v_0$, where v_0 solves the ϕ-independent problem

$$-v_0'' = 2 \text{ on }]1,2[,\ v_0(1) = 0,\ v_0(2) = 0,$$

which has the solution $v_0(x) = -(x-1.5)^2 + 0.25$ with derivative $v_0'(1) = 1$ at $x = 1$. The second part $\tilde{v}$ satisfies the following differential equation

$$\begin{gathered}-\tilde{v}'' + \phi_1\tilde{v} = 0 \text{ on } [0,1],\ \tilde{v}'(0) = 0 \\ -\tilde{v}'' = 0 \text{ on }]1,2],\ \tilde{v}(2) = 0 \\ \tilde{v}_-'(1) = \tilde{v}_+'(1) + v_0'(1) = \tilde{v}_+'(1) + 1,\end{gathered}$$

where $\tilde{v}_-'(1)$ and $\tilde{v}_+'(1)$ denote the left and right limit of $\tilde{v}'$ at $x = 1$, respectively. Obviously, $\tilde{v}$ is a linear polynomial on $]1,2]$, so that by our boundary conditions we have $\tilde{v}_+'(1) = -\tilde{v}(1)$, and it remains to solve the following problem on $[0,1]$:

$$-\tilde{v}'' + \phi_1\tilde{v} = 0 \text{ on } [0,1],\ \tilde{v}'(0) = 0,\ \tilde{v}'(1) = 1 - \tilde{v}(1).$$

By a classical ansatz, this equation has a solution of the form

$$\begin{aligned}\tilde{v}(x) &= a\cosh(\sqrt{\phi_1}x) \\ \tilde{v}'(x) &= \sqrt{\phi_1}a\sinh(\sqrt{\phi_1}x)\end{aligned}$$

which already has $\tilde{v}'(0) = 0$ built in, so that we only have to determine a from the condition $\tilde{v}'(1) = 1 - \tilde{v}(1)$. A short computation yields:

$$a = \frac{1}{\sqrt{\phi_1}\sinh(\sqrt{\phi_1}) + \cosh(\sqrt{\phi_1})},$$

so that we can compute (recall $\phi_0 = 0$)

$$\begin{aligned}\|\phi v\|^2_{L_2([0,2])} = \|\phi_1 \tilde{v}\|^2_{L_2([0,1])} &= \phi_1^2 a^2 \int_0^1 \cosh^2(\sqrt{\phi_1}x)\,dx \\ &= \frac{\phi_1^2}{(\sqrt{\phi_1}\sinh(\sqrt{\phi_1}) + \cosh(\sqrt{\phi_1}))^2}\frac{\sqrt{\phi_1} + \cosh(\sqrt{\phi_1})\sinh(\sqrt{\phi_1})}{2\sqrt{\phi_1}}.\end{aligned}$$

Taking into consideration that

$$\lim_{t\to\infty}\frac{\cosh(t)}{e^t} = \lim_{t\to\infty}\frac{\sinh(t)}{e^t} = \frac{1}{2},$$

we obtain for large ϕ_1:

$$\lim_{\phi_1\to\infty}\|\phi_1\tilde{v}\|_{L_2([0,1])}\phi_1^{-1/4} = \lim_{\phi_1\to\infty}\sqrt{\phi_1^{3/2}\frac{e^{\sqrt{\phi_1}}}{2\phi_1 e^{\sqrt{\phi_1}}}}\phi_1^{-1/4} = \frac{1}{\sqrt{2}}.$$

Since $\|f\|_{L_2([0,2])}$ is fixed, we obtain the asymptotics

$$\|\Phi v\|_{L_2([0,2])} \sim \phi_1^{1/4}\|f\|_{L_2([0,2])}.$$

We finally remark that this problem can be lifted by parallel translation to a Poisson problem on $]0,2[^d$ for $d > 1$, if the newly created boundaries are equipped with homogeneous Neumann boundary conditions. In this case $\partial_{\kappa\nu}v_0(x) = 1$ is constant along the set $\{x \in]0,2[^d : x_1 = 1\}$, so that our estimate cannot be improved, even if $\partial_{\kappa\nu}v_0$ is assumed to be in a more regular space than $L_2(\partial J_0)$.

Thus, taking also into account our lower bound on the condition number (19) we can be quite sure that the estimates in the following section will be sharp.

6.1 Application to distributed control with control bounds

Let us come back to the problem, defined by the operator (4). In this setting we have $C^*v = \alpha^{-1/2}\chi_{\mathcal{I}}(p)E_S v$, and $M = E_S : H^1(\Omega) \hookrightarrow L_2(\Omega)$ and $I = Id$. Thus, we may set $\phi(x) = \alpha^{-1}\chi_{\mathcal{I}}(p)$, and thus $\phi_1 = \alpha^{-1/2}$ and $\phi_0 = 0$. Then $(CIMv)(x) = E_S^*\phi(x)(E_S v)(x) = \Phi$ in the notation of (29).

Application of Lemma 6.3 yields the following result:

Proposition 6.5. *Consider the preconditioner $\langle\cdot,\cdot\rangle_Q$ applied to the block operator* (4)*. Assume that the boundary between active and inactive set satisfies the Assumption 6.1. Then we obtain the following condition number:*

$$\kappa_Q \le c(1 + \alpha^{-1/4}).$$

Proof. In order to apply Lemma 4.6 we set

$$f = M^{-*}(A + CIM)^*v = (A + \Phi)^*v,$$

so that v is the solution of the following problem

$$a(v,w) + \int_\Omega \phi(x) v w \, dx = \int_\Omega f w \, dx \quad \forall w \in H_0^1(\Omega).$$

Hence, Lemma 6.3 yields

$$\|C^* v\|_{L_2(\Omega)} \le (1 + c\alpha^{-1/8}) \|f\|_{L_2(\Omega)}.$$

Inserting this estimate into (16) we obtain the desired result. □

6.2 Application to distributed control with state constraints

For state constraints, we assume that $b(x)$ is piecewise constant taking two values $\|b\|_\infty = b_1 > b_0 > 0$. In penalty methods, as in Example 2.4 we have $b(x) = 1 + \gamma\chi_{y<0}(x)$. As usual in distributed control $U = H = L_2(\Omega)$ and $C = \alpha^{-1/2} E_S^*$, $I = Id$, $M = \sqrt{b(x)} E_S$, so that

$$(CIMv)(x) = E_S^* \alpha^{-1/2} \sqrt{b(x)} (E_S v)(x).$$

So we can define $\phi(x) := \alpha^{-1/2}\sqrt{b(x)}$ such that $CIM = \Phi$ in the notation of (29). By our assumption, ϕ is also piecewise constant, and we set $\phi_1 = b_1 \alpha^{-1/2}$ and $\phi_0 = b_0 \alpha^{-1/2}$. Application of Lemma 6.3 then yields:

Proposition 6.6. *Assume that $p \in L_\infty$. Consider the preconditioner $\langle \cdot, \cdot \rangle_Q$ applied to the operator* (22). *Then we obtain the following condition number:*

$$\kappa_Q \le c(1 + \|b\|_\infty^{1/4} \alpha^{-1/4}).$$

Proof. We proceed similar as in the control constrained case, defining

$$f = (A + CIM)^* v = (A + \Phi)^* v.$$

Then $\langle v, v \rangle_Q = \|M^{-*} f\|^2_{L_2(\Omega)}$. By Lemma 6.3 we conclude

$$\begin{aligned}
\|C^* v\|_{L_2(\Omega)} &= \alpha^{-1/2} \|v\|_{L_2(\Omega)} \le \alpha^{-1/2}(1 + c\phi_1^{1/4}) \|\Phi^{-1} f\|_{L_2(\Omega)} \\
&= \alpha^{-1/2}(1 + c\|b\|_\infty^{1/8} \alpha^{-1/8}) \|\sqrt{\alpha} M^{-*} f\|_{L_2(\Omega)} \\
&= (1 + c\|b\|_\infty^{1/8} \alpha^{-1/8}) \|M^{-*} f\|_{L_2(\Omega)}
\end{aligned}$$

Inserting this into (16) we obtain the desired result. □

In penalty methods for state constrained problems one considers a homotopy, which results in $\gamma \to \infty$ and thus $\|b\|_\infty \to \infty$. In practical applications this leads to values of γ in the order of 10^8 to 10^{12}. Concerning parameters of such high magnitude, our improved condition number estimate is of particular value, since the number of required cg iterations then only grows with the 8^{th} square-root of $\|b\|_\infty$. Compared to the preconditioner Q_0, which may require hundreds or thousands of cg iterations (proportional to $\sqrt{\|b\|_\infty}$), Q merely takes a very limited number of iterations.

7 Small (in)active sets

In some situations it is likely that those sets where M and C are large have small Lebesgue measure. Applications comprise regularized bang-bang control, where the inactive set becomes small for small regularization parameters, and regularized state constrained control, if the active set tends to a Lebesgue null set.

We approach our problem via an L_∞-estimate due to Stampacchia:

Lemma 7.1. *Let $\Omega \subset \mathbb{R}^d$ for $d \le 3$ be a bounded Lipschitz domain and consider the following elliptic equation in weak form:*

$$v \in H_0^1(\Omega): \quad a(v,w) + \int_\Omega \phi(x) v w \, dx = \int_\Omega f w \, dx,$$

where $\phi(x)$ is a positive function in $L_\infty(\Omega)$ and $f \in L_2(\Omega)$.

Then we have the estimates

$$\begin{aligned} \|v\|_{L_\infty} &\le c\|f\|_{L_2} \\ \|v\|_{L_2} &\le c\|f\|_{L_1}, \end{aligned}$$

where c is independent of ϕ.

Proof. Our first estimate is due to Stampacchia [14], and our second estimate follows via a duality technique, similar to the one, used in the proof of Lemma 6.3. □

Proposition 7.2. *Consider the preconditioners Q_0 from* (12) *and Q from* (15) *applied to the block operator* (4)*. We have the condition number estimates*

$$\begin{aligned} \kappa_{Q_0} &\le c(1 + \alpha^{-1}\|\chi_\mathcal{I}(p)\|_{L_1}) \\ \kappa_Q &\le c(1 + \alpha^{-1}\|\chi_\mathcal{I}(p)\|_{L_1}). \end{aligned}$$

Proof. In both cases we have

$$\langle C^* v, C^* v\rangle = \int_\Omega \alpha^{-1}\chi_\mathcal{I}(p) v^2 \, dx \le \|\alpha^{-1}\chi_\mathcal{I}(p)\|_{L_1}\|v\|_{L_\infty}^2.$$

By Lemma 7.1 we obtain $\|v\|_{L_\infty}^2 \le c\langle v, v\rangle_{Q_0}$ and also $\|v\|_{L_\infty}^2 \le c\langle v, v\rangle_Q$. Inserting these estimates into (13) and (16), respectively, we obtain the desired result. □

In bang-bang control, one frequently encounters an assumption of the form (cf. e.g. [5]):

$$|\{x \in \Omega : |p(x)| < \varepsilon\}| \le c\varepsilon.$$

If such a problem is regularized by a homotopy $\alpha \to 0$, and one assumes that the corresponding adjoint states p_α uniformly satisfy such an assumption, too, then one obtains $|\chi_\mathcal{I}(p_\alpha)| \le c\alpha$. In this context, the condition number of the preconditioners is bounded.

Proposition 7.3. *Consider the preconditioners $\langle\cdot,\cdot\rangle_{Q_0}$ and $\langle\cdot,\cdot\rangle_Q$ applied to the block operator* (22)*. We have the condition number estimates*

$$\begin{aligned} \kappa_{Q_0} &\le c(1 + \alpha^{-1}\|b\|_{L_1}) \\ \kappa_Q &\le c(1 + \alpha^{-1}\|b\|_{L_1}). \end{aligned}$$

Proof. In both cases we have

$$\langle C^* v, C^* v\rangle \le \alpha^{-1}\|v\|^2_{L_2},$$

and it remains to derive a bound of the form

$$\|v\|_{L_2} \le c\|M^{-*}\underbrace{(A+CIM)^* v}_{f}\|_{L_2} = c\|M^{-*}f\|_{L_2},$$

for Q and the corresponding one for Q_0, where the term CIM is missing. In our case we have $(M^*w)(x) = \sqrt{b(x)}w(x)$. Since $(A+CIM)^*v = f$, we conclude with Lemma 7.1 that

$$\|v\|_{L_2} \le \|f\|_{L_1} = \|\sqrt{b}M^{-*}f\|_{L_1} \le \|\sqrt{b}\|_{L_2}\|M^{-*}f\|_{L_2} \le \sqrt{\|b\|_{L_1}}\|M^{-*}f\|_{L_2}.$$

Inserting this estimate (and the corresponding one for Q_0) into (13) and (16), respectively, we obtain the desired result. □

8 Numerical examples

In this section we perform a study numerical study of the pcg method with our preconditioners. We consider two examples. The first is related to a control constrained problem, the second is related to a regularized state constrained problem.

The pcg iteration is terminated, after the estimated error in energy norm has dropped below 10^{-8}, where an estimator in the spirit of [6, Sec. 5.3.3(c)] is used.

Our simple implementation is based on matlab, and the discretization of our optimality system was done with 5-point star finite differences. For the solution of the single PDE blocks, the built-in sparse direct Cholesky factorization has been used.

In both cases the computational domain is the unit-square, and the current active set $\mathcal{A}$ is a disc with center $(0.5, 0.5)$ and radius 0.4. As right hand side we choose the function $r \equiv 1$.

$h \setminus \alpha$	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}	10^{-12}
2^{-6}	8	13	18	22	24	23
2^{-7}	7	13	18	22	29	32
2^{-8}	7	13	17	22	31	41
2^{-9}	7	12	17	22	31	52
2^{-10}	7	12	17	22	33	54

$h \setminus \alpha$	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}	10^{-12}
2^{-6}	5	9	33	234	798	953
2^{-7}	5	9	33	236	1749	3222
2^{-8}	5	9	31	233	2095	> 10000
2^{-9}	4	9	31	231	2122	> 10000

Figure 1: Number of pcg iterations for preconditioner Q (top) and Q_0 (bottom) for control constraints with varying grid size h and Tychonov parameter α.

In our first problem, we thus solve a problem of the form

$$\begin{pmatrix} E^*E & A^* \\ A & -E^*\alpha^{-1}\chi_{\Omega\setminus\mathcal{A}}E \end{pmatrix}\begin{pmatrix} \delta y \\ \delta p \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{35}$$

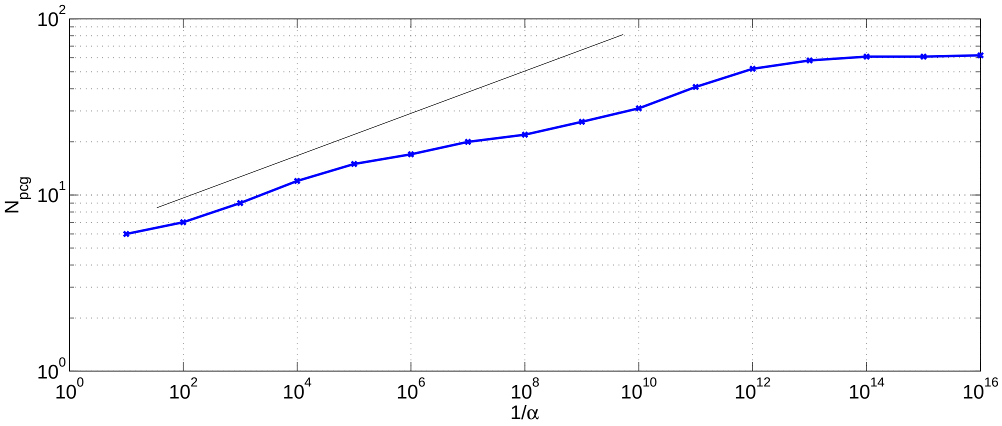

Figure 2: Number N_{pcg} of computed pcg iterations for problem (35) and theoretical prediction $N_{pcg} \sim \alpha^{-1/8}$ plotted against α^{-1} on a 512×512 grid.

with varying Tychonov parameter α. Here A corresponds to the weak form of $-\Delta$ on $H_0^1(]0,1[\times]0,1[)$, and E is the Sobolev embedding. Of particular interest is the case, where α is very small. In the second example we solve a problem of the form

$$\begin{pmatrix} E^*(1+\gamma\chi_{\mathcal{A}})E & A^* \\ A & -E^*E \end{pmatrix} \begin{pmatrix} \delta y \\ \delta p \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{36}$$

with varying penalty parameter γ. Here γ can become very large during the course of a path-following method. It can be observed in both examples that the preconditioner Q is vastly superior to Q_0 for large α^{-1} or γ. Moreover, for very large parameters discretization effects tend to yield smaller numbers of pcg iterations for coarse grids than for fine grids.

$h \setminus \gamma$	10^2	10^4	10^6	10^8	10^{10}	10^{12}
2^{-6}	8	12	17	23	28	29
2^{-7}	8	12	17	23	34	38
2^{-8}	8	12	16	23	36	49
2^{-9}	8	12	16	23	37	58
2^{-10}	7	12	16	23	36	61

$h \setminus \gamma$	10^2	10^4	10^6	10^8	10^{10}	10^{12}
2^{-6}	5	11	41	300	1204	1638
2^{-7}	5	11	39	297	2464	5323
2^{-8}	5	11	39	293	2727	> 10000
2^{-9}	5	11	40	293	2722	> 10000

Figure 3: Number of pcg iterations for preconditioner Q (top) and Q_0 (bottom) for regularized state constraints with varying grid size h and penalty parameter γ.

If we compare the observed number of pcg iterations with the predicted number of iterations (cf. Figure 8) we observe two things. First, the average increase of iterations seems to be slightly slower than predicted. However, there are regions (i.e. $\alpha^{-1} \in [10^2, 10^4]$ and $\alpha^{-1} \in [10^{10}, 10^{12}]$) where the slopes seem to fit. For very large $\alpha^{-1} > 10^{12}$ we observe a

$\alpha \setminus$	N_{Newton}	N_{pcg}^{total}	N_{pcg}^{avg}
10^{-2}	3	8	2.667
10^{-4}	6	28	4.667
10^{-6}	10	59	5.9
10^{-8}	9	79	8.778
10^{-10}	9	119	13.22
10^{-12}	8	142	17.75

Figure 4: Convergence history of semi-smooth Newton method for varying α

saturation of the number if iterations. This effect is most probably due to the discretization of the problem, and can also be observed in the last columns of the top table in Figure 1. Such behavior is usually a clear indication that the discretization of the problem is too coarse.

Of course, this iterative solver can also be used as an inner loop inside a semi-smooth Newton method. Let us give an example in the context of control constraints $u \geq 0$. Here we use again a control problem with distributed control and corresponds to the weak form of $\mathcal{A} = -\Delta$ on $H_0^1(]0,1[\times]0,1[)$. As desired state, we choose $y_d = A^{-1}\sin(\pi x_1 x_2)$. State y and adjoint state p are both discretized by standard finite differences with mesh size $h = 2^{-8}$. To compute the solution for this problem with varying, up to very small α we reuse the computed solution for the last (larger) α as an initial guess for the next (smaller) α. This acts as a simplistic path-following method for $\alpha \to 0$ and compensates for the inefficient global convergence behavior of semi-smooth Newton in case of small α.

9 Conclusion and Outlook

We have proposed and analyzed block preconditioners for systems that arise in certain optimal control problems with PDEs. It can be used effectively for control constraints, if the domain of observation contains the domain of control. For state constraints the approximate active constraint set should be contained in the control domain.

In these cases the condition numbers of the resulting systems are in general only the square root of the condition numbers, that are obtained via a simple preconditioner Q_0, asymptotically with respect to critical parameters. Under additional structural assumptions, it can be shown that the condition number grows even slower, like the fourth root of the condition number of Q_0. In the unconstrained case one obtains a condition numbers independent of the critical parameter.

Our results are of particular interest in state constrained optimal control problems, where up to now the robustness of available preconditioners with respect to regularization parameters was poor. The class of state constrained problems seems to be divided into two subclasses. The first, where the control can act on the active set of the constraints, and the class of remaining problems, where the control acts more indirectly. The first class seems to be tractable more easily than the second, and new ideas are needed for the second class.

Finally, our results are valid for exact solutions of the modified PDEs. This is already a significant progress, since direct solvers for elliptic problems are much more efficient than for coupled systems. In the parabolic case the advantage is even larger. It is straightforward to replace direct solvers by multigrid preconditioners. However, showing optimality and robustness of these preconditioners seems an open non-trivial theoretical issue that needs to be addressed in the future. The usual H^1 techniques cannot be used because the natural space for the preconditioners is D_K. Finally, from an algorithmic point of view it is desirable to apply our preconditioners in an adaptive multilevel method in the spirit of [20], where inexactness of Newton steps caused by the iterative solver and the adaptive grid refinement is appropriately handled within an adaptive inexact Newton path-following method in function space.

References

[1] A. Battermann and M. Heinkenschloss. Preconditioners for Karush-Kuhn-Tucker matrices arising in the optimal control of distributed systems. In *Control and estimation of distributed parameter systems (Vorau, 1996)*, volume 126 of *Internat. Ser. Numer. Math.*, pages 15–32. Birkhäuser, 1998.

[2] A. Battermann and E. W. Sachs. Block preconditioners for KKT systems in PDE-governed optimal control problems. In *Fast solution of discretized optimization problems (Berlin, 2000)*, volume 138 of *Internat. Ser. Numer. Math.*, pages 1–18. Birkhäuser, Basel, 2001.

[3] A. Borzì. Smoothers for control- and state-constrained optimal control problems. *Comput. Vis. Sci.*, 11(1):59–66, 2008.

[4] A. Borzi and V. Schulz. Multigrid methods for PDE optimization. *SIAM Rev.*, 51(2):361–395, 2009.

[5] K. Deckelnick and M. Hinze. A note on the approximation of elliptic control problems with bang-bang controls. *Comput. Optim. Appl.*, pages 931–939, 2012.

[6] P. Deuflhard and M. Weiser. *Adaptive Numerical Solution of Partial Differential Equations.* de Gruyter, 2012.

[7] H. S. Dollar, N. I. M. Gould, M. Stoll, and A. J. Wathen. Preconditioning saddle-point systems with applications in optimization. *SIAM J. Sci. Comput.*, 32(1):249–270, 2010.

[8] H. Gajewski, K. Gröger, and K. Zacharias. *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen.* Akademie-Verlag, Berlin, 1974. Mathematische Lehrbücher und Monographien, II. Abteilung, Mathematische Monographien, Band 38.

[9] A. Günnel, R. Herzog, and E. Sachs. A note on preconditioners and scalar products for Krylov methods in Hilbert space. Technical report, TU Chemnitz, July 2011.

[10] R. Herzog and E. Sachs. Preconditioned conjugate gradient method for optimal control problems with control and state constraints. *SIAM J. Matrix Anal. Appl.*, 31(5):2291–2317, 2010.

[11] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semi-smooth Newton method. *SIAM J. Optim.*, 13:865–888, 2003.

[12] M. Hintermüller and K. Kunisch. Feasible and non-interior path-following in constrained minimization with low multiplier regularity. *SIAM J. Control Optim.*, 45(4):1198–1221, 2006.

[13] M. Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case. *Comput. Optim. Appl.*, 30:45–63, 2005.

[14] D. Kinderlehrer and G. Stampacchia. *An Introduction to Variational Inequalities and their Applications*. Academic Press, New York, 1980.

[15] K. Krumbiegel and A. Rösch. A virtual control concept for state constrained optimal control problems. *Comput. Optim. Appl.*, 43(2):213–233, 2009.

[16] J. W. Pearson, M. Stoll, and A. J. Wathen. Preconditioners for state constrained optimal control problems with Moreau-Yosida penalty function. Technical Report NA-12/05, Oxford University, Mathematical Institute, March.

[17] T. Rees, M. Stoll, and A. Wathen. All-at-once preconditioning in PDE-constrained optimization. *Kybernetika (Prague)*, 46(2):341–360, 2010.

[18] A. Schiela. A simplified approach to semismooth Newton methods in function space. *SIAM J. Optim.*, 19(3):1417–1432, 2008.

[19] A. Schiela. Barrier methods for optimal control problems with state constraints. *SIAM J. Optim.*, 20(2):1002–1031, 2009.

[20] A. Schiela and A. Günther. An interior point algorithm with inexact step computation in function space for state constrained optimal control. *Numer. Math.*, 119(2):373–407, 2011.

[21] J. Schöberl, R. Simon, and W. Zulehner. A robust multigrid method for elliptic optimal control problems. *SIAM J. Numer. Anal.*, 49(4):1482–1503, 2011.

[22] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM J. Matrix Anal. Appl.*, 29(3):752–773, 2007.

[23] M. Stoll and A. Wathen. Preconditioning for partial differential equation constrained optimization with control constraints. *Numerical Linear Algebra with Applications*, 19(1):53–71, 2012.

[24] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. AMS, Providence, 2010.

[25] M. Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.*, 13:805–842, 2003.

[26] D. Wachsmuth and G. Wachsmuth. Necessary conditions for convergence rates of regularizations of optimal control problems. RICAM Report 2012-04, Johann Radon Institute for Computational and Applied Mathematics, January 2012.

[27] W. Zulehner. Nonstandard norms and robust estimates for saddle point problems. *SIAM J. Matrix Anal. Appl.*, 32(2):536–560, 2011.