HAL
archives-ouvertes.fr

Primal-Dual Formulations for Parameter Estimation Problems

Guy Chavent, Karl Kunisch, Jean Roberts

▶ **To cite this version:**

Guy Chavent, Karl Kunisch, Jean Roberts. Primal-Dual Formulations for Parameter Estimation Problems. [Research Report] RR-2891, INRIA. 1996. <inria-00073799>

HAL Id: inria-00073799
https://hal.inria.fr/inria-00073799

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Primal-Dual Formulations for Parameter Estimation Problems

Guy Chavent, Karl Kunisch and Jean E. Roberts

N° 2891

May 1996

THÈME 4

ISSN 0249-6399

INRIA
ROCQUENCOURT

Primal-Dual Formulations for Parameter Estimation Problems

Guy Chavent[*] Karl Kunisch [†] and Jean E. Roberts[*]

Thème 4 — Simulation et optimisation
de systèmes complexes
Projets Estime et Ondes

Rapport de recherche no 2891 — May 1996 — 51 pages

Abstract: A new method for formulating and solving parameter estimation problems based on Fenchel duality is presented. The partial differential equation is considered as a contraint in a least squares type formulation and is realized as a penalty term involving the primal and dual energy functionals associated with the differential equation. Splitting algorithms and mixed finite element discretizations are discussed and some numerical examples are given.

Key-words: parameter estimation, Fenchel duality, inverse problems, regularization

(Résumé : tsvp)

[*] {Guy.Chavent} {Jean.Roberts}@inria.fr
[†] Technische Universität Berlin, Fachbereich 3 / Mathematik - Sekr. MA 6-2, Straße des 17. Juni 136, 10623 Berlin, Germany kunisch@math.tu-berlin.de

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : (33 1) 39 63 55 11 – Télécopie : (33 1) 39 63 53 30

Formulations Primales-Duales pour des Problemes d'Estimation de Parametres

Résumé : On présente une méthode nouvelle pour la formulation et la résolution des problèmes d'estimation de paramètres basée sur la dualité de Fenchel. L'équation différentielle est traitée comme une contrainte dans une formulation du type moindre carrés et cette contrainte est mise en œuvre comme un terme de pénalité construit à partir des fonctionelles d'énergies primale et duale associées à l'équation différentielle. Des algorithmes de splitting et une discrétisation via des éléments finis mixtes sont proposés, et des exemples numériques sont présentés.

Mots-clé : estimation des parametres, dualité de Fenchel, problèmes inverse, regularisation

1 Introduction

In this paper, we describe a general framework for a new approach to the formulation of parameter estimation problems. This approach is based on Fenchel duality and is applicable to linear as well as nonlinear problems of monotone type.

Let us first describe some concepts involved in formulating parameter estimation problems in a manner that is suitable for computations. Many of the currently used methods are modifications and/or combinations of the equation error and the least squares formulations. Within the class of least squares formulations, we can distinguish those which treat the partial differential equation as an implicit constraint and others which impose the differential equation as an explicit constraint.

To be more concrete, let $e : Q \times X \to \tilde{X}$, with Q, X, and $\tilde{X}$ Banach spaces, describe the partial differential equation

$$e(a, u) = 0, \tag{1.1}$$

where a denotes the parameter and u the state variable of the differential equation. Let z denote the observation (or data) in the data space Z, and let $B : X \to Z$ be the observation operator. With the *equation error* approach one replaces u in (1.1) by z and solves

$$e(a, z) = 0, \tag{1.2}$$

for a. The least squares approach is based on the optimization problem

$$\text{minimize } \frac{1}{2}|Bu - z|_Z^2 \quad \text{over } Q \tag{1.3}$$

(or over some appropriately defined subset of Q), where u is a solution to (1.1). If (1.1) is treated as an implicit constraint, then $u = u(a)$ is a dependent variable in (1.3). Alternatively, (1.1) can be treated as an explicit constraint, and a as well as u becomes an independent variable in this case. We refer to $[A, BK, C1, C2, KL, IK1, IK2]$ and the references given therein for a more detailed discussion.

Concerning the equation error approach, (1.2), here we only mention that this approach requires that distributed data be available and further that the data be differentiable. However, for linear differential equations with e affine in a and in u, (1.2) has the advantage of being affine with respect to the unknown a.

The output least squares approach on the other hand is versatile with respect to the type of data required. If, for example, only point-wise or boundary data are available one can easily find an appropriate observation B and choose an appropriate norm for Z. But, the simple structure yielded by the equation error approach is lost when a least squares formulation is chosen. Even if e is affine in a and u, problem (1.3) is highly nonlinear if (1.1) is considered as an implicit constraint. If (1.1) is realized as an explicit constraint then (1.3) is quadratic in u with a bilinear constraint.

Because it requires the differentiation of data, a pure equation error technique will not be the method of choice for most applications. The least squares approach with explicit constraints appears at present to be one of the most efficient methods for solving parameter estimation problems numerically. With this approach, for example, the gradient of the fit-to-data function, which is required in every iterative technique for solving (1.3) numerically, may be calculated in a straightforward manner. Moreover, since a and u are independent variables, an equation error method can be used to generate an initial guess for the parameter a, thus combining the advantages of the equation error and the least squares formulations.

It seems appropriate to mention here as well the adaptive control technique, also referred to as the asymptotic embedding method [AHS, BS]. The idea of this method is to introduce a dynamical system having the fit-to-data-term $Bu - z$ as inhomogeneity and having the solution to $e(u, a) = 0$ as stationary solution. The adaptive control technique can also be interpreted as a continuous version of a gradient method for solving the least squares problem (1.3).

The new framework that we propose in this paper belongs to the class of least squares formulations with explicit treatment of the differential equa-

tion as constraint.

In §2 we present the method and give several applications to second order differential equations of elliptic and parabolic type. The subsequent sections are devoted to the analysis of a particular formulation for a problem proposed in §2. Basic properties are developed in §3, splitting algorithms are derived in §4 and a mixed finite element implementation is described in §5. Numerical results are given in §6.

Acknowledgement The authors would like to express their gratitude to Guillaume Vigo who very efficiently and graciously carried out the numerical experiments reported in §6.

2 Primal-dual formulation based on Fenchel duality

We consider the case of a system whose equilibrium state $u \in X$ is obtained by minimization, over the Hilbert space X of states, of an *energy functional* $E_a(u)$:

$$\min E_a(u) \qquad \text{over } u \in X. \tag{$\mathcal{E}$}$$

We suppose that this energy functional depends on an unknown parameter a in a set C of admissible parameters in a Banach space Q:

$$a \in C \subset Q.$$

In order to estimate this unknown parameter, we suppose that we have at our disposal a measurement z of the observation $B(u)$. Here $B \in \mathcal{L}(X, Z)$ is the observation operator, and Z is the observation space which is assumed to be a Hilbert space. The classical least squares formulation for parameter identification problems is given by

$$\min \frac{1}{2}|z - Bu|_Z^2 \qquad \text{over } a \in C, \tag{$\mathcal{P}$}$$

where $u = u(a)$ is a solution to $(\mathcal{E})$ with $a \in C$.

Hence $(\mathcal{E})$ appears in $(\mathcal{P})$ as a constraint between $a \in C$ and $u \in X$. In particular, each evaluation of the objective function in $(\mathcal{P})$ requires a full solution of $(\mathcal{E})$. However, solving $(\mathcal{E})$ precisely in the first steps of an iterative technique for $(\mathcal{P})$, when the parameter a is still far from its converged value, may be inefficient. Moreover, in situations where the observation operator B is rich enough, one can build up, by interpolation or smoothing of the available data z, a (possibly rough) estimate $\tilde{u}$ of the state variable u, but the fact that $u = u(a)$ is a hard constraint in $(\mathcal{P})$ makes it impossible to use this estimate. One has to chose an initial value a_0 for the parameter a, which determines the first approximation $u(a_0)$ to the state variable which may be quite far from z. The idea presented here is to relax, in $(\mathcal{P})$, the constraint that a and u satisfy the state equation by imposing it through penalization.

A first realization of this idea would consist in taking advantage of the fact that the state equation is defined via the minimization problem $(\mathcal{E})$, and to replace $(\mathcal{P})$, for $\varepsilon > 0$, by

$$\min \left\{ \frac{1}{2}|z - Bu|_Z^2 \; + \; \frac{1}{\varepsilon} E_a(u) \right\} \qquad \text{over } (a,u) \in C \times X. \qquad (\tilde{\mathcal{P}}_\varepsilon)$$

We would expect this problem to be a perturbation of (P) if the set of minimizers of the penalization function $E_a(u)$ were made up of all couples (a,u) which satisfy the constraint $a \in C, u = u(a)$. But the energy

$$\min_{u \in C} \; E_a(u) = E_a\left(u(a)\right)$$

of the state $u(a)$ associated to a given parameter a depends, in general, on this parameter ! Since under appropriate hypotheses

$$\min_{(a,u) \in C \times X} \; E_a(u) = \min_{a \in C} \; E_a\left(u(a)\right) = E_{\min}$$

the set of minimizers of $E_a(u)$ over $C \times X$ is made up of only those couples (a,u) for which the constraint $a \in C, u = u(a)$ is satisfied and which produce

states with minimum energy $E_{\min}$! So $E_a(u)$ is not a penalization function for the constraint in (P), and $(\tilde{P}_\varepsilon)$ is not an appropriate perturbation of (P).

The above considerations suggest, however, that the penalization approach would work if we could replace, as characterization of the equilibrium state, $E_a(u)$ by a different energy functional whose minimum with respect to u is independent of the parameter a. Fenchel duality provides us with a systematic way of doing this - at the price of an enlargement of the state space - provided that the energy functional $E_a(u)$ can be written as the sum of two convex functionals. So we shall suppose from now on that

$$E_a(u) = F_a(u) + G_a(Au) \qquad \text{for every } u \in X, \tag{2.1}$$

where for each $a \in C$, $F_a : X \to \overline{\mathrm{R}}$ and $G_a : Y \to \overline{\mathrm{R}}$ are proper, convex and lower semi-continuous with $\overline{R} = \mathrm{R} \cup \{+\infty\}$, Y is a Hilbert space, $A \in \mathcal{L}(X;Y)$. We suppose as well that

$$\begin{array}{l} \text{for each } a \in C, \text{ there exists } u_o = u_o(a) \text{ such that} \\ F_a(u_o) < +\infty,\ G_a(Au_o) < +\infty \text{ and } G_a \text{ is continuous at } Au_o. \end{array} \tag{2.2}$$

The Fenchel duality theorem $[BP, ET]$ asserts that, with these hypotheses, for every $a \in C$

$$\inf_{u \in X} \{F_a(u) + G_a(Au)\} + \min_{q \in Y} \{F_a^*(-A^*q) + G_a^*(q)\} = 0, \tag{FD}$$

where

$$A^* \in \mathcal{L}(Y;X)$$

is the adjoint of A,

$$F_a^* : X \to \overline{R}, \quad \text{and} \quad G_a^* : Y \to \overline{R}$$

are the convex conjugates of F_a and G_a respectively, defined by

$$F_a^*(u) = \sup_{v \in X} \{< u, v >_X - F_a(v)\}, \quad \text{for all } u \in X,$$

and

$$G_a^*(p) = \sup_{q \in Y} \{< p, q >_Y - G_a(q)\}, \quad \text{for all } p \in Y.$$

If in addition to (2.1) and (2.2) we suppose that

$$\text{for every } a \in C, \qquad \lim_{|u| \to \infty} \{F_a(u) + G_a(Au)\} = +\infty, \tag{2.3}$$

then the inf in (FD) becomes a min (which means that the original problem (E) has at least one solution), and any pair of minimizers (u, q), $u = u(a), q = q(a)$, of the left hand side of (FD) satisfies the extremality conditions

$$\begin{aligned} -A^*q &\in \partial F_a(u) \\ q &\in \partial G_a(Au) \end{aligned} \tag{EC}$$

where ∂ denotes the sub-differential operator for convex functions.

We refer to $u \in X$ as the primal state variable, and to $q \in Y$ as the dual state variable. In view of (FD) it is natural to associate to the dual variable q a dual energy function $E_a^*(q)$ defined by

$$E_a^*(q) = F_a^*(-A^*q) + G_a^*(q) \quad \text{for every } q \in Y \tag{2.4}$$

and to consider the dual problem

$$\min \; E_a^*(q) \qquad \text{over } q \in Y. \tag{$\mathcal{E}^*$}$$

To any couple (u, q) of primal/dual state variables, one associates the total energy $E_a(u) + E_a^*(q)$, and one considers the corresponding minimization problem

$$\min \{E_a(u) + E_a^*(q)\} \qquad \text{over } (u, q) \in X \times Y \tag{$\mathcal{E}\mathcal{E}^*$}$$

which, due to the Fenchel duality formula (FD), unlike (E), has the property that

$$\text{for each } a \in C, \quad \min_{(u,q)\in X\times Y} \{E_a(u) + E_a^*(q)\} = 0. \tag{2.5}$$

We summarize the above discussion in the following proposition.

Proposition 2.1 *Let (2.1) (2.2) and (2.3) hold. Then far all $a \in C$ one has*

i) the primal, dual and total energy minimization problems $(\mathcal{E})$, $(\mathcal{E}^*)$, and $(\mathcal{E}\mathcal{E}^*)$ admit solutions

ii) solving $(\mathcal{E}\mathcal{E}^*)$ is equivalent to solving $(\mathcal{E})$ and $(\mathcal{E}^*)$

iii) any pair of solutions u and q to $(\mathcal{E})$ and $(\mathcal{E}^*)$ satisfies $(\mathcal{E}\mathcal{E}^*)$

iv) the minimum of the total energy is zero.

The problem of the estimation of $a \in C$ from the measurement $z \in Z$ of Bu, can therefore be approximated by penalizing $(\mathcal{P})$ by the total energy, which has a minimum (equal to zero) if and only if (a, u, q) satisfies the constraint $a \in C, u = u(a)$ and $q = q(a)$. This leads to the sought *primal-dual formulation of the parameter estimation problem*:

$$\min \left\{ \frac{1}{2}|z - Bu|_Z^2 \; + \; \frac{1}{\varepsilon}\left(E_a(u) + E_a^*(q)\right) \right\} \quad \text{over } (a, u, q) \in C \times X \times Y, \tag{$\mathcal{P}_\varepsilon$}$$

where $\varepsilon > 0$ is the penalization parameter. Note that when the data $z \in Z$ are *attainable* - i.e. when there exists $a \in C$ such that $z = Bu(a)$ - the problem $(\mathcal{P}_\varepsilon)$ is an *exact penalty formulation*. That is to say that for any $\varepsilon > 0$, the solutions to $(\mathcal{P}_\varepsilon)$ and $(\mathcal{P})$ coincide. Moreover, the minimum value of the cost functional in $(\mathcal{P}_\varepsilon)$ is zero in this case.

We give now some examples of parameter estimation problems for which the primal-dual formulation can be applied. It is assumed throughout that Ω is a bounded domain in R^n with sufficiently smooth boundary $\partial\Omega$, and that $B \in \mathcal{L}(X; Z)$.

Example 2.1 : Estimation of the diffusion coefficient a in

$$\begin{aligned} -\nabla(a\nabla u) &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega \end{aligned} \tag{2.6}$$

where $f \in L^2(\Omega)$ is given, and where the set C of admissible parameters satisfies :

$$C \subset \{a \in L^\infty(\Omega) \mid 0 < \nu \leq a(x)\}\,. \tag{2.7}$$

Equation (2.6) corresponds to the minimization of

$$E_a(u) = \frac{1}{2}\int_\Omega a \mid \nabla u \mid^2 - \int_\Omega fu \tag{2.8}$$

over $X = H_0^1(\Omega)$. A first way of casting $E_a(u)$ in the form (2.1) consists in choosing :

$$\begin{aligned} &Y = L_n^2(\Omega), && Au = -\nabla u \text{ for all } u \in X, \\ &F_a(u) = -\int_\Omega fu, && G_a(q) = \frac{1}{2}\int_\Omega a \mid q \mid^2_{R^n}, \end{aligned} \tag{2.9}$$

which clearly satisfies the hypotheses (2.2) and (2.3). Then

$$\begin{aligned} &A^*q = (-\Delta)^{-1}\nabla q \qquad \text{for } q \in Y, \\ &F_a^*(u) = \begin{cases} 0 & \text{if } -\Delta u + f = 0 \\ +\infty & \text{if } -\Delta u + f \neq 0 \end{cases}, \qquad G_a^*(q) = \frac{1}{2}\int_\Omega \frac{1}{a} \mid q \mid^2_{R^n} \end{aligned} \tag{2.10}$$

so that the dual energy $E_a^*(q)$, (2.4), is

$$E_a^*(q) = \begin{cases} \dfrac{1}{2}\displaystyle\int_\Omega \frac{1}{a}\,|q|^2_{R^n} & \text{if} \quad \nabla q = f, \\ +\infty & \text{if} \quad \nabla q \neq f, \end{cases} \tag{2.11}$$

and the *total energy* is

$$E_a(u) + E_a^*(q) = \begin{cases} \frac{1}{2}\int_\Omega a \mid \nabla u \mid^2 - \int_\Omega fu + \frac{1}{2}\int_\Omega \frac{1}{a} \mid q \mid^2 & \text{if} \quad \nabla q = f \\ +\infty & \text{if} \quad \nabla q \neq f \end{cases} \tag{2.12}$$

which, by the Fenchel duality formula (FD), is necessarily nonnegative, and is equal to zero if and only if u is the minimizer $u(a)$ of $E_a(u)$ (*i.e.* the solution of the elliptic equation (2.6) and q is the minimizer $q(a)$ of $E_a^*(q)$). The penalization of $(\mathcal{P})$ by the total energy (2.12) produces the sought *primal-dual formulation* $(\mathcal{P}_\varepsilon)$ for the estimation of $a \in C$ from a measurement z of Bu :

$$\begin{aligned} &\min \left\{ \frac{1}{2}|z - Bu|_Z^2 \; + \; \frac{1}{\varepsilon}\left[\frac{1}{2}\int_\Omega a \mid \nabla u \mid^2 - \int_\Omega fu + \frac{1}{2}\int_\Omega \frac{1}{a} \mid q \mid^2\right]\right\} \\ &\text{over } (a,u,q) \in C \times H_0^1 \times L_n^2, \nabla q = f. \end{aligned} \tag{2.13}$$

As expected, this formulation allows us to take advantage of the knowledge of an a-priori guess $\tilde{u}$ of the state (for example if the observation z is rich enough) by starting the inversion process by minimizing, in (2.13) with respect to a and q for $u = \tilde{u}$ fixed. The space Q for a will be specified in Section 3

For Example 2.1, the properties of the total energy summarized in Proposition 2.1 can be obtained directly by using (2.12) as a definition, and by noticing that

$$\text{if } \nabla q = f, \quad \text{then}$$

$$\frac{1}{2}\int_\Omega a \mid \nabla u \mid^2 - \int_\Omega fu + \frac{1}{2}\int_\Omega \frac{1}{a} \mid q \mid^2 = \frac{1}{2}\int_\Omega \frac{1}{a} \mid q + a\nabla u \mid^2 \geq 0. \tag{2.14}$$

Remark 2.1 The above penalization of equation (2.6) is indeed valid when the right hand side f is only in $H^{-1}(\Omega)$. We shall use this fact in Example 2.7 for the parabolic case.

Example 2.2 : We revisit Example 2.1 with the same set C of admissible parameters satisfying (2.7), the same state space $X = H_0^1(\Omega)$ and the same energy $E_a(u)$ given by (2.8). But here we use a different choice for casting $E_a(u)$ in the form (2.1). We choose

$$Y = H_0^1(\Omega), A = I, F_a = 0, G_a(u) = \frac{1}{2}\int_\Omega a \mid \nabla u \mid^2 dx - \int_\Omega f u \, dx.$$

The convex conjugates of F_a and G_a are found to be

$$F_a^*(q) = \begin{cases} \infty \text{ if } q \neq 0 \\ 0 \quad \text{if } q = 0 \end{cases}$$

and

$$G_a^*(q) = \frac{1}{2} < A_a^{-1}(f+q), f+q >_{H_0^1, H^{-1}}$$

where $A_a : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by

$$A_a(u) = -\nabla \cdot (a \nabla u).$$

In this case problem $(\mathcal{P}_\varepsilon)$ becomes

$$\min \left\{ \frac{1}{2} \mid z - Bu \mid^2 + \frac{1}{\varepsilon} \left[\frac{1}{2}\int_\Omega a \mid \nabla u \mid^2 dx - \int_\Omega f u \, dx + \frac{1}{2}\int_\Omega A_a^{-1}(f) \cdot f \, dx \right] \right\}$$
$$\text{over } (a,u) \in C \times H_0^1. \tag{2.15}$$

Clearly (2.2) and (2.3) are satisfied. Once again, if an a-priori guess $\tilde{u}$ for u is available, one can make use of it by optimizing first, in (2.15) with respect to a for $u = \tilde{u}$ fixed, which produces a first guess for the parameter which is coherent with the data.

Example 2.3 : Estimation of the nonlinear diffusion in

$$\begin{aligned} &-\text{div}\,(\partial\sigma(\nabla u)) \ni f \text{ in } \Omega \\ &u_{|\partial\Omega} = 0. \end{aligned} \tag{2.16}$$

Here the spaces X and Y and the operator A are chosen as in Example 2.1 above, i.e.

$$X = H_0^1(\Omega), \quad Y = L_n^2(\Omega), \quad Au = -\nabla u.$$

The parameter σ is chosen from the set

$$\begin{aligned} C = \; & \{\sigma : \mathrm{R}^n \to \mathrm{R} \; : \; \sigma \text{ is convex, is continuous and satisfies} \\ & \beta_1 \mid r \mid^2_{\mathrm{R}^n} + \delta_1 \leq \sigma(r) \leq \beta_2 \mid r \mid^2_{\mathrm{R}^n} + \delta_2 \text{ for all } r \in \mathrm{R}^n\} \end{aligned}$$

for $0 < \beta_1 < \beta_2 < \infty$ and δ_1 and $\delta_2 \in \mathrm{R}$. The parameterized mappings are defined by

$$F(u) = -\int_\Omega f u \, dx \text{ and } G_\sigma(q) = \frac{1}{2}\int_\Omega \sigma(q) dx$$

with the convex conjugate F^* given in Example 2.1 and

$$G_\sigma^*(q) = \frac{1}{2}\int_\Omega \sigma^*(q) dx.$$

The resulting problems $(\mathcal{P}_\varepsilon)$ have the form

$$\begin{aligned} & \min \left\{ \frac{1}{2} \mid z - Bu \mid^2 + \frac{1}{\varepsilon} \left[\frac{1}{2}\int_\Omega \sigma(\nabla u)^2 \, dx - \int_\Omega f u \, dx + \frac{1}{2}\int_\Omega \sigma^*(q) dx \right] \right\} \\ & \text{over } (\sigma, u, q) \in C \times H_o^1 \times L_n^2, \; div \; q = f. \end{aligned} \tag{2.17}$$

An analysis of this formulation of the identification of nonlinear diffusion operators is given in [BaK]. The motivation for the choice of the cost functional in [BaK] is based on the convex conjugacy formula. We note that (2.13) is not a special case of (2.17).

Example 2.4 : Estimation of the potential c in

$$\begin{aligned} & -\Delta u + cu = f \text{ in } \Omega \\ & u \mid \partial\Omega = 0. \end{aligned} \tag{2.18}$$

One can proceed in a manner analogous to Example 2.2, with

$$C = \{c \in L^p(\Omega) : c(x) \geq 0\}, \text{ with } p \geq \min(\frac{n}{2}, 2)$$
$$X = Y = H_0^1(\Omega), A = I, F = 0, \text{ and}$$
$$G_a(u) = \frac{1}{2}\int_\Omega \mid \nabla u \mid^2 dx + \frac{1}{2}\int cu^2\, dx - \int_\Omega fu\, dx.$$

We find

$$G_a^*(q) =< A_c^{-1}(f+q), f+q >_{H_0^1, H^{-1}}$$

where $A_c : H_0^1(\Omega) \to H^{-1}$ is defined by

$$A_c(u) = -\Delta u + cu.$$

The penalized problems (P_ε) are

$$\min \left\{ \frac{1}{2} \mid z - Bu \mid^2 \frac{1}{\varepsilon} \left[\frac{1}{2}\int_\Omega \mid \nabla u \mid^2 dx + \frac{1}{2}\int_\Omega cu^2\, dx - \int_\Omega fu\, dx + \frac{1}{2}\int_\Omega (A_c^{-1}f)f\, dx \right] \right\} \tag{2.19}$$

over $(c, u) \in C \times H_0^1$.

Example 2.5 : We turn to nonlinear potential problems and consider

$$\begin{aligned} -\Delta u + \partial\sigma(u) &\ni f \text{ in } \Omega \\ u_{|\partial u} &= 0 \end{aligned} \tag{2.20}$$

associated with the energy

$$E_\sigma(u) = \frac{1}{2}\int_\Omega \mid \nabla u \mid^2 - \int_\Omega fu + \int_\Omega \sigma(u)$$

defined on $X = H_0^1(\Omega)$. For any parameter $\sigma \in C$ with C as in Example 2.3 but with $n = 1$, we cast $E_\sigma(u)$ in the form (2.1) by choosing

$$Y = L^2(\Omega), A = H_0^1 \to L^2 \text{ embedding}, \quad f \in L^2,$$

$$F(u) = \frac{1}{2} \int_\Omega \mid \nabla u \mid^2 - \int_\Omega fu \qquad \text{for } u \in H_0^1(\Omega)$$

$$G_\sigma(u) = \int_\Omega \sigma(u) \qquad \text{for } u \in L^2(\Omega).$$

For the convex conjugate functions we find

$$F^*(u^*) = \frac{1}{2} \mid u^* + f \mid^2_{H^{-1}} \qquad \text{for } \; u^* \in H^{-1}(\Omega)$$

and

$$G^*_\sigma(u) = \int_\Omega \frac{1}{2}\sigma^*(u) \qquad \text{for } \; u \in L^2(\Omega).$$

Hence the penalized problems $(\mathcal{P}_\varepsilon)$ are

$$\min \left\{ \frac{1}{2} \mid z - Bu \mid^2 + \frac{1}{\varepsilon} \left[\frac{1}{2} \int_\Omega \mid \nabla u \mid^2 - \int_\Omega fu + \int_\Omega \sigma(u) \right.\right.$$
$$\left.\left. + \int_\Omega \sigma^*(q) + \frac{1}{2} \mid -q + f \mid^2_{H^{-1}} \right] \right\} \qquad (2.21)$$

over $(\sigma, u, q) \in C \times H_0^1 \times L^2$.

Remark 2.2 For the above example, the extremality conditions (EC) have the form

$$q = \Delta u + f \qquad (2.22)$$

$$q \in \partial\sigma(u) \qquad (2.23)$$

Hence if we choose $u \in H_0^1(\Omega)$ with $\Delta u \in L^2(\Omega)$ and if we use the value of $q \in L^2(\Omega)$ given by (2.22) in the expression of the total energy, we find that, for $\sigma \in C$

$$\inf_{u \in H_0^1, q \in L^2} \{E_\sigma(u) + E_\sigma^*(q)\}$$
$$= \inf_{u \in H_0^1, \Delta u \in L^2} \left\{ \int_\Omega \mid \nabla u \mid^2 - \int_\Omega fu + \int_\Omega \sigma(u) + \int_\Omega \sigma^*(\Delta u + f) \right\}. \quad (2.24)$$

This shows that the functional on the right hand side can be used as a penalization function for equation (2.17), which gives rise to the following penalized least squares formulation :

$$\min \left\{ \frac{1}{2} \mid z - Bu \mid_Z^2 + \frac{1}{\varepsilon} \int_\Omega \left[\mid \nabla u \mid^2 - fu + \sigma(u) + \sigma^*(\Delta u + f) \right] \right\}$$
$$\text{over } (\sigma, u) \in C \times H_0^1, \Delta u \in L^2 \quad (2.25)$$

which was analyzed in [BaK].

Example 2.6 : We revisit the nonlinear potential problem of Example 2.5 but with a different choice of duality: the dual variable q, instead of being linked to u, at the optimum, by $q = \Delta u + f$ as in the previous Example 2.5, will satisfy $q = -\nabla u$, in a way similar to the mixed finite element formulation of Example 2.1. This will be achieved by keeping the same $X = H_0^1(\Omega)$ and C as in Example 2.5, and by choosing

$$Y = L_n^2(\Omega), \qquad Au = -\nabla u,$$
$$F_\sigma(u) = \int_\Omega (\sigma(u) - fu), \qquad G_\sigma(q) = \tfrac{1}{2} \textstyle\int_\Omega \mid q \mid^2 .$$

The convex conjugates are found to be

$$F^*(u) = \begin{cases} \displaystyle\int_\Omega \sigma^*(u + f) \text{ if } u \in L^2(\Omega) \\ \infty \qquad\qquad \text{if } u \in H^{-1}(\Omega) \text{ but } u \text{ is not in } L^2(\Omega) \end{cases}$$
$$G^* = G.$$

The corresponding penalized least squares problems $(\mathcal{P}_\varepsilon)$ are

$$\min \left\{ \frac{1}{2} \mid z - Bu \mid^2 + \frac{1}{\varepsilon} \left[\int_\Omega (\sigma(u) - fu) + \frac{1}{2} \int_\Omega \mid \nabla u \mid^2 \right.\right. \\ \left.\left. + \int_\Omega \sigma^*(-\text{div } q + f) + \frac{1}{2} \int_\Omega \mid q \mid^2_{R^n} \right] \right\} \tag{2.26}$$

over $(\sigma, u, q) \in C \times H_0^1(\Omega) \times H(\text{div}\,, \Omega)$.

One checks easily that, according to the second extremality condition in (EC), the relationship between primal and dual variables at the minimum is given by $q = -\nabla u$.

Example 2.7 : Here we discuss the applicability of the primal-dual formulation to a parabolic estimation problem. We treat the estimation of a in

$$\begin{aligned} &\frac{\partial u}{\partial t} - \nabla(a \nabla u) = f \text{ in } \Omega \times (0,T) \\ &u = 0 \qquad\qquad\qquad \text{on } \partial\Omega \times (0,T) \\ &u = u_0 \qquad\qquad\qquad \text{on } \Omega \text{ for } t = 0, \end{aligned} \tag{2.27}$$

where $u_0 \in L^2(\Omega)$ and $f \in L^2(0,T; H^{-1}(\Omega))$, from data z. As in Example 2.1 we choose

$$C \subset \{a \in L^\infty(\Omega) : a(x) \geq \nu > 0 \, a.e. \text{ in } \Omega\} .$$

For every $a \in C$ equation (2.27) has a unique solution u in $W(0,T)$, where

$$W(0,T) = \left\{ u \in L^2\left(0,T; H_0^1(\Omega)\right) : \frac{du}{dt} \in L^2\left(0,T; H^{-1}(\Omega)\right) \right\} .$$

In particular this implies that

$$q = -a\nabla u \in L^2\left(0,T; L^2(\Omega)\right) = L_n^2(Q) \tag{2.28}$$

and

$$f(t) - \frac{du}{dt}(t) \in H^{-1}(\Omega) \text{ for almost every } t \in (0,T).$$

Of course, (2.27) does not correspond to the minimization of an energy functional, and we cannot apply directly the duality results. So we proceed by treating (2.27) as a family of "elliptic equations" with right-hand side given by $f(t) - \frac{du}{dt}(t)$ and use the total energy functional of Example 2.1.

As was seen in Example 2.1, for any $(v,s) \in H_0^1(\Omega) \times L_n^2(\Omega)$ and any $\tilde{f} \in H^{-1}(\Omega)$

$$\int_\Omega \left(\frac{a}{2} \mid \nabla v \mid^2 - \tilde{f} v + \frac{1}{2a} \mid s \mid^2 \right) \geq 0 \tag{2.29}$$

provided that $\nabla s = \tilde{f}$. Moreover equality holds in (2.29) if and only if $-a\nabla v = s$. Taking $\tilde{f} = f - \frac{du}{dt}$ in (2.29) and integrating with respect to t, we obtain that for every $(v,s) \in W(0,T) \times L_n^2(Q)$

$$\mathcal{B}_a(v,s) = \int_0^T \int_\Omega \left(\frac{a}{2} \mid \nabla v(t) \mid^2 - \left(f - \frac{dv}{dt}(t) \right) v(t) + \frac{1}{2a} \mid s(t) \mid^2 \right) \geq 0 \tag{2.30}$$

provided that $\nabla s = f - \frac{dv}{dt}$ in $L^2(0,T; H^{-1}(\Omega))$. As in (2.29), equality holds in (2.30) if $-a\nabla v = s$. Thus the pair (u,q), with u the solution to (2.27) and q defined in (2.28), is a solution to

$$\min \mathcal{B}_a(v,s) \quad \text{over} \quad (v,s) \in W(0,T) \times L_n^2(Q),\ v(0) = u_0,\ \frac{dv}{dt} + \nabla s = f. \tag{2.31}$$

It is also simple to see that (u,q) is the unique solution to (2.31). In fact, if v and s are coupled by $\frac{dv}{dt} + \nabla s = f$, then

$$\mathcal{B}_a(v,s) = \frac{1}{2}\int_0^T \int_\Omega \frac{1}{a} \mid a\nabla v + s \mid^2$$

and uniqueness in the affine variety

$$X = \left\{ (u,s) \in W(0,T) \times L^2_n(Q) : v(0) = u_0,\ \frac{dv}{dt} + \nabla s = f \right\}$$

follows. Let us also note that on the tangent space

$$\delta X = \left\{ (\delta v, \delta s) \in W(0,T) \times L^2_n(Q) : \delta v(0) = 0,\ \frac{d\delta v}{dt} + \nabla \delta s = 0 \right\}$$

to the affine variety X, the second Frechet derivative of $\mathcal{B}_a$ with respect to (v,s) in the direction $(\delta v, \delta s)^2$ is

$$\begin{aligned} \mathcal{B}_a''(\delta v, \delta s)^2 &= \int_0^T \int_\Omega a \mid \nabla \delta v \mid^2 + \int_\Omega \mid \delta v(T) \mid^2 + \int_0^T \int_\Omega \frac{1}{a} \mid \delta s \mid^2 \\ &\geq \nu \mid \delta v \mid^2_{L^2(0,T;H^1_o)} + \mid \delta v(T) \mid^2_{L^2(\Omega)} + \frac{1}{\mu} \mid \delta s \mid^2_{L^2_n(Q)}, \end{aligned} \tag{2.32}$$

where it is assumed that $a(x) \leq \mu$ a.e. in Ω. Due to the fact that there is no energy associated to the evolution in time for the parabolic equation (2.27), $\mathcal{B}_a''(\delta v, \delta s)^2$ cannot be bounded from below by $\mid \frac{d\,\delta v}{dt} \mid^2_{L^2(0,T;H^{-1})}$. Hence $\mathcal{B}_a$ is not uniformly convex with respect to the natural norm induced by $W(0,T) \times L^2_n(Q)$ on X.

Returning to (2.30) we note that it provides the desired property that the minimal value of the energy functional is zero independently of $a \in C$. Thus $\mathcal{B}_a(u,q) = 0$ (with (u,q) satisfying the constraints of (2.31)) characterizes the set of pairs $(a,u) \in C \times W(0,T)$ which satisfy (2.27). This suggests the following formulation for the parameter estimation problem :

$$\begin{aligned} &\min \left\{ \frac{1}{2} \mid z - Bv \mid^2 + \frac{1}{\varepsilon}\mathcal{B}_a(v,s) \right\} \\ &\text{over } (v,s) \in W(0,T) \times L^2_n(Q), v(0) = u_0, \frac{dv}{dt} + \nabla s = f. \end{aligned}$$

3 Convergence as $\varepsilon \longrightarrow 0$

We return now to the estimation of the diffusion coefficient in an elliptic equation in the setting of Example 2.1, and consider its regularized version:

$$\begin{aligned} &\min \left\{ \tfrac{1}{2}|z - Bu|_Z^2 + \tfrac{\beta}{2}|a - a^\sharp|_Q^2 \right\} \\ &\text{over } (a,u) \in C \times H_0^1 \text{ which satisfy equation (2.6),} \end{aligned} \tag{$\mathcal{P}$}$$

where $a^\sharp$ is an a-priori guess for the true parameter, and β is a regularization parameter assumed to satisfy $\beta > 0$. This ensures the existence of a solution to $(\mathcal{P})$.

As was seen in §2, the primal-dual formulation $(\mathcal{P}_\varepsilon)$ of $(\mathcal{P})$ is, cf (2.13),

$$\begin{aligned} &\min \left\{ \frac{1}{2}|z - Bu|_Z^2 + \frac{\beta}{2}|a - a^\sharp|_Q^2 + \frac{1}{\varepsilon}[\frac{1}{2}\int_\Omega a|\nabla u|^2 - \int_\Omega fu + \frac{1}{2}\int_\Omega \frac{1}{a}|q|^2] \right\} \\ &\text{over } (a,u,q) \in C \times H_0^1 \times H_{div}, \nabla q = f \in L^2, \end{aligned} \tag{$\mathcal{P}_\varepsilon$}$$

where

$$H_{div} = \{q \in L_n^2(\Omega) : \text{ div } q \in L^2(\Omega)\}$$

and

$$C = \{a \in Q : 0 < \nu \leq a \leq \mu \text{ a.e. on } \Omega\}, \tag{3.1}$$

with ν and μ known lower and upper bounds for a and Q a Hilbert space that embeds compactly in $L^\infty(\Omega)$.

It will be convenient to introduce the following notation:

$$\mathcal{X} = Q \times H_0^1(\Omega) \times H_{div},$$

$J_1 : H_0^1(\Omega) \times H_{div} \longrightarrow R$, is the least squares functional

$$J_1(a,u) = \frac{1}{2}|z - Bu|_Z^2 + \frac{\beta}{2}|a - a^\sharp|_Q^2, \tag{3.2}$$

and $J_2 : Q \times H_0^1(\Omega) \times H_{div} \longrightarrow R$, is the total energy functional

$$J_2(a,u,q) = \frac{1}{2}\int_\Omega a|\nabla u|^2 - \int_\Omega fu + \frac{1}{2}\int_\Omega \frac{1}{a}|q|^2. \tag{3.3}$$

Of course J_1 is non-negative, and, as we saw in §2, J_2 is non-negative whenever (a,u,q) satisfies $\nabla q = f$, and vanishes only when (a,u,q) also satisfies $q + a\nabla u = 0$, i.e. when the elliptic equation (2.6) is satisfied. The functional in $(\mathcal{P}_\varepsilon)$ may now be written

$$J_\varepsilon(a,u,q) = J_1(a,u) + \frac{1}{\varepsilon}J_2(a,u,q). \tag{3.4}$$

Proposition 3.1 *For every $\varepsilon > 0$ there exists a solution $(a_\varepsilon, u_\varepsilon, q_\varepsilon) \in \mathcal{X}$ to $(\mathcal{P}_\varepsilon)$.*

Proof : Let $\{(a_n,u_n,q_n)\}_{n\in N} \subset \mathcal{X}$ be a minimizing sequence such that for each $n \in N$

$$\begin{aligned} &\text{div } q_n = f \\ &\alpha \le J_1(a_n,u_n) + \tfrac{1}{\varepsilon}J_2(a_n,u_n,q_n) \le \alpha + \tfrac{1}{n}, \end{aligned} \tag{3.5}$$

where α denotes the infimum of the cost in $(\mathcal{P}_\varepsilon)$. From (3.5) it is simple to argue that

$$\{(a_n,u_n,q_n)\}_{n\in N} \text{ is bounded in } \mathcal{X}.$$

Hence, there exists a subsequence still denoted $\{(a_n,u_n,q_n)\}_{n\in N}$, and $(a_\varepsilon,u_\varepsilon,q_\varepsilon) \in \mathcal{X}$ such that

$$(a_n,u_n,q_n) \rightharpoonup (a_\varepsilon,u_\varepsilon,q_\varepsilon) \text{ weakly in} \mathcal{X}.$$

In particular this implies that

$$(a_n,u_n) \to (a_\varepsilon,u_\varepsilon) \text{ strongly in } L^\infty \times L^2,$$

and moreover div $q_\varepsilon = f$. It also follows that

$$\nabla u_n \rightharpoonup \nabla u_\varepsilon \text{ weakly in } L^2_n,$$

and hence

$$\sqrt{a_n}\nabla u_n \rightharpoonup \sqrt{a_\varepsilon}\nabla u_\varepsilon \text{ weakly in } L^2_n,$$

as well. Similarly

$$\frac{1}{\sqrt{a_n}} q_n \rightharpoonup \frac{1}{\sqrt{a_\varepsilon}} q_\varepsilon \text{ weakly in } L^2_n.$$

We therefore find that

$$\int_\Omega a_\varepsilon |\nabla u_\varepsilon|^2 \le \underline{\lim} \int_\Omega a_n |\nabla u_n|^2 \tag{3.6}$$

and

$$\int_\Omega \frac{1}{\sqrt{a_\varepsilon}} |q_\varepsilon|^2 \le \underline{\lim} \int_\Omega \frac{1}{\sqrt{a_n}} |q_n|^2. \tag{3.7}$$

Using (3.6) and (3.7) in (3.5) we obtain

$$\begin{aligned} J_1(a_\varepsilon, u_\varepsilon) + \frac{1}{\varepsilon} J_2(a_\varepsilon, u_\varepsilon, q_\varepsilon) & \\ \le \underline{\lim} \frac{1}{2} |z - Bu_n|^2_Z + \underline{\lim} \frac{\beta}{2} |a - a^\sharp|^2_Q + \underline{\lim} \frac{1}{\varepsilon} \int_\Omega \frac{a_n}{2} |\nabla u_n|^2 & \\ - \lim \frac{1}{\varepsilon} \int_\Omega f u_n + \underline{\lim} \frac{1}{\varepsilon} \int_\Omega \frac{1}{2a_n} |q_n|^2 & \\ \le \underline{\lim} (J_1(a_n, u_n) + \frac{1}{\varepsilon} J_2(a_n, u_n, q_n)) \quad \le \quad \alpha. & \end{aligned}$$

This implies that $(a_\varepsilon, u_\varepsilon, g_\varepsilon)$ is a solution to $\mathcal{P}_\varepsilon$.

■

We next turn to the convergence of $(a_\varepsilon, u_\varepsilon, q_\varepsilon)$ to a solution of $\mathcal{P}$ as $\varepsilon \to 0^+$.

Proposition 3.2 *For every $\varepsilon > 0$ let $(a_\varepsilon, u_\varepsilon, q_\varepsilon)$ denote a solution to $(\mathcal{P}_\varepsilon)$. Then $\{(a_\varepsilon, u_\varepsilon, q_\varepsilon)\}$ contains a weakly convergent subsequence as $\varepsilon \to 0^+$, and every weak cluster point $(\bar{a}, \bar{u}, \bar{q})$ is a solution to $(\mathcal{P})$.*

Proof : Let $(a, u, q) \in \mathcal{X}$ satisfy $\operatorname{div} q = f$ and $q = -a\nabla u$. Then, $J_2(a, u, q) = 0$, and, for every $\varepsilon > 0$,

$$J_1(a_\varepsilon, u_\varepsilon) + \frac{1}{\varepsilon} J_2(a_\varepsilon, u_\varepsilon, q_\varepsilon) \le J_1(a, u). \tag{3.8}$$

Since $\operatorname{div} q_\varepsilon = f$ we know that $J_2(a_\varepsilon, u_\varepsilon, q_\varepsilon) \ge 0$, and hence by (3.8)

$$0 \le \int_\Omega (\frac{a_\varepsilon}{2} |\nabla u_\varepsilon|^2 + \frac{1}{2a_\varepsilon} |q_\varepsilon|^2 - f u_\varepsilon) dx \le \varepsilon J_1(a, u). \tag{3.9}$$

It follows that $\{(a_\varepsilon, u_\varepsilon, q_\varepsilon)\}_{\varepsilon>0}$ is bounded in $\mathcal{X}$. Hence there exists a weakly convergent subsequence (denoted by the same symbols) and $(\bar{a}, \bar{u}, \bar{q}) \in \mathcal{X}$ such that

$$(a_\varepsilon, u_\varepsilon, q_\varepsilon) \rightharpoonup (\bar{a}, \bar{u}, \bar{q}) \in \mathcal{X}.$$

As in the proof of Proposition 2.1 this implies that

$$(a_\varepsilon, u_\varepsilon) \to (\bar{a}, \bar{u}) \text{ strongly in } L^\infty \times L^2,$$

$$\sqrt{a_\varepsilon}\nabla u_\varepsilon \rightharpoonup \sqrt{\bar{a}}\nabla \bar{u} \text{ weakly in } L^2_n$$

and

$$\frac{1}{\sqrt{a_\varepsilon}} q_\varepsilon \rightharpoonup \frac{1}{\sqrt{\bar{a}}}\bar{q} \text{ weakly in } L^2_n.$$

Taking the lim inf in (3.9) we obtain

$$\begin{aligned} 0 \;&\leq \int_\Omega (\frac{\bar{a}}{2}|\nabla \bar{u}|^2 + \frac{1}{2\bar{a}}|\bar{q}|^2 - f\bar{u}) \;\; = \int_\Omega \frac{\bar{a}}{2}(|\nabla \bar{u}|^2 + \frac{1}{2\bar{a}}|\bar{q}|^2 + \bar{q}\nabla \bar{u}) \\ &= \int_\Omega \frac{1}{2\bar{a}}|\bar{q} + \bar{a}\nabla \bar{u}|^2 dx \qquad\qquad = 0, \end{aligned}$$

and hence $\bar{q} = -\bar{a}\nabla \bar{u}$. Consequently $(\bar{a}, \bar{u}, \bar{q})$ satisfies all constraints of $(\mathcal{P})$. Moreover by (3.8) we have

$$J_1(\bar{a}, \bar{u}) \leq J_1(a, u)$$

for all (a, u) such that there exists $q \in C$ with (a, u, q) admissible for $(\mathcal{P})$. It follows that $(\bar{a}, \bar{u}, \bar{q})$ is a solution of $(\mathcal{P})$.

■

Proposition 3.3 *Let $\beta > 0$ and let $(\bar{a}, \bar{u}, \bar{q})$ be a weak cluster point of $\{(a_\varepsilon, u_\varepsilon, q_\varepsilon)\}_{\varepsilon>0}$ as $\varepsilon \to 0^+$. Then*

$$(a_\varepsilon, Bu_\varepsilon) \to (\bar{a}, B\bar{u}) \text{ strongly in } Q \times Z \tag{3.10}$$

and

$$\left| \nabla u_\varepsilon + \frac{q_\varepsilon}{a_\varepsilon} \right|_{L^2_n} \leq \frac{\varepsilon}{\nu} |B^*(Bu_\varepsilon - z)|_{H^1_0} \tag{3.11}$$

where ν is the lower bound of the elements of C.

In particular:

- if $Z = H_0^1$ and B is the identity, then $u_\varepsilon \to \bar{u}$ strongly in H_0^1, and

$$\left| \nabla u_\varepsilon + \frac{q_\varepsilon}{a_\varepsilon} \right|_{L_n^2} \leq \frac{\varepsilon}{\nu} |u_\varepsilon - z|_{H_0^1}$$

- if $Z = L^2$ and B is the canonical injection from H_0^1 into L^2, then $u_\varepsilon \to \bar{u}$ strongly in L^2 only and, as in this case $B^* = (-\Delta)^{-1}$, one has

$$\left| \nabla u_\varepsilon + \frac{q_\varepsilon}{a_\varepsilon} \right|_{L_n^2} \leq \frac{\varepsilon}{\nu} |u_\varepsilon - z|_{H^{-1}}.$$

Proof : From the definition of $(a_\varepsilon, u_\varepsilon, q_\varepsilon)$ we obtain

$$J_1(a_\varepsilon, u_\varepsilon) + \frac{1}{\varepsilon} J_2(a_\varepsilon, u_\varepsilon, q_\varepsilon) \leq J_1(\bar{a}, \bar{u}) + \frac{1}{\varepsilon} J_2(\bar{a}, \bar{u}, \bar{q}).$$

Proposition 2.1 implies that $J_2(a_\varepsilon, u_\varepsilon, q_\varepsilon) \geq 0$ and that $J_2(\bar{a}, \bar{u}, \bar{q}) = 0$ as $(\bar{a}, \bar{u}, \bar{q})$ satisfies the elliptic equation. Hence

$$J_1(a_\varepsilon, u_\varepsilon) \leq J_1(\bar{a}, \bar{u})$$

and, taking the $\overline{\lim}$

$$\overline{\lim}\ J_1(a_\varepsilon, u_\varepsilon) \leq J_1(\bar{a}, \bar{u}).$$

On the other hand, J_1 is convex and $(a_\varepsilon, Bu_\varepsilon)$ is weakly convergent in $Q \times Z$ to $(\bar{a}, B\bar{u})$ so that

$$J_1(\bar{a}, \bar{u}) \leq \underline{\lim}\ J_1(a_\varepsilon, u_\varepsilon),$$

and

$$J_1(a_\varepsilon, u_\varepsilon) \to J(\bar{a}, \bar{u}) \text{ when } \varepsilon \to 0.$$

This proves that the sequences $|a_\varepsilon - a^\sharp|_Q$ and $|Bu_\varepsilon - z|_Z$ converge to $|\bar{a} - a^\sharp|_Q$ and $|B\bar{u} - z|_Z$. Hence the weakly convergent sequence $(a_\varepsilon - a^\sharp, Bu_\varepsilon - z)$ converges strongly in $Q \times Z$ to $(\bar{a} - a^\sharp, B\bar{u} - z)$, and the first result (3.10) of Proposition 3.3 is verified.

In order to prove the second result (3.11), we write the first order necessary conditions which are satisfied by the solution $(a_\varepsilon, u_\varepsilon, q_\varepsilon)$ of $(\mathcal{P}_\varepsilon)$:

$$\varepsilon\beta(a_\varepsilon - a^\sharp, h)_Q + \frac{1}{2}\int_\Omega [|\nabla u_\varepsilon|^2 - \frac{|q_\varepsilon|^2}{a_\varepsilon^2}]h \geq 0, \tag{3.12}$$

$$\text{for } a_\varepsilon + \lambda h \in C \text{ and } \lambda \text{ small enough}$$

$$\varepsilon(Bu_\varepsilon - z, Bv) + \int_\Omega (a_\varepsilon \nabla u_\varepsilon, \nabla v) - \int_\Omega fv = 0, \quad \text{for } v \in H_0^1(\Omega) \tag{3.13}$$

$$\int_\Omega (\frac{q_\varepsilon}{a_\varepsilon}, p) = 0, \qquad \text{for } p \in H_{div} \text{ with } \nabla p = 0 \tag{3.14}$$

$$\nabla q_\varepsilon = f. \tag{3.15}$$

Transposing B in (3.13), substituting ∇q_ε for f, using (3.15) and integrating by parts, we obtain

$$\varepsilon(B^*(Bu_\varepsilon - z), v)_{H_0^1} + \int_\Omega (a_\varepsilon \nabla u_\varepsilon, \nabla v) + \int_\Omega (q_\varepsilon, \nabla v) = 0, \qquad v \in H_0^1(\Omega).$$

As H_0^1 is equipped with the scalar product $(u, v)_{H_0^1} = \int_\Omega (\nabla u, \nabla v)$ we get

$$\int_\Omega (\varepsilon\nabla[B^*(Bu_\varepsilon - z)] + a_\varepsilon \nabla u_\varepsilon + q_\varepsilon, \nabla v) = 0, \qquad v \in H_0^1(\Omega). \tag{3.16}$$

We use (3.16) in two different ways :

- first, we choose $v = u_\varepsilon$:

$$\int_\Omega (\varepsilon\nabla[B^*(Bu_\varepsilon - z)] + a_\varepsilon \nabla u_\varepsilon + q_\varepsilon, \nabla u_\varepsilon) = 0. \tag{3.17}$$

- second, we see from (3.16) that the vector field $\varepsilon\nabla[B^*(Bu_\varepsilon - z)] + a_\varepsilon \nabla u_\varepsilon + q_\varepsilon$ has zero divergence. Hence we can chose p in (3.14) equal to this vector field, which gives

$$\int_\Omega (\varepsilon\nabla[B^*(Bu_\varepsilon - z)] + a_\varepsilon \nabla u_\varepsilon + q_\varepsilon, \frac{q_\varepsilon}{a_\varepsilon}) = 0. \tag{3.18}$$

By adding (3.17) and (3.18) we obtain

$$\int_\Omega (\varepsilon \nabla [B^*(Bu_\varepsilon - z)] + a_\varepsilon \nabla u_\varepsilon + q_\varepsilon, \nabla u_\varepsilon + \frac{q_\varepsilon}{a_\varepsilon}) = 0;$$

i.e.

$$\int_\Omega a_\varepsilon \left| \nabla u_\varepsilon + \frac{q_\varepsilon}{a_\varepsilon} \right|^2 + \varepsilon \int_\Omega (\nabla [B^*(Bu_\varepsilon - z)], \nabla u_\varepsilon + \frac{q_\varepsilon}{a_\varepsilon}) = 0;$$

and, as $a_\varepsilon \geq \nu > 0$,

$$\nu \left| \nabla u_\varepsilon + \frac{q_\varepsilon}{a_\varepsilon} \right|_{L^2_n} \leq \varepsilon |B^*(Bu_\varepsilon - z)|_{H^1_0},$$

which is (3.11). This ends the proof of Proposition 3.3. ∎

We conclude this paragraph with the analysis of the convexity properties of the cost functional

$$J_\varepsilon(a, u, q) = J_1(a, u) + \frac{1}{\varepsilon} J_2(a, u, q)$$

in $(\mathcal{P}_\varepsilon)$. We give first the second derivative of J_2.

Proposition 3.4 *For any $(a, u, q) \in \mathcal{X}$ such that $\nabla q = f$ and any $(h, v, p) \in \mathcal{X}$ such that $\nabla p = 0$ one has*

$$J_2''(a, u, q)(h, v, p)^2 = \int_\Omega \frac{1}{a} | - \frac{h}{a} q + a \nabla v + p|^2 + 2 \int_\Omega h(\nabla u + \frac{q}{a}, \nabla v). \tag{3.19}$$

Proof : Differentiating twice in (3.3), the definition of J_2, we obtain

$$J_2''(a, u, q)(h, v, p)^2 = \int_\Omega \frac{|q|^2}{a^3} h^2 + 2 \int_\Omega (\nabla u, \nabla v) h - 2 \int_\Omega \frac{(q, p)}{a^2} h + \int_\Omega a |\nabla v|^2 + \int_\Omega \frac{1}{a} |p|^2.$$

Define

$$p_0 = -\frac{h}{a} q + a \nabla v + p$$

and substitute for p in the third term of the right-hand side

$$\begin{aligned} J_2''(a,u,q)(h,v,p)^2 &= \int_\Omega \frac{|q|^2}{a^3}h^2 + 2\int_\Omega (\nabla u, \nabla v)h \\ &\quad -2\int_\Omega (\frac{q}{a^2}, p_0 + \frac{h}{a}q - a\nabla v)h + \int_\Omega a|\nabla v|^2 + \int_\Omega \frac{1}{a}|p|^2. \end{aligned}$$

Expanding and rearranging terms we obtain

$$\begin{aligned} J_2''(a,u,q)(h,v,p)^2 &= -\int_\Omega \frac{|q|^2}{a^3}h^2 - 2\int_\Omega (q,p_0)\frac{h}{a^2} + 2\int_\Omega (\nabla u + \frac{q}{a}, \nabla v)h \\ &\quad + \int_\Omega a|\nabla v|^2 + \int_\Omega \frac{1}{a}|p|^2. \end{aligned}$$

Noting that the first two terms of the right-hand side form part of a perfect square, we write

$$\begin{aligned} J_2''(a,u,q)(h,v,p)^2 &= -\int_\Omega \frac{1}{a}\left|\frac{qh}{a} + p_0\right|^2 + \int_\Omega \frac{1}{a}|p_0|^2 + 2\int_\Omega h(\nabla u + \frac{q}{a}, \nabla u) \\ &\quad + \int_\Omega a|\nabla v|^2 + \int_\Omega \frac{1}{a}|p|^2. \end{aligned}$$

Using the definition of p_0, we can rewrite the first term as

$$\int_\Omega \frac{1}{a}\left|\frac{qh}{a} + p_0\right|^2 = \int_\Omega \frac{1}{a}|a\nabla v + p|^2 = \int_\Omega a|\nabla v|^2 + \int_\Omega (\nabla v, p) + \int_\Omega \frac{1}{a}|p|^2.$$

But the central term vanishes, as $\nabla p = 0$. Plugging the two remaining terms into the last formula for J_2'' produces the announced result. ∎

We remark that (3.19) is not unexpected given the properties of $J_2(a,u,q)$ seen in §2: if (a,u,q) is a minimizer of J_2, then $q + a\nabla u = 0$ (the equation is satisfied) and (3.19) reduces to

$$J_2''(a,u,q)(h,v,p)^2 = \int_\Omega \frac{1}{a}|h\nabla u + a\nabla v + p|^2$$

which is always positive and vanishes in the directions (h,v,p) in which the equation $q + a\nabla u = 0$ is satisfied up to the first order (such directions are "tangent" to the set of minimizers of J_2, on which J_2 has the constant value zero).

Corollary 3.1 *(partial convexity of J_2 and hence of J_ε)*

- *for any fixed $a \in C$, J_2 and J_ε are globally convex with respect to (u, q)*
- *for any fixed $u \in H_0^1(\Omega)$, J_2 and J_ε are globally convex with respect to (a, q).*

Proof : These results follow immediately from (3.19) and from the fact that a function whose Hessian is positive everywhere is necessarily convex. ∎

We can now investigate the coercivity of the primal-dual objective function J_ε at minimizers $(a_\varepsilon, u_\varepsilon, q_\varepsilon)$ of $(\mathcal{P}_\varepsilon)$. We shall need the following hypotheses:

$$\text{the observation space } Z \text{ is } H_0^1, \text{ hence } B = Id, \tag{3.20}$$

$$\begin{aligned} &\text{the data } z \in H_0^1 \text{ is attainable, i.e. there exists} \\ &(\bar{a}, \bar{u}, \bar{q}) \in C \times H_0^1 \times H_{div} \text{ such that } \nabla \bar{q} = f,\ \bar{q} + \bar{a}\nabla\bar{u} = 0 \text{ and } \bar{u} = z. \end{aligned} \tag{3.21}$$

Before giving the uniform coercivity result for J_ε, let us remark that one could replace hypothesis (3.20) by the inclusion of a regularization proportional to $|\nabla u|^2$ in J_ε (state space regularization).

Proposition 3.5 *Let hypotheses (3.20) and (3.21) hold, and let $\{(a_\varepsilon, u_\varepsilon, q_\varepsilon)\}_{\varepsilon>0}$ be any sequence of minimizers of $(\mathcal{P}_\varepsilon)$. Then there exists $\bar{\beta} > 0$ and for every $\beta \in]0, \bar{\beta}[$ an $\bar{\varepsilon}(\beta) > 0$ such that for every $\varepsilon \in]0, \bar{\varepsilon}(\beta)[$ there exists a convex neighborhood $V(a_\varepsilon) \times V(u_\varepsilon) \times V(q_\varepsilon)$ in $Q \times H_0^1 \times L_n^2$ of $(a_\varepsilon, u_\varepsilon, q_\varepsilon)$ and $\gamma > 0$ such that*

$$\begin{aligned} &\textit{for all } (a, u, q) \in V(a_\varepsilon) \times V(u_\varepsilon) \times V(q_\varepsilon) \\ &\textit{and for all } (h, v, p) \in Q \times H_0^1 \times L_n^2 \textit{with } \nabla q = f \textit{ and } \nabla p = 0 \\ &J_\varepsilon''(a, u, q)(h, v, p)^2 \geq \gamma(|\nabla v|_{L_n^2}^2 + |h|_Q^2 + |p|_{L_n^2}^2). \end{aligned} \tag{3.22}$$

Proof : Due to the attainability assumption, one has for any $\varepsilon > 0$ and $\beta > 0$:

$$\frac{1}{2}|\nabla(u_\varepsilon - z)|_{L_n^2}^2 + \frac{\beta}{2}|a_\varepsilon - a^\sharp|_Q^2 + \frac{1}{2\varepsilon}\int_\Omega a_\varepsilon |\nabla u_\varepsilon + \frac{q_\varepsilon}{a_\varepsilon}|^2 \leq \frac{\beta}{2}|\bar{a} - a^\sharp|^2.$$

Thus it is simple to argue that

$$\begin{aligned}&\text{there exists } M>0, \text{ such that for } 0<\beta\le 1 \text{ and } 0<\varepsilon\le 1,\\ &|a_\varepsilon|_Q \le M \text{ and } |q_\varepsilon|_{L^2_n} \le M.\end{aligned} \tag{3.23}$$

Now Proposition 3.4 implies that, for any $a \in B_Q(a_\varepsilon, 1), \varepsilon > 0$ and $\beta > 0$,

$$\begin{aligned}J''_\varepsilon(a,u,q)(h,v,p)^2 \ge |\nabla v|^2_{L^2_n} + \beta|h|^2_Q + \frac{1}{\varepsilon k_\infty(M+1)}|p_0|^2_{L^2_n}\\ +\frac{2}{\varepsilon}\int_\Omega h(\nabla u + \frac{q}{a}, \nabla v),\end{aligned} \tag{3.24}$$

where

$$p_0 = -h\frac{q}{a} + a\nabla v + p. \tag{3.25}$$

Using the Cauchy-Schwarz inequality in the last term of (3.24) we obtain

$$|\frac{2}{\varepsilon}\int_\Omega h(\nabla u + \frac{q}{a}, \nabla v)| \le \frac{\beta}{2}|h|^2_Q + \frac{2k^2_\infty}{\varepsilon^2\beta}|\nabla u + \frac{q}{a}|^2|\nabla v|^2$$

so that (3.24) becomes

$$\begin{aligned}J''(a,u,q)(h,v,p)^2 \ge (1 - \frac{2k^2_\infty}{\varepsilon^2\beta}|\nabla u + \frac{q}{a}|^2_{L^2_n})|\nabla v|^2_{L^2_n} + \frac{\beta}{2}|h|^2_Q\\ +\frac{1}{\varepsilon k_\infty(M+1)}|p_0|^2_{L^2_n}\end{aligned} \tag{3.26}$$

We now choose β, ε and the neighborhood $V(a_\varepsilon) \times V(u_\varepsilon) \times V(q_\varepsilon)$ in such a way that

$$\frac{2k^2_\infty}{\varepsilon^2\beta}\left|\nabla u + \frac{q}{a}\right|^2_{L^2_n} \le \frac{1}{2},$$

i.e. such that

$$\frac{2k_\infty}{\varepsilon\beta^{\frac{1}{2}}}\left|\nabla u + \frac{q}{a}\right|_{L^2_n} \le 1.$$

We have

$$\frac{2k_\infty}{\varepsilon\beta^{\frac{1}{2}}}\left|\nabla u + \frac{q}{a}\right|_{L^2_n} \le \frac{2k_\infty}{\varepsilon\beta^{\frac{1}{2}}}\left|\nabla u_\varepsilon + \frac{q_\varepsilon}{a_\varepsilon}\right|_{L^2_n} + \frac{2k_\infty}{\varepsilon\beta^{\frac{1}{2}}}\left|\nabla(u-u_\varepsilon) + \frac{q}{a} - \frac{q_\varepsilon}{a_\varepsilon}\right|_{L^2_n};$$

i.e., using Proposition 3.3,

$$\frac{2k_\infty}{\varepsilon\beta^{\frac{1}{2}}}\left|\nabla u+\frac{q}{a}\right| \le \frac{2k_\infty}{\nu\beta^{\frac{1}{2}}}|u_\varepsilon - z|_{H_0^1} + \frac{2k_\infty}{\varepsilon\beta^{\frac{1}{2}}}\left|\nabla(u-u_\varepsilon)+\frac{q}{a}-\frac{q_\varepsilon}{a_\varepsilon}\right|_{L_n^2}. \tag{3.27}$$

Let $(\bar{a}^\beta, \bar{u}^\beta) \in C \times H_0^1$ be the solution of the regularized problem:

$$\begin{aligned} &\min \frac{1}{2}|\nabla(u-z)|^2_{L_n^2} + \frac{\beta}{2}|a-a^\sharp|^2_Q \\ &\text{over } (a,u) \in C \times H_0^1 \text{ and } \nabla(a\nabla u) = f. \end{aligned} \tag{3.28}$$

Due to the attainability assumption (3.21), it is well-known [14] that there exists a $\tilde{\beta} > 0$ such that, for any $\beta \in]0, \tilde{\beta}]$, the solution $(\bar{a}^\beta, \bar{u}^\beta)$ of (3.28) is unique. Hence we see from Proposition 3.3 that the sequence $(a_\varepsilon, u_\varepsilon)$, itself, converges to $(\bar{a}^\beta, \bar{u}^\beta)$ in $Q \times H_0^1$. Hence for any $\beta \in]0, \tilde{\beta}]$, there exists a function $\rho_\beta(\varepsilon) > 0$ with $\rho_\beta(\varepsilon) \to 0$ as $\varepsilon \to 0$, such that

$$|u_\varepsilon - \bar{u}^\beta|_{H_0^1} \le \rho_\beta(\varepsilon).$$

Another classical consequence of the attainability assumption (3.21) is the following rate of convergence of $\bar{u}^\beta$ to z in $H_0^1(\Omega)$:

$$|\bar{u}^\beta - z|_{H_0^1} = \beta^{1/2}\rho(\beta),$$

where $\rho(\beta) > 0$ and $\rho(\beta) \to 0$ as $\beta \to 0$. Knowing this, we can write, for any $\beta \in]0, \tilde{\beta}[$ and $\varepsilon > 0$:

$$|u_\varepsilon - z|_{H_0^1} \le |u_\varepsilon - \bar{u}^\beta|_{H_0^1} + |\bar{u}^\beta - z|_{H_0^1} \le \rho_\beta(\varepsilon) + \beta^{1/2}\rho(\beta).$$

Thus (3.27) becomes

$$\begin{aligned} \frac{2k_\infty}{\varepsilon\beta^{1/2}}|\nabla u+\frac{q}{a}| \;\le\; & \frac{2k_\infty}{\nu}\rho(\beta) + \frac{2k_\infty}{\nu\beta^{1/2}}\rho(\beta)(\varepsilon) \\ & + \frac{2k_\infty}{\varepsilon\beta^{1/2}}|\nabla(u-u_\varepsilon)+\frac{q}{a}-\frac{q_\varepsilon}{a_\varepsilon}|_{L_n^2}. \end{aligned} \tag{3.29}$$

We can now chose $0 < \bar{\beta} \le \tilde{\beta}$ such that

$$\frac{2k_\infty}{\nu}\rho(\beta) \le 1/3 \quad \text{for each } \beta \in]0, \bar{\beta}],$$

and, for any such β, choose $\bar{\varepsilon}(\beta)$ such that

$$\frac{2k_\infty}{\nu\beta^{1/2}}\rho_\beta(\varepsilon) \leq 1/3 \quad \text{for each } \varepsilon \in]0, \bar{\varepsilon}(\beta)].$$

Then, for β and ε chosen as above, one can choose the neighborhood $V(a_\varepsilon) \times V(u_\varepsilon) \times V(q_\varepsilon)$ in such a way that $V(a_\varepsilon) \subset B_Q(a_\varepsilon, 1)$, $V(q_\varepsilon) \subset B_{L^2_n}(q_\varepsilon, 1)$, and that

$$\frac{2k_\infty}{\varepsilon\beta^{\frac{1}{2}}}\left|\nabla(u-u_\varepsilon) + \frac{q}{a} - \frac{q_\varepsilon}{a_\varepsilon}\right|_{L^2_n} \leq 1/3,$$
$$\text{for each } (a,u,q) \in V(a_\varepsilon) \times V(u_\varepsilon) \times V(q_\varepsilon).$$

Then for $0 < \beta \leq \bar{\beta}$ and $0 < \varepsilon \leq \bar{\varepsilon}(\beta)$, inequality (3.26) may be rewritten as:

$$J''_\varepsilon(a,u,q)(h,v,p)^2 \geq \frac{\beta}{2}|h|^2_Q + \frac{1}{2}|\nabla u|^2_{L^2_n} + \frac{1}{\varepsilon k_\infty(M+1)}|p_0|^2_{L^2_n} \tag{3.30}$$

We now estimate the continuity constant of the mapping $(h,v,p_0) \rightsquigarrow (h,v,p)$ when $a \in B_Q(a_\varepsilon,1)$ and $q \in B_{L^2_n}(q_\varepsilon,1)$. It will be convenient to define the following weighted norm:

$$|(h,v,p)|^2 = \beta|h|^2_Q + |\nabla v|^2_{L^2_n} + \frac{1}{\varepsilon k_\infty(M+1)}|p|^2_{L^2_n}. \tag{3.31}$$

From (3.25) we obtain immediately that

$$|p|^2_{L^2_n} \leq 3|h\frac{q}{a}|^2_{L^2_n} + 3|a\nabla v|^2_{L^2_n} + 3|p_0|^2_{L^2_n},$$

i.e., using (3.23) and the fact that $a \in C$ defined in (3.1) and $q \in B_{L^2_n}(q_\varepsilon,1)$,

$$|p|^2_{L^2_n} \leq \frac{3k^2_\infty(M+1)^2}{\nu^2}|h|^2_Q + 3k^2_\infty(M+1)^2|\nabla v|^2_{L^2_n} + 3|p_0|^2_{L^2_n}.$$

Plugging this estimate for $|p|^2_{L^2_n}$ into (3.31), we get

$$|(h,v,p)|^2 \leq \left(1 + 3\frac{k^2_\infty(M+1)^2}{\nu^2}\right)|h|^2_Q + (1 + 3k^2_\infty(M+1)^2)|\nabla v|^2_{L^2_n} + 3|p_0|^2_{L^2_n}.$$

Thus

$$|(h,v,p)|^2$$
$$\leq \max\left\{[2+6\tfrac{k_\infty^2(M+1)^2}{\nu^2}]/\beta\ ,\ 2+6k_\infty^2(M+1)^2\ ,\ 3\varepsilon k_\infty(M+1)\right\}|(h,v,p_0)|_0^2,$$

or

$$|(h,v,p_0)|^2 \geq \gamma|(h,v,p)|^2,$$

where

$$\gamma = \min\left\{\frac{\beta}{2+6k_\infty^2(M+1)^2/\nu^2}\ ,\ \frac{1}{2+6k_\infty^2(M+1)^2}\ ,\ \frac{1}{3\varepsilon k_\infty(M+1)}\right\}.$$

Hence (3.30) becomes

$$J_\varepsilon''(a,u,q)(h,v,p)^2 \geq \gamma(|h|_Q^2+|\nabla v|_{L_n^2}^2+|p|_{L_n^2}^2),$$

which completes the proof of Proposition 3.5. ∎

4 Splitting algorithms for the numerical resolution

The uniform convexity of the cost functional J_ε in each of the variables separately suggests solving $(\mathcal{P}_\varepsilon)$ by splitting algorithms. The functional J_ε is not jointly convex in the variables (a,u,q), however, and hence the convergence of the splitting algorithms must be considered locally. We shall carry out the analysis on the basis of Proposition 3.5 which asserts the local convexity of J_ε in some neighborhood of each solution $(a_\varepsilon,u_\varepsilon,q_\varepsilon)$ to $(\mathcal{P}_\varepsilon)$.

Throughout this section it is assumed that (3.19) and (3.20) hold and that β and ε are chosen such that for a fixed solution $(a_\varepsilon,u_\varepsilon,q_\varepsilon)$ the conclusion (3.21) of Proposition 3.5 holds. Further $U(a_\varepsilon)\times U(u_\varepsilon)\times U(q_\varepsilon)$, is a convex neighborhood of $(a_\varepsilon,u_\varepsilon,q_\varepsilon)$ satisfying $\overline{U}(a_\varepsilon)\times\overline{U}(u_\varepsilon)\times\overline{U}(q_\varepsilon)\subset V(a_\varepsilon)\times V(u_\varepsilon)\times V(q_\varepsilon)$. In the first algorithm that we analyze, minimization is carried out separately with respect to a,q and u.

Algorithm 4.1

(i) Set $u_0 = z \in \overline{U}(u_\varepsilon)$, set $n = 1$ and choose $q_0 \in \overline{U}(q_\varepsilon)$.
(ii) $a_n =$ argmin $J_\varepsilon(a, u_{n-1}, q_{n-1})$ over $a \in C \cap \overline{U}(a_\varepsilon)$.
(iii) $q_n =$ argmin $J_\varepsilon(a_n, u_{n-1}, q)$ over $q \in \overline{U}(q_\varepsilon)$, div $q = f$.
(iv) $u_n =$ argmin $J_\varepsilon(a_n, u, q_n)$ over $u \in \overline{U}(u_\varepsilon)$.
(v) check convergence, stop or set $n = n + 1$ and go to (ii).

Note that the cost functions in (iii) and (iv) are quadratic. Also the cost functional is separable with respect to u and q, the only coupling occuring in the $\int_\Omega a|\nabla u|^2$-term. Hence (iii) and (iv) can be solved in parallel. We shall prove that $(a_n, u_n, q_n) \to (a_\varepsilon, u_\varepsilon, q_\varepsilon)$ in $\mathcal{X}$ so that after finitely many steps of the iteration the constraints $a \in \overline{U}(a_\varepsilon)$, $q \in \overline{U}(q_\varepsilon)$ and $u \in \overline{U}(u_\varepsilon)$ become inactive. In the statement of the following theorem the notation of Proposition 3.5 is used.

Theorem 4.1 *Assume that (3.20) and (3.21) hold, and let* $\beta \in]0, \bar{\beta}]$, *and* $\varepsilon \in]0, \bar{\varepsilon}(\beta)]$. *Then the sequence* (a_n, u_n, q_n) *generated by Algorithm 4.1 converges in* $\mathcal{X}$ *to* $(a_\varepsilon, u_\varepsilon, q_\varepsilon)$.

Proof : It is simple to argue the existence of unique solutions (a_n, u_n, q_n) in $(ii), (iv)$ of the algorithm. Concerning the convergence of (a_n, u_n, q_n), the proof relies on arguments that are similar to standard ones in the context of splitting algorithms [4]. The special structure of the problem, however, does not allow us to refer directly to known results. The solutions of (ii)-(iv) satisfy

$$\frac{\partial}{\partial a} J_1(a_n, u_{n-1})(a - a_n) + \frac{1}{\varepsilon}(J_2(a, u_{n-1}, q_{n-1}) - J_2(a_n, u_{n-1}, q_{n-1})) \geq 0$$
$$\text{for all } a \in C \cap \overline{U}(a_\varepsilon) \quad (4.1)$$

$$\frac{\partial}{\partial q} J_2(a_n, u_{n-1}, q_n)(q - q_n) \geq 0$$
$$\text{for all } q \in \overline{U}(q_\varepsilon) \text{ such that div } q = f, \quad (4.2)$$

$$\frac{\partial}{\partial u} J_1(a_n, u_n)(u - u_n) + \frac{1}{\varepsilon}(J_2(a_n, u, q_n) - J_2(a_n, u_n, q_n)) \geq 0$$
$$\text{for all } u \in \overline{U}(u_\varepsilon). \quad (4.3)$$

Note that for every $n = 2, 3, ...$

$$J_\varepsilon(a_{n-1}, u_{n-1}) - J_\varepsilon(a_n, u_n) = J_\varepsilon(a_{n-1}, u_{n-1}, q_{n-1}) - J_\varepsilon(a_n, u_{n-1}, q_{n-1})$$
$$+ J_\varepsilon(a_n, u_{n-1}, q_{n-1}) - J_\varepsilon(a_n, u_{n-1}, q_n) + J_\varepsilon(a_n, u_{n-1}, q_n) - J_\varepsilon(a_n, u_n, q_n) \geq 0,$$

and therefore $J_\varepsilon(a_n, u_n, q_n)$ is monotonically decreasing with respect to n. Since this sequence is also bounded from below it is necessarily convergent. Moreover we find from (4.1) - (4.3) that

$$\begin{aligned}
&J_\varepsilon(a_{n-1}, u_{n-1}, q_{n-1}) - J_\varepsilon(a_n, u_n, q_n) \\
&\quad = J_1(a_{n-1}, u_{n-1}) - J_1(a_n, u_{n-1}) - \frac{\partial}{\partial a} J_1(a_n, u_{n-1})(a_{n-1} - a_n) \\
&\qquad + \frac{\partial}{\partial a} J_1(a_n, u_{n-1})(a_{n-1} - a_n) + \frac{1}{\varepsilon}(J_2(a_{n-1}, u_{n-1}, q_{n-1}) \\
&\qquad - J_2(a_n, u_{n-1}, q_{n-1})) + \frac{1}{\varepsilon}(J_2(a_n, u_{n-1}, q_{n-1}) - J_2(a_n, u_{n-1}, q_n)) \\
&\qquad + J_1(a_n, u_{n-1}) - J_1(a_n, u_n) - \frac{\partial}{\partial u} J_1(a_n, u_n)(u_{n-1} - u_n) \\
&\qquad + \frac{\partial}{\partial u} J_1(a_n, u_n)(u_{n-1} - u_n) + \frac{1}{\varepsilon}(J_2(a_n, u_{n-1}, q_n) - J_2(a_n, u_n, q_n)) \\
&\quad \geq J_1(a_{n-1}, u_{n-1}) - J_1(a_n, u_{n-1}) - \frac{\partial}{\partial a} J_1(a_n, u_{n-1})(a_{n-1} - a_n) \\
&\qquad + \frac{1}{\varepsilon} \int_\Omega \frac{1}{a_n} |q_{n-1} - q_n|^2 + J_1(a_n, u_{n-1}) - J_1(a_n, u_n) \\
&\qquad - \frac{\partial}{\partial u} J_1(a_n, u_n)(u_{n-1} - u_n) \\
&\quad \geq \frac{\beta}{2} |a_n - a_{n-1}|_Q^2 + \frac{1}{2} |\nabla(u_{n-1} - u_n)|_{L^2}^2 + \frac{1}{\varepsilon} \int \frac{1}{a_n} |q_{n-1} - q_n|^2.
\end{aligned}$$

In particular it follows that

$$\lim_{n \to \infty} |(a_n, u_n, q_n) - (a_{n-1}, u_{n-1}, q_{n-1})|_{\mathcal{X}} = 0. \tag{4.4}$$

In the following step we use (3.21) and the fact that by construction $(a_n, u_n, q_n) \in V(a_\varepsilon) \times V(u_\varepsilon) \times V(q_\varepsilon)$:

$$\begin{aligned}
&(J'_\varepsilon(a_n, u_n, q_n) - J'_\varepsilon(a_\varepsilon, u_\varepsilon, q_\varepsilon))(a_n - a_\varepsilon, u_n - u_\varepsilon, q_n - q_\varepsilon) \\
&\quad = \int_0^1 J''_\varepsilon(t(a_n, u_n, q_n) + (1-t)(a_\varepsilon, u_\varepsilon, q_\varepsilon))((a_n, u_n, q_n) - (a_\varepsilon, u_\varepsilon, q_\varepsilon))^2 dt \\
&\quad \geq \gamma |(a_n, u_n, q_n) - (a_\varepsilon, u_\varepsilon, q_\varepsilon)|_{\mathcal{X}}^2.
\end{aligned}$$

Since $J'_\varepsilon(a_\varepsilon, u_\varepsilon, q_\varepsilon)(a_n - a_\varepsilon, u_n - u_\varepsilon, q_n - q_\varepsilon) \geq 0$, this implies that

$$\begin{aligned}(\frac{\partial}{\partial a}J_\varepsilon(a_n, u_n, q_n) - \frac{\partial}{\partial a}J_\varepsilon(a_n, u_{n-1}, q_{n-1}))(a_n - a_\varepsilon) \\ +\frac{\partial}{\partial q}J_\varepsilon(a_n, u_{n-1}, q_{n-1})(a_n - a_\varepsilon) \\ +(\frac{\partial}{\partial q}J_\varepsilon(a_n, u_n, q_n) - \frac{\partial}{\partial q}J_\varepsilon(a_n, u_{n-1}, q_n))(q_n - q_\varepsilon) \\ +\frac{\partial}{\partial q}J_\varepsilon(a_n, u_{n-1}, q_n)(q_n - q_\varepsilon) + \frac{\partial}{\partial u}J_\varepsilon(a_n, u_n, q_n)(u_n - u_\varepsilon) \\ \geq \gamma|(a_n, u_n, q_n) - (a_\varepsilon, u_\varepsilon, q_\varepsilon)|^2_{\mathcal{X}}.\end{aligned} \tag{4.5}$$

But $\frac{\partial}{\partial a}J_\varepsilon(a_n, u_{n-1}, q_{n-1})(a_n - a_\varepsilon) \leq 0$, $\frac{\partial}{\partial q}J_\varepsilon(a_n, u_{n-1}, q_n))(q_n - q_\varepsilon) \leq 0$ and $\frac{\partial}{\partial u}J_\varepsilon(a_n, u_n, q_n)(u_n - u_\varepsilon) \leq 0$, so that

$$\begin{aligned}\frac{1}{2\varepsilon}\int_\Omega (|\nabla u_n|^2 - |\nabla u_{n-1}|^2)(a_n - a_\varepsilon)dx - \frac{1}{2\varepsilon}\int_\Omega \frac{1}{a_n^2}(|q_n|^2 - |q_{n-1}|^2)(a_n - a_\varepsilon)dx \\ \geq \gamma|(a_n, u_n, q_n) - (a_\varepsilon, u_\varepsilon, q_\varepsilon)|^2_{\mathcal{X}}.\end{aligned} \tag{4.6}$$

The boundedness of $\{a_n\}$ in Q together with (4.4) implies that

$$\lim_{n\to\infty}(a_n, u_n, q_n) = (a_\varepsilon, u_\varepsilon, q_\varepsilon) \text{ in } \mathcal{X}.$$

■

The second splitting algorithm that we discuss requires only an initial guess for u_0 and z is a good choice.

Algorithm 4.2

- (i) Set $u_0 = z \in \overline{U}(u_\varepsilon)$ and let $n = 1$
- (ii) $(a_n, q_n) =$ argmin $J_\varepsilon(a, u_{n-1}, q)$ over $(a, q) \in (C \cap \overline{U}(u_\varepsilon)) \times \overline{U}(q_\varepsilon)$, div $q = f$,
- (iii) $u_n =$ argmin $J_\varepsilon(a_n, u, q_n)$ over $u \in \overline{U}(u_\varepsilon)$.
- (iv) check convergence, stop or set $n = n + 1$ and go to (ii).

Theorem 4.2 *Under the hypotheses of Theorem 4.1 the sequence* (a_n, u_n, q_n) *generated by Algorithm 4.2 converges in* $\mathcal{X}$ *to* $(a_\varepsilon, u_\varepsilon, q_\varepsilon)$.

Proof : Necessary optimality conditions for (a_n, q_n) and for u_n, respectively, are given by

$$J_\varepsilon(a, u_{n-1}, q) - J_\varepsilon(a_n, u_{n-1}, q_n) \geq 0$$
$$\text{for all } (a, q) \in (\overline{U}(a_\varepsilon) \cap C) \times \overline{U}(q_\varepsilon), \text{ div } q = f, \quad (4.7)$$

and

$$\frac{\partial}{\partial u} J_1(a_n, u_n, q_n)(u - u_n) + \frac{1}{\varepsilon}(J_2(a_n, u, q_n) - J_2(a_n, u_n, q_n)) \geq 0$$
$$\text{for all } u \in \overline{U}(u_\varepsilon). \quad (4.8)$$

Further, for all $n = 1, 2, ...$ we have

$$\begin{aligned} &J_\varepsilon(a_{n-1}, u_{n-1}, q_{n-1}) - J_\varepsilon(a_n, u_n, q_n) \\ &\quad \geq J_\varepsilon(a_{n-1}, u_{n-1}, q_{n-1}) - J_\varepsilon(a_n, u_{n-1}, q_n) + J_\varepsilon(a_n, u_{n-1}, q_n) - J_\varepsilon(a_n, u_n, q_n) \\ &\quad \geq J_1(a_n, u_{n-1}) - J_1(a_n, u_n) - \frac{\partial}{\partial u} J_1(a_n, u_n)(u_{n-1} - u_n) \\ &\qquad + \frac{\partial}{\partial u} J_1(a_n, u_n)(u_{n-1} - u_n) + \frac{1}{\varepsilon}[J_2(a_n, u_{n-1}, q_n) - J_2(a_n, u_n, q_n)] \\ &\quad \geq \frac{1}{2}|\nabla(u_{n-1} - u_n)|^2, \end{aligned}$$

and thus

$$\lim_{n \to \infty} |u_n - u_{n-1}|_{H_0^1} = 0. \quad (4.9)$$

The estimate corresponding to (4.5) is

$$\begin{aligned} &(\frac{\partial}{\partial a} J_\varepsilon(a_n, u_n, q_n) - \frac{\partial}{\partial a} J_\varepsilon(a_n, u_{n-1}, q_n)(a_n - a_\varepsilon) + \frac{\partial}{\partial a} J_\varepsilon(a_n, u_{n-1}, q_n)(a_n - a_\varepsilon) \\ &\qquad + (\frac{\partial}{\partial q} J_\varepsilon(a_n, u_n, q_n) - \frac{\partial}{\partial q} J_\varepsilon(a_n, u_{n-1}, q_n))(q_n - q_\varepsilon) \\ &\qquad + \frac{\partial}{\partial q} J_\varepsilon(a_n, u_{n-1}, q_n)(q_n - q_\varepsilon) + \frac{\partial}{\partial u} J_\varepsilon(a_n, u_n, q_n) \\ &\quad \geq \gamma |(a_n, u_n, q_n) - (a_\varepsilon, u_\varepsilon, q_\varepsilon)|_{\mathcal{X}}^2, \end{aligned}$$

and hence

$$\frac{1}{2\varepsilon}\int_\Omega (|\nabla u_n|^2 - |\nabla u_{n-1}|^2)(a_n - a_\varepsilon) \geq \gamma |(a_n, u_n, q_n) - (a_\varepsilon, u_\varepsilon, q_\varepsilon)|^2_{\mathcal{X}}. \tag{4.10}$$

It is simple to argue that $\{a_n\}$ is bounded in Q and hence combining (4.9) and (4.10) it follows that $\lim_{n\to\infty}(a_n, u_n, q_n) = (a_\varepsilon, u_\varepsilon, q_\varepsilon)$. ■

5 Mixed finite element implementation

We describe here the numerical discretization used for the implementation of the primal-dual formulation for the estimation of the diffusion coefficient in the elliptic equation

$$\begin{aligned} -\nabla(a\nabla u) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{5.1}$$

More precisely, we construct a primal-dual formulation for a discretization of (5.1) rather than discretize one of the primal dual formulations given in the examples of §2. We have used for the discretization a mixed finite element scheme as with such a scheme both the primal and dual energy functionals are readily calculated. The primal-dual formulation that we shall use for the discretized problem follows the primal-dual formulation given in Example 2.1 of §2 and analyzed in §3:

$$\begin{aligned} &\min \left\{ \frac{1}{2}|z - Bu|^2_Z + \frac{\beta}{2}|a - a^\sharp|^2_Q + \frac{1}{\varepsilon}\left(\frac{1}{2}\int_\Omega a|\nabla u|^2 - \int_\Omega fu + \frac{1}{2}\int_\Omega \frac{1}{a}|q|^2\right)\right\} \\ &\text{over } (a, u, q) \in C \times H^1_0 \times H_{div}, \nabla q = f \in L^2. \end{aligned} \tag{$\mathcal{P}_\varepsilon$}$$

We first describe briefly the mixed finite element numerical scheme that we use to discretize (5.1). For more details concerning the numerical method see [9] for the implementation or [6, 15] for a more theoretical treatment. For simplicity we suppose that $\Omega \subset \mathrm{R}^2$. Extension to higher dimension is straightforward as is extension to other types of boundary conditions.

Let $\mathcal{T}_h$ be a triangulation of Ω by triangles and/or rectangles, and let $\mathcal{E}_h$ be the set of edges of elements of $\mathcal{T}_h$. We shall denote by K, respectively E, the typical element of $\mathcal{T}_h$, respectively $\mathcal{E}_h$, and by NT, respectively NE, its cardinality.

The mixed finite element method that we shall use for the discretization is based on the following mixed variational formulation of (5.1):

$$
\begin{aligned}
&(u,q) \in L^2(\Omega) \times H_{div} \\
&\int_\Omega \frac{1}{a} q\, q' - \int_\Omega u \,\mathrm{div}\, q' = 0, \qquad \text{for } q' \in H_{div} \\
&\int_\Omega \mathrm{div}\, q\, u' = \int_\Omega f\, u', \qquad \text{for } v' \in L^2(\Omega).
\end{aligned}
$$

We shall thus approximate the state variable u in a finite-dimensional subspace of $L^2(\Omega)$. The space of discretized state variables, X_h, will be the space of piecewise constant functions, functions constant on each element K of $\mathcal{T}_h$. The dual state variable q will be approximated in the lowest-order Raviart-Thomas space for the approximation of H_{div}-functions, which in keeping with the notation of §2 we shall denote Y_h. A basis for X_h is the set of characteristic functions χ_K of the elements K of $\mathcal{T}_h$.

$$X_h = \text{ span}\{\chi_K; K \in \mathcal{T}_h\}.$$

To specify a basis for Y_h we choose for each edge $E \in \mathcal{E}_h$ a unit vector ν_E normal to E and define the basis function ω_E to be the unique function satisfying

- each component of ω_E is linear on each element K of $\mathcal{T}_h$
- $\omega_E \in H_{div}$ i.e. $\omega_{E|K} \cdot \nu_E = \omega_{E|K'} \cdot \nu_E$ if E is an edge of K and of K'
- for each edge $E' \in \mathcal{E}_h$, $\omega_E \cdot \nu_{E'}$ is constant on E'
- ω_E has 0 flux across each edge in $\mathcal{E}_h$ other than E itself where it has flux equal to one:

$$\int_{E'} \omega_E \cdot \nu_{E'} = \begin{cases} 1 \text{ if } E = E' \\ 0 \text{ if } E \neq E' \end{cases} .$$

Now

$$Y_h = \text{ span}\{\omega_E : E \in \mathcal{E}_h\}.$$

Thus we may write

$$u_h = \sum_{K \in \mathcal{T}_h} U_K \chi_K \qquad q_h = \sum_{E \in \mathcal{E}_h} Q_E \omega_E, \tag{5.2}$$

with U_K giving the constant value of u_h on K, and $Q_E = \int_E q_h \cdot \nu_E$ the flux of q_h across E in the direction ν_E. Thus u_h is identified with the vector $\{U_K\}_{K \in \mathcal{T}_h} \in \mathrm{R}^{\mathrm{NT}}$ and q_h with $\{Q_E\}_{E \in \mathcal{E}_h} \in \mathrm{R}^{\mathrm{NE}}$. The source function f and the diffusion coefficient a will be assumed to be piecewise constant and thus can be written as

$$f = \sum_{K \in \mathcal{T}_h} f_K \chi_K \qquad a = \sum_{K \in \mathcal{T}_h} a_K \chi_K.$$

The mixed finite element method yields an approximation (u_h, q_h) of (u, q) satisfying

$$\begin{aligned} &(u_h, q_h) \in X_h \times Y_h \\ &\int_\Omega \frac{1}{a} q_h \, q_h' - \int_\Omega u_h \, \mathrm{div}\, q_h' = 0, \qquad \text{for } q_h' \in Y_h \\ &\int_\Omega \mathrm{div}\, q_h \, u_h' = \int_\Omega f \, u_h', \qquad \text{for } u_h' \in X_h. \end{aligned} \tag{5.3}$$

Defining the $NE \times NE$ matrix M and the $NT \times NE$ matrix D by

$$M_{E,E'} = \int_\Omega \frac{1}{a} \omega_E \, \omega_{E'} \qquad D_{K,E} = \int_\Omega \mathrm{div}\, \omega_E \, \chi_K, \tag{5.4}$$

and letting F denote the vector in R^{NT} with coordinates

$$F_K = f_K |K|,$$

we may write (5.3) as a linear system

$$\begin{aligned} MQ - D^T U &= 0 \\ DQ &= F. \end{aligned} \tag{5.5}$$

The matrix M is symmetric and positive definite. Thus we can use the first equation of (5.5) to express Q in terms of U, and plug this expression into the second equation of (5.5) to obtain the problem

$$DM^{-1}D^TU = F. \tag{5.6}$$

Remark 5.1 The (K, E) entry of the divergence matrix D is 0 unless E is an edge of K in which case $D_{K,E}$ is 1 if ν_E points outward from K and is -1 if ν_E points inward.

Remark 5.2 If $\mathcal{T}_h$ is a set of triangles, the symmetric matrix M has five nonzero entries in rows corresponding to interior edges and three for those corresponding to boundary edges. In case $\mathcal{T}_h$ is made up exclusively of rectangles, M is tridiagonal (with any reasonable ordering of the edges). Further, if we calculate the integrals used to define the matrix M by the numerical quadrature formula that approximates an integral over an element K by the average of the values at the vertices multiplied by the area of the element, then M becomes a diagonal matrix with $M_{E,E} = \dfrac{1}{2\,|E|}(\dfrac{|K|}{a_K} + \dfrac{|K'|}{a_{K'}})$, for E an edge of K and of K'.

Having described the discretization of (5.1), we turn now to the construction of the primal-dual formulation of the identification problem. The solution U of problem (5.6) is characterized as the solution of the minimization problem

$$\begin{aligned} &U \in X_h \\ &E_{ah}(U) = \inf_{U' \in X_h} E_{ah}(U'), \end{aligned} \tag{5.7}$$

where the primal energy functional $E_{ah} : X_h \longrightarrow \mathrm{R}$ is defined by

$$E_{ah}(U') = \tfrac{1}{2} M^{-1} D^T U' \cdot D^T U' - F \cdot U', \tag{5.8}$$

and the solution Q of problem (5.5) is characterized as the solution of the minimization problem

$$\begin{aligned} &Q \in Y_h \\ &E^*_{ah}(Q) = \inf_{Q' \in X_h, DQ = F} E^*_{ah}(Q'), \end{aligned} \tag{5.9}$$

where the dual energy functional $E^*_{ah} : Y_h \longrightarrow \mathrm{R}$ is defined by

$$E^*_{ah}(Q') = \tfrac{1}{2} MQ' \cdot Q'. \tag{5.10}$$

(Note that the U calculated in (5.5) is considered here simply as a multiplier used to solve the constrained problem (5.9) and does not appear in either energy functional.)

In the notation of §2 the space C_h of permissible parameters is R^{NT}, the state space X_h is R^{NT}, the space of dual states Y_h is R^{NE}, and

$$A_h U = D^T U,$$

$$F_{ah}(U) = -F \cdot U, \qquad G_{ah}(Q) = \tfrac{1}{2} M^{-1} Q \cdot Q.$$

Thus

$$A^*_h Q = -DQ,$$

$$F^*_{ah}(U) = \begin{cases} 0 & \text{if } U + F = 0 \\ +\infty & \text{if } U + F \neq 0 \end{cases}, \qquad G^*_{ah}(Q) = \tfrac{1}{2} MQ \cdot Q.$$

We remark that F_{ah} and G_{ah} are convex functions and that

$$Q \in \partial F(D^T U)$$

i.e.

$$Q = M^{-1}(D^T U).$$

Further, the Fenchel duality theorem guarantees that

$$\inf_{U \in \mathrm{R}^{\mathrm{NT}}} E_{ah}(U) + \inf_{Q \in \mathrm{R}^{\mathrm{NE}}, \mathrm{DQ=F}} E^*_{ah}(Q) = 0.$$

In the mixed formulations (5.2) and (5.3) the diffusion coefficient a appears only by means of its inverse $\dfrac{1}{a}$ and in (5.6) and (5.5) only by means of the matrix M which is defined in terms of $\dfrac{1}{a}$. We have thus, for the numerical experiments, chosen to identify not $a = \sum_{K \in \mathcal{T}_h} a_K \chi_K$ itself but $b = \dfrac{1}{a} = \displaystyle\sum_{K \in \mathcal{T}_h} b_K \chi_K$,

where $b_K = \dfrac{1}{a_K}$. The vector $B = \{b_K\}_{K\in\mathcal{T}_h}$ like $\{a_K\}_{K\in\mathcal{T}_h}$ is in R_{NT}.

Remark 5.3 Identifying the riciprocal of a is in fact suggested by the linear structure in which a and u appear in (5.2) and (5.1). In particular, for the case in which a is constant, u depends linearly on $\dfrac{1}{a}$ not on a.

To estimate the diffusion coefficient in (5.3) or (5.5) we must define the observation space Z_h and the observation operator from the state space $X_h = \mathrm{R}^{\mathrm{NT}}$ to Z_h. In the examples that we consider Z_h is taken to be $X_h = \mathrm{R}^{\mathrm{NT}}$ and the observation operator to be the identity. The space used for the regularization (denoted Q is §2 but not denoted Q_h here for obvious notational reasons) will be taken to be R^{NT} but with the semi-norm $|B|^2 = GB\cdot GB$, where G is the gradient matrix obtained from D^T by eliminating the rows corresponding to edges E contained in the boundary of Ω. Thus the regularized least squares functional $J_1 : C_h \times X_h \longrightarrow \mathrm{R}$ is

$$J_1(B,U) = \frac{1}{2}(Z-U)\cdot(Z-U) + \frac{\beta}{2}G(B-B^\sharp)\cdot G(B-B^\sharp),$$

and the total energy functional $J_2 : C_h \times X_h \times Y_h \longrightarrow \mathrm{R}$ is

$$J_2(B,U,Q) = \frac{1}{2}M^{-1}D^TU\cdot D^TU - F\cdot U + \frac{1}{2}MQ\cdot Q.$$

The discretized version of $(\mathcal{P}_\varepsilon)$ may now be given:

$$\begin{aligned}\min\Bigg\{&\frac{1}{2}(Z-U)\cdot(Z-U) + \frac{\beta}{2}G(B-B^\sharp)\cdot G(B-B^\sharp)\\ &+\frac{1}{\varepsilon}\left(\frac{1}{2}M^{-1}D^TU\cdot D^TU - F\cdot U + \frac{1}{2}MQ\cdot Q\right)\Bigg\}\\ &\text{over } (B,U,Q)\in C_h\times X_h\times Y_h,\ DQ = F.\end{aligned}\qquad(\mathcal{P}_{\varepsilon h})$$

We write J_ε for the functional to be minimized in $(\mathcal{P}_{\varepsilon h})$:

$$J_\varepsilon(B,U,Q) = J_1(B,U) + \frac{1}{\varepsilon}J_2(B,U,Q).$$

In all of the experiments that we consider in §6 $B^\sharp$ is chosen to be 0; thus, the regularization penalizes oscillations in B.

To use one of the splitting algorithms of §4 we note that to minimize J_ε with respect to B for U and Q fixed is to minimize

$$J_{\varepsilon B} = \frac{\beta}{2} G(B - B^\sharp) \cdot G(B - B^\sharp) + \frac{1}{\varepsilon} \left(\frac{1}{2} M^{-1} D^T U \cdot D^T U + \frac{1}{2} MQ \cdot Q \right)$$

over $B \in C_h$.

To minimize J_ε with respect to Q for B and U fixed is to minimize

$$J_{\varepsilon Q} = \frac{1}{2} MQ \cdot Q$$

over $Q \in Y_h, DQ = F$, which is equivalent to solving the dual problem (5.5). The linear system (5.5) may be solved by using a hybridization of the mixed method; see [9]. (Recall that the vector U produced by the resolution of (5.5) serves only as an auxiliary variable.)

Finally, to minimize J_ε with respect to U for B and Q fixed is to minimize

$$J_{\varepsilon U} = \frac{1}{2} (Z - U) \cdot (Z - U) + \frac{1}{\varepsilon} \left(\frac{1}{2} M^{-1} D^T U \cdot D^T U - F \cdot U \right)$$

over $U \in X_h$, which is equivalent to solving the following regularization of the primal problem (5.6):

$$[DM^{-1}D^T + \varepsilon I]U \quad = \quad F + \varepsilon Z \tag{5.11}$$

Thus in the discretized context, Algorithm 4.1, for example, becomes

Algorithm

i) Set $U_0 = Z \in X_h = \mathrm{R}^{\mathrm{NT}}$, choose $\mathrm{Q}_0 \in \mathrm{Y_h} = \mathrm{R}^{\mathrm{NE}}$, set $\mathrm{n} = 1$
ii) $B_n = \text{argmin } J_\varepsilon(B, U_{n-1}, Q_{n-1})$
iii) $Q_n =$ solution to (5.5) with M formed using B_n
iv) $U_n =$ solution to (5.11) with M formed using B_n
v) check convergence, stop or set $n = n + 1$ and go to (*ii*).

6 Numerical results

In the numerical examples presented here we identify the diffusion coefficient, or rather its reciprocal $b = \dfrac{1}{a}$ (see Remark 5.3), in the equation

$$\begin{aligned} -\nabla(a\nabla u) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{6.1}$$

In all of the experiments, Ω is taken to be the unit square in $\mathrm{R}^2 : \Omega = [0,1] \times [0,1]$, and the source function is $f = \Delta(x(1-x)y(1-y)\exp(x\,y)$. The same grid is used for the discretization of the reciprocal of the diffusion coefficient b, the state variable u, and the dual state variable q. We take a regular grid of squares of side length 0.05, so $\mathcal{T}_h$ contains 400 squares, $(NT = 400)$, and $\mathcal{E}_h$ contains 420 vertical edges and 420 horizontal edges, $(NE = 840)$.

In terms of the preceding section, §5, we solve the discretized minimization problem

$$\begin{aligned} \min \Bigg\{ &\frac{1}{2}(Z-U)\cdot(Z-U) + \frac{\beta}{2} G(B-B^\sharp)\cdot G(B-B^\sharp) \\ &+\frac{1}{\varepsilon}\left(\frac{1}{2}M^{-1}D^TU\cdot D^TU - F\cdot U + \frac{1}{2}MQ\cdot Q\right)\Bigg\} \qquad (\mathcal{P}_{\varepsilon h}) \\ \text{over } &(B,U,Q) \in R^{NT}\times R^{NT}\times R^{NE},\ DQ = F, \end{aligned}$$

where $U \in \mathrm{R}^{\mathrm{NT}}$ is the state variable, $Q \in \mathrm{R}^{\mathrm{NE}}$ is the dual state variable, $B \in \mathrm{R}^{\mathrm{NT}}$ is the reciprocal of the piecewise constant diffusion coefficient: $b_K = \dfrac{1}{a_K}$. The matrix D is the divergence matrix described in §5 and D^T is its transpose. The gradient matrix G used in the regularization term is also given in §5. As we have used a regular square grid, we use the numerical quadrature rule described in Remark 2.2 to obtain the matrix M. Thus M is a diagonal matrix with entry corresponding to the edge E between the rectangles K and K' equal to $|E|\,\dfrac{b_K + b_{K'}}{2}$ and with entry corresponding to the edge E of K on the boundary of Ω equal to $|E|\,b_K$. The vector $F \in \mathrm{R}^{\mathrm{NT}}$ has component corresponding to the element $K \in \mathcal{T}_h$, $\int_K f\,dxdy$. Here $B^\sharp$ is taken to be 0; the penalization parameter ε^2 is 10^{-1}.

What will change with the experiment are the observation Z, the amount of noise and the amount of regularization β. The observation Z will be determined as follows:

$$Z = U_{exact} + 2N\,|U_{max} - U_{min}|\,R,$$

where U_{exact} is the solution of (5.6) for F as given above and M computed using the sought diffusion coefficient B, the noise level N is taken to be either 0 or 0.1, U_{max} and U_{min} are the maximum and minimum values of U_{exact}, and $R \in \mathrm{R}^{\mathrm{NT}}$ is a uniform random distribution of numbers in $[-1, 1]$. The sought coefficient will be $B = \{b_K\}_{K \in |\mathcal{T}_h}$ given by $b_K = 2 - x_K + y_K$, where (x_K, y_K) is the coordinate of the center or K. We also consider the case $b_K = 1$ if $X_K < .5$; $b_K = 6$ if $X_K > .5$ We will show results of experiments with no regularization $\beta = 0$, a small amount of regularization, $\beta = 2.5 \times 10^{-5}$, for the case in which there is no noise and the sought B is affine, and larger amounts of regularization, $\beta = 10^{-2}$ or $\beta = 2.5 \times 10^{-1}$, for cases in which there is noise or the sought B is highly discontinuous.

The algorithm follows Algorithm 4.1 as well as that given at the end of §5.

Algorithm

- *initialization*
  - choose $B_\star$ arbitrarily
  - set $U_0 = Z$
  - set $Q_0 = \dfrac{1}{B_\star} D^T U_0$
  - calculate B_0 using a minimization routine to minimize with respect to B with $U = U_0$ and $Q = Q_0$ fixed.
  - set $n = 1$
- *main loop*
  - minimize with respect to B by using a minimization routine to obtain B_n
  - minimize with respect to U by solving (5.11) with matrix M calculated using B_n to obtain U_n

- minimize with respect to Q by solving (5.5) with matrix M calculated using B_n to obtain Q_n
- check convergence, stop or set $n = n + 1$ and continue.

For all of the numerical calculations we used the library SCILAB which is very convenient to use and produces an efficient code due to its ability to compile the entire code prior to beginning execution. The minimization routine used in the initialization and in the main loop to obtain B is the minimization routine **optim** of SCILAB. It is used with the quasi-Newton, low memory option **qnm**. The number of iterations of the minimization algorithm used before updating U and Q was fixed and in the results reported here was taken to be 1.

In the first experiment we seek to identify an affine coefficient function $b_K = 2 - x_K + y_K$. The observation is from noiseless data, $N = 0$, so that Z is U_{exact}. In Figure 1 we see four graphs of the coefficient function B. The two graphs on the left show B after convergence, after 30 iterations of the main loop. The upper one is without regularization and we see oscillation around the singular point near the center of Ω. For the lower one, the regularization coefficient is $\beta = 2.5 \times 10^{-5}$, and the oscillation has essentially disappeared. The graph on the lower right shows B_0, i.e. B after initialization, (with regularization). Thirty iterations were allowed in the minimization routine to obtain B_0. The graph on the upper right is obtained by a classical least squares method without regularization, starting from the same initial guess, $B_\star$ as in the algorithm above. Again 30 iterations were used to obtain the result. The calculation time used to obtain each of the three results on a Digital 3000/900 was of the order of one minute.

In the second experiment we have added noise. The coefficient B which we seek to identify is the same as in the preceeding experiment so that U_{exact} is the same as before but here the noise level N is taken to be 0.1. The two upper graphs in Figure 2 show B after convergence, the one on the left obtained without regularization, the one on the right with regularization coefficient $\beta = 10^{-2}$. The graph on the lower left shows the noisy observation Z while the graph on the lower right shows U, the pressure, calculated with the coefficient B shown on the upper right; i.e. U after convergence.

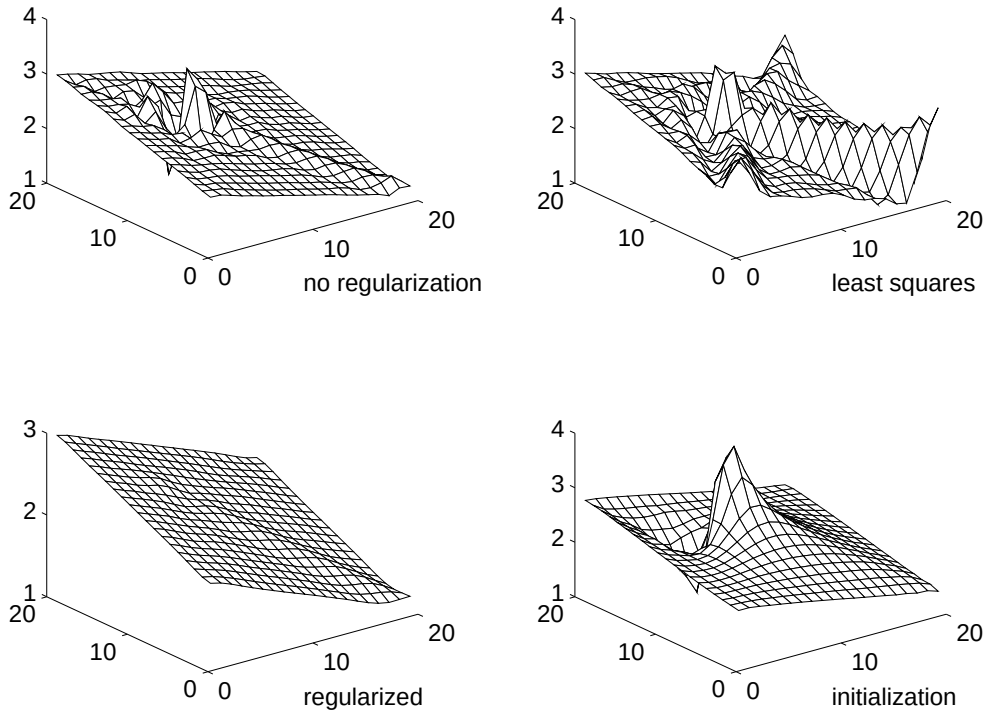

Figure 6.1: Identification of an affine coefficient function by the primal-dual method, with and without regularization, compared with identification by a least squares method.

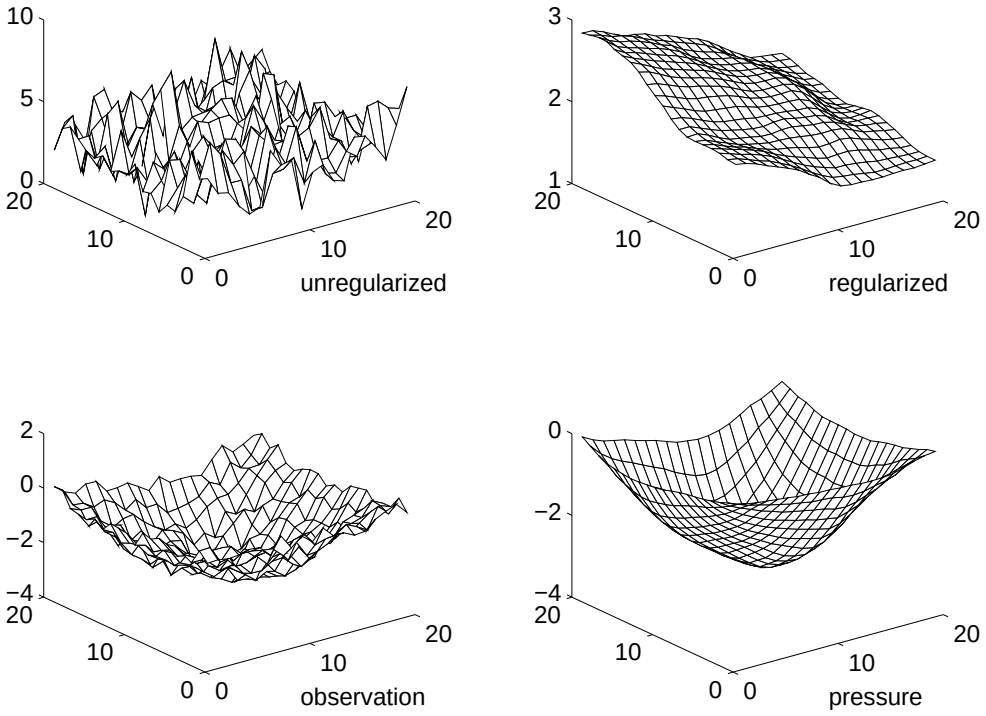

Figure 6.2: Identification with noisy data.

In the final experiment, we have tried to identify a strongly discontinuous function B, $b_K = 1$ if $X_K < .5$; $b_K = 6$ if $X_K > .5$. Both graphs show B after convergence. The graph on the left was obtained from an observation with no noise, $N = 0$, using a regularization coefficient, $\beta = 10^{-2}$. The right-hand side corresponds to a coefficient B obtained from noisy data, $N = 0.1$, but with more regularization; $\beta = 25 \times 10^{-2}$. We see the effect of the greater amount of regularization in the latter case in that the discontinuity is not represented as well as in the former.

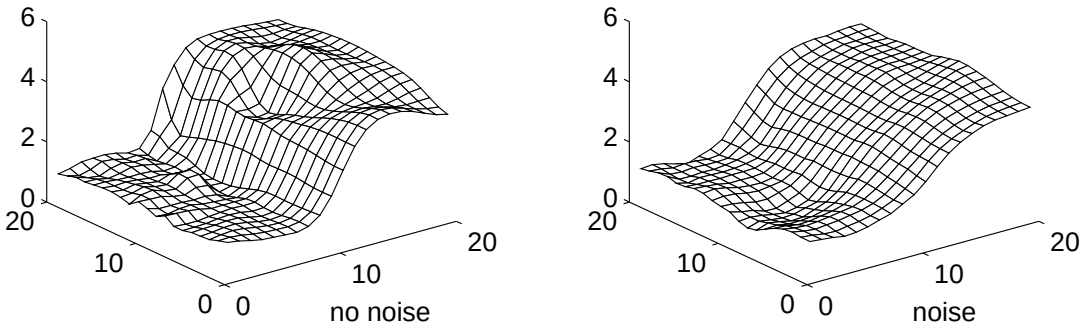

Figure 6.3: Identification of a discontinuous coefficient function.

In conclusion, the primal-dual method works well for the experiments we have carried out on a model problem and in general is superior to the least squares method. The fact that information from the observation can be exploited for initialization as well as for regularization makes the primal-dual method efficient and robust. In particular, it works well even with noisy data.

References

[1] R. Accar, *Identification of coefficients in elliptic equations*, SIAM Journal on Control and Optimization, 31 (1993), pp. 1221–1244.

[2] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, in Systems and Control, Foundations and Applications, Basel, ed., Birkhäuser, Boston, 1989.

[3] V. Barbu and K. Kunisch, *Identification of non linear elliptic equations*, Applied Mathematics and Optimization. to appear.

[4] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, Reidl Publishing, Dodrecht, 1986.

[5] J. Baumeister and W. Scondo, *Adaptive methods for parameter identification*, in Methoden und Verfahren der Mathematischen Physik Vol. 34, Verlag P. Lang, 1987, pp. 87–116.

[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer series in Computational Mathematics n° 15, Springer-Verlag New York Inc., New York, 1991.

[7] G. Chavent, *Identification de Coefficients Répartis dans les Equations aux Dérivées Partielles*, Thèse de Doctorat d'Etat, Faculté des Sciences de Paris, 1971.

[8] ——, *On the theory and practice of nonlinear least squares*, Advances in Water Resources, 14 (1991), pp. 55–63.

[9] G. Chavent and J. E. Roberts, *A unified physical presentation of mixed, mixed-hybrid finite elements and standard finite difference approximations for the determination of velocities in waterflow problems*, Advances in Water Ressources, 14 (1991), pp. 329–348. (Preprint in report INRIA n° 1107, Oct. 89).

[10] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnels*, Etudes Mathematiques, Dunod, Paris, 1974.

[11] K. Ito and K. Kunisch, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM Journal on Control and Optimization, 28 (1990), pp. 113–136.

[12] ——, *Sensitivity analysis to optimization problems in Hilbert spaces with application to optimal control and estimation*, J. Differential Equations, 99 (1992), pp. 1–40.

[13] R. Kohn and B. Lowe, *A variationnal method for parameters estimation*, RAIRO, M2AN, 22 (1988), pp. 119–158.

[14] K. Kunisch and X. Tai, *Sequential and parallel splitting methods for linear control problems in Hilbert spaces*, SIAM J. Num. Analysis. to appear.

[15] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., vol. 2 Finite Element Methods–Part 1, Elsevier Science Publishers B.V. (North–Holland), Amsterdam, 1991.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399