Worst case error bounds for the solution of uncertain Poisson equations with mixed boundary conditions

Tanveer Iqbal[1], Arnold Neumaier

Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Wien, Austria

Abstract

Given linear elliptic partial differential equations with mixed boundary conditions, with uncertain parameters constrained by inequalities, we show how to use finite element approximations to compute worst case a posteriori error bounds for linear response functionals determined by the solution. All discretization errors are taken into account.

Our bounds are based on the dual weighted residual (DWR) method of Becker & Rannacher [1], and treat the uncertainties with the optimization approach described in Neumaier [8].

We implemented the method for Poisson-like equations with an uncertain mass distribution and mixed Dirichlet/Neumann boundary conditions on arbitrary polygonal domains. To get the error bounds, we use a first order formulation whose solution with linear finite elements produces compatible piecewise linear approximations of the solution and its gradient. We need to solve nine related boundary value problems, from which we produce the bounds. No knowledge of domain-dependent a priori constants is necessary.

Keywords:
linear elliptic partial differential equation, dual weighted residual, uncertain parameters, global optimization.

1. Introduction

In practical applications, partial differential equations represent an approximate model of the real life situation. In many applications, partial differential equations depend on parameters which are only approximately known. Modeling errors can also be accounted for by adding parameters (constants or functions) to the model and specifying the uncertainty in these parameters. Each parameter then represents a particular scenario from the set of possibilities. In practice, one can solve the equation for a particular scenario or for just a few scenarios. But one is interested in how the solution varies over the full set of allowed scenarios.

For partial differential equations, one needs not only to consider the uncertainty due to parameters but also the errors introduced by discretization. In a traditional sensitivity analysis, one usually neglects the discretization errors, and ignores higher order terms in the sensitivity analysis. These two types of errors may however significantly affect the validity of the resulting bounds.

The work by Nakao & Plum [6, 7, 10] presents rigorous error bounds for linear elliptic equations using interval analysis. It is mathematically rigorous and also accounts for roundoff errors and errors in the numerical integrations. The parameter-dependent case is also studied by Plum [11] and Yamamoto et al. [15]. The methods apply to Dirichlet boundary conditions, compute error estimation in global norms like the energy norm or L^2 norm, and assume the knowledge of domain-dependent a priori constants for key inequalities used. These are known only for a few domains.

In many applications, the error in the global norm does not provide useful bounds for the errors in the quantities of real physical interest. Here work exists only in the nonparametric case (no uncertainties). Bertsimas & Caramanis [2] present a method based on semidefinite optimization to get bounds on linear functionals of the solutions of elliptic equations with Dirichlet boundary conditions. In the work by Repin [12], a posteriori estimates have been derived

Email addresses: iqbal.tanveer@univie.ac.at (Tanveer Iqbal), arnold.neumaier@univie.ac.at (Arnold Neumaier)

[1] Corresponding author.

with the help of duality theory from the calculus of variations. Work by the group of Peraire [9, 13, 14] used finite elements and a piecewise polynomial form of the coefficients to derive a posteriori error bounds for problems without uncertainty. Numerical integration errors and rounding errors are not taken into account.

No computable error bounds seem to be available in the case of mixed Dirichlet–Neumann boundary conditions treated in the present work.

Overview. In the present work, we discuss an approach that provides bounds on a linear response functional for a solution of mass-weighted Poisson equations with mixed boundary conditions on polygonal domains, with uncertain mass distribution. No a priori information is needed. We rigorously bound all discretization errors using new techniques, and bound the errors in the sensitivity analysis using the optimization approach outlined in Neumaier [8]. On the other hand, we shall assume that, compared to these errors, errors in the global optimization, errors in numerical integrations, and rounding errors can be neglected. For fully rigorous bounds, these would have to be taken into account, too.

We derive optimization-based error bounds for the discretization error, using a variant of the dual weighted residual (DWR) method by Becker & Rannacher [1]. For given uncertain parameters in the mass distribution m_θ, we compute the worst case error of a given linear response functional of the solution.

The first part of the present paper treats the problem in an abstract functional analytic setting. In Section 2, we discuss the spaces needed, and introduce the concept of a quasi-adjoint, used in Section 3 to derive abstract error bounds. Section 4 then discusses how we handle uncertainty in the differential equation.

The second part treats more specifically the mass-weighted Poisson equation. A first order formulation of the primal and adjoint equations is derived in Section 5. The θ-dependent operators M, N and E from the abstract theory are constructed for the mass-weighted Poisson equation in Section 6. The formulas for evaluating the dual norm of the residual of the adjoint problem are found in Section 7. Section 8 formulates an optimization problem whose solution defines suitable values of β, e, and $\mathbf{e}$ needed in the bounds.

The resulting algorithm was implemented in Matlab for the mass-weighted Poisson equation with mixed Dirichlet and Neumann boundary conditions on a polygonal domain in 2 dimensions. We illustrate the method with results for a particular example in Section 9.

2. Spaces and quasi-adjoint

Let U be a vector space and let V be a subspace of a Hilbert space $\overline{V}$. We write V^* for the dual space of V; thus $V \subseteq \overline{V} \subseteq V^*$. We write v^*v' for the inner product of $v, v' \in V$ and the bilinear pairing of $v \in V^*$ and $v' \in V$. Let $L\colon U \to \overline{V}$ be a linear operator mapping U into $\overline{V}$.

In the applications, U and V are spaces of locally differentiable functions. Thus V^* is a space of distributions obtained by differentiation of a square integrable function. L is composed of a first order differential operator and an associated boundary value mapping. We introduce the norm

$$\|v\|_V := \sqrt{v^*v} \qquad \text{for } v \in \overline{V} \tag{1}$$

in $\overline{V}$. If L is injective then

$$\|u\|_U := \|Lu\|_V \tag{2}$$

is a norm on U, and the completion $\overline{U}$ of U is a Hilbert space with inner product

$$\langle u, u'\rangle := (Lu)^*Lu'.$$

Clearly, $L\colon U \to \overline{V}$ can be completed to $L\colon \overline{U} \to \overline{V}$.

Proposition 1. *Suppose that*

$$\textit{for } v \in \overline{V},\ \ v^*L\tilde{u} = 0 \ \ \textit{for all } \tilde{u} \in \overline{U} \textit{ implies } v = 0. \tag{3}$$

Then for $f \in \overline{V}$, $Lu = f$ is solvable for some $u \in \overline{U}$.

Proof. For $f \in \overline{V}$, we define the linear functional mapping $\tilde{u} \in \overline{U}$ to $f^* L\tilde{u}$. By the representation theorem of Riesz [3], we find a $u \in \overline{U}$ such that

$$f^* L\tilde{u} = \langle u, \tilde{u} \rangle = (Lu)^* L\tilde{u} \qquad \text{for all } \tilde{u} \in \overline{U}.$$

Therefore,

$$(f - Lu)^* L\tilde{u} = 0.$$

If we use (3) with $v := f - Lu$, we get $v = 0$. Therefore, $Lu = f$. □

The assumed injectivity of L now implies that for given $f \in \overline{V}$ the operator equation

$$Lu = f, \tag{4}$$

is uniquely solvable for $u \in \overline{U}$. We are interested in bounding this solution u in terms of a computed approximation $\tilde{u} \in U$.

Under appropriate conditions, the construction of bounds for an approximate solution to (4) will be done with the help of suitable additional maps and the formulation of an adjoint problem. Let W be a Hilbert space and $w^* w'$ the inner product for $w, w' \in W$. The linear operator $L' \colon V \to W$ is called **quasi-adjoint** to L if

$$L = L'^* J, \tag{5}$$

for some linear embedding operator $J \colon U \to W$. In Section 4, the maps J and L' are specified for the mass-weighted Poisson equation with mixed Dirichlet-Neumann boundary conditions, rewritten as a system of first order PDEs.

To construct the bounds, we shall also need a linear operator $E \colon U \to V$ and symmetric linear operators $M : W \to W$, $N : W \to W$ such that M is positive semidefinite. A graphical representation is shown in Figure 1.

$$\begin{tikzcd}
W & & W \arrow[ll, "{M, N}"'] \arrow[dd, "L'^*"] \\
& U \arrow[ur, "J"] \arrow[dl, "E"'] \arrow[d, "L"'] & \\
V \arrow[uu, "L'"] \arrow[r, "\subset"'] & \overline{V} \arrow[r, "\subset"'] & V^*
\end{tikzcd}$$

Figure 1: Spaces and maps

Lemma 1. *If*

$$Mw = 0 \quad \Rightarrow \quad Nw = 0$$

and

$$\sup\left\{ \frac{w^* N w}{w^* M w} \;\middle|\; Mw \neq 0 \right\} \leq \beta^* < \infty, \tag{6}$$

then

$$M_\beta := \beta M - N \tag{7}$$

is symmetric and positive semidefinite for all $\beta > \beta^*$, and $w^* M_\beta w > 0$ if $M_\beta w \neq 0$.

Proof. As M is positive semidefinite, we have $w^*Mw = 0$ only if $Mw = 0$. For w with $Mw \neq 0$, we have $w^*Mw > 0$, and by (6), we get

$$\frac{w^*Nw}{w^*Mw} \leq \beta^* < \beta,$$

which implies

$$w^*M_\beta w = w^*(\beta M - N)w > 0 \qquad \text{if } Mw \neq 0.$$

If $Mw = 0$ then $Nw = 0$, so $M_\beta w = 0$ and $w^*(\beta M - N)w = 0$. Thus $\beta M - N$ is positive semidefinite. □

We define for $\beta > \beta^*$ a seminorm $\|\cdot\|_\beta$ on W by

$$\|w\|_\beta := \sqrt{w^*M_\beta w} \qquad \text{if } w \in W. \tag{8}$$

The inverse M_β^{-1} is only defined on the range

$$W_\beta := \text{Range } M_\beta$$

up to an element in the kernel of M_β, but the value $w^*M_\beta^{-1}w$ is independent of the choice of which solution w_β of $M_\beta w_\beta = w$ is taken as $M_\beta^{-1}w$. Indeed, if w'_β, w''_β are any two solutions of $M_\beta w_\beta = w$, then, since M_β is symmetric,

$$\begin{aligned} w^*w'_\beta - w^*w''_\beta &= (M_\beta w_\beta)^*w'_\beta - (M_\beta w_\beta)^*w''_\beta = w_\beta^*(M_\beta^*w'_\beta - M_\beta^*w''_\beta) \\ &= w_\beta^*(M_\beta w'_\beta - M_\beta w''_\beta) = w_\beta^*(w - w) = 0, \end{aligned}$$

which implies $w^*w'_\beta = w^*w''_\beta$. So, we can define a dual norm

$$\|w\|_{W_\beta} = \sqrt{w^*M_\beta^{-1}w} \qquad \text{for } w \in W_\beta. \tag{9}$$

3. Abstract error bounds from a quasi-adjoint

We are interested in bounding the solution $u \in \overline{U}$ of equation (4) for given $f \in \overline{V}$ in terms of a computed approximation $\tilde{u} \in U$. Theorem 2 below ensures that the error $J(u - \tilde{u})$ can be bounded in terms of the norm of the residual of (4).

Theorem 2. *Let $E\colon U \to V$ and $J\colon U \to W$ be operators such that, for $u \in U$*

$$\begin{aligned} 2(Eu)^*Lu &= (Ju)^*MJu, & (10) \\ (Eu)^*Eu &= (Ju)^*NJu. & (11) \end{aligned}$$

Then J can be extended to a map from $\overline{U}$ to W such that

$$\|Ju\|_\beta \leq \beta\,\|Lu\|_V \qquad \text{for } u \in \overline{U}. \tag{12}$$

Equality holds when

$$Lu = \beta^{-1}Eu. \tag{13}$$

Proof. For $u \in U$, equations (7), (10) and (11) imply

$$\begin{aligned} \|Ju\|_\beta^2 &= (Ju)^*(\beta M - N)(Ju) = 2\beta(Eu)^*Lu - (Eu)^*Eu \\ &= \beta^2(Lu)^*Lu - \Big((Eu)^*Eu - 2\beta(Eu)^*Lu + \beta^2(Lu)^*Lu\Big) \\ &= \beta^2\,\|Lu\|_V^2 - \|Eu - \beta Lu\|_V^2\,, \end{aligned}$$

which implies (12) for $u \in U$. Let u_l be a sequence of functions in U with

$$\lim_{l\to\infty} u_l \to u \in \overline{U}.$$

Then

$$\|Ju_l - Ju_m\|_\beta = \|J(u_l - u_m)\|_\beta \le \beta\,\|L(u_l - u_m)\|_V\,,$$

which implies by (2) that

$$\|Ju_l - Ju_m\|_\beta \le \beta\,\|u_l - u_m\|_U \to 0 \quad \text{if } l, m \to \infty.$$

Therefore, $\|Ju_l - Ju_m\|_\beta \to 0$ as $l, m \to \infty$. Therefore, the Ju_l form a Cauchy sequence in W and the limit $Ju := \lim\limits_{l\to\infty} Ju_l$ exists. Clearly (12) also holds for $u \in \overline{U}$. Equality holds if $Eu - \beta Lu = 0$, i.e., if (13) holds. □

For an operator L associated to a first order formulation of the mass-weighted Poisson equation, we show in Section 6 how to construct M, N, and E satisfying the conditions (10), (11).

From Theorem 2, we get the computable first order error estimate

$$\|Ju - J\tilde{u}\|_\beta \le \beta\,\|f - L\tilde{u}\|_V\,. \tag{14}$$

Corollary 1. *If $Ju \in W$, $w \in W_\beta$ then*

$$|w^* Ju| \le \beta\,\|w\|_{W_\beta}\,\|Lu\|_V\,. \tag{15}$$

Proof. Let λ be some scalar, then

$$\begin{aligned} 0 \le \left\|w - \lambda M_\beta Ju\right\|^2_{W_\beta} &= (w - \lambda M_\beta Ju)^* M_\beta^{-1}(w - \lambda M_\beta Ju) \\ &= w^* M_\beta^{-1} w - 2\lambda w^* Ju + \lambda^2 (Ju)^* M_\beta Ju. \end{aligned}$$

If $Ju = 0$, (15) is trivial. If $Ju \neq 0$, then $(Ju)^* M_\beta Ju \neq 0$, and by choosing $\lambda = w^* Ju/(Ju)^* M_\beta Ju$, we get $|w^* Ju|^2 \le (w^* M_\beta^{-1} w)(Ju)^* M_\beta Ju = \|w\|^2_{W_\beta}\,\|Ju\|^2_\beta$, which implies $|w^* Ju| \le \|w\|_{W_\beta}\,\|Ju\|_\beta$. By using (12), we get (15). □

Often one is not interested in the accuracy of the complete solution but just the accuracy of some important response functional $R(u)$ of the solution. We consider a linear response functional of the form

$$R(u) = g^* Ju, \tag{16}$$

for some $g \in W_\beta$. The first order bounds

$$|R(u) - R(\tilde{u})| = |g^*(Ju - J\tilde{u})| \le \beta \|g\|_{W_\beta}\,\|f - L\tilde{u}\|_{\overline{V}}$$

for $R(u)$ derivable from the corollary when $g \in W_\beta$ are not very accurate. The main result of this section are better second order bounds obtained (without the restriction $g \in W_\beta$) by bounding the error in the linear functional $g^* Ju$ of the solution in terms of the dual norms:

Theorem 3. *Let $\tilde{u} \in U$ be an approximation to a solution $u \in \overline{U}$ of $Lu = f$ and $\tilde{v} \in V$ be an approximation to a solution v of $L'v = g$, where L' is quasi-adjoint to L. Let*

$$r := f - L\tilde{u} \in \overline{V}, \qquad s := g - L'\tilde{v} \in W_\beta, \tag{17}$$

$$\gamma = \tilde{v}^* r + g^* J\tilde{u}, \qquad \delta = \beta\,\|s\|_{W_\beta}\,\|r\|_V\,. \tag{18}$$

Then

$$\|J(u - \tilde{u})\|_\beta \le \beta\,\|r\|_V\,, \tag{19}$$

$$g^* Ju \in [\gamma - \delta, \gamma + \delta]. \tag{20}$$

Proof. (12) implies (19). We compute

$$\begin{aligned} g^*(Ju - J\tilde{u}) &= (s + L'\tilde{v})^* J(u - \tilde{u}) = (s^* + \tilde{v}^* L'^*) J(u - \tilde{u}) \\ &= s^* J(u - \tilde{u}) + \tilde{v}^* L'^* J(u - \tilde{u}), \end{aligned}$$

as L' is quasi-adjoint to L, therefore, $g^*(Ju - J\tilde{u}) = s^* J(u - \tilde{u}) + \tilde{v}^* L(u - \tilde{u})$. Since $L(u - \tilde{u}) = Lu - L\tilde{u} = f - L\tilde{u} = r$, we get $g^*(Ju - J\tilde{u}) = s^* J(u - \tilde{u}) + \tilde{v}^* r$. By using (15), we have

$$|g^* Ju - g^* J\tilde{u} - \tilde{v}^* r| = |s^* J(u - \tilde{u})| \le \beta\,\|s\|_{W_\beta}\,\|L(u - \tilde{u})\|_{\overline{V}}\,,$$

and we get (20) from $|g^* Ju - \gamma| \le \beta\,\|s\|_{W_\beta}\,\|r\|_{\overline{V}} = \delta$. □

The dual norms needed to evaluate the bounds are defined by (1) and (9). The computation of (9) in concrete cases is the dominant effort for the application of Theorem 3, since it requires to define M_β and hence β that satisfy all requirements.

4. Equations with uncertain parameters

We now discuss how to treat uncertain partial differential equations on the abstract level.

We assume that the uncertainty is specified by a condition $\theta \in \Theta$, where Θ is the region of uncertainty. We use optimization techniques to get bounds that cover the errors in solving the differential equation and the errors caused by the uncertainty in the parameters of the Poisson equation. The problem to be solved is assumed to be parametrized by the parameter θ and will be of the form

$$L(\theta)u = f(\theta), \qquad \text{for some } \theta \in \Theta. \tag{21}$$

Therefore, the solution $u = u(\theta)$, the computed approximations and the bounds derived in Section 3, will depend on θ. Therefore, the error bounds (20) take the form

$$g(\theta)^* Ju(\theta) \in [\gamma(\theta) - \delta(\theta), \gamma(\theta) + \delta(\theta)],$$

where

$$\gamma(\theta) = \tilde{v}(\theta)^* r(\theta) + g^*(\theta)Ju(\theta), \qquad \delta(\theta) = \beta \, \|s(\theta)\|_{W_\beta} \, \|r(\theta)\|_V ,$$

and $r(\theta) = f(\theta) - L(\theta)u(\theta)$, $s(\theta) = g(\theta) - L'(\theta)v(\theta)$. Since there are infinitely many scenarios in the uncertainty region Θ, we need some extra preparation to work with a finite and controlled amount. An initial approximation $\tilde{u} = \tilde{u}(\theta)$ for $u = u(\theta)$ is computed by interpolation from approximate solutions of (21) for a small number of scenarios $\theta_l \in \Theta$.

Now we define

$$\begin{aligned} \underline{e} &:= \min\{\gamma(\theta) - \delta(\theta) \mid \theta \in \Theta\}, \\ \overline{e} &:= \max\{\gamma(\theta) + \delta(\theta) \mid \theta \in \Theta\}, \end{aligned}$$

and find the desired rigorous error bounds

$$g(\theta)^* Ju(\theta) \in [\underline{e}, \overline{e}] \qquad \text{for all } \theta \in \Theta.$$

The computation of $\underline{e}$ and $\overline{e}$ is a global optimization problem which is computationally tractable if θ is not too high-dimensional. The optimization problem can be stated specifically by using an algebraic modeling language such as AMPL [4] and can be solved with a variety of solvers.

5. The mass-weighted Poisson equation

Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain with Lipschitz boundary Γ. Let $\Gamma_D \neq \emptyset$ be that part of the boundary where Dirichlet boundary conditions are imposed while we have Neumann boundary conditions on the remaining part $\Gamma_N := \Gamma \setminus \Gamma_D$ of the boundary. In particular, vertices at which the Dirichlet conditions are imposed do not belong to Γ_N. We consider the mass-weighted Poisson equation

$$-\nabla \cdot (m_\theta \nabla u_0) = f_0 - \nabla \cdot \mathbf{f}, \tag{22}$$

where $0 < m_\theta \in L^\infty(\overline{\Omega})$ is a scalar positive function, $f_0 \in L_2(\Omega)$ and $\mathbf{f} \in L_2(\Omega)^d$, with Dirichlet and Neumann boundary conditions

$$u_0|_{\Gamma_D} = 0 \ \text{ on } \Gamma_D, \qquad \mathbf{n} \cdot \nabla u_0|_{\Gamma_N} = f_N \ \text{ on } \Gamma_N. \tag{23}$$

We use the notation

$$\int f \mathrm{d}\Omega = \int_\Omega f(x)\mathrm{d}x, \quad \int f \mathrm{d}\Gamma = \int_\Gamma f(x)\mathrm{d}x,$$

and extend functions h defined only on Γ_N to Γ by setting $h(x) = 0$ for $x \in \Gamma_D$.

Then the right hand side of (22) is in the dual space $H^{-1}(\Omega)$ of the Sobolev space $H^1(\Omega)$ and the solution u_0 is in $H^1(\Omega)$.

Definition 1.

(*i*) We define the spaces

$$\begin{aligned}
U &:= \left\{u = \begin{pmatrix} u_0 \\ \mathbf{u} \end{pmatrix} \in C^{0,1}(\overline{\Omega}) \times C^{0,1}(\overline{\Omega})^d \;\middle|\; u_0|_{\Gamma_D} = 0\right\}, \\
V &:= C^{0,1}(\overline{\Omega}) \times C^{0,1}(\overline{\Omega})^d \times L_2(\Gamma_N), \\
\overline{V} &:= L_2(\Omega) \times L_2(\Omega)^d \times L_2(\Gamma_N), \\
W &:= L_2(\Omega) \times L_2(\Omega)^d \times L_2(\Gamma_N) \times L_2(\Gamma).
\end{aligned}$$

(*ii*) We define the inner product of

$$v = \begin{pmatrix} v_0 \\ \mathbf{v} \\ v_1 \end{pmatrix} \in \overline{V}, \qquad v' = \begin{pmatrix} v'_0 \\ \mathbf{v}' \\ v'_1 \end{pmatrix} \in \overline{V},$$

$$w = \begin{pmatrix} w_0 \\ \mathbf{w} \\ w_1 \\ w_2 \end{pmatrix} \in W, \qquad w' = \begin{pmatrix} w'_0 \\ \mathbf{w}' \\ w'_1 \\ w'_2 \end{pmatrix} \in W$$

by

$$\begin{aligned}
v^* v' &:= \int (v_0 v'_0 + \mathbf{v} \cdot \mathbf{v}') \mathrm{d}\Omega + \int v_1 v'_1 \mathrm{d}\Gamma_N, \\
w^* w' &:= \int (w_0 w'_0 + \mathbf{w} \cdot \mathbf{w}') \mathrm{d}\Omega + \int (w_1 w'_1 + w_2 w'_2) \mathrm{d}\Gamma.
\end{aligned}$$

(*iii*) We define the linear differential operators $L(\theta)\colon U \to \overline{V}$

$$L(\theta)u = \begin{pmatrix} \nabla \cdot \mathbf{u}|_\Omega \\ (m_\theta \nabla u_0 + \mathbf{u})|_\Omega \\ \mathbf{n} \cdot \mathbf{u}|_{\Gamma_N} \end{pmatrix} \qquad \text{for } u = \begin{pmatrix} u_0 \\ \mathbf{u} \end{pmatrix} \in U. \tag{24}$$

Proposition 2. *$L(\theta)$ can be extended to an injective operator $L(\theta)\colon \overline{U} \to \overline{V}$.*

Proof. By (24) if $u \in U$ and $L(\theta)u = 0$ then

$$\begin{aligned}
\nabla \cdot \mathbf{u} &= 0 \qquad \text{in } \Omega, && (25) \\
\mathbf{u} + m_\theta \nabla u_0 &= 0 \qquad \text{in } \Omega, && (26) \\
\mathbf{n} \cdot \mathbf{u} &= 0 \qquad \text{on } \Gamma_N, && (27)
\end{aligned}$$

(25), (26) imply

$$\nabla \cdot (m_\theta \nabla u_0) = 0 \qquad \text{in } \Omega, \tag{28}$$

and (26), (27) imply

$$\mathbf{n} \cdot \nabla u_0 = 0 \qquad \text{on } \Gamma_N. \tag{29}$$

(28) gives

$$\int u_0 \nabla \cdot (m_\theta \nabla u_0) \mathrm{d}\Omega = 0,$$

therefore

$$\int m_\theta (\nabla u_0)^2 \mathrm{d}\Omega = \int u_0 (\mathbf{n} \cdot \nabla u_0) m_\theta \mathrm{d}\Gamma_D + \int u_0 (\mathbf{n} \cdot \nabla u_0) m_\theta \mathrm{d}\Gamma_N. \tag{30}$$

By using $u_0|_{\Gamma_D} = 0$ and (29), we conclude that

$$\int m_\theta (\nabla u_0)^2 \mathrm{d}\Omega = 0,$$

which implies $\nabla u_0 = 0$ in $\overline{\Omega}$. Therefore, u_0 is constant in $\overline{\Omega}$. Since, $u_0|_{\Gamma_D} = 0$, the constant vanishes; therefore $u_0 = 0$. Now (26) implies $\mathbf{u} = 0$.

Thus $L(\theta)$ is injective on U, and the assumptions of the abstract theory are satisfied. Thus we can extend $L(\theta)$ to an operator $L(\theta)\colon \overline{U} \to \overline{V}$. Repetition of the argument now shows that $L(\theta)$ is injective on $\overline{U}$. □

Proposition 3. *For the right hand side*

$$f = \begin{pmatrix} f_0 \\ \mathbf{f} \\ f_1 \end{pmatrix} \in \overline{V} \ \textit{with}\ f_1 = \mathbf{n} \cdot \mathbf{f} - m_\theta f_N,$$

the primal equation

$$L(\theta)u = f, \qquad u = \begin{pmatrix} u_0 \\ \mathbf{u} \end{pmatrix} \in \overline{U},$$

is equivalent to the Poisson equation (22) with Dirichlet and Neumann boundary conditions (23) together with

$$\mathbf{u} = \mathbf{f} - m_\theta \nabla u_0 \qquad \textit{in}\ \Omega. \tag{31}$$

Proof. By (24), $L(\theta)u = f$ implies

$$\begin{aligned} \nabla \cdot \mathbf{u} &= f_0 && \text{in } \Omega, && (32)\\ m_\theta \nabla u_0 + \mathbf{u} &= \mathbf{f} && \text{in } \Omega, && (33)\\ \mathbf{n} \cdot \mathbf{u} &= f_1 && \text{on } \Gamma_N, && (34) \end{aligned}$$

and $u = 0$ on Γ_D by definition of U. (33) implies (31), inserting this into (32) gives the Poisson equation (22). Inserting (31) into (34) gives $\mathbf{n} \cdot (\mathbf{f} - m_\theta \nabla u_0) = f_1$, which implies $\mathbf{n} \cdot (\mathbf{f} - m_\theta \nabla u_0)|_{\Gamma_N} = f_1|_{\Gamma_N} = \mathbf{n} \cdot \mathbf{f}|_{\Gamma_N} - m_\theta f_N|_{\Gamma_N}$ on Γ_N, hence (23). □

The solution will be constructed in the Hilbert space $\overline{U}$. By using the theory of Section 2, we first construct the quasi-adjoint.

Proposition 4. *Define the embedding operator $J\colon U \to W$ by*

$$Ju := \begin{pmatrix} u_0|_\Omega \\ \mathbf{u}|_\Omega \\ u_0|_{\Gamma_N} \\ \mathbf{n} \cdot \mathbf{u}|_\Gamma \end{pmatrix} \in W \qquad \textit{for}\ u = \begin{pmatrix} u_0 \\ \mathbf{u} \end{pmatrix} \in U. \tag{35}$$

Then the mapping $L'(\theta)\colon V \to W$ defined by

$$L'(\theta)v = \begin{pmatrix} -\nabla \cdot m_\theta \mathbf{v}|_\Omega \\ (\mathbf{v} - \nabla v_0)|_\Omega \\ m_\theta \mathbf{n} \cdot \mathbf{v}|_{\Gamma_N} \\ v_1 + v_0|_\Gamma \end{pmatrix} \qquad \textit{for}\ v = \begin{pmatrix} v_0 \\ \mathbf{v} \\ v_1 \end{pmatrix} \in V \tag{36}$$

is a quasi-adjoint of $L(\theta)$:

$$L(\theta) = L'(\theta)^* J.$$

Moreover, for $v \in \overline{V}$, $v^* L(\theta)u = 0$ for all $u \in U$ implies $v = 0$.

Proof. (*i*) For $v \in V$, $u \in U$, we compute

$$\begin{aligned} v^*(L(\theta)u) &= \begin{pmatrix} v_0 \\ \mathbf{v} \\ v_1 \end{pmatrix}^* \begin{pmatrix} \nabla \cdot \mathbf{u}|_\Omega \\ (m_\theta \nabla u_0 + \mathbf{u})|_\Omega \\ \mathbf{n} \cdot \mathbf{u}|_{\Gamma_N} \end{pmatrix} \\ &= \int \Big(v_0 \nabla \cdot \mathbf{u} + \mathbf{v} \cdot (m_\theta \nabla u_0 + \mathbf{u}) \Big) d\Omega + \int v_1 \mathbf{n} \cdot \mathbf{u}\, d\Gamma_N \\ &= I_\Omega + I_\Gamma. \end{aligned} \tag{37}$$

Using integration by parts, we rewrite I_Ω as

$$\begin{aligned} I_\Omega &= \int \Big(v_0 \nabla \cdot \mathbf{u} + \mathbf{v} \cdot (m_\theta \nabla u_0 + \mathbf{u}) \Big) \mathrm{d}\Omega \\ &= \int \Big(-(\nabla \cdot m_\theta \mathbf{v}) u_0 + (\mathbf{v} - \nabla v_0) \cdot \mathbf{u} \Big) \mathrm{d}\Omega + \int \Big(m_\theta u_0 \mathbf{n} \cdot \mathbf{v} + v_0 \mathbf{n} \cdot \mathbf{u} \Big) \mathrm{d}\Gamma. \end{aligned}$$

Since $u_0 = 0$ on Γ_D, we have

$$I_\Omega = \int \Big(-(\nabla \cdot m_\theta \mathbf{v}) u_0 + (\mathbf{v} - \nabla v_0) \cdot \mathbf{u} \Big) \mathrm{d}\Omega + \int m_\theta u_0 \mathbf{n} \cdot \mathbf{v} \, \mathrm{d}\Gamma_N + \int v_0 \mathbf{n} \cdot \mathbf{u} \, \mathrm{d}\Gamma.$$

Since $v_1|_{\Gamma_D} = 0$, (37) becomes

$$\begin{aligned} v^*(L(\theta)u) &= I_\Omega + I_\Gamma \\ &= \int \Big(-(\nabla \cdot m_\theta \mathbf{v}) u_0 + (\mathbf{v} - \nabla v_0) \cdot \mathbf{u} \Big) \mathrm{d}\Omega \\ &\quad + \int \Big((m_\theta \mathbf{n} \cdot \mathbf{v}) u_0 + (v_0 + v_1) \mathbf{n} \cdot \mathbf{u} \Big) \mathrm{d}\Gamma, \end{aligned}$$

hence

$$v^*(L(\theta)u) = (L'(\theta)v)^* J u. \tag{38}$$

From this we get $v^* L(\theta) u = (L'(\theta)v)^*(Ju) = v^*(L'(\theta)^* J)u$, which implies $L(\theta) = L'(\theta)^* J$.

(*ii*) Suppose $v \in \overline{V}$ satisfies $v^*(L(\theta)u) = 0$ for all $u \in U$. There is a sequence $v_l \in V$, $l = 1, 2, \ldots$, with twice continuously differentiable components, converging to v. Then (38) implies

$$(L'(\theta)v_l)^* J u = v_l^*(L(\theta)u) \to 0 \qquad \text{for all } u \in U.$$

Using (35) and (36) we find

$$\begin{aligned} \mathbf{v}_l - \nabla v_{l0} &\to 0 \qquad \text{in } \Omega, \qquad (39)\\ \nabla \cdot (m_\theta \mathbf{v}_l) &\to 0 \qquad \text{in } \Omega, \qquad (40)\\ v_{l0} + v_{l1} &\to 0 \qquad \text{on } \Gamma, \qquad (41)\\ \mathbf{n} \cdot \mathbf{v}_l &\to 0 \qquad \text{on } \Gamma_N. \qquad (42) \end{aligned}$$

From (39), (42), we get

$$\mathbf{n} \cdot \nabla v_{l0} \to 0 \qquad \text{on } \Gamma_N. \tag{43}$$

From (39), (40), we get

$$\nabla \cdot (m_\theta \nabla v_{l0}) \to 0 \qquad \text{in } \Omega,$$

which implies

$$\int v_{l0} \nabla \cdot (m_\theta \nabla v_{l0}) \mathrm{d}\Omega \to 0. \tag{44}$$

By (44), we get

$$\int m_\theta (\nabla v_{l0})^2 \mathrm{d}\Omega - \int m_\theta v_{l0} \mathbf{n} \cdot \nabla v_{l0} \, \mathrm{d}\Gamma_D - \int m_\theta v_{l0} \mathbf{n} \cdot \nabla v_{l0} \, \mathrm{d}\Gamma_N \to 0. \tag{45}$$

As $v_{l1}|_{\Gamma_D} = 0$, therefore, by (41)

$$v_{l0} \to 0 \qquad \text{on } \Gamma_D. \tag{46}$$

Now by using (43), (45) implies

$$\int m_\theta (\nabla v_{l0})^2 \mathrm{d}\Omega \to 0,$$

which implies $\nabla v_{l0} \to 0$ in $\overline{\Omega}$. Therefore, v_{l0} is converging to a constant in $\overline{\Omega}$ and by (46) this constant vanishes. Now (41) implies $v_{l1} \to 0$ on Γ_N and (39) implies $\mathbf{v}_l \to 0$ in $\overline{\Omega}$. Therefore, $v = \lim_{l\to\infty} v_l = 0$. □

The abstract theory now implies

Corollary 2. *For $f \in \overline{V}$, the equation $L(\theta)u = f$ is uniquely solvable for some $u \in \overline{U}$.*

Proposition 5. *For the right hand side*

$$g = \begin{pmatrix} g_0 \\ \mathbf{g} \\ g_1 \\ g_2 \end{pmatrix} \in W, \tag{47}$$

the dual equation

$$L'(\theta)v = g, \qquad v = \begin{pmatrix} v_0 \\ \mathbf{v} \\ v_1 \end{pmatrix} \in V,$$

is equivalent to the Poisson equation

$$-\nabla \cdot (m_\theta \nabla v_0) = g_0 + \nabla \cdot (m_\theta \mathbf{g}), \tag{48}$$

with Dirichlet and Neumann boundary conditions

$$v_0|_{\Gamma_D} = g_2 \ \text{ on } \Gamma_D, \qquad \mathbf{n} \cdot \nabla v_0|_{\Gamma_N} = m_\theta^{-1} g_1 - \mathbf{n} \cdot \mathbf{g} \ \text{ on } \Gamma_N, \tag{49}$$

together with

$$\begin{aligned} \mathbf{v} &= \mathbf{g} + \nabla v_0 && \text{in } \Omega, && (50) \\ v_1 &= g_2 - v_0 && \text{on } \Gamma_N. && (51) \end{aligned}$$

Proof. By (36), $L'(\theta)v = g$ implies

$$\begin{aligned} -\nabla \cdot (m_\theta \mathbf{v}) &= g_0 && \text{in } \Omega, && (52) \\ \mathbf{v} - \nabla v_0 &= \mathbf{g} && \text{in } \Omega, && (53) \\ m_\theta \mathbf{n} \cdot \mathbf{v} &= g_1 && \text{on } \Gamma_N, && (54) \\ v_0 + v_1 &= g_2 && \text{on } \Gamma. && (55) \end{aligned}$$

(53) implies (50), inserting this into (52) gives the Poisson equation (48). Inserting (50) into (54) gives $m_\theta \mathbf{n} \cdot (\mathbf{g} + \nabla v_0) = g_1$ on Γ_N, which implies $\mathbf{n} \cdot \nabla v_0|_{\Gamma_N} = m_\theta^{-1} g_1 - \mathbf{n} \cdot \mathbf{g}$ on Γ_N. Finally, (55) is equivalent to (51) together with $v_0|_{\Gamma_D} = g_2$. □

Using (47) and (35) in equation (16) gives

$$R(u) = \int (g_0 u_0 + \mathbf{g} \cdot \mathbf{u}) \mathrm{d}\Omega + \int (g_1 u_0 + g_2 \mathbf{n} \cdot \mathbf{u}) \mathrm{d}\Gamma. \tag{56}$$

The functions g_0, $\mathbf{g}$, g_1, g_2 must be chosen such that (56) defines the response of interest.

6. Defining $M(\theta)$, $N(\theta)$ and $E(\theta)$

In this section, we define the operators $M(\theta)$, $N(\theta)$ and $E(\theta)$ for the Poisson equation with $L(\theta)$ defined by (24). The conditions (10), (11) from Theorem 2 will be used for their construction. The proofs of the following propositions are omitted; they are straightforward consequences of the definitions, the linearity of the integral and integration by parts.

Proposition 6. *Let $e, \mathbf{e} \in C^{0,1}(\Omega)$ and $c_0, c \in L^\infty(\Gamma_N)$. For $L(\theta)$ defined in (24) and the operator $E(\theta)\colon U \to V$ defined by*

$$E(\theta)u = \begin{pmatrix} m_\theta e u_0|_\Omega \\ (\mathbf{e}u_0 + e\mathbf{u})|_\Omega \\ (c_0 u_0 + c\mathbf{n}\cdot\mathbf{u})|_{\Gamma_N} \end{pmatrix} \in V \qquad \textit{for } u = \begin{pmatrix} u_0 \\ \mathbf{u} \end{pmatrix} \in U, \tag{57}$$

we have

$$(E(\theta)u)^*(L(\theta)u) = I_\Omega + I_\Gamma,$$

where

$$I_\Omega \;=\; \int \Big((\mathbf{e} - \nabla m_\theta e)\cdot \mathbf{u}u_0 + e\mathbf{u}^2 - \frac{1}{2}(\nabla \cdot m_\theta \mathbf{e})u_0^2\Big)\mathrm{d}\Omega, \tag{58}$$

$$I_\Gamma \;=\; \int \Big((m_\theta e + c_0)(\mathbf{n}\cdot\mathbf{u})u_0 + \frac{1}{2}m_\theta(\mathbf{n}\cdot\mathbf{e})u_0^2 + c(\mathbf{n}\cdot\mathbf{u})^2\Big)\mathrm{d}\Gamma_N. \tag{59}$$

Proposition 7. *For $L(\theta)$ and $E(\theta)$ defined in and (24) and J defined in (35), we have*

$$2(E(\theta)u)^*(L(\theta)u) = (Ju)^* M(\theta)(Ju),$$

where $M(\theta)\colon W \to W$ is given by multiplication with the symmetric matrix

$$M(\theta) = \begin{pmatrix} M_\Omega(\theta) & 0 & 0 \\ 0 & M_{\Gamma_N}(\theta) & 0 \\ 0 & 0 & M_{\Gamma_D}(\theta) \end{pmatrix},$$

where

$$M_\Omega(\theta) = \begin{pmatrix} d(\theta) & \mathbf{d}(\theta)^T \\ \mathbf{d}(\theta) & 2eI \end{pmatrix}, \quad M_{\Gamma_N}(\theta) = \begin{pmatrix} m_0(\theta) & m_1(\theta) \\ m_1(\theta) & m_2(\theta) \end{pmatrix}, \quad M_{\Gamma_D}(\theta) = 0, \tag{60}$$

$$d(\theta) \;=\; -\nabla\cdot(m_\theta \mathbf{e}), \tag{61}$$

$$\mathbf{d}(\theta) \;=\; \mathbf{e} - \nabla(m_\theta e), \tag{62}$$

$$m_0(\theta) \;=\; m_\theta \mathbf{n}\cdot\mathbf{e}, \tag{63}$$

$$m_1(\theta) \;=\; m_\theta e + c_0, \tag{64}$$

$$m_2(\theta) \;=\; 2c. \tag{65}$$

Here $M_\Omega(\theta)$ acts on $\binom{w_0}{\mathbf{w}}$, $M_{\Gamma_N}(\theta)$ acts on $\binom{w_1}{w_2|_{\Gamma_N}}$, and $M_{\Gamma_D}(\theta)$ acts on $w_2|_{\Gamma_D}$.

Proposition 8. *For $E(\theta)$ defined in (57), we have*

$$(E(\theta)u)^* E(\theta)u = I_\Omega + I_{\Gamma_N}, \tag{66}$$

where

$$I_\Omega \;=\; \int \Big((m_\theta^2 e^2 + \mathbf{e}^2)u_0^2 + e^2\mathbf{u}^2 + 2e\mathbf{e}\cdot u_0\mathbf{u}\Big)\mathrm{d}\Omega, \tag{67}$$

$$I_\Gamma \;=\; \int \Big(c_0^2 u_0^2 + 2c_0 c(\mathbf{n}\cdot\mathbf{u})u_0 + c^2(\mathbf{n}\cdot\mathbf{u})^2\Big)\mathrm{d}\Gamma_N. \tag{68}$$

Proposition 9. *For J, $E(\theta)$ defined in (35), (57) we have*

$$(E(\theta)u)^* E(\theta)u = (Ju)^* N(\theta) Ju,$$

where $N(\theta) : W \to W$ is given by multiplication with the symmetric matrix

$$N(\theta) = \begin{pmatrix} N_\Omega(\theta) & 0 & 0 \\ 0 & N_{\Gamma_N}(\theta) & 0 \\ 0 & 0 & N_{\Gamma_D}(\theta), \end{pmatrix},$$

where

$$N_\Omega(\theta) = \begin{pmatrix} h(\theta) & \mathbf{h}(\theta)^T \\ \mathbf{h}(\theta) & e^2 I \end{pmatrix}, \quad N_{\Gamma_N}(\theta) = \begin{pmatrix} h_0(\theta) & h_1(\theta) \\ h_1(\theta) & h_2(\theta) \end{pmatrix}, \quad N_{\Gamma_D}(\theta) = 0,$$

and

$$\begin{aligned} h(\theta) &= m_\theta^2 e^2 + \mathbf{e}^2, \tag{69} \\ \mathbf{h}(\theta) &= e\mathbf{e}, \tag{70} \\ h_0(\theta) &= c_0^2, \tag{71} \\ h_1(\theta) &= c_0 c, \tag{72} \\ h_2(\theta) &= c^2. \tag{73} \end{aligned}$$

Here $N_\Omega(\theta)$ acts on $\binom{w_0}{\mathbf{w}}$, $N_{\Gamma_N}(\theta)$ acts on $\binom{w_1}{w_2|_{\Gamma_N}}$, and $N_{\Gamma_D}(\theta)$ acts on $w_2|_{\Gamma_D}$.

7. Finding β^* and the dual norm

This section provides the information about the functions involved in the construction of the operators $M(\theta)$ and $N(\theta)$ in the previous section. That will lead us to formulate the dual norm defined by (9) and the optimization problems for β and β^* in the next section.
By (7), we have

$$M_\beta(\theta) = \beta M(\theta) - N(\theta) = \begin{pmatrix} M_{\beta\Omega}(\theta) & 0 & 0 \\ 0 & M_{\beta N}(\theta) & 0 \\ 0 & 0 & M_{\beta D}(\theta) \end{pmatrix},$$

where

$$M_{\beta\Omega}(\theta) = \begin{pmatrix} d_\beta(\theta) & \mathbf{d}_\beta(\theta)^T \\ \mathbf{d}_\beta(\theta) & (2e\beta - e^2)I \end{pmatrix}, \quad M_{\beta N}(\theta) = \begin{pmatrix} m_{0\beta}(\theta) & m_{1\beta}(\theta) \\ m_{1\beta}(\theta) & m_{2\beta}(\theta) \end{pmatrix}, \quad M_{\beta D}(\theta) = 0,$$

with

$$\begin{aligned} d_\beta(\theta) &= \beta d(\theta) - h(\theta) = -\beta \nabla \cdot (m_\theta \mathbf{e}) - m_\theta^2 e^2 - \mathbf{e}^2 && \text{in } \Omega, \\ \mathbf{d}_\beta(\theta) &= \beta \mathbf{d}(\theta) - \mathbf{h}(\theta) = (\beta - e)\mathbf{e} - \beta \nabla(m_\theta e) && \text{in } \Omega. \end{aligned}$$

On Γ_N, we have

$$\begin{aligned} m_{0\beta}(\theta) &= \beta m_0(\theta) - h_0(\theta) = \beta m_\theta \mathbf{n} \cdot \mathbf{e} - c_0^2, \tag{74} \\ m_{1\beta}(\theta) &= \beta m_1(\theta) - h_1(\theta) = \beta(m_\theta e + c_0) - c c_0, \tag{75} \\ m_{2\beta}(\theta) &= \beta m_2(\theta) - h_2(\theta) = (2\beta - c)c, \tag{76} \end{aligned}$$

where $d(\theta)$, $\mathbf{d}(\theta)$, $m_0(\theta)$, $m_1(\theta)$, $m_2(\theta)$ are given by (61)–(65) and $h(\theta)$, $\mathbf{h}(\theta)$, $h_0(\theta)$, $h_1(\theta)$, $h_2(\theta)$ are given by (69)–(73).

Proposition 10. *For $s_\beta = \begin{pmatrix} s_{0\beta} \\ \mathbf{s}_\beta \\ s_{1\beta} \\ s_{2\beta} \end{pmatrix} \in W$ and $s = \begin{pmatrix} s_0 \\ \mathbf{s} \\ s_1 \\ s_2 \end{pmatrix} \in W$,*

$$M_\beta(\theta) s_\beta = s \tag{77}$$

implies $s_2 = 0$ on Γ_D. If this holds, the general solution of (77) is

$$
\begin{aligned}
s_{0\beta} &= \frac{(2e\beta - e^2)s_0 - \mathbf{d}_\beta \cdot \mathbf{s}}{(2e\beta - e^2)d_\beta - \mathbf{d}_\beta^2} \qquad \textit{in } \Omega, && (78)\\
\mathbf{s}_\beta &= \frac{(2e\beta - e^2)(d_\beta \mathbf{s} - \mathbf{d}_\beta s_0) + (\mathbf{s}\cdot \mathbf{d}_\beta)\mathbf{d}_\beta - \mathbf{d}_\beta^2 \mathbf{s}}{(2e\beta - e^2)((2e\beta - e^2)d_\beta - \mathbf{d}_\beta^2)} \qquad \textit{in } \Omega, && (79)\\
s_{1\beta} &= \frac{m_{2\beta}s_1 - m_{1\beta}s_2}{m_{2\beta}m_{0\beta} - m_{1\beta}^2} \qquad \textit{on } \Gamma_N, && (80)\\
s_{2\beta} &= \frac{m_{0\beta}s_2 - m_{1\beta}s_1}{m_{2\beta}m_{0\beta} - m_{1\beta}^2} \qquad \textit{on } \Gamma_N. && (81)
\end{aligned}
$$

Proof. By (77), we get

$$
\begin{aligned}
d_\beta s_{0\beta} + \mathbf{d}_\beta \cdot \mathbf{s}_\beta &= s_0 \qquad \text{in } \Omega, && (82)\\
\mathbf{d}_\beta s_{0\beta} + (2e\beta - e^2)I\mathbf{s}_\beta &= \mathbf{s} \qquad \text{in } \Omega, && (83)\\
m_{0\beta}s_{1\beta} + m_{1\beta}s_{2\beta} &= s_1 \qquad \text{on } \Gamma_N, && (84)\\
m_{1\beta}s_{1\beta} + m_{2\beta}s_{2\beta} &= s_2 \qquad \text{on } \Gamma_N. && (85)
\end{aligned}
$$

(78), (79) are obtained from (82), (83) and (80), (81) are obtained from (84) and (85). Since $M_{\beta D}(\theta) = 0$, we find $s_2 = 0$ on Γ_D. □

In particular, we see that $W_\beta = \{s \in W \mid s_2 = 0 \text{ on } \Gamma_D\}$. So, the Dirichlet conditions for the adjoint equations in Proposition 5 must be satisfied exactly in order to be able to apply Theorem 3.

Proposition 11. *$M_{\beta\Omega}(\theta)$, $M_{\beta N}(\theta)$ are positive semidefinite if*

$$
\begin{aligned}
A^\theta &:= (2e\beta - e^2)d_\beta(\theta) - \mathbf{d}_\beta(\theta)^2 > 0 \qquad \textit{in } \Omega, && (86)\\
B^\theta &:= 2e\beta - e^2 > 0 \qquad \textit{in } \Omega, && (87)\\
C^\theta &:= m_{2\beta}(\theta)m_{0\beta}(\theta) - m_{1\beta}(\theta)^2 > 0 \qquad \textit{on } \Gamma_N, && (88)\\
m_{0\beta}(\theta) &= \beta m_\theta \mathbf{n}\cdot\mathbf{e} - c_0^2 > 0 \qquad \textit{on } \Gamma_N. && (89)
\end{aligned}
$$

In this case

$$||s||^2_{W_\beta(\theta)} = |s|^2_{\beta\Omega(\theta)} + |s|^2_{\beta N(\theta)} ,$$

where

$$
\begin{aligned}
|s|^2_{\beta\Omega(\theta)} &= \int \Big(\frac{(B^\theta s_0 - \mathbf{s}\cdot \mathbf{d}_\beta(\theta))^2}{A^\theta} + \mathbf{s}^2\Big)\frac{d\Omega}{B^\theta}, && (90)\\
|s|^2_{\beta N(\theta)} &= \int \Big(m_{2\beta}(\theta)s_1^2 + m_{0\beta}(\theta)s_2^2 - 2m_{1\beta}(\theta)s_1 s_2\Big)\frac{d\Gamma_N}{C^\theta}. && (91)
\end{aligned}
$$

Proof. By (9), for $s \in W_\beta$ the dual norm is

$$||s||^2_{W_\beta(\theta)} = s^* M_\beta(\theta)^{-1} s = s^* s_\beta = |s|^2_{\beta\Omega(\theta)} + |s|^2_{\beta N(\theta)} .$$

Now $|s|^2_{\beta\Omega(\theta)} = \int(s_0 s_{0\beta} + \mathbf{s}\cdot\mathbf{s}_\beta)d\Omega$, and using (78), (79), we get (90). Similarly, $|s|^2_{\beta N(\theta)} = \int(s_1 s_{1\beta} + s_2 s_{2\beta})d\Gamma_N$, and using (80) and (81),we get (91). □

8. An optimization problem for β

To compute the dual norm $\|s\|_{W_\beta}$ by Proposition 11, one needs to have A^θ, B^θ, C^θ given by (86)–(88). Therefore, we can apply Theorem 3 giving the error bounds. To find A^θ, B^θ and C^θ we need to solve some optimization problems for β and β^*. Section 7 gives the information about the functions involved in the construction of the operators $M(\theta)$, $N(\theta)$ and $E(\theta)$ in Section 6. This is used in the construction of the optimization problems for β and β^*.

Theorem 4. *Suppose that there exist functions $e \in C^{0,1}(\overline{\Omega})$ and $\mathbf{e} \in C^{0,1}(\overline{\Omega})^d$ such that*

$$\left.\begin{array}{rcll} e & > & 0 & \text{in } \Omega, \\ -2e\nabla\cdot(m_\theta\mathbf{e}) - (\nabla(m_\theta e) - \mathbf{e})^2 & > & 0 & \text{in } \Omega, \\ \mathbf{n}\cdot\mathbf{e} & > & 0 & \text{on } \Gamma_N, \end{array}\right\} \tag{92}$$

and

$$c_0 = -m_\theta e, \; c = \beta \qquad \text{on } \Gamma_N. \tag{93}$$

Then $M_\beta(\theta)$ is positive semidefinite if

$$\beta > \beta^* = \max\left(\beta_A(\theta), \beta_B(\theta), \beta_C(\theta)\right), \tag{94}$$

where

$$\begin{aligned} \beta_A(\theta) &= \max_{x\in\overline{\Omega}} \frac{e}{[2k_1(\theta)]_+}\Big(- k_2(\theta) + \sqrt{k_2^2(\theta) - 4k_1(\theta)m_\theta^2 e^2}\Big), \\ \beta_B(\theta) &= \max_{x\in\overline{\Omega}} \frac{e}{2}, \\ \beta_C(\theta) &= \max_{x\in\Gamma_N} \frac{2m_\theta e^2}{[\mathbf{n}\cdot\mathbf{e}]_+}, \end{aligned}$$

and

$$\begin{aligned} k_1(\theta) &= -2e\nabla\cdot(m_\theta\mathbf{e}) - (\nabla(m_\theta e) - \mathbf{e})^2, & (95)\\ k_2(\theta) &= e\nabla\cdot(m_\theta\mathbf{e}) - 2m_\theta^2 e^2 - 2\mathbf{e}\cdot\nabla(m_\theta e). & (96) \end{aligned}$$

Proof. In Ω, (86) can be written as

$$A^\theta = k_1(\theta)\beta^2 + ek_2(\theta)\beta + m_\theta^2 e^4 > 0.$$

Since $e > 0$ and $k_1(\theta) > 0$ in Ω by (92), condition (86) is satisfied if $\beta > \beta_A(\theta)$. Since $e > 0$ in Ω, therefore, (87) is satisfied if $\beta > \beta_B(\theta)$. By (74)–(76) and (93), we get

$$\begin{aligned} m_{2\beta}(\theta)m_{0\beta}(\theta) - m_{1\beta}(\theta)^2 &= \beta^2 m_\theta(\beta\mathbf{n}\cdot\mathbf{e} - 2m_\theta e^2) \quad \text{on } \Gamma_N, & (97)\\ m_{0\beta}(\theta) &= m_\theta(\beta\mathbf{n}\cdot\mathbf{e} - m_\theta e^2) \qquad \text{on } \Gamma_N. & (98) \end{aligned}$$

By (97) and (98), we see that (88) is a stronger condition than (89). Therefore, (88) is satisfied if $\beta > \beta_C(\theta)$. Thus if we take

$$\beta > \beta^* = \max\left(\beta_A(\theta), \beta_B(\theta), \beta_C(\theta)\right),$$

then (86)–(89) hold. Therefore, $M_\beta(\theta)$ is positive semidefinite. □

To ensure that (92) and (94) are satisfied, we solve the optimization problem

$$\left.\begin{array}{lrcl} \min & \beta & & \\ s.t. & e & \geq 1 & \text{in } \Omega, \\ & 2\beta - e & \geq 1 & \text{in } \Omega, \\ & k_1(\theta)\beta^2 + ek_2(\theta)\beta + m_\theta^2 e^4 & \geq 1 & \text{in } \Omega, \\ & \beta\mathbf{n}\cdot\mathbf{e} - 2m_\theta e^2 & \geq 1 & \text{on } \Gamma_N, \end{array}\right\} \tag{99}$$

where $k_1(\theta)$ and $k_2(\theta)$ are given by (95) and (96). We can then apply Theorem 3 and obtain error bounds for any $\beta > \beta^*$.

9. Numerical results

The method discussed in this paper was implemented in Matlab for mass-weighted Poisson equations with mixed Dirichlet and Neumann boundary conditions on arbitrary polygonal domains in 2 dimensions, using linear finite elements on a uniformly refined grid constructed from the initial triangulation.

A discretized version of the optimization problem (99) was encoded in the modeling language AMPL [4] and solved by IPOPT [5], which gives $\beta =: \beta_{ampl}$, e and $\mathbf{e}$.To account for the discretization errors in the AMPL formulation, we then solve for the linear finite element solution e, $\mathbf{e}$ interpolated from the grid small many optimization problems for β, three over each triangle and one over each boundary edge, corresponding to the four inequalities in (99). This is done using the routine `fmincon` from the Matlab optimization tool box. The maximum β_{fmin} of the resulting maxima is then a valid upper bound for β^*. Choosing $\beta > \beta_{fmin}$ therefore gives valid bounds. However, a choice very close to β_{fmin} will lead to some very small denominators in the expressions in the integrals, causing large error bounds. We found that choosing $\beta = 1.01\beta_{ampl}$ gives reasonable error bounds.

The method works well for convex domains. It also produced bounds in the nonconvex case, but these bounds for an L-shaped domain were not very tight, and bounds for domains where an angle $\alpha < 90°$ was cut out deteriorated rapidly when the angle was decreased. Some of these difficulties are likely due to the limitations of using linear finite elements, others due to the limitations of using a uniformly refined grid. Using adaptive higher order finite elements would probably improve the bounds.

In the following, we give results for a square domain. We consider the Poisson equation

$$-\nabla \cdot (m_\theta \nabla u_0) = 2x - 3y \qquad \text{in } \Omega = [0,1] \times [0,1], \tag{100}$$

in 2 dimensions, with mixed Dirichlet and Neumann boundary conditions as specified in Figure 2 and uncertain mass distribution

$$m_\theta = 1 + \theta_1 x + \theta_2 y \qquad \text{for } \theta_1, \theta_2 \in \Theta.$$

The uncertain parameter domain was taken to be $\theta \in \Theta = [-0.1, 0.1] \times [-0.1, 0.1]$. The response functional was chosen to be

$$R(u) = \int u_0 \mathrm{d}\Omega,$$

which is of the form (56).

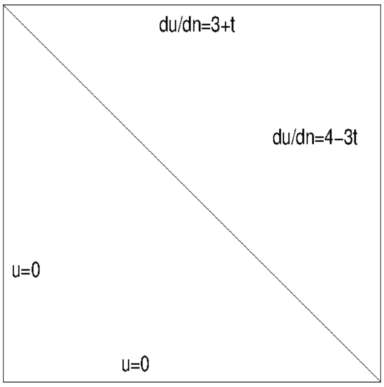

Figure 2: Initial grid and boundary conditions for test problem

We solve (22)–(23) numerically by the least squares finite element method, using linear finite elements on a uniform grid. So, we minimize the norm of the residual defined by (17), for $L(\theta)u$ defined by (24).

For given $f \in \overline{V}$, let $\tilde{u} = \begin{pmatrix} \tilde{u}_0 \\ \tilde{\mathbf{u}} \end{pmatrix} \in U$ be an approximate solution of $L(\theta)u = f$, where $\tilde{u}_0$ is an approximate solution of the Poisson equation (22)–(23). The residual r is then computed as

$$r = f - L(\theta)\tilde{u} = \begin{pmatrix} (f_0 - \nabla \cdot \tilde{\mathbf{u}})|_\Omega \\ (\mathbf{f} - m_\theta \nabla \tilde{u}_0 - \tilde{\mathbf{u}})|_\Omega \\ (f_1 - \mathbf{n} \cdot \tilde{\mathbf{u}})|_{\Gamma_N} \end{pmatrix} \in \overline{V}. \tag{101}$$

Approximate solutions u_l of (100) for a small number of scenarios $\theta_l \in \Theta$ are computed. An initial approximation $\tilde{u}(\theta)$ is being computed by the bilinear interpolation of u_l in Θ-space. In a similar way, a θ-dependent approximation $\tilde{v}(\theta)$ of the dual problem is computed. The optimization problem (99) was encoded in the modeling language AMPL [4] and solved with the nonlinear constrained optimization solver IPOPT [5] for the 4 scenarios θ_l at the corners of the square Θ. This gives $\beta_{ampl} = 7.79333$ and the computed bounds at the corners (where the worst cases are attained) are given in Table 1.

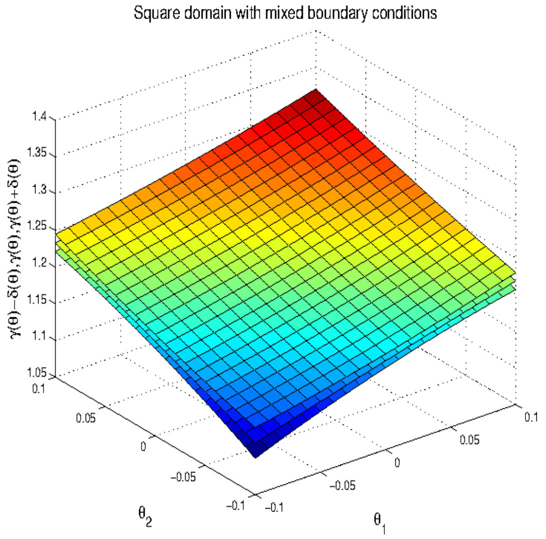

Figure 3: Bounds for $\theta_1, \theta_2 \in \Theta$ with mesh size $h = 2^{-12}$

For the global solution, we used the solution of the problem (99) as an initial guess for (94), and solved the latter in Θ-space with the optimization program FMINCON from the MATLAB optimization tool box. The result – cf. (94) – was $\beta_{fmin} = \beta^* = 7.561897$. Thus any $\beta > \beta^*$ satisfies the assumptions for our bounds. We chose $\beta = 1.01 * \beta_{ampl} = 7.8713$ for our numerical computation.

The computed bounds as a function of $\theta \in \Theta$ are shown in Figure 3.

corners and center of Θ	$\gamma(\theta) - \delta(\theta)$	$\gamma(\theta)$	$\gamma(\theta) + \delta(\theta)$
$(\theta_1, \theta_2) = (-0.1, -0.1)$	1.0996	1.1205	1.1415
$(\theta_1, \theta_2) = (-0.1, 0.1)$	1.2207	1.2328	1.2448
$(\theta_1, \theta_2) = (0.1, -0.1)$	1.2066	1.2183	1.2300
$(\theta_1, \theta_2) = (0.1, 0.1)$	1.2968	1.3084	1.3200
$(\theta_1, \theta_2) = (0, 0)$	1.2209	1.2230	1.2250

Table 1: Bounds at the corners and at the center of the region Θ

References

[1] R. Becker and R. Rannacher, An optimal control approach to a posteriori error estimation in finite element methods, pp. 1–102 in: Acta Numerica 2001 (A. Iserles, ed.), Cambridge Univ. Press 2001.

[2] D. Bertsimas and C. Caramanis, Bounds on linear PDEs via semidefinite optimization, Mathematical Programming, Ser. A. 108 (2006), 135–158.

[3] D. Gilbarg and N.S. Trudinger, Elliptic Partial Differential Equations of Second Order, Springer, Berlin, 2001.

[4] R. Fourer, D.M. Gay and B.W. Kernighan, AMPL: A Modeling Language for Mathematical Programming, Duxbury Press, Brooks/Cole Publishing Company, 1993.

[5] IPOPT home page, Web document, 2013, https://projects.coin-or.org/Ipopt.

[6] M.T. Nakao, A numerical approach to the proof of existence of solutions for elliptic problems I, Japan J. Appl. Math. 5 (1988), 313–332.

[7] M.T. Nakao, A numerical approach to the proof of existence of solutions for elliptic problems II, Japan J. Appl. Math. 7 (1990), 477–488.

[8] A. Neumaier, Certified error bounds for uncertain elliptic equations, J. Comput. Appl. Math. 218 (2008), 125–136.

[9] N. Pares, J. Bonet, A. Huerta, and J. Peraire, The computation of bounds for linear-functional outputs of weak solutions to the two-dimensional elasticity equations, Comp. Meth. Appl. Mech. Engin. 195 (2006), 406–429.

[10] M. Plum, Explicit H_2-estimates and pointwise bounds for solutions of second-order elliptic boundary value problems, J. Math. Anal. Appl. 165 (1992), 36–61.

[11] M. Plum, Existence and enclosure results for continua of solutions of parameter-dependent nonlinear boundary value problems, J. Comput. Appl. Math. 60 (1995), 187–200.

[12] S. Repin, A posteriori error estimation for nonlinear variational problems by duality theory, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) 243 (1997), 201–214.

[13] A.M. Sauer-Budge, J. Bonet, A. Huerta, and J. Peraire, Computing bounds for linear functionals of exact weak solutions to Poisson's equation, SIAM J. Numer. Anal. 42 (2004), 1610–1630.

[14] A.M. Sauer-Budge and J. Peraire, Computing Bounds for Linear Functionals of Exact Weak Solutions to the Advection-Diffusion-Reaction Equation, SIAM Journal on Scientific Computing 26 (2004), 636–652.

[15] N. Yamamoto, M.T. Nakao, M.T., and Y. Watanabe, Validated computation for a linear elliptic problem with a parameter, pp. 155–162 in: Advances in Numerical Mathematics (Kawarada et al., eds.), GAKUTO International Series, Mathematical Sciences and Applications Vol. 12 (1999), 155–162.