Noname manuscript No.
(will be inserted by the editor)

The R-linear convergence rate of an algorithm arising from the semi-smooth Newton method applied to 2D contact problems with friction

Radek Kučera · Kristina Motyčková · Alexandros Markopoulos

Received: date / Accepted: date

Abstract The goal is to analyze the semi-smooth Newton method applied to the solution of contact problems with friction in two space dimensions. The primal-dual algorithm for problems with the Tresca friction law is reformulated by eliminating primal variables. The resulting dual algorithm uses the conjugate gradient method for inexact solving of inner linear systems. The globally convergent algorithm based on computing a monotonously decreasing sequence is proposed and its R-linear convergence rate is proved. Numerical experiments illustrate the performance of different implementations including the Coulomb friction law.

Keywords Contact problem · Friction · Semi-smooth Newton method · Conjugate gradient method · Gradient projection · Convergence rate

Mathematics Subject Classification (2000) 65K10 · 65N22 · 49M29 · 74M15

1 Introduction

Finite element approximations of frictional contact problems lead typically to nonsmooth equations that are equivalent, in many cases, to a constrained minimization. Algorithms based on active sets belong to the most efficient iterative methods for solving such problems in two (2D) as well as three (3D) space dimensions. There are at least two strategies how to introduce active set

R. Kučera, A. Markopoulos
Centre of Excellence IT4I, VŠB-Technical University of Ostrava
Tř. 17. listopadu 15, CZ-78033 Ostrava-Poruba
Tel.: +420-597324126, E-mail: radek.kucera@vsb.cz E-mail: alexandros.markopoulos@vsb.cz

K. Motyčková
Department of Applied Mathematics, VŠB-Technical University of Ostrava
Tř. 17. listopadu 15, CZ-78033 Ostrava-Poruba
Tel.: +420-597326291, E-mail: kristina.motyckova@vsb.cz

algorithms in context of contact problems. The first one have been developed for dual contact problems given by the minimization of a strictly quadratic cost function subject to separable inequality constraints that is the case of Tresca friction; see [10,5] for 2D and [18,19,7] for 3D case, respectively. Here, the active set is the index subset of components, for which the constraints are satisfied by equalities in the current iteration. The conjugate gradient method (CGM) generates iterations with respect to non-active components and, when the progress is not sufficient, the active set is changed by a gradient projection step. Thus, the algorithm seeks for the active set of the solution so that it generates monotonously decreasing iterations lying in the feasible set and enjoys the R-linear convergence rate [10,5,7,19]. The paper [8] presents an alternative duality and active set based domain decomposition approach with the proof of a global linear complexity for 2D semicoercive contact problems with friction.

Another class of active set algorithms arises from the use of the semismooth Newton method (SSNM). The starting point is the primal-dual formulation of a contact problem, in which contact conditions are reformulated by nonsmooth equations [33] as proposed already in [1]. Later on, it was recognized that the SSNM may be interpreted as a primal-dual active set method [25,14]. This approach is widely used for direct solving contact problems with different friction laws; see [32,22,23] for the 2D and [24] for the 3D case, respectively. The paper [34] comprises an up to date comprehensive exposition of the SSNM oriented research for contact problems. The standard convergence analysis is usually based on the slant differentiability concept [4, 25] leading to the local superlinear convergence rate. This convergence result assumes exact solutions of inner linear systems that is, however, unrealistic for large-scale problems. Another drawback consists in the fact that an initial iteration "sufficiently close" to the solution is required; see [28] for a sophisticated choice. The Newton-type methods for solving abstract discrete variational inequalities are summarized in [11]. An inexact Newton method applied to nonsmooth equations were analyzed already in [29] including a globally convergent result. However, according to our knowledge, there is no analysis of a *global convergence rate* in cases, when the inner subproblems of the SSNM are solved inexactly.

The aim of our paper is to analyze discrete active set algorithms arising from the SSNM applied to the solution of contact problems with Tresca friction in 2D. The outline of the paper is as follows. The PDEs describing the problem are given in the rest of Section 1. The algebraic problem resulting from a finite element approximation is introduced in Section 2. The primal-dual formulation given by nonsmooth equations depends on a positive parameter ρ. In Section 3, we derive the active-set algorithm in terms of dual variables (Lagrange multipliers) and we recall the superlinear convergence rate valid for the exact SSNM. The implementation introduced in Section 4 is based on the use of the CGM with an adaptive precision control for inexact solving inner subproblems. Main results are summarized in Section 5. First, we analyze sufficient conditions guaranteeing that a modified algorithm will generate a

monotonously decreasing sequence of the cost function values. For that, one needs: *(i)* feasibility of iterations and *(ii)* the upper bound $\rho < 2\sigma_{\max}^{-1}$, where $\sigma_{\max}$ is the largest eigenvalue of the dual Hessian. The inequality in *(ii)* plays an important role in the active-set algorithms for the constrained minimization problems developed in [10,5,9]. The proof of the R-linear convergence rate of our (modified) algorithm uses the auxiliary result on the decrease of the cost function along the projected gradient path from [2]. Numerical experiments reported in Section 6 confirm the analysis. This section contains also a short extension to contact problems with Coulomb friction, where we combine the SSNM with the TFETI domain decomposition method [6]. The geometry of the model problem arises in mining industry [8].

Since the SSNM starts from an algebraic counterpart of PDEs, we introduce formulation of the problem in terms of the Lame equations. We consider two elastic bodies represented by non-overlapping bounded domains $\Omega_k \subset \mathbb{R}^2$ with Lipschitzian boundaries $\partial\Omega_k$, $k = 1, 2$ (see Figure 1.a). Each boundary splits on three non-empty disjoint parts γ_k^u, γ_k^p, and γ_k^c so that $\partial\Omega_k = \overline{\gamma_k^u} \cup \overline{\gamma_k^p} \cup \overline{\gamma_k^c}$. The zero displacements are prescribed on γ_k^u, surface tractions $\boldsymbol{p}_k \in (L^2(\gamma_k^p))^2$ act on γ_k^p, and the bodies Ω_k are subject to volume forces $\boldsymbol{f}_k \in (L^2(\Omega_k))^2$, $k = 1, 2$. We seek displacement fields $\boldsymbol{u}_k$ in Ω_k satisfying the following equations

$$\left.\begin{aligned} -\operatorname{div}\boldsymbol{\sigma}_k(\boldsymbol{u}_k) &= \boldsymbol{f}_k \quad \text{in} \quad \Omega_k \\ \boldsymbol{u}_k &= \boldsymbol{0} \quad \text{on} \quad \gamma_k^u \\ \boldsymbol{\sigma}_k(\boldsymbol{u}_k)\boldsymbol{\nu}_k &= \boldsymbol{p}_k \quad \text{on} \quad \gamma_k^p \end{aligned}\right\} \quad k = 1, 2, \tag{1.1}$$

where $\boldsymbol{\nu}_k$ is the unit outward normal vector to the boundary $\partial\Omega_k$. The stress field $\boldsymbol{\sigma}_k(\boldsymbol{u}_k)$ in Ω_k is related to the linearized strain tensor $\boldsymbol{\varepsilon}_k(\boldsymbol{u}_k) = 1/2(\nabla\boldsymbol{u}_k + \nabla^\top\boldsymbol{u}_k)$ by the linearized Hooke law for elastic, homogeneous, and isotropic materials: $\boldsymbol{\sigma}_k(\boldsymbol{u}_k) = a_k \operatorname{tr}(\boldsymbol{\varepsilon}_k(\boldsymbol{u}_k))I + 2b_k\boldsymbol{\varepsilon}_k(\boldsymbol{u}_k)$, where "tr" is the trace of matrices, $I \in \mathbb{R}^{2\times 2}$ is the identity matrix, and $a_k, b_k > 0$ are the Lamé constants. We predefine one-to one transfer mapping $\boldsymbol{\chi} : \gamma_1^c \mapsto \gamma_2^c$, by means of which we define the initial distance between the bodies $\delta(\boldsymbol{x}) = \|\boldsymbol{\chi}(\boldsymbol{x}) - \boldsymbol{x}\|$, $\boldsymbol{x} \in \gamma_1^c$ and the critical direction: $\boldsymbol{\nu}(\boldsymbol{x}) = (\boldsymbol{\chi}(\boldsymbol{x}) - \boldsymbol{x})/\delta(\boldsymbol{x})$, if $\delta(\boldsymbol{x}) \neq 0$, or $\boldsymbol{\nu}(\boldsymbol{x}) = \boldsymbol{\nu}_1(\boldsymbol{x})$, if $\delta(\boldsymbol{x}) = 0$. The unit vector orthogonal to $\boldsymbol{\nu} = \boldsymbol{\nu}(\boldsymbol{x})$ is denoted by $\boldsymbol{\tau} = \boldsymbol{\tau}(\boldsymbol{x})$, $\boldsymbol{x} \in \gamma_1^c$. On the contact interface, we consider three contact conditions:

$$u_\nu - \delta \le 0,\ \sigma_\nu \le 0,\ \sigma_\nu(u_\nu - \delta) = 0 \quad \text{on} \quad \gamma_1^c, \tag{1.2}$$

$$\boldsymbol{\sigma}_1(\boldsymbol{u}_1)\boldsymbol{\nu}_1 = \boldsymbol{\sigma}_2(\boldsymbol{u}_2 \circ \boldsymbol{\chi})\boldsymbol{\nu}_1 \quad \text{on} \quad \gamma_1^c, \tag{1.3}$$

$$\left.\begin{aligned} &|\sigma_\tau| \le g \\ &|\sigma_\tau| < g \Rightarrow u_\tau = 0 \\ &|\sigma_\tau| = g \Rightarrow \exists c_\tau \ge 0 :\ u_\tau = -c_\tau\sigma_\tau \end{aligned}\right\} \quad \text{on} \quad \gamma_1^c, \tag{1.4}$$

where $u_\nu = (\boldsymbol{u}_1 - \boldsymbol{u}_2 \circ \boldsymbol{\chi})^\top\boldsymbol{\nu}$, $u_\tau = (\boldsymbol{u}^1 - \boldsymbol{u}^2 \circ \boldsymbol{\chi})^\top\boldsymbol{\tau}$, $\sigma_\nu = \boldsymbol{\nu}^\top\boldsymbol{\sigma}_1(\boldsymbol{u}_1)\boldsymbol{\nu}_1$, $\sigma_\tau = \boldsymbol{\tau}^\top\boldsymbol{\sigma}_1(\boldsymbol{u}^1)\boldsymbol{\nu}_1$, and $g \in L^2(\gamma_1^c)$, $g \ge 0$ is the slip bound. Here, (1.2) is the unilateral contact law, (1.3) describes the transmission of contact stresses, and (1.4) is the Tresca friction law. For the weak form of (1.1)-(1.4), we refer to [16, 27]. The algebraic problem arising from the finite element approximation is the starting point of our analysis in the next section. Note that there is no

algebraic counterpart of (1.3), since the transmission of contact stresses is inherently hidden in the weak form.

Let us introduce notation used trough the whole paper. All vectors and matrices are denoted below by mathematical italic. $\mathbb{R}^{p\times q}$ is the space of $p\times q$ matrices, $\mathbb{R}^p = \mathbb{R}^{p\times 1}$ is the space of p-dimensional vectors, and $\mathbb{R}^p_+$ is the non-negative orthant in $\mathbb{R}^p$. 0 stands for a zero matrix, a zero vector, or the zero number, while I is a square identity matrix. The caligraphic symbols are used for sets of indices, e.g., $\mathcal{M} = \{1, 2, \ldots, p\}$ and $\mathcal{A}, \mathcal{I} \subseteq \mathcal{M}$. Let $N \in \mathbb{R}^{p\times q}$. Then, $N_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|\times q}$ is composed by those rows of N whose indices belong to $\mathcal{A}$. Let $A \in \mathbb{R}^{p\times p}$ be symmetric, positive definite with the smallest and largest eigenvalues $0 < \sigma_{\min} \le \sigma_{\max}$, respectively. The submatrix $A_{\mathcal{A}\mathcal{I}} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{I}|}$ is composed by those entries of A whose row, column indices belong to $\mathcal{A}$, $\mathcal{I}$, respectively. If $\mathcal{I} \neq \emptyset$, then

$$\sigma_{\min} v^\top v \le v^\top A_{\mathcal{I}\mathcal{I}} v \le \sigma_{\max} v^\top v \quad \forall v \in \mathbb{R}^{|\mathcal{I}|}. \tag{1.5}$$

Finally, the Euclidean norm of $v \in \mathbb{R}^p$, $v = (v_1, v_2, \ldots, v_p)^\top$, is defined by $\|v\|^2 = v^\top v = v_1^2 + v_2^2 + \cdots + v_p^2$.

2 Algebraic contact problems with Tresca friction

Let us consider the problem to find $(u^*, \lambda_\nu^*, \lambda_\tau^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ satisfying:

$$Ku + N^\top \lambda_\nu + T^\top \lambda_\tau - f = 0, \tag{2.1}$$

$$Nu - d \le 0,\ \lambda_\nu \ge 0,\ \lambda_\nu^\top (Nu - d) = 0, \tag{2.2}$$

$$\left.\begin{array}{l} |\lambda_{\tau,i}| \le g_i \\ |\lambda_{\tau,i}| < g_i \Rightarrow (Tu)_i = 0 \\ |\lambda_{\tau,i}| = g_i \Rightarrow \exists c_i \ge 0 :\ (Tu)_i = c_i \lambda_{\tau,i} \end{array}\right\} \quad i \in \mathcal{M}, \tag{2.3}$$

where $\mathcal{M} = \{1, \ldots, m\}$ is the index set, $K \in \mathbb{R}^{n\times n}$ is symmetric and positive definite, $N, T \in \mathbb{R}^{m\times n}$ have full row-rank, $f \in \mathbb{R}^n$, $d \in \mathbb{R}^m_+$, and g_i are entries of $g \in \mathbb{R}^m_+$. This problem is the algebraic *primal-dual* formulation of the contact problem with Tresca friction arising from a finite element discretization. The primal unknown u^* approximates displacements, while the dual unknowns λ_ν^*, λ_τ^* approximate the (negative) normal, tangential contact stresses, respectively. Note that (2.1)-(2.3) can be considered as the Karush-Khun-Tucker conditions to a (primal) contact problem so that λ_ν^*, λ_τ^* play also the role of the Lagrange multipliers.

First of all we show how to eliminate the primal unknown. Let us introduce the Lagrangian $L : \mathbb{R}^n \times \Lambda_\nu \times \Lambda_\tau \mapsto \mathbb{R}$ to (2.1)-(2.3) by

$$L(u, \lambda_\nu, \lambda_\tau) = \frac{1}{2} u^\top K u - u^\top f + \lambda_\nu^\top (Nu - d) + \lambda_\tau^\top T u,$$

where $\Lambda_\nu = \mathbb{R}^m_+$ and $\Lambda_\tau = \{\lambda_\tau \in \mathbb{R}^m : |\lambda_{\tau,i}| \le g_i,\ i \in \mathcal{M}\}$. It is easy to show that $(u^*, \lambda_\nu^*, \lambda_\tau^*)$ is the saddle-point of L, i.e., this point satisfies

$$L(u^*, \lambda_\nu, \lambda_\tau) \le L(u^*, \lambda_\nu^*, \lambda_\tau^*) \le L(u, \lambda_\nu^*, \lambda_\tau^*) \quad \forall (u, \lambda_\nu, \lambda_\tau) \in \mathbb{R}^n \times \Lambda_\nu \times \Lambda_\tau. \tag{2.4}$$

The second inequality in (2.4) is equivalent to

$$u^* = K^{-1}(f - N^\top \lambda_\nu^* - T^\top \lambda_\tau^*). \tag{2.5}$$

Inserting (2.5) in the first inequality in (2.4), one can derive the problem in terms of the Lagrange multipliers. Before giving this formulation, we introduce some notation.

Let $q : \mathbb{R}^{2m} \mapsto \mathbb{R}$ be the quadratic cost function defined by

$$q(\lambda) = \frac{1}{2}\lambda^\top A\lambda - \lambda^\top b, \tag{2.6}$$

where $\lambda = (\lambda_\nu^\top, \lambda_\tau^\top)^\top$, $A = BK^{-1}B^\top$ with $B = \left(N^\top, T^\top\right)^\top$ is symmetric and positive definite, $b = BK^{-1}f - c$, and $c = \left(d^\top, 0^\top\right)^\top$. The gradient $r : \mathbb{R}^{2m} \mapsto \mathbb{R}$ to q at $\lambda \in \mathbb{R}^{2m}$ is given by

$$r(\lambda) = A\lambda - b. \tag{2.7}$$

Denote $\lambda^* = ({\lambda_\nu^*}^\top, {\lambda_\tau^*}^\top)^\top$ and $\Lambda = \Lambda_\nu \times \Lambda_\tau$. This notation enables us to express the first inequality in (2.4), after eliminating of u^*, as

$$(\lambda - \lambda^*)^\top r(\lambda^*) \geq 0 \quad \forall \lambda \in \Lambda. \tag{2.8}$$

Since λ^* satisfying (2.8) minimizes q on Λ [30], we arrive at the *dual* formulation of the contact problems with Tresca friction:

$$\lambda^* = \arg\min q(\lambda) \quad \text{subject to } \lambda \in \Lambda. \tag{2.9}$$

As q is strictly quadratic and Λ is the closed and convex set, there is the unique solution $\lambda^* \in \Lambda$ to (2.9) [30]. It proves also that our original problem (2.1)-(2.3) has the unique solution with the first component u^* determined by (2.5).

Now we reformulate (2.1)-(2.3) as a projective equation represented by nonsmooth functions. Let $P_{\Lambda_\nu} : \mathbb{R}^m \mapsto \Lambda_\nu$, $P_{\Lambda_\tau} : \mathbb{R}^m \mapsto \Lambda_\tau$ be the projections onto Λ_ν, Λ_τ defined by the max-function as follows:

$$P_{\Lambda_\nu,i}(\lambda_\nu) = \max\{0, \lambda_{\nu,i}\}, \tag{2.10}$$

$$P_{\Lambda_\tau,i}(\lambda_\tau) = \max\{0, \lambda_{\tau,i} + g_i\} - \max\{0, \lambda_{\tau,i} - g_i\} - g_i, \tag{2.11}$$

respectively. Let us introduce the function $G : \mathbb{R}^{n+2m} \mapsto \mathbb{R}^{n+2m}$ with $y = (u^\top, \lambda_\nu^\top, \lambda_\tau^\top)^\top$ given by

$$G(y) = \begin{pmatrix} Ku + N^\top\lambda_\nu + T^\top\lambda_\tau - f \\ \lambda_\nu - P_{\Lambda_\nu}(\lambda_\nu + \rho(Nu - d)) \\ \lambda_\tau - P_{\Lambda_\tau}(\lambda_\tau + \rho Tu) \end{pmatrix}, \tag{2.12}$$

where $\rho > 0$ is an arbitrary but fixed parameter. It is easy to verify that (2.1)-(2.3) and the equation

$$G(y) = 0, \tag{2.13}$$

have the same solution $y^* = ({u^*}^\top, {\lambda_\nu^*}^\top, {\lambda_\tau^*}^\top)^\top$. The function G is nonsmooth due to the presence of the max-function. Fortunately, it is semi-smooth in the

sense of [4] so that the SSNM can be used. We will analyze it in the next sections.

In the rest of this section, we summarize some auxiliary results. We denote $P_\Lambda : \mathbb{R}^{2m} \mapsto \Lambda$ the projection onto Λ given by $P_\Lambda = (P_{\Lambda_\nu}^\top, P_{\Lambda_\tau}^\top)^\top$. The *reduced gradient* $\widetilde{r}_\alpha : \Lambda \mapsto \mathbb{R}$ to q for $\alpha > 0$ is defined by:

$$\widetilde{r}_\alpha(\lambda) = \frac{1}{\alpha}(\lambda - P_\Lambda(\lambda - \alpha r(\lambda))). \tag{2.14}$$

It is well-known [19] that $\widetilde{r}_\alpha$ is the optimality criterion to (2.9) in the sense that $\lambda^* \in \Lambda$ solves (2.9) iff $\widetilde{r}_\alpha(\lambda^*) = 0$. Therefore, the reduced gradient will be used as the stopping criterion in our algorithms. The following lemma is the key ingredient of our analysis.

Lemma 1 *Let $\sigma_{\min}$, $\sigma_{\max}$ be the smallest, largest eigenvalue of A, respectively, and let λ^* be the solution to (2.9). The following statement holds:*

$$q(P_\Lambda(\lambda - \alpha r(\lambda))) - q(\lambda^*) \leq \eta(\alpha)\,(q(\lambda) - q(\lambda^*)) \tag{2.15}$$

for all $\lambda \in \Lambda$, where

$$\eta(\alpha) = \begin{cases} 1 - \alpha\sigma_{\min} & \text{for } \alpha \in [0, \sigma_{\max}^{-1}], \\ 1 - (2\sigma_{\max}^{-1} - \alpha)\sigma_{\min} & \text{for } \alpha \in [\sigma_{\max}^{-1}, 2\sigma_{\max}^{-1}]. \end{cases}$$

Proof See [10,5] for the first case and [2] for the second case. □

3 Semi-smooth Newton method

The concept of semismoothness uses slant differentiability of a function. Here, we recall basic results related to our problem [4].

Let Y, Z be Banach spaces with the norms $\|\cdot\|_Y$, $\|\cdot\|_Z$, respectively, and $\mathcal{L}(Y,Z)$ denote the set of all bounded linear mappings of Y into Z with the norm $\|\cdot\|_{\mathcal{L}(Y,Z)}$. Let $U \subseteq Y$ be an open subset and $G : U \mapsto Z$ be a function.

Definition 1 *(i) The function G is called slantly differentiable at $y \in U$ if there exists a mapping $G^o : U \mapsto \mathcal{L}(Y,Z)$ such that $\{G^o(y+h)\}$ are uniformly bounded for sufficiently small $h \in Y$ and*

$$\lim_{h\to 0} \frac{\|G(y+h) - G(y) - G^o(y+h)h\|_Z}{\|h\|_Y} = 0.$$

The function G^o is called a slanting function for G at y.
(ii) The function G is called slantly differentiable in U, if there exists $G^o : U \mapsto \mathcal{L}(Y,Z)$ such that G^o is a slanting function for G at every point $y \in U$. The function G^o is called a slanting function for G in U.

Theorem 1 *Let G be slantly differentiable in U with a slanting function G^o. Suppose that $y^* \in U$ is a solution to the nonlinear equation $G(y) = 0$. If $G^o(y)$ is non-singular for all $y \in U$ and $\{\|G^o(y)^{-1}\|_{\mathcal{L}(Y,Z)} : y \in U\}$ is bounded, then the Newton iterations*

$$y^{k+1} = y^k - G^o(y^k)^{-1} G(y^k) \tag{3.1}$$

converge superlinearly to y^* provided that $\|y^0 - y^*\|_Y$ is sufficiently small.

Proof See [4]. □

Let us focus on the max-function $\phi(y) = \max\{0, y\}$ with $Y = Z = \mathbb{R}$. This function is slantly differentiable and

$$\phi^o(y) = \begin{cases} 1 & \text{for } y > 0 \\ \sigma & \text{for } y = 0 \\ 0 & \text{for } y < 0 \end{cases}$$

is the slanting function in $\mathbb{R}$ for an arbitrary real number σ (below we use $\sigma = 1$). As the function G given by (2.12) is defined by finitely many max-functions, it is also slantly differentiable with $Y = Z = \mathbb{R}^{n+2m}$. A convenient setting of its slanting function G^o uses an active set terminology.

Recall that $\mathcal{M} = \{1, 2, \ldots, m\}$. Let $\mathcal{A}_\nu, \mathcal{A}_\tau \subseteq \mathcal{M}$ denote the *active* sets and $\mathcal{I}_\nu, \mathcal{I}_\tau^+, \mathcal{I}_\tau^- \subseteq \mathcal{M}$ be the respective *inactive* sets at $y = (u^\top, \lambda_\nu^\top, \lambda_\tau^\top) \in \mathbb{R}^{n+2m}$:

$$\begin{aligned} \mathcal{A}_\nu &= \{i \in \mathcal{M} : 0 \le \lambda_{\nu,i} + \rho(Nu - d)_i\}, \\ \mathcal{I}_\nu &= \mathcal{M} \setminus \mathcal{A}_\nu, \\ \mathcal{I}_\tau^+ &= \{i \in \mathcal{M} : g_i < \lambda_{\tau,i} + \rho(Tu)_i\}, \\ \mathcal{I}_\tau^- &= \{i \in \mathcal{M} : \lambda_{\tau,i} + \rho(Tu)_i < -g_i\}, \\ \mathcal{A}_\tau &= \mathcal{M} \setminus \left(\mathcal{I}_\tau^+ \cup \mathcal{I}_\tau^-\right). \end{aligned}$$

For $\mathcal{S} \subseteq \mathcal{M}$, we define the indicator matrix $D(\mathcal{S}) = diag(s_1, s_2, \ldots, s_m)$ with $s_i = 1$ for $i \in \mathcal{S}$ and $s_i = 0$ for $i \notin \mathcal{S}$. The function G given by (2.12) can be written as:

$$G(y) = \begin{pmatrix} Ku - f + N^\top \lambda_\nu + T^\top \lambda_\tau \\ \lambda_\nu - D(\mathcal{A}_\nu)(\lambda_\nu + \rho(Nu - d)) \\ \lambda_\tau - D(\mathcal{A}_\tau)(\lambda_\tau + \rho Tu) - D(\mathcal{I}_\tau^+)g + D(\mathcal{I}_\tau^-)g \end{pmatrix}.$$

Differentiating with respect to y, we derive the following slanting function:

$$G^o(y) = \begin{pmatrix} K & N^\top & T^\top \\ -\rho D(\mathcal{A}_\nu)N & D(\mathcal{I}_\nu) & 0 \\ -\rho D(\mathcal{A}_\tau)T & 0 & D(\mathcal{I}_\tau^+ \cup \mathcal{I}_\tau^-) \end{pmatrix}.$$

Each iterative step of (3.1) consists in solving the linear system with the matrix $G^o(y^k)$ and with the right hand-side vector

$$G^o(y^k)y^k - G(y^k) = \begin{pmatrix} f \\ -\rho D(\mathcal{A}_\nu)d \\ (D(\mathcal{I}_\tau^+) - D(\mathcal{I}_\tau^-))g \end{pmatrix}.$$

From the block structure, one can easily deduce that the components of λ_ν^{k+1}, λ_τ^{k+1} corresponding to the inactive sets are known à-priori as

$$\lambda_{\nu,\mathcal{I}_\nu}^{k+1} = 0,\ \lambda_{\tau,\mathcal{I}_\tau^+}^{k+1} = g_{\mathcal{I}_\tau^+},\ \lambda_{\tau,\mathcal{I}_\tau^-}^{k+1} = -g_{\mathcal{I}_\tau^-}. \tag{3.2}$$

The remaining components of y^{k+1} satisfy the reduced saddle-point linear system:

$$\begin{pmatrix} K & N_{\mathcal{A}_\nu}^\top & T_{\mathcal{A}_\tau}^\top \\ N_{\mathcal{A}_\nu} & 0 & 0 \\ T_{\mathcal{A}_\tau} & 0 & 0 \end{pmatrix} \begin{pmatrix} u^{k+1} \\ \lambda_{\nu,\mathcal{A}_\nu}^{k+1} \\ \lambda_{\tau,\mathcal{A}_\tau}^{k+1} \end{pmatrix} = \begin{pmatrix} f - T_{\mathcal{I}_\tau^+}^\top g_{\mathcal{I}_\tau^+} + T_{\mathcal{I}_\tau^-}^\top g_{\mathcal{I}_\tau^-} \\ d_{\mathcal{A}_\nu} \\ 0 \end{pmatrix}. \tag{3.3}$$

In the rest of this section, we show how to implement the iterative process (3.1) without necessity to generate the sequence $\{u^k\}$. To this end, we denote

$$\mathcal{A} = \mathcal{A}_\nu \cup \{i+m|\ i \in \mathcal{A}_\tau\} \quad \text{and} \quad \mathcal{I} = \{1, 2, \ldots, 2m\} \setminus \mathcal{A}. \tag{3.4}$$

We need also the matrices A, B and the vectors b, c introduced in (2.6). If $\mathcal{A} = \emptyset$, λ^{k+1} is fully determined by (3.2). Let $\mathcal{A}$ be non-empty. The second and third block equations in (3.3) read as

$$B_{\mathcal{A}} u^{k+1} = c_{\mathcal{A}}. \tag{3.5}$$

The first block equation in (3.3) yields

$$u^{k+1} = K^{-1}(f - B_{\mathcal{A}}^\top \lambda_{\mathcal{A}}^{k+1} - B_{\mathcal{I}}^\top \lambda_{\mathcal{I}}^{k+1}), \tag{3.6}$$

where $\lambda_{\mathcal{I}}^{k+1}$ is given by (3.2). Substituting (3.6) into (3.5), one can see that $\lambda_{\mathcal{A}}^{k+1}$ solves the linear system

$$A_{\mathcal{A}\mathcal{A}} \lambda_{\mathcal{A}}^{k+1} = \hat{b}_{\mathcal{A}} \tag{3.7}$$

with $\hat{b}_{\mathcal{A}} = b_{\mathcal{A}} - A_{\mathcal{A}\mathcal{I}} \lambda_{\mathcal{I}}^{k+1}$. Consequently, λ^{k+1} minimizes the cost function q subject to the constraints (3.2). Finally, note that

$$r^k = r(\lambda^k) = BK^{-1}B^\top \lambda^k - BK^{-1}f + c = -Bu^k + c = \begin{pmatrix} -Nu^k + d \\ -Tu^k \end{pmatrix}.$$

Therefore, one can omit $u = u^k$ also from the definitions of the active/inactive sets replacing it by the gradient values:

$$(Nu^k - d)_i = -r_i^k, \quad (Tu^k)_i = -r_{i+m}^k, \quad i \in \mathcal{M}.$$

We arrive at the following algorithmic scheme, in which α used in the definition of the reduced gradient (2.14) is chosen by ρ.

Algorithm SSNM

Given $\lambda^0 \in \mathbb{R}^{2m}$, $\varepsilon \geq 0$, and $\rho > 0$. For $k \geq 0$, compute:

(*Step 1*) If $\|\widetilde{r}_\rho(P_\Lambda(\lambda^k))\| \leq \varepsilon$, return $\lambda = P_\Lambda(\lambda^k)$, else go to step *Step 2*.

(*Step 2*) Assembly the active/inactive sets at λ^k:

$$\mathcal{A}_\nu = \{i \in \mathcal{M} : \lambda_i^k - \rho r_i^k \geq 0\}, \tag{3.8}$$
$$\mathcal{I}_\nu = \mathcal{M} \setminus \mathcal{A}_\nu, \tag{3.9}$$
$$\mathcal{I}_\tau^+ = \{i \in \mathcal{M} : g_i < \lambda_{i+m}^k - \rho r_{i+m}^k\}, \tag{3.10}$$
$$\mathcal{I}_\tau^- = \{i \in \mathcal{M} : \lambda_{i+m}^k - \rho r_{i+m}^k < -g_i\}, \tag{3.11}$$
$$\mathcal{A}_\tau = \mathcal{M} \setminus (I_\tau^+ \cup \mathcal{I}_\tau^-). \tag{3.12}$$

(*Step 3*) Find λ^{k+1} so that

$$\lambda^{k+1} = \arg\min q(\lambda) \quad \text{subject to (3.2)}. \tag{3.13}$$

The following theorem summarizes the standard convergence results.

Theorem 2 *Let $y^0 \in \mathbb{R}^{2m}$, $\varepsilon = 0$, and $\rho > 0$. Let $\{\lambda^k\}$ be the sequence generated by* Algorithm SSNM *and let λ^* be the solution to (2.9). (i) Then, $\{\lambda^k\}$ converges to λ^* superlinearly provided that $\|\lambda^* - \lambda^0\|$ is sufficiently small. (ii) If $\{\lambda^k\}$ converges, then it is the finite sequence and its last element is λ^*.*

Proof As Algorithm SSNM is equivalent to (3.1), we verify the assumptions of Theorem 1 to prove *(i)*. There is an appropriate permutation matrix $Q = Q(y)$ enabling us to transform $G^o(y)$ as follows:

$$QG^o(y)Q^\top = \left(\begin{array}{ccc|cc} K & N_{\mathcal{A}_\nu}^\top & T_{\mathcal{A}_\tau}^\top & N_{\mathcal{I}_\nu}^\top & T_{\mathcal{I}_\tau^+ \cup \mathcal{I}_\tau^-}^\top \\ -\rho N_{\mathcal{A}_\nu} & 0 & 0 & 0 & 0 \\ -\rho T_{\mathcal{A}_\tau} & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{array}\right).$$

We get the block upper triangular matrix with non-singular diagonal blocks. Therefore, $G^o(y)$ is non-singular for any $y \in \mathbb{R}^{n+2m}$. The boundedness of $\{\|G^o(y)^{-1}\| : y \in \mathbb{R}^{n+2m}\}$ yields from the fact that this set is finite and its largest element is the desired bound. The statement *(ii)* follows immediately from the finite number of the active/inactive sets. The theorem is proved. □

Remark 1 Note that ρ can be discarded from $\mathcal{A}_\nu$ and $\mathcal{I}_\nu$, when the inner subproblems in *Step 3* are solved exactly (and $\lambda^0 = 0$, e.g.), since either $\lambda_i^k = 0$ or $r_i^k = 0$. A similar observation is valid also for $\mathcal{A}_\tau$, $\mathcal{I}_\tau^+$, and $\mathcal{I}_\tau^-$ provided that λ^k is sufficiently close to λ^* and g is sufficiently large. We will see in next sections that ρ plays a significant role in our globally convergent algorithm.

4 Inexact implementation

The computational efficiency of the SSNM depends on a way how the inner subproblems are implemented. We propose to accept inexact solutions to (3.13), denoted again by λ^{k+1}, that are computed by few CGM iterations (see Appendix (A2)). It is referred by

$$\lambda^{k+1} = \mathrm{CGM}(A, b, \mathcal{A}, \lambda^{k+1,0}, tol^{k+1}), \tag{4.1}$$

where $\lambda^{k+1,0}$ is the initial CGM iteration and tol^{k+1} denotes the stopping tolerance. The implementation ideas are summarized by Algorithm ISSNM, where $err^k = \|\widetilde{r}_\rho(P_\Lambda(\lambda^k))\|$ stands for the precision achieved on the outer level. The value tol^{k+1} in *Step 3.1* respects err^k but, when the progress is not sufficient, it improves the previous tolerance tol^k. The inner initialization $\lambda^{k+1,0}$ in *Step 3.2* is chosen by the previous iteration λ^k and by the constraints (3.2).

Algorithm ISSNM (Inexact SSNM)

Given $\lambda^0 \in \mathbb{R}^{2m}$, $\varepsilon \geq 0$, $\rho > 0$, and $r_{tol}, c_{fact} \in (0,1)$.

Set $err^0 = \|\widetilde{r}_\rho(P_\Lambda(\lambda^0))\|$, $tol^0 = r_{tol}/c_{fact}$, and $k = 0$.

(*Step 1*) If $err^k \leq \varepsilon$, return $\lambda = P_\Lambda(\lambda^k)$, else go to step *Step 2*.

(*Step 2*) Assembly the active/inactive sets at λ^k by (3.8)-(3.12) and (3.4).

(*Step 3.1*) $tol^{k+1} = \min\{r_{tol} \times err^k/err^0, c_{fact} \times tol^k\}$

(*Step 3.2*) $\lambda^{k+1,0}_{\mathcal{A}} = \lambda^k_{\mathcal{A}}$, $\lambda^{k+1,0}_{\nu,\mathcal{I}_\nu} = 0$, $\lambda^{k+1,0}_{\tau,\mathcal{I}^+_\tau} = g_{\mathcal{I}^+_\tau}$, $\lambda^{k+1,0}_{\tau,\mathcal{I}^-_\tau} = -g_{\mathcal{I}^-_\tau}$

(*Step 3.3*) $\lambda^{k+1} = \mathrm{CGM}(A, b, \mathcal{A}, \lambda^{k+1,0}, tol^{k+1})$

(*Step 3.4*) $err^{k+1} = \|\widetilde{r}_\rho(P_\Lambda(\lambda^{k+1}))\|$, $k = k+1$, and go to *Step 1*.

Note that the convergence is not guaranteed by Theorem 2, since the inner subproblems are solved inexactly.

5 Globally convergent algorithm

The iterative process of Algorithm ISSNM may be interpreted as the restarted minimization procedure searching for the minima of (2.9). The key idea leading to the globally convergent result consists in modifying the algorithm so that the sequence $\{q(\lambda^k)\}$ generated by the new algorithm is monotonously decreasing. The following lemma proves the sufficient conditions, under which the restart of the CGM does not increase the value of q

Lemma 2 *Let $\lambda^k \in \Lambda$ and $\rho \in (0, 2\sigma^{-1}_{\max})$. If $\lambda^{k+1,0}$ is given by Step 3.2 of* Algorithm ISSNM*, then*

$$q(\lambda^k) \geq q(\lambda^{k+1,0}). \tag{5.1}$$

Proof Denote $\lambda = \lambda^k$, $\mu = \lambda^{k+1,0}$. Using $\lambda_{\mathcal{A}} = \mu_{\mathcal{A}}$ and (1.5), we get

$$q(\lambda^k) - q(\lambda^{k+1,0}) = (\lambda_{\mathcal{I}} - \mu_{\mathcal{I}})^\top r_{\mathcal{I}}(\lambda) - \frac{1}{2}(\lambda_{\mathcal{I}} - \mu_{\mathcal{I}})^\top A_{\mathcal{I}\mathcal{I}}(\lambda_{\mathcal{I}} - \mu_{\mathcal{I}})$$
$$\geq (\lambda_{\mathcal{I}} - \mu_{\mathcal{I}})^\top r_{\mathcal{I}}(\lambda) - \frac{1}{2}\sigma_{\max}\|\lambda_{\mathcal{I}} - \mu_{\mathcal{I}}\|^2. \tag{5.2}$$

The first assumption $\lambda^k \in \Lambda$ yields $0 \leq \lambda_i$ and $-g_i \leq \lambda_{i+m} \leq g_i$, $i \in \mathcal{M}$. For $i \in \mathcal{I}_\nu = \mathcal{I}_\nu(\lambda)$, we have $0 \leq \lambda_i < \rho r_i(\lambda)$ and $\mu_i = 0$ that implies

$$(\lambda_i - \mu_i)^2 \leq \rho(\lambda_i - \mu_i) r_i(\lambda).$$

For $i \in \mathcal{I}_\tau^+ = \mathcal{I}_\tau^+(\lambda)$, we have $0 \geq \lambda_{i+m} - g_i > \rho r_{i+m}(\lambda)$ and $\mu_{i+m} = g_i$ that implies

$$(\lambda_{i+m} - \mu_{i+m})^2 \leq \rho(\lambda_{i+m} - \mu_{i+m}) r_{i+m}(\lambda).$$

The same inequality holds also for $i \in \mathcal{I}_\tau^- = \mathcal{I}_\tau^-(\lambda)$, since in this case $0 \leq \lambda_{i+m} + g_i < \rho r_{i+m}(\lambda)$ and $\mu_{i+m} = -g_i$. We arrive at

$$\|\lambda_{\mathcal{I}} - \mu_{\mathcal{I}}\|^2 \leq \rho(\lambda_{\mathcal{I}} - \mu_{\mathcal{I}})^\top r_{\mathcal{I}}(\lambda).$$

Using this result in (5.2), we obtain

$$q(\lambda^k) - q(\lambda^{k+1,0}) \geq \left(\rho^{-1} - \frac{1}{2}\sigma_{\max}\right) \|\lambda_{\mathcal{I}} - \mu_I\|^2.$$

The right hand-side of the last inequality is non-negative due to $\rho \in (0, 2\sigma_{\max}^{-1})$. The lemma is proved. □

Since the inequality $q(\lambda^{k+1,0}) \geq q(\lambda^{k+1})$ yields immediately from the CGM, Lemma 2 guarantees that $\{q(\lambda^k)\}$ is (at least) non-increasing. Before introducing the globally convergent algorithm, we interpret *Step 3.2*.

Lemma 3 *Let $\lambda^{k+1,0}$ be given by Step 3.2 of* Algorithm ISSNM. *The following statement holds:*

$$\lambda_{\mathcal{I}}^{k+1,0} = P_{\Lambda,\mathcal{I}}\left(\lambda^k - \rho r(\lambda^k)\right). \tag{5.3}$$

Proof For $i \in \mathcal{I}_\nu = \mathcal{I}_\nu(\lambda^k)$, we have $\lambda_i^k - \rho r_i(\lambda^k) < 0$. Using (2.10), we get

$$P_{\Lambda,i}(\lambda^k - \rho r(\lambda^k)) = P_{\Lambda_\nu,i}(\lambda^k - \rho r(\lambda^k)) = 0 = \lambda_i^{k+1,0}.$$

For $i \in \mathcal{I}_\tau^+ = \mathcal{I}_\tau^+(\lambda^k)$, we have $g_i < \lambda_{i+m}^k - \rho r_{i+m}(\lambda^k)$. Using (2.11), we get

$$P_{\Lambda,i+m}(\lambda^k - \rho r(\lambda^k))) = P_{\Lambda_\tau,i}(\lambda^k - \rho r(\lambda^k)) = g_i = \lambda_{i+m}^{k+1,0}.$$

For $i \in \mathcal{I}_\tau^-(\lambda^k)$, we get the same result. The lemma is proved.

We modify ALGORITHM ISSNM in three points. First, we introduce the upper bound for ρ. Second, we ensure that all iterations belong to Λ. For that, we choose $\lambda^0 \in \Lambda$ and terminate the CGM loops before an iteration outside of Λ is generated. It is referred by the CGM_{feas} (see Appendix (A1)). Finally, the projection (5.3) in *Step 3.2* is extended onto all components that simplifies considerably the convergence analysis (see Appendix (A3) for further comments). The resulting algorithm reads as follows.

ALGORITHM GISSNM (Globally convergent ISSNM)

Given $\lambda^0 \in \Lambda$, $\varepsilon \geq 0$, $\rho \in (0, 2\sigma_{\max}^{-1})$, and $r_{tol}, c_{fact} \in (0,1)$.

Set $err^0 = \|\widetilde{r}_\rho(\lambda^0)\|$, $tol^0 = r_{tol}/c_{fact}$, and $k = 0$.

(*Step 1*) If $err^k \leq \varepsilon$ is small , return $\lambda = \lambda^k$, else go to step *Step 2*.

(*Step 2*) Assembly the active/inactive sets at λ^k by (3.8)-(3.12) and (3.4).

(*Step 3.1*) $tol^{k+1} = \min\{r_{tol} \times err^k/err^0, c_{fact} \times tol^k\}$

(*Step 3.2*) $\lambda^{k+1,0} = P_\Lambda(\lambda^k - \rho r(\lambda^k))$

(*Step 3.3*) $\lambda^{k+1} = \mathrm{CGM}_{feas}(A, b, \mathcal{A}, \lambda^{k+1,0}, tol^{k+1})$

(*Step 3.4*) $err^{k+1} = \|\widetilde{r}_\rho(\lambda^{k+1})\|$, $k = k+1$, and go to *Step 1*.

The following theorem proves the R-linear convergence rate.

Theorem 3 *Let $\lambda^0 \in \Lambda$, $\varepsilon = 0$, $\rho \in (0, 2\sigma_{\max}^{-1})$, and $r_{tol}, c_{fact} \in (0,1)$. Let $\sigma_{\min}$, $\sigma_{\max}$ be the smallest, largest eigenvalue of A, respectively, and let λ^* be the solution to (2.9). Let $\{\lambda^k\}$ denote the sequence generated by* ALGORITHM GISSNM. *The following statement holds:*
(i) the sequence $\{q(\lambda^k)\}$ decreases so that

$$q(\lambda^{k+1}) - q(\lambda^*) \leq \eta(\rho)\,(q(\lambda^k) - q(\lambda^*)), \tag{5.4}$$

where

$$\eta(\rho) = \begin{cases} 1 - \rho\sigma_{\min} & \text{for } \rho \in (0, \sigma_{\max}^{-1}], \\ 1 - (2\sigma_{\max}^{-1} - \rho)\sigma_{\min} & \text{for } \rho \in [\sigma_{\max}^{-1}, 2\sigma_{\max}^{-1}); \end{cases}$$

(ii) if $\{\lambda^k\}$ is finite, then its last element is λ^*;
(iii) if $\{\lambda^k\}$ is infinite, then it converges to λ^* R-linearly so that

$$\|\lambda^k - \lambda^*\| \leq C\,\eta(\rho)^{k/2}, \tag{5.5}$$

where $C = \sqrt{2(q(\lambda^0) - q(\lambda^*))/\sigma_{\min}}$.

Proof First of all, we show that each iteration is well-defined, if the stopping criterion in *Step 1* is not fulfilled (for $\varepsilon = 0$). As $\|\widetilde{r}_\rho(\lambda^k)\| > 0$, we have $\widetilde{r}_\rho(\lambda^k) \neq 0$ so that the definition of the reduced gradient (2.14) with $\lambda = \lambda^k$ and $\alpha = \rho$ yields

$$\lambda^{k+1,0} = P_\Lambda(\lambda^k - \rho r(\lambda^k)) = \lambda^k - \rho\widetilde{r}_\rho(\lambda^k).$$

Therefore, *Step 3.2* generates $\lambda^{k+1,0} \in \Lambda$ so that $\lambda^{k+1,0} \neq \lambda^k$. Then, the CGM_{feas} in *Step 3.3* produces λ^{k+1} such that

$$q(\lambda^{k+1}) \leq q(\lambda^{k+1,0}). \tag{5.6}$$

The equality in (5.6) occurs iff $\lambda^{k+1} = \lambda^{k+1,0}$ that is the case, when the CGM_{feas} does not generate any inner iteration due to a large value of tol^{k+1}. Otherwise, $\lambda^{k+1} \neq \lambda^{k+1,0}$ and the inequality (5.6) is strict. In both cases, we get $\lambda^{k+1} \in \Lambda$ so that $\lambda^{k+1} \neq \lambda^k$.

Lemma 1 with $\lambda = \lambda^k$ and $\alpha = \rho$ implies

$$q(\lambda^{k+1,0}) - q(\lambda^*) = q(P_\Lambda(\lambda^k - \rho r(\lambda^k))) - q(\lambda^*) \leq \eta(\rho)\,(q(\lambda^k)) - q(\lambda^*)). \tag{5.7}$$

Combining (5.7) with (5.6), we arrive at (5.4). As $\eta(\rho) < 1$, the sequence $\{q(\lambda^k)\}$ decreases, which proves the statement *(i)*.

Let $\{\lambda^k\}$ be finite. The last element of this sequence denoted by λ in *Step 1* fulfils $\|\widetilde{r}_\rho(\lambda)\| = \varepsilon = 0$ so tha $\lambda = \lambda^*$. The statement *(ii)* is proved.

Let $\{\lambda^k\}$ be infinite. To prove (5.5), we start from

$$q(\lambda^k) - q(\lambda^*) = (\lambda^k - \lambda^*)^\top r(\lambda^*) + \frac{1}{2}(\lambda^k - \lambda^*)^\top A(\lambda^k - \lambda^*).$$

Using (2.8) with $\lambda = \lambda^k$, we get

$$(\lambda^k - \lambda^*)^\top A(\lambda^k - \lambda^*) \leq 2\left(q(\lambda^k) - q(\lambda^*)\right).$$

Estimating the left hand-side using the smallest eigenvalue of A (see (1.5)) and the right hand-side inductively by (5.4), we obtain

$$\sigma_{\min}\|\lambda^k - \lambda^*\|^2 \leq 2\eta(\rho)^k\left(q(\lambda^0) - q(\lambda^*)\right).$$

The bound (5.5) is an easy consequence. The convergence of $\{\lambda^k\}$ to λ^* follows immediately from (5.5) that proves *(iii)*. □

Remark 2 For $\rho = \sigma_{\max}^{-1}$, we get the smallest value of the convergence factor:

$$\eta(\rho) = 1 - \kappa(A)^{-1},$$

where $\kappa(A) = \sigma_{\max}/\sigma_{\min}$ is the condition number of A.

Remark 3 The parameters r_{tol}, c_{fact} do not influence the upper bound in (5.5), but affect the performance of the CGM iterations. Their optimal values may be found experimentally.

Remark 4 Algorithm GISSNM is closely related to the algorithm MPRGP (Modified Proportioning with Reduced Gradient Projections) investigated in [10] for dual formulations of contact problems (see also [5,19,7]). A principal difference consists in a different definition of the active/inactive sets. As the active/inactive sets arising from the SSNM define a complementary decomposition of the index set, i.e. $\mathcal{A} \cap \mathcal{I} = \emptyset$ and $\mathcal{A} \cup \mathcal{I} = \{1, 2, \ldots, 2m\}$, one can use them also in the algorithm MPRGP. Moreover, the same convergence rate can be proved (see Remark 3.2 in [9]).

Remark 5 The statements of Theorem 3 remain valid even, if the CGM is replaced by an arbitrary (inner) solver satisfying (5.6). Taking $\lambda^{k+1} = \lambda^{k+1,0}$, we arrive at the projected gradient method [3]:

$$\lambda^{k+1} = P_\Lambda(\lambda^k - \rho r(\lambda^k)), \quad k = 0, 1, \dots$$

with $\lambda^0 \in \Lambda$ and $\rho \in (0, 2\sigma_{\max}^{-1})$.

6 Numerical experiments

We will assess the performance of our algorithms for contact of two steal bricks. The second benchmark represents more realistic contact problem with Coulomb friction.

6.1 Two steal bricks

Let us consider the problem (1.1)-(1.4) with the following data. Two plane elastic bodies $\Omega_1 = (0,3) \times (1,2)$, $\Omega_2 = (0,3) \times (0,1)$ are characterized by the Young modulus $E_k = 21.19 \cdot 10^{10}$ and the Poisson ratio $\nu_k = 0.277$ (steal) and $a_k = E_k\nu_k/(1-\nu_k^2)$, $b_k = E_k/(2(1+\nu_k))$, $k = 1, 2$. The decompositions of $\partial\Omega_1$ and $\partial\Omega_2$ read as follows: $\gamma_1^u = \{0\} \times (1,2)$, $\gamma_1^c = (0,3) \times \{1\}$, $\gamma_1^p = \partial\Omega_1 \setminus \overline{\gamma_1^u \cup \gamma_1^c}$ and $\gamma_2^u = \{0\} \times (0,1)$, $\gamma_2^c = (0,3) \times \{1\}$, $\gamma_2^p = \partial\Omega_2 \setminus \overline{\gamma_2^u \cup \gamma_2^c}$, respectively. The volume forces vanish for both bodies. The non-vanishing surface tractions $\boldsymbol{p}_1 = (p_{1x}, p_{1y})$ act on γ_1^p so that

$$\begin{aligned}
p_{1y}(s,2) &= p_{1y,L} + p_{1y,R}\, s, \quad s \in (0,3),\\
p_{1x}(3,s) &= p_{1x,B}(2-s) + p_{1x,U}\,(s-1), \quad s \in (1,2),\\
p_{1y}(3,s) &= p_{1y,B}(2-s) + p_{1y,U}\,(s-1), \quad s \in (1,2),
\end{aligned}$$

where $p_{1y,L} = -6 \cdot 10^7$, $p_{1y,R} = -1 \cdot 10^7$, $p_{1x,B} = p_{1x,U} = p_{1y,U} = 2 \cdot 10^7$, and $p_{1y,B} = 4 \cdot 10^7$. The constant slip bound is given as $g = 1.7 \cdot 10^7$ on γ_1^c. The problem is approximated using linear finite elements over regular triangulations, as it is seen in Figure 1.(a). The computed results are drawn in Figure 1.(b)-(d).

In tables below we report the total number of outer iterations *iter* (i.e. the last value of k) and the total number of matrix-vector multiplications with A denoted by n_A for different sizes of unknowns n and m. Note that n_A characterizes the computational complexity. A sufficiently accurate value of $\sigma_{\max}$ is estimated, if it is necessary, by few steps of the power method. The algorithms are initialized by $\lambda^0 = 0$ and terminated by $\varepsilon = 10^{-4} \cdot \|b\|$. This value of ε leads to sufficiently small relative equilibria in the algebraic contact problems (2.1)-(2.3) that are below the level 10^{-6}. The algorithms are written in Matlab 2009a. The computations are performed by Intel Core i5 (2.53GHz) with RAM 4GB.

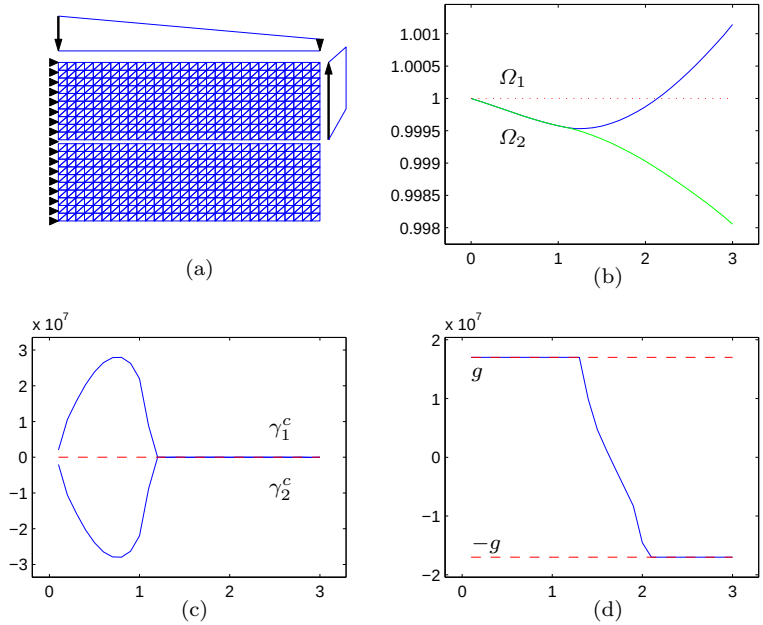

Fig. 1 (a) Bodies Ω_1 and Ω_2, the applied tractions, and the triangulations ($n = 1320$, $m = 30$). (b) Contact interface for the solution. (c) Distribution of the normal contact stress. (d) Distribution of the tangential contact stress.

Example 1. We simulate exact solving of subproblems (3.13) using Algorithm ISSNM with $r_{tol} = 1e{-}8$ (and $c_{fact} = 0.8$). Table 1 demonstrates the fact that computations do not depend on ρ except of the extremely large value of ρ, for which the algorithm oscillates; see Remark 1.

Table 1 SSNM: different ρ

ρ	$1e{-}2$	1	$1e2$	$1e4$	$1e12$
n/m	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$
5040/120	10/312	10/312	10/312	10/312	200/5373
19680/240	10/427	10/427	10/427	10/427	200/8403
43920/360	12/574	12/574	12/574	12/574	200/10275
77760/480	11/607	11/607	11/607	11/607	200/11876
121200/600	12/699	12/699	12/699	12/699	200/13061

if $iter = 200$, the default number of outer iterations is achieved

Example 2. The aim of this example is to find experimentally optimal values of r_{tol} and c_{fact} for Algorithm ISSNM. In Tables 2 and 3, we summarize some of our tests, from which one can propose: $r_{tol} = 0.1$, $c_{fact} = 0.8$. This choice will be used below including Algorithm GISSNM.

Table 2 ISSNM: $\rho = \sigma_{\max}^{-1}$, $r_{tol} = 0.1$, different c_{fact}

c_{fact}	0.2	0.6	0.8	0.9	0.99
n/m	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$
1320/60	07/73	06/46	07/35	07/35	07/38
11160/180	10/178	10/124	09/49	09/49	09/49
30600/300	10/207	10/136	09/48	09/48	09/48
59640/420	10/239	10/147	09/49	09/49	09/49
98280/540	11/295	11/177	09/51	09/51	09/51
146520/660	11/324	11/213	10/57	09/51	09/51
204360/780	12/381	12/244	10/59	10/59	10/59

Table 3 ISSNM: $\rho = \sigma_{\max}^{-1}$, $c_{fact} = 0.8$, different r_{tol}

r_{tol}	0.01	0.05	0.1	0.5	0.9
n/m	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$
1320/60	08/59	08/48	07/35	12/43	14/43
11160/180	10/78	09/54	09/49	12/50	17/50
30600/300	10/70	09/52	09/48	13/48	18/56
59640/420	09/63	09/56	09/49	13/49	20/58
98280/540	09/64	10/64	09/51	13/59	23/69
146520/660	10/75	09/55	10/57	15/58	16/60
204360/780	11/81	09/57	10/59	14/59	17/59

Example 3. Table 4 shows how Algorithm ISSNM behaves with respect to the value ρ taken as β multiple of $\sigma_{\max}^{-1}$. We observe that the dependence on β is weak.

Table 4 ISSNM: $\rho = \beta \cdot \sigma_{\max}^{-1}$

β	0.05	1	1.9	20	100
n/m	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$
1320/60	07/35	07/35	07/35	07/36	10/48
11160/180	09/49	09/49	09/49	08/41	10/52
30600/300	09/48	09/48	09/48	08/43	10/55
59640/420	09/49	09/49	09/49	10/59	11/62
98280/540	09/51	09/51	09/51	09/45	12/72
146520/660	09/53	10/57	10/57	10/59	10/54
204360/780	10/59	10/59	10/59	10/61	10/56

Example 4. It is seen from Table 5 that the dependence of Algorithm GISSNM on β is more significant. Although the algorithm is globally convergent for $0 < \beta < 2$, the best performance is achieved by $\beta = 15$. The convergence rate is slow for $\beta = 0.05$ and the algorithm oscillates for $\beta = 20$.

Table 5 GISSNM: $\rho = \beta \cdot \sigma_{\max}^{-1}$

β	0.05	1	1.9	15	20
n/m	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$	$iter/n_A$
1320/60	36/91	31/80	27/71	13/40	18/55
11160/180	109/237	82/184	64/146	26/69	200/410
30600/300	176/371	114/247	91/200	33/82	200/1827
59640/420	200/409	138/296	101/222	38/87	200/2013
98280/540	200/409	164/345	118/253	45/103	200/414
146520/660	200/409	172/364	126/270	47/105	200/2290
204360/780	200/409	186/386	138/290	49/108	200/2398

if $iter = 200$, the default number of outer iterations is achieved

Example 5. We compare Algorithm GISSNM with related algorithms. The MPRGP denotes the algorithm of [10]. The role of ρ is played by the stepsize $\widetilde{\alpha}$ in the MPRGP. Therefore, we set $\widetilde{\alpha} = \rho$. The MPRGP-S is the MPRGP modified so that the active/inactive sets are taken as in the SSNM (see Remark 4). In Table 6, we introduce the matrix-vector multiplications n_A leading to comparable relative equilibria on the level 10^{-6}. One can conclude that the performance of Algorithm GISSNM and the MPRGP is comparable. The use of the active/inactive sets from the SSNM may stabilize computations in the sense that the MPRGP-S converges for $\beta = 20$ (with the best performance) while the MPRGP oscillates in this case.

Table 6 Different algorithms: $\rho = \beta \cdot \sigma_{\max}^{-1}$

	GISSNM	MPRGP-S	MPRGP	GISSNM	MPRGP-S	MPRGP
β	1.9	1.9	1.9	15	20	15
n/m	n_A	n_A	n_A	n_A	n_A	n_A
30600/300	**200**	**252**	**185**	**82**	**78**	**78**
43920/360	**203**	**262**	**198**	**86**	**85**	**89**
59640/420	**222**	**278**	**219**	**87**	**90**	**83**
77760/480	**240**	**323**	**227**	**101**	**95**	**89**
98280/540	**253**	**339**	**250**	**103**	**102**	**96**
121200/600	**270**	**339**	**253**	**96**	**95**	**105**
146520/660	**270**	**356**	**262**	**105**	**109**	**115**
174240/720	**279**	**361**	**265**	**113**	**106**	**106**
204360/780	**290**	**381**	**286**	**108**	**109**	**110**

Example 7. Figure 2 shows iteration history for different variants of the method. We plot two graphs for each test: values of the cost function q (left) and values of the reduced gradient norm $\|\widetilde{r}_\rho\|$ (right). We observe values corresponding to the union of all CGM iterations, i.e. to the sequence $\bigcup_k \{\lambda^{k,j}\}$, that is drawn by blue solid lines. The values corresponding to the Newton iterations $\{\lambda^k\}$ are depicted by red stars. Here, we use notation introduced in the figure

caption. It is seen from SSNM, SSNM(5), and ISSNM(1) that the sequence $\{q(\lambda^k) : k \geq 1\}$ is monotonously increasing, if the globally convergent result does not hold. If ρ is to large, the algorithm may oscillate between two positions with the same reduced gradient norms, as it is seen in ISSNM(10^3). The globally convergent algorithm GISSNM(1.9) leads to the monotonously decreasing sequences of the cost function values, as it follows from the theory. However, the total efficiency is higher for GISSNM(15), when the convergence is non-monotonous due to $\rho > 2\sigma_{\max}^{-1}$. Finally, note that the finite termination property proved in Theorem 2.*(ii)* is confirmed by the high jump in values of the reduced gradient norm of SSNM after the last Newton iteration.

6.2 An extension to Coulomb friction

The benchmark in this subsection is a simplification of this one originally used in [8]. Let us consider the clamp joint of the support of mines; see Figure 3 (left). These joints are placed in the tunnels to prevent buckling. The arch of the support of mines consists of several segments that are tied by clamp clips. In the center of attention is the place of contact, where these segments overlap. For computations we interpret the situation as the planar problem. The bodies A and D are the clamp elements and bodies B and C are the supports. The parts of clamp join are steel with the Young moduls 210 000 MPa and the Poisson ratio 0.3. The thicknesses are different. Body A and D have 10 mm, B and C have 40 mm. Different thickness simulates different stiffness of the parts. Body C is fixed in space. The zero Dirichlet boundary conditions are depicted by the red lines in the figure. The surface tractions 10 kN simulate the action of bolts, see the red arrows in the figure. The rests of the boundaries are equipped by zero tractions and contact conditions. Contact conditions including the Coulomb friction law are prescribed where necessary with the constant coefficient of friction $\mathcal{F} = 0.3$. On Figure 3 (right) resulting total displacements are shown.

The contact problem with Coulomb friction is given by (1.1)-(1.4) with the following minor change in (1.4): g is replaced by $-\mathcal{F}\sigma_\nu$. The algebraic problem arising from the finite element approximation takes the form (2.1)-(2.3) with g_i replaced by $\mathcal{F}(\lambda_\nu + \rho(Nu - d))_i^+$, where the superscript "+" denotes the non-negative part of a number [1]. In our algorithm, we combine the SSNM with the TFETI domain decomposition method [6]. Each Newton iteration is given by the primal-dual linear system $\boldsymbol{A}y^{k+1} = \boldsymbol{b}$, where

$$\boldsymbol{A} = \left(\begin{array}{c|c|c|c} K & N^\top & T^\top & B_e^\top \\ \hline -\rho D(\mathcal{A}_\nu)N & D(\mathcal{I}_\nu) & 0 & 0 \\ \hline -\rho D(\mathcal{A}_\tau)T & \mathcal{F}D(\mathcal{A}_\nu)(D(\mathcal{I}_\tau^-) - D(\mathcal{I}_\tau^+)) & D(\mathcal{I}_\tau^+ \cup \mathcal{I}_\tau^-) & 0 \\ \hline B_e & 0 & 0 & 0 \end{array}\right),$$

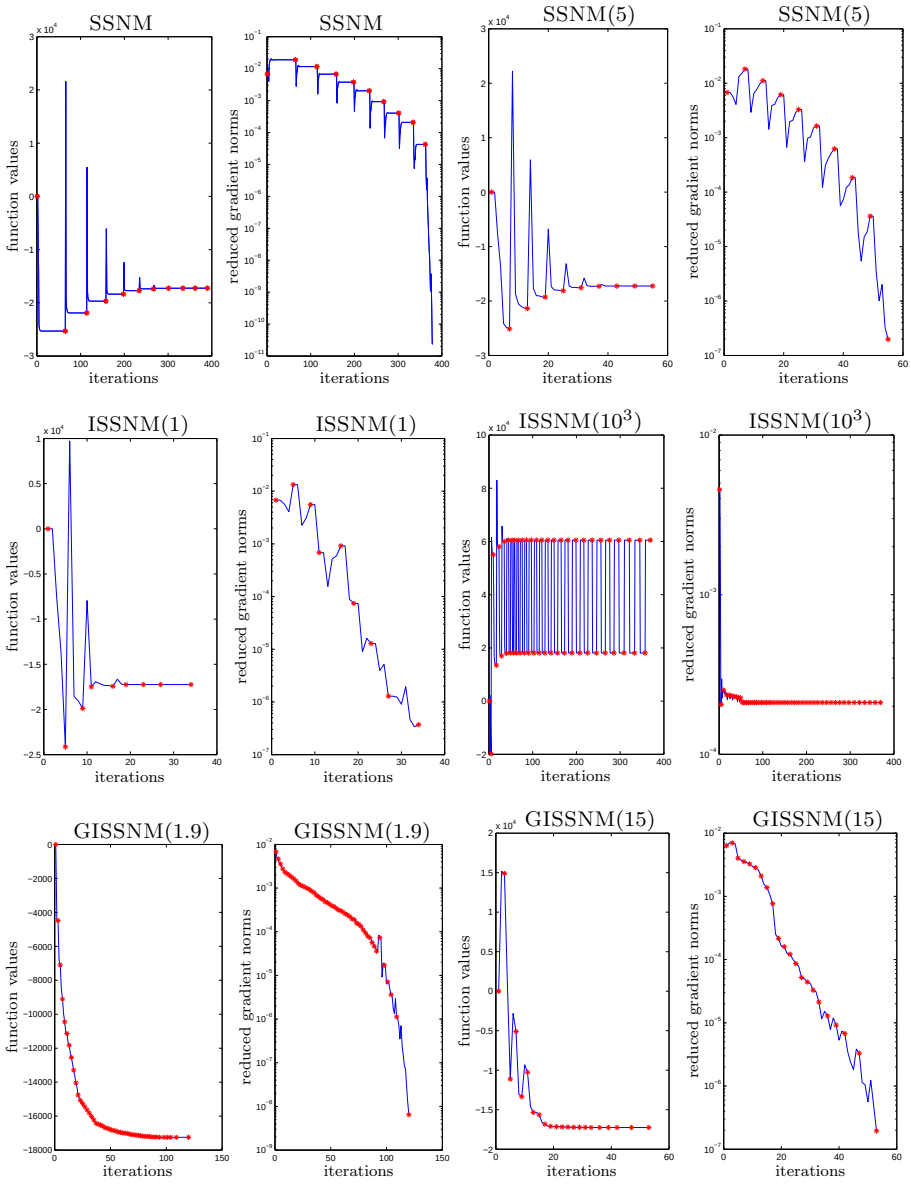

Fig. 2 Iteration history. SSNM: exact solving of inner subproblems; SSNM(5): five CGM iterations performed in each inner subproblem; ISSNM(β): $\beta = 1$ and $\beta = 10^3$; GISSNM(β): $\beta = 1.9$ and $\beta = 15$; where $\rho = \beta \cdot \sigma_{\max}^{-1}$.

$$y^{k+1} = \begin{pmatrix} u^{k+1} \\ \lambda_\nu^{k+1} \\ \lambda_\tau^{k+1} \\ \lambda_e^{k+1} \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} f \\ -\rho D(\mathcal{A}_\nu)d \\ 0 \\ 0 \end{pmatrix}.$$

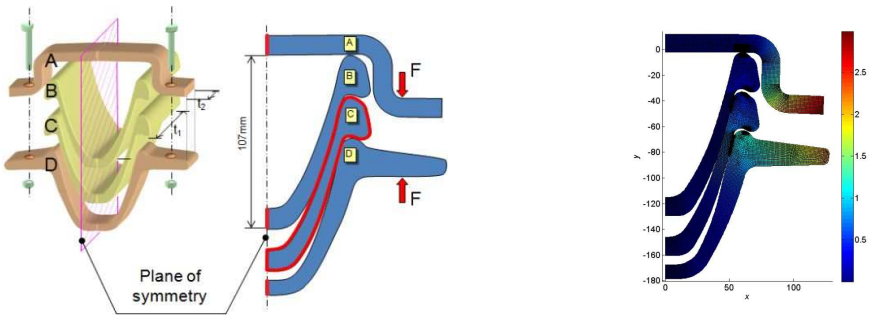

Fig. 3 Planar problem of the clamp joint (left). Total displacements (mm), $n/m = 78230/12718$ (right).

The active/inactive sets are defined by (3.8)-(3.12) with $\mathcal{F}(\lambda_i^k - \rho r_i^k)^+$ instead of g_i. B_e plays the role of the "gluing" matrix interconnecting solution components from artificial subdomains and λ_e^{k+1} approximates the respective Lagrange multiplier. The remaining entries are analogous to the case of Tresca friction. Eliminating u^{k+1}, we get the reduced linear system with two conceptual distinctions: as the block diagonal matrix K is singular, an efficient procedure form computing actions of a generalized inverse $K^\dagger$ is needed (see [20] for more details); as $\boldsymbol{A}$ is non-symmetric, the reduced linear systems are solved by the projected BiCGSTAB algorithm for non-symmetric matrices [13,21]. The adaptive inner precision control terminating BiCGSTAB iterations is used as in Algorithm ISSNM.

We will compare the computational efficiency of the proposed algorithm with the fixed point approach (FPA) combining the successive approximations with the augmented Lagrangian method (see [5] and references therein). As the symmetry of inner subproblems is guaranteed in the FPA, the MPRGP [10] is used. The stopping criteria are chosen so that comparable accuracies of the computed solutions are achieved (the relative equilibria are on the level 10^{-6}). Optimal values of other parameters are used, e.g. $\rho = \widetilde{\alpha} = 1.9 \cdot \sigma_{\max}^{-1}$, where $\sigma_{\max}$ is the largest eigenvalue of $A = BK^\dagger B^\top$, $B = (N^\top, T^\top, B_e^\top)^\top$, $r_{tol} = 0.5$, and $c_{fact} = 0.9$. All computations are performed by our Matsol library implemented in Matlab [17]. Decompositions onto artificial subdomains are made by METIS [15]. In Table 7, s denotes the number of subdomains, n is the number of primal variables (displacements), and m is the number of dual variables (Lagrange multipliers). The total number of outer iterations *iter* represents the Newton ones for the SSNM or combined the fixed point and the augmented Lagrangian ones for the FPA. The computational complexity is characterized by the total number n_K of matrix-vector multiplications by $K^\dagger$ that is the most expensive operation. We can conclude that the SSNM is more efficient than the FPA in 10 cases while the opposite situations appears in 4 cases. Note that the complexity of computations (characterized by *iter* and n_K) is not monotonous when the size of the problem increases. This observation is typical when solving problems with a complicated geometry, since,

e.g., contact nodes with a weak contact may differ on different discretization levels.

It is known that the discrete contact problem with Coulomb friction has a solution for any $\mathcal{F}$, but its uniqueness is not guaranteed for $\mathcal{F}$ large [26]. When the problem has a continuum of solutions [31], the matrix $\boldsymbol{A}$ is singular for the active/inactive sets in a solution.

Table 7 Coulomb friction: comparison of the algorithms

		FPA	SSNM
s	n/m	$iter/n_K$	$iter/n_K$
10	67454/1942	23/697	29/1317
20	68236/2724	53/2170	24/898
40	69332/3820	13/766	30/1424
100	71936/6424	39/962	27/1865
150	73716/8204	33/779	22/690
200	75376/9864	31/890	23/499
260	77262/11750	57/1112	19/575
360	79814/14302	97/4674	21/459
400	80548/15036	56/1310	27/929
460	81908/16396	18/1268	36/982
500	82742/17230	31/1374	34/3168
600	84632/19120	31/995	19/335
900	89660/24148	80/2010	37/1389
1000	91086/25568	102/2020	26/994

7 Conclusions and comments

We have analyzed the SSNM applied to the solution of contact problems with friction in 2D. We have shown for the Tresca friction law that the method is equivalent to a restarted minimization procedure. The globally convergent variant (Algorithm GISSNM) is closely related to the active set algorithms originally developed for dual formulations of contact problems. The analogous R-linear convergence rate holds. The numerical experiments indicate that the heuristic implementation (Algorithm ISSNM) is more efficient. Namely, it does not depend on the parameter determining the steplength along the projected gradient. We have experimentally tested the SSNM also for the Coulomb friction law, when the inner linear systems are given by nonsymmetric matrices. Although the convergence rate is not valid in this case, the experiments are promising. A computational difficulty is hidden in possible floating bodies (for Tresca as well as Coulomb friction), since an inappropriate choice of the active/inactive sets may lead to a singular inner linear system. It is the reason for prescribing Dirichlet boundary conditions for all bodies in our model problems. An appropriate regularization is possible in such a case.

Our analysis can be extended for 3D contact problems. However, it is more involved due to projections onto circles in $\mathbb{R}^2$ describing the friction law in 3D. The slanting function to such projection is constant outside the circle only along lines going to the center. We postpone the analysis of 3D case to a future work.

The SSNM is often used in function spaces, i.e. for problems in a continuous setting. It implies immediately a scalable behavior of the outer (Newton) loop, since each finite element approximation is a realization of the continuous method with the same convergence property (superlinear convergence rate). This result assumes, as we have already mentioned, exact solutions of inner subproblems. We have surprisingly observed scalable computations also for Algorithm ISSNM with the inexact implementation of the inner loop. A sophisticated theory enabling to prove the scalability of discrete algorithms related to our Algorithm GISSNM can be found in [5,8] (see also references therein). The use of the TFETI domain decomposition method plays an important role in this case.

Appendix

(A1) Let $\mathcal{A},\mathcal{I}$ be index sets such that $\mathcal{A}\cup\mathcal{I}=\{1,2,\dots,2m\}$ and $\mathcal{A}\cap\mathcal{I}=\emptyset$. We consider the problem

$$\bar\lambda = \arg\min q(\lambda) \quad \text{subject to } \lambda_{\mathcal{I}} = \lambda_{\mathcal{I}}^{k+1,0}, \tag{7.1}$$

where $q(\lambda)=\frac{1}{2}\lambda^\top A\lambda-\lambda^\top b$ is given by $A\in\mathbb{R}^{2m\times 2m}$ being symmetric, positive definite, $b\in\mathbb{R}^{2m}$, and $\lambda^{k+1,0}\in\Lambda$. Recall that $\Lambda=\{\lambda\in\mathbb{R}^{2m}: 0\le\lambda_i,\ |\lambda_{i+m}|\le g_i,\ i\in\mathcal{M}\}$ and $r=r(\lambda)=A\lambda-b$. We introduce the CGM_{feas} with the stopping tolerance $tol=tol^{k+1}>0$.

$\underline{\mathrm{CGM}_{feas}(A,b,\mathcal{A},\lambda^{k+1,0},tol)}$

(1) $r=A\lambda^{k+1,0}-b$, $p_{\mathcal{A}}=r_{\mathcal{A}}$, $p_{\mathcal{I}}=0$, $j=0$
(2) *while* $\|r_{\mathcal{A}}\|>tol\,\|b\|$ *and* $\lambda^{k+1,j}\in\Lambda$
(3) $\quad w=Ap,\ \alpha_{cg}=r^\top p/p^\top w$
(4) $\quad \lambda^{k+1,j+1}=\lambda^{k+1,j}-\alpha_{cg}p,\ j=j+1$
(5) $\quad r=r-\alpha_{cg}w,\ \gamma=r_{\mathcal{A}}^\top w_{\mathcal{A}}/p^\top w,\ p_{\mathcal{A}}=r_{\mathcal{A}}-\gamma p_{\mathcal{A}}$
(6) *endwhile*
(7) *if* $\lambda^{k+1,j}\in\Lambda$
(8) $\quad$ return $\lambda^{k+1}=\lambda^{k+1,j}$
(9) *elseif* $\lambda^{k+1,j}\notin\Lambda$
(10) $\quad \alpha_f=\max\{\alpha\in[0,1):\ \alpha\lambda^{k+1,j}-(1-\alpha)\lambda^{k+1,j-1}\in\Lambda\}$
(11) $\quad$ return $\lambda^{k+1}=\alpha_f\lambda^{k+1,j}-(1-\alpha_f)\lambda^{k+1,j-1}$
(12) *endif*

Steps (*2*)-(*6*) represent the standard CGM loop with added the feasibility test $\lambda^{k+1,j} \in \Lambda$ in step (*2*). Steps (*7*)-(*12*) define the result $\lambda^{k+1} \in \Lambda$ returned by the CGM_{feas}. If $\lambda^{k+1,j} \notin \Lambda$, λ^{k+1} is determined by the largest feasible steplength α_f in the last conjugate gradient direction. This is called the *half step* in [10,5,19,9,7]. Note that the CGM_{feas} performs typically few CGM iterations.

(A2) The CGM for solving (7.1) differs from the CGM_{feas} as follows: the feasibility test is omitted from step (*2*) and steps (*7*)-(*12*) are replaced by one return step $\lambda^{k+1} = \lambda^{k+1,j}$.

(A3) Let us replace the initial CGM iteration in *Step 3.2* of Algorithm GISSNM so that

$$\lambda_{\mathcal{A}}^{k+1,0} = \lambda_{\mathcal{A}}^{k}, \quad \lambda_{\mathcal{I}}^{k+1,0} = P_{\Lambda,\mathcal{I}}\left(\lambda^k - \rho r(\lambda^k)\right). \tag{7.2}$$

The first CGM iteration (given by step (*4*) of the CGM_{feas} with $j = 0$) satisfies:

$$\lambda_{\mathcal{I}}^{k+1,1} = \lambda_{\mathcal{I}}^{k+1,0} = P_{\Lambda,\mathcal{I}}\left(\lambda^k - \rho r\left(\lambda^k\right)\right)$$

and

$$\begin{aligned}
\lambda_{\mathcal{A}}^{k+1,1} &= \lambda_{\mathcal{A}}^{k+1,0} - \alpha_{cg} r_{\mathcal{A}}\left(\lambda^{k+1,0}\right) \\
&= \lambda_{\mathcal{A}}^{k+1,0} - \alpha_{cg}\left(A_{\mathcal{A}\mathcal{A}}\lambda_{\mathcal{A}}^{k+1,0} + A_{\mathcal{A}\mathcal{I}}\lambda_{\mathcal{I}}^{k+1,0} - b_{\mathcal{A}}\right) \\
&= \lambda_{\mathcal{A}}^{k} - \alpha_{cg}\left(A_{\mathcal{A}\mathcal{A}}\lambda_{\mathcal{A}}^{k} + A_{\mathcal{A}\mathcal{I}}P_{\Lambda,\mathcal{I}}\left(\lambda^k - \rho r\left(\lambda^k\right)\right) - b_{\mathcal{A}}\right) \\
&= \lambda_{\mathcal{A}}^{k} - \alpha_{cg}\left(r_{\mathcal{A}}\left(\lambda^k\right) + A_{\mathcal{A}\mathcal{I}}\left(P_{\Lambda,\mathcal{I}}\left(\lambda^k - \rho r\left(\lambda^k\right)\right) - \lambda_{\mathcal{I}}^{k}\right)\right) \\
&= \lambda_{\mathcal{A}}^{k} - \alpha_{cg} r_{\mathcal{A}}\left(\lambda^k\right) + \alpha_{cg}\rho A_{\mathcal{A}\mathcal{I}}\widetilde{r}_{\rho,\mathcal{I}}\left(\lambda^k\right) \\
&= \lambda_{\mathcal{A}}^{k} - \rho r_{\mathcal{A}}\left(\lambda^k\right) - (\alpha_{cg} - \rho) r_{\mathcal{A}}\left(\lambda^k\right) + \alpha_{cg}\rho A_{\mathcal{A}\mathcal{I}}\widetilde{r}_{\rho,\mathcal{I}}\left(\lambda^k\right),
\end{aligned}$$

where $\widetilde{r}_{\rho}$, denotes the reduced gradient (2.14) with $\alpha = \rho$. Due to the definition of the active/inactive sets (3.8)-(3.12), we have $P_{\Lambda,\mathcal{A}}\left(\lambda^k - \rho r\left(\lambda^k\right)\right) = \lambda_{\mathcal{A}}^{k} - \rho r_{\mathcal{A}}\left(\lambda^k\right)$ so that

$$\lambda^{k+1,1} = P_{\Lambda}\left(\lambda^k - \rho r\left(\lambda^k\right)\right) - s(\lambda^k) \tag{7.3}$$

with $s_{\mathcal{I}}(\lambda^k) = 0$ and $s_{\mathcal{A}}(\lambda^k) = (\alpha_{cg} - \rho) r_{\mathcal{A}}\left(\lambda^k\right) - \alpha_{cg}\rho A_{\mathcal{A}\mathcal{I}}\widetilde{r}_{\rho,\mathcal{I}}\left(\lambda^k\right)$. The following inequalities follows from the CGM:

$$q(\lambda^k) \geq q(\lambda^{k+1,0}) \geq q(\lambda^{k+1,1}).$$

Under assumptions that $\rho \leq \alpha_{cg}$ and $\widetilde{r}_{\rho,\mathcal{I}}\left(\lambda^k\right)$ is sufficiently small, one can prove:

$$q(\lambda^{k+1,0}) \geq q\left(P_{\Lambda}\left(\lambda^k - \rho r\left(\lambda^k\right)\right)\right) \geq q(\lambda^{k+1,1}). \tag{7.4}$$

The following interpretation yields from (7.3) and (7.4): if the initial CGM iteration is given by (7.2), the full projection $P_{\Lambda}\left(\lambda^k - \rho r\left(\lambda^k\right)\right)$ is inherently included in the SSNM after the first CGM iteration.

Acknowledgements This work was supported by the European Development Fund in the IT4Innovations Centre of Excellence project CZ.1.05/1.1.00/02.0070 (RK,AM), by the project Opportunity for young researchers CZ.1.07/2.3.00/30.0016 (AM) and by the grants P201/12/0671 (RK) and 13-30657P (AM) of the Grant Agency of the Czech Republic.

References

1. *P. Alart, A. Curnier*: A mixed formulation for frictional contact problems prone to Newton like solution methods. Comput. Methods Appl. Mech. Engrg. **92** (1991), 353–375.
2. *J. Bouchala, Z. Dostál, P. Vodstrčil*: Separable spherical constraints and the decrease of a quadratic function in the gradient projection step. J. Optim. Theory Appl. **157** (2013), 132–140.
3. *P. H. Calamai, J. J. Moré*: Projected gradient methods for linearly constrained problems. Mathematical Programming **39** (1987), 93–116.
4. *X. Chen, Z. Nashed, L. Qi*: Smoothing methods and semismooth methods for nondifferentiable operator equations. SIAM J. Numer. Anal. **38** (2000), 1200–1216.
5. *Z. Dostál*: Optimal quadratic programming algorithms: with applications to variational inequalities. SOIA 23, Springer US, New York 2009.
6. *Z. Dostál, D. Horák, R. Kučera*: Total FETI - an easier implementable variant of the FETI method for numerical solution of elliptic PDE. Communications in Numerical Methods in Engineering **22** (2006), 1155–1162.
7. *Z. Dostál, T. Kozubek*: An optimal algorithm and superrelaxation for minimization of a quadratic function subject to separable convex constraints with applications. Mathematical Programming **135** (2012), 195–220.
8. *Z. Dostál, T. Kozubek, A. Markopoulos, T. Brzobohatý, V. Vondrák, P. Horyl*: Scalable TFETI algorithm for two dimensional multibody contact problems with friction. Journal of Computational and Applied Mathematics, **235** (2010), 403–418.
9. *Z. Dostál, R. Kučera*: An optimal algorithm for minimization of quadratic functions with bounded spectrum subject to separable convex inequality and linear equality constraints. SIAM J. Optim. **20** (2010), 2913–2938.
10. *Z. Dostál, J. Schöberl*: Minimizing quadratic functions over non-negative cone with the rate of convergence and finite termination. Comput. Optim. Appl. **30** (2005), 23–44.
11. *F. Facchinei, J. S. Pang*: Finite-dimensional variational inequalities and complementarity problems. Vol. 1 and Vol. 2, Springer Series in Operations Research and Financial Engineering, Springer-Verlag, New York, 2003.
12. *G. H. Golub, C. F. Van Loan*: Matrix computation. The Johns Hopkins University Press: Baltimore, 1996.
13. *J. Haslinger, T. Kozubek, R. Kučera, G. Peichl*: Projected Schur complement method for solving non-symmetric saddle-point systems arising from fictitious domain approach. Numerical Linear Algebra with Applications **14** (2007), 713–739.
14. *K. Ito, K. Kunisch*: Semismooth Newton methods for variational inequalities of the first kind. ESAIM: Mathematical Modelling and Numerical Analysis **30** (2010), 41–62.
15. *G. Karypis, V. Kumar*: MeTis: Unstructured Graph Partitioning and Sparse Matrix Ordering System, Version 4.0. At *http://www.cs.umn.edu/∼metis*, University of Minnesota, Minneapolis, MN, 2009.
16. *N. Kikuchi, J. T. Oden*: Contact problems in elasticity: A study of variational inequalities and finite element methods. SIAM Studies in Applied Mathematics 8, SIAM, Philadelphia, PA, 1988.
17. *T. Kozubek, A. Markopoulos, T. Brzobohatý, R. Kučera, V. Vondrák, Z. Dostál*: MatSol - MATLAB efficient solvers for problems in engineering. At *http://matsol.vsb.cz/*.
18. *R. Kučera*: Minimizing quadratic functions with separable quadratic constraints. Optim. Methods Soft. **22** (2007), 453–467.
19. *R. Kučera*: Convergence rate of an optimization algorithm for minimizing quadratic functions with separable convex constraints. SIAM J. Optim. **19** (2008), 846–862.
20. *R. Kučera, T. Kozubek, A. Markopoulos*: On large-scale generalized inverses in solving two-by-two block linear systems. Linear Algebra and Its Applications **438** (2013), 3011–3029.

21. *R. Kučera, T. Kozubek, A. Markopoulos, J. Haslinger, L. Mocek*: Projected Krylov methods for solving non-symmetric two-by-two block linear systems arising from fictitious domain formulations. Advances in Electrical and Electronic Engineering **12** (2014), 131–143.
22. *K. Kunisch, G. Stadler*: Generalized Newton methods for the 2D-Signorini contact problem with friction in function space. M2AN Math. Model. Numer. Anal. **39** (2005), 827-854.
23. *S. Hüeber, A. Matei, B. Wohlmuth*: Efficient algorithms for problems with friction. SIAM J. Sci. Comput. **29** (2007), 70-92.
24. *S. Hüeber, G. Stadler, B. Wohlmuth*: A primal-dual active set algorithm for three-dimensional contact problems with Coulomb friction. SIAM J. Sci. Comput. **30** (2008), 572-596.
25. *M. Hintermüller, K. Ito, K. Kunisch*: The primal-dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13** (2003), 865-888.
26. *J. Haslinger*: Approximation of the Signorini problem with friction, obeying Coulomb law. Math. Meth. Appl. **5** (1983), 422–437.
27. *J. Haslinger, I. Hlaváček, J. Nečas*: Numerical methods for for unilateral problems in solid mechanics. In Handbook of Numerical Analysis, vol. IV, Ciarlet, P.G. and Lions, J.L. eds., North-Holland, Amsterdam, 1996, 313-485.
28. *J. Lee*: A strategy of finding an initial active set for inequality constrained quadratic programming problems. Submitted to Optim. Methods Soft. (2013).
29. *J. M. Martinez, L. Qi* : Inexact Newton methods for solving nonsmooth equations. Journal of Computational and Applied Mathematics **60** (1999), 127–145.
30. *J. Nocedal, S. J. Wright*: Numerical Optimization. Springer-Verlag, New York, 1999.
31. *S. Scholtes*: Introduction to Piecewise Differentiable Equations, Springer Briefs in Optimalization, Springer, Berlin, 2012.
32. *G. Stadler*: Semismooth Newton and augmented Lagrangian methods for a simplified friction problem. SIAM J. Optim. **15** (2004), 39-62.
33. *D. Sun, L. Sun*: On NCP-fuctions. Comput. Optim. Appl. **13** (1999), 201-220.
34. *B. I. Wohlmuth*: Variationally consistent discretization schemes and numerical algorithms for contact problems. Acta Numerica (2011), 569–734.