

On the finite element solution of frictionless contact problems using an exact penalty approach

Fabian SEWERIN¹ and Panayiotis PAPADOPOULOS^{2*}

¹*Institute of Solid Mechanics, Technische Universität Braunschweig, GERMANY*

²*Department of Mechanical Engineering, University of California, Berkeley, CA 94720-1740, USA*

Table of contents

1	Introduction	2
2	Elements of contact mechanics	5
3	Discrete formulation	7
4	Contact equality constraints	8
5	Contact inequality constraints	13
6	Numerical examples	18
6.1	Beam bending over an elastic foundation	19
6.2	Ironing of an elastic foundation	20
7	Conclusions	24
Appendix A Equality constraints and Dirichlet boundary conditions		29

Abstract

In this article, an exact penalty method is introduced for the enforcement of the impenetrability constraint in problems of frictionless two-body contact solved by the finite element method. A complete algorithmic implementation is presented, including an automated scheme for the selection of the penalty parameter. Numerical examples are employed to assess the accuracy and robustness of the exact penalty method in comparison to classical penalty and Lagrange multiplier alternatives.

Keywords: contact; finite element method; exact penalty; Lagrange multipliers

*Corresponding author

1 Introduction

The enforcement of inequality constraints in implicit dual finite element formulations of contact problems continues to present a challenge to both commercial and research codes. Indeed, the inherent non-smoothness of the impenetrability constraint is known to test the robustness of Newton-based iterative methods and occasionally lead to “flip-flop” and even divergent sequences. Various regularization techniques have long been a practical means of enforcing this constraint by removing it from the admissible solution space and imposing it instead as part of the solution of a suitably defined unconstrained extremization problem. The advantages and disadvantages of such regularization techniques are well-understood. Indeed, the classical penalty method, attributed to Courant [5], is easy to implement, but enforces the constraints only approximately before the increase of the penalty parameter leads to ill-conditioning of the tangent stiffness matrix of the underlying global Newton-based iteration scheme. Moreover, the choice of a suitable penalty parameter is problem-specific and an optimal balance between accurate enforcement and good conditioning is often elusive. Augmented Lagrangian methods, *e.g.*, with the standard Uzawa implementation, on the other hand, yield a sequence of unconstrained minimization problems for a given penalty parameter embedded in an update loop for the estimates of the Lagrange multipliers [18]. Although augmented Lagrangian methods lead to exact enforcement of the constraints for a finite value of the penalty parameter, the need for a top-level iteration increases the solution cost, especially in light of the linear rate of convergence [12]. As a countermeasure, Alart and Curnier [1, Section 4.1.2] explored interleaving the equilibrium iterations and the multiplier updates. At the other end of the spectrum, the classical Lagrange multiplier approach, in which the constrained minimization problem is reformulated as a saddle-point problem, involves no top-level iteration, yet leads to an indefinite linear system and an increase in the number of unknowns with an associated increase in the overall solution cost. In principle, however, the unknown Lagrange multipliers may be eliminated from the global linear system by static condensation, an operation that can be rendered efficient by the adoption of suitable dual shape functions, albeit at the expense of a more involved linearization procedure [13, 26].

Yet another approach for the enforcement of the impenetrability constraint that has recently gained momentum is based on Nitsche’s idea [22]. Here, the normal stress on the contact surface is approximated in terms of the displacements by means of the strong constitutive and kinematic laws. In order for the resulting unconstrained formulation to

satisfy coercivity, the weak form of the equilibrium equations is augmented by a penalty term that is reminiscent of classical penalty-based constraint enforcement techniques [2]. A key property of Nitsche formulations is that the penalty parameter is strongly tied to the underlying discretization, particularly the element size at the contact interface [19]. Apart from special cases, however, the discrete contact constraints are not satisfied exactly [19, 30]. Since, on a more abstract level, the material behavior of the contacting bodies and the contact kinematics are deeply intertwined in Nitsche’s approach, the contact formulation may not be as general and transferable as those resulting from more traditional constraint enforcement [30].

In the context of a weak enforcement of Dirichlet boundary conditions, Baiges *et al.* [3] showed that the additional penalty term in Nitsche’s formulation can be omitted if the normal stress is considered as an unknown (akin to the multiplier field in the Lagrange multiplier method) and computed by imposing a weak counterpart of the constitutive law, also see [17]. As with the Lagrange multiplier method, the discrete formulation obtained from this approach may also be cast into a purely displacement-based system by statically condensing the degrees of freedom associated with the normal stress at the interface.

Contrary to the methods described above, this article documents an exact penalty regularization method based on earlier work by Fletcher [7, 8] and Fletcher and Lill [10] on optimization. While a penalty parameter appears in the formulation, it is possible to show that the constrained minimization problem can be reformulated as an equivalent unconstrained one for a finite value of this parameter. As argued in Mukai and Polak [20, Chapter 1], the main difference between augmented Lagrangian and exact penalty methods is that the latter involve a *continuous and simultaneous* approximation of the Lagrange multipliers, thus completely circumventing the need for a costly top-level iteration. In the context of contact mechanics, this means that the Lagrange multipliers are explicitly computed from the displacement degrees of freedom by some formula, hence are eliminated as unknowns from the overall system. The original exact penalty algorithm in Mukai and Polak [20], which was intended for equality-constrained minimization problems, was later extended to problems involving equality and inequality constraints by Glad and Polak [11]. The latter forms the basis for the algorithmic development in the present work. A more general framework for exact penalty methods is presented in Bertsekas [4, Chapter 4.3] from a unified perspective.

A critical challenge with exact penalty methods is that a suitable value of the penalty parameter is essential for convergence, yet difficult to estimate *a priori*. For the special

case of quadratic objective functions with linear constraints, Fletcher [8] established lower bounds on the penalty parameter. However, these results do not readily carry over to more general minimization problems. As an alternative, Mukai and Polak [20] adopted the idea of automatically adjusting the penalty parameter as the unconstrained minimization algorithm progresses. Indeed, the analysis may commence with a small arbitrarily chosen penalty parameter which is automatically adjusted as soon as the unconstrained minimizer appears to approach a stationary point of the exact penalty function at which the constraints are violated. This idea was subsequently cast into an abstract algorithm model by Polak [25] and generalized to the case of inequality constraints by Glad and Polak [11].

Traditionally, constraint enforcement techniques for inequalities are based on some form of an active set strategy, thus allowing a reformulation of inequality constraints as equalities. The exact penalty approach proposed here adopts this strategy. Commensurate with the continuous approximation of the multipliers in terms of the displacements, both the multiplier iterations and active set updates are merged into the equilibrium scheme, obviating any top-level iteration scheme.

In this article, a complete implementation of the exact penalty constraint enforcement technique due to Glad and Polak [11] is presented and adapted to the finite element solution of the frictionless two-body contact problem. This includes an automatic update scheme for the penalty parameter and an averaging scheme for the global tangent stiffness to alleviate the active/inactive flip-flop of constraints in the course of the Newton-based equilibrium scheme. Furthermore, it is shown that Dirichlet boundary conditions for nodes that lie on the contact surface can be enforced in the standard fashion. Two numerical examples are employed to demonstrate the behavior of the exact penalty method both in absolute terms and also in comparison to other constraint enforcement methods. The evaluation of the proposed exact penalty method is very positive, on the basis of robustness, efficiency, and accuracy.

The article is structured as follows: Section 2 contains a brief review of contact mechanics fundamentals with emphasis on the dual formulation of the two-body problem in finite deformations. This is followed in Section 3 by a derivation of the discrete equations emanating from a displacement-based finite element approximation. Classical enforcement techniques are reviewed for the case of kinematic equality constraints in Section 4 and the exact penalty method is introduced in connection to these techniques. The extension to the general inequality constrained contact formulation is presented in Section 5. Two selected numerical examples are introduced and analyzed in Section 6, before concluding remarks

are offered in Section 7.

2 Elements of contact mechanics

Let $\mathcal{B}^{(1)}$ and $\mathcal{B}^{(2)}$ be two distinct bodies whose material points $X^{(\alpha)}$, $\alpha = 1, 2$, occupy distinct positions $\mathbf{X}^{(\alpha)}$, $\alpha = 1, 2$, in the reference configurations $\Omega^{(\alpha)}$, where $\Omega^{(\alpha)}$, $\alpha = 1, 2$, are taken as open and connected subsets of the Euclidean point-space $\mathcal{E}^3$. The deformation $\chi^{(\alpha)}$ is understood as a map on $\Omega^{(\alpha)}$ which associates bijectively each $\mathbf{X}^{(\alpha)} \in \Omega^{(\alpha)}$ to its respective position $\mathbf{x}^{(\alpha)}$ in the deformed configuration $\omega^{(\alpha)}$, such that

$$\chi^{(\alpha)} : \Omega^{(\alpha)} \rightarrow \omega^{(\alpha)}, \quad \mathbf{X}^{(\alpha)} \mapsto \mathbf{x}^{(\alpha)} = \chi^{(\alpha)}(\mathbf{X}^{(\alpha)}), \quad (2.1)$$

where $J^{(\alpha)} = \det \frac{\partial \chi^{(\alpha)}(\mathbf{X}^{(\alpha)})}{\partial \mathbf{X}^{(\alpha)}} > 0$. Given Equation (2.1), corresponding displacement fields $\mathbf{u}^{(\alpha)}(\mathbf{X}^{(\alpha)})$ may be defined as

$$\mathbf{u}^{(\alpha)}(\mathbf{X}^{(\alpha)}) = \chi^{(\alpha)}(\mathbf{X}^{(\alpha)}) - \mathbf{X}^{(\alpha)}. \quad (2.2)$$

In contact mechanics, it is common to assume that the boundary $\partial\Omega^{(\alpha)}$ of $\Omega^{(\alpha)}$ is decomposed into three open subsets $\Gamma_u^{(\alpha)}$, $\Gamma_q^{(\alpha)}$ and $\Gamma_c^{(\alpha)}$, such that

$$\partial\Omega^{(\alpha)} = \overline{\Gamma_u^{(\alpha)} \cup \Gamma_q^{(\alpha)} \cup \Gamma_c^{(\alpha)}}, \quad (2.3)$$

where the overbar denotes closure. Here, $\Gamma_u^{(\alpha)}$ is the Dirichlet boundary on which displacements are prescribed according to

$$\mathbf{u}^{(\alpha)}(\mathbf{X}^{(\alpha)}) = \bar{\mathbf{u}}^{(\alpha)}(\mathbf{X}^{(\alpha)}). \quad (2.4)$$

Moreover, $\Gamma_q^{(\alpha)}$ represents the Neumann boundary on which the surface tractions are imposed according to

$$\mathbf{t}^{(\alpha)}(\mathbf{X}^{(\alpha)}) = \bar{\mathbf{t}}^{(\alpha)}(\mathbf{X}^{(\alpha)}), \quad (2.5)$$

while $\Gamma_c^{(\alpha)}$ encompasses the reference locations of those material points which may come into contact with material points of the opposing body. If no contact occurs or if the contact constraints only apply to a subset of $\Gamma_c^{(\alpha)}$, then the remainder of $\Gamma_c^{(\alpha)}$ is assumed, with negligible loss of generality, to be subject to homogeneous Neumann boundary conditions. Furthermore, the images of $\Gamma_u^{(\alpha)}$, $\Gamma_q^{(\alpha)}$ and $\Gamma_c^{(\alpha)}$ under $\chi^{(\alpha)}$ are denoted $\gamma_u^{(\alpha)}$, $\gamma_q^{(\alpha)}$ and $\gamma_c^{(\alpha)}$, respectively, and are also subject to the decomposition

$$\partial\omega^{(\alpha)} = \overline{\gamma_u^{(\alpha)} \cup \gamma_q^{(\alpha)} \cup \gamma_c^{(\alpha)}}. \quad (2.6)$$

The contact boundary conditions are typically formulated by invoking the principle of impenetrability of matter, according to which

$$\omega^{(1)} \cap \omega^{(2)} = \emptyset . \quad (2.7)$$

A practical equivalent to this principle is deduced by way of a gap function on $\gamma_c^{(2)}$ which quantifies the (local) separation of the two bodies. Following Jones and Papadopoulos [15], suppose that $\gamma_c^{(\alpha)}$, $\alpha = 1, 2$, are oriented surfaces with unique outward unit normals $\mathbf{n}^{(\alpha)}$. Then, for a given point $\mathbf{x}^{(2)} \in \gamma_c^{(2)}$, introduce the point set

$$\mathcal{V}(\mathbf{x}^{(2)}) = \left\{ \mathbf{x}^{(1)} \in \gamma_c^{(1)} \mid (\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) \times \mathbf{n}^{(2)}(\mathbf{x}^{(2)}) = \mathbf{0}, \mathbf{n}^{(1)}(\mathbf{x}^{(1)}) \cdot \mathbf{n}^{(2)}(\mathbf{x}^{(2)}) \leq 0 \right\} \quad (2.8)$$

and, when $\mathcal{V}(\mathbf{x}^{(2)}) \neq \emptyset$, define a normal projection¹

$$\boldsymbol{\pi} : \gamma_c^{(2)} \rightarrow \gamma_c^{(1)}, \quad \mathbf{x}^{(2)} \mapsto \bar{\mathbf{x}}^{(1)} = \arg \min_{\mathbf{x}^{(1)} \in \mathcal{V}(\mathbf{x}^{(2)})} \left\| \mathbf{x}^{(1)} - \mathbf{x}^{(2)} \right\| . \quad (2.9)$$

Based on Equations (2.8) and (2.9), a gap function $g : \gamma_c^{(2)} \rightarrow \mathbb{R} \cup \{\infty\}$ may now be constructed as

$$\mathbf{x}^{(2)} \mapsto g(\mathbf{x}^{(2)}) = \begin{cases} (\boldsymbol{\pi}(\mathbf{x}^{(2)}) - \mathbf{x}^{(2)}) \cdot \mathbf{n}^{(2)}(\mathbf{x}^{(2)}) & \text{if } \mathcal{V}(\mathbf{x}^{(2)}) \neq \emptyset \\ \infty & \text{otherwise} \end{cases} , \quad (2.10)$$

as depicted in Figure 1. With reference to the potential contact surface $\gamma_c^{(2)}$, the impenetrability condition (2.7) may now be expressed by the inequality constraint condition

$$g(\mathbf{x}^{(2)}) \geq 0 , \quad \forall \mathbf{x}^{(2)} \in \gamma_c^{(2)} . \quad (2.11)$$

Clearly, the gap function in Equation (2.10) is geometrically biased in admitting that surface distances are measured from body 2 (the “slave”) to body 1 (the “master”), although nothing inherently prevents the reversal of the two roles. Also, explicit reference of the dependence of g on the geometry and deformation of body 1 by means of the projection $\boldsymbol{\pi}$ is suppressed here for notational brevity.

Adopting a dual approach, the Galerkin form of the equilibrium equations subject to the impenetrability constraint may be expressed as

$$\begin{aligned} A(\mathbf{u}, \mathbf{v}) - f(\mathbf{v}) - B(\lambda, \mathbf{v}) &= 0 , \\ C(\mu - \lambda, \mathbf{X} + \mathbf{u}) &\geq 0 , \end{aligned} \quad (2.12)$$

¹Note that Equation (2.9) does not, in general, constitute a closest-point projection.

for all weighting functions $\mathbf{v}$ and $\mu \leq 0$, where the displacement field $\mathbf{u} = (\mathbf{u}^{(1)}, \mathbf{u}^{(2)})$ and the Lagrange multiplier field $\lambda \leq 0$ are to be determined. The latter may be physically interpreted as the normal stress or negative contact pressure on $\gamma_c^{(2)}$ [16, Section 12.2]. Also, $\mathbf{v} = (\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$, $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, while the non-linear form $A(\mathbf{u}, \mathbf{v}) = \sum_{\alpha=1}^2 A^{(\alpha)}(\mathbf{u}^{(\alpha)}, \mathbf{v}^{(\alpha)})$ represents the combined stress-divergence contribution and $f(\mathbf{v}) = \sum_{\alpha=1}^2 f^{(\alpha)}(\mathbf{v}^{(\alpha)})$ the combined external force contribution in the two bodies. In addition, the non-linear form C associated with contact is defined as

$$C(\mu, \mathbf{X} + \mathbf{u}) = - \int_{\gamma_c^{(2)}} \mu(\mathbf{x}^{(2)}) g(\mathbf{x}^{(2)}) da \quad (2.13)$$

and

$$B(\lambda, \mathbf{v}) = DC(\lambda, \mathbf{X} + \mathbf{u})(\mathbf{x}, \mathbf{v}) , \quad (2.14)$$

where $D(\cdot)(\mathbf{x}, \mathbf{v})$ denotes the Gâteaux differential of $(\cdot)$ at $\mathbf{x} = \mathbf{X} + \mathbf{u}$ in the direction $\mathbf{v}$. Formal definitions of the admissible spaces for the displacement and multiplier fields may be found elsewhere, *e.g.*, in [15, 28].

Equations (2.12) may be shown to imply the complementarity condition

$$\lambda(\mathbf{x}^{(2)}) g(\mathbf{x}^{(2)}) = 0 , \quad (2.15)$$

for all $\mathbf{x}^{(2)} \in \gamma_c^{(2)}$, as well as the kinematic contact constraints in Equation (2.11). With the aid of Equations (2.13) and (2.15), Equation (2.12)₂ implies that

$$- \int_{\gamma_c^{(2)}} \mu(\mathbf{x}^{(2)}) g(\mathbf{x}^{(2)}) da \geq 0 , \quad (2.16)$$

for all $\mu \leq 0$. Strong (point-on-surface) or weak (surface-on-surface) enforcement of the impenetrability constraint may be imposed by means of Equations (2.11) or (2.16), respectively.

3 Discrete formulation

The finite element discretization of the governing equations follows a standard approach that yields piecewise polynomial displacement fields $\mathbf{u}_h^{(\alpha)}$ in $\Omega_h^{(\alpha)}$, whose nodal values are aggregated into a vector $\mathbf{d}$. Likewise, the Lagrange multiplier field λ is approximated on $\gamma_{c,h}^{(2)}$ as

$$\lambda_h(\mathbf{x}_h^{(2)}) = \sum_{j \in \mathcal{S}} \Phi_j(\mathbf{x}_h^{(2)}) \lambda_j , \quad (3.1)$$

where $\mathbf{x}_h^{(\alpha)} = \mathbf{X}_h^{(\alpha)} + \mathbf{u}_h^{(\alpha)}$ are the (interpolated) position fields in the current configuration, $\mathcal{S}$ is the index set of all nodes on the discretized boundary $\gamma_{c,h}^{(2)}$, Φ_j are the interpolation functions, and λ_j are the nodal values of λ_h . Again, all these nodal values are conveniently aggregated into a vector $\boldsymbol{\lambda} \leq \mathbf{0}$. As is standard Galerkin practice, analogous approximations are afforded to the weighting functions $\mathbf{v}$ and μ .

Given the preceding approximations, Equations (2.12)₁ and (2.16) give rise to the algebraic system

$$\begin{aligned} \mathbf{f}(\mathbf{d}) + \nabla^T \mathbf{g}(\mathbf{d}) \boldsymbol{\lambda} &= \mathbf{0} , \\ \mathbf{g}(\mathbf{d}) &\geq \mathbf{0} . \end{aligned} \quad (3.2)$$

Here, $\mathbf{f}(\mathbf{d}) = \mathbf{f}^i(\mathbf{d}) - \mathbf{f}^e$, with the internal force vector $\mathbf{f}^i$ emanating from the form A and the external force vector $\mathbf{f}^e$ from the form f . In addition, the vector function $\mathbf{g}$, with components

$$g_j(\mathbf{d}) = \int_{\gamma_c^{(2)}} \Phi_j(\mathbf{x}_h^{(2)}) g(\mathbf{x}_h^{(2)}) da \quad , \quad j \in \mathcal{S} , \quad (3.3)$$

represents the (weighted) gaps obtained with the aid of a smooth unit normal field $\mathbf{n}_h^{(2)}$ on $\gamma_{c,h}^{(2)}$, while $\nabla \mathbf{g}$ is the gradient of the vector function $\mathbf{g}$ with respect to $\mathbf{d}$. Equations (3.2) are supplemented by the discrete counterpart of Equation (2.15), in the form

$$\lambda_i g_i = 0 \quad , \quad i \in \mathcal{S} . \quad (3.4)$$

Surface-on-surface formulations of the contact problem are obtained by selecting piecewise smooth interpolation functions $\Phi_j(\mathbf{x}_h^{(2)})$, $j \in \mathcal{S}$. Alternatively, node-on-surface formulations result from letting $\Phi_j(\mathbf{x}_h^{(2)}) = \delta(\mathbf{x}_h^{(2)} - \mathbf{x}_j^{(2)})$, $j \in \mathcal{S}$, where $\mathbf{x}_j^{(2)}$ are the current locations of the nodes on $\gamma_{c,h}^{(2)}$ and δ stands for Dirac's δ -distribution.

4 Contact equality constraints

By way of introduction to the exact penalty method, attention is initially confined to the special case in which persistent contact is effected throughout $\gamma_{c,h}^{(2)}$. In this case, Equation (3.2)₂ holds as an equality, so that the discrete equilibrium equations (3.2)₁ are now supplemented by the constraint equation

$$\mathbf{g}(\mathbf{d}) = \mathbf{0} . \quad (4.1)$$

Note that Equation (4.1) trivially satisfies the discrete complementarity condition in Equation (3.4). Furthermore, the inequality condition $\boldsymbol{\lambda} \leq \mathbf{0}$ is ignored, which is tantamount to allowing the generation of tensile contact tractions.

In this simpler setting, the Lagrange multiplier method reduces to directly solving the coupled system of Equations (3.2)₁ and (4.1). However, the purpose of the present section is to motivate the exact penalty method within the broader class of penalty-based constraint enforcement methods, which are meant to yield approximate solutions without the added cost of explicitly solving for the Lagrange multipliers. To this end, start by replacing Equations (3.2)₁ and (4.1) with the modified equilibrium equations

$$\mathbf{r}_c(\mathbf{d}) = \mathbf{f}(\mathbf{d}) + \nabla(\boldsymbol{\lambda}_c^T(\mathbf{d})\mathbf{g}(\mathbf{d})) = \mathbf{0} , \quad (4.2)$$

where $\boldsymbol{\lambda}_c(\mathbf{d})$ represents an approximation of $\boldsymbol{\lambda}$ in terms of the displacement vector $\mathbf{d}$ for a given penalty parameter $c \geq 0$. It is clear that if $\mathbf{d}^*$ is a solution of Equation (4.2), such that $\mathbf{g}(\mathbf{d}^*) = \mathbf{0}$, then the pair $(\mathbf{d}^*, \boldsymbol{\lambda}_c(\mathbf{d}^*))$ constitutes a solution of the original Equations (3.2)₁ and (4.1).

The so-called classical penalty method [23] corresponds to

$$\boldsymbol{\lambda}_c(\mathbf{d}) = \frac{c}{2}\mathbf{g}(\mathbf{d}) . \quad (4.3)$$

In this case, higher values of c yield more accurate imposition of the contact constraint, yet render the algebraic system (4.2) increasingly ill-conditioned. A standard remedy is to set the value of c based on some prior knowledge of its desired range. Alternatively, one may prescribe a monotonically increasing sequence of penalty parameters $\{c_k\}_{k \geq 1}$, and solve the system for each c_k , $k = 1, 2, \dots$. In this iterative version of the classical penalty method, if the ratio $c_{k+1}/c_k > 1$ is kept sufficiently small, then the solution $\mathbf{d}^{k+1}$ of the $(k+1)$ -st system is close to that of the k -th system, and the impact of ill-conditioning on the non-linear system solver is partially alleviated by the proximity of the two successive approximations [4, Section 2.1].

As a starting point for the exact penalty method, consider the augmented Lagrangian method with an Uzawa update [12], which calls for Equation (4.2) to be solved repeatedly by setting

$$\begin{aligned} \boldsymbol{\lambda}_{c_k}(\mathbf{d}^k) &= \mathbf{z}^k + \frac{c_k}{2}\mathbf{g}(\mathbf{d}^k) \\ \mathbf{z}^{k+1} &= \boldsymbol{\lambda}_{c_k}(\mathbf{d}^k) \end{aligned} , \quad k = 0, 1, \dots , \quad (4.4)$$

with an initial guess $\mathbf{z}^0$, in order to obtain sequences $\{\mathbf{z}^k\}_{k \geq 0}$ and $\{\mathbf{d}^k\}_{k \geq 0}$ for a specified sequence of penalty parameters $\{c_k\}_{k \geq 0}$. Under mild assumptions, there exists a finite $\bar{c} > 0$, such that for $c_k \geq \bar{c}$, $k = 0, 1, \dots$, any limit point of the sequence $\{\mathbf{d}^k\}_{k \geq 0}$ obtained from Equations (4.2) and (4.4) is also a solution of the original Equations (3.2)₁ and (4.1)

[14, Section 2]. It is important to stress here that the initial penalty c_0 for which the method converges depends on the proximity of the initial guess $\mathbf{z}^0$ to the exact multipliers $\boldsymbol{\lambda}$. In fact, it can be shown that a poor choice for $\mathbf{z}^0$ may be counteracted by a sufficiently large value of c_0 [4, Section 2.2.4]. Some recommendations for constructing an efficient sequence $\{c_k\}_{k \geq 0}$ are summarized in [4, Section 2.2.5]. In computational practice, two strategies have been developed for the solution of Equations (4.2) and (4.4) [27]. In the first, Equations (4.2) and (4.4)₁ are solved to within a prescribed accuracy and the discrete Lagrange multipliers are subsequently updated using Equation (4.4)₂. Algorithmically, this leads to an equilibration process which is nested within a top-level iteration for the updated Lagrange multipliers. In the second, Equation (4.4)₂ may be invoked before a sufficiently accurate solution of Equations (4.2) and (4.4)₁ has been computed or, indeed, after every iteration of the solution scheme for Equations (4.2) and (4.4)₁ [1]. Clearly, in the former strategy the convergence rate of the non-linear solution method is preserved, although each update of the multipliers requires a new solution. In contrast, the latter strategy does not generally preserve the convergence rate of the non-linear solution scheme.

To circumvent the challenge of updating the multipliers in the augmented Lagrangian method, Fletcher [7, 8] and Fletcher and Lill [10] devised a class of so-called differentiable exact penalty functions. With particular reference to Equations (4.4), the underlying argument is that if sequences of multipliers $\{\mathbf{z}^k\}_{k \geq 0}$ and displacements $\{\mathbf{d}^k\}_{k \geq 0}$ may be constructed which converge, under certain conditions, to a solution $(\mathbf{d}^*, \boldsymbol{\lambda}^*)$, then it ought to be possible to construct a *continuous* approximation of the multipliers in terms of the displacement vector $\mathbf{d}$. To this end, Fletcher [8] proposed the formula

$$\boldsymbol{\lambda}_c(\mathbf{d}) = \boldsymbol{\psi}(\mathbf{d}) + \frac{c}{2} \mathbf{g}(\mathbf{d}) , \quad (4.5)$$

where

$$[\nabla \mathbf{g}(\mathbf{d}) \nabla^T \mathbf{g}(\mathbf{d})] \boldsymbol{\psi}(\mathbf{d}) = -\nabla \mathbf{g}(\mathbf{d}) \mathbf{f}(\mathbf{d}) . \quad (4.6)$$

Here, $\boldsymbol{\psi}(\mathbf{d})$ can be interpreted as a vector of nodal estimates for the normal stress on the contact interface, whereas the second term on the right-hand side of Equation (4.5) is required to restore coercivity of the unconstrained equilibrium equations (4.2). Conceptually, this penalty-like term acts as a safeguard against an excessively inaccurate estimate of the multipliers computed from Equation (4.6). The function $\boldsymbol{\psi}(\mathbf{d})$ in Equations (4.5) and (4.6) is chosen such that, for a given displacement vector $\mathbf{d}$,

$$\boldsymbol{\psi}(\mathbf{d}) = \arg \min_{\mathbf{z}} \|\mathbf{f}(\mathbf{d}) + \nabla^T \mathbf{g}(\mathbf{d}) \mathbf{z}\|^2 , \quad (4.7)$$

where $\|\cdot\|$ denotes the standard L_2 vector norm [8, 20]. This means that $\psi(\mathbf{d})$ is a minimizer, in the sense of least squares, of the constrained equilibrium equations $(3.2)_1$. A crucial observation here is that if $(\mathbf{d}^*, \boldsymbol{\lambda}^*)$ is a solution of Equations $(3.2)_1$ and (4.1) , then $\mathbf{d}^*$ is also a solution of Equation (4.2) with $\boldsymbol{\lambda}_c$ given by Equations (4.5) and (4.6) [20, Proposition 2.5]. Indeed, in this case the solution of Equation (4.2) reduces to

$$\mathbf{f}(\mathbf{d}^*) + \nabla^T \mathbf{g}(\mathbf{d}^*) \boldsymbol{\lambda}_c(\mathbf{d}^*) = \mathbf{0} , \quad (4.8)$$

where $\boldsymbol{\lambda}_c(\mathbf{d}^*) = \psi(\mathbf{d}^*)$ and, in view of Equations $(3.2)_1$ and (4.7) ,

$$\psi(\mathbf{d}^*) = \arg \min_{\mathbf{z}} \|\mathbf{f}(\mathbf{d}^*) + \nabla^T \mathbf{g}(\mathbf{d}^*) \mathbf{z}\|^2 = \boldsymbol{\lambda}^* . \quad (4.9)$$

This shows that the penalty method is exact. Fletcher [9] also offered an alternative derivation of Equations (4.5) and (4.6) , which includes the penalty term. In it, $\boldsymbol{\lambda}_c(\mathbf{d})$ is shown to be the Lagrange multiplier of the constrained quadratic minimization problem

$$\begin{aligned} \min_{\boldsymbol{\delta}} \left[\mathbf{f}^T(\mathbf{d}) \boldsymbol{\delta} + \frac{c}{2} \|\boldsymbol{\delta}\|^2 \right] , \\ \nabla \mathbf{g}(\mathbf{d}) \boldsymbol{\delta} + \mathbf{g}(\mathbf{d}) = \mathbf{0} \end{aligned} \quad (4.10)$$

for fixed $\mathbf{d}$ and any $c > 0$, [9, Theorem 2.1].

In analogy to the augmented Lagrangian method, there exists a *finite* penalty parameter $\bar{c} > 0$, such that, if $\mathbf{d}^*$ represents a solution of Equations (4.2) , (4.5) , and (4.6) for any $c \geq \bar{c}$, then $(\mathbf{d}^*, \boldsymbol{\lambda}_c(\mathbf{d}^*))$ constitutes also a solution of Equations $(3.2)_1$ and (4.1) [20, Proposition 2.6]. As already shown above, the converse of this result holds for all $c \geq 0$, that is, if a pair $(\mathbf{d}^*, \boldsymbol{\lambda}^*)$ forms a solution to Equations $(3.2)_1$ and (4.1) , then $\mathbf{d}^*$ is a solution of the unconstrained equilibrium equations (4.2) and $\boldsymbol{\lambda}_c(\mathbf{d}^*) = \boldsymbol{\lambda}^*$. As with the augmented Lagrangian method, there does not appear to exist a practical way to *a priori* estimate the parameter $\bar{c}$ [25]. To address this, Mukai and Polak [20] put forth the idea of employing an iterative solver for Equations (4.2) , (4.5) , and (4.6) starting with some initial value of c (say, c_0), and increasing c in the course of the iterations if its value is somehow deemed too small. This potential deficiency in the penalty parameter c is quantified with the aid of a continuous scalar test function $t_c(\mathbf{d})$ which, (i) attains a non-positive value at a solution $\mathbf{d}^*$ of Equations (4.2) , (4.5) , and (4.6) only if Equation (4.1) is satisfied, and (ii) for sufficiently large values of c , attains a non-positive value for all $\mathbf{d}$ in the vicinity of any iterate.

An iterative scheme for solving Equations (4.2) and (4.5) based on the Newton-Raphson method and featuring an automatic penalty adjustment guided by a test function $t_c(\mathbf{d})$ is

briefly outlined next. Specifically, given initial estimates $\mathbf{d}^0$ and c_0 , the iterative scheme is defined as

$$\begin{aligned} c_{k+1} &= \arg \min_c \left\{ c_k \leq c \mid t_c(\mathbf{d}^k) \leq 0 \right\} , \\ \left[\nabla \mathbf{f}(\mathbf{d}^k) + \mathbf{K}_{c_{k+1}}(\mathbf{d}^k) \right] \Delta \mathbf{d}^{k+1} &= -\nabla \mathbf{r}_{c_{k+1}}(\mathbf{d}^k) , \\ \mathbf{d}^{k+1} &= \mathbf{d}^k + \Delta \mathbf{d}^{k+1} , \end{aligned} \quad (4.11)$$

for $k = 0, 1, 2, \dots$. Here, $\mathbf{K}_c(\mathbf{d})$ is the stiffness matrix associated with the contact equality constraints, which, in view of Equation (4.5), is given by

$$\mathbf{K}_c = (\nabla^T \psi \nabla \mathbf{g} + \nabla^T \mathbf{g} \nabla \psi) + \psi^T \nabla(\nabla \mathbf{g}) + \mathbf{g}^T \nabla(\nabla \psi) + c \nabla^T \mathbf{g} \nabla \mathbf{g} + c \mathbf{g}^T \nabla(\nabla \mathbf{g}) , \quad (4.12)$$

where, with a slight abuse of notation, the term $\psi^T \nabla(\nabla \mathbf{g})$ above (and like terms henceforth) represents a two-dimensional square matrix and the gradient of $\psi(\mathbf{d})$ is obtained from Equation (4.6) as

$$[\nabla \mathbf{g} \nabla^T \mathbf{g}] \nabla \psi = -\nabla(\nabla \mathbf{g}) \mathbf{f} - \nabla \mathbf{g} \nabla \mathbf{f} - [\nabla(\nabla \mathbf{g}) \nabla^T \mathbf{g}] \psi - \nabla \mathbf{g} [\psi^T \nabla(\nabla \mathbf{g})] . \quad (4.13)$$

Note that the stiffness matrix in Equation (4.12) contains a term involving second derivatives of ψ and, hence, third derivatives of $\mathbf{g}$ which are excessively involved to derive analytically. However, since the term vanishes as $\mathbf{d}$ approaches a configuration for which the contact equality constraints in Equation (4.1) are satisfied, it may be ignored in practical computations, thereby leading to an inexact Newton method. Indeed, in line with the observations in [20], computational experience (see Section 6.2) suggests that, despite omitting this term in the update formula of Equation (4.11)₂, the rate of convergence of the iterative equilibrium scheme in Equations (4.11) is not appreciably retarded.

For an equality-constrained problem, as in this section, Mukai and Polak [20] proposed a specific form of the test function $t_c(\mathbf{d})$ in terms of the magnitude of $\mathbf{g}$ and the angle subtended by the residual vector $-\mathbf{r}_c(\mathbf{d})$ in Equation (4.2) (akin to a steepest-descent direction) and the normal projection

$$\boldsymbol{\pi}_{\mathbf{g}}(\mathbf{d}) = -\nabla^T \mathbf{g}(\mathbf{d}) [\nabla \mathbf{g}(\mathbf{d}) \nabla \mathbf{g}^T(\mathbf{d})]^{-1} \mathbf{g}(\mathbf{d}) \quad (4.14)$$

of $\mathbf{d}$ onto the surface $\mathbf{g}(\mathbf{d}) = \mathbf{0}$, according to

$$t_c(\mathbf{d}) = \|\mathbf{g}(\mathbf{d})\|^2 - \boldsymbol{\pi}_{\mathbf{g}}^T(\mathbf{d}) \mathbf{r}_c(\mathbf{d}) . \quad (4.15)$$

This test function readily satisfies condition (i), as defined earlier. Likewise, condition (ii) is satisfied upon observing that the matrix $\nabla \mathbf{g}(\mathbf{d}) \nabla^T \mathbf{g}(\mathbf{d})$ is positive-definite since the vectors

$\nabla g_i(\mathbf{d})$ are linearly independent. To restore dimensional consistency between the two terms on the right-hand side of Equation (4.15) (an implicit assumption in [20]), one needs to slightly reformulate the preceding test function by replacing the residual vector $-\mathbf{r}_c(\mathbf{d})$ with an approximate Newton-Raphson increment $-\tilde{\mathbf{K}}_c^{-1}\mathbf{r}_c(\mathbf{d})$ and inserting the positive definite matrix $\nabla \mathbf{g}(\mathbf{d})\nabla^T \mathbf{g}(\mathbf{d})$ in the definition of the norm of $\mathbf{g}(\mathbf{d})$. With these modifications, the test function becomes

$$t_c(\mathbf{d}) = \mathbf{g}^T(\mathbf{d}) [\nabla \mathbf{g}(\mathbf{d})\nabla^T \mathbf{g}(\mathbf{d})]^{-1} \mathbf{g}(\mathbf{d}) - \boldsymbol{\pi}_{\mathbf{g}}^T(\mathbf{d}) \tilde{\mathbf{K}}_c(\mathbf{d})^{-1} \mathbf{r}_c(\mathbf{d}) . \quad (4.16)$$

Here, $\tilde{\mathbf{K}}_c(\mathbf{d})$ represents an approximation of the tangent stiffness matrix $\nabla \mathbf{f}(\mathbf{d}) + \mathbf{K}_c(\mathbf{d})$, taken for computational simplicity to be

$$\tilde{\mathbf{K}}_c(\mathbf{d}) = \text{diag} [\nabla \mathbf{f}(\mathbf{d}) + \mathbf{K}_c(\mathbf{d})] . \quad (4.17)$$

Owing to the positive-definiteness of $\nabla \mathbf{g}(\mathbf{d})\nabla^T \mathbf{g}(\mathbf{d})$ and $\tilde{\mathbf{K}}_c(\mathbf{d})$, the modified test function in Equation (4.16) continues to observe the two desired conditions stated earlier.

Table 1 briefly summarizes the multiplier approximations inherent in the equality constraint enforcement techniques based on penalization reviewed above.

Method	Lagrange multipliers	Parameters
Penalty	$\boldsymbol{\lambda}_c(\mathbf{d}) = \frac{c}{2} \mathbf{g}(\mathbf{d})$	$c > 0$
Penalty (iterative)	$\boldsymbol{\lambda}_{c_k}(\mathbf{d}) = \frac{c_k}{2} \mathbf{g}(\mathbf{d})$	$\{c_k > 0\}_{k \geq 0}$
Augmented Lagrangian	$\boldsymbol{\lambda}_{c_k}(\mathbf{d}^k) = \mathbf{z}^k + \frac{c_k}{2} \mathbf{g}(\mathbf{d}^k), \mathbf{z}^{k+1} = \boldsymbol{\lambda}_{c_k}(\mathbf{d}^k)$	$\{c_k > 0\}_{k \geq 0}$
Exact Penalty	$\boldsymbol{\lambda}_{c_k}(\mathbf{d}) = - [\nabla \mathbf{g}(\mathbf{d})\nabla^T \mathbf{g}(\mathbf{d})]^{-1} \nabla \mathbf{g}(\mathbf{d}) \mathbf{f}(\mathbf{d}) + \frac{c_k}{2} \mathbf{g}(\mathbf{d})$	$\{c_k > 0\}_{k \geq 0}$

Table 1: Overview of different techniques for enforcing the contact equality constraints $\mathbf{g}(\mathbf{d}) = \mathbf{0}$ by considering an augmented discrete force balance and estimating the discrete Lagrange multipliers $\boldsymbol{\lambda}$ in terms of the displacements $\mathbf{d}$ (note that the sequences $\{c_k > 0\}_{k \geq 0}$ are monotonically increasing)

5 Contact inequality constraints

The extension of the exact penalty formulation presented in Section 4 to the general case involving inequality constraints is almost immediate if one takes into account that the constraints in Equations (3.2)₂ and (3.4), together with the condition $\boldsymbol{\lambda} \leq \mathbf{0}$, are equivalent to the equality constraints

$$\mathbf{a}_c(\mathbf{d}, \boldsymbol{\lambda}) = \frac{1}{c} [\min(\mathbf{0}, \boldsymbol{\lambda} + c\mathbf{g}(\mathbf{d})) - \boldsymbol{\lambda}] = \mathbf{0} , \quad (5.1)$$

for any $c > 0$ [11, Lemma 1]. In view of Equations (4.2) and (4.5), this reformulation suggests a modified set of equilibrium equations in the form

$$\mathbf{r}_c(\mathbf{d}) = \mathbf{f}(\mathbf{d}) + \nabla (\boldsymbol{\lambda}_c^T(\mathbf{d}) \mathbf{a}_c(\mathbf{d}, \boldsymbol{\phi}(\mathbf{d}))) = \mathbf{0} , \quad (5.2)$$

where the Lagrange multipliers are approximated according to

$$\boldsymbol{\lambda}_c(\mathbf{d}) = \boldsymbol{\phi}(\mathbf{d}) + \frac{c}{2} \mathbf{a}_c(\mathbf{d}, \boldsymbol{\phi}(\mathbf{d})) , \quad (5.3)$$

in terms of a function $\boldsymbol{\phi}$ (akin to the function $\boldsymbol{\psi}$ in Equation (4.5)) which renders $\boldsymbol{\lambda}_c$ a smooth function of $\mathbf{d}$. As in the equality-constrained case, $\boldsymbol{\phi}(\mathbf{d})$ can be interpreted as an approximation of the normal stress at the nodes, while the stabilizing penalty term in Equation (5.3) maintains coercivity. Postponing momentarily the exact definition of $\boldsymbol{\phi}$, proceed with the introduction of the active set $\mathcal{A}(\mathbf{d})$ and the inactive set $\mathcal{I}(\mathbf{d})$ as

$$\mathcal{A}(\mathbf{d}) = \{i \in \mathcal{S} \mid \phi_i(\mathbf{d}) + cg_i(\mathbf{d}) \leq 0\} , \quad (5.4)$$

$$\mathcal{I}(\mathbf{d}) = \{i \in \mathcal{S} \mid \phi_i(\mathbf{d}) + cg_i(\mathbf{d}) > 0\} , \quad (5.5)$$

where ϕ_i denotes the i -th component of $\boldsymbol{\phi}$. Since, clearly, $\mathcal{A}(\mathbf{d}) \cup \mathcal{I}(\mathbf{d}) = \mathcal{S}$ and $\mathcal{A}(\mathbf{d}) \cap \mathcal{I}(\mathbf{d}) = \emptyset$, Equation (5.1) readily implies that the i -th component of the function $\mathbf{a}_c(\mathbf{d}, \boldsymbol{\phi}(\mathbf{d}))$ is

$$a_{c,i}(\mathbf{d}, \boldsymbol{\phi}(\mathbf{d})) = \begin{cases} g_i(\mathbf{d}) & \text{if } i \in \mathcal{A} \\ -\frac{1}{c}\phi_i(\mathbf{d}) & \text{if } i \in \mathcal{I} \end{cases} , \quad (5.6)$$

hence also

$$\nabla a_{c,i}(\mathbf{d}, \boldsymbol{\phi}(\mathbf{d})) = \begin{cases} \nabla g_i(\mathbf{d}) & \text{if } i \in \mathcal{A} \\ -\frac{1}{c}\nabla \phi_i(\mathbf{d}) & \text{if } i \in \mathcal{I} \end{cases} . \quad (5.7)$$

Taking into account Equations (5.3), (5.6), and (5.7), the equilibrium equations (5.2) become

$$\mathbf{r}_c(\mathbf{d}) = \mathbf{f}(\mathbf{d}) + \underbrace{\nabla^T \mathbf{g}(\mathbf{d}) (\boldsymbol{\phi}(\mathbf{d}) + c \mathbf{a}_c(\mathbf{d}, \boldsymbol{\phi}(\mathbf{d}))) + \nabla^T \boldsymbol{\phi}(\mathbf{d}) \mathbf{a}_c(\mathbf{d}, \boldsymbol{\phi}(\mathbf{d}))}_{\mathbf{f}_c(\mathbf{d})} = \mathbf{0} . \quad (5.8)$$

This equation appears to have been first obtained by Glad and Polak [11]. Since $\mathbf{a}_c(\mathbf{d}, \boldsymbol{\phi}(\mathbf{d}))$ is continuous, albeit not differentiable in $\mathbf{d}$ and $\mathbf{f}(\mathbf{d})$, $\mathbf{g}(\mathbf{d})$, and $\boldsymbol{\phi}(\mathbf{d})$ vary smoothly in $\mathbf{d}$ by assumption, the augmented equation (5.8) is everywhere continuous and, thus, well-defined. In the context of frictionless contact, the vector $\mathbf{f}_c(\mathbf{d})$ in Equation (5.8) represents the discrete force contributions arising from the contact constraints.

Equation (5.8) is complemented by a rule for defining the multipliers $\lambda_c(\mathbf{d})$ or, equivalently, the function $\phi(\mathbf{d})$ in Equation (5.3). Starting from the pioneering efforts of Fletcher [9], Glad and Polak [11] proposed to adapt Equation (4.7) for inequality constraints according to

$$\phi(\mathbf{d}) = \arg \min_{\mathbf{z}} \left\{ \|\mathbf{f}(\mathbf{d}) + \nabla^T \mathbf{g}(\mathbf{d}) \mathbf{z}\|^2 + \mathbf{z}^T \mathbf{G}(\mathbf{d}) \mathbf{z} \right\} , \quad (5.9)$$

where $\mathbf{G}(\mathbf{d}) = (\text{diag } \mathbf{g})^2$, that is a diagonal square matrix with entries $g_i^2(\mathbf{d})$, $i \in \mathcal{S}$. Equation (5.9) yields a unique $\phi(\mathbf{d})$, which, in analogy to Equation (4.6), satisfies

$$[\nabla \mathbf{g} \nabla^T \mathbf{g} + \mathbf{G}] \phi = -\nabla \mathbf{g} \mathbf{f} . \quad (5.10)$$

Note that explicit reference to the argument $\mathbf{d}$ is omitted here and henceforth, when possible. Given the definition of $\mathbf{G}$, the gradient of $\phi(\mathbf{d})$ may now be readily determined from Equation (5.10) as

$$\begin{aligned} [\nabla \mathbf{g} \nabla^T \mathbf{g} + \mathbf{G}] \nabla \phi &= -\nabla(\nabla \mathbf{g}) \mathbf{f} - \nabla \mathbf{g} \nabla \mathbf{f} - [\nabla(\nabla \mathbf{g}) \nabla^T \mathbf{g}] \phi - \nabla \mathbf{g} [\phi^T \nabla(\nabla \mathbf{g})] \\ &\quad - 2\phi [\mathbf{g}^T \nabla \mathbf{g}] . \end{aligned} \quad (5.11)$$

Compared to more established constraint enforcement techniques, the linear systems in Equations (5.10) and (5.11) represent an additional source of computational expense for the exact penalty method. Furthermore, Equation (5.11) demonstrates that the discrete equilibrium equations (5.8) involve second derivatives of the discrete gap vector $\mathbf{g}(\mathbf{d})$. This is in contrast to the standard penalty, augmented Lagrangian, and Lagrange multiplier methods in which second derivatives of $\mathbf{g}(\mathbf{d})$ contribute only to the tangent stiffness matrix obtained from a consistent linearization. Thus, the exact penalty method draws on the second derivatives of $\mathbf{g}(\mathbf{d})$ as an additional source of information to effect augmentation at a finite value of the penalty parameter.

Although the equality constraint function $\mathbf{a}_c(\mathbf{d}, \phi(\mathbf{d}))$ varies continuously with $\mathbf{d}$, its gradient $\nabla \mathbf{a}_c(\mathbf{d}, \phi(\mathbf{d}))$ encounters a discontinuity at points $\mathbf{d}$ where the strict complementarity condition $\phi_i(\mathbf{d}) + c g_i(\mathbf{d}) \neq 0$ is violated for some $i \in \mathcal{S}$. This is known to impede (if not outright prevent) convergence of iterative methods used to solve the equilibrium equations. In order to develop a rational approach toward linearizing Equation (5.8), one may start by separately considering the cases in which *all* constraints are either active or inactive according to the criteria set in Equations (5.4) and (5.5). In the former case, the contact forces $\mathbf{f}_c(\mathbf{d})$, denoted $\mathbf{f}_c^{(a)}(\mathbf{d})$, are defined as in Equation (5.8) with $\mathbf{a}_c$ replaced by $\mathbf{g}$, as stipulated by Equation (5.6), while the corresponding stiffness $\mathbf{K}_c^{(a)}(\mathbf{d})$ is given by

Equation (4.12) with $\psi(\mathbf{d})$ replaced by $\phi(\mathbf{d})$, that is,

$$\mathbf{K}_c^{(a)} = \nabla^T \phi \nabla \mathbf{g} + \nabla^T \mathbf{g} \nabla \phi + \phi^T \nabla (\nabla \mathbf{g}) + \mathbf{g}^T \nabla (\nabla \phi) + c \nabla^T \mathbf{g} \nabla \mathbf{g} + c \mathbf{g}^T \nabla (\nabla \mathbf{g}) . \quad (5.12)$$

By contrast, if all constraints are inactive, then Equations (5.6) and (5.8) imply that the contact forces, denoted now $\mathbf{f}_c^{(i)}(\mathbf{d})$, read

$$\mathbf{f}_c^{(i)} = -\frac{1}{c} \nabla^T \phi \phi . \quad (5.13)$$

Taking the gradient of Equation (5.13) leads to an expression for the contact stiffness $\mathbf{K}_c^{(i)}$ in the inactive case, as

$$\mathbf{K}_c^{(i)} = -\frac{1}{c} \nabla^T \phi \nabla \phi - \frac{1}{c} [\phi^T \nabla (\nabla \phi)] . \quad (5.14)$$

As previously argued for the equality-constrained case in Section 4, one may neglect the terms that involve second derivatives of $\phi(\mathbf{d})$ (hence, third derivatives of $\mathbf{g}(\mathbf{d})$) in the contact stiffnesses of Equations (5.12) and (5.14).

For the general case of partial contact, the discontinuity in the contact stiffness may be circumvented by constructing an approximate stiffness which is always continuous in $\mathbf{d}$ and smoothly conforms to the exact stiffness as $\mathbf{d}$ approaches a point at which the constraints of Equation (5.1) are satisfied. To this end, Glad and Polak [11] introduced a weighting vector $\mathbf{b}(\mathbf{d})$ with components $b_i(\mathbf{d}) \in [0, 1]$, $i \in \mathcal{S}$, and defined such an approximate stiffness as an appropriately weighted average of the limiting stiffnesses $\mathbf{K}_c^{(a)}$ and $\mathbf{K}_c^{(i)}$ of Equations (5.12) and (5.14). In particular, assuming strict complementarity for all displacements $\mathbf{d}$, Glad and Polak [11] proposed (to within a minor modification) the formula

$$b_i = \beta(g_i, \phi_i) = \frac{|\phi_i|}{|cg_i| + |\phi_i|} . \quad (5.15)$$

Clearly, as an iterate $\mathbf{d}$ approaches a point $\mathbf{d}^*$ at which the constraints in Equation (5.1) are satisfied, the weight $b_i(\mathbf{d})$ approaches either the value 1 (for an active constraint, $i \in \mathcal{A}(\mathbf{d}^*)$) or 0 (for an inactive constraint, $i \in \mathcal{I}(\mathbf{d}^*)$), while always remaining bounded in $[0, 1]$. The weighted stiffness is subsequently defined as

$$\begin{aligned} \mathbf{K}_c &= \nabla^T \phi \text{diag}(\mathbf{b}) \nabla \mathbf{g} + \nabla^T \mathbf{g} \text{diag}(\mathbf{b}) \nabla \phi + c \nabla^T \mathbf{g} \text{diag}(\mathbf{b}) \nabla \mathbf{g} \\ &\quad + (\phi + c\mathbf{g})^T \text{diag}(\mathbf{b}) \nabla (\nabla \mathbf{g}) - \frac{1}{c} \nabla^T \phi (\mathbf{I} - \text{diag}(\mathbf{b})) \nabla \phi , \end{aligned} \quad (5.16)$$

where $\text{diag}(\mathbf{b})$ represents the diagonal matrix with entries b_i , $i \in \mathcal{S}$, $\mathbf{I}$ is the identity matrix in the dimension of $\mathcal{S}$, and all the stiffness terms involving third derivatives of $\mathbf{g}$ are omitted.

In practice, there are instances in which active contact forces are accompanied by a purely inactive stiffness and *vice versa*. This well-known pathology can have a negative effect on the convergence of the equilibrium scheme. By way of example, suppose that, for some $i \in \mathcal{S}$, an iterate $\mathbf{d}$ yields $g_i(\mathbf{d}) < 0$ and $\phi_i = 0$, hence $\phi_i(\mathbf{d}) + cg_i(\mathbf{d}) < 0$, $i \in \mathcal{A}(\mathbf{d})$, and the corresponding contact forces are active. Yet, Equation (5.15) mandates that $b_i(\mathbf{d}) = 0$, so that the contribution of the i -th constraint to the contact stiffness reduces to its associated inactive stiffness term. Conversely, let an iterate $\mathbf{d}$ yield $g_i(\mathbf{d}) = 0$ and $\phi_i(\mathbf{d}) > 0$. This case frequently arises when a body is about to be released from contact. Here, owing to $\phi_i(\mathbf{d}) + cg_i(\mathbf{d}) > 0$ and $b_i(\mathbf{d}) = 1$, the inactive contact forces are accompanied by an active contact stiffness. In the test cases considered in this work, such situations are prevented by redefining the weights $b_i(\mathbf{d})$ as

$$b_i(\mathbf{d}) = \beta(\theta_i(\mathbf{d})) = \begin{cases} 0 & \text{if } \theta_i \in (-\pi, -\frac{\pi}{2}] \\ (\theta_i + \frac{\pi}{2})\frac{2}{\pi} & \text{if } \theta_i \in (-\frac{\pi}{2}, 0) \\ 1 & \text{if } \theta_i \in [0, \frac{\pi}{2}] \\ (\pi - \theta_i)\frac{2}{\pi} & \text{if } \theta_i \in (\frac{\pi}{2}, \pi] \end{cases}. \quad (5.17)$$

Here, $\theta_i(\mathbf{d}) = \text{Arg}(cg_i(\mathbf{d}) + \iota\phi_i(\mathbf{d}))$ represents the polar angle of the point $(cg_i(\mathbf{d}), \phi_i(\mathbf{d}))$ on the complex plane, where “Arg” is the complex argument function with principal values in $(-\pi, \pi]$ and ι is the imaginary unit. For illustration, both the original function β in Equation (5.15) and the alternative one of Equation (5.17) are depicted in Figure 2.

Extending the work in [20] to account for inequality constraints, Glad and Polak [11] devised a test function $t_c(\mathbf{d})$ for controlling the growth of c in the course of seeking a solution to the equilibrium equations (5.2). Recall from Section 4 that for a given $\mathbf{d}$ that satisfies Equation (5.2), the condition $t_c(\mathbf{d}) > 0$ indicates that the contact constraints in Equation (5.1) are violated. Moreover, by its construction, there must exist a penalty parameter $\tilde{c}$ such that, for any displacement sufficiently close to an iterate $\mathbf{d}$, t_c turns non-positive for all $c \geq \tilde{c}$. In a slight departure from the algorithm in Glad and Polak [11], a test function for the inequality-constrained problem is constructed by replacing, in Equation (4.16), the function $\mathbf{a}_c(\mathbf{d}, \phi(\mathbf{d}))$ for $\mathbf{g}(\mathbf{d})$, which leads to

$$t_c = \mathbf{a}_c^T (\nabla \mathbf{a}_c \nabla^T \mathbf{a}_c)^{-1} \mathbf{a}_c - \mathbf{p}_c^T \tilde{\mathbf{K}}_c^{-1} \mathbf{r}_c \quad (5.18)$$

where $\mathbf{p}_c = -\nabla^T \mathbf{a}_c (\nabla \mathbf{a}_c \nabla^T \mathbf{a}_c)^{-1} \mathbf{a}_c$ represents the Newton-Raphson step towards the constraint surface $\mathbf{a}_c(\mathbf{d}, \phi(\mathbf{d})) = \mathbf{0}$ and $\tilde{\mathbf{K}}_c(\mathbf{d})$ is in the form of Equation (4.17).

Strict complementarity near a solution to Equations (3.2)₁ and (5.1) is required for Equations (5.15) and (5.17) to be well-defined. This is a common assumption, as it ensures

that there is an open neighborhood in which the equilibrium solution $\mathbf{d}^*$ of Equation (5.8) satisfying the constraints in Equation (5.1) is unique [11, p. 148], thus preventing loss in regularity of $\nabla \mathbf{f}(\mathbf{d}) + \mathbf{K}_c(\mathbf{d})$.

Box 1 briefly summarizes the main algorithmic steps of the exact penalty equilibrium scheme with automatic penalty adjustment. Also, Appendix A addresses the case of simultaneous contact and Dirichlet boundary conditions in the context of the proposed exact penalty method.

6 Numerical examples

In this section, two numerical examples are presented to assess the properties of the proposed exact penalty solution scheme in comparison with conventional constraint enforcement techniques. Specifically, focus is placed on the accuracy with which the discrete impenetrability constraints are enforced and the conditioning of the tangent stiffness matrix.

In both numerical examples, the contacting bodies are assumed to be in a state of plane stress. Also, the material is taken to obey the compressible neo-Hookean elastic law with a standard logarithmic representation for the volumetric effect, such that the strain energy W is given by

$$W(I_1, I_3) = \frac{C_1}{2} (I_1 - 3 - \ln I_3) + \frac{C_2}{8} \ln^2 I_3 , \quad (6.1)$$

in terms of the first and third invariants, I_1 and I_3 , of the right Cauchy-Green deformation tensor and the material constants C_1 and C_2 . All simulations are conducted using 4-node isoparametric elements with standard displacement interpolation and 2×2 Gaussian quadrature. Also, all contact integrals are computed using surface-to-surface elements and the multipliers are interpolated using the restriction of the displacement shape functions on $\gamma_c^{(2)}$ (see Supplemental material). This finite element approximation has been shown to be LBB-stable for infinitesimal deformations [28] and pass the contact patch test postulated in Taylor and Papadopoulos [29]. The weighted gap of Equation (3.3) leads to a locally rational polynomial integrand on $\gamma_{c,h}^{(2)}$, and four-point Gauss integration quadrature is employed for its evaluation in each contact cell. The absolute and relative convergence tolerances ε_a and ε_r of the equilibrium scheme (see Box 1) are set to 10^{-8} . In the case of Uzawa's algorithm, a tolerance of 10^{-3} is chosen in the convergence criterion for the top-level multiplier (and penalty) iteration. The conditioning of the stiffness matrix is

monitored and reported by means of the condition number κ_1 associated with the 1-norm.

All computations are carried out using a base step-size, which is recursively halved if convergence cannot be attained. Unless otherwise stated, results reported for the exact penalty method are obtained using the automatic penalty update scheme. Here, the initial value c_0 of the penalty parameter (see Box 1) is reset to unity in the beginning of each step.

6.1 Beam bending over an elastic foundation

Here, a softer elastic beam ($C_1 = 385$ Pa, $C_2 = 577$ Pa) is bent by applied end-moments M such that it gradually comes into contact with a substantially stiffer elastic foundation ($C_1 = 385$ kPa, $C_2 = 577$ kPa), as in Figure 3. This example is inspired by earlier analytical work [21] and has been previously investigated using a finite element method, albeit in a three-dimensional setting, in [24]. It is included here for two reasons. First, the number of active constraints increases gradually as the beam advances toward the foundation. Second, in later stages of deformation, the pressure vanishes in the center of the contact region, thus yielding touching contact (for which strict complementarity is violated) or release. Thus, at least in part, the example serves as a comprehensive test of the active set search and the stiffness interpolation method described in Section 5.

Figure 4 depicts the 250-element mesh in the reference configuration as well as the deformed configurations for $M = 11.7, 21.7$ and 33.3 Nm computed using the exact penalty method. For these deformed configurations, Figure 5 shows the distribution of pressure along the slave surface. As the bending moment increases, the pressure in the center of the contact interface vanishes, as predicted in [21]. Note the non-smoothness in the pressure distribution near the edge of the active contact interface, which is attributed to the sharp drop of the pressure at the edge.

Figure 6 depicts the minimum separation of the two bodies $\min_{i \in \mathcal{S}} g_i(\mathbf{d})$ for inactive contact and the maximum violation of impenetrability $\max_{i \in \mathcal{A}} |g_i(\mathbf{d})|$ admitted by different constraint enforcement techniques for active contact as a function of the applied bending moment. Both the Lagrange multiplier and exact penalty techniques achieve constraint enforcement to computing precision. By contrast, the penalty regularization allows for significant constraint violation. For values of the penalty parameter c larger than 10^7 convergence became increasingly tenuous even with very small step-size.

Figure 7 shows the 1-norm condition number associated with the equilibrium tangent

stiffness matrix as a function of the bending moment for different constraint enforcement techniques. For reference, the condition number of the tangent stiffness matrix $\nabla \mathbf{f}(\mathbf{d})$ without any contact stiffness contributions (yet still subject to the displacement boundary conditions of Figure 3) and evaluated at the equilibrium configurations obtained from the exact penalty calculation is also included in Figure 7. Evidently, the enforcement of the impenetrability constraints strongly influences the conditioning of the tangent stiffness matrix. For the exact penalty method, it is observed that the condition number is affected even in the absence of any active constraints. When contact is activated, the condition number is smaller than the one obtained for the Lagrange multiplier method in many, albeit not all steps.² On the other hand, the classical penalty regularization for the chosen penalty parameters appears to improve on the conditioning of the tangent stiffness relative to the reference tangent stiffness matrix without contact contributions. However, Figure 8 indicates that the penalty methods is not able to capture the release of active contact in the center of the contact surface for large bending moments. This is at variance with results obtained from the Lagrange and exact penalty formulations for which the size of the active contact area and the number of active constraints remain approximately constant (compare with Figure 5).

6.2 Ironing of an elastic foundation

A stiffer non-linearly elastic indenter ($C_1 = 2612$ Pa, $C_2 = 4644$ Pa) is initially pressed onto a softer non-linearly elastic foundation ($C_1 = 261.2$ Pa, $C_2 = 464.4$ Pa) and, subsequently, slides across it. The geometry of the problem is illustrated in Figure 9, see also References [6, 26]. The upper edge of the indenter, after having been displaced downwards by $v = 1.8$ m onto the foundation, is driven rightwards for $u = 10$ m. The resulting “ironing” of the foundation is effected by eight equal vertical displacement steps ($\Delta v = 0.225$ m), followed by fifty equal horizontal displacement steps ($\Delta u = 0.2$ m).

For a finite element mesh comprising 328 elements, the reference configuration as well as selected deformed configurations computed by the exact penalty method and corresponding to displacement steps 8 (end of pressing phase), 32, and 50 are depicted in Figure 10.

²Strictly, the impact of ill-conditioning on part of the tangent stiffness matrix in the Lagrange multiplier formulation can be mitigated by employing dedicated preconditioners and (iterative) linear algebraic system solvers or by statically condensing the multipliers. In the examples considered here and in Section 6.2, static condensation of the multipliers from the tangent stiffness matrix entails a reduction in κ_1 of two to three orders of magnitude.

Likewise, Figure 11 depicts the contact pressure field along the slave surface for a finer mesh of 5800 elements (not shown) at steps 8, 32 and 50. Owing to the resolution of the mesh, the boundary effects which were pronounced in the previous example, see Figure 5, have almost disappeared.

In analogy to the example in Section 6.1, Figure 12 illustrates the accuracy to within which the active contact constraints are enforced in the course of the analysis. Again, the Lagrange multiplier and exact penalty approaches achieve perfect enforcement of the impenetrability constraints up to computing precision, while the penalty method entails a constraint violation which depends on the choice of penalty parameter c .

Figure 13 depicts the results for the evolution of the 1-norm condition number κ_1 associated with the tangent stiffness matrices evaluated at equilibrium for different constraint enforcement techniques. As in Section 6.1, the condition number of the tangent stiffness matrix without the contact stiffness terms and evaluated at the exact penalty equilibrium configurations is included for comparison. The Lagrange multiplier method leads to tangent stiffness matrices whose condition exceeds that of the stiffness without contact contributions by several orders of magnitude. For the penalty method, the conditioning of the tangent stiffness matrices in the present example remains comparable to that of the tangent stiffness without contact terms for low penalties ($c = 10^4$ and $c = 10^5$), but starts increasing notably for $c = 10^6$. While the exact penalty method leads to converged solutions for any initial penalty value $c \geq 1$, it was observed that the conditioning of the equilibrium stiffness matrix depends quite strongly on the choice of c . For the present example, it appears that there exists an optimal penalty parameter, roughly on the order of 10^4 to 10^6 , for which the equilibrium stiffness matrix is well-conditioned and the condition number κ_1 exceeds that of the reference stiffness matrix without contact contributions by only one order of magnitude. On the other hand, choosing c outside this range causes significant deterioration in the conditioning of the equilibrium stiffness matrix ; for example, for $c = 1$, the condition number attains values similar to those of the direct Lagrange multiplier method. The automatic penalty update scheme terminates in penalty values that result in a wide range of condition numbers for the tangent stiffness matrix, ranging from nearly optimal to as poor as in the case of the Lagrange multiplier method. Upon closer examination, it is concluded that the active and inactive contact stiffnesses in Equations (5.12) and (5.14) can adversely affect the conditioning of the cumulative stiffness matrix in Equation (5.16), such that there exists an optimal penalty value for which the effects on the conditioning of either contribution are balanced.

Figure 14 shows the change in the terminal penalty value reached by the automatic penalty update scheme over the course of the analysis. The terminal penalty parameter varies significantly from one step to the next and, in relation to Figure 13, it is inferred that small values of this penalty parameter do not necessarily imply a better conditioned tangent stiffness matrix.

Figure 15 illustrates the change in the number of active constraints for different enforcement methods. Again, the Lagrange multiplier and the exact penalty methods show identical responses in this measure. Also considering Figure 13(a), it is observed that the conditioning of the tangent stiffness matrix obtained from the Lagrange multiplier method increases during the initial displacement steps in which the indenter gradually comes into contact with the foundation and the number of active contact constraints increases, while the conditioning of the penalty tangent stiffness is significantly less affected. By contrast, the condition number for the exact penalty method with automatic penalty adjustment in Figure 13(b) varies from step to step and seems to be affected by the fine balance between the active and inactive contact stiffnesses mediated by the penalty parameter.

Table 2 documents the convergence behavior of the exact penalty equilibrium scheme for a representative displacement step. Note that asymptotic quadratic convergence is recovered as soon as the correct active set is determined.

Iteration	Penalty	Maximum constraint violation	Residual norm	Relative displacement change
0	1	2.0×10^{-14}	1.172×10^3	—
1*	2×10^1	2.7×10^{-2}	9.982×10^4	1.000
2*	1×10^2	1.8×10^{-2}	1.766×10^3	1.028×10^{-1}
3*	3×10^2	7.5×10^{-5}	3.995×10^2	2.694×10^{-2}
4	3×10^2	8.0×10^{-7}	2.365	2.522×10^{-3}
5	3×10^2	5.0×10^{-10}	6.679×10^{-4}	2.626×10^{-5}
6	3×10^2	3.0×10^{-13}	5.096×10^{-7}	1.637×10^{-8}
7	3×10^2	2.6×10^{-16}	2.886×10^{-9}	4.906×10^{-12}

Table 2: *Ironing of an elastic foundation: Convergence of the exact penalty iterative solution (automatic penalty update; see Box 1) for displacement step 50 from $(u, v) = (8.2, 1.8)$ to $(u, v) = (8.4, 1.8)$ (the superscript * in entries of the first column indicates that the active set changed in the current iteration)*

Finally, Table 3 summarizes certain simulation statistics for runs using different con-

straint enforcement methods. The exact penalty method with automatic penalty update requires the same number of steps as the Lagrange multiplier method, but one more equilibrium iteration per step on average. However, if the exact penalty method is applied with a fixed penalty parameter, then the number of required steps significantly increases for larger values of the penalty parameter, while the number of equilibrium iterations in each step remains approximately constant. This observation also carries over to the penalty method and to Uzawa's scheme. As the penalty parameter increases, the number of step-size refinements grows significantly, limiting, in our present case, the maximum practical value of the penalty more than the conditioning of the tangent stiffness matrix. Returning to the exact penalty method, it appears that the automatic penalty update scheme entails, in each equilibrium iteration, penalty sequences that promote robustness and lead to a rapid convergence. For the simulations reported here, the total run-time per equilibrium iteration varies insignificantly between the exact penalty and the classical methods, as it is dominated by the formation of the finite element arrays and the solution of the global algebraic system.

In practice, the condition $t_c(\mathbf{d}) > 0$ is replaced in both examples by $t_c(\mathbf{d}) > \varepsilon_a$, where ε is the absolute convergence tolerance on $\|\mathbf{r}_c(\mathbf{d})\|$ (see Box 1). The choice of ε_a appears to affect the terminal penalty value deduced by the automatic penalty update scheme. Likewise, the terminal penalty value is influenced by the choice of scaling matrices in Equation (5.18), particularly, by the approximation for $\tilde{\mathbf{K}}_c^{-1}$, as these control the delicate balance of the two terms in the test function t_c in Equation (5.18).

In the two examples presented in this section, convergence to an equilibrium solution for which the constraints are exactly enforced is obtained for any value of the penalty parameter $c \geq 1$, except, of course, for very large ones which may lead to ill-conditioning. This underscores the fact that the automatic penalty adjustment scheme does not necessarily terminate on the smallest penalty value required to attain the exact solution to the constrained equilibrium problem. Rather, the scheme acts as a safeguard, in the sense that if the equilibrium iterations appear to converge to a solution that violates the constraints, then the test function check forces an increase of the penalty parameter so that the exact solution to the constrained equilibrium problem is guaranteed. However, since the test function check is conservative, the terminal penalty reached by the scheme at a given equilibrium step is not, in general, a good estimate for the threshold penalty above which the exact penalty formulation is valid. Indeed, computational experiments on the two numerical examples suggest that this terminal penalty value is one for which the tangent stiffness

Method	Penalty c	Number of steps	Number of iterations per step
Exact penalty	Automatic	58	6.78
Exact penalty	1	58	7.00
Exact penalty	10^2	58	6.83
Exact penalty	10^4	58	6.38
Exact penalty	10^6	145	8.63
Exact penalty	10^8	552	8.77
Lagrange multiplier	1	58	5.93
Penalty	10^4	58	6.17
Penalty	10^5	111	6.05
Penalty	10^6	341	5.48
Uzawa's algorithm	10^4	58	4.00×3.96
Uzawa's algorithm	10^5	132	3.67×3.98

Table 3: *Ironing of an elastic foundation: Comparison of the average number of displacement steps and the average number of iterations per step for different constraint enforcement methods (in the Lagrange multiplier method, c is the complementarity parameter of Equation (5.1), which may slightly influence the rate of convergence [26]; also, the final column for Uzawa's algorithm shows the average number of top-level iterations per step times the average number of equilibrium iterations per top-level iteration; in addition, no line search is used to improve convergence of the equilibrium scheme)*

matrices are typically, albeit not always, well-conditioned and such that convergence is rapidly attained.

7 Conclusions

This article documents the use of a novel exact penalty method adapted from the work of Glad and Polak [11] for the solution of the frictionless, non-linearly elastic two-body contact problem in conjunction with a finite element-based approximation. The exact penalty method was shown to have definite advantages over both the Lagrange multiplier method (no additional equations) and its various regularizations, including those based on penalty (no inexact enforcement of constraints) and augmented Lagrangian approaches (no need for multiple cycles of iterations for exact enforcement of constraints). Although

the exact penalty function includes a penalty parameter, there exists a finite lower bound above which the unconstrained minimization problem defined by the exact penalty function is valid and leads to exact solutions. Moreover, the penalty can be updated automatically, thus circumventing the need for a problem-specific selection.

While the exact penalty method does not involve a top-level active-set search, the distinction between active and inactive constraints is reflected as a discontinuity in the contact stiffness. A critical idea in the algorithm proposed by Glad and Polak [11] is to bridge this discontinuity by combining the active and the inactive contact stiffnesses by a suitable interpolation. Another key feature of the implementation is that only first and second derivatives of the contact constraints are computed exactly. While, strictly speaking, consistent linearization also requires third derivatives, no deterioration is detected in the ultimate rate of convergence by omitting these terms. Furthermore, the numerical tests demonstrated that the conditioning of the tangent stiffness matrix is strongly influenced by the current penalty value which seems to control the balance of the active and inactive contact stiffness contributions. In particular, a small penalty does not imply a well-conditioned tangent stiffness matrix in general and, depending on the length scales and elastic moduli inherent in a specific equilibrium problem, large penalties may be required for a well-conditioned tangent stiffness matrix. Irrespective of conditioning, it is observed that the penalty sequences automatically constructed promote a robust and rapid convergence, reducing the need for step-size refinement. In addition, the exact penalty method performs very well in terms of the maximum step-size allowed and the average number of equilibrium iterations required for each step when compared to the standard constraint enforcement methods.

Acknowledgments

One of us (FS) wishes to acknowledge the support of the German-American Fulbright Association and the Institute of International Education, which supported the initial phase of this research at the University of California, Berkeley. Further support to FS was provided by a research project at Bombardier Transportation GmbH, Hennigsdorf. Both authors wish to thank Professor Elijah Polak of the University of California, Berkeley, for valuable discussions on optimization.

References

- [1] P. Alart and A. Curnier. “A mixed formulation for frictional contact problems prone to Newton like solution methods”. *Computer Methods in Applied Mechanics and Engineering* 92.3 (1991), pp. 353–375.
- [2] C. Annavarapu, M. Hautefeuille, and J. E. Dolbow. “A Nitsche stabilized finite element method for frictional sliding on embedded interfaces. Part I: Single interface”. *Computer Methods in Applied Mechanics and Engineering* 268 (2014), pp. 417–436.
- [3] J. Baiges, R. Codina, F. Henke, S. Shahmiri, and W. A. Wall. “A symmetric method for weakly imposing Dirichlet boundary conditions in embedded finite element meshes”. *International Journal for Numerical Methods in Engineering* 90.5 (2012), pp. 636–658.
- [4] D. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1996.
- [5] R. Courant. “Variational methods for the solution of problems of equilibrium and vibration”. *Bulletin of the American Mathematical Society* 49 (1943), pp. 1–23.
- [6] K. A. Fischer and P. Wriggers. “Mortar based frictional contact formulation for higher order interpolations using the moving friction cone”. *Computer Methods in Applied Mechanics and Engineering* 195.37 (2006), pp. 5020–5036.
- [7] R. Fletcher. “A class of methods for nonlinear programming III: Rates of convergence”. *Numerical Methods for Nonlinear Optimization*. Ed. by F. A. Lootsma. New York: Academic Press, 1972, pp. 371–382.
- [8] R. Fletcher. “A class of methods for nonlinear programming with termination and convergence properties”. *Integer and Nonlinear Programming*. Ed. by J. Abadie. Amsterdam: North-Holland Publishing Company, 1970. Chap. 6, pp. 157–175.
- [9] R. Fletcher. “An exact penalty function for nonlinear programming with inequalities”. *Mathematical Programming* 5.1 (1973), pp. 129–150.
- [10] R. Fletcher and S. A. Lill. “A class of methods for non-linear programming II: Computational experience”. *Nonlinear Programming*. Ed. by J. B. Rosen, O. L. Mangasarian, and K. Ritter. New York: Academic Press, 1971, pp. 67–92.
- [11] T. Glad and E. Polak. “A multiplier method with automatic limitation of penalty growth”. *Mathematical Programming* 17.1 (1979), pp. 140–155.
- [12] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and Operator-Splitting Methods in Non-linear Mechanics*. Vol. 8. Philadelphia: SIAM, 1989.
- [13] S. Hartmann, S. Brunssen, E. Ramm, and B. Wohlmuth. “Unilateral non-linear dynamic contact of thin-walled structures using a primal-dual active set strategy”. *International Journal for Numerical Methods in Engineering* 70.8 (2007), pp. 883–912.
- [14] M. R. Hestenes. “Multiplier and gradient methods”. *Journal of Optimization Theory and Applications* 4.5 (1969), pp. 303–320.

- [15] R. E. Jones and P. Papadopoulos. “A novel three-dimensional contact finite element based on smooth pressure interpolations”. *International Journal for Numerical Methods in Engineering* 51.7 (2001), pp. 791–811.
- [16] N. Kikuchi and J. T. Oden. *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*. Philadelphia: SIAM, 1988.
- [17] S. Kollmannsberger, A. Özcan, J. Baiges, M. Ruess, E. Rank, and A. Reali. “Parameter-free, weak imposition of Dirichlet boundary conditions and coupling of trimmed and non-conforming patches”. *International Journal for Numerical Methods in Engineering* 101.9 (2015), pp. 670–699.
- [18] J. A. Landers and R. L. Taylor. *An Augmented Lagrangian Formulation for the Finite Element Solution of Contact Problems*. SESM Report 85/09. University of California at Berkeley, 1985.
- [19] R. Mlika, Y. Renard, and F. Chouly. “An unbiased Nitsche’s formulation of large deformation frictional contact and self-contact”. *Computer Methods in Applied Mechanics and Engineering* 325 (2017), pp. 265–288.
- [20] H. Mukai and E. Polak. “A quadratically convergent primal-dual algorithm with global convergence properties for solving optimization problems with equality constraints”. *Mathematical Programming* 9.1 (1975), pp. 336–349.
- [21] P. M. Naghdi and M. B. Rubin. “On the Significance of Normal Cross-sectional Extension in Beam Theory with Application to Contact Problems”. *International Journal of Solids and Structures* 25.3 (1989), pp. 249–265.
- [22] J. Nitsche. “Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind”. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg* 36.1 (1971), pp. 9–15.
- [23] J. T. Oden. “Exterior Penalty Methods for Contact Problems in Elasticity”. *Europe-US Workshop: Nonlinear Finite Element Analysis in Structural Mechanics*. Ed. by E. S. K.-J. Bathe and W. Wunderlich. Berlin: Springer-Verlag, 1980.
- [24] P. Papadopoulos and R. L. Taylor. “A Simple Algorithm for Three-dimensional Finite Element Analysis of Contact Problems”. *Computers & Structures* 46 (1993), pp. 1107–1118.
- [25] E. Polak. “On the global stabilization of locally convergent algorithms”. *Automatica* 12.4 (1976), pp. 337–342.
- [26] A. Popp, M. W. Gee, and W. A. Wall. “A finite deformation mortar contact formulation using a primal-dual active set strategy”. *International Journal for Numerical Methods in Engineering* 79.11 (2009), pp. 1354–1391.
- [27] J. C. Simo and T. A. Laursen. “An augmented Lagrangian treatment of contact problems involving friction”. *Computers & Structures* 42.1 (1992), pp. 97–116.
- [28] J. M. Solberg and P. Papadopoulos. “An analysis of dual formulations for the finite element solution of two-body contact problems”. *Computer Methods in Applied Mechanics and Engineering* 194.25-26 (2005), pp. 2734–2780.

- [29] R. L. Taylor and P. Papadopoulos. “On a Patch Test for Contact Problems in Two Dimensions”. *Nonlinear Computational Mechanics*. Ed. by P. Wriggers and W. Wagner. Berlin: Springer-Verlag, 1991, pp. 690–702.
- [30] P. Wriggers and G. Zavarise. “A formulation for frictionless contact problems using a weak form introduced by Nitsche”. *Computational Mechanics* 41.3 (2008), pp. 407–420.

Appendices

A Equality constraints and Dirichlet boundary conditions

In this section, an extension of the exact penalty method detailed in Sections 4 and 5 is presented for solving an equilibrium problem subject to both kinematic inequality and equality constraints. This forms the basis for an examination of how to treat nodes which are simultaneously subject to Dirichlet and contact boundary conditions. In a direct extension of Equation (3.2), consider the following discrete Lagrange multiplier formulation

$$\begin{aligned} \mathbf{f}(\mathbf{d}) + \nabla \mathbf{h}(\mathbf{d})^T \boldsymbol{\mu} + \nabla^T \mathbf{g}(\mathbf{d}) \boldsymbol{\lambda} &= \mathbf{0} \\ \mathbf{h}(\mathbf{d}) &= \mathbf{0} \ , \\ \mathbf{g}(\mathbf{d}) &\geq \mathbf{0} \end{aligned} \tag{A.1}$$

where, as earlier, $\boldsymbol{\lambda} \leq \mathbf{0}$, and Equation (3.4) holds. Also, $\mathbf{h}(\mathbf{d}) = [h_i(\mathbf{d})]_{i \in \mathcal{E}}$, where $\mathcal{E}$ is the index set for the equality constraints, and $\boldsymbol{\mu}$ is the corresponding Lagrange multiplier vector. Substituting $\mathbf{h}$ for $\mathbf{g}$ in Equations (4.2) and (4.5) and combining the results with Equation (5.8) leads to the exact penalty equilibrium equations

$$\mathbf{r}_c(\mathbf{d}) = \mathbf{f}(\mathbf{d}) + \mathbf{f}_c(\mathbf{d}) + \nabla \psi(\mathbf{d})^T \mathbf{h}(\mathbf{d}) + \nabla \mathbf{h}(\mathbf{d})^T (\psi(\mathbf{d}) + c\mathbf{h}(\mathbf{d})) = \mathbf{0} \ . \tag{A.2}$$

Following Glad and Polak [11], the Lagrange multipliers $\boldsymbol{\mu}(\mathbf{d})$ and $\boldsymbol{\lambda}(\mathbf{d})$ can be obtained from Equations (4.5) and (5.3), where ψ and ϕ are solutions of the least-squares problem (compare with Equations (4.7) and (5.9))

$$(\psi(\mathbf{d}), \phi(\mathbf{d})) = \arg \min_{\mathbf{y}, \mathbf{z}} \left\{ \|\mathbf{f}(\mathbf{d}) + \nabla^T \mathbf{h}(\mathbf{d}) \mathbf{y} + \nabla^T \mathbf{g}(\mathbf{d}) \mathbf{z}\|^2 + \mathbf{z}^T \mathbf{G}(\mathbf{d}) \mathbf{z} \right\} \ , \tag{A.3}$$

whence, in line with Equation (5.10),

$$\begin{pmatrix} \nabla \mathbf{g} \nabla^T \mathbf{g} + \mathbf{G} & \nabla \mathbf{g} \nabla^T \mathbf{h} \\ \nabla \mathbf{h} \nabla^T \mathbf{g} & \nabla \mathbf{h} \nabla^T \mathbf{h} \end{pmatrix} \begin{pmatrix} \phi \\ \psi \end{pmatrix} = - \begin{pmatrix} \nabla \mathbf{g} \\ \nabla \mathbf{h} \end{pmatrix} \mathbf{f} \ . \tag{A.4}$$

In practice, Equation (A.2) may be solved using a Newton-Raphson method, such that

$$(\nabla \mathbf{f} + \mathbf{J}_c + \mathbf{K}_c) \Delta \mathbf{d} = -\mathbf{r}_c \ , \tag{A.5}$$

(also see Box 1), where $\mathbf{K}_c$ is given by Equation (5.16) and $\mathbf{J}_c$ by Equation (4.12) (with $\mathbf{h}$ substituted for $\mathbf{g}$). As in Sections 4 and 5, the terms involving third derivatives of $\mathbf{h}$ and $\mathbf{g}$

may be omitted. In addition, the test function in Equation (5.18) is slightly amended by replacing $\mathbf{a}_c$ by the concatenation of $\mathbf{a}_c$ and $\mathbf{h}$.

It is now possible to show that the standard manner of imposing Dirichlet boundary conditions on an equilibrium step remains valid also for an exact penalty formulation if either the contact and Dirichlet nodes are distinct or if, for situations in which this cannot be guaranteed, the formulae for estimating the Lagrange multipliers $\phi(\mathbf{d})$ and their gradients $\nabla\phi(\mathbf{d})$ are slightly amended. Dirichlet and contact nodes may coincide in practice, for instance, if the Dirichlet and contact boundaries are adjacent or if one of the contacting bodies features a spatially fixed contact surface. In order to investigate the inclusion of Dirichlet boundary conditions in an exact penalty approach, let the displacement vector $\mathbf{d}$ be decomposed, after some potential reordering, according to

$$\mathbf{d} = \begin{bmatrix} \bar{\mathbf{d}} \\ \tilde{\mathbf{d}} \end{bmatrix}, \quad (\text{A.6})$$

where $\bar{\mathbf{d}}$ is the vector of displacement degrees of freedom prescribed by Dirichlet boundary conditions, while the vector $\tilde{\mathbf{d}}$ contains the remaining (unspecified) displacement degrees of freedom. Therefore, the equality constraint associated with the Dirichlet boundary conditions takes the form

$$\mathbf{h}(\mathbf{d}) = \bar{\mathbf{d}}_n - \begin{bmatrix} \bar{\mathbf{I}} & \bar{\mathbf{0}} \end{bmatrix} \mathbf{d} = \mathbf{0}, \quad (\text{A.7})$$

where $\bar{\mathbf{d}}_n$ is the prescribed value of $\bar{\mathbf{d}}$ at step n , while $\bar{\mathbf{I}}$ is a square identity matrix of the dimension of $\bar{\mathbf{d}}$.

Upon introducing the constraint condition (A.7) in Equation (A.4), one immediately finds that the equality constraint function ψ is given by

$$\psi = \bar{\nabla}^T \mathbf{g} \phi + \bar{\mathbf{f}}, \quad (\text{A.8})$$

hence its gradient is

$$\nabla\psi = \bar{\nabla}^T \mathbf{g} \nabla\phi + \sum_{i \in S} \phi_i \nabla \bar{\nabla}^T g_i + \nabla \bar{\mathbf{f}}, \quad (\text{A.9})$$

where $\bar{\nabla}$ denotes differentiation with respect to $\bar{\mathbf{d}}$ and $\bar{\mathbf{f}}$ is the force vector with components corresponding to the degrees of freedom in $\bar{\mathbf{d}}$. It follows from the above and from Equation (A.4) that the inequality constraint function ϕ satisfies the equation

$$\left(\tilde{\nabla} \mathbf{g} \tilde{\nabla}^T \mathbf{g} + \mathbf{G} \right) \phi = -\tilde{\nabla} \mathbf{g} \tilde{\mathbf{f}} \quad (\text{A.10})$$

and, likewise, its gradient satisfies

$$\begin{aligned} \left(\tilde{\nabla} \mathbf{g} \tilde{\nabla}^T \mathbf{g} + \mathbf{G} \right) \nabla \phi &= - \tilde{\nabla} \mathbf{g} \left[\nabla \tilde{\mathbf{f}} + \sum_{i \in \mathcal{S}} \phi_i \tilde{\nabla} \nabla g_i \right] - 2\phi [\mathbf{g}^T \nabla \mathbf{g}] \\ &\quad - \begin{pmatrix} \left(\tilde{\mathbf{f}}^T + \phi^T \tilde{\nabla} \mathbf{g} \right) \tilde{\nabla} \nabla g_{i_1} \\ \vdots \\ \left(\tilde{\mathbf{f}}^T + \phi^T \tilde{\nabla} \mathbf{g} \right) \tilde{\nabla} \nabla g_{i_s} \end{pmatrix}, \end{aligned} \quad (\text{A.11})$$

where $\tilde{\nabla}$ denotes differentiation with respect to $\tilde{\mathbf{d}}$, $\tilde{\mathbf{f}}$ is the force vector with components corresponding to the degrees of freedom in $\tilde{\mathbf{d}}$, and $\mathcal{S} = \{i_1, \dots, i_s\}$. Upon substitution of Equations (A.7) to (A.9) into Equation (A.2), the discrete force residual can be expressed as

$$\mathbf{r}_c = \mathbf{f} + \mathbf{f}_c + \left(\nabla^T \phi \bar{\nabla} \mathbf{g} + \sum_{i \in \mathcal{S}} \phi_i \nabla^T \bar{\nabla} g_i + \nabla^T \tilde{\mathbf{f}} \right) \mathbf{h} - \begin{pmatrix} \bar{\nabla}^T \mathbf{g} \phi + \bar{\mathbf{f}} + c\mathbf{h} \\ \mathbf{0} \end{pmatrix}. \quad (\text{A.12})$$

Introducing Equations (A.6) to (A.9) and (A.12) into Equation (A.5) and explicitly enforcing the condition $\Delta \bar{\mathbf{d}} = \mathbf{h}(\mathbf{d}) = \bar{\mathbf{d}}_n - \bar{\mathbf{d}}$, leads to a linear system of the form

$$\begin{pmatrix} \bar{\mathbf{I}} & \mathbf{0} \\ \mathbf{0} & \tilde{\nabla} \tilde{\mathbf{f}} + \tilde{\nabla} \tilde{\mathbf{f}}_c \end{pmatrix} \begin{pmatrix} \Delta \bar{\mathbf{d}} \\ \Delta \tilde{\mathbf{d}} \end{pmatrix} = - \begin{pmatrix} -\mathbf{h} \\ \tilde{\mathbf{f}} + \tilde{\mathbf{f}}_c + \left(\bar{\nabla} \tilde{\mathbf{f}} + \bar{\nabla} \tilde{\mathbf{f}}_c \right) \mathbf{h} \end{pmatrix}, \quad (\text{A.13})$$

where $\Delta \bar{\mathbf{d}}$ and $\Delta \tilde{\mathbf{d}}$ denote the Newton-Raphson increments. Equations (A.12) and (A.13) indicate that the Dirichlet boundary conditions may be imposed on an equilibrium iteration in the standard way if Equations (5.10) and (5.11) for evaluating the Lagrange multipliers associated with the contact inequality constraints are replaced by Equations (A.10) and (A.11). Concomitantly, the test function t_c in Equation (5.18) for the automatic penalty adjustment scheme is evaluated based on the right-hand side of Equation (A.13) (as $\mathbf{r}_c$) and the diagonal of the left-hand side matrix (as $\tilde{\mathbf{K}}_c$). For the case where contact nodes are distinct from Dirichlet nodes, $\bar{\nabla} \mathbf{g} = \mathbf{0}$, whence Equations (A.10) and (A.11) coincide with Equations (5.10) and (5.11). Equation (A.13) remains valid regardless of whether contact nodes are constrained by Dirichlet boundary conditions.

Input: $\alpha > 1$, $\varepsilon_a > 0$, $\varepsilon_r > 0$, $c_0 > 0$, $\mathbf{d}^0$, $\mathbf{f}$, $\mathbf{g}$

Output: $\mathbf{d}^*$, λ^*

1. Set $k = 0$
2. Evaluate $\mathbf{g}(\mathbf{d}^k)$, $\nabla \mathbf{g}(\mathbf{d}^k)$ and $\nabla(\nabla \mathbf{g}(\mathbf{d}^k))$
3. Evaluate $\mathbf{f}(\mathbf{d}^k)$ and $\nabla \mathbf{f}(\mathbf{d}^k)$
4. Evaluate $\phi(\mathbf{d}^k)$ and $\nabla \phi(\mathbf{d}^k)$ using Equations (5.10) and (5.11)
5. Compute discrete force residual and penalty adjustment
 - 5.1 Evaluate $\mathbf{r}_{c_k}(\mathbf{d}^k)$ from Equation (5.8)
 - 5.2 If $\|\mathbf{r}_{c_k}(\mathbf{d}^k)\| \leq \varepsilon_a$, set $\mathbf{d}^* = \mathbf{d}^k$, $\lambda^* = \phi(\mathbf{d}^*)$ and terminate
 - 5.3 Evaluate $t_{c_k}(\mathbf{d}^k)$ from Equation (5.18)
 - 5.4 If $t_{c_k}(\mathbf{d}^k) > 0$, set $c_k \leftarrow \alpha c_k$ and return to Step 4.1
6. Compute $\mathbf{K}_{c_k}(\mathbf{d}^k)$ using Equations (5.16) and (5.17)
7. Execute an inexact Newton step

$$\begin{aligned} \left[\nabla \mathbf{f}(\mathbf{d}^k) + \mathbf{K}_{c_k}(\mathbf{d}^k) \right] \Delta \mathbf{d}^{k+1} &= -\mathbf{r}_{c_k}(\mathbf{d}^k) \\ \mathbf{d}^{k+1} &= \mathbf{d}^k + \Delta \mathbf{d}^{k+1} \end{aligned}$$

8. If $\|\Delta \mathbf{d}^{k+1}\| / \|\mathbf{d}^{k+1}\| \leq \varepsilon_r$, set $\mathbf{d}^* = \mathbf{d}^{k+1}$, $\lambda^* = \phi(\mathbf{d}^*)$ and terminate; else set $c_{k+1} = c_k$, set $k \leftarrow k + 1$ and return to Step 2

Box 1: Summary of the algorithmic steps for solving the general discrete contact problem using the exact penalty constraint enforcement technique due to Glad and Polak [11] (here, ε_a and ε_r represent absolute and relative convergence tolerances, respectively)

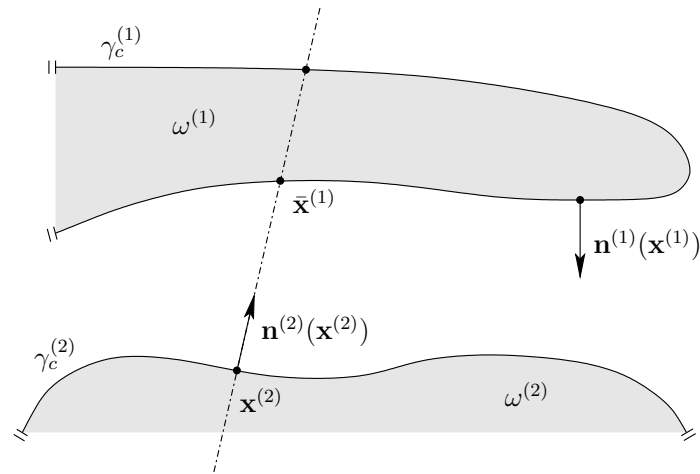


Figure 1: Schematic illustration of the definition of the gap function $g(\mathbf{x}^{(2)})$ on $\gamma_c^{(2)}$ in Equation (2.10)

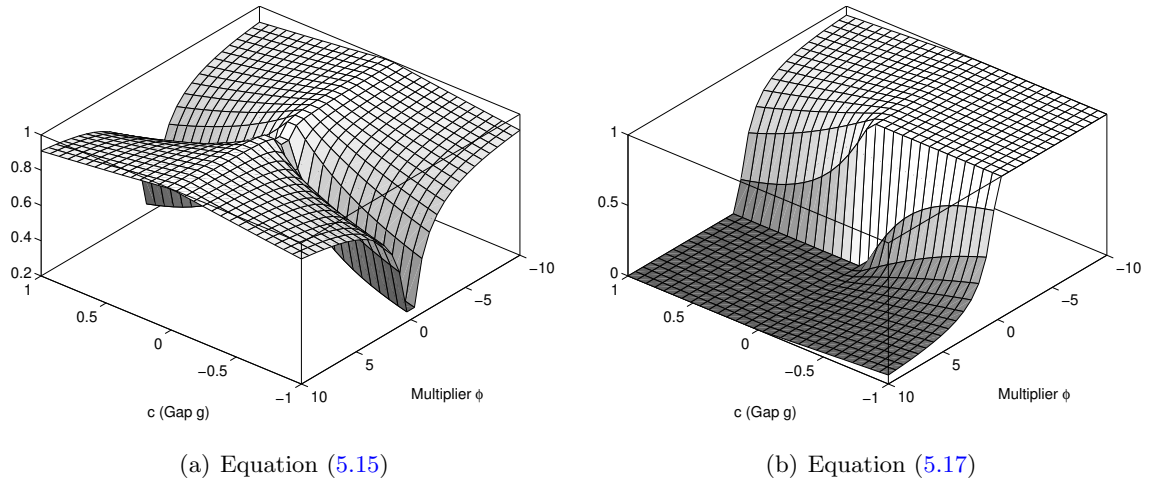


Figure 2: Comparison of two choices for the weight function b evaluated in terms of the gap g and multiplier ϕ : (a) the original function due to Glad and Polak [11], (b) the alternative function adopted in this work

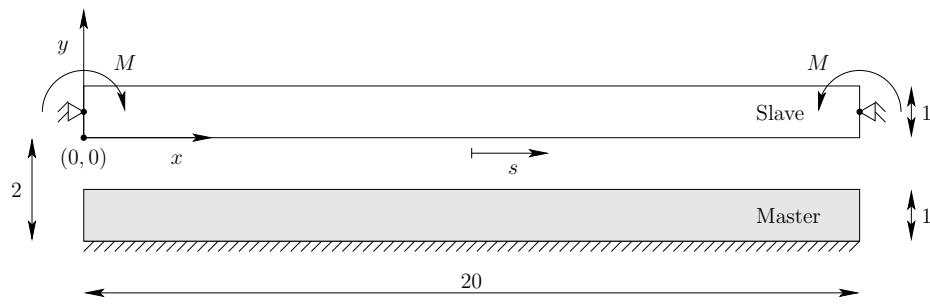


Figure 3: Beam bending over an elastic foundation: Geometry (all lengths are in meters)

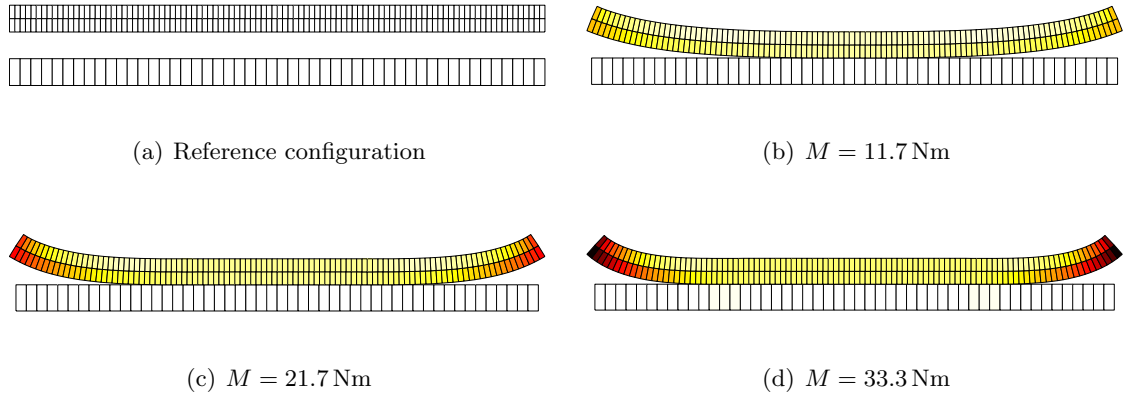


Figure 4: Beam bending over an elastic foundation: Reference configuration and configurations for $M = 11.7, 21.7, 33.3 \text{ Nm}$. The color scheme indicates the von Mises stress ranging from 0 (white) to a maximum value of 121.2 Pa (dark red)

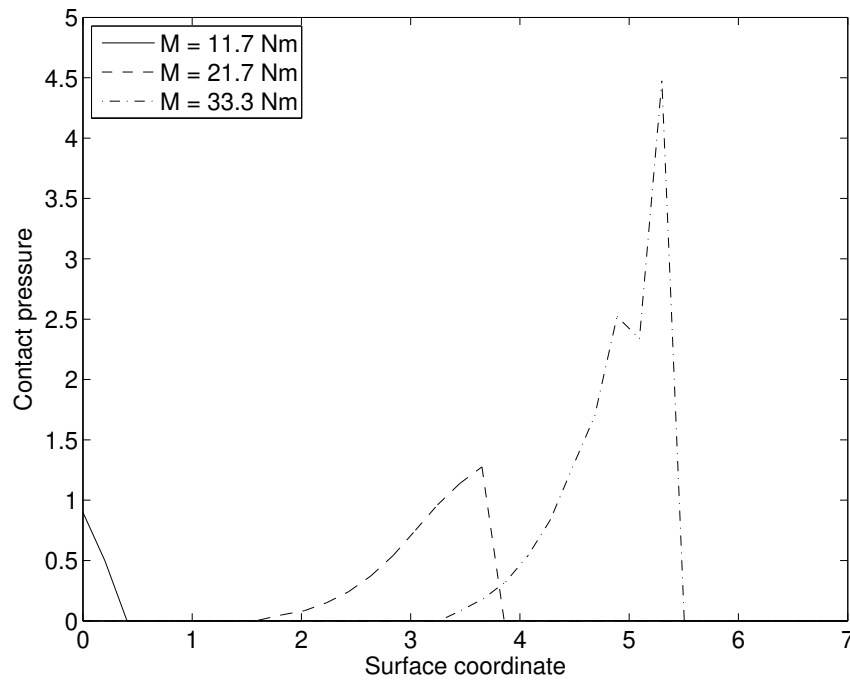


Figure 5: Beam bending over an elastic foundation: Contact pressure field on slave surface for $M = 11.7, 21.7$ and 33.3 Nm as a function of the arc-length s in the current configuration ($s = 0$ on the symmetry line)

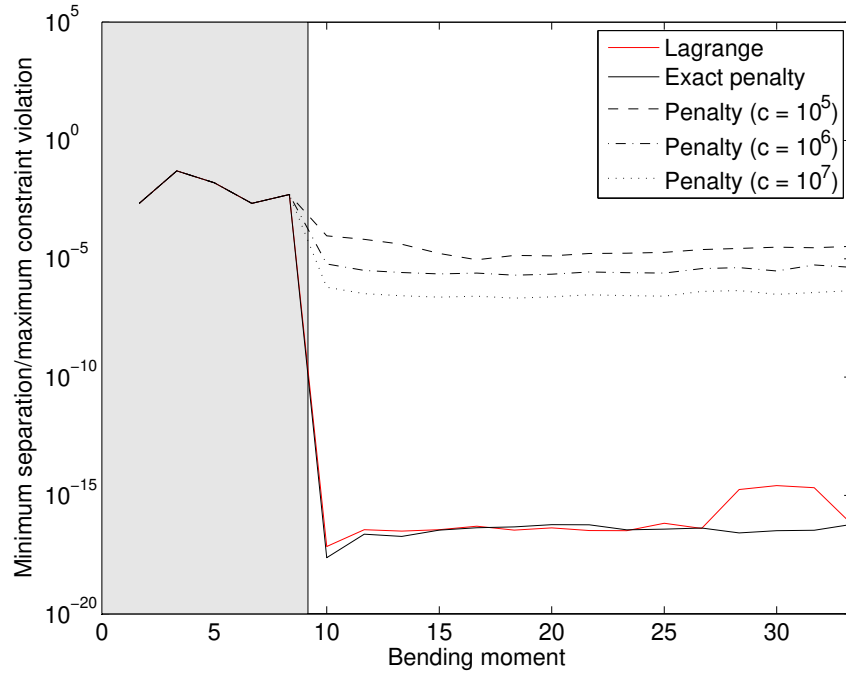


Figure 6: Beam bending over an elastic foundation: Minimum separation (all constraints inactive) or maximum constraint violation (any constraint active) at equilibrium as a function of the bending moment M (the shaded area corresponds to instants for which contact has not yet taken place)

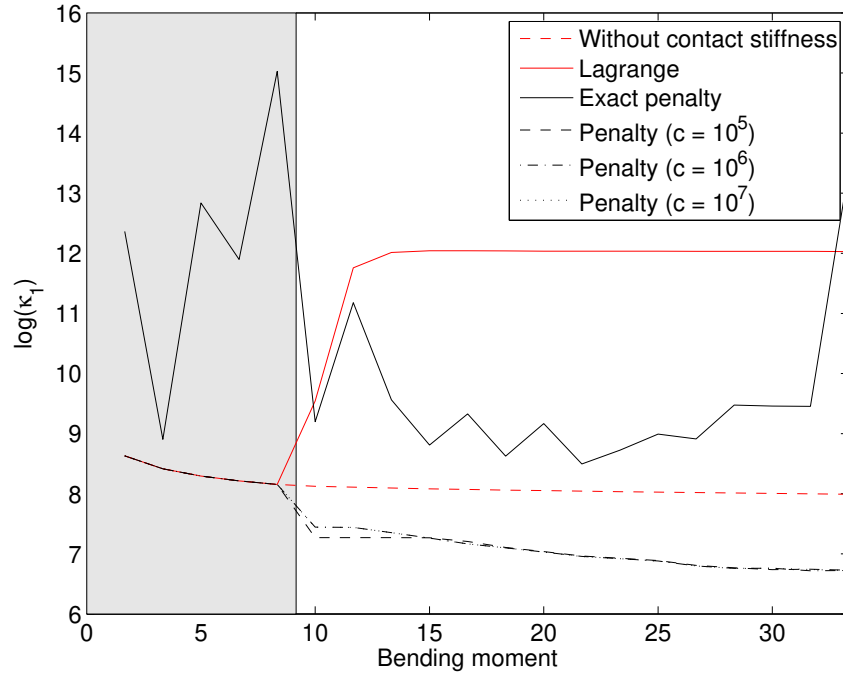


Figure 7: Beam bending over an elastic foundation: Evolution of the 1-norm condition number κ_1 of the tangent stiffness matrices at equilibrium for different constraint enforcement techniques as a function of the applied moment M (the shaded area corresponds to instants for which contact has not yet taken place; the curves for the penalty cases $c = 10^6$ and $c = 10^7$ are practically coincident)

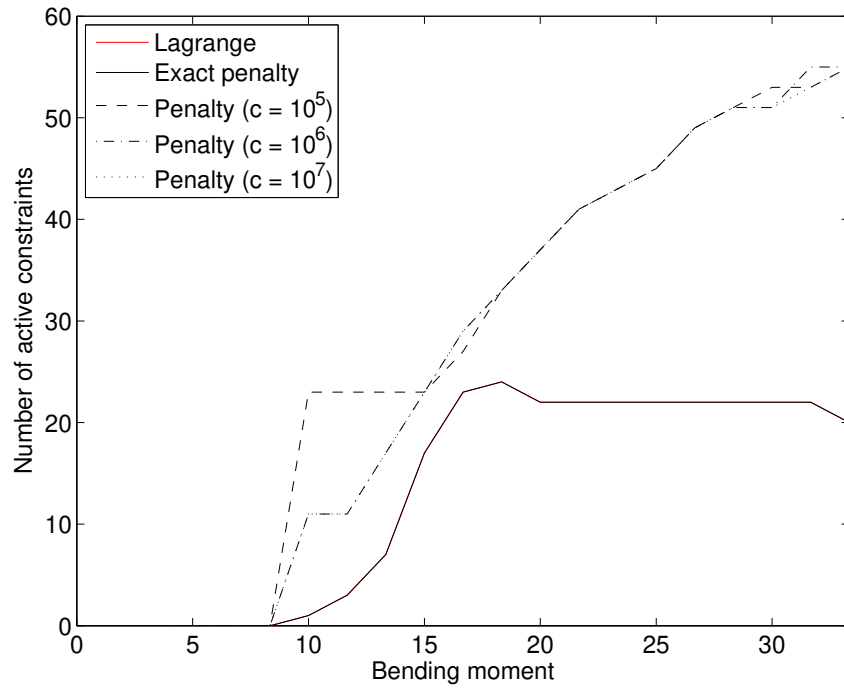


Figure 8: Beam bending over an elastic foundation: Number of active constraints as a function of the applied bending moment M (the curves for the exact penalty and Lagrange multiplier methods are coincident)

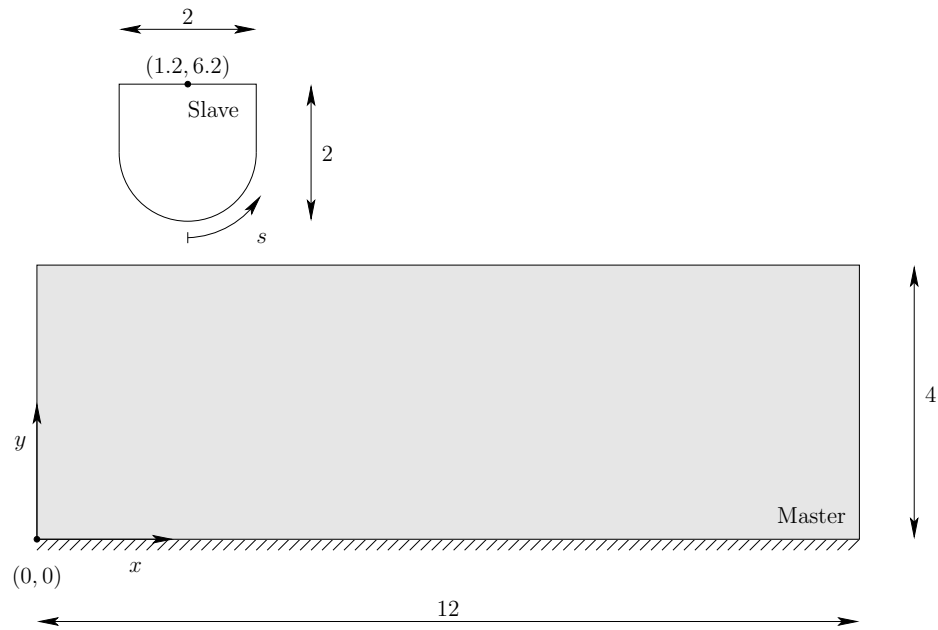


Figure 9: Ironing of an elastic foundation: Geometry (all coordinates and lengths are in meters)

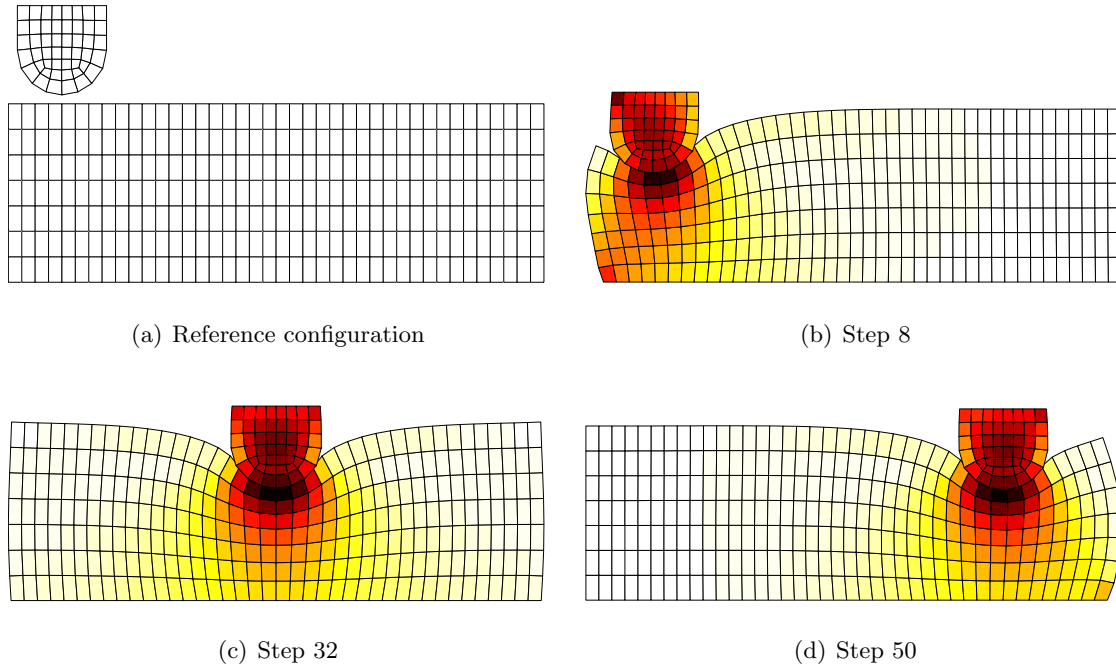


Figure 10: Ironing of an elastic foundation: Reference configuration and current configurations with von Mises stress distribution at steps 8, 32 and 50 (the mesh comprises 328 elements and the maximum von Mises stress is approximately 691 Pa)

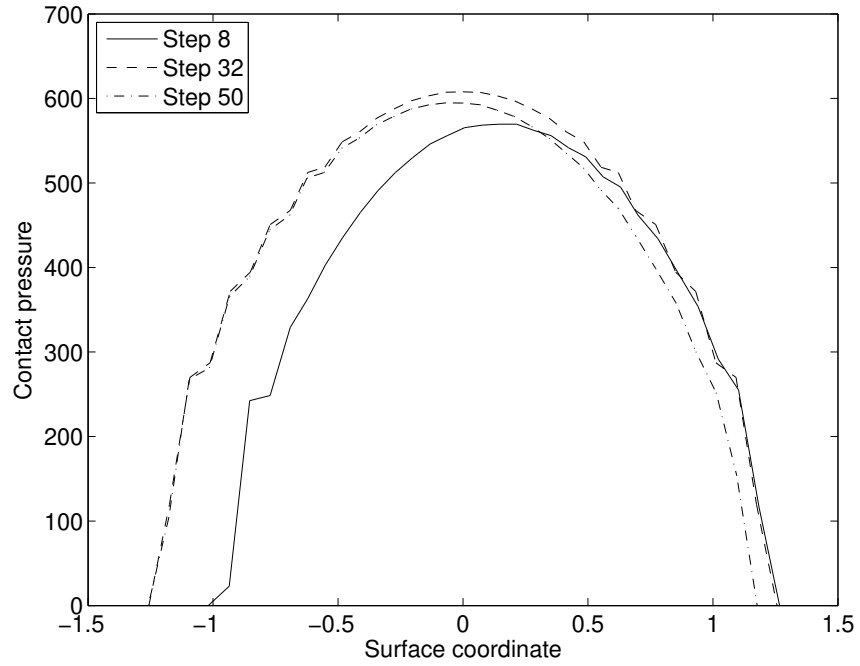


Figure 11: Ironing of an elastic foundation: Contact pressure distribution on the slave surface at the displacement steps 8, 32 and 50 as a function of the arc-length s of the faceted slave surface measured from the center of the curved arc forming the lower boundary of the indenter (the mesh comprises 5800 elements)

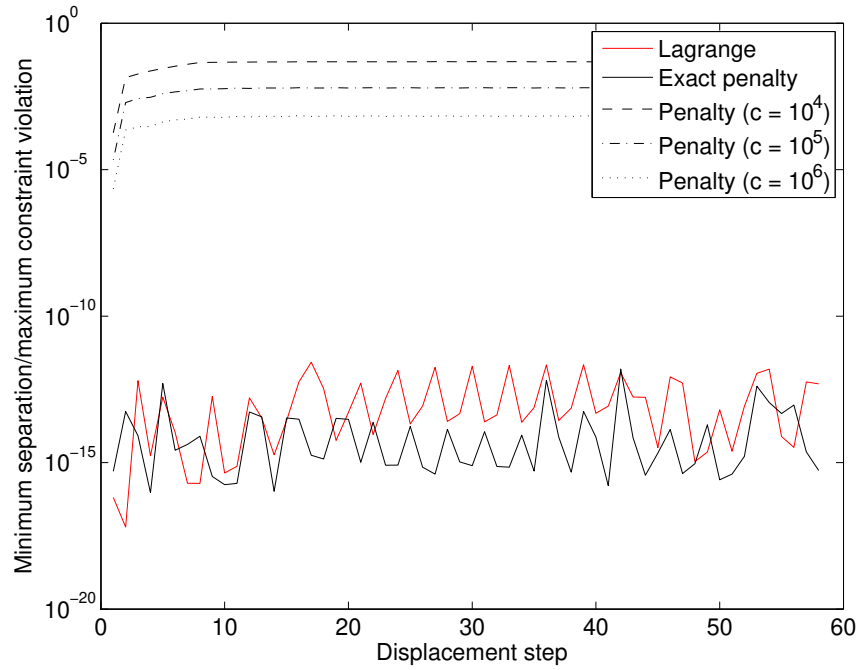


Figure 12: Ironing of an elastic foundation: Change in minimum separation (all constraints inactive) or maximum constraint violation (any constraint active) at equilibrium as the indenter slides on the foundation

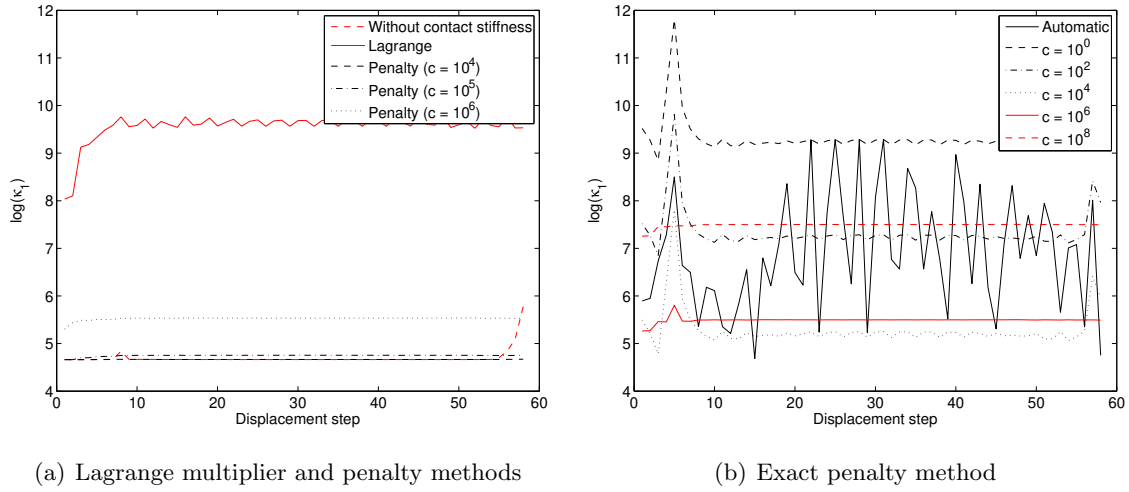


Figure 13: Ironing of an elastic foundation: Evolution of the 1-norm condition number of the equilibrium tangent stiffness matrices as a function of the displacement step for different constraint enforcement techniques

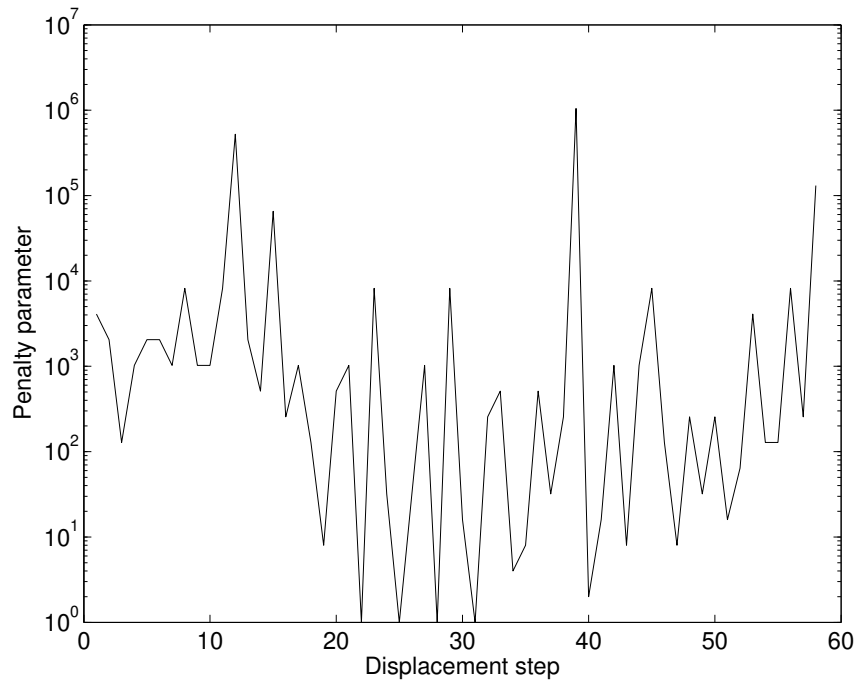


Figure 14: Ironing of an elastic foundation: Penalty parameter at equilibrium, as determined by the automatic update scheme (Steps 5.1 to 5.4 in Box 1), as a function of the displacement step

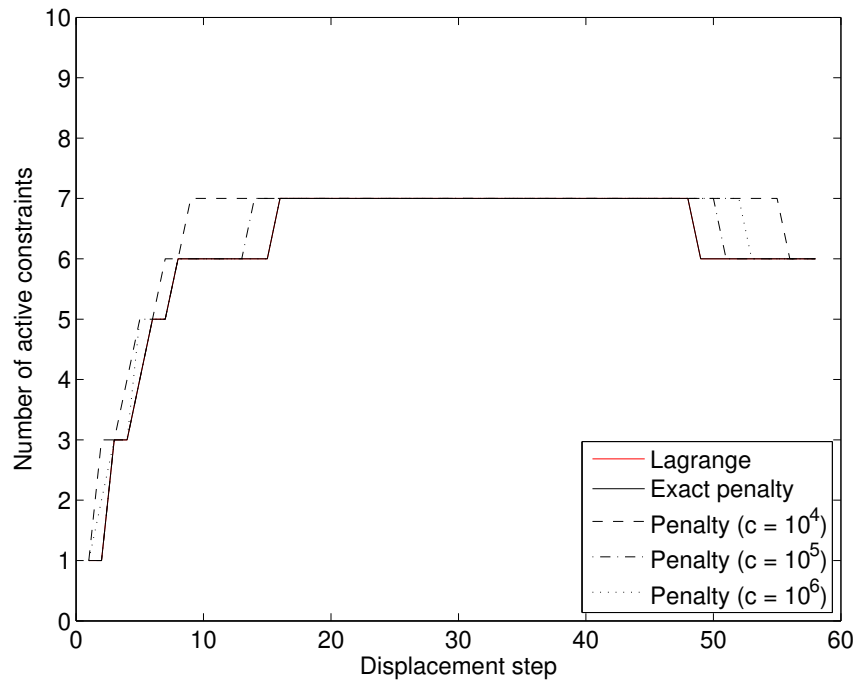


Figure 15: Ironing of an elastic foundation: Number of active contact constraints as a function of the displacement step (the curves for the exact penalty and Lagrange multiplier methods are coincident)