JYX JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Haslinger, Jaroslav; Blaheta, Radim; Mäkinen, Raino A. E.

Title: Parameter identification for heterogeneous materials by optimal control approach with flux cost functionals

Year: 2021

Version: Accepted version (Final draft)

Copyright: © 2021 Elsevier

Rights: CC BY-NC-ND 4.0

Rights url: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the original version:

Haslinger, J., Blaheta, R., & Mäkinen, R. A. E. (2021). Parameter identification for heterogeneous materials by optimal control approach with flux cost functionals. Mathematics and Computers in Simulation, 189, 55-68. https://doi.org/10.1016/j.matcom.2020.06.009

Parameter identification for heterogeneous materials by optimal control approach with flux cost functionals

Jaroslav Haslinger[1], Radim Blaheta[2]

Institute of Geonics of the Czech Academy of Sciences, Studentská 1768, 708 00 Ostrava-Poruba, Czech Republic

Raino A. E. Mäkinen*

Faculty of Information Technology, University of Jyvaskyla, P.O. Box 35, 40014 Jyvaskyla, Finland

Abstract

The paper deals with the identification of material parameters characterizing components in heterogeneous geocomposites provided that the interfaces separating different materials are known. We use the optimal control approach with flux type cost functionals. Since solutions to the respective state problems are not regular, in general, the original cost functionals are expressed in terms of integrals over the computational domain using the Green formula. We prove the existence of solutions to the optimal control problem and establish convergence results for appropriately defined discretizations. The rest of the paper is devoted to computational aspects, in particular how to handle high sensitivity of the problem on the accuracy of data gained by measurements.

Keywords: identification of conductivity coefficients, optimal control of PDEs, sensitivity analysis
2010 MSC: 49J20, 35J86, 65K15, 65N30, 90C30

1. Introduction

Inverse problems of material parameter identification play a significant role in many fields of engineering, especially in situations when material testing by classical procedures fails to provide material parameters necessary for various mathematical models based on partial differential equations. Using flow in porous media as an example, the classical material testing consists of taking material samples and testing permeability by laboratory devices. There are several drawbacks of this classical procedure, especially when we want to solve geotechnical problems:

*Corresponding author
Email addresses: `hasling@karlin.mff.cuni.cz` (Jaroslav Haslinger), `blaheta@ugn.cas.cz` (Radim Blaheta), `raino.a.e.makinen@jyu.fi` (Raino A. E. Mäkinen)

[1]J. Haslinger and R. Blaheta acknowledge the support of the grant 19-11441S of the Grant Agency of the Czech Republic.

[2]The work of R. Blaheta was supported by the EURAD project, European Joint Programme (EJP) Cofund Action 847593.

- The heterogeneity of geomaterials existing even on the small scale causes that different permeability values are obtained from different samples despite that they correspond to the same geological type. To avoid this effect, the size of the sample should correspond to a representative volume of the material which can have too large dimensions for laboratory testing. In this case, large-scale in situ tests have to be performed in dimensions proper to the heterogeneity and their evaluation needs to solve inverse identification problems.

- For understanding of the role of the microstructure of laboratory size samples, the standard tests are not sufficient as they only allow to evaluate global (effective) response without possibility to identify the local material characteristics.

For geotechnics dealing with processes in heterogeneous geomaterials and geological environment, the use of inverse identification problems is therefore very desirable. But usually we meet another problem - the lack of measurements necessary as input for the identification. The in-situ tests usually concern the existing geological situation and the measurements can be done only on the earth surface, the surface of underground openings or in specially prepared boreholes. The material properties are evaluated from a response to external influences as mechanical loads, sources of fluid or heat etc. An example can be the pumping test in hydrogeology, i.e. pumping water to some boreholes and measuring the reaction in pressure or outflow in other boreholes. Another example can be the measurement of deformations due to excavation or blasting in some distance from the measurements. In general, for in-situ tests there is only little and localized input information, moreover frequently corrupted by measurements inaccuracy (a noise).

Even classical laboratory tests on smaller size samples do not provide enough information for finding local material properties and understanding the influence of inner microstructures to the overall behaviour. As a particular problem of this type, we can mention the understanding of the influence of grouting to mechanical and hydraulic properties, see e.g. [3].

To reduce the noise (accuracy) of measurements, we shall prefer the quantities excluding very local effects by averaging. To balance the small amount of input information provided by measurements with the number of required outputs, we take two measures. If possible, the amount of measured data is increased by repeating the tests with different external influences (loading and sources inside the investigated domain or on its boundary). On the other hand, the reduction of the amount of the required output information is done by introducing apriori knowledge about the partition of the considered domain into several parts where homogeneous material can be assumed. Consequently, the identification concerns only a (small) number of parameters which represent the material properties in homogeneous parts. Note that we do not require that the homogeneous parts are continuous, on the contrary, they can be very discontinuous e.g. when a binary material (a mixture of two materials) is considered.

The accuracy of the identification is still influenced by the noise in the measurements and accuracy of the provided decomposition of the investigated domain into homogeneous parts. In geotechnical applications, the above domain decomposition can be done by extrapolation of geophysical investigations for in-situ applications or by computer tomography in the case of laboratory tests. The geophysical investigation or tomography can be used not only for the determination of the material interfaces but also for getting a guess of the type of material in individual parts of the considered domain and consequently a guess of their material properties. As shown later, such guess is beneficiary for both regularization and starting an iterative

optimization procedure. Note that the example of layered material indicates that the identified parameters can be more influenced by the volume fractions of the individual homogeneous subdomains than by the exact position of the material interfaces.

In this paper, we consider a model problem of the Darcy flow in saturated piecewise homogeneous, isotropic porous media governed by the equation $-\operatorname{div}(k\nabla u) = f$, where the unknown u has the physical meaning of the pressure and f is a source term (fluid flow rate). We are interested in identification of the unknown permeability k.

The identification is made by measuring the amount of water entering and leaking out of the sample. We start with a single experiment clarifying how to get data needed in the identification problem. Consider a very simple situation depicted in Figure 1: Γ_1 and Γ_2 are inflow segments. Above them there are water columns of a constant height determining a constant pressure on each Γ_i, i=1, 2. At the same time we measure the total amount of the water penetrating into the sample, i.e. the constants c_1, c_2 (to keep a constant height, water is continuously refilled). On the outflow segment Γ_3 we prescribe the ambient pressure, e.g. u=0 and measure the amount of leaked out water, i.e. the constant c_3.

As the flow velocity is given by $\boldsymbol{v} = -k\nabla u$ the measured amount of inflow and outflow determines $\int_{\Gamma_i} k\frac{\partial u}{\partial n}\, ds$ on Γ_i, i=1, 2, 3.

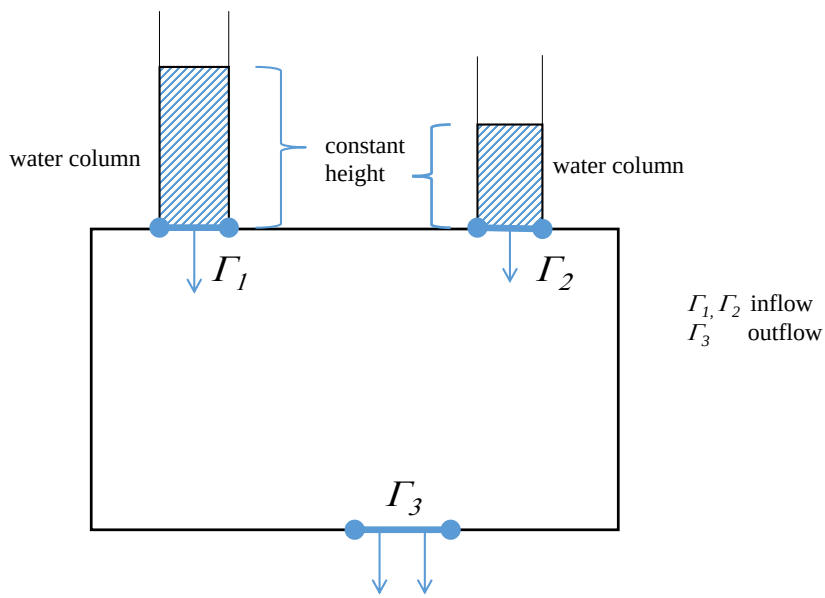

Figure 1: Schematic diagram of the experimental setup. The flux $\int_{\Gamma_i} k\frac{\partial u}{\partial n}\, ds$ is known on the inflow and outflow boundary segments Γ_1, Γ_2, and Γ_3.

Note that the classical identification experiment considers only inflow in one segment (the whole upper side) and outflow from another segment (the whole bottom side). Such experiment is able to identify the permeability of isotropic homogeneous material or effective permeability in one direction of heterogeneous material. For identification of the permeabilities of several parts of a material sample we need more experiments, which are realized by changing the location (and possibly also the number) of inflow/outflow segments along the boundary. To simplify our presentation we shall restrict ourselves to one experiment, which is enough for theoretical analysis of the problem.

The corresponding mathematical model leads to an optimal control problem driven by the scalar, second order elliptic equation whose permeability coefficient k plays the role of the control variable. Optimal control of systems by coefficients of PDE's is the topic which is nowadays very well studied in the literature (e.g. [5], [9], [8]) and the references therein. The overall majority of papers deals with least-squares type approaches using measured data which are available in the whole computational domain or its boundary or their substantial parts. Let us observe that the cost functional in our problem is not of this type. It is defined by a sum of integral mean values of fluxes on a system of inflow and outflow boundary segments. The individual functionals contain only modest information for identification purposes with consequences for the numerical realization mentioned above. On the other hand, the theoretical analysis of this problem is standard. It is based on continuity and compactness arguments to prove the existence of a solution and density and basic properties of finite element spaces to establish convergence results [7].

The main difficulty we face in this problem is a high sensitivity of the final result on the accuracy of measurements. If the measurements are noise-free then the material parameters can be identified practically exactly. On the other hand, if the measurements are polluted by a noise, then some material components can be very far from the true value. To involve uncertainties of measured data the authors used in [2] the Bayesian inverse which seems to be robust and providing more information about the identified parameters. The present paper uses the Tikhonov regularization, the advantage of which is an easy implementation and cheaper computations comparing with the Bayesian inverse. In this paper, we consider a linear model. Nonlinear problems can arise e.g. due to the presence of fractures in the material, see [1]. When considering time evolution, then nonlinearity can appear due to modelling of flow in unsaturated or variably saturated porous material. In mechanics, the coal-polyurethane composite from [3] was also investigated by nonlinear models. In [12] perfect plasticity and limit analysis have been used to determine uniaxial composite strength and the related failure zones.

The paper is organized as follows In Section 2 the optimal control problem is formulated with the cost functional based on averaged flux over boundary segments. Since the coefficients of the state equation are piecewise constant, one can not expect high global regularity of the solution. For this reason, the original cost functional is expressed in terms of integrals over the whole domain in which the state problem is defined. This is done in Section 3 where the new expression for the cost functional is also used for the proof that the identification problem has a solution. Section 4 is devoted to the discretization of the problem and convergence analysis. The rest of the paper deals with computational aspects. In Section 5 the algebraic form of the optimal control problem is presented. Finally, in Section 6 two model examples are solved numerically demonstrating that the proposed method works very well in practice.

2. Formulation of the problem

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with the Lipschitz boundary $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, where Γ_N and Γ_D are non-empty and disjoint. In addition, Γ_D consists of segments Γ_j, $j = 1, ..., m$:

$$\overline{\Gamma}_D = \bigcup_{j=1}^{m} \overline{\Gamma}_j, \ \exists \delta > 0 : \ \mathrm{dist}(\Gamma_i, \Gamma_j) \geq \delta, \ i \neq j, \ \text{and } m \geq 2. \tag{2.1}$$

Finally, Ω is decomposed into q subdomains $\Omega_i,\ i = 1, ..., q$:

$$\overline{\Omega} = \bigcup_{i=1}^{q} \overline{\Omega}_i, \quad \Omega_i \cap \Omega_j = \emptyset,\ i \neq j, \tag{2.2}$$

see Figure 2.

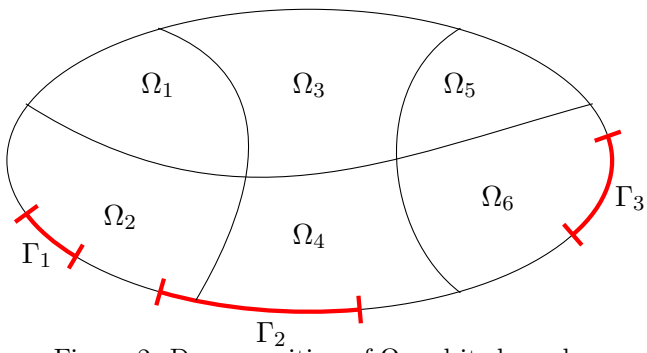

Figure 2: Decomposition of Ω and its boundary

For any positive function k such that $k_{|\Omega_i} \in P_0(\Omega_i)$, $i = 1, ..., q$ (i.e. k is piecewise constant over $\{\Omega_i\}$) we define the mixed Dirichlet–Neumann problem: Find $u := u(k)$ satisfying

$$\begin{cases} -\operatorname{div}(k\nabla u) = f & \text{in } \Omega \\ u = u_i & \text{on } \Gamma_i,\ i = 1, ..., m \\ k\dfrac{\partial u}{\partial n} = 0 & \text{on } \Gamma_N, \end{cases} \tag{$\widetilde{\mathcal{P}}(k)$}$$

where u_i are given Dirichlet data on Γ_i, $i = 1, ..., m$. The weak formulation of $(\widetilde{\mathcal{P}}(k))$ reads as follows:

$$\begin{cases} \text{Find } u := u(k) \in u_0 + V(\Omega) \quad \text{such that} \\ a(k, u, v) = (f, v)_{0,\Omega} \quad \forall v \in V(\Omega). \end{cases} \tag{$\mathcal{P}(k)$}$$

Here

$$V(\Omega) = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_D\},$$
$$a(k, u, v) = \int_\Omega k\nabla u \cdot \nabla v \, dx, \quad f \in L^2(\Omega),$$

and $u_0 \in H^1(\Omega)$ is such that $u_0|_{\Gamma_i} = u_i,\ i = 1, ..., m$.

Our aim will be to identify the function k on the basis of flux measurements on Γ_i, $i = 1, ..., m$. To this end we introduce the set

$$\mathcal{U}_{ad} = \{k \in L^\infty(\Omega) \mid 0 < k_{\min} \leq k \leq k_{\max},\ k_{|\Omega_i} \in P_0(\Omega_i),\ i = 1, ..., q\} \tag{2.3}$$

and the cost functional

$$J(k) = \tfrac{1}{2}\sum_{i=1}^{m} J_i(k) := \tfrac{1}{2}\sum_{i=1}^{m} \left(\int_{\Gamma_i} k\frac{\partial u(k)}{\partial n}\, ds - c_i\right)^2, \tag{2.4}$$

where $k_{\min}, k_{\max}$ are given, $u(k)$ is the solution of $(\mathcal{P}(k))$, $k \in \mathcal{U}_{ad}$, and $c_i \in \mathbb{R}$ are the flux measurements on Γ_i, $i = 1, ..., m$.

The identification problem is formulated as follows:

$$\begin{cases} \text{Find } k^* \in \mathcal{U}_{ad} \text{ such that} \\ J(k^*) \leq J(k) \quad \forall k \in \mathcal{U}_{ad}. \end{cases} \tag{$\mathbb{P}$}$$

3. Existence result with an equivalent computable expression of J

Instead of the curvilinear integrals defining the cost functional J (which needs an additional regularity of solution u) we give an equivalent expression of J in terms of integrals using Green's formula.

Let $u := u(k),\ k \in \mathcal{U}_{ad}$ be the solution to $(\mathcal{P}(k))$ and define the functional $L_u \in (H^1(\Omega))'$ by

$$L_u(v) = a(k, u(k), v) - (f, v)_{0,\Omega} \quad \forall v \in H^1(\Omega). \tag{3.1}$$

From the definition of $(\mathcal{P}(k))$ it follows that $L_u(v) = 0\ \forall v \in V(\Omega)$ implying $L_u(v_1) = L_u(v_2)$ for any $v_1, v_2 \in H^1(\Omega)$, $v_1 = v_2$ on Γ_D. Thus L_u can be considered as a linear functional on the trace space $H^{1/2}(\Gamma_D)$, where

$$H^{1/2}(\Gamma_D) = \{\psi \in L^2(\Gamma_D) \mid \exists v \in H^1(\Omega),\ v = \psi \text{ on } \Gamma_D\}.$$

Instead of $L_u(v)$ in (3.1) we shall write $\langle \mu_u, \psi \rangle$, $\mu_u \in (H^{1/2}(\Gamma_D))'$, where $\psi = v$ on Γ_D or simply $\langle \mu_u, v \rangle$. Thus (3.1) becomes

$$a(k, u(k), v) = (f, v)_{0,\Omega} + \langle \mu_u, v \rangle \quad \forall v \in H^1(\Omega). \tag{3.2}$$

In particular, choosing $v \equiv 1$ in Ω we have

$$\langle \mu_u, 1 \rangle = -\int_\Omega f\, dx. \tag{3.3}$$

Let us observe that if $u(k)$ is sufficiently regular then

$$\langle \mu_u, v \rangle = \int_{\Gamma_D} \frac{\partial u(k)}{\partial n} v\, ds.$$

In what follows we extend the additive property

$$\int_{\Gamma_D} k \frac{\partial u(k)}{\partial n} v\, ds = \sum_{i=1}^{m} \int_{\Gamma_i} k \frac{\partial u(k)}{\partial n} v\, ds$$

valid for regular solutions $u(k)$ of $(\mathcal{P}(k))$ to $\mu_u \in (H^{1/2}(\Gamma_D))'$ from (3.2). To this end define the spaces

$$V_i(\Omega) = \{v \in H^1(\Omega) \mid v = 0 \text{ on any } \Gamma_j,\ j \neq i\},\ i = 1, ..., m.$$

From (3.2) it follows that

$$\langle \mu_u, v \rangle = a(k, u(k), v) - (f, v)_{0,\Omega} \quad \forall v \in V_i(\Omega),\ i = 1, ..., m.$$

As before one can show that μ_u depends only on the trace of v on Γ_i provided that $v \in V_i(\Omega)$. Let $\mu_u^i \in (H^{1/2}(\Gamma_i))'$, $i = 1, ..., m$ be the functional over $H^{1/2}(\Gamma_i)$, where

$$H^{1/2}(\Gamma_i) = \{\psi \in L^2(\Gamma_i) \mid \exists v \in V_i(\Omega),\ v = \psi \text{ on } \Gamma_i\}.$$

Then[3]

$$a(k, u(k), v) = (f, v)_{0,\Omega} + \langle \mu_u^i, v\rangle \quad \forall v \in V_i(\Omega). \tag{3.4}$$

It is easy to show that any function $v \in H^1(\Omega)$ can be written in the form

$$v = v_0 + \sum_{i=1}^{m} v_i, \tag{3.5}$$

where $v_0 \in H_0^1(\Omega)$ and $v_i \in V_i(\Omega)$, $i = 1, ..., m$. To prove (3.5) we use the partition of unity technique. Let $\{Q_i\}_{i=0}^m$ be a covering of $\overline{\Omega}$ such that $\overline{\Gamma}_i \subset Q_i$, $Q_i \cap \overline{\Gamma}_j = \emptyset$, $\forall j \neq i$, $i = 1, ..., m$, and $\overline{Q_0} \subset \Omega$. Then there exist functions φ_i, $i = 0, ..., m$ such that

$$\varphi_i \in C_0^\infty(Q_i),\ 0 \le \varphi_i \le 1 \text{ in } Q_i,\ i = 0, ..., m \quad \text{and} \quad \sum_{i=0}^{m} \varphi_i \equiv 1 \text{ in } \Omega. \tag{3.6}$$

Let $v \in H^1(\Omega)$ and define

$$v_i = v\varphi_i, \quad i = 0, ..., m. \tag{3.7}$$

Then $v_0 \in H_0^1(\Omega)$, $\operatorname{supp} v_i \subset Q_i$ so that $v_i \in V_i(\Omega)$, $v_i = v$ on Γ_i for $i = 1, ..., m$ and (3.5) is satisfied. Inserting (3.5) into (3.2) we get:

$$\begin{aligned}\langle \mu_u, v\rangle &= a(k, u(k), v) - (f, v)_{0,\Omega} = a(k, u(k), v_0) - (f, v_0)_{0,\Omega} \\ &+ \sum_{i=1}^{m} [a(k, u(k), v_i) - (f, v_i)_{0,\Omega}] \overset{(3.4)}{=} \sum_{i=1}^{m} \langle \mu_u^i, v_i\rangle,\end{aligned} \tag{3.8}$$

using that $a(k, u(k), v_0) - (f, v_0)_{0,\Omega} = 0\ \forall v_0 \in H_0^1(\Omega)$. This proves the additive property of μ_u. From (3.8) we see that

$$\langle \mu_u^i, v_i\rangle = a(k, u(k), v_i) - (f, v_i)_{0,\Omega} = \int_{\operatorname{supp} v_i} (k\nabla u(k) \cdot \nabla v_i - f v_i)\ dx \tag{3.9}$$

holds for any v_i defined by (3.7). In particular, if $v \equiv 1$ in Ω, the decomposition (3.5) reads:

$$\varphi_0 + \sum_{i=1}^{m} \varphi_i = 1 \quad \text{in } \Omega.$$

Since $\varphi_i \in V_i(\Omega)$ and $\varphi_i = 1$ on Γ_i, $i = 1, ..., m$ we obtain from (3.9):

$$\langle \mu_u^i, 1\rangle = \int_{\operatorname{supp} \varphi_i} (k\nabla u(k) \cdot \nabla \varphi_i - f\varphi_i)\ dx. \tag{3.10}$$

[3]To simplify notation, the duality between $(H^{1/2}(\Gamma_i))'$ and $H^{1/2}(\Gamma_i)$ is still denoted by $\langle\ ,\ \rangle$.

In addition,

$$\sum_{i=1}^{m} \langle \mu_u^i, 1 \rangle = -\int_\Omega f\, dx. \tag{3.11}$$

This leads to another expression of $J(k)$ which will be used in computations and also in the forthcoming existence and convergence analysis, namely

$$J(k) = \tfrac{1}{2} \sum_{i=1}^{m} \left(\int_{\operatorname{supp} \varphi_i} \left(k \nabla u(k) \cdot \nabla \varphi_i - f \varphi_i \right) dx - c_i \right)^2. \tag{3.12}$$

Now we are ready to prove the existence of a solution to $(\mathbb{P})$ with the cost functional J defined by (3.12).

Define the control-to-state mapping $\Phi : \mathcal{U}_{ad} \to u_0 + V(\Omega)$ by

$$\Phi(k) = u(k) \in u_0 + V(\Omega), \quad k \in \mathcal{U}_{ad}$$

with $u(k)$ being solution to $(\mathcal{P}(k))$.

Lemma 3.1. *The mapping Φ is continuous in $\mathcal{U}_{ad}$:*

$$k_n \to k \text{ in } L^\infty(\Omega),\ k_n, k \in \mathcal{U}_{ad} \implies u(k_n) \to u(k) \text{ in } H^1(\Omega), \tag{3.13}$$

where $u(k_n), u(k)$ is the solution to $(\mathcal{P}(k_n))$, and $(\mathcal{P}(k))$, respectively.

Proof. Proof is standard and it can be omitted. □

From Lemma 3.1, compactness of $\mathcal{U}_{ad}$ in $L^\infty(\Omega)$ and continuity of J defined by (3.12) we obtain the following existence result.

Theorem 3.1. *Problem $(\mathbb{P})$ has a solution.*

4. Discretization of $(\mathbb{P})$ and convergence analysis

In this section we shall define the discrete version $(\mathbb{P}_h)$ of $(\mathbb{P})$ and study the mutual relation between $(\mathbb{P})$ and $(\mathbb{P}_h)$ if $h \to 0+$.

Next we shall suppose that Ω and all Ω_i, $i = 1, ..., q$ are polygonal domains. Let $\{\mathcal{T}_h\}$ be a regular family of triangulations of $\overline{\Omega}$ such that each $\mathcal{T}_h \in \{\mathcal{T}_h\}$ is consistent with the decomposition of $\partial\Omega$ into Γ_N, Γ_D and Ω into Ω_i, $i = 1, ..., q$. With any $\mathcal{T}_h$ we associate the finite element spaces

$$\widetilde{V}_h(\Omega) = \{v_h \in C(\overline{\Omega}) \mid v_h|_{T_i} \in P_1(T)\ \forall T_i \in \mathcal{T}_h\},$$
$$V_h(\Omega) = \{v_h \in \widetilde{V}_h(\Omega) \mid v_h = 0 \text{ on } \Gamma_D\}.$$

In addition, the function u_0 defining the Dirichlet data on Γ_D is supposed to belong to $H^2(\Omega)$. The discretization of $(\mathcal{P}(k))$, $k \in \mathcal{U}_{ad}$ reads as follows:

$$\begin{cases} \text{Find } u_h := u_h(k) \in u_{h0} + V_h(\Omega) \quad \text{such that} \\ a(k, u_h, v_h) = (f, v_h)_{0,\Omega} \quad \forall v_h \in V_h(\Omega), \end{cases} \tag{$\mathcal{P}_h(k)$}$$

where $u_{h0} = r_h u_0$ and r_h is the piecewise linear Lagrange interpolation operator on $\mathcal{T}_h$.

The discretization of the identification problem $(\mathbb{P})$ is defined by

$$\begin{cases} \text{Find } k^*(h) \in \mathcal{U}_{ad} \text{ such that} \\ J_h(k^*(h)) \le J_h(k) \quad \forall k \in \mathcal{U}_{ad}, \end{cases} \tag{$\mathbb{P}_h$}$$

where

$$J_h(k) = \tfrac{1}{2} \sum_{i=1}^{m} \left(\int_{\operatorname{supp} \varphi_i} (k \nabla u_h(k) \cdot \nabla r_h \varphi_i - f r_h \varphi_i) \, dx - c_i \right)^2, \tag{4.1}$$

$u_h(k) \in u_{h0} + V_h(\Omega)$ is the solution to $(\mathcal{P}_h(k))$ and $\{\varphi_i\}_{i=1}^m$ are the functions from (3.6). The pair $(u_h(k^*(h)), k^*(h)) \in (u_{h0} + V_h(\Omega)) \times \mathcal{U}_{ad}$ will be called an optimal pair of $(\mathbb{P}_h)$.

The following existence result is readily seen.

Theorem 4.1. *Problem $(\mathbb{P}_h)$ has a solution for any $h > 0$.*

In the remaining part of this section we shall study the relation between $(\mathbb{P})$ and $(\mathbb{P}_h)$ if $h \to 0+$.

To this end we shall need the following auxiliary result.

Lemma 4.1. *Let $k_n \to k$ in $L^\infty(\Omega)$, $n \to \infty$, $k_n, k \in \mathcal{U}_{ad}$ and $\{u_{h_n}(k_n)\}$ be the sequence of solutions to $(\mathcal{P}_{h_n}(k_n))$, where $h_n \to 0+$ if $n \to \infty$. Then*

$$u_{h_n}(k_n) \to u(k) \quad \text{in } H^1(\Omega) \tag{4.2}$$

$$J_{h_n}(k_n) \to J(k) \quad \text{as } n \to \infty. \tag{4.3}$$

In addition, $u(k)$ is the solution to $(\mathcal{P}(k))$.

Proof. We use the definition of $(\mathcal{P}_{h_n}(k_n))$:

$$\begin{cases} \text{Find } u_{h_n}(k_n) \in u_{h_n 0} + V_{h_n}(\Omega) \quad \text{such that} \\ a(k_n, u_{h_n}(k_n), v_{h_n}) = (f, v_{h_n})_{0,\Omega} \quad \forall v_{h_n} \in V_{h_n}(\Omega). \end{cases} \tag{4.4}$$

Since $u_0 \in H^2(\Omega)$ it holds that

$$r_{h_n} u_0 \to u_0 \quad \text{in } H^1(\Omega), \ n \to \infty. \tag{4.5}$$

From this and (4.4) we immediately obtain that $\{u_{h_n}(k_n)\}$ is bounded in $H^1(\Omega)$. Therefore there exists a subsequence of $\{u_{h_n}(k_n)\}$ (denoted by the same symbol) and a function $u \in u_0 + V(\Omega)$ such that

$$u_{h_n}(k_n) \rightharpoonup u \text{ (weakly) in } H^1(\Omega), \ n \to \infty. \tag{4.6}$$

To prove that u solves $(\mathcal{P}(k))$ we need the following density result (see [4]):

$$\forall v \in V(\Omega) \ \exists \{v_{h_n}\}, \ v_{h_n} \in V_{h_n}(\Omega): \quad v_{h_n} \to v \text{ in } H^1(\Omega), \ n \to \infty. \tag{4.7}$$

Letting $n \to \infty$ in (4.4) and using (4.5), (4.6), and (4.7) we prove that $u := u(k)$ solves $(\mathcal{P}(k))$. Owing to the uniqueness of the solution to $(\mathcal{P}(k))$, (4.5) holds for the whole sequence. Strong convergence of $\{u_{h_n}(k_n)\}$ to $u(k)$ can be proven in a standard way. To prove (4.3) we use (4.1), (4.2) and the fact that $r_{h_n} \varphi_i \to \varphi_i$ in $H^1(\Omega)$, $n \to \infty$, $\forall i = 1, ..., m$. □

Remark 4.1. Observe that if $\{\varphi_i\}_{i=0}^m$ is the system of functions satisfying (3.6) then $\{r_h\varphi_i\}_{i=0}^m$ shares the same properties.

The main result of this section is the following theorem.

Theorem 4.2. *For any sequence of optimal pairs* $\{(u_h(k^*(h)), k^*(h))\}$ *of* $(\mathbb{P}_h)$, $h \to 0+$ *there exists a subsequence* $\{(u_{h_n}(k^*(h_n)), k^*(h_n))\}$ *and a couple* $(u(k^*), k^*) \in (u_0 + V(\Omega)) \times \mathcal{U}_{ad}$ *such that*

$$u_{h_n}(k^*(h_n)) \to u(k^*) \quad \text{in } H^1(\Omega) \tag{4.8}$$

$$k^*(h_n) \to k^* \quad \text{in } L^\infty(\Omega),\ n \to \infty. \tag{4.9}$$

In addition, $(u(k^*), k^*)$ *is an optimal pair of* $(\mathbb{P})$. *Any accumulation point of* $\{(u_h(k^*(h)), k^*(h))\}$ *in the sense of* (4.8) *and* (4.9) *possesses this property.*

Proof. The existence of a subsequence $\{(u_{h_n}(k^*(h_n)), k^*(h_n))\}$ and a pair $(u(k^*), k^*)$ satisfying (4.8) and (4.9) follows from compactness of $\mathcal{U}_{ad}$ in $L^\infty(\Omega)$ and (4.2). To prove that $(u(k^*), k^*)$ is an optimal pair we use the definition of $(\mathbb{P}_{h_n})$:

$$J_{h_n}(k^*(h_n)) \le J_{h_n}(k) \quad \forall k \in \mathcal{U}_{ad} \tag{4.10}$$

Passing to the limit with $n \to \infty$, i.e. also $h_n \to 0+$ in (4.10) we obtain

$$J(k^*) \le J(k) \quad \forall k \in \mathcal{U}_{ad}$$

making use of (4.3). □

5. Numerical realization

Let $\mathbf{k} = (k_1, ..., k_q) \in \mathbb{R}^q$ be the vector containing the parameters defining the diffusion coefficient k. The nodal values of the approximate solution of the state problem $(\mathcal{P}(k))$ is obtained by the solution of the linear algebraic system

$$\mathbf{A}(\mathbf{k})\mathbf{u} = \mathbf{b},$$

where $\mathbf{A}(\mathbf{k})$ is the stiffness matrix (depending on $\mathbf{k}$) and $\mathbf{b}$ is the force vector. Let $\mathcal{J}_i(\mathbf{k})$ denote the discretized objective function $J_i(k)$. The evaluation of $\mathcal{J}_i(\mathbf{k})$ is done using (3.12) with φ_i given by the sum of the Courant basis functions at the nodes placed on the $\overline{\Gamma}_i$. Let $\mathbf{z}^i$ be the vector of nodal values of φ_i. Then the objective function in matrix form reads as

$$\mathcal{J}_i(\mathbf{k}) = \tfrac{1}{2}\left(\mathbf{u}^{\mathrm{T}}\widehat{\mathbf{A}(\mathbf{k})}\mathbf{z}^i - \widehat{\mathbf{b}}^{\mathrm{T}}\mathbf{z}^i - c_i\right)^2,$$

where $\widehat{\mathbf{A}(\mathbf{k})}$ and $\widehat{\mathbf{b}}$ are the stiffness matrix and force vector corresponding the state problem with the pure Neumann condition on $\partial\Omega$.

The matrix form of adjoint problem (A.2) is

$$\mathbf{A}(\mathbf{k})\mathbf{p}^i = \widehat{\mathbf{A}(\mathbf{k})}\mathbf{z}^i. \tag{5.1}$$

Finally the partial derivatives are computed using (A.3) resulting in

$$\frac{\partial \mathcal{J}_i(\mathbf{k})}{\partial k_j} = \left(\mathbf{u}^{\mathrm{T}}\widehat{\mathbf{A}(\mathbf{k})}\mathbf{z}^i - \widehat{\mathbf{b}}^{\mathrm{T}}\mathbf{z}^i - c_i\right)\mathbf{u}^{\mathrm{T}}\frac{\partial \mathbf{A}(\mathbf{k})}{\partial k_j}(\mathbf{z}^i - \mathbf{p}^i), \quad j = 1, ..., q, \quad i = 1, ..., m. \tag{5.2}$$

In practice the number of integral flux measurements per experiment is much less than the number of the subdomains. Therefore, it is very likely that we cannot determine the coefficient k by using one experiment only. Instead several experiments need to be done under different input conditions generating optimization problems $(\mathbb{P}_h)_\ell,\ \ell = 1, ..., L$. Here L stands for the total number of experiments corresponding to different positions (and possibly the number) of segments Γ_i which define their own cost functionals whose algebraic counterparts are denoted by $\mathcal{J}^1, ..., \mathcal{J}^L$. We can then use e.g. the scalarization approach and minimize their weighted sum

$$\mathcal{J}(\mathbf{k}) = \sum_{\ell=1}^{L} w_\ell \mathcal{J}^\ell(\mathbf{k}), \tag{5.3}$$

where $w_\ell > 0$ are suitable weights. The same approach has been used to a closely related problem in [6]. In the rest of the paper we assume $w_\ell = 1,\ \ell = 1, ..., L$.

6. Numerical examples

Let $\Omega =]0, 1[\times]0, 1[$ and let $f = 0$ in Ω. The state problem was discretized using a uniform triangular mesh. The finite element solver was implemented in MATLAB [10]. The mesh size $h = \frac{1}{40}$ was used. This mesh is fine enough and the use of slightly coarser or denser mesh does not result in essentially different results.

Throughout this section we assume that in addition to the knowledge of the exact boundaries of the subdomains $\{\Omega_i\}$ we know a reasonably good initial guess vector $\mathbf{k}^0$ that predicts the order of magnitudes of the components of the true $\mathbf{k}^{\mathrm{ref}}$. The availability of good $\mathbf{k}^0$ allows us to reduce the set of admissible parameters. Therefore we assume that $\mathbf{k} \in \mathbb{U}_{ad}$, where

$$\mathbb{U}_{ad} = \{\mathbf{k} \in \mathbb{R}^q \mid k_i^{\min} \le k_i \le k_i^{\max}, \quad i = 1, ..., q\}$$

is the discrete analogue of $\mathcal{U}_{ad}$ and $\mathbf{k}^{\min}, \mathbf{k}^{\max} \in \mathbb{R}^q$ are two given positive vectors.

As inverse problems are generally ill-posed, some form of an additional regularization is needed especially if there is some sort of noise present in the observations [11]. In what follows we utilize the following Tikhonov regularization with a weight $\rho \ge 0$:

$$\mathcal{J}_r(\mathbf{k}) = \mathcal{J}(\mathbf{k}) + \frac{\rho}{2}\sum_{i=1}^{q}(k_i - k_i^0)^2, \tag{6.1}$$

where $\mathcal{J}(\mathbf{k})$ is defined by (5.3). To choose the weight ρ optimally in the Tikhonov regularization is generally an unsolved problem. We experimented with the L-curve method [11]. For some subdomain decomposition/experiment/noise combinations it gave useful information for choosing ρ but for some other not. However, our experiences might not reflect the situation with real data. Therefore, further elaboration of this topic was left to future studies and in the following examples ρ was determined ad hoc.

Example 1. The setting of the (synthetic) identification problem is the same as in [2] (Section 3). The subdomain decomposition is depicted in Figure 3. We have three experiments with different locations for the inflow/outflow boundaries (see Figure 4). As suggested in [2] we used logarithmic transformation $\boldsymbol{\kappa} = \log(\mathbf{k})$ in the minimization of $\mathcal{J}$. This transformation brings the advance that the optimizers "see" variables that have the same order of magnitude. We simulate the measurements by computing the boundary fluxes c_j, $j = 1, ..., m$ from the numerical solution corresponding to the following known vector of permeabilities $\mathbf{k}^{\text{ref}} = \log([9000, 100, 5000, 4, 300, 2, 200])$. As the initial guess vector we used $\mathbf{k}^0 = [9, 5, 9, 1, 5, 1, 5] \approx \log([8103, 148, 8103, 3, 148, 3, 148])$. The lower and upper bounds were set to $\mathbf{k}^{\min} = \log([2000, 50, 2000, 1, 50, 1, 50])$, $\mathbf{k}^{\max} = \log([10000, 500, 10000, 5, 500, 5, 500])$, respectively. As the initial guess for the optimizer, the vector $\mathbf{k}^0$ was used.

Figure 3: Subdomain topology

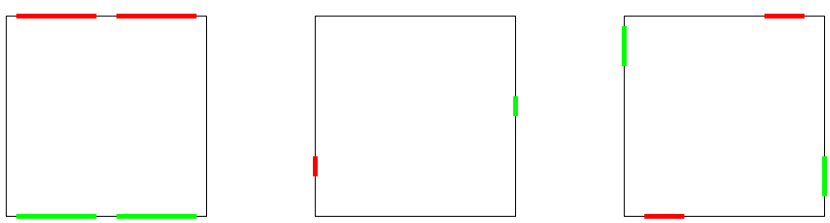

Figure 4: Location the of boundary measurements for three experiments. Input boundary segments ($u_i = 1$) are marked with red and output boundary segments ($u_i = 0$) are marked with green.

In optimization we used `fmincon` procedure with `active-set` option from the MATLAB Optimization Toolbox. The procedure applies a variant of the gradient based sequential quadratic progamming algorithm with a quasi-Newton approximation of the Hessian. The gradients were hand coded using the formulae (5.1), (5.2) derived in Section 5 and supplied to the optimizer. In the case of measurements without noise (and without regularization, i.e. $\rho = 0$), the optimizer found the correct parameter vector with a very high accuracy.

Next we introduced noisy measurements for each experiment. Let $\{c_j^{\text{ref}}\}$ be the set of fluxes

Table 1: Initial and final objective function values as well as the number of iterations and the number of objective function evaluations for Example 1.

	init.cost	final cost	iter/feval
clean	1.73×10^4	1.91×10^{-15}	37/90
noisy	1.79×10^4	1.05×10^{-5}	33/86
regul.	1.79×10^4	0.844	27/102

corresponding to the reference solution $\mathbf{k}^{\text{ref}}$. Then we define the perturbed fluxes $\{\tilde{c}_j\}$ as follows:

$$\tilde{c}_j = c_j + \eta_j, \quad \eta_j \sim \mathcal{N}(0, \sigma_j^2), \quad \sigma_j = p c_j^{\text{ref}}, \quad p \geq 0, \quad j = 1, ..., m.$$

Next the problem was solved with noisy data with the parameter $p = 0.02$. The effect of noise and regularization is demonstrated in Figure 5. The summary of the cost function value evolution and the number of iterations (and cost function evaluations) is shown in Table 1.

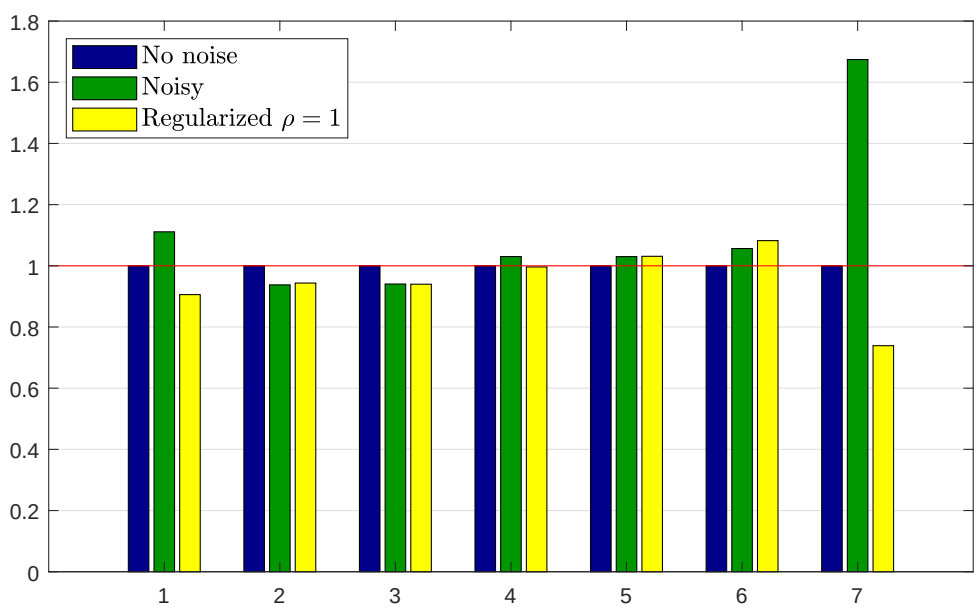

Figure 5: Ratio of the identified parameters k_i^* to the exact values k_i^{ref}. Blue column – no noise, green column – noise with $p = 0.02$ without the regularization; yellow column – noise with $p = 0.02$ and the regularization with the parameter $\rho = 1$.

Let us comment on the number of experiments needed. If there is q unknown coefficients k_i, then clearly we should have at least q extra equations for unique solvability. Due to the flux balance (our noisy data by its construction satisfies it, too) one of the fluxes is a linear combination of others (within one experiment). So in reality, if we have e.g. m flux measurements per experiment, we only have $m - 1$ new conditions per experiment. Thus, if we have $q = 9$ unknowns and we use two measurements per experiment, we should perform at least nine

experiments. So from the computational point of view, having as many flux measurements per experiment is advantageous, but not necessarily feasible in the laboratory environment.

Example 2. This is yet another synthetic model problem. The division of Ω into the subdomains is depicted in Figure 6. Unlike the previous example now there are two subdomains that do not meet the boundary $\partial\Omega$. Again the measurements are simulated by computing the fluxes corresponding to the vector $\mathbf{k}^{\text{ref}} = \log([9000, 100, 2, 2, 300, 200, 5000, 3000, 4])$. As the initial guess vector we used $\mathbf{k}^0 = [9, 5, 1, 1, 5, 5, 9, 8, 1] \approx \log([8103, 148, 3, 3, 148, 148, 8103, 2981, 3])$. The upper and lower bounds are $\mathbf{k}^{\min} = \log([2000, 50, 1, 1, 50, 50, 2000, 2000, 1])$ and $\mathbf{k}^{\max} = \log([10000, 500, 5, 5, 500, 500, 10000, 10000, 5])$. We performed four experiments with six or five flux boundaries depicted in Figure 7. That setup simulates the situation where fluxes can be measured only at the top and bottom of the specimen. The first and the second experiment (similarly the third and the fourth experiment) correspond to the case where we rotate the specimen 90 degrees between experiments.

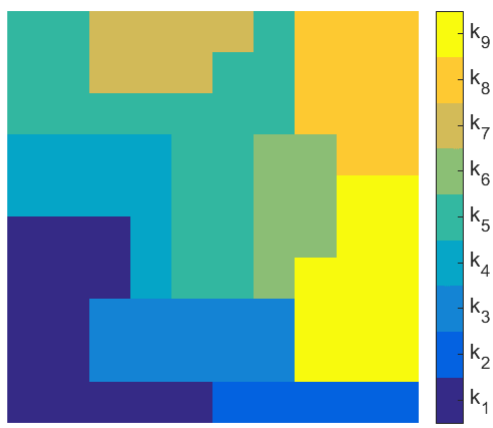

Figure 6: Subdomain topology

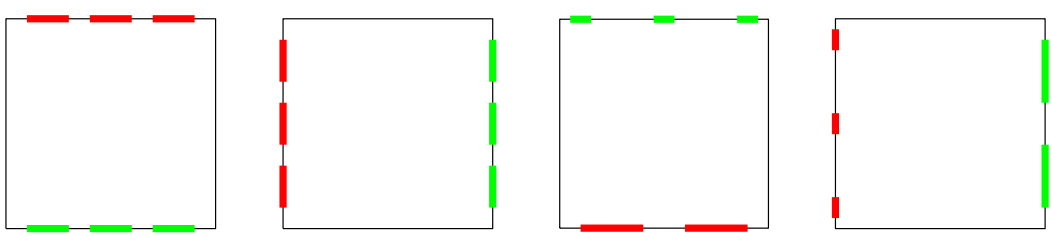

Figure 7: Locations of the boundary the measurements for four experiments. Input boundary segments ($u_i = 1$) are marked with red and output boundary segments ($u_i = 0$) are marked with green.

The results of optimization in the absense of noise and with noise are shown in Figure 8. The summary of the cost function value evolution and the number of iterations (and cost function evaluations) is shown in Table 2.

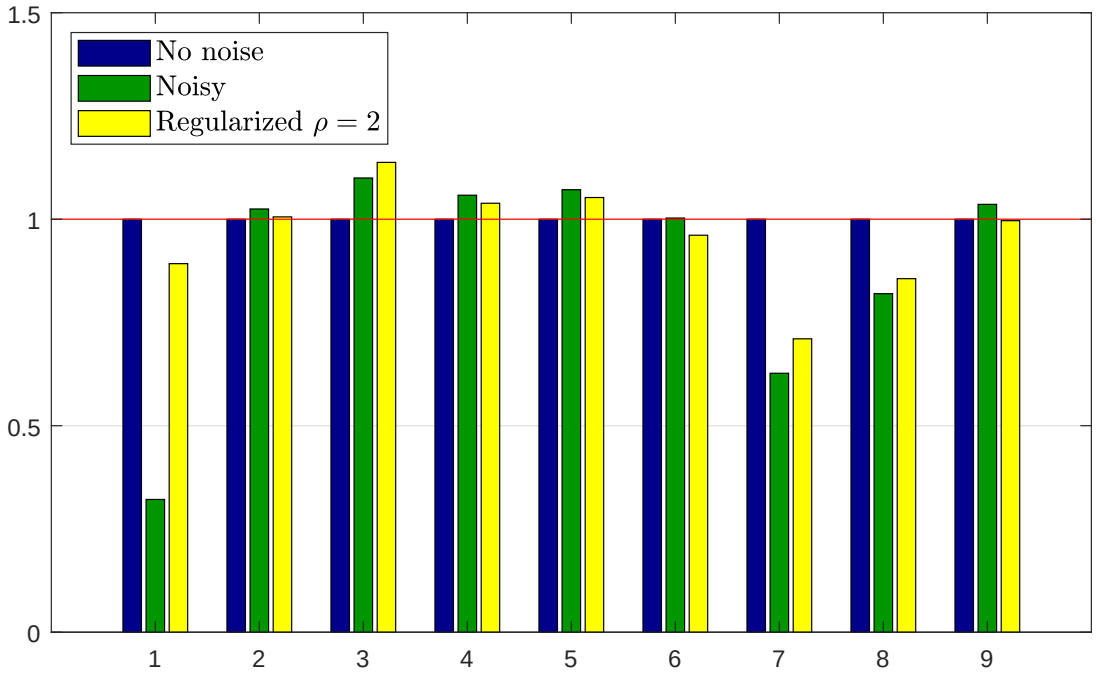

Figure 8: Ratio of the identified parameters k_i^* to the exact values k_i^{ref}. Blue column – no noise, green column – noise with $p = 0.02$ without the regularization; yellow column – noise with $p = 0.02$ and the Tikhonov regularization with the parameter $\rho = 0.05$.

7. Conclusions

The aim of this paper is to identify the unknown permeability coefficient k in a piecewise homogeneous, isotropic media by using an optimal control approach. Data needed to define cost functionals are obtained by measuring the amount of injected and leaked out water from the sample. In the first theoretical part the problem is formulated and the original cost functionals are expressed in a way which is more convenient for computations. After the discretization of the state equation by a standard finite element method, the resulting algebraic problem leads to a non-convex but smooth minimization problem. Since the problem turned out to be sensitive to the accuracy of measured data, the Tikhonov regularization was used to suppress this occurrence.

The numerical examples demonstrate that with a reasonable number of noiseless measure-

Table 2: Initial and final objective function values as well as the number of iterations and the number of objective function evaluations for Example 2.

	init.cost	final cost	iter/feval
clean	1.76×10^4	2.44×10^{-13}	49/145
noisy	1.79×10^4	0.265	52/130
regul.	1.79×10^4	2.50	32/54

ments, the optimizer supplied with exact gradients easily finds the exact solution. With noisy measurements the situation is, of course, more subtle. As the main contribution of this paper is the clever evaluation of the cost function and its gradient using the standard finite element method, the noise reduction was considered to be left to other studies.

References

[1] R. Blaheta, M. Béreš, S. Domesová, D. Horák, Bayesian inversion for steady flow in fractured porous media with contact on fractures and hydro-mechanical coupling, Computational Geosciences (2020). `https://doi.org/10.1007/s10596-020-09935-8`.

[2] R. Blaheta, M. Béreš, S. Domesová, P. Pan, A comparison of deterministic and bayesian inverse with application in micromechanics, Applications of Mathematics 63 (2018) 665–686.

[3] R. Blaheta, R. Kohut, A. Kolcun, K. Souek, L. Sta, L. Vavro, Digital image based numerical micromechanics of geocomposites with application to chemical grouting, Journal of Rock Mechanics and Mining Sciences 77 (2015) 77–88.

[4] P. Doktor, On the density of smooth functions in certain subspaces of sobolev space, Commentationes Mathematicae Universitatis Carolinae 14 (1973) 609–622.

[5] M.S. Gockenbach, A.A. Khan, An abstract framework for elliptic inverse problems: Part 1. An output least squares approach, Math. Mech. Solids 12 (2007) 259–276.

[6] J. Haslinger, R. Blaheta, R. Hrtus, Identification problems with given material interfaces, Journal of Computational and Applied Mathematics 310 (2017) 129–142.

[7] J. Haslinger, R.A.E. Mäkinen, Introduction to Shape Optimization: Theory, Approximation, and Computation, volume 07 of *Advances in Design and Control*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003.

[8] M. Hinze, R. Pinnau, M. Ulbrich, S. Ulbrich, Optimization with PDE constraints, volume 23 of *Mathematical Modelling: Theory and Applications*, Springer, 2009.

[9] M. Hinze, T.N.T. Quyon, Matrix coefficient identification in an elliptic equation with the convex energy functional method, Inverse problems 32 (2016) 1–29.

[10] MATLAB, Release R2016b with Optimization Toolbox 7.5, The MathWorks Inc., Natick, Massachusetts, 2016.

[11] J.L. Mueller, S. Siltanen, Linear and Nonlinear Inverse Problems with Practical Applications, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2012.

[12] S. Sysala, R. Blaheta, A. Kolcun, J. Ščučka, K. Souček, P. Pan, Computation of composite strength by limit analysis, Key Engineering Materials 810 (2019) 137–142.

Appendix A. Sensitivity analysis

Let

$$J(k) = \sum_{i=1}^{m} J_i(k)$$

be defined by (3.12), where

$$J_i(k) = \tfrac{1}{2}\left(\int_\Omega \big(k\nabla u(k)\cdot\nabla\varphi_i - f\varphi_i\big)\,dx - c_i\right)^2 - \tfrac{1}{2}\underbrace{\Big(\int_\Omega \big(k\nabla u(k)\cdot\nabla q - fq\big)\,dx\Big)}_{=0}^2 \quad \forall q\in V(\Omega).$$

We shall compute the directional derivative $J_i'(k,h)$ of J_i at $k\in\mathcal{U}_{ad}$ and direction $h\in L^\infty(\Omega)$, $h_{|\Omega_i}\in P_0(\Omega_i)$, $i=1,...,q$. From (3.10) it follows

$$\begin{aligned} J_i'(k,h) &= \lim_{t\to 0+}\frac{J_i(k+th)-J_i(k)}{t} \\ &= (\langle\mu_u^i,1\rangle - c_i)\int_\Omega \big(h\nabla u(k)\cdot\nabla\varphi_i + k\nabla u'(k,h)\cdot\nabla\varphi_i\big)\,dx \\ &\quad - (\langle\mu_u^i,1\rangle - c_i)\int_\Omega \big(h\nabla u(k)\cdot\nabla q + k\nabla u'(k,h)\cdot\nabla q\big)\,dx \\ &= (\langle\mu_u^i,1\rangle - c_i)\int_\Omega \big(h\nabla u(k)\cdot\nabla\varphi_i - h\nabla u(k)\cdot\nabla q\big)\,dx \\ &\quad + (\langle\mu_u^i,1\rangle - c_i)\int_\Omega \big(k\nabla u'(k,h)\cdot\nabla\varphi_i - k\nabla u'(k,h)\cdot\nabla q\big)\,dx \end{aligned} \tag{A.1}$$

holds for any $q\in V(\Omega)$.

Let $p_i\in V(\Omega)$ be the adjoint state defined by the adjoint equation

$$\int_\Omega k\nabla p_i\cdot\nabla z\,dx = \int_\Omega k\nabla\varphi_i\cdot\nabla z\,dx \quad \forall z\in V(\Omega). \tag{A.2}$$

In particular, if $z := u'(k,h)\in V(\Omega)$ in (A.2) we see that

$$\int_\Omega \big(k\nabla p_i\cdot\nabla u'(k,h) - k\nabla\varphi_i\cdot\nabla u'(k,h)\big)\,dx = 0.$$

Choosing $q=p_i$ in (A.1) we finally obtain:

$$J_i'(k,h) = (\langle\mu_u^i,1\rangle - c_i)\int_\Omega h\nabla u(k)\cdot(\nabla\varphi_i - \nabla p_i)\,dx, \tag{A.3}$$

where $p_i\in V(\Omega)$ solves (A.2), $i=1,...,m$.