Deutsche Forschungsgemeinschaft
Priority Programme 1962

A Composite Step Method with Inexact Step Computations for PDE Constrained Optimization

Manuel Schaller, Anton Schiela, Matthias Stöcklein

SPP 1962 | Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization

Preprint Number SPP1962-098

received on October 31, 2018

Edited by
SPP1962 at Weierstrass Institute for Applied Analysis and Stochastics (WIAS)
Leibniz Institute in the Forschungsverbund Berlin e.V.
Mohrenstraße 39, 10117 Berlin, Germany
E-Mail: spp1962@wias-berlin.de

World Wide Web: http://spp1962.wias-berlin.de/

A COMPOSITE STEP METHOD WITH INEXACT STEP COMPUTATIONS FOR PDE CONSTRAINED OPTIMIZATION

MANUEL SCHALLER, ANTON SCHIELA, AND MATTHIAS STÖCKLEIN

Abstract. We consider a composite step algorithm for equality constrained optimization problems. The arising linear systems are inexactly solved with a conjugate gradient method using a constraint preconditioner. The influence of the error on the damping parameters of the underlying Newton scheme and algorithmic parameters is discussed and specialized termination criteria for the conjugate gradient method are given. Application to optimal control of a quasilinear heat equation and nonlinear elasticity illustrates the numerical performance of the resulting algorithm.

AMS MSC 2000: 49M37, 90C55, 90C06

Keywords: optimization with PDEs, cubic regularization, affine covariance, composite step methods, nonlinear elasticity

1. Introduction

In a Hilbert space X with scalar product $\langle\cdot,\cdot\rangle$ and a reflexive space P we consider minimization problems of the form

$$\min_{x\in X} f(x) \quad \text{s.t.} \quad c(x) = 0, \tag{1}$$

where $f : X \to \mathbb{R}$ is a twice continuously Fréchet differentiable functional and $c : X \to P^\star$ is a twice continuously Fréchet differentiable nonlinear operator that could model a differential equation in weak form:

$$c(x) = 0 \quad \text{in } P^\star \quad \Longleftrightarrow \quad c(x)v = 0 \quad \forall v \in P.$$

This setting applies e.g. in the case of optimal control, where $x = (y, u)$ and

$$c(x) = A(y) - Bu$$

with $A : Y \to P^\star$ being a possibly nonlinear differential operator with continuous inverse and $B : U \to P^\star$ a linear and continuous operator. In the following, the Riesz isomorphism is represented by a linear operator $M : X \to X^\star$, defined by $(Mv)w = \langle v, w\rangle$.

We denote the Lagrangian function corresponding to (1) by $L(x, p) = f(x)+pc(x)$. Proceeding formally, we obtain the first order optimality conditions at a minimizer $x^\star$ with multiplier $p^\star$ by $L'(x^\star, p^\star)(v, q) = 0 \quad \forall (v, q) \in X \times P$. For the derivation of this condition, we refer to [22]. Solving this nonlinear equation with a Newton algorithm yields the linear system for the Newton update $(\delta x, \delta p) \in X \times P$:

$$\begin{pmatrix} L_{xx}(x,p) & c'(x)^\star \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \delta x \\ \delta p \end{pmatrix} + \begin{pmatrix} L_x(x,p) \\ c(x) \end{pmatrix} = 0. \tag{2}$$

For the solution of the minimization problem (1) we consider a composite step algorithm based on the preceding work [21]. The idea of composite step methods is to split the Newton update δx into a normal and a tangential step for a precise treatment of optimality and feasibility.

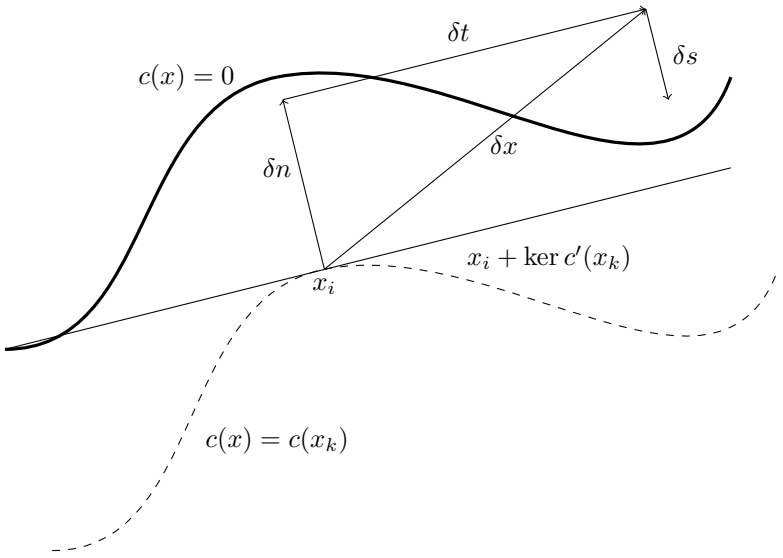


FIGURE 1. Sketch of a composite step $\delta x = \delta n + \delta t$ with second order correction δs at iteration point x_k

A normal step δn satisfies $\delta n \in \ker c'(x)^\perp$ and aims for feasibility, and a tangential step δt satisfies $\delta t \in \ker c'(x)$ and aims for a decrease of the functional value. Moreover, as presented in [21], a second correction δs is used, which is defined as a minimum norm solution of a simplified Newton equation. The motivation for this step is twofold. First it is used for globalization and second to to avoid the Maratos effect. Figure 1 illustrates the steps that are computed in each outer iteration of the composite step algorithm.

In terms of globalization of the normal step, we extend an idea proposed in [11, Section 4.4], where the region of Newton contraction is estimated and normal step damping is performed if necessary. We aim to preserve affine covariance of the algorithm, hence residual norms $\|c(x)\|$ will be avoided. This is motivated by the observation that residual norms are often hard to interpret in the context of partial differential equations. It is contrary to the approach of Byrd-Omojokun [4, 5, 24], where the normal steps minimize $\|c(x)\|$ in a trust region. We will compute normal steps, denoted by δn as possibly damped minimum norm solutions of the Newton-equation $c(x) + c'(x)\delta n = 0$. To assure the solvability of this equation, we need to assume surjectivity of $c'(x)$, which is often given in an optimal control context. The approach employed for the computation of the normal step thus resembles more the ideas of Vardi in [32].

In terms of optimality, a tangential step $\delta t \in \ker c'(x)$ will be computed, minimizing a local quadratic model. For globalization of this matter, we will use a cubic regularization approach as proposed in [6] for the case of unconstrained optimization.

The computation of the steps results in the solution of saddle point systems that are typically of very large scale, in particular if $c(x)$ models a time-dependent PDE or a nonlinearly elastic material. In these cases, the use of iterative solvers is mandatory. In this paper, we will discuss the conjugate gradient method with a constraint preconditioner, also called the projected preconditioned conjugate gradient method (PPCG) [15] as a solution technique. It has the advantage to preserve the optimization structure of the linear systems and treats the equality contraints exactly. Together with a block lower triangular constraint preconditioner the solution of the saddle point system can be reduced to succesive solutions of the underlying linearized

PDEs and their adjoint equations. This strategy is particularly well suited for large scale time dependent problems. In this case one application of the preconditioner can be interpreted as a forward-backward solve in time.

For reasons of efficiency, overly accurate solution of the arising linear systems should be avoided. In other words, inexact steps should be taken. Within the approach of Byrd-Omojokun, mentioned above, inexactness was considered in [17, 16, 25] and, in the context of spatial adaptivity [34].

Here we consider inexact steps for our affine covariant method. The arising residuals influence several algorithmic quantities, which are e.g. necessary for the computation of the damping parameters, and thus the globalization mechanism of the overall algorithm. In broad terms, less accurate steps lead to smaller damping factors and thus to slower convergence of the outer iteration. Thus, a good adaptive stategy to achieve a reasonable trade-off has to be devised. Moreover, we aim to derive requirements on the residuals, such that local superlinear convergence is preserved. This is the main objective of this paper.

2. Definition of exact substeps

In this section, we will introduce normal and tangential step as components of the update, the Lagrange multiplier and the simplified normal step. First, we introduce the Lagrange multiplier at an iterate (x, p) defined as the solution of

$$\underbrace{\begin{pmatrix} M & c'(x)^\star \\ c'(x) & 0 \end{pmatrix}}_{=:H_n} \begin{pmatrix} g_x \\ p_x \end{pmatrix} + \begin{pmatrix} f'(x) \\ 0 \end{pmatrix} = 0. \tag{3}$$

The second equation yields $g_x \in \ker c'(x)$. Together with the first equation, this implies that g_x is the projected gradient of $f(x)$ onto $\ker c'(x)$.

Remark 2.1. Note, that this multiplier is dependent on the current iterate x. Under continuity assumptions on $f'(x)$ and a continuity and surjectivity assumption on $c'(x)$, this multiplier is given as a continuous implicit function in a neighborhood of x, cf. [21, Lemma 2.4]. For algorithmic purposes, we compute a correction for the Lagrange multiplier each iteration via

$$\begin{pmatrix} M & c'(x)^\star \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} g_x \\ \delta p \end{pmatrix} + \begin{pmatrix} L_x(x, p_-) \\ 0 \end{pmatrix} = 0 \tag{4}$$

and set $p_x := p_- + \delta p$, which yields the advantage of the right hand side approaching zero when we get to a local minimizer [21, Section 3.1.2].

The normal step δn_{ex} aiming for feasibility is a minimum norm Gauss-Newton step for the solution of the (underdetermined) problem $c(x) = 0$. The undamped normal step is denoted by Δn_{ex}, whereas the damped normal step is denoted by $\delta n_{\mathrm{ex}} = \nu \Delta n_{\mathrm{ex}}$ with a damping factor $\nu \in]0, 1]$. We compute Δn_{ex} as a solution of the minimum norm problem

$$\min_{\Delta n_{\mathrm{ex}} \in X} \quad \frac{1}{2} \langle \Delta n_{\mathrm{ex}}, \Delta n_{\mathrm{ex}} \rangle_M \quad \text{s.t.} \quad c(x) + c'(x) \Delta n_{\mathrm{ex}} = 0. \tag{5}$$

The stationarity condition of above problem reads

$$\begin{pmatrix} M & c'(x)^\star \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta n_{\mathrm{ex}} \\ q \end{pmatrix} + \begin{pmatrix} 0 \\ c(x) \end{pmatrix} = 0. \tag{6}$$

Lemma 2.2 characterizes an important property of this normal step.

Lemma 2.2. *The normal step δn_{ex} defined by (6) satisfies $\delta n_{\mathrm{ex}} \in \ker c'(x)^\perp$.*

Proof. Let $v \in \ker c'(x)$. Then

$$0 = \langle \Delta n_{\mathrm{ex}}, v\rangle_M + (c'(x)^\star q)v = \langle \Delta n_{\mathrm{ex}}, v\rangle_M + qc'(x)v = \langle \Delta n_{\mathrm{ex}}, v\rangle_M.$$

The same holds for $\delta n_{\mathrm{ex}} = \nu \Delta n_{\mathrm{ex}}$. □

For the computation of the tangential step direction Δt, we aim to minimize a quadratic model of the objective function $q(\Delta t + \delta n_{\mathrm{ex}})$ in the linearized kernel of the equality constraint, where we define

$$\begin{aligned} q(\delta x) :=& f(x) + f'(x)\delta x + \frac{1}{2}L_{xx}(x, p_x)(\delta x)^2 \\ =& f(x) + f'(x)\delta x + \frac{1}{2}(f''(x) + p_x c''(x))(\delta x)^2 \end{aligned} \tag{7}$$

with $\delta x = \Delta t + \delta n_{\mathrm{ex}}$. Note that δn_{ex} is known beforehand, such that we minimize only over the tangential step Δt. Ignoring globalization issues we define the undamped tangential step as a solution of the quadratic problem

$$\min_{\Delta t \in X} q(\delta n_{\mathrm{ex}} + \Delta t) \quad \text{s.t.} \quad c'(x)\Delta t = 0,$$

which is, after adding $p_x c'(x)\Delta t = 0$, equivalent to

$$\min_{\Delta t \in X} (L_x(x, p_x) + L_{xx}(x, p_x)\delta n_{\mathrm{ex}})\Delta t + \frac{1}{2}L_{xx}(x, p_x)(\Delta t)^2 \quad \text{s.t.} \quad c'(x)\Delta t = 0. \tag{8}$$

Therefore, sufficiently close to the solution $(x^\star, p^\star)$ of the optimal control problem, with $L_{xx}(x, p_x)$ being elliptic on $\ker c'(x)$, the first order optimality conditions for (8) read

$$\begin{pmatrix} L_{xx}(x, p_x) & c'(x)^\star \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta t \\ \Delta p \end{pmatrix} + \begin{pmatrix} L_x(x, p_x) + L_{xx}(x, p_x)\Delta n_{\mathrm{ex}} \\ 0 \end{pmatrix} = 0. \tag{9}$$

Thus, defining our update $(\Delta x, \Delta p) := (\Delta t + \Delta n_{\mathrm{ex}}, \Delta p)$ we observe the equivalence to computing the full Lagrange-Newton step (2). Eventually, we define the second order correction δs_{ex} as a solution to the minimum norm problem

$$\min_{\delta s_{\mathrm{ex}} \in X} \quad \frac{1}{2}\langle \delta s_{\mathrm{ex}}, \delta s_{\mathrm{ex}}\rangle_M \quad \text{s.t.} \quad c(x + \delta x) - c(x) - c'(x)\delta x + c'(x)\delta s_{\mathrm{ex}} = 0. \tag{10}$$

This minimization problem (10) is strongly connected to (5) as we conclude the first order optimality conditions

$$\begin{pmatrix} M & c'(x)^\star \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \delta s_{\mathrm{ex}} \\ q \end{pmatrix} + \begin{pmatrix} 0 \\ r(x, \delta x) \end{pmatrix} = 0, \tag{11}$$

where $r(x, \delta x) := c(x+\delta x) - c(x) - c'(x)\delta x$. Just as the normal step, cf. Lemma 2.2, the simplified normal step satisfies $\ker c'(x)^\perp$. We will denote the solution of above problem (10) by

$$\delta s_{\mathrm{ex}} = -c'(x)^-(c(x + \delta x) - c(x) - c'(x)\delta x) \tag{12}$$

where $v = c'(x)^- r$ denotes the least norm solution of $c'(x)v = r$.

3. Iterative Solution of the Saddle Point Systems

In this section, we present the solution algorithm for the computation of the steps at an iteration point (x, p). We use a conjugate gradient algorithm with a constraint preconditioner which solves the constraint exactly and to which we will refer to as the projected preconditioned conjugate gradient (PPCG) method. We first present the solution method for the computation of the Lagrange multiplier, the normal and the simplified normal step. Second, for the computation of the tangential step with possibly indefinite system operator, we propose a modification of the PPCG-method to allow for indefinite systems.

3.1. **Lagrange multiplier, normal and simplified normal step.** In this subsection we are interested in the linear system

$$\underbrace{\begin{pmatrix} M & c'(x)^\star \\ c'(x) & 0 \end{pmatrix}}_{H_n} + \begin{pmatrix} r_1 \\ 0 \end{pmatrix} = 0 \tag{13}$$

for different $r_1 \in X^\star$, which is to be solved to compute the Lagrange multiplier, the normal and the simplified normal step.

Remark 3.1. We note, that the right-hand-side of the normal step system (6) and the simplified normal step system (11) is not of the above form. In the case of optimal control however, one can transform the system to obtain a right-hand-side $\begin{pmatrix} r_1 \\ 0 \end{pmatrix}$. We will come back to this issue in section 7, when we analyze the case of optimal control.

As we will use a particular conjugate gradient method, we introduce the constraint preconditioner $P : X \times P \to X^\star \times P^\star$

$$P := \begin{pmatrix} \tilde{M} & c'(x)^\star \\ c'(x) & 0 \end{pmatrix}, \tag{14}$$

where $\tilde{M}$ is an approximation of M, e.g. the diagonal of M. As an iterative solution method for a general linear system with a a so called saddle point operator as in (13), we employ the following algorithm, cf. [9], where P is the above defined constraint preconditioner. We now

Algorithm 1 Projected preconditioned conjugate gradient method

Require: $z = \begin{pmatrix} v \\ q \end{pmatrix}$ satisfying $c'(x)v = 0, r = Hz + b, g = P^{-1}r, d = -g$

1: **while** convergence test failed **do**
2: $\sigma \leftarrow r^T g$
3: $\alpha \leftarrow \sigma / d^T H d$ ▷ linesearch along d
4: $z \leftarrow z + \alpha d$ ▷ iterate update
5: $r \leftarrow r + \alpha H d$ ▷ derivative/residual update
6: $g \leftarrow P^{-1} r$ ▷ preconditioning/computation of gradient
7: $\beta \leftarrow r^T g / \sigma$
8: $d \leftarrow -g + \beta d$ ▷ search direction update
9: **end while**
10: **return** z

present several properties of the above algorithm which can be proven straighforwardly.

Proposition 3.2. *Consider Algorithm 1 with constraint preconditioner (14) for the solution of (13). Then*

(1) *all primal components of the CG-iterates are in* $\ker c'(x)$*, and*
(2) H_n *as defined in (3) is positive definite on the iterates* (x_i, p_i) *and* $\|(x_i, p_i)\|_{H_n} = \|x_i\|_M$*, i.e. the energy norm coincides with the Hilbert space norm.*

Proof. The first part follows easily by induction, as the initial iterate is in $\ker c'(x)$, the right-hand-side is of the form $(r_1, 0)$ and the preconditioner yields $g \in \ker c'(x)$ if $r = (r_1, 0)$ for $r_1 \in X^\star$. For the second part, we use that all PPCG-iterates are in $\ker c'(x)$ and hence $(H_n)(x_i, p_i)^2 = \|x_i\|_M^2 + 2p_i c'(x)x_i = \|x_i\|_M^2$. □

3.2. **Tangential step.** For the computation of the tangential step, we aim to solve the possibly indefinite system

$$\underbrace{\begin{pmatrix} L_{xx}(x,p_x) & c'(x)^\star \\ c'(x) & 0 \end{pmatrix}}_{=:H_t} \begin{pmatrix} \Delta t \\ q \end{pmatrix} + \begin{pmatrix} L_x(x,p_x) + L_{xx}(x,p_x)\Delta n \\ 0 \end{pmatrix} = 0.$$

and employ a modified version of Algorithm 1 and use the same preconditioner P as introduced in (14). However far away from the solution, $L_{xx}(x,p)$ can be indefinite and nonconvexities can occur, i.e. search directions d, such that $H_t(d,d) \leq 0$. In this case, we will use the positive definite operator H_n introduced in (3) as regularization. As this operator is positive definite on $\ker c'(x)$ and by the properties of the preconditioner, see Proposition 3.2, we can find $\theta > 0$, such that $H_t + \theta H_n$ is positive definite and restart the iteration. However, far from a local minimum we do not want to compute tangential steps overly accurately and thus we truncate the PPCG-iteration in the spirit of cf. [27, 31]. For an in detail description of different methods for the computation of the tangential step, we refer to [18, Section 4.3].

4. Inexact computation of Steps

In this section, we assume that the normal step, the simplified normal step and the Lagrange multiplier are computed by the PPCG-algorithm, described in the previous section. If we initialize the PPCG-algorithm with an iterate in $\ker c'(x)$, the residual in the second component will be zero, whereas an error term in the first component will remain. We will denote residuals by $\varepsilon_1^{p_x}, \varepsilon_1^{\Delta n}$ and $\varepsilon_1^{\delta s}$ for the linear system of the Lagrange multiplier, the normal step and the simplified normal step respectively. The following equations characterize the solutions of the saddle-point systems solved via the PPCG-method:

$$\begin{pmatrix} M & c'(x)^\star \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} g_x \\ p_x \end{pmatrix} + \begin{pmatrix} f'(x) \\ 0 \end{pmatrix} = \begin{pmatrix} \varepsilon_1^{p_x} \\ 0 \end{pmatrix}, \tag{15}$$

$$\begin{pmatrix} M & c'(x)^\star \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \Delta n \\ q \end{pmatrix} + \begin{pmatrix} 0 \\ c(x) \end{pmatrix} = \begin{pmatrix} \varepsilon_1^{\Delta n} \\ 0 \end{pmatrix} \tag{16}$$

and

$$\begin{pmatrix} M & c'(x)^\star \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} \delta s \\ q \end{pmatrix} + \begin{pmatrix} 0 \\ r(x,\delta x) \end{pmatrix} = \begin{pmatrix} \varepsilon_1^{\delta s} \\ 0 \end{pmatrix}, \tag{17}$$

where $r(x,\delta x) = c(x+\delta x) - c(x) - c'(x)\delta x$.

Due to the constraint preconditioner, Δn and δs still satisfy the underdetermined Newton equations:

$$\begin{aligned} c'(x)\Delta n + c(x) &= 0 \\ c'(x)\delta s + (c(x+\delta x) - c(x) - c'(x)\delta x) &= 0, \end{aligned} \tag{18}$$

however, their minimal norm property and their orthogonality to $\ker c'(x)$ are lost.

We have the following relationship between the inexactly computed quantities δn and δs defined by (16), (17) and the exact solutions δn_{ex} and δs_{ex} characterized by (6), and (11). The difference between the exactly and inexactly computed normal step $\Delta n_{\text{err}} := \Delta n - \Delta n_{\text{ex}}$ fulfills the identity $M\Delta n_{\text{err}} = \varepsilon_1^{\Delta n}$, i.e. Δn_{err} is the primal quantity corresponding to the residual $\varepsilon_1^{\Delta n}$. Similarly, defining $\delta s_{\text{err}} := \delta s - \delta s_{\text{ex}}$, it holds that $M\delta s_{\text{err}} = \varepsilon_1^{\delta s}$.

Moreover, by subtracting (16) from (6) and (17) from (11), we obtain

$$\delta n_{\text{err}} \in \ker c'(x), \quad \delta s_{\text{err}} \in \ker c'(x), \tag{19}$$

respectively, and thus

$$\delta n_{\text{ex}} \perp \delta n_{\text{err}}, \quad \delta s_{\text{ex}} \perp \delta s_{\text{err}}. \tag{20}$$

In the remainder of this section, we recapitulate the globalization scheme from [21], study the effect of inexact computations on this mechanism, and derive appropriate algorithmic measures. This includes modified definitions of the algorithmic quantities and practical termination criteria for the inner solvers.

4.1. **Consistency of the quadratic model.** In this section we study the influence of inexactness on the consistency of the quadratic model $q_x(\delta x)$ with the true function value $f(x+\delta x+\delta s)$. In [21] a third order consistency result has been derived, giving an interpretation of δs as a second-order correction that helps to avoid the Maratos effect.

Theorem 4.1. *Assume that there are constants $\omega_c, \omega_{f'}$ and ω_L, such that*

$$\|c'(x)^-(c'(x+v)-c'(x))v\| \le \omega_c\|v\|^2, \tag{21}$$

$$|(L_{xx}(x+v,p_x)-L_{xx}(x,p_x))(v,v)| \le \omega_L\|v\|^3, \tag{22}$$

$$|(f'(x+v)-f'(x))w| \le \omega_{f'}\|v\|\|w\|, \tag{23}$$

where (x,p_x) are taken among the iterates, and v,w are arbitrary. Then, for arbitrary δx and corresponding inexact and exact simplified normal steps δs and δs_{ex}, where $\frac{\|\delta s_{err}\|}{\|\delta s_{ex}\|} < \gamma$ for $\gamma > 0$, we have the estimates:

$$\|\delta s_{ex}\| \le \frac{\omega_c}{2}\|\delta x\|^2 \tag{24}$$

$$\|\delta s\| \le \frac{\tilde{\omega}_c}{2}\|\delta x\|^2, \quad \textit{where } \tilde{\omega}_c := \sqrt{1+\gamma^2}\omega_c \tag{25}$$

$$|f(x+\delta x+\delta s)-q(\delta x)-\varepsilon_1^{p_x}\delta s+\varepsilon_1^{\delta s}g_x| \le \left(\frac{\omega_L}{6}+\frac{\omega_{f'}\tilde{\omega}_c}{2}\left(1+\frac{\tilde{\omega}_c}{4}\|\delta x\|\right)\right)\|\delta x\|^3, \tag{26}$$

Here, $q(\delta x)$ is the quadratic model as defined in (7), and $\varepsilon_1^{p_x}$, δs, $\varepsilon_1^{\delta s}$, and g_x are defined via (15) *and* (17).

Proof. Inequality (24) has been shown in [21, Theorem 3.4]. Since $\delta s_{\text{ex}} \perp \delta s - \delta s_{\text{ex}}$, we conclude by the Pythagoras theorem:

$$\|\delta s\| = \sqrt{\|\delta s_{\text{ex}}\|^2 + \|\delta s_{\text{err}}\|^2} \le \sqrt{1+\gamma^2}\|\delta s_{\text{ex}}\|,$$

which yields (25).

Finally, (26) can be shown using the identity

$$L_x(x,p_x)\delta s = f'(x)\delta s + p_x c'(x)\delta s = \varepsilon_1^{p_x}\delta s - \varepsilon_1^{\delta s}v$$

and proceeding analogously to the proof of [21, Theorem 3.4]. □

The left hand side of (26) yields a measure for the nonlinearity of f. The term $-\varepsilon_1^{p_x}\delta s+\varepsilon_1^{\delta s}g_x$ can be evaluated algorithmically and represents the effect of the inexactness on the quality of the quadratic model. The right hand side consists of higher order terms only and is independent of the inexactness.

4.2. **Globalization Strategy.** In this section, we sketch the globalization mechanism for the algorithm. More details can be found in [21]. One popular globalization strategy is to measure $\|c(x)\|$ and require descent in some sense. This leads to merit functions and filter methods.

However, our premise of affine covariance does not permit the evaluation of residual norms. We therefore employ an idea coming from affine covariant Newton methods, cf. [11]. A point x is considered to be sufficiently close to the feasible set, if a Newton algorithm for $c(x) = 0$ started at x yields fast convergence. By their definition, the normal and the simplified normal step satisfy, in the absence of damping,

$$\begin{aligned} c(x) + c'(x)\delta n &= 0, \\ c(x + \delta x) + c'(x)\delta s &= 0 \end{aligned}$$

and thus can be interpreted as two steps of a simplified Newton algorithm starting at x. Motivated by this observation, we introduce the contraction factor

$$\Theta(\delta x) := \frac{\|\delta s\|}{\|\delta x\|}$$

which allows us to quantify the vicinity of the feasible region: If $\Theta(\delta x)$ is small, $\|\delta s\| \ll \|\delta x\|$ and hence fast local convergence of the Newton scheme takes place. In the general case, if we have a damping factor $\nu < 1$ for the normal step, we use the same argumentation for the relaxed problem $x(\xi_\nu) = (1 - \nu)c(x)$, cf. [21, Section 4.1]. Hence, we will accept δx if $\Theta(\delta x) \leq \Theta_{\text{acc}}$ is observed, where $0 < \Theta_{\text{acc}} < 1$ is given. By Theorem 4.1 we have the estimate

$$\Theta(\delta x) \leq \frac{\tilde{\omega}_c}{2}\|\delta x\|, \tag{27}$$

where we, however, have to estimate the theoretical quantity $\tilde{\omega}_c$ by an estimate $[\tilde{\omega}_c]$ and hence introduce

$$[\Theta](\xi) := \frac{[\tilde{\omega}_c]}{2}\|\xi\|,$$

where $[\tilde{\omega}_c]$ is an estimate from below for $\tilde{\omega}_c$ defined in Theorem 4.1. The full step δx is then computed, such that

$$[\Theta](\delta x) := \frac{[\tilde{\omega}_c]}{2}\|\delta x\| \leq \Theta_{\text{aim}},$$

and where Θ_{aim} is chosen in $]0, \Theta_{\text{acc}}[$. This can be interpreted as a trust-region constraint

$$\|\delta x\| \leq \Gamma := \frac{2\Theta_{\text{aim}}}{[\tilde{\omega}_c]}. \tag{28}$$

After computing a step δx and the corresponding δs, we update the Lipschitz estimate $[\tilde{\omega}_c]$ via

$$[\tilde{\omega}_c] := \frac{2\Theta(\delta x)}{\|\delta x\|} = \frac{2\|\delta s\|}{\|\delta x\|^2}, \tag{29}$$

where it follows from (25) that $[\tilde{\omega}_c] \leq \tilde{\omega}_c$.

Damping of the normal step. Given a normal step direction Δn, we compute a damping factor ν, such that

$$\|\nu\Delta n\| \leq \Gamma_n := \frac{2\rho_{\text{elbow}}\Theta_{\text{aim}}}{[\tilde{\omega}_c]} \tag{30}$$

where $\rho_{\text{elbow}} \in]0, 1[$ is an "elbow-space" parameter for the tangential step δt to prevent δn from occupying the the full trust region radius.

Damping of the tangential step. For the globalization of tangential step, we employ a cubic regularization approach as proposed in [6] for unconstrained problems, i.e. we minimize the cubic model

$$\begin{aligned} m_{[\omega_f]} &:= q(\delta x) + \frac{[\omega_f]}{6}\|\delta x\|^3 \\ &= f(x) + f'(x)\delta x + \frac{1}{2}L_{xx}(x,p)(\delta x)^2 + \frac{[\omega_f]}{6}\|\delta x\|^3 \end{aligned}$$

over $\delta t \in X$ under the conditions

$$\begin{aligned} c'(x)\delta t &= 0 \\ \text{and} \qquad \frac{[\tilde{\omega}_c]}{2}\|\delta x\| &= \frac{[\tilde{\omega}_c]}{2}\|\delta n + \delta t\| \leq \Theta_{\text{aim}}, \end{aligned}$$

where $[\omega_f]$ is an affine covariant estimate of the prefactor on the right hand side of (26). To this end, we will first compute a direction Δt, which minimizes the quadratic model (7) on $\ker c'(x)$. With this direction Δt, we employ a simple linesearch method for the cubic model, i.e.

$$\tau := \arg\min_t m_{[\omega_f]}(\nu\Delta n + t\Delta t) \tag{31}$$

$$\text{s.t.} \quad \|\delta n + \tau\Delta t\| \leq \Gamma. \tag{32}$$

Remark 4.2. For the exactly computed normal step δn_{ex}, we have $\|\delta n_{\text{ex}} + \delta t\|^2 = \nu^2\|\delta n_{\text{ex}}\|^2 + \tau^2\|\delta t\|^2$ and hence monotonicity of $\|\delta x\|$ in τ. This does not hold for inexact step computations, as certainly $\delta t \in \ker c'(x)$, but $\delta n \notin \ker c'(x)^\perp$.

4.3. **Influence of errors on algorithmic parameters.** In this section, we will discuss the influence of the residuals defined by (15), (16) and (17) on the quantities $\Theta(\delta x)$, $[\tilde{\omega}_c]$ and $[\omega_f]$. The error will be again denoted by the subscript *err*, i.e. $\delta s_{\text{err}} := \delta s - \delta s_{\text{ex}}$, where δs solves (17) and δs_{ex} solves (11). By Lemma 2.2, it holds that $\delta s_{\text{ex}} \in \ker c'(x)^\perp$. This orthogonality property allows us to estimate the influence of the relative error on the contraction factor $\Theta(\delta x)$.

Lemma 4.3. *Let* $\frac{\|\delta s_{err}\|}{\|\delta s\|} < \gamma$ *and* $\gamma > 0$. *Then with* $\Theta(\delta x) = \frac{\|\delta s\|}{\|\delta x\|} = \frac{\|\delta s_{ex} + \delta s_{err}\|}{\|\delta x\|}$ *and* $\Theta_{ex}(\delta x) := \frac{\|\delta s_{ex}\|}{\|\delta x\|}$, *we have*

$$\sqrt{1-\gamma^2}\Theta(\delta x) < \Theta_{ex}(\delta x) \leq \Theta(\delta x).$$

and

$$\sqrt{1-\gamma^2}[\tilde{\omega}_c] < [\omega_c] \leq [\tilde{\omega}_c].$$

where $[\omega_c]$ *is defined by definition (29) with* δs_{ex} *instead of* δs.

Proof. By using the orthogonality of δs_{ex} and δs_{err}, we have $\|\delta s\|^2 = \|\delta s_{\text{ex}}\|^2 + \|\delta s_{\text{err}}\|^2$ and thus $\gamma^2 > \frac{\|\delta s_{\text{err}}\|^2}{\|\delta s\|^2} = \frac{\|\delta s_{\text{err}}\|^2}{\|\delta s_{\text{ex}}\|^2 + \|\delta s_{\text{err}}\|^2}$

$$\frac{\Theta_{\text{ex}}^2(\delta x)}{\Theta^2(\delta x)} = \frac{\|\delta s_{\text{ex}}\|^2}{\|\delta s_{\text{ex}} + \delta s_{\text{err}}\|^2} = \frac{\|\delta s_{\text{ex}}\|^2 + \|\delta s_{\text{err}}\|^2 - \|\delta s_{\text{err}}\|^2}{\|\delta s_{\text{ex}}\|^2 + \|\delta s_{\text{err}}\|^2} = 1 - \frac{\|\delta s_{\text{err}}\|^2}{\|\delta s_{\text{ex}}\|^2 + \|\delta s_{\text{err}}\|^2} > 1 - \gamma^2.$$

For the second inequality, we conclude

$$\frac{\Theta_{\text{ex}}^2(\delta x)}{\Theta^2(\delta x)} = \frac{\|\delta s_{\text{ex}}\|^2}{\|\delta s_{\text{ex}}\|^2 + \|\delta s_{\text{err}}\|^2} \leq \frac{\|\delta s_{\text{ex}}\|^2}{\|\delta s_{\text{ex}}\|^2} = 1.$$

The estimate for $[\tilde{\omega}_c]$ immediately follows by its definition (29). □

As a consequence of Lemma 4.3, the contraction factor and the Lipschitz estimate of the constraint will always be overestimated, yet only by a factor depending on the relative accuracy in the simplified normal step computation.

For the influence of inexactness on the second Lipschitz estimate $[\omega_f]$ we recall the estimate Theorem 4.1, inequality (26). Therefore, we update $[\omega_f]$ via

$$[\omega_f] := \frac{6}{\|\delta x\|^3} \left(f(x + \delta x + \delta s) - q(\delta x) - \varepsilon_1^{p_x} \delta s + \varepsilon_1^{\delta s} v \right) \tag{33}$$

to obtain an estimate for the nonlinearity independent of the residuals and $\mathrm{d}v$, $\varepsilon_1^{p_x}$, δs, $\varepsilon_1^{\delta s}$ are given by (3) and (17) respectively. This quantity is computable, after δs has been computed. In case of exact computations we obtain the same quantity as in [21].

Acceptable decrease. We employ the strategy described in [21, Section 4.4]. The acceptance criterion for decrease in the functional value is motivated by unconstrained cubic regularization approaches, see [6], replacing $m_{[\omega_f]}(0)$ by $m_{[\omega_f]}(\delta n)$:

$$\eta := \frac{f(x + \delta x + \delta s) - m_{[w_f]}(\delta n)}{m_{[w_f]}(\delta x) - m_{[w_f]}(\delta n)}.$$

We accept a tangential step if

$$\eta \geq \underline{\eta} \tag{34}$$

for a user-defined $\underline{\eta} \in]0, 1[$. Combining these ideas, we arrive at the following algorithm.

Algorithm 2 Raw outline of the composite step algorithm

Require: initial iterate x, $[\tilde{\omega}_c]$, $[\omega_f]$.
  repeat// NLP-loop
    repeat// step computation loop
      compute Lagrange multiplier p_x via (4)
      compute normal step direction Δn via (6)
      compute normal step damping factor ν via (30)
      compute tangential step direction Δt (9)
      compute damping factors via (31) and (32)
      compute simplified normal step δs via (11)
      update Lipschitz constants $[\tilde{\omega}_c], [\omega_f]$ via (29) and (33)
    until δx accepted
  until converged

4.4. **Influence of errors on damping parameters.** In this section, we will analyze the influence of the inexact normal step computation on the damping parameters ν and τ. First, as the normal step is computed as a minimum norm solution, we have that the inexactly computed step satisfies $\|\Delta n\| \geq \|\Delta n_{\mathrm{ex}}\|$ due to the nesting of the Krylov spaces in the CG-method. Thus, the normal step damping parameter defined in (30) will increase with a decreasing residual of the PPCG-algorithm used for the solution of the minimum norm problem. In this section, a bound for the normal step damping factor ν depending on the relative error will be given. Secondly, a longer normal step could decrease the tangential step damping factor due to the trust region constraint (28) on the total step. Therefore, we will to quantify this decrease of the tangential damping factor and formulate an adaptive stopping criterion of the PPCG-iteration for the normal step system. Since the computation of the tangential step is done after the normal step, the stopping criterion has to be independent of the tangential step.

We will again denote the error by Δn_{err}, where $\Delta n_{\text{err}} := \Delta n - \Delta n_{\text{ex}}$ and the inexactly computed normal step Δn is defined by (16), whereas the exactly normal step Δn_{ex} solves (6). After a direction Δn is computed, a damping factor is chosen via

$$\nu = \min\{1, \frac{2\rho_{\text{elbow}}\Theta_{\text{aim}}}{[\tilde{\omega}_c]\|\Delta n\|}\},$$

where $\rho_{\text{elbow}} \in]0, 1[$ and $\Theta_{\text{aim}} \in]0, \Theta_{\text{acc}}[$ are algorithmic quantities and $[\tilde{\omega}_c]$ is our estimate of the Lipschitz constant of the equality constraint. Furthermore, we will denote the damping factor of the exact normal step by

$$\nu_{\text{ex}} = \min\{1, \frac{2\rho_{\text{elbow}}\Theta_{\text{aim}}}{[\tilde{\omega}_c]\|\Delta n_{\text{ex}}\|}\}.$$

Influence on damping of the normal step. First, we derive a result similar to Lemma 4.3 which bounds the fraction of the damping parameters by the relative error.

Lemma 4.4. *Let $\Gamma\rho_{elbow} \leq \|\Delta n_{ex}\|$, i.e. normal step damping would occur with exact step computation and with inexact step computation as $\|\Delta n\| \geq \|\Delta n_{ex}\|$. Then*

$$\frac{\nu}{\nu_{ex}} = \frac{1}{\sqrt{1 + \frac{\|\Delta n_{err}\|^2}{\|\Delta n_{ex}\|^2}}}. \tag{35}$$

Proof. We compute using $\Delta n_{\text{err}} \perp \Delta n_{\text{ex}}$ and Pythagoras:

$$\frac{\nu_{\text{ex}}^2}{\nu^2} = \frac{\|\Delta n\|^2}{\|\Delta n_{\text{ex}}\|^2} = 1 + \frac{\|\Delta n_{\text{err}}\|^2}{\|\Delta n_{\text{ex}}\|^2}.$$

□

Influence on damping of the tangential step. The composite step $\delta x = \nu\Delta n + \tau\Delta t$ has to fullfill the trust region bound

$$\|\delta x\| \leq \frac{2\Theta_{\text{aim}}}{[\tilde{\omega}_c]} =: \Gamma. \tag{36}$$

To achieve this, after computation of the normal step damping factor ν, a damping factor τ is chosen such that

$$\Gamma^2 \geq \|\tau\Delta t + \delta n\|^2 = \tau^2\|\Delta t\|^2 + 2\tau\langle\Delta t, \delta n\rangle + \|\delta n\|^2.$$

For this inequality, we obtain an upper bound for τ with the solution of

$$\bar{\tau} = \frac{-\langle\Delta t, \delta n\rangle + \sqrt{\langle\Delta t, \delta n\rangle^2 + \|\Delta t\|^2(\Gamma^2 - \|\delta n\|^2)}}{\|\Delta t\|^2} \tag{37}$$

as $\|\delta n\| < \Gamma$ in every case by definition of the normal step damping.

We consider the case where both Δn and Δn_{ex} remain undamped, which is the case if $\Gamma\rho_{\text{elbow}} > \|\Delta n\|$ and therefore also $\Gamma\rho_{\text{elbow}} > \|\Delta n_{\text{ex}}\|$ due to $\|\Delta n_{\text{ex}}\| \leq \|\Delta n\|$. First, we derive a lower bound on the tangential step damping factor following an inexactly computed normal step.

Lemma 4.5. *Let $\bar{\tau}$ be defined by (37). Then*

$$\bar{\tau} \geq \frac{-\|\Delta n_{err}\| + \sqrt{\Gamma^2 - \|\Delta n\|^2}}{\|\Delta t\|}.$$

Proof. Using $\Delta n_{\text{ex}} \in \ker c'(x)^\perp, \Delta t \in \ker c'(x)$ and the Cauchy-Schwarz inequality, we get

$$-\langle \Delta t, \Delta n\rangle = -\langle \Delta t, \Delta n_{\text{err}}\rangle \geq -\|\Delta t\|\|\Delta n_{\text{err}}\|. \tag{38}$$

Then we can bound the numerator of (37) from below by

$$\begin{aligned}
&-\langle \Delta t, \Delta n\rangle + \sqrt{\langle \Delta t, \Delta n\rangle^2 + \|\Delta t\|^2(\Gamma^2 - \|\Delta n\|^2)} \\
\geq &-\|\Delta t\|\|\Delta n_{\text{err}}\| + \sqrt{\langle \Delta t, \Delta n\rangle^2 + \|\Delta t\|^2(\Gamma^2 - \|\Delta n\|^2)} \\
\geq &-\|\Delta t\|\|\Delta n_{\text{err}}\| + \sqrt{\|\Delta t\|^2(\Gamma^2 - \|\Delta n\|^2)} = \|\Delta t\|\left(-\|\Delta n_{\text{err}}\| + \sqrt{(\Gamma^2 - \|\Delta n\|^2)}\right).
\end{aligned}$$

Cancelling the term $\|\Delta t\|$ yields the result. □

The above estimate includes the norm of the tangential step, which is not at hand when computing the normal step. Therefore, we present a bound independently of the tangential step in the following Lemma. To eliminate the dependence on $\|\Delta t\|$, we need to assume, that the tangential step norm following the inexact normal step and the tangential step norm following the exact normal step is the same, i.e. the norm of the solution of

$$\min_{\delta t \in X} q(\delta n + \Delta t) \quad \text{s.t.} \quad c'(x)\Delta t = 0 \tag{39}$$

is the same for preceding normal steps δn and δn_{ex}. This would be the case if the derivatives are the same, i.e. $q'(\delta n) = q'(\delta n_{\text{ex}})$. However, this can neither be verified nor estimated since the influence of the error part of δn on the quadratic model is not clear. We only have an estimate of the norm of δn_{err} during the PPCG process.

Lemma 4.6. *Assume that the norm of the tangential step defined by (39) is the same for a preceding exact and inexact normal step. Then*

$$\frac{\bar\tau}{\bar\tau_{ex}} \geq \frac{-\|\Delta n_{err}\| + \sqrt{\Gamma^2 - \|\Delta n\|^2}}{\sqrt{\Gamma^2 - \|\Delta n_{ex}\|^2}}. \tag{40}$$

Proof. With Lemma 4.5 we compute

$$\frac{\bar\tau}{\bar\tau_{\text{ex}}} \geq \frac{\frac{-\|\Delta n_{\text{err}}\| + \sqrt{\Gamma^2 - \|\Delta n\|^2}}{\|\Delta t\|}}{\frac{\sqrt{\Gamma^2 - \|\Delta n_{\text{ex}}\|}}{\|\Delta t\|}} = \frac{-\|\Delta n_{\text{err}}\| + \sqrt{\Gamma^2 - \|\Delta n\|^2}}{\sqrt{\Gamma^2 - \|\Delta n_{\text{ex}}\|^2}}.$$

□

This bound on the ratio of the damping factors is independent of the computed tangential step. This is crucial, as we want to employ an adaptive stopping criterion for the computation of the normal step with the PPCG-Method, where we do not have the quantity Δt at hand. Furthermore, we can not readjust the accuracy of the normal step by performing some more PPCG-iterations for the normal step after the computation of the tangential step as the right-hand side depends on Δn.

Remark 4.7. Rewriting the bound of Lemma 4.6 yields

$$\frac{\bar\tau}{\bar\tau_{\text{ex}}} \geq \frac{\sqrt{\Gamma^2 - \|\Delta n\|^2}}{\sqrt{\Gamma^2 - \|\Delta n_{\text{ex}}\|^2}} - \frac{\|\Delta n_{\text{err}}\|}{\sqrt{\Gamma^2 - \|\Delta n_{\text{ex}}\|^2}} \tag{41}$$

and using boundedness of Γ by $\|\delta x\|$ from below, see (36), we observe that if $\frac{\|\Delta n\|}{\|\delta x\|}$ approaches zero, which usually happens towards the end of the solution process near the optimal value of the Newton iteration, the first term approaches one from below and the last term approaches zero from above for a fixed error.

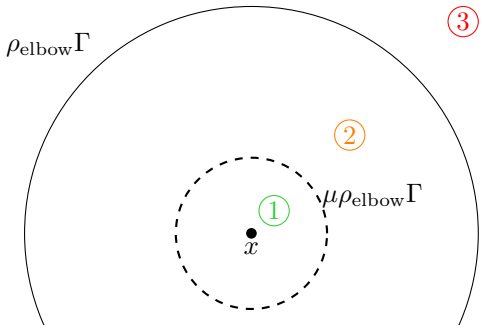


FIGURE 2. Depiction of the division of the space into three regions around the outer iteration point x.

5. Adaptive Termination Criteria

In the previous section we studied, how inexactness influences the damping factors, chosen by our algorithm. Based on this information we will now formulate adaptive stopping criteria for the computation of the normal step and the simplified normal step. The idea is to find a trade-off between early termination and the requirement that ν and $\tilde{\tau}$ do not deteriorate too much, according to (35) and (40).

5.1. **Normal step.** Motivated by the considerations of subsection 4.4, we can formulate two adaptive stopping criteria for the PPCG-method applied to the linear system of the normal step as alternatives to the pure relative error termination criterion one usually has for CG-methods. Recalling the definition of the normal step damping factor

$$\nu = \min\{1, \frac{\rho_{\text{elbow}}\Gamma}{\|\Delta n\|}\}, \qquad \text{where} \qquad \Gamma = \frac{2\Theta_{\text{aim}}}{[\tilde{\omega}_c]},$$

we can rewrite this as a trust-region constraint

$$\nu\|\Delta n\| \leq \rho_{\text{elbow}}\Gamma.$$

Moreover, with $\mu \in]0,1[$, we define a stricter trust region with radius $\mu\rho_{\text{elbow}}\Gamma$. We will consider three cases for the length of the normal step, illustrated in Figure 2. First, if normal steps Δn and Δn_{ex} are very small and contained in the stricter trust-region, we will stop the PPCG-iteration since the normal step has very little influence and might not even be necessary in this iteration as the feasible region is very close. Therefore, it is natural not to spend too much effort in computing this very small step.

Secondly, we discuss the case where the normal step is contained in the trust-region with radius $\rho_{\text{elbow}}\Gamma$ and hence remains undamped. In this context, we will refer to subsection 4.4, more specifically formula (41). The term

$$\frac{\sqrt{\Gamma^2 - \|\Delta n\|^2}}{\sqrt{\Gamma^2 - \|\Delta n_{\text{ex}}\|^2}} - \frac{\|\Delta n_{\text{err}}\|}{\sqrt{\Gamma^2 - \|\Delta n_{\text{ex}}\|^2}}$$

approaches one from below during the iterative solution process and gives a lower bound to the fraction $\frac{\bar{\tau}}{\bar{\tau}_{\text{ex}}}$. Therefore, we aim to iterate the PPCG-Algorithm as long as the inexactness of the normal step occupies too much space from the total step due to the trust-region constraint imposed on the total step. More precisely, we stop the iteration if the above estimate yields an acceptable bound on the fraction of the maximal tangential step damping parameters.

Thirdly, we consider the case of the normal step leading towards the outside of the trust-region. In this case, the normal step will be damped and we do not want to exit the PPCG-iteration early as we are away from the feasible region. This is motivated by our idea, stressing feasibility first, when we are far away from the feasible region. Moreover, the premise on the estimate for the tangential step damping factor was that the norm of the computed tangential step is similar for both Δn and Δn_{ex}, see (39). For a large normal step at a iteration point with a small trust region, this assumption might be violated if Δn and Δn_{ex} differ by a large amount. For those reasons, the only termination criterion for the PPCG-method is a relative error criterion, where we aim to satisfy

$$\frac{\|\Delta n_{\text{err}}\|}{\|\Delta n_{\text{ex}}\|} < \gamma$$

for a desired accuracy $\gamma > 0$.

Remark 5.1. Note that the second and the third case use the quantities $\|\delta n_{\text{ex}}\|$ and $\|\delta n_{\text{err}}\|$ which have to be estimated. This can be done by methods of error estimation for the conjugate gradient method, i.e. [1, 28, 29] which, however, use the current CG-iterate as an approximation of the exact solution. These algorithms often include warm-up phases with no error estimate. Therefore, as long as no error estimate is available, we use a worst case estimate for (41) and assume that $\Delta n_{\text{ex}} = 0$ and hence $\Delta n_{\text{err}} = \Delta n$.

Algorithm 3 PPCG-iteration for the normal step Δn

Require: initial offset Δn_0, $\eta_\tau \in]0, 1[$, $\gamma > 0$.
  repeat// PPCG-loop
    update cg-iterate $\Delta \tilde{n}_k$
    compute $\|\Delta n_k\| = \|\Delta n_0 + \Delta \tilde{n}_k\|$
    if $\|\Delta n_k\| \leq \mu \Gamma \rho_{\text{elbow}}$ **then** // in region ①
      break
    end if
    if $\|\Delta n_k\| \leq \Gamma \rho_{\text{elbow}}$ **then** // in region ① or ②
      if have error estimate **then**
        $l \leftarrow \frac{\sqrt{\Gamma^2 - \|\Delta n\|^2}}{\sqrt{\Gamma^2 - \|\Delta n_{\text{ex}}\|^2}} - \frac{\|\Delta n_{\text{err}}\|}{\sqrt{\Gamma^2 - \|\Delta n_{\text{ex}}\|^2}}$ // see (41)
      else
        $l \leftarrow \frac{\sqrt{\Gamma^2 - \|\Delta n\|^2}}{\Gamma} - \frac{\|\Delta n\|}{\Gamma}$
      end if
      if $\eta_\tau \leq l$ **then** // $\frac{\tau}{\tau_{\text{ex}}}$ large enough
        break
      end if
    end if
  until $\frac{\|\Delta n_{\text{err}}\|}{\|\Delta n_{\text{ex}}\|} < \gamma$ // in region ①, ② or ③.

5.2. **Simplified normal step.** The simplified normal step fulfills two tasks in our context. First, to avoid the Maratos effect and second, to assist the globalization. By (25) we see that close to the solution, the simplified normal steps get very small. Therefore, we want to avoid to spend too much effort on guaranteeing orthogonality of a small step whose purpose is mainly for globalization. The monotonicity properties of the PPCG-method yield a decrease of the norm of δs, i.e. also a decrease of $\Theta(\delta x) := \frac{\|\delta s\|}{\|\delta x\|}$ in every PPCG-iteration. Motivated by these

Algorithm 4 Summary of composite step algorithm with inexact step computations

Require: initial iterate $x, [\tilde{\omega}_c], [\omega_f], \gamma, \overline{\gamma}_{\Delta t,p_x}, \underline{\gamma}_{\Delta t,p_x}, \Theta_{\delta s}, \eta_\tau$
  repeat// NLP loop
    compute Δn by the solution of (6) with rel. error γ and Algorithm 3 with η_τ for the alternative stopping criteria.
    $\nu \leftarrow \min\{1, \frac{2\rho_{\text{elbow}}\Theta_{\text{aim}}}{[\tilde{\omega}_c]}\}$
    if $\nu < 1$ **then**
      $\gamma_{\Delta t,p_x} \leftarrow \overline{\gamma}_{\Delta t,p_x}$
    else
      $\gamma_{\Delta t,p_x} \leftarrow \underline{\gamma}_{\Delta t,p_x}$
    end if
    compute p_x by the solution of (4) with rel. error $\gamma_{\Delta t,p_x}$.
    compute Δt by the solution of (9) with rel. error $\min\{\Theta(\delta s), \gamma_{\Delta t,p_x}\}$.
    repeat// Inner loop
      $DiscardTangentialStep \leftarrow false$
      $\nu \leftarrow \min\{1, \frac{2\rho_{\text{elbow}}\Theta_{\text{aim}}}{[\tilde{\omega}_c]}\}$
      compute τ via (31) and (32).
      $\delta x \leftarrow \nu\Delta n + \tau\Delta t$
      compute δs by the solution of (11) with rel. error γ and the alternative stopping criterion with $\Theta_{\delta s}$ as in Subsection 5.2.
      update $[\tilde{\omega}_c]$ via (29), $[\omega_f]$ via (33).
      if Stagnating update of $[\omega_f]$ **then**
        $DiscardTangentialStep \leftarrow true$
        $\delta x \leftarrow \nu\Delta n$
      end if
    until $\Theta(\delta x) \leq \Theta_{\text{acc}} \wedge ((34) \vee DiscardTangentialStep)$
    $x \leftarrow x + \delta x + \delta s$
  until converged

observations, we will define $\Theta_{\delta s} \in]0, \Theta_{\text{acc}}]$ and terminate the PPCG-iteration for δs if $\Theta(\delta x) \leq \Theta_{\delta s}$. This assures, that the simplified normal step is just computed exactly enough, i.e. its norm is small enough, such that the step δx is not rejected. If we do not fall below this bound during the PPCG-iteration, we only terminate via a relative error criterion.

5.3. **Tangential step.** For the tangential step, we use another adaptive mechanism. If a strong damping of the total step Δx is to be expected, we will increase the tolerance for the relative error of the undamped tangential step Δt and for the Lagrange multiplier correction δp. This is motivated by the construction of the algorithm which focuses on feasibility first. Therefore, we will consider two parameters, $0 < \underline{\gamma}_{\Delta t,p_x} < \overline{\gamma}_{\Delta t,p_x} < 1$, which we will set as relative accuracy, depending on the normal step being damped. If the normal step is damped, we will choose $\overline{\gamma}_{\Delta t,p_x}$ as relative accuracy and hence obtain a less exact solution. For the other case of an undamped normal step, we pick the more accurate solution of the tangential step system with $\underline{\gamma}_{\Delta t,p_x}$.

Additionally, as it will be shown in the proof of convergence in Section 6, we need the error of the Newton step to approach zero as we converge to the solution. To achieve this, we will set the final relative accuracy for the iterative solver to $\min\{\Theta(\delta x), \gamma_{\Delta t,p_x}\}$, where $\gamma_{\Delta t,p_x}$ is chosen as stated above and δx is the previous step. This results in the desired convergence of the error

since $\Theta(\delta x) = \frac{\|\delta s\|}{\|\delta x\|}$ and $\|\delta s\| = o(\|\delta x\|)$. Moreover do not need the normal step error norm to approach zero because we solve the Newton equation $c(x) + c'(x)\Delta n = 0$ exactly.

6. Local Convergence analysis

In this section, the proof of local superlinear convergence given in [21, Section 6] is adapted to fit the setting with inexact step computations. It is shown that only the relative error of the tangential step system has to approach zero to guarantee local superlinear convergence, whereas the relative error for the computation of the Lagrange multiplier has to be constant. In addition, it is shown that the norm minimizing property of the normal step does not contribute to the Newton step as a whole. Thus, all normal steps that satisfy the Newton equation $c(x)+c'(x)\delta n = 0$ guarantee local superlinear convergence. Eventually, when looking at the convergence of the damping parameters in the third part of this section, we observe that if the relative error of the normal step is constant, we achieve convergence of our computed step $\|\Delta n\| \to 0$.

6.1. **Transition to local superlinear convergence.**

In this subsection, we show local superlinear convergence of the Newton scheme. In the course of this, we assume sufficient smoothness and second order sufficient optimality conditions at the local minimzer. In the absence of damping, the tangential step Δt is computed as a solution of (9). Furthermore, let Δn be an arbitrary normal step satisfying

$$c(x) + c'(x)\Delta n = 0.$$

Then $\Delta x = \Delta t + \Delta n$ satisfies the Newton equation (2). It was shown in [21, Proposition 6.2] that the iteration

$$(x_{k+1}, w) = (x_k, p_{x_k}) - L''(x_k, p_{x_k})^{-1} L'(x_k, p_{x_k})$$

converges locally superlinearly. Note that this is not the general Newton-Lagrange-Method as w is a dummy variable that is discarded after each update. However, with the help of the implicit function theorem, one can show that the above iteration converges anyway. In the following, p will be the inexactly computed Lagrange multiplier fulfilling (15), whereas $p_\star$ denotes the Lagrange multiplier at the optimal value $(x_\star, p_\star)$ of the optimization problem (1).

Lemma 6.1. *Let f' and c' depend continuously on x and assume that $c'(x_\star) : X \to P^\star$ is a bounded and surjective operator. Let (v, p_{ex}) be the solution of*

$$\begin{pmatrix} M & c'(x)^\star \\ c'(x) & 0 \end{pmatrix} \begin{pmatrix} v \\ p_{ex} \end{pmatrix} = \begin{pmatrix} f'(x) \\ 0 \end{pmatrix}.$$

Then p_{ex} is given as a continuous implicit function of x in a neighborhood of $x_\star$ and

$$\|p_{ex} - p_\star\| = \omega(\|x - x_\star\|).$$

Proof. See [21, Lemma 2.4]. □

Lemma 6.2. *Let the assumptions of Lemma 6.1 hold. Then, if $c''(x)$ is bounded, $L_{xx}(x, p_\star)$ is positive definite and $\|p - p_\star\| = \omega(\|x - x_\star\|)$, it holds that*

$$\|(L''(x,p)^{-1}L'(x,p))_x - (L''(x,p_\star)^{-1}L'(x,p_\star))_x\| \le \|\Delta x\|\omega(\|x - x_\star\|).$$

Proof. Similar to [21, Lemma 6.1], we conclude the inequality

$$\alpha\|\Delta x - \Delta x_\star\|^2 \le -(p - p_\star)c''(x)(\Delta x, \Delta x - \Delta x_\star), \tag{42}$$

where $\Delta x_\star = (L''(x,p_\star)^{-1}L'(x,p_\star))_x$ and $\Delta x = (L''(x,p)^{-1}L'(x,p))_x$. By taking norms we obtain

$$-(p - p_\star)c''(x)(\Delta x, \Delta x - \Delta x_\star) \le \|p - p_\star\|\|c''(x)\|_{X\times X\to P^\star}\|\|\Delta x\|\|\Delta x - \Delta x_\star\|.$$

With (42), the claim follows by cancelling $\|\Delta x - \Delta x_\star\|$. □

In the following, we will consider an inexact tangential step $\Delta t = \Delta t_{\text{ex}} + \Delta t_{\text{err}}$, where Δt_{ex} is the solution of (9). Since we use a constraint preconditioner, we have $\Delta t,\ \Delta t_{\text{err}} \in \ker c'(x)$.

Theorem 6.3. *Assume that the assumptions of Lemma 6.2 hold. Additionally, let $L''(x, p_\star)$ be continuous w.r.t. x, and $\frac{\|\Delta t_{err}\|}{\|\Delta x\|} = \omega(\|x - x_\star\|)$. Then, the method converges local superlinearly, i.e.*

$$\|x^+ - x_\star\| = o(\|x - x_\star\|).$$

Proof. The proof follows similar to [21, Proposition 6.2], as we compute

$$\begin{aligned} x^+ - x_\star =& (x^+ - x) + (x - x_\star) = \Delta x - (x - x_\star) \\ =& -(L''(x,p)^{-1}L'(x,p))_x + \Delta t_{\text{err}} \ - (x - x_\star, 0)_x \\ =& -\left[(L''(x,p)^{-1}L'(x,p))_x - (L''(x,p_\star)^{-1}L'(x,p_\star))_x\right] + \Delta t_{\text{err}} \\ & - \left(L''(x,p_\star)^{-1}\left(L'(x,p_\star) - L'(x_\star,p_\star) + L''(x,p_\star)(x - x_\star, 0)\right)\right)_x. \end{aligned}$$

The first term can be estimated with Lemma 6.2:

$$\left[(L''(x,p)^{-1}L'(x,p))_x - (L''(x,p_\star)^{-1}L'(x,p_\star))_x\right] \leq \|\Delta x\|\omega(\|x - x_\star\|),$$

whereas for the third term we use continuity of $L''(x, p_\star)$ and the fundamental theorem of calculus. Together with the assumption for the error Δt_{err} and $\|\Delta x\| \leq \|x - x_\star\| + \|x_\star - x^+\|$, we get

$$\begin{aligned} \|x^+ - x_\star\| &\leq \omega(\|x - x_\star\|)(\|\Delta x\| + \|x - x_\star\|) \\ &\leq \omega(\|x - x_\star\|)(\|x_\star - x^+\| + 2\|x - x_\star\|). \end{aligned}$$

Eventually, if $\omega(\|x - x_\star\|) \leq \varepsilon < 1$, we obtain

$$\|x^+ - x_\star\|(1 - \varepsilon) \leq \omega(\|x - x_\star\|)\|x - x_\star\|.$$

□

6.2. **Convergence of the damping parameters.** In this subsection, the convergence $\nu, \tau \to 1$ of the damping parameters will be assured to guarantee the transition to the local superlinear convergence analyzed in the previous chapter. We will consider two different total steps. First, $\Delta x_{\text{ex}} = \Delta t_{\text{ex}} + \Delta n_{\text{ex}}$, where Δn_{ex} solves (6) and Δt_{ex} solves (9) with Δn_{ex} on the right-hand side. Secondly, we consider $\Delta x = \Delta t + \Delta n$, where Δn solves (16) and Δt solves equation (9) with Δn on the right-hand side. We note that, as described in the beginning of Section 6.1, for any Δn fulfilling the equation $c(x) + c'(x)\Delta n = 0$, the step $\Delta x = \Delta t + \Delta n$ satisfies the Newton equation (2), and hence both Δx_{ex} and Δx are primal components of Newton steps for the first order optimality condition. In the following, we will assume that the iterates x converge to the SSC-point $x_\star$.

Lemma 6.4. *Let $\|\Delta x_{ex}\| \to 0$. Then it follows that $\|\Delta t_{ex}\| \to 0$ and $\|\Delta n_{ex}\| \to 0$.*

Proof. Using orthogonality of $\Delta t_{\text{ex}} \in \ker c'(x)$ and $\Delta n_{\text{ex}} \in \ker c'(x)^\perp$, we get

$$\|\Delta x_{\text{ex}}\|^2 = \|\Delta t_{\text{ex}} + \Delta n_{\text{ex}}\|^2 = \|\Delta t_{\text{ex}}\|^2 + \|\Delta n_{\text{ex}}\|^2.$$

If the left-hand side approaches zero and both summands are positive, every summand has to approach zero as well. □

Lemma 6.5. *Let $\|\Delta x_{ex}\| \to 0$. Then, if $\frac{\|\Delta n - \Delta n_{ex}\|}{\|\Delta n_{ex}\|} < \gamma$, it holds that $\|\Delta n\| \to 0$ and $\|\Delta t\| \to 0$.*

Proof. First, for the normal step, we conclude with Lemma 6.4 that

$$\|\Delta n\|^2 \leq \|\Delta n\|^2 + \|\Delta n_{\text{ex}}\|^2 \leq \|\Delta n_{\text{ex}}\|^2(\gamma^2+1) \to 0.$$

Moreover, as Δt_{ex} and Δt solve (9) with Δn_{ex} andΔn respectively, we get

$$\begin{aligned}\|\Delta t_{\text{ex}} - \Delta t\| \leq& \|L''(x,p)^{-1}(-L_x(x,p) + L_{xx}(x,p_x)\Delta n_{\text{ex}}) - L''(x,p)^{-1}(-L_x(x,p) + L_{xx}(x,p_x)\Delta n)\| \\ =& \| \left(-L''(x,p)^{-1}\begin{pmatrix} L_{xx}(x,p_x)\Delta n_{\text{ex}} \\ 0 \end{pmatrix}\right)_x + \left(L''(x,p)^{-1}\begin{pmatrix} L_{xx}(x,p_x)\Delta n \\ 0 \end{pmatrix}\right)_x \| \\ \leq& C\|\Delta n_{\text{ex}} - \Delta n\| = C\|\Delta n_{\text{err}}\| \leq C\gamma\|\Delta n_{\text{ex}}\|\end{aligned}$$

by continuity of $L''(x,p)^{-1}$ and $L_{xx}(x,p_x)$, for a constant $C > 0$. Using Lemma 6.4, we obtain $\|\Delta t\| \to 0$. The result for Δx follows by the triangle inequality. □

Corollary 6.6. *Let the assumptions of Lemma 6.5 hold. Then, $\nu \to 1$.*

Proof. Lemma 6.5 and the definition of the normal step damping factor $\nu := \min\{1, \frac{\rho_{\text{elbow}}}{\|\Delta n\|}\}$ yield the result. □

Where $\Delta x = \Delta t + \Delta n$ denoted the undamped step without residual in the tangential step, we introduce the damped version with a possible error term in the tangential step, $\delta x := \tau(\Delta t + \Delta t_{\text{err}}) + \nu\Delta n$. The term Δt_{err} stems from an inexactly solved tangential step equation.

Lemma 6.7. *Let the assumptions of Lemma 6.5 hold and $\|\Delta t_{err}\| \to 0$. Then, $\tau \to 1$.*

Proof. We use the minimizing property of δx along Δt and get

$$\begin{aligned}0 &= m'_{[\omega_f]}(\delta x)\Delta t \\ &= (f'(x) + L_{xx}(x,p)\delta n)\Delta t + L_{xx}(x,p)(\delta t, \Delta t) + \frac{[\omega_f]}{2}\|\delta x\|\langle \delta x, \Delta t\rangle \\ &= (f'(x) + L_{xx}(x,p)\delta n)\Delta t + \tau(L_{xx}(x,p)(\Delta t, \Delta t) + \frac{[\omega_f]}{2}\|\delta x\|\langle \Delta t, \Delta t\rangle) \\ &\quad + \frac{[\omega_f]}{2}\|\delta x\|\langle \delta n, \Delta t\rangle.\end{aligned}$$

Since the full tangential step Δt minimizes m_0, we also obtain

$$0 = m'_0(\delta n)\Delta t = (f'(x) + L_{xx}(x,p)\delta n)\Delta t + L_{xx}(x,p)(\Delta t, \Delta t).$$

Subtracting these two equations, we get

$$\begin{aligned}1 \geq \tau &= \frac{L_{xx}(x,p)(\Delta t, \Delta t) - \frac{[\omega_f]}{2}\|\delta x\|\langle \delta n, \Delta t\rangle}{L_{xx}(x,p)(\Delta t, \Delta t) + \frac{[\omega_f]}{2}\|\delta x\|\langle \Delta t, \Delta t\rangle} \\ &= \frac{L_{xx}(x,p)(\Delta t, \Delta t) - \frac{[\omega_f]}{2}\|\delta x\|\nu\langle \Delta n - \Delta n_{\text{ex}}, \Delta t\rangle}{L_{xx}(x,p)(\Delta t, \Delta t) + \frac{[\omega_f]}{2}\|\delta x\|\langle \Delta t, \Delta t\rangle} \\ &\geq \frac{L_{xx}(x,p)(\Delta t, \Delta t) - \frac{[\omega_f]}{2}\|\delta x\|\nu\|\Delta n - \Delta n_{\text{ex}}\|\|\Delta t\|}{L_{xx}(x,p)(\Delta t, \Delta t) + \frac{[\omega_f]}{2}\|\delta x\|\langle \Delta t, \Delta t\rangle} \to 1,\end{aligned}$$

as $\|\Delta n - \Delta n_{\text{ex}}\| < \gamma\|\Delta n_{\text{ex}}\| \to 0$ and $[\omega_f]\|\delta x\| \to 0$ by boundedness of $[\omega_f]$ and $\|\delta x\| \leq (\nu+\tau)(\|\Delta t\| + \|\Delta t_{\text{err}}\| + \|\Delta n\|) \to 0$ using the assumptions and Lemma 6.5. □

With the convergence of both of the damping parameters, we can now proof the following lemma.

Lemma 6.8. *Let the assumptions of Lemma 6.5 hold. Then, if* $\frac{\|\Delta t_{err}\|}{\|\Delta x\|} \to 0$,

$$\begin{aligned}\|\Delta x - \delta x\| &= o(\|\Delta x\|)\\ \|\delta s\| &= o(\|\Delta x\|).\end{aligned}$$

Proof. For the first claim, we compute

$$\begin{aligned}\frac{\|\Delta x - \delta x\|}{\|\Delta x\|} =& \frac{\|(1-\nu)\Delta n + (1-\tau)\Delta t + \Delta t_{\text{err}}\|}{\|\Delta x\|} \leq \|\max\{(1-\nu),(1-\tau)\}\| \underbrace{\frac{\|\Delta n + \Delta t\|}{\|\Delta x\|}}_{=1} + \frac{\|\Delta t_{\text{err}}\|}{\|\Delta x\|}\\ =& \|\max\{(1-\nu),(1-\tau)\}\| + \frac{\|\Delta t_{\text{err}}\|}{\|\Delta x\|} \to 0.\end{aligned}$$

Moreover, it follows by Theorem 4.1, that $\|\delta s\| = o(\|\delta x\|)$. Thus, we get

$$\|\delta x\| = \|\Delta x - (\Delta x - \delta x)\| \leq \|\Delta x\| + \|(\Delta x - \delta x)\| = \|\Delta x\| \left(1 + \frac{\|(\Delta x - \delta x)\|}{\|\Delta x\|}\right).$$

Therefore, together with the the first part of this proof, it follows that

$$\frac{\|\delta s\|}{\|\Delta x\|} \leq \frac{\|\delta s\|}{\|\delta x\|} \left(1 + \frac{\|(\Delta x - \delta x)\|}{\|\Delta x\|}\right) \to 0.$$

□

Eventually, we obtain locally superlinear convergence of our iterates to the Newton iterate Δx, and hence inherit the superlinear convergence of the classical Newton scheme.

Theorem 6.9. *Let the assumptions of Lemma 6.5 hold. Then, if* $\frac{\|\Delta t_{err}\|}{\|\Delta x\|} \to 0$,

$$\|\Delta x - (x_{k+1} - x_k)\| = o(\|\Delta x\|).$$

Proof. Follows with Lemma 6.8 and

$$\|\Delta x - (x_{k+1} - x_k)\| = \|\Delta x - (\delta x + \delta s)\| \leq \|\Delta x - \delta x\| + \|\delta s\|.$$

□

7. Application to optimal control and numerical results

In this section, we use the presented algorithm for the solution of two fundamentally different optimal control problems. In the first example, we consider a quasilinear heat equation and in the second example an obstacle problem in the context of elasticity.

First, we discuss a constraint preconditioner for optimal control problems which we employed for the examples. We will move from the general setting to an optimal control context and consider an optimal control problem with $X = Y \times U$ and $x = (y, u)$, where $y \in Y$ is the state and $u \in U$ the corresponding control. Let $\langle \cdot, \cdot \rangle_Y$ and $\langle \cdot, \cdot \rangle_U$ denote the scalar product in Y and U respectively. The norm on the product space is then given by the norms on Y and U and denoted by

$$\|x\|_M^2 = \|(y,u)\|_M^2 := \|y\|_{M_y}^2 + \|u\|_{M_u}^2.$$

Here $M_y : Y \to Y^\star$ and $M_u : U \to U^\star$ correspond to the scalar products of Y and U respectively.

We consider a constraint $c(x) = A(y) - Bu = 0$, where $A(y)$ is continuously invertible, twice continuously Fréchet differentiable and B is linear. Let $A := A'(y_k)$ for a fixed Newton-iterate

$x_k = (y_k, u_k)$. Then, the saddle point matrix for the normal step, the simplified normal step and the Lagrange multiplier is given by

$$H_n = \begin{pmatrix} M_y & 0 & A^T \\ 0 & M_u & -B^T \\ A & -B & 0 \end{pmatrix}.$$

For the solution of the normal and simplified normal step and the Lagrange multiplier system with Algorithm 1, we employ the preconditioner

$$P := \begin{pmatrix} 0 & 0 & A^T \\ 0 & \operatorname{diag} M_u & -B^T \\ A & -B & 0 \end{pmatrix}.$$

Hence, $\tilde{M}$ in (14) is chosen as $\begin{pmatrix} 0 & 0 \\ 0 & \operatorname{diag} M_u \end{pmatrix}$. This choice of $\tilde{M}$ is motivated by the spectral equivalence of M_u and M_y by boundedness of the operator $\|A^{-1}B\|_{U\to Y}$:

$$\begin{aligned} \langle u, u\rangle_U &\le \langle x, x\rangle_{Y\times U} = \langle y, y\rangle_Y + \langle u, u\rangle_U \\ &= \langle A^{-1}Bu, A^{-1}Bu\rangle_Y + \langle u, u\rangle_U \le \left(1 + \|A^{-1}B\|^2_{U\to Y}\right)\langle u, u\rangle_U. \end{aligned}$$

Furthermore for mass matrices, the diagonal is a reasonable preconditioner and we refer to [33] for this matter.

This choice of the preconditioner allows for an efficient solution if we factorize the block containing A, as we can solve equations in a row-wise fashion, starting from the first. As indicated in Remark 3.1, the transformation in [18, Chapter 4] remedies the problem of the right-hand side of the normal step and the simplified normal step system, see (6) and (11). In this case, the search space is not $\ker c'(x_k) = \ker\left(A'(y_k), -B\right)$. A solution to this is presented in [18, Chapter 4] with computing $z = z_0 + \tilde{z}$, where $z_0 = \left(A'(y_k)^{-1}r_p, 0, 0\right)^T$, $r_p = A(y_k) - Bu_k$ for the normal step and $r_p = r(x, \delta x)$ for the simplified normal step. The auxiliary solution $\tilde{z}$ can be computed by

$$H_n\tilde{z} = r - H_n z_0 = \begin{pmatrix} -M_y A'(y_k)^{-1}r_p \\ 0 \\ 0 \end{pmatrix}. \tag{43}$$

Therefore, in an optimal control framework, we can transform the system to obtain right-hand sides of the form $r = (r_y, r_u, 0)^T$.

All following examples were implemented in the C++-library Spacy[1], using the finite element library Kaskade7 [14]. For the factorizations we applied the sparse direct solver UMFPACK [10]. Additionally, the automatic differentiation library FunG [19] was used to compute the derivatives of the total energy functional in the case of elasticity.

7.1. **Quasilinear heat equation.** As a first example, we consider a quasilinear heat equation with distributed control. For this, we introduce a time interval $[0, 1]$ and $\Omega = [0, 3] \times [0, 1]$ and define the space-time cylinder $Q := [0, 1] \times \Omega$. For the variable's spaces, we have $Y \times U := W([0, 1]) \times L_2([0, 1] \times \Omega)$ and $P = W([0, 1])$. We aim to minimize the tracking-type cost-functional

$$f(x) = f(y, u) = \frac{1}{2}\|y - y_d\|^2_{L_2(Q)} + \frac{\alpha}{2}\|u\|^2_{L_2(Q)}$$

[1] https://spacy-dev.github.io/Spacy/

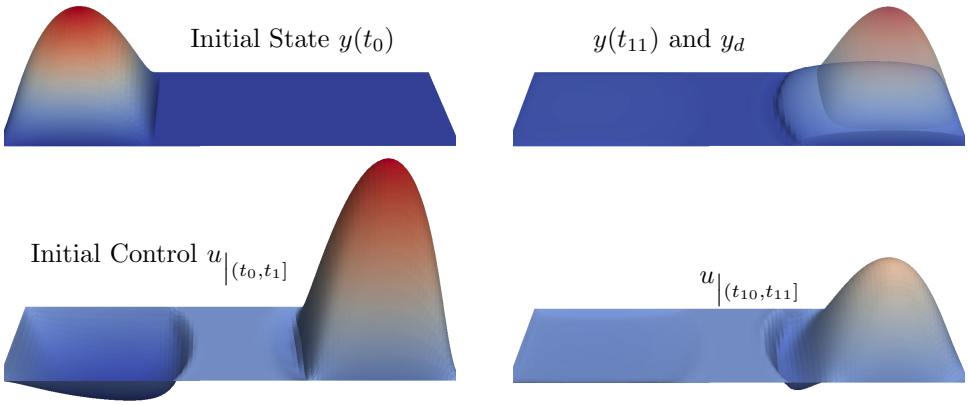


FIGURE 3. Setting $c = 100$, $d = 0.01$. Top left: Initial state. Top right: State at timestep 11 and the reference state (transparent). Bottom left: Control at timestep 1. Bottom right: Control at timestep 11.

with constant reference $y_d \in H_0^1(\Omega)$, subject to the quasilinear dynamics

$$0 = c(y,u)(v,v_0) = \int_0^1 \langle y', v\rangle_{V^\star,V}\, dt + \langle \kappa(y)\nabla y, \nabla v\rangle_{L_2(Q)} - \langle u, v\rangle_{L_2(Q)} + \langle y(0) - y_0, v_0\rangle_{L_2(\Omega)}$$

for test functions $v \in L_2(0,1;H_0^1(\Omega))$ and $v_0 \in L_2(\Omega)$. The prescribed initial datum is denoted by $y_0 \in L_2(\Omega)$ and $\kappa(y)(t,x) := (c|y(t,x)|^2 + d)I$ models an isotropic heat conduction tensor. With the parameters $c, d > 0$ we can control the nonlinearity and singularity of the problem. The higher c, the more nonlinear the problem gets and for smaller d, the diffusion term is less elliptic. For analysis on optimal control problems governed by quasilinear parabolic PDEs, we refer to [3] and the references therein. Where a result on maximal parabolic regularity and second order optimality conditions are derived, even for the control constrained case, as well as a result on Hölder continuity of the state which implies boundedness of $\kappa(y)$.

We discretize the equations arising in the algorithm with a Galerkin method, constant and discontinuous in time and linear, continuous in space, cf. [13, 12, 30]. We discretize the time interval $[0,1]$ with 101 equidistant points $0 = t_0 < t_1 \cdots < t_{100} = 1$. For the spatial discretization, we use a grid with 3201 degrees of freedom.

At an iterate (y_k, u_k), we employ the local scalar product

$$\langle (y,u),(z,v)\rangle = \langle y, z\rangle_{M_y(y_k)} + \langle u, v\rangle_{M_u}$$

where

$$\langle y, z\rangle_{M_y(y_k)} = \alpha\langle \kappa(y_k)\nabla y, \nabla z\rangle_{L_2(Q)} + \langle y, z\rangle_{L_2(Q)}$$

$$\langle u, v\rangle_{M_u} = \alpha\langle u, v\rangle_{L_2(Q)}.$$

In Figure 3 we illustrate the initial state $y(t_0)$, the state at time t_{11} and the (constant in time) desired state y_d as well as the initial control $u\big|_{(t_0,t_1]}$ and the control at time t_{11}. It is observed,

TABLE 1. Number of composite step iterations and total computations time in seconds for different parameters of the heat conduction tensor.

c, d	$1, 1$	$10^1, 10^{-1}$	$10^2, 10^{-2}$	$10^3, 10^{-3}$
Iterations / time(sec.)	4/204	7/464	38/4672	144/18424
Reference	4/325	6/1108	41/16295	

that at the time $t_{11} = 0.11$, a steady state is approached, being an optimal compromise between the tracking term in y and the penalization of the control. This steady state is depicted in the right column. This behavior is well known for time dependent systems and is often called the turnpike property. Table 1 shows the iterations for different parameters of the heat conduction tensor $\kappa(y) = (c|y|^2 + d)$, increasing in difficulty from left to right. It can be seen, that for harder problems, more iterations and hence more time is needed. However, in every case, the optimal solution was found, which reflects that even for strongly nonlinear and almost singular problems, e.g. $c = 10^3$, $d = 10^{-3}$, the globalization mechanism and the preemptive adaptive termination of the PPCG-iterations described in Section 5 lead to stable performance of the algorithm. Moreover, for time-dependent problems, the special block-diagonal structure of the differential operator A in the case of a discontinuous Galerkin method in time allows for a time-step wise factorization, which results in high saving of memory as opposed to factorization of the full matrix A. In the two settings $c = d = 1$ and $c = 10$, $d = 0.1$, no nonconvexities were encountered. This indicates, together with the low iteration numbers, that the starting value $(y, u) = (0, 0)$ is close to a local solution. For the other two cases, nonconvexities and stronger nonlinearities led to higher computations times and to all mechanisms of the algorithm being used.

Figure 4 depicts the evolution of several quantities over the course of the composite step iterations for two of the above settings. On the left-hand side of Figure 4, we observe the behaviour of the algorithm for a moderate difficulty. No nonconvexities are encountered and the algorithm converges in seven steps. After three iterations, the steps are undamped and superlinear convergence is observed. On the right-hand side, the same quantities for a more difficult problem are depicted. For the eleven iterations, the normal step is damped, i.e. the focus is laid on admissibility. In this phase, the effort for the computation of the tangential step and the Lagrange multiplier is kept low. Afterwards we focus on optimality while staying close to the admissible set. In iteration 22, we enter the final phase, with the functional being convex, and the steps being undamped. Again, we observe the local superlinear convergence proven in Section 6.

We used 0.1 as relative accuracy for the PPCG-iteration for the normal and the simplified normal step. The truncation of the Lagrange multiplier's system's solution was only performed due to a relative accuracy criterion. For the normal step termination, we formulated an adaptive termination in Algorithm 3. The behavior of this adaptive criterion for $c = 100$, $d = 0.01$ is shown in Figure 5. We observe, that in the first third of the iterations, away from the admissible set with the normal step being damped, only termination via relative error is permitted. This is consistent with the main idea of our algorithm, aiming for feasibility first, hence focusing on the normal step. After this phase, optimality is pursued. Thus, only small normal steps are needed to stay in the area close to the feasible set. These smaller normal steps are computed by less conjugate gradient iterations due to our adaptive criterion. Often, the normal step computation is exited in the first few iterations, leading to a negligible computational effort without interfering with fast local convergence.

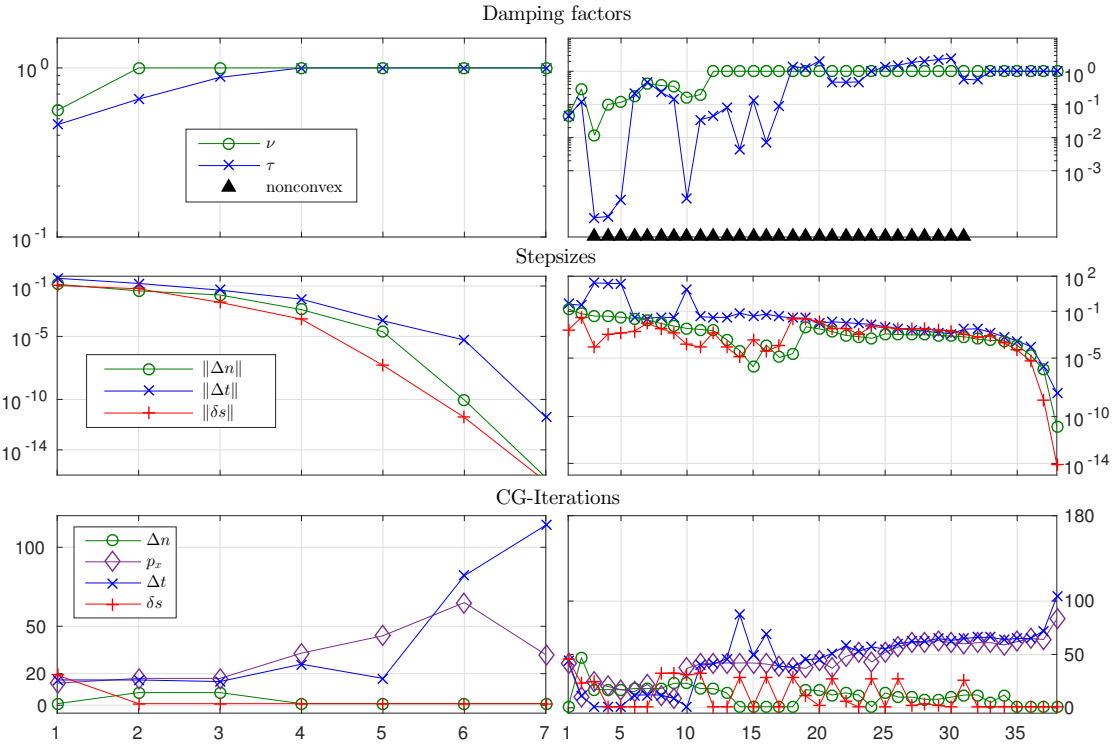


FIGURE 4. Left: $c = 10$, $d = 0.1$. Right: $c = 100$, $d = 0.01$. Damping factors, Stepsizes and CG-Iterations are depicted for each outer composite step iteration. If several computations of the simplified normal step δs are computed in one outer iteration, the number shown is the sum over all computations.

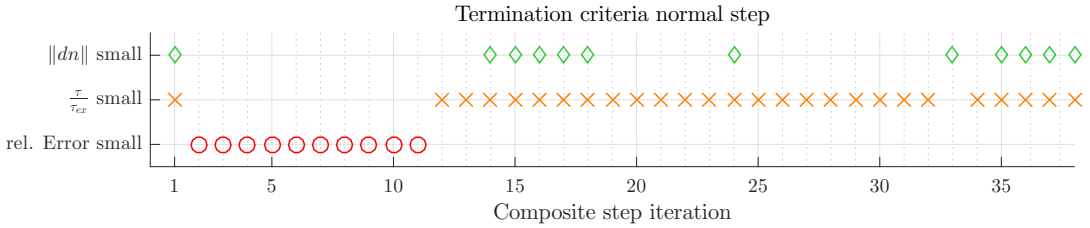


FIGURE 5. Depiction of the criteria which led to the termination of the normal step computation for each composite step iteration. For a formulation of the criteria, see Algorithm 3.

Secondly, we proposed an adaptive termination criterion for the simplified normal step cf. (5.2), i.e. $\Theta(\delta x) = \frac{\|\delta s\|}{\|\delta x\|} \leq \Theta_{\delta s}$. This assures that no rejection of δx is caused solely by inaccuracy of the simplified normal step. In some outer iterations, the rejection of the step δx and hence a decrease of the damping parameters led to repeated computations of the simplified normal step. By construction, the termination of the simplified normal step was not terminated early, if a rejection due to the contraction $\Theta(\delta s)$ being too large is to be expected. However, every

simplified normal step that led to acceptance of the total step was terminated by the adaptive criterion formulated in Subsection 5.2.

7.2. **Static nonlinear elastic contact problem.** In the second example, we study the optimal control of a static nonlinear elastic contact problem. In this setting, the state is required to be a deformation of an elastic body that results from applying some external boundary force. This boundary force also acts as the control. Additionally, the deformation is constrained by some obstacle which the body cannot penetrate.

We briefly introduce the required notation and assumptions. In our setting, the domain $\Omega \subset \mathbb{R}^3$ is required to be a Lipschitz domain which represents the nonlinear elastic body. Its boundary Γ consists of three disjoint subsets such that

$$\Gamma = \overline{\Gamma_D \cup \Gamma_N \cup \Gamma_C}.$$

Theses segments denote the Dirichlet, Neumann and contact boundary, respectively.

In this example, Ω is described by a discretized cuboid $\overline{\Omega} = [0, 2] \times [0, 2] \times [0, 0.2]$, whereby the respective grid is uniform. The grid is displayed in Figure 6.

For the state space, we choose $Y = H^1(\Omega)$ and for the control space we choose $U = L^2(\Gamma_N)$. Additionally, in order to reasonable describe deformations, each state y is required to satisfy the local injectivity condition

$$\det \nabla y > 0 \text{ a.e. in } \Omega. \tag{44}$$

In the case of hyperelasticity, deformations can be modeled as energy minimizers w.r.t. the state of the total energy functional

$$I(y, u) := \int_\Omega \hat{W}(x, \nabla y(x)) \, \mathrm{d}x - \int_{\Gamma_N} uy \, \mathrm{d}s,$$

where $u \in U$ is some fixed boundary force. Here, the stored energy function $\hat{W}$ is chosen as a compressible Mooney-Rivilin model

$$\hat{W}(\nabla y) = a\|\nabla y\|^2 + b\|\mathrm{Cof}\, \nabla y\|^2 + c(\det \nabla y)^2 - d \log \det \nabla y$$

with the respective parameters

$$a = 3.69, \quad b = 0.41, \quad c = 2.09, \quad d = 13.20.$$

For a detailed introduction into elasticity, we refer here to [7].

Furthermore, the deformation of the body is restricted by the following constraint

$$y_z \geq 0 \text{ a.e. on } \Gamma_C.$$

Those kinds of problems were first analyzed in [8] in a more general setting. However, since contact constraints are difficult to handle theoretically and numerically, we instead apply a regularization approach, the so called normal compliance method [23]. In this approach, we consider the penalty function

$$P(y) := \frac{1}{4}[-y_z]_+^4$$

which is evaluated on the contact boundary Γ_C. This function locally penalizes the violation of the constraints. Consequently, the regularized total energy functional reads as follows:

$$I_\gamma(y, u) := I(y, u) + \frac{\gamma}{4} \int_{\Gamma_C} [-y_z]_+^4 \, \mathrm{d}s,$$

for some penalty parameter $\gamma \geq 0$. For the numerical example, we choose $\gamma = 10^7$. In order to apply the composite step method, we have to replace the minimizing property by the formal first order optimality condition

$$c_\gamma(y,u)v = \int_\Omega \hat{W}'(x, \nabla y(x))\nabla v \,\mathrm{d}x - \int_{\Gamma_N} uv \,\mathrm{d}s - \gamma \int_{\Gamma_C} [-y_z]_+^3 v_z \,\mathrm{d}s = 0, \quad \forall v \in Y.$$

This can only be done in a formal way, since in the current setting, it cannot be proven whether minimizers satisfy this condition or not. We assume here that the expression is well defined throughout our analysis and we refer to [2] for a more detailed discussion.

For the objective functional, we choose a standard tracking type functional

$$f(y,u) := \frac{1}{2}\|y - y_\mathrm{d}\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Gamma_N)}.$$

Here, $y_\mathrm{d} \in L^2(\Omega)$ denotes the desired state and we used the Tikhonov parameter $\alpha = 6 \cdot 10^{-5}$. A detailed theoretical analysis of optimal control in hyperelasticity can be found in [20].

In the case of elasticity, we perform two modifications to the composite step method. First, at each iterate (y_k, u_k), we apply the problem adjusted scalar product

$$\langle (y,u),(z,v)\rangle = \langle y, z\rangle_{H^1(\Omega)} + \gamma \langle y, P''(y_k)z\rangle_{L^2(\Gamma_C)} + \alpha \langle u, v\rangle_{L^2(\Gamma_N)}.$$

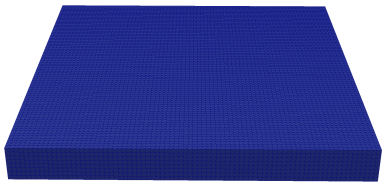
Second, in case that a new iterate violates the local injectivity condition (44), we reduce the normal and tangential step damping factor by the factor $\frac{1}{2}$ each.

The resulting solution is depicted in Figure 6. There, the desired state and the optimal solution in relation to the obstacle are displayed. It can be seen that the optimal solution already represents a good approximation of a contact constrained solution. Therefore, this approach can be used for a more sophisticated path-following approach with a composite step method as inner solver. More details to this approach can be found in [26]. In Figure 7, the respective quantities of the composite step method are displayed. In case of nonlinear elasticity, the problem is highly nonlinear and nonconvex. Nevertheless, we observe that the globalization mechanism of the composite step method again can overcome non-convexities and steer the algorithm towards a feasible solution. In contrast to before, we observe additional damping due to violation of the local injectivity condition (44). Only in iteration four, the normal step is damped solely due to nonlinearity. After that, damping only occurs occasionally to maintain condition (44). As a result, we do not obtain the separation between first achieving feasibility and after that approaching optimality as clearly as in the example in Subsection 7.1. This behavior is partially caused by the fact that in case of rejection due to condition (44), the entire step is damped. Nevertheless, we observe that most of the damping of the normal step occurs at the beginning of the composite step method and that overall the tangential step has to be damped more significantly throughout most parts of the computation. At the end, when the problem is convex, we see a significant reduction in damping and we again enter the final phase of superlinear convergence.

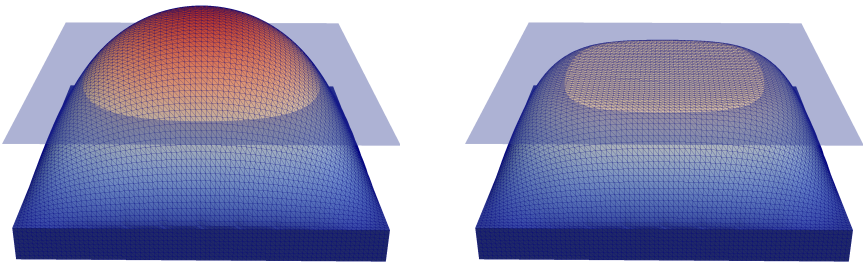
We also note the computational effort for the normal and simplified normal step are high before the algorithm enters the phase of superlinear convergence. Especially, the number of iterations required for the computation of the simplified normal step is significant, due to multiple rejections at the beginning. However, close to the feasible solution, only a few CG-iterations are required since even small normal steps keep the iterate sufficiently close to the feasible set.

In contrast to the previous example in Subsection 7.1, we observe that the computation of the Lagrange multiplier requires significantly more iterations in comparison to the computation of the tangential step.

In Figure 8, the reason for normal step termination is depicted for all outer iterations. The situation is quite different to Figure 5. Only in three outer iterations the normal step system is



(A) Undeformed domain



(B) Desired deformation

(C) Optimal deformation

FIGURE 6. Top: Undeformed domain. Bottom left: Desired deformation with obstacle (transparent). Bottom right: Optimal deformation with obstacle (transparent).

solved up to the relative accuracy. For the majority of the steps, the termination was due to the estimated fraction of tangential step damping parameters being small enough. After iteration 19, when the phase of local superlinear convergence is entered, see also Figure 7, every normal step computation is exited as the norm of the normal step is very small, indicating that we are very close to the feasible set.

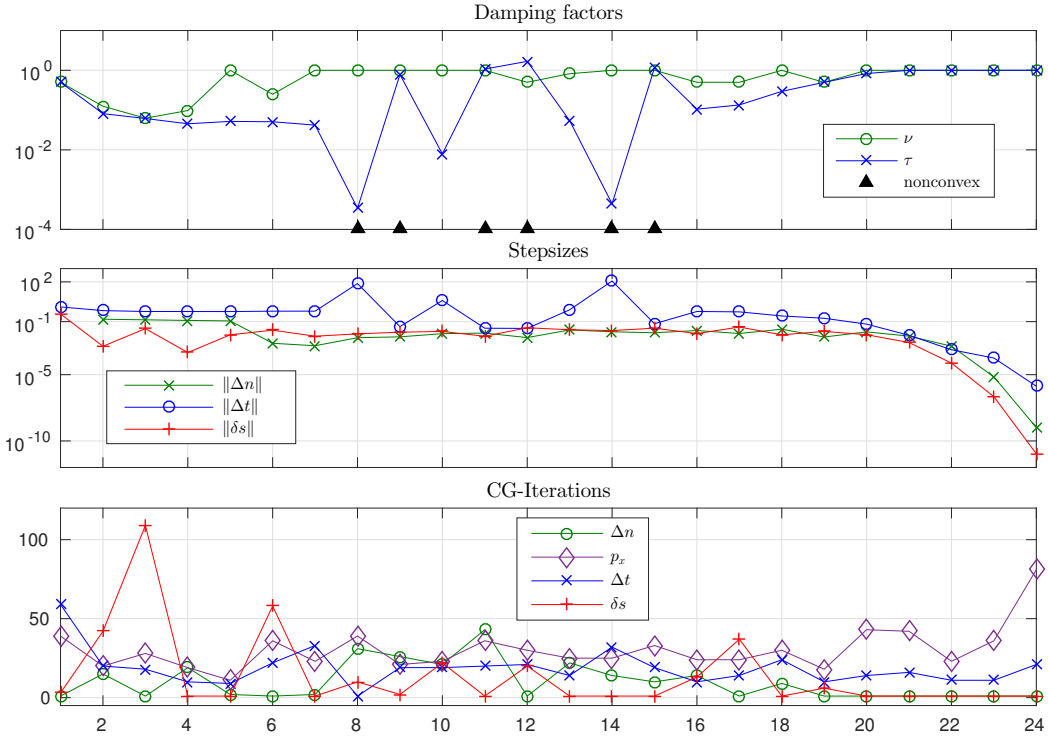


FIGURE 7. Damping factors, Stepsizes and CG-Iterations are depicted for each outer composite step iteration. If several computations of the simplified normal step δs are computed in one outer iteration, the number shown is the sum over all computations.

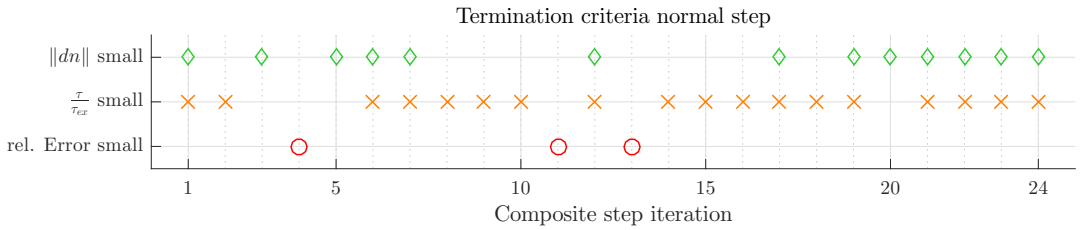


FIGURE 8. Depicition of the criteria which led to the termination of the normal step computation for each composite step iteration. For a formulation of the criteria, see Algorithm 3.

Acknowledgements. The first author was supported by the DFG Grants GR 1569/17-1 and SCHI 1379/5-1. The third author was supported by the DFG Grant SCHI 1379/2-1 within the Priority Programme SPP 1962.

REFERENCES

[1] Arioli, M.: A stopping criterion for the conjugate gradient algorithm in a finite element method framework. Numerische Mathematik **97**(1), 1–24 (2002)

[2] Ball, J.M.: Some open problems in elasticity. In: Geometry, mechanics, and dynamics, pp. 3–59. Springer (2002)
[3] Bonifacius, L., Neitzel, I.: Second order optimality conditions for optimal control of quasilinear parabolic equations. Mathematical Control & Related Fields **8**(1), 1–34 (2018)
[4] Byrd, R., Gilbert, J., Nocedal, J.: A trust region method based on interior point techniques for nonlinear programming. Math. Program. **89**(1, Ser. A), 149–185 (2000). DOI 10.1007/PL00011391. URL http://dx.doi.org/10.1007/PL00011391
[5] Byrd, R., Hribar, M., Nocedal, J.: An interior point algorithm for large-scale nonlinear programming. SIAM J. Optim. **9**(4), 877–900 (1999). DOI 10.1137/S1052623497325107. URL http://dx.doi.org/10.1137/S1052623497325107. Dedicated to John E. Dennis, Jr., on his 60th birthday
[6] Cartis, C., Gould, N.I., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. Mathematical Programming **127**(2), 245–295 (2011)
[7] Ciarlet, P.: Mathematical Elasticity: Three-dimensional elasticity. Bd. 1. North-Holland (1994)
[8] Ciarlet, P.G., Nečas, J.: Unilateral problems in nonlinear, three-dimensional elasticity. Archive for Rational Mechanics and Analysis **87**(4), 319–338 (1985). DOI 10.1007/BF00250917
[9] Coleman, T.F.: Linearly constrained optimization and projected preconditioned conjugate gradients. Proc. Fifth SIAM Conference on Applied Linear Algebra pp. 118–122 (1994)
[10] Davis, T., Duff, I.: An unsymmetric-pattern multifrontal method for sparse LU factorization. SIAM Journal on Matrix Analysis and Applications **18**(1), 140–158 (1997)
[11] Deuflhard, P.: Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms. Springer Publishing Company, Incorporated (2011)
[12] Eriksson, K., Johnson, C.: Adaptive finite element methods for parabolic problems i: A linear model problem. SIAM Journal on Numerical Analysis **28**(1), 43–77 (1991). DOI 10.1137/0728003. URL http://dx.doi.org/10.1137/0728003
[13] Eriksson, K., Johnson, C., Thomé, V.: Time discretization of parabolic problems by the discontinuous galerkin method. ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique **19**(4), 611–643 (1985). URL http://eudml.org/doc/193462
[14] Götschel, S., Weiser, M., Schiela, A.: Solving optimal control problems with the Kaskade 7 finite element toolbox (2010)
[15] Gould, N., Hribar, M., Nocedal, J.: On the solution of equality constrained quadratic programming problems arising in optimization. SIAM Journal on Scientific Computing **23**(4), 1376–1395 (2001)
[16] Heinkenschloss, M., Ridzal, D.: A matrix-free trust-region SQP method for equality constrained optimization. SIAM J. Optim. **24**(3), 1507–1541 (2014). DOI 10.1137/130921738. URL http://dx.doi.org/10.1137/130921738
[17] Heinkenschloss, M., Vicente, L.: Analysis of inexact trust-region SQP algorithms. SIAM J. Optim. **12**(2), 283–302 (2001/02). DOI 10.1137/S1052623499361543. URL http://dx.doi.org/10.1137/S1052623499361543
[18] Lubkoll, L.: An optimal control approach to implant shape design. Ph.D. thesis, Universität Bayreuth (2015)
[19] Lubkoll, L.: Fung - invariant-based modeling. Archive of Numerical Software **5**(1) (2017)
[20] Lubkoll, L., Schiela, A., Weiser, M.: An optimal control problem in polyconvex hyperelasticity. SIAM J. Control Opt. **52**(3), 1403 – 1422 (2014)
[21] Lubkoll, L., Schiela, A., Weiser, M.: An affine covariant composite step method for optimization with pdes as equality constraints. Optimization Methods and Software **32**(5), 1132–1161 (2017). DOI 10.1080/10556788.2016.1241783. URL https://doi.org/10.1080/10556788.2016.1241783
[22] Luenberger, D.G.: Optimization by vector space methods. John Wiley & Sons (1997)
[23] Martins, J., Oden, J.: Existence and uniqueness results for dynamic contact problems with nonlinear normal and friction interface laws. Nonlinear Analysis: Theory, Methods and Applications **11**(3), 407 – 428 (1987)
[24] Omojokun, E.O.: Trust region algorithms for optimization with nonlinear equality and inequality constraints. Ph.D. thesis, University of Colorado Boulder, Boulder, CO, USA (1989). UMI Order No: GAX89-23520
[25] Ridzal, D.: Trust-region SQP methods with inexact linear system solves for large-scale optimization. ProQuest LLC, Ann Arbor, MI (2006). Thesis (Ph.D.)–Rice University
[26] Schiela, A., Stöcklein, M.: Optimal control of static contact in finite strain elasticity (2018). Preprint SPP1962-097
[27] Steihaug, T.: The conjugate gradient method and trust regions in large scale optimization. SIAM Journal on Numerical Analysis **20**(3), 626–637 (1983). DOI 10.1137/0720042. URL https://doi.org/10.1137/0720042
[28] Strakoš, Z., Tichý, P.: On error estimation in the conjugate gradient method and why it works in finite precision computations. Electronic Transactions on Numerical Analysis **13**, 56–80 (2002)
[29] Strakoš, Z., Tichý, P.: Error estimation in preconditioned conjugate gradients. BIT Numerical Mathematics **45**, 789–817 (2005)
[30] Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems. Springer (2006)

[31] Toint, P.: Towards an Efficient Sparsity Exploiting Newton Method for Minimization, pp. 57–88. Academic press (1981). Publication editors : I.S. Duff

[32] Vardi, A.: A trust region algorithm for equality constrained minimization: convergence properties and implementation. SIAM J. Numer. Anal. **22**(3), 575–591 (1985). DOI 10.1137/0722035. URL http://dx.doi.org/10.1137/0722035

[33] Wathen, A.: Realistic eigenvalue bounds for the Galerkin mass matrix. IMA Journal of Numerical Analyisis **7**, 449–457 (1987)

[34] Ziems, J.C., Ulbrich, S.: Adaptive multilevel inexact SQP methods for PDE-constrained optimization. SIAM J. Optim. **21**(1), 1–40 (2011)