UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

Dizertační práce

Computational methods of shape design optimization with nonlinear semicoercive state problems: An elastic beam on a unilateral foundation

Školitel:
prof. RNDr. Ing. Lubomír Kubáček, DrSc.
Rok odevzdání: 2012

Vypracoval:
Mgr. Roman Šimeček
Apl. matematika, VI. ročník

Prohlášení

Prohlašuji, že jsem na základě zadání vytvořil tuto disertační práci samostatně pod vedením prof. RNDr. Ing. Lubomíra Kubáčka, DrSc., Dr. h. c., a že jsem v seznamu použité literatury uvedl všechny zdroje použité při zpracování práce.

V Olomouci dne 20. května 2012

Poděkování

Rád bych na tomto místě poděkoval panu Jiřímu V. Horákovi za čas, který věnoval konzultacím, a za jeho cenné připomínky při vzniku této práce.

Abstrakt

V dizertační práci se zabýváme problémem optimalizace rozměrů elastického nosníku na jednostranně pružném podloží. Stavová úloha má tvar okrajového problému s nelineární obyčejnou diferenciální rovnicí čtvrtého řádu. Budeme se zabývat dvěma konkrétními případy okrajových podmínek, pro které je úloha semikoercivní. Předmětem optimalizace bude tloušťka nosníku a koeficient tuhosti jeho podloží. Výsledná úloha potom spočívá v minimalizaci cenového funkcionálu na množině přípustných návrhových proměnných.

Nejprve stanovíme nutné a postačující podmínky existence a jednoznačnosti řešení stavového problému. Dokázána bude i spojitá závislost takového řešení na návrhové proměnné a existence alespoň jednoho řešení úlohy optimalizace.

Poté aproximujeme úlohu metodou konečných prvků. Hodnoty bilineární formy odpovídající podloží a cenového funkcionálu nemohou být vyčísleny přesně, proto je aproximujeme užitím vhodné kvadraturní formule pro numerickou integraci. Dokážeme existenci řešení aproximované úlohy a provedeme konvergenční analýzu.

Vzhledem k jednostrannosti podloží je algebrická forma stavové úlohy nelineární a pravděpodobně nediferencovatelná. Proto využijeme přístup založený na převodu a řešení takové úlohy ve formě problému smíšené lineární komplementarity. Diskrétní optimalizační úloha má potom tvar minimalizace nelineární, nediferencovatelné a možná i nekonvexní funkce na množině dané lineárními podmínkami ve tvaru rovnosti a nerovnosti.

V závěrečné části představujeme vhodný postup numerické realizace výsledného problému. Provádíme tzv. analýzu citlivosti a navrhujeme vzorce pro efektivní výpočet gradientu (subgradientu) cenového funkcionálu. Součástí práce je i kód (vytvořený ve jazycích C a Fortran), který implementuje navržený postup řešení. Jeho použití je demonstrováno na několika příkladech.

Klíčová slova: elastický nosník, optimalizace rozměrů, semikoercivní úloha, jednostranné podloží, analýza citlivosti, nehladká optimalizace, adjungovaná úloha

Abstract

A design optimization of an elastic beam with an elastic unilateral foundation will be studied in the thesis. The state problem is here represented by a nonlinear ordinary differential equation of 4-th order with boundary conditions. We will deal with two special cases of boundary conditions which cause semicoercivity of the state problem. The object of optimization will be the thickness of the beam and the stiffness coefficient of its foundation. The optimization problem is then formulated as a minimization of a cost functional over a set of all admissible design variables.

Firstly, we establish necessary and sufficient conditions for the existence and uniqueness of a solution to the state problem. The continuous dependence of the state problem solution on the design variable and the existence of at least one solution to the design optimization problem are proved.

After that, we approximate the problem using the finite element method. The bilinear form representing the foundation as well as the cost functional can not be evaluated exactly and therefore we approximate them making use of suitable quadrature formula for numerical integration. The existence of at least one solution to the approximated design optimization problem is established. Convergence analysis is made.

In view of the unilaterality of the foundation the algebraic form of the state problem is nonlinear and possibly nonsmooth. We make use of the approach which considers the state problem in a mixed linear complementarity form. The discrete optimization problem then leads to a minimization of a nonlinear, nonsmooth and possibly nonconvex function with respect to linear equality and inequality constraints.

Finally, we propose a suitable approach for numerical realization of the optimization problem. We make the design sensitivity analysis and propose a formula for efficient computation of a subgradient of the cost functional. We present a code (created in C/C++ and Fortran languages) which implements the approach presented in the thesis and we demonstrate how to use the program on several examples.

Keywords: elastic beam, shape design optimization, semicoercive beam problem, unilateral foundation, sensitivity analysis, nonsmooth optimization, adjoint problem

Contents

Introduction

0.1. Outline of the thesis

A design optimization of an elastic beam rested on an elastic unilateral foundation (subsoil) will be studied in the thesis. Shape design optimization has been the subject of considerable research and is of concern in many engineering applications. Let us mention civil and railway engineering for example.

We will mainly focus on long thin beams, therefore the well known Euler - Bernoulli mathematical model of the beam will be considered. This model is based on the theory of elasticity and if some required assumptions are satisfied (size of the beam, orientation of the load etc.) then it is represented by a boundary value problem for an ordinary differential equation of 4-th order. It takes the advantage of dimensional reduction and the problem is then described by an 1-D model (see e.g. [35]).

For the purpose of modeling of the contact between the beam and the foundation we will not consider the beam and the foundation as two elastic mutually non-penetrated bodies as it is usual in standard models of classical contact problems. The influence of the subsoil is represented in the model by adding the so called *response function* s which is in general dependent on the stiffness coefficient q of the subsoil, on the deflection u and its derivatives. The variant of linear (bilateral) subsoil with response function $s = qu$ is usually used and is well known from literature. This model has the advantage that the final mathematical model is linear and has a unique solution (see e.g. [24], [36]). Unfortunately, in some cases the linear model is not suitable. Especially, when the foundation is not firmly connected to the beam. Then the nonlinear (unilateral) model is more precise. In the thesis we consider one-parametric unilateral subsoil of Winkler's type with response function qu^+. The state problem is then described by a nonlinear differential equation. This kind of foundation is from the theoretical and practical point of view examined e.g. in [24], [50], [51]. It is also possible to use a two-parametric model of the foundation. Mostly used is the two-parametric Pasternak's model with response function $s = qu - ku''$, where the second parameter k relates to the shear forces in the subsoil. Special case is a unilateral rigid subsoil (rigid obstacle). The mathematical model then leads to a variational inequality (see e.g. [17], [24]).

The next important aspect of the problem are the boundary conditions. In the thesis we will deal with two particular cases of boundary conditions which cause semicoercivity of the state problem. To ensure the coercivity and therefore the existence and uniqueness of a solution, we will formulate some additional assumption on the beam load. Admissible rigid displacements with zero potential energy will no longer be allowed and the existence of nonzero contact zone between the beam and the foundation will be enforced.

The object of optimization will be the thickness t of the beam and the stiffness

coefficient q of its foundation. They appear in the problem as coefficients of the differential operator defining the state problem. The thickness will be represented by Lipschitz continuous bounded functions. The stiffness coefficient will be represented by Lebesgue integrable bounded functions. The optimization problem is then formulated as a minimization of a cost functional over a set of all admissible design variables. Many works have been done in this field. Firstly let us mention [16], [17] or [22]. Optimization of beams with linear foundation is studied e.g. in [25], [26] or [37]. Design optimization of a beam with unilateral supports is presented in [21]. A related problem, optimization of an axisymmetric plate on elastic foundation is treated in [45], [46]. But none of these works is concerning a beam optimization with semicoercive state problem.

0.2. Main tasks of the thesis

- The first task of the thesis is to make a complete mathematical analysis of the given optimization problem. There are two steps in this analysis. Firstly try to formulate necessary and sufficient condition for the existence and uniqueness of a solution to the state problem $(\mathcal{P}(e))$. And secondly we prove the existence of a solution to the optimization problem $(\mathcal{P}(e))$.

- The next task is the approximation of (P) and the convergence analysis. Define the approximated state problem and again formulate necessary and sufficient condition for the existence and uniqueness of its solution. Next we should define the approximation of $(\mathcal{P}(e))$ and prove that there exists at least one solution of it. Finally we need to study the relation between continuous and discrete solutions for $h \to 0$.

- Define the algebraic form of (P) and try to propose a suitable and efficient solution approach for it. Make the sensitivity analysis and establish a formula for subgradient computation.

- Implement the proposed solution algorithm in a programming language and present its functionality on several examples.

The thesis is organized as follows: In Section 1 the mathematical model of the problem is defined. In Section 2 we study the optimization problem (P) for the first case of boundary conditions. Firstly we analyze the state problem $(\mathcal{P}(e))$, where e is the design variable. We formulate necessary and sufficient condition for the existence and uniqueness of a solution to $(\mathcal{P}(e))$. We make use of decomposition of the space of kinematically admissible displacements to a closed convex cone of rigid displacements and its negative polar cone. Then using a modification of the well known Poincaré inequality we prove the coercivity of the problem. Secondly, we will turn our attention to the optimization problem (P). Uniform boundedness of a solution to $(\mathcal{P}(e))$ and its continuous dependence

on the design variable e will be established. Finally the existence of a solution to (P) is proved. Section 2 then continues by the approximation of the problem. The finite element approximation (P_h) of (P) is presented here. The existence and uniqueness analysis of the approximated problem is made and it will be shown that there exists at least one solution to (P_h), $\forall h > 0$. The final part of Section 2 contains the convergence analysis. It will be established that solutions of (P_h) are close on subsequences to the solution of (P) as $h \to 0$.

In Section 3 the optimization (P) problem for the second case of boundary conditions is studied. We proceed similarly as in Section 2.

In Section 4 we define the algebraic form of the problem. For the state problem we make use of the approach presented in [36] which is based on application of Gauss-Lobatto quadrature formula and decomposition of the deflection in integration nodes into positive and negative part. The discrete state problem then takes a form of mixed linear complementarity problem. The discrete optimization problem leads to a minimization of a nonlinear, nonsmooth and possibly nonconvex function with respect to linear equality and inequality constraints. The second part of Section 4 is dedicated to the sensitivity analysis. The cost functional, as a composite mapping, can be nondifferentiable. We will show its Lipschitz continuity and the existence of at least one subgradient in each point. At the end of the section we propose an approach of efficient computing of these subgradients. This approach is based on the definition of the so called adjoint problem and on a decomposition of the constraint set on active, inactive and semi-active constraints.

Section 5 contains a brief summary of numerical methods used to solve the state problem in the mixed linear complementarity form and the optimization problem in the nonlinear nonsmooth mathematical programming form. We have chosen (in cooperation with prof. Mäkela from University of Turku, Finland) methods MPBNGC, PBUN, PVAR and PNEW for nonsmooth and nonconvex optimization, see [28], [29], [30], [32] and [33].

In Section 6 we present the code created in C/C++ and Fortran languages (the code is available on the attached CD) and in Section 7 we demonstrate its functionality on several examples.

0.3. Basic notation

Through the thesis we will use the following notation:

$\mathbb{N}$ ... Set of all positive integers.

$\mathbb{R}^n$, $n \geqslant 1$ ... real n - dimensional Euclidean space. The corresponding norm will be denoted by $\|\cdot\|_n$ and the scalar product by $(\cdot,\cdot)_n$.

$\Omega \subset \mathbb{R}^1$ ... open, nonempty and bounded interval in $\mathbb{R}^1$. The closure of Ω will be denoted by $\bar{\Omega}$.

$C^k(\bar{\Omega})$ ... spaces of functions whose derivatives up to order k, $(k = 0, 1, \dots)$ are continuous in $\bar{\Omega}$. The corresponding norm will be denoted by $\|\cdot\|_{C^k(\bar{\Omega})}$. For more detailed information see [27].

$L^p(\Omega),\ p \geqslant 1$ ... Lebesgue spaces. We will denote the norm of $L^p(\Omega)$ by $\|\cdot\|_{p,\Omega}$. The standard scalar product in $L^2(\Omega)$ will be denoted by $(\cdot,\cdot)_{2,\Omega}$. See e.g. [2], [27].

$W^{k,p}(\Omega)$ ... Sobolev spaces $(k, p = 1, 2, \dots)$. Standard norm of the space $W^{k,p}(\Omega)$ will be denoted as $\|\cdot\|_{k,p,\Omega}$ and the i-th seminorm we will denote by $|\cdot|_{i,p,\Omega}$, $i = 1, 2, \dots, k$. Especially for $p = 2$ we will use the notation $W^{k,2}(\Omega) = H^k(\Omega)$. The space $H^k(\Omega),\ k = 0, 1, 2, \dots$ is a Hilbert space and its scalar product will be denoted by $(\cdot,\cdot)_{k,2,\Omega}$. For more information about Sobolev spaces see e.g. [2], [27].

P_k ... space of polynomials of k-th degree $(k = 0, 1, 2, \dots)$.

1. Mathematical model of the optimization problem

Let us consider an elastic beam of length l which is situated in the interval $\Omega := (0, l)$. The beam has a rectangular cross section and its thickness is represented by function t. The beam is subject to a vertical load f. The well known one-dimensional Euler-Bernoulli model for long thin beams will be used to compute the deflection. This 1-D model is obtained under some assumption (size of the beam, orientation of the load etc.) from the general 3-D elasticity problem by dimensional reduction, see [35].

Along its entire length the beam is supported by a unilateral elastic foundation of Winkler's type. The influence of the subsoil is added to the model by the so called response function dependent on the stiffness coefficient q and the deflection u. The response function for the unilateral Winkler's subsoil is defined by $s = qu^+$ and therefore the foundation is active only if the beam deflects against it (see e.g. [53]).

In the thesis we will consider two variants of boundary conditions. Both cases allow the existence of rigid beam displacements and cause the semicoercivity of the state problem.

The classical formulation of the beam bending problem has the form of nonlinear differential equation of 4-th order with mixed boundary conditions:
Find $u \in C^4(\Omega) \cap C^3(\bar{\Omega})$ such that

$$\begin{cases} (\beta(x)t^3(x)u''(x))'' + q(x)u^+(x) \;=\; f(x) & \forall x \in \Omega, \\ \text{a)} \quad u'(0) = u'''(0) = u''(l) = u'''(l) = 0, \\ \text{b)} \quad u(0) = u'''(0) = u''(l) = u'''(l) = 0, \end{cases} \tag{1}$$

where t, q and f are functions corresponding to the beam thickness, the foundation stiffness and the intensity of the vertical load. Function u represents the deflection of the beam and u^+ is its positive part

$$u^+(x) = \frac{u(x) + |u(x)|}{2}, \; x \in \Omega.$$

Function β has the following form:

$$\beta(x) = \frac{2}{3} b(x) E(x),$$

where E denotes the Young's modulus of elasticity and b is a function representing the width of the beam. In the sequel we will consider β to be a constant.

Remark 1.1. *Instead of Euler-Bernoulli mathematical model it is possible to consider the Timoshenko model, where the plane normal to the beam axis before deformation remains plane after deformation, but not necessarily normal to*

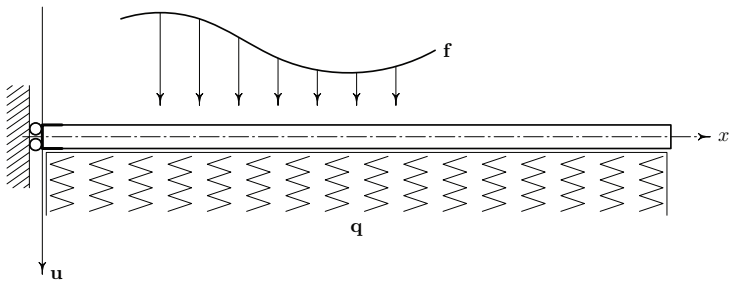

Figure 1: Outline of the beam with axes orientation.

the deformed axis as it is in Euler-Bernoulli case, see e.g. [37], [47] or [42]. Transverse shear deformations are considered in this model. But we have to keep in mind that the Timoshenko model is suitable especially for short, equivalently thick, beams but Euler-Bernoulli is valid only for long span, equivalently thin, beams.

Optimizaion of a Timoshenko beam with a linear elastic foundation and comparison to the results for Euler-Bernoulli model is treated in [37].

The thickness t and the stiffness coefficient q will be the subject of optimization. Unlike the standard optimization problems (see e.g. [16], [17], [41]) in our case the design variables appear as coefficients of the differential operator of the state equation and the integration area remain fixed. The set of all admissible thicknesses t will be defined as follows:

$$U_{ad}^t = \big\{ t \in C^{0,1}(\bar{\Omega}) \ : 0 < t_0 \leq t(x) \leq t_1 \text{ in } \Omega, \\ \int_\Omega t(x)\,\mathrm{d}x = \gamma_1, \ \ |t'(x)| \leqslant \gamma_2 \text{ in } \Omega \ \big\}.$$

The thickness is represented by Lipschitz continuous and bounded functions. Constants t_0, t_1, γ_1 and γ_2 are chosen in such a way that U_{ad}^t is nonempty. Constraints from the set U_{ad}^t are reasonable from the physical point of view as well as they play an important role in the mathematical analysis of the problem. For example the constraint $|t'(x)| \leqslant \gamma_2$ in Ω prevents thickness oscillation and ensures that thickness functions are uniformly continuous.

The set of all admissible stiffness coefficients q can be defined in a similar way:

$$U_{ad}^q = \big\{ q \in L^2(\Omega) \ : \ q_0 \leq q(x) \leq q_1 \ \text{ a.e. in } \Omega \ \big\}. \tag{2}$$

The optimal foundation stiffness will be chosen from the set of Lebesgue integrable function that are bounded in the interval Ω. Constants q_0, q_1 are set in such a way that $U_{ad}^q \neq \emptyset$.

Finally the set of all admissible design variables is defined as the Cartesian product

$$U_{ad} = U_{ad}^t \times U_{ad}^q. \tag{3}$$

Elements of U_{ad} will be denoted by $e = \{t, q\}$.

The classical formulation (1) can be used only if the input data β, f, t and q are sufficiently smooth. In practical applications we often can not guarantee this smoothness. Therefore we define the variational formulation that is based on the minimum potential energy principle (see e.g. [35]). It enable us to weaken the assumptions on the input data. Let $e \in U_{ad}$ is arbitrary but fixed, $\beta \in L^\infty(\Omega)$ and let there exist a constant β_0 such that $0 < \beta_0 \leqslant \beta(x)$ a.e. in Ω. Now we define spaces of kinematically admissible displacements for two variants of boundary conditions in (1):

$$V_1 = \{v \in H^2(\Omega) : \ v'(0) = 0\},$$
$$V_2 = \{v \in H^2(\Omega) : \ v(0) = 0\}.$$

In what follows we will consider $V = V_1$ resp. $V = V_2$. Forms $a_t : H^2 \times H^2 \to \mathbb{R}$ and $b_q : H^1 \times H^1 \to \mathbb{R}$ representing inner energy and work of the foundation are defined as follows:

$$a_t(u, v) := \int_\Omega \beta\, t^3 u'' v''\, \mathrm{d}x, \qquad b_q(u, v) := \int_\Omega q\, u\, v\, \mathrm{d}x.$$

It is clear that these forms are bilinear $\forall e \in U_{ad}$. Work of outer forces is represented by a linear functional $F : H^2 \to \mathbb{R}$. If we denote $L(v) := \int_\Omega f(x)v(x)\, \mathrm{d}x$ then

$$F(v) := L(v) + \sum_i F_i v(x_i) - \sum_j M_j v'(x_j).$$

Values F_i, M_j correspond to generalized forces in points $x_i, x_j \in \bar{\Omega}$. Functional of total potential energy then reads as follows:

$$\mathcal{E}_e(v) = \frac{1}{2}(a_t(v, v) + b_q(v^+, v^+)) - F(v), \qquad v \in H^2(\Omega),\ e \in U_{ad}.$$

By the variational formulation of the state problem corresponding to $e \in U_{ad}$ we mean the following problem:

$$\text{Find } u \in V : \ \mathcal{E}_e(u) \leqslant \mathcal{E}_e(v) \quad \forall v \in V. \tag{$\mathcal{P}(e)$}$$

As the last part of the optimization problem we define the cost functional. In general it is a mapping $I : U_{ad} \times V \to \mathbb{R}^1$. Let us denote $J(e) \equiv I(e, u(e))$, where $u(e)$ solves $(\mathcal{P}(e))$. Let us now present some practical examples of cost

functionals:

$$I_1(e,\, u(e)) \;\equiv\; J_1(e) = \int_\Omega f\, u(e)\, \mathrm{d}x, \tag{4}$$

$$I_2(e,\, u(e)) \;\equiv\; J_2(e) = \int_\Omega u^2(e)\, \mathrm{d}x, \tag{5}$$

$$I_3(e,\, u(e)) \;\equiv\; J_3(e) = \int_\Omega t^2 (u''(e))^2\, \mathrm{d}x. \tag{6}$$

The cost functional (4) represents the compliance of the beam and in fact it is closely related to the potential energy of the beam. Indeed

$$2\mathcal{E}_e(u(e)) = a_e(u(e),\, u(e)) + b_e(u^+(e), u^+(e)) - 2F(u(e)) \;=\; -F(u(e)) \;=\; -J(e).$$

In fact the minimization of the compliance is equivalent to a maximization of the total potential energy evaluated in the equilibrium state $u(e)$. The functional (5) corresponds to the distance between the deflection u and the zero function in the sense of least square method. The functional (6) corresponds to normal stresses at the extreme fiber of the beam. The last functional is the only one from them, which is explicitly dependent on the design variable $e = \{t,\, q\}$.

At this point we have everything what is needed for the definition of the design optimization problem:

$$\text{Find } e^* \in U_{ad}: \quad J(e^*) \;\leqslant\; J(e) \quad \forall e \in U_{ad}, \tag{P}$$

where $J(e) \equiv I(e,\, u(e))$ with $u(e)$ being a solution to $(\mathcal{P}(e))$.

From the following outline we can see the form of the cost functional and the specific scheme of design optimization problems.

$$e \longmapsto u(e) \longmapsto I(e, u(e)). \tag{7}$$

Shape design optimization problems can be difficult, especially its numerical realization. Every time we want to evaluate the cost functional, we have to solve the state problem $(\mathcal{P}(e))$ first. Usually we need to solve $(\mathcal{P}(e))$ several times in every iteration of the optimization algorithm. Moreover the cost functional obtained as a composition of two mappings (see (7)), does not need to be necessarily convex and continuously differentiable. As an example we can mention design optimization with state problems governed by variational inequalities, see e.g. [16], [17], [14]. In order to choose a suitable approach for the numerical realization we have to analyze the problem properly.

2. Natural boundary condition $u'(0) = 0$

2.1. Existence analysis of (P)

In this section we shall be aimed at the existence of a solution to the optimization problem (P) with

$$V = V_1 = \{v \in H^2(\Omega) : v'(0) = 0\}.$$

Firstly we establish a necessary and sufficient condition for existence and uniqueness of a solution to $(\mathcal{P}(e))$. Then we will proceed standardly by the proof of compactness of U_{ad} and by the proof of continuous dependence of the solution to $(\mathcal{P}(e))$ on the design variable e.

2.1.1. Existence and uniqueness of a solution to $(\mathcal{P}(e))$

The boundary conditions have the following form: $u'(0) = u'''(0) = u''(l) = u'''(l) = 0$. From the physical point of view it causes that the right end of the beam is free. The left end it is fixed in such a way that it can move in vertical the direction but it can not slope. The beam is allowed to float and the state problem

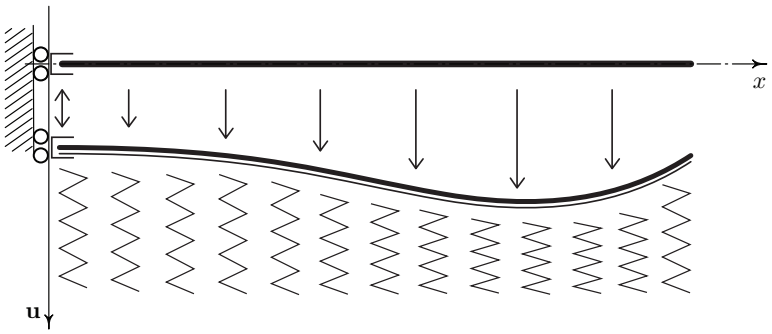

Figure 2: Outline of the beam with boundary condition $u'(0) = 0$.

is not coercive. For semicoercive problems it is typical that the input data of the problem must satisfy some additional conditions to enforce the coercivity (see e.g. [19], [23]). Therefore we must suppose that the resultant of the beam load acts against the subsoil, what prevents rigid displacements and ensures the solvability of $(\mathcal{P}(e))$.

In this subsection we will consider that the pair $e = \{t, q\} \in U_{ad}$ is arbitrary but fixed and $\beta \in L^\infty(\Omega)$, $0 < \beta_0 \leqslant \beta(x)$ a.e. v Ω. First let us recall the boundedness of forms a_t, b_q.

Lemma 2.1. *There exist positive constants c_1, c_2 such that*

$$|a_t(u, v)| \leqslant c_1 \|u\|_{2,2,\Omega} \|v\|_{2,2,\Omega} \quad \forall u, v \in H^2(\Omega), \forall e \in U_{ad},$$
$$|b_q(u^+, v)| \leqslant c_2 \|u\|_{2,2,\Omega} \|v\|_{2,2,\Omega} \quad \forall u, v \in H^2(\Omega), \forall e \in U_{ad}.$$

Proof. By using the well–known Cauchy–Schwarz inequality (see Theorem 8.1), Theorem 8.4, (3) and Lemma 8.3 we easily receive the assertion of the lemma. ■

In what follows we will prove some important properties of the functional $\mathcal{E}_e$. Its Gâteaux differentiability, convexity and coercivity on V will be sufficient for the existence of a solution to $(\mathcal{P}(e))$.

Lemma 2.2. *The functional $\mathcal{E}_e$ is Gâteaux differentiable and convex on $H^2(\Omega)$. Its Gâteaux derivative in arbitrary point $u \in H^2(\Omega)$ and arbitrary direction $v \in H^2(\Omega)$ reads*

$$\mathcal{E}_e^{'}(u;\, v) = a_t(u,\, v) + b_q(u^+, v) - F(v) \quad \forall e \in U_{ad}. \tag{8}$$

Proof. It is easy to prove that the following auxiliary relation holds:

$$\lim_{\epsilon \to 0^+} \frac{[(s+\epsilon t)^+]^2 - [s^+]^2}{\epsilon} = 2s^+ t, \quad \forall s, t \in \mathbb{R}^1.$$

From there we have:

$$\lim_{\epsilon \to 0} \frac{b_q((u+\epsilon v)^+, (u+\epsilon v)^+) \, - \, b_q(u^+, u^+)}{\epsilon} \, = \, 2b_q(u^+, v)$$

$\forall u, v \in H^2(\Omega)$, $\forall e \in U_{ad}$. In case of the bilinear form a_t and the linear functional F we proceed in a standard way. Then we directly obtain (8). In view of the fact that $\mathcal{E}_e^{'}(u;\, \cdot)$ is linear and continuous on $H^2(\Omega)$, the functional $\mathcal{E}_e$ is Gâteaux differentiable on $H^2(\Omega)$.

Next we prove the convexity of $\mathcal{E}_e$. Using assumptions $e \in U_{ad}$, $0 < \beta_0 \leqslant \beta$ a.e. in Ω and the inequality

$$(s^+ - t^+)(s - t) \, \geqslant \, (s^+ - t^+)^2 \quad \forall s, t \in \mathbb{R}^1 \tag{9}$$

we obtain

$$\begin{aligned} \mathcal{E}_e^{'}(u;\, u-v) \, - \, \mathcal{E}_e^{'}(v;\, u-v) &= a_t(u-v,\, u-v) \, + \, b_q(u^+ - v^+,\, u-v) \geqslant \\ &\geqslant a_t(u-v,\, u-v) \, + \, b_q(u^+ - v^+,\, u^+ - v^+) \geqslant \\ &\geqslant \beta_0 t_0^3 \, |u-v|^2_{2,2,\Omega} \, + \, q_0 \, \left\|u^+ - v^+\right\|^2_{2,\Omega} \, \geqslant \, 0 \\ &\quad \forall u, v \in H^2(\Omega),\, \forall e \in U_{ad}. \end{aligned}$$

This estimate is a sufficient condition of convexity of Gâteaux differentiable functional $\mathcal{E}_e$, see e.g. [12], [8].

■

According to Lemma 2.2 we can introduce the equivalent weak formulation of $(\mathcal{P}(e))$:

$$\text{Find } u \in V: \quad a_t(u,\, v) \,+\, b_q(u^+, v) \,=\, F(v) \quad \forall v \in V. \tag{$\mathcal{P}'(e)$}$$

The weak formulation can be obtained from the classical formulation (1) by multiplying by a testing function $v \in V$, integrating over the interval Ω, using boundary conditions and suitable Green's formula for integration per partes. If the solution u and the input data of the problem are smooth enough we can also pass from the weak formulation to the classical one.

In what follows we will focus on the key property of $\mathcal{E}_e$, its coercivity. In our case $\mathcal{E}_e$ is only semicoercive on V, i.e. there exists a constant $c > 0$ such that

$$a_t(v, v) \,+\, b_q(v^+, v) \,\geqslant\, c\,|v|^2_{2,2,\Omega} \quad \forall e \in U_{ad},\, \forall v \in V. \tag{10}$$

By prescribing suitable conditions on the beam load we eliminate rigid displacements from the problem, enforce the coercivity of $\mathcal{E}_e$ and the existence of a solution to $(\mathcal{P}(e))$. In our case we make use of the approach based on orthogonal decomposition of V into a convex closed cone of rigid displacements and its negative polar cone (see e.g. [24]).

The set of rigid displacements is generally given by linear polynomials in the form $p = ax + b \in P_1$. If we take into account the natural boundary condition $u'(0) = 0$, it reduces to the set of all constant polynomials $p \in P_0$. All rigid displacements for which the subsoil is inactive must be eliminated from the problem. A set of such displacements is defined as

$$\mathcal{R}_V \,=\, \{v \in V \cap P_1 : a_t(v, v) + b_q(v^+, v) = 0\} \,=\, \{p \in P_0 : p \leqslant 0\}. \tag{11}$$

It is easy to prove that $\mathcal{R}_V$ is closed convex cone. Using the definition of the standard scalar product on $H^2(\Omega)$ the negative polar cone is defined by

$$\mathcal{R}_V^{\ominus} = \{v \in V : (v, p)_{2,2,\Omega} \leqslant 0 \;\; \forall p \in \mathcal{R}_V\} = \{v \in V : (v, 1)_{2,\Omega} \geqslant 0\}. \tag{12}$$

From (12) it again follows that $\mathcal{R}_V^{\ominus}$ is convex and closed cone.

Theorem 2.1. (Necessary condition for the existence of a solution to $(\mathcal{P}(e))$.) *Let there exist at least one solution to $(\mathcal{P}(e))$, then the condition*

$$F(1) \,=\, L(1) \,+\, \sum_i F_i \,\geqslant\, 0 \tag{S1}$$

must be satisfied.

Proof. Let $u \in V$ be a solution to $(\mathcal{P}(e))$. Then

$$a_t(u, v) \,+\, b_q(u^+, v) \,=\, F(v) \quad \forall v \in V. \tag{13}$$

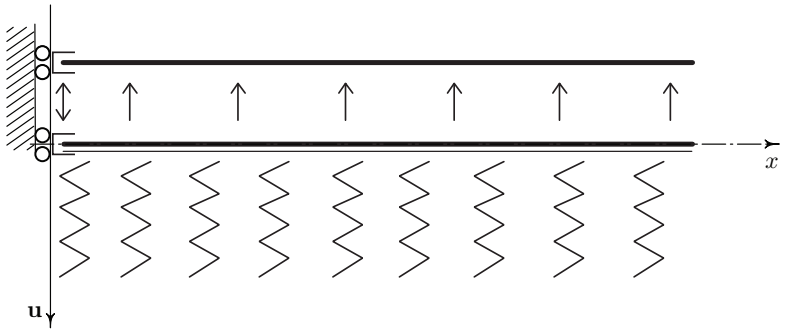

Figure 3: Rigid beam motions belonging to $\mathcal{R}_V$.

Inserting $v \equiv p \in \mathcal{R}_V$ into (13) we have:

$$0 \geqslant b_q(u^+, p) \;=\; a_t(u, p) \;+\; b_q(u^+, p) \;=\; F(p) \;=\; p\left(L(1) \;+\; \sum_i F_i\right).$$

■

The next lemma says that the space V can be decomposed into a direct orthogonal sum of cones $\mathcal{R}_V$ and $\mathcal{R}_V^{\ominus}$.

Lemma 2.3. *Let $\mathcal{R}_V$, $\mathcal{R}_V^{\ominus}$ be defined by (11) and (12), respectively. Then V can be decomposed as follows:*

$$V = \mathcal{R}_V \oplus \mathcal{R}_V^{\ominus}.$$

In addition $\forall v \in V \;\exists!\; \{p, \bar{v}\} \in \mathcal{R}_V \times \mathcal{R}_V^{\ominus}$ such that

$$v \;=\; p \oplus \bar{v}, \quad (p, \bar{v})_{2,2,\Omega} \;=\; p(1, \bar{v})_{2,\Omega} \;=\; 0. \tag{14}$$

Proof. For the proof we refer to [3].

■

In view of the definition of $\mathcal{R}_V$, $\mathcal{R}_V^{\ominus}$ and the orthogonality (14) we can deduce that only one of the following two variants can occur:

$$p \;=\; 0 \quad \text{and} \quad (\bar{v}, 1)_{2,\Omega} \;\geqslant\; 0, \tag{A1}$$
$$p \;\leqslant\; 0 \quad \text{and} \quad (\bar{v}, 1)_{2,\Omega} \;=\; 0. \tag{A2}$$

The next lemma will play an important role in the proof of coercivity of $\mathcal{E}_e$. In fact it is a suitable modification of the well known Poincaré inequality, see e.g. [43], [27] or [2].

Lemma 2.4. (Poincaré type inequality) *Let* $V = \{v \in H^2(\Omega) : v'(0) = 0\}$, *then there exists a positive constant* c_P *depending only on the length of interval* $\Omega := (0,\, l)$ *such that*

$$\|v\|^2_{2,2,\Omega} \leqslant c_P \left(|v|^2_{2,2,\Omega} + (v,\, 1)^2_{2,\Omega} \right) \quad \forall v \in V. \tag{15}$$

Proof. Suppose that (15) does not hold. Then one can find a sequence $\{v_n\} \subset V$ such that

$$\frac{1}{n} \|v_n\|^2_{2,2,\Omega} > |v_n|^2_{2,2,\Omega} + (v_n,\, 1)^2_{2,\Omega} \geqslant 0 \quad \forall n \geqslant 1. \tag{16}$$

First, let us divide the inequality (16) by $\|v_n\|^2_{2,2,\Omega}$ and pass to the limit for $n \to \infty$. Then

$$\lim_{n\to\infty} |w_n|^2_{2,2,\Omega} = 0 \quad \text{and} \quad \lim_{n\to\infty} (w_n,\, 1)^2_{2,\Omega} = 0, \tag{17}$$

where we denote $w_n = v_n / \|v_n\|_{2,2,\Omega}$. Clearly $\|w_n\|_{2,2,\Omega} = 1$ and $\{w_n\}$ is bounded in $H^2(\Omega)$. Hence one can find a subsequence of $\{w_n\}$ (denoted by the same sequence) and an element $w \in V$ such that $w_n \rightharpoonup w$ in V. In view of (17) it holds that $w_n \to w$ in V. Therefore we have

$$0 = \liminf_{n\to\infty} |w_n|^2_{2,2,\Omega} = |w|^2_{2,2,\Omega} \geqslant 0.$$

Then $|w|^2_{2,2,\Omega} = 0$ and $w \equiv p \in P_0$. From the inequality

$$0 = \liminf_{n\to\infty} (w_n,\, 1)^2_{2,\Omega} = (w,\, 1)^2_{2,\Omega} \geqslant 0,$$

it follows that $p = 0$. But it leads to a contradiction with $\|w_n\|_{2,2,\Omega} = 1$ and $w_n \to p$ in V.

■

Let us now approach to the proof of coercivity of $\mathcal{E}_e$. Theorem 2.1 says that the necessary condition is in form $F(1) \geqslant 0$. Unfortunately it is not a sufficient condition which ensures the coercivity.

Lemma 2.5. *Let the condition*

$$F(1) = L(1) + \sum_i F_i > 0 \tag{S2}$$

be fulfilled. Then the functional $\mathcal{E}_e$ *is coercive on* V.

Proof. Let (S2) hold. In view of (14) we can write:

$$\begin{aligned} 2\mathcal{E}_e(v) = 2\mathcal{E}_e(p+\bar v) &= a_t(\bar v, \bar v) + b_q(v^+, v^+) - 2F(p) - 2F(\bar v) \geqslant \\ &\geqslant \beta_0 t_0^3 \, |\bar v|^2_{2,2,\Omega} + q_0 \left| (p+\bar v)^+ \right|^2_{0,2,\Omega} + 2|p|F(1) - 2F(\bar v). \end{aligned}$$

We know that only one of the variants (A1), (A2) can occur. Firstly we will focus on variant (A1) for which it holds $p = 0$, $v \equiv \bar{v}$ and $(\bar{v}, 1)_{2,\Omega} \geqslant 0$. The following inequality is a consequence of properties of function $\bar{v}^+$:

$$0 \leqslant (\bar{v}, 1)^2_{2,\Omega} \leqslant (\bar{v}^+, 1)^2_{2,\Omega} \leqslant l \left\|\bar{v}^+\right\|^2_{2,\Omega}. \tag{18}$$

Using (15) and (18) it reads

$$\begin{aligned} 2\mathcal{E}_e(v) = 2\mathcal{E}_e(\bar{v}) \;&=\; a_t(\bar{v}, \bar{v}) \;+\; b_q(\bar{v}^+, \bar{v}^+) \;-\; 2F(\bar{v}) \geqslant \\ &\geqslant \beta_0 t_0^3 \; |\bar{v}|^2_{2,2,\Omega} \;+\; q_0 \left\|\bar{v}^+\right\|^2_{2,\Omega} \;-\; 2F(\bar{v}) \geqslant \\ &\geqslant \beta_0 t_0^3 \; |\bar{v}|^2_{2,2,\Omega} \;+\; \frac{q_0}{l}(\bar{v}, 1)^2_{2,\Omega} \;-\; 2F(\bar{v}) \geqslant \\ &\geqslant \|\bar{v}\|_{2,2,\Omega} \, (c_1 \, \|\bar{v}\|_{2,2,\Omega} \;-\; 2 \, \|f\|_{2,\Omega}), \end{aligned} \tag{19}$$

denoting $c_1 := (1/c_P) \min\{\beta_0 t_0^3, \, q_0/l\}$.

Secondly, in the case of variant (A2) we have $(\bar{v}, 1)_{2,\Omega} = 0$ and $p \leqslant 0$. Again by using (15) we obtain

$$\begin{aligned} 2\mathcal{E}_e(v) = 2\mathcal{E}_e(p + \bar{v}) \;&=\; a_t(\bar{v}, \bar{v}) \;+\; b_q(v^+, v^+) \;-\; 2F(p) \;-\; 2F(\bar{v}) \geqslant \\ &\geqslant \beta_0 t_0^3 \; |\bar{v}|^2_{2,2,\Omega} \;+\; q_0 \left\|(p + \bar{v})^+\right\|^2_{2,\Omega} \;+\; 2|p|F(1) \;-\; 2F(\bar{v}) \geqslant \\ &\geqslant \beta_0 t_0^3 \, |\bar{v}|^2_{2,2,\Omega} \;+\; 2|p|F(1) \;-\; 2F(\bar{v}) = \\ &= \beta_0 t_0^3 \, |\bar{v}|^2_{2,2,\Omega} \;+\; (\bar{v}, 1)^2_{2,\Omega} \;+\; 2|p|F(1) \;-\; 2F(\bar{v}) \geqslant \\ &\geqslant c_2 \, \|\bar{v}\|^2_{2,2,\Omega} \;+\; 2|p|F(1) \;-\; 2 \, \|f\|_{2,\Omega} \, \|\bar{v}\|_{2,2,\Omega}, \end{aligned} \tag{20}$$

where $c_2 := (1/c_P) \min\{\beta_0 t_0^3, \, 1\}$. Due to (14) it holds that $\|v\|^2_{2,2,\Omega} = \|\bar{v}\|^2_{2,2,\Omega} + \|p\|^2_{2,2,\Omega}$. Therefore if $\|v\|_{2,2,\Omega} \to +\infty$ then at least one part of the function v in appropriate norm has to converge to $+\infty$. And finally we make use of condition (S2) that ensures the coercivity.

■

Now we can establish the main results of this subsection. Coercivity of $\mathcal{E}_e$ enable us to introduce the following theorem.

Theorem 2.2. (Necessary and sufficient condition for the existence and uniqueness of a solution to $(\mathcal{P}(e))$.) *The state problem $(\mathcal{P}(e))$ has a unique solution if and only if the condition (S2) is fulfilled. Such a solution $u \in V$ can be characterized in the following way:*

$$\mu(M_u) > 0, \tag{M1}$$

where $\mu(M_u)$ denotes the one–dimensional Lebesgue measure of the set

$$M_u = \{x \in \Omega : \, u(x) > 0\}.$$

Proof. *Necessity.* This part of the proof will be made by contradiction. Assume that $u \in V$ is a unique solution of $(\mathcal{P}(e))$ and (S2) does not hold. According to Theorem 2.1, the equality $F(1) = 0$ have to hold. By inserting $v = p \in \mathcal{R}_V$ into $(\mathcal{P}'(e))$ we obtain

$$a_t(u,p) + b_q(u^+,p) = F(p) = pF(1) \quad \forall p \in \mathcal{R}_V,\ p \neq 0, \tag{21}$$

$$b_q(u^+,p) = 0 \quad \forall p \in \mathcal{R}_V,\ p \neq 0. \tag{22}$$

The equality (22) implies $u^+ = 0$, thus $u \leqslant 0$ a.e. in Ω. Then it clearly holds that $u + p < 0$ a.e. in Ω, $\forall p \in \mathcal{R}_V,\ p \neq 0$. From there $b_q((u+p)^+, v) = 0\ \forall p \in \mathcal{R}_V,\ p \neq 0$ and $\forall v \in V$. It is not difficult to see that $u + p$ is another solution of $(\mathcal{P}(e))$ being in contradiction with the uniqueness of u. The condition (S2) must be then fulfilled.
Sufficiency. Let (S2) be satisfied. According to Lemma 2.2 and Lemma 2.5 we know that $\mathcal{E}_e$ is Gâteaux differentiable, convex and coercive on V implying the existence of a solution to $(\mathcal{P}(e))$, see e.g. [12].

Further let $u \in V$, $u \leqslant 0$ a.e. in Ω, solve $(\mathcal{P}(e))$. Then by setting $v \equiv p \in \mathcal{R}_V,\ p \neq 0$ we receive

$$0 = b_q(u^+,p) = p\,F(1)$$

what is in contradiction with (S2). Therefore the set M_u must have a positive Lebesgue measure.

Finally suppose that $u_1,\ u_2 \in V$ are solutions to $(\mathcal{P}(e))$. Then

$$a_t(u_1,v) + b_q(u_1^+,v) = F(v) \quad \forall v \in V, \tag{23}$$

$$a_t(u_2,v) + b_q(u_2^+,v) = F(v) \quad \forall v \in V. \tag{24}$$

Subtracting (24) from (23) and putting $v = u_1 - u_2$ yield

$$a_t(u_1 - u_2, u_1 - u_2) + b_q(u_1^+ - u_2^+, u_1 - u_2) = 0.$$

Making use of the definition of a_t, b_q we obtain

$$u_1 - u_2 = p \in P_0 \quad \text{and} \quad u_1^+ - (u_1 - p)^+ = 0 \text{ a.e. in } \Omega.$$

Taking into account (M1), we have $p = 0$ and therefore $u_1 = u_2$ a.e. in Ω. Solution of the problem $(\mathcal{P}(e))$ is unique.

■

2.1.2. Existence of solutions to (P)

In this subsection we shall focus on the proof of existence of a solution to the optimization problem (P). Firstly we prove the compactness of U_{ad}. The next task will be the analysis of the mapping $u : e \mapsto u(e)$. We shall prove that $u(e)$ depends continuously on e what will be sufficient for the existence of

an optimal solution to (P). In addition we will prove the Lipschitz continuity of this mapping. It ensures stability of $(\mathcal{P}(e))$, i.e. small change of the design variable will produce a small change of the solution. Lipschitz continuity is also important for the numerical realization. Due to the unilaterality of the foundation there might occur a situation where the cost functional (composite mapping) will not be continuously differentiable. Then the Lipschitz continuity enable us to use a subgradient based nonsmooth optimization algorithm, see e.g. [31], [32], [28], [30].

We start by the definition of convergence in U_{ad}. After that we will focus on the compactness of U_{ad}. Convergence in the set U_{ad}^t will be defined as uniform convergence of continuous functions in the interval Ω:

$$t_n \to t \text{ in } U_{ad}^t \;\Leftrightarrow\; t_n \rightrightarrows t \text{ in } C(\bar{\Omega}). \tag{25}$$

Lemma 2.6. *The set U_{ad}^t with convergence defined by (25) is a compact subset of $C(\bar{\Omega})$.*

Proof. Functions belonging to U_{ad}^t are uniformly bounded and due to the condition $|t'(x)| \leqslant T_3$ uniformly continuous in Ω. Then according to Arzela - Ascoli theorem (see e.g. [17], [27]) U_{ad}^t with the convergence (25) is a compact subset of $C(\bar{\Omega})$.

■

Convergence in U_{ad}^q will be defined as weak convergence in the Lebesgue space $L^2(\Omega)$:

$$q_n \to q \text{ in } U_{ad}^q \;\Leftrightarrow\; q_n \rightharpoonup q \text{ in } L^2(\Omega). \tag{26}$$

Lemma 2.7. *The set U_{ad}^q with convergence defined by (26) is weakly compact subset of the space $L^2(\Omega)$.*

Proof. The set U_{ad}^q is closed and bounded in the reflexive Banach space $L^2(\Omega)$. According to Eberlein–Šmuljan theorem (see e.g. [12]) the set U_{ad}^q is weakly compact subset of $L^2(\Omega)$. ■

Finally we can introduce the convergence in U_{ad}:

$$e_n \to e \text{ in } U_{ad} \;\Leftrightarrow\; t_n \rightrightarrows t \text{ in } \Omega \;\wedge\; q_n \rightharpoonup q \text{ in } L^2(\Omega), \tag{27}$$

where $e_n = \{t_n, q_n\}$, $e = \{t, q\}$.

Lemma 2.8. *The set U_{ad} with convergence introduced in Definition 27 is a compact subset of $C(\bar{\Omega}) \times L^2(\Omega)$.*

Proof. The assertion follows from properties of the Cartesian product, Lemma 2.6 and Lemma 2.7.

■

Assume that $\beta \in L^\infty(\Omega)$, $0 < \beta_0 \leqslant \beta(x)$ a.e. in Ω and $F \in V^*$, $F(1) > 0$ are given. Then we know that for any $e \in U_{ad}$ there exists a unique solution to $(\mathcal{P}(e))$ with the property (M1). The set of all such solutions will be denoted by W:

$$W := \{\{u,t,q\} \in V \times U_{ad}^t \times U_{ad}^q \,:\, u := u(e) \text{ solves } (\mathcal{P}(e)),\ e = \{t,q\}\}.$$

Lemma 2.9. *There exists a positive constant c_1 such that*

$$c_1 \, \|u\|^2_{2,2,\Omega} \leqslant a_t(u,u) + b_q(u^+,u) \qquad \forall \{u,t,q\} \in W. \tag{28}$$

The constant c_1 does not depend on $\{u,t,q\} \in W$.

Proof. Let us suppose that (28) does not hold. Then one can find a sequence $\{u_n,t_n,q_n\} \subset W$ such that

$$\frac{1}{n} \, \|u_n\|^2_{2,2,\Omega} > a_{t_n}(u_n,u_n) + b_{q_n}(u_n^+,u_n) \geqslant 0 \quad \forall n \geqslant 1. \tag{29}$$

Dividing (29) by $\|u_n\|^2_{2,2,\Omega}$ and passing to the limit for $n \to \infty$ we obtain:

$$\lim_{n\to\infty} a_{t_n}(w_n,w_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} b_{q_n}(w_n^+,w_n) = 0,$$

where $w_n := u_n / \|u_n\|_{2,2,\Omega}$. Clearly $\|w_n\|_{2,2,\Omega} = 1$. Hence there exists a subsequence of $\{w_n\}$ (denoted by the same sequence) and an element $w \in V$ such that $w_n \rightharpoonup w$ in V. Therefore

$$0 = \lim_{n\to\infty} a_{t_n}(w_n,w_n) \geqslant t_0 \liminf_{n\to\infty} |w_n|^2_{2,2,\Omega} \geqslant t_0 |w|^2_{2,2,\Omega} \geqslant 0.$$

Thus $|w|^2_{2,2,\Omega} = 0$, $w \equiv p \in P_0$ and $|w_n|^2_{2,2,\Omega} \to 0$. Therefore $w_n \to p$ in V. From

$$0 = \lim_{n\to\infty} b_{q_n}(w_n^+,w_n^+) \geqslant q_0 \liminf_{n\to\infty} \|w_n^+\|^2_{2,\Omega} = q_0 \|w^+\|^2_{2,\Omega} \geqslant 0,$$

it follows that $w \equiv p \leqslant 0$ in Ω. Since

$$w_n \to p \text{ in } V, \tag{30}$$

due to the compact embedding of $H^2(\Omega)$ into $C(\bar{\Omega})$ we have that $w_n \rightrightarrows p$ in Ω. We know that $\forall n \geqslant 1$, $\exists x_n \in \Omega$ such that $w_n(x_n) > 0$. Without loss of generality we may assume $x_n \to x$ in $\bar{\Omega}$. Then $w_n(x_n) \to p(x) \geqslant 0$. Therefore $p = 0$. But this leads to a contradiction with $\|w_n\|_{2,2,\Omega} = 1$ and (30).

■

The next theorem shows the continuous dependence of $u(e)$ on the design variable e.

Lemma 2.10. (Continuous dependence.) *Let e_n, $e \in U_{ad}$, $e_n \to e$. Further let $u_n := u(e_n) \in V$ be a solution to $(\mathcal{P}(e_n))$ and let (S2) be fulfilled. Then there exists a function $u \in V$ such that*

$$u_n \to u \ \text{ in } V$$

and moreover $u = u(e)$ is a solution to $(\mathcal{P}(e))$.

Proof. Let $\{u(e_n), t_n, q_n\} \in W$. Using the definition of $(\mathcal{P}(e_n))$ and setting $v = u_n$ we have

$$c_1 \|u_n\|^2_{2,2,\Omega} \leqslant a_{t_n}(u_n, u_n) + b_{q_n}(u_n^+, u_n) = F(u_n) \leqslant \|f\|_{2,\Omega} \|u_n\|_{2,2,\Omega},$$

making use of (28). Thus the sequence $\{u_n\}$ is bounded in $H^2(\Omega)$:

$$\|u_n\|_{2,2,\Omega} \leqslant c, \tag{31}$$

where $c > 0$ does not depend on $n \in \mathbb{N}$. Consequently one can pass to a subsequence of $\{u_n\}$ (denoted by the same sequence) such that

$$u_n \rightharpoonup u \quad \text{in } V, \tag{32}$$

for some $u \in V$. In order to show that u solves $(\mathcal{P}(e))$ we pass to the limit for $n \to \infty$ in $(\mathcal{P}(e_n))$

$$a_{t_n}(u_n, v) + b_{q_n}(u_n^+, v) = F(v) \quad \forall v \in V. \tag{33}$$

First of all we will focus on the term $a_{t_n}(u_n, v)$. We employ (27), (31) and (32). It is readily seen that $\lim\limits_{n\to\infty} (a_{t_n}(u_n, v) - a_t(u_n, v)) = 0$ so that

$$\lim_{n\to\infty} a_{t_n}(u_n, v) = \lim_{n\to\infty} (a_{t_n}(u_n, v) - a_t(u_n, v)) + \lim_{n\to\infty} a_t(u_n, v) = a_t(u, v).$$

In the analysis of the term $b_{q_n}(u_n^+, v)$ we make use of (27), (32) and Lemma 8.4. It is easy to see that $\lim\limits_{n\to\infty} (b_{q_n}(u_n^+, v) - b_{q_n}(u^+, v)) = 0$ so that

$$\lim_{n\to\infty} b_{q_n}(u_n^+, v) = \lim_{n\to\infty} (b_{q_n}(u_n^+, v) - b_{q_n}(u^+, v)) + \lim_{n\to\infty} b_{q_n}(u^+, v) = b_q(u^+, v).$$

Thus the limit element $u \in V$ satisfies

$$a_t(u, v) + b_q(u^+, v) = F(v) \quad \forall v \in V, \tag{34}$$

i.e. u solves $(\mathcal{P}(e))$.

Since $u(e)$ is unique, not only the subsequence, but the whole sequence $\{u_n\}$ tends weakly to u in V. Since $u_n \rightharpoonup u$ in V due to the Rellich theorem (Theorem 8.4) we have that $u_n \to u$ in $H^1(\Omega)$. Now it is sufficient to prove the convergence

in the seminorm $|u|_{a_t,\Omega} := \sqrt{a_t(u,u)}$, i.e. $a_t(u_n,u_n) \to a_t(u,u)$ as $n \to \infty$. From (34) and definition of $(\mathcal{P}(e))$, $(\mathcal{P}(e_n))$ it follows that

$$a_{t_n}(u_n,u_n) + b_{q_n}(u_n^+,u_n) = F(u_n) \to F(u) = a_t(u,u) + b_q(u^+,u) \tag{35}$$

as $n \to \infty$. It is not difficult to see that $\lim_{n\to\infty}(b_{q_n}(u_n^+,u_n) - b_q(u^+,u)) = 0$. Therefore $\lim_{n\to\infty}(a_{t_n}(u_n,u_n) - a_t(u,u)) = 0$ and consequently

$$a_t(u_n,u_n) = a_t(u_n,u_n) \pm a_{t_n}(u_n,u_n) \to a_t(u,u), \quad n \to \infty. \tag{36}$$

The assertion of the lemma is now proved. ■

Lemma 2.11. *There exists a constant $c_2 > 0$ such that $\forall\{u_i,t_i,q_i\} \in W, i = 1,2$:*

$$c_2 \|u_1 - u_2\|^2_{2,2,\Omega} \leqslant a_{t_1}(u_1-u_2,u_1-u_2) + b_{q_1}(u_1^+ - u_2^+, u_1-u_2). \tag{37}$$

The constant c_2 does not depend on $\{u_i,t_i,q_i\} \in W, i = 1,2$.

Proof. Assume that (37) does not hold. Then there exist sequences $\{u_{1,n},t_{1,n},q_{1,n}\}$, $\{u_{2,n},t_{2,n},q_{2,n}\} \subset W$ such that

$$\begin{aligned}\frac{1}{n}\|u_{1,n}-u_{2,n}\|^2_{2,2,\Omega} > a_{t_{1,n}}(u_{1,n}-u_{2,n},u_{1,n}-u_{2,n}) + \\ + b_{q_{1,n}}(u_{1,n}^+ - u_{2,n}^+, u_{1,n}-u_{2,n}) \geqslant 0 \quad \forall n \geqslant 1.\end{aligned} \tag{38}$$

According to (31) the sequences $\{u_{1,n}\}$, $\{u_{2,n}\}$ are bounded in $H^2(\Omega)$. Thus one can find its subsequences (denoted by the same sequences) and functions $\hat{u}_1$, $\hat{u}_2$ such that $u_{i,n} \rightharpoonup \hat{u}_i$ in $H^2(\Omega)$, $i = 1,2$. Due to Theorem 8.4 and Theorem 8.3 (see e.g. [27]) it holds $u_{i,n} \rightrightarrows \hat{u}_i$ in $\bar{\Omega}$, $i = 1,2$. By setting $v = 1$ in $(\mathcal{P}(e_{i,n}))$, $i = 1,2$ and passing to the limit for $n \to \infty$ we obtain for $i = 1,2$

$$q_1 \int_\Omega \hat{u}_i^+ \,\mathrm{d}x = q_1 \lim_{n\to\infty}\int_\Omega u_{i,n}^+ \,\mathrm{d}x \geqslant \lim_{n\to\infty}\int_\Omega q_{i,n}u_{i,n}^+ \,\mathrm{d}x = F(1) > 0.$$

Hence, we can find sets M_1, M_2 with the positive one–dimensional Lebesgue measure, such that $u_{i,n} > 0, \hat{u}_i > 0$ in M_i, $i = 1,2$ for n large enough.

Dividing (38) by $\|u_{1,n}-u_{2,n}\|^2_{2,2,\Omega}$ we have

$$a_{t_{1,n}}(w_{1,n}-w_{2,n},w_{1,n}-w_{2,n}) \to 0 \quad \text{and} \quad b_{q_{1,n}}(w_{1,n}^+ - w_{2,n}^+, w_{1,n}-w_{2,n}) \to 0, \tag{39}$$

where $w_{i,n} := u_{i,n}/\|u_{1,n}-u_{2,n}\|_{2,2,\Omega}$, $i = 1,2$. Clearly $\{w_{1,n}-w_{2,n}\}$ is bounded in $H^2(\Omega)$ and $\|w_{1,n}-w_{2,n}\|_{2,2,\Omega} = 1$. Hence there exist subsequences of $\{w_{i,n}\}$, $i = 1,2$ (denoted by the same sequences) and an element $w \in V$ such that $w_{1,n} - w_{2,n} \rightharpoonup w$ in V. Thus

$$0 = \lim_{n\to\infty} a_{t_{1,n}}(w_{1,n}-w_{2,n},w_{1,n}-w_{2,n}) \geqslant t_0 \liminf_{n\to\infty}|w_{1,n}-w_{2,n}|^2_{2,2,\Omega} \geqslant t_0|w|^2_{2,2,\Omega} \geqslant 0.$$

Therefore $|w_{1,n} - w_{2,n}|^2_{2,2,\Omega} \to 0$, $|w|^2_{2,2,\Omega} = 0$, i.e. $w \equiv p \in P_0$ and $w_{1,n} - w_{2,n} \to p$ in $H^2(\Omega)$. Consequently (39) reads

$$w_{1,n} - w_{2,n} \to p \text{ in } H^2(\Omega) \quad \text{and} \quad w^+_{1,n} - w^+_{2,n} \to 0 \text{ in } L^2(\Omega). \tag{40}$$

Firstly consider:

$$\exists c > 0 : \|u_{1,n} - u_{2,n}\|_{2,2,\Omega} \geqslant c \qquad \forall n. \tag{41}$$

Then $\{w_{1,n}\}$, $\{w_{2,n}\}$ are bounded in $H^2(\Omega)$ and there exist subsequences (denoted by the same sequences) converging weakly to $\hat{w}_1$, $\hat{w}_2$ in $H^2(\Omega)$. Hence (40) leads to

$$\hat{w}_1 - \hat{w}_2 = p \quad \text{and} \quad \hat{w}^+_1 - (\hat{w}^+_1 - p) = 0 \text{ a.e. in } \Omega. \tag{42}$$

As $\hat{u}_1 > 0$ in M_1, also $\hat{w}_1 > 0$ in M_1. From this and (42), $p = 0$ a.e. in Ω on the one hand and $\|p\|_{2,2,\Omega} = 1$ on the other hand as follows from (40) and the fact that $\|w_{1,n} - w_{2,n}\|_{2,2,\Omega} = 1$.

If (41) is not satisfied, then $\|u_{1,n} - u_{2,n}\|_{2,2,\Omega} \to 0$. Thus $\hat{u}_1 = \hat{u}_2$ in Ω. Denote by $M_{1,2} \subseteq \Omega$ the subinterval where $\hat{u}_1, u_{i,n}$, $i = 1, 2$ are positive for n large enough. This implies that $w_{i,n} > 0$, $i = 1, 2$ in $M_{1,2}$. Then

$$w_{1,n} - w_{2,n} = w^+_{1,n} - w^+_{2,n} \to 0 \text{ a.e. in } M_{1,2} \quad \text{as} \quad n \to \infty. \tag{43}$$

From (43) and (40) it follows that $p = 0$, being in contradiction with

$$\|w_{1,n} - w_{2,n}\|_{2,2,\Omega} = 1.$$

■

Let us now mention that the optimization problems with the state described by a variational inequality are in general nonsmooth, see e.g. [16], [17], [14]. Our state problem ($\mathcal{P}(e)$) is represented by a nonlinear variational equation, which is very close to problems governed by inequalities. Accordingly we can assume that the problem (P) will be nondifferentiable as well. In Lemma 2.10 the continuity of the mapping $u : e \longmapsto u(e)$ is established and in what follows we shall prove its Lipschitz continuity.

Lemma 2.12. *Let (S2) be satisfied. Then the mapping $u : e \mapsto u(e)$, where $u(e)$ solves ($\mathcal{P}(e)$), is Lipschitz continuous in U_{ad}, i.e. there exists a constant $K_1 > 0$ such that $\forall e_1 = \{t_1, q_1\}, e_2 = \{t_2, q_2\} \in U_{ad}$:*

$$\|u(e_1) - u(e_2)\|_{2,2,\Omega} \leqslant K_1 \left(\|t_1 - t_2\|_{C(\bar{\Omega})} + \|q_1 - q_2\|_{2,\Omega} \right).$$

Proof. Let e_1, $e_2 \in U_{ad}$ and $u_1 := u(e_1)$, $u_2 := u(e_2)$ be solutions of ($\mathcal{P}(e_1)$), ($\mathcal{P}(e_2)$), respectively. Subtracting ($\mathcal{P}'(e_2)$) from ($\mathcal{P}'(e_1)$) we have:

$$a_{t_1}(u_1, v) - a_{t_2}(u_2, v) + b_{q_1}(u^+_1, v) - b_{q_2}(u^+_2, v) = 0 \quad \forall v \in V. \tag{44}$$

Adding and subtracting $a_{t_2}(u_1, v)$, $b_{q_2}(u_1^+, v)$ to the left hand side of (44) yield

$$\begin{aligned} a_{t_2}(u_1 - u_2, v) + b_{q_2}(u_1^+ - u_2^+, v) \; &= \\ &= (a_{t_2} - a_{t_1})(u_1, v) \; + \; (b_{q_2} - b_{q_1})(u_1^+, v) \qquad \forall v \in V. \end{aligned} \tag{45}$$

Inserting $v = u_1 - u_2$ into (45) and using (37) we have that

$$c \, \|u_1 - u_2\|_{2,2,\Omega}^2 \leqslant a_{t_2}(u_1 - u_2, u_1 - u_2) \; + \; b_{q_2}(u_1^+ - u_2^+, u_1 - u_2), \tag{46}$$

where c is a positive constant independent on $\{u_1, t_1, q_1\}, \{u_2, t_2, q_2\} \in W$. The right hand side of (45) can be estimated as follows:

$$(a_{t_2} - a_{t_1})(u_1, u_1 - u_2) \leqslant c \, \|t_1 - t_2\|_{C(\bar{\Omega})} \, \|u_1\|_{2,2,\Omega} \, \|u_1 - u_2\|_{2,2,\Omega} \, , \tag{47}$$

$$(b_{q_2} - b_{q_1})(u_1^+, u_1 - u_2) \leqslant \|q_1 - q_2\|_{2,\Omega} \, \left\|u_1^+\right\|_{1,2,\Omega} \|u_1 - u_2\|_{2,2,\Omega} \, . \tag{48}$$

Therefore the assertion of the lemma is a consequence of (45) - (48) and the uniform boundedness of $u(e)$, $e \in U_{ad}$.

■

To ensure the existence of a solution to (P), it remains to assume the lower semicontinuity of the cost functional I:
(I1) If e, $e_n \in U_{ad}$, $e_n \to e$ in U_{ad} and v, $v_n \in V$, $v_n \to v$ in V, then

$$\liminf_{n \to \infty} I(e_n, \, v_n) \geqslant I(e, \, v).$$

Theorem 2.3. *Let the cost functional I satisfy* (I1)*, then there exists at least one solution of* (P)*.*

Proof. Let us denote $\lambda = \inf_{e \in U_{ad}} I(e, u(e))$. Then there exists a minimization sequence $\{e_n\} \subset U_{ad}$ such that

$$\lambda = \lim_{n \to \infty} I(e_n, u_n(e_n)).$$

The compactness of U_{ad}, proved by Lemma 2.8 implies the existence of a subsequence (denoted by the same sequence) $\{e_n\} \subset U_{ad}$ and an element $e^* \in U_{ad}$ such that $e_n \to e^*$ in U_{ad}. Therefore by making use of Lemma 2.10 we obtain $u_n(e_n) \to u^*(e^*)$ in V, where $u_n(e_n), u^*(e^*)$ solve $(\mathcal{P}(e_n))$ and $(\mathcal{P}(e^*))$, respectively. Due to the lower semicontinuity of the cost functional, we have

$$\lambda = \lim_{n \to \infty} I(e_n, u_n(e_n)) \geqslant I(e^*, u^*(e^*)).$$

Then $I(e^*, u^*(e^*)) = \min_{e \in U_{ad}} I(e, u(e))$ and the assertion of the theorem is proved.

■

In addition, let us suppose that I is Lipschitz continuous in $U_{ad} \times V$:

(I2) There exist a constant $c > 0$ such that $\forall e_1, e_2 \in U_{ad}$ and $\forall v_1, v_2 \in V$:

$$|I(e_1, v_1) - I(e_2, v_2)| \leqslant c \left(\|v_1 - v_2\|_{2,2,\Omega} + \|t_1 - t_2\|_{C(\bar{\Omega})} + \|q_1 - q_2\|_{2,\Omega} \right).$$

Lemma 2.13. *Let I satisfy* (I2). *Then $J(e) := I(e, u(e))$, with $u(e)$ being a solution of $(\mathcal{P}(e))$, is Lipschitz continuous in U_{ad}, i.e. there exists a constant $K_2 > 0$ such that*

$$|J(e_1) - J(e_2)| \leqslant K_2 \left(\|t_1 - t_2\|_{C(\bar{\Omega})} + \|q_1 - q_2\|_{2,\Omega} \right) \qquad \forall e_1, e_2 \in U_{ad}.$$

Proof. The assertion directly follows from (I2) and Lemma 2.12. ■

At the end of this section we shall show that the cost functionals (4), (5) and (6) have the required properties.

Lemma 2.14. *Cost functionals* (4), (5) *and* (6), *with $u(e)$ being a solution to $(\mathcal{P}(e))$, satisfy* (I1) *and* (I2).

Proof. Let $e_1, e_2 \in U_{ad}$ and let $u(e_1), u(e_2) \in V$ be solutions to $(\mathcal{P}(e_1))$ and $(\mathcal{P}(e_2))$, respectively. We start with the cost functional (4). Condition (I1) is a direct consequence of the continuous dependence of $J_1(e) = I_1(e, u(e))$ on u. Using the Cauchy–Schwarz inequality, it directly reads

$$|J_1(e_1) - J_1(e_2)| \leqslant \|f\|_{2,\Omega} \, \|u(e_1) - u(e_2)\|_{2,2,\Omega}.$$

Therefore $J_1(e) = I_1(e, u(e))$ also satisfies (I2).

Let us now continue with the cost functional (5). The lower semicontinuity again follows from the continuity of $J_2(e) = I_2(e, u(e))$. By using the Cauchy–Schwarz inequality and boundedness of solution to $(\mathcal{P}(e))$ we obtain

$$\begin{aligned} |J_2(e_1) - J_2(e_2)| &\leqslant \|u(e_1) + u(e_2)\|_{2,2,\Omega} \, \|u(e_1) - u(e_2)\|_{2,2,\Omega} \leqslant \\ &\leqslant (\|u(e_1)\|_{2,2,\Omega} + \|u(e_2)\|_{2,2,\Omega}) \, \|u(e_1) - u(e_2)\|_{2,2,\Omega} \leqslant \\ &\leqslant c \, \|u(e_1) - u(e_2)\|_{2,2,\Omega}, \end{aligned}$$

where c is a positive constant which does not depend on e_1, e_2. Thus $J_2(e) = I_2(e, u(e))$ satisfies (I2).

Finally we can approach to the functional (6). Let t, $t_n \in U_{ad}^t$, $t_n \rightrightarrows t$ in Ω, then $u(e_n) \to u(e)$ in V. Therefore

$$\liminf_{n\to\infty} I_3(e_n, u_n) = \liminf_{n\to\infty} \int_\Omega t_n^2 (u_n'')^2 \, \mathrm{d}x = \int_\Omega t^2 (u'')^2 \, \mathrm{d}x = I_3(e, u).$$

By using the Cauchy–Schwarz inequality and boundedness of solution to $(\mathcal{P}(e))$ we have

$$\begin{aligned}|J_3(e_1) - J_3(e_2)| &\leqslant t_1^2 \, \|u(e_1) + u(e_2)\|_{2,2,\Omega} \, \|u(e_1) - u(e_2)\|_{2,2,\Omega} + \\ &+ c \, \|t_1 - t_2\|_{C(\bar{\Omega})} \, \|u(e_2)\|_{2,2,\Omega} \leqslant \\ &\leqslant t_1^2 (\|u(e_1)\|_{2,2,\Omega} + \|u(e_2)\|_{2,2,\Omega}) \, \|u(e_1) - u(e_2)\|_{2,2,\Omega} + \\ &+ c \, \|t_1 - t_2\|_{C(\bar{\Omega})} \, \|u(e_2)\|_{2,2,\Omega} \leqslant \\ &\leqslant c \Big(\, \|u(e_1) - u(e_2)\|_{2,2,\Omega} \, + \, \|t_1 - t_2\|_{C(\bar{\Omega})} \, \Big),\end{aligned}$$

where c, c_1 are positive constants which do not depend on e_1, e_2. Thus $J_3(e) = I_3(e, u(e))$ satisfies (I2).

■

2.2. Approximation of (P)

Optimization problem (P) in the continuous form, as it is introduced in the previous section, is not suitable for numerical realization. In this section we will pay attention to the approximation of the state problem ($\mathcal{P}(e)$) and corresponding design optimization problem (P). We will study the existence of a solution to (P_h) as well as we shall prove that discrete (approximated) problems are close to the original problem in the sense of subsequences. In other words, discrete problems approximate the original problem well if we let the approximation parameter converge to zero.

2.2.1. Approximation of U_{ad}

Firstly we will be aimed at the approximation of design variables, i.e. approximation of the set $U_{ad} = U_{ad}^t \times U_{ad}^q$. We define a partition of the interval Ω into subintervals $K_i = [x_{i-1}, x_i]$, where the nodes satisfy

$$0 = x_0 < x_1 < ... < x_n = l. \tag{49}$$

Without loss of generality we will restrict ourselves to equidistant partition, i.e. $x_i - x_{i-1} = h,\ h > 0,\ l = nh,\ x_i = ih,\ \forall i = 0, 1, \dots, n$. The admissible set U_{ad}^t is approximated by Lipschitz continuous and piecewise linear functions. Similarly we approximate the set U_{ad}^q by piecewise constant functions, i.e., we define

$$U_{ad,h}^t = \left\{ t_h \in C^{0,1}(\Omega) : t_h|_{K_i} \in P_1(K_i),\ \forall i = 1, \dots, n \right\} \cap U_{ad}^t, \tag{50}$$

$$U_{ad,h}^q = \left\{ q_h \in L^2(\Omega) : q_h|_{K_i} \in P_0(K_i),\ \forall i = 1, \dots, n \right\} \cap U_{ad}^q. \tag{51}$$

Approximation of the set U_{ad} then has the following form:

$$U_{ad}^h = U_{ad,h}^t \times U_{ad,h}^q.$$

The set U_{ad}^h is an inner approximation of U_{ad}, i.e. $U_{ad}^h \subset U_{ad}$.

Lemma 2.15. *The set U_{ad}^h is a compact subset of $C(\bar{\Omega}) \times L^2(\Omega)$ with regard to the convergence defined by (27).*

Proof. The assertion follows from the fact that every closed subset of a compact set is also compact.

■

Remark 2.1. *Instead of piecewise linear approximation of the thickness t we can consider piecewise constant functions, the so called stepped beam. Such an approximation is no longer represented by continuous functions and the set $U_{ad,h}$ is not a subset of U_{ad}. However, for a stepped beam it is possible to reach similar convergence results as for the continuous case (see e.g. [17]).*

2.2.2. Approximation of $(\mathcal{P}(e))$

Now we can approach to the approximation of the state problem. We use the finite element method with the partition (49). Assume that $\{t_h,\, q_h\} = e_h \in U_{ad}^h$. We define the following finite dimensional approximations of V as it is usual for beam problems:

$$V_h = \left\{v_h \in C^1(\bar{\Omega}) :\, v_h\,|_{K_i} \in P_3(K_i),\, \forall i = 1, ..., n,\; v_h'(0) = 0\right\}$$

Space $V_h \subset V$ contains of piecewise cubic polynomials that are continuous together with their first derivatives in Ω. These polynomials satisfy the same natural boundary condition as functions from V. Using the classical Ritz method we approximate $(\mathcal{P}(e))$ as follows:

$$\text{Find } u_h \in V_h :\;\; \mathcal{E}_{e_h}(u_h) \leqslant \mathcal{E}_{e_h}(v_h) \quad \forall v_h \in V_h, \qquad (\mathcal{P}_h(e_h))$$

where $\mathcal{E}_{e_h}(v_h) = \frac{1}{2}(a_{t_h}(v_h,\, v_h) + b_{q_h}(v_h^+, v_h^+)) - F(v_h)$. The integrand defining the form a_{t_h} is piecewise polynomial of order 5 at most, therefore we can evaluate the corresponding integral exactly and no additional approximation of a_{t_h} is needed. Due to the nonlinear term v_h^+ we can not evaluate the form $b_{q_h}(v_h^+, v_h^+)$ exactly. The same issue occurs in the case of the continuous linear functional $F(v_h)$, $v_h \in V_h$ because $f \in L^2(\Omega)$ is a general function that is defined everywhere in Ω.

Therefore terms b_{q_h} and F will be approximated using the numerical quadrature (176). Let Φ_i, $i = 1, \ldots, n$ be a transformation of the interval K_i onto the reference interval $[-1, 1]$ defined by (178) with $s = x_{i-1}$, $t = x_i$. For the sake of simplicity the generalized forces F_i, M_j will no longer be considered. Further let $z_{j,i} = \Phi_i^{-1}(\hat{z}_j)$ a $\omega_j = (h/2)\hat{\omega}_j$. Approximations then have the following form:

$$b_{q_h}^h(u, v) := \sum_{i=1}^{n} \left(q_{h,i} \sum_{j=1}^{m} \omega_j u(z_{j,i}) v(z_{j,i}) \right), \tag{52}$$

$$F^h(v) := \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j f(z_{j,i}) v(z_{j,i}), \tag{53}$$

where $q_{h,i} = q_h|_{K_i}$. For arbitrary $e_h \in U_{ad}$ these terms are defined on $H^1(\Omega)$ and associated with the reference quadrature formula (176) and the partition (49). We will adopt the notation $b_{q_h}^h \in \mathcal{Q}_h^k$, $F^h \in \mathcal{Q}_h^k$ to express the fact that the used formula is exact for polynomials of degree k at least. In addition let us denote by

$$Q_h = \{z_{j,i} = \Phi_i^{-1}(\hat{z}_j) \in \Omega,\, j = 1, \ldots, m,\, i = 1, \ldots, n\}$$

the set containing all nodes of the quadrature formula on Ω. The approximated state problem then reads as follows:

$$\text{Find } u_h \in V_h :\;\; \mathcal{E}_{e_h}^h(u_h) \leqslant \mathcal{E}_{e_h}^h(v_h) \quad \forall v_h \in V_h, \qquad (\mathcal{P}_h(e_h))$$

where $\mathcal{E}_{e_h}^h(v_h) = \frac{1}{2}(a_{t_h}(v_h,\, v_h) + b_{q_h}^h(v_h^+, v_h^+)) - F^h(v_h)$.

2.2.3. Approximation of the cost functional and the optimization problem

Let $I_h : U_{ad}^h \times V_h \to \mathbb{R}^1$ be an approximation of I and denote $J_h(e_h) = I_h(e_h, u_h(e_h))$ with $u_h(e_h)$ being a solution to $(\mathcal{P}_h(e_h))$. The approximation of (P) then reads as follows:

$$\text{Find } e_h^* \in U_{ad}^h : \quad J_h(e_h^*) \leqslant J_h(e_h) \quad \forall e_h \in U_{ad}^h. \tag{P_h}$$

Cost functionals (4), (5) and (6) have an integral form. Therefore we can also use the formula for numerical integration for their approximation.

$$J_{h,1}(e_h) \equiv I_{h,1}(e_h, u_h(e_h)) = \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j f(z_{j,i}) u_h(z_{j,i}), \tag{54}$$

$$J_{h,2}(e_h) \equiv I_{h,2}(e_h, u_h(e_h)) = \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j u_h^2(z_{j,i}), \tag{55}$$

$$J_{h,3}(e_h) \equiv I_{h,3}(e_h, u_h(e_h)) = \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j t^2(z_{j,i}) (u_h''(z_{j,i}))^2, \tag{56}$$

where $u_h(e_h)$ solves $(\mathcal{P}_h(e_h))$.

2.2.4. Existence and uniqueness of a solution to $(\mathcal{P}_h(e_h))$

This subsection will be devoted to the existence analysis of $(\mathcal{P}_h(e_h))$ for $h > 0$ fixed. We will proceed in a similar way as for the continuous problem (see [52]), i.e. we define orthogonal decomposition of V_h, prove the coercivity of $\mathcal{E}_{e_h}^h$ on V_h using a modified Poincaré inequality and suitable assumption on the beam load. In the sequel we will assume that $f \in W^{1,1}(\bar{K}_i)$, $i = 1, \ldots, n$, $e_h \in U_{ad}^h$, $b_{q_h}^h \in \mathcal{Q}_h^0$ and $F^h \in \mathcal{Q}_h^0$. Firstly we prove the uniform boundedness of $b_{q_h}^h$ and F^h.

Lemma 2.16. *There exist positive constants c_1, c_2 such that*

$$\left|b_{q_h}^h(u^+, v)\right| \leqslant c_1 \|u\|_{2,2,\Omega} \|v\|_{2,2,\Omega} \qquad \forall u, v \in H^2(\Omega),\ \forall e_h \in U_{ad}^h, \tag{57}$$

$$|F^h(v)| \leqslant c_2 \|v\|_{2,2,\Omega} \qquad \forall v \in H^2(\Omega). \tag{58}$$

Constants c_1, c_2 depend only on the length of the interval Ω, definition of the set U_{ad}^h and the load f.

Proof. Let $b^h_{q_h} \in \mathcal{Q}^0_h$, then

$$\left|b^h_{q_h}(u,v)\right| = \left|\sum_{i=1}^{n}\left(q_{h,i}\sum_{j=1}^{m}\omega_j u^+(z_{j,i})v(z_{j,i})\right)\right| \leqslant q_1\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j\left|u^+(z_{j,i})\right|\left|v(z_{j,i})\right| \leqslant$$

$$\leqslant q_1\left\|u^+\right\|_{C(\bar\Omega)}\|v\|_{C(\bar\Omega)}\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j = q_1\left\|u^+\right\|_{C(\bar\Omega)}\|v\|_{C(\bar\Omega)}\int_\Omega 1\,\mathrm{d}x =$$

$$= lq_1\left\|u^+\right\|_{C(\bar\Omega)}\|v\|_{C(\bar\Omega)} \leqslant c_1\left\|u^+\right\|_{1,2,\Omega}\|v\|_{2,2,\Omega} \leqslant c_1\|u\|_{2,2,\Omega}\|v\|_{2,2,\Omega}$$

$$\forall u,v \in H^2(\Omega),\ \forall e_h \in U^h_{ad}.$$

Next we can pass to the proof of the estimate (58). Let $F^h \in \mathcal{Q}^0_h$, then

$$\left|F^h(v)\right| = \left|\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j f(z_{j,i})v(z_{j,i})\right| \leqslant \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j\left|f(z_{j,i})\right|\left|v(z_{j,i})\right| \leqslant$$

$$\leqslant c\|v\|_{C(\bar\Omega)}\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j = lc\|v\|_{C(\bar\Omega)} \leqslant c_2\|v\|_{2,2,\Omega} \quad \forall v \in H^2(\Omega),$$

where $c := \max\limits_i \|f\|_{C(\bar K_i)}$. In both estimates we used the definition of U^h_{ad}, the compactness of embedding of $H^1(\Omega)$ into $C(\bar\Omega)$ (see Theorem 8.3) and the compactness of embedding of $H^2(\Omega)$ into $H^1(\Omega)$ (see Remark 8.1). It is easy to see that values $u^+(x), v(x)$ are defined correctly $\forall x \in \bar\Omega$ and $\forall v \in H^2(\Omega)$.

■

Lemma 2.17. *There exist positive constants c_1, c_2 such that*

$$\left|b^h_{q_h}(u^+_h,v_h) - b_{q_h}(u^+_h,v_h)\right| \leqslant c_1 h\|u_h\|_{2,2,\Omega}\|v_h\|_{2,2,\Omega} \quad \forall u_h,v_h \in V_h, \tag{59}$$

$$\left|F^h(v_h) - F(v_h)\right| \leqslant c_2 h\|v_h\|_{2,2,\Omega} \quad \forall v_h \in V_h. \tag{60}$$

Constant c_1, c_2 depend only on the length of the interval Ω, definition of U^h_{ad} and the load f.

Proof. Let $b^h_{q_h} \in \mathcal{Q}^0_h$. It is known that $u_h, v_h \in H^2(\Omega)$ implies $(u^+_h v_h)|_{K_i} \in W^{1,1}(K_i)$. Therefore by Lemma 8.5, Cauchy–Schwarz inequality and compact embedding of $H^2(\Omega)$ into $H^1(\Omega)$ we have

$$\left|b^h_{q_h}(u^+_h,v_h) - b_{q_h}(u^+_h,v_h)\right| \leqslant q_1\sum_{i=1}^{n}\left|\sum_{j=1}^{m}\omega_j u^+_h(z_{j,i})v_h(z_{j,i}) - \int_{K_i}u^+_h v_h\,\mathrm{d}x\right| \leqslant$$

$$\leqslant ch\sum_{i=1}^{n}\left|u^+_h v_h\right|_{1,1,K_i} \leqslant ch\sum_{i=1}^{n}\left(\left\|u^+_h\right\|_{2,K_i}|v_h|_{1,2,K_i} + \left|u^+_h\right|_{1,2,K_i}\|v_h\|_{2,K_i}\right) \leqslant$$

$$\leqslant ch\sum_{i=1}^{n}\left\|u^+_h\right\|_{1,2,K_i}\|v_h\|_{1,2,K_i} \leqslant ch\sum_{i=1}^{n}\left\|u^+_h\right\|_{1,2,K_i}\sum_{i=1}^{n}\|v_h\|_{1,2,K_i} \leqslant$$

$$\leqslant ch\left\|u^+_h\right\|_{1,2,\Omega}\|v_h\|_{1,2,\Omega} \leqslant c_1 h\|u_h\|_{2,2,\Omega}\|v_h\|_{2,2,\Omega} \quad \forall u_h,v_h \in V_h,\ \forall e_h \in U_{ad}.$$

Next we will prove the relation (60). Let $F^h \in \mathcal{Q}_h^0$. In view of $f \in W^{1,1}(\bar{K}_i)$, $i = 1, \ldots, n$, $v_h \in V_h$ it holds that $(fv_h)|_{K_i} \in W^{1,1}(K_i)$, $i = 1, \ldots, n$. Then we have

$$\begin{aligned}
\left|F^h(v_h) - F(v_h)\right| &\leqslant \sum_{i=1}^{n} \left| \sum_{j=1}^{m} \omega_j f(z_{j,i}) v_h(z_{j,i}) - \int_{K_i} f v_h \,\mathrm{d}x \right| \leqslant \\
&\leqslant ch \sum_{i=1}^{n} |fv_h|_{1,1,K_i} \leqslant ch \,\|v_h\|_{C^1(\bar{K}_i)} \sum_{i=1}^{n} \|f\|_{1,2,K_i} \leqslant \\
&\leqslant ch \,\|v_h\|_{2,2,\Omega} \,\|f\|_{1,2,\Omega} \leqslant c_2 h \,\|v_h\|_{2,2,\Omega} \qquad \forall v \in V_h.
\end{aligned}$$

■

Lemma 2.18. *The functional $\mathcal{E}^h_{e_h}$ is Gâteaux differentiable and convex on $H^2(\Omega)$. Its Gâteaux derivative in arbitrary point $u \in H^2(\Omega)$ and arbitrary direction $v \in H^2(\Omega)$ has the following form:*

$$\mathcal{E}^{h\,'}_{e_h}(u;\, v) = a_{t_h}(u,\, v) + b^h_{q_h}(u^+, v) - F^h(v) \quad \forall u, v \in H^2(\Omega),\ \forall e_h \in U^h_{ad}. \tag{61}$$

Proof. We proceed analogically as in the proof of Lemma 2.2. We have

$$\lim_{\epsilon \to 0} \frac{b^h_{q_h}((u+\epsilon v)^+, (u+\epsilon v)^+) \,-\, b^h_{q_h}(u^+, u^+)}{\epsilon} \;=\; 2b^h_{q_h}(u^+, v) \tag{62}$$

$\forall u, v \in H^2(\Omega)$, $\forall e_h \in U^h_{ad}$.

The convexity follows from

$$\begin{aligned}
\mathcal{E}^{h\,'}_{e_h}(u;\, u-v) \,-\, \mathcal{E}^{h\,'}_{e_h}(v;\, u-v) &= a_{t_h}(u-v,\, u-v) \,+\, b^h_{q_h}(u^+ - v^+,\, u-v) \geqslant \\
&\geqslant a_{t_h}(u-v,\, u-v) \,+\, b^h_{q_h}(u^+ - v^+,\, u^+ - v^+) \geqslant \\
&\geqslant \beta_0 t_0^3 \;|u-v|^2_{2,2,\Omega} \,+ \\
&+ q_0 \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j (u^+(z_{j,i}) - v^+(z_{j,i}))^2 \;\geqslant\; 0 \\
&\quad \forall u, v \in H^2(\Omega),\ \forall e_h \in U^h_{ad}.
\end{aligned}$$

■

Lemma 2.18 enable us to introduce the equivalent weak formulation of the problem $(\mathcal{P}_h(e_h))$.

$$\text{Find } u_h \in V_h :\; a_{t_h}(u_h,\, v_h) + b^h_{q_h}(u_h^+, v_h) \;=\; F^h(v_h) \quad \forall v_h \in V_h. \qquad (\mathcal{P}'_h(e_h))$$

It remains to prove the coercivity of $\mathcal{E}^h_{e_h}$ on V_h. We make decomposition of V_h into a convex cone of rigid displacements and its negative polar cone.

$$\mathcal{R}_{V_h} \;=\; \{v_h \in V_h \cap P_1 : a_{t_h}(v_h, v_h) + b^h_{q_h}(v_h^+, v_h) = 0\} \;=\; \{p \in P_0 : p \leqslant 0\}.$$

From there by the definition of the following scalar product on $H^2(\Omega)$

$$((u,v))_{2,\Omega} = \int_\Omega u''v''\,\mathrm{d}x + \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j u(z_{j,i})v(z_{j,i}), \tag{63}$$

the negative polar cone $\mathcal{R}^{\ominus}_{V_h}$ reads as follows:

$$\begin{aligned}\mathcal{R}^{\ominus}_{V_h} &= \{v_h \in V_h : ((v_h,p))_{2,\Omega} \leqslant 0 \ \ \forall p \in \mathcal{R}_{V_h}\} = \\ &= \{v_h \in V_h : \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j v_h(z_{j,i}) \geqslant 0\}.\end{aligned}$$

Lemma 2.19. (Necessary condition for the existence of a solution to $(\mathcal{P}_h(e_h))$.) *Let there exist a solution to $(\mathcal{P}_h(e_h))$, then the condition*

$$F^h(1) \geqslant 0 \tag{S1$_h$}$$

must be satisfied.

Proof. The assertion can be obtained by inserting $v = p \in \mathcal{R}_{V_h}$ into $(\mathcal{P}'_h(e_h))$. ■

The space V_h can be uniquely decomposed into the orthogonal sum $\mathcal{R}_{V_h} \oplus \mathcal{R}^{\ominus}_{V_h}$. In addition $\forall v_h \in V_h \ \exists!\ \{p, \bar{v}_h\} \in \mathcal{R}_{V_h} \times \mathcal{R}^{\ominus}_{V_h}$ such that

$$v_h \ = \ p \oplus \bar{v}_h, \quad ((p,\bar{v}_h))_{2,\Omega} \ = \ p\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j \bar{v}_h(z_{j,i}) \ = \ 0. \tag{64}$$

In view of definitions of $\mathcal{R}_{V_h}$, $\mathcal{R}^{\ominus}_{V_h}$ and properties of the decomposition (64), only one of the following variants can occur :

$$p \ = \ 0 \quad \text{and} \quad \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j \bar{v}_h(z_{j,i}) \ \geqslant \ 0, \tag{A1$_h$}$$

$$p \ \leqslant \ 0 \quad \text{and} \quad \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j \bar{v}_h(z_{j,i}) \ = \ 0. \tag{A2$_h$}$$

Lemma 2.20. (Poincaré type inequality). *Let $V = \{v \in H^2(\Omega) : v'(0) = 0\}$, then there exists a positive constant c_P dependent only on the interval Ω such that*

$$\|v\|^2_{2,2,\Omega} \leqslant c_P\left(|v|^2_{2,2,\Omega} + \left(\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j v(z_{j,i})\right)^2\right) \quad \forall v \in V, \tag{65}$$

where $\omega_j > 0$ are wages and $z_{j,i} \in Q_h$ nodes of the integration formula.

Proof. Let (65) do not hold. Then there exists a sequence $\{v_k\} \subset V$ such that

$$\frac{1}{k} \|v_k\|^2_{2,2,\Omega} > |v_k|^2_{2,2,\Omega} + \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j v_k(z_{j,i}) \right)^2 \geqslant 0 \quad \forall k \geqslant 1. \tag{66}$$

Divide (66) by $\|v_k\|^2_{2,2,\Omega}$ and pass to the limit for $k \to \infty$. Then

$$\lim_{k\to\infty} |w_k|^2_{2,2,\Omega} = 0 \quad \text{and} \quad \lim_{k\to\infty} \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j w_k(z_{j,i}) \right)^2 = 0, \tag{67}$$

where $w_k := v_k / \|v_k\|_{2,2,\Omega}$. Clearly $\|w_k\|_{2,2,\Omega} = 1$ and we can find its subsequence (denoted by the same sequence) and an element $w \in V$ such that $w_k \rightharpoonup w$ in V. Due to (67) it holds that $w_k \to w$ in V and moreover $w_k \rightrightarrows w$ in $\bar{\Omega}$. Thus

$$0 = \liminf_{k\to\infty} |w_k|^2_{2,2,\Omega} = |w|^2_{2,2,\Omega} \geqslant 0.$$

Then $|w|^2_{2,2,\Omega} = 0$ and $w \equiv p \in P_0$. From

$$0 = \liminf_{k\to\infty} \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j w_k(z_{j,i}) \right)^2 = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j w(z_{j,i}) \right)^2 \geqslant 0,$$

it follows that $p = 0$ being in contradiction with $\|w_k\|_{2,2,\Omega} = 1$ and $w_k \to p$ in V. ■

Lemma 2.21. *Let the condition*

$$F^h(1) > 0 \tag{S2$_h$}$$

be fulfilled. Then the functional $\mathcal{E}^h_{e_h}$ is coercive on V_h.

Proof. Let (S2$_h$) hold. Firstly consider the alternative (A1$_h$), then $p = 0$, $v_h \equiv \bar{v}_h$ and $\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \bar{v}_h(z_{j,i}) \geqslant 0$. For every function $\bar{v}_h \in \mathcal{R}^{\ominus}_{V_h}$ it holds the following inequality:

$$l \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \left(\bar{v}^+_h(z_{j,i}) \right)^2 \geqslant \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \bar{v}^+_h(z_{j,i}) \right)^2 \geqslant \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \bar{v}_h(z_{j,i}) \right)^2, \tag{68}$$

where we used $b^h_{q_h} \in \mathcal{Q}^0_h$ and the discrete form of the Cauchy-Schwarz inequality (see Theorem 8.2). By using $F^h \in \mathcal{Q}^0_h$, (68) and (65) we can rewrite the functional $\mathcal{E}^h_{e_h}$ as follows:

$$\begin{aligned} 2\mathcal{E}^h_{e_h}(v_h) = 2\mathcal{E}^h_{e_h}(\bar{v}_h) &= a_{t_h}(\bar{v}_h, \bar{v}_h) + b^h_{q_h}(\bar{v}^+_h, \bar{v}^+_h) - 2F^h(\bar{v}_h) \geqslant \\ &\geqslant \beta_0 t_0^3 \, |\bar{v}_h|^2_{2,2,\Omega} + \frac{q_0}{l} \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \bar{v}_h(z_{j,i}) \right)^2 - 2F^h(\bar{v}_h) \geqslant \\ &\geqslant \|\bar{v}_h\|_{2,2,\Omega} \left(c_1 \|\bar{v}_h\|_{2,2,\Omega} - 2c_2 \right), \end{aligned}$$

where $c_1 := (1/c_P)\min\{\beta_0 t_0^3,\, q_0/l\}$ and c_2 is the constant from Lemma 2.16.

If the alternative ($A2_h$) occurs, we have $\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j \bar{v}_h(z_{j,i}) = 0$ and $p \leqslant 0$. Then we obtain:

$$\begin{aligned}
2\mathcal{E}^h_{e_h}(v_h) = 2\mathcal{E}^h_{e_h}(p+\bar{v}_h) &= a_{t_h}(\bar{v}_h,\bar{v}_h) + b^h_{q_h}(v_h^+, v_h^+) - 2F^h(p) - 2F^h(\bar{v}_h) \geqslant \\
&\geqslant \beta_0 t_0^3\, |\bar{v}_h|^2_{2,2,\Omega} + q_0 \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j \left(v_h^+(z_{j,i})\right)^2 + 2|p|F^h(1) - 2F^h(\bar{v}_h) \geqslant \\
&\geqslant \beta_0 t_0^3\, |\bar{v}_h|^2_{2,2,\Omega} + \left(\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j \bar{v}_h(z_{j,i})\right)^2 + 2|p|F^h(1) - 2F^h(\bar{v}_h) \geqslant \\
&\geqslant c_1\, \|\bar{v}_h\|^2_{2,2,\Omega} + 2|p|F^h(1) - 2c_2\, \|\bar{v}_h\|_{2,2,\Omega},
\end{aligned}$$

denoting $c_1 := (1/c_P)\min\{\beta_0 t_0^3,\, 1\}$. Due to the orthogonality of (64) it holds that $\|v_h\|^2_{2,2,\Omega} = \|\bar{v}_h\|^2_{2,2,\Omega} + \|p\|^2_{2,2,\Omega}$. Therefore $\|v_h\|_{2,2,\Omega} \to +\infty$ causes that at least one part of $v_h = p + \bar{v}_h$ converges to $+\infty$ in appropriate norm. By using ($S2_h$) we arrive at the assertion of the lemma.

■

Theorem 2.4. (Necessary and sufficient condition for the existence and uniqueness of a solution to $(\mathcal{P}_h(e_h))$.) *There exists a unique solution to $(\mathcal{P}_h(e_h))$ if and only if ($S2_h$) is fulfilled. In addition for such a solution $u_h \in V_h$ it holds*

$$M_{u_h} = \{z_{j,i} \in Q_h :\, u_h(z_{j,i}) > 0\} \neq \emptyset. \qquad (M1_h)$$

Proof. *Necessity.* Assume that $u_h \in V_h$ is a unique solution to $(\mathcal{P}_h(e_h))$ and ($S2_h$) does not hold. Due to Lemma 2.19 it follows that $F^h(1) = 0$. Inserting $v_h \equiv p \in \mathcal{R}_{V_h}$ into $(\mathcal{P}_h'(e_h))$ yields

$$b^h_{q_h}(u_h^+, p) = 0 \quad \forall p \in \mathcal{R}_{V_h},\, p \neq 0. \qquad (69)$$

Then (69) implies $u_h^+(z_{j,i}) = 0,\ u_h(z_{j,i}) \leqslant 0\ \forall z_{j,i} \in Q_h$. Consequently $u_h(z_{j,i}) + p < 0\ \forall z_{j,i} \in Q_h$ and $\forall p \in \mathcal{R}_{V_h},\ p \neq 0$. From there $b^h_{q_h}((u_h+p)^+, v_h) = 0\ \forall p \in \mathcal{R}_{V_h},\ p \neq 0$ and $\forall v_h \in V_h$. Then it is not difficult to see that $u_h + p$ is another solution to $(\mathcal{P}_h(e_h))$, what is in contradiction with the uniqueness of u_h. The condition ($S2_h$) must be satisfied.
Sufficiency. Let ($S2_h$) be fulfilled. We know that $\mathcal{E}^h_{e_h}$ is Gâteaux differentiable, convex and coercive on V_h, therefore the existence of a solution $u_h \in V_h$ is ensured, see e.g. [12], [8].

Next we prove ($M1_h$). Let $u_h \in V_h$, $u_h(z_{j,i}) \leqslant 0\ \forall z_{j,i} \in Q_h$, solve $(\mathcal{P}_h(e_h))$. Then by inserting $v_h \equiv p \in \mathcal{R}_{V_h},\ p \neq 0$ into $(\mathcal{P}_h'(e_h))$ we obtain

$$0 = b^h_{q_h}(u_h^+, 1) = F^h(1). \qquad (70)$$

But (70) is in contradiction with ($S2_h$) and the solution u_h must satisfy ($M1_h$).

In the rest of the proof we show the uniqueness of the solution. Let $u_{h,1},\ u_{h,2} \in V_h$ be solutions to $(\mathcal{P}_h(e_h))$. Subtracting corresponding weak formulations and setting $v = u_{h,1} - u_{h,2}$ yield

$$a_{t_h}(u_{h,1} - u_{h,2}, u_{h,1} - u_{h,2}) \ + \ b^h_{q_h}(u^+_{h,1} - u^+_{h,2}, u_{h,1} - u_{h,2}) \ = \ 0.$$

Therefore

$$u_{h,1} - u_{h,2} \ = \ p \in P_0 \quad \text{and} \quad u^+_{h,1}(z_{j,i}) - (u_{h,1}(z_{j,i}) - p)^+ \ = \ 0 \ \forall z_{j,i} \in Q_h.$$

Taking into account ($M1_h$) we obtain $p = 0$ and $u_{h,1} = u_{h,2}$ in Ω.

■

Remark 2.2. *Notice that satisfying of the condition $F(1)$ generally does not directly imply satisfying of the discrete condition $F^h(1)$. One must pay attention to the choice of the discretization parameter h, it should be small enough, such that the numerical quadrature is able to evaluate the condition as exact as possible.*

2.2.5. Existence of solutions to (P_h)

The next part of the thesis will be devoted to the existence analysis of (P_h). Analogically as in the continuous case it is possible to prove that $u_h(e_h)$ depends continuously on e_h and that the approximated optimization problem (P_h) has at least one solution.

Assume that $\beta \in L^\infty(\Omega)$, $0 < \beta_0 \leqslant \beta(x)$ a.e. in Ω and $F^h \in V_h^*$, $F^h(1) > 0$. Then for arbitrary $e_h \in U^h_{ad}$ there exists a unique solution of $(\mathcal{P}_h(e_h))$ with the property ($M1_h$). A set of all such solutions will be denoted by W_h. Recall the notation $e_h = \{t_h, q_h\}$.

$$W_h := \{\{u_h, t_h, q_h\} \in V_h \times U^t_{ad,h} \times U^q_{ad,h} : \ u_h \text{ solves } (\mathcal{P}_h(e_h))\}.$$

In the next lemma we will consider a whole class of problems $(\mathcal{P}_h(e_h))$ for $0 < h \leqslant h_0$ in order to use it also in the convergence analysis. Therefore we suppose that $F^h \in V_h^*$, $F^h(1) > 0$ for all $0 < h \leqslant h_0$. It implies that there exists a solution of $(\mathcal{P}_h(e_h))$ for all $e_h \in \bigcup\limits_{0<h\leqslant h_0} U^h_{ad}$.

Lemma 2.22. *There exists a positive constant c_1 such that*

$$c_1 \ \|u_h\|^2_{2,2,\Omega} \ \leqslant \ a_{t_h}(u_h, u_h) \ + \ b^h_{q_h}(u^+_h, u_h) \qquad \forall \{u_h, t_h, q_h\} \in \bigcup_{0<h\leqslant h_0} W_h, \tag{71}$$

where the constant c_1 does not depend on $\{u_h, t_h, q_h\} \in \bigcup\limits_{0<h\leqslant h_0} W_h$.

Proof. Let us suppose that (71) does not hold. Then there exists a sequence $\{u_{h_k}, t_{h_k}, q_{h_k}\} \subset \bigcup_{0<h\leqslant h_0} W_h$ such that

$$\frac{1}{k}\,\|u_{h_k}\|^2_{2,2,\Omega} \;>\; a_{t_{h_k}}(u_{h_k}, u_{h_k}) \;+\; b^{h_k}_{q_{h_k}}(u^+_{h_k}, u_{h_k}) \;\geqslant\; 0 \quad \forall k \geqslant 1. \tag{72}$$

Dividing (72) by $\|u_{h_k}\|^2_{2,2,\Omega}$ and passing to the limit for $k \to \infty$ lead to

$$\lim_{k\to\infty} a_{t_{h_k}}(w_{h_k}, w_{h_k}) = 0 \quad \text{and} \quad \lim_{k\to\infty} b^{h_k}_{q_{h_k}}(w^+_{h_k}, w_{h_k}) = 0,$$

where $w_{h_k} := u_{h_k}/\,\|u_{h_k}\|_{2,2,\Omega}$. Clearly $\|w_{h_k}\|_{2,2,\Omega} = 1$ and $\{w_{h_k}\}$ is bounded. Hence one can find a subsequence of $\{w_{h_k}\}$(denoted by the same sequence) and an element $w_h \in V$ such that $w_{h_k} \rightharpoonup w_h$ in V. Without loss of generality we may suppose that $q_{h_k} \to q_h$ and $h_k \to h$. It is not difficult to prove that $|w_h|^2_{2,2,\Omega} = 0$ and $w_h \equiv p \in P_0$ and $w_{h_k} \to p$ in V, see [52]. Due to the compact embedding of $H^1(\Omega)$ into $C(\bar{\Omega})$ it holds $w_{h_k} \rightrightarrows p$ in Ω. Then

$$0 \leqslant b^h_{q_0}(p^+, p^+) = \lim_{k\to\infty} b^{h_k}_{q_0}(w^+_{h_k}, w^+_{h_k}) \leqslant \lim_{k\to\infty} b^{h_k}_{q_{h_k}}(w^+_{h_k}, w^+_{h_k}) = 0.$$

Therefore $p \leqslant 0$. In view of ($M1_h$) we know that $\forall k$ there exists a node $z_k := z_{j,i}(k) \in Q_{h_k}$, such that $w_{h_k}(z_k) > 0$. Without loss of generality we may suppose that $z_k \to z \in \Omega$. Therefore $p(z) \geqslant 0$ and we obviously have $p = 0$. But it leads to a contradiction with $1 = \|w_{h_k}\|_{2,2,\Omega}$ and $w_{h_k} \to p$ in V.

■

Lemma 2.23. (Continuous dependence.) *Let $h > 0$ be fixed, $e_{h,n}$, $e_h \in U^h_{ad}$ and $e_{h,n} \to e_h$ in U_{ad}. Further let $u_{h,n} := u_h(e_{h,n}) \in V_h$ be a solution to $(\mathcal{P}_h(e_{h,n}))$ and let ($S2_h$) be fulfilled. Then there exists a function $u_h \in V_h$ such that*

$$u_{h,n} \to u_h \;\text{ in } V$$

and moreover $u_h = u_h(e_h)$ is a solution to $(\mathcal{P}_h(e_h))$.

Proof. Based on Lemma 2.16 and Lemma 2.22 we know that the sequence $\{u_{h,n}\}$ is bounded in $H^2(\Omega)$, i.e.

$$\|u_{h,n}\|_{2,2,\Omega} \;\leqslant\; c, \tag{73}$$

where the positive constant c does not depend on $n \in \mathbb{N}$. Therefore we can pass to a subsequence of $\{u_{h,n}\}$ (denoted by the same sequence) such that

$$u_{h,n} \rightharpoonup u_h \quad \text{in } V. \tag{74}$$

To prove that u_h solves $(\mathcal{P}_h(e_h))$ we pass to the limit for $n \to \infty$ in $(\mathcal{P}_h(e_{h,n}))$.

$$a_{t_{h,n}}(u_{h,n}, v_h) \;+\; b^h_{q_{h,n}}(u^+_{h,n}, v_h) \;=\; F^h(v_h) \quad \forall v_h \in V_h.$$

It holds that $\lim\limits_{n\to\infty} a_{t_{h,n}}(u_{h,n}, v_h) = a_{t_h}(u_h, v_h)$, see [52]. In the analysis of the term $b^h_{q_{h,n}}(u^+_{h,n}, v_h)$ we make use of (74) (it implies $u_{h,n} \rightrightarrows u_h$ in Ω). Hence $\lim\limits_{n\to\infty}(b^h_{q_{h,n}}(u^+_{h,n}, v_h) - b^h_{q_{h,n}}(u^+_h, v_h)) = 0$ so that

$$\lim_{n\to\infty} b^h_{q_{h,n}}(u^+_{h,n}, v_h) = \lim_{n\to\infty}(b^h_{q_{h,n}}(u^+_{h,n}, v_h) - b^h_{q_{h,n}}(u^+_h, v_h)) + \lim_{n\to\infty} b^h_{q_{h,n}}(u^+_h, v_h) = \\ = b^h_{q_h}(u^+_h, v_h).$$

Thus the limit element $u_h \in V_h$ satisfies

$$a_{t_h}(u_h, v_h) + b^h_{q_h}(u^+_h, v_h) = F^h(v_h) \quad \forall v_h \in V_h,$$

i.e. u_h is a solution of $(\mathcal{P}_h(e_h))$. Since $u_h(e_h)$ is unique, not only the subsequence, but the whole sequence $\{u_{h,n}\}$ tends weakly to u_h in V.

It remains to prove the strong convergence. Since $u_{h,n} \rightharpoonup u_h$ in V we have $u_{h,n} \to u_h$ in $H^1(\Omega)$. It is sufficient to prove the convergence in the seminorm $|u|_{a_{t_h},\Omega} := \sqrt{a_{t_h}(u,u)}$, i.e. $a_{t_h}(u_{h,n}, u_{h,n}) \to a_{t_h}(u_h, u_h)$ as $n \to \infty$. From (74), $(\mathcal{P}_h(e_h))$ and $(\mathcal{P}_h(e_{h,n}))$ it follows that

$$a_{t_{h,n}}(u_{h,n}, u_{h,n}) + b^h_{q_{h,n}}(u^+_{h,n}, u_{h,n}) = \\ = F^h(u_{h,n}) \to F^h(u_h) = a_{t_h}(u_h, u_h) + b^h_{q_h}(u^+_h, u_h) \tag{75}$$

as $n \to \infty$. It is not difficult to see that $\lim\limits_{n\to\infty}(b^h_{q_{h,n}}(u^+_{h,n}, u_{h,n}) - b^h_{q_h}(u^+_h, u_h)) = 0$. Then (75) implies $\lim\limits_{n\to\infty}(a_{t_{h,n}}(u_{h,n}, u_{h,n}) - a_{t_h}(u_h, u_h)) = 0$ and consequently

$$a_{t_h}(u_{h,n}, u_{h,n}) = a_{t_h}(u_{h,n}, u_{h,n}) \pm a_{t_{h,n}}(u_{h,n}, u_{h,n}) \to a_{t_h}(u_h, u_h), \quad n \to \infty.$$

■

Lemma 2.24. *There exists a constant $c_2 > 0$ such that $\forall\{u_{h,i}, t_{h,i}, q_{h,i}\} \in W_h$, $i = 1, 2$ it holds*

$$c_2 \|u_{h,1} - u_{h,2}\|^2_{2,2,\Omega} \leqslant a_{t_{h,1}}(u_{h,1} - u_{h,2}, u_{h,1} - u_{h,2}) + \\ + b^h_{q_{h,1}}(u^+_{h,1} - u^+_{h,2}, u_{h,1} - u_{h,2}). \tag{76}$$

The constant c_2 does not depend on $\{u_{h,i}, t_{h,i}, q_{h,i}\} \in W_h, i = 1, 2$.

Proof. Suppose that (76) is not fulfilled. Then there exist sequences $\{u_{h,1,n}, t_{h,1,n}, q_{h,1,n}\}$, $\{u_{h,2,n}, t_{h,2,n}, q_{h,2,n}\} \subset W_h$ such that $\forall n \geqslant 1$ it holds

$$\frac{1}{n}\|u_{h,1,n} - u_{h,2,n}\|^2_{2,2,\Omega} > a_{t_{h,1,n}}(u_{h,1,n} - u_{h,2,n}, u_{h,1,n} - u_{h,2,n}) + \\ + b^h_{q_{h,1,n}}(u^+_{h,1,n} - u^+_{h,2,n}, u_{h,1,n} - u_{h,2,n}) \geqslant 0. \tag{77}$$

According to (73), sequences $\{u_{h,1,n}\}$, $\{u_{h,2,n}\}$ are bounded in $H^2(\Omega)$. Then there exist their subsequences (denoted by the same sequences) and functions $\hat{u}_{h,1}$, $\hat{u}_{h,2}$ such that $u_{h,i,n} \rightharpoonup \hat{u}_{h,i}$ in $H^2(\Omega)$, $i = 1, 2$. In fact it holds $u_{h,i,n} \rightrightarrows \hat{u}_{h,i}$ in Ω, $i = 1, 2$. By setting $v_h = 1$ in $(\mathcal{P}_h(e_{h,1,n}))$ and passing to the limit for $n \to \infty$ we obtain

$$q_1 \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j \hat{u}^+_{h,1}(z_{j,i}) = q_1 \lim_{n\to\infty} \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j u^+_{h,1,n}(z_{j,i}) \geqslant$$
$$\geqslant \lim_{n\to\infty} \sum_{i=1}^{n} q_{h,1,n} \sum_{j=1}^{m} \omega_j u^+_{h,1,n}(z_{j,i}) = F^h(1) > 0.$$

The same estimate holds also for $\{u_{h,2,n}\}$. Hence, we can find nodes z_1, $z_2 \in Q_h$ such that $u_{h,1,n}(z_1) > 0, \hat{u}_{h,1}(z_1) > 0$ (resp. $u_{h,2,n}(z_2) > 0, \hat{u}_{h,2}(z_2) > 0$) for n large enough.

Dividing (77) by $\|u_{h,1,n} - u_{h,2,n}\|^2_{2,2,\Omega}$ yields

$$a_{t_{h,1,n}}(w_{h,1,n} - w_{h,2,n}, w_{h,1,n} - w_{h,2,n}) \to 0, \quad n \to \infty \tag{78}$$

and

$$b^h_{q_{h,1,n}}(w^+_{h,1,n} - w^+_{h,2,n}, w_{h,1,n} - w_{h,2,n}) \to 0, \quad n \to \infty, \tag{79}$$

where we denote $w_{h,i,n} := u_{h,i,n} / \|u_{h,1,n} - u_{h,2,n}\|_{2,2,\Omega}$, $i = 1, 2$. Clearly $\{w_{h,1,n} - w_{h,2,n}\}$ is bounded in $H^2(\Omega)$. Hence there exist subsequences of $\{w_{h,i,n}\}$, $i = 1, 2$ (denoted by the same sequences) and an element $w_h \in V_h$ such that $w_{h,1,n} - w_{h,2,n} \rightharpoonup w_h$ in $H^2(\Omega)$. From (78) it follows that $|w_h|^2_{2,2,\Omega} = 0$, $w_h \equiv p \in P_0$. Then

$$w_{h,1,n} - w_{h,2,n} \to p \text{ in } H^2(\Omega) \text{ and } w^+_{h,1,n}(z_{j,i}) - w^+_{h,2,n}(z_{j,i}) \to 0 \ \ \forall z_{j,i} \in Q_h. \tag{80}$$

Let us first consider that:

$$\exists c > 0 : \|u_{h,1,n} - u_{h,2,n}\|_{2,2,\Omega} \geqslant c \qquad \forall n. \tag{81}$$

Therefore $\{w_{h,1,n}\}$, $\{w_{h,2,n}\}$ are bounded in $H^2(\Omega)$ and there exist their subsequences (denoted by the same sequences) converging weakly to $\hat{w}_{h,1}$, $\hat{w}_{h,2}$ in $H^2(\Omega)$. Hence (80) leads to

$$\hat{w}_{h,1} - \hat{w}_{h,2} = p \quad \text{and} \quad \hat{w}^+_{h,1}(z_{j,i}) - (\hat{w}^+_{h,1}(z_{j,i}) - p) = 0 \quad \forall z_{j,i} \in Q_h.$$

As the sequence $\{u_{h,1,n}\}$ is bounded and $\hat{u}_{h,1}(z_1) > 0$, also $\hat{w}_{h,1}(z_1) > 0$. From there obviously $p = 0$ in Ω.

If (81) is not satisfied, then $\|u_{h,1,n} - u_{h,2,n}\|_{2,2,\Omega} \to 0$, $n \to \infty$. Thus $\hat{u}_{h,1} = \hat{u}_{h,2}$ a.e. in Ω and we can denote by $z_{1,2}$ the point where $\hat{u}_{h,1}, u_{h,1,n}, \hat{u}_{h,2}$ and $u_{h,2,n}$

are positive for n large enough. It implies that $w_{h,1,n}(z_{1,2}) > 0$ and $w_{h,2,n}(z_{1,2}) > 0$. Then

$$(w_{h,1,n} - w_{h,2,n})(z_{1,2}) = (w^+_{h,1,n} - w^+_{h,2,n})(z_{1,2}) \to 0, \quad n \to \infty \tag{82}$$

what again implies $p = 0$ being in contradiction with $1 = \|w_{h,1,n} - w_{h,2,n}\|_{2,2,\Omega}$ and $w_{h,1,n} - w_{h,2,n} \to p$ in $H^2(\Omega)$.

■

To ensure the existence of a solution to (P_h), it remains to assume the lower semicontinuity of I_h:

($\mathrm{I1}_h$) If $e_h,\ e_{h,n} \in U^h_{ad},\ e_{h,n} \to e_h$ in U^h_{ad} and $v_h,\ v_{h,n} \in V_h,\ v_{h,n} \to v_h$ in V, then

$$\liminf_{n\to\infty} I_h(e_{h,n},\, v_{h,n}) \geqslant I_h(e_h,\, v_h).$$

Theorem 2.5. (Existence of a solution to (P_h)) *Let I_h satisfy* ($\mathrm{I1}_h$). *Then* (P_h) *has at least one solution for every $h > 0$.*

Proof. The assertion follows from ($\mathrm{I1}_h$), Lemma 2.23 and Lemma 2.15.

■

In view of Lemma 2.23 and Lemma 2.24, the mapping $e_h \longmapsto u_h(e_h)$ is Lipschitz continuous, i.e. there exists $K_1 > 0$ such that $\forall e_{h,1} = \{t_{h,1}, q_{h,1}\}, e_{h,2} = \{t_{h,2}, q_{h,2}\} \in U^h_{ad}$ it holds

$$\|u_h(e_{h,1}) - u_h(e_{h,2})\|_{2,2,\Omega} \leqslant K_1 \left(\|t_{h,1} - t_{h,2}\|_{C(\bar{\Omega})} + \|q_{h,1} - q_{h,2}\|_{2,\Omega} \right).$$

In addition let us suppose that I_h is Lipschitz continuous on $U^h_{ad} \times V_h$:

($\mathrm{I2}_h$) There exists a constant $c > 0$ such that $\forall e_{h,1}, e_{h,2} \in U^h_{ad}$ and $\forall v_{h,1}, v_{h,2} \in V_h$ it holds:

$$|I_h(e_{h,1}, v_{h,1}) - I_h(e_{h,2}, v_{h,2})| \leqslant c \Big(\|v_{h,1} - v_{h,2}\|_{2,2,\Omega} + \\ + \|t_{h,1} - t_{h,2}\|_{C(\bar{\Omega})} + \|q_{h,1} - q_{h,2}\|_{2,\Omega} \Big).$$

Lemma 2.25. *Let I_h satisfy* ($\mathrm{I2}_h$). *Then $J_h(e_h) = I_h(e_h, u_h(e_h))$, with $u_h(e_h)$ being a solution to ($\mathcal{P}_h(e_h)$), is Lipschitz continuous in U^h_{ad}, i.e. there exists a constant $K_2 > 0$ such that $\forall e_{h,1}, e_{h,2} \in U^h_{ad}$:*

$$|J_h(e_{h,1}) - J_h(e_{h,2})| \leqslant K_2 \left(\|t_{h,1} - t_{h,2}\|_{C(\bar{\Omega})} + \|q_{h,1} - q_{h,2}\|_{2,\Omega} \right).$$

Proof. The assertion directly follows from ($\mathrm{I2}_h$) and the Lipschitz continuity of $e_h \longmapsto u_h(e_h)$.

■

To end this section we will show that the approximations of cost functionals defined by (54), (55) and (56) have the required properties.

Lemma 2.26. *Let $I_{h,i} \in \mathcal{Q}_h^0$, $i = 1, 2, 3$. Then the cost functionals (54), (55) and (56) satisfy conditions* ($\mathrm{I1}_h$), ($\mathrm{I2}_h$).

Proof. We start with the cost functional (54). Condition ($\mathrm{I1}_h$) follows from the fact that $u_{h,n} \to u_h$ in V implies $u_{h,n} \rightrightarrows u_h$ in Ω. Let $e_{h,1}, e_{h,2} \in U_{ad}^h$ and $u_h(e_{h,1}), u_h(e_{h,2}) \in V_h$ solve $(\mathcal{P}_h(e_{h,1}))$ and $(\mathcal{P}_h(e_{h,2}))$, respectively. Therefore it holds

$$|J_{h,1}(e_{h,1}) - J_{h,1}(e_{h,2})| \leqslant l \max_i \|f\|_{C(\bar{K}_i)} \|u_h(e_{h,1}) - u_h(e_{h,2})\|_{2,2,\Omega}\,.$$

Thus $J_{h,1}(e_h) = I_{h,1}(e_h, u_h(e_h))$ satisfies ($\mathrm{I2}_h$).

Let us now continue with the cost functional (55). It is again not difficult to prove the lower semicontinuity of $J_{h,2}(e_h) = I_{h,2}(e_h, u_h(e_h))$ using the uniform convergence as in the previous case. By using the discrete Cauchy–Schwarz inequality (see e.g. [6]) and boundedness (73) we obtain

$$\begin{aligned}
|J_{h,2}(e_{h,1}) - J_{h,2}(e_{h,2})| &\leqslant \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j \,|u_{h,1}(z_{j,i}) - u_{h,2}(z_{j,i})|\,|u_{h,1}(z_{j,i}) + u_{h,2}(z_{j,i})| \leqslant \\
&\leqslant l\,\|u_{h,1} + u_{h,2}\|_{2,2,\Omega}\,\|u_{h,1} - u_{h,2}\|_{2,2,\Omega} \leqslant \\
&\leqslant l(\|u_{h,1}\|_{2,2,\Omega} + \|u_{h,2}\|_{2,2,\Omega})\,\|u_{h,1} - u_{h,2}\|_{2,2,\Omega} \leqslant \\
&\leqslant c\,\|u_{h,1} - u_{h,2}\|_{2,2,\Omega}\,,
\end{aligned}$$

where c is a positive constant which does not depend on $e_{h,1}, e_{h,2}$. Thus $J_{h,2}(e) = I_{h,2}(e_h, u_h(e_h))$ satisfies ($\mathrm{I2}_h$).

Finally we can approach to the functional (56). Let $e_{h,n} \to e_h$ in U_{ad}, then $u_h(e_{h,n}) \to u_h(e_h) \in V$. Then

$$\begin{aligned}
\liminf_{n\to\infty} I_{h,3}(e_{h,n},\, u_{h,n}) &= \liminf_{n\to\infty} \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j t_{h,n}^2(z_{j,i})(u_{h,n}'')^2(z_{j,i}) \;\geqslant \\
&\geqslant \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j t_h^2(z_{j,i})(u_h'')^2(z_{j,i}) \;=\; I_{3,h}(e_h,\, u_h).
\end{aligned}$$

Further it holds

$$\begin{aligned}
|J_{h,3}(e_{h,1}) - J_{h,3}(e_{h,2})| &\leqslant t_1^2 \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j \left|(u_{h,1}'' - u_{h,2}'')(z_{j,i})\right| \left|(u_{h,1}'' + u_{h,2}'')(z_{j,i})\right| + \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j \left|(t_{h,1}^2 - t_{h,2}^2)(z_{j,i})\right| \left|(u_{h,2}'')^2(z_{j,i})\right| \leqslant \\
&\leqslant l t_1^2 \max_i \|u_{h,1} + u_{h,2}\|_{C^2(\bar{K}_i)}\,\|u_{h,1} - u_{h,2}\|_{2,2,\Omega} + \\
&+ lc\,\|t_{h,1} - t_{h,2}\|_{C(\bar{\Omega})} \max_i \|u_{h,2}\|_{C^2(K_i)} \leqslant \\
&\leqslant c\Big(\,\|u_{h,1} - u_{h,2}\|_{2,2,\Omega} \,+\, \|t_{h,1} - t_{h,2}\|_{C(\bar{\Omega})}\,\Big),
\end{aligned}$$

where c is a positive constant which does not depend on $e_{h,1}, e_{h,2}$. Thus $J_{h,3}(e_h) = I_{h,3}(e_h, u_h(e_h))$ satisfies (I2$_h$).

■

2.3. Convergence analysis

This subsection will be devoted to the analysis of the relation between solutions of (P$_h$) and the solution of (P) for $h \to 0^+$. The convergence analysis starts with the following lemma:

Lemma 2.27. *Let $e_h \in U_{ad}^h, e \in U_{ad}$, $e_h \to e$ in U_{ad} as $h \to 0^+$. Further let $u_h(e_h) \in V_h$ be a solution to $(\mathcal{P}_h(e_h))$ and let (S2$_h$) be fulfilled for every $0 < h \leqslant h_0$. Then there exists a function $u \in V$ such that*

$$u_h \to u \ \text{ in } V$$

and in addition $u = u(e)$ is a solution to $(\mathcal{P}(e))$.

Proof. Firstly we prove the uniform boundedness of $\{u_h\}$. Let u_h be a solution to $(\mathcal{P}_h(e_h))$. We insert $v_h = u_h$ into $(\mathcal{P}_h'(e_h))$ and according to (S2$_h$), Lemma 2.22 and (58) there exist positive constants c_1, c_2 such that

$$c_1 \|u_h\|_{2,2,\Omega}^2 \leqslant a_{t_h}(u_h, u_h) + b_{q_h}^h(u_h^+, u_h) = F^h(u_h) \leqslant c_2 \|u_h\|_{2,2,\Omega}.$$

Therefore

$$\exists c > 0 : \qquad \|u_h\|_{2,2,\Omega} \leqslant c \qquad \forall 0 < h \leqslant h_0.$$

Thus one can find a subsequence of $\{u_h\}$ (denoted by the same sequence) and a function $u \in V$ such that

$$u_h \rightharpoonup u \quad \text{in } V \text{ as } h \to 0^+. \tag{83}$$

From the definition of V_h it follows that for arbitrary function $v \in V$ one can find a sequence $\{v_h\}, v_h \in V_h$ such that $v_h \to v$ in V as $h \to 0^+$.

Next we prove that u solves $(\mathcal{P}(e))$. The state problem $(\mathcal{P}_h(e_h))$ reads

$$a_{t_h}(u_h, v_h) + b_{q_h}^h(u_h^+, v_h) = F^h(v_h) \quad \forall v_h \in V_h.$$

We will pass to the limit for $h \to 0^+$. Firstly we focus on $a_{t_h}(u_h, v_h)$. Making use of (83) we have

$$\lim_{h\to 0^+} a_{t_h}(u_h, v_h) = \lim_{h\to 0^+} (a_{t_h} - a_t)(u_h, v_h) + \lim_{h\to 0^+} a_t(u_h, v_h) = a_t(u, v).$$

Passing to the limit for $h \to 0^+$ in $b_{q_h}^h(u_h^+, v_h)$ we obtain

$$\begin{aligned}\lim_{h\to 0^+} b_{q_h}^h(u_h^+, v_h) &= \lim_{h\to 0^+} \left(b_{q_h}^h(u_h^+, v_h) - b_{q_h}(u_h^+, v_h)\right) + \lim_{h\to 0^+} b_{q_h}(u_h^+, v_h) = \\ &= \lim_{h\to 0^+} (b_{q_h} - b_q)(u_h^+, v_h) + \lim_{h\to 0^+} b_q(u_h^+, v_h) = b_q(u^+, v).\end{aligned}$$

We used (59), (83) and the boundedness of $\{u_h\}, \{v_h\}$. It remains to pass to the limit in $F^h(v_h)$. Estimate (60) then implies:

$$\lim_{h\to 0^+} F^h(v_h) = \lim_{h\to 0^+} \left(F^h - F\right)(v_h) + \lim_{h\to 0^+} F(v_h) = \lim_{h\to 0^+} F(v_h) = F(v).$$

Summarizing the previous results yields

$$a_t(u,v) + b_q(u^+, v^+) = F(v) \quad \forall v \in V.$$

The limit function $u \in V$ is a solution to $(\mathcal{P}(e))$. Since the solution u is unique, then the whole sequence $\{u_h\}$ tends weakly to u in V.

Next we will prove the strong convergence. Since $u_h \rightharpoonup u$ in V implies $u_h \to u$ in $H^1(\Omega)$, it is sufficient to prove that $a_t(u_h, u_h) \to a_t(u,u)$. We know that

$$a_{t_h}(u_h, u_h) + b^h_{q_h}(u_h^+, u_h) = F^h(u_h) \to F(u) = a_t(u,u) + b_q(u^+, u^+). \tag{84}$$

It is not difficult to prove that $b^h_{q_h}(u_h^+, u_h) \to b_q(u^+, u^+)$. From (84) then follows $a_{t_h}(u_h, u_h) \to a_t(u,u)$ and

$$a_t(u_h, u_h) = a_t(u_h, u_h) \pm a_{t_h}(u_h, u_h) \to a_t(u,u).$$

■

Now we turn our attention to the relation between cost functionals I, I_h. Let us assume that I, I_h have the following properties:

(I3$_h$) There exists a constant $c > 0$ such that

$$|I_h(e_h, v_h) - I(e_h, v_h)| \leqslant ch \, \|v_h\|_{2,2,\Omega} \quad \forall v_h \in V_h, \, \forall e_h \in U^h_{ad}.$$

(I4$_h$) Let $e_h \to e$ in U_{ad}, $v_h \to v$ in V, where $e_h \in U^h_{ad}$, $e \in U_{ad}$, $v_h \in V_h$, $v \in V$, then

$$\lim_{h\to 0^+} I(e_h, v_h) = I(e, v).$$

Theorem 2.6. *Let I, I_h satisfy* (I3$_h$), (I4$_h$). *Then for arbitrary sequence $\{e^*_h\}$, where $e^*_h \in U^h_{ad}$ is a solution to* (P$_h$) *and $u_h(e^*_h)$ solves $(\mathcal{P}_h(e^*_h))$, one can find a subsequence $\{e^*_{h_j}\}$ such that*

$$e^*_{h_j} \to e^* \ \text{ in } U_{ad}, \tag{85}$$

$$u_{h_j}(e^*_{h_j}) \to u(e^*) \ \text{ in } V, \tag{86}$$

where $\{e^*, u(e^*)\}$ is a solution of (P).

Proof. For arbitrary $\bar{e} \in U_{ad}$ there exists a sequence $\{\bar{e}_h\} \subset U^h_{ad}$ such that $\bar{e}_h \to \bar{e}$ in U_{ad} (see e.g. [5]). From $U^h_{ad} \subset U_{ad}$ and compactness of U_{ad} it follows the existence of $\{e^*_{h_j}\}$, $\{\bar{e}_{h_j}\}$ and $e^* \in U_{ad}$ such that

$$e^*_{h_j} \to e^* \ \text{ in } U_{ad},$$
$$\bar{e}_{h_j} \to \bar{e} \ \text{ in } U_{ad}.$$

Let us denote by $u_{h_j}(e^*_{h_j})$, $u_{h_j}(\bar{e}_{h_j})$ solutions to $(\mathcal{P}_{h_j}(e^*_{h_j}))$ resp. $(\mathcal{P}_{h_j}(\bar{e}_{h_j}))$. Using Lemma 2.27 it yields

$$u_{h_j}(e^*_{h_j}) \to u(e^*) \ \text{ in } V,$$
$$u_{h_j}(\bar{e}_{h_j}) \to u(\bar{e}) \ \text{ in } V,$$

where $u(e^*)$, $u(\bar{e})$ solves $(\mathcal{P}(e^*))$ resp. $(\mathcal{P}(\bar{e}))$. The definition of (P_h) implies

$$I_{h_j}(e^*_{h_j}, u_{h_j}(e^*_{h_j})) \leqslant I_{h_j}(\bar{e}_{h_j}, u_{h_j}(\bar{e}_{h_j})) \ \ \forall \bar{e}_{h_j} \in U^{h_j}_{ad}. \tag{87}$$

If we pass to the limit for $h_j \to 0^+$ in (87), use $(\mathrm{I3}_h)$ and $(\mathrm{I4}_h)$, we obtain

$$\lim_{h_j \to 0^+} I_{h_j}(e^*_{h_j}, u_{h_j}(e^*_{h_j})) = \lim_{h_j \to 0^+} I_{h_j}(e^*_{h_j}, u_{h_j}(e^*_{h_j})) \pm I(e^*_{h_j}, u_{h_j}(e^*_{h_j})) =$$
$$= \lim_{h_j \to 0^+} I(e^*_{h_j}, u_{h_j}(e^*_{h_j})) = I(e, u(e)).$$

We proceed similarly for the right hand side of (87) and we finally obtain

$$I(e^*, u(e^*)) \leqslant I(\bar{e}, u(\bar{e})) \ \ \forall \bar{e} \in U_{ad}.$$

Therefore e^* is an optimal solution of (P).

■

Condition $(\mathrm{I4}_h)$ is clearly satisfied for the cost functionals (4), (5) and (6) and its approximations $I_{h,i} \in \mathcal{Q}^0_h$, $i = 1, 2, 3$ defined by (54), (55) and (56).

Lemma 2.28. *Let $I_{h,i} \in \mathcal{Q}^0_h$, $i = 1, 2, 3$. Then (54) and (55) satisfy $(\mathrm{I3}_h)$.*

Proof. For functional (54) we proceed analogically as in the proof of inequality (60).

Let us now continue by the cost functional (55). Assume that $e_h \in U^h_{ad}$, $u_h(e_h)$ solves $(\mathcal{P}_h(e_h))$. In view of $u_h \in H^2(\Omega)$ it holds that $u_h^2 \in W^{1,1}(K_i)$, $i = 1, \ldots, n$. Due to the compact embedding of $H^2(\Omega)$ into $C^1(\bar{\Omega})$, (73) and Cauchy–Schwarz inequality, we have

$$|I_{h,2}(e_h, u_h) - I_2(e_h, u_h)| \leqslant \sum_{i=1}^{n} \left| \sum_{j=1}^{m} \omega_j u_h^2(z_{j,i}) - \int_{K_i} u_h^2 \, \mathrm{d}x \right| \leqslant$$
$$\leqslant ch \sum_{i=1}^{n} \left| u_h^2 \right|_{1,1,K_i} \leqslant ch \sum_{i=1}^{n} \|u_h\|_{1,2,K_i} \leqslant$$
$$\leqslant ch \, \|u_h\|_{1,2,\Omega} \leqslant c_2 h \, \|u_h\|_{2,2,\Omega} \, .$$

■

3. Natural boundary condition $u(0) = 0$

3.1. Existence analysis of (P)

In this section the existence of a solution to the problem (P) with

$$V = V_2 = \{v \in H^2(\Omega) : \ v(0) = 0\}$$

will be studied. We will proceed using the same approach as in the case of natural boundary condition $u'(0) = 0$.

3.1.1. Existence and uniqueness of a solution to ($\mathcal{P}(e)$)

Through the subsection we will assume that $e \in U_{ad}$ is arbitrary but fixed. The following boundary conditions are prescribed in this case:

$$u(0) = u'''(0) = u''(l) = u'''(l) = 0. \tag{88}$$

Conditions (88) define a beam that is free at the right end ($x = l$) and its left end ($x = 0$) can slope but can not move in the vertical direction (see Fig. 4).

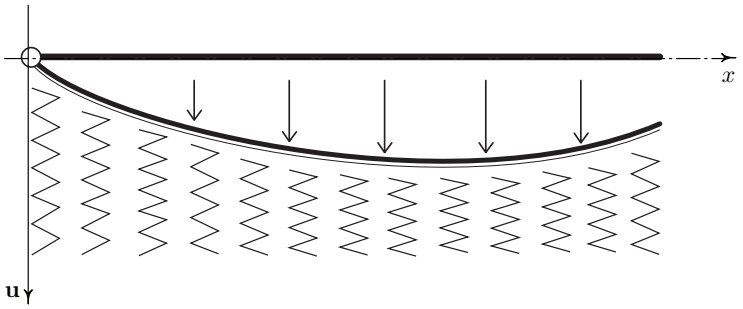

Figure 4: Outline of the beam with boundary condition $u(0) = 0$.

Therefore there are allowed rigid motions of the beam for which the foundation is not active (see Fig. 5) and only the estimate (10) holds. Let us now define a scalar product on $H^2(\Omega)$:

$$((u, v))_{2,2,\Omega} := (u, v)_{2,\Omega} \ + \ (u'', v'')_{2,\Omega}. \tag{89}$$

The decomposition of V to $\mathcal{R}_V$ and $\mathcal{R}_V^{\ominus}$ with regard to (89) then reads

$$\mathcal{R}_V \ = \ \{v \in V \cap P_1 : a_t(v, v) + b_q(v^+, v) = 0\} \ = \ \{p \in P_1 : p = ax, \ a \leqslant 0\}. \tag{90}$$

The negative polar cone has the following form:

$$\mathcal{R}_V^{\ominus} = \{v \in V_2 : ((v, p))_{2,2,\Omega} \leqslant 0 \ \ \forall p \in \mathcal{R}_V\} = \{v \in V : (v, x)_{2,\Omega} \geqslant 0\}. \tag{91}$$

It is easy to prove that cones (90), (91) are convex and closed. The norm $[|v|]_{2,2,\Omega} = ((v, v))_{2,2,\Omega}^{1/2}$ induced by the scalar product (89) is equivalent to the standard norm on the Sobolev space $H^2(\Omega)$, see e.g. [27].

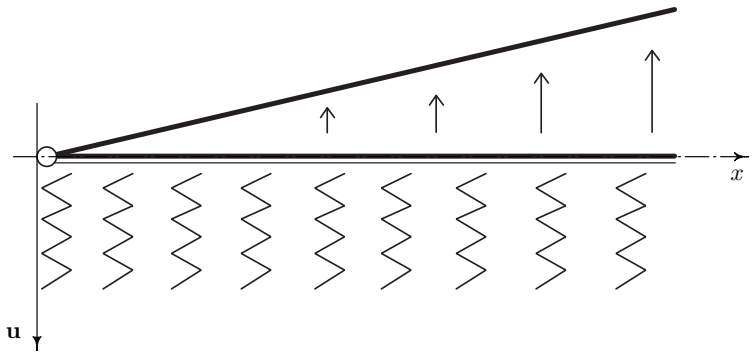

Figure 5: Rigid beam motions belonging to $\mathcal{R}_V$.

Theorem 3.1. (Necessary condition for the existence of a solution to $(\mathcal{P}(e))$.) *Let there exist a solution to $(\mathcal{P}(e))$, then the condition*

$$F(x) \;=\; L(x) \;+\; \sum_i F_i x_i \;-\; \sum_j M_j \;\geqslant\; 0 \tag{S3}$$

must be satisfied.

Proof. Let $u \in V$ be a solution of $(\mathcal{P}(e))$. Inserting $v \;=\; p \in \mathcal{R}_V$ into $(\mathcal{P}'(e))$ we obtain:

$$0 \geqslant b_q(u^+, p) \;=\; F(p) \;=\; a\left(L(x) \;+\; \sum_i F_i x_i \;-\; \sum_j M_j\right).$$

■

Lemma 3.1. *Let $\mathcal{R}_V$, $\mathcal{R}_V^{\ominus}$ be defined by (90),(91). Then*

$$V = \mathcal{R}_V \oplus \mathcal{R}_V^{\ominus}. \tag{92}$$

Moreover $\forall v \in V \;\exists!\; \{p, \bar{v}\} \in \mathcal{R}_V \times \mathcal{R}_V^{\ominus}$ such that

$$v \;=\; p \oplus \bar{v}, \quad (\bar{v}, p)_{2,2,\Omega} \;=\; a \cdot (\bar{v}, x)_{2,\Omega} \;=\; 0. \tag{93}$$

Proof. For the proof we refer to [3].

■

In view of (90),(91) and (93) we easily deduce that only one of the following variants can occur:

$$a \;=\; 0 \quad \text{and} \quad (\bar{v}, x)_{2,\Omega} \;\geqslant\; 0, \tag{A3}$$

$$a \;\leqslant\; 0 \quad \text{and} \quad (\bar{v}, x)_{2,\Omega} \;=\; 0. \tag{A4}$$

In the proof of coercivity of $\mathcal{E}_e$ will be necessary to use the following modification of the Poincaré inequality.

Lemma 3.2. (Poincaré type inequality) *Let $V = \{v \in H^2(\Omega) : v(0) = 0\}$, then there exists a positive constant c_P dependent only on the interval Ω such that*

$$\|v\|^2_{2,2,\Omega} \leqslant c_P \left(|v|^2_{2,2,\Omega} + (v,\, x)^2_{2,\Omega} \right) \quad \forall v \in V. \tag{94}$$

Proof. We proceed similarly as in the proof of Lemma 2.4. ■

Next we can pass to the proof of coercivity of $\mathcal{E}_e$.

Lemma 3.3. *Let the condition*

$$F(x) \,=\, L(x) \,+\, \sum_i F_i x_i \,-\, \sum_j M_j \,>\, 0 \tag{S4}$$

be fulfilled. Then the functional $\mathcal{E}_e$ is coercive on V.

Proof. Let (S4) be fulfilled. Firstly, in the case of variant (A3), we have $a = 0$, $v \equiv \bar{v}$ and $(\bar{v}, x)_{2,\Omega} \geqslant 0$. The following inequality is a consequence of properties of $\bar{v}^+$:

$$0 \,\leqslant\, (\bar{v}, x)^2_{2,\Omega} \,\leqslant\, (\bar{v}^+, x)^2_{2,\Omega} \,\leqslant\, l^3 \left\|\bar{v}^+\right\|^2_{2,\Omega}. \tag{95}$$

Then we can rewrite $\mathcal{E}_e$, with use of (93), (94) and (95), as follows:

$$\begin{aligned} 2\mathcal{E}_e(v) = 2\mathcal{E}_e(\bar{v}) \,&=\, a_t(\bar{v}, \bar{v}) \,+\, b_q(\bar{v}^+, \bar{v}^+) \,-\, 2F(\bar{v}) \,\geqslant \\ &\geqslant \beta_0 t_0^3 \, |\bar{v}|^2_{2,2,\Omega} \,+\, q_0 \left\|\bar{v}^+\right\|^2_{2,\Omega} \,-\, 2F(\bar{v}) \,\geqslant \\ &\geqslant \beta_0 t_0^3 \, |\bar{v}|^2_{2,2,\Omega} \,+\, \frac{2q_0}{l^2} (\bar{v}, x)^2_{2,\Omega} \,-\, 2F(\bar{v}) \,\geqslant \\ &\geqslant \|\bar{v}\|_{2,2,\Omega} \, (c_1 \, \|\bar{v}\|_{2,2,\Omega} \,-\, 2 \, \|f\|_{2,\Omega}), \end{aligned}$$

where $c_1 := (1/c_P) \min\{\beta_0 t_0^3,\, 2q_0/l^2\}$.

Secondly in the case of variant (A4), it holds that $(\bar{v}, x)_{2,\Omega} = 0$ and $a \leqslant 0$. Using (94) we have

$$\begin{aligned} 2\mathcal{E}_e(v) = 2\mathcal{E}_e(p + \bar{v}) \,&=\, a_t(\bar{v}, \bar{v}) \,+\, b_q(v^+, v^+) \,-\, 2F(p) \,-\, 2F(\bar{v}) \,\geqslant \\ &\geqslant \beta_0 t_0^3 \, |\bar{v}|^2_{2,2,\Omega} \,+\, q_0 \left\|(p + \bar{v})^+\right\|^2_{2,\Omega} \,+\, 2|a|F(x) \,-\, 2F(\bar{v}) \,\geqslant \\ &\geqslant \beta_0 t_0^3 \, |\bar{v}|^2_{2,2,\Omega} \,+\, 2|a|F(x) \,-\, 2F(\bar{v}) \,\geqslant \\ &\geqslant \beta_0 t_0^3 \, |\bar{v}|^2_{2,2,\Omega} \,+\, (\bar{v}, x)^2_{2,\Omega} \,+\, 2|a|F(x) \,-\, 2F(\bar{v}) \,\geqslant \\ &\geqslant c_2 \, \|\bar{v}\|^2_{2,2,\Omega} \,+\, 2|a|F(x) \,-\, 2 \, \|f\|_{2,\Omega} \, \|\bar{v}\|_{2,2,\Omega}, \end{aligned}$$

where $c_2 := (1/c_P) \min\{\beta_0 t_0^3,\, 1\}$. The orthogonality of the decomposition (92) ensures that $\|v\|^2_{2,2,\Omega} = \|\bar{v}\|^2_{2,2,\Omega} + \|ax\|^2_{2,2,\Omega}$. Thus if $\|v\|_{2,2,\Omega} \to +\infty$, then at least one part of the function $v = \bar{v} + ax$ converge to infinity in appropriate norm. Finally we make use of (S4) which ensures the coercivity.

■

According to Lemma 2.2 we know that $\mathcal{E}_e$ is convex and Gâteaux differentiable on V. The coercivity of $\mathcal{E}_e$ on V enable us to introduce the following existence theorem.

Theorem 3.2. (Necessary and sufficient condition for the existence and uniqueness of a solution to $(\mathcal{P}(e))$.) *The state problem $(\mathcal{P}(e))$ has a unique solution if and only if (S4) is fulfilled. Such a solution $u \in V$ can be characterized as follows:*

$$\mu(M_u) > 0, \tag{M2}$$

where $\mu(M_u)$ is the one–dimensional Lebesgue measure of the set $M_u = \{x \in \Omega : u(x) > 0\}$.

Proof. *Necessity.* This part of the proof will be done by contradiction. Assume that $u \in V$ is a unique solution to $(\mathcal{P}(e))$ and (S4) does not hold. Then according to Theorem 3.1 we have $F(x) = 0$. By setting $v \equiv p \in \mathcal{R}_V$ in $(\mathcal{P}'(e))$ we obtain:

$$a_t(u,p) + b_q(u^+, ax) = F(ax) = aF(x) \quad \forall ax \in \mathcal{R}_V,\ ax \neq 0, \tag{96}$$

$$b_q(u^+, ax) = 0 \qquad \forall ax \in \mathcal{R}_V,\ p \neq 0. \tag{97}$$

From (97) one can deduce that $u + p$ is another solution of $(\mathcal{P}(e))$ what is in contradiction with the uniqueness of u (see the proof of Theorem 2.2). Thus (S4) must be satisfied.

Sufficiency. Let (S4) hold. In view of Lemma 2.2 and Lemma 3.3 we know that $\mathcal{E}_e$ is Gâteaux differentiable, convex and coercive on V. It implies that there exists at least one function $u \in V$ solving $(\mathcal{P}(e))$, see e.g. [12].

Further let $u \in V$, $u \leqslant 0$ a.e. in Ω solves $(\mathcal{P}(e))$. Setting $v \equiv p \in \mathcal{R}_V$, $p \neq 0$ in $(\mathcal{P}'(e))$ leads to

$$0 = b_q(u^+, ax) = a\,F(x). \tag{98}$$

But (98) is in contradiction with (S4). Thus M_u must have a positive Lebesgue measure.

It remains to prove the uniqueness. It can be proved exactly in the same way as in the proof of Theorem 2.2. ■

3.1.2. Existence of solutions to (P)

Let us consider that $\beta \in L^\infty(\Omega)$, $0 < \beta_0 \leqslant \beta(x)$ a.e. v Ω and $F \in S_\delta$, where

$$S_\delta = \{F \in V^* : F(x) \geqslant \delta > 0\}.$$

We know that for any $e \in U_{ad}$ there exists a unique solution to $(\mathcal{P}(e))$ with the property (M2). Then we denote the set of all such solutions by W:

$$W := \{\{u,t,q\} \in V \times U^t_{ad} \times U^q_{ad} : u = u(e) \text{ solves } (\mathcal{P}(e)),\ e = \{t,q\}\}.$$

Lemma 3.4. *There exists a positive constant $c_1 = c_1(\delta)$ such that*

$$c_1 \, \|u\|^2_{2,2,\Omega} \leqslant a_t(u,u) + b_q(u^+,u) \qquad \forall \{u,t,q\} \in W. \tag{99}$$

The constant c_1 does not depend on $\{u,t,q\} \in W$.

Proof. Let us suppose that (99) does not hold. Then one can find a sequence $\{u_n, t_n, q_n\} \subset W$ such that

$$\frac{1}{n} \, \|u_n\|^2_{2,2,\Omega} > a_{t_n}(u_n,u_n) + b_{q_n}(u_n^+,u_n) \geqslant 0 \quad \forall n \geqslant 1. \tag{100}$$

Dividing (100) by $\|u_n\|^2_{2,2,\Omega}$ and passing to the limit for $n \to \infty$ lead to

$$\lim_{n\to\infty} a_{t_n}(w_n,w_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} b_{q_n}(w_n^+,w_n) = 0,$$

where $w_n := u_n / \|u_n\|_{2,2,\Omega}$. Clearly $\|w_n\|_{2,2,\Omega} = 1$. Hence there exists a subsequence of $\{w_n\}$ (denoted by the same sequence) and an element $w \in V$ such that $w_n \rightharpoonup w$ in V. In a similar manner as in the proof of Lemma 2.9 we can show that $|w|^2_{2,2,\Omega} = 0$, $w \equiv p = ax \in P_1, a \leqslant 0$ and $w_n \to w$ in V. It implies that $w_n \rightrightarrows ax$ in Ω.

From $F \in S_\delta$ it follows that

$$\int_0^l u_n^+ \, \mathrm{d}x \geqslant \delta / q_1 l \tag{101}$$

which implies that there exists $\varepsilon = \varepsilon(\delta) > 0$ such that $\mu(M_{u_n}) \geqslant \varepsilon, \ \forall n \geqslant 1$. Then there exists a sequence $\{x_n\}, x_n \in (\varepsilon, l]$ such that $w_n(x_n) > 0, \forall n \geqslant 1$.

Without loss of generality we will suppose that $x_n \to x$ in $\mathbb{R}^1$, then $x \in [\epsilon, l]$ and $w_n(x_n) \to p(x) \geqslant 0$. Therefore we obviously have $a = 0$. But it leads to a contradiction with $1 = \|w_n\|_{2,2,\Omega}$ and $w_n \to w$ in V.

■

We proceed by the continuous dependence of the solution to $(\mathcal{P}(e))$ on the design variable e.

Lemma 3.5. (Continuous dependence.) *Let e_n, $e \in U_{ad}$, $e_n \to e$ in U_{ad}. Further let $u_n := u(e_n) \in V$ be a solution to $(\mathcal{P}(e_n))$ and let (S4) be fulfilled. Then there exists $u \in V$ such that*

$$u_n \to u \ \text{ in } V$$

and moreover $u := u(e)$ is a solution to $(\mathcal{P}(e))$.

Proof. Using Lemma 3.4 and the Cauchy–Schwarz inequality we can easily prove the boundedness of u_n in V. Therefore one can find a subsequence (denoted by the same sequence) such that $u_n \rightharpoonup u$ in V. In the same way as in the proof of Lemma 2.10 we can show that u solves $(\mathcal{P}(e))$ and $u_n \to u$ in V.

■

Lemma 3.6. *There exists a constant $c_2 = c_2(\delta) > 0$ such that $\forall \{u_i, t_i, q_i\} \in W, i = 1, 2$ it holds*

$$c_2 \, \|u_1 - u_2\|^2_{2,2,\Omega} \leqslant a_{t_1}(u_1 - u_2, u_1 - u_2) \, + \, b_{q_1}(u_1^+ - u_2^+, u_1 - u_2). \tag{102}$$

The constant c_2 does not depend on $\{u_i, t_i, q_i\} \in W, i = 1, 2$.

Proof. Assume that (102) does not hold. Then one can find sequences $\{u_{1,n}, t_{1,n}, q_{1,n}\}$, $\{u_{2,n}, t_{2,n}, q_{2,n}\} \subset W$ such that

$$\begin{aligned} \frac{1}{n} \, \|u_{1,n} - u_{2,n}\|^2_{2,2,\Omega} > a_{t_{1,n}}(u_{1,n} - u_{2,n}, u_{1,n} - u_{2,n}) + \\ + \, b_{q_{1,n}}(u_{1,n}^+ - u_{2,n}^+, u_{1,n} - u_{2,n}) \, \geqslant \, 0 \quad \forall n \geqslant 1. \end{aligned} \tag{103}$$

Due to Lemma 3.5 and condition (S4), sequences $\{u_{1,n}\}$, $\{u_{2,n}\}$ are bounded in $H^2(\Omega)$. Then one can find their subsequences (denoted by the same sequences) and functions $\hat{u}_1$, $\hat{u}_2$ such that $u_{i,n} \rightharpoonup \hat{u}_i$ in $H^2(\Omega)$, $i = 1, 2$. In fact it holds $u_{i,n} \rightrightarrows \hat{u}_i$ in Ω, $i = 1, 2$. By setting $v = x$ in $(\mathcal{P}(e_{i,n}))$, $i = 1, 2$ and passing to the limit for $n \to \infty$ we have

$$q_1 l \int_0^l \hat{u}_i^+ \, \mathrm{d}x = q_1 l \lim_{n\to\infty} \int_0^l u_{i,n}^+ \, \mathrm{d}x \geqslant \lim_{n\to\infty} \int_0^l q_{i,n} u_{i,n}^+ x \, \mathrm{d}x = F(x) > 0.$$

Hence, we can find sets M_1, M_2 with a positive one–dimensional Lebesgue measure, such that $u_{i,n} > 0, \hat{u}_i > 0$ in M_i, $i = 1, 2$.

Dividing (103) by $\|u_{1,n} - u_{2,n}\|^2_{2,2,\Omega}$ and taking into account properties of $a_{t_{1,n}}$, $b_{q_{1,n}}$ yield

$$a_{t_{1,n}}(w_{1,n} - w_{2,n}, w_{1,n} - w_{2,n}) \to 0 \quad \text{and} \quad b_{q_{1,n}}(w_{1,n}^+ - w_{2,n}^+, w_{1,n} - w_{2,n}) \to 0, \tag{104}$$

where $w_{i,n} := u_{i,n} / \|u_{1,n} - u_{2,n}\|_{2,2,\Omega}$, $i = 1, 2$. Clearly $\{w_{1,n} - w_{2,n}\}$ is bounded in $H^2(\Omega)$ and $\|w_{1,n} - w_{2,n}\|_{2,2,\Omega} = 1$. Hence there exist subsequences of $\{w_{i,n}\}$, $i = 1, 2$ (denoted by the same sequences) and an element $w \in V$ such that $w_{1,n} - w_{2,n} \rightharpoonup w$ in V. Similarly as in Lemma 2.11 we can show that $|w|^2_{2,2,\Omega} = 0$, $w \equiv p = ax \in P_1$ and (104) then takes the following form:

$$w_{1,n} - w_{2,n} \to ax \ \text{ in } \ H^2(\Omega) \quad \text{and} \quad w_{1,n}^+ - w_{2,n}^+ \to 0 \ \text{ in } \ L^2(\Omega). \tag{105}$$

Firstly consider:

$$\exists c > 0 : \ \|u_{1,n} - u_{2,n}\|_{2,2,\Omega} \geqslant c \qquad \forall n. \tag{106}$$

Therefore $\{w_{1,n}\}$, $\{w_{2,n}\}$ are bounded in $H^2(\Omega)$ and there exist their subsequences (denoted by the same sequences) converging weakly to $\hat{w}_1$, $\hat{w}_2$ in $H^2(\Omega)$. Hence (105) leads to

$$\hat{w}_1 - \hat{w}_2 = ax \quad \text{and} \quad \hat{w}_1^+ - (\hat{w}_1^+ - ax) = 0 \ \text{ a.e. in } \ \Omega. \tag{107}$$

As $\hat{u}_1 > 0$ in M_1, also $\hat{w}_1 > 0$ in M_1. From there and (107) we have $a = 0$ a.e. in Ω being in contradiction with $\|w_{1,n} - w_{2,n}\|_{2,2,\Omega} = 1$ and (105).

If (106) is not satisfied, then similarly as in the Lemma 2.11 we can prove that $a = 0$, what is in contradiction with $\|w_{1,n} - w_{2,n}\|_{2,2,\Omega} = 1$ and (105).

■

Next we prove the Lipschitz continuity of the mapping $u : e \longmapsto u(e)$, where $u(e)$ solves $(\mathcal{P}(e))$.

Lemma 3.7. *Let (S4) be satisfied. Then the mapping $u : e \mapsto u(e)$, where $u(e)$ solves $(\mathcal{P}(e))$, is Lipschitz continuous in U_{ad}, i.e. there exists a constant $K_1 > 0$ such that $\forall e_1 = \{t_1, q_1\}, e_2 = \{t_2, q_2\} \in U_{ad}$:*

$$\|u(e_1) - u(e_2)\|_{2,2,\Omega} \leqslant K_1 \left(\|t_1 - t_2\|_{C(\bar{\Omega})} + \|q_1 - q_2\|_{2,\Omega} \right).$$

Proof. We proceed exactly in the same way as in the proof of Lemma 2.12. We make use of Lemma 3.6.

■

Theorem 3.3. *Let the cost functional I satisfy* (I1)*, then there exists at least one solution of* (P)*.*

Proof. The assertion follows from Lemma 2.8 and Lemma 3.5.

■

Lemma 3.8. *Let I satisfy* (I2)*. Then $J(e) := I(e, u(e))$, with $u(e)$ being a solution of $(\mathcal{P}(e))$, is Lipschitz continuous in U_{ad}, i.e. there exists a constant $K_2 > 0$ such that:*

$$|J(e_1) - J(e_2)| \leqslant K_2 \left(\|t_1 - t_2\|_{C(\bar{\Omega})} + \|q_1 - q_2\|_{2,\Omega} \right) \qquad \forall e_1, e_2 \in U_{ad}.$$

Proof. The assertion directly follows from (I2) and Lemma 3.7.

■

3.2. Approximation of (P)

3.2.1. Approximation of $(\mathcal{P}(e))$

Let us consider the partition of Ω defined by (49). The set U_{ad} will be approximated similarly as in the previous section. The following finite dimensional approximation of V will be used:

$$V_h = \left\{ v_h \in C^1(\bar{\Omega}) : v_h|_{K_i} \in P_3(K_i),\ \forall i = 1, ..., n,\ v_h(0) = 0 \right\} \subset V.$$

Using the classical Ritz method we approximate $(\mathcal{P}(e))$ as follows:

$$\text{Find } u_h \in V_h : \ \mathcal{E}_{e_h}(u_h) \leqslant \mathcal{E}_{e_h}(v_h) \quad \forall v_h \in V_h, \qquad (\mathcal{P}_h(e_h))$$

where $\mathcal{E}_{e_h}(v_h) = \frac{1}{2}(a_{t_h}(v_h,\, v_h) + b_{q_h}(v_h^+, v_h^+)) - F(v_h)$. The terms b_{q_h} and F will be approximated by applying the quadrature formula (176), see (52), (53). The approximated state problem then reads as follows:

$$\text{Find } u_h \in V_h : \;\; \mathcal{E}^h_{e_h}(u_h) \leqslant \mathcal{E}^h_{e_h}(v_h) \quad \forall v_h \in V_h, \qquad (\mathcal{P}_h(e_h))$$

where $\mathcal{E}^h_{e_h}(v_h) = \frac{1}{2}(a_{t_h}(v_h,\, v_h) + b^h_{q_h}(v_h^+, v_h^+)) - F^h(v_h)$. The equivalent weak formulation of the problem $(\mathcal{P}_h(e_h))$ has the following form:

$$\text{Find } u_h \in V_h : \; a_{t_h}(u_h,\, v_h) + b^h_{q_h}(u_h^+, v_h) \;=\; F^h(v_h) \quad \forall v_h \in V_h. \qquad (\mathcal{P}_h'(e_h))$$

3.2.2. Approximation of the cost functional and the optimization problem

Let $I_h : U^h_{ad} \times V_h \to \mathbb{R}^1$ be the approximation of I. The approximation of the whole optimization problem then reads as follows:

$$\text{Find } e_h^* \in U^h_{ad} : \;\; J_h(e_h^*) \;\leqslant\; J_h(e_h) \quad \forall e_h \in U^h_{ad}, \qquad (\mathrm{P}_h)$$

denoting $J_h(e_h) \equiv I_h(e_h,\, u_h(e_h))$ with $u_h(e_h)$ being a solution to $(\mathcal{P}_h(e_h))$.

3.2.3. Existence and uniqueness of a solution to $(\mathcal{P}_h(e_h))$

Firstly recall that in this subsection $e_h \in U^h_{ad}$ is arbitrary but fixed and $b^h_{q_h} \in \mathcal{Q}^0_h$, $F^h \in \mathcal{Q}^0_h$. We will make the decomposition of V_h into the convex cone of rigid displacements and its negative polar cone. After that we will prove the coercivity of $\mathcal{E}^h_{e_h}$ on V_h.

$$\mathcal{R}_{V_h} \;=\; \{v_h \in V_h \cap P_1 : a_{t_h}(v_h, v_h) + b^h_{q_h}(v_h^+, v_h) = 0\} \;=\; \{ax \in P_1 : a \leqslant 0\}.$$

From there by using (63) we define the negative polar cone $\mathcal{R}^{\ominus}_{V_h}$ as follows:

$$\begin{aligned}\mathcal{R}^{\ominus}_{V_h} &= \{v_h \in V_h : ((v_h, p))_{2,\Omega} \leqslant 0 \;\; \forall p \in \mathcal{R}_{V_h}\} = \\ &= \{v_h \in V_h : \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j v_h(z_{j,i})\, z_{j,i} \geqslant 0\}.\end{aligned}$$

Lemma 3.9. (Necessary condition for the existence of a solution to $(\mathcal{P}_h(e_h))$.) *Let there exist a solution of $(\mathcal{P}_h(e_h))$, then the condition*

$$F^h(x) \;\geqslant\; 0 \qquad (\mathrm{S3}_h)$$

must be fulfilled.

Proof. If we insert $v_h = p \in \mathcal{R}_{V_h}$ into $(\mathcal{P}_h'(e_h))$, we directly obtain the assertion of the lemma.

■

The space V_h can be uniquely decomposed into the orthogonal sum $\mathcal{R}_{V_h} \oplus \mathcal{R}^{\ominus}_{V_h}$. Moreover $\forall v_h \in V_h \; \exists! \; \{p, \bar{v}_h\} \in \mathcal{R}_{V_h} \times \mathcal{R}^{\ominus}_{V_h}$ such that

$$v_h \;=\; p \oplus \bar{v}_h, \quad ((p, \bar{v}_h))_{2,\Omega} \;=\; a \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \bar{v}_h(z_{j,i}) z_{j,i} \;=\; 0. \tag{108}$$

It is clear that only one of the following alternatives can occur:

$$a \;=\; 0 \quad \text{and} \quad \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \bar{v}_h(z_{j,i}) z_{j,i} \;\geqslant\; 0, \tag{A3$_h$}$$

$$a \;\leqslant\; 0 \quad \text{and} \quad \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \bar{v}_h(z_{j,i}) z_{j,i} \;=\; 0. \tag{A4$_h$}$$

Lemma 3.10. (Poincaré type inequality) *Let $V = \{v \in H^2(\Omega) : \; v(0) = 0\}$, then there exists a positive constant c_P dependent only on the interval Ω such that*

$$\|v\|^2_{2,2,\Omega} \leqslant c_P \left(|v|^2_{2,2,\Omega} \;+\; \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j v(z_{j,i}) z_{j,i} \right)^2 \right) \quad \forall v \in V, \tag{109}$$

where $\omega_j > 0$ are wages and $z_{j,i} \in Q_h$ nodes of the quadrature formula (176).

Proof. The assertion can be proved exactly in the same way as Lemma 2.20. ■

Lemma 3.11. *Let the condition*

$$F^h(x) \;>\; 0 \tag{S4$_h$}$$

be satisfied. Then the functional $\mathcal{E}^h_{e_h}$ is coercive on V_h.

Proof. Let us suppose that (S4$_h$) is fulfilled. If the alternative (A3$_h$) occurs, then $a = 0$, $v_h \equiv \bar{v}_h$ and $\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{m} \omega_j \bar{v}_h(z_{j,i}) z_{j,i} \geqslant 0$. For arbitrary function $\bar{v}_h \in \mathcal{R}^{\ominus}_{V_h}$ it holds the following inequality:

$$\begin{aligned} l^3 \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \left(\bar{v}^+_h(z_{j,i}) \right)^2 \geqslant \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \bar{v}^+_h(z_{j,i}) z_{j,i} \right)^2 \geqslant \\ \geqslant \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j \bar{v}_h(z_{j,i}) z_{j,i} \right)^2, \end{aligned} \tag{110}$$

where we make use of $b^h_{q_h} \in \mathcal{Q}^0_h$ and the discrete Cauchy-Schwarz inequality (see Theorem 8.2). Using $F^h \in \mathcal{Q}^0_h$, (109) and (110) the functional $\mathcal{E}^h_{e_h}$ can be rewritten as follows:

$$\begin{aligned} 2\mathcal{E}^h_{e_h}(v_h) = 2\mathcal{E}^h_{e_h}(\bar{v}_h) = a_{t_h}(\bar{v}_h, \bar{v}_h) + b^h_{q_h}(\bar{v}^+_h, \bar{v}^+_h) - 2F^h(\bar{v}_h) \geqslant \\ \geqslant \beta_0 t_0^3 \, |\bar{v}_h|^2_{2,2,\Omega} + \frac{q_0}{l^3}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j \bar{v}_h(z_{j,i}) z_{j,i}\right)^2 - 2F^h(\bar{v}_h) \geqslant \\ \geqslant \|\bar{v}_h\|_{2,2,\Omega}\left(c_1 \|\bar{v}_h\|_{2,2,\Omega} - 2c_2\right), \end{aligned}$$

where $c_1 := (1/c_P)\min\{\beta_0 t_0^3, \, q_0/l^3\}$ and c_2 is the constant from Lemma 2.16.

In the case ($A2_h$), we have $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\omega_j \bar{v}_h(z_{j,i}) z_{j,i} = 0$ and $a \leqslant 0$. Then we obtain:

$$\begin{aligned} 2\mathcal{E}^h_{e_h}(v_h) = 2\mathcal{E}^h_{e_h}(p + \bar{v}_h) = a_{t_h}(\bar{v}_h, \bar{v}_h) + b^h_{q_h}(v^+_h, v^+_h) - 2F^h(p) - 2F^h(\bar{v}_h) \geqslant \\ \geqslant \beta_0 t_0^3 \, |\bar{v}_h|^2_{2,2,\Omega} + q_0 \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j \left(v^+_h(z_{j,i})\right)^2 + 2|a|F^h(x) - 2F^h(\bar{v}_h) \geqslant \\ \geqslant \beta_0 t_0^3 \, |\bar{v}_h|^2_{2,2,\Omega} + \left(\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_j \bar{v}_h(z_{j,i}) z_{j,i}\right)^2 + 2|a|F^h(x) - 2F^h(\bar{v}_h) \geqslant \\ \geqslant c_1 \|\bar{v}_h\|^2_{2,2,\Omega} + 2|a|F^h(x) - 2c_2 \|\bar{v}_h\|_{2,2,\Omega}, \end{aligned}$$

denoting $c_1 := (1/c_P)\min\{\beta_0 t_0^3, \, 1\}$. Taking into account the orthogonality of (108) we have $\|v_h\|^2_{2,2,\Omega} = \|\bar{v}_h\|^2_{2,2,\Omega} + \|p\|^2_{2,2,\Omega}$. Thus $\|v_h\|_{2,2,\Omega} \to +\infty$ implies that at least one part of the function $v_h = \bar{v}_h + ax$ converges to $+\infty$ in appropriate norm. Making use of the assumption ($S4_h$) the assertion is proved.

■

Theorem 3.4. (Necessary and sufficient condition for the existence and uniqueness of a solution to $(\mathcal{P}_h(e_h))$.) *There exists a unique solution of $(\mathcal{P}_h(e_h))$ if and only if ($S4_h$) is fulfilled. In addition for such a solution $u_h \in V_h$ it holds*

$$M_{u_h} = \{z_{j,i} \in Q_h : \, u_h(z_{j,i}) > 0\} \neq \emptyset. \qquad (M2_h)$$

Proof. *Necessity.* Let $u_h \in V_h$ be a unique solution of $(\mathcal{P}_h(e_h))$ and let ($S4_h$) be not satisfied. Due to Lemma 3.9 we have $F^h(x) = 0$. Inserting $v_h \equiv ax \in \mathcal{R}_{V_h}$ into $(\mathcal{P}'_h(e_h))$ reads

$$b^h_{q_h}(u^+_h, x) = 0. \qquad (111)$$

Thus $u^+_h(z_{j,i}) = 0, \; u_h(z_{j,i}) \leqslant 0 \; \forall z_{j,i} \in Q_h$. Then $\forall p \in \mathcal{R}_{V_h}, \; ax \neq 0$ it holds that $u_h(z_{j,i}) + ax \leqslant 0 \, \forall z_{j,i} \in Q_h$. From there $b^h_{q_h}((u_h + ax)^+, v_h) = 0 \; \forall p \in \mathcal{R}_{V_h}, \; ax \neq 0$ and $\forall v_h \in V_h$. Thus $u_h + ax$ is a solution to $(\mathcal{P}_h(e_h))$ being in contradiction with the uniqueness of u_h. Therefore the condition ($S4_h$) must be fulfilled.

Sufficiency. Let the condition ($S4_h$) hold. From Lemma 2.18 and Lemma 3.11 it follows that $\mathcal{E}^h_{e_h}$ is Gâteaux differentiable, convex and coercive on V_h, thus the existence of a solution $u_h \in V_h$ is ensured, see e.g. [12].

Further we prove that the solution can be characterized by ($M2_h$). Let $u_h \in V_h$, $u_h(z_{j,i}) \leqslant 0$ $\forall z_{j,i} \in Q_h$ be a solution of $(\mathcal{P}_h(e_h))$. By inserting $v_h \equiv p \in \mathcal{R}_{V_h}$, $p \neq 0$ into $(\mathcal{P}'_h(e_h))$ we obtain

$$0 \;=\; b^h_{q_h}(u^+_h, x) = F^h(x). \tag{112}$$

But (112) is in contradiction with ($S4_h$), therefore u_h satisfies ($M2_h$).

Let us assume that there exist solutions $u_{h,1}$, $u_{h,2} \in V_h$ of $(\mathcal{P}_h(e_h))$. Subtracting corresponding weak formulations and setting $v = u_{h,1} - u_{h,2}$ yield

$$a_{t_h}(u_{h,1} - u_{h,2}, u_{h,1} - u_{h,2}) \;+\; b^h_{q_h}(u^+_{h,1} - u^+_{h,2}, u_{h,1} - u_{h,2}) \;=\; 0. \tag{113}$$

In view of definitions of a_{t_h}, $b^h_{q_h}$ it holds

$$u_{h,1} - u_{h,2} \;=\; ax \in P_1 \quad \text{and} \quad u^+_{h,1}(z_{j,i}) - (u_{h,1}(z_{j,i}) - ax)^+ \;=\; 0 \;\; \forall z_{j,i} \in Q_h.$$

making use of ($M2_h$) we obtain $a = 0$ and $u_{h,1} = u_{h,2}$ in Ω. ■

3.2.4. Existence of solutions to (P_h)

Similarly as in the previous case we will show that $u_h(e_h)$ depends continuously on e_h and that (P_h) has at least one solution.

Let $\beta \in L^\infty(\Omega)$, $0 < \beta_0 \leqslant \beta(x)$ a.e. in Ω and $F^h \in S_{h,\delta}$, where

$$S_{h,\delta} = \{F^h \in V^*_h : \; F^h(x) \geqslant \delta > 0\}.$$

We know that for any $e_h \in U^h_{ad}$ there exists a unique solution to $(\mathcal{P}_h(e_h))$ with the property ($M2_h$). We will denote the set (for fixed $h > 0$) of all such solutions again by W_h:

$$W_h := \{\{u_h, t_h, q_h\} \in V_h \times U^t_{ad,h} \times U^q_{ad,h} \;:\; u_h = u_h(e_h) \text{ solves } (\mathcal{P}_h(e_h))\}.$$

In order to use the next lemma in the convergence analysis we will consider whole class of problems $(\mathcal{P}_h(e_h))$, $0 < h \leqslant h_0$. Therefore we assume $F^h \in S_{h,\delta}$ for each $0 < h \leqslant h_0$.

Lemma 3.12. *There exists a positive constant $c_1 := c_1(\delta)$ such that*

$$c_1 \;\|u_h\|^2_{2,2,\Omega} \;\leqslant\; a_{t_h}(u_h, u_h) \;+\; b^h_{q_h}(u^+_h, u_h) \qquad \forall \{u_h, t_h, q_h\} \in \bigcup_{0<h\leqslant h_0} W_h. \tag{114}$$

The constant c_1 does not depend on $\{u_h, t_h, q_h\} \in \bigcup\limits_{0<h\leqslant h_0} W_h$.

Proof. Let us suppose that (114) does not hold. Then one can find a sequence $\{u_{h_k}, t_{h_k}, q_{h_k}\} \subset \bigcup_{0<h\leqslant h_0} W_h$ such that

$$\frac{1}{k}\,\|u_{h_k}\|^2_{2,2,\Omega} > a_{t_{h_k}}(u_{h_k}, u_{h_k}) + b^{h_k}_{q_{h_k}}(u^+_{h_k}, u_{h_k}) \geqslant 0 \quad \forall k \geqslant 1. \tag{115}$$

Dividing the inequality (115) by $\|u_{h_k}\|^2_{2,2,\Omega}$ and passing to the limit for $k \to \infty$ lead to

$$\lim_{k\to\infty} a_{t_{h_k}}(w_{h_k}, w_{h_k}) = 0 \quad \text{and} \quad \lim_{k\to\infty} b^{h_k}_{q_{h_k}}(w^+_{h_k}, w_{h_k}) = 0,$$

where we denote $w_{h_k} := u_{h_k}/\|u_{h_k}\|_{2,2,\Omega}$. Clearly $\|w_{h_k}\|_{2,2,\Omega} = 1$ and $\{w_{h_k}\}$ is bounded. Hence there exists a subsequence of $\{w_{h_k}\}$ (denoted by the same sequence) and an element $w_h \in V$, $h > 0$ such that $w_{h_k} \rightharpoonup w_h$ in V. Without loss of generality we may suppose that $h_k \to h$ for $k \to \infty$. In a similar manner as in the proof of Lemma 2.9 we can show that $|w_h|^2_{2,2,\Omega} = 0$, $w_h \equiv p = ax \in P_1$ and $w_{h_k} \to p$ in Ω. Due to embedding theorems (see e.g. [27]) it holds that $w_{h_k} \rightrightarrows ax$ in Ω. We know that

$$0 \leqslant b^h_{q_0}(p^+, p^+) = \lim_{k\to\infty} b^{h_k}_{q_0}(w^+_{h_k}, w^+_{h_k}) \leqslant \lim_{k\to\infty} b^{h_k}_{q_{h_k}}(w^+_{h_k}, w^+_{h_k}) = 0 \tag{116}$$

which implies $a \leqslant 0$. We shall show now that we can find a point $z_P > 0$ such that $p(z_P) = az_P \geqslant 0$. From $F^{h_k} \in S_{h_k,\delta}$ it follows that there exist $\varepsilon = \varepsilon(\delta) > 0$ such that for each $u_{h_k} \in \bigcup_{0<h\leqslant h_0} W_h$ it exists at least one $z^k \in Q_{h_k}$, $z^k \in (\varepsilon, l]$ satisfying $u_{h_k}(z^k) > 0$ and $w_{h_k}(z^k) > 0$. Without loss of generality we may suppose that $z^k \to z_P$ for $k \to \infty$. Then $w_{h_k}(z^k) \to az_P \geqslant 0$. Therefore we obviously have $a = 0$ being in contradiction with $1 = \|w_{h_k}\|_{2,2,\Omega}$ and $w_{h_k} \to ax$ in Ω.

■

We proceed by the continuous dependence of the solution $u_h(e_h)$ on the approximated design variable e_h.

Lemma 3.13. (Continuous dependence.) *Let $e_{h,n}$, $e_h \in U^h_{ad}$, $e_{h,n} \to e_h$ in U_{ad}. Further let $u_{h,n} := u_h(e_{h,n}) \in V_h$ be a solution to $(\mathcal{P}_h(e_{h,n}))$ and let the condition $(S4_h)$ be fulfilled. Then there exists a function $u_h \in V_h$ such that*

$$u_{h,n} \to u_h \ \text{ in } V$$

and moreover $u_h = u_h(e_h)$ is a solution to $(\mathcal{P}_h(e_h))$.

Proof. Using Lemma 3.12 we can easily prove the boundedness of $\{u_{h,n}\}$ in V. Therefore one can find a subsequence (denoted by the same sequence) such that $u_{h,n} \rightharpoonup u_h$ in V_h. In the same way as in the proof of Lemma 2.23 we show that u_h solves $(\mathcal{P}_h(e_h))$ and that $u_{h,n} \to u_h$ in V.

■

Lemma 3.14. *There exists a constant* $c_2 := c_2(\delta) > 0$ *such that* $\forall \{u_{h,i}, t_{h,i}, q_{h,i}\} \in W_h, i = 1, 2$ *it holds*

$$c_2 \, \|u_{h,1} - u_{h,2}\|^2_{2,2,\Omega} \leqslant a_{t_{h,1}}(u_{h,1} - u_{h,2}, u_{h,1} - u_{h,2}) + \\ + \, b^h_{q_{h,1}}(u^+_{h,1} - u^+_{h,2}, u_{h,1} - u_{h,2}). \tag{117}$$

The constant c_2 *does not depend on* $\{u_{h,i}, t_{h,i}, q_{h,i}\} \in W_h, i = 1, 2$.

Proof. Assume that (117) does not hold. Then one can find sequences $\{u_{h,1,n}, t_{h,1,n}, q_{h,1,n}\}$, $\{u_{h,2,n}, t_{h,2,n}, q_{h,2,n}\} \subset W_h$ such that

$$\frac{1}{n} \|u_{h,1,n} - u_{h,2,n}\|^2_{2,2,\Omega} > a_{t_{h,1,n}}(u_{h,1,n} - u_{h,2,n}, u_{h,1,n} - u_{h,2,n}) + \\ + \, b^h_{q_{h,1,n}}(u^+_{h,1,n} - u^+_{h,2,n}, u_{h,1,n} - u_{h,2,n}) \geqslant 0 \quad \forall n \geqslant 1. \tag{118}$$

Due to Lemma 3.13, sequences $\{u_{h,1,n}\}$, $\{u_{h,2,n}\}$ are bounded in $H^2(\Omega)$. Then one can find its subsequences (denoted by the same sequences) and functions $\hat{u}_{h,1}$, $\hat{u}_{h,2}$ such that $u_{h,i,n} \rightharpoonup \hat{u}_{h,i}$ in $H^2(\Omega)$, $i = 1, 2$. In fact it holds $u_{h,i,n} \rightrightarrows \hat{u}_{h,i}$ in Ω, $i = 1, 2$. By setting $v_h = x$ in $(\mathcal{P}_h(e_{h,1,n}))$ and passing to the limit for $n \to \infty$ we have

$$q_1 l \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j u^+_{h,1}(z_{j,i}) = q_1 l \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_j u^+_{h,1,n}(z_{j,i}) \geqslant \\ \geqslant \lim_{n\to\infty} \sum_{i=1}^{n} q_{h,1,n} \sum_{j=1}^{m} \omega_j u^+_{h,1,n}(z_{j,i}) z_{j,i} = F^h(x) > 0.$$

A similar estimate holds also for $\{u_{h,2,n}\}$. Hence, we can find points z_1, $z_2 \in Q_h$ such that $u_{h,1,n}(z_1) > 0, \hat{u}_{h,1}(z_1) > 0$ resp. $u_{h,2,n}(z_2) > 0, \hat{u}_{h,2}(z_2) > 0$ for n large enough.

Let us now divide (118) by $\|u_{h,1,n} - u_{h,2,n}\|^2_{2,2,\Omega}$ and use the following notation $w_{h,i,n} := u_{h,i,n} / \|u_{h,1,n} - u_{2,n}\|_{2,2,\Omega}$, $i = 1, 2$. According to Lemma 2.24 there exists $w_h \in V_h$ such that $w_{h,1,n} - w_{h,2,n} \to w_h$ in V_h and $|w_h|^2_{2,2,\Omega} = 0$, $w_h \equiv p = ax \in P_1$. Thus

$$w_{h,1,n} - w_{h,2,n} \to ax \text{ in } H^2(\Omega) \quad \text{and} \quad w^+_{h,1,n}(z_{j,i}) - w^+_{h,2,n}(z_{j,i}) \to 0 \quad \forall z_{j,i} \in Q_h.$$

Following the approach presented in Lemma 2.24 we again have that $a = 0$ what is in contradiction with $1 = \|w_{h,1,n} - w_{h,2,n}\|_{2,2,\Omega}$ and $w_{h,1,n} - w_{h,2,n} \to ax$ in $H^2(\Omega)$. ■

Making use of the Lemma 3.14 we can prove the Lipschitz continuity of the mapping $e_h \longmapsto u_h(e_h)$, where $u_h(e_h)$ solves $(\mathcal{P}_h(e_h))$, i.e. there exists $K_1 > 0$ such that $\forall e_{h,1} = \{t_{h,1}, q_{h,1}\}, e_{h,2} = \{t_{h,2}, q_{h,2}\} \in U^h_{ad}$ it holds

$$\|u_h(e_{h,1}) - u_h(e_{h,2})\|_{2,2,\Omega} \leqslant K_1 \left(\|t_{h,1} - t_{h,2}\|_{C(\bar{\Omega})} + \|q_{h,1} - q_{h,2}\|_{2,\Omega} \right). \tag{119}$$

Theorem 3.5. (Existence of solutions to (P_h)) *Let I_h satisfy ($\mathrm{I1}_h$), then (P_h) has at least one solution for every $h > 0$.*

Proof. The assertion follows from ($\mathrm{I1}_h$), Lemma 2.15 and Lemma 3.13. ■

Lemma 3.15. *Let I_h satisfy ($\mathrm{I2}_h$). Then the functional $J_h(e_h) = I_h(e_h, u_h(e_h))$, with $u_h(e_h)$ being a solution to $(\mathcal{P}_h(e_h))$, is Lipschitz continuous in U_{ad}^h, i.e. there exists $K_2 > 0$ such that $\forall e_{h,1}, e_{h,2} \in U_{ad}^h$:*

$$|J_h(e_{h,1}) - J_h(e_{h,2})| \leqslant K_2 \left(\|t_{h,1} - t_{h,2}\|_{C(\bar{\Omega})} + \|q_{h,1} - q_{h,2}\|_{2,\Omega} \right).$$

Proof. The assertion directly follows from ($\mathrm{I2}_h$) and (119). ■

3.3. Convergence analysis

Lemma 3.16. *Let $u_h := u_h(e_h) \in V_h$ be a solution to $(\mathcal{P}_h(e_h))$ and let ($\mathrm{S4}_h$) be satisfied for every $0 < h \leqslant h_0$. Then there exists a function $u \in V$ such that*

$$u_h \to u \ \text{ in } V$$

and in addition $u = u(e)$ is a solution to $(\mathcal{P}(e))$.

Proof. The boundedness of $\{u_h\}$ can be proved by using ($\mathrm{S4}_h$), Lemma 3.12 and (58). Thus one can find a subsequence of $\{u_h\}$ (denoted by the same sequence) and a function $u \in V$ such that

$$u_h \rightharpoonup u \quad \text{in } V \text{ as } h \to 0^+. \tag{120}$$

In the same manner as in the proof of Lemma 2.27 we show that u solves $(\mathcal{P}(e))$ and $u_h \to u$. ■

Theorem 3.6. *Let I, I_h satisfy ($\mathrm{I3}_h$), ($\mathrm{I4}_h$). Then for arbitrary sequence $\{e_h^*\}$, where $e_h^* \in U_{ad}^h$ is an optimal solution to (P_h) and $u_h(e_h^*)$ is a solution of $(\mathcal{P}_h(e_h^*))$, one can find a subsequence $\{e_{h_j}^*\}$ such that*

$$e_{h_j}^* \to e^* \ \text{ in } U_{ad}, \tag{121}$$

$$u_{h_j}(e_{h_j}^*) \to u(e^*) \ \text{ in } V, \tag{122}$$

where e^* is an optimal solution of (P) with $u(e^*)$ being a solution to $(\mathcal{P}(e^*))$.

Proof. See the proof of Theorem 2.6. ■

4. Numerical realization

In the first part of this section we will show one of the possible approaches for solving of problem (P_h), where the solution of the state problem ($\mathcal{P}_h(e_h)$) is based on transformation of a system of nonlinear algebraic equations into a mixed linear complementarity problem, see [36]. The second part of the section will be devoted to design sensitivity analysis that in general deals with differentiability in shape design optimization problems, see e.g. [16], [17] or [48].

4.1. Algebraic formulation of (P_h)

Let us start with design variables. From (50), (51) it follows that t_h and q_h can be uniquely determined by $n+1$ and n – dimensional vectors, respectively. The thickness will be represented by nodal values of t_h, i.e. $t_i = t_h(x_i)$, $i = 1, \dots, n+1$. Vector representing the stiffness consists of values of q_h on subintervals K_i, i.e. $q_i = q_h|_{K_i}$, $i = 1, \dots, n$. Then the algebraic form of U_{ad}^h is

$$\mathcal{U}^h = \mathcal{U}_h^t \times \mathcal{U}_h^q, \tag{123}$$

where

$$\mathcal{U}_h^t = \left\{ \mathbf{t} \in \mathbb{R}^{n+1} : t_0 \leqslant t_i \leqslant t_1, \sum_{i=1}^{n} \frac{h}{2}(t_i + t_{i+1}) = \gamma_1, \ |t_{i+1} - t_i| \leqslant h\gamma_2 \right\},$$

$$\mathcal{U}_h^q = \left\{ \mathbf{q} \in \mathbb{R}^{n+1} : q_0 \leqslant q_i \leqslant q_1, \sum_{i=1}^{n} h q_i = \gamma_3, \right\}.$$

Pair $\{\mathbf{t}, \mathbf{q}\}$ will be denoted by $\mathbf{e}$. Notice that for the numerical realization we have added the constraint $\sum_{i=1}^{n} h q_i = \gamma_3$ into $\mathcal{U}_h^q$. It has no influence on the existence and convergence analysis, but it is important for the practical computations. It prevents the stiffness from jumping to the upper bound q_1.

Now we can pass to the algebraic form of ($\mathcal{P}_h(e_h)$) and its transformation into a mixed linear complementarity problem. Unfortunately the standard finite element method (see [7], [18], [55], [56] or [49]) does not work with functions u_h^+ and therefore the current form of ($\mathcal{P}_h(e_h)$) is not suitable for the numerical realization. But we can modify the problem by using

$$q_h(x)\, u_h^+(x) = \bar{q}_h(x)\, u_h(x) \qquad \forall x \in \Omega,$$

where

$$\bar{q}_h(x) = \begin{cases} q_h(x) & u_h(x) > 0, \\ 0 & u_h(x) \leqslant 0. \end{cases}$$

The bilinear form $b^h_{q_h}$ can be then rewritten as follows:

$$b^h_{q_h}(u^+_h,\, v_h) = \sum_{i=1}^{n}\sum_{j=1}^{m} \omega_j \bar{q}_h(z_{j,i}) u_h(z_{j,i}) v_h(z_{j,i}).$$

Now we can use the standard finite element method. But notice that the function $q_h(x)$ is here replaced by a piecewise constant unknown function $\bar{q}_h(x)$. Since V_h is finite dimensional we can define its finite Hermite basis $\{\varphi_{2i-1},\, \varphi_{2i}\}$, $i = 1, \ldots, n+1$, see e.g. [7], [36] or [55].

$$\varphi_{2i-1}(x) = \begin{cases} 0 & x < x_{i-1}, \\ -\frac{2}{h^3}(x - x_{i-1})^2(x - x_i - h/2) & x \in [x_{i-1}, x_i], \\ \frac{2}{h^3}(x - x_{i+1})^2(x - x_i + h/2) & x \in [x_i, x_{i+1}], \\ 0 & x > x_{i+1}, \end{cases} \tag{124}$$

$$\varphi_{2i}(x) = \begin{cases} 0 & x < x_{i-1}, \\ \frac{1}{h^2}(x - x_{i-1})^2(x - x_i) & x \in [x_{i-1}, x_i], \\ \frac{1}{h^2}(x - x_{i+1})^2(x - x_i) & x \in [x_i, x_{i+1}], \\ 0 & x > x_{i+1}. \end{cases} \tag{125}$$

Each function $u_h \in V_h$ can be written as a linear combination of basis functions in the form $u_h = \sum_{i=1}^{n+1} u_i\, \varphi_{2i-1} + u'_i\, \varphi_{2i}$. Using the well known Galerkin method we obtain the following system of nonlinear algebraic equations for unknown coefficients u_i, u'_i:

$$\mathbf{K}\mathbf{u} + \mathbf{P}^T\mathbf{Q}(\mathbf{u})\mathbf{P}\mathbf{u} = \mathbf{F}, \tag{126}$$

where $\mathbf{u} = \{u_1, u'_1, \ldots, u_{n+1}, u'_{n+1}\}$, $\mathbf{K} \in \mathbb{R}^{N\times N}$ is the stiffness matrix of the beam, $\mathbf{P} \in \mathbb{R}^{M\times N}$ is a matrix that transforms the function values and the values of the first derivatives in the nodal points x_i, onto nodes $z_{j,i} \in Q_h$, $j = 1, \ldots, m$, $i = 1, \ldots, n$ and $\mathbf{Q}(\mathbf{u}) \in \mathbb{R}^{M\times M}$ is a diagonal matrix containing products of the weights of the numerical quadrature and the stiffness coefficients of the subsoil. We denote $N = 2n + 2$, $M = nm$. $\mathbf{F} \in \mathbb{R}^N$ is a vector corresponding to the load of the beam.

In order to simplify the notation we will denote the nodes $z_{j,i} \in Q_h$ by z_k and corresponding wages by ω_k, $k = 1, \ldots, M$, then

$$\begin{aligned} \mathbf{K}(i,j) &= \int_0^l \beta\, t_h^3 \varphi_i{}'' \varphi_j{}'' \mathrm{d}x, \qquad i,j = 1,\ldots,N, \\ \mathbf{P}(k,i) &= \varphi_i(z_k), \qquad i = 1,\ldots,N,\ k = 1,\ldots,M, \\ \mathbf{Q}(\mathbf{u})(k,k) &= \omega_k \bar{q}_h(z_k), \qquad k = 1,\ldots,M. \\ \mathbf{F}(i) &= F^h(\varphi_i), \qquad i = 1,\ldots,N. \end{aligned}$$

Let the polynomial $-1 \in \mathcal{R}_{V_h}$ (for the first case of boundary conditions) be represented by vector $p_1 \in \mathbb{R}^N$, then p_1 represents all polynomials belonging to $\mathcal{R}_{V_h}$. In addition it holds that $\mathbf{K}p_1 = 0$. This vector creates the kernel of $\mathbf{K}$. Similarly $-x \in \mathcal{R}_{V_h}$ (for the second case of boundary conditions) is represented by vector $p_x \in \mathbb{R}^N$. It represents all polynomials belonging to $\mathcal{R}_{V_h}$ and moreover $\mathbf{K}p_x = 0$. From the construction of matrices $\mathbf{K}$ and $\mathbf{Q}$ is clear that mappings $\mathbf{e} \longmapsto \mathbf{K}(\mathbf{e})$ and $\mathbf{e} \longmapsto \mathbf{Q}(\mathbf{e})$ are continuous.

The necessary and sufficient condition for the existence and uniqueness of a solution to (126) has the following algebraic form:

$$p_1^T\mathbf{F} < 0, \qquad p_x^T\mathbf{F} \leqslant 0.$$

Every solution then can be characterized by $M_{\mathbf{u}} = \{i \in \mathcal{I} : (\mathbf{P}\mathbf{u})_i > 0\} \neq \emptyset$, where

$$\mathcal{I} = \{1, \ldots, M.\}$$

Now we know that $\forall \mathbf{e} \in \mathcal{U}^h$ there exists a solution of the nonlinear system (126) and that the mapping $\mathbf{e} \longmapsto \mathbf{u}(\mathbf{e})$ is Lipschitz continuous.

In practical computations there are usually used shape functions $\mathcal{N}_i(\xi)$, $i = 1, \ldots, 4$ (see e.g. [36]). Using the transformation (178) we obtain shape functions on the reference interval $(0, 1)$ in the following form:

$$\begin{aligned}
\mathcal{N}_1(\xi) &= \frac{1}{4}(1-\xi)^2(2+\xi) \\
\mathcal{N}_2(\xi) &= \frac{h}{8}(1-\xi)^2(1+\xi) \\
\mathcal{N}_3(\xi) &= \frac{1}{4}(1+\xi)^2(2-\xi) \\
\mathcal{N}_4(\xi) &= -\frac{h}{8}(1+\xi)^2(1-\xi).
\end{aligned}$$

Matrices $\mathbf{K}$, $\mathbf{K}_f(\mathbf{u}) = \mathbf{P}^T\mathbf{Q}(\mathbf{u})\mathbf{P}$ can be assembled from element matrices $\mathbf{K}^{(i)} \in \mathbb{R}^{4\times4}, \mathbf{K}_f^{(i)}(\mathbf{u}) \in \mathbb{R}^{4\times4}$, $i = 1, \ldots, n$ as it is usual in the finite element method.

$$\mathbf{K}^{(i)}(k,\, l) \;=\; \frac{8\beta}{h^3}\int_{-1}^{1} t_h^3(\xi)\,\mathcal{N}_k''(\xi)\,\mathcal{N}_l''(\xi)\,\mathrm{d}\xi,$$

$$\mathbf{K}_f^{(i)}(\mathbf{u})(k,l) \;=\; \frac{h}{2}\sum_{j=1}^{m} \bar{q}_h(\hat{z}_j)\,\hat{\omega}_j\,\mathcal{N}_k(\hat{z}_j)\,\mathcal{N}_l(\hat{z}_j), \qquad \mathbf{F}^{(i)}(l) \;=\; \frac{h}{2}\sum_{j=1}^{m} \hat{\omega}_j\,\hat{f}(\hat{z}_j)\,\mathcal{N}_l(\hat{z}_j),$$

where $\hat{\omega}_j$, $\hat{z}_j$ correspond to the reference quadrature formula (176) and $\hat{f}$ is a transformation of f onto $[-1,\, 1]$. The system (126) assembled in the way described above is nonlinear and due to the form of $\bar{q}_h(x)$ the dependence of $\mathbf{Q}$ on $\mathbf{u}$ is not continuously differentiable. In such a case we can not solve the problem

by standard methods like Newton's method or Quasi–Newton methods (see e.g. [4], [11], [13]).

In what follows we will introduce the approach based on transformation of (126) onto a mixed linear complementarity problem. This approach was firstly published in [36]. We know that the influence of the foundation for an element is represented in the model by coefficients that are given as follows:

$$\frac{h}{2}\int_{-1}^{1}\bar{q}_h(\xi)\,\mathcal{N}_k(\xi)\,\mathcal{N}_l(\xi)\;\mathrm{d}\xi \approx \frac{h}{2}\sum_{j=1}^{m}\bar{q}_h(\hat{z}_j)\,\hat{\omega}_j\,\mathcal{N}_k(\hat{z}_j)\,\mathcal{N}_l(\hat{z}_j).$$

Without loss of generality we will consider only the first element $K_1 = [0, h]$, then the i–th row of the stiffness matrix $\mathbf{K}_f^{(1)}(\mathbf{u})$ affects the global system as follows:

$$\begin{aligned}
&\frac{h}{2}\int_{-1}^{1}\bar{q}_h(\xi)\,\mathcal{N}_i(\xi)\,\mathcal{N}_1(\xi)\;\mathrm{d}\xi\;u_1 \;+\; \frac{h}{2}\int_{-1}^{1}\bar{q}_h(\xi)\,\mathcal{N}_i(\xi)\,\mathcal{N}_2(\xi)\;\mathrm{d}\xi\;u_1' \;+\\
&+\frac{h}{2}\int_{-1}^{1}\bar{q}_h(\xi)\,\mathcal{N}_i(\xi)\,\mathcal{N}_3(\xi)\;\mathrm{d}\xi\;u_2 \;+\; \frac{h}{2}\int_{-1}^{1}\bar{q}_h(\xi)\,\mathcal{N}_i(\xi)\,\mathcal{N}_4(\xi)\;\mathrm{d}\xi\;u_2' \;\approx \qquad (127)\\
&\approx \frac{h}{2}\sum_{j=1}^{m}\bar{q}_h(\hat{z}_j)\,\hat{\omega}_j\,\mathcal{N}_i(\hat{z}_j)\,u_h(\hat{z}_j).
\end{aligned}$$

Now it is necessary to decompose function values $u_h(\hat{z}_j)$ to their positive and negative parts. For each $j = 1, \ldots, m$ we define

$$u_h(\hat{z}_j) \;=\; v_j \;-\; w_j, \qquad (128)$$

where v_j, w_j are $u_h^+(\hat{z}_j)$ and $u_h^-(\hat{z}_j)$, respectively. Substituting (128) into (127) we obtain

$$\frac{h}{2}\sum_{j=1}^{m}\bar{q}_h(\hat{z}_j)\,\hat{\omega}_j\,\mathcal{N}_i(\hat{z}_j)\,u_h(\hat{z}_j) = \frac{h}{2}\sum_{j=1}^{m}q_h(\hat{z}_j)\,\hat{\omega}_j\,\mathcal{N}_i(\hat{z}_j)\,v_j. \qquad (129)$$

From (129) it is clear that we are now able to compute all the unknowns explicitly. To connect both parts of the decomposed system we have to add following equations describing the relation between original values $u_h(\hat{z}_j)$ and new variables v_j and w_j:

$$u_h(\hat{z}_j) - v_j + w_j = \mathcal{N}_1(\hat{z}_j)\;u_1 + \mathcal{N}_2(\hat{z}_j)\;u_1' + \mathcal{N}_3(\hat{z}_j)\;u_2 + \mathcal{N}_4(\hat{z}_j)\;u_2' - v_j + w_j \;=\; 0.$$

If we use a reference quadrature formula with integration points -1 and 1 then it is not necessary to add any more equations. In the previous chapter we considered a general quadrature formula. An efficient choice satisfying property $b_{q_h}^h \in \mathcal{Q}_h^0$

seems to be the 4-point Gauss–Lobatto formula with integration points ± 1, $\pm\frac{1}{\sqrt{5}}$ and wages $\frac{1}{6}$, $\frac{5}{6}$, which is exact for polynomials of degree 5 at most, see [20], [1] or [36]. The influence of the unilateral subsoil for the first element is then given by element matrices $\mathbf{S}^{(1)}$ a $\mathbf{D}^{(1)}$.

$$\mathbf{S}^{(1)}(j,i) = \frac{h}{2} q_h(\hat{z}_j)\,\hat{\omega}_j\,\mathcal{N}_i(\hat{z}_j), \quad i = 1,\dots,4,\ j = 1,\dots,m.$$
$$\mathbf{D}^{(1)}(j,j) = \frac{h}{2} q_h(\hat{z}_j)\,\hat{\omega}_j, \quad j = 1,\dots,m.$$

Matrices $\mathbf{S} \in \mathbb{R}^{M\times N}$ and $\mathbf{D} \in \mathbb{R}^{M\times M}$ can be obtained again by assembling of element matrices as it is usual for matrix $\mathbf{K}$. The nonlinear system (126) then transforms into the following problem of mixed linear complementarity:

$$\begin{pmatrix} \mathbf{K} & \mathbf{S}^T & \mathbf{0}^T \\ \mathbf{S} & -\mathbf{D} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{0} \end{pmatrix} \qquad \text{(mLCP(e))}$$
$$\mathbf{v}^T\mathbf{w} = 0, \quad \mathbf{v}, \mathbf{w} \geqslant 0,$$

where $\mathbf{v} = (v_1,\dots,v_m)$, $\mathbf{w} = (w_1,\dots,w_m)$. In fact we replaced the problematic part $\mathbf{P}^T\mathbf{Q}(\mathbf{u})\mathbf{P}\mathbf{u}$ of (126) by the linear product $\mathbf{S}^T\mathbf{v}$ and we added some equations describing the relation between the old variable $\mathbf{u}$ and new vectors $\mathbf{v}, \mathbf{w}$.

At the end of this subsection we will focus on the algebraic expression of the cost functional. In the previous sections we considered cost functionals approximated by a general formula for numerical integration satisfying $I_{h,i} \in \mathcal{Q}_h^0$, $i = 1,2,3$. Now we will approximate them using the trapezoidal formula for numerical integration, which satisfies all the requirements. Then

$$J_{h,1}(\mathbf{e}) = I_{h,1}(\mathbf{e},\,\mathbf{u}(\mathbf{e})) = \sum_{i=1}^{n+1} \frac{h}{2}\left(u_{i-1}\,f(x_{i-1}) + u_i\,f(x_i)\right) = \mathbf{u}^T\mathbf{B}\,\mathbf{f},$$
$$J_{h,2}(\mathbf{e}) = I_{h,2}(\mathbf{e},\,\mathbf{u}(\mathbf{e})) = \sum_{i=1}^{n+1} \frac{h}{2}\left(u_{i-1}^2 + u_i^2\right) = \mathbf{u}^T\mathbf{B}\,\mathbf{u},$$
$$J_{h,3}(\mathbf{e}) = I_{h,3}(\mathbf{e},\,\mathbf{u}(\mathbf{e})) = \sum_{i=1}^{n+1} \frac{h}{2}\left(t_{i-1}^2(u_h''(x_{i-1}))^2 + t_i^2(u_h''(x_i))^2\right) = \mathbf{u}^T\mathbf{\Phi}^T\mathbf{E}^T\mathbf{E}\,\mathbf{\Phi}\,\mathbf{u},$$

where $\mathbf{u}(\mathbf{e})$ is a solution to (mLCP(e)) and

$$\mathbf{B} = h\,\mathrm{diag}\,(1/2,\,0,\,1,\,0,\,1,\,0,\,\dots,\,1,\,0,\,1/2,\,0) \in \mathbb{R}^{(2n+2)\times(2n+2)},$$
$$\mathbf{f} = (f(x_0),\,0,\,f(x_1),\,0,\,\dots,\,f(x_n),\,0) \in \mathbb{R}^{2n+2},$$
$$\mathbf{E} = h\,\mathrm{diag}\left(\frac{1}{2}t_0,\,t_1,\,t_2,\,\dots,\,t_{n-1},\,\frac{1}{2}t_n\right) \in \mathbb{R}^{(n+1)\times(n+1)}.$$

The matrix $\mathbf{\Phi} \in \mathbb{R}^{(n+1)\times(2n+2)}$ is defined as follows:

$$\mathbf{\Phi}_{i,j} = \varphi_j''(x_i) \quad i = 0, \ldots, n,\ j = 1, \ldots, 2n+2.$$

The discrete optimization problem then turns to the following nonlinear programming problem:

$$\text{Find} \qquad \mathbf{e}^* \in \mathcal{U}^h : \ J_h(\mathbf{e}^*) \leqslant J_h(\mathbf{e}) \quad \forall \mathbf{e} \in \mathcal{U}^h. \tag{P_h}$$

4.2. Design sensitivity analysis

In this subsection we shall make the design sensitivity analysis which in general deals with differentiability in shape design optimization problems.

Every cost function evaluation requires solution of (mLCP(**e**)), which is costly. Therefore we would like to use as few evaluations of J_h as possible during the optimization process. Thus zero order methods (see e.g. [11], [4]) are not suitable as they usually use many function evaluations in each iteration to find a direction of decrease. We shall prepare the problem for application of a gradient (subgradient) optimization method. These methods usually proceed fast and do not require so many function evaluations.

We know that the mapping $\mathbf{e} \longmapsto \mathbf{u}(\mathbf{e})$ is Lipschitz continuous. Due to the fact that $\mathbf{v} = (\mathbf{P}\mathbf{u})^+$ and $\mathbf{w} = (\mathbf{P}\mathbf{u})^-$, the mappings $\mathbf{e} \longmapsto \mathbf{v}(\mathbf{e})$, $\mathbf{e} \longmapsto \mathbf{w}(\mathbf{e})$ are Lipschitz continuous as well.

From there by the Rademacher theorem (see e.g. [31]) it follows that these mappings are differentiable almost everywhere in $\mathcal{U}^h$. Next we prove that $\mathbf{u}, \mathbf{v}$ and $\mathbf{w}$ are in fact directionally differentiable.

Theorem 4.1. *A solution* $\{\mathbf{u}(\mathbf{e}), \mathbf{v}(\mathbf{e}), \mathbf{w}(\mathbf{e})\}$ *of the state problem* (mLCP(**e**)) *is directionally differentiable at any point* $\mathbf{e} \in \mathcal{U}^h$ *and in any direction* $\mathbf{d} \in \mathbb{R}^N$. *Moreover the directional derivatives* $\mathbf{u}'(\mathbf{e}) := \mathbf{u}'(\mathbf{e}; \mathbf{d})$, $\mathbf{v}'(\mathbf{e}) := \mathbf{v}'(\mathbf{e}; \mathbf{d})$, $\mathbf{w}'(\mathbf{e}) := \mathbf{w}'(\mathbf{e}; \mathbf{d})$ *can be computed from the following problem:*

$$\begin{pmatrix} \mathbf{K} & \mathbf{S}^T & \mathbf{0}^T \\ \mathbf{S} & -\mathbf{D} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{u}'(\mathbf{e}) \\ \mathbf{v}'(\mathbf{e}) \\ \mathbf{w}'(\mathbf{e}) \end{pmatrix} = \begin{pmatrix} -\mathbf{K}'(\mathbf{e}) & (\mathbf{S}')^T(\mathbf{e}) & \mathbf{0}^T \\ -\mathbf{S}'(\mathbf{e}) & \mathbf{D}'(\mathbf{e}) & -\mathbf{D}'(\mathbf{e}) \end{pmatrix} \begin{pmatrix} \mathbf{u}(\mathbf{e}) \\ \mathbf{v}(\mathbf{e}) \\ \mathbf{w}(\mathbf{e}) \end{pmatrix},$$

$$\mathbf{v}_i'(\mathbf{e})\, \mathbf{w}_i'(\mathbf{e}) = 0,\ \mathbf{v}_i'(\mathbf{e}) \geqslant 0,\ \mathbf{w}_i'(\mathbf{e}) \geqslant 0,\ i \in \mathcal{I}_0(\mathbf{e}),$$
$$\mathbf{w}_i'(\mathbf{e}) = 0,\ i \in \mathcal{I}_+(\mathbf{e}),\ \mathbf{v}_i'(\mathbf{e}) = 0,\ i \in \mathcal{I}_-(\mathbf{e}),$$

where the index sets $\mathcal{I}_+(\mathbf{e})$, $\mathcal{I}_-(\mathbf{e})$ *and* $\mathcal{I}_0(\mathbf{e})$ *are defined as follows:*

$$\mathcal{I}_+(\mathbf{e}) = \{i \in \mathcal{I} : \mathbf{v}_i(\mathbf{e}) > 0\},$$
$$\mathcal{I}_-(\mathbf{e}) = \{i \in \mathcal{I} : \mathbf{v}_i(\mathbf{e}) = 0,\ \mathbf{w}_i(\mathbf{e}) > 0\},$$
$$\mathcal{I}_0(\mathbf{e}) = \{i \in \mathcal{I} : \mathbf{v}_i(\mathbf{e}) = 0,\ \mathbf{w}_i(\mathbf{e}) = 0\}.$$

Proof. Let us denote by $\{\mathbf{u}(\mathbf{e}+\epsilon\mathbf{d}), \mathbf{v}(\mathbf{e}+\epsilon\mathbf{d}), \mathbf{w}(\mathbf{e}+\epsilon\mathbf{d})\}$ a solution of (mLCP($\mathbf{e}+\epsilon\mathbf{d}$)), where $\mathbf{e} \in \mathcal{U}^h$, $\mathbf{d} \in \mathbb{R}^N$ and $\epsilon > 0$. Next we define difference quotients

$$\frac{\mathbf{u}(\mathbf{e}+\epsilon\mathbf{d}) - \mathbf{u}(\mathbf{e})}{\epsilon}, \frac{\mathbf{v}(\mathbf{e}+\epsilon\mathbf{d}) - \mathbf{v}(\mathbf{e})}{\epsilon}, \frac{\mathbf{w}(\mathbf{e}+\epsilon\mathbf{d}) - \mathbf{w}(\mathbf{e})}{\epsilon}.$$

Due to the Lipschitz continuity of $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, these quotients are bounded for $\epsilon > 0$ and one can find a sequence $\{\epsilon_n\}_n$, $\epsilon_n \to 0^+$ and vectors $\dot{\mathbf{u}}$, $\dot{\mathbf{v}}$, $\dot{\mathbf{w}}$ such that

$$\frac{\mathbf{u}(\mathbf{e}+\epsilon_n\mathbf{d}) - \mathbf{u}(\mathbf{e})}{\epsilon_n} \to \dot{\mathbf{u}}, \frac{\mathbf{v}(\mathbf{e}+\epsilon_n\mathbf{d}) - \mathbf{v}(\mathbf{e})}{\epsilon_n} \to \dot{\mathbf{v}}, \frac{\mathbf{w}(\mathbf{e}+\epsilon_n\mathbf{d}) - \mathbf{w}(\mathbf{e})}{\epsilon_n} \to \dot{\mathbf{w}},$$

where $\dot{\mathbf{u}} = \dot{\mathbf{u}}(\{\epsilon_n\})$, $\dot{\mathbf{v}} = \dot{\mathbf{v}}(\{\epsilon_n\})$, $\dot{\mathbf{w}} = \dot{\mathbf{w}}(\{\epsilon_n\})$. We know that the mappings $\mathbf{e} \longmapsto \mathbf{K}(\mathbf{e})$, $\mathbf{e} \longmapsto \mathbf{S}(\mathbf{e})$ and $\mathbf{e} \longmapsto \mathbf{D}(\mathbf{e})$ are continuously differentiable, therefore the operation "$\dot{\cdot}$" coincide with the classical derivative for these mappings. Now we apply the operation "$\dot{\cdot}$" to the state problem (mLCP($\mathbf{e}$)):

$$\sum_{j=0}^{N} \mathbf{K}_{i,j}\, \dot{\mathbf{u}}_j + \sum_{j=0}^{M} \mathbf{S}_{j,i}\, \dot{\mathbf{v}}_j = \mathbf{F}'_i - \sum_{j=0}^{N} \mathbf{K}'_{i,j}\, \mathbf{u}_j - \sum_{j=0}^{M} \mathbf{S}'_{j,i}\, \mathbf{v}_j,$$

$$\sum_{j=0}^{N} \mathbf{S}_{i,j}\, \dot{\mathbf{u}}_j + \mathbf{D}_{i,i}\, (\dot{\mathbf{w}}_i - \dot{\mathbf{v}}_i) = \mathbf{D}'_{i,i}\, (\mathbf{v}_i - \mathbf{w}_i) - \sum_{j=0}^{N} \mathbf{S}'_{i,j}\, \mathbf{u}_j,$$

where $i = 1, \ldots, n$.

As the next step we will approach to the constraints. We will analyze their behavior for small parameter perturbations $\mathbf{e}+\epsilon\mathbf{d}$, $\epsilon \to 0^+$. Firstly let us suppose that $i \in \mathcal{I}_+(\mathbf{e})$. Then because the mapping $\mathbf{e} \longmapsto \mathbf{v}(\mathbf{e})$ is continuous, we have $\mathbf{v}_i(\mathbf{e}) > 0 \Rightarrow \mathbf{v}_i(\mathbf{e}+\epsilon\mathbf{d}) > 0$ and $i \in \mathcal{I}_+(\mathbf{e}+\epsilon\mathbf{d})$ for $|\epsilon| < \delta$, $\delta > 0$ small enough. From (mLCP($\mathbf{e}$)) it follows that $\mathbf{w}_i(\mathbf{e}) = 0$, $\mathbf{w}_i(\mathbf{e}+\epsilon\mathbf{d}) = 0$ and

$$\frac{\mathbf{w}_i(\mathbf{e}+\epsilon\mathbf{d}) - \mathbf{w}_i(\mathbf{e})}{\epsilon} = 0 \Rightarrow \dot{\mathbf{w}}_i(\mathbf{e}) = 0.$$

Further we consider $i \in \mathcal{I}_-(\mathbf{e})$. Then owing the continuity of mapping $\mathbf{e} \longmapsto \mathbf{w}(\mathbf{e})$ we obtain $\mathbf{w}_i(\mathbf{e}+\epsilon\mathbf{d}) > 0$ for $|\epsilon| < \delta$, $\delta > 0$ small enough. State problem (mLCP($\mathbf{e}$)) implies $\mathbf{v}_i(\mathbf{e}) = 0$, $\mathbf{v}_i(\mathbf{e}) + \epsilon\mathbf{d}) = 0$ and consequently $\dot{\mathbf{v}}_i(\mathbf{e}) = 0$ have to be fulfilled. The third case is $i \in \mathcal{I}_0(\mathbf{e})$. For all $\epsilon > 0$ it holds that $\mathbf{v}_i(\mathbf{e}+\epsilon\mathbf{d}) \geqslant \mathbf{v}_i(\mathbf{e}) = 0$ and $\mathbf{w}_i(\mathbf{e}+\epsilon\mathbf{d}) \geqslant \mathbf{w}_i(\mathbf{e}) = 0$. Therefore $\dot{\mathbf{v}}_i(\mathbf{e}) \geqslant 0$ and $\dot{\mathbf{w}}_i(\mathbf{e}) \geqslant 0$. Let us now show that

$$\dot{\mathbf{v}}_i(\mathbf{e})\, \dot{\mathbf{w}}_i(\mathbf{e}) = 0 \ \ \forall i \in \mathcal{I}. \tag{130}$$

For $i \in \mathcal{I}_+(\mathbf{e}) \cup \mathcal{I}_-(\mathbf{e})$ it is clear that (130) holds since either $\dot{\mathbf{v}}_i(\mathbf{e}) = 0$ or $\dot{\mathbf{w}}_i(\mathbf{e}) = 0$. It remains to prove the relation for $i \in \mathcal{I}_0(\mathbf{e})$. If $\dot{\mathbf{w}}_i(\mathbf{e}) = 0$ then (130) holds. If $\dot{\mathbf{w}}_i(\mathbf{e}) > 0$ then $\mathbf{w}_i(\mathbf{e}+\epsilon\mathbf{d}) > 0$, $\mathbf{v}_i(\mathbf{e}+\epsilon\mathbf{d}) = 0$ and consequently $\dot{\mathbf{v}}_i(\mathbf{e}) = 0$.

From the previous analysis it follows that $\dot{\mathbf{u}} = \dot{\mathbf{u}}(\{\epsilon_n\})$, $\dot{\mathbf{v}} = \dot{\mathbf{v}}(\{\epsilon_n\})$, $\dot{\mathbf{w}} = \dot{\mathbf{w}}(\{\epsilon_n\})$ satisfy the following system.

$$\sum_{j=0}^{N} \mathbf{K}_{i,j}\,\dot{\mathbf{u}}_j \;+\; \sum_{j=0}^{M} \mathbf{S}_{j,i}\,\dot{\mathbf{v}}_j = \mathbf{F}'_i \;-\; \sum_{j=0}^{N} \mathbf{K}'_{i,j}\,\mathbf{u}_j \;-\; \sum_{j=0}^{M} \mathbf{S}'_{j,i}\,\mathbf{v}_j,$$

$$\sum_{j=0}^{N} \mathbf{S}_{i,j}\,\dot{\mathbf{u}}_j \;+\; \mathbf{D}_{i,i}\,(\dot{\mathbf{w}}_i \;-\; \dot{\mathbf{v}}_i) = \mathbf{D}'_{i,i}\,(\mathbf{v}_i \;-\; \mathbf{w}_i) \;-\; \sum_{j=0}^{N} \mathbf{S}'_{i,j}\,\mathbf{u}_j,$$

$$\dot{\mathbf{v}}_i(\mathbf{e})\,\dot{\mathbf{w}}_i(\mathbf{e}) \;=\; 0,\; \dot{\mathbf{v}}_i(\mathbf{e}) \;\geqslant\; 0,\; \dot{\mathbf{w}}_i(\mathbf{e}) \;\geqslant\; 0,\; i \in \mathcal{I}_0(\mathbf{e}),$$

$$\dot{\mathbf{w}}_i(\mathbf{e}) \;=\; 0,\; i \in \mathcal{I}_+(\mathbf{e}),\; \dot{\mathbf{v}}_i(\mathbf{e}) \;=\; 0,\; i \in \mathcal{I}_-(\mathbf{e}).$$

Any accumulation point of $\frac{\mathbf{u}(\mathbf{e}+\epsilon\mathbf{d})-\mathbf{u}(\mathbf{e})}{\epsilon}$, $\frac{\mathbf{v}(\mathbf{e}+\epsilon\mathbf{d})-\mathbf{v}(\mathbf{e})}{\epsilon}$, $\frac{\mathbf{w}(\mathbf{e}+\epsilon\mathbf{d})-\mathbf{w}(\mathbf{e})}{\epsilon}$ has this property. Then if the difference quotients have a limit (from practical computations it follows that the limit exists) then $\dot{\mathbf{u}}(\{\epsilon_n\}) = \mathbf{u}'(\mathbf{e};\mathbf{d})$, $\dot{\mathbf{v}}(\{\epsilon_n\}) = \mathbf{v}'(\mathbf{e};\mathbf{d})$, $\dot{\mathbf{w}}(\{\epsilon_n\}) = \mathbf{w}'(\mathbf{e};\mathbf{d})$ and the assertion is proved.

■

Despite the fact that the control state mappings are directionally differentiable they need not to be continuously differentiable in $\mathcal{U}^h$. If $\mathcal{I}_0(\mathbf{e})$ is nonempty then $\mathbf{u}'(\mathbf{e};\,\mathbf{d})$, $\mathbf{v}'(\mathbf{e};\,\mathbf{d})$ and $\mathbf{w}'(\mathbf{e};\,\mathbf{d})$ are nonlinear in $\mathbf{d}$ and therefore $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are only directionally differentiable at $\mathbf{e}$. Their Lipschitz continuity implies that there exists at least one subgradient of these mappings at any $\mathbf{e} \in \mathcal{U}^h$. The mapping J_h as a composite mapping of $\mathbf{e} \mapsto \mathbf{u}(\mathbf{e})$ and $I_h(\mathbf{e}, \mathbf{u}(\mathbf{e}))$ is Lipschitz continuous and possibly nonsmooth in $\mathcal{U}^h$. Therefore to solve (P_h) one needs to use suitable nonsmooth optimization method.

The evaluation of J_h involves computing of a solution to the nonlinear problem (mLCP(**e**)). Consequently, the optimization algorithm should use as few function evaluations as possible. Thus some gradient (subgradient) information is needed. In what follows we shall evaluate the subgradient of $J_h(\mathbf{e}) = I_h(\mathbf{e}, \mathbf{u}(\mathbf{e}))$ with respect to the design variable $\mathbf{e} = \{\mathbf{t},\, \mathbf{q}\}$.

It holds the following formula for computation of the subgradient of $I_h(\mathbf{e}, \mathbf{u}(\mathbf{e}))$ (see Theorem 8.6 or [31]):

$$\nabla_{\mathbf{e}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})) \;+\; \xi_{\mathbf{u}}^T(\mathbf{e})\, \nabla_{\mathbf{u}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})) \;\in\; \partial I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})). \tag{131}$$

In practice it can be difficult to compute a representative $\xi_{\mathbf{u}}$ from the generalized Jacobian $\partial\mathbf{u}(\mathbf{e})$. Therefore we will go forward by applying the adjoint state technique to eliminate the term $\xi_{\mathbf{u}}$ from (131). Firstly we define the adjoint state

problem (if it has a solution):

$$\begin{pmatrix} \mathbf{K} & \mathbf{S}^T & \mathbf{0}^T \\ \mathbf{S} & -\mathbf{D} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \mathbf{r} \\ \mathbf{s} \end{pmatrix} = \begin{pmatrix} \nabla_{\mathbf{u}} I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})) \\ \mathbf{0} \end{pmatrix}, \qquad (\mathcal{A}(\mathbf{e}))$$
$$\mathbf{r}_i(\mathbf{e})\, \mathbf{s}_i(\mathbf{e}) \,=\, 0,\ \mathbf{r}_i(\mathbf{e}) \,\geqslant\, 0,\ \mathbf{s}_i(\mathbf{e}) \,\geqslant\, 0,\ i \in \mathcal{I}_0(\mathbf{e}),$$
$$\mathbf{s}_i(\mathbf{e}) \,=\, 0,\ i \in \mathcal{I}_+(\mathbf{e}),\ \mathbf{r}_i(\mathbf{e}) \,=\, 0,\ i \in \mathcal{I}_-(\mathbf{e}).$$

As $\mathbf{e} = \{\mathbf{t},\, \mathbf{q}\}$, the subgradient of I_h contains of two parts $\partial_{\mathbf{t}} I_h$ and $\partial_{\mathbf{q}} I_h$. We will analyze both parts separately. Firstly we shall focus on the subgradient with respect to the thickness.

Theorem 4.2. *Let* $\{\mathbf{u}(\mathbf{e}),\, \mathbf{v}(\mathbf{e}),\, \mathbf{w}(\mathbf{e})\}$ *be a solution to* (mLCP(**e**)) *and let* $\{\mathbf{p}(\mathbf{e}),\, \mathbf{r}(\mathbf{e}),\, \mathbf{s}(\mathbf{e})\}$ *solve* ($\mathcal{A}(\mathbf{e})$). *Then*

$$\nabla_{\mathbf{e}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})) \,+\, (-\nabla_{\mathbf{t}} \mathbf{K}\, \mathbf{u})^T\, \mathbf{p} \,\in\, \partial_{\mathbf{t}} I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})), \qquad (132)$$

where the multilinear mapping $\nabla_{\mathbf{t}} \mathbf{K}$ *is defined as follows:*

$$\nabla_{\mathbf{t}} \mathbf{K}\, \mathbf{u} \,=\, \left(\sum_{k=1}^{N} \frac{\partial}{\partial t_j} \mathbf{K}_{i,k}\, \mathbf{u}_k \right)_{i,j\,=\,1}^{N,\, n+1} \,\in\, \mathbb{R}^{N \times (n+1)}.$$

Proof. From (131) we obtain

$$\nabla_{\mathbf{e}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})) \,+\, (\xi^{\mathbf{t}}_{\mathbf{u}})^T\, \nabla_{\mathbf{u}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})) \,\in\, \partial_{\mathbf{t}} I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})).$$

By differentiating (mLCP(**e**)) we have

$$\mathbf{K}\, \xi^{\mathbf{t}}_{\mathbf{u}} \,=\, -\nabla_{\mathbf{t}} \mathbf{K}\, \mathbf{u} \,-\, \mathbf{S}^T \xi^{\mathbf{t}}_{\mathbf{v}}. \qquad (133)$$

Substituting from ($\mathcal{A}(\mathbf{e})$) into $(\xi^{\mathbf{t}}_{\mathbf{u}})^T\, \nabla_{\mathbf{u}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e}))$ we receive

$$(\xi^{\mathbf{t}}_{\mathbf{u}})^T \left(\mathbf{K}\, \mathbf{p} \,+\, \mathbf{S}^T \mathbf{r} \right) \,=\, \left(\mathbf{K}\, \xi^{\mathbf{t}}_{\mathbf{u}} \right)^T \mathbf{p} \,+\, (\xi^{\mathbf{t}}_{\mathbf{u}})^T \mathbf{S}^T \mathbf{r}.$$

Making use of (133) we obtain

$$\left(-\nabla_{\mathbf{t}} \mathbf{K}\, \mathbf{u} \,-\, \mathbf{S}^T \xi^{\mathbf{t}}_{\mathbf{v}} \right)^T \mathbf{p} \,+\, (\xi^{\mathbf{t}}_{\mathbf{u}})^T \mathbf{S}^T \mathbf{r}.$$

It remains to prove that

$$(\xi^{\mathbf{t}}_{\mathbf{u}})^T \mathbf{S}^T \mathbf{r} \,-\, (\xi^{\mathbf{t}}_{\mathbf{v}})^T \mathbf{S}\, \mathbf{p} \,=\, 0. \qquad (134)$$

From (mLCP(**e**)) and ($\mathcal{A}(\mathbf{e})$) it follows that

$$\mathbf{S}\, \xi^{\mathbf{t}}_{\mathbf{u}} \,=\, \mathbf{D}\, \xi^{\mathbf{t}}_{\mathbf{v}} \,-\, \mathbf{D}\, \xi^{\mathbf{t}}_{\mathbf{w}}, \quad \mathbf{S}\, \mathbf{p} \,=\, \mathbf{D}\, \mathbf{r} \,-\, \mathbf{D}\, \mathbf{s}. \qquad (135)$$

In view of (134) and (135) we have

$$(\xi^{\mathbf{t}}_{\mathbf{u}})^T \mathbf{S}^T \mathbf{r} \,-\, (\xi^{\mathbf{t}}_{\mathbf{v}})^T \mathbf{S}\,\mathbf{p} \,=\, (\xi^{\mathbf{t}}_{\mathbf{v}})^T \mathbf{D}\,\mathbf{s} \,-\, (\xi^{\mathbf{t}}_{\mathbf{w}})^T \mathbf{D}\,\mathbf{r}.$$

Let $i \in \mathcal{I}_+(\mathbf{e})$, then $\mathbf{s}_i = 0$ and $\dot{\mathbf{w}}_i(\mathbf{e};\, \mathbf{d}) = 0$. If $i \in \mathcal{I}_-(\mathbf{e})$, then $\mathbf{r}_i = 0$ and $\dot{\mathbf{v}}_i(\mathbf{e};\, \mathbf{d}) = 0$. The same situation arises if $i \in \mathcal{I}_0(\mathbf{e})$, it can be proved under some technical assumptions, see [40]. From there we may conclude that (134) holds and the assertion is proved.

■

Next we can approach to the subgradient with regard to the foundation stiffness.

Theorem 4.3. *Let* $\{\mathbf{u}(\mathbf{e}),\, \mathbf{v}(\mathbf{e}),\, \mathbf{w}(\mathbf{e})\}$ *be a solution to* (mLCP(e)) *and let* $\{\mathbf{p}(\mathbf{e}),\, \mathbf{r}(\mathbf{e}),\, \mathbf{s}(\mathbf{e})\}$ *solve* ($\mathcal{A}$(e)). *Then*

$$\begin{aligned} \nabla_{\mathbf{e}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})) \,+\, \left(-\nabla_{\mathbf{q}} \mathbf{S}^T\, \mathbf{v}\right)^T \mathbf{p} \,- \\ -\, (\nabla_{\mathbf{q}}\, \mathbf{D}\, \mathbf{v} \,-\, \nabla_{\mathbf{q}}\, \mathbf{S}\, \mathbf{u} \,-\, \nabla_{\mathbf{q}}\, \mathbf{D}\, \mathbf{w})^T\, \mathbf{r} \,\in\, \partial_{\mathbf{q}} I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})), \end{aligned} \tag{136}$$

where the multilinear mappings $\nabla_{\mathbf{q}}\mathbf{S}$, $\nabla_{\mathbf{q}}\mathbf{S}^T$ *and* $\nabla_{\mathbf{q}}\mathbf{D}$ *are defined in the following way:*

$$\nabla_{\mathbf{q}}\mathbf{S}\,\mathbf{u} = \left(\sum_{k=1}^{N} \frac{\partial}{\partial q_j} \mathbf{S}_{i,k}\, \mathbf{u}_k\right)_{i,j=1}^{M,n} \in \mathbb{R}^{M\times n},$$

$$\nabla_{\mathbf{q}}\mathbf{S}^T\,\mathbf{v} = \left(\sum_{k=1}^{M} \frac{\partial}{\partial q_j} \mathbf{S}_{k,i}\, \mathbf{v}_k\right)_{i,j=1}^{N,n} \in \mathbb{R}^{N\times n},$$

$$\nabla_{\mathbf{q}}\mathbf{D}\,\mathbf{v} = \left(\sum_{k=1}^{M} \frac{\partial}{\partial q_j} \mathbf{D}_{i,k}\, \mathbf{v}_k\right)_{i,j=1}^{M,n} \in \mathbb{R}^{M\times n}.$$

Proof. The generalized chain rule (131) narrowed only to the components corresponding to design variable $\mathbf{q}$ reads

$$\nabla_{\mathbf{e}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})) \,+\, (\xi^{\mathbf{q}}_{\mathbf{u}})^T\, \nabla_{\mathbf{u}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})) \,\in\, \partial_{\mathbf{q}} I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e})).$$

By differentiating (mLCP(e)) we obtain

$$\mathbf{K}\,\xi^{\mathbf{q}}_{\mathbf{u}} \,=\, -\nabla_{\mathbf{q}}\mathbf{S}^T\,\mathbf{v} \,-\, \mathbf{S}^T \xi^{\mathbf{q}}_{\mathbf{v}}. \tag{137}$$

From ($\mathcal{A}$(e)) we can substitute into $(\xi^{\mathbf{q}}_{\mathbf{u}})^T\, \nabla_{\mathbf{u}}\, I_h(\mathbf{e},\, \mathbf{u}(\mathbf{e}))$.

$$(\xi^{\mathbf{q}}_{\mathbf{u}})^T \left(\mathbf{K}\,\mathbf{p} \,+\, \mathbf{S}^T\mathbf{r}\right) \,=\, (\mathbf{K}\,\xi^{\mathbf{q}}_{\mathbf{u}})^T\,\mathbf{p} \,+\, (\xi^{\mathbf{q}}_{\mathbf{u}})^T \mathbf{S}^T\mathbf{r}. \tag{138}$$

Then by substituting (137) into (138) we have

$$\left(-\nabla_{\mathbf{q}}\mathbf{S}^T\,\mathbf{v}\right)^T - (\xi_{\mathbf{v}}^{\mathbf{q}})^T\mathbf{S}\,\mathbf{p} + (\xi_{\mathbf{u}}^{\mathbf{q}})^T\mathbf{S}^T\mathbf{r}.$$

In what follows we shall prove that

$$(\xi_{\mathbf{u}}^{\mathbf{q}})^T\mathbf{S}^T\mathbf{r} - (\xi_{\mathbf{v}}^{\mathbf{q}})^T\mathbf{S}\,\mathbf{p} = \left(\nabla_{\mathbf{q}}\,\mathbf{D}\,\mathbf{v} - \nabla_{\mathbf{q}}\,\mathbf{S}\,\mathbf{u} - \nabla_{\mathbf{q}}\,\mathbf{D}\,\mathbf{w}\right)^T\mathbf{r}. \tag{139}$$

Definitions of ($\mathcal{A}(\mathbf{e})$) and (mLCP($\mathbf{e}$)) imply

$$\begin{aligned}\mathbf{S}\,\xi_{\mathbf{u}}^{\mathbf{q}} &= \mathbf{D}\,\xi_{\mathbf{v}}^{\mathbf{q}} - \mathbf{D}\,\xi_{\mathbf{w}}^{\mathbf{q}} + \left(\nabla_{\mathbf{q}}\,\mathbf{D}\,\mathbf{v} - \nabla_{\mathbf{q}}\,\mathbf{S}\,\mathbf{u} - \nabla_{\mathbf{q}}\,\mathbf{D}\,\mathbf{w}\right),\\ \mathbf{S}\,\mathbf{p} &= \mathbf{D}\,\mathbf{r} - \mathbf{D}\,\mathbf{s}.\end{aligned} \tag{140}$$

Then by (140) we receive

$$\begin{aligned}(\xi_{\mathbf{u}}^{\mathbf{q}})^T\mathbf{S}^T\mathbf{r} - (\xi_{\mathbf{v}}^{\mathbf{q}})^T\mathbf{S}\,\mathbf{p} &= (\xi_{\mathbf{v}}^{\mathbf{q}})^T\mathbf{D}\,\mathbf{s} - (\xi_{\mathbf{w}}^{\mathbf{q}})^T\mathbf{D}\,\mathbf{r} +\\ &+ \left(\nabla_{\mathbf{q}}\,\mathbf{D}\,\mathbf{v} - \nabla_{\mathbf{q}}\,\mathbf{S}\,\mathbf{u} - \nabla_{\mathbf{q}}\,\mathbf{D}\,\mathbf{w}\right)^T\mathbf{r}.\end{aligned}$$

In the rest of the proof we can proceed similarly as in the proof of (134) and finally we may conclude that (139) holds.

■

The main advantage of this approach consists in the possibility of computing the subgradient of the cost functional using only the solution of (126) and ($\mathcal{A}(\mathbf{e})$). There is no longer necessary to compute the representative $\xi_{\mathbf{u}}$ from the generalized Jacobian, what would be costly.

5. Methods

The algebraic form of the discrete optimization problem leads to the following nonlinear programming problem:

$$\text{Find} \qquad \mathbf{e}^* \in \mathcal{U}^h : \;\; J_h(\mathbf{e}^*) \leqslant J_h(\mathbf{e}) \quad \forall \mathbf{e} \in \mathcal{U}^h, \tag{P_h}$$

where $J_h(\mathbf{e}) := I_h(\mathbf{e}, \mathbf{u}(\mathbf{e}))$ with $\mathbf{u}(\mathbf{e})$ being a solution to the following mixed linear complementarity problem:

$$\begin{pmatrix}\mathbf{K} & \mathbf{S}^T & \mathbf{0}^T\\ \mathbf{S} & -\mathbf{D} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\\ \mathbf{w}\end{pmatrix} = \begin{pmatrix}\mathbf{F}\\ \mathbf{0}\end{pmatrix} \tag{141}$$
$$\mathbf{v}^T\mathbf{w} = 0, \quad \mathbf{v}, \mathbf{w} \geqslant 0.$$

5.1. Numerical solution of (mLCP(e))

Firstly we will focus on numerical methods available for solving of the state problem (mLCP(e)) and possibly for the adjoint problem ($\mathcal{A}$(e)). There are several possibilities how to solve such a case of problems. One can use the so called Lemke method or the Gauss–Seidel method with projection, see e.g. (mLCP(e)), [36], [38]. But for boundary conditions which we are working with in the thesis the stiffness matrix $\mathbf{K}$ is only positive semidefinite and the mentioned algorithms may have some issues with such a kind of problems. Therefore we decided to use an approach based on interior point methods (IPM). See [36], for example. These methods are much more robust and more suitable for this case. As nowadays interior point methods are studied widely we will present here the method only briefly. The reader can find more information about interior point methods e.g. in [9], [10], [39], [44] or [54]. A primal-dual interior point method which works with a linearized step equation has been used. The key is a linearization of the scalar product $\mathbf{v}^T\mathbf{w}$. It can be written as $\mathbf{W}\,\mathbf{V}\,\bar{e}$ and the general formula for the path-following step is then defined as follows

$$\begin{pmatrix} \mathbf{K} & \mathbf{S}^T & \mathbf{0}^T \\ \mathbf{S} & -\mathbf{D} & \mathbf{D} \\ \mathbf{0} & \mathbf{W} & \mathbf{V} \end{pmatrix} \begin{pmatrix} \Delta\mathbf{u} \\ \Delta\mathbf{v} \\ \Delta\mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ -\mathbf{V}\mathbf{W}\bar{e} + \sigma_k\,\mu_k\,\bar{e} \end{pmatrix}, \tag{142}$$

where σ_k, μ_k are parameters known from the theory of interior point methods. Value $\sigma_k \in [0,1]$ is a so called centering parameter and $\mu_k = \mathbf{v}^T\mathbf{w}/M$ represents a duality measure. By $\bar{e}$ we denote the unit vector (do not interchange with the design variable $\mathbf{e}$), $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_M)$, $\mathbf{V} = \mathrm{diag}(v_1, \ldots, v_M)$ and $\mathbf{r}_1 = \mathbf{F} - \mathbf{K}\mathbf{u} - \mathbf{S}^T\mathbf{v}$, $\mathbf{r}_2 = \mathbf{D}\mathbf{v} - \mathbf{D}\mathbf{w} - \mathbf{S}\mathbf{u}$ are residual vectors. The new iteration is then obtained as

$$(\mathbf{u}^{k+1}, \mathbf{v}^{k+1}, \mathbf{w}^{k+1}) = (\mathbf{u}^k, \mathbf{v}^k, \mathbf{w}^k) + \alpha_k\,(\Delta\mathbf{u}, \Delta\mathbf{v}, \Delta\mathbf{w}),$$

where using suitable step length α_k and parameter values σ_k, μ_k are such that $v_i, w_i > 0$, $i = 1, \ldots, M$.

Most of the computational effort in primal-dual interior point methods is taken up in solving linear systems (142). Moreover the coefficient matrix is large and sparse, because the matrices $\mathbf{D}$, $\mathbf{V}$ and $\mathbf{W}$ are large and sparse themself. The special structure of the system matrix enable us to rewrite it in a much more compact form, that is easier and cheaper to solve than the original system. First let us eliminate $\Delta\mathbf{w}$ and add $-\mathbf{D}\,\mathbf{V}^{-1}$ times the third equation to the second equation of the system. It is possible because $\mathbf{v}$ and $\mathbf{w}$ are strictly positive, so that the matrices $\mathbf{V}$ and $\mathbf{W}$ are nonsingular. If we denote $\mathbf{p}_2 = \mathbf{r}_2 + \mathbf{D}\,\mathbf{W}\bar{e} - \mathbf{D}\,\mathbf{V}^{-1}\sigma_k\,\mu_k\,\bar{e}$ then we obtain

$$\begin{pmatrix} \mathbf{K} & \mathbf{S}^T \\ \mathbf{S} & -\mathbf{D}(\mathbf{I} - \mathbf{V}^{-1}\mathbf{W}) \end{pmatrix} \begin{pmatrix} \Delta\mathbf{u} \\ \Delta\mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{p}_2 \end{pmatrix}, \tag{143}$$
$$\Delta\mathbf{w} = -\mathbf{W}\,\bar{e} - \mathbf{V}^{-1}\mathbf{W}\,\Delta\mathbf{v} + \mathbf{V}^{-1}\sigma_k\,\mu_k\,\bar{e}.$$

Further we can eliminate $\Delta\mathbf{v}$ and add $\mathbf{S}^T[\mathbf{D} - \mathbf{D}\,\mathbf{V}^{-1}\mathbf{W}]^{-1}$ times the second equation to the first equation of (143).

$$[\mathbf{K} + \mathbf{S}^T[\mathbf{D} - \mathbf{D}\,\mathbf{V}^{-1}\mathbf{W}]^{-1}\mathbf{S}]\,\Delta\mathbf{u} = \mathbf{r}_1 + \mathbf{S}^T[\mathbf{D} - \mathbf{D}\,\mathbf{V}^{-1}\mathbf{W}]^{-1}\mathbf{p}_2 \tag{144}$$

$$\Delta\mathbf{v} = [\mathbf{D} - \mathbf{D}\,\mathbf{V}^{-1}\mathbf{W}]\,[\mathbf{S}\,\Delta\mathbf{u} - \mathbf{p}_2]. \tag{145}$$

$$\Delta\mathbf{w} = -\mathbf{W}\,\bar{e} - \mathbf{V}^{-1}\mathbf{W}\,\Delta\mathbf{v} + \mathbf{V}^{-1}\sigma_k\,\mu_k\,\bar{e}. \tag{146}$$

The dimension of the linear system (144) is N while the dimension of the original linear system was $N + 2\,M$. In addition the system is now symmetric and positive definite. The products and matrix inversions in (144)-(146) are cheap due to the diagonal structure of matrices $\mathbf{D}$, $\mathbf{V}$ and $\mathbf{W}$.

In practice it is an important issue to choose a starting point. A poor choice of starting point may lead to a failure of convergence. We present here an heuristic starting point selection procedure. This approach is based on [39]. Firstly let us choose $\mathbf{v} > 0$. Then find a solution of the following problem

$$\min_{\mathbf{u}} \frac{1}{2} r(\mathbf{u})^T\, r(\mathbf{u}), \quad r(\mathbf{u}) = \mathbf{F} - \mathbf{S}^T\mathbf{v} - \mathbf{K}\,\mathbf{u}. \tag{147}$$

Further by solving the following linear system we obtain vector $\bar{\mathbf{w}}$.

$$\mathbf{D}\,\bar{\mathbf{w}} = \mathbf{D}\,\mathbf{v} - \mathbf{S}\,\mathbf{u}. \tag{148}$$

In general $\bar{\mathbf{w}}$ obtained from (148) has nonpositive components. Such a vector is not suitable for us as the starting point. We define

$$\delta_{\mathbf{w}} = \max(-(3/2)\,\min_i w_i,\, 0) \tag{149}$$

and set

$$\hat{\mathbf{w}} = \bar{\mathbf{w}} + \delta_{\mathbf{w}}\,\bar{e}. \tag{150}$$

Now clearly $\hat{\mathbf{w}} > 0$ and to ensure that $\mathbf{v}$ and $\mathbf{w}$ are not too close to zero and not too dissimilar we define

$$\hat{\delta}_{\mathbf{w}} = \frac{1}{2}\frac{\hat{\mathbf{w}}^T\mathbf{v}}{\bar{e}^T\mathbf{v}}, \quad \hat{\delta}_{\mathbf{v}} = \frac{1}{2}\frac{\mathbf{v}^T\hat{\mathbf{w}}}{\bar{e}^T\hat{\mathbf{w}}} \tag{151}$$

and set

$$\mathbf{v}_0 = \mathbf{v} + \hat{\delta}_{\mathbf{v}}\,\bar{e}, \quad \mathbf{w}_0 = \hat{\mathbf{w}} + \hat{\delta}_{\mathbf{w}}\,\bar{e}. \tag{152}$$

Now we can finally introduce the outline of the IPM algorithm.

IPM Algorithm:

Calculate $(\mathbf{u}^0, \mathbf{v}^0, \mathbf{w}^0)$ as described above, set $\mu_0 = (\mathbf{v}^0)^T\mathbf{w}^0/M$, $\sigma_0 \in [0, 1]$;

for $k = 0, 1, 2, \ldots$

Set $(\mathbf{u}, \mathbf{v}, \mathbf{w}) = (\mathbf{u}^k, \mathbf{v}^k, \mathbf{w}^k)$.

Compute residuals $\mathbf{r}_1, \mathbf{r}_2$ and solve the linear system (144)-(146) for $(\Delta\mathbf{u}, \Delta\mathbf{v}, \Delta\mathbf{w})$;

Calculate $\alpha_k = \max\{\alpha \in (0, 1] : (\mathbf{v}, \mathbf{w}) + \alpha(\Delta\mathbf{v}, \Delta\mathbf{w}) \geqslant 0\}$;

Set $(\mathbf{u}^{k+1}, \mathbf{v}^{k+1}, \mathbf{w}^{k+1}) = (\mathbf{u}^k, \mathbf{v}^k, \mathbf{w}^k) + \alpha_k\,(\Delta\mathbf{u}, \Delta\mathbf{v}, \Delta\mathbf{w})$;

Compute $\mu_{k+1} = (\mathbf{v}^{k+1})^T\mathbf{w}^{k+1}/M$;

end(for) The algorithm stops when the duality measure is small enough.

5.2. Numerical solution of (P_h)

This subsection will be devoted to the solution algorithm of the nonlinear optimization problem (P_h). As we mentioned above, the objective function J_h can be nonsmooth and possibly nonconvex.

5.2.1. Bundle methods for nonsmooth optimization

Let us suppose a general nonsmooth optimization problem in the following form:

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g(x) = (g_1(x), \dots, g_m(x)) \leqslant 0, \end{aligned} \tag{153}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$ are Lipschitz continuous. In this section we briefly describe a general bundle method for solution of (153). It produces a sequence $\{x_k\}_{k=1}^{\infty}\mathbb{R}^n$ converging to the global minimum of the objective function, if it exists.

First let us consider an improvement function defined by:

$$H(x, y) := \max\{f(x) - f(y), g_j(x),\ j = 1, \dots, m\}.$$

Let x_k be the current approximation to the solution of (153) at the k-th iteration. We seek for the search direction d_k as a solution of the following unconstrained optimization problem:

$$\begin{aligned} &\text{minimize } H(x_k + d, x_k) \\ &\text{subject to } d \in \mathbb{R}^n. \end{aligned} \tag{154}$$

But (154) is still a nonsmooth problem. Therefore we need to approximate it in some way. Firstly let us suppose for a moment that the objective function f is convex and in addition to the current iteration we have some trial points $y_l \in \mathbb{R}^n$ (from the past iterations), subgradients $\xi_f^l \in \partial f(y_l)$ and subgradients $\xi_{g_j}^l \in \partial g_j(y_l)$ for $l \in I_k$, $j = 1, \dots, m$, where I_k is a nonempty subset of $\{1, \dots, k\}$. The cutting plane model of the improvement function linearizing both the objective and the constraint functions is defined by

$$\hat{H}_k(x) := \max\{\hat{f}_l(x) - f(x_k), \hat{g}_{j,l}(x),\ j = 1, \dots, m,\ l \in I_k\},$$

where

$$\begin{aligned} \hat{f}_l(x) &= f(y_l) + (\xi_f^l)^T(x - y_l), \qquad l \in I_k, \\ \hat{g}_{j,l}(x) &= g(y_l) + (\xi_{g_j}^l)^T(x - y_l), \qquad l \in I_k,\ j = 1, \dots, m. \end{aligned}$$

The approximation of (154) then reads as follows:

$$\begin{aligned} &\text{minimize } \hat{H}_k(x_k + d) + \tfrac{1}{2} d^T M_k d \\ &\text{subject to } d \in \mathbb{R}^n, \end{aligned} \tag{155}$$

where the regular and symmetric $n \times n$ matrix M_k is intended to accumulate information about the curvature of f in a ball around x_k.

Notice that (155) is still nonsmooth optimization problem, but due to its piecewise linear nature it is equivalent to the following (smooth) quadratic programming problem:

$$\begin{array}{ll} \text{minimize} & \nu + \frac{1}{2} d^T M_k d \\ \text{subject to} & -\alpha_{f,l}^k + (\xi_f^l)^T d \leqslant \nu, \qquad l \in I_k, \\ & -\alpha_{g_j,l}^k + (\xi_{g_j}^l)^T d \leqslant \nu, \qquad l \in I_k, j = 1, \ldots, m, \end{array} \tag{156}$$

where

$$\begin{array}{l} \alpha_{f,l}^k = f(x_k) - \hat{f}_l(x_k), \quad l \in I_k, \\ \alpha_{g_j,l}^k = -\hat{g}_{j,l}(x), \quad l \in I_k,\ j = 1, \ldots, m \end{array} \tag{157}$$

are so-called linearization errors. To avoid the difficulties caused by nonconvexity we replace linearization errors by so-called subgradient locality measures

$$\begin{array}{l} \beta_{f,l}^k = \max\{|\alpha_{f,l}^k|,\ \gamma_f \|x_k - y_l\|^2\}, \\ \beta_{g_j,l}^k = \max\{|\alpha_{g_j,l}^k|,\ \gamma_{g_j} \|x_k - y_l\|^2\}, \end{array} \tag{158}$$

where $\gamma_f,\ \gamma_{g_j} \geqslant 0,\ j = 1, \ldots, m$ are so-called distance measure parameters ($\gamma_f = 0$ if f is convex, $\gamma_{g_j} = 0$ if g_j is convex). In what follows, we shortly present several versions of bundle methods, which are slight modifications of the general bundle algorithm presented above. We focus on their main differences in the choice of the cutting plane approximation $\hat{f}_k$ and the stabilizing matrix M_k.

5.2.2. Diagonal variable metric bundle methods

A weighting parameter was added to the quadratic term of the objective function in (156) in order to accumulate some second-order information about the curvature of f around x_k. Thus the variable metric matrix M_k took the diagonal form

$$M_k = u_k I_k \tag{159}$$

with the weighting parameter $u_k > 0$. Based on the proximal point algorithm the proximal bundle method was derived. Also an adaptive safeguarded quadratic interpolation technique for updating u_k was introduced. For more detailed information see e.g. [32].

5.2.3. Variable metric bundle methods

The development of second-order methods has been in the center of interest for many researchers in nonsmooth optimization. Several attempts to employ

$$M_k \quad \text{as a full matrix} \tag{160}$$

with some updating scheme have been proposed by various authors. One of the most recent variable metric bundle methods using BFGS update was derived in [28]. The idea of the method is to use only three subgradients (two calculated at x_k and y_{k+1}, one aggregated, containing information from past iterations). This means that the dimension of the normally time-consuming quadratic programming subproblem (156) is only three and it can be solved with simple calculations. For details we refer e.g. to [28], [32].

5.2.4. Bundle-Newton method

The most recent advance in the development of the second-order bundle method was made in [29], where the bundle-Newton method was derived. Instead of the piecewise linear cutting plane model, it uses a quadratic model of the form

$$\hat{f}_k(x) := \max_{l \in I_k} \{f(y_l) + (\xi_f^l)^T (x - y^l) + \frac{1}{2}\rho_l (x - y^l)^T M_l (x - y^l)\}, \tag{161}$$

where $\rho_l \in [0, 1]$ is a damping parameter. When we compare the bundle-Newton method to the earlier variable metric bundle methods, we can state that the bundle-Newton method is the "real" second-order method, since every part of the model contains the second-order information in the form of the stabilizing matrix M_l. For the approximation

$$M_l \approx \nabla^2 f(y_l) \tag{162}$$

the authors proposed optionally analytic or finite-difference approximations. Under some additional assumptions it can be shown to maintain superlinear convergence rate. Although the operations with full matrix demand more storage and time. More detailed information can be found e.g. in [29], [32].

6. Computer implementation in C/C++ and Fortran

The Multiobjective Proximal Bundle method (`MPBNGC` 2.0 by M.M. Mäkelä) for nonsmooth, Nonconvex and Generally Constrained optimization, the Proximal Bundle algorithms for nonsmooth optimization(`PBUN` by L. Lukšan and J. Vlček), the Variable metric bundle method (`PVAR` by L. Lukšan and J. Vlček) and the bundle-Newton algorithm for nonsmooth optimization (`PNEW` by L. Lukšan and J. Vlček) briefly described in the previous section will be used as the optimization algorithm for (P_h). For detailed description of the algorithms see [33], [28], [29] and [30].

The interior point approach (IPM) presented in Section 5 is used for computing of a solution to the state problem in the mixed linear complementarity

form (mLCP(**e**)). Reduced linear systems (144) arising in the IPM method will be solved using an algorithm based on the well known Gaussian elimination.

In this section we will present all subroutines contained in the program developed for solution of the discrete optimization problem (P_h). Setting of all the input parameters will be described. To get a better view how the program proceeds and which subroutines are called, we will show a graphical scheme of the code (see Fig. 6).

The code is divided into several files. File `main.cpp` is the main file of the code. All the optimization parameters and state problem parameters are defined there and all the nonsmooth optimization algorithms are called from there. Subroutines `MPBNGC` 2.0, `PBUN`, `PVAR`, `PNEW` are originally written in FORTRAN 77 while the rest of the code (the state problem solver, the adjoint problem solver etc.) is written in C/C++. Therefore it was needed to connect the FORTRAN subroutines contained in files `mpbngc.f`, `pbun.f`, `pvar.f`, `pnew.f`, `msubs.f` and `psubs.f` with the rest of the C/C++ code using the interface defined in `methods_cpp.cpp`, `methods_cpp.h`. Conversely the objective function written in C/C++ needed to be converted in order to work with the FORTRAN subroutines, it was made in `objfunc_wrapper.f`. Files `beam_defl.cpp`, `beam_defl.h` contain function `fun_and_grad` which calls the solver of the state problem named `state_solver` and the adjoint problem solver `sensitivity_analysis` (with subgradient computation). The state problem solver is based on the interior point approach presented in Section 5, it uses the Gausian elimination (see e.g. [15]) for solution of the symmetric linear systems (144) and GMRES algorithm in the starting point computation in the IPM method.

The subgradient of the objective function is computed using the adjoint problem approach described in Section 4.2. Finally the files `functions.cpp`, `functions.h`, `print.cpp` and `print.h` contain of definitions of some auxiliary functions and print subroutines.

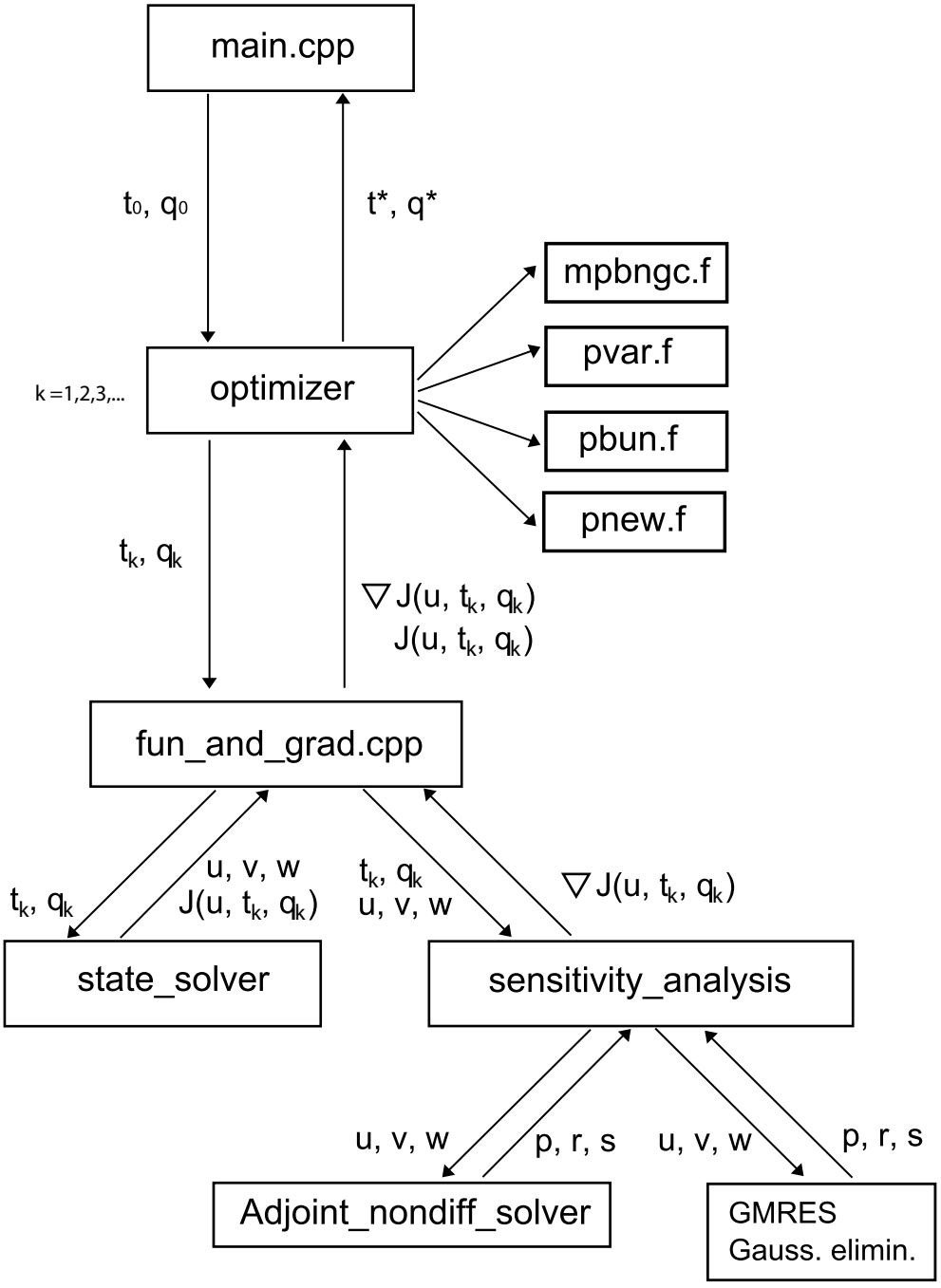

Figure 6: Scheme of the optimization code

The C/C++ equivalent functions of `MPBNGC`, `PBUN`, `PVAR` and `PNEW` defined in the interface files `methods_cpp.cpp`, `methods_cpp.h` are called by the following statements:

```
MPBNGC_setup setup(n,1,2*N+2,uad,h,imethod,rmethod,istate,state);
MPBNGC_results results(setup);
MPBNGC_solve(&beam_deflection_wrapper,x0,n,setup,results);

PVARL_setup setup(n,0,1,2*N+2,uad,h,imethod,rmethod,istate,state);
PVARL_results results(setup);
PVARL_solve(&beam_deflection_wrapper,x0,n,setup,results);

PBUNL_setup setup(n,0,1,2*N+2,uad,h,imethod,rmethod,istate,state);
PBUNL_results results(setup);
PBUNL_solve(&beam_deflection_wrapper,x0,n,setup,results);

PNEWL_setup setup(n,0,1,2*N+2,uad,h,imethod,rmethod,istate,state);
PNEWL_results results(setup);
PNEWL_solve(&beam_deflection_wrapper,x0,n,setup,results);
```

In what follows we will summarize all the input arguments of the code needed to be initialized in the `main.cpp` file. The following abbreviations are used: II - integer, input, RI - real, input, IU - integer, input, output, RU - real, input, output.

Argument	Type	Significance
`N`	II	Number of finite elements.
`n`	II	Total number of the design variables ($n = 2N + 1$).
`h`	RI	Element length ($h = L/N$).
`x0(n)`	RU	On input, vector with the initial estimate to the optimal solution. On output, the approximation of the optimal thickness and subsoil stiffness. `x0[0]-x0[N]` contains vector $\mathbf{t}$ and `x0[N+1]-x0[2N]` contains $\mathbf{q}$.
`uad(7)`	RI	Vector containing parameters defining the set $\mathcal{U}^h$. `uad[0]` - t_0, $0 < t_0$. `uad[1]` - t_1, $t_0 \leqslant t_1$. `uad[2]` - γ_1, $0 < \gamma_1$. `uad[3]` - γ_2, $0 < \gamma_2$. `uad[4]` - q_0, $0 < q_0$. `uad[5]` - q_1, $q_0 \leqslant q_1$. `uad[6]` - γ_3, $0 < \gamma_3$.
`state(6)`	RI	Vector containing real input parameters of the state problem ($\mathcal{P}_h(e_h)$). `state[0]` - E - Young's modulus of elasticity ($0 < E$). `state[1]` - b - Width of the beam ($0 < b$). `state[2]` - L - Length of the beam ($0 < L$). `state[3]` - ϵ - Final accuracy for the GMRES method ($0 < \epsilon$). `state[4]` - $\epsilon_{\text{contact}}$ - Tolerance parameter for activity of the subsoil ($0 < \epsilon_{\text{contact}}$).
`istate(2)`	II	Vector containing integer input parameters of the state problem ($\mathcal{P}_h(e_h)$). `istate[0]` - natural boundary condition. 0: $u(0) = u'(0) = 0$. 1: $u'(0) = 0$. 2: $u(0) = 0$. `istate[1]` - cost functional. 0: $I_1(e, u(e)) = J_1(e) := \int_\Omega fu \,\mathrm{d}x$. 1: $I_2(e, u(e)) = J_2(e) := \int_\Omega u^2 \,\mathrm{d}x$. 2: $I_3(e, u(e)) = J_3(e) := \int_\Omega t^2(u'')^2 \,\mathrm{d}x$.

`method` II Parameter defining the nonsmooth optimization method used for solution of (P_h).
1: `MPBNGC` - Proximal Bundle method for Nonsmooth, nonconvex and Generally Constrained optimization (by Mäkelä, M.M.).
2: `PVAR` - Variable metric bundle method (by Lukšan, L., Vlček, J.).
3: `PBUN` - Proximal bundle algorithm for nonsmooth optimization (by Lukšan, L., Vlček, J.).
4: `PNEW` - Bundle-Newton algorithm for nonsmooth optimization (by Lukšan, L., Vlček, J.).

The nonsmooth optimization methods have the following input parameters

`imethod` II Vector containing the integer input parameters of the optimization method.

`rmethod` RI Vector containing the real input parameters of the optimization method.

These parameters are listed in the following table:

Parameter	MPBNGC	PVAR	PBUN	PNEW
imethod[0]	iprint	MIT	MIT	MIT
imethod[1]	lmax	MFV	MFV	MFV
imethod[2]	jmax	MEX	MET	-
imethod[3]	niter	MTESX	MTESX	MTESX
imethod[4]	nfasg	MTESF	MTESF	MTESF
imethod[5]	-	IPRNT	IPRNT	IPRNT
imethod[6]	-	-	-	IHES
rmethod[0]	gam	XMAX	XMAX	XMAX
rmethod[1]	rl	TOLX	TOLX	TOLX
rmethod[2]	eps	TOLF	TOLF	TOLF
rmethod[3]	feas	TOLB	TOLB	TOLB
rmethod[4]	-	TOLG	TOLG	TOLG
rmethod[5]	-	ETA	ETA	ETA

The arguments have the following meaning:

Argument	Type	Significance
`iprint`	II	Printout control parameter. -1: No printout. 0: Only the error messages. 1: The final value of the objective function. 2: The whole final solution. 3: At each iteration value of the objective function. 4: At each iteration the whole solution.
`lmax`	II	The maximum number of the objective function calls in line search.
`jmax`	II	The maximum number of stored subgradients.
`niter`	IU	Input : The maximum number of iterations. Output: Number of used iterations.
`nfasg`	IU	Input : The maximum number of the objective function calls. Output: Number of the objective function calls.
`gam`	RI	Distance measure parameter.
`rl`	RI	Line search parameter.
`eps`	RI	Tolerance for constraint feasibility.
`feas`	RI	Final objective function accuracy parameter.
`MIT`	II	Variable that specifies the maximum number of iterations; the choice `MIT=0` causes that the default value 200 will be taken.
`MFV`	II	Variable that specifies the maximum number of function evaluations; the choice `|MFV|=0` causes that the default value 500 will be taken.
`MEX`	II	Version of nonsmooth variable metric method: 0: Convex version. 1: Nonconvex version.
`MET`	II	Variable that specifies the weight updating method: 0: quadratic interpolation. 1: local minimization. 2: quasi-Newton condition.
`MTESX`	II	Variable that specifies the maximum number of iterations with changes of the coordinate vector `X` smaller than `TOLX`; the choice `MTESX=0` causes that the default value `MTESX=20` will be taken.
`MTESF`	II	variable that specifies the maximum number of iterations with changes of function values smaller than `TOLF`; the choice `MTESF=0` causes that the default value `MTESX=2` will be taken.

`IPRNT`	II	Variable that specifies print. 0: Print is suppressed. 1: Basic print of final results. -1: Extended print of final results. 2: Basic print of intermediate and final results. -2: Extended print of intermediate and final results.
`IHES`	II	Variable that specifies a way for computing second derivatives: 0: Numerical computation. 1: analytical computation by the user supplied subroutine `HES`.
`XMAX`	RI	Maximum stepsize; the choice `XMAX=0` causes that the default value 10^{-3} will be taken.
`TOLX`	RI	Tolerance for the change of the coordinate vector `X`; the choice `TOLX=0` causes that the default value 10^{-16} will be taken.
`TOLF`	RI	Tolerance for the change of function values; the choice `TOLF=0` causes that the default value 10^{-8} will be taken.
`TOLB`	RI	Minimum acceptable function value; the choice `TOLB=0` causes that the default value -10^{60} will be taken.
`TOLG`	RI	Tolerance for the termination criterion; the choice `TOLG=0` causes that the default value 10^{-6} will be taken.
`ETA`	RI	Distance measure parameter.

Let us now present a sample initialization of the input arguments of the code in the file `main.cpp`:

```
int main(int argc, char **argv){

 const int N = 10;
 const int n = 2*N+1;
 double uad[7];
 double rmethod[6];
 int    imethod[7];
 double state[6];
 int    istate[5];
 //---------------------------------------------------------//
 //        Selection of the optimization method              //
 //---------------------------------------------------------//
 int    method = 3;
 //---------------------------------------------------------//
 //     Set of admissible design variables Uad              //
 //---------------------------------------------------------//
 uad[0] = 0.2;   //  T0
```

```
uad[1] = 0.8;   //  T1
uad[2] = 5;     //  T2
uad[3] = 0.3;   //  T3
uad[4] = 0.5;   //  Q0
uad[5] = 1.5;   //  Q1
uad[6] = 10;    //  Q2
//-----------------------------------------------------//
//     State problem constants                         //
//-----------------------------------------------------//
state[0] = 2.19e+6;  // E
state[1] = 0.4;      // b_w
state[2] = 10;       // L
state[3] = 10e-10;   // epsilon
state[4] = 10e-6;    // epsilon_contact
state[5] = 10e+6;    // f_m
//-----------------------------------------------------//
//     Boundary conditions and Cost functional         //
//-----------------------------------------------------//
istate[0] = 1;      // bc
istate[1] = 0;      // cf
//-----------------------------------------------------//
//     Definition of a starting point                  //
//-----------------------------------------------------//
double  h  = state[2]/N;
double x0[n];
for(int i = 0; i < n; i++){
   if(i < N+1)
     x0[i] = 0.5;  // t[i]
   else
 x0[i] = 1;      // q[i]
}
//*****************************************************//
// method==1  -->   MPBNGC method will be used         //
//*****************************************************//
if(method==1){
//-----------------------------------------------------//
//     Parameters of the MPBNGC method                 //
//-----------------------------------------------------//
imethod[0] = 3;       //iprint
imethod[1] = 100;     //lmax
imethod[2] = n;       //jmax
imethod[3] = 1000;    //niter
imethod[4] = 1000;    //nfasg
```

```
rmethod[0] = 0.3;      // gam
rmethod[1] = 0.0001;    // rl
rmethod[2] = 1e-6;     // eps
rmethod[3] = 1e-9;     // feas
//---------------------------------------------------------//
//     MPBNGC setup                                        //
//---------------------------------------------------------//
MPBNGC_setup setup(n,1,2*N+2,0,uad,h,imethod,rmethod,
                   istate,state);
MPBNGC_results results(setup);


//---------------------------------------------------------//
//     Executing of the MPBNGC method                      //
//---------------------------------------------------------//
MPBNGC_solve(&beam_deflection_wrapper,x0,n,setup,results);
}
//*********************************************************//
//     method==2 -->   PVAR method will be used            //
//*********************************************************//
if(method==2){
//---------------------------------------------------------//
//      Parameters of the PVARL method                     //
//---------------------------------------------------------//
imethod[0] =  4000;         // MIT
imethod[1] =  4000;         // MFV
imethod[2] =  1;            // MET
imethod[3] =  4000;         // MTESX
imethod[4] =  4000;         // MTESF
imethod[5] =  -2;           // IPRNT
rmethod[0] =  1;            // XMAX
rmethod[1] =  0;            // TOLX
rmethod[2] =  0;            // TOLF
rmethod[3] =  0;            // TOLB
rmethod[4] =  0;            // TOLG
rmethod[5] =  0.3;          // ETA
//---------------------------------------------------------//
//      PVARL setup                                        //
//---------------------------------------------------------//
PVARL_setup setup(n,0,1,2*N+2,uad,h,imethod,rmethod,
                  istate,state);
PVARL_results results(setup);
//---------------------------------------------------------//
```

```
//      Executing of the PVARL method                    //
//-------------------------------------------------------//
PVARL_solve(&beam_deflection_wrapper,x0,n,setup,results);
}
//*******************************************************//
//     method==3  -->  PBUN method will be used          //
//*******************************************************//
if(method==3){
//-------------------------------------------------------//
//     Parameters of the PBUNL method                    //
//-------------------------------------------------------//
imethod[0] =  4000;         // MIT
imethod[1] =  4000;         // MFV
imethod[2] =  2;            // MET
imethod[3] =  4000;         // MTESX
imethod[4] =  4000;         // MTESF
imethod[5] =  -2;           // IPRNT
rmethod[0] =  1;            // XMAX
rmethod[1] =  0;            // TOLX
rmethod[2] =  0;            // TOLF
rmethod[3] =  0;            // TOLB
rmethod[4] =  0;            // TOLG
rmethod[5] =  0.5;          // ETA
//-------------------------------------------------------//
//      PBUNL setup                                      //
//-------------------------------------------------------//
PBUNL_setup setup(n,0,1,2*N+2,uad,h,imethod,rmethod,
                  istate,state);
PBUNL_results results(setup);
//-------------------------------------------------------//
//      Executing of the PBUNL method                    //
//-------------------------------------------------------//
PBUNL_solve(&beam_deflection_wrapper,x0,n,setup,results);
}
//*******************************************************//
//       method==4   -->  PNEW method will be used       //
//*******************************************************//
 if(method==4){
//-------------------------------------------------------//
//       Parameters of the PNEW method                   //
//-------------------------------------------------------//
imethod[0] =  4000;         // MIT
imethod[1] =  4000;         // MFV
```

```
  imethod[2] =  2;            // MET
  imethod[3] =  4000;         // MTESX
  imethod[4] =  4000;         // MTESF
  imethod[5] =  -2;           // IPRNT
  imethod[6] =  0;            // IHES
  rmethod[0] =  1;            // XMAX
  rmethod[1] =  0;            // TOLX
  rmethod[2] =  0;            // TOLF
  rmethod[3] =  0;            // TOLB
  rmethod[4] =  0;            // TOLG
  rmethod[5] =  0.5;          // ETA
  //-----------------------------------------------------------//
  //        PNEW setup                                         //
  //-----------------------------------------------------------//
  PNEWL_setup setup(n,0,1,2*N+2,uad,h,imethod,rmethod,
                    istate,state);
  PNEWL_results results(setup);
  //-----------------------------------------------------------//
  //        Executing of the PNEW method                       //
  //-----------------------------------------------------------//
  PNEWL_solve(&beam_deflection_wrapper,x0,n,setup,results);
  }
  return EXIT_SUCCESS;
}
```

The output from the code may have the following form:

```
 ENTRY TO PBUN :
 NIT=    0  NFV=    1  NFG=    1  F=  98.8348549      G= 0.100E+61
 NIT=    1  NFV=    2  NFG=    2  F=  55.8645670      G= 0.500E+00
 NIT=    2  NFV=    3  NFG=    3  F=  45.8273444      G= 0.543E-01
 NIT=    3  NFV=    4  NFG=    4  F=  44.6196036      G= 0.661E-02
 NIT=    4  NFV=    5  NFG=    5  F=  43.6140929      G= 0.535E-02
 NIT=    5  NFV=    7  NFG=    7  F=  43.6065349      G= 0.668E-02
 NIT=    6  NFV=    9  NFG=    9  F=  43.6032432      G= 0.988E-02
 NIT=    7  NFV=   11  NFG=   11  F=  43.5908066      G= 0.352E-01
 NIT=    8  NFV=   12  NFG=   12  F=  43.5739112      G= 0.239E-01
 NIT=    9  NFV=   14  NFG=   14  F=  43.5724029      G= 0.165E-01
 NIT=   10  NFV=   15  NFG=   15  F=  43.5706204      G= 0.138E-01
 NIT=   11  NFV=   17  NFG=   17  F=  43.5697041      G= 0.100E-01
 NIT=   12  NFV=   18  NFG=   18  F=  43.5690191      G= 0.832E-02
 NIT=   13  NFV=   19  NFG=   19  F=  43.5690191      G= 0.675E-02
 EXIT FROM PBUN :
```

```
 NIT= 13  NFV= 19  NFG= 19  F= 43.5690191  G= 0.675E-02  ITERM= 4
 t= 0.8000000   0.8000000   0.7913685   0.6896902   0.5312745
    0.2876668   0.2000000   0.2000000   0.2000000
 q= 0.6724032   0.7069785   0.7080681   0.6958103   0.7167398
    1.500000    1.500000    1.500000
-----------------------------------------------------------------
Total time:                                | 0.231 sec | 100%
-----------------------------------------------------------------
State problem:                             | 0.103 sec | 44.5887%
-----------------------------------------------------------------
  - starting point:                        | 0.008 sec | 3.4632%
  - assembling the linear system in IPM:   | 0.07 sec  | 30.303%
  - solution of the linear system in IPM:  | 0.024 sec | 10.3896%
-----------------------------------------------------------------
Adjoint problem, Sensitivity analysis:     | 0.045 sec | 19.4805%
-----------------------------------------------------------------
  - adjoint problem:                       | 0.015 sec | 6.49351%
  - sensitivity analysis:                  | 0.03 sec  | 12.987%
-----------------------------------------------------------------
The rest of the code:                      | 0.084 sec | 36.3636%
-----------------------------------------------------------------
```

This output is different for different values of the input argument `IPRNT` resp. `iprint`. In this case we have set `IPRNT`= 2. For sample output for the other `IPRNT` (resp. `iprint`) options see the documentation of `MPBNGC`, `PVAR`, `PBUN` and `PNEW` on the attached CD.

7. Numerical experiments

In this section we shall present results of several numerical examples. Firstly we will try to compare the efficiency of methods `MPBNGC`, `PVAR`, `PBUN` and `PNEW` on a particular example. Setting of the parameters of $\mathcal{U}^h$, the cost functional J_h, boundary conditions, number of finite elements and the vertical load f certainly have an influence on the optimal solution (P_h). It will be demonstrated on the following examples. In all examples we will consider a beam of length

$$\mathrm{L} = 10$$

with an equidistant partition. We will use the 4-point Gauss–Lobatto formula with integration points ± 1, $\pm\frac{1}{\sqrt{5}}$ and wages $\frac{1}{6}$, $\frac{5}{6}$, see e.g. [1] or [20]. The parameters related to the material properties and the cross sectional area of the beam will be defined as follows: $b = 0.4$, $E = 2.19 \cdot 10^6$.

7.1. Nonsmooth optimization methods

In the *first example* the load function f is piecewise constant and given by

$$f(x) = \begin{cases} -50 & x < 5, \\ 100 & x \geqslant 5. \end{cases} \tag{163}$$

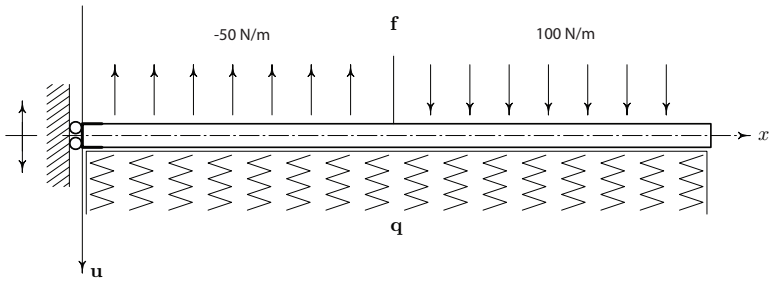

Figure 7: Outline of the beam with the load.

The cost functional is the compliance of the beam:

$$J(e) = \int_\Omega f\, u(e)\, \mathrm{d}x \approx \sum_{i=1}^{n+1} \frac{h}{2} \left(u_{i-1}\, f(x_{i-1}) + u_i\, f(x_i)\right) = \mathbf{u}^T \mathbf{B}\, \mathbf{f}.$$

Let $\mathcal{U}^h$ be defined by the following parameters: $t_0 = 0.2$, $t_1 = 0.8$, $\gamma_1 = 5$, $\gamma_2 = 0.3$, $q_0 = 500$, $q_1 = 1500$, $\gamma_3 = 10000$ and let the initial guess be $\mathbf{e}_0 = \{\mathbf{t}_0, \mathbf{q}_0\}$, where $t_i^0 = 0.5$ for $i = 0, \ldots, N$ and $q_i^0 = 1000$ for $i = 0, \ldots, N-1$. We used 32 finite elements in discretization; i.e., $N = 32$ and $h = 5/16$. The following boundary condition is prescribed: $u'(0) = 0$. It is clear that F^h with f defined by (163) fulfills the condition ($\mathrm{S2}_h$).

The nonsmooth and possibly nonconvex nonlinear mathematical programming problem (P_h) was solved using optimization codes `MPBNGC`, `PVAR`, `PBUN` and `PNEW`. Optimal results are dependent on the setting of input arguments of these algorithms. Inputs are slightly different for `MPBNGC` and the three remaining methods.

Algorithms have been run with the following arguments:
`MPBNGC`: `iprint` $= 3$, `lmax` $= 100$, `jmax` $= 2n+1$, `niter` $= 1000$, `nfasg` $= 1000$, `gam` $= 0.3$, `rl` $= 0.1$, `eps` $= 10^{-4}$, `feas` $= 10^{-9}$.
`PVAR`: `MIT` $= 1000$, `MFV` $= 1000$, `MEX` $= 1$, `MTESX` $= 1000$, `MTESF` $= 1000$, `IPRNT` $= -2$, `XMAX` $= 0.7$, `TOLX` $= 0$, `TOLF` $= 0$, `TOLB` $= 0$, `TOLG` $= 10^{-4}$, `ETA` $= 0.3$.
`PBUN`: `MIT` $= 1000$, `MFV` $= 1000$, `MET` $= 1$, `MTESX` $= 1000$, `MTESF` $= 1000$, `IPRNT` $= -2$, `XMAX` $= 0.7$, `TOLX` $= 0$, `TOLF` $= 0$, `TOLB` $= 0$, `TOLG` $= 10^{-4}$, `ETA` $= 0.3$.

`PNEW`: `MIT` = 1000, `MFV` = 1000, `IHES` = 0, `MTESX` = 1000, `MTESF` = 1000, `IPRNT` = −2, `XMAX` = 1, `TOLX` = 0, `TOLF` = 0, `TOLB` = 0, `TOLG` = 10^{-4}, `ETA` = 0.3.

The optimal cost functional values and number of iterations are summarized in Table 1. The following abbreviations are used: *Algorithm* = nonsmooth optimization algorithm, *Final* = optimal value of the cost functional, *Iter* = number of iterations, *Feval* = number of `fun_and_grad` calls, *Ctime* = solution time (in seconds).

Table 1: Cost functional values and number of iterations

Algorithm	Final	Iter	Feval	Ctime
`MPBNGC`	40.2276673	52	165	27.042
`PVAR`	40.2295349	80	81	13.356
`PBUN`	40.2270780	25	35	6.151
`PNEW`	42.3717060	104	6930	1187.235

The best cost functional value was reached by the `PBUN` method. Also the number of function evaluations is very small in comparison to the other algorithms. Therefore if the dimension of the problem is higher, it will be probably efficient to use the `PBUN` method. The difference between the final cost functional values for `PBUN`, `PVAR` and `MPBNGC` is minimal.

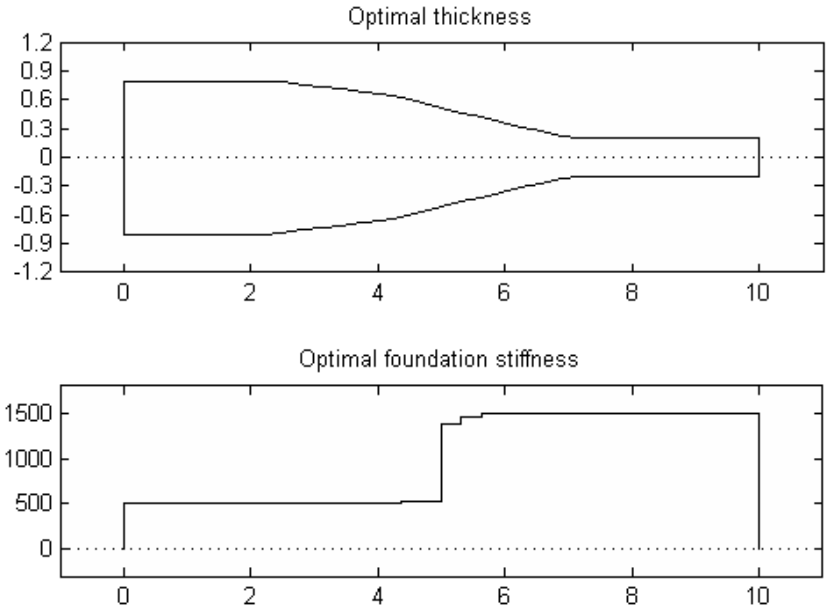

Figure 8: The optimal results obtained by the PBUN method.

In Figure 8 the optimal thickness of the beam and the optimal stiffness of its foundation reached by `PBUN` are shown. The optimal deflection of the beam is

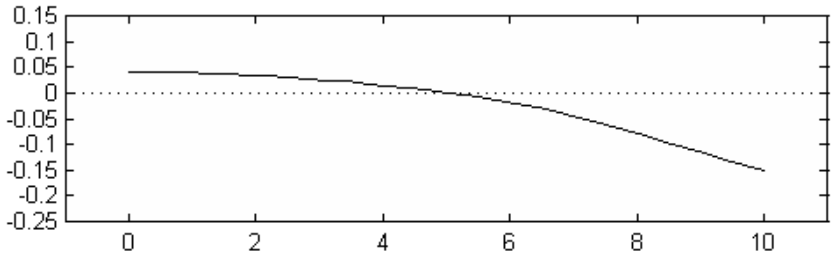

Figure 9: A deflection of the beam

shown on Figure 9. The graphical display of the results obtained by `PVAR`, `PNEW` and `MPBNGC` are almost identical to Figures 8 and 9.

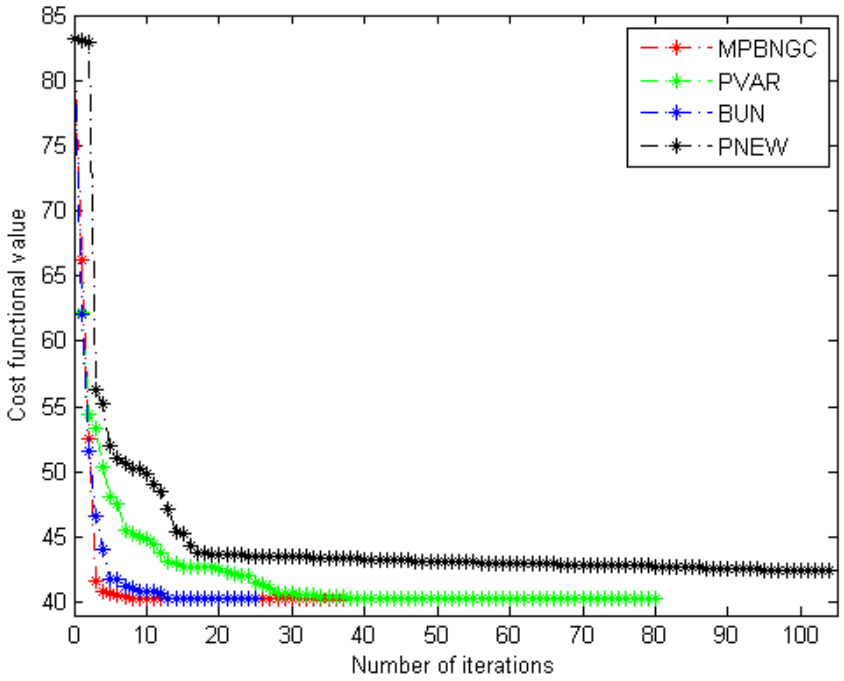

Figure 10: Cost functional values

In Figure 10 the cost functional values during the iterations of methods `MPBNGC`, `PVAR`, `PBUN` and `PNEW` are plotted. Methods `MPBNGC` and `PBUN` proceed vey quickly to the optimum and after first 5 iterations they almost reached the optimal value. The codes `PVAR` and `PNEW` decreased slower. `MPBNGC`, `PVAR`, `PBUN` reached almost the same cost functional value, the fastest from these three algorithms is `PBUN` which converged in 25 iterations.

The evaluation of the cost functional and its gradient (subgradient) involves

solving of the nonlinear state problem (mLCP(**e**)) and the adjoint problem ($\mathcal{A}$(**e**)). Consequently, the optimization algorithm should use as few `fun_and_grad` calls as possible. In Table 2 the dependence of the number of function evaluations on the discretization parameter is shown. All the other parameters are the same as before, only the dimension changes ($N = 8, 16, 32, 64, 128$).

Table 2: Number of cost functional evaluations

Algorithm	8	16	32	64	128
`MPBNGC`	40	187	165	177	220
`PVAR`	24	43	81	106	257
`PBUN`	33	42	35	36	40
`PNEW`	1260	20686	6930	–	–

On the one hand it is obvious that the method `PBUN` is the most efficient method for this problem and for the actual setting of the input parameters. On the other hand the method `PNEW` with the options `IHES` =0 is certainly not suitable for this type of problems. The high number of `fun_and_grad` calls is caused by the fact that the method uses the value of the gradient (subgradient) for numerical computation of second order derivatives.

We have to notice that the number of iterations depends on the setting of input arguments of the algorithms, especially setting of the distance measure parameter `gam` resp. `ETA` and the line search parameter `rl` resp. `XMAX`. Therefore a setting which seems to be optimal for one problem configuration does not need to be optimal for other problem configurations and we must be careful when setting these arguments.

7.2. The influence of the discretization parameter h

In the *second example* we shall analyze the dependence of the optimal solution to (P_h) on the discretization parameter h (resp. N). The load function f is piecewise polynomial and given by

$$f(x) = \begin{cases} x^3 - 20 & x < \frac{50}{8}, \\ -(x-8)^4 - 30 & x \geqslant \frac{50}{8}. \end{cases} \tag{164}$$

We will minimize the compliance of the beam:

$$J(e) = \int_\Omega f\, u(e)\, \mathrm{d}x \approx \sum_{i=1}^{n+1} \frac{h}{2} \left(u_{i-1}\, f(x_{i-1}) + u_i\, f(x_i)\right) = \mathbf{u}^T \mathbf{B}\, \mathbf{f}.$$

Let the set $\mathcal{U}^h$ be defined by the following parameters: $t_0 = 0.2$, $t_1 = 0.8$, $\gamma_1 = 5$, $\gamma_2 = 0.4$, $q_0 = 500$, $q_1 = 1500$, $\gamma_3 = 10000$ and let the initial guess be

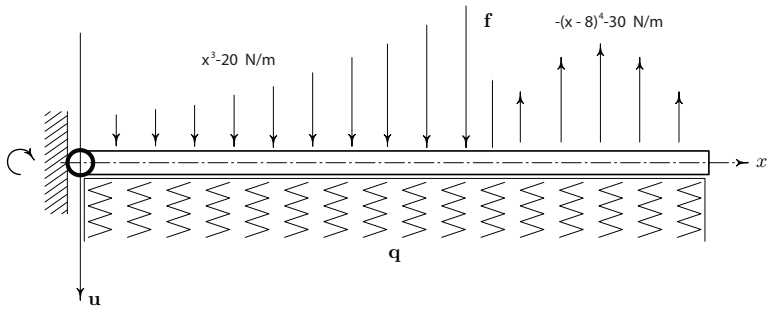

Figure 11: Outline of the beam with its load.

Table 3: Results

Dim	Initial	Final	Iter	Eval
8	5.12908953	2.50448124	33	42
16	4.99601988	2.30372504	55	69
32	5.00690169	2.35593981	43	58
64	5.10577388	2.43685832	33	47

$\mathbf{e}_0 = \{\mathbf{t}_0, \mathbf{q}_0\}$, where $t_i^0 = 0.5$, $i = 0, \ldots, N$ and $q_i^0 = 1000$, $i = 0, \ldots, N-1$. The following boundary condition is prescribed: $u(0) = 0$.

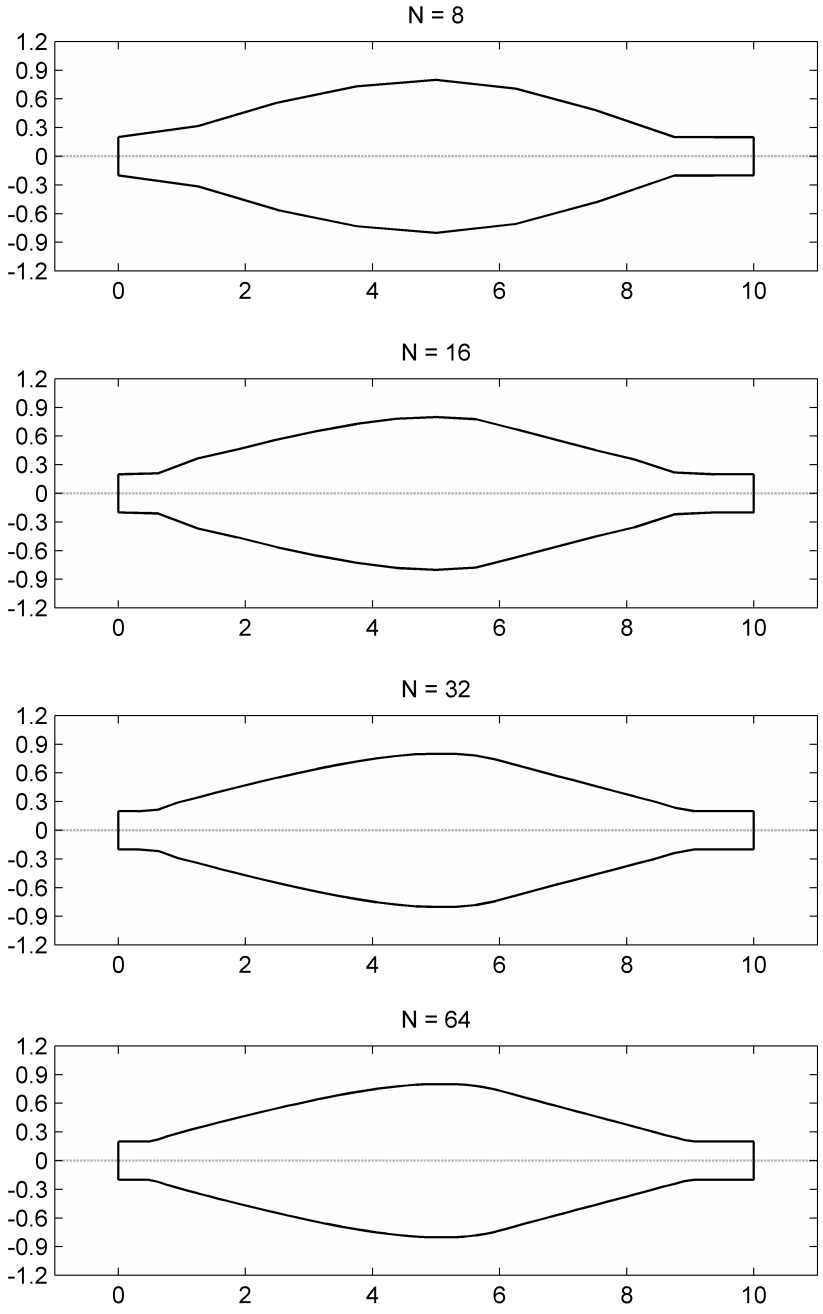

Figure 12: Optimal thickness

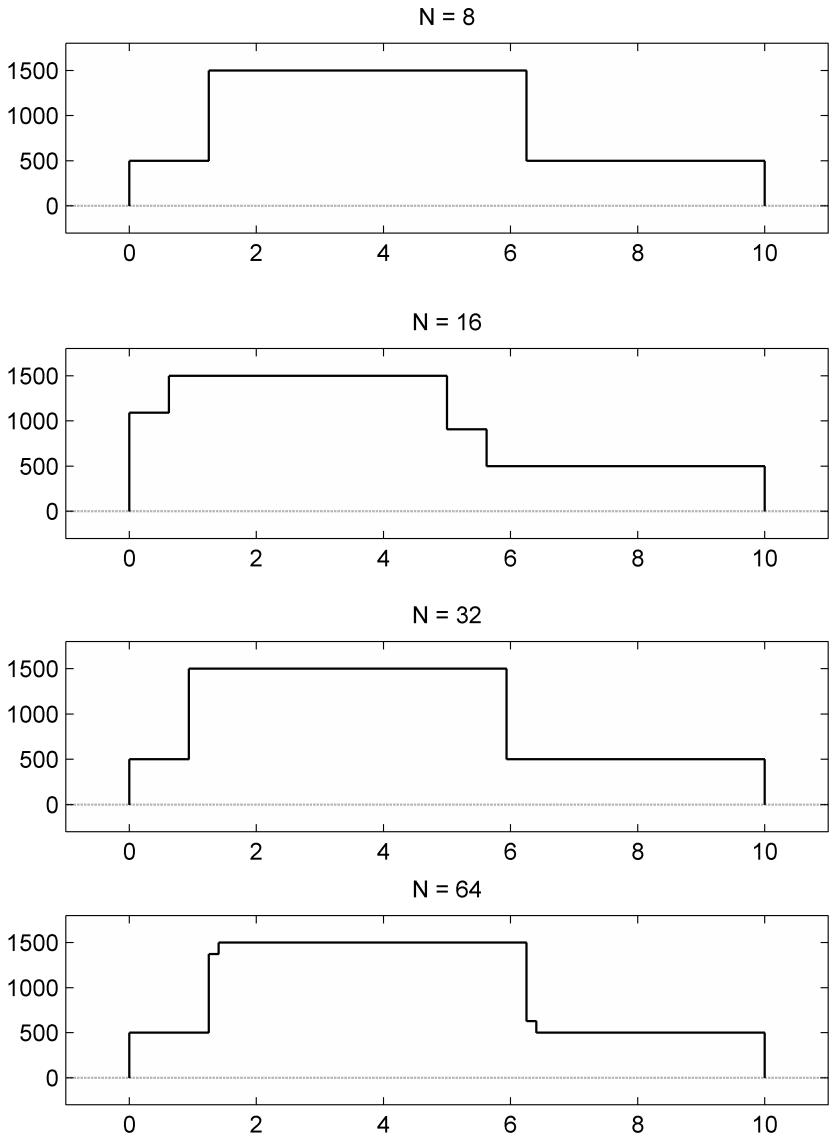

Figure 13: Optimal stiffness of the foundation

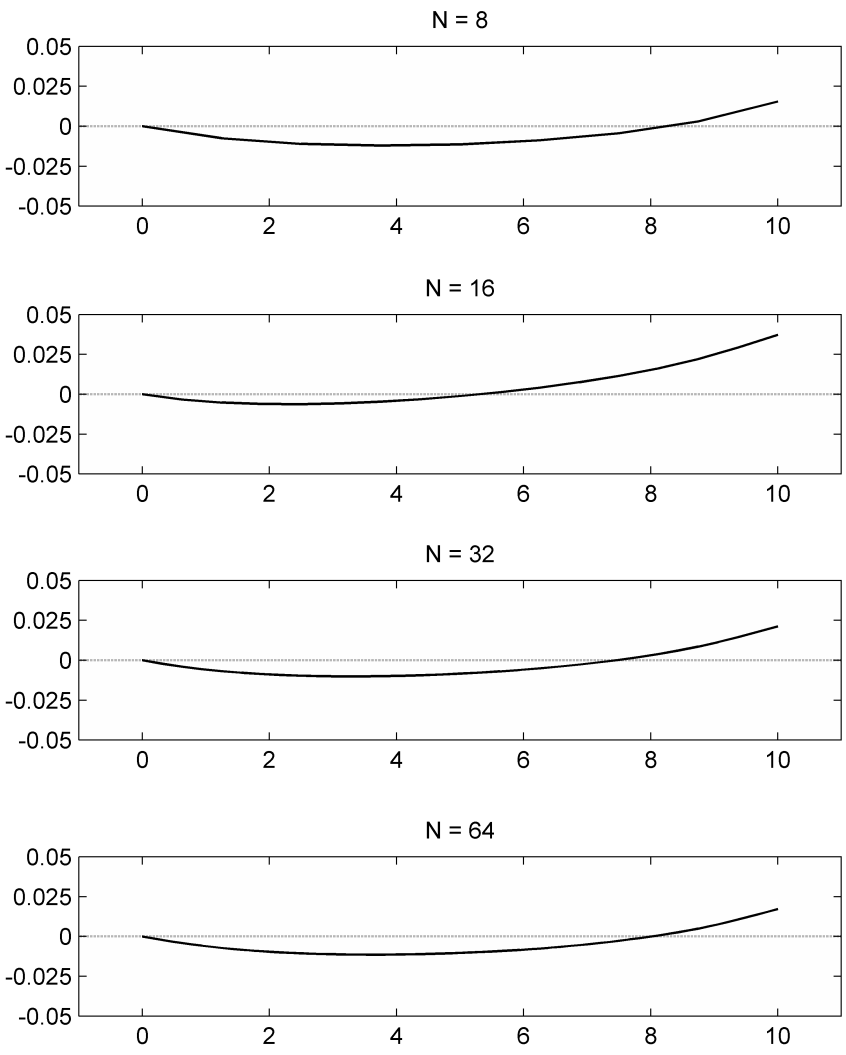

Figure 14: Optimal deflection of the beam

We have solved the problem using the PBUN algorithm with $N = 8, 16, 32, 64$. F^h with f defined by (164) fulfills the condition ($S4_h$) for all h. The algorithm has been run with the following arguments:
PBUN: MIT $= 1000$, MFV $= 1000$, MET $= 2$, MTESX $= 1000$, MTESF $= 1000$, IPRNT $= -2$, XMAX $= 0.5, 0.6$, TOLX $= 0$, TOLF $= 0$, TOLB $= 0$, TOLG $= 10^{-4}$, ETA $= 0.1, 0.2$.
The results are summarized in Figures 12-14 and Table 3. The optimal solutions are slightly different. It can be seen especially on the optimal subsoil stiffness, see Figure 13. The optimal thicknesses and optimal deflections look similar at the first sight, nevertheless there are also small deviations for different N.

7.3. The influence of the definition of U_{ad}

In the *third example* we shall illustrate the dependence of the optimal design on the parameters appearing in the definition of the set $\mathcal{U}^h$. Especially we will change the parameter γ_1. The load function f is piecewise constant and given by

$$f(x) = \begin{cases} 100 & x \leqslant \frac{15}{8} \vee x \geqslant \frac{65}{8}, \\ -33 & \frac{15}{8} < x < \frac{65}{8}. \end{cases} \tag{165}$$

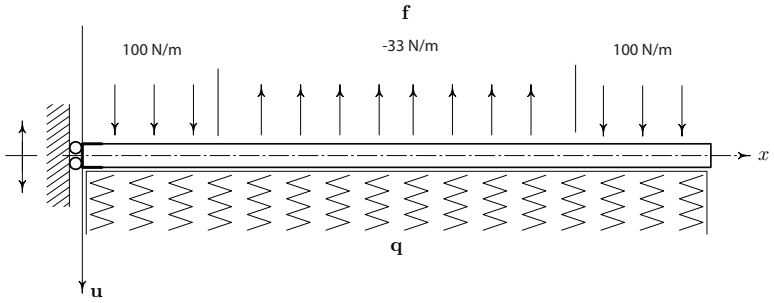

Figure 15: Outline of the beam with the load.

Let the set $\mathcal{U}^h$ be defined by the following parameters: $t_0 = 0.2$, $t_1 = 0.8$, $\gamma_1 = 5$, $q_0 = 500$, $q_1 = 1500$, $\gamma_3 = 10000$ and let the initial guess be $\mathbf{e}_0 = \{\mathbf{t}_0, \mathbf{q}_0\}$ where $t_i^0 = 0.5$, $i = 0, \ldots, N$ and $q_i^0 = 1000$, $i = 0, \ldots, N-1$. The initial cost functional value is 11.1931345. We used 32 finite elements in discretization; i.e., $N = 32$ and $h = 5/16$. The following boundary condition is prescribed: $u'(0) = 0$. The functional F^h with f given by (165) fulfills the condition ($S2_h$) and the cost functional corresponds to the compliance of the beam:

$$J(e) = \int_\Omega f\, u(e)\, \mathrm{d}x \approx \sum_{i=1}^{n+1} \frac{h}{2} \left(u_{i-1}\, f(x_{i-1}) + u_i\, f(x_i)\right) = \mathbf{u}^T \mathbf{B}\, \mathbf{f}.$$

The problem will be solved for $\gamma_2 = 0.1, 0.2, 0.3$. The other parameters remains the same. The input arguments of `PBUN` will be set as follows:
`PBUN`: `MIT` $= 1000$, `MFV` $= 1000$, `MET` $= 2$, `MTESX` $= 1000$, `MTESF` $= 1000$, `IPRNT` $= -2$, `XMAX` $= 0.7$, `TOLX` $= 0$, `TOLF` $= 0$, `TOLB` $= 0$, `TOLG` $= 10^{-4}$, `ETA` $= 0.1$.
The results are summarized in Table 4 and Figures 16-18.

The optimal foundation stiffness is the same for all three choices of γ_1. It can be seen that the final shapes of the beam are slightly different. If γ_1 grows the mass of the the beam is allowed to be re-distributed more efficiently and we are able to reach a better cost functional values (see Table 4 and Figures 16-18).

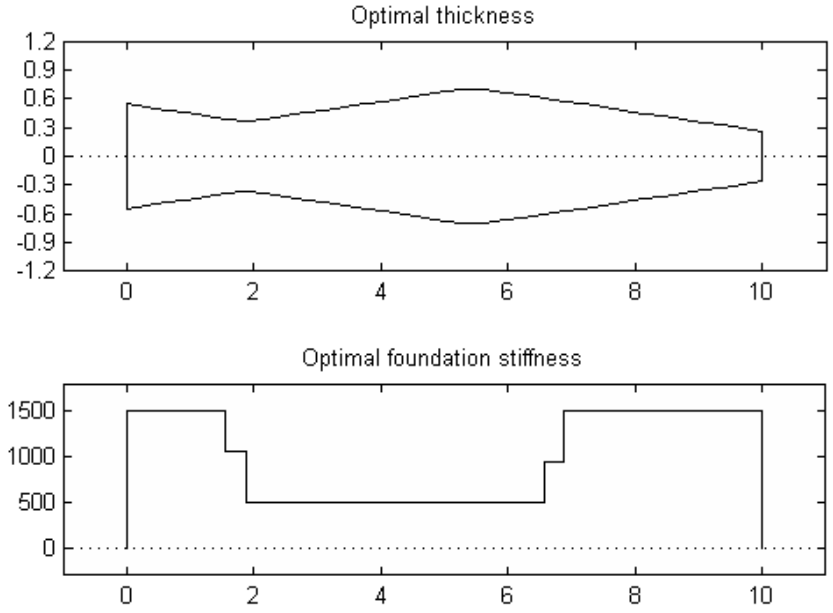

Figure 16: Optimal results for $\gamma_2 = 0.1$

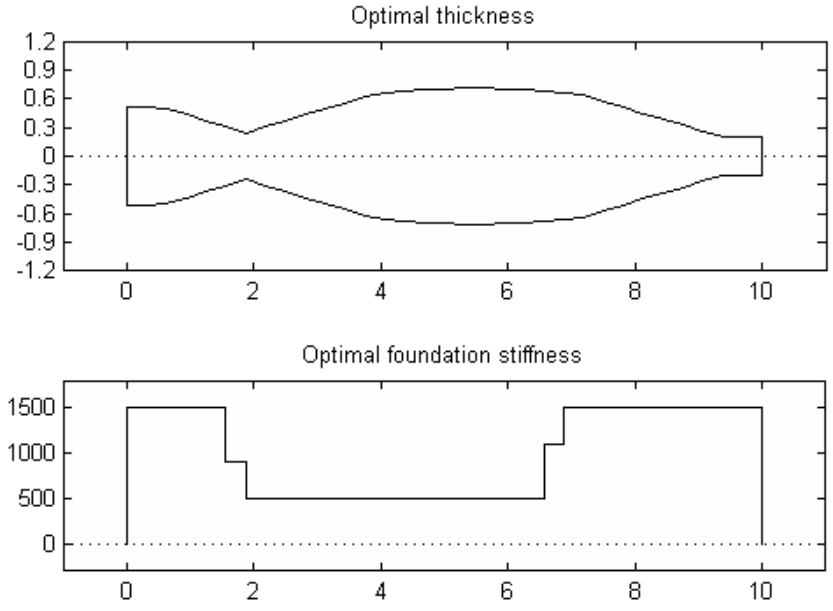

Figure 17: Optimal results for $\gamma_2 = 0.2$

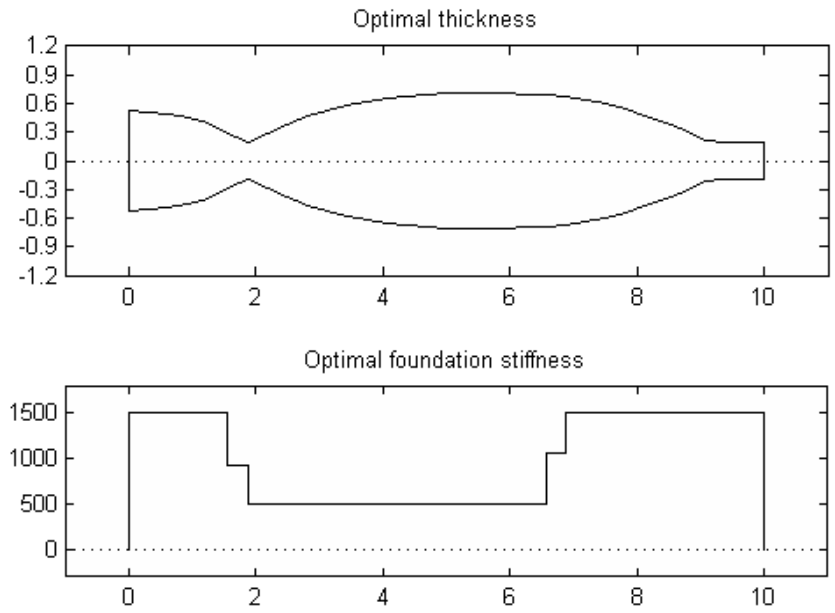

Figure 18: Optimal results for $\gamma_2 = 0.3$

Table 4: Results

γ_2	Final	Iter	Eval
0.1	8.50544174	23	35
0.2	8.20262861	25	34
0.3	8.16788551	28	33

7.4. The dependence of the optimal solution on the cost functional

In the *fourth example* we will present optimal results for three different cost functionals. The load f is piecewise constant and given by

$$f(x) = \begin{cases} -70 & x < 5, \\ 100 & x \geqslant 5. \end{cases} \tag{166}$$

Let the set $\mathcal{U}^h$ be defined by the following parameters: $t_0 = 0.2$, $t_1 = 0.8$, $\gamma_2 = 0.3$, $\gamma_1 = 5$, $q_0 = 500$, $q_1 = 1500$, $\gamma_3 = 10000$ and let the initial guess be $\mathbf{e}_0 = \{\mathbf{t}_0, \mathbf{q}_0\}$, where $t_i^0 = 0.5$ for $i = 0, \ldots, N$ and $q_i^0 = 1000$ for $i = 0, \ldots, N-1$. We used 32 finite elements in discretization; i.e., $N = 32$ and $h = 5/16$. The following boundary condition is prescribed: $u'(0) = 0$. Functional F^h clearly satisfies the condition ($S2_h$).

The optimal solution will be found with respect to the following three cost

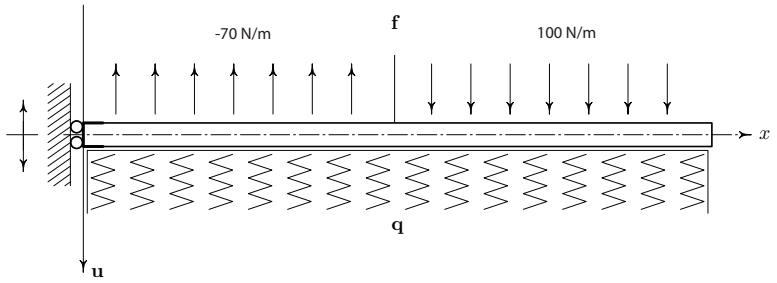

Figure 19: Outline of the beam with the load.

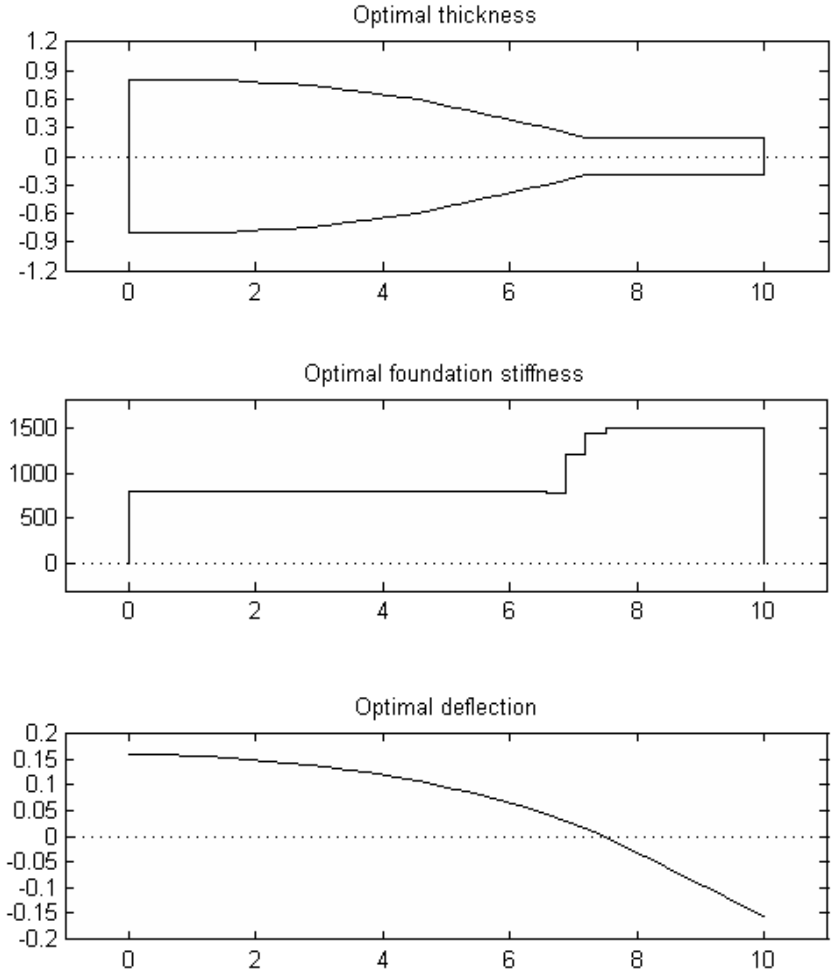

Figure 20: Optimal results for J_1

functionals:

$$J_1(e) = \int_\Omega f\, u(e)\,\mathrm{d}x \approx \sum_{i=1}^{n+1} \frac{h}{2}\,(u_{i-1}\, f(x_{i-1}) \,+\, u_i\, f(x_i)) \,=\, \mathbf{u}^T\mathbf{B}\,\mathbf{f},$$

$$J_2(e) = \int_\Omega u^2(e)\,\mathrm{d}x \approx \sum_{i=1}^{n+1} \frac{h}{2}\,(u_{i-1}^2 \,+\, u_i^2) \,=\, \mathbf{u}^T\mathbf{B}\,\mathbf{u},$$

$$J_3(e) = \int_\Omega t^2(u''(e))^2\,\mathrm{d}x \approx \sum_{i=1}^{n+1} \frac{h}{2}\,(t_{i-1}^2 u_h''(x_{i-1}) \,+\, t_i^2 u_h''(x_i)) \,=\, \mathbf{u}^T\mathbf{\Phi}^T\mathbf{E}^T\mathbf{E}\,\mathbf{\Phi}\,\mathbf{u}.$$

The other parameters of the problem remains the same.

We have solved the problem using the `PBUN` algorithm with the following arguments:

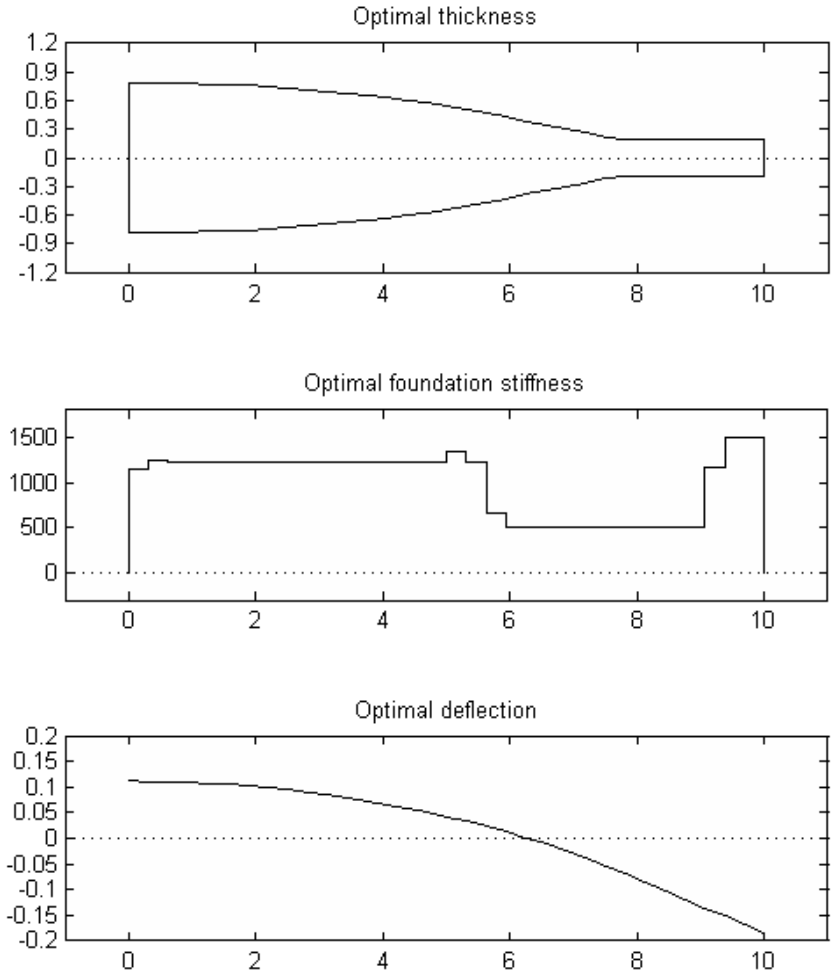

Figure 21: Optimal results for J_2

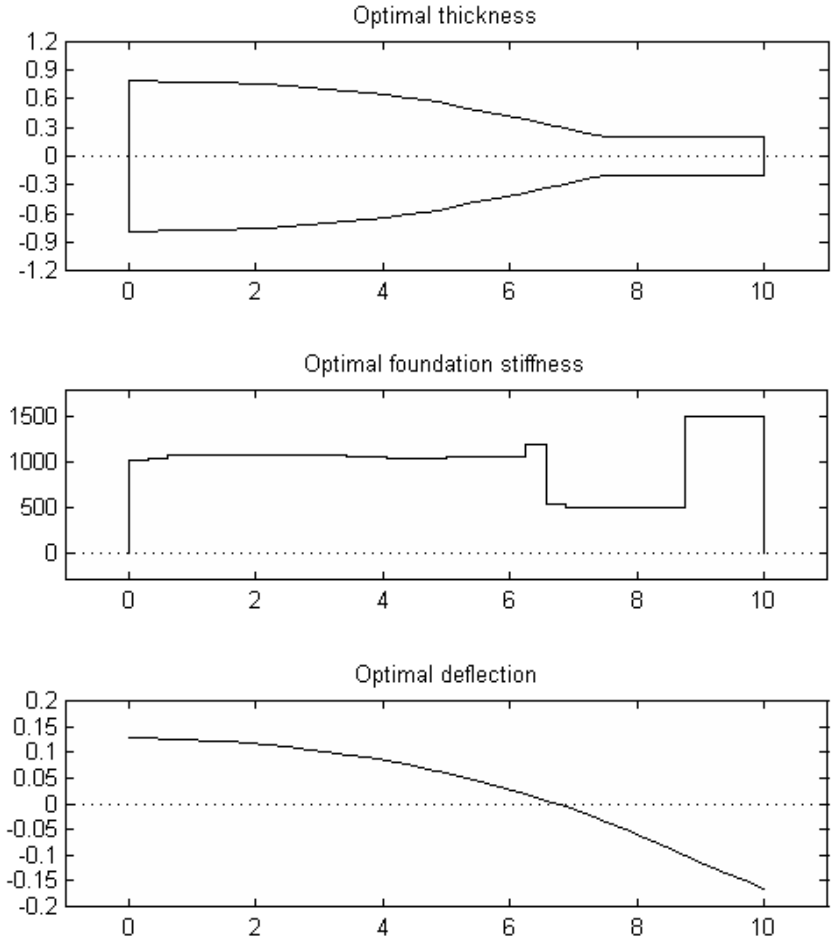

Figure 22: Optimal results for J_3

PBUN: MIT $= 1000$, MFV $= 1000$, MET $= 2$, MTESX $= 1000$, MTESF $= 1000$, IPRNT $= -2$, XMAX $= 1$, TOLX $= 0$, TOLF $= 0$, TOLB $= 0$, TOLG $= 10^{-4}$, ETA $= 0.3$.
The results are summarized in Table 5 and Figures 20-22.

Table 5: Results

Cfun	Initial	Final	Iter	Eval
J_1	132.235808	55.1465208	33	49
J_2	0.802462135	0.082413541	60	114
J_3	2.21450923	0.622296156	40	57

7.5. The influence of boundary conditions

In the *fifth example* we shall show the dependence of the optimal solution on the choice of boundary condition. The load function f is given in the following form:

$$f(x) = \begin{cases} -60 & x < 5, \\ 100 & x \geqslant 5. \end{cases} \tag{167}$$

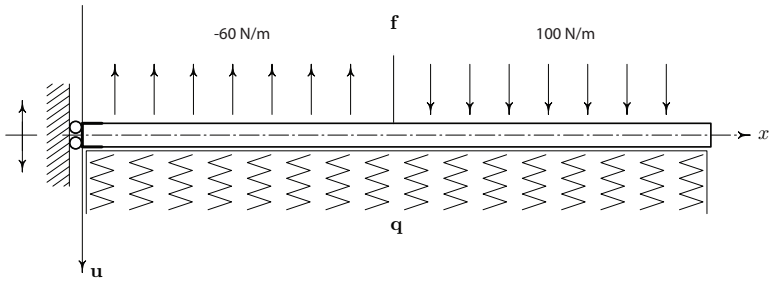

Figure 23: Outline of the beam with the load.

The cost functional is defined as follows:

$$J(e) = \int_{\Omega} u^2(e)\,\mathrm{d}x \approx \sum_{i=1}^{n+1} \frac{h}{2}\,(u_{i-1}^2 + u_i^2) = \mathbf{u}^T \mathbf{B}\,\mathbf{u}.$$

Let the set $\mathcal{U}^h$ be defined by the following parameters: $t_0 = 0.2$, $t_1 = 0.8$, $\gamma_2 = 0.3$, $\gamma_1 = 5$, $q_0 = 500$, $q_1 = 1500$, $\gamma_3 = 10000$ and let the initial guess be $\mathbf{e}_0 = \{\mathbf{t}_0, \mathbf{q}_0\}$, where $t_i^0 = 0.5$, $i = 0, \ldots, N$ and $q_i^0 = 1000$, $i = 0, \ldots, N-1$. We used 32 finite elements in discretization; i.e., $N = 32$ and $h = 5/16$. The problem will be solved with the boundary condition $u'(0) = 0$ and the boundary condition $u(0) = 0$. The condition ($S2_h$) and ($S4_h$) are clearly satisfied.

We have solved the problem by the `MPBNGC` algorithm with the following arguments:
`MPBNGC`: `iprint` $= 3$, `lmax` $= 100$, `jmax` $= 2n+1$, `niter` $= 1000$, `nfasg` $= 1000$, `gam` $= 0.3$, `rl` $= 0.1$, `eps` $= 10^{-4}$, `feas` $= 10^{-9}$.
The results are summarized in Table 6 and Figures 24, 25.

Table 6: Results

Bcondition	Initial	Final	Iter	Eval
$u'(0) = 0$	0.377676653	0.046572618	41	56
$u(0) = 0$	0.110807054	0.054800595	50	62

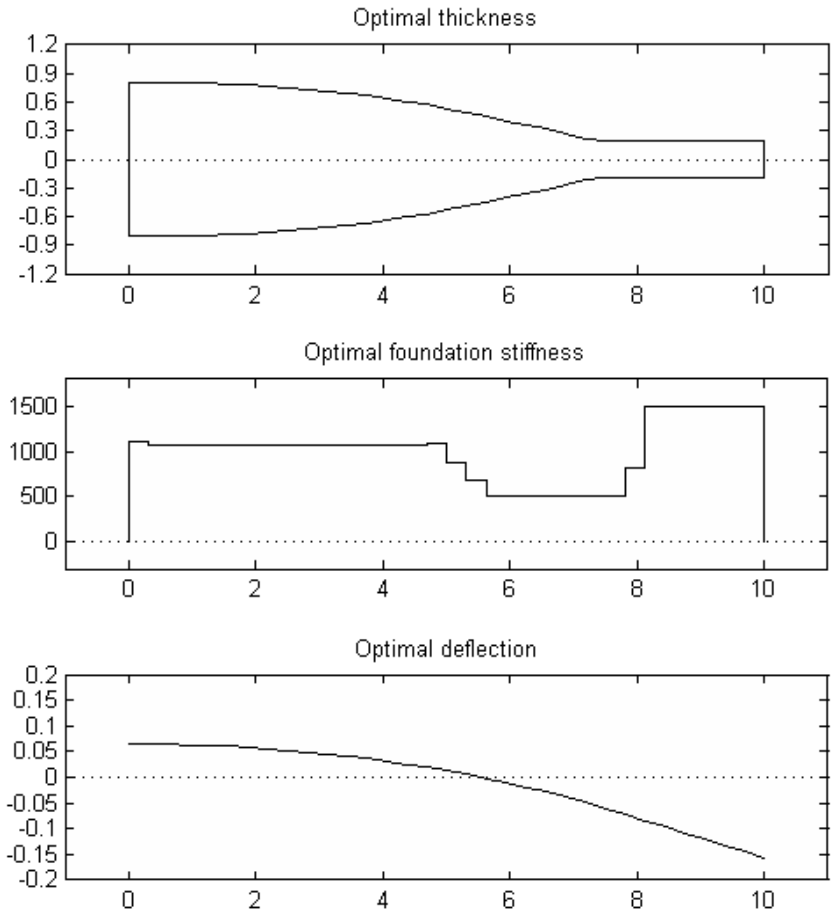

Figure 24: Optimal results for $u'(0) = 0$

The constraint $\int_\Omega t(x)\,\mathrm{d}x = \gamma_1$ keeps the beam volume fixed during the optimization process. The algorithm is allowed to prevent the deflection only by moving the mass of the beam. It can be seen from Figures 24, 25 that the optimal results are different. In the first case the mass of the beam is concentrated on the left end where the boundary condition is prescribed. It is typical for beams with free right end. In the second case the mass of the beam is on the one hand concentrated near the left end of the beam to take a benefit from the support. But on the other hand the mass is also concentrated near the middle and the right end of the beam to prevent the biggest deflection. The distribution of the mass of the beam depends on many aspects such as load distribution, boundary conditions etc.

It can be seen from the figures above that for given cost functional, the subsoil is usually distributed such that it is the stiffest at locations where the beam deflects at most.

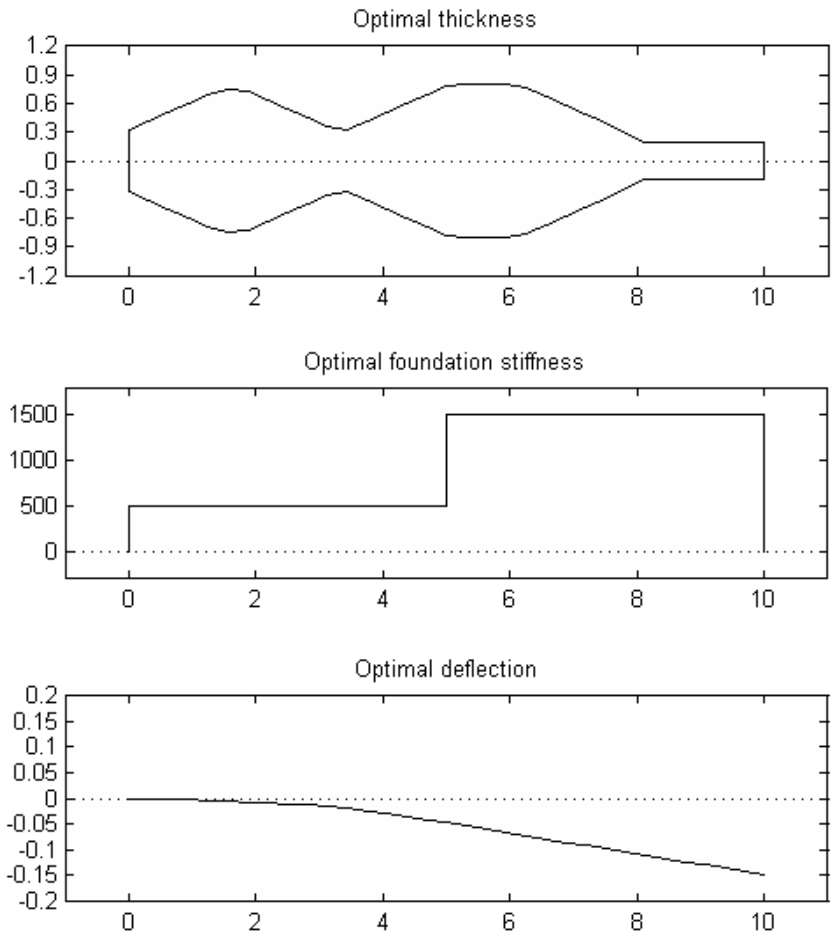

Figure 25: Optimal results for $u(0) = 0$

7.6. Computational time of some particular parts of the algorithm

In the *sixth example* we shall illustrate how much of the total computational effort take some particular pieces of the optimization code, especially computing of a solution to (mLCP(**e**)), computing of a solution to ($\mathcal{A}$(**e**)) and the computing of the gradient (subgradient) using the approach presented in Section 4.2. The load function f is given by

$$f(x) = \begin{cases} -50 & x < 5, \\ 100 & x \geqslant 5. \end{cases} \tag{168}$$

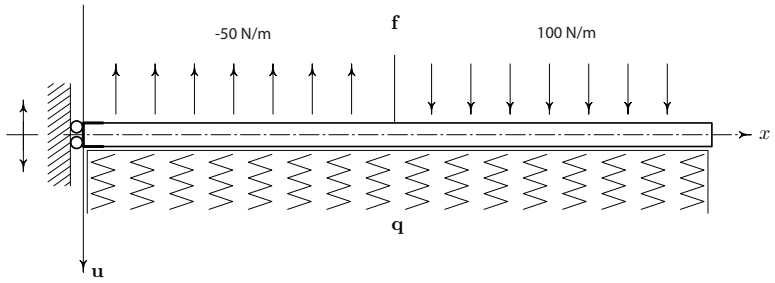

Figure 26: Outline of the beam with the load.

The cost functional is defined as follows:

$$J(e) = \int_{\Omega} f u(e)\,\mathrm{d}x \approx \sum_{i=1}^{n+1} \frac{h}{2}\left(u_{i-1}\, f(x_{i-1}) \,+\, u_i\, f(x_i)\right) \,=\, \mathbf{u}^T \mathbf{B}\,\mathbf{f}. \tag{169}$$

Let the set $\mathcal{U}^h$ be defined by the following parameters: $t_0 = 0.2$, $t_1 = 0.8$, $\gamma_2 = 0.3$, $\gamma_1 = 5$, $q_0 = 500$, $q_1 = 1500$, $\gamma_3 = 10000$ and let the initial guess be $\mathbf{e}_0 = \{\mathbf{t}_0, \mathbf{q}_0\}$ where $t_i^0 = 0.5$ for $i = 0, \ldots, N$ and $q_i^0 = 1000$ for $i = 0, \ldots, N-1$. We will run the algorithm with $N = 8, 16, 32, 64, 128, 256$. The boundary condition is $u'(0) = 0$. The condition ($S2_h$) is satisfied. We have solved the problem using the `PBUN` algorithm with the following arguments:
`PBUN`: `MIT` $= 1000$, `MFV` $= 1000$, `MET` $= 2$, `MTESX` $= 1000$, `MTESF` $= 1000$, `IPRNT` $= -2$, `XMAX` $= 1$, `TOLX` $= 0$, `TOLF` $= 0$, `TOLB` $= 0$, `TOLG` $= 10^{-4}$, `ETA` $= 0.3$.

Table 7: Results

	N					
Part of the code	8	16	32	64	128	256
SolveSP	0.145	0.422	0.848	11.269	40.255	602.442
SPinit	0.011	0.037	0.125	1.726	9.878	226.342
MakeLS	0.099	0.265	0.4	4.141	9.225	69.203
SolveLS	0.026	0.113	0.318	5.386	21.137	306.834
SolveAP	0.013	0.042	0.11	1.851	13.089	322.272
GradComp	0.047	0.174	0.423	9.485	53.379	768.593
RestC	0.171	0.21	0.397	1.032	3.171	30.359
Total	0.376	0.848	1.773	23.621	109.879	1723.67

Solution times in seconds are shown in Table 7. In Figures 27, 28 is presented the percentage of the computational time for particular pieces of the code. The following abbreviations are used: *SolveSP* = time needed to solve (mLCP(**e**)),

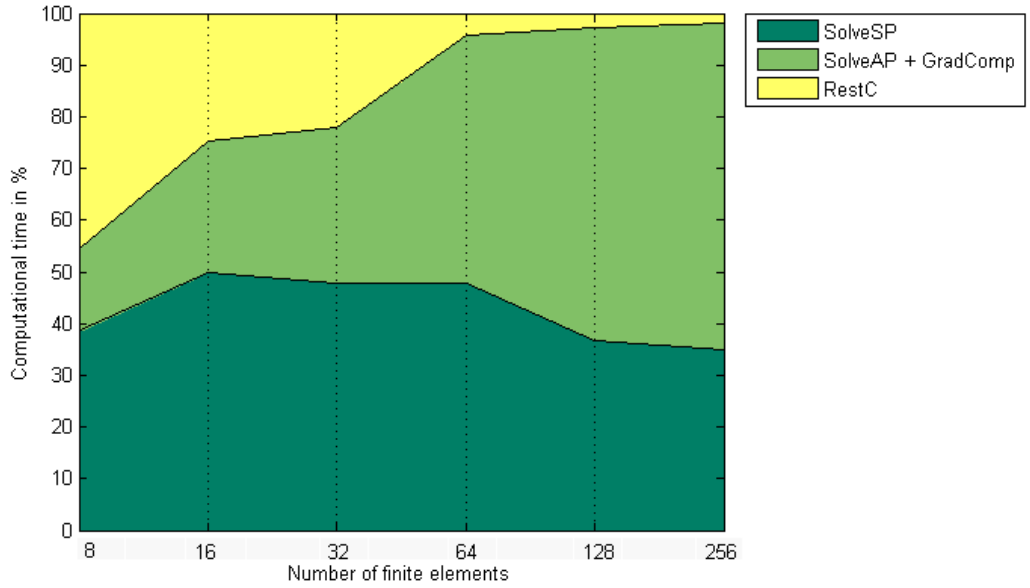

Figure 27: The amount of time spent by the algorithm (in percents)

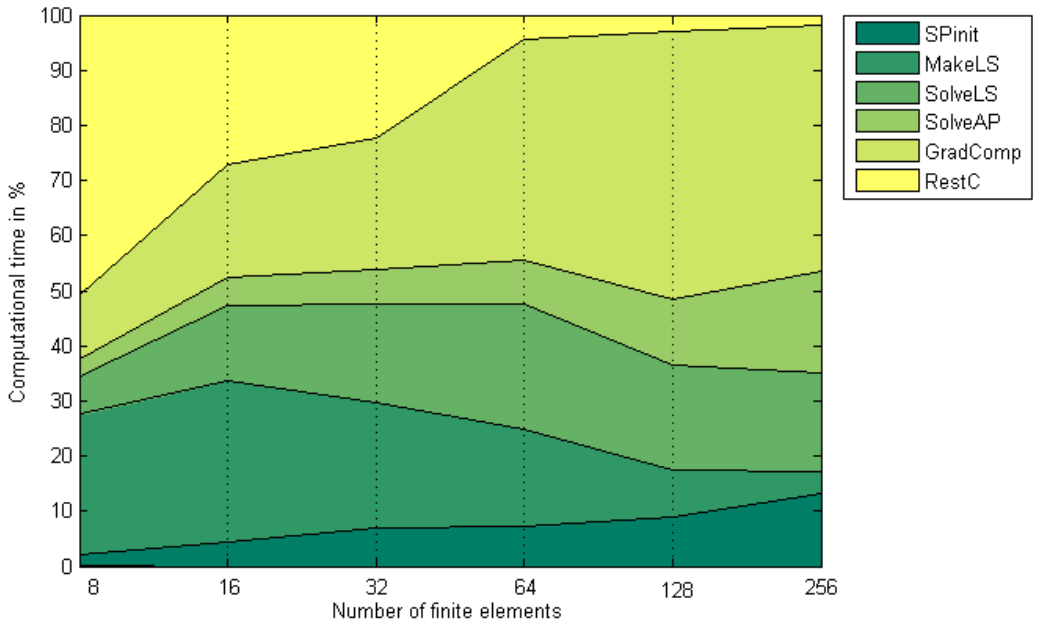

Figure 28: The amount of time spent by the algorithm

SPinit = time needed to compute the initial guess for the state problem solution algorithm (IPM), *MakeLS* = time needed to assemble the linear algebraic system for computation of the direction in IPM, *SolveLS* = time needed to solve the linear

algebraic system for computation of the direction in IPM, *SolveAP* = time needed to solve ($\mathcal{A}(\mathbf{e})$), *GradComp* = time needed to compute the gradient(subgradient) of the cost functional using relations (132), (136), *RestC* = time spent by of the rest of the algorithm (initializations, auxiliary computations), *Total* = The total computational time of the algorithm.

Firstly we divide the code into three parts SolveSP, SolveAP + GradComp and RestC. If $N = 8$ then the time needed to solve the state problem, the adjoint problem and to compute the subgradient (in the code it is represented by calls of functions `state_solver` and `sensitivity_analysis`), takes only about 50 percents of the total solution time. As the FEM mesh is getting finer and the number of elements is increasing, the amount of time spent by this two parts of the code is growing in comparison to the rest of the code. It can be clearly seen in Figure 27. This fact is reasonable because solution of the state problem and the design sensitivity analysis involve multiple solution of linear algebraic systems and many matrix multiplications. Therefore the number of operations grows bigger if the dimension of the problem rises.

In Table 7 and Figure 28 more detailed results are shown. The solution of the state problem (mLCP($\mathbf{e}$)) is here divided into three parts: SPinit, MakeLS, SolveLS. And the design sensitivity analysis is divided into two parts: SolveAP, GradComp. We can see that in case of low problem dimension ($N = 8, 16$), the biggest part of the computational time is taken by assembling the linear algebraic system (144) for IPM, computing the gradient using (132), (136) and the rest of the code. While in case of $N = 128, 256$ the biggest amount of operations is spent on computation of the subgradient, solution of the linear system (144) in IPM and solution of the adjoint problem ($\mathcal{A}(\mathbf{e})$).

8. Conclusions

In the thesis we have dealed with an application of mathematics in mechanics. Particularly we have considered the optimization of an elastic beam with a unilateral elastic foundation of Winkler type. The state problem was here represented by a boundary value problem for nonlinear ordinary differential equation of fourth order. Due to particular choice of boundary conditions and due to the unilaterality of the foundation the state problem was semicoercive. The objects of optimization were the beam thickness and the stiffness coefficient of the foundation.

Between the main results of the thesis it belongs the establishing of necessary and sufficient conditions for the existence and uniqueness of a solution to the state problem ($\mathcal{P}(e)$). We have also proved the continuous dependence of the state problem solution u on the design variable e and the existence of at least one solution to the design optimization problem (P). The Lipschitz continuity of the mapping $u : e \longmapsto u(e)$ and consequently the Lipschitz continuity of considered

cost functionals have been shown.

The problem has been approximated using the finite element method. The linear form F and the bilinear form b_q appearing in the variational formulation of the state problem have been approximated by a formula for numerical integration. As in the continuous case we have introduced necessary and sufficient conditions for the approximated state problem ($\mathcal{P}_h(e_h)$) and the existence of at least one solution of the approximated optimization problem (P_h) have been established.

Finally, for the numerical solution we have proposed the approach based on use of a nonsmooth optimization method for nonlinear programming. We have made the design sensitivity analysis and we have proposed a procedure for efficient computing of a gradient (subgradient) of the cost functional J_h. The main points of the design sensitivity analysis were the definition and solution of the adjoint problem and establishing of formula needed for computation of the subgradient from the state and adjoint solutions.

The whole procedure (with usage four different nonsmooth optimization methods `MPBNGC`, `PVAR`, `PBUN`, `PNEW`) have been implemented in C/C++ and Fortran. The code and its practical usage have been described. The results obtained by these four optimization methods have been presented, compared and analyzed. The influence of boundary conditions, cost functional, definition of $\mathcal{U}^h$ and the discretization parameter on the optimal design have been illustrated on several examples.

The results obtained in the thesis can be useful in the technical practice. Beams are widely used especially in the civil or railway engineering and in many other engineering applications. The studied unilateral (nonlinear) model of foundation is in some cases more precise as the widely used linear model of foundation. Therefore in some situations it is more correct to use the model which is in detail described and analyzed in the thesis. The main originality of the thesis insist in passing through the issues caused by the semicoercivity of the state problem and also in the possible nondifferentiability of the resulting optimization problem.

The thesis can be extended in many different ways. For example we can consider the foundation only on a part of the interval $[0, l]$ or we can consider a system of subsoils and topsoils situated in certain subintervals of $[0, l]$. Instead of Winkler model we can also consider the so call Pasternak's model of the foundation with response function $s(x) = q(x)u(x) - k(x)u''(x)$, where the second parameter $k(x)$ relates to the shear forces in the subsoil. The state problem can be also generalized to a 2D problem of a thin plate. The optimization part of the problem can be extended for example by adding the material coefficient or width of the beam as the design variables. We can optimize the beam with respect to many other cost functionals.

Appendix

In this section we will introduce some preliminary results that are used through the thesis.

Theorem 8.1. (Cauchy–Schwarz inequality.) *Let $f, g \in L^2(\Omega)$, where Ω is a nonempty open interval in $\mathbb{R}^1$. Then*

$$\int_\Omega fg \,\mathrm{d}x \leqslant \left(\int_\Omega f^2 \,\mathrm{d}x\right)^{1/2} \left(\int_\Omega g^2 \,\mathrm{d}x\right)^{1/2}. \tag{170}$$

Proof. For the proof we refer to [6]. ■

Theorem 8.2. (Discrete Cauchy–Schwarz inequality.) *Let $\mathbf{a} = (a_1, \ldots, a_n)$, $\mathbf{b} = (b_1, \ldots, b_n)$ be real vectors, then*

$$\Big(\sum_{i=1}^n a_i b_i\Big)^2 \leqslant \sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2. \tag{171}$$

Proof. For the proof see e.g. [27]. ■

Lemma 8.1. *Let Ω be a nonempty open interval in $\mathbb{R}^1$ and let the subspace of $H^2(\Omega)$ be defined as $V = \{v \in H^2(\Omega) : v'(0) = 0\}$. Further let $\{u_n\} \subset V$ be a sequence bounded in $H^2(\Omega)$ such that*

$$|u_n|_{2,2,\Omega} \to 0, \quad n \to +\infty.$$

Then there exists a subsequence $\{u_{n_j}\} \subset \{u_n\}$ and a polynomial $p \in P_0$ such that

$$u_{n_j} \to p \text{ in } H^2(\Omega), \quad j \to +\infty.$$

Proof. Since $\{u_n\}$ is bounded in $H^2(\Omega)$, there exists its subsequence $\{u_{n_j}\}$ and a function $u \in V$ such that $u_{n_j} \rightharpoonup u$ in $H^2(\Omega)$. By the well–known Rellich theorem (see e.g. [27]) we have $u_{n_j} \to u$ in $H^1(\Omega)$ and

$$\begin{aligned} \|u_{n_i} - u_{n_j}\|_{2,2,\Omega} &\leqslant |u_{n_i} - u_{n_j}|_{2,2,\Omega} + \|u_{n_i} - u_{n_j}\|_{1,2,\Omega} \leqslant \\ &\leqslant |u_{n_i}|_{2,2,\Omega} + |u_{n_j}|_{2,2,\Omega} + \|u_{n_i} - u_{n_j}\|_{1,2,\Omega} \to \\ &\to 0 \qquad i, j \to +\infty. \end{aligned}$$

Thus $u_{n_j} \to u$ in $H^2(\Omega)$ and $|u|_{2,2,\Omega} = 0$. Owing the fact that $u \in V$ we obviously have $u = p \in P_0$. ■

Lemma 8.2. *Let Ω be a nonempty open interval in $\mathbb{R}^1$ and let the subspace of $H^2(\Omega)$ be defined as $V = \{v \in H^2(\Omega) : v(0) = 0\}$. Further let $\{u_n\} \subset V$ be a sequence bounded in $H^2(\Omega)$ such that*

$$|u_n|_{2,2,\Omega} \to 0, \quad n \to +\infty.$$

Then there exists a subsequence $\{u_{n_j}\} \subset \{u_n\}$ *and a polynomial* $p = ax \in P_1$ *such that*

$$u_{n_j} \to p \text{ in } H^2(\Omega), \quad j \to +\infty.$$

Proof. See the proof of Lemma 8.1. ■

Theorem 8.3. (Sobolev embedding Theorem.) *Let* Ω *be a nonempty bounded interval in* $\mathbb{R}^1$. *Then the embedding of the space* $H^{k+1}(\Omega)$, $k = 1, 2, \ldots$ *into the space* $C^k(\bar{\Omega})$ *is continuous, i.e. there exists a constant* $c > 0$ *such that*

$$\|u\|_{C^k(\bar{\Omega})} \leqslant c \|u\|_{k+1,2,\Omega} \quad \forall u \in H^{k+1}(\Omega). \tag{172}$$

The constant c *is dependent only on the length of the interval* Ω *and on the parameter* k.

Proof. For the proof see e.g. [27]. ■

Theorem 8.4. (Rellich Theorem.) *Let* Ω *be a nonempty bounded interval in* $\mathbb{R}^1$. *Then the embedding of the space* $H^1(\Omega)$ *into the space* $L^2(\Omega)$ *is compact.*

Proof. For the proof we refer to [27]. ■

Remark 8.1. *Analogous inclusion can be derived by "translating" the derivatives. Therefore we have that* $H^{k+1}(\Omega)$ *is compactly embedded into* $H^k(\Omega)$ *for* $k = 1, 2, 3, \ldots$ *and there exists a constant* $c > 0$ *such that*

$$\|u\|_{k,2,\Omega} \leqslant c \|u\|_{k+1,2,\Omega} \quad \forall u \in H^{k+1}(\Omega). \tag{173}$$

Next we introduce some properties of the positive part u^+ of a function $u \in H^2(\Omega)$.

Lemma 8.3. *Let* Ω *be a nonempty open interval in* $\mathbb{R}^1$ *and* $u \in H^2(\Omega)$, *then the positive part*

$$u^+(x) = (u(x) + |u(x)|)/2, \quad x \in \Omega \tag{174}$$

belongs to the space $H^1(\Omega)$ *and* $\|u^+\|_{1,2,\Omega} \leqslant \|u\|_{1,2,\Omega}$. *Moreover, the following inequality holds:*

$$|u^+(x) - v^+(x)| \leqslant |u(x) - v(x)| \quad \forall u, v \in C(\bar{\Omega}),\ x \in \bar{\Omega}. \tag{175}$$

Proof. For the proof we refer to [50]. ■

Consequently, if there exists a constant $c_1 > 0$ such that $\|u\|_{2,2,\Omega} \leqslant c_1$, then there exists $c_2 > 0$ such that $\|u^+\|_{1,2,\Omega} \leqslant c_2$.

Lemma 8.4. *Let* Ω *be a nonempty open interval in* $\mathbb{R}^1$ *and* $u_n, u \in H^2(\Omega)$, $n \in \mathbb{N}$ *such that* $u_n \rightharpoonup u$ in $H^2(\Omega)$. *Then* $u_n \to u$ in $L^2(\Omega)$ *and in addition* $u_n^+ \to u^+$ in $L^2(\Omega)$.

Proof. The first part of the assertion is a consequence of the compactness of embedding $H^2(\Omega)$ into $L^2(\Omega)$, see e.g. [27]. The second part follows from the definition of the positive part (174). ■

Let us now summarize some basic properties of numerical quadrature that will be used for the approximation of the optimization problem. We define the numerical quadrature on the reference interval $[-1, 1]$ as follows

$$\int_{-1}^{1} \hat{\varphi}(\xi)\,\mathrm{d}\xi \;\approx\; \sum_{j=1}^{m} \hat{\omega}_j \hat{\varphi}(\hat{z}_j) \qquad \forall \hat{\varphi} \in W^{1,1}([-1,1]), \tag{176}$$

where $\hat{\omega}_j > 0$ and the points $\hat{z}_j$ belong to the reference interval $\forall j = 1, 2, \ldots, m$. We say that the quadrature formula is exact for polynomials of degree k at least if

$$\int_{-1}^{1} \hat{p}(\xi)\,\mathrm{d}\xi \;=\; \sum_{j=1}^{m} \hat{\omega}_j \hat{p}(\hat{z}_j) \qquad \forall \hat{p} \in P_k([-1,1]). \tag{177}$$

Next we can approach to the definition of a numerical quadrature on general interval $[s, t]$ with length $h > 0$. Transformation of the interval $[s, t]$ onto $[-1, 1]$ is given by

$$\Phi(x) \;:=\; \xi \;=\; \frac{h}{2}(x - s) - 1, \quad \forall x \in [s, t]. \tag{178}$$

Therefore

$$\int_{s}^{t} \varphi(x)\,\mathrm{d}x \;=\; \frac{h}{2} \int_{-1}^{1} \hat{\varphi}(\xi)\,\mathrm{d}\xi, \forall \varphi \in W^{1,1}([s,t]), \tag{179}$$

where $\hat{\varphi}(\xi) = \varphi(\Phi^{-1}(\xi))$. Corresponding numerical quadrature on $[s, t]$ is defined in the following way

$$\int_{s}^{t} \varphi(x)\,\mathrm{d}x \;\approx\; \sum_{j=1}^{m} \omega_j \varphi(z_j) \qquad \forall \varphi \in W^{1,1}([s,t]), \tag{180}$$

denoting $z_j := \Phi^{-1}(\hat{z}_j)$ and $\omega_j := (h/2)\hat{\omega}_j$.

Lemma 8.5. *Let Ω be a nonempty open interval in $\mathbb{R}^1$ with length $h > 0$. Let the numerical quadrature formula be exact for polynomials of degree $k \geqslant 0$ at least. Then there exists a constant $c > 0$ such that*

$$\left| \sum_{i=1}^{m} \omega_j \varphi(z_j) \;-\; \int_{\Omega} \varphi(x)\,\mathrm{d}x \right| \;\leqslant\; c h^{k+1} \,|\varphi|_{k+1,1,\Omega} \quad \forall \varphi \in W^{k+1,1}(\Omega). \tag{181}$$

Proof. For the proof see e.g. [7]. ■

And finally we introduce a which is a generalization of the classical chain rule of differentiation. This lemma plays an important role in establishing of formulas for subgradient calculation.

Lemma 8.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function defined by*

$$f(x) = J(x,\, y(x)), \tag{182}$$

where $J : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a continuously differentiable function and $y : \mathbb{R}^n \to \mathbb{R}^m$ is a locally Lipschitz mapping. If $\xi_y(x) \in \partial y(x)$, then

$$\nabla_x J(x,\, y(x)) \,+\, \xi_y^T(x)\, \nabla_y J(x,\, y(x)) \,\in\, \partial f(x). \tag{183}$$

Proof. For the proof we refer to [31]. ■

Bibliography

[1] Abramowitz, M. and Stegun, I.A.(Eds): *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* 9th printing. New York: Dover, pp. 888-890, 1972.

[2] Adams, R.: *Sobolev spaces.* Academic Press, New York, 1975.

[3] Aubin, J. P.: *Applied Functional Analysis, 2nd edition.* J. Wiley and Sons, New York, 2000.

[4] Bazaraa, M. S., Sherali , H. D., Shetty , C. M.: *Nonlinear programming: theory and algorithms, 3rd edition.* J. Wiley and Sons, New York, 2006.

[5] Begis, D., Glowinski, R.: *Application de la metode des elements finis a l'approximation d'un probleme de domaine optimal.* Appl. Math. Optim. 2, 130-169, 1975.

[6] Ciarlet, P.G.: *Introduction to Numerical Linear Algebra and Optimization.* Cambridge, Cambridge University Press, 1989.

[7] Ciarlet, P.G.: *The Finite Element Method for Elliptic Problems.* North Holland, 1979.

[8] Ekeland, I., Temam, R.: *Convex Analysis and Variational Problems.* North Holland, 1976.

[9] Facchinei, F., Pang, J. S.: *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I.* Springer-Verlag, New York, 2003.

[10] Facchinei, F., Pang, J. S.: *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume II.* Springer-Verlag, New York, 2003.

[11] Fletcher, R.: *Practical Methods of Optimization, 2nd edition.* J. Wiley and Sons, 2000.

[12] Fučík, S., Kufner, A.: *Nelineární diferenciální rovnice.* Praha, SNTL, 1978.

[13] Gill, P. E., Murray, W., Wright, M.H.: *Practical Optimization.* Academic Press, 1981.

[14] Glowinski, R.: *Numerical Methods for Nonlinear Variational Problems.* Springer-Verlag, Berlin, 2008.

[15] Golub, G. H., Van Loan, Ch. F.: *Matrix Computations.* John Hopkins University Press, 1996.

[16] Haslinger, J., Mäkinen, R. A. E.: *Introduction to Shape Optimization: Theory, Approximation and Computation*. SIAM, Philadelphia, 2003.

[17] Haslinger, J., Neittaanmäki, P.: *Finite Element Approximation for Optimal Shape, Material and Topology Design*. Second edition, John Wiley and Sons, Chichester, 1997.

[18] Haslinger, J.: *Metoda konečných prvků pro řešení eliptických rovnic a nerovnic*. Praha, SPN, 1980.

[19] Haslinger, J.: *A note on contact shape optimization with semicoercive state problems*. Applications of Mathematics, vol. 47 (2002), issue 5, pp. 397-410.

[20] Hildebrand, F.B.: *Introduction to Numerical Analysis*. New York: McGraw-Hill, pp. 343-345, 1956.

[21] Hlaváček, I.,Bock, I., Lovíšek, J.: *Optimal control of a variational inequality with applications to structural analysis. I. Optimal design of a beam with unilateral supports*. Appl. Math. Optimization 11, 1984, 111-143.

[22] Hlaváček, I.: *Weight Minimization of an Elastic Beam with Classical and Non-Classical Boundary Conditions*. ZAAM - Journal of Applied Mathematics and Mechanics, Volume 67 Issue, 8, pages 345 - 408, 1987.

[23] Horák, J. V., Fibinger, P.: *O rešitelnosti semikoercivních úloh ohybu desek – nerovnice*. Sborník konference Olomoucké dny aplikované matematiky ODAM2001, 75–100, KMAaAM, PřF UP Olomouc, 2001.

[24] Horák, J. V., Netuka, H.: *Mathematical model of pseudointeractive set: 1D body on nonlinear subsoil: I. Theoretical aspects*. Engineering Mechanics, Vol. 14, 2007.

[25] Horák, J.V., Šimeček, R.: *ANSYS implementation of shape design optimization problems*. 1. ANSYS conference 2008, 16. ANSYS FEM Users' Meeting, Luhačovice 5.-7. November 2008, 32 pages, released on CD, published by SVS-FEM, Brno, 2008.

[26] Chleboun, J.: *Optimal design of an elastic beam on an elastic basis*. Applications of Mathematics, Vol. 31 (1986), issue 2, pp. 118-140.

[27] Kufner, A., John, O, Fučík, S.: *Function Spaces*. Academia, Praha, 1977.

[28] Lukšan, L., Vlček, J.: *Globally convergent variable metric method for convex nonsmooth unconstrained minimization*. Journal of Optimization Theory and Applications, 102, 593–613, 1999.

[29] Lukšan, L., Vlček, J.: *A bundle-Newton method for nonsmooth unconstrained minimization*. Mathematical Programming, 83, 373–391, 1998.

[30] Lukšan, L., Vlček, J.: *NDA: Algorithms for Nondifferentiable Optimization*. Research Report V-797, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic, 2000.

[31] Mäkelä, M.M., Neittaanmäki, P.: *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control*. World Scientific Publishers, 1992.

[32] Mäkelä, M.M.: *Survey of bundle methods for nonsmooth optimization*. Optimization Methods and Software, Vol. 17(1), pp. 1–29.

[33] Mäkelä, M.M.: *Multiobjective proximal bundle method for nonconvex nonsmooth optimization: Fortran subroutine MPBNGC 2.0*. Reports of the Department of Mathematical Information Technology, Series B, Scientific computing, No. B 13/2003, University of Jyväskylä, Jyväskylä, 2003.

[34] Murty, K.G.: *Linear complementarity, linear and nonlinear programming*. Sigma Series in Applied Mathematics. 3. Berlin: Heldermann Verlag. ISBN 3-88538-403-5.

[35] Nečas, H., Hlaváček, I.: *Úvod do matematické teorie pružných a pružně plastických těles*. Praha, SNTL, 1981

[36] Netuka, H., Horák, J. V.: *Soustava nosník – pružiny – podloží po dvou letech*. Sborník konference Olomoucké dny aplikované matematiky ODAM2007, 18–42, KMAaAM, PřF UP Olomouc, 2007

[37] Machalová, J., Netuka, H., Šimeček, R.: *Shape optimization of a Timoshenko beam together with an elastic foundation*. Applied and Computational Mechanics, Vol 4, No 2 (2010), pp. 179 - 190. ISSN 1802-680X.

[38] Morales, J.L., Nocedal, J., Smelyanskiy, M.: *An algorithm for the fast solution of symmetric linear complementarity problems*. Numerische Mathematik, Volume 111 Issue 2, November 2008

[39] Nocedal, J., Wright, S. J.: *Numerical Optimization*. 2nd edition, Springer, 2006.

[40] Outrata, J. V.: *On the Numerical Solution of a Class of Stackelberg Problems*. ZOR - Methods and Models of Operations Research 34, 1990, 255-277.

[41] Pironneau, O.: *Optimal Shape Design for Elliptic Systems*. Springer Series in Computational Physics, Springer-Verlag, New York, 1984.

[42] Reddy, J.N.:*On locking free shear deformable beam finite element.* Computational Methods in Applied Mechanics and Engineering, 149 (1997), 113-132.

[43] Rektorys, K.: *Variationsmethoden in Mathematik, Physik und Technik.* Carl Hanser Verlag, Munich, 1984.

[44] Renegar, J.: *A mathematical view of interior-point methods in convex optimization.* SIAM, Philadelphia, 2001.

[45] Salač, P.: *Optimal Design of an Elastic Circular Plate on a Unilateral Elastic Foundation. I: Continuous Problems.* ZAMM - Journal of Applied Mathematics and Mechanic, Volume **82**, Issue 1, pages 21-32, January 2002. Zbl 1060.74054

[46] Salač, P.: *Shape optimization of elastic axisymmetric plate on an elastic foundation.* Applications of Mathematics, vol. **40** (1995), issue 4, pp. 319–338. Zbl 0839.73036

[47] Shrikhande, M.: *Finite element method and computational structural dynamics.* Textbook in preparation, Indian Institute of Technology Roorkee, Roorkee, 2008.

[48] Sokolowski, J., Zolesio, J.P.: *Introduction to Shape Optimization. Shape sensitivity analysis.* (Springer Series in Computational Mathematics 16.) Springer-Verlag, New York 1992.

[49] Solin, P.: *Partial Differential Equations and the Finite Element Method.* John Wiley and Sons,New Jersey, 2006.

[50] Sysala, S.: *Unilateral subsoil of Winkler's type: Semi-coercive beam problem.* Applications of Mathematics, Vol. 53 (2008), issue 4, pp. 347–379.

[51] Sysala, S.: *Numerical modelling of of semi–coercive beam problem with unilateral elastic subsoil of Winkler's type.* Applications of Mathematics.

[52] Šimeček, R.: *Optimal Design of an Elastic Beam with Unilateral Elastic Foundation.* Applications of Mathematics.

[53] Winkler, E.: *Die Lehre von der Elastizität und Festigkeit.* Dominicus, Prag, 1867.

[54] Wright, S.J.: *Primal-Dual Interior-Point Methods.* SIAM, Philadelphia, 1997.

[55] Zienkiewicz, O.C., Taylor, R.L.: *The Finite Element Method, Volume I: The Basis.* Butterworth-Heinemann, Oxford, 2000.

[56] Zienkiewicz, O.C., Taylor, R.L.: *The Finite Element Method, Volume II: Solid Mechanics.* Butterworth-Heinemann, Oxford, 2000.

Curriculum Vitae

Adresa

Roman Šimeček
Lískovec 268
798 07 Brodek u Prostějova
Česká republika
e-mail: simecekr@seznam.cz

Vzdělání

2001-2006: **Magisterské studium**, Univerzita Palackého v Olomouci, Přírodovědecká fakulta, Katedra matematické analýzy a aplikací matematiky

- Studijní obor: Aplikovaná matematika, matematické a počítačové modelování
- Téma diplomové práce: Řešení úlohy kvadratického programování metodou Goldfarba a Idnaniho
- Vedoucí diplomové práce: RNDr. Horymír Netuka, Phd.

od 2006: **Doktorské studium**, Univerzita Palackého v Olomouci, Přírodovědecká fakulta, Katedra matematické analýzy a aplikací matematiky

- Studijní obor: Aplikovaná matematika
- Téma dizertační práce: Design Optimization of a Beam on a Unilateral Foundation: Semicoercive State Problem
- Školitel: Jiří V. Horák, Lubomír Kubáček

Akademická stáž

březen - květen, 2010: **Univerzita v Turku, Finsko**, Numerická realizace úlohy optimalizace nosníku na jednostranném podloží použitím algoritmu pro nehladkou a nekonvexní optimalizaci.

Publikační činnost

1. Šimeček, R.: *A note on design optimization of a beam: Algebraic sensitivity analysis*. In: Book of Abstracts, ODAM 2011, Olomoucian Days of Applied Mathematics, Olomouc, January 26 - 28, 2011. ISBN 978- 80-244-2684-6.

2. Šimeček, R.: *Optimal Design of an Elastic Beam with an Unilateral Elastic Foundation: Semicoercive State Problem.* Applications of Mathematics, přijato k publikaci.

3. Machalová, J., Netuka, H., Šimeček, R.: *Shape optimization of a Ti- moshenko beam together with an elastic foundation.* Applied and Com- putational Mechanics, Vol 4, No 2 (2010), pp. 179 - 190. ISSN 1802- 680X.

4. Šimeček, R.: *Optimalizace nosníku na jednostranném podloží: Numerická real- izace.* In: Sborník: Moderní matematické metody v inženýrství, Dolní Lomná, 31. 5.-2. 6. 2010, JCMF a katedra MaDG VŠB-TU Ostrava, 2010. ISBN 978-80-248-2342-3 31.

5. Šimeček, R.: *Optimalizace nosníku na jednostranném podloží: Existence řešení.* str. 68-84, Sborník konference Olomoucké dny aplikované matematiky ODAM 2009, KMAaAM, PřF UP Olomouc, 2009.

6. Šimeček, R.: *Sizing Optimization of an Elastic Beam with a Rigid Obstacle: Numerical Realization.* Sborník SVOČ, Olomouc, 2009.

7. Horák, J. V., Šimeček, R.: *ANSYS Implementation of Shape Design Optimization Problems.* 32 stran, vyšlo na CD ve sborníku mezinárodní konference 16. ANSYS FEM User's Meeting and ANSYS CFD User's Meeting, 5.-7. listopadu v Luhačovicích, vydalo SVS-FEM Brno 2008.