Control in the coefficients of an elliptic differential operator: topological derivatives and Pontryagin maximum principle

Daniel Wachsmuth*

May 8, 2024

Abstract. We consider optimal control problems, where the control appears in the main part of the operator. We derive the Pontryagin maximum principle as a necessary optimality condition. The proof uses the concept of topological derivatives. In contrast to earlier works, we do not need continuity assumptions for the coefficient or gradients of solutions of partial differential equations. Following classical proofs, we consider perturbations of optimal controls by multiples of characteristic functions of sets, whose scaling factor is send to zero. For $2d$ problems, we can perform an optimization over the elliptic shapes of such sets leading to stronger optimality conditions involving a variational inequality of a new type.

Keywords. Optimal control, control in the coefficients, Pontryagin maximum principle, topological derivatives

MSC (2020) classification. 49K20, 35J15

1 Introduction

In this article, we are interested in proving the Pontryagin maximum principle maximum principle for the following problem: Minimize

$$\frac{1}{2}\int (y(x) - y_d(x))^2 \,\mathrm{d}x + \int_\Omega g(a(x))\,\mathrm{d}x \tag{1.1}$$

over all

$$a \in \mathcal{A} \subseteq L^\infty(\Omega;\mathbb{R}^{d,d}), \tag{1.2}$$

where $y \in H^1_0(\Omega)$ is the weak solution of

$$-\operatorname{div}(a\nabla y) = f \text{ a.e. in } \Omega. \tag{1.3}$$

Hence, the optimization variable is the coefficient in the main part of the differential operator. In this problem, $\Omega \subseteq \mathbb{R}^d$ is a bounded domain, $f \in H^{-1}(\Omega)$ is

*Institut für Mathematik, Universität Würzburg, 97074 Würzburg, Germany, daniel.wachsmuth@mathematik.uni-wuerzburg.de. This research was partially supported by the German Research Foundation DFG under project grant Wa 3626/5-1.

a given source term, $y_d \in L^2(\Omega)$ is the desired state, while $g : \mathbb{R}^{d,d} \to \mathbb{R} \cup \{+\infty\}$ models the cost of choosing a certain coefficient matrix. In addition, $\mathcal{A}$ is a feasible set, that contains matrices with uniformly positive definite symmetric part. For the precise statement of the assumptions, we refer to Section 2.1.

Problem (1.1)–(1.3) is a classical problem, and lead to the study of H-convergence, [14]. One cannot prove existence of solutions, and examples without solutions can be found, e.g., in [2, 13].

Here, we are interested in proving the Pontryagin maximum principle, which is a classical necessary optimality condition in optimal control theory. Let a be a solution of the problem above. Then we consider a perturbation of the type

$$a_r := a + \chi_{r\omega}(b - a), \tag{1.4}$$

where $b \in \mathbb{R}^{d,d}$ is a constant matrix, ω is the unit ball in $\mathbb{R}^d$. And we are interested in passing to the limit in the difference quotient

$$\frac{1}{r^d}(J(y_r, a_r) - J(y, a)), \tag{1.5}$$

where y_r is the solution to the elliptic equation with coefficient a_r. For the state-dependent part of the cost functional J, we have the following expansion

$$\begin{aligned}\frac{1}{2}\|y_r - y_d\|^2_{L^2(\Omega)} - \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} \\ = -\int_\Omega (a_r - a)\nabla y \cdot \nabla p \,\mathrm{d}x - \int_\Omega (a_r - a)\nabla y \cdot \nabla(\tilde{p}_r - p)\,\mathrm{d}x,\end{aligned}$$

where p and $\tilde{p}_r$ are certain adjoint states. The first term in the expansion represents the Fréchet derivative of the map $a \mapsto y$ from $L^\infty(\Omega)$ to $H^1_0(\Omega)$, while the second term is of higher-order in $\|a_r - a\|_{L^\infty(\Omega)}$. Due to the choice of the perturbation, we do not get that $\|a_r - a\|_{L^\infty(\Omega)} \to 0$ for $r \to 0$. And the second term in the expansion does not vanish when passing to the limit in the difference quotient (1.5).

As one would expect, we will encounter topological derivatives of solutions of the elliptic partial differential equations. In contrast to earlier works, we prove the corresponding results under much weaker assumptions than in the literature:

1. The coefficient function a is assumed to belong to $L^\infty(\Omega; \mathbb{R}^{d,d})$ such that $-\operatorname{div}(a\nabla\cdot)$ is a uniformly elliptic operator. We do not assume that a is piecewise constant [2, 3, 6] or continuous [11],

2. We work with weak solutions in H^1. We do not assume that the weak solutions y or their gradients ∇y are continuous.

3. That is, our proof works under the same set of assumptions than the Lax-Milgram theorem, and the proof only uses regularity of solutions as provided by Lax-Milgram.

The assumption of piecewise constant coefficients a makes sense in material or topology optimization. However, this assumption is too restrictive for the optimization problem (1.1)–(1.3). Our proof follows the developments of [6, 11, 22]. The main improvement compared to these earlier works is the consequent

use of the celebrated Lebesgue differentiation theorem, which allows to dispense with continuity assumptions. This derivation is done in Section 2 with the main result being the asymptotic expansion of the cost functional in Theorem 2.12.

Utilizing these concepts, we obtain the following statement of the Pontryagin maximum principle. Let a be (locally) optimal in $L^1(\Omega)$ with associated state y and adjoint p. Then for almost all $x_0 \in \Omega$ and all feasible perturbations $b \in \mathbb{R}^{d,d}$ we have

$$0 \leq -(b - a(x_0))\nabla y(x_0) \cdot \left(\nabla p(x_0) + \frac{1}{|\omega|} \int_\omega \nabla_{x'} Q \, \mathrm{d}x' \right) + g(b) - g(a(x_0)).$$

Here, Q is a solution of a certain adjoint equation on $\mathbb{R}^d$, which depends on $a(x_0)$, b, and the set ω, which is used in the perturbation (1.4). The precise statement can be found in Theorem 4.1. If a would be locally optimal in $L^\infty(\Omega)$, then one could use Fréchet derivatives, and one would obtain a similar inequality but with $Q = 0$. For certain shapes ω (balls in $\mathbb{R}^d$, ellipses in $\mathbb{R}^2$), explicit formulas for Q are available. Similar results can be found in the work of Raitums [17, 18, 19, 20], which deserves to be better known, but which seems to be only available in Russian.

In the special case that the coefficient a is scalar and $d = 2$, one can optimize the above formula for elliptic shapes ω to obtain the following strengthened version

$$\begin{aligned} &- (b - a(x_0))\nabla y(x_0) \cdot \nabla p(x_0) + g(b) - g(a(x_0)) \\ &\qquad + \frac{1}{2} \frac{(b - a(x_0))^2}{b} \left(\nabla y(x_0) \cdot \nabla p(x_0) - \|\nabla y(x_0)\|_2 \|\nabla p(x_0)\|_2 \right) \geq 0. \end{aligned}$$

A related result can be found in [5] for the study of a material optimization problem, where a is allowed to take only two different values. These inequalities are stronger than the related inequalities one gets using Fréchet derivatives in $L^\infty(\Omega)$.

The plan of the paper is as follows. The sensitivity analysis of the cost functional J with respect to perturbations of the coefficient is performed in Section 2, where the main result is Theorem 2.12. The special cases of perturbations with characteristic functions of balls and ellipses are considered in Section 3. These results are applied in Section 4 to an optimal control problem with control in the coefficients.

Notation Given $v \in \mathbb{R}^d$, we denote its Euclidean norm by $\|v\|_2$. The set $B(x, r) \subseteq \mathbb{R}^d$ is the open ball centered at x with radius r. The characteristic function of a set $A \subseteq \mathbb{R}^d$ is denoted by χ_A. The Lebesgue measure of a measurable set $A \subseteq \mathbb{R}^d$ is denoted by $|A|$, the measure of a ball with radius r in $\mathbb{R}^d$ is denoted by $|B(r)|$.

As is customary in the literature on partial differential equations, we will denote the inner product in $\mathbb{R}^d$ of gradients by dots, i.e., $\nabla y(x) \cdot \nabla p(x) := \nabla p(x)^T \nabla y(x)$.

2 Sensitivity analysis with respect to perturbations on general sets

2.1 Setup of the problem

Throughout the paper we assume the following about the data of the problem.

Assumption 1. *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain. Let $\alpha > 0$ be given. In addition, let $y_d, f \in L^2(\Omega)$ be given.*

Let us define the set of admissible coefficient functions by

$$\mathcal{A} := \{a \in L^\infty(\Omega; \mathbb{R}^{d,d}) : \ a(x) \in \mathcal{M} \ \text{ f.a.a. } x \in \Omega\}, \tag{2.1}$$

where

$$\mathcal{M} := \{a \in \mathbb{R}^{d,d} : \ \xi^T a \xi \geq \alpha |\xi|^2 \ \forall \xi \in \mathbb{R}^d\}. \tag{2.2}$$

In the sequel, we will work with coefficient functions from the set $\mathcal{A}$. Note that $a(x) \in \mathcal{A}$ for almost all $x \in \Omega$ is the minimum requirement in order that the Lax-Milgram theorem guarantees existence of weak solutions. We will not assume more regularity of a and Ω, and we will not rely on any elliptic regularity results beyond basic H^1-regularity.

Let us fix a reference coefficient $a \in \mathcal{A}$. We denote the corresponding solution of the state equation by y, i.e., $y \in H_0^1(\Omega)$ solves

$$\int_\Omega a \nabla y \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x \quad \forall v \in H_0^1(\Omega). \tag{2.3}$$

Here, we used the notation

$$a \nabla y \cdot \nabla v := \sum_{i,j=1}^d a_{ij} \frac{\partial}{\partial x_j} y \frac{\partial}{\partial x_i} v.$$

By Lax-Milgram theorem, the equation (2.3) is uniquely solvable. Given a, we define

$$J(a) := \frac{1}{2} \|y - y_d\|^2_{L^2(\Omega)},$$

where y is the solution of (2.3) to $a \in A$. We are interested in the sensitivity analysis of J with respect to perturbations of a. Here, we will use perturbations by characteristic functions, which is a well-known concept in optimal control, inverse problems, or material optimization. To this end, let $\omega \subseteq \mathbb{R}^d$ be an open bounded set with $0 \in \omega$. Given a point $x_0 \in \Omega$, a value $b \in \mathcal{M}$, and a radius (or scaling parameter) $r > 0$, we define

$$a_r := a + \chi_{x_0 + r\omega}(b - a). \tag{2.4}$$

The goal of this section is to compute the variation of J at a with respect to b, ω, which is defined as

$$\delta J(a; b, x_0, \omega) := \lim_{r \to 0} \frac{J(a_r) - J(a)}{r^d |\omega|}.$$

Note that $r^d|\omega|$ is the Lebesgue measure of $r\omega$, and it is larger or equal to the L^0- or Ekeland distance between a_r and a.

The exposition in the following subsections follows earlier work [11, 22, 6]. As one would expect, the statements of the main results are identical. However, we do not use or assume any regularity beyond L^∞ for the coefficients and H^1 for the weak solutions of state and adjoint equations.

2.2 Basic expansion of the functional

Recall the definition of a_r in (2.4). Let y_r be the corresponding solution of the state equation, i.e.,

$$\int_\Omega a_r \nabla y_r \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega). \tag{2.5}$$

Note that the difference $y_r - y$ satisfies the following equation

$$\int_\Omega a_r \nabla (y_r - y) \cdot \nabla v \,\mathrm{d}x + \int_\Omega (a_r - a) \nabla y \cdot \nabla v \,\mathrm{d}x = 0 \quad \forall v \in H_0^1(\Omega). \tag{2.6}$$

We define the averaged adjoint $\tilde{p}_r \in H_0^1(\Omega)$, see, e.g., [22], as the solution of

$$\int_\Omega a_r \nabla v \cdot \nabla \tilde{p}_r \,\mathrm{d}x = \frac{1}{2} \int_\Omega [(y_r - y_d) + (y - y_d)] v \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega). \tag{2.7}$$

In addition, let the adjoint $p \in H_0^1(\Omega)$ be given as the solution of

$$\int_\Omega a \nabla v \cdot \nabla p \,\mathrm{d}x = \int_\Omega (y - y_d) v \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega). \tag{2.8}$$

Then we have the following result.

Lemma 2.1. *Let a_r as in (2.4) with the notation from Section 2.1. Then it holds*

$$J(a_r) - J(a) = -\int_\Omega (a_r - a) \nabla y \cdot \nabla p \,\mathrm{d}x - \int_\Omega (a_r - a) \nabla y \cdot \nabla (\tilde{p}_r - p) \,\mathrm{d}x,$$

where y, $\tilde{p}_r$, p solve (2.3), (2.7), (2.8).

Proof. Using the equations (2.7) and (2.6), we find

$$\begin{aligned}
J(a_r) - J(a) &= \frac{1}{2} \|y_r - y_d\|^2_{L^2(\Omega)} - \frac{1}{2} \|y - y_d\|^2_{L^2(\Omega)} \\
&= \frac{1}{2} (y_r - y,\ (y_r - y_d) + (y - y_d)) \\
&= \int_\Omega a_r \nabla (y_r - y) \cdot \nabla \tilde{p}_r \,\mathrm{d}x \\
&= -\int_\Omega (a_r - a) \nabla y \cdot \nabla \tilde{p}_r \,\mathrm{d}x \\
&= -\int_\Omega (a_r - a) \nabla y \cdot \nabla p \,\mathrm{d}x - \int_\Omega (a_r - a) \nabla y \cdot \nabla (\tilde{p}_r - p) \,\mathrm{d}x,
\end{aligned}$$

which is the claim. □

Here the second term in the expansion of Lemma 2.1 seems to be of second order. In fact, it corresponds to the remainder term in a Taylor expansion of J using the Fréchet differentiability of $a \mapsto y$ from $L^\infty(\Omega)$ to $H_0^1(\Omega)$. Consequently it is of the order $\|a_r - a\|^2_{L^\infty(\Omega)}$. However, the second term is *not* of higher order with respect to $r \searrow 0$. In fact, the second term converges with the same order as the first for $r \searrow 0$.

The key for the asymptotic analysis of the expansion of Lemma 2.1 is the following coordinate transform. This well-known idea is to transform the small set $x_0 + r\omega$ to ω. To this end, let us define the following functions:

$$T_r(x') := x_0 + rx', \quad T_r^{-1}(x) := (x - x_0)/r. \tag{2.9}$$

2.3 Lebesgue differentiation theorem

Before continuing with the analysis of the expansion, let us recall the Lebesgue differentiation theorem in the following form.

Theorem 2.2. *Let $u \in L^p_{loc}(\mathbb{R}^d)$ for some $p \in [1, \infty)$. Then*

$$\lim_{r \to 0} \frac{1}{|B(x,r)|} \int_{B(x,r)} |u(y) - u(x)|^p \, \mathrm{d}y = 0$$

for almost all $x \in \mathbb{R}^d$. These points x are called p-Lebesgue points *of u.*

Proof. See [10, Section 1.7, Corollary 1]. □

Note that the claim is slightly different from the standard formulation of the theorem given by

$$\lim_{r \to 0} \frac{1}{|B(x,r)|} \int_{B(x,r)} u(y) \, \mathrm{d}y = u(x)$$

for almost all $x \in \mathbb{R}^d$. In the sequel, the Lebesgue differentiation theorem in the form of Theorem 2.2 will be a valuable tool. It will serve as a replacement of continuity assumptions frequently encountered in the literature.

A well-known result says that if there is $C > 0$ and $\alpha \in (0, 1]$ such that

$$\frac{1}{|B(x,r)|} \int_{B(x,r)} |u(y) - u(x)| \, \mathrm{d}y \leq C r^\alpha$$

for all x, then u is Hölder continuous with order α. For a precise formulation, see [16, Theorem 4.3]. This might explain the heavy use of Hölder continuity assumptions in the literature on topological derivatives for perturbations in the coefficients of the differential operator. As we will show, convergence to zero as in Theorem 2.2 is enough, no faster convergence with respect to $r \searrow 0$ is needed.

It is well-known that the above theorem can be generalized to take means on sets of bounded eccentricity. Here, we will use the following modification.

Corollary 2.3. *Let $\omega \subseteq \mathbb{R}^d$ be such that there are $\rho_1, \rho_2 > 0$ with*

$$B(0, \rho_1) \subseteq \omega \subseteq B(0, \rho_2).$$

Let x be a p-Lebesgue point of u. Then

$$\lim_{r \to 0} \frac{1}{r^d |\omega|} \int_{x + r\omega} |u(y) - u(x)|^p \, \mathrm{d}y = 0.$$

Proof. The claim follows from

$$\frac{1}{r^d|\omega|}\int_{x+r\omega}|u(y)-u(x)|^p\,\mathrm{d}y \le \frac{1}{|B(x,r\rho_1)|}\int_{B(x,\rho_2 r)}|u(y)-u(x)|^p\,\mathrm{d}y \to 0,$$

see also [21, Section 3.1.2]. □

The interplay between the Lebesgue differentiation theorem and the coordinate transform (2.9) is made precise in the next result.

Corollary 2.4. *Let $u \in L^p(\Omega)$, $p \in [1,\infty)$. Let $x_0 \in \Omega$ be a p-Lebesgue point of u. Let T_r be given by (2.9). Then*

$$\lim_{r\to 0}\int_{s\omega}|u\circ T_r - u(x_0)|^p\,\mathrm{d}x' = 0$$

for all $s \ge 0$.

Proof. By elementary calculations, we get

$$\int_{s\omega}|u\circ T_r - u(x_0)|^p\,\mathrm{d}x' = r^{-d}\int_{x_0+sr\omega}|u(x)-u(x_0)|^p\,\mathrm{d}x \to 0,$$

where we have used Corollary 2.3. □

2.4 Transformed linearized state equation

In this section, we will investigate the asymptotics of $\frac{1}{r^d|\omega|}(y_r - y)$. Let us recall from (2.6) that the difference $y_r - y$ satisfies

$$\int_\Omega a_r\nabla(y_r-y)\cdot\nabla v\,\mathrm{d}x + \int_\Omega (a_r-a)\nabla y\cdot\nabla v\,\mathrm{d}x = 0 \quad \forall v \in H_0^1(\Omega).$$

Following earlier works, e.g., [22], we define

$$K_r(x') := \frac{1}{r}(y_r - y)\circ T_r(x'). \tag{2.10}$$

Due to construction, it follows $\nabla_{x'}K_r(x') = ((\nabla(y_r-y))\circ T_r)(x')$. Then K_r satisfies the transformed equation

$$\int_{T_r^{-1}(\Omega)} a_r\circ T_r\nabla_{x'}K_r\cdot\nabla_{x'}v\,\mathrm{d}x' + \int_\omega((b-a)\nabla y)\circ T_r\cdot\nabla_{x'}v\,\mathrm{d}x' = 0 \tag{2.11}$$

for all $v \in H_0^1(T_r^{-1}(\Omega))$. Note, $K_r \in H_0^1(T_r^{-1}(\Omega))$ implies that its extension by zero belongs to $H^1(\mathbb{R}^d)$ without assumptions on the regularity of Ω, see [1, Lemma 3.22]. Hence, in the sequel we will assume K_r is extended by zero to $\mathbb{R}^d$. The next goal is to pass to the limit $r \searrow 0$ in (2.11). We have the following bound of K_r.

Lemma 2.5. *Let K_r be given by (2.10). Let x_0 be a 2-Lebesgue point of $|(b-a)\nabla y|$. Then there is $M > 0$ such that*

$$r^2\|K_r\|^2_{L^2(\mathbb{R}^d)} + \|\nabla_{x'}K_r\|^2_{L^2(\mathbb{R}^d)} \le M.$$

Proof. Testing (2.6) with $y_r - y$ yields

$$\begin{aligned}\alpha\|\nabla(y_r - y)\|^2_{L^2(\Omega)}\,\mathrm{d}x &\leq -\int_\Omega (a_r - a)\nabla y \cdot \nabla(y_r - y)\,\mathrm{d}x \\ &\leq \frac{\alpha}{2}\|\nabla(y_r - y)\|^2_{L^2(\Omega)} + \frac{1}{2\alpha}\int_{x_0+r\omega} |(b-a)\nabla y|^2\,\mathrm{d}x.\end{aligned}$$

By Poincare inequality, we get

$$\|y_r - y\|^2_{L^2(\Omega)} + \|\nabla(y_r - y)\|^2_{L^2(\Omega)} \leq c\int_{x_0+r\omega} |(b-a)\nabla y|^2\,\mathrm{d}x$$

for some $c > 0$ only depending on Ω and α. Due to Theorem 2.2, we have that

$$\frac{1}{r^d|\omega|}\int_{x_0+r\omega} |(b-a)\nabla y|^2\,\mathrm{d}x$$

converges for $r \searrow 0$. This shows that there is $M > 0$ such that

$$\|y_r - y\|^2_{L^2(\Omega)} + \|\nabla(y_r - y)\|^2_{L^2(\Omega)} \leq Mr^d|\omega|$$

for all $r > 0$ sufficiently small. Applying the coordinate transform T_r to this inequality and using the definition of K_r in (2.10) proves the claim. □

Corollary 2.6. *Let K_r be given by (2.10). Let x_0 be a 2-Lebesgue point of $|(b-a)\nabla y|$. Then*

$$rK_r \rightharpoonup 0 \text{ in } L^2(\mathbb{R}^d).$$

Proof. By Lemma 2.5, we have that $(rK_r)_{r>0}$ is uniformly bounded in $H^1(\mathbb{R}^d)$ with $r\nabla_{x'}K_r \to 0$ in $L^2(\mathbb{R}^d)$ for $r \searrow 0$. Let $r_k \searrow 0$ such that $r_k K_{r_k} \rightharpoonup w$ in $L^2(\mathbb{R}^d)$. It follows $\nabla_{x'} w = 0$ in $\mathbb{R}^d$, so that w has to be equal to a constant function. Since $w \in L^2(\mathbb{R}^d)$, it follows $w = 0$. □

Let us define the coefficient function

$$\tilde{a} := \chi_\omega b + \chi_{\omega^c} a(x_0). \tag{2.12}$$

It turns out that $\tilde{a}$ is the limit of the transformed coefficients $a_r \circ T_r$ in the following sense.

Lemma 2.7. *Let x_0 be a 2-Lebesgue point of a. We have the convergence*

$$a_r \circ T_r \to \tilde{a}$$

in $L^2(n\omega)$ for all $n \in \mathbb{N}$.

Proof. Observe that $a_r \circ T_r = \tilde{a}$ on ω. Then we get

$$\int_{n\omega} |a_r \circ T_r - \tilde{a}|^2\,\mathrm{d}x' = \int_{n\omega\setminus\omega} |(a \circ T_r)(x') - a(x_0)|^2\,\mathrm{d}x',$$

and the claim is a direct consequence of Corollary 2.4. □

Let us introduce the so-called homogeneous Sobolev space (or Beppo-Levi space). Here, we define it as the space of equivalence classes under the relation $u \sim v \Leftrightarrow \nabla(u-v) = 0$. That is, the constant functions are quotiented out:

$$\dot{H}^1(\mathbb{R}^d) := \{z \in H^1_{\mathrm{loc}}(\mathbb{R}^d) : \ \nabla z \in L^2(\mathbb{R}^d)\} \ / \ \mathbb{R}.$$

It is a Hilbert space when supplied with the norm $\|z\|_{\dot{H}^1(\mathbb{R}^d)} := \|\nabla z\|_{L^2(\mathbb{R}^d)}$. In addition, equivalence classes of test functions from $C_c^\infty(\mathbb{R}^d)$ are dense in $\dot{H}^1(\mathbb{R}^d)$. For the proofs, we refer [15, Section 3]. Note that [15] denotes the Beppo-Levi space $\dot{H}^1(\mathbb{R}^d)$ by $H^1(\mathbb{R}^d)$, [15, eq. (10)].

This space is the proper space to look for the weak limit of K_r for $r \searrow 0$.

Lemma 2.8. *Let x_0 be a 2-Lebesgue point of $|(b-a)\nabla y|$ and a. Then we have $K_r \rightharpoonup K$ in $\dot{H}^1(\mathbb{R}^d)$, where K is the unique solution of*

$$\int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} K \cdot \nabla_{x'} v \,\mathrm{d}x' + (b - a(x_0)) \nabla y(x_0) \cdot \int_\omega \nabla_{x'} v \,\mathrm{d}x' = 0 \quad \forall v \in \dot{H}^1(\mathbb{R}^d). \tag{2.13}$$

Proof. Existence and uniqueness of the solution of (2.13) follows directly from the Lax-Milgram theorem. For each $r > 0$, the function K_r belongs to the spaces $H^1_0(T_r^{-1}(\Omega))$ and $H^1(\mathbb{R}^d)$, hence $K_r \in \dot{H}^1(\mathbb{R}^d)$. By Lemma 2.5, (K_r) is bounded in $\dot{H}^1(\mathbb{R}^d)$.

Let $r_k \searrow 0$ be a sequence such that $K_{r_k} \rightharpoonup \tilde{K}$ in $\dot{H}^1(\mathbb{R}^d)$. It remains to pass to the limit in the equation (2.10). Let $v \in C_c^\infty(\mathbb{R}^d)$ be given. Let $n \in \mathbb{N}$ be such that $n\omega \supseteq \operatorname{supp} v$. Let $\rho > 0$ be such that $\rho n < \operatorname{dist}(x_0, \partial\Omega)$. Then $T_\rho(\operatorname{supp} v) \subseteq \Omega$, and v can be used as test function in (2.11) if $r < \rho$. Due to Lemma 2.7, we have $a_r \circ T_r \to \tilde{a}$ in $L^2(n\omega)$. This allows us to pass to the limit in the first integral of (2.11):

$$\int_{T_{r_k}^{-1}(\Omega)} a_{r_k} \circ T_{r_k} \nabla_{x'} K_{r_k} \cdot \nabla_{x'} v \,\mathrm{d}x' \to \int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} \tilde{K} \cdot \nabla_{x'} v \,\mathrm{d}x'.$$

The convergence of the second integral follows from Corollary 2.4, and we have

$$\int_\omega ((b-a)\nabla y) \circ T_{r_k} \cdot \nabla_{x'} v \,\mathrm{d}x' \to (b - a(x_0)) \nabla y(x_0) \cdot \int_\omega \nabla_{x'} v \,\mathrm{d}x'.$$

Since test functions from $C_c^\infty(\mathbb{R}^d)$ are dense in $\dot{H}^1(\mathbb{R}^d)$, it follows that $\tilde{K}$ satisfies (2.13), hence $\tilde{K} = K$. □

The convergence of K_r to K is even strong. A similar result can be found in [6, Proposition 4.1] and [11, Theorem 4.3], where they assumed continuity of ∇y at x_0. In addition, the reference coefficient a was assumed to be constant, so that (in our notation) $a_r \circ T_r = \tilde{a}$, compare also Lemma 2.7. In order to deal with the convergence $a_r \circ T_r \to \tilde{a}$, we follow an idea of [9, Theorem 3.1].

Lemma 2.9. *Under the assumptions of Lemma 2.8, we have $K_r \to K$ in $\dot{H}^1(\mathbb{R}^d)$.*

Proof. We will use the Cholesky decomposition. Let M^+ denote the set of matrices with positive definite symmetric part, i.e.,

$$M^+ := \{A \in \mathbb{R}^{d,d} : \ \frac{1}{2}(A + A^T) \text{ positive definite } \}.$$

Then M^+ is open in $\mathbb{R}^{d,d}$. We will denote the map from the symmetric part of a matrix to its lower triangular Cholesky factor by L, i.e., $L(A)$ is such that $\frac{1}{2}(A + A^T) = L(A)L(A)^T$ for matrices $A \in M^+$. The map $A \mapsto L(A)$ is continuous from M^+ to $\mathbb{R}^{d,d}$. In addition, we have $\|L(A)\|_F^2 = \operatorname{tr}(L(A)L(A)^T) = \operatorname{tr}(A)$, so that the superposition operator induced by L is nicely behaved. In particular, it is continuous from $L^2(\mathbb{R}^d)$ to $L^4(\mathbb{R}^d)$.

Since all relevant matrices are elements of $\mathcal{M}$, see (2.2), we have the following bound on $L(A)^{-1}$. For $A \in \mathcal{M}$, we have that $\|L(A)^{-1}\|_F^2 = \operatorname{tr}(2(A + A^T)^{-1})$. The eigenvalues of $\frac{1}{2}(A + A^T)$ are bounded from below by α, so that $\|L(A)^{-1}\|_F^2 \leq d\alpha^{-1}$. In addition, $A \mapsto L(A)^{-1}$ is continuous on $\mathcal{M}$.

By extending $a_r \circ T_r$ with zero to $\mathbb{R}^d$, we can consider it as function on the domain $\mathbb{R}^d$. We are going to use the equality $a\nabla v \cdot \nabla v = \frac{1}{2}(a + a^T)\nabla v \cdot \nabla v = |L(a)^T \nabla v|^2$. Testing (2.11) with K_r, we get

$$\int_{\mathbb{R}^d} |L(a_r \circ T_r)^T \nabla_{x'} K_r|^2 \,\mathrm{d}x' = -\int_\omega ((b-a)\nabla y) \circ T_r \cdot \nabla_{x'} K_r \,\mathrm{d}x'.$$

Due to Corollary 2.4 and the weak convergence $\nabla_{x'} K_r \rightharpoonup \nabla_{x'} K$ in $L^2(\mathbb{R}^d)$, the right-hand side converges as

$$\begin{aligned} -\int_\omega ((b-a)\nabla y) \circ T_r \cdot \nabla_{x'} K_r \,\mathrm{d}x' &\to -(b - a(x_0))\nabla y(x_0) \cdot \int_\omega \nabla_{x'} K \,\mathrm{d}x' \\ &= \int_{\mathbb{R}^d} |L(\tilde{a})^T \nabla_{x'} K|^2 \,\mathrm{d}x' \end{aligned}$$

where we have used (2.13). Hence, $L(a_r \circ T_r)^T \nabla_{x'} K_r$ is bounded in $L^2(\mathbb{R}^d)$, and in addition $\|L(a_r \circ T_r)^T \nabla_{x'} K_r\|_{L^2(\mathbb{R}^d)}$ converges to $\|L(\tilde{a})^T \nabla_{x'} K\|_{L^2(\mathbb{R}^d)}$.

Due to Lemma 2.7, $L(a_r \circ T_r)^T$ converges in $L^4(n\omega)$ to $L(\tilde{a})^T$. Together with the weak convergence $\nabla_{x'} K_r \rightharpoonup \nabla_{x'} K$, we get that the weak limit of $L(a_r \circ T_r)^T \nabla_{x'} K_r$ in $L^2(\mathbb{R}^d)$ is equal to $L(\tilde{a})^T \nabla_{x'} K$. Due to the convergence of the norms, it follows

$$L(a_r \circ T_r)^T \nabla_{x'} K_r \to L(\tilde{a})^T \nabla_{x'} K \text{ in } L^2(\mathbb{R}^d).$$

We will prove the desired convergence with the celebrated dominated convergence theorem. Let $r_k \searrow 0$ be a sequence. Then after extracting a subsequence if necessary, we have the pointwise convergence $L(a_{r_k} \circ T_{r_k})^T \nabla_{x'} K_{r_k} \to L(\tilde{a})^T \nabla_{x'} K$ a.e. on $\mathbb{R}^d$. Applying a diagonal sequence argument to the result of Lemma 2.7, we can extract another subsequence (still denoted the same) such that $a_{r_k} \circ T_{r_k} \to \tilde{a}$ a.e. on $\mathbb{R}^d$. Using the continuity of $a \mapsto L(a)^{-T}$, we get the pointwise a.e. convergence

$$\nabla_{x'} K_{r_k} = L(a_{r_k} \circ T_{r_k})^{-T} L(a_{r_k} \circ T_{r_k})^T \nabla_{x'} K_{r_k} \to L(\tilde{a})^{-T} L(\tilde{a})^T \nabla_{x'} K = \nabla_{x'} K.$$

Since the L^∞-norms of $L(a_{r_k} \circ T_{r_k})^{-T}$ are uniformly bounded, the claim follows with the dominated convergence theorem applied to

$$\begin{aligned} \nabla_{x'} K_{r_k} - \nabla_{x'} K = L(a_{r_k} \circ T_{r_k})^{-T} &(L(a_r \circ T_r)^T \nabla_{x'} K_r - L(\tilde{a})^T \nabla_{x'} K) \\ &+ (L(a_{r_k} \circ T_{r_k})^{-T} L(\tilde{a})^T - I)\nabla_{x'} K. \end{aligned}$$

The convergence for $r \to 0$ follows by a standard subsequence-subsequence argument. □

2.5 Transformed adjoint equations

Let us recall the definition of the averaged adjoint $\tilde{p}_r$,

$$\int_\Omega a_r \nabla v \cdot \nabla \tilde{p}_r \,\mathrm{d}x = \frac{1}{2}\int_\Omega [(y_r - y_d) + (y - y_d)]v \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega).$$

Let us define

$$Q_r(x') := \frac{1}{r}(\tilde{p}_r - p) \circ T_r(x'), \tag{2.14}$$

where p satisfies the adjoint equation

$$\int_\Omega a \nabla v \cdot \nabla p \,\mathrm{d}x = \int_\Omega (y - y_d)v \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega).$$

Lemma 2.10. *Let x_0 be a 2-Lebesgue point of $(b-a)\nabla p$. Let Q_r be given by (2.14). Then there is $M > 0$ such that*

$$r^2 \|Q_r\|^2_{L^2(\mathbb{R}^d)} + \|\nabla_{x'} Q_r\|^2_{L^2(\mathbb{R}^d)} \le M$$

for all $r > 0$ small enough.

Proof. We proceed as in the proof of Lemma 2.5. Note that the difference $\tilde{p}_r - p$ satisfies

$$\int_\Omega a_r \nabla v \cdot \nabla(\tilde{p}_r - p)\,\mathrm{d}x + \int_\Omega (a_r - a)\nabla v \cdot \nabla p \,\mathrm{d}x = \frac{1}{2}\int_\Omega (y_r - y)v \,\mathrm{d}x \quad \forall v \in H_0^1(\Omega).$$

Testing this equation with $\tilde{p}_r - p$, we obtain using Poincare inequality and standard estimation procedures

$$\|\tilde{p}_r - p\|^2_{L^2(\Omega)} + \|\nabla(\tilde{p}_r - p)\|^2_{L^2(\Omega)} \le c\left(\int_{x_0 + r\omega} |(b-a)\nabla p|^2 \,\mathrm{d}x + \int_\Omega (y - y_r)^2 \,\mathrm{d}x\right),$$

where c is independent of r. Due to Theorem 2.2 and Lemma 2.5, the right-hand side is bounded by $Mr^d|\omega|$. The proof follows using the definition of Q_r. □

Lemma 2.11. *Let x_0 be a 2-Lebesgue point of $(b-a)\nabla p$ and a. We have $Q_r \rightharpoonup Q$ in $\dot{H}^1(\mathbb{R}^d)$, where Q is the unique solution of*

$$\int_{\mathbb{R}^d} \tilde{a}\nabla_{x'} v \cdot \nabla_{x'} Q \,\mathrm{d}x' + (b - a(x_0)) \int_\omega \nabla_{x'} v \,\mathrm{d}x' \cdot \nabla p(x_0) = 0 \quad \forall v \in \dot{H}^1(\mathbb{R}^d). \tag{2.15}$$

Proof. The proof follows the lines of the proof of Lemma 2.8. Unique solvability of (2.15) follows from Lax-Milgram theorem. After applying the coordinate transform, we find that Q_r satisfies

$$\int_{T_r^{-1}(\Omega)} a_r \circ T_r \nabla_{x'} v \cdot \nabla_{x'} Q_r \,\mathrm{d}x' + \int_\omega ((b-a)\circ T_r)\nabla_{x'} v \cdot (\nabla p) \circ T_r \,\mathrm{d}x' = \int_{T_r^{-1}(\Omega)} r K_r v \,\mathrm{d}x \tag{2.16}$$

for all $v \in H_0^1(T_r^{-1}(\Omega))$.

Let $v \in C_c^\infty(\mathbb{R}^d)$ be given. Then for $r > 0$ small enough the above equation is fulfilled. Passing to the limit in the integrals on the left-hand side of (2.16) can be done exactly as in the proof of Lemma 2.8. The integral on the right-hand side vanishes for $r \searrow 0$ due to Corollary 2.6. And the claim is proven. □

Strong convergence $Q_r \to Q$ in $\dot{H}^1(\mathbb{R}^d)$ can be proven similarly as in Lemma 2.9.

2.6 First sensitivity result

Theorem 2.12. *Let $\omega \subseteq \mathbb{R}^d$ be an open bounded set with $0 \in \omega$. Let $b \in \mathcal{M}$. Then for almost all $x_0 \in \Omega$, we have*

$$\begin{aligned}\delta J(a; b, x_0, \omega) &= \lim_{r \searrow 0} \frac{J(a_r) - J(a)}{r^d |\omega|} \\ &= -(b - a(x_0)) \nabla y(x_0) \cdot \left(\nabla p(x_0) + \frac{1}{|\omega|} \int_\omega \nabla_{x'} Q \, \mathrm{d}x' \right),\end{aligned}$$

where $Q \in \dot{H}^1(\mathbb{R}^d)$ is the solution of (2.15). The function Q depends solely on $a(x_0)$, b, $\nabla p(x_0)$, and ω.

Proof. Let x_0 be a 1-Lebesgue point of the integrable functions $(b-a)\nabla y \cdot \nabla p$, and a 2-Lebesgue point of the L^2-functions $(b-a)\nabla y$, $(b-a)\nabla p$, and a. Due to Theorem 2.2, the set of such points x_0 differs from Ω by a set of measure zero.

We will use the expansion from Lemma 2.1

$$J(a_r) - J(a) = -\int_\Omega (a_r - a) \nabla y \cdot \nabla p \, \mathrm{d}x - \int_\Omega (a_r - a) \nabla y \cdot \nabla (\tilde{p}_r - p) \, \mathrm{d}x.$$

Due to Theorem 2.2, we have

$$\lim_{r \searrow 0} \frac{1}{r^d |\omega|} \int_\Omega (a_r - a) \nabla y \cdot \nabla p \, \mathrm{d}x = (b - a(x_0)) \nabla y(x_0) \cdot \nabla p(x_0).$$

Using the coordinate transform T_r and the definition of Q_r, cf., (2.9) and (2.14), we have

$$\frac{1}{r^d |\omega|} \int_\Omega (a_r - a) \nabla y \cdot \nabla (\tilde{p}_r - p) \, \mathrm{d}x = |\omega|^{-1} \int_\omega ((a_r - a) \nabla y) \circ T_r \cdot \nabla_{x'} Q_r \, \mathrm{d}x'.$$

Due to Corollary 2.4, the term $((a_r - a)\nabla y) \circ T_r$ converges in $L^2(\omega)$ to the constant function $(b - a(x_0))\nabla y(x_0)$. By Lemma 2.11, we have $\nabla_{x'} Q_r \rightharpoonup \nabla_{x'} Q$ in $L^2(\mathbb{R}^d)$. Hence, it follows

$$\lim_{r \searrow 0} |\omega|^{-1} \int_\omega ((a_r - a) \nabla y) \circ T_r \cdot \nabla_{x'} Q_r \, \mathrm{d}x' = |\omega|^{-1} (b - a(x_0)) \nabla y(x_0) \cdot \int_\omega \nabla_{x'} Q \, \mathrm{d}x',$$

and the claim is proven. □

Using the definitions of Q and K, we have the following identity

$$\begin{aligned}(b - a(x_0)) \nabla y(x_0) \cdot \int_\omega \nabla_{x'} Q \, \mathrm{d}x' &= -\int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} K \cdot \nabla_{x'} Q \, \mathrm{d}x' \\ &= (b - a(x_0)) \int_\omega \nabla_{x'} K \, \mathrm{d}x' \cdot \nabla p(x_0),\end{aligned}$$

which can be used to equivalently rewrite the result of Theorem 2.12.

In case that $a(x_0)$ and b are multiples of the indentity matrix, the formula in Theorem 2.12 can be written in terms of the so-called polarization matrix M, i.e,

$$\delta J(a; b, x_0, \omega) = -\frac{1}{|\omega|} (b - a(x_0)) \frac{a(x_0)}{b} \nabla y(x_0) \cdot M \nabla p(x_0). \tag{2.17}$$

The scaling of the polarization matrix is not unique across the literature, here we followed [7, Theorem 2.1]. Many details on polarization matrices can be found in the monograph [3].

2.7 Sensitivity matrix

For fixed $a(x_0)$ and b, the mapping $\nabla p(x_0) \mapsto \int_\omega \nabla_{x'} Q \, \mathrm{d}x'$ is a linear mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$. Hence, the mapping $(\nabla y(x_0), \nabla p(x_0)) \mapsto (b - a(x_0))\nabla y(x_0) \cdot \int_\omega \nabla_{x'} Q \, \mathrm{d}x'$ can be written in terms of a matrix multiplication.

Given $p \in \mathbb{R}^d$ let us define $Q_p \in \dot{H}^1(\mathbb{R}^d)$ as the solution of

$$\int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} v \cdot \nabla_{x'} Q \, \mathrm{d}x' + (b - a(x_0)) \int_\omega \nabla_{x'} v \, \mathrm{d}x' \cdot p = 0 \quad \forall v \in \dot{H}^1(\mathbb{R}^d). \tag{2.18}$$

Let us define the matrix $R \in \mathbb{R}^{d,d}$ by

$$p^T R y := -(b - a(x_0)) y \cdot \int_\omega \nabla_{x'} Q_p \, \mathrm{d}x'. \tag{2.19}$$

Of course, R depends on the coefficients $a(x_0)$ and b. The matrix R has the following properties.

Lemma 2.13. *Suppose $a(x_0)$ and b are symmetric. Then R defined in (2.19) is symmetric.*

Proof. Let $Q_y \in \dot{H}^1(\mathbb{R}^d)$ be the solution of

$$\int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} v \cdot \nabla_{x'} Q \, \mathrm{d}x' + (b - a(x_0)) \int_\omega \nabla_{x'} v \, \mathrm{d}x' \cdot y = 0 \quad \forall v \in \dot{H}^1(\mathbb{R}^d).$$

Then using the symmetry assumption and testing the equation for Q_y with Q_p results in

$$\begin{aligned} p^T R y &= -(b - a(x_0)) y \cdot \int_\omega \nabla_{x'} Q_p \, \mathrm{d}x' \\ &= \int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} Q_p \cdot \nabla_{x'} Q_y \, \mathrm{d}x' \\ &= -(b - a(x_0)) \int_\omega \nabla_{x'} Q_y \, \mathrm{d}x' \cdot p \\ &= y^T R p, \end{aligned}$$

which proves the symmetry of R. □

Lemma 2.14. *Suppose $b - a(x_0)$ is symmetric. Then $p^T R p \geq 0$ forall p, and $p^T R p = 0$ if and only if $(b - a(x_0))p = 0$.*

Proof. Take $p \in \mathbb{R}^d$. Since $b - a(x_0)$ is symmetric, it follows

$$-(b - a(x_0))p \cdot \int_\omega \nabla_{x'} v = -(b - a(x_0)) \int_\omega \nabla_{x'} v \, \mathrm{d}x' \cdot p = \int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} v \cdot \nabla_{x'} Q_p \, \mathrm{d}x'$$

for all $v \in \dot{H}^1(\mathbb{R}^d)$, which implies

$$\begin{aligned} p^T R p &= -(b - a(x_0))p \cdot \int_\omega \nabla_{x'} Q_p \, \mathrm{d}x' \\ &= \int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} Q_p \cdot \nabla_{x'} Q_p \, \mathrm{d}x' \geq \alpha \|\nabla_{x'} Q_p\|^2_{L^2(\mathbb{R}^d)} \geq 0 \end{aligned}$$

since $\tilde{a}(x') \in \mathcal{M}$ for almost all x'.

Clearly, $p^T R p = 0$ if $(b - a(x_0))p = 0$. Let us assume $p^T R p = 0$. This implies $\nabla_{x'} Q_p = 0$ and $(b - a(x_0)) \int_\omega \nabla_{x'} v \, \mathrm{d}x' \cdot p = 0$ forall $v \in \dot{H}^1(\mathbb{R}^d)$. Setting $v(x') := p^T (b - a(x_0))x'$ proves $(b - a(x_0))p = 0$, and definiteness of R is established as claimed. □

For anisotropic variants of the polarization matrix introduced in (2.17), we refer to [3, Section 4.12] and [12].

3 Perturbations on balls and ellipses

Here, we will work in the isotropic case, that is, $a(x_0)$ and b are assumed to be positive multiples of the identity matrix. In this case, we get explicit expressions for the variation $\delta J(a; b, x_0, \omega)$. With a little abuse of notation, we will assume

$$a(x_0), b \in \mathbb{R}, \ a(x_0), b \geq \alpha.$$

Polarization matrices for the anisotropic case and $d = 2, 3$ were computed in [12].

3.1 Balls

Here, we will set

$$\omega := B(0, 1).$$

Interestingly, in this case the solutions of (2.13) and (2.15) can be computed explicitly.

Lemma 3.1. *Let $\omega := B(0,1)$. Let $g \in \mathbb{R}^d$ be given, $a(x_0), b \in \mathbb{R}$. Then the solution $G \in \dot{H}^1(\mathbb{R}^d)$ of the equation*

$$\int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} G \cdot \nabla_{x'} v \, \mathrm{d}x' + g \cdot \int_\omega \nabla_{x'} v \, \mathrm{d}x' = 0 \quad \forall v \in \dot{H}^1(\mathbb{R}^d).$$

is given by

$$G(x') := g \cdot x' \frac{1}{\max(1, |x'|^d)} \cdot \frac{-1}{b + a(x_0)(d-1)}.$$

Proof. Note that $\Delta G(x') = 0$ for all x' with $|x'| \neq 1$. Define

$$G_0(x') = g \cdot x' \frac{1}{\max(1, |x'|^d)}.$$

Let $v \in C_c^\infty(\mathbb{R}^d)$. Then

$$\begin{aligned}
\int_{\mathbb{R}^d} \tilde{a} \nabla_{x'} G_0 \cdot \nabla_{x'} v \, \mathrm{d}x' &= \int_{\omega^c} a(x_0) \nabla_{x'} G_0 \cdot \nabla_{x'} v \, \mathrm{d}x' + \int_\omega b \nabla_{x'} G_0 \cdot \nabla_{x'} v \, \mathrm{d}x' \\
&= \int_{\partial \omega^c} a(x_0) \nabla_{x'} G_0 \cdot (-x') \, v \, \mathrm{d}x' + \int_\omega b \nabla_{x'} G_0 \cdot \nabla_{x'} v \, \mathrm{d}x' \\
&= \int_{\partial \omega^c} a(x_0)(g - d(g \cdot x')x') \cdot (-x') \, v \, \mathrm{d}x' + \int_{\partial \omega} bg \cdot x' v \, \mathrm{d}x' \\
&= \int_{\partial \omega} (b + a(x_0)(d-1)) g \cdot x' v \, \mathrm{d}x'
\end{aligned}$$

and

$$g \cdot \int_\omega \nabla_{x'} v \, \mathrm{d}x' = \int_\omega \nabla_{x'}(g \cdot x') \cdot \nabla_{x'} v \, \mathrm{d}x' = 0 + \int_{\partial\omega} g \cdot x' v \, \mathrm{d}x.$$

And G satisfies the integral equation when tested with test functions from $C_c^\infty(\mathbb{R}^d)$. Let us argue that $G \in \dot{H}^1(\mathbb{R}^d)$. For $d > 1$, we have $|\nabla G| \sim |x|^{-d}$ for $|x| > 1$. For $d = 1$, it holds $\nabla G(x) = 0$ for $|x| > 1$. It follows $\nabla G \in L^2(\mathbb{R}^d)$. By density, G satisfies the equation for all test functions. □

Theorem 3.2. *Let $a(x_0), b \in \mathcal{M}$ be multiples of the identity matrix. Then for almost all $x_0 \in \Omega$, we have*

$$\delta J(a; b, x_0, B(0,1)) = -\nabla y(x_0) \cdot \nabla p(x_0) \frac{a(x_0) d}{b + a(x_0)(d-1)} (b - a(x_0)).$$

Proof. Due to Lemma 3.1 and (2.15), we have for $x' \in \omega$

$$Q(x') = (b - a(x_0)) \nabla p(x_0) \cdot x' \cdot \frac{-1}{b + a(x_0)(d-1)},$$

which implies

$$\int_\omega \nabla_{x'} Q_r \, \mathrm{d}x = (b - a(x_0)) \nabla p(x_0) |\omega| \cdot \frac{-1}{b + a(x_0)(d-1)}.$$

Using this identity in the result of Theorem 2.12, we find

$$\begin{aligned} \lim_{r \to 0} \frac{1}{|B(r)|} (J(y_r) - J(y)) &= -(b - a(x_0)) \nabla y(x_0) \cdot \nabla p(x_0) \cdot \left(1 - \frac{b - a(x_0)}{b + a(x_0)(d-1)}\right) \\ &= -\nabla y(x_0) \cdot \nabla p(x_0) \frac{a(x_0) d}{b + a(x_0)(d-1)} (b - a(x_0)). \end{aligned}$$

□

This result coincides with those of [4, Theorem 6.1], [8, Theorem 3.1], which were derived under much stronger assumptions.

3.2 Ellipses

Now let $H \in \mathbb{R}^{2,2}$ be symmetric, positive definite with $\det H = 1$. Then

$$\omega = \{x \in \mathbb{R}^2 : \ x^T H x \leq 1\} \tag{3.1}$$

is an ellipse with $|\omega| = |B(1)|$. We will now study perturbations on elliptic sets (instead of on balls). We restrict to the 2d case, where explicit formulas for the polarization matrix are available from [7, 12]. Such explicit formulas are available for $d = 3$ as well [12], but are much more technical to analyze.

Here, we have the following result for axis-aligned ellipses. It was derived in [7] using an explicitly constructed solution to the equation (2.15) in terms of elliptic coordinates.

Lemma 3.3. *Let $H = \mathrm{diag}(\lambda, \lambda^{-1}) \in \mathbb{R}^{2,2}$ with $\lambda > 0$ be given, and define ω as in (3.1). Let $a(x_0), b \in \mathcal{M}$ be multiples of the identity matrix. Then for almost all $x_0 \in \Omega$, we have*

$$\delta J(a; b, x_0, \omega) = -(b - a(x_0)) a(x_0) \nabla p(x_0) \cdot \begin{pmatrix} \frac{\lambda+1}{a(x_0)\lambda + b} & 0 \\ 0 & \frac{\lambda+1}{a(x_0) + b\lambda} \end{pmatrix} \nabla y(x_0).$$

Proof. From Theorem 2.12, we get

$$\delta J(a;b,x_0,\omega) = -(b-a(x_0))\nabla y(x_0)\cdot\left(\nabla p(x_0) - \int_\omega \nabla_{x'} Q\,\mathrm{d}x'\right),$$

where Q solves (2.15). Using the polarization matrix M, cf., (2.17), we can write

$$\delta J(a;b,x_0,\omega) = -\frac{1}{|\omega|}(b-a(x_0))\frac{a(x_0)}{b}\nabla y(x_0)\cdot M\nabla p(x_0).$$

The matrix M was computed in [7]. Using [7, eq. (A.4)] with $\kappa := b$, $\gamma := a(x_0)$, $a := \lambda^{1/2}$, $b := \lambda^{-1/2}$, $|\omega| = 1$, we find

$$M = |\omega|\begin{pmatrix} \frac{\kappa(a+b)}{\gamma a+\kappa b} & 0 \\ 0 & \frac{\kappa(a+b)}{\gamma b+\kappa a}\end{pmatrix} = \begin{pmatrix} \frac{b(\lambda+1)}{a(x_0)\lambda+b} & 0 \\ 0 & \frac{b(\lambda+1)}{a(x_0)+b\lambda}\end{pmatrix},$$

and the claim follows. □

Note that in the case of $H = I_2$ and $\lambda = \lambda^{-1} = 1$ this reduces to the result of Theorem 3.2.

Corollary 3.4. *Let $H = R^T\operatorname{diag}(\lambda,\lambda^{-1})R \in \mathbb{R}^{2,2}$ with $\lambda > 0$, and $R^TR = I_2$, and define ω as in (3.1). Let $a(x_0), b \in \mathcal{M}$ be multiples of the identity matrix. Then for almost all $x_0 \in \Omega$, we have*

$$\delta J(a;b,x_0,\omega) = \\ -(b-a(x_0))a(x_0)\nabla p(x_0)\cdot R^T\begin{pmatrix} \frac{\lambda+1}{a(x_0)\lambda+b} & 0 \\ 0 & \frac{\lambda+1}{a(x_0)+b\lambda}\end{pmatrix}R\nabla y(x_0).$$

for almost all x_0.

Proof. Define $\omega' := \{x : \ x^T\operatorname{diag}(\lambda,\lambda^{-1})x \le 1\}$, which implies $\omega = R^T\omega'$. If M_ω and $M_{\omega'}$ are the polarization matrices associated to ω and ω', then it holds $M_\omega = R^TM_{\omega'}R$ by [3, Lemma 4.5]. Using the matrix M from the proof of Lemma 3.3, the claim follows. □

We will now compute the range of

$$\omega \mapsto \delta J(a;b,x_0,\omega),$$

where ω varies over elliptic shapes as considered in the above results. Similar considerations are done in [5] with a different approach and different notation.

Lemma 3.5. *Let $y, p \in \mathbb{R}^2$ be given. Let $D = \operatorname{diag}(\lambda_1,\lambda_2)$ be a diagonal matrix. For $R \in \mathbb{R}^{2,2}$ define*

$$F(R) := p^TR^TDRy.$$

Then the range of F is given by

$$F(\mathrm{O}(2)) = F(\mathrm{SO}(2)) = \frac{\lambda_1+\lambda_2}{2}y^Tp + [-1,+1]\cdot\frac{|\lambda_1-\lambda_2|}{2}\|y\|_2\|p\|_2,$$

where $\mathrm{O}(2) = \{R \in \mathbb{R}^{2,2} : \ R^TR = I_2\}$, $\mathrm{SO}(2) = \{R \in \mathrm{O}(2) : \det(R) = 1\}$.

Proof. Wlog we can assume $\|y\|_2 = \|p\|_2 = 1$. Let $R \in \mathrm{O}(2)$ with $\det(R) = -1$ be given, and set $S := \mathrm{diag}(1, -1)$. Then $SR \in \mathrm{SO}(2)$ and $F(R) = F(SR)$ since $SDS = D$. Now, define R_0 to be the rotation matrix that rotates p onto the first unit vector e_1, i.e.,

$$R_0 = \begin{pmatrix} p_1 & p_2 \\ -p_2 & p_1 \end{pmatrix}.$$

Set $\tilde{y} := R_0 y$. Then it is enough to compute the range of $\tilde{F}(R) := F(RR_0)$. Given $v \in \mathbb{R}^2$ with $\|v\|_2 = 1$, we parametrize R as follows

$$R(v) := \begin{pmatrix} v_1 & v_2 \\ -v_2 & v_1 \end{pmatrix} = v_1 I_2 + v_2 J, \quad J := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then we get

$$\begin{aligned} \tilde{F}(R(v)) &= e_1^T (v_1 I_2 + v_2 J)^T D (v_1 I_2 + v_2 J) \tilde{y} \\ &= v^T \begin{pmatrix} \lambda_1 \tilde{y}_1 & \frac{\lambda_1 - \lambda_2}{2} \tilde{y}_2 \\ \frac{\lambda_1 - \lambda_2}{2} \tilde{y}_2 & \lambda_2 \tilde{y}_1 \end{pmatrix} v. \qquad =: v^T M v \end{aligned}$$

Hence, the range of $v \mapsto F(R(v))$ is equal to the interval determined by the eigenvalues of the matrix M. A short calulation shows that the eigenvalues of M are given by

$$t_{1,2} := \frac{\lambda_1 + \lambda_2}{2} \tilde{y}_1 \pm \frac{|\lambda_1 - \lambda_2|}{2}.$$

Since $\tilde{y}_1 = y^T p$, the claim follows. □

Lemma 3.6. *Let $y, p \in \mathbb{R}^2$ and $a, b > 0$ be given. For $R \in \mathbb{R}^{2,2}$ and $\lambda > 0$ define*

$$G(R, \lambda) := p^T R^T \begin{pmatrix} \frac{\lambda+1}{a\lambda+b} & 0 \\ 0 & \frac{\lambda+1}{a+b\lambda} \end{pmatrix} Ry.$$

Then

$$\mathrm{cl}\, G(\mathrm{O}(2),\ \mathbb{R}^+) = \frac{1}{2}\left(\frac{1}{a} + \frac{1}{b}\right) y^T p + [-1, +1] \cdot \frac{1}{2}\left(\frac{1}{a} - \frac{1}{b}\right) \cdot \|y\|_2 \|p\|_2.$$

Proof. Wlog we can assume $\|y\|_2 = \|p\|_2 = 1$. In addition, we can assume $b \geq a$. In case $b < a$, we can consider the function $G(R, \lambda^{-1})$, which is equal to $G(R, \lambda)$ but with the roles of a, b exchanged. Define

$$\lambda_1(\lambda) := \frac{\lambda + 1}{a\lambda + b},\ \lambda_2(\lambda) := \frac{\lambda + 1}{a + b\lambda}.$$

Note that $\lambda_2(\lambda) = \lambda_1(\lambda^{-1})$. Since $b > a$ it follows that λ_1 is monotonically increasing, and λ_2 is monotonically decreasing. This implies that $|\lambda_1(\lambda) - \lambda_2(\lambda)|$ is maximal at $\lambda = 0$ and $\lambda \to \infty$, with maximum $\frac{1}{a} - \frac{1}{b}$.

In addition, we find

$$\lambda_1(\lambda) + \lambda_2(\lambda) = (a + b) \frac{(\lambda + 1)^2}{ab\lambda^2 + (a^2 + b^2)\lambda + ab}$$

with derivative

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}(\lambda_1(\lambda)+\lambda_2(\lambda)) = (a+b)(a-b)^2\frac{\lambda^2-1}{(a\lambda+b)^2(a+b\lambda)^2}.$$

Hence, the minimum and maximum of $\lambda_1(\lambda)+\lambda_2(\lambda)$ are attained at $\lambda=1$ and $\lambda=0$, respectively, which means

$$\lambda_1(1)+\lambda_2(1) = \frac{4}{a+b} \leq \lambda_1(\lambda)+\lambda_2(\lambda) \leq \frac{1}{a}+\frac{1}{b} = \lambda_1(0)+\lambda_2(0). \tag{3.2}$$

By Lemma 3.5, we have

$$G(\mathrm{O}(2),\lambda) = \frac{\lambda_1(\lambda)+\lambda_2(\lambda)}{2}y^Tp + \left[-\frac{|\lambda_1(\lambda)-\lambda_2(\lambda)|}{2}, +\frac{|\lambda_1(\lambda)-\lambda_2(\lambda)|}{2}\right]. \tag{3.3}$$

Let us only consider the case $y^Tp \geq 0$, the case $y^Tp < 0$ can be proven by a simple change of sign. Then using that the supremum of both addends is attained at $\lambda=0$, see (3.2), we get

$$\begin{aligned}\sup_{\lambda>0}\sup(G(\mathrm{O}(2),\lambda)) &= \sup_{\lambda>0}\frac{\lambda_1(\lambda)+\lambda_2(\lambda)}{2}y^Tp + \frac{|\lambda_1(\lambda)-\lambda_2(\lambda)|}{2}\\ &= \frac{(\lambda_1(0)+\lambda_2(0))}{2}y^Tp + \frac{|\lambda_1(0)-\lambda_2(0)|}{2}\\ &= \frac{1}{2}\left(\frac{1}{a}+\frac{1}{b}\right)y^Tp + \frac{1}{2}\left(\frac{1}{a}-\frac{1}{b}\right).\end{aligned}$$

To compute the infimum, we observe that the lower bound in (3.3) is invariant under the transform $\lambda \mapsto \lambda^{-1}$. Hence, it is sufficient to consider $\lambda \geq 1$. Here, we find

$$\begin{aligned}\inf_{\lambda>0}\inf(G(\mathrm{O}(2),\lambda)) &= \inf_{\lambda\geq 1}\inf(G(\mathrm{O}(2),\lambda))\\ &= \inf_{\lambda\geq 1}\frac{\lambda_1(\lambda)+\lambda_2(\lambda)}{2}y^Tp - \frac{\lambda_1(\lambda)-\lambda_2(\lambda)}{2}\\ &= \frac{1}{2}\inf_{\lambda\geq 1}\Big(\lambda_1(\lambda)\underbrace{(y^Tp-1)}_{\leq 0} + \lambda_2(\lambda)\underbrace{(y^Tp+1)}_{\geq 0}\Big)\\ &= \frac{1}{2}\lim_{\lambda\to\infty}\big(\lambda_1(\lambda)(y^Tp-1)+\lambda_2(\lambda)(y^Tp+1)\big)\\ &= \frac{1}{2}\left(\frac{1}{a}+\frac{1}{b}\right)y^Tp - \frac{1}{2}\left(\frac{1}{a}-\frac{1}{b}\right)\end{aligned}$$

due to the monotonicity properties of λ_1 and λ_2. And the claim is proven. □

Using these results, we obtain the following statement about the range of the topol derivative, if the shape of the perturbation ω varies over all possible ellipses. The infimum in the next result will be useful for necessary optimality conditions.

Theorem 3.7. *Let $a(x_0), b \in \mathcal{M}$ be multiples of the identity matrix. Let us define*

$$\begin{aligned}\mathcal{H} &:= \{H \in \mathbb{R}^{2,2} \text{ positive definite with } \det H = 1\}\\ \omega(H) &:= \{x : \ x^THx \leq 1\}.\end{aligned}$$

Then for almost all $x_0 \in \Omega$, the closure of the range of δJ with respect to variations of ellipses ω is given by

$$\begin{aligned}&\operatorname{cl}\{\delta J(a;b,x_0,\omega(H)) : H \in \mathcal{H}\}\\ &\quad = -(b-a(x_0))\nabla y(x_0)\cdot\nabla p(x_0)\\ &\qquad + \frac{1}{2}\frac{(b-a(x_0))^2}{b}\left([-1,+1]\cdot\|\nabla y(x_0)\|_2\|\nabla p(x_0)\|_2 + \nabla y(x_0)\cdot\nabla p(x_0)\right).\end{aligned}$$

In particular,

$$\begin{aligned}&\inf\{\delta J(a;b,x_0,\omega(H)) : H \in \mathcal{H}\}\\ &\qquad = -(b-a(x_0))\nabla y(x_0)\cdot\nabla p(x_0)\\ &\qquad\quad + \frac{1}{2}\frac{(b-a(x_0))^2}{b}\left(\nabla y(x_0)\cdot\nabla p(x_0) - \|\nabla y(x_0)\|_2\|\nabla p(x_0)\|_2\right).\end{aligned}$$

Proof. First, we apply Corollary 3.4 to matrices $H \in \mathcal{H}$ with rational eigenvalues and eigenvectors. Such matrices are dense in $\mathcal{H}$. Then for almost all $x_0 \in \Omega$ it holds

$$\operatorname{cl}\{\delta J(a;b,x_0,\omega(H)) : H \in \mathcal{H}\} = -(b-a(x_0))a(x_0)\operatorname{cl} G(\mathrm{O}(2),\ \mathbb{R}^+, x_0),$$

where G is defined by

$$G(R,\lambda,x_0) := \nabla p(x_0)^T R^T \begin{pmatrix} \frac{\lambda+1}{a(x_0)\lambda+b} & 0 \\ 0 & \frac{\lambda+1}{a(x_0)+b\lambda}\end{pmatrix} R\nabla y(x_0).$$

Due to Lemma 3.6, we get

$$\begin{aligned}\operatorname{cl} G(\mathrm{O}(2),\ \mathbb{R}^+, x_0) &= \frac{1}{2}\left(\frac{1}{a(x_0)}+\frac{1}{b}\right)\nabla y(x_0)\cdot\nabla p(x_0)\\ &\quad + [-1,+1]\cdot\frac{1}{2}\left(\frac{1}{a(x_0)}-\frac{1}{b}\right)\cdot\|\nabla y(x_0)\|_2\|\nabla p(x_0)\|_2.\end{aligned}$$

The claim follows now from elementary computations. □

4 Control in the coefficients

In this section, we will apply the results on the δJ to derive the Pontryagin maximum principle for the following optimization problem: Minimize

$$\tilde{J}(y,a) := \frac{1}{2}\|y-y_d\|^2_{L^2(\Omega)} + \int_\Omega g(a(x))\,\mathrm{d}x, \tag{4.1}$$

subject to $a \in \mathcal{A}$, and $y \in H^1_0(\Omega)$ solves

$$\int_\Omega a\nabla y\cdot\nabla v\,\mathrm{d}x = \int_\Omega fv\,\mathrm{d}x \quad \forall v \in H^1_0(\Omega).$$

This is a classical problem, and lead to the study of H-convergence, [14]. Solutions do not exist in general. The data of the problem is assumed to satisfy the

basic Assumption 1. In addition, $g : \mathbb{R}^{2,2} \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous, which implies that $g(a)$ is measurable for all $a \in \mathcal{A}$.

We will formulate necessary optimality conditions in terms of the Pontryagin maximum principle. Using the results from the previous sections, we find that the maximum principle holds in the following form.

Theorem 4.1. *Let a be a local solution of (4.1) with respect to the $L^1(\Omega)$-norm. Let y and p be the corresponding solution of the state equation (2.3) and (2.8). Let $\omega \subseteq \mathbb{R}^d$ be open and bounded with $0 \in \omega$. Then for almost all $x_0 \in \Omega$ and all $b \in \mathcal{M}$ (see (2.2)) it holds*

$$-(b - a(x_0))\nabla y(x_0) \cdot \left(\nabla p(x_0) + \frac{1}{|\omega|} \int_\omega \nabla_{x'} Q \, \mathrm{d}x' \right) + g(b) - g(a(x_0)) \geq 0,$$

where $Q \in \dot{H}^1(\mathbb{R}^d)$ is the solution of (2.15).

Proof. Fix $b \in \mathcal{M}$ and $x_0 \in \Omega$. Define a_r as in (2.4) and y_r as in (2.5). Then $a_r \to a$ in $L^1(\Omega)$ for $r \searrow 0$. Hence, $\tilde{J}(y_r, a_r) - \tilde{J}(y, a) \geq 0$ for all $r > 0$ small enough. Using Theorem 2.12 and Theorem 2.2, we get

$$\begin{aligned} 0 &\leq \lim_{r \searrow 0} \frac{1}{r^d |\omega|} (\tilde{J}(y_r, a_r) - \tilde{J}(y, a)) \\ &= -(b - a(x_0))\nabla y(x_0) \cdot \left(\nabla p(x_0) + \frac{1}{|\omega|} \int_\omega \nabla_{x'} Q \, \mathrm{d}x' \right) + g(b) - g(a(x_0)), \end{aligned}$$

where the limit exists for almost all $x_0 \in \Omega$. Let G be a countable and dense subset of $\operatorname{gph} g = \{(b, g(b)) : b \in \mathcal{M}\}$. Then using the above arguments, the claim follows for all $(b, g(b)) \in G$. By continuity, the claim holds for all $b \in \mathcal{M}$. □

4.1 The scalar case

Now let us investigate the case when a is a multiple of the identity. That is, we will prove necessary optimality condition for the problem

$$\min \tilde{J}(y, a) \text{ subject to } a(x) \in \mathcal{M} \cap \operatorname{span}(I_d) \text{ f.a.a. } x. \tag{4.2}$$

With little abuse of notation, we will consider now $a : \Omega \to \mathbb{R}$ and $b \in \mathbb{R}$. Using the explicit expression from Theorem 3.2, we get the following form of the maximum principle:

Theorem 4.2. *Let a be a local solution of (4.2) with respect to the $L^1(\Omega)$-norm. Let y and p be the corresponding solutions of the state equation (2.3) and adjoint equation (2.8). Then for almost all $x_0 \in \Omega$ and all $b \geq \alpha$ (see (2.2)) it holds*

$$-(b - a(x_0))\nabla y(x_0) \cdot \nabla p(x_0) \frac{a(x_0) d}{b + a(x_0)(d - 1)} + g(b) - g(a(x_0)) \geq 0. \tag{4.3}$$

Proof. The proof is similar to that of Theorem 4.1 and uses Theorem 3.2 for the expression of δJ. □

4.2 The scalar case, $d = 2$

In the two-dimensional case, we have the following expression for the maximum principle associated with (4.2).

Theorem 4.3. *Let a be a local solution of (4.2) with respect to the $L^1(\Omega)$-norm. Let y and p be the corresponding solutions of the state equation (2.3) and adjoint equation (2.8). Then for almost all $x_0 \in \Omega$ and all $b \geq \alpha$ (see (2.2)) it holds*

$$-(b - a(x_0))\nabla y(x_0) \cdot \nabla p(x_0) + g(b) - g(a(x_0)) + \frac{1}{2}\frac{(b - a(x_0))^2}{b}\left(\nabla y(x_0) \cdot \nabla p(x_0) - \|\nabla y(x_0)\|_2 \|\nabla p(x_0)\|_2\right) \geq 0. \quad (4.4)$$

Proof. Follows from Theorem 3.7. □

Let us remark that the conclusion of Theorem 4.3 is stronger than that of Theorem 4.2.

Corollary 4.4. *Let a be feasible for (4.2). Let y and p be the corresponding solutions of the state equation (2.3) and adjoint equation (2.8). Suppose that for almost all $x_0 \in \Omega$ and all $b \geq \alpha$ (see (2.2)) the inequality (4.4) is satisfied. Then (4.3) is true for almost all $x_0 \in \Omega$ and all $b \geq \alpha$.*

Proof. Clearly, (4.4) implies

$$-(b - a(x_0))\nabla y(x_0) \cdot \nabla p(x_0) + g(b) - g(a(x_0)) \geq 0.$$

Since the coeffiecients of $\nabla y(x_0) \cdot \nabla p(x_0)$ in this inequality and in (4.3) satisfy

$$-(b - a(x_0))\left(\frac{a(x_0)d}{b + a(x_0)(d-1)} - 1\right) = -(b - a(x_0))\frac{-(b - a(x_0))}{b + a(x_0)(d-1)} \geq 0,$$

the inequality (4.3) follows as claimed. □

4.3 Relation to Fréchet derivative

The mapping $a \mapsto y$ is Fréchet differentiable from $L^\infty(\Omega)$ to $H_0^1(\Omega)$. If a is a local minimum of (4.2) and g is continuously differentiable, then a satisfies the necessary optimality condition

$$(-\nabla y(x_0) \cdot \nabla p(x_0) + g'(a))(b - a) \geq 0 \quad \forall b \geq \alpha. \quad (4.5)$$

Naturally, the results of Theorem 4.2 and Theorem 4.3 are stronger: replacing b by $a(x_0) + t(b - a(x_0))$ in those inequalities, dividing by t, and passing to the limit $t \searrow 0$ yields (4.5).

4.4 Example with linear g

Let us consider the following example, which is motivated by material optimization problems, [5]. We consider the feasible set given by

$$\mathcal{M} := \{aI_d : \ a \in [\alpha, \beta]\},$$

where $0 < \alpha < \beta$ are real numbers. In addition, we chose

$$g(a) := \ell \cdot a,$$

where $\ell \in \mathbb{R}$, which could be used to model the cost of materials. In this simplified situation, we can analyze the optimality conditions (4.4) and (4.5). Here, we have the following characterization of solutions of the variational inequality (4.4). This should be compared to the characterization of solutions of the condition (4.5). which is given by

$$\begin{aligned} \ell > \nabla y(x_0) \cdot \nabla p(x_0) \;&\Rightarrow\; a(x_0) = \alpha \;\Rightarrow\; \ell - \nabla y(x_0) \cdot \nabla p(x_0) \geq 0, \\ \ell < \nabla y(x_0) \cdot \nabla p(x_0) \;&\Rightarrow\; a(x_0) = \beta \;\Rightarrow\; \ell - \nabla y(x_0) \cdot \nabla p(x_0) \leq 0. \end{aligned}$$

Corollary 4.5. *Let a be feasible for (4.2). Let y and p be the corresponding solutions of the state equation (2.3) and adjoint equation (2.8).*

Suppose that for almost all $x_0 \in \Omega$ and all $b \geq \alpha$ (see (2.2)) the inequality (4.4) is satisfied. Then for almost all x_0 the following implications hold:

$$\begin{aligned} \ell = \nabla y(x_0) \cdot \nabla p(x_0) \;&\Rightarrow\; \|\nabla y(x_0)\|_2 \|\nabla p(x_0)\|_2 = \nabla y(x_0) \cdot \nabla p(x_0) \\ \ell > \nabla y(x_0) \cdot \nabla p(x_0) \;&\Rightarrow\; a(x_0) = \alpha \\ \ell < \nabla y(x_0) \cdot \nabla p(x_0) \;&\Rightarrow\; a(x_0) = \beta. \end{aligned}$$

In addition, if $a(x_0) = \alpha$ then

$$\begin{aligned} &\ell - \nabla y(x_0) \cdot \nabla p(x_0) \\ &\qquad \geq \frac{1}{2} \frac{\beta - \alpha}{\beta} (\|\nabla y(x_0)\|_2 \|\nabla p(x_0)\|_2 - \nabla y(x_0) \cdot \nabla p(x_0)) \geq 0. \end{aligned}$$

If $a(x_0) = \beta$ then

$$\begin{aligned} &\ell - \nabla y(x_0) \cdot \nabla p(x_0) \\ &\qquad \leq \frac{1}{2} \frac{\alpha - \beta}{\alpha} (\|\nabla y(x_0)\|_2 \|\nabla p(x_0)\|_2 - \nabla y(x_0) \cdot \nabla p(x_0)) \leq 0. \end{aligned}$$

Proof. Let us define for abbreviation

$$s(x_0) := \nabla y(x_0) \cdot \nabla p(x_0), \quad n(x_0) := \|\nabla y(x_0)\|_2 \|\nabla p(x_0)\|_2,$$

which implies $|s| \leq n$. Then (4.4) is equivalent to

$$-(b - a(x_0))s(x_0) + \ell(b - a(x_0)) + \frac{1}{2} \frac{(b - a(x_0))^2}{b} (s(x_0) - n(x_0)) \geq 0$$

and

$$(b - a(x_0))(\ell - s(x_0)) \geq \frac{1}{2} \frac{(b - a(x_0))^2}{b} (n(x_0) - s(x_0)) \quad \forall b \in [\alpha, \beta]. \tag{4.6}$$

Now let us assume that (4.4) is true, and hence the inequality is satisfied for almost all x_0. In case, $\ell - s(x_0) = 0$ it follows $n(x_0) = s(x_0)$.

If $\ell - s(x_0) \neq 0$ it follows $a(x_0) \in \{\alpha, \beta\}$ as the left-hand side of (4.6) changes sign at $b = a(x_0)$, while the right-hand side is non-negative. Suppose $a(x_0) = \alpha$. Then (4.6) implies

$$\ell - s(x_0) \geq \frac{1}{2} \frac{\beta - \alpha}{\beta} (n(x_0) - s(x_0)).$$

If $a(x_0) = \beta$ we get the reverse inequality

$$\ell - s(x_0) \leq \frac{1}{2} \frac{\alpha - \beta}{\alpha} (n(x_0) - s(x_0)).$$

□

5 Conclusion and outlook

We developed the Pontryagin maximum principle for control in the coefficients using quite elementary methods. It would be interesting to consider more complicated settings using general cost functionals semilinar or quasilinear equations. Also the case $b = 0$ could be considered following [6]. Another interesting question is, whether the Ekeland variational principle could be used to prove existence of ϵ-solutions of the Pontryagin maximum principle. The maximum principle in the $2d$ case was written in terms of a variational inequality (4.4) of a new type, whose solution theory is completely open.

References

[1] Robert A. Adams. *Sobolev spaces*. Pure and Applied Mathematics, Vol. 65. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975.

[2] Grégoire Allaire. *Shape optimization by the homogenization method*, volume 146 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2002.

[3] Habib Ammari and Hyeonbae Kang. *Polarization and moment tensors*, volume 162 of *Applied Mathematical Sciences*. Springer, New York, 2007. With applications to inverse problems and effective medium theory.

[4] Samuel Amstutz. Sensitivity analysis with respect to a local perturbation of the material property. *Asymptot. Anal.*, 49(1-2):87–108, 2006.

[5] Samuel Amstutz. Connections between topological sensitivity analysis and material interpolation schemes in topology optimization. *Struct. Multidiscip. Optim.*, 43(6):755–765, 2011.

[6] Samuel Amstutz. An introduction to the topological derivative. *Engineering Computations*, 39:3–33, 2021.

[7] Martin Brühl, Martin Hanke, and Michael S. Vogelius. A direct impedance tomography algorithm for locating small inhomogeneities. *Numer. Math.*, 93(4):635–654, 2003.

[8] Ana Carpio and María-Luisa Rapún. Solving inhomogeneous inverse problems by topological derivative methods. *Inverse Problems*, 24(4):045014, 32, 2008.

[9] Eduardo Casas and Luis Alberto Fernández. Distributed control of systems governed by a general class of quasilinear elliptic equations. *J. Differential Equations*, 104(1):20–47, 1993.

[10] Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[11] Peter Gangl and Kevin Sturm. A simplified derivation technique of topological derivatives for quasi-linear transmission problems. *ESAIM Control Optim. Calc. Var.*, 26:Paper No. 106, 20, 2020.

[12] Hyeonbae Kang and Kyoungsun Kim. Anisotropic polarization tensors for ellipses and ellipsoids. *J. Comput. Math.*, 25(2):157–168, 2007.

[13] François Murat. Contre-exemples pour divers problèmes où le contrôle intervient dans les coefficients. *Ann. Mat. Pura Appl. (4)*, 112:49–68, 1977.

[14] François Murat and Luc Tartar. On the control of coefficients in partial differential equations [MR0428166 (55 #1193)]. In *Topics in the mathematical modelling of composite materials*, volume 31 of *Progr. Nonlinear Differential Equations Appl.*, pages 1–8. Birkhäuser Boston, Boston, MA, 1997.

[15] Dirk Praetorius. Analysis of the operator Δ^{-1}div arising in magnetic models. *Z. Anal. Anwendungen*, 23(3):589–605, 2004.

[16] Humberto Rafeiro, Natasha Samko, and Stefan Samko. Morrey-Campanato spaces: an overview. In *Operator theory, pseudo-differential equations, and mathematical physics*, volume 228 of *Oper. Theory Adv. Appl.*, pages 293–323. Birkhäuser/Springer Basel AG, Basel, 2013.

[17] U. E. Raĭtum. Necessary optimality conditions for systems described by nonlinear elliptic equations. I. *Z. Anal. Anwendungen*, 3(1):65–79, 1984.

[18] U. Ë. Raĭtum. Necessary optimality conditions for systems described by nonlinear elliptic equations. II. *Z. Anal. Anwendungen*, 3(2):133–152, 1984.

[19] U. E. Raĭtum. The maximum principle in optimal control problems for an elliptic equation. *Z. Anal. Anwendungen*, 5(4):291–306, 1986.

[20] U. Ë. Raĭtum. *Optimal Control Problems for Elliptic Equations (in Russian)*. Zinatne, Riga, 1989.

[21] Elias M. Stein and Rami Shakarchi. *Real analysis*, volume 3 of *Princeton Lectures in Analysis*. Princeton University Press, Princeton, NJ, 2005. Measure theory, integration, and Hilbert spaces.

[22] Kevin Sturm. Topological sensitivities via a Lagrangian approach for semilinear problems. *Nonlinearity*, 33(9):4310–4337, 2020.