Numerical Verification of Optimality Conditions in Optimal Control Problems

Dissertation zur Erlangung
des naturwissenschaftlichen Doktorgrades

der Julius-Maximilians-Universität Würzburg

vorgelegt von

Saheed Ojo Akindeinde

Eingereicht am: 16. November 2012
1. Gutachter: Professor Dr. Daniel Wachsmuth, Universität Würzburg
2. Gutachter: Professor Dr. Arnd Rösch, Universität Duisburg-Essen

Numerical Verification of Optimality Conditions in Optimal Control Problems

A dissertation submitted to the
Institute of Mathematics of the

Julius-Maximilians-Universität Würzburg

for awarding the degree

Doctor of Natural Sciences
(Doctor rerum naturalium, Dr. rer. nat.)

by

Saheed Ojo Akindeinde

Date of Submission:	16th November, 2012
First referee:	Professor Dr. Daniel Wachsmuth, Universität Würzburg
Second referee:	Professor Dr. Arnd Rösch, Universität Duisburg-Essen

Dedication

To my parents, Chief Ramoni A. Akindeinde and Mrs Kudirat A. Ajayi

Abstract

This thesis is devoted to numerical verification of optimality conditions for non-convex optimal control problems. In the first part, we are concerned with a-posteriori verification of sufficient optimality conditions. It is a common knowledge that verification of such conditions for general non-convex PDE-constrained optimization problems is very challenging. We propose a method to verify second-order sufficient conditions for a general class of optimal control problem. If the proposed verification method confirms the fulfillment of the sufficient condition then a-posteriori error estimates can be computed. A special ingredient of our method is an error analysis for the Hessian of the underlying optimization problem. We derive conditions under which positive definiteness of the Hessian of the discrete problem implies positive definiteness of the Hessian of the continuous problem. The results are complemented with numerical experiments.

In the second part, we investigate adaptive methods for optimal control problems with finitely many control parameters. We analyze a-posteriori error estimates based on verification of second-order sufficient optimality conditions using the method developed in the first part. Reliability and efficiency of the error estimator are shown. We illustrate through numerical experiments, the use of the estimator in guiding adaptive mesh refinement.

Acknowledgments

Above all, I give all praises and adoration to almighty God for His blessings.

My deep gratitude goes to my supervisor Professor Daniel Wachsmuth for giving me the opportunity to work on this project. I owe him a lot of thanks for introducing me to the field of optimization and optimal control of partial differential equations. I am specifically grateful for his patient and professional guidance during the three years of working on this project. His encouragement and support have been very invaluable to the successful completion of this work. Coming up with this piece would not have been possible without his painstaking discussions, reviewing and useful critiques of this research work.

My sincere and special appreciation goes to my wife for her support and encouragement throughout the duration of working on this project. Her love and soothing words have kept me going during the strenuous times.

I owe many thanks to the people of Johann Radon Institute for computational and applied mathematics (RICAM), Linz, Austria where this research work started. I am particularly indebted to Professor Sven Beuchler for his role in securing the opportunity to work on this project. Furthermore, to the people and staff of the department of mathematics, University of Wuerzburg where this thesis is completed, thanks for the conducive and friendly working environment.

It would be erroneous not to mention the important contributions of the Austrian Academy of Sciences for the generous funding of this research work under the Austrian Science Fund (FWF) grant P21564-N18.

Finally, I would like to thank all my friends who have contributed in one way or the other to the successful completion of this work.

Contents

List of Tables

List of Figures

Chapter 1

General introduction

This thesis is concerned with numerical verification of optimality conditions for optimal control problems. The theme of the thesis is best introduced with a model example. As a model problem we are going to study optimal control problems of the following type: Minimize the cost functional J given by

$$J(y,u) = g(y) + j(u) \tag{P}$$

over all $(y,u) \in Y \times U$ that satisfy the nonlinear elliptic partial differential equation

$$E(y,u) = 0$$

and the control constraints

$$u \in U_{ad}.$$

Here, the variable y denotes the state and u is the control. The state space Y and the control space U are real Banach spaces. Furthermore the functions g, j are mappings from $Y \to \mathbb{R}$, $U \to \mathbb{R}$ respectively. The set U_{ad} is a non-empty, convex and closed subset of U. Examples that are covered by this framework include parameter identification and optimization problems (with finitely many parameters) as for instance least-square problems as given by e.g.

$$J(y,u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} \tag{1.1}$$

over all $(y,u) \in H^1_0(\Omega) \times \mathbb{R}^n$ that satisfy the elliptic equation

$$\begin{aligned} -\Delta y + d(u;y) &= b \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{1.2}$$

Here, $d(u;y)$ denotes a nonlinear function of y parameterized by parameters $u \in \mathbb{R}^n$. The parameters have to be recovered by fitting the state y to the measured state y_d.

Another application is the optimization of material parameters by minimizing (1.1) subject to

$$\begin{aligned} -\operatorname{div}(a\nabla y) &= b \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.3}$$

where the coefficient function a is given by $a = \sum_{k=1}^{n} \chi_k u_k$ with $\chi_k = \chi_{\Omega_k}$ being characteristic functions of subsets $\Omega_k \subseteq \Omega$. Both problems are complemented by the constraint $u \in U_{ad}$.

We are interested in the numerical solution and the solution accuracy for such type of problems. Given a numerical solution u_h of a discretization of (P), we are asking, under which conditions u_h is near a local solution $\bar{u}$ of (P). This is a non-trivial question, since the optimization problem (P) is non-convex due to the nonlinearity of the elliptic equation. Hence, all results on discretization errors are subject to a second-order sufficient optimality condition (SSC) usually specified in terms of the Lagrange functional $\mathcal{L}$ in the form: there exists $\delta > 0$ such that

$$\mathcal{L}''(\bar{y}, \bar{u})[z, v] \geq \delta \|v\|_U^2 \tag{1.4}$$

for every $v \in U_{ad}$ where z is the solution of a linearized counterpart of the nonlinear PDE. This statement applies to both a-priori error estimates [6, 14] as well as a-posteriori error estimates [9, 10, 56]. The aim of this work is to develop a method that allows to verify the fulfillment of the sufficient condition a-posteriori. As a side result, we obtain reliable a-posteriori error estimates.

Asides the fact that the fulfillment of SSC confirms optimality of a solution, many important results in the numerics of partial differential equations can be proved under the assumption of SSC. For instance, they are used in proving stability of local solution under small perturbations, as can be found in e.g. [39], [40], [26]. More so, second-order sufficient conditions are essential ingredients in proving convergence results for optimization methods e.g. semi-smooth Newton method [37], [53], [27], [33], and the SQP methods [17], [51], [7], [28], [30], [31], [49], [60].

Let us now comment on why it is difficult to verify the second-order sufficient optimality condition (1.4). One obstacle is that the sufficient condition is required at an unknown solution $\bar{u}$ of the original undiscretized problem (P). Even if $\bar{u}$ would be known, it would still be tedious to check that the SSC holds, as it requires the exact solution z of linearized partial differential equations. Without the knowledge of $\bar{u}$ the check for SSC appears to be of the same difficulty as the check whether (P) is convex.

Earlier work on verification of sufficient optimality conditions can be found for instance in the works of Rösch and Wachsmuth [47, 48]. There the optimal control of semilinear elliptic equations was studied. The principal idea was to verify the fulfillment of a infinite-dimensional second-order condition at the (known) discrete solution u_h, and by a careful analysis confirm that this property carries over to the unknown solution $\bar{u}$ of (P). This infinite-dimensional second-order condition at u_h was relatively easy to check due to the special structure of the Lagrangian associated to the considered problem, which makes it impossible to generalize the results to e.g. (1.3). Numerical studies on second-order sufficient conditions can be found in the works of Mittelmann [42, 43]. There the second-order sufficient optimality condition was checked for the *discrete* problem. The fulfillment of the discrete SSC is a strong indication for the fulfillment of the SSC for the continuous problem but not sufficient. In the present work we will combine both strategies: first check the discrete SSC as in [42, 43], then develop conditions, under which the discrete SSC implies a continuous second-order condition, and finally with an

analysis as in [47, 48] conclude that this second-order condition carries over to the unknown solution $\bar{u}$ of (P).

As we also want to develop an a-posteriori error analysis, let us comment on available a-posteriori error estimators in the literature. A-posteriori error estimates for nonlinear control and identification problems can be found for instance in [9, 25, 36, 56]. Both the dual-weighted residual type [10] and the residual type error estimators are available. However, they depend on two crucial *a-priori* assumptions: the first is that a second-order sufficient condition has to hold at the solution of the continuous problem. With this assumption, error estimates of the type

$$\|\bar{u} - u_h\|_U \leq c\,\eta + \mathcal{R}$$

can be derived, where η is a computable error indicator and $\mathcal{R}$ is a second-order remainder term. Here, the second a-priori assumption comes into play: one has to assume that $\mathcal{R}$ is small enough in order to guarantee that mesh refinement solely based on η is meaningful. A different approach with respect to mesh refinement was followed in [64]. There the residuals in the first-order necessary optimality condition were used to derive an adaptive procedure. However, smallness of residuals does not imply smallness of errors without any further assumption. Here again, SSC as well as smallness of remainder terms is essential to draw this conclusion. In this thesis, we will present conditions that allow reliable a-posteriori error estimates that *verify* these two conditions *a-posteriori* and that do not require them a-priori.

The goals of this work are thus the following: first, we aim at verifying the sufficient condition a-posteriori. This allows to derive a-posteriori error estimators in a second step, see our main results Theorems 3.26 and 3.27. In the last step, we will establish reliability and efficiency of the error estimator from which different adaptive methods are derived.

We remark that our analysis will be done only for problems with finite dimensional control space. For some class of problems considered, SSC verification in the infinite dimensional case is still not available. Some preliminary results in this direction are however given in Chapter 5.

Organization of the thesis

The first chapter describes some basics of optimal control of nonlinear partial differential equations. There, the existence result and the derivation of optimality conditions are discussed. By writing the optimality system as a non-smooth equation, we discuss semi-smooth Newton method for solving the resulting non-smooth system.

The third chapter constitutes the major part of the thesis, being our original contribution and the main theme of this work. There, a method for the numerical verification of second-order sufficient optimality condition is derived. As mentioned before, our method complements the relevant ideas of preceding publications on this subject with eigenvalue error analysis of the Hessian of the underlying optimization problem. The main task is to prove that the second-order sufficient condition is fulfilled at an unknown solution. It turns out that if the fulfillment of SSC is confirmed, a-posteriori estimate for the error in the control is additionally obtained. The performance of the method is tested with numerical examples.

Efficiency of the obtained error estimator is proved in the following chapter. The derivation of a lower bound for the error estimator as well as its application in mesh adaptivity are detailed in Chapter 4. The results of the different mesh adaptive procedures are illustrated with numerical examples. The results contained in this chapter are original contributions.

In the concluding chapter, we attempt to extend the verification results of Chapter 3 to problems with infinite dimensional control space. The main result there is the regularity result for the eigenfunctions of the associated eigenvalue problem. We mention possible challenges in extending our method of verification to infinite dimensional SSC as a closing remark.

Chapter 2
2 Optimal control of nonlinear PDE

This chapter contains relevant basics of optimal control of partial differential equations. We will start by recalling some useful notions from functional analysis. This leads us to the proof of existence result for abstract problem (P). The optimality conditions for the abstract problem are then derived. The chapter ends with discussion on semi-smooth Newton method for solving the optimality system.

2.1 Functional analytic preliminaries

Equivalence of norms

Let X be a vector space. The two norms $\|\cdot\|_a, \|\cdot\|_b$ defined on X are said to be equivalent if there exists positive constants $c_l, c_u \in \mathbb{R}$ such that

$$c_l\|x\|_a \leq \|x\|_b \leq c_u\|x\|_a$$

holds for every $x \in X$.

As an example, on a finite dimensional space $X = \mathbb{R}^n$ the mappings

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, \ 1 \leq p < \infty$$

and

$$\|x\|_\infty = \max_i |x_i|$$

define norms on X. Furthermore these norms are equivalent: For every $x \in X$ it hold

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2, \quad \|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty, \quad \|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty.$$

More generally all norms on finite dimensional space $\mathbb{R}^n$ are equivalent, see e.g. [16] or [21] .

Matrix norms

Let us consider a matrix $A = (a_{ij}) \in \mathbb{R}^{n\times m}$. The Frobenius norm of A is defined as

$$\|A\|_F = \sum_{i,j} |a_{ij}|^2 = \text{trace}(A^T A) \tag{2.1}$$

where A^T denotes the transpose of matrix A. Another notion of norm on matrices is the Euclidean norm which is defined for matrix A as

$$\|A\|_2 = \left\{ \max \frac{\|Ax\|_2}{\|x\|_2} : x \in \mathbb{R}^n, x \neq 0 \right\}. \tag{2.2}$$

If A is a square matrix then the Euclidean norm is equivalent to the spectral norm, which is defined as the square root of the largest eigenvalue of the symmetric matrix $A^T A$. That is, if λ denotes the maximum eigenvalue of matrix $A^T A$, then the spectral norm $\|A\|_\rho$ is defined by

$$\|A\|_\rho = \sqrt{\lambda(A^T A)}. \tag{2.3}$$

Let us mention the following well-known relation [24, Section 2.3.2] between (2.1) and (2.2) or (2.3) which will be needed in later chapters

$$\|A\|_2 \leq \|A\|_F. \tag{2.4}$$

Sobolev spaces and Embeddings

A famous reference for the materials presented below is [1].

Definition 2.1 (L^p spaces)**.** *Let $\Omega \subseteq \mathbb{R}^n$, $n \in \{2, 3\}$ denote a Lebesgue measurable domain. The Lebesgue spaces $L^p(\Omega)$ are defined as the spaces of real valued measurable functions $f : \Omega \to \mathbb{R}$ with $\int_\Omega |f(x)|^p \mathrm{d}x < \infty$. They are normed spaces with norms $\|\cdot\|_{L^p}$ defined by*

$$\|f\|_{L^p} := \left(\int_\Omega |f(x)|^p \mathrm{d}x \right)^{\frac{1}{p}}, \quad 1 \leq p < \infty$$

and for $p = \infty$ it is defined as

$$\|f\|_{L^\infty} := \operatorname{ess\,sup}\{|f(x)| : x \in \Omega\}.$$

The spaces $L^p(\Omega)$, $p \geq 1$ are Banach spaces and in particular, $L^2(\Omega)$ is a Hilbert space with scalar product $\langle f, g\rangle := \int_\Omega fg\mathrm{d}x \ \forall f, g \in L^2(\Omega)$. For functions $f \in L^p, g \in L^q$ with $1/p + 1/q = 1$, we have $fg \in L^1(\Omega)$ and the Hölder inequality

$$\|fg\|_{L^1} \leq \|f\|_{L^p} \|g\|_{L^q}$$

holds. The case $p = q = 2$ is the famous Cauchy-Schwarz inequality

$$|\langle f, g\rangle| \leq \|f\|_{L^2} \, \|g\|_{L^2}.$$

Let k be a non-negative integer and $p \geq 1$. The Sobolev space $W^{k,p}(\Omega)$ is defined as

$$W^{k,p}(\Omega) = \{f \in L^p(\Omega) : |\alpha| \leq k, D^\alpha f \in L^p(\Omega)\}$$

where $D^\alpha f$ denotes the generalized weak derivative of function f. The space $W^{k,p}(\Omega)$ equipped with the norm

$$\|f\|_{W^{k,p}} = \left(\sum_{|\alpha| \leq k} \|D^\alpha f\|_{L^p}^p \right)^{1/p} \tag{2.5}$$

is a Banach space. For the case $p = 2$ we write $H^k(\Omega) := W^{k,2}(\Omega)$ which is again a Hilbert space if furnished with the inner product

$$\langle f, g\rangle_{H^k} = \sum_{|\alpha| \leq k} \langle D^\alpha f, D^\alpha g\rangle.$$

Theorem 2.2 (Sobolev Embedding)**.** *Let $\Omega \subset \mathbb{R}^n$, $n \in \{2, 3\}$ be a bounded Lipschitz domain. If $kp < n$, then the embedding*

$$W^{k,p}(\Omega) \hookrightarrow L^q(\Omega)$$

is continuous for $1 \leq q \leq \frac{np}{n-kp}$. Furthermore, for $kp > n$, it holds

$$W^{k,p}(\Omega) \hookrightarrow C(\bar{\Omega}).$$

In particular for two dimensional domains Ω, we have the embedding $H^1(\Omega) \hookrightarrow L^q(\Omega), 1 \leq q < \infty$. Denoting the norms of these embeddings by I_q, we have the inequality $\|v\|_{L^q(\Omega)} \leq I_q \|v\|_{H^1(\Omega)}$ for every $v \in H^1(\Omega)$. For $n = 3$ the embedding $H^1(\Omega) \hookrightarrow L^6(\Omega)$ holds true.

Operator norm, Adjoint operator

Let X, Y be Banach spaces. An operator $T : X \to Y$ is said to be bounded if there is a constant $c \geq 0$ such that $\|Tx\|_Y \leq c\|x\|_X \quad \forall x \in X$. We denote by $\mathcal{L}(X, Y)$ the space of bounded linear operators from the space X into Y. This space is a Banach space if equipped with the norm

$$\|T\|_{\mathcal{L}(X,Y)} = \sup_{x \in X, \ \|x\|_X \leq 1} \|Tx\|_Y. \tag{2.6}$$

Equation (2.6) defines the norm of a bounded operator T.

Let H be real Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Let $T : H \to H$ be a bounded operator on H. An operator $T^* : H \to H$ satisfying

$$\langle Tu, v \rangle = \langle u, T^* v \rangle \quad \forall u, v \in H$$

is called the adjoint of T. A bounded operator T is called self-adjoint if $T^* = T$. For a normed vector space X we shall denote by X^* the dual space of X which by definition is the space of all bounded linear functionals on X. That is

$$X^* = \mathcal{L}(X, \mathbb{R}) = \Big\{ f : X \to \mathbb{R} \Big| \sup_{\|x\| \leq 1} \{|f(x)|\} < \infty \Big\}$$

where $f(x) := \langle f, x \rangle_{X^*,X}$. The notation $\langle \cdot, \cdot \rangle_{X^*,X}$ denotes the duality paring of X and X^*.

Differentiability

Here we want to define the notion of Fréchet differentiability of functions. It is therefore natural to start with the following definitions. Let X, Y be Banach spaces.

Definition 2.3. *The mapping $f : X \to Y$ is called directionally differentiable at x if for all $x \in X$ and $h \in X$ the limit*

$$f'(x; h) = \lim_{t \to 0} \frac{f(x + th) - f(x)}{t} \quad \textit{exists in } Y.$$

Definition 2.4. *Let $f : X \to Y$. If f is directionally differentiable at x and there exists a linear and bounded operator $A \in \mathcal{L}(X, Y)$ such that*

$$f'(x; h) = Ah$$

then f is said to be Gâteaux differentiable at x and we write $f'(x) = A$.

Definition 2.5 (Fréchet-differentiability)**.** *Let $U \subset X$ be an open subset of X. The mapping $f : X \to Y$ is said to be Fréchet-differentiable on U if for every $x \in U$, f is Gâteaux differentiable at x and*

$$\lim_{h \to 0} \frac{\|f(x + h) - f(x) - f'(x)h\|_Y}{\|h\|_X} = 0.$$

Example 1. *Let H be a Hilbert space. The functional $f : H \to \mathbb{R}$ given by $f(x) = \|x\|^2$ is Fréchet- differentiable.*

Proof. Firstly we will show that f is Gâteaux differentiable at $x \in X$. For $h \in X$ let us compute the directional derivative $f'(x; h)$ at x. It holds

$$f(x + th) - f(x) = \langle x + th, x + th \rangle - \langle x, x \rangle = \langle 2x, th \rangle + \|th\|^2.$$

Next dividing through by t, passing to the limit $t \to 0$ and employing Cauchy-Schwarz inequality we obtain

$$f'(x;h) = \langle 2x, h\rangle \leq 2\|x\| \cdot \|h\|.$$

This shows that the directional derivative $f'(x;h)$ exists and is linear and bounded so that f is Gâteaux differentiable at x.

Finally we look at

$$|f(x+h) - f(x) - f'(x)h| = |\langle x+h, x+h\rangle - \langle x, x\rangle - \langle 2x, h\rangle| = \|h\|^2$$

from which the required property follows. □

Now having introduced the relevant preliminaries from functional analysis, we will now derive some basic results for the abstract optimal control problem. These include the existence result for the state equation and the optimal control problem, and differentiability properties of the solution mapping.

2.2 Optimal control problem

We will consider optimal control problem (P) which is restated below for convenience:
Minimize

$$J(y,u) = g(y) + j(u)$$

over all (y,u) in the Banach space $Y \times U$ satisfying respectively the state equation and the control constraint

$$E(y,u) = 0, \quad u \in U_{ad}.$$

Let us proceed by fixing the assumptions on the abstract problem (P). The assumptions are assumed to hold for the rest of the chapter except where otherwise stated. Firstly, the domain $\Omega \subseteq \mathbb{R}^n$ is a two or three dimensional domain, the control space $U = L^2(\Omega_s), \Omega_s \subseteq \Omega$ and the state space Y is a real Banach space to be made precise shortly. Moreover, we assume that

Assumption 1. *1. The mapping $E : Y \times U \to Y^*$ is twice continuously Fréchet-differentiable. Furthermore, we assume that the mapping E is strongly monotone with respect to the first variable, i.e. there is a constant $\delta > 0$ such that*

$$\langle E(y_1,u) - E(y_2,u),\, y_1 - y_2\rangle_{Y^*,Y} \geq \delta \|y_1 - y_2\|_Y^2$$

for all $u \in U_{ad}, y_1, y_2 \in Y$.

2. The functions $g : Y \to \mathbb{R}$ and $j : U \to \mathbb{R}$ are twice continuously Fréchet-differentiable.

Remark 2.6. *The computations that follow are carried out with the choice $Y = H^1(\Omega)$ in mind. However the results can be adapted to stronger spaces which embed in Y e.g. $L^\infty \cap H^1$ by relaxing the differentiability assumptions on E and g, see Remark 3.1 on page 45.*

Proposition 2.7. *Under Assumption 1, for each admissible control $u \in U$ the state equation $E(y,u) = 0$ is uniquely solvable. Furthermore, the partial derivative of operator E with respect to y, denoted by E_y, is continuously invertible, i.e. $E_y(y_0,u_0)^{-1} \in \mathcal{L}(Y^*, Y)$ for all $(y_0,u_0) \in Y \times U_{ad}$, and it holds $\|E_y^{-1}(y_0,u_0)\|_{\mathcal{L}(Y^*,Y)} \leq \delta^{-1}$ for all $(y_0,u_0) \in Y \times U_{ad}$. Additionally, the solution mapping $S : U \to Y$ which maps every admissible control to corresponding state y is Lipschitz continuous.*

Proof. The result is standard and its proof can be found for instance in [63, Theorem 26.A, p 557]. □

The result of the above proposition implies the unique solvability of the linearized state equation.

Corollary 2.8. *Let Assumption 1 holds. Let $\bar{u} \in U_{ad}$ and $\bar{y}$ denotes the associated state. Then for every $v \in U$ the linearized state equation*

$$E_y(\bar{y},\bar{u})z + E_u(\bar{y},\bar{u})v = 0$$

is uniquely solvable in Y.

Proof. Due to monotonicity of operator E, the operator E_y, which is the Fréchet derivative of E with respect to the state y, is itself monotone. The result is then a direct consequence of Proposition 2.7. □

Let us now turn our attention to the existence result for abstract problem (P). We begin with relevant definitions.

Definition 2.9. *A pair $(\bar{y},\bar{u}) \in Y \times U_{ad}$ is called optimal for problem (P) if $E(\bar{y},\bar{u}) = 0$ and it holds*

$$J(\bar{y},\bar{u}) \leq J(y,u) \quad \forall\ (y,u) \in Y \times U_{ad} \text{ with } E(y,u) = 0.$$

The state equation $E(y,u) = 0$ is in general a nonlinear partial differential equation. Thus the associated optimal control problem will be non-convex even if the cost functional J and the admissible set U_{ad} are both convex. The notion of local solution will therefore play a major role in the subsequent analysis.

Definition 2.10. *Let $\bar{u} \in U_{ad}$ and let $\bar{y}$ be the corresponding state. A control $\bar{u} \in U_{ad}$ is said to be locally optimal if there exists $\rho > 0$ such that*

$$J(\bar{y},\bar{u}) \leq J(y,u)$$

holds for all $(y,u) \in Y \times U_{ad}$ with $||u - \bar{u}||_U \leq \rho$ and $E(y,u) = 0$.

Essential in proving the existence result for problem (P) is lower semi-continuity of the cost functional J. This concept is defined below.

Definition 2.11. *Let H be a Banach space and $f : H \to \mathbb{R}$ be a functional on H. The functional f is called*

- *lower semi-continuous if*

$$f(x) \leq \liminf_{n\to\infty} f(x_n) \tag{2.7}$$

  for all convergent sequences $x_n \to x$ in H.

- *weakly lower semi-continuous if (2.7) holds true for all weakly convergent sequences $x_n \rightharpoonup x$ in H.*

The following lemma on lower semi-continuity of convex, continuous functions is well known and can be found in e.g. [62] or [61].

Lemma 2.12. *Every continuous, convex functional is weakly lower semi-continuous.*

Next, we will prove some nice properties of the admissible set U_{ad} which will also be essential in proving existence result for problem (P).

Lemma 2.13. *Let $U = L^2(\Omega)$ and $u_a, u_b \in U$ with $u_a \leq u_b$. The set*

$$U_{ad} := \{u \in U : u_a \leq u \leq u_b\}$$

is non-empty, closed, convex and bounded subset of U.

Proof. Clearly $u_a \in U_{ad}$ so that $U_{ad} \neq \emptyset$. Furthermore, let $u \in U_{ad}$ then we can estimate $\|u\|_{L^2(\Omega)} \leq \|\max(|u_a|, |u_b|)\|_{L^2(\Omega)}$, thereby proving boundedness of the set. Let $u_1, u_2 \in U_{ad}$. Then for any constant $\alpha \in (0, 1)$ the trivial inequality $u_a \leq \alpha u_1 + (1-\alpha)u_2 \leq u_b$ shows convexity of the set. Lastly for closedness, let $u_n \in U_{ad}$ and $u_n \to u$. The convergence holds in L^2 sense so that $u_n(x) \to u(x)$ a.e. on Ω. Now since $u_a(x) \leq u_n(x) \leq u_b(x)$ it must hold $u_a(x) \leq u(x) \leq u_b(x)$ a.e on Ω and hence $u \in U_{ad}$. □

2.2.1 Existence of optimal control

In addition to the above preliminaries, in order to prove existence result for the optimal control problem (P), we further require the following.

Assumption 2. *The control to state map $S : U \to Y$ is compact and the functional $j : U \to \mathbb{R}$ is convex.*

Assumption 3. *For all sequences $u_n \rightharpoonup \bar{u} \in U_{ad}$ and $y_n \to \bar{y} \in Y$ it holds*

$$\langle E(y_n, u_n), \tilde{y}\rangle_{Y^*,Y} \to \langle E(\bar{y}, \bar{u}), \tilde{y}\rangle_{Y^*,Y} \quad \forall \tilde{y} \in Y.$$

Proposition 2.14 (Existence result)**.** *Let $U = L^2(\Omega)$. Let Assumptions 1, 2 and 3 be fulfilled. Then the optimal control problem (P) is solvable.*

Proof. Let us define the quantity

$$\bar{J} := \inf_{u \in U_{ad}} J(y(u), u).$$

Then we follow the standard argument of proof by taking a minimizing sequence $(y_n, u_n) \in Y \times U_{ad}$ where for $u_n \in U_{ad}$ we define $y_n = S(u_n)$. This minimizing sequence realizes the infimum of the functional J on the admissible space $Y \times U_{ad}$, that is

$$\lim_{n \to \infty} J(y_n, u_n) = \bar{J}.$$

By Lemma 2.13 the set of admissible control U_{ad} is non-empty, closed, convex and bounded. Therefore U_{ad} is weakly sequentially compact. Hence there exists a weakly convergent subsequence $u_n \in U_{ad}$ (which is chosen here as the sequence itself for convenience) such that

$$u_n \rightharpoonup \bar{u}, \ n \to \infty$$

with $\bar{u} \in U_{ad}$. Since the sequence $u_n \in U_{ad}$ is bounded and the solution mapping S is compact by assumption, we have that the sequence $y_n = S(u_n)$ is bounded in the state space Y. Furthermore it holds

$$y_n = S(u_n) \to \bar{y} \quad \text{in} \quad Y.$$

Now we have to show that $\bar{y}$ is the state corresponding to the control $\bar{u}$, i.e. $\bar{y} = S(\bar{u})$. This would be the case if $(\bar{y}, \bar{u})$ satisfies the state equation $E(\bar{y}, \bar{u}) = 0$.

To check that $(\bar{y}, \bar{u})$ fulfills the state equation $E(\bar{y}, \bar{u}) = 0$, let $y_n = S(u_n)$ and $y = S(u)$. Then for each n, the sequence $(y_n, u_n) \in Y \times U_{ad}$ is admissible and fulfills $E(y_n, u_n) = 0$. Thanks to the monotonicity of operator E and Assumption 3 it holds for all $y \in Y$

$$\begin{aligned}
0 &\leq \langle E(\bar{y}, \bar{u}) - E(y, \bar{u}), \ \bar{y} - y \rangle_{Y^*, Y} \\
&= \lim_{n \to \infty} \langle E(y_n, u_n) - E(y, \bar{u}), \ \bar{y} - y \rangle_{Y^*, Y} \\
&= \lim_{n \to \infty} \langle E(y_n, u_n), \ \bar{y} - y \rangle_{Y^*, Y} + \langle -E(y, \bar{u}), \ \bar{y} - y \rangle_{Y^*, Y} \\
&= \langle -E(y, \bar{u}), \ \bar{y} - y \rangle_{Y^*, Y}.
\end{aligned}$$

That is

$$\langle -E(y, \bar{u}), \ \bar{y} - y \rangle_{Y^*, Y} \geq 0. \tag{2.8}$$

We will apply Minty-trick to prove the other direction of (2.8). Let us write $y = \bar{y} + \epsilon w$, where $w \in Y$ is arbitrary. Now using $y = \bar{y} + \epsilon w$ in (2.8), we obtain

$$\langle -E(\bar{y} + \epsilon w, \bar{u}),\ -\epsilon w \rangle_{Y^*,Y} \geq 0.$$

On dividing by $-\epsilon$ and passing to the limit $\epsilon \to 0$ we obtain

$$\langle -E(\bar{y}, \bar{u}),\ w \rangle_{Y^*,Y} \leq 0$$

by the continuity of E. As w is arbitrary, one obtains $E(\bar{y}, \bar{u}) = 0$ as required.

To complete the proof of the proposition, it remains to show that the pair $(\bar{y}, \bar{u})$ indeed minimizes the optimal control problem, that is we have to prove $J(\bar{y}, \bar{u}) \leq J(y, u)$ for every $(y, u) \in Y \times U_{ad}$ with $y = S(u)$. Owing to Lemma 2.12, the functional j is weakly lower semi-continuous and therefore fulfills $j(\bar{u}) \leq \lim_{n\to\infty} \inf j(u_n)$. Also, as $\bar{y} = \lim_{n\to\infty} y_n$ the continuity of g implies $g(\bar{y}) = \lim_{n\to\infty} g(y_n)$. Altogether we obtain

$$\begin{aligned} \bar{J} = \lim_{n\to\infty} J(y_n, u_n) &= \lim_{n\to\infty} g(y_n) + \lim_{n\to\infty} \inf j(u_n) \\ &\geq g(\bar{y}) + j(\bar{u}) \\ &= J(\bar{y}, \bar{u}). \end{aligned}$$

By the definition of infimum $\bar{J}$, we know that $\bar{J} \leq J(\bar{y}, \bar{u})$. Hence, we must have $\bar{J} = J(\bar{y}, \bar{u})$.

Altogether, we have therefore shown that $(\bar{y}, \bar{u})$ realizes the infimum of J on $Y \times U_{ad}$ and fulfills both the state and the control constraint. This completes the proof.

□

Let us now prove differentiability properties of the solution mapping S.

2.2.2 Fréchet derivatives of the control-to-state map

It is quite a common practice in PDE-constrained optimization to eliminate the occurrence of the state variable by introducing the control-to-state map $S : U \to Y$. We write $y = S(u)$ if and only if $E(y, u) = 0$. By doing so, the optimization parameters reduce to only the control variable u and one obtains a control-reduced problem

$$\min_{u \in U_{ad}} f(u) = J(S(u), u) \tag{2.9}$$

as an equivalent problem to (P). To be able to derive the necessary and sufficient optimality conditions for the above problem, the Fréchet differentiability of the mapping S needs to be established. Therefore in the sequel, we shall prove that the mapping S is twice Fréchet differentiable at every admissible control $\bar{u} \in U_{ad}$.

Lemma 2.15. *Let Assumption 1 be fulfilled. Then the control-to-state map $S : U \to Y$ is*

Fréchet differentiable and its first derivative $z = S'(\bar{u})v$ at $\bar{u}$ in the direction $v \in U$ solves the linearized equation

$$E_y(\bar{y}, \bar{u})z + E_u(\bar{y}, \bar{u})v = 0$$

where $\bar{y} = S(\bar{u})$.

Proof. It suffices to show the existence of a continuous linear operator $D : U \to Y$ such that

$$S(\bar{u} + v) - S(\bar{u}) = D(v) + r(\bar{u}, v) \tag{2.10}$$

and

$$\frac{\|r(\bar{u}, v)\|_Y}{\|v\|_U} \to 0, \text{ as } \|v\|_U \to 0. \tag{2.11}$$

If these conditions hold, we can then set $S'(\bar{u}) = D$. Let us now establish the claim of the lemma.

Firstly by definition $\bar{y} = S(\bar{u})$ if and only if $E(\bar{y}, \bar{u}) = 0$. Similarly $\tilde{y} = S(\bar{u} + v) := S(\tilde{u})$ if and only if $(\tilde{y}, \tilde{u})$ solves $E(\tilde{y}, \tilde{u}) = 0$. Therefore the difference $\tilde{y} - \bar{y}$ solves

$$E(\tilde{y}, \tilde{u}) - E(\bar{y}, \bar{u}) = 0. \tag{2.12}$$

Since by Assumption 1, operator E is continuously Fréchet differentiable, we therefore obtain through Taylor expansion

$$0 = E(\tilde{y}, \tilde{u}) - E(\bar{y}, \bar{u}) = E_y(\bar{y}, \bar{u})(\tilde{y} - \bar{y}) + E_u(\bar{y}, \bar{u})(\tilde{u} - \bar{u}) + r_d(\tilde{y} - \bar{y}, \tilde{u} - \bar{u}) \tag{2.13}$$

with r_d satisfying the remainder property

$$\frac{\|r_d(\tilde{y} - \bar{y}, \tilde{u} - \bar{u})\|_{Y^*}}{(\|\tilde{u} - \bar{u}\|_U + \|\tilde{y} - \bar{y}\|_Y)} \to 0, \text{ as } \|\tilde{u} - \bar{u}\|_U + \|\tilde{y} - \bar{y}\|_Y \to 0. \tag{2.14}$$

By Proposition 2.7, the operator S is Lipschitz continuous. Thus there exists $L > 0$ such that

$$\|\tilde{y} - \bar{y}\|_Y = \|S(\tilde{u}) - S(\bar{u})\|_Y \leq L\|\tilde{u} - \bar{u}\|_U.$$

Hence it holds $\|\tilde{y} - \bar{y}\|_Y \to 0$ as $\|\tilde{u} - \bar{u}\|_U \to 0$. Consequently the condition

$$\frac{\|r_d(\tilde{y} - \bar{y}, \tilde{u} - \bar{u})\|_{Y^*}}{\|\tilde{u} - \bar{u}\|_U} \to 0, \text{ as } \|\tilde{u} - \bar{u}\|_U \to 0 \tag{2.15}$$

follows from (2.14). We will then use (2.15) instead of (2.14) in the computations that follow.

Let us go back to equation (2.13) and set $\tilde{y} - \bar{y} = z + r$ where z, r solve simultaneously

$$E_y(\bar{y}, \bar{u})z + E_u(\bar{y}, \bar{u})v = 0 \tag{2.16}$$

$$E_y(\bar{y}, \bar{u})r = -r_d. \tag{2.17}$$

If r fulfills the remainder condition (2.11), and $z \in Y$, then we can finish the proof by setting $z = D(v)$.

Now thanks to the monotonicity of operator E and Corollary 2.8 we obtain from (2.17) that

$$\|r\|_Y \leq (1/\delta)\|r_d\|_{Y^*}$$

with constant δ as defined in Assumption 1. Recall that $v = \tilde{u} - \bar{u}$ by definition. Therefore using property (2.15) we obtain

$$\frac{\|\tilde{y} - \bar{y} - z\|_Y}{\|v\|_U} = \frac{\|r\|_Y}{\|v\|_U} \leq (1/\delta)\frac{\|r_d\|_{Y^*}}{\|v\|_U} \to 0 \text{ as } \|v\|_U \to 0$$

so that $r(\bar{u}, v)$ fulfills the remainder condition (2.11). It then remains to prove that $v \mapsto z$ is a continuous linear mapping from U to Y. This follows from the estimate

$$\|z\|_Y \leq (1/\delta)\|E_u(\bar{y}, \bar{u})\|_{\mathcal{L}(U,Y^*)}\|v\|_U$$

where we have applied the result of Corollary 2.8 on (2.16). We complete the proof by setting $D(v) = z$. □

Next, we will prove similarly that $v \mapsto z = S'(\bar{u})v$ itself is a Fréchet differentiable mapping. Note that since E is twice continuously Fréchet differentiable, E_y is locally Lipschitz continuous. This means for every $(u_1, y_1), (u_2, y_2)$ in a bounded, open subset $D \subseteq U_{ad} \times Y$, there exists a positive constant L_E depending on D such that

$$\|E_y(y_1, u_1) - E_y(y_2, u_2)\|_{\mathcal{L}(Y,Y^*)} \leq L_E \left(\|y_1 - y_2\|_Y + \|u_1 - u_2\|_U\right). \tag{2.18}$$

Lemma 2.16. *Let the hypothesis and result of Lemma 2.15 hold. The mapping $S' : U \to \mathcal{L}(U, Y)$ is Fréchet differentiable from the space U into $\mathcal{L}(U, Y)$. Its derivative $z = S''(\bar{u})[v_1, v_2]$ at $\bar{u}$ in the directions $v_1, v_2 \in U$ solves the equation*

$$E_y(\bar{y}, \bar{u})z + E_{yy}(\bar{y}, \bar{u})[y_1, y_2] + E_{yu}(\bar{y}, \bar{u})[y_1, v_2] + E_{uy}(\bar{y}, \bar{u})[v_1, y_2] + E_{uu}(\bar{y}, \bar{u})[v_1, v_2] = 0$$

where $\bar{y} = S(\bar{u}),\ y_i = S'(\bar{u})v_i,\ i = 1, 2.$

Proof. Again, we have to show the existence of a continuous linear map G such that

$$S'(\bar{u} + v_2) - S'(\bar{u}) = G(v_2) + r(\bar{u}, v_2)$$

with $r(\bar{u}, v_2)$ fulfilling the remainder property

$$\frac{\|r(\bar{u}, v_2)\|_{\mathcal{L}(U,Y)}}{\|v_2\|_U} \to 0 \text{ as } \|v_2\|_U \to 0. \tag{2.19}$$

Let $u^* = \bar{u} + v_2$, $\tilde{y} = S'(\bar{u} + v_2)v_1 = S'(u^*)v_1$ and $y_1 = S'(\bar{u})v_1$. Furthermore let $\bar{y} = S(\bar{u})$ as before and $y^* = S(u^*)$. Then by the result of Lemma 2.15 we have that $\tilde{y} = S'(\bar{u} + v_2)v_1, y_1 = S'(\bar{u})v_1$ are the solution of

$$\begin{aligned} E_y(y^*, u^*)\tilde{y} + E_u(y^*, u^*)v_1 = 0, \\ E_y(\bar{y}, \bar{u})y_1 + E_u(\bar{y}, \bar{u})v_1 = 0 \end{aligned}$$

respectively. The difference $\tilde{y} - y_1$ therefore solves

$$\begin{aligned} E_y(\bar{y}, \bar{u})(\tilde{y} - y_1) + \Big(E_y(y^*, u^*) - E_y(\bar{y}, \bar{u})\Big)y_1 + \Big(E_y(y^*, u^*) - E_y(\bar{y}, \bar{u})\Big)(\tilde{y} - y_1) \\ + \Big(E_u(y^*, u^*) - E_u(\bar{y}, \bar{u})\Big)v_1 = 0. \end{aligned} \tag{2.20}$$

In order to estimate the solution of the above equation like in the previous lemma, it is useful to have the estimate $\|y^* - \bar{y}\|_Y$ at hand. Using the Lipschitz property of S (cf. Lemma 2.15) one obtains

$$\|y^* - \bar{y}\|_Y = \|S(\bar{u} + v_2) - S(\bar{u})\|_Y \leq L\|v_2\|_U. \tag{2.21}$$

Also since the solution mapping S is Fréchet differentiable, we write

$$\begin{aligned} y^* - \bar{y} = S(u^*) - S(\bar{u}) &= S(\bar{u} + v_2) - S(\bar{u}) \\ &= S'(\bar{u})v_2 + r_2(\bar{u}, u^* - \bar{u}) \\ &=: y_2 + r_2(\bar{u}, v_2) \end{aligned} \tag{2.22}$$

with r_2 fulfilling

$$\frac{\|r_2(\bar{u}, v_2)\|_Y}{\|v_2\|_U} \to 0, \text{ as } \|v_2\|_U \to 0. \tag{2.23}$$

Since E is twice continuously Fréchet differentiable by assumption, Taylor expansion gives

$$\mathcal{L}(Y, Y^*) \ni E_y(y^*, u^*) - E_y(\bar{y}, \bar{u}) = E_{yy}(\bar{y}, \bar{u})(y^* - \bar{y}) + E_{yu}(\bar{y}, \bar{u})v_2 + r_1^y(v_2, y^* - \bar{y}) \tag{2.24}$$

where the remainder term r_1^y fulfills

$$\frac{\|r_1^y(v_2, y^* - \bar{y})\|_{\mathcal{L}(Y,Y^*)}}{(\|y^* - \bar{y}\|_Y + \|v_2\|_U)} \to 0, \text{ as } \|y^* - \bar{y}\|_Y + \|v_2\|_U \to 0. \tag{2.25}$$

Now using (2.22) in the first addend of (2.24) we obtain

$$E_y(y^*,u^*)-E_y(\bar y,\bar u)=E_{yy}(\bar y,\bar u)y_2+E_{yy}(\bar y,\bar u)r_2(\bar u,v_2)+E_{yu}(\bar y,\bar u)v_2+r_1^y(v_2,y^*-\bar y). \quad (2.26)$$

In a similar manner one obtains

$$E_u(y^*,u^*)-E_u(\bar y,\bar u)=E_{uu}(\bar y,\bar u)v_2+E_{uy}(\bar y,\bar u)y_2+E_{uy}(\bar y,\bar u)r_2(\bar u,v_2)+r_1^u(v_2,y^*-\bar y) \quad (2.27)$$

with r_1^u satisfying the remainder property

$$\frac{\|r_1^u(v_2,y^*-\bar y)\|_{\mathcal{L}(U,Y^*)}}{(\|y^*-\bar y\|_Y+\|v_2\|_U)}\to 0,\ \text{as } \|y^*-\bar y\|_Y+\|v_2\|_U\to 0. \quad (2.28)$$

Then substituting (2.26) and (2.27) in (2.20) yields

$$\begin{aligned}&E_y(\bar y,\bar u)(\tilde y-y_1)+E_{yy}(\bar y,\bar u)(y_1,y_2)+E_{yy}(\bar y,\bar u)r_2(\bar u,v_2)y_1+E_{yu}(\bar y,\bar u)(v_2,y_1)+r_1^y(v_2,y^*-\bar y)y_1\\&\quad+E_{uu}(\bar y,\bar u)(v_2,v_1)+E_{uy}(\bar y,\bar u)(y_2,v_1)+E_{uy}(\bar y,\bar u)r_2(\bar u,v_2)v_1+r_1^u(v_2,y^*-\bar y)v_1\\&\quad+\left(E_y(y^*,u^*)-E_y(\bar y,\bar u)\right)(\tilde y-y_1)=0.\end{aligned}$$

Let us write $\tilde y-y_1=z+\tilde r$, where z solves

$$E_y(\bar y,\bar u)z+E_{yy}(\bar y,\bar u)(y_1,y_2)+E_{yu}(\bar y,\bar u)(v_2,y_1)+E_{uu}(\bar y,\bar u)(v_2,v_1)+E_{uy}(\bar y,\bar u)(v_1,y_2)=0 \quad (2.29)$$

and $\tilde r$ solves

$$\begin{aligned}&E_y(\bar y,\bar u)\tilde r+E_{yy}(\bar y,\bar u)r_2(\bar u,v_2)y_1+E_{uy}(\bar y,\bar u)r_2(\bar u,v_2)v_1+r_1^y(v_2,y^*-\bar y)y_1+r_1^u(v_2,y^*-\bar y)v_1\\&\quad+\left(E_y(y^*,u^*)-E_y(\bar y,\bar u)\right)(\tilde y-y_1)=0.\end{aligned} \quad (2.30)$$

Thanks to the result of Proposition 2.7, the operator $E_y(\cdot,\cdot)$ is continuously invertible on $Y\times U_{ad}$. Hence, from (2.30) we estimate

$$\begin{aligned}\delta\|\tilde r\|_Y\le&\|E_{yy}(\bar y,\bar u)\|_{\mathcal{L}(Y,\mathcal{L}(Y,Y^*))}\|r_2(\bar u,v_2)\|_Y\|y_1\|_Y\\&+\|E_{uy}(\bar y,\bar u)\|_{\mathcal{L}(Y,\mathcal{L}(U,Y^*))}\|r_2(\bar u,v_2)\|_Y\|v_1\|_U\\&+\|r_1^y(v_2,y^*-\bar y)\|_{\mathcal{L}(Y,Y^*)}\|y_1\|_Y+\|r_1^u(v_2,y^*-\bar y)\|_{\mathcal{L}(U,Y^*)}\|v_1\|_U\\&+\|E_y(y^*,u^*)-E_y(\bar y,\bar u)\|_{\mathcal{L}(Y,Y^*)}\|\tilde y-y_1\|_Y.\end{aligned} \quad (2.31)$$

Let us estimate the last term on the right side of (2.31). The norm $\|y^*-\bar y\|_Y$ is already estimated

through (2.21) by $\|y^* - \bar{y}\|_Y \leq L\|v_2\|_U$. Therefore using the Lipschitz estimate (2.18) we obtain

$$\begin{aligned}\|E_y(y^*,u^*) - E_y(\bar{y},\bar{u})\|_{\mathcal{L}(Y,Y^*)} &\leq L_{E_y}\left(\|y^* - \bar{y}\|_Y + \|u^* - \bar{u}\|_U\right)\\ &\leq L_{E_y}\left(L\|v_2\|_U + \|v_2\|_U\right)\\ &= L_{E_y}\left(L+1\right)\|v_2\|_U.\end{aligned} \tag{2.32}$$

Now for the estimate $\|\tilde{y} - y_1\|_Y$, by the continuous invertibility of E_y (c.f Proposition 2.7), we estimate through (2.20)

$$\|\tilde{y} - y_1\|_Y \leq \delta^{-1}\Big(\|E_y(y^*,u^*) - E_y(\bar{y},\bar{u})\|_{\mathcal{L}(Y,Y^*)}\|y_1\|_Y + \|E_u(y^*,u^*) - E_u(\bar{y},\bar{u})\|_{\mathcal{L}(U,Y^*)}\|v_1\|_U\Big).$$

Similarly as in (2.32), it holds $\|E_u(y^*,u^*) - E_u(\bar{y},\bar{u})\|_{\mathcal{L}(U,Y^*)} \leq L_{E_u}(L+1)\|v_2\|_U$. We also estimate

$$\|y_1\|_Y = \|S'(\bar{u})v_1\|_Y \leq \|S'(\bar{u})\|_{\mathcal{L}(U,Y^*)}\|v_1\|_U \leq c\|v_1\|_U. \tag{2.33}$$

On inserting the above estimates in the last addend of (2.31) we obtain

$$\|E_y(y^*,u^*) - E_y(\bar{y},\bar{u})\|_{\mathcal{L}(Y,Y^*)}\|\tilde{y} - y_1\|_Y \leq L'\|v_1\|_U\|v_2\|_U^2 \tag{2.34}$$

where the constant $L' = L_{E_y}(L+1)^2(cL_{E_y} + L_{E_u})$. Using (2.33), (2.34) in (2.31) we obtain

$$\begin{aligned}\delta\|\tilde{r}\|_Y \leq\ & c\|E_{yy}(\bar{y},\bar{u})\|_{\mathcal{L}(Y,\mathcal{L}(Y,Y^*))}\|r_2(\bar{u},v_2)\|_Y\|v_1\|_U + \|E_{uy}(\bar{y},\bar{u})\|_{\mathcal{L}(Y,\mathcal{L}(U,Y^*))}\|r_2(\bar{u},v_2)\|_Y\|v_1\|_U\\ &+ c\|r_1^y(v_2,y^* - \bar{y})\|_{\mathcal{L}(Y,Y^*)}\|v_1\|_U + \|r_1^u(v_2,y^* - \bar{y})\|_{\mathcal{L}(U,Y^*)}\|v_1\|_U + L'\|v_1\|_U\|v_2\|_U^2.\end{aligned} \tag{2.35}$$

Now observe that

$$\|r(\bar{u},v_2)\|_{\mathcal{L}(U,Y)} = \sup_{v_1\in U}\frac{\|r(\bar{u},v_2)v_1\|_Y}{\|v_1\|_U} = \sup_{v_1\in U}\frac{\|\tilde{r}\|_Y}{\|v_1\|_U}.$$

Hence, invoking properties (2.25),(2.28), we divide the above expression by $\|v_2\|_U$ and passing to the limit $\|v_2\|_U \to 0$ to obtain the remainder property (2.19).

It remains to show that z defined by (2.29) is a linear and continuous operator. Using monotonicity of E_y, from (2.29) we derive

$$\begin{aligned}\delta\|z\|_Y \leq\ & \|E_{yy}(\bar{y},\bar{u})\|_{\mathcal{L}(Y,\mathcal{L}(Y,Y^*))}\|y_1\|_Y\|y_2\|_Y + \|E_{yu}(\bar{y},\bar{u})\|_{\mathcal{L}(Y,\mathcal{L}(U,Y^*))}\|y_1\|_Y\|v_2\|_U\\ &+ \|E_{uu}(\bar{y},\bar{u})\|_{\mathcal{L}(U,\mathcal{L}(U,Y^*))}\|v_2\|_U\|v_1\|_U + \|E_{uy}(\bar{y},\bar{u})\|_{\mathcal{L}(Y,\mathcal{L}(U,Y^*))}\|y_2\|_Y\|v_1\|_U.\end{aligned} \tag{2.36}$$

By definition, the functions $y_i, i = 1, 2$ are bounded:

$$\|y_i\|_Y = \|S'(\bar{u})v_i\|_Y \leq \|S'(\bar{u})\|_{\mathcal{L}(U,Y^*)}\|v_i\|_U.$$

Hence, applying the above estimate in (2.36), the boundedness of z follows. The proof ends by setting $G(v_2)v_1 = z$. □

Observe that since the functional J is twice continuously Fréchet differentiable, then by the results of Lemmas 2.15, 2.16 the reduced cost functional f in (2.9) is also twice Fréchet differentiable. This allows us to derive the optimality conditions below.

2.3 Optimality conditions

The roles of optimality conditions are to identify (necessary condition) and verify (sufficient conditions) optimality of solutions to a given optimal control problem. Optimality conditions are conveniently derived with the aid of Lagrange functional. For optimal control problem (P), we define the associated Lagrange functional $\mathcal{L} : Y \times U_{ad} \times Y \to \mathbb{R}$ by

$$\mathcal{L}(y, u, p) = g(y) + j(u) + \langle E(y, u),\ p\rangle_{Y,Y^*}. \tag{2.37}$$

Let us now discuss the optimality conditions characterizing the solutions of problem (P).

2.3.1 Optimality system

The following characterization of the optimal solution of (P) is standard, see e.g. [52, Sections 2.10, 2.13]. Here, the continuous invertibility of operator E_y from Proposition 2.7 is essential.

Theorem 2.17. *Let $\mathcal{L}$ be given by (2.37). Let $(\bar{y}, \bar{u})$ be a locally optimal solution of (P). Then the first-order necessary optimality conditions hold: there exists a Lagrange multiplier $\bar{p} \in Y$ such that it hold*

$$\begin{aligned}
\mathcal{L}_y(\bar{y}, \bar{u}, \bar{p})h &= 0,\\
\mathcal{L}_p(\bar{y}, \bar{u}, \bar{p})h &= 0,\\
\langle \mathcal{L}_u(\bar{y}, \bar{u}, \bar{p}), u - \bar{u}\rangle_{U^*,U} &\geq 0,\ \forall u \in U_{ad}
\end{aligned}$$

for every $h \in Y$.

The above system translates to

$$E_y(\bar{y}, \bar{u})^*\bar{p} = -g'(\bar{y}), \tag{2.38}$$

$$E(\bar{y}, \bar{u}) = 0, \tag{2.39}$$

$$\langle j'(\bar{u}) + E_u(\bar{y}, \bar{u})^*\bar{p},\ u - \bar{u}\rangle_{U^*,U} \geq 0,\ \forall u \in U_{ad}. \tag{2.40}$$

Equation (2.40) is sometimes referred to as variational inequality. The optimality system above can as well be written in terms of the reduced cost functional f defined in (2.9), namely

$$f'(\bar{u})(u - \bar{u}) \geq 0 \quad \forall u \in U_{ad} \tag{2.41}$$

where $f'(\bar{u})v = J_y(S(u), u)S'(u)v + J_u(S(u), u)v$. This latter form will be employed in the proof of sufficiency of the second-order condition to be introduced shortly.

As a consequence of Corollary 2.8 the adjoint state $\bar{p}$ is uniquely determined from (2.38) for every $g'(\bar{y}) \in Y^*$.

Let us conclude our discussion on the first-order necessary condition with the following observation.

Lemma 2.18. *Let $\bar{u} \in U_{ad}$ be given with corresponding state $\bar{y}$ and adjoint state $\bar{p}$. Let $(\bar{y}, \bar{u}, \bar{p})$ satisfies (2.40). Then almost everywhere $x \in \Omega$ it holds*

$$\bar{u}(x) = \begin{cases} u_a(x) & \text{if } j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^* \bar{p})(x) > 0, \\ u_b(x) & \text{if } j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^* \bar{p})(x) < 0 \end{cases} \tag{2.42}$$

and

$$j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^* \bar{p})(x) = 0 \quad \text{if } u_a(x) < \bar{u}(x) < u_a(x).$$

Proof. We start the proof by defining the sets

$$I_a = \{x \in \Omega : j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^* \bar{p})(x) > 0\},$$
$$I_b = \{x \in \Omega : j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^* \bar{p})(x) < 0\}.$$

Let us now establish the first statement of the lemma. We prove the result by contradiction.

Suppose the first hypothesis is false, that is, there is a measurable set $S_a \subset I_a$ with positive measure on which $\bar{u}(x) > u_a(x)$ for almost every $x \in S_a$. Now consider the function u defined as

$$u(x) = \begin{cases} u_a(x) & \text{for } x \in S_a \\ \bar{u}(x) & \text{for } x \in \Omega \setminus S_a. \end{cases} \tag{2.43}$$

Clearly $u_a \in U_{ad}$, and by assumption $\bar{u} \in U_{ad}$. It is then immediate that u defined by (2.43)

belongs to U_{ad}. Furthermore, $u(x) - \bar{u}(x) < 0$ on S_a and $u(x) - \bar{u}(x) = 0$ on its compliment. Now using this u as a test function in (2.40), we obtain

$$\begin{aligned}\langle j'(\bar{u}) + E_u(\bar{y},\bar{u})^*\bar{p},\ u-\bar{u}\rangle_{U^*,U} &= \int_\Omega (j'(\bar{u}(x)) + (E_u(\bar{y},\bar{u})^*\bar{p})(x))(u(x)-\bar{u}(x))\mathrm{d}x \\ &= \int_{S_a} (j'(\bar{u}(x)) + (E_u(\bar{y},\bar{u})^*\bar{p})(x))(u(x)-\bar{u}(x))\mathrm{d}x \\ &\quad + \int_{\Omega\setminus S_a} (j'(\bar{u}(x)) + (E_u(\bar{y},\bar{u})^*\bar{p})(x))(u(x)-\bar{u}(x))\mathrm{d}x \\ &= \int_{S_a} (j'(\bar{u}(x)) + (E_u(\bar{y},\bar{u})^*\bar{p})(x))(u(x)-\bar{u}(x))\mathrm{d}x \\ &< 0\end{aligned}$$

which contradicts the first order necessary condition (2.40). Hence we must have $\bar{u} = u_a$ on I_a as claimed. By a similar technique with appropriate modifications, the required result on the set I_b is obtained.

It is then easy to see that on $I_a \cap I_b = \{x \in \Omega : u_a(x) < \bar{u} < u_b(x)\}$, we must have $j'(\bar{u}(x)) + (E_u(\bar{y},\bar{u})^*\bar{p})(x) = 0$ almost everywhere $x \in \Omega$. □

2.3.2 Second-order conditions

As the name suggests, the second-order sufficient conditions are essential in establishing the optimality of a solution candidate fulfilling the first-order condition. Let us now derive the second-order sufficient conditions for the abstract problem (P).

Second-order necessary condition

Motivated by the characterization of the optimal control $\bar{u}$ by (2.42), let us introduce the set

$$\mathcal{A}_0(\bar{u}) = \{x \in \Omega : |j'(\bar{u}(x)) + (E_u(\bar{y},\bar{u})^*\bar{p})(x)| > 0\}.$$

This set $\mathcal{A}_0$ will play an important role in specifying the second-order optimality conditions in the subsequent discussion. Let us define the critical cone

Definition 2.19. *The cone $C_0(\bar{u})$ is the set of all $v \in U_{ad}$ such that*

$$v(x)\begin{cases} = 0 & \text{if } x \in \mathcal{A}_0(\bar{u}), \\ \geq 0 & \text{if } \bar{u} = u_a(x) \text{ and } x \notin \mathcal{A}_0(\bar{u}), \\ \leq 0 & \text{if } \bar{u} = u_b(x) \text{ and } x \notin \mathcal{A}_0(\bar{u}). \end{cases}$$

Then in view of the abstract problem (P), the second-order necessary condition satisfied by the solution $(\bar{y},\bar{u})$ of (P) together with the corresponding adjoint state $\bar{p}$ is given by

$$\mathcal{L}''(\bar{y},\bar{u},\bar{p})[z,v]^2 \geq 0 \quad \forall v \in C_0(\bar{u}) \tag{2.44}$$

where z is the solution of the linearized equation

$$E_y(\bar{y},\bar{u})z + E_u(\bar{y},\bar{u})v = 0$$

see [52, Lemma 4.28].

As a justification to expanding the set $C_0(\bar{u})$ for specifying the second-order sufficient condition, observe that for all functions v satisfying $v \neq 0$ on $\mathcal{A}_0(\bar{u})$, the necessary second-order condition (2.44) delivers no information for such directions. In fact, an example given in [19] (and restated in [13, Example 2.1]) showed that second-order condition based on $C_0(\bar{u})$ is not even sufficient for local optimality of solution in the general case. Consequently the cone of critical directions $C_0(\bar{u})$ has to be enlarged. For that purpose, let us define a new cone $C(\bar{u})$ as the set of all $v \in U_{ad}$ such that

$$v(x)\begin{cases} \geq 0 & \text{if } \bar{u}(x) = u_a(x), \\ \leq 0 & \text{if } \bar{u}(x) = u_b(x). \end{cases}$$

Now using the new cone $C(\bar{u})$, the second-order sufficient condition is given by the variational inequality (2.38)-(2.40) and the condition: there exists $\delta > 0$ such that

$$\mathcal{L}''(\bar{y},\bar{u},\bar{p})[z,v]^2 \geq \delta\|v\|_U^2 \quad \forall v \in C(\bar{u}) \tag{2.45}$$

where z is the solution of the linearized equation $E_y(\bar{y},\bar{u})z + E_u(\bar{y},\bar{u})v = 0$.

Let us now establish the sufficiency of condition (2.45) together with the first optimality condition (2.40). Before proceeding to the proof, observe that using $\bar{y} = S(\bar{u})$, the second-order condition (2.45) is equivalent to the condition: there exists $\delta > 0$ such that

$$f''(\bar{u})[v,v] \geq \delta\|v\|_U^2 \quad \forall v \in C(\bar{u}). \tag{2.46}$$

The corresponding necessary first-order condition in terms of the reduced functional f is given by (2.41). We will now prove the sufficiency of condition (2.46) together with (2.41) instead.

The following property [52, Lemma 4.26] will be essential in the proof of the theorem that follows.

Assumption 4. *For each $M > 0$ there exists a constant $L_{f''}(M) > 0$ such that*

$$|f''(u+h)[u_1,u_2] - f''(u)[u_1,u_2]| \leq L_{f''}(M)\|h\|_{L^\infty}\|u_1\|_U\|u_2\|_U \tag{2.47}$$

for all $u, u_1, u_2, h \in U_{ad}$ with $\max\{\|u\|_{L^\infty}, \|h\|_{L^\infty}\} \leq M$.

Theorem 2.20. *Let $U_{ad} \subseteq L^\infty(\Omega)$ and let $\bar{u} \in U_{ad}$ fulfills the first-order necessary condition*

(2.41). *If $\bar{u}$ additionally fulfills the second-order condition* (2.46), *then there exist $\epsilon > 0, \delta > 0$ such that for all $u \in U_{ad}$ with $\|u - \bar{u}\|_{L^\infty} \leq \epsilon$ we have the quadratic growth condition*

$$f(u) \geq f(\bar{u}) + \frac{\delta}{4}\|u - \bar{u}\|_U^2.$$

Proof. Using (2.41) we can estimate

$$\begin{aligned} f(u) - f(\bar{u}) &= f'(\bar{u})(u - \bar{u}) + \frac{1}{2}f''(\bar{u} + \theta(u - \bar{u}))(u - \bar{u})^2, \quad \theta \in (0, 1) \\ &\geq \frac{1}{2}f''(\bar{u} + \theta(u - \bar{u}))(u - \bar{u})^2 \\ &= \frac{1}{2}f''(\bar{u})(u - \bar{u})^2 + \frac{1}{2}[f''(\bar{u} + \theta(u - \bar{u})) - f''(\bar{u})](u - \bar{u})^2. \end{aligned} \tag{2.48}$$

Since $u \in U_{ad}$, it is easy to see that $u - \bar{u} \in C(\bar{u})$. Thus the second-order sufficient condition (2.46) applies to the first addend above. We estimate the second addend using (2.47). Altogether, if $\|u - \bar{u}\|_U \leq \epsilon$ with a sufficiently small $\epsilon > 0$ we obtain

$$\begin{aligned} f(u) - f(\bar{u}) &= \frac{1}{2}f''(\bar{u})(u - \bar{u})^2 + \frac{1}{2}[f''(\bar{u} + \theta(u - \bar{u})) - f''(\bar{u})](u - \bar{u})^2 \\ &\geq \frac{\delta}{2}\|u - \bar{u}\|_U^2 - \frac{1}{2}L_{f''}\|u - \bar{u}\|_{L^\infty}\|u - \bar{u}\|_U^2 \\ &\geq \frac{\delta}{4}\|u - \bar{u}\|_U^2. \end{aligned}$$

In the above, $\epsilon > 0$ has been chosen in such a way that $\epsilon \leq \frac{\delta}{2L_{f''}}$. □

We observe that the second-order sufficient condition (2.45) is in general an overly restrictive condition. It appears the cone $C(\bar{u})$ is too large compared to $C_0(\bar{u})$. The gap thus created between these sets can be partially closed by introducing the strongly active constraints. This idea was first introduced in [18]. The goal of the next section is therefore to discuss strongly active sets in the context of the model problem (P).

2.3.3 Strongly active constraints

Let us consider the set

$$\mathcal{A}_\tau(\bar{u}) = \{x \in \Omega : |j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^* \bar{p})(x)| > \tau\}$$

which shall be called the strongly active set. To specify the second-order sufficient condition in this case, let us define

Definition 2.21. *The $\tau-$critical cone $C_\tau(\bar{u})$ is the set of $v \in U$ such that*

$$v(x)\begin{cases} = 0 & \text{if } x \in \mathcal{A}_\tau(\bar{u}), \\ \geq 0 & \text{if } \bar{u} = u_a(x) \text{ and } x \notin \mathcal{A}_\tau(\bar{u}), \\ \leq 0 & \text{if } \bar{u} = u_b(x) \text{ and } x \notin \mathcal{A}_\tau(\bar{u}). \end{cases}$$

For proving the optimality of $\bar{u}$ it is then sufficient to require the second-order condition with respect to the critical directions $v \in C_\tau(\bar{u})$, see [15], [52]. Note that to specify a second-order sufficient condition, the choice $\tau = 0$ is not allowed due to a counter example given in [19], see also [13, Example 2.1]. The less restrictive second-order sufficient condition in this case therefore states: there exists $\delta > 0$ and $\tau > 0$ such that

$$\mathcal{L}''(\bar{y}, \bar{u}, \bar{p})(z, v)^2 \geq \delta \|v\|_U^2 \quad \forall v \in C_\tau(\bar{u}) \tag{2.49}$$

where z solves the linearized equation $E_y(\bar{y}, \bar{u})z + E_u(\bar{y}, \bar{u})v = 0$.

Let us now prove that (2.49) is sufficient for local optimality of $\bar{u}$ fulfilling the first order condition (2.40). Again we prefer to work with the reduced functional f and write (2.49) as

$$f''(\bar{u})(v, v) \geq \delta \|v\|_U^2 \quad \forall v \in C_\tau(\bar{u}). \tag{2.50}$$

Remark 2.22. *In the proof to be given, we have to deal with the so-called two-norm discrepancy. Two-norm discrepancy is a common occurrence when specifying second-order sufficient conditions for optimal control of nonlinear PDEs. It is a situation where the cost functional J is differentiable in the L^∞ norm whereas the second-order sufficient condition is only specified in the weaker norm of the control space U, see e.g [41]. Consequently we will prove that (2.46) implies local optimality in the L^∞ sense.*

In line with the above remark, let us suppose the cost functional J is twice Fréchet differentiable with respect to the $L^\infty-$norm. With the setting $f(u) = J(y(u), u)$, the same property will hold true for the reduced functional f.

Theorem 2.23. *Let $U_{ad} \subseteq L^\infty(\Omega)$ and let Assumption 4 hold. Let $\bar{u} \in U_{ad}$, $\bar{y} = S(\bar{u})$ together with the associated adjoint state $\bar{p}$ satisfy the first order necessary condition (2.40). Suppose in addition that $(\bar{y}, \bar{u})$ fulfills (2.49) for some $\tau > 0, \delta > 0$. Then there exist $\epsilon > 0$ and $\sigma > 0$ such that for all $u \in U_{ad}, y = S(u)$ with $\|u - \bar{u}\|_{L^\infty} \leq \epsilon$ we have the quadratic growth condition*

$$J(y, u) \geq J(\bar{y}, \bar{u}) + \sigma \|u - \bar{u}\|_U^2.$$

Proof. We will follow the method of proof given in [52]. Again we set $f(u) = J(S(u), u)$. Taylor expansion of f gives

$$f(u) - f(\bar{u}) = f'(\bar{u})h + \frac{1}{2} f''(\bar{u})h^2 + r^f \tag{2.51}$$

where $h = u - \bar{u}$ and r^f denotes the second order remainder term in the expansion. We now have to estimate the first two addends of the above expansion. Let us start with the following preliminary computations. Through the Lagrange functional $\mathcal{L}$ we derive

$$f''(\bar{u})[v_1, v_2] = \mathcal{L}''(\bar{y}, \bar{u}, \bar{p})[(z_1, v_1), (z_2, v_2)]. \tag{2.52}$$

By Corollary 2.8, the mapping $S'(\bar{u})$ is continuous. As a result, we can estimate the norm of $z_i = S'(\bar{u})v_i$ by $\|z_i\|_Y \leq c\|v_i\|_U$. Therefore using the representation (2.52) we estimate

$$\begin{aligned}
|f''(\bar{u})[v_1, v_2]| &\leq |\mathcal{L}''(\bar{y}, \bar{u}, \bar{p})[(z_1, v_1), (z_2, v_2)]| \\
&\leq |\mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p})[z_1, z_2] + |\mathcal{L}_{uy}(\bar{y}, \bar{u}, \bar{p})[v_1, z_2]| \\
&\quad + |\mathcal{L}_{yu}(\bar{y}, \bar{u}, \bar{p})[z_1, v_2]| + |\mathcal{L}_{uu}(\bar{y}, \bar{u}, \bar{p})[v_1, v_2]| \\
&\leq \tilde{C}\Big(\|z_1\|_Y\|z_2\|_Y + \|v_1\|_U\|z_2\|_Y + \|z_1\|_Y\|v_2\|_U + \|v_1\|_U\|v_2\|_U\Big) \\
&\leq C\|v_1\|_U\|v_2\|_U,
\end{aligned} \tag{2.53}$$

where C is a generic constant.

Let us now estimate the term $f'(\bar{u})h$ in (2.51). Note that by the first-order condition (2.40), the pointwise variational inequality

$$\Big(j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^*\bar{p})(x)\Big)h(x) \geq 0 \quad a.e \text{ in } \Omega \ \forall h = u - \bar{u}, u \in U_{ad}$$

holds. Then using the definition of set $\mathcal{A}_\tau$ we have

$$\begin{aligned}
f'(\bar{u})h &= \int_\Omega \Big(j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^*\bar{p})(x)\Big)h(x)\mathrm{d}x \\
&= \int_\Omega |j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^*\bar{p})(x)|\ |h(x)|\mathrm{d}x \\
&= \int_{\mathcal{A}_\tau} |j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^*\bar{p})(x)|\ |h(x)|\mathrm{d}x \\
&\quad + \int_{\Omega\setminus\mathcal{A}_\tau} |j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^*\bar{p})(x)|\ |h(x)|\mathrm{d}x \\
&\geq \int_{\mathcal{A}_\tau} |j'(\bar{u}(x)) + (E_u(\bar{y}, \bar{u})^*\bar{p})(x)|\ |h(x)|\mathrm{d}x \\
&\geq \int_{\mathcal{A}_\tau} \tau|h(x)|\mathrm{d}x = \tau\|h\|_{L^1(\mathcal{A}_\tau)}.
\end{aligned} \tag{2.54}$$

We continue our proof with the computation of the second derivative $\frac{1}{2}f''(\bar{u})h^2$. Let us split $h := h_0 + h_1$ such that

$$h_0(x) = \begin{cases} h(x) & \text{if } x \notin \mathcal{A}_\tau, \\ 0 & \text{if } \ x \in \mathcal{A}_\tau. \end{cases}$$

Therefore we obtain

$$\frac{1}{2}f''(\bar{u})h^2 = \frac{1}{2}f''(\bar{u})(h_0+h_1)^2 = \frac{1}{2}f''(\bar{u})h_0^2 + f''(\bar{u})[h_0,h_1] + \frac{1}{2}f''(\bar{u})h_1^2. \tag{2.55}$$

By definition, $h_0 \in C_\tau(\bar{u})$ and thus the second-order sufficient condition (2.49) (or equivalently (2.50)) applies to the first term of the above equation. It follows that

$$\frac{1}{2}f''(\bar{u})h_0^2 \geq \frac{\delta}{2}\|h_0\|_U^2. \tag{2.56}$$

To estimate the term $f''(\bar{u})[h_0,h_1]$ in (2.55), we will make use of the interpolation inequality (see [12])

$$\|h_1\|_{L^2(\Omega)}^2 \leq \|h_1\|_{L^1(\Omega)}\|h_1\|_{L^\infty(\Omega)}. \tag{2.57}$$

Now for sufficiently small $\epsilon \in (0,1)$, with $\|h\|_{L^\infty(\Omega)} \leq \epsilon$ it clearly holds $\|h_1\|_{L^\infty(\Omega)} \leq \|h\|_{L^\infty(\Omega)} \leq \epsilon$. Then applying (2.53) yields

$$|f''(\bar{u})[h_0,h_1]| \leq C\|h_0\|_{L^2(\Omega)}\|h_1\|_{L^2(\Omega)}. \tag{2.58}$$

Employing Young inequality and (2.57), we estimate (2.58) further as

$$\begin{aligned}
|f''(\bar{u})[h_0,h_1]| &\leq \frac{\delta}{4}\|h_0\|_{L^2(\Omega)}^2 + c_1\|h_1\|_{L^2(\Omega)}^2 \\
&\leq \frac{\delta}{4}\|h_0\|_{L^2(\Omega)}^2 + c_1\|h_1\|_{L^1(\Omega)}\|h_1\|_{L^\infty(\Omega)} \\
&\leq \frac{\delta}{4}\|h_0\|_{L^2(\Omega)}^2 + c_1\epsilon\|h_1\|_{L^1(\Omega)}
\end{aligned}$$

from which we obtain

$$f''(\bar{u})[h_0,h_1] \geq -\Big(\frac{\delta}{4}\|h_0\|_{L^2(\Omega)}^2 + c_1\epsilon\|h_1\|_{L^1(\Omega)}\Big). \tag{2.59}$$

Similarly we estimate

$$\Big|\frac{1}{2}f''(\bar{u})h_1^2\Big| \leq C\|h_1\|_{L^2(\Omega)}^2 \leq c_2\|h_1\|_{L^1(\Omega)}\|h_1\|_{L^\infty(\Omega)} \leq c_2\epsilon\|h_1\|_{L^1(\Omega)}$$

so that

$$\frac{1}{2}f''(\bar{u})h_1^2 \geq -c_2\epsilon\|h_1\|_{L^1(\Omega)}. \tag{2.60}$$

Altogether using (2.56), (2.59) and (2.60) in (2.55) we obtain

$$\begin{aligned}\frac{1}{2}f''(\bar{u})h^2 &\geq \frac{\delta}{4}\|h_0\|^2_{L^2(\Omega)} - (c_1+c_2)\epsilon\|h_1\|_{L^1(\Omega)} \\ &\geq \frac{\delta}{4}\|h_0\|^2_{L^2(\Omega)} - \frac{\tau}{2}\|h_1\|_{L^1(\Omega)}\end{aligned}$$

where $\epsilon > 0$ has been chosen in such a way that $(c_1+c_2)\epsilon \leq \frac{\tau}{2}$. By definition $h_1 = 0$ on $\Omega \setminus \mathcal{A}_\tau$ and $h_0 = 0$ on $\mathcal{A}_\tau$. It therefore holds

$$\|h_1\|_{L^1(\Omega)} = \|h\|_{L^1(\mathcal{A}_\tau)}.$$

Hence

$$\begin{aligned}\frac{1}{2}f''(\bar{u})h^2 &\geq \frac{\delta}{4}\|h_0\|^2_{L^2(\Omega)} - \frac{\tau}{2}\|h_1\|_{L^1(\Omega)} \\ &= \frac{\delta}{4}\|h_0\|^2_{L^2(\Omega)} - \frac{\tau}{2}\|h\|_{L^1(\mathcal{A}_\tau)}.\end{aligned} \tag{2.61}$$

Now using (2.61) and (2.54) in (2.51) we arrive at

$$\begin{aligned}f(u) - f(\bar{u}) &\geq \tau\|h\|_{L^1(\mathcal{A}_\tau)} + \frac{\delta}{4}\|h_0\|^2_{L^2(\Omega)} - \frac{\tau}{2}\|h\|_{L^1(\mathcal{A}_\tau)} + r^f \\ &\geq \frac{\tau}{2}\|h\|_{L^1(\mathcal{A}_\tau)} + \frac{\delta}{4}\|h_0\|^2_{L^2(\Omega)} + r^f.\end{aligned} \tag{2.62}$$

We estimate the addend on the right side of the above inequality through the following arguments. Without loss of generality, we choose $\epsilon \leq 1$. By the hypothesis of the theorem, we have $\|h\|_{L^\infty(\Omega)} \leq \epsilon \leq 1$. As a result, it holds $|h(x)| \geq h(x)^2$. Also observe that

$$\|h_0\|^2_{L^2(\Omega)} = \int_{\Omega\setminus\mathcal{A}_\tau} h_0^2 = \int_{\Omega\setminus\mathcal{A}_\tau} h^2 = \|h\|^2_{L^2(\Omega\setminus\mathcal{A}_\tau)}.$$

Hence we continue our estimation in (2.62) with

$$\begin{aligned}f(u) - f(\bar{u}) &\geq \frac{\tau}{2}\|h\|_{L^1(\mathcal{A}_\tau)} + \frac{\delta}{4}\|h_0\|^2_{L^2(\Omega)} + r^f \\ &\geq \frac{\tau}{2}\int_{\mathcal{A}_\tau} |h(x)|\,\mathrm{d}x + \frac{\delta}{4}\|h\|^2_{L^2(\Omega\setminus\mathcal{A}_\tau)} + r^f \\ &\geq \frac{\tau}{2}\int_{\mathcal{A}_\tau} h(x)^2\,\mathrm{d}x + \frac{\delta}{4}\|h\|^2_{L^2(\Omega\setminus\mathcal{A}_\tau)} + r^f \\ &\geq \min\left\{\frac{\tau}{2}, \frac{\delta}{4}\right\}\|h\|^2_{L^2(\Omega)} + r^f.\end{aligned}$$

For the second-order remainder term r^f we have by (2.47)

$$\begin{aligned} r^f(\bar{u},h) &= f(u) - f(\bar{u}) - f'(\bar{u})h - \frac{1}{2}f''(\bar{u})h^2 \\ &= \int_0^1 f'(\bar{u}+sh)h\mathrm{d}s - f'(\bar{u})h - \frac{1}{2}f''(\bar{u})h^2 \\ &\leq \int_0^1 \int_0^s |f''(\bar{u}+th)h^2 - f''(\bar{u})h^2|\mathrm{d}t\mathrm{d}s \\ &\leq L_{f''}(M)\|h\|_{L^\infty(\Omega)}\|h\|^2_{L^2(\Omega)}. \end{aligned}$$

This gives

$$\frac{r^f(\bar{u},h)}{\|h\|^2_{L^2(\Omega)}} \leq L_{f''}(M)\|h\|_{L^\infty(\Omega)} \to 0 \quad \text{as} \quad \|h\|_{L^\infty(\Omega)} \to 0.$$

Finally based on the above estimations, we conclude that for sufficiently small $\epsilon > 0$ and $\|h\|_{L^\infty(\Omega)} \leq \epsilon$, it holds

$$J(y,u) - J(\bar{y},\bar{u}) = f(u) - f(\bar{u}) \geq \frac{1}{2}\min\left\{\frac{\tau}{2}, \frac{\delta}{4}\right\}\|h\|^2_{L^2(\Omega)} = \sigma\|h\|^2_{L^2(\Omega)}.$$

□

2.4 Finding the optimal solution

In this section we will be concerned with the task of finding the solution of (P) by solving the optimality system below for $(\bar{y},\bar{u})$

$$E_y(\bar{y},\bar{u})^*\bar{p} + g'(\bar{y}) = 0, \tag{2.63}$$

$$\langle j'(\bar{u}) + E_u(\bar{y},\bar{u})^*\bar{p},\ u-\bar{u}\rangle \geq 0 \quad \forall u \in U_{ad}, \tag{2.64}$$

$$E(\bar{y},\bar{u}) = 0. \tag{2.65}$$

Every solution to the above system is a solution candidate for the optimization problem (P). However, the obtained solution can only be judged optimal if it fulfills the second-order sufficient condition (2.45).

To be able to solve (2.63)-(2.65), the system has to be transformed into a system of equations by getting rid of the variational inequality. This is achieved with the aid of the projection operator $P_{U_{ad}}$ (defined by (2.67) below). The projection operator $P_{U_{ad}}$ transforms the variational inequality into a corresponding equation. As a necessity, the choice of the functional

$$j(u) = \frac{\alpha}{2}\|u\|^2_U,\ \alpha > 0 \tag{2.66}$$

is needed for the subsequent analysis.

Projection onto the space of admissible controls

For a box-type control constraint of the form $\{u \in U : u_a \leq u \leq u_b\}$, the projection $P_{U_{ad}} : U \to U_{ad}$ is defined as

$$\begin{aligned} P_{U_{ad}}(u) &= \begin{cases} u_a & \text{for } u < u_a \\ u & \text{for } u \in (u_a, u_b) \\ u_b & \text{for } u_b < u \end{cases} \\ &= \max\Big(u_a, \min(u_b, u)\Big). \end{aligned} \tag{2.67}$$

The function max above denotes the point-wise maximum operation.

Now using the introduced function $P_{U_{ad}}$, we will derive an equivalent equation for (2.64). The results are summarized in the following lemma.

Lemma 2.24. *Let $U_{ad} \subseteq U$ be given as in Lemma 2.13 and j be given by the quadratic functional (2.66). If $U = U_{ad}$ then the variational inequality (2.64) is equivalent to*

$$\bar{u} = -\frac{1}{\alpha} E_u(\bar{y}, \bar{u})^* \bar{p}.$$

On the other hand if $U_{ad} \subset U$, the variational inequality is equivalent to

$$\bar{u} = P_{U_{ad}}\Big(-\frac{1}{\alpha} E_u(\bar{y}, \bar{u})^* \bar{p}\Big).$$

where $P_{U_{ad}}$ is defined by (2.67).

Proof. Since $\alpha > 0$, the variational inequality (2.64) is the same as

$$\langle \bar{u} + \frac{1}{\alpha} E_u(\bar{y}, \bar{u})^* \bar{p},\ u - \bar{u} \rangle_{U^*, U} \geq 0,\ \forall u \in U_{ad}. \tag{2.68}$$

Let $\bar{u} \in U_{ad} \subset U$ and let us set $v := -\frac{1}{\alpha} E_u(\bar{y}, \bar{u})^* \bar{p}$. Then by the projection theorem in Hilbert spaces

$$\langle \bar{u} - v, u - \bar{u} \rangle \geq 0 \ \ \forall u \in U_{ad}$$

holds if and only if $\bar{u} = P_{U_{ad}}(v)$, which gives the first claim.

For the other claim, observe that if $U = U_{ad}$, the projection operator $P_{U_{ad}}$ is an identity mapping. The claim then follows easily. □

Example 2. *Let $\Omega \subseteq \mathbb{R}^n$. Consider the optimization problem: Minimize*

$$J(y, u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}, \quad y_d \in L^2(\Omega),\ \alpha > 0$$

subject to the constraint: $u \in L^2(\Omega)$

$$-\Delta y = u \text{ in } \Omega, \quad y = 0 \text{ on } \partial\Omega$$

and $u \in U_{ad} = \{u \in L^2(\Omega) : u_a \leq u \leq u_b\}$.

The solution $(\bar{y}, \bar{u})$ *to the above problem is obtained as a solution of the optimality system*

$$\begin{aligned} -\Delta \bar{y} &= \bar{u}, \\ \langle \alpha \bar{u} + \bar{p}, u - \bar{u} \rangle_U &\geq 0, \ \forall u \in U_{ad}, \\ -\Delta \bar{p} &= \bar{y} - y_d \end{aligned}$$

which by Lemma 2.24 can be replaced with a system of equations

$$\begin{aligned} -\Delta \bar{y} &= \bar{u}, \\ -\Delta \bar{p} &= \bar{y} - y_d, \\ \bar{u} &= P_{U_{ad}}\Big(-\frac{1}{\alpha}\bar{p}\Big). \end{aligned}$$

Now let us go back to the abstract problem. As in the example above, the optimality system (2.63)-(2.65) can be written as

$$\begin{aligned} E_y(\bar{y}, \bar{u})^* \bar{p} + g'(\bar{y}) &= 0, \\ \bar{u} - P_{U_{ad}}\Big(-\frac{1}{\alpha} E_u(\bar{y}, \bar{u})^* \bar{p}\Big) &= 0, \\ E(\bar{y}, \bar{u}) &= 0. \end{aligned} \tag{2.69}$$

The above system now forms the basis of the next section.

2.4.1 Semi-smooth Newton method

The goal of this section is to solve the system (2.69). Although we got rid of the inequality in (2.64) by introducing the map $P_{U_{ad}}$, nevertheless its presence in the optimality system (2.63)-(2.65) poses another difficulty in solving the resulting system with classical Newton method. This is due to non-differentiability of the projection $P_{U_{ad}}$ in the classical sense. However in the sequel by recalling a differentiability concept that covers non-smooth functions, the differentiability of $P_{U_{ad}}$ will be proved in a sense to be made clear shortly.

Let us mention the following property of the projection operator $P_{U_{ad}}$.

Lemma 2.25. *Let* $U = L^2(\Omega)$ *and let its subset* U_{ad} *be as defined in Lemma 2.13. The mapping* $P_{U_{ad}} : U \to U$ *is Lipschitz continuous and non-expansive:*

$$\|P_{U_{ad}}(u_1) - P_{U_{ad}}(u_2)\|_U \leq \|u_1 - u_2\|_U \quad \forall u_1, u_2 \in U.$$

Proof. Let $u_1, u_2 \in U$. The projection theorem in Hilbert space implies

$$\langle P_{U_{ad}}(u_1) - u_1,\ z - P_{U_{ad}}(u_1)\rangle \geq 0 \quad \forall z \in U_{ad}. \tag{2.70}$$

Similarly it holds

$$\langle P_{U_{ad}}(u_2) - u_2,\ z - P_{U_{ad}}(u_2)\rangle \geq 0 \quad \forall z \in U_{ad}. \tag{2.71}$$

Then choosing $z = P_{U_{ad}}(u_2)$ in (2.70), $z = P_{U_{ad}}(u_1)$ in (2.71) and adding up gives

$$0 \leq \langle P_{U_{ad}}(u_1) - P_{U_{ad}}(u_2) + (u_2 - u_1),\ P_{U_{ad}}(u_2) - P_{U_{ad}}(u_1)\rangle.$$

The result then follows from

$$\begin{aligned}
0 &\leq \langle P_{U_{ad}}(u_1) - P_{U_{ad}}(u_2) + (u_2 - u_1),\ P_{U_{ad}}(u_2) - P_{U_{ad}}(u_1)\rangle \\
&= \langle P_{U_{ad}}(u_1) - P_{U_{ad}}(u_2),\ P_{U_{ad}}(u_2) - P_{U_{ad}}(u_1)\rangle + \langle u_2 - u_1,\ P_{U_{ad}}(u_2) - P_{U_{ad}}(u_1)\rangle \\
&\leq -\|P_{U_{ad}}(u_2) - P_{U_{ad}}(u_1)\|_U^2 + \|u_1 - u_2\|_U \|P_{U_{ad}}(u_2) - P_{U_{ad}}(u_1)\|_U.
\end{aligned}$$

□

Let us now define a notion of Newton differentiability that allows Newton methods to accommodate non-smooth functions. This concept was introduced in [34].

Definition 2.26. *Let X, Y be Banach spaces and let U be an open subset of X. The mapping $f : U \to Y$ is called Newton differentiable at $u \in U$ if there exists a family of mappings $G : U \to \mathcal{L}(X, Y)$ such that*

$$\lim_{h \to 0} \frac{\|f(u+h) - f(u) - G(u+h)h\|_Y}{\|h\|_X} = 0.$$

The mapping G is called the Newton derivative of f at u.

The following are useful examples of Newton differentiable mappings.

Example 3. *Let function f be as given above. Then a continuously Fréchet differentiable function f is Newton differentiable.*

Proof. Let f be continuously Fréchet differentiable at $u \in U$. Taking $G(u) = f'(u)$ we see that

$$\begin{aligned}
\frac{\|f(u+h) - f(u) - G(u+h)h\|_Y}{\|h\|_X} \leq{}& \frac{\|f(u+h) - f(u) - G(u)h\|_Y}{\|h\|_X} \\
&+ \left(\frac{\|G(u)h - G(u+h)h\|_Y}{\|h\|_X}\right).
\end{aligned}$$

Since G is continuous by assumption, passing to the limit $h \to 0$ yields the claim. □

Example 4. *[34, p 236] Let X be a Hilbert space with inner product $\langle x, x\rangle = \|x\|^2$. Then the norm functional $f(x) = \|x\|$ is Newton differentiable.*

We remark that Newton differentiability respects the chain rule of composition of mappings. This is formally stated below.

Lemma 2.27. *Let X, Y, Z be Banach spaces and D an open subset of X. Suppose $f : D \subset X \to Y$ is continuously Fréchet differentiable at $x \in D$, and let $g : Y \to Z$ be Newton differentiable at $f(x)$ with Newton derivative g'. Then $\phi = g(f) : D \subset X \to Z$ is Newton differentiable at x with Newton derivative $g'(f(x+h))f'(x+h) \in \mathcal{L}(X, Z)$ for a sufficiently small h.*

Proof. The proof to the lemma can be found on page 238 of [34]. □

Analogous result holds for the summation of Newton differentiable mappings, see e.g. [32, Theorem 2.10].

Let us now establish Newton differentiability of $P_{U_{ad}}$ which will enable us to solve (2.69) with semi-smooth Newton method (see e.g. [34]). The min-max representation of the projection operator can be simplified as

$$\begin{aligned} P_{U_{ad}}(u) &= \max\Big(u_a, \min(u_b, u)\Big) \\ &= \max\Big(0, \min(0, u - u_b) + u_b - u_a\Big) + u_a. \end{aligned} \tag{2.72}$$

Hence, in order to establish Newton differentiability of $P_{U_{ad}}$, it suffices to prove Newton differentiability of the maximum function. Analogous result will hold for the minimum function min . We have the following result [29, Proposition 4.1] for the differentiability of the max function in L^q spaces.

Lemma 2.28. *The mapping $\max(0, \cdot) : L^q(\Omega) \to L^p(\Omega)$ with $1 \le p < q \le \infty$ is Newton differentiable in $L^q(\Omega)$ and the function G_{max} given by*

$$G_{max}(u)(x) = \begin{cases} 1 & \text{if } u(x) > 0 \\ \delta & \text{if } u(x) = 0,\ \delta \text{ arbitrary} \\ 0 & \text{otherwise} \end{cases} \tag{2.73}$$

is a Newton derivative of the max *function.*

The derivative of the min function is analogously defined, namely

$$G_{min}(u)(x) = \begin{cases} 1 & \text{if } u(x) < 0 \\ \delta & \text{if } u(x) = 0,\ \delta \text{ arbitrary} \\ 0 & \text{otherwise.} \end{cases} \tag{2.74}$$

Remark 2.29. *As pointed out in [29] the function G_{max} in (2.73) cannot be chosen as a Newton derivative of* max *as a function from $L^p(\Omega) \to L^p(\Omega)$. The gap $q > p$ is in fact vital for the*

validity of the result of Lemma 2.28. In the lemma that follows, we will fix the arbitrary constant $\delta = 0$ *in the derivative of the max and min functions.*

A direct computation of the Newton derivative of $P_{U_{ad}}$ from (2.72) using (2.73) and (2.74) with $\delta = 0$ is given below.

Lemma 2.30. *If we define the set* $I = \{x \in \Omega : u_a(x) < u(x) < u_b(x)\}$ *then the Newton derivative* $G_{U_{ad}}$ *of* $P_{U_{ad}}$ *at* $u \in U$ *in the direction* $v \in U$ *is given by*

$$\Big(G_{U_{ad}}(u)\Big)(v) = \chi_I v$$

where χ_I *denotes the characteristic function of set* I*.*

The above computation of the derivative of $P_{U_{ad}}$ is only valid if the norm gap in Lemma 2.28 is respected. This concern shall be taken care of shortly. Let us go back to the optimality system (2.69). Since $U = L^2(\Omega)$ it is desirable that the mapping

$$u \to P_{U_{ad}}\Big(-\frac{1}{\alpha}E_u(y,u)^*p\Big)$$

be Newton differentiable from $L^2(\Omega) \to L^2(\Omega)$. From Lemma 2.28, the projection operator $P_{U_{ad}}$ is Newton differentiable only from $L^{q>2} \to L^2$. However if the term $-\frac{1}{\alpha}E_u(y,u)^*p$ is continuously Fréchet differentiable from L^2 to $L^{q>2}$, then the Newton differentiability of the map $u \to P_{U_{ad}}\Big(-\frac{1}{\alpha}E_u(y,u)^*p\Big)$ from L^2 to L^2 will follow by composition of Newton differentiable mappings. To ensure that the desired property holds for the component $-\frac{1}{\alpha}E_u(y,u)^*p$, we therefore assume

Assumption 5. *The operator* E_u^* *is a continuously Fréchet differentiable mapping from* $Y \times U \to \mathcal{L}(Y, L^{p>2})$.

Observe that the above condition is especially fulfilled by the problem of Example 2 due to the embedding of $Y = H^1(\Omega)$ into $L^p(\Omega)$ spaces. Assumption 5 allows us to prove the following.

Lemma 2.31. *Let* $Y = H^1(\Omega)$ *and let Assumption 5 holds. Then the mapping*

$$u \to P_{U_{ad}}\Big(-\frac{1}{\alpha}E_u(y,u)^*p(u)\Big)$$

is Newton differentiable from $L^2(\Omega)$ *to* $L^2(\Omega)$*.*

Proof. We intend to prove the result using Lemma 2.27. Firstly the projection $P_{U_{ad}}$ is Newton differentiable from $L^{p>2} \to L^2$ by Lemma 2.28. Next, applying Corollary 2.8 on (2.38), we have that the mapping $u \to p(u)$ is continuous as a mapping from $L^2 \to Y$. Therefore using Assumption 5 the product $E_u(y(u),u)^*p(u)$ is continuously Fréchet differentiable from $L^2 \to L^{p>2}$. The claim then follows by applying Lemma 2.27. □

We now want to apply semi-smooth Newton method to the optimality system (2.69). First

we have to establish that the system as a whole is Newton differentiable. For that purpose, let us introduce an operator

$$F(y,u,p) = u - P_{U_{ad}}\Big(-\frac{1}{\alpha}E_u(y,u)^*p\Big).$$

Then the optimality system (2.69) can be expressed as

$$\Sigma(\bar{y},\bar{u},\bar{p}) := \begin{pmatrix} \mathcal{L}_y(\bar{y},\bar{u},\bar{p}) \\ F(\bar{y},\bar{u},\bar{p}) \\ E(\bar{y},\bar{u}) \end{pmatrix} = 0 \tag{2.75}$$

where $\mathcal{L}$ is given by

$$\mathcal{L}(y,u,p) = g(y) + \frac{\alpha}{2}\|u\|_U^2 + \langle E(y,u),\ p\rangle_{Y^*,Y}. \tag{2.76}$$

Assumption 1 on page 9 ensures that E, g and consequently $\mathcal{L}$ are twice continuously Fréchet differentiable. Hence by the result of Example 3, we conclude that

Lemma 2.32. *Operator E and the Lagrange functional $\mathcal{L}$ (as well as its first derivative $\mathcal{L}_y$) defined through (2.76) are Newton differentiable.*

Furthermore

Lemma 2.33. *The system (2.75) is Newton differentiable.*

Proof. By the sum rule for Newton differentiable mappings and Lemma 2.31, it follows that $F(y,u,p)$ is Newton differentiable. Then invoking Lemma 2.32 completes the proof. □

As a result of the above lemma, the semi-smooth Newton method can be applied to solving (2.75). A semi-smooth Newton step to solve (2.75) is given by

$$G_{\Sigma'}(y_k,u_k,p_k)(y-y_k,u-u_k,p-p_k) + \Sigma(y_k,u_k,p_k) = 0, \quad k = 0,1,... \tag{2.77}$$

where $G_{\Sigma'}$ denotes the Newton derivative of Σ. The super-linear convergence of the above scheme is well known and stated in the following theorem. We set $x = (y,u,p)$ for convenience.

Theorem 2.34. *[29] Let $\Sigma : U \subseteq X \to Y$ be a nonlinear map on an open subset $U \subseteq X$. Let $\bar{x} \in U$ be a solution of $\Sigma(x) = 0$ and let Σ be Newton differentiable with Newton derivative $G_{\Sigma'}$ in an open neighborhood $\mathcal{O}$ containing $\bar{x}$ and*

$$\{\|G_{\Sigma'}(x)^{-1}\|_{\mathcal{L}(X,Y)} : x \in \mathcal{O}\} \tag{2.78}$$

is bounded. Then for any $x_0 \in U$ the Newton iteration (2.77) converges super-linearly to $\bar{x}$ provided that $\|x_0 - \bar{x}\|_X$ is sufficiently small.

In the next section, we will derive the condition(s) required for the fulfillment of the hypotheses of the above theorem.

2.4.2 SSC and convergence of semi-smooth Newton method

Note that the Newton differentiability requirement of Theorem 2.34 is already fulfilled through Lemma 2.33. It remains then to derive the Newton derivative $G_{\Sigma'}$ of Σ and prove that it satisfies (2.78) under a certain condition to be made precise shortly.

The derivative $G_{\Sigma'}$ which is defined through the gradient of Σ is obtained in the following lemma.

Theorem 2.35. *Let Assumption 1 hold. Furthermore let the results of Lemmas 2.25, 2.31 hold. Then the mapping $\Sigma : Y \times U \times Y \to Y^* \times U^* \times Y^*$ defined by (2.75) is locally Lipschitz continuous and Newton differentiable. The Newton derivative $G_{\Sigma'} := M \in \mathcal{L}(Y \times U \times Y, Y^* \times U^* \times Y^*)$ at $(y, u, p) \in Y \times U \times Y$ is of the form*

$$M(y,u,p) = \begin{pmatrix} \mathcal{L}_{yy}(y,u,p) & \mathcal{L}_{yu}(y,u,p) & E_y(y,u)^* \\ F_y(y,u,p) & F_u(y,u,p) & F_p(y,u,p) \\ E_y(y,u) & E_u(y,u) & 0 \end{pmatrix}$$

with

$$\begin{aligned} F_y(y,u,p) &= \chi_I \alpha^{-1} E_{uy}(y,u)^* p, \\ F_u(y,u,p) &= I + \chi_I \alpha^{-1} E_{uu}(y,u)^* p, \\ F_p(y,u,p) &= \chi_I \alpha^{-1} E_u(y,u)^*, \end{aligned}$$

where the set

$$I = I(u) := \{x \in \Omega : -\frac{1}{\alpha}(E_u(y(u),u)^* p(u))(x) \in (u_a(x), u_b(x))\}. \tag{2.79}$$

Proof. The proof is given in [54, Theorem 5.21]. There, the Lipschitz continuity property of $P_{U_{ad}}$ (cf. Lemma 2.25) is applied in the proof. □

It now remains to investigate under what condition does the operator matrix M above fulfill (2.78). Here we have to keep in mind the strongly active constraints introduced in the earlier sections. Including the strongly active constraints will allow us to separate the active and inactive components of the Newton system before examining the bounded invertibility of M.

The semi-smooth Newton scheme (2.77) can now be written as

$$M(y_k, u_k, p_k)s = -\Sigma(y_k, u_k, p_k) \tag{2.80}$$

where the operator matrix M is as given in Theorem 2.35 and $s = (y - y_k, u - u_k, p - p_k)^T$. For

computational importance, we shall make matrix M symmetric through elementary operations. First multiplying the second row of (2.80) by α yields an equivalent system

$$\begin{pmatrix} \mathcal{L}_{yy}(y_k,u_k,p_k) & \mathcal{L}_{yu}(y_k,u_k,p_k) & E_y(y_k,u_k)^* \\ \chi_I E_{uy}(y_k,u_k)^* p & \alpha I + \chi_I E_{uu}(y_k,u_k)^* p & \chi_I E_u(y_k,u_k)^* \\ E_y(y_k,u_k) & E_u(y_k,u_k) & 0 \end{pmatrix} \begin{pmatrix} y-y_k \\ u-u_k \\ p-p_k \end{pmatrix} = - \begin{pmatrix} (\mathcal{L}_y(y_k,u_k,p_k) \\ \alpha F(y_k,u_k) \\ E(y_k,u_k) \end{pmatrix}. \tag{2.81}$$

Let us define a set $A = \Omega \setminus I$. Then multiplying the second row of (2.81) by χ_A gives

$$\chi_A\Big(\alpha(u-u_k) + \alpha F(y_k,u_k)\Big) = 0.$$

This implies that, on the active set A, the next iterate u is known and is given by

$$\begin{aligned} u &= u_k - F(y_k,u_k) \\ &= u_k - u_k + P_{U_{ad}}\Big(-\frac{1}{\alpha} E_u(y_k,u_k)^* p_k\Big) \\ &= P_{U_{ad}}\Big(-\frac{1}{\alpha} E_u(y_k,u_k)^* p_k\Big). \end{aligned}$$

Now using the splitting

$$u - u_k = \chi_I(u-u_k) + \chi_A(u-u_k)$$

in (2.81) and subsequently multiplying the second column from the right by χ_I one obtains

$$M \begin{pmatrix} y-y_k \\ \chi_I(u-u_k) \\ p-p_k \end{pmatrix} = \rho \tag{2.82}$$

where

$$M = \begin{pmatrix} \mathcal{L}_{yy}(y_k,u_k,p_k) & \mathcal{L}_{yu}(y_k,u_k,p_k)\chi_I & E_y(y_k,u_k)^* \\ \chi_I E_{uy}(y_k,u_k)^* p & (\alpha I + \chi_I E_{uu}(y_k,u_k)^* p)\chi_I & \chi_I E_u(y_k,u_k)^* \\ E_y(y_k,u_k) & E_u(y_k,u_k)\chi_I & 0 \end{pmatrix}$$

and

$$\rho = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{pmatrix} := - \begin{pmatrix} \mathcal{L}_y(y_k,u_k,p_k) \\ \alpha F(y_k,u_k) + (\alpha I + \chi_I E_{uu}(y_k,u_k)^* p_k)\chi_A(u-u_k) \\ E(y_k,u_k) + E_u(y_k,u_k)\chi_A(u-u_k) \end{pmatrix}.$$

Note that ρ is a known quantity. Let us comment briefly on the adjoint notation for the second derivatives of operator E which is not quite obvious. The notation is to be understood in the following sense: e.g. for $v_1, v_2 \in U$

$$\Big(E_{uu}(y_k,u_k)^* p_k\Big)(v_1,v_2) := \langle E_{uu}(y_k,u_k)(v_1,v_2),\ p_k\rangle = \langle (v_1,v_2),\ E_{uu}(y_k,u_k)^* p_k\rangle_{(U\times U),(U\times U)^*}.$$

Then due to $\mathcal{L}_{yu}(y,u,p) = E_{yu}(y,u)p = E_{uy}(y,u)^* p$, the operator matrix M is now symmetric as desired.

With cost effectiveness in mind, let us derive an equivalent reduced system for (2.82). The reduced system will be obtained by applying block elimination on the augmented matrix of system (2.82). To shorten notations in the sequel, we will drop the arguments $(y_k,u_k,p_k),(y_k,u_k)$ in the derivatives of $\mathcal{L}$ and E. We will also denote the entries of M by the corresponding derivatives of $\mathcal{L}$ where convenient. For example, we assign $\mathcal{L}_{uu}(y,u,p) =: \alpha I + E_{uu}(y,u)^* p$. Thus by the definition of χ_I we derive

$$\begin{aligned}(\alpha I + \chi_I E_{uu}(y_k,u_k)^* p)\chi_I &= \alpha I\chi_I + \chi_I(\mathcal{L}_{uu} - \alpha I)\chi_I \\ &= \chi_I \mathcal{L}_{uu}\chi_I + (1-\chi_I)\alpha I\chi_I \\ &= \chi_I \mathcal{L}_{uu}\chi_I.\end{aligned}$$

More conveniently we now write

$$M = \begin{pmatrix} \mathcal{L}_{yy} & \mathcal{L}_{yu}\chi_I & E_y^* \\ \chi_I\mathcal{L}_{uy} & \chi_I\mathcal{L}_{uu}\chi_I & \chi_I E_u^* \\ E_y & E_u\chi_I & 0 \end{pmatrix}.$$

Let us now consider the augmented matrix of system (2.82) on which we intend to perform block

elimination process. Essential in the elimination process is the result of Proposition 2.7, namely that operator E_y^{-1} is bounded. Thus, the following elementary row operations are valid:

$$\left(\begin{array}{ccc|c} \mathcal{L}_{yy} & \mathcal{L}_{yu}\chi_I & E_y^* & \rho_1 \\ \chi_I\mathcal{L}_{uy} & \chi_I\mathcal{L}_{uu}\chi_I & \chi_I E_u^* & \rho_2 \\ E_y & E_u\chi_I & 0 & \rho_3 \end{array}\right)$$

$$\Updownarrow \quad (\text{row } 1 - \mathcal{L}_{yy}E_y^{-1} \times \text{row } 3)$$

$$\left(\begin{array}{ccc|c} 0 & \mathcal{L}_{yu}\chi_I - \mathcal{L}_{yy}E_y^{-1}E_u\chi_I & E_y^* & \rho_1 - \mathcal{L}_{yy}E_y^{-1}\rho_3 \\ \chi_I\mathcal{L}_{uy} & \chi_I\mathcal{L}_{uu}\chi_I & \chi_I E_u^* & \rho_2 \\ E_y & E_u\chi_I & 0 & \rho_3 \end{array}\right)$$

$$\Updownarrow \quad (\text{row } 2 - \chi_I\mathcal{L}_{uy}E_y^{-1} \times \text{row } 3)$$

$$\left(\begin{array}{ccc|c} 0 & \mathcal{L}_{yu}\chi_I - \mathcal{L}_{yy}E_y^{-1}E_u\chi_I & E_y^* & \rho_1 - \mathcal{L}_{yy}E_y^{-1}\rho_3 \\ 0 & \chi_I\mathcal{L}_{uu}\chi_I - \chi_I\mathcal{L}_{uy}E_y^{-1}E_u\chi_I & \chi_I E_u^* & \rho_2 - \chi_I\mathcal{L}_{uy}E_y^{-1}\rho_3 \\ E_y & E_u\chi_I & 0 & \rho_3 \end{array}\right)$$

$$\Updownarrow \quad (\text{row } 2 - \chi_I E_u^*(E_y^*)^{-1} \times \text{row } 1)$$

$$\left(\begin{array}{ccc|c} 0 & \mathcal{L}_{yu}\chi_I - \mathcal{L}_{yy}E_y^{-1}E_u\chi_I & E_y^* & \rho_1 - \mathcal{L}_{yy}E_y^{-1}\rho_3 \\ 0 & H & 0 & \rho_2' \\ E_y & E_u\chi_I & 0 & \rho_3 \end{array}\right)$$

where

$$\rho_2' = \rho_2 - \chi_I E_u^*(E_y^*)^{-1}\rho_1 + (\chi_I E_u^*(E_y^*)^{-1}\mathcal{L}_{yy} - \chi_I\mathcal{L}_{uy})E_y^{-1}\rho_3$$

and the reduced Hessian matrix

$$H = \chi_I\mathcal{L}_{uu}\chi_I - \chi_I\mathcal{L}_{uy}E_y^{-1}E_u\chi_I - \chi_I E_u^*(E_y^*)^{-1}\mathcal{L}_{yu}\chi_I + \chi_I E_u^*(E_y^*)^{-1}\mathcal{L}_{yy}E_y^{-1}E_u\chi_I.$$

A simple computation shows that the Hessian H can be expressed as

$$H = Z^T \begin{pmatrix} \mathcal{L}_{yy} & \mathcal{L}_{yu}\chi_I \\ \chi_I\mathcal{L}_{uy} & \chi_I\mathcal{L}_{uu}\chi_I \end{pmatrix} Z \tag{2.83}$$

where the function

$$Z = \begin{pmatrix} -E_y^{-1} E_u \chi_I \\ I \end{pmatrix}.$$

Let us set, as the unknowns in (2.82)

$$\delta_{u_I} = \chi_I(u - u_k),\ \delta_y = y - y_k,\ \delta_p = p - p_k$$

and the known quantity $\delta_{u_A} = \chi_A(u - u_k)$. What we have thus derived through the elimination process is that a semi-smooth Newton iteration step of solving the optimality system $\Sigma(y, u, p) = 0$ is equivalent to the reduced system

$$M_r \delta = \rho' \tag{2.84}$$

where

$$M_r = \begin{pmatrix} 0 & \mathcal{L}_{yu}\chi_I - \mathcal{L}_{yy}E_y^{-1}E_u\chi_I & E_y^* \\ 0 & H & 0 \\ E_y & E_u\chi_I & 0 \end{pmatrix}, \quad \delta = \begin{pmatrix} \delta_y \\ \delta_{u_I} \\ \delta_p \end{pmatrix}, \quad \rho' = \begin{pmatrix} \rho_1' \\ \rho_2' \\ \rho_3' \end{pmatrix}$$

with

$$\begin{aligned} \rho' &= \rho_1 - \mathcal{L}_{yy}E_y^{-1}\rho_3, \\ \rho_2' &= \rho_2 - \chi_I E_u^*(E_y^*)^{-1}\rho_1 + (\chi_I E_u^*(E_y^*)^{-1}\mathcal{L}_{yy} - \chi_I\mathcal{L}_{uy})E_y^{-1}\rho_3, \\ \rho_3' &= \rho_3. \end{aligned}$$

Note that in the above, $M_r = M_r(y_k, u_k, p_k)$ and $\rho' = \rho'(y_k, u_k, p_k)$.

Let us now examine the well-posedness of the system (2.84) which is the same as

$$\begin{aligned} (\mathcal{L}_{yu}\chi_I - \mathcal{L}_{yy}E_y^{-1}E_u\chi_I)\ \delta_{u_I} + E_y^*\ \delta_p &= \rho_1', \\ H\ \delta_{u_I} &= \rho_2', \\ E_y\ \delta_y + E_u\chi_I\ \delta_{u_I} &= \rho_3'. \end{aligned} \tag{2.85}$$

A close look at the structure of the above system reveals that the bounded invertibility of matrix M_r is related to that of the reduced Hessian matrix H. In fact the above system is uniquely solvable if the equation involving H is. Therefore in the sequel, our goal is to derive the condition on H which ensures unique solvability of the above system and hence, well-posedness of (2.77) in the sense of (2.78). We begin with the following observation.

Lemma 2.36. *Let $f(u) = J(S(u), u)$ where the functional J is as given in (P). For $u \in U_{ad}$ and all $v \in U$ it holds*

$$f''(u)(\chi_I v, \chi_I v) = H(y(u), u, p(u))(\chi_I v, \chi_I v).$$

Proof. If $y = S(u)$ is a solution of $E(y, u) = 0$ we obtain the representation $f(u) = J(S(u), u) = \mathcal{L}(S(u), u, p)$. The claim then follows from the following computations. For every $p \in Y$, straightforward computations yield

$$\begin{aligned} f''(u)(\chi_I v_1, \chi_I v_2) = {} & \mathcal{L}_y(y, u, p) \cdot S'(u)(\chi_I v_1, \chi_I v_2) + \mathcal{L}_{yy}(y, u, p)[S'(u)\chi_I v_1, S'(u)\chi_I v_2] \\ & + \mathcal{L}_{yu}(y, u, p)[S'(u)\chi_I v_1, \chi_I v_2] + \mathcal{L}_{uy}(y, u, p)[\chi_I v_1, S'(u)\chi_I v_2] \\ & + \mathcal{L}_{uu}(y, u, p)[\chi_I v_1, \chi_I v_2]. \end{aligned}$$

If we define the adjoint state p as the solution of $\mathcal{L}_y(S(u), u, p) = 0$ we obtain

$$\begin{aligned} f''(u)(\chi_I v_1, \chi_I v_2) = {} & \mathcal{L}_{yy}(y, u, p)[S'(u)\chi_I v_1, S'(u)\chi_I v_2] + \mathcal{L}_{yu}(y, u, p)[S'(u)\chi_I v_1, \chi_I v_2] \\ & + \mathcal{L}_{uy}(y, u, p)[\chi_I v_1, S'(u)\chi_I v_2] + \mathcal{L}_{uu}(y, u, p)[\chi_I v_1, \chi_I v_2]. \end{aligned}$$

Now recall that $S'(u)\chi_I v_i$ is the solution of the linearized system (cf. Lemma 2.15)

$$E_y(S(u), u)S'(u)\chi_I v_i + E_u(S(u), u)\chi_I v_i = 0.$$

Then by setting

$$Z(y, u, p) = \begin{pmatrix} -E_y^{-1}(y, u)E_u(y, u)\chi_I v \\ I \end{pmatrix}$$

and using (2.83), we obtain

$$\begin{aligned} f''(u)(\chi_I v, \chi_I v) &= \mathcal{L}_{yy}(y, u, p)[S'(u)\chi_I v, S'(u)\chi_I v] + \mathcal{L}_{yu}(y, u, p)[S'(u)\chi_I v, \chi_I v] \\ &\quad + \mathcal{L}_{uy}(y, u, p)[\chi_I v, S'(u)\chi_I v] + \mathcal{L}_{uu}(y, u, p)[\chi_I v, \chi_I v] \\ &= Z^T(y, u, p) \begin{pmatrix} \mathcal{L}_{yy}(y, u, p) & \mathcal{L}_{yu}(y, u, p)\chi_I \\ \chi_I \mathcal{L}_{uy}(y, u, p) & \chi_I \mathcal{L}_{uu}(y, u, p)\chi_I \end{pmatrix} Z(y, u, p) \\ &= H(y(u), u, p(u))(\chi_I v, \chi_I v). \end{aligned}$$

□

Let us now assume a second-order sufficient condition.

Assumption 6. *The second-order sufficient condition holds at the optimal solution $\bar{u} \in U_{ad}$. That is, there exits $\delta > 0$ such that*

$$f''(\bar{u})(v,v) \geq \delta \|v\|_U^2 \quad \forall v \in U.$$

Under the above assumption we obtain

Lemma 2.37. *Let the pair $(\bar{y}, \bar{u})$ together with the corresponding adjoint state $\bar{p}$ be a solution of the optimality system $\Sigma(y,u,p) = 0$. Furthermore let $(y(u_k), u_k, p(u_k))$ be a semi-smooth Newton iterate solving (2.77). If Assumption 6 holds and u_k is sufficiently close to $\bar{u}$, then the reduced system (2.85) is uniquely solvable.*

Proof. Let $v \in U$ and let $I = I(u_k)$ be defined by (2.79). We set $y_k := y(u_k), p_k := p(u_k)$. By Lemma 2.36, Assumption 6 and continuity of f'', we estimate

$$\begin{aligned} H(y_k,u_k,p_k)(\chi_I v, \chi_I v) &= f''(u_k)(\chi_I v, \chi_I v) \\ &= f''(u_k)(\chi_I v, \chi_I v) - f''(\bar{u})(\chi_I v, \chi_I v) + f''(\bar{u})(\chi_I v, \chi_I v) \\ &\geq -L_{f''}\|\bar{u} - u_k\|_U \|\chi_I v\|_U^2 + \delta \|\chi_I v\|_U^2 \\ &= \left(\delta - L_{f''}\|\bar{u} - u_k\|_U\right)\|\chi_I v\|_U^2 \\ &\geq \frac{\delta}{2}\|\chi_I v\|_U^2 \end{aligned}$$

for sufficiently small $\|\bar{u} - u_k\|_U$. The reduced Hessian matrix $H(y_k, u_k, p_k)$ is thus positive definite and hence invertible for every iterates (y_k, u_k, p_k). Consequently the matrix M_r (or equivalently $G_{\Sigma'} = M$) is as well invertible at every (y_k, u_k, p_k). Therefore the system (2.85) is uniquely solvable as claimed. □

Remark 2.38. *In place of Assumption 6 in the above proof, one can also use the following condition: There exist $\delta > 0, \tau > 0$ such that it holds*

$$f''(\bar{u})(\chi_{I_\tau} v, \chi_{I_\tau} v) \geq \delta \|\chi_{I_\tau} v\|_U^2 \quad \forall v \in C_\tau(\bar{u})$$

where the set

$$I_\tau = I_\tau(\bar{u}) := \{x \in \Omega : -\frac{1}{\alpha}(E_u(\bar{y}, \bar{u})^* \bar{p})(x) \in (u_a(x) - \tau, u_b(x) + \tau)\}$$

and $C_\tau(\bar{u})$ is given by (2.21) with $A_\tau(\bar{u}) = \Omega \setminus I_\tau(\bar{u})$. For the arguments in the proof of Lemma 2.37 to be valid in this case, one has to show that $I(u_k) \subseteq I(\bar{u})$ whenever u_k is close to $\bar{u}$. For that purpose, the continuity of $u \mapsto E_u(y(u), u)^* p(u)$ as a mapping from U to L^∞ would be required.

Lemma 2.39. *Let the assumption and result of Lemma 2.37 hold. Then the bounded invertibility condition (2.78) is fulfilled and hence the solutions $(y_k, u_k, p_k), k = 0, 1, \ldots$ of the semi-smooth Newton iteration (2.77) converge superlinearly.*

Proof. As mentioned earlier, the nonlinear map Σ is Newton differentiable by Lemma 2.33. Furthermore, by the result of Lemma 2.37, we obtained the invertibility of its Newton derivative $G_{\Sigma'} = M$ or M_r at the iterates (y_k, u_k, p_k). Also note that $G_{\Sigma'}(y_k, u_k, p_k)$ is bounded by Theorem 2.35. Hence we conclude by bounded inverse theorem [1] that $G_{\Sigma'}(y_k, u_k, p_k)$ is boundedly invertible, thereby satisfying (2.78). The superlinear convergence then follows by Theorem 2.34. □

Remark 2.40. *As a final remark for this chapter, it becomes clear from the above result that the second-order sufficient condition of Assumption 6 is the required condition for the fulfillment of (2.78) and consequently the superlinear convergence of semi-smooth Newton system applied to (2.75). It therefore confirms our earlier indication of the usefulness of second-order sufficient condition in convergence analysis of Newton-type methods.*

[1] Let X, Y be Banach spaces. Let $T : X \to Y$ be an invertible bounded operator. Then T^{-1} is also bounded. See [21, Theorem 4.12]

3 Chapter 3

A posteriori verification of optimality conditions for optimal control problems with finite dimensional control space

This chapter constitutes the major part of this thesis. A method to verify second-order sufficient optimality conditions based solely on a-posteriori information is developed. In addition to verifying second-order sufficient conditions, we obtain a computable upper bound for error in the control.

We remark that the results presented in this chapter are published in [4].

3.1 Introduction

We will consider the already introduced abstract problem (P):

$$\min \ J(y,u) = g(y) + j(u)$$

over all $(y,u) \in Y \times U$ satisfying the nonlinear elliptic partial differential equation

$$E(y,u) = 0$$

and the control constraints

$$u \in U_{ad}.$$

Here and for the rest of this chapter, the state space Y is a real Banach space and the control space $U = \mathbb{R}^n$. The set $U_{ad} \subset U$ is a non-empty, convex and closed set given by

$$U_{ad} = \{u \in U : \ u_a \leq u \leq u_b\},$$

where the inequalities are to be understood component-wise. Here, the cases $u_a = -\infty$ and $u_b = +\infty$ are allowed, such that problems with one-sided constraints or without control constraints are included in the analysis as well. We follow the notations of Chapter 1 as regards definition of spaces and variables. Furthermore we will pursue the goals set therein: Given a numerical solution u_h of a discretization of (P), we are asking, under which conditions is u_h near a local solution $\bar{u}$ of (P) and under which conditions is SSC fulfilled at $\bar{u}$.

Notational convention

Throughout this chapter we will use the following convention when naming constants: M_f and c_f will denote global bounds and Lipschitz constants for a function f; ϵ_x and r_x will denote error estimates and residuals of a quantity x. Moreover, x^h will denote auxiliary quantities that have certain similarities to a discrete quantity x_h but do not need to be explicitly known.

3.1.1 The abstract framework

We will complement problem (P) with Assumption 1 on page 9. Since $U = \mathbb{R}^n$ for the present problem, the compactness requirement of Assumption 2 on page 11 is already fulfilled. The assumptions on E are met for instance by semilinear elliptic equations with monotone nonlinearities, e.g. it is fulfilled for problem of the form

$$-\Delta y + d(u; y) = b \quad \text{in } \Omega,\ b \in Y^*$$

if $d(u;\cdot)$ is monotonically increasing for all admissible u and the assumption is fulfilled for the type

$$-\operatorname{div}(u\nabla y) = b \quad \text{in } \Omega$$

if the coefficients u are strictly positive. Moreover, the assumptions on g and j are met by the prototypical functional given in (1.1).

Since E maps to the dual of Y, we can consider the weak formulation of the state equation

$$\langle E(y,u),\, v\rangle_{Y^*,Y} = 0 \qquad \forall v \in Y.$$

The solution mapping $u \to y$ that assigns to every control u a state y is denoted by S, i.e. $y = S(u)$. Its Fréchet derivative $S'(u)v$ for $v \in U$ is the unique solution of the linearized equation (cf. Lemma 2.15)

$$E_y(S(u),u)S'(u)v + E_u(S(u),u)v = 0.$$

We define the Lagrange functional for the abstract problem

$$\mathcal{L}(u,y,p) = g(y) + j(u) - \langle E(y,u),\, p\rangle_{Y^*,Y}.$$

Let $(\bar y, \bar u)$ be locally optimal for (P). Then the first-order necessary optimality conditions can be expressed as $\mathcal{L}_y(\bar y, \bar u, \bar p) = 0$ and $\mathcal{L}_u(\bar y, \bar u, \bar p)(u - \bar u) \geq 0$ for all $u \in U_{ad}$, which is equivalent to

$$\begin{aligned} E_y(\bar u, \bar y)^* \bar p &= g'(\bar y), \\ \langle j'(\bar u) - E_u(\bar u, \bar y)^* \bar p,\, u - \bar u\rangle_{U^*,U} &\geq 0 \quad \forall u \in U_{ad}. \end{aligned}$$

Since the problem (P) is in general non-convex, the fulfillment of these necessary conditions does not imply optimality. In order to guarantee optimality, one needs additional sufficient optimality conditions. One particular (and strong) instance is given by: There exists $\alpha > 0$ such that

$$\mathcal{L}''(\bar{u}, \bar{y}, \bar{p})[(z, v)^2] \geq \alpha \|v\|_U^2 \tag{3.1}$$

holds for all $v = u - \bar{u}$, $u \in U_{ad}$, and z solves the linearized equation $E_y(\bar{u}, \bar{y})z + E_u(\bar{u}, \bar{y})v = 0$. Later we will work with a weaker sufficient condition, where the subspace on which $\mathcal{L}''$ is required to be positive is shrunk taking strongly active inequality constraints into account.

As hinted earlier in the general introduction (cf. Chapter 1), the above condition is difficult to check numerically even in the case when $(\bar{u}, \bar{y}, \bar{p})$ are given. The main difficulty here is that the function z appearing in (3.1) is given as solution of a partial differential equation, which cannot be solved explicitly. Any discretization of this equation introduces another error that has to be analyzed.

Remark 3.1. *We can relax the differentiability requirements of Assumption 1 as follows: let Y_∞ be a Banach space with a continuous embedding in Y. Then it is sufficient to require Fréchet-differentiability of E and g with respect to y in the stronger topology of Y_∞ as long as the derivatives of E and g with respect to y satisfy the Lipschitz estimates in Assumption 9 below with respect to the weaker topology of Y. This would allow for instance to choose $Y = H_0^1(\Omega)$ and $Y_\infty = L^\infty(\Omega) \cap Y$ or $Y_\infty = H^2(\Omega) \cap Y$. See also the comments after Assumption 9 and in Section 3.4.3 below.*

3.1.2 Discretization

In order to solve (P) the problem has to be discretized. Let Y_h be a finite-dimensional subspace of Y. Here and in the following, the index h denotes a discrete quantity. Then a discretization of the state equation can be obtained in the following way: A function $y_h \in Y_h$ is a solution of the discretized equation for given $u \in U_{ad}$ if and only if

$$\langle E(y_h, u),\, \phi_h \rangle_{Y^*,Y} = 0 \qquad \forall \phi_h \in Y_h.$$

Note that due to the monotonicity assumption on operator E, the discretized state equation is also uniquely solvable for every $u \in U_{ad}$. The discrete optimization problem is then given by: Minimize the functional $J(y_h, u_h)$ over all $(y_h, u_h) \in Y_h \times U_{ad}$, where y_h solves the discrete equation.

Let $(\bar{y}_h, \bar{u}_h)$ be a local solution of the discrete problem. Then it fulfills the discrete first-order necessary optimality condition, which is given as: there exists a uniquely determined discrete adjoint state $\bar{p}_h \in Y_h$ such that it holds

$$\begin{aligned} \langle E_y(\bar{y}_h, \bar{u}_h)^* \bar{p}_h,\, \phi_h \rangle_{Y^*,Y} &= \langle g'(\bar{y}_h),\, \phi_h \rangle_{Y^*,Y} \quad \forall \phi_h \in Y_h \\ \langle j'(\bar{u}_h) - E_u(\bar{y}_h, \bar{u}_h)^* \bar{p}_h, u - \bar{u}_h \rangle_{U^*,U} &\geq 0 \ \forall u \in U_{ad}. \end{aligned} \tag{3.2}$$

Throughout this work, we will assume that errors in discretizing the optimality system are controllable in the following sense. We will not make any further assumptions on the discretization, in particular, we do not assume a sufficiently fine discretization.

Assumption 7. *For a fixed finite-dimensional subspace Y_h, let (u_h, y_h, p_h) be approximations of the discrete optimal control and the corresponding state and adjoint. There are positive constants r_y, r_p, r_u such that the following holds*

$$\|E(y_h, u_h)\|_{Y^*} \leq r_y, \tag{3.3}$$

$$\|g'(y_h) - E_y(y_h, u_h)^* p_h\|_{Y^*} \leq r_p, \tag{3.4}$$

$$\langle j'(u_h) - E_u(y_h, u_h)^* p_h, u - u_h\rangle \geq -r_u \|u - u_h\|_U \ \forall u \in U_{ad}. \tag{3.5}$$

If $(\bar{y}_h, \bar{u}_h, \bar{p}_h)$ fulfills the first-order necessary optimality system (3.2) of the discrete problem then $r_u = 0$. The residuals r_y and r_p cannot be expected to vanish as they reflect the discretization error of the partial differential equation. We report on the computation of these residuals in Section 3.4.1.

As already mentioned in Chapter 1, without any further assumption, smallness of the residuals in (3.3)–(3.5) does not imply smallness of the error $\|u-u_h\|_U$ in the control. In order to establish such a bound, it is essential to check that a second-order sufficient optimality condition is satisfied.

Here it is important to recognize that sufficient optimality conditions for the *discrete* problem alone are still not enough. The sufficient optimality condition for the discrete problem is given by: There exists $\alpha_h > 0$ such that

$$\mathcal{L}''(\bar{u}_h, \bar{y}_h, \bar{p}_h)[(z_h, v)^2] \geq \alpha_h \|v\|_U^2$$

holds for all $v = u - \bar{u}$, $u \in U_{ad}$, and z_h solves the linearized *discrete* equation

$$\langle E_y(\bar{u}_h, \bar{y}_h) z_h + E_u(\bar{u}_h, \bar{y}_h) v,\ \phi_h\rangle = 0 \quad \forall \phi_h \in Y_h.$$

This condition is equivalent to positive definiteness of a certain computable matrix, see Section 3.3.5. Hence, this condition can be checked computationally, see e.g. [42, 43]. Under conditions to be worked out in the sequel, the fulfillment of this discrete SSC implies the fulfillment of the continuous SSC. These conditions are fulfilled if α_h is large compared to the residuals r_y, r_p, r_u, see Section 3.3.5 below.

If we can verify that the SSC (3.1) holds then we have a reliable error bound for the error in the optimal controls and states

$$\|\bar{y} - y_h\|_Y + \|\bar{u} - u_h\|_U \leq C\,(r_y + r_p + r_u)\,,$$

where C depends mildly on the discrete quantities, see Theorem 3.27. This allows to devise an

adaptive refinement scheme, which refines elements with relatively large local error components in r_y, r_p, cf. Chapter 4.

We will first derive such an error representation for the reduced problem in Section 3.2, where the unknown y is eliminated in terms of $y = S(u)$. These results are then applied to the problem (P), the required error and Lipschitz estimates are carried out in Section 3.3. The main result on the relation between discrete and continuous second-order conditions can be found in Section 3.3.5, Theorem 3.26. Finally, we state the a-posteriori error estimate and verification of optimality in Section 3.3.6, Theorem 3.27.

3.2 Verification of optimality for reduced functional

Let us introduce the reduced objective functional $f : U \to \mathbb{R}$ by

$$f(u) = g(S(u)) + j(u). \tag{3.6}$$

Since J and S are twice Fréchet-differentiable, the reduced functional f inherits this property as well. This allows us to write the original abstract minimization problem in the control-reduced form:

$$\min_{u \in U_{ad}} f(u). \tag{3.7}$$

For the control-reduced problem (3.7), the first-order optimality condition to be fulfilled by every optimal solution candidate $\bar{u}$ states

$$f'(\bar{u})(u - \bar{u}) \geq 0 \qquad \forall u \in U_{ad}.$$

The corresponding second-order sufficient optimality condition to be fulfilled by a locally optimal solution $\bar{u}$ is given by the existence of $\alpha > 0$ such that

$$f''(\bar{u})[v, v] \geq \alpha \|v\|_U^2 \qquad \forall v = u - \bar{u} : \ u \in U_{ad}, \ f'(\bar{u})v = 0. \tag{3.8}$$

Let us define the strongly active set A as

$$A(u) = \{k \in \{1, ..., n\} : |f'(u)_k| > 0\}$$

and the corresponding inactive set as $I = \{1, ..., n\} \setminus A$. Here the notion of active set comes from the fact that for the solution $\bar{u}$ it holds $\bar{u}_k \in \{u_{a,k}, u_{b,k}\}$ for $k \in A(\bar{u})$. That is, the inequality constraints are active for these components. Moreover, for strongly active constraints $k \in A$ the first-order condition $|f'(u)_k| > 0$ is also sufficient. That is, the following condition, which is of

first-order for active constraints and of second-order for inactive indices, is sufficient for local optimality: there exist $\alpha, \sigma > 0$ such that

$$|f'(\bar{u})_k| \geq \sigma \text{ for all } k \in A \tag{3.9}$$

and

$$f''(\bar{u})[v,v] \geq \alpha \|v\|_U^2 \qquad \forall v \in U : v = u - \bar{u},\ u \in U_{ad},\ v_k = 0\ \forall k \in A. \tag{3.10}$$

Here, (3.10) is equivalent to (3.8). One of the tasks in this chapter is to verify conditions (3.9) and (3.10) numerically for the reduced problem (3.7).

Of course, since the control space is finite-dimensional, the requirement (3.8) is equivalent to assuming $f''(\bar{u})[v,v] > 0$ for the mentioned test functions $v \neq 0$. Similarly, the requirement $|f'(\bar{u})_k| \geq \sigma$ can be replaced by $|f'(\bar{u})_k| > 0$. However, we will need later a quantification of these bounds. So we opted to present the sufficient optimality condition in this way.

Let us define the following notation that will be useful in this section. Let $A \subset \{1 \dots n\}$, $I = \{1 \dots n\} \setminus A$ be given. Then the restriction of a vector v to A is given by

$$(v|_A)_k := \begin{cases} v_k & \text{if } k \in A, \\ 0 & \text{if } k \notin A. \end{cases}$$

Additionally, we can split the norm on U as

$$\|u\|_U^2 = \sum_{k \in A} |u_k|^2 + \sum_{k \in I} |u_k|^2 = \|u|_A\|_U^2 + \|u|_I\|_U^2 =: \|u\|_A^2 + \|u\|_I^2.$$

A similar splitting of control variables is used to prove second-order optimality conditions for optimal control with infinite-dimensional control space, see e.g. [15].

Now let us suppose we computed an approximate solution u_h of the reduced problem (3.7). We want to verify that this approximation is close to a local solution of the reduced problem. In order to prove this we assume that u_h fulfills the following:

Assumption 8. *There is a subset $A \subset \{1 \dots n\}$ with $I = \{1 \dots n\} \setminus A$, and positive constants ϵ, α, σ such that the following hold*

$$f'(u_h)(u - u_h) \geq \sigma \|u - u_h\|_A - \epsilon \|u - u_h\|_I \qquad \forall u \in U_{ad}, \tag{3.11}$$

$$f''(u_h)[v,v] \geq \alpha \|v\|_U^2 \qquad \forall v \in U : v_k = 0\ \forall k \in A. \tag{3.12}$$

Additionally there is $R > 0$ and positive constants $c_{f'}$, $c_{f''}$, $M_{f''}$ depending on R such that it holds

$$\|f'(u) - f'(u_h)\|_{U^*} \leq c_{f'}\|u - u_h\|_U, \tag{3.13}$$

$$\|f''(u) - f''(u_h)\|_{(U\times U)^*} \leq c_{f''}\|u - u_h\|_U, \tag{3.14}$$

$$\|f''(u)\|_{(U\times U)^*} \leq M_{f''} \tag{3.15}$$

for all $u \in U_{ad}$ with $\|u - u_h\| \leq R$.

Some comments are in order. Inequality (3.11) means that the derivative $f'(u_h)$ has the right sign on the active set A, while $f'(u_h)$ is bounded by ϵ on the inactive set. We expect for computations that ϵ tends to zero with decreasing mesh size of the discretization, while σ should be bounded away from zero. Assumption (3.12) is exactly the second-order requirement (3.10) of the sufficient optimality condition for the reduced problem. The second part of Assumption 8 simply names the Lipschitz constants of f. An assumption similar to Assumption 8 was used in [48] without the notion of a set A, i.e. there $A = \emptyset$, $\sigma = 0$ was used.

Let us remark that Assumption 8 is fulfilled for the solution $\bar{u}$, if $\bar{u}$ satisfies the sufficient condition (3.9)–(3.10). Here it has to be noted that (3.9) implies $f'(\bar{u})(u - \bar{u}) \geq \sigma\|u - \bar{u}\|_{l^1(A)}$. Due to the finite-dimensional setting, the l^1-norm dominates the l^2-norm, which gives (3.11) with $\epsilon = 0$. In order to transfer the results to the infinite-dimensional case, in particular to $U = L^2(\Omega)$, one would have to work with two different norms of $L^1(\Omega)$ and $L^2(\Omega)$ type.

Let $u \in U_{ad}$ be an arbitrary feasible point. In the following, we want to analyze $f(u) - f(u_h)$ in terms of $\|u - u_h\|_A$ and $\|u - u_h\|_I$. To this end, let us introduce an auxiliary admissible control $\tilde{u}$ defined by

$$\tilde{u}_k = \begin{cases} u_{h,k} & k \in A, \\ u_k & k \in I. \end{cases}$$

Furthermore, we define the abbreviations

$$r_I := \|u - u_h\|_I, \quad r_A := \|u - u_h\|_A.$$

Then we have $\|u - \tilde{u}\|_U = \|u - u_h\|_A$ and $\|\tilde{u} - u_h\|_U = \|u - u_h\|_I$. With the aid of Assumption 8 we will estimate the difference $f(u) - f(u_h)$. First we make use of the new control variable $\tilde{u}$ to split the difference

$$f(u) - f(u_h) = (f(u) - f(\tilde{u})) + (f(\tilde{u}) - f(u_h)). \tag{3.16}$$

Applying Taylor expansion up to the second order on the first addend of (3.16), we have

$$\begin{aligned} f(u) - f(\tilde{u}) &= f'(\tilde{u})(u-\tilde{u}) + \int_0^1 \int_0^s f''(\tilde{u} + t(u-\tilde{u}))[(u-\tilde{u})^2] \,\mathrm{d}t\mathrm{d}s \\ &= f'(u_h)(u-\tilde{u}) + (f'(\tilde{u}) - f'(u_h))(u-\tilde{u}) \\ &\quad + \int_0^1 \int_0^s f''(\tilde{u} + t(u-\tilde{u}))[(u-\tilde{u})^2] \,\mathrm{d}t\mathrm{d}s, \end{aligned}$$

which, due to (3.11), (3.13) and (3.15), implies

$$f(u) - f(\tilde{u}) \geq \sigma r_A - c_{f'} r_A r_I - \frac{M_{f''}}{2} r_A^2.$$

Employing (3.11) and (3.12), the second addend of (3.16) is estimated as

$$\begin{aligned} f(\tilde{u}) - f(u_h) &\geq f'(u_h)(\tilde{u} - u_h) + \frac{1}{2} f''(u_h)(\tilde{u} - u_h)^2 \\ &\quad + \int_0^1 \int_0^s (f''(u_h + t(\tilde{u} - u_h)) - f''(u_h))[(u-\tilde{u})^2] \,\mathrm{d}t\mathrm{d}s \\ &\geq -\epsilon r_I + \frac{\alpha}{2} r_I^2 - \frac{c_{f''}}{6} r_I^3. \end{aligned}$$

Altogether, we arrived at

$$f(u) - f(u_h) \geq \sigma r_A - c_{f'} r_A r_I - \frac{M_{f''}}{2} r_A^2 - \epsilon r_I + \frac{\alpha}{2} r_I^2 - \frac{c_{f''}}{6} r_I^3. \tag{3.17}$$

We can now prove a first result for the reduced problem: Under assumptions on the constants in Assumption 8 we obtain the existence of a local solution $\bar{u}$ near u_h.

Theorem 3.2. *Let Assumption 8 be satisfied. If there exist $r_I, r_A > 0$ with $r_I^2 + r_A^2 < R^2$ such that*

$$\min(\sigma r_A, \frac{\alpha}{2} r_I^2) - c_{f'} r_A r_I - \frac{M_{f''}}{2} r_A^2 - \epsilon r_I - \frac{c_{f''}}{6} r_I^3 > 0 \tag{3.18}$$

then there exists a local solution $\bar{u}$ to the control-reduced problem (3.7) satisfying

$$\|\bar{u} - u_h\|_A < r_A, \quad \|\bar{u} - u_h\|_I < r_I.$$

If moreover with $r_+ := \sqrt{r_I^2 + r_A^2}$

$$\sigma - c_{f'} r_+ > 0, \quad \alpha - c_{f''} r_I > 0 \tag{3.19}$$

hold, then $\bar{u} = u_h$ on A and the second-order sufficient optimality conditions (3.9)–(3.10) are fulfilled at $\bar{u}$.

Proof. Let us define

$$B := \{u \in U : \ \|u - u_h\|_I \leq r_I, \ \|u - u_h\|_A \leq r_A\}.$$

Then B is bounded, closed, and non-empty. Hence, by the Weierstraß-Theorem, we have that the minimization problem

$$\min_{u \in U_{ad} \cap B} f(u)$$

admits a solution $\bar{u}$. Let us show that $\bar{u}$ does not lie on the boundary of B. Let us define

$$\rho_I := \|u_h - \bar{u}\|_I, \quad \rho_A := \|u_h - \bar{u}\|_A.$$

At first, we assume that $\|\bar{u} - u_h\|_A = r_A$ holds. Then according to (3.17) and (3.18) we have

$$\begin{aligned} f(\bar{u}) - f(u_h) &\geq \sigma r_A - c_{f'} r_A \rho_I - \frac{M_{f''}}{2} r_A^2 - \epsilon \rho_I + \frac{\alpha}{2} \rho_I^2 - \frac{c_{f''}}{6} \rho_I^3 \\ &\geq \sigma r_A - c_{f'} r_A r_I - \frac{M_{f''}}{2} r_A^2 - \epsilon r_I - \frac{c_{f''}}{6} r_I^3 > 0, \end{aligned}$$

which yields a contradiction, since by optimality of $\bar{u}$ we have $f(\bar{u}) - f(u_h) \leq 0$.

Second, suppose that it holds $\|\bar{u} - u_h\|_I = r_I$. Similarly as above, we get

$$f(\bar{u}) - f(u_h) \geq \frac{\alpha}{2} r_I^2 - c_{f'} r_A r_I - \frac{M_{f''}}{2} r_A^2 - \epsilon r_I - \frac{c_{f''}}{6} r_I^3 > 0,$$

which gives a contradiction as well. This proves that $\bar{u}$ lies in the interior of B, making it a local solution of the original problem. It remains to show that $\bar{u}$ satisfies SSC.

Let us take $u \in U_{ad}$ and define

$$\hat{u}_k = \begin{cases} u_k & k \in A, \\ u_{h,k} & k \in I. \end{cases}$$

Then we can estimate due to (3.13) and (3.19)

$$\begin{aligned} f'(\bar{u})(\hat{u} - \bar{u})|_A \geq f'(\bar{u})(\hat{u} - u_h)|_A &= (f'(\bar{u}) - f'(u_h))(\hat{u} - u_h)|_A + f'(u_h)(\hat{u} - u_h)|_A \\ &\geq (-c_{f'} r_+ + \sigma) \|u - u_h\|_A. \end{aligned}$$

Hence, $|f'(\bar{u})_k| > 0 \ \forall k \in A$. Moreover, by optimality of $\bar{u}$

$$\begin{aligned} 0 &\leq f'(\bar{u})(u_h - \bar{u})|_A \\ &\leq (f'(\bar{u}) - f'(u_h))(u_h - \bar{u})|_A + f'(u_h)(u_h - \bar{u})|_A \\ &\leq (c_{f'} r_+ - \sigma) \|u_h - \bar{u}\|_A \leq 0, \end{aligned}$$

which proves $\bar{u} = u_h$ on the active set A. Hence $\rho_A = 0$ and $\rho_I = r_+$ holds.

Similarly by (3.14) and (3.19), we obtain

$$f''(\bar{u})[(v,v)] \geq (\alpha - c_{f''}\rho_I)\|v\|_U^2 \geq (\alpha - c_{f''}r_I)\|v\|_U^2$$

implying the fulfillment of the positivity condition (3.9) and the coercivity condition (3.10) at the unknown solution $\bar{u}$. □

In the theorem, we proved that constraints that are active at u_h stay active at the continuous solution $\bar{u}$. It is certainly possible that inactive constraints for u_h become active at $\bar{u}$. However, this case cannot be predicted.

Under a condition slightly different from that of Theorem 3.2, a similar result can be obtained. Moreover, the conditions are more accessible than the ones from Theorem 3.2.

Theorem 3.3. *Let Assumption 8 be satisfied. Let us suppose that there exist $r_I, r_A > 0$ such that with $r_+ := \sqrt{r_I^2 + r_A^2} < R$ it hold*

$$-\epsilon + \alpha r_I - \frac{c_{f''}}{2} r_I^2 > 0, \tag{3.20}$$

$$\sigma - c_{f'} r_+ > 0. \tag{3.21}$$

Then there exists a local optimal control $\bar{u}$ to the original problem (3.7) with

$$\|\bar{u} - u_h\|_U < r_+.$$

Moreover $\bar{u} = u_h$ on A and the second-order sufficient optimality conditions (3.9)–(3.10) hold at $\bar{u}$.

Proof. Let us define $B := \{u \in U : \ \|u - u_h\|_I \leq r_I, \ \|u - u_h\|_A \leq r_A\}$. As argued in the proof of Theorem 3.2, we have that the minimization problem $\min_{u \in U_{ad} \cap B} f(u)$ admits a solution $\bar{u}$. Again, it remains to prove that $\bar{u}$ is a local solution of the original problem. To this end, we will show that $\bar{u}$ lies not on the boundary of the ball B. We write

$$f'(\bar{u})(u_h - \bar{u}) = f'(\bar{u})(u_h - \bar{u})|_I + f'(\bar{u})(u_h - \bar{u})|_A,$$

and we will estimate the derivative of f on the active and inactive set separately. Let us define

$$\rho_+ := \|u_h - \bar{u}\|, \quad \rho_I := \|u_h - \bar{u}\|_I, \quad \rho_A := \|u_h - \bar{u}\|_A,$$

which implies $\rho_+ \leq r_+, \rho_I \leq r_I$, and $\rho_A \leq r_A$. For the active set, we obtain

$$\begin{aligned} f'(\bar{u})(u_h - \bar{u})|_A &= f'(u_h)(u_h - \bar{u})|_A + (f'(\bar{u}) - f'(u_h))(u_h - \bar{u})|_A \\ &\leq \rho_A(-\sigma + c_{f'}\rho_+). \end{aligned}$$

The contribution on the inactive set can be estimated as

$$\begin{aligned} f'(\bar{u})(u_h-\bar{u})|_I &= f'(u_h)(u_h-\bar{u})|_I + f''(u_h)[(u_h-\bar{u})|_I,\, \bar{u}-u_h] \\ &\quad + \int_0^1 (f''(u_h+s(\bar{u}-u_h)) - f''(u_h))[(u_h-\bar{u})|_I,\, \bar{u}-u_h]ds \\ &\leq \epsilon\rho_I - \alpha\rho_I^2 + M_{f''}\rho_I\rho_A + \frac{1}{2}c_{f''}\rho_I\rho_+^2 \\ &\leq \rho_I(\epsilon - \alpha\rho_I + M_{f''}\rho_A + \frac{1}{2}c_{f''}\rho_+^2). \end{aligned}$$

Furthermore, by the necessary optimality conditions, we have

$$f'(\bar{u})(u_h-\bar{u}) \geq 0.$$

First, let us assume that the constraint $\|u-u_h\|_A \leq r_A$ is active at $\bar{u}$, i.e. $\|\bar{u}-u_h\|_A = r_A > 0$. Then we obtain

$$0 \leq f'(\bar{u})(u_h-\bar{u})|_A \leq \rho_A(-\sigma + c_{f'}\rho_+) \leq r_A(-\sigma + c_{f'}r_+) < 0$$

by assumption (3.21), which is a contradiction. Moreover, this estimate implies $\rho_A = 0$, which means that $\bar{u} = u_h$ on the active set. Hence $\rho_+ = \rho_I$.

Second, let us assume that the the constraint $\|u-u_h\|_I \leq r_I$ is active at $\bar{u}$, i.e. $\|\bar{u}-u_h\|_I = r_I > 0$. Then it holds

$$0 \leq f'(\bar{u})(u_h-\bar{u})|_I \leq \rho_I(\epsilon - \alpha r_I + \frac{1}{2}c_{f''}r_I^2) < 0$$

by (3.20), which proves $\|\bar{u}-u_h\|_I < \rho_I$ by a similar reasoning as on the active set.

This implies that $\bar{u}$ is an interior point of B, and hence a local solution of the original problem. The inequality (3.20) implies the convexity condition (3.9), which can be proven as in [48, Theorem 2.5]. □

Let us now prove a much more explicit error bound.

Corollary 3.4. *Let the assumptions of Theorem 3.3 be satisfied. Then it holds*

$$\|\bar{u}-u_h\|_U \leq \frac{2\epsilon}{\alpha}.$$

Moreover,

$$f''(\bar{u})[v,v] \geq (\alpha - \|\bar{u}-u_h\|_U c_{f''})\,\|v\|_U^2 > 0$$

for all $v \in U$ with $v_k = 0,\ k \in A$.

Here, one can see clearly that our results imply $\|\bar{u}-u_h\|_U \to 0$ provided that $\epsilon \to 0$ and α is bounded away from zero, which can be expected if SSC holds at $\bar{u}$.

Proof. By assumption, the polynomial $-\epsilon + \alpha r_I - \frac{c_{f''}}{2} r_I^2$ in (3.20) has a positive root. Since the polynomial is negative at $r_I = 0$, this implies that all roots are positive. The smallest root can be computed as

$$\tilde{r}_I = \frac{\alpha - \sqrt{\alpha^2 - 2\epsilon c_{f''}}}{c_{f''}},$$

which implies $\alpha^2 - 2\epsilon c_{f''} > 0$. If $\alpha^2 - 2\epsilon c_{f''} < 0$ then the polynomial would be strictly negative, which is a contradiction to the assumption (3.20). By elementary calculations, we find

$$\tilde{r}_I = \frac{\alpha - \sqrt{\alpha^2 - 2\epsilon c_{f''}}}{c_{f''}} \frac{\alpha + \sqrt{\alpha^2 - 2\epsilon c_{f''}}}{\alpha + \sqrt{\alpha^2 - 2\epsilon c_{f''}}} = \frac{2\epsilon}{\alpha + \sqrt{\alpha^2 - 2\epsilon c_{f''}}} \leq \frac{2\epsilon}{\alpha}.$$

The claim on the second derivative follows immediately. □

Let us close this section with the following observation, which gives a sufficient condition for the assumptions of Theorem 3.3. Moreover, these conditions are easier to check and independent of $M_{f''}$. In addition, they highlight the fact that if $\alpha, \sigma, c_{f'}, c_{f''}$ stay bounded, the assumption of Theorem 3.3 is satisfied if the discretization error ϵ goes to zero, which is guaranteed at least for uniform mesh refinement.

Corollary 3.5. *Let Assumption 8 be satisfied. If*

$$\alpha^2 - 2c_{f''}\epsilon > 0, \tag{3.22}$$

$$\alpha\sigma - 2c_{f'}\epsilon > 0 \tag{3.23}$$

hold then the assumptions of Theorem 3.3 are satisfied.

Proof. As argued in the proof of the previous Corollary 3.4, condition (3.22) is sufficient for (3.20). Moreover, there the bound $r_I \leq 2\epsilon/\alpha$ was proven. Then we obtain

$$\sigma - c_{f'} r_+ \geq \sigma - c_{f'}(\frac{2\epsilon}{\alpha} + r_A) = \alpha^{-1}(\alpha\sigma - 2c_{f'}\epsilon) - c_{f'} r_A,$$

which shows that due to (3.23) we can choose $r_A > 0$ to satisfy (3.21). □

3.3 Application to the abstract problem

Here, we will transform Assumption 8 in the previous section to assumptions on the solution (y_h, u_h, p_h) of the abstract problem (P). First we derive the Fréchet derivatives of the reduced functional f involving the abstract PDE operator $E(y, u)$.

Let us recall the definition of the reduced functional (3.6): $f(u) = g(S(u)) + j(u)$. Then the first derivative of the reduced functional is obtained as

$$f'(u)v = g'(S(u))z + j'(u)v \qquad \forall v \in U, \tag{3.24}$$

where $z = S'(u)v$. With p being the solution of

$$E_y(S(u), u)^* p = g'(S(u)), \tag{3.25}$$

an obvious computation gives the equality $g'(S(u))z = -\langle E_u(S(u), u)p,\, v\rangle_{U^*,U}$, which we use in (3.24) to obtain

$$f'(u)v = \langle -E_u(S(u), u)^* p + j'(u), v\rangle_{U^*,U}. \tag{3.26}$$

Similarly for $v_1, v_2 \in U$, taking the derivative of (3.26) yields

$$f''(u)[v_1,\, v_2] = g''(S(u))[S'(u)v_1,\, S'(u)v_2] + g'(S(u))S''(u)[v_1,\, v_2] + j''(u)[v_1,\, v_2]. \tag{3.27}$$

Since $S'(u)v = z$ solves

$$E_y(S(u), u)z + E_u(S(u), u)v = 0$$

it follows that $S''(u)[v_1, v_2] = \zeta$ is a solution of

$$E_y(S(u), u)\zeta + E''(S(u), u)[(v_1, S'(u)v_1),\, (v_2, S'(u)v_2)] = 0. \tag{3.28}$$

Again, by testing (3.25) by ζ, (3.28) by p, and comparing the resulting equations, one obtains

$$g'(S(u))\zeta = -\langle E''(S(u), u)[(v_1, S'(u)v_1),\, (v_2, S'(u)v_2)],\, p\rangle_{Y^*,Y}. \tag{3.29}$$

Now using (3.29) in (3.27) yields

$$f''(u)[v_1,\, v_2] = g''(S(u))[z_1,\, z_2] + j''(u)[v_1,\, v_2] - \langle E''(S(u), u)[(v_1, z_1),\, (v_2, z_2)],\, p\rangle_{Y^*,Y}$$

with $z_i = S'(u)v_i$, $i = 1, 2$.

By Assumption 1, the functions E, j, g are twice Fréchet-differentiable with Lipschitz continuous second derivatives. In the sequel, we will need the associated Lipschitz constants. In order to get a compact notation, we will introduce a short-hand notation of bounds of bilinear forms. If $G : X \times X \to Z$ is a bounded bilinear form, then

$$\|G\|_{\mathcal{B}(X,Z)} := \|G\|_{\mathcal{L}(X,\mathcal{L}(X,Z))} = \sup_{\|x_1\|_X = \|x_2\|_X = 1} \|G(x_1, x_2)\|_Z$$

is the associated bound of the bilinear form.

Let us fix the Lipschitz constants and bounds of derivatives with the following assumption.

Assumption 9. *Let R be a positive constant. We assume there are positive constants c_E, c_{E_y}, c_{E_u}, $c_{E''}$, $c_{g'}$, $c_{g''}$, $c_{j'}$, $c_{j''}$ depending on R such that the estimates*

$$\|E(y,u) - E(y^h,u_h)\|_{Y^*} \leq c_E(\|y^h - y\|_Y + \|u - u_h\|_U), \tag{3.30a}$$
$$\|E_y(y,u) - E_y(y^h,u_h)\|_{\mathcal{L}(Y,Y^*)} \leq c_{E_y}(\|y^h - y\|_Y + \|u - u_h\|_U), \tag{3.30b}$$
$$\|E_u(y,u) - E_u(y^h,u_h)\|_{\mathcal{L}(U,Y^*)} \leq c_{E_u}(\|y^h - y\|_Y + \|u - u_h\|_U), \tag{3.30c}$$
$$\|E''(y,u) - E''(y^h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)} \leq c_{E''}(\|y^h - y\|_Y + \|u - u_h\|_U), \tag{3.30d}$$
$$\|g'(y) - g'(y^h)\|_{Y^*} \leq c_{g'}\|y - y^h\|_Y, \tag{3.30e}$$
$$\|g''(y) - g''(y^h)\|_{(Y\times Y)^*} \leq c_{g''}\|y - y^h\|_Y, \tag{3.30f}$$
$$\|j'(u) - j'(u_h)\|_{(U\times U)^*} \leq c_{j'}\|u - u_h\|_U, \tag{3.30g}$$
$$\|j''(u) - j''(u_h)\|_{(U\times U)^*} \leq c_{j''}\|u - u_h\|_U \tag{3.30h}$$

hold for all $u \in U_{ad}$, $\|u - u_h\|_U \leq R$, $y = S(u)$, and $y^h = S(u_h)$ or $y^h = y_h$.

As already indicated in Remark 3.1, we can relax the differentiability requirements for E and g, i.e. it is sufficient to have E and g to be Fréchet-differentiable with respect to a stronger space $Y_\infty \hookrightarrow Y$. Here, we have in mind to take $Y = H_0^1(\Omega)$ and $Y_\infty = H_0^1(\Omega) \cap L^\infty(\Omega)$. However, we still need the Lipschitz continuity w.r.t. y in the (weaker) space Y. Otherwise it would be necessary to have computable errors of y and p in Y_∞, which seems to be impossible for e.g. $Y_\infty = H_0^1(\Omega) \cap L^\infty(\Omega)$.

Let us recall the statement of Assumption 7:

$$\|E(y_h,u_h)\|_{Y^*} \leq r_y,$$
$$\|g'(y_h) - E_y(y_h,u_h)^* p_h\|_{Y^*} \leq r_p,$$
$$\langle j'(u_h) - E_u(y_h,u_h)^* p_h, u - u_h\rangle \geq -r_u\|u - u_h\|_U \ \forall u \in U_{ad}.$$

In the remainder of this section, we will express the constants in Assumption 8 by means of the residuals of Assumption 7 and the constants of Assumption 9.

3.3.1 Error estimates for state and adjoint equation, estimates for auxiliary functions

Let y^h and p^h be auxiliary variables that solve

$$E(y^h,u_h) = 0, \tag{3.31}$$
$$E_y(y^h,u_h)^* p^h = g'(y^h), \tag{3.32}$$

respectively. The following estimates hold for the introduced state and adjoint variables.

Lemma 3.6. *Let y^h, p^h be given by (3.31) and (3.32) respectively. Then it holds*

$$\|y^h - y_h\|_Y \leq \epsilon_y, \tag{3.33}$$
$$\|p^h - p_h\|_Y \leq \epsilon_p \tag{3.34}$$

with $\epsilon_y = \delta^{-1} r_y$ and $\epsilon_p = \delta^{-1}\left(c_{g'}\epsilon_y + r_p + c_{E_y}\epsilon_y \|p_h\|_Y\right)$.

Proof. Since y^h is a solution of the nonlinear equation, we have $E(y^h, u_h) - E(y_h, u_h) = -E(y_h, u_h)$. Testing this equation with $y^h - y_h$ and using the strong monotonicity of E we obtain

$$\delta \|y^h - y_h\|_Y^2 \leq \|E(y_h, u_h)\|_{Y^*} \, \|y^h - y_h\|_Y.$$

The result then follows by using the residual estimate (3.3).

For estimating the adjoint state, observe that $p^h - p_h$ fulfills

$$E_y(y^h, u_h)^*(p^h - p_h) = g'(y^h) - g'(y_h) + (g'(y_h) - E_y(y_h, u_h)^* p_h) + \left(E_y(y_h, u_h)^* - E_y(y^h, u_h)^*\right) p_h.$$

Hence by Lipschitz properties (3.30e) and (3.30b) of g' and E_y, respectively, and the residual estimates (3.4) and (3.33), we obtain

$$\begin{aligned}
\|p^h - p_h\|_Y &\leq \delta^{-1}\left(\|g'(y^h) - g'(y_h)\|_{Y^*} + \|g'(y_h) - E_y(y_h, u_h)^* p_h\|_{Y^*}\right) \\
&\quad + \delta^{-1}\|E_y(y_h, u_h) - E_y(y^h, u_h)\|_{\mathcal{L}(Y, Y^*)}\|p_h\|_Y \\
&\leq \delta^{-1}\left(c_{g'}\|y_h - y^h\|_Y + r_p + c_{E_y}\|y^h - y_h\|_Y \|p_h\|_Y\right) \\
&\leq \delta^{-1}\left(c_{g'}\epsilon_y + r_p + c_{E_y}\epsilon_y\|p_h\|_Y\right).
\end{aligned}$$

□

As one can see in the estimates above, it hold $\|y^h - y_h\|_Y \to 0$ and $\|p^h - p_h\|_Y \to 0$ if $r_y, r_p \to 0$. In addition to these results above, we derive bounds for the norms of y^h and p^h, which will turn out useful in the sequel.

Corollary 3.7. *Let y^h, p^h be as defined in (3.31) (3.32) respectively. Then it holds*

$$\|y^h\|_Y \leq M_y, \tag{3.35}$$
$$\|p^h\|_Y \leq M_p \tag{3.36}$$

with $M_y = \epsilon_y + \|y_h\|_Y$ and $M_p = \epsilon_p + \|p_h\|_Y$.

Proof. The claim is an easy consequence of Lemma 3.6 and the triangle inequality. □

3.3.2 Lipschitz estimate of f', computation of $c_{f'}$

Let $u \in U_{ad}$ be given with $\|u - u_h\|_U \leq R$. We define the associated state y and adjoint state p through

$$E(y, u) = 0, \tag{3.37}$$
$$E_y(y, u)^* p = g'(y). \tag{3.38}$$

In order to obtain the Lipschitz estimates for f' and f'' we have to estimate the differences $y - y^h$ and $p - p^h$.

Lemma 3.8. *Let $u \in U_{ad}$ be given with $\|u - u_h\|_U \leq R$, where R is as in Assumption 9. Let y, p be the associated state and adjoint state solving (3.37) and (3.38) respectively. Then it holds*

$$\|y - y^h\|_Y \leq c_y \|u - u_h\|_U, \tag{3.39}$$
$$\|p - p^h\|_Y \leq c_p \|u - u_h\|_U \tag{3.40}$$

with

$$c_y = \delta^{-1} c_E,$$
$$c_p = \delta^{-1} \left(c_{g'} c_y + M_p c_{E_y} (c_y + 1) \right).$$

Proof. The functions y and y^h are the solutions of $E(y, u) = 0$ and $E(y^h, u_h) = 0$ respectively. Therefore we can write

$$\langle E(y, u) - E(y^h, u), y - y^h \rangle_{Y^*, Y} = \langle E(y^h, u_h) - E(y^h, u), y - y^h \rangle_{Y^*, Y}.$$

By the monotonicity assumption on E we obtain using (3.30a)

$$\begin{aligned} \delta \|y - y^h\|_Y^2 &\leq \|E(y^h, u_h) - E(y^h, u)\|_{Y^*} \|y - y^h\|_Y \\ &\leq c_E \|u - u_h\|_U \|y - y^h\|_Y, \end{aligned}$$

which gives the first estimate.

For the second estimate, recall that p and p^h are the solutions to the equations $E_y(y, u)^* p = g'(y)$ and $E_y(y^h, u_h)^* p^h = g'(y^h)$, respectively. Hence, the difference $p - p^h$ fulfills

$$E_y(y, u)^* (p - p^h) = g'(y) - g'(y^h) + \left(E_y(y^h, u_h)^* - E_y(y, u)^* \right) p^h.$$

Using the Lipschitz estimates (3.30e),(3.30b) together with (3.36) and (3.39) we obtain the a-priori estimate

$$\begin{aligned}\|p-p^h\|_Y &\le \delta^{-1}\left(\|g'(y)-g'(y^h)\|_{Y^*}+\|E_y(y^h,u_h)-E_y(y^h,u)\|_{\mathcal{L}(Y,Y^*)}\|p^h\|_Y\right)\\ &\le \delta^{-1}\left(c_{g'}\|y-y^h\|_Y+c_{E_y}\left(\|y-y^h\|_Y+\|u-u_h\|_U\right)M_p\right)\\ &\le \delta^{-1}\left(c_{g'}c_y\|u-u_h\|_U+c_{E_y}(c_y+1)\|u-u_h\|_U M_p\right).\end{aligned}$$

□

Lemma 3.9. *Let $u\in U_{ad}$ be given with $\|u-u_h\|_U\le R$, where R is as in Assumption 9. Then the first derivative f' of the reduced functional satisfies the Lipschitz estimate*

$$\|f'(u)-f'(u_h)\|_{U^*}\le c_{f'}\|u-u_h\|_U$$

with

$$c_{f'}=c_{E_u}(c_y+1)(Rc_p+M_p)+c_p\left(\epsilon_y c_{E_u}+\|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}\right)+c_{j'}.$$

Proof. Let p^h be as in (3.32). Let us set $v:=u-u_h$. By previous computation

$$f'(u_h)v=\langle -E_u(y^h,u_h)^*p^h+j'(u_h),v\rangle_{U^*,U}.$$

Using the Lipschitz estimate (3.30g) we have

$$\begin{aligned}\|f'(u)-f'(u_h)\|_{U^*} &\le \|E_u(y^h,u_h)^*p^h-E_u(y,u)^*p+j'(u)-j'(u_h)\|_{U^*}\\ &\le \|E_u(y^h,u_h)^*p^h-E_u(y,u)^*p\|_{U^*}\|v\|_U+\|j'(u)-j'(u_h)\|_{U^*}\\ &\le \|E_u(y^h,u_h)^*p^h-E_u(y,u)^*p\|_{U^*}\|v\|_U+c_{j'}\|u-u_h\|_U.\end{aligned}\tag{3.41}$$

By adopting the splitting

$$\begin{aligned}E_u(y^h,u_h)^*p^h-E_u(y,u)^*p &= E_u(y,u)^*(p^h-p)+\left(E_u(y^h,u_h)^*-E_u(y,u)^*\right)p^h\\ &= \left(E_u(y,u)-E_u(y^h,u_h)+E_u(y^h,u_h)-E_u(y_h,u_h)\right)^*(p^h-p)\\ &\quad +E_u(y_h,u_h)^*(p^h-p)+\left(E_u(y^h,u_h)^*-E_u(y,u)^*\right)p^h,\end{aligned}$$

we estimate

$$\begin{aligned}\|E_u(y^h,u_h)^*p^h-E_u(y,u)^*p\|_{U^*} &\le \|E_u(y,u)-E_u(y^h,u_h)\|_{\mathcal{L}(U,Y^*)}\|p^h-p\|_Y\\ &\quad+\left(\|E_u(y^h,u_h)-E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}+\|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}\right)\|p^h-p\|_Y\\ &\quad+\|E_u(y^h,u_h)^*-E_u(y,u)^*\|_{\mathcal{L}(U,Y^*)}\|p^h\|_Y.\end{aligned}$$

Thanks to property (3.30c) of E_u, (3.39) and the residual estimate (3.33), it hold

$$\|E_u(y,u) - E_u(y^h,u_h)\|_{\mathcal{L}(U,Y^*)} \leq c_{E_u}(\|y-y^h\|_Y + \|u-u_h\|_U) \leq c_{E_u}(c_y+1)R \tag{3.42}$$

and

$$\|E_u(y^h,u_h) - E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)} \leq c_{E_u}\|y^h-y_h\|_Y \leq c_{E_u}\epsilon_y. \tag{3.43}$$

Hence by (3.40) and (3.36)

$$\begin{aligned}\|E_u(y^h,u_h)^*p^h - E_u(y,u)^*p\|_{U^*} \leq \Big(c_{E_u}(c_y+1)R + c_{E_u}\epsilon_y + \|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}\Big)\, c_p\|u-u_h\|_U \\ + c_{E_u}(c_y+1)M_p\|u-u_h\|_U.\end{aligned} \tag{3.44}$$

Finally substituting (3.44) in (3.41) yields the desired estimate. □

3.3.3 Estimates for $f'(u_h)$, computation of ϵ and σ

After having computed the Lipschitz constant of f', we will now derive bounds for the constants ϵ and σ appearing in the first-order part (3.11) of Assumption 8.

At first, we will estimate the difference between $f'(u_h)$ and the gradient of the discrete problem defined by

$$f_h'(u_h) := -E_u(y_h,u_h)^*p_h + j'(u_h). \tag{3.45}$$

Here, we have the following.

Lemma 3.10. *Let $f_h'(u_h)$ be defined by (3.45). Then it holds*

$$\|f'(u_h) - f_h'(u_h)\|_{U^*} \leq \epsilon_{f'} \tag{3.46}$$

with

$$\epsilon_{f'} := c_{E_u}\epsilon_y M_p + \epsilon_p\|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}.$$

Proof. Let $v \in U$. We estimate the difference

$$\begin{aligned}\|f'(u_h) - f_h'(u_h)\|_{U^*} &= \|E_u(y_h,u_h)^*p_h - E_u(y^h,u_h)^*p^h\|_{U^*} \\ &\leq \|E_u(y_h,u_h) - E_u(y^h,u_h)\|_{\mathcal{L}(U,Y^*)}\|p^h\|_Y + \|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}\|p_h-p^h\|_Y.\end{aligned}$$

Using (3.43), and the bounds of p^h and $p_h - p^h$ in (3.36) and (3.34), respectively, we have the estimate

$$\|E_u(y_h,u_h)^*p_h - E_u(y^h,u_h)^*p^h\|_{U^*} \leq c_{E_u}\epsilon_y M_p + \epsilon_p\|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}, \tag{3.47}$$

which proves (3.46). □

The estimate for the positivity constant σ can now be computed easily thanks to the result of the previous Lemma 3.10.

Lemma 3.11. *For all admissible controls $u \in U_{ad}$ with $\|u - u_h\|_U < R$, where R is as in Assumption 9, the following inequality holds on the active set of u_h*

$$f'(u_h)(u - u_h)|_A \geq \sigma \|u - u_h\|_A$$

where

$$\begin{aligned} \sigma &:= \sigma_h - \epsilon_{f'}, \\ \sigma_h &:= \min_{k \in A} |f_h'(u_h)_k|. \end{aligned}$$

Proof. We write

$$f'(u_h)(u-u_h)|_A = (f'(u_h)-f_h'(u_h))(u-u_h)|_A + f_h'(u_h)(u-u_h)|_A \geq (\sigma_h - \|f'(u_h)-f_h'(u_h)\|)\|u-u_h\|_A,$$

which yields the result upon applying the estimate (3.46) provided by Lemma 3.10. □

Furthermore, we have the following estimate for the first derivative of f on the inactive set.

Lemma 3.12. *For all admissible controls $u \in U_{ad}$ with $\|u - u_h\|_U < R$, where R is as in Assumption 9, the following inequality holds on the inactive set of u_h*

$$f'(u_h)(u - u_h)|_I \geq -\epsilon \|u - u_h\|_I$$

where

$$\epsilon := c_{E_u} \epsilon_y M_p + \epsilon_p \|E_u(y_h, u_h)\|_{\mathcal{L}(U, Y^*)} + r_u.$$

Proof. Applying the residual estimate (3.5) it holds

$$\begin{aligned} f'(u_h)(u - u_h)|_I &= \langle -E_u(y^h, u_h)^* p^h + j'(u_h), u - u_h \rangle_{U^*,U} \\ &= \langle E_u(y_h, u_h)^* p_h - E_u(y^h, u_h)^* p^h, u - u_h \rangle_{U^*,U} \\ &\quad + \langle j'(u_h) - E_u(y_h, u_h)^* p_h, u - u_h \rangle_{U^*,U} \\ &\geq -\|E_u(y_h, u_h)^* p_h - E_u(y^h, u_h)^* p^h\|_{U^*} \|u - u_h\|_U - r_u \|u - u_h\|_U. \end{aligned}$$

The estimate for the leading term is given by (3.47), which gives

$$f'(u_h)(u - u_h)|_I \geq -\left(c_{E_u} \epsilon_y M_p + \epsilon_p \|E_u(y_h, u_h)\|_{\mathcal{L}(U, Y^*)} + r_u \right) \|u - u_h\|_U.$$

□

3.3.4 Estimates for f'', computation of $c_{f''}$ and $M_{f''}$

Let us now turn to the relevant estimate for the second derivative f'', which was derived above as

$$f''(u)[v_1, v_2] = g''(S(u))[z_1, z_2] + j''(u)[v_1, v_2] - \langle E''(S(u), u)[(v_1, z_1), (v_2, z_2)], p\rangle_{Y^*,Y}$$

with p solving (3.38). Obviously, any change in u will not only change the argument of g'', j'', and E'', but also it will change the point, where the linearization of the solution operator S is made. This necessitates an analysis of $S'(u) - S'(u_h)$ in order to be able to estimate $f''(u) - f''(u_h)$.

Let $v_i \in U$, $i = 1, 2$ be given. In the sequel, unless otherwise stated, z_i, z_i^h are defined as the solutions of

$$E_y(y, u)z_i + E_u(y, u)v_i = 0, \tag{3.48}$$
$$E_y(y^h, u_h)z_i^h + E_u(y^h, u_h)v_i = 0, \tag{3.49}$$

respectively. That is, we have $z_i = S'(u)v_i$ and $z_i^h = S'(u_h)v_i$. To be able to find the Lipschitz estimate for the second derivative f'', we derive the following useful estimates.

Lemma 3.13. *Let $v_i \in U$ be given. Let z_i, z_i^h be defined by (3.48) and (3.49) respectively. Then for $u \in U_{ad}$, $\|u - u_h\|_U \leq R$, where R is as in Assumption 9, it holds*

$$\|z_i\|_Y \leq M_z\|v_i\|_U, \tag{3.50}$$
$$\|z_i^h\|_Y \leq M_{z^h}\|v_i\|_U, \tag{3.51}$$
$$\|z_i - z_i^h\|_Y \leq c_z\|u - u_h\|_U\|v_i\|_U. \tag{3.52}$$

The bounds are derived in the course of the proof.

Proof. First, testing (3.49) by z_i^h one finds the a-priori estimate

$$\begin{aligned}\|z_i^h\|_Y &\leq \delta^{-1}\|E_u(y^h, u_h)\|_{\mathcal{L}(U,Y^*)}\|v_i\|_U \\ &\leq \delta^{-1}\left(\|E_u(y^h, u_h) - E_u(y_h, u_h)\|_{\mathcal{L}(U,Y^*)} + \|E_u(y_h, u_h)\|_{\mathcal{L}(U,Y^*)}\right)\|v_i\|_U.\end{aligned}$$

Due to (3.43), we obtain

$$\|z_i^h\|_Y \leq \delta^{-1}\left(c_{E_u}\epsilon_y + \|E_u(y_h, u_h)\|_{\mathcal{L}(U,Y^*)}\right)\|v_i\|_U,$$

which implies $M_{z^h} = \delta^{-1}\left(c_{E_u}\epsilon_y + \|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}\right)$. Similarly by using the already obtained estimates (3.42) and (3.43), we obtain from (3.48)

$$\begin{aligned}\|z_i\|_Y &\le \delta^{-1}\|E_u(y,u)\|_{\mathcal{L}(U,Y^*)}\|v_i\|_U\\ &\le \delta^{-1}\Big(\|E_u(y,u)-E_u(y^h,u_h)\|_{\mathcal{L}(U,Y^*)} + \|E_u(y^h,u_h)-E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}\\ &\qquad +\|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}\Big)\|v_i\|_U\\ &\le \delta^{-1}\left(c_{E_u}R(c_y+1)+c_{E_u}\epsilon_y+\|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}\right)\|v_i\|_U,\end{aligned}$$

which gives $M_z = \delta^{-1}\left(c_{E_u}R(c_y+1)+c_{E_u}\epsilon_y+\|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)}\right)$. To estimate $z_i - z_i^h$, observe that $z_i - z_i^h$ fulfills

$$E_y(y,u)(z_i-z_i^h) = \Big(E_u(y^h,u_h)-E_u(y,u)\Big)v_i + \Big(E_y(y^h,u_h)-E_y(y,u)\Big)z_i^h.$$

This implies the a-priori estimate

$$\|z_i-z_i^h\|_Y \le \delta^{-1}\|E_u(y^h,u_h)-E_u(y,u)\|_{\mathcal{L}(U,Y^*)}\|v_i\|_U + \|E_y(y^h,u_h)-E_y(y,u)\|_{\mathcal{L}(Y,Y^*)}\|z_i^h\|_Y.$$

Employing (3.30c) and (3.30b) in estimating the first and second addend respectively, and using the estimates (3.39), (3.51) gives

$$\begin{aligned}\|z_i-z_i^h\|_Y &\le \delta^{-1}c_{E_u}\Big(\|y^h-y\|_Y+\|u_h-u\|_U\Big)\|v_i\|_U + c_{E_y}\Big(\|y^h-y\|_Y+\|u_h-u\|_U\Big)M_{z^h}\|v_i\|_U\\ &\le \delta^{-1}c_{E_u}(c_y+1)\|u_h-u\|_U\|v_i\|_U + c_{E_y}(c_y+1)M_{z^h}\|u_h-u\|_U\|v_i\|_U,\end{aligned}$$

which yields the last estimate (3.52) with $c_z = \delta^{-1}(c_y+1)(c_{E_u}+c_{E_y}M_{z^h})$. □

Now we are ready to do the first step in estimating $f''(u) - f''(u_h)$.

Lemma 3.14. *Let $u \in U_{ad}$, $\|u-u_h\|_U \le R$, where R is as in Assumption 9, be given. Let $v_i \in U$, $i=1,2$, be given. Let z_i, z_i^h, $i=1,2$ be defined as in the previous lemma. Then it holds*

$$|g''(y^h)[z_1^h,z_2^h] - g''(y)[z_1,z_2]| \le C_{g''}\|u-u_h\|_U\|v_1\|_U\|v_2\|_U$$

with

$$C_{g''} = c_{g''}c_yM_z^2 + c_z\left(c_{g''}\epsilon_y + \|g''(y_h)\|_{(Y\times Y)^*}\right)(M_{z^h}+M_z).$$

Proof. We split

$$\begin{aligned} &g''(y^h)[z_1^h, z_2^h] - g''(y)[z_1, z_2] \\ &\quad = \left(g''(y^h) - g''(y)\right)[z_1, z_2] + \left(g''(y^h) - g''(y_h) + g''(y_h)\right)\left([z_1^h, z_2^h] - [z_1, z_2]\right) \\ &\quad = \left(g''(y^h) - g''(y)\right)[z_1, z_2] + \left(g''(y^h) - g''(y_h) + g''(y_h)\right)\left([z_1^h - z_1, z_2^h] + [z_1, z_2^h - z_2]\right), \end{aligned}$$

so that we can estimate

$$\begin{aligned} &|g''(y^h)[z_1^h, z_2^h] - g''(y)[z_1, z_2]| \\ &\leq \|g''(y^h) - g''(y)\|_{(Y\times Y)^*}\|z_1\|_Y\|z_2\|_Y \\ &\quad + \left(\|g''(y^h) - g''(y_h)\|_{(Y\times Y)^*} + \|g''(y_h)\|_{(Y\times Y)^*}\right)\left(\|z_1^h - z_1\|_Y\|z_2^h\|_Y + \|z_2^h - z_2\|_Y\|z_1\|_Y\right). \end{aligned}$$

Now upon applying the estimates (3.50)-(3.52) in Lemma 3.13, and the Lipschitz estimate (3.30f) of g'', we obtain

$$\begin{aligned} |g''(y^h)[z_1^h, z_2^h] - g''(y)[z_1, z_2]| &\leq c_{g''}\|y^h - y\|_Y M_z^2\|v_1\|_U\|v_2\|_U \\ &\quad + \left(c_{g''}\|y^h - y_h\|_Y + \|g''(y_h)\|_{(Y\times Y)^*}\right)\left(c_z M_z^h + c_z M_z\right)\|u - u_h\|_U\|v_1\|_U\|v_2\|_U. \end{aligned}$$

Employing the estimates (3.39) and (3.33) for $\|y^h - y\|_Y$ and $\|y^h - y_h\|_Y$ we obtain finally

$$\begin{aligned} &|g''(y^h)[z_1^h, z_2^h] - g''(y)[z_1, z_2]| \\ &\qquad \leq \left(c_{g''}c_y M_z^2 + c_z\left(c_{g''}\epsilon_y + \|g''(y_h)\|_{(Y\times Y)^*}\right)\left(M_{z^h} + M_z\right)\right)\|u - u_h\|_U\|v_1\|_U\|v_2\|_U, \end{aligned}$$

which is the desired result. □

In order to simplify the exposition of the Lipschitz estimates of the part of f'' that involves E'', we present the next lemma, where we have in mind to choose $G = E''(y, u)$.

Lemma 3.15. *Let G be a bounded bilinear form on the space $U\times Y$, i.e., $G : (U\times Y)\times(U\times Y) \mapsto Y^*$. Let $v_i \in U$ and for $i = 1, 2$ let z_i, z_i^h be defined by (3.48) and (3.49) respectively. Then for the pairs $d^h = [(v_1, z_1^h), (v_2, z_2^h)]$ and $d = [(v_1, z_1), (v_2, z_2)]$ of directions it holds*

$$\|G(d)\|_{Y^*} \leq M_d\,\|G\|_{\mathcal{B}(U\times Y, Y^*)}\|v_1\|_U\|v_2\|_U, \tag{3.53}$$

$$\|G(d^h) - G(d)\|_{Y^*} \leq c_d\,\|G\|_{\mathcal{B}(U\times Y, Y^*)}\|u - u_h\|_U\|v_1\|_U\|v_2\|_U \tag{3.54}$$

where

$$\begin{aligned} M_d &= (1 + M_z)^2, \\ c_d &= c_z\,(2 + M_z + M_{z^h}). \end{aligned}$$

Proof. It holds

$$\begin{aligned}\|G(d)\|_{Y^*} &\leq \|G\|_{\mathcal{B}(U\times Y,Y^*)}(\|v_1\|_U + \|z_1\|_Y)(\|v_2\|_U + \|z_2\|_Y)\\ &\leq \|G\|_{\mathcal{B}(U\times Y,Y^*)}\left(\|v_1\|_U\|v_2\|_U + \|v_1\|_U\|z_2\|_Y + \|z_1\|_Y\|v_2\|_U + \|z_1\|_Y\|z_2\|_Y\right)\\ &\leq \|G\|_{\mathcal{B}(U\times Y,Y^*)}\left(1 + 2M_z + M_z^2\right)\|v_1\|_U\|v_2\|_U.\end{aligned}$$

For the second estimate, we have due to the bilinearity of G

$$\begin{aligned}G(d^h) - G(d) &= G[(v_1, z_1^h), (v_2, z_2^h)] - G[(v_1, z_1), (v_2, z_2)]\\ &= G[(v_1, 0), (v_2, 0)] + G[(v_1, 0), (0, z_2^h)] + G[(0, z_1^h), (v_2, 0)] + G[(0, z_1^h), (0, z_2^h)]\\ &\quad - G[(v_1, 0), (v_2, 0)] - G[(v_1, 0), (0, z_2)] - G[(0, z_1), (v_2, 0)] - G[(0, z_1), (0, z_2)]\\ &= G[(v_1, 0), (0, z_2^h - z_2)] + G[(0, z_1^h - z_1), (v_2, 0)]\\ &\quad + G[(0, z_1^h - z_1), (0, z_2^h)] + G[(0, z_1), (0, z_2^h - z_2)].\end{aligned}$$

Hence, we can estimate

$$\begin{aligned}\|G(d^h) - G(d)\|_{Y^*} &\leq \|G\|_{\mathcal{B}(U\times Y,Y^*)}(\|v_1\|_U\|z_2^h - z_2\|_Y + \|v_2\|_U\|z_1^h - z_1\|_Y\\ &\qquad + \|z_2^h - z_2\|_Y\|z_1\|_Y + \|z_1^h - z_1\|_Y\|z_2^h\|_Y)\\ &\leq \|G\|_{\mathcal{B}(U\times Y,Y^*)}c_z\left(2 + M_z + M_{z^h}\right)\|u - u_h\|_U\|v_1\|_U\|v_2\|_U,\end{aligned}$$

where we used the estimates (3.50)–(3.52) provided by Lemma 3.13. □

Lemma 3.16. *Let $v_i \in U$, $i = 1, 2$, be given. Let z_i, z_i^h, $i = 1, 2$ be defined as in Lemma 3.13, i.e. z_i and z_i^h solve (3.48) and (3.49), respectively. For $i = 1, 2$ let z_i, z_i^h be defined by (3.48) and (3.49) respectively. Let $u \in U_{ad}$, $\|u - u_h\|_U \leq R$, where R is as in Assumption 9, with associated adjoint state p that solves (3.38) be given. Then it holds*

$$\begin{aligned}&\left|\left\langle E''(y^h, u_h)[(v_1, z_1^h)(v_2, z_2^h)],\, p^h\right\rangle_{Y^*,Y} - \langle E''(y, u)[(v_1, z_1)(v_2, z_2)],\, p\rangle_{Y^*,Y}\right|\\ &\qquad\qquad \leq C_{E''}\|u - u_h\|_U\|v_1\|_U\|v_2\|_U\end{aligned}$$

with

$$C_{E''} := \left(c_{E''}\epsilon_y + \|E''(y_h, u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\right)(c_d M_p + c_p M_d) + M_d c_{E''}(c_y + 1)(M_p + Rc_p).$$

Proof. Let $d^h = [(v_1, z_1^h), (v_2, z_2^h)]$ and $d = [(v_1, z_1), (v_2, z_2)]$. We write

$$\begin{aligned}&\left\langle E''(y^h, u_h)(d^h),\, p^h\right\rangle_{Y^*,Y} - \langle E''(y, u)(d),\, p\rangle_{Y^*,Y}\\ &\qquad = \left\langle E''(y, u)(d),\, p^h - p\right\rangle_{Y^*,Y} + \left\langle E''(y^h, u_h)(d^h) - E''(y, u)(d),\, p^h\right\rangle_{Y^*,Y}\end{aligned} \tag{3.55}$$

and

$$E''(y^h,u_h)(d^h)-E''(y,u)(d)=E''(y^h,u_h)(d^h)-E''(y^h,u_h)(d)\ +\ E''(y^h,u_h)(d)-E''(y,u)(d). \tag{3.56}$$

The first term on the right-hand side of (3.55) is estimated using the estimate (3.53) and (3.40) as

$$\begin{aligned}\langle E''(y,u)d,p^h-p\rangle_{Y^*,Y}&\leq \|E''(y,u)d\|_{Y^*}\|p^h-p\|_Y\\&\leq M_d\|E''(y,u)\|_{\mathcal{B}(U\times Y,Y^*)}\|p^h-p\|_Y\|v_1\|_U\|v_2\|_U\\&\leq M_dc_p\|E''(y,u)\|_{\mathcal{B}(U\times Y,Y^*)}\|u-u_h\|_U\|v_1\|_U\|v_2\|_U.\end{aligned}$$

Applying the Lipschitz property (3.30d) of E'' we obtain

$$\begin{aligned}\|E''(y,u)\|_{\mathcal{B}(U\times Y,Y^*)}&\leq \|E''(y,u)-E''(y^h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}+\|E''(y^h,u_h)-E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\\&\quad+\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\\&\leq c_{E''}\left(\|y-y^h\|_Y+\|u-u_h\|_U+\|y^h-y_h\|_Y\right)+\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\\&\leq c_{E''}((c_y+1)R+\epsilon_y)+\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)},\end{aligned} \tag{3.57}$$

where we have used the error estimates (3.39) and (3.33). Hence, we get

$$\begin{aligned}&\langle E''(y,u)d,p^h-p\rangle_{Y^*,Y}\\&\quad\leq M_dc_p\left(c_{E''}((c_y+1)R+\epsilon_y)+\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\right)\|u-u_h\|_U\|v_1\|_U\|v_2\|_U.\end{aligned} \tag{3.58}$$

To estimate the first addend on the right-hand side of (3.56), we employ Lemma 3.15

$$\|E''(y^h,u_h)(d^h)-E''(y^h,u_h)(d)\|_{Y^*}\leq c_d\|E''(y^h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\|u-u_h\|_U\|v_1\|_U\|v_2\|_U. \tag{3.59}$$

Applying again the Lipschitz property (3.30d) of E'' and the residual estimate (3.33) we have the estimate

$$\begin{aligned}\|E''(y^h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}&\leq \|E''(y^h,u_h)-E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}+\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\\&\leq c_{E''}\|y^h-y_h\|_Y+\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\\&\leq c_{E''}\epsilon_y+\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}.\end{aligned} \tag{3.60}$$

Now substituting the estimate (3.60) in (3.59) we obtain

$$\begin{aligned}&\|E''(y^h,u_h)(d^h)-E''(y^h,u_h)(d)\|_{Y^*}\\&\qquad\leq c_d\left(c_{E''}\epsilon_y+\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\right)\|u-u_h\|_U\|v_1\|_U\|v_2\|_U.\end{aligned} \tag{3.61}$$

The second addend in (3.56) is estimated as in (3.57) and applying (3.53)

$$\|E''(y^h, u_h)(d) - E''(y, u)(d)\|_{Y^*} \leq c_{E''}(c_y + 1)M_d\|u - u_h\|_U\|v_1\|_U\|v_2\|_U. \tag{3.62}$$

Putting (3.58), (3.61), and (3.62) together the claim follows after simple factorization, with the bound $\|p^h\|_Y \leq M_p$, cf. (3.36). □

We have now obtained all the necessary ingredients to compute the Lipschitz constant of f''. The estimate is given in the following lemma.

Lemma 3.17. *Let $u \in U_{ad}$ with $\|u - u_h\|_U \leq R$, where R is as in Assumption 9, and $v_i \in U$, $i = 1, 2$, be given. Then the estimate*

$$|(f''(u) - f''(u_h))[v_1, v_2]| \leq c_{f''}\|u - u_h\|_U\|v_1\|_U\|v_2\|_U$$

holds with $c_{f''} = c_{j''} + C_{g''} + C_{E''}$.

Proof. The second derivative f'' is given by

$$f''(u_h)[v_1, v_2] = j''(u_h)[v_1, v_2] + g''(y^h)[z_1^h, z_2^h] - \langle E''(y^h, u_h)[(v_1, z_1^h), (v_2, z_2^h)], p^h\rangle_{Y^*,Y}.$$

Then we obtain

$$\begin{aligned}
&|\left(f''(u) - f''(u_h)\right)[v_1, v_2]| \\
&\qquad \leq \|j''(u) - j''(u_h)\|_{(U\times U)^*}\|v_1\|_U\|v_2\|_U + |g''(y)[z_1, z_2] - g''(y^h)[z_1^h, z_2^h]| \\
&\qquad\quad + \langle E''(y^h, u_h)[(v_1, z_1^h), (v_2, z_2^h)], p^h\rangle_{Y^*,Y} - \langle E''(y, u)[(v_1, z_1), (v_2, z_2)], p\rangle_{Y^*,Y}.
\end{aligned}$$

With the help of Lipschitz estimate (3.30h) for j'', Lemma 3.14 and Lemma 3.16 one finds

$$|\left(f''(u) - f''(u_h)\right)[v_1, v_2]| \leq (c_{j''} + C_{g''} + C_{E''})\|u - u_h\|_U\|v_1\|_U\|v_2\|_U,$$

which completes the proof. □

Using similar arguments and estimates, we derive now a uniform bound of f''.

Lemma 3.18. *Let $u \in U_{ad}$ with $\|u - u_h\|_U \leq R$, where R is as in Assumption 9, be given. Then there is a positive constant $M_{f''}$ such that*

$$\|f''(u)\|_{(U\times U)^*} \leq M_{f''}$$

holds with

$$M_{f''} := c_{j''}R + \|j''(u_h)\|_{(U\times U)^*} + c_{g''}c_yR + c_{g''}\epsilon_y + \|g''(y_h)\|_{(Y\times Y)^*} + M_d\left(c_{E''}\left((c_y+1)R+\epsilon_y\right) + \|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\right)(c_pR+M_p)\,.$$

Proof. Let us take $v_i \in U$ with $\|v_i\|_U = 1, i = 1,2$. Using (3.53) we have

$$\begin{aligned} |f''(u)[v_1,v_2]| &\le \|j''(u)\|_{(U\times U)^*} + \|g''(y)\|_{(Y\times Y)^*} + \left|\langle E''(y,u)[(v_1,z_1),(v_2,z_2)],p\rangle_{Y^*,Y}\right| \\ &\le \|j''(u)\|_{(U\times U)^*} + \|g''(y)\|_{(Y\times Y)^*} + M_d\|E''(y,u)\|_{\mathcal{B}(U\times Y,Y^*)}\|p\|_Y. \end{aligned} \tag{3.63}$$

We estimate each of the norms separately as follows. First, making use of the Lipschitz property (3.30h) it holds

$$\begin{aligned} \|j''(u)\|_{(U\times U)^*} &\le \|j''(u) - j''(u_h)\|_{(U\times U)^*} + \|j''(u_h)\|_{(U\times U)^*} \\ &\le c_{j''}R + \|j''(u_h)\|_{(U\times U)^*}, \end{aligned} \tag{3.64}$$

where we have estimated the norm $\|u-u_h\|_U$ by its upper bound R. Secondly, applying property (3.30f) of g'', estimate (3.39) and the residual estimate (3.33) we obtain

$$\begin{aligned} \|g''(y)\|_{(Y\times Y)^*} &\le \|g''(y) - g''(y^h)\|_{(Y\times Y)^*} + \|g''(y^h) - g''(y_h)\|_{(Y\times Y)^*} + \|g''(y_h)\|_{(Y\times Y)^*} \\ &\le c_{g''}c_yR + c_{g''}\epsilon_y + \|g''(y_h)\|_{(Y\times Y)^*}. \end{aligned} \tag{3.65}$$

With the aid of estimates (3.40) and (3.36), the norm of the adjoint state p is estimated as

$$\|p\|_Y \le \|p-p^h\|_Y + \|p^h\|_Y \le c_pR + M_p.$$

Finally putting the already obtained estimate (3.57) of $\|E''(y,u)\|_{\mathcal{B}(U\times Y,Y^*)}$, (3.64) and (3.65) in (3.63) we obtain

$$|f''(u)[v_1,v_2]| \le c_{j''}R + \|j''(u_h)\|_{(U\times U)^*} + c_{g''}c_yR + c_{g''}\epsilon_y + \|g''(y_h)\|_{(Y\times Y)^*} + M_d\left(c_{E''}\left((c_y+1)R+\epsilon_y\right) + \|E''(y_h,u_h)\|_{\mathcal{L}((U\times Y)^2,Y^*)}\right)(c_pR+\epsilon_p+\|p_h\|_Y)\,.$$

□

3.3.5 Computation of the coercivity constant α

Let us now describe how to determine the lower bound α in (3.10). The particular challenge is to find a computable estimate. Due to the finite-dimensional control space, the inequality

$$f''(u)[v,v] \ge \alpha\|v\|_U^2 \quad \forall v \in U \tag{3.66}$$

is equivalent to the inequality $\lambda_i \geq \alpha$ for all eigenvalues λ_i of all possible matrix realizations of f''. Let us choose

$$\{v_1 \dots v_n\} \text{ to be the canonical basis of } U = \mathbb{R}^n.$$

This choice also fits to the inequality constraints in U_{ad}: as they are posed component-wise, they are equivalent to inequality bounds on the coordinates of control vectors with respect to the chosen basis. With the basis $\{v_1 \dots v_n\}$ fixed, the inequality (3.66) is equivalent to the statement that all eigenvalues of the symmetric matrix F, $F_{ij} = f''(u)[v_i, v_j]$, $i, j = 1 \dots n$, are greater than α.

Let us recall the structure of f''. Let $u \in U_{ad}$ be given with the associated state y and adjoint p. Then we have

$$f''(u)[v_i, v_j] = j''(u)[v_i, v_j] + g''(y)[z_i, z_j] - \langle E''(y, u)[(v_i, z_i), (v_j, z_j)], p\rangle_{Y^*,Y} \tag{3.67}$$

where the functions z_i are the solutions of the linearized problem

$$E_y(y, u)z_i + E_u(y, u)v_i = 0.$$

As a consequence of representation (3.67), we have that the constant α in the inequality (3.12), which reads

$$f''(u_h)[w, w] \geq \alpha \|w\|_U^2 \quad \forall w \in U : \; w_k = 0 \; k \in A,$$

is equal to the smallest eigenvalue of the matrix L^h given by

$$L^h := \left(j''(u_h)[v_i, v_j] + g''(y^h)[z_i^h, z_j^h] - \langle E''(y^h, u_h)[(v_i, z_i^h), (v_j, z_j^h)], p^h\rangle_{Y^*,Y}\right)_{i,j\in I} \tag{3.68}$$

where the functions z_i^h solve

$$E_y(y^h, u_h)z_i^h + E_u(y^h, u_h)v_i = 0. \tag{3.69}$$

However, since y^h as well as p^h and z_i^h are solutions of the infinite-dimensional operator equations, the entries of L^h are not computable. We will overcome this difficulty by making use of the computable matrix

$$L_h := \left(j''(u_h)[v_i, v_j] + g''(y_h)[z_{i,h}, z_{j,h}] - \langle E''(y_h, u_h)[(v_i, z_{i,h}), (v_j, z_{j,h})], p_h\rangle_{Y^*,Y}\right)_{i,j\in I}, \tag{3.70}$$

where the elements $z_{i,h} \in Y_h$ are solutions of discrete linearized equations

$$\langle E_y(y_h, u_h)z_{i,h} + E_u(y_h, u_h)v_i, \psi_h\rangle_{Y^*,Y} = 0 \quad \forall \psi_h \in Y_h. \tag{3.71}$$

Let us emphasize that all the involved functions are discrete quantities and as such can be computed, which makes L_h computable as well. By construction, L^h and L_h are the matrix representations of $f''(u_h)$ and $f_h''(u_h)$, respectively, where f_h denotes the cost functional of the discrete problem.

We define α_h to be the minimum eigenvalue of matrix L_h. We will later on assume that $\alpha_h > 0$. This is a verifiable assumption since the entries and the eigenvalues of matrix L_h are computable. Moreover, if u_h is a solution of the discrete problem then $\alpha_h \geq 0$ holds by second-order necessary optimality conditions.

To be able to estimate the error between L^h and L_h, we make the following assumption on the discrete functions $z_{i,h}$ and the residuals of (3.71).

Assumption 10. *Let (y_h, u_h, p_h) be as in Assumption 7. Let $\{v_1 \dots v_n\}$ be the canonical basis of $\mathbb{R}^n$. Let $z_{i,h}$ be approximations of solutions of (3.71). We suppose that the upper bounds on the associated residuals are available as*

$$\|E_y(y_h, u_h)z_{i,h} + E_u(y_h, u_h)v_i\|_{Y^*} \leq r_{z,i} \quad \forall i = 1 \dots n.$$

Analogous to Lemma 3.6, we have the following result on the error between $z_{i,h}$ and z_i^h.

Lemma 3.19. *Let Assumption 10 be satisfied. Let z_i^h be given as solution of (3.69). Then it holds*

$$\|z_i^h - z_{i,h}\|_Y \leq \epsilon_{z,i} \tag{3.72}$$

with $\epsilon_{z,i} = \delta^{-1}((c_{E_u} + c_{E_y}\|z_{i,h}\|_Y)\epsilon_y + r_{z,i})$, $i = 1 \dots n$.

Proof. The difference $z_i^h - z_{i,h}$ satisfies

$$\begin{aligned} E_y(y^h, u_h)(z_i^h - z_{i,h}) &= -E_u(y^h, u_h)v_i - E_y(y^h, u_h)z_{i,h} \\ &= -(E_u(y^h, u_h) - E_u(y_h, u_h))v_i + (E_y(y^h, u_h) - E_y(y_h, u_h))z_{i,h} \\ &\quad + E_y(y_h, u_h)z_{i,h} + E_u(y_h, u_h))v_i. \end{aligned}$$

Using the Lipschitz estimates of E_u and E_y and the result of Lemma 3.6, we obtain

$$\|z_i^h - z_{i,h}\|_Y \leq \delta^{-1}((c_{E_u} + c_{E_y}\|z_{i,h}\|_Y)\epsilon_y + r_{z,i}).$$

□

Corollary 3.20. *Let Assumption 10 be satisfied. Then it holds*

$$\|z_i^h\|_Y \leq M_{z_i^h} \tag{3.73}$$

with $M_{z_i^h} := \epsilon_{z,i} + \|z_{i,h}\|_Y$, $i = 1 \dots n$.

The following lemmas, analogous to Lemma 3.14 and Lemma 3.15, are required in the subsequent computation.

Lemma 3.21. *Let z_i^h be defined by (3.49) and $z_{i,h}$ by (3.71). Then the following inequality holds*

$$\left|g''(y^h)[z_i^h, z_j^h] - g''(y_h)[z_{i,h}, z_{j,h}]\right| \leq \epsilon_{g''_{i,j}}$$

for $i, j = 1 \dots n$, where

$$\epsilon_{g''_{i,j}} := c_{g''} M_{z_i^h} M_{z_j^h} \epsilon_y + \|g''(y_h)\|_{(Y\times Y)^*} \left(M_{z_j^h}\epsilon_{z,i} + \|z_{i,h}\|_Y \epsilon_{z,j}\right).$$

Proof. We can write

$$\begin{aligned} &g''(y^h)[z_i^h, z_j^h] - g''(y_h)[z_{i,h}, z_{j,h}] \\ &\qquad = (g''(y^h) - g''(y_h))[z_i^h, z_j^h] + g''(y_h)([z_i^h, z_j^h] - [z_{i,h}, z_{j,h}]) \\ &\qquad = (g''(y^h) - g''(y_h))[z_i^h, z_j^h] + g''(y_h)\left([z_i^h - z_{i,h}, z_j^h] + [z_{i,h}, z_j^h - z_{j,h}]\right). \end{aligned} \tag{3.74}$$

We estimate the first addend of (3.74) using the Lipschitz estimate (3.30f) of g'', (3.73) and (3.33) to obtain

$$\begin{aligned} \left|(g''(y^h) - g''(y_h))[z_i^h, z_j^h]\right| &\leq \|g''(y^h) - g''(y_h)\|_{(Y\times Y)^*} \|z_i^h\|_Y \|z_j^h\|_Y \\ &\leq c_{g''} \|y^h - y_h\|_Y M_{z_i^h} M_{z_j^h} \\ &\leq c_{g''} M_{z_i^h} M_{z_j^h} \epsilon_y. \end{aligned} \tag{3.75}$$

The second addend is likewise estimated using (3.72) and (3.73) as

$$\begin{aligned} &\left|g''(y_h)\left([z_i^h - z_{i,h}, z_j^h] + [z_{i,h}, z_j^h - z_{j,h}]\right)\right| \\ &\qquad \leq \|g''(y_h)\|_{(Y\times Y)^*} \|z_i^h - z_{i,h}\|_Y \|z_j^h\|_Y + \|z_{i,h}\|_Y \|z_j^h - z_{j,h}\|_Y \\ &\qquad \leq \|g''(y_h)\|_{(Y\times Y)^*} \left(M_{z_j^h}\epsilon_{z,i} + \|z_{i,h}\|_Y \epsilon_{z,j}\right). \end{aligned} \tag{3.76}$$

Now putting (3.75) and (3.76) in (3.74) yields the claim. □

Lemma 3.22. *Let G be a bounded bilinear form on the space $U \times Y$, i.e.*

$$G : (U \times Y) \times (U \times Y) \mapsto Y^*.$$

Let $d^h_{i,j} := [(v_i, z^h_i), (v_j, z^h_j)]$ *and* $d_{i,j,h} := [(v_i, z_{i,h}), (v_j, z_{j,h})]$, $i, j = 1 \dots n$. *Then it holds*

$$\|G(d_{i,j,h})\|_{Y^*} \leq M_{d_{i,j,h}} \|G\|_{\mathcal{B}(U\times Y, Y^*)}, \tag{3.77}$$

$$\|G(d^h_{i,j})\|_{Y^*} \leq M_{d_{i,j}} \|G\|_{\mathcal{B}(U\times Y, Y^*)}, \tag{3.78}$$

$$\|G(d^h_{i,j}) - G(d_{i,j,h})\|_{Y^*} \leq \epsilon_{d_{i,j}} \|G\|_{\mathcal{B}(U\times Y, Y^*)} \tag{3.79}$$

for all $i, j = 1 \dots n$, *where the constants are given by*

$$\begin{aligned} M_{d_{i,j,h}} &:= (1 + \|z_{i,h}\|_Y)(1 + \|z_{j,h}\|_Y), \\ M_{d_{i,j}} &:= (1 + M_{z^h_i})(1 + M_{z^h_j}), \\ \epsilon_{d_{i,j}} &:= \epsilon_{z,i}(1 + \|z_{j,h}\|_Y) + \epsilon_{z,j}(1 + M_{z^h_i}). \end{aligned} \tag{3.80}$$

Proof. For the first two estimates, we follow the steps of the proof of (3.53) in Lemma 3.15. Since $\|v_i\|_U = 1$, we have

$$\begin{aligned} \|G(d_{i,j,h})\|_{Y^*} &= \|G[(v_i, z_{i,h}), (v_j, z_{j,h})]\|_{Y^*} \\ &\leq \|G\|_{\mathcal{B}(U\times Y, Y^*)}(\|v_i\|_U + \|z_{i,h}\|_Y)(\|v_j\|_U + \|z_{j,h}\|_Y) \\ &\leq \|G\|_{\mathcal{B}(U\times Y, Y^*)}(1 + \|z_{i,h}\|_Y)(1 + \|z_{j,h}\|_Y). \end{aligned}$$

Similarly using (3.73) we obtain (3.78). For the last estimate, the proof is analogous to that of (3.54). Applying the estimates (3.72) and (3.73) we obtain

$$\begin{aligned} \|G(d^h_{i,j}) - G(d_{i,j,h})\|_{Y^*} &= \|G[(v_i, z^h_i), (v_j, z^h_j)] - G[(v_i, z_{i,h}), (v_j, z_{j,h})]\|_{Y^*} \\ &\leq \|G\|_{\mathcal{B}(U\times Y, Y^*)}(\|v_i\|_U \|z^h_j - z_{j,h}\|_Y + \|v_j\|_U \|z^h_i - z_{i,h}\|_Y \\ &\qquad + \|z^h_j - z_{j,h}\|_Y \|z^h_i\|_Y + \|z^h_i - z_{i,h}\|_Y \|z_{j,h}\|_Y) \\ &\leq \|G\|_{\mathcal{B}(U\times Y, Y^*)} \left(\epsilon_{z,i}(1 + \|z_{j,h}\|_Y) + \epsilon_{z,j}(1 + M_{z^h_i})\right). \end{aligned}$$

□

Please note that the estimates $\epsilon_{g''_{i,j}}$ and $\epsilon_{d_{i,j}}$ are symmetric, e.g. it holds $\epsilon_{g''_{i,j}} = \epsilon_{g''_{j,i}}$, which follows from the structure of the bound $M_{z^h_i}$ given by (3.73). This is a nice coincidence as these error estimates are error bounds for symmetric perturbations of symmetric matrices.

Lemma 3.23. *Let* $d^h_{i,j}$ *and* $d_{i,j,h}$ *be as defined in Lemma 3.22. Then it holds*

$$\left|\langle E''(y_h, u_h)(d_{i,j,h}),\, p_h\rangle_{Y^*,Y} - \langle E''(y^h, u_h)(d^h_{i,j}),\, p^h\rangle_{Y^*,Y}\right| \leq \epsilon_{E''_{i,j}}$$

with

$$\epsilon_{E''_{i,j}} := M_{d_{i,j}} c_{E''} \epsilon_y M_p + \|E''(y_h, u_h)\|_{\mathcal{B}(U\times Y, Y^*)} \left(\epsilon_{d_{i,j}} M_p + M_{d_{i,j,h}} \epsilon_p\right).$$

Proof. It holds

$$\begin{aligned}
&\langle E''(y_h,u_h)(d_{i,j,h}),\, p_h\rangle_{Y^*,Y} - \langle E''(y^h,u_h)(d^h_{i,j}),\, p^h\rangle_{Y^*,Y}\\
&\quad= \langle E''(y_h,u_h)(d_{i,j,h}) - E''(y^h,u_h)(d^h_{i,j}),\, p^h\rangle_{Y^*,Y} + \langle E''(y_h,u_h)(d_{i,j,h}),\, p_h - p^h\rangle_{Y^*,Y}\\
&\quad\le \|E''(y_h,u_h)d_{i,j,h} - E''(y^h,u_h)(d^h_{i,j})\|_{Y^*}\|p^h\|_Y + \|E''(y_h,u_h)(d_{i,j,h})\|_{Y^*}\|p_h - p^h\|_Y.
\end{aligned} \tag{3.81}$$

We employ a similar splitting as in (3.56) to obtain

$$\begin{aligned}
E''(y_h,u_h)(d_{i,j,h}) - E''(y^h,u_h)(d^h_{i,j}) = &\left(E''(y_h,u_h) - E''(y^h,u_h)\right)(d^h_{i,j})\\
&+ E''(y_h,u_h)(d_{i,j,h}) - E''(y_h,u_h)(d^h_{i,j}).
\end{aligned}$$

Hence, by applying the estimates (3.78), (3.79), (3.30d) and (3.33) we obtain

$$\begin{aligned}
&\|E''(y_h,u_h)(d_{i,j,h}) - E''(y^h,u_h)(d^h_{i,j})\|_{Y^*}\\
&\quad\le \|\left(E''(y_h,u_h) - E''(y^h,u_h)\right)(d^h_{i,j})\|_{Y^*} + \|E''(y_h,u_h)(d_{i,j,h}) - E''(y_h,u_h)(d^h_{i,j})\|_{Y^*}\\
&\quad\le M_{d_{i,j}}\|E''(y_h,u_h) - E''(y^h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)} + \epsilon_{d_{i,j}}\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\\
&\quad\le M_{d_{i,j}}c_{E''}\|y_h - y^h\|_Y + \epsilon_{d_{i,j}}\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\\
&\quad\le M_{d_{i,j}}c_{E''}\epsilon_y + \epsilon_{d_{i,j}}\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}.
\end{aligned} \tag{3.82}$$

Due to (3.77), it holds

$$\|E''(y_h,u_h)(d_{i,j,h})\|_{Y^*} \le M_{d_{i,j,h}}\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}. \tag{3.83}$$

Altogether, substituting (3.82) and (3.83) in (3.81) we arrive at

$$\begin{aligned}
&\left|\langle E''(y_h,u_h)(d_{i,j,h}),\, p_h\rangle_{Y^*,Y} - \langle E''(y^h,u_h)(d^h_{i,j}),\, p^h\rangle_{Y^*,Y}\right|\\
&\qquad\le M_p\left(M_{d_{i,j}}c_{E''}\epsilon_y + \epsilon_{d_{i,j}}\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}\right) + \epsilon_p M_{d_{i,j,h}}\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)},
\end{aligned}$$

where we have applied (3.36) and (3.34) to estimate the norms $\|p^h\|_Y$ and $\|p^h - p_h\|_Y$, respectively. □

Now we are in the position to prove the following bounds for the entries of the error matrix $L^h - L_h$.

Lemma 3.24. *Let the matrices L^h and L_h be given by (3.68) and (3.70), respectively. Then it holds*

$$|L^h_{i,j} - L_{h_{i,j}}| \le \mathcal{E}_{i,j} := \epsilon_{g''_{i,j}} + \epsilon_{E''_{i,j}}, \quad i,j \in I.$$

Proof. By the definitions (3.68) and (3.70), the elements of the error matrix e_{ij} fulfill

$$\begin{aligned} e_{ij} &= j''(u_h)(v_i,v_j) + g''(y^h)(z_i^h,z_j^h) - \langle E''(y^h,u_h)[(v_i,z_i^h),(v_j,z_j^h)],p^h\rangle_{Y^*,Y} \\ &\quad - j''(u_h)(v_i,v_j) - g''(y_h)(z_{i,h},z_{j,h}) + \langle E''(y_h,u_h)[(v_i,z_{i,h}),(v_j,z_{j,h})],p_h\rangle_{Y^*,Y} \\ &\leq \left|g''(y^h)(z_i^h,z_j^h) - g''(y_h)(z_{i,h},z_{j,h})\right| \\ &\quad + \left|\langle E''(y_h,u_h)[(v_i,z_{i,h}),(v_j,z_{j,h})],\, p_h\rangle_{Y^*,Y} - \langle E''(y^h,u_h)[(v_i,z_i^h),(v_j,z_j^h)],\, p^h\rangle_{Y^*,Y}\right|. \end{aligned}$$

Applying the results of Lemma 3.21 and Lemma 3.23 completes the proof. □

We finalize the computation of the coercivity constant by recalling the following result from matrix perturbation theory.

Theorem 3.25. *Let the matrix A be perturbed by a symmetric matrix $\mathcal{E}$, and denote by $\lambda_k(A)$ the k-th eigenvalue of A. If A and $A+\mathcal{E}$ are $n\times n$ symmetric matrices, then*

$$|\lambda_k(A+\mathcal{E}) - \lambda_k(A)| \leq \|\mathcal{E}\|_2$$

for $k = 1\dots n$.

Proof. The simple proof can be found in [24, Corollary 8.1.6] or [16, Theorem 5.1]. □

Theorem 3.26. *Let α, α_h be the minimum eigenvalues of matrices L^h and L_h respectively. Then it holds*

$$\alpha \geq \alpha_h - \|\mathcal{E}\|_2,$$

where the error matrix $\mathcal{E} = (\mathcal{E}_{i,j})$ is given in Lemma 3.24.

Proof. The claim follows from the previous Theorem 3.25, as L^h, L_h, as well as $\mathcal{E}$ are symmetric matrices. Moreover, $\mathcal{E}_{i,j}$ is an upper bound of $|L^h_{i,j} - L_{h_{i,j}}|$, which implies $\|L^h - L_h\|_2 \leq \|\mathcal{E}\|_2$. □

If $\bar{u}$ is a solution of the optimization problem satisfying the SSC, then the bound α_h will eventually become positive. If mesh refinement is done in such a way that the residuals r_y, r_p, r_u vanish and $u_h \to \bar{u}$, then the error $\|\mathcal{E}\|_2$ will tend to zero as well.

3.3.6 Main result

Let us summarize the results obtained in this section so far. The goal of all these works was to derive bounds to apply the results of Section 3.2. Let us recall that these results were given in terms of quantities ϵ, α, σ, and $c_{f'}, c_{f''}, M_{f''}$, cf. Assumption 8 on page 48. All these constants were derived in the previous subsections. It remains to collect them and to present the main result, which is an error estimate for the error in the solution. Moreover, it allows to verify the fulfillment of the second-order sufficient condition a-posteriori.

Theorem 3.27. *Let ϵ, α, σ, and $c_{f'}, c_{f''}, M_{f''}$ be computed according to the results in this section. Let us suppose that these constants satisfy the assumptions of Corollary 3.5. Let us assume that*

$\frac{2\epsilon}{\alpha} < R$, *where* R *is as in Assumption 9. Then there exists a local solution* $\bar{u}$ *of* (P) *that satisfies the error bound*

$$\|\bar{u} - u_h\|_U \leq \frac{2\epsilon}{\alpha}.$$

The solutions $\bar{u}$ *fulfills the second-order sufficient condition given by* (3.9)–(3.10)*. Moreover, we have the a-posteriori error representation in terms of the residuals in Assumption 7*

$$\|\bar{u} - u_h\|_U \leq \frac{2}{\alpha}\left(r_u + \omega_y\, r_y + \omega_p\, r_p\right)$$

with weights given by

$$\omega_y = c_{E_u}\delta^{-1}\left(\delta^{-2}\left(c_{g'} + c_{E_y}\|p_h\|_Y\right) r_y + \|p_h\|_Y\right) + \|E_u(y_h, u_h)\|_{\mathcal{L}(U,Y^*)}\delta^{-1}\left(c_{g'} + c_{E_y}\|p_h\|_Y\right),$$
$$\omega_p = \delta^{-1}\left(c_{E_u} r_y + \|E_u(y_h, u_h)\|_{\mathcal{L}(U,Y^*)}\right).$$

Proof. The claim follows from Corollaries 3.4 and 3.5 as well as from the representation of ϵ in terms of r_u, r_y, r_p derived in Lemma 3.12. By assumption, the control $\bar{u}$ satisfies $\|\bar{u} - u_h\|_U < R$, hence it is within the neighborhood of u_h, where the local Lipschitz properties according to Assumption 9 are satisfied. □

3.4 Application to parameter optimization problems

In this section, we apply the developed abstract framework to the parameter optimization problems (1.2) and (1.3), which were introduced on page 1. First, we fix the following settings, which are common to both problems.

Throughout this section, Ω is a two or three dimensional bounded Lipschitz domain with polygonal boundary $\partial\Omega$. The state space is $Y = H_0^1(\Omega)$, its dual $Y^* = H^{-1}(\Omega)$ and the control space is $U = \mathbb{R}^n$. The functions y_d, u_a, u_b, the regularization parameter $\kappa > 0$ and the source term b are all given in appropriate spaces. Furthermore, the functionals g, j in (P) are given by

$$g(y) := \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)}, \qquad j(u) := \frac{\kappa}{2}\|u\|^2_{\mathbb{R}^n}.$$

At first, let us argue that the cost functional, in particular the functions g and j met the requirements of Assumption 9. Indeed, we find that (3.30e)–(3.30h) are satisfied with

$$c_{g'} = I_2^2,\ c_{j'} = \kappa,\ c_{g''} = 0,\ c_{j''} = 0.$$

Moreover, it holds $\|g''(y)\|_{(Y\times Y)^*} \leq I_2^2$ and $\|j''(u)\|_{(U\times U)^*} = \kappa$ for all $u \in U$, $y \in Y$, uniformly. Here, we denoted by I_2 the norm of the embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$. In the sequel, let us denote norm of the embedding $H_0^1(\Omega) \hookrightarrow L^p(\Omega)$ by I_p if such a continuous embedding exists.

We will argue in the sequel that the resulting optimization problems subject to the nonlinear elliptic equations (1.2) or (1.3) fulfill Assumption 1 on page 9. Let us first describe the em-

ployed discretization procedure. Afterwards, we will report on how the remaining estimates in Assumption 9 are computed for each problem.

3.4.1 Discretization and computation of residuals

We used standard finite element techniques to discretize the problem. The domain is split into triangles. The finite element space Y_h is the classical spaces of piecewise quadratic and continuous elements (P2).

The critical part is the computation of the residuals. Here, 'computation' refers to the fact that we need constant-free error estimates, i.e. we have to determine r_y satisfying $\|E(y_h, u_h)\|_{Y^*} \leq r_y$, no extra constants involved. That means, we cannot use standard residual-type a-posteriori error estimates. Nevertheless, there are quite a few options available, as for instance the so-called hyper-circle method, see e.g. [11], estimates based on local $H(\mathrm{div})$-error representations [55], equilibrated residuals [2], or functional error estimates [46].

We used a related technique, as described in [45]. Let $\sigma \in H(\mathrm{div})$ be given, i.e. $\sigma \in L^2(\Omega)^d$ with $\mathrm{div}(\sigma) \in L^2(\Omega)$. The residual in the equation $-\Delta y + d(y) = b$ at a discrete function y_h can be estimated using [45, inequality (65)] as

$$\begin{aligned}\|-\Delta y_h + d(y_h) - b\|_{H^{-1}} &\leq \|\nabla y_h - \sigma\|_{L^2} + \|\,\mathrm{div}(\sigma) + d(y_h) - b\|_{H^{-1}} \\ &\leq \|\nabla y_h - \sigma\|_{L^2} + I_2\|\,\mathrm{div}(\sigma) + d(y_h) - b\|_{L^2},\end{aligned}$$

where this inequality holds for all functions $\sigma \in H(\mathrm{div})$. In our computations, we used the Raviart-Thomas elements RT_1 to discretize the space $H(\mathrm{div})$. Let us denote the discrete space consisting of RT_1-elements by Σ_h, which is a conforming discretization by $\Sigma_h \subset H(\mathrm{div})$. In a post-processing step, we then computed $\sigma_h \in \Sigma_h$ as minimizer of

$$\min_{\sigma_h \in \Sigma_h} \|\nabla y_h - \sigma_h\|_{L^2}^2 + I_2^2\|\,\mathrm{div}(\sigma_h) + d(y_h) - b\|_{L^2}^2.$$

This problem is a quadratic minimization problem, which can be solved efficiently by e.g. the conjugate gradients method. Using the so-obtained $\sigma_h \in \Sigma_h$, the residual was computed as

$$\|-\Delta y_h + d(y_h) - b\|_{H^{-1}} \leq \|\nabla y_h - \sigma_h\|_{L^2} + I_2\|\,\mathrm{div}(\sigma_h) + d(y_h) - b\|_{L^2}.$$

We applied this technique to compute bounds of the residuals for the state and adjoint equations as well as for the linearized equations appearing in the eigenvalue problem associated to f''.

In Chapter 4, we will employ a slightly different technique of [59] to compute residuals in the state and adjoint equations.

3.4.2 Identification of coefficient in the main part of elliptic equation

Let us verify the assumptions for the optimization problems involving parameters in the differential operator (cf. (1.3)). To this end, let disjoint, measurable sets $\Omega_k \subset \Omega$ be given, $k = 1 \dots n$, with $\Omega = \bigcup_{k=1}^n \Omega_k$. In order to make the resulting differential operator coercive, the lower bound

on the coefficients u_k (in the representation of a) is a positive number, $u_{a,k} = \tau > 0$. Upper bounds are taken into account as well.

We set $Y = H_0^1(\Omega)$ with norm $\|y\|_Y^2 := \|\nabla y\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Omega)}^2$. The mapping E is now defined as

$$\langle E(y,u),\, v\rangle = \sum_{k=1}^{n} u_k \int_{\Omega_k} \nabla y \cdot \nabla v \,\mathrm{d}x - \int_\Omega bv \,\mathrm{d}x.$$

Here, $b \in L^2(\Omega)$ is a given data function. With this definition, we have that the differentiability requirements of Assumption 1 on page 9 are met. The following lemma states that also the strong monotonicity as well as Lipschitz continuity conditions of Assumptions 1 and 9 are fulfilled.

Lemma 3.28. *Let $u \in U_{ad}$ and $y \in Y$ be given. Then it holds*

$$\|E_y^{-1}(y,u)\|_{\mathcal{L}(Y^*,Y)} \le \delta^{-1}$$

with $\delta = \tau(1 - I_2^2)$, τ being the lower bound on the parameters u. Moreover, we have that the inequalities (3.30a)–(3.30d) of Assumption 9 hold with

$$c_E = \max(\|y^h\|_Y,\, \|u_h\|_U + R), \quad c_{E_y} = c_{E_u} = 1, \quad c_{E''} = 0.$$

In addition, the inequalities $\|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)} \le \|y_h\|_Y$ and $\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)} \le 1$ are satisfied.

Proof. First since the coefficients $u_k > 0$, it holds

$$\langle y, E_y(y,u)y\rangle_{Y^*,Y} = \sum_{k=1}^{n} u_k \int_{\Omega_k} |\nabla y|^2 \ge u_a \|\nabla y\|_{L^2(\Omega)}^2 \ge u_a(\|y\|_Y^2 - \|y\|_{L^2(\Omega)}^2) \ge u_a(1 - I_2^2)\|y\|_Y^2.$$

By denoting the lower bound on the coefficients $u_a = \tau$, the first statement of the lemma then follows from the inequality

$$\tau(1 - I_2^2)\|y\|_Y^2 \le \langle y, E_y(y,u)y\rangle_{Y^*,Y} \le \|y\|_Y \|E_y(y,u)y\|_{Y^*}.$$

Let $u, u_h \in U_{ad}$, $y^h \in Y$ be given with $\|u - u_h\| \le R$. Let $y = S(u)$ be the solution of $E(y,u) = 0$. At first, let us determine the Lipschitz constant of E:

$$\begin{aligned}\|E(y,u) - E(y^h,u_h)\|_{Y^*} &\le \sum_{k=1}^{n} \|(u_k - u_{h,k})\nabla y^h\|_{L^2(\Omega_k)} + \|u_k(\nabla y - \nabla y^h)\|_{L^2(\Omega_k)} \\ &\le \|u - u_h\|_U \|y^h\|_Y + (\|u_h\| + R)\|y - y^h\|_Y,\end{aligned}$$

so that (3.30a) holds with $c_E := \max(\|y^h\|_Y,\ \|u_h\|_U + R)$. Let us take $z \in Y$. Since $E_y(y,u)z = -\mathrm{div}(u\nabla z)$, it follows that

$$\|(E_y(y,u) - E_y(y^h,u_h))z\|_{Y^*} \le \sum_{k=1}^{n} \|(u_k - u_{h,k})\nabla z\|_{L^2(\Omega_k)} \le \|u - u_h\|_U \|z\|_Y,$$

which implies $c_{E_y} = 1$. A similar computation gives with $v \in U$

$$\|(E_u(y,u) - E_u(y^h,u_h))v\|_{Y^*} \le \sum_{k=1}^{n} \|v_k(\nabla y - \nabla y^h)\|_{L^2(\Omega_k)} \le \|y - y^h\|_Y \|v\|_U,$$

implying $c_{E_u} = 1$. With a similar estimate we immediately obtain $\|E_u(y_h,u_h)\|_{\mathcal{L}(U,Y^*)} \le \|y_h\|_Y$. Since E is bilinear with respect to (u,y), the second derivative $E''(y,u)$ is independent of (y,u), hence $c_{E''} = 0$. More precisely $\langle E_{yu}(y_h,u_h)[z,w],\ v\rangle_{Y^*,Y} = \sum_{k=1}^{n} w_k \int_{\Omega_k} \nabla z \nabla v$. This implies

$$\|E_{yu}(y_h,u_h)[z,w]\|_{Y^*} \le \sum_{k=1}^{n} \|w_k \nabla z\|_{L^2(\Omega_k)} \le \|w\|_U \|z\|_Y.$$

Noting that the second derivatives E_{yy}, E_{uu} vanish, we obtain $\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)} \le 1$. □

3.4.3 Parameter identification problem

Let us consider the elliptic problem (1.2). We will make the special choice

$$d(u;y) = \sum_{k=1}^{n} u_k d_k(y),$$

where $d_k : \mathbb{R} \to \mathbb{R}$ are assumed to be twice continuously differentiable, $k = 1 \dots n$, with the second derivatives being Lipschitz continuous on intervals $[-M, M]$ for all $M \ge 0$. Furthermore, we assume d_k to be monotonically increasing. Here we have in mind to work with $d_k(y) = y|y|^{k-2}$.

As a result, we define the nonlinear operator E as

$$E(y,u) := -\Delta y + \sum_{k=1}^{n} u_k d_k(y) - b, \tag{3.84}$$

where $b \in L^2(\Omega)$ is a given function. In order to make the resulting operator monotonic with respect to y we impose positivity requirements on u, i.e. we set

$$U_{ad} = \{u \in \mathbb{R}^n :\ u_k \ge 0 \quad \forall k = 1 \dots n\}.$$

Due to the choice of functions d_k, the operator E is Fréchet-differentiable from $H^1(\Omega) \cap L^\infty(\Omega) \to H^{-1}(\Omega)$. That is we have to work with the framework $Y = H_0^1(\Omega)$, $Y_\infty = L^\infty(\Omega) \cap Y$. We will use the following norm in Y: $\|y\|_Y^2 := \|\nabla y\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Omega)}^2$.

Lipschitz estimates

At first, let us consider the Nemytskii (superposition) operators induced by the functions d_k. For simplicity, we will denote them by d_k, too.

Lemma 3.29. *The Nemytskii operators d_k are twice Fréchet-differentiable from $L^\infty(\Omega)$ to $L^\infty(\Omega)$. Moreover, we have*

$$\begin{aligned}
\|d_k(y) - d_k(y^h)\|_{L^p(\Omega)} &\leq \Phi_k(\max(\|y\|_{L^\infty(\Omega)}, \|y^h\|_{L^\infty(\Omega)}))\|y - y^h\|_{L^p(\Omega)},\\
\|d_k'(y) - d_k'(y^h)\|_{L^p(\Omega)} &\leq \Phi_k'(\max(\|y\|_{L^\infty(\Omega)}, \|y^h\|_{L^\infty(\Omega)}))\|y - y^h\|_{L^p(\Omega)},\\
\|d_k''(y) - d_k''(y^h)\|_{L^p(\Omega)} &\leq \Phi_k''(\max(\|y\|_{L^\infty(\Omega)}, \|y^h\|_{L^\infty(\Omega)}))\|y - y^h\|_{L^p(\Omega)}
\end{aligned}$$

for all $y, y^h \in L^\infty(\Omega)$, $p \in [1, +\infty]$. Here we used the functions

$$\Phi_k(M) := \max_{x\in[-M,M]} |d_k'(x)|, \Phi_k'(M) := \max_{x\in[-M,M]} |d_k''(x)|,$$

and $\Phi_k''(M)$ denotes the Lipschitz modulus of d_k'' on the interval $[-M, M]$, i.e.

$$|d_k''(x_1) - d_k''(x_2)| \leq \Phi_k''(M)|x_1 - x_2|$$

for $x_1, x_2 \in [-M, M]$.

Proof. Due to the assumptions on the functions d_k the Nemytskii operators d_k are twice continuously Fréchet differentiable. By the mean value theorem, we have

$$|d_k(x_1) - d_k(x_2)| \leq \Phi_k(\max(|x_1|, |x_2|))|x_1 - x_2|$$

for all $x_1, x_2 \in \mathbb{R}$. Hence,

$$\|d_k(y) - d_k(y^h)\|_{L^p(\Omega)} \leq \Phi_k(\max(\|y\|_{L^\infty(\Omega)}, \|y^h\|_{L^\infty(\Omega)}))\|y - y^h\|_{L^p(\Omega)}.$$

With analogous arguments we obtain the estimates for the derivatives of d_k. □

In order to prove Lipschitz estimates of E, we need Lipschitz estimates and global bounds for solutions of the equation $E(y, u) = 0$.

Lemma 3.30. *Let $u \in U_{ad}$, $y \in Y_\infty$ be given. Then it holds*

$$\|E_y^{-1}(y, u)\|_{\mathcal{L}(Y^*,Y)} \leq \delta^{-1}$$

with $\delta = 1 - I_2^2$.

Proof. This follows from the fact that $d_y(u; y) \in L^\infty(\Omega)$ as well as u is non-negative and from the simple inequality $\|\nabla y\|_{L^2(\Omega)}^2 = \|y\|_Y^2 - \|y\|_{L^2(\Omega)}^2 \geq (1 - I_2^2)\|y\|_Y^2$. □

Lemma 3.31. *Let $E : Y \times \mathbb{R}^n \to Y^*$ be given as in (3.84). Let $u \in U_{ad}$ be given. Then it holds for $y = S(u)$ being the solution of $E(y,u) = 0$*

$$\|y\|_{L^\infty(\Omega)} \leq M_{L^\infty}\|b\|_{L^2(\Omega)} + M_{u,L^\infty}\|u\|_U$$

with

$$M_{L^\infty} = 4\frac{I_6^2}{1-I_2^2}|\Omega|^{1/6}, \qquad M_{u,L^\infty} = M_{L^\infty}|\Omega|^{1/2}(\sum_{k=1}^n |d_k(0)|^2)^{1/2}.$$

Proof. Due to Stampacchia [50], we have the estimate

$$\|y\|_{L^\infty(\Omega)} \leq M_{L^\infty}\left(\|b\|_{L^2(\Omega)} + \sum_{k=1}^n |u_k| \cdot \|d_k(0)\|_{L^2(\Omega)}\right)$$

with $M_{L^\infty} = 4\frac{I_6^2}{1-I_2^2}|\Omega|^{1/6}$ computed in [47]. □

Lemma 3.32. *Let $u, u_h \in U_{ad}$, $\|u - u_h\|_U \leq R$, $y = S(u) \in Y_\infty$ and $y^h \in Y_\infty$ be given. Then it holds*

$$\|E(y,u) - E(y^h,u_h)\|_{Y^*} \leq c_E(\|y^h - y\|_Y + \|u - u_h\|_U)$$

with

$$c_E := \max\left(1 + I_2^2\|\Phi(u_h, y^h, R)\|_U(\|u_h\|_U + R),\, I_2\|d(y^h)\|_U\right),$$

where we used the abbreviations

$$\|d(y^h)\|_U := \left(\sum_{k=1}^n \|d_k(y^h)\|^2_{L^2(\Omega)}\right)^{1/2},$$
$$\|\Phi(u_h, y^h, R)\|_U := \left(\sum_{k=1}^n \Phi_k\left(\max(\|y^h\|_{L^\infty(\Omega)},\, M_{L^\infty} + M_{u,L^\infty}(R + \|u_h\|_U))\right)\right)^{1/2}.$$

This implies inequality (3.30a).

Proof. The claim follows from the splitting

$$u_k d_k(y) - u_{h,k} d_k(y^h) = (u_k - u_{h,k})d_k(y^h) + u_k(d_k(y) - d_k(y^h))$$

and applying Lemma 3.29. E.g. we have using the embedding $Y \hookrightarrow L^2(\Omega)$

$$\|u_k(d_k(y) - d_k(y^h))\|_{Y^*} \leq I_2^2|u_k| \cdot \Phi_k(\max(\|y\|_{L^\infty(\Omega)},\, \|y^h\|_{L^\infty(\Omega)}))\|y - y^h\|_Y.$$

Applying the L^∞-bound given by Lemma 3.31 finishes the proof. □

Lemma 3.33. *Let $u, u_h \in U_{ad}$, $\|u - u_h\|_U \leq R$, $y = S(u) \in Y_\infty$ and $y^h \in Y_\infty$ be given. Then the inequality (3.30b), i.e.*

$$\|E_y(y,u) - E_y(y^h,u_h)\|_{\mathcal{L}(Y,Y^*)} \leq c_{E_y}(\|y^h - y\|_Y + \|u - u_h\|_U),$$

is fulfilled with

$$c_{E_y} := \max\left(I_3^3 \|\Phi'(u_h, y^h, R)\|_U(\|u_h\|_U + R),\, I_3^2 \|d'(y^h)\|_U\right),$$

where we used the abbreviations

$$\|d'(y^h)\|_U := \left(\sum_{k=1}^n \|d_k'(y^h)\|_{L^3(\Omega)}^2\right)^{1/2},$$
$$\|\Phi'(u_h, y^h, R)\|_U := \left(\sum_{k=1}^n \Phi_k'\left(\max(\|y^h\|_{L^\infty(\Omega)},\, M_{L^\infty} + M_{u,L^\infty}(R + \|u_h\|_U))\right)\right)^{1/2}.$$

Proof. The proof is analogous to the proof of the previous Lemma 3.32. □

Lemma 3.34. *Let $u, u_h \in U_{ad}$, $\|u - u_h\|_U \leq R$, $y = S(u) \in Y_\infty$ and $y^h \in Y_\infty$ be given. Then the inequalities*

$$\|E_u(y,u) - E_u(y^h,u_h)\|_{\mathcal{L}(U,Y^*)} \leq c_{E_u}\|y^h - y\|_Y,$$
$$\|E_{yu}(y,u) - E_{yu}(y^h,u_h)\|_{\mathcal{L}(Y,\mathcal{L}(U,Y^*))} \leq c_{E_{yu}}\|y^h - y\|_Y$$

are fulfilled with

$$c_{E_u} := I_2^2 \|\Phi(u_h, y^h, R)\|_U(\|u_h\|_U + R),$$
$$c_{E_{yu}} := I_3^3 \|\Phi'(u_h, y^h, R)\|_U(\|u_h\|_U + R),$$

where we used the notation of Lemma 3.32 and 3.33. This gives (3.30b).

Proof. Let $w \in U$ be given. By construction, we have $E_u(y,u)w = \sum_{k=1}^n w_k d_k(y)$, which obviously implies $(E_u(y,u) - E_u(y^h,u_h))w = \sum_{k=1}^n w_k(d_k(y) - d_k(y^h))$. Using Lemma 3.29, we obtain

$$\|(E_u(y,u) - E_u(y^h,u_h))w\|_{Y^*} \leq \|w\|_U \|\Phi_k(\max(\|y\|_{L^\infty(\Omega)},\, \|y^h\|_{L^\infty(\Omega)}))\|y - y^h\|_Y.$$

The claim follows with the same argumentation as in the proof of Lemma 3.32. By analogous considerations we obtain the Lipschitz estimate for E_{yu}. □

Lemma 3.35. *Let $u, u_h \in U_{ad}$, $\|u - u_h\|_U \leq R$, $y = S(u) \in Y_\infty$ and $y^h \in Y_\infty$ be given. Then the inequality*

$$\|E_{yy}(y,u) - E_{yy}(y^h,u_h)\|_{\mathcal{B}(Y,Y^*)} \leq c_{E_{yy}}(\|y^h - y\|_Y + \|u - u_h\|_U),$$

is fulfilled with

$$c_{E_{yy}} := \max\left(I_4^4\|\Phi''(u_h, y^h, R)\|_U(\|u_h\|_U + R),\, I_4^3\|d''(y^h)\|_U\right),$$

where we used the abbreviations

$$\|d''(y^h)\|_U := \left(\sum_{k=1}^{n} \|d_k''(y^h)\|_{L^4(\Omega)}^2\right)^{1/2},$$
$$\|\Phi''(u_h, y^h, R)\|_U := \left(\sum_{k=1}^{n} \Phi_k''\left(\max(\|y^h\|_{L^\infty(\Omega)},\, M_{L^\infty} + M_{u,L^\infty}(R + \|u_h\|_U))\right)\right)^{1/2}.$$

Proof. The proof is analogous to the proof of Lemma 3.32. □

Corollary 3.36. *Let $u, u_h \in U_{ad}$, $\|u - u_h\|_U \le R$, $y = S(u) \in Y_\infty$ and $y^h \in Y_\infty$ be given. Then the inequality (3.30d), i.e.*

$$\|E''(y,u) - E''(y^h, u_h)\|_{\mathcal{B}(U\times Y, Y^*)} \le c_{E''}(\|y^h - y\|_Y + \|u - u_h\|_U)$$

is satisfied with

$$c_{E''} = c_{E_{yy}} + c_{E_{yu}}$$

where $c_{E_{yu}}$ and $c_{E_{yy}}$ are given by Lemmata 3.34 and 3.35, respectively.

Proof. The constant $c_{E''}$ can be determined as the spectral norm of the matrix $\begin{pmatrix} c_{E_{yy}} & c_{E_{yu}} \\ c_{E_{yu}} & 0 \end{pmatrix}$. The largest eigenvalue of this matrix is given by $\frac{1}{2}\left(c_{E_{yy}} + \sqrt{c_{E_{yy}}^2 + 4c_{E_{yu}}^2}\right) \le c_{E_{yy}} + c_{E_{yu}}$, which gives the bound $c_{E''}$. □

Let us close these considerations with stating the Lipschitz constants for the choice $d_1 = 1$, $d_k = y|y|^{k-2}$ for $k \ge 2$. Special care has to be taken as $d_3 = y|y|$ is not twice continuously differentiable. If we restrict all the considerations to positive values of y, then the previous results still hold.

Corollary 3.37. *Let the functions d_k be given by $d_1 = 1$, $d_k = y|y|^{k-2}$ for $k = 2 \dots n$. Then we have*

$$\Phi_k(M) = (k-1)M^{k-2},$$
$$\Phi_k'(M) = \max(0,\ (k-1)(k-2)M^{k-3}),$$
$$\Phi_k''(M) = \max(0,\ (k-1)(k-2)(k-3)M^{k-4}) \quad k \ne 3.$$

If $y, y^h \in Y_\infty$ are non-negative then the claims of Lemma 3.35 and Corollary 3.36 are true with $\Phi_k''(M) = \max(0,\ (k-1)(k-2)(k-3)M^{k-4})$ for $k = 1 \dots n$.

Lemma 3.38. *It holds*

$$\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)} \le \sum_{k=1}^{n} \left(I_2^2 \|d_k'(y_h)\|_{L^\infty} + I_2 I_4^2 u_{h,k} \|d_k''(y_h)\|_{L^\infty} \right).$$

Proof. Let $(w,z) \in U \times Y$. The following estimates hold

$$\begin{aligned}
\|E_{yu}(y_h,u_h)[w,z]\|_{Y^*} &\le \sum_{k=1}^{n} \|w_k d_k'(y_h) z\|_{Y^*} \le I_2^2 \sum_{k=1}^{n} \|d_k'(y_h)\|_{L^\infty} \|w\|_U \|z\|_Y, \\
\|E_{yy}(y_h,u_h)[z,z]\|_{Y^*} &\le \sum_{k=1}^{n} \|u_{h,k}\, d_k''(y_h)[z,z]\|_{Y^*} \le I_2 I_4^2 \sum_{k=1}^{n} u_{h,k} \|d_k''(y_h)\|_{L^\infty} \|z\|_Y^2.
\end{aligned}$$

Since the second derivative of E with respect to u vanishes, the claim then follows from the inequality $\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)} \le \|E_{yy}(y_h,u_h)\|_{\mathcal{L}(Y\times Y,Y^*)} + \|E_{yu}(y_h,u_h)\|_{\mathcal{L}(U\times Y,Y^*)}$. □

3.4.4 Numerical results

Let us report about the outcome of our numerical experiments. The first example is concerned with the optimization of coefficients in the main part of the operator, see Section 3.4.2.

Let us comment briefly on the computation of the safety radius R, which appears in the previous sections. An adaptive procedure was employed. Starting with initial guess $R = 0$, the control error bound r_+ was computed according to Corollary 3.4. If $r_+ > R$ the safety radius R was updated as $R := \theta r_+$ with $\theta = 1.01$. The computation of r_+ was then repeated until the condition $r_+ \le R$ is fulfilled.

Example 1

The domain was chosen as $\Omega = (0,1)^2$. The domain was split into four sub-domains

$$\Omega_1 = (0,0.5)^2,\ \Omega_2 = (0,0.5)\times(0.5,1),\ \Omega_3 = (0.5,1)\times(0,0.5),\ \Omega_4 = (0.5,1)^2.$$

The problem data was given as

$$u_a = 0.1,\ u_b = +\infty,\ \kappa = 10^{-1},\ y_d(x_1,x_2) = (1-x_1)^2(1-x_2)x_1x_2^2,\ b = 10.0001.$$

We solved the discretized problem on a sequence of uniformly refined grids. The solution vector on the finest grid was computed to

$$\bar{u}_h = (0.8047,\ 0.8062,\ 0.8020,\ 0.8047).$$

As can be seen, the inequality constraints are not active at $\bar{u}_h$, hence $A = \emptyset$.

The results for the verification process are shown in Tables 3.1 and 3.2. As can be seen from the second column of Table 3.1, the discrete sufficient optimality condition is satisfied on all grids, as α_h is uniformly positive. However, we can only verify that $f''(u_h)$ is positive definite

for the grids with $h \leq 0.0177$, since the error $\|\mathcal{E}\|_2$ is larger than α_h on the coarser grids. But since $\|\mathcal{E}\|_2$ decays like h^2 the condition $\alpha = \alpha_h - \|\mathcal{E}\|_2$ is eventually satisfied. In this example, we see that as soon as the bound α is positive, the conditions of Theorem 3.27 are satisfied, and hence an error estimate $\|\bar{u} - u_h\|_U \leq r_+$ is available. These include the statements that we are able to verify the existence of a local solution $\bar{u}$ of (P) in the neighborhood of u_h and the fulfillment of second-order sufficient optimality condition at $\bar{u}$. For $h = 0.0022$ the latter condition holds with

$$f''(\bar{u})(v,v) \geq (\alpha - c_{f''} r_+)\|v\|_U^2 = 4.2600 \cdot 10^{-2}\|v\|_U^2 \quad \forall v \in U.$$

Moreover, the error bound r_+ decays with h^2, which is expected, since $r_+ \leq 2\epsilon/\alpha$ holds and ϵ tends to zero with the rate h^2, as can be seen Table 3.1.

h	α_h	$\|\mathcal{E}\|_2$	α	ϵ	r_+
0.0707	$1.4644 \cdot 10^{-1}$	$8.0093 \cdot 10^{-1}$	$-6.5449 \cdot 10^{-1}$	$3.1342 \cdot 10^{-2}$	–
0.0354	$1.4644 \cdot 10^{-1}$	$2.1288 \cdot 10^{-1}$	$-6.6443 \cdot 10^{-2}$	$8.3960 \cdot 10^{-3}$	–
0.0177	$1.4644 \cdot 10^{-1}$	$5.6723 \cdot 10^{-2}$	$8.9718 \cdot 10^{-2}$	$2.2376 \cdot 10^{-3}$	$4.9880 \cdot 10^{-2}$
0.0088	$1.4644 \cdot 10^{-1}$	$1.5054 \cdot 10^{-2}$	$1.3139 \cdot 10^{-1}$	$5.9291 \cdot 10^{-4}$	$9.0255 \cdot 10^{-3}$
0.0044	$1.4644 \cdot 10^{-1}$	$3.9754 \cdot 10^{-3}$	$1.4247 \cdot 10^{-1}$	$1.5627 \cdot 10^{-4}$	$2.1938 \cdot 10^{-3}$
0.0022	$1.4644 \cdot 10^{-1}$	$1.0453 \cdot 10^{-3}$	$1.4540 \cdot 10^{-1}$	$4.0995 \cdot 10^{-5}$	$5.6391 \cdot 10^{-4}$

Table 3.1: Example 1: verification results, α, ϵ, r_+

For convenience, we also report about the other quantities involved in the verification process, namely the Lipschitz constants of the reduced functional f. As can be seen in Table 3.2, the Lipschitz constants $c_{f'}$ and $c_{f''}$ as well as the bound $M_{f''}$ are bounded uniformly for all discretizations. They are monotonically decreasing due to their computation, which is the expected behavior in the light of the derivation in Section 3.3. Of course, all these constants are expected to be bounded away from zero.

h	$\epsilon_{f'}$	$c_{f'}$	$c_{f''}$	$M_{f''}$
0.0707	$3.1342 \cdot 10^{-2}$	8.8765	$1.8794 \cdot 10^{2}$	9.7845
0.0354	$8.3960 \cdot 10^{-3}$	8.7049	$1.8383 \cdot 10^{2}$	9.4297
0.0177	$2.2376 \cdot 10^{-3}$	8.6589	$1.8273 \cdot 10^{2}$	9.3346
0.0088	$5.9291 \cdot 10^{-4}$	8.6466	$1.8244 \cdot 10^{2}$	9.3092
0.0044	$1.5627 \cdot 10^{-4}$	8.6433	$1.8236 \cdot 10^{2}$	9.3025
0.0022	$8.0962 \cdot 10^{-5}$	8.6424	$1.8234 \cdot 10^{2}$	9.3007

Table 3.2: Example 1: verification results, Lipschitz constants

Example 2

Let us present the results of the computations for a problem similar to Example 1, but with changed parameters

$$u_a = 0.4572,\ u_b = +\infty,\ \kappa = 10^{-1},\ y_d(x_1, x_2) = \sin(15x_1x_2)e^{\frac{x_1}{2}+\frac{x_2}{3}}.$$

The discrete solution was computed to

$$\bar{u}_h = (0.4572,\ 0.7669,\ 0.7780,\ 0.8871).$$

As one can see, the inequality constraint $u_{a,1} \leq u_1$ is active, giving rise to the choice $A = \{1\}$ of the active set.

h	α_h	$\|\mathcal{E}\|_2$	α	ϵ	r_+
0.0707	$1.3108 \cdot 10^{-1}$	1.0209	$-8.8984 \cdot 10^{-1}$	$6.2166 \cdot 10^{-2}$	–
0.0354	$1.3107 \cdot 10^{-1}$	$2.7589 \cdot 10^{-1}$	$-1.4482 \cdot 10^{-1}$	$1.6510 \cdot 10^{-2}$	–
0.0177	$1.3107 \cdot 10^{-1}$	$7.8072 \cdot 10^{-2}$	$5.2997 \cdot 10^{-2}$	$4.4240 \cdot 10^{-3}$	$1.6695 \cdot 10^{-1}$
0.0088	$1.3107 \cdot 10^{-1}$	$2.4100 \cdot 10^{-2}$	$1.0697 \cdot 10^{-1}$	$1.2418 \cdot 10^{-3}$	$2.3217 \cdot 10^{-2}$
0.0044	$1.3107 \cdot 10^{-1}$	$8.5596 \cdot 10^{-3}$	$1.2251 \cdot 10^{-1}$	$3.8949 \cdot 10^{-4}$	$6.3585 \cdot 10^{-3}$

Table 3.3: Example 2: verification results, α, ϵ, r_+

Table 3.3 depicts the computed error bounds for different mesh sizes h. Similar to example 1, the discrete sufficient optimality condition $\alpha_h > 0$ is satisfied for all meshes, while the positive definiteness of $f''(u_h)$ can be proven only for fine meshes. As in example 1, we get the convergence of ϵ and r_+ like h^2.

h	σ_h	$\epsilon_{f'}$	σ	$\sigma - c_{f'}r_+$
0.0707	$1.1202 \cdot 10^{-2}$	$6.2166 \cdot 10^{-2}$	$-5.0964 \cdot 10^{-2}$	–
0.0354	$1.1206 \cdot 10^{-2}$	$1.6510 \cdot 10^{-2}$	$-5.3042 \cdot 10^{-3}$	–
0.0177	$1.1207 \cdot 10^{-2}$	$4.4240 \cdot 10^{-3}$	$6.7826 \cdot 10^{-3}$	-1.9550
0.0088	$1.1207 \cdot 10^{-2}$	$1.2418 \cdot 10^{-3}$	$9.9649 \cdot 10^{-3}$	$-2.6231 \cdot 10^{-1}$
0.0044	$1.1207 \cdot 10^{-2}$	$3.8949 \cdot 10^{-4}$	$1.0817 \cdot 10^{-2}$	$-6.3714 \cdot 10^{-2}$

Table 3.4: Example 2: verification results, strongly active constraints

Now, let us have a closer inspection of the results with respect to the strongly active inequality constraints, the associated numbers can be found in Table 3.4. As can be seen, the active constraints are strongly active for the discrete problem, i.e. $\sigma_h > 0$. Moreover, they become strongly active for the continuous problem too, as σ is positive for the fine meshes. However, we were not to be able to verify that the constraints are active at the solution of the continuous problem, too. This would require to find $\sigma - c_{f'}r_+ > 0$, which was not the case in our computations. We expect, that this condition will become true for even finer discretizations, since σ and $c_{f'}$ converge to some fixed positive value, while r_+ decays for uniform refinement. For the Lipschitz constants of f we observe a similar behavior like in Example 1, see Table 3.5.

h	$c_{f'}$	$c_{f''}$	$M_{f''}$
0.0707	$1.2160 \cdot 10^{+1}$	$2.7978 \cdot 10^{+2}$	$1.4084 \cdot 10^{+1}$
0.0354	$1.1836 \cdot 10^{+1}$	$2.7135 \cdot 10^{+2}$	$1.3347 \cdot 10^{+1}$
0.0177	$1.1750 \cdot 10^{+1}$	$2.6911 \cdot 10^{+2}$	$1.3152 \cdot 10^{+1}$
0.0088	$1.1728 \cdot 10^{+1}$	$2.6852 \cdot 10^{+2}$	$1.3100 \cdot 10^{+1}$
0.0044	$1.1721 \cdot 10^{+1}$	$2.6836 \cdot 10^{+2}$	$1.3086 \cdot 10^{+1}$

Table 3.5: Example 2: verification results, Lipschitz constants

Example 3

For the identification problem analyzed in Section 3.4.3, the following choices were made

$$\Omega = (0,1)^2,\ u_a = 0,\ u_b = 0.5,\ \kappa = 10^{-2},\ b = 10.0001,\ y_d(x_1, x_2) = 0.5 \sin(2\pi x_1 x_2).$$

The discrete solution was computed to

$$\bar{u}_h = (0.5000,\ 0.2640,\ 0.1363,\ 0.0750),$$

which necessitates the choice $A = \{1\}$.

The computed bounds and constants can be found in Tables 3.6 and 3.7. As can be seen from Table 3.6, the verification assumptions were satisfied already on the coarsest mesh.

h	α_h	$\|\mathcal{E}\|_2$	α	ϵ	r_+
0.0707	$9.9631 \cdot 10^{-3}$	$4.7558 \cdot 10^{-3}$	$5.2073 \cdot 10^{-3}$	$5.3201 \cdot 10^{-4}$	$2.0433 \cdot 10^{-1}$
0.0354	$9.9631 \cdot 10^{-3}$	$1.2338 \cdot 10^{-3}$	$8.7293 \cdot 10^{-3}$	$1.3759 \cdot 10^{-4}$	$3.1523 \cdot 10^{-2}$
0.0177	$9.9631 \cdot 10^{-3}$	$3.2033 \cdot 10^{-4}$	$9.6427 \cdot 10^{-3}$	$3.5641 \cdot 10^{-5}$	$7.3924 \cdot 10^{-3}$
0.0088	$9.9631 \cdot 10^{-3}$	$8.3029 \cdot 10^{-5}$	$9.8800 \cdot 10^{-3}$	$9.2241 \cdot 10^{-6}$	$1.8672 \cdot 10^{-3}$
0.0044	$9.9631 \cdot 10^{-3}$	$2.1468 \cdot 10^{-5}$	$9.9416 \cdot 10^{-3}$	$2.3825 \cdot 10^{-6}$	$4.7931 \cdot 10^{-4}$

Table 3.6: Example 3: verification results, α, ϵ, r_+

Before discussing the observed behavior with respect to decreasing mesh-size, let us turn to the inspection of the results for one fixed discretization. For $h = 0.0177$, we obtained the fulfillment of Assumption 8 with $\sigma = 7.0463 \cdot 10^{-4}$, $\epsilon = 3.5641 \cdot 10^{-5}$, and $\alpha = 9.6427 \cdot 10^{-3}$. The corresponding values of $c_{f'}$ and $c_{f''}$ can be found in Table 3.7. By Theorem 3.27, there exists an optimal control $\bar{u}$ in the neighborhood of the discrete solution $\bar{u}_h$ with

$$\|\bar{u} - \bar{u}_h\|_U \leq 7.3924 \cdot 10^{-3}.$$

The safety radius was adaptively computed to $R = 7.5 \cdot 10^{-3}$. Hence the assumption $r_+ \leq R$ in Theorem 3.3 is fulfilled. Furthermore the condition (3.21) is satisfied with $\sigma - c_{f'} r_{+_1} =$

$3.5291 \cdot 10^{-4} > 0$, which implies $\bar{u} = \bar{u}_h$ on the active set A. Additionally, we have that the second-order sufficient optimality condition is fulfilled with

$$f''(\bar{u})[v, v] \geq 7.7625 \cdot 10^{-3} \|v\|_U^2$$

for all $v \in U$ with $v|_A = 0$.

h	$c_{f'}$	$c_{f''}$	$\sigma - c_{f'} r_+$	$\alpha - c_{f''} r_+$
0.0707	$5.0265 \cdot 10^{-2}$	$2.7562 \cdot 10^{-1}$	$-1.0063 \cdot 10^{-2}$	$-5.1112 \cdot 10^{-2}$
0.0354	$4.8313 \cdot 10^{-2}$	$2.5972 \cdot 10^{-1}$	$-9.2030 \cdot 10^{-4}$	$5.4221 \cdot 10^{-4}$
0.0177	$4.7579 \cdot 10^{-2}$	$2.5434 \cdot 10^{-1}$	$3.5291 \cdot 10^{-4}$	$7.7625 \cdot 10^{-3}$
0.0088	$4.7278 \cdot 10^{-2}$	$2.5238 \cdot 10^{-1}$	$6.4277 \cdot 10^{-4}$	$9.4088 \cdot 10^{-3}$
0.0044	$4.7144 \cdot 10^{-2}$	$2.5159 \cdot 10^{-1}$	$7.1530 \cdot 10^{-4}$	$9.8210 \cdot 10^{-3}$

Table 3.7: Example 3: verification results, Lipschitz constants

Similarly as in the previous examples, we observe a convergence rate $r_+ \sim h^2$. Moreover, for fine grids, we find that the sufficient second-order condition is satisfied at the *still unknown* local solution $\bar{u}$ of (P). First, the inequality constraints are strongly active at $\bar{u}$ on A, see the column '$\sigma - c_{f'} r_+$' in Table 3.7, which contains an estimate $|f'(\bar{u})_k| \geq \sigma - c_{f'} r_+$ for $k \in A$. And second, also $f''(\bar{u})$ is positive definite, as we have the lower bound $f''(\bar{u})[v, v] \geq (\alpha - c_{f''} r_+)\|v\|_U^2$ for all $v \in U$ with $v_k = 0, k \in A$, where $\alpha - c_{f''} r_+$ can be found in Table 3.7 as well.

Chapter 4

4 Adaptive methods for control problem with finite dimensional control space

The goal of this chapter is to prove efficiency of the error estimator obtained in Theorem 3.27. This will be achieved by deriving a lower bound for the error estimator. Using the then efficient estimator, different adaptive methods will be presented to illustrate its performance in mesh refinement procedure.

As a model problem for this chapter, we will again consider the optimal control problem (P) with finite dimensional control space $U = \mathbb{R}^n$. The analysis presented here also cover classes of optimization problems with PDE constraints of the forms

$$-\Delta y + d(u, y) = b \in Y^*, \quad -\mathrm{div}(u\nabla y) = b.$$

It is worth mentioning that some of the results presented in this chapter are contained in [5].

4.1 Introduction

Adaptive mesh refinement remains a valuable tool in scientific computation. The main objective of an adaptive procedure is to find a discrete solution to a problem while maintaining as few as possible numbers of unknowns with respect to a desired error estimate. As the solution and hence the error distributions on the mesh are unknown a-priori, one has to rely on a-posteriori error estimates.

A-posteriori error estimates for nonlinear control and identification problems can be found for instance in [9, 25, 36, 56]. However, they depend on two crucial *a-priori* assumptions: the first is that a second-order sufficient condition (SSC) has to hold at the solution of the continuous problem. With this assumption, error estimates of the type $\|\bar{u} - u_h\|_U \leq c\,\eta + \mathcal{R}$ can be derived, where η is a computable error indicator and $\mathcal{R}$ is a second-order remainder term. Here, the second a-priori assumption comes into play: one has to assume that $\mathcal{R}$ is small enough, in order to guarantee that mesh refinement solely based on η is meaningful. A different approach with respect to mesh refinement was followed in [64]. There the residuals in the first-order necessary optimality condition were used to derive an adaptive procedure. However, smallness of residuals does not imply smallness of errors without any further assumption. Here again, SSC as well as smallness of remainder terms is essential to draw this conclusion.

In Chapter 3, the sufficient optimality condition as well as smallness of remainders is checked *a-posteriori*. If both conditions are fulfilled, an error-estimator of the form

$$\|u - u_h\|_U \leq \frac{2}{\alpha}(\omega_y r_y + \omega_p r_p)$$

is available, cf. Theorem 3.27 on page 74. This error estimator is localizable if r_y and r_p are localizable error estimates for the norm of the residual in the state and adjoint equations, respectively.

In this chapter, we will prove a lower bound of the error estimator. For the setting $Y = H_0^1(\Omega)$, we obtain

$$r_y + r_p \leq c(\|u - u_h\|_U + \|y - y_h\|_Y + \|\nabla y - \sigma_h\|_{L^2(\Omega)} + \|p - p_h\|_Y + \|\nabla p - \tau_h\|_{L^2(\Omega)} + \tilde{o}),$$

where y, p and y_h, p_h are solutions of continuous and discrete state and adjoint equations, respectively, and σ_h and τ_h are approximations of ∇y and ∇p in $H(\mathrm{div})$. The term $\tilde{o}$ is a higher-order oscillation term. In addition, we have localized lower bounds for the residuals in the state and adjoint equations, respectively. These justify the use of the a-posteriori estimator above in an adaptive mesh-refinement procedure.

4.1.1 The abstract framework

Let Ω be a polygonal domain in $\mathbb{R}^m$, $m = 2, 3$. The function space for the states of the optimal control problem is chosen as $Y := H_0^1(\Omega)$. Abstract operator E is assumed for fulfill Assumption 1 on page 9.

4.1.2 Discretization

Let the abstract problem be discretized as described in Section 3.1.2 of Chapter 3 where a function $y_h \in Y_h$ is defined a solution of the discretized equation for a given $u \in U_{ad}$ if and only if

$$\langle E(y_h, u),\, \phi_h \rangle_{Y^*,Y} = 0 \qquad \forall \phi_h \in Y_h. \tag{4.1}$$

Furthermore let Assumption 7 on page 46 be fulfilled by the residuals r_y, r_u, r_p.

In Chapter 3 the following estimate

$$\alpha \geq \alpha_h - \|\mathcal{E}\|_2,$$

relating the coercivity constants α and α_h was obtained where $\|\mathcal{E}\|_2$ is the norm of an error matrix taking the discretization error in the linearized equation $E_y(\bar{u}, \bar{y})z + E_u(\bar{u}, \bar{y})v = 0$ into account. If the computable lower bound $\alpha_h - \|\mathcal{E}\|_2$ of α is positive, then it follows that the second-order condition (3.1) is satisfied. Moreover, we have the following result.

Theorem 4.1 (Upper bound of the error)**.** *Let Assumptions 1 and 7 be satisfied. Let (y_h, u_h, p_h) be a solution of the discrete optimal control problem. If $\alpha_h - \|\mathcal{E}\|_2 > 0$ holds and the residuals r_y and r_p are small enough, then there exists a local solution $\bar{u}$ of (P) that satisfies the error bound*

$$\|\bar{u} - u_h\|_U \leq \frac{2}{\alpha_h - \|\mathcal{E}\|_2} \left(\omega_y r_y + \omega_p r_p\right), \tag{4.2}$$

where the weights ω_y, ω_p depend on the discrete solution (y_h, u_h, p_h). If for different discretizations the discrete solutions $\{(y_h, u_h, p_h)\}_{h>0}$ are uniformly bounded in $Y \times U \times Y$ then the weights ω_y, ω_p are bounded as well.

Corollary 4.2. *Let the assumptions of Theorem 4.1 be satisfied. Let $\bar{y}$, $\bar{p}$ denote the solutions of the state and adjoint equations to $\bar{u}$, respectively. Then it holds*

$$\begin{aligned}
\|\bar{y} - y_h\|_Y &\leq \upsilon_{yu} \|\bar{u} - u_h\|_U + \delta^{-1} r_y, \\
\|\bar{p} - p_h\|_Y &\leq \upsilon_{pu} \|\bar{u} - u_h\|_U + \delta^{-1} r_p + \upsilon_{py} r_y,
\end{aligned}$$

with δ^{-1} being the global bound of $\|E_y^{-1}(y,u)\|_{\mathcal{L}(Y^*,Y)}$, and weights υ_{yu}, υ_{pu}, and υ_{py} depending on (y_h, u_h, p_h) in the same way as the weights ω_y, ω_p in Theorem 4.1.

Proof. Using triangle inequality, the result is a consequence of Theorem 4.1 and Lemmas 3.6, 3.8. □

4.2 Main result: Lower error bounds

We will distinguish between two classes of elliptic nonlinear PDEs representing the state equation. The derivation of the lower bounds for the error in the two cases are slightly different so that this separate treatment becomes necessary. The state space is chosen here as $Y = H_0^1(\Omega)$.

4.2.1 Problem class $E(y,u) = -\Delta y + d(y,u)$

In this case, the elliptic operator E is given as $E(y,u) = -\Delta y + d(y,u)$ with d being a superposition operator induced by a smooth function $d : \mathbb{R}^2 \to \mathbb{R}$.

The weak formulation of the state equation $E(y,u) = 0$ is given by: Find $y \in H_0^1(\Omega)$ satisfying

$$\mathcal{B}(y,v) = \int_\Omega \nabla y \nabla v \, \mathrm{d}x = \langle -d(y,u), v \rangle_{L^2}$$

or in vector form

$$\mathcal{B}_v(\nabla y, \nabla v) = \int_\Omega \nabla y \nabla v \, \mathrm{d}x = \langle -d(y,u), v \rangle_{L^2} \tag{4.3}$$

for all $v \in H_0^1(\Omega)$. By Cauchy-Schwarz inequality, the bilinear forms $\mathcal{B}, \mathcal{B}_v$ are continuous. The following important abstract estimate is curled from [59, Theorem 3.1].

Theorem 4.3. *Let $v, w, t \in L^2(\Omega)^m$ be arbitrary. Then it holds*

$$\|v-w\|_{L^2(\Omega)} \leq \|w-t\|_{L^2(\Omega)} + \left|\mathcal{B}_v\left(v-w, \frac{v-t}{\|v-t\|_{L^2(\Omega)}}\right)\right|.$$

Proof. Using bilinearity and continuity of $\mathcal{B}_v$ we estimate

$$\begin{aligned}\|v-t\|_{L^2}^2 = \mathcal{B}_v(v-t, v-t) &= \mathcal{B}_v(v-w, v-t) + \mathcal{B}_v(w-t, v-t)\\ &\leq \|v-t\|_{L^2}\, \mathcal{B}_v\left(v-w, \frac{v-t}{\|v-t\|_{L^2(\Omega)}}\right) + \|w-t\|_{L^2}\|v-t\|_{L^2}\\ &\leq \|v-t\|_{L^2}\, \left|\mathcal{B}_v\left(v-w, \frac{v-t}{\|v-t\|_{L^2(\Omega)}}\right)\right| + \|w-t\|_{L^2}\|v-t\|_{L^2}.\end{aligned}$$

If $\|v-w\|_{L^2} \leq \|v-t\|_{L^2}$, the result follows.

On the other hand, if $\|v-t\|_{L^2} \leq \|v-w\|_{L^2}$, as in the above computation it holds

$$\begin{aligned}\|v-w\|_{L^2}^2 = \mathcal{B}_v(v-w, v-w) &= \mathcal{B}_v(v-w, v-t) + \mathcal{B}_v(v-w, t-w)\\ &\leq \|v-t\|_{L^2}\, \mathcal{B}_v\left(v-w, \frac{v-t}{\|v-t\|_{L^2(\Omega)}}\right) + \|v-w\|_{L^2}\|w-t\|_{L^2}\\ &\leq \|v-w\|_{L^2}\, \left|\mathcal{B}_v\left(v-w, \frac{v-t}{\|v-t\|_{L^2(\Omega)}}\right)\right| + \|v-w\|_{L^2}\|w-t\|_{L^2}\end{aligned}$$

which yields the result in this case. □

We will work with a classical finite-element discretization: The discrete space Y_h is the classical space of piecewise quadratic and continuous elements (P2) on a given conforming triangulation $\mathcal{T}_h$ of Ω. The diameter of an element $T \in \mathcal{T}_h$ is denoted by h_T.

Let us endow $Y = H_0^1(\Omega)$ with the norm $\|y\|_Y^2 := \|\nabla y\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Omega)}^2$. In the sequel, let us denote the norm of the embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ by I_2.

Lemma 4.4. *Let I_2 denotes the norm of the embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$. Then for any $y \in H_0^1(\Omega)$, it holds*

$$\|y\|_{L^2(\Omega)}^2 \leq \frac{I_2^2}{1-I_2^2}\|\nabla y\|_{L^2(\Omega)}^2.$$

Proof. Due to the embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ and the definition of Y-norm it holds

$$\|y\|_{L^2(\Omega)}^2 \leq I_2^2\|y\|_Y^2 = I_2^2(\|\nabla y\|_{L^2(\Omega)}^2 + \|y\|_{L^2(\Omega)}^2).$$

The result follows on re-arranging the terms. □

Now let us report on the computation of the residual r_y in the state equation. As required by Assumption 7 on page 46, we are interested in constant-free error estimates, i.e. all constants

appearing in the a-posteriori error estimate must be computable. Here, we apply the results of Vohralík [59].

Theorem 4.5. *Let $y_h \in Y_h \subset H_0^1(\Omega)$, $u_h \in U_{ad}$ satisfy the discrete equation (4.1). Let $\sigma_h \in H(\mathrm{div})$ be given such that*

$$(\mathrm{div}\sigma_h,\, 1)_{L^2(T)} = (d(y_h, u_h),\, 1)_{L^2(T)} \quad \text{for all cells } T \in \mathcal{T}_h. \tag{4.4}$$

Let us define the cell-wise indicator $\eta_{y,T}$, $T \in \mathcal{T}_h$,

$$\eta_{y,T} := 2\|\nabla y_h - \sigma_h\|_{L^2(T)} + \pi^{-1} h_T \|d(y_h, u_h) - \mathrm{div}\sigma_h\|_{L^2(T)}. \tag{4.5}$$

Then it holds

$$\|-\Delta y_h + d(y_h, u_h)\|^2_{H^{-1}(\Omega)} \leq (1 - I_2^2)^{-1} \sum_{T \in \mathcal{T}_h} \eta_{y,T}^2 =: r_y^2. \tag{4.6}$$

If moreover, $\mathcal{T}_h$ is shape-regular, then it holds

$$\eta_{y,T} \leq C\|\nabla(\tilde{y} - y_h)\|_{L^2(T)} + \hat{c}\|\sigma_h - \nabla\tilde{y}\|_{L^2(T)} + c\; h_T \|d(y_h, u_h) - \Pi d(y_h, u_h\|_{L^2(T)} \tag{4.7}$$

where $\tilde{y} := \Delta^{-1} d(y_h, u_h)$ and Π denotes the orthogonal L^2-projection onto Y_h. The constants $C, c, \hat{c}$ depend only on the spatial dimension m and the shape regularity of the triangulation.

Proof. The upper bound (4.6) is a consequence of [59, Thm. 6.8, 6.12] taking [59, Remark 6.3] into account for σ_h satisfying $(\mathrm{div}\sigma_h,\, 1)_{L^2(T)} = (d(y_h, u_h),\, 1)_{L^2(T)}$, $T \in \mathcal{T}_h$. The proof goes thus: First, we estimate

$$\|\nabla\tilde{y} - \nabla y_h\|_{L^2(\Omega)} \leq \|\nabla\tilde{y} - \sigma_h\|_{L^2(\Omega)} + \|\sigma_h - \nabla y_h\|_{L^2(\Omega)}. \tag{4.8}$$

Now we use the idea of [59] to estimate the term $\|\nabla\tilde{y} - \sigma_h\|_{L^2}$. By setting $v = \nabla\tilde{y}, w = \sigma_h$ in Theorem 4.3 and for some vector $s \in H_0^1(\Omega)$, $t = \nabla s$ one immediately have

$$\begin{aligned} \|\nabla\tilde{y} - \sigma_h\|_{L^2(\Omega)} &\leq \|\sigma_h - \nabla s\|_{L^2(\Omega)} + \left|\mathcal{B}_v\left(\nabla\tilde{y} - \sigma_h, \frac{\nabla\tilde{y} - \nabla s}{\|\nabla\tilde{y} - \nabla s\|_{L^2(\Omega)}}\right)\right|, \quad \tilde{y} \neq s \\ &\leq \|\sigma_h - \nabla s\|_{L^2(\Omega)} + |\mathcal{B}_v\left(\nabla\tilde{y} - \sigma_h, \nabla\phi\right)| \end{aligned} \tag{4.9}$$

where $\phi = \frac{\tilde{y} - s}{\|\nabla(\tilde{y} - s)\|_{L^2(\Omega)}}$. Clearly, $\|\nabla\phi\|_{L^2(\Omega)} = 1$ and since $\tilde{y}, s \in H_0^1(\Omega)$ trivially $\phi \in H_0^1(\Omega)$.

The second addend on the right side of (4.9) is estimated as follows. Let π_0 be an L^2-orthogonal projection onto the space of piecewise constant polynomials. For an element $T \in \mathcal{T}_h$, let $\psi \in H^1(T)$. The Poincare inequality (see [8], [44]) implies

$$\|\psi - \pi_0\psi\|^2_{L^2(T)} \leq C_p h_T^2 \|\nabla\psi\|^2_{L^2(T)}, \;\; \forall \psi \in H^1(T) \tag{4.10}$$

with constant $C_p = \frac{1}{\pi^2}$. Since $\tilde{y} = \Delta^{-1}(d(y_h, u_h))$, the definition of the bilinear form associated with the abstract problem gives

$$\mathcal{B}_v(\nabla\tilde{y}, \nabla\phi) = \int_\Omega \nabla\tilde{y}\nabla\phi \, \mathrm{d}x = \langle -d(y_h, u_h), \phi\rangle_{L^2}. \tag{4.11}$$

Since $\phi \in H_0^1(\Omega)$, then using (4.11) and applying divergence theorem yield

$$\begin{aligned} \mathcal{B}_v(\nabla\tilde{y} - \sigma_h, \nabla\phi) &= \int_\Omega \nabla\tilde{y}\nabla\phi \, \mathrm{d}x - \int_\Omega \sigma_h\nabla\phi \, \mathrm{d}x \\ &= \int_\Omega -d(y_h, u_h)\phi \, \mathrm{d}x + \int_\Omega \mathrm{div}\sigma_h\phi \, \mathrm{d}x. \end{aligned} \tag{4.12}$$

To estimate the right hand side of (4.12), we will make use of condition (4.4). Observe that for any $\phi \in H^1(T)$, (4.4) implies

$$\int_T \Big(\mathrm{div}\sigma_h - d(y_h, u_h)\Big)\pi_0\phi \, \mathrm{d}x = 0. \tag{4.13}$$

Then by (4.13), Cauchy-Schwarz inequality, (4.10) and Poincare inequality, we continue estimating (4.12) with

$$\begin{aligned} |\mathcal{B}_v(\nabla\tilde{y} - \sigma_h, \nabla\phi)| &= \Big| \int_\Omega \Big(\mathrm{div}\sigma_h - d(y_h, u_h)\Big)\phi \, \mathrm{d}x\Big| \\ &= \Big| \sum_{T\in\mathcal{T}_h} \int_T \Big(\mathrm{div}\sigma_h - d(y_h, u_h)\Big)\phi \, \mathrm{d}x\Big| \\ &= \Big| \sum_{T\in\mathcal{T}_h} \int_T \Big(\mathrm{div}\sigma_h - d(y_h, u_h)\Big)(\phi - \pi_0\phi) \, \mathrm{d}x\Big| \\ &\leq \sum_{T\in\mathcal{T}_h} \|\mathrm{div}\sigma_h - d(y_h, u_h)\|_{L^2(T)}\|\phi - \pi_0\phi\|_{L^2(T)} \\ &\leq \sum_{T\in\mathcal{T}_h} \|\mathrm{div}\sigma_h - d(y_h, u_h)\|_{L^2(T)}\pi^{-1}h_T\|\nabla\phi\|_{L^2(T)}. \end{aligned}$$

By the choice of ϕ, the norm $\|\nabla\phi\|_{L^2(\Omega)} = 1$. We therefore obtain

$$\begin{aligned} |\mathcal{B}_v(\nabla\tilde{y} - \sigma_h, \nabla\phi)| &\leq \sum_{T\in\mathcal{T}_h} \|\mathrm{div}\sigma_h - d(y_h, u_h)\|_{L^2(T)}\pi^{-1}h_T\|\nabla\phi\|_{L^2(T)} \\ &\leq \Big[\sum_{T\in\mathcal{T}_h} \Big(\pi^{-1}h_T\|\mathrm{div}\sigma_h - d(y_h, u_h)\|_{L^2(T)}\Big)^2\Big]^{\frac{1}{2}} \Big[\sum_{T\in\mathcal{T}_h} \|\nabla\phi\|^2_{L^2(T)}\Big]^{\frac{1}{2}} \\ &\leq \sum_{T\in\mathcal{T}_h} \pi^{-1}h_T\|\mathrm{div}\sigma_h - d(y_h, u_h)\|_{L^2(T)}\|\nabla\phi\|_{L^2(\Omega)} \\ &= \sum_{T\in\mathcal{T}_h} \pi^{-1}h_T\|\mathrm{div}\sigma_h - d(y_h, u_h)\|_{L^2(T)}. \end{aligned} \tag{4.14}$$

On substituting back (4.14) into (4.9) and choosing $s = y_h$, one finally arrive at

$$\|\nabla\tilde{y} - \sigma_h\|_{L^2(\Omega)} \leq \sum_{T\in\mathcal{T}_h} \|\sigma_h - \nabla y_h\|_{L^2(T)} + \pi^{-1}h_T\|\mathrm{div}\sigma_h - d(y_h,u_h)\|_{L^2(T)}.$$

Altogether, we have by (4.8)

$$\|\nabla\tilde{y} - \nabla y_h\|_{L^2(\Omega)} \leq \sum_{T\in\mathcal{T}_h} 2\|\sigma_h - \nabla y_h\|_{L^2(T)} + \pi^{-1}h_T\|\mathrm{div}\sigma_h - d(y_h,u_h)\|_{L^2(T)}. \tag{4.15}$$

Let us now relate (4.15) to the residual in the state equation. With the aid of Lemma 4.4 we derive

$$\begin{aligned}
\|-\Delta y_h + d(y_h,u_h)\|^2_{H^{-1}(\Omega)} &= \|-(\Delta)^{-1}(-\Delta y_h + d(y_h,u_h))\|^2_{H^1(\Omega)}\\
&= \|y_h - \tilde{y}\|^2_{H^1(\Omega)}\\
&= \|y_h - \tilde{y}\|^2_{L^2(\Omega)} + \|\nabla(y_h - \tilde{y})\|^2_{L^2(\Omega)}\\
&\leq \frac{I_2^2}{1-I_2^2}\|\nabla(y_h - \tilde{y})\|^2_{L^2(\Omega)} + \|\nabla(y_h - \tilde{y})\|^2_{L^2(\Omega)}\\
&\leq (1-I_2^2)^{-1}\|\nabla(y_h - \tilde{y})\|^2_{L^2(\Omega)}.
\end{aligned}$$

The estimate (4.6) then follows on applying (4.15).

For the second part of the result, the lower bound (4.7) is a consequence of [59, Thm. 6.16], see also [58, Lemma 7.6]. The arguments of the proof are as follow. First observe that the bilinear form $\mathcal{B}$ is continuous and for all $\phi \in H^1_0(\Omega)$ it holds

$$\mathcal{B}(\tilde{y}-y_h,\phi) = \int_\Omega \nabla(\tilde{y}-y_h)\nabla\phi \leq \|\nabla(\tilde{y}-y_h)\|_{L^2(\Omega)}\|\nabla\phi\|_{L^2(\Omega)} \leq \|\nabla(\tilde{y}-y_h)\|_{L^2(\Omega)}\|\phi\|_{H^1(\Omega)} \tag{4.16}$$

by Cauchy-Schwarz and Poincare inequalities. Since ϕ vanishes on the boundary of Ω, integration by parts gives

$$\mathcal{B}(\tilde{y}-y_h,\phi) = \int_\Omega -d(y_h,u_h)\phi - \nabla y_h\nabla\phi = \sum_{T\in\mathcal{T}_h}\left(\int_T(\Delta y_h - d(y_h,u_h))\phi + \sum_{E\subset\partial T}\int_E\Big[\frac{\partial y_h}{\partial\nu}\Big]\phi\right) \tag{4.17}$$

where E denotes the edges of the elements of the triangulation and ν is an outward pointing unit normal on the element's boundaries. Now recall that $\eta_{y,T} := \|\nabla y_h - \sigma_h\|_{L^2(T)} + \pi^{-1}h_T\|d(y_h,u_h) - \mathrm{div}\sigma_h\|_{L^2(T)}$. Since $d(y_h,u_h)$ is in general not in the discrete space Y_h, we obtain an additional oscillation term $h_T\|d(y_h,u_h) - \Pi d(y_h,u_h)\|_{L^2(T)}$ as follows. By triangle inequality we have

$$\begin{aligned}
\eta_{y,T} &= \|\nabla y_h - \sigma_h\|_{L^2(T)} + \pi^{-1}h_T\|d(y_h,u_h) - \mathrm{div}\sigma_h\|_{L^2(T)}\\
&\leq \|\nabla(y_h - \tilde{y})\|_{L^2(T)} + \|\nabla\tilde{y} - \sigma_h\|_{L^2(T)} + \pi^{-1}h_T\|\mathrm{div}\sigma_h - \Pi d(y_h,u_h)\|_{L^2(T)}\\
&\quad + \pi^{-1}h_T\|\Pi d(y_h,u_h) - d(y_h,u_h)\|_{L^2(T)}
\end{aligned} \tag{4.18}$$

where Π denotes the L^2-orthogonal projection onto the discretized state space Y_h and $\tilde{y} = \Delta^{-1}d(y_h, u_h)$. It remains to estimate the norm of $v = \mathrm{div}\sigma_h - \Pi d(y_h, u_h)$ appearing in (4.18).

Let ψ_T be an interior bubble function on element $T \in \mathcal{T}_h$. Since ψ_T vanishes outside of element T as well as its boundary, then $\psi_T v$ localizes v to element T. By norm equivalence on finite dimensional space [3, Lemma 2.1 and Theorem 2.2] it holds

$$c_e\|v\|^2_{L^2(T)} \le \langle v, \psi_T v\rangle_{L^2(T)}, \tag{4.19}$$

$$\|\psi_T v\|_{H^1(T)} \le h_T^{-1}\|v\|_{L^2(T)}, \tag{4.20}$$

$$\|\psi_T v\|_{L^2(T)} \le \|v\|_{L^2(T)} \tag{4.21}$$

where c_e is a constant depending only on the element T. Thanks to (4.20), it is immediate from (4.16) that

$$\mathcal{B}(\tilde{y} - y_h, \psi_T v) \le \|\nabla(\tilde{y} - y_h)\|_{L^2(T)}\|\psi_T v\|_{H^1(T)} \le \|\nabla(\tilde{y} - y_h)\|_{L^2(T)} h_T^{-1}\|v\|_{L^2(T)}. \tag{4.22}$$

Using $\phi = \psi_T v$ in (4.17) we obtain

$$\begin{aligned}
\mathcal{B}(\tilde{y} - y_h, \psi_T v) &= \langle \Delta y_h - d(y_h, u_h), \psi_T v\rangle_{L^2(T)} + \sum_{E\subset\partial T}\int_E \Big[\frac{\partial y_h}{\partial \nu}\Big]\psi_T v \\
&= \langle \Delta y_h - d(y_h, u_h), \psi_T v\rangle_{L^2(T)} \\
&= \langle \Delta y_h - \mathrm{div}\sigma_h + \mathrm{div}\sigma_h - \Pi d(y_h, u_h) + \Pi d(y_h, u_h) - d(y_h, u_h), \psi_T v\rangle_{L^2(T)} \\
&= \langle v, \psi_T v\rangle_{L^2(T)} - \langle \mathrm{div}\sigma_h - \Delta y_h, \psi_T v\rangle_{L^2(T)} - \langle d(y_h, u_h) - \Pi d(y_h, u_h), \psi_T v\rangle_{L^2(T)}.
\end{aligned} \tag{4.23}$$

The middle term on the right hand side of (4.23) is estimated using integration by parts and the fact that the bubble function ψ_T vanishes on the boundary ∂T of element T. That is, by Cauchy-Schwarz inequality and (4.20) it holds

$$\begin{aligned}
\langle \mathrm{div}\sigma_h - \Delta y_h, \psi_T v\rangle_{L^2(T)} &= \langle \sigma_h - \nabla y_h, \nabla(\psi_T v)\rangle_{L^2(T)} + \langle (\sigma_h - \nabla y_h)\cdot\nu, \psi_T v\rangle_{L^2(\partial T)} \\
&= \langle \sigma_h - \nabla y_h, \nabla(\psi_T v)\rangle_{L^2(T)} \\
&\le \|\sigma_h - \nabla y_h\|_{L^2(T)}\|\nabla(\psi_T v)\|_{L^2(T)} \\
&\le \|\sigma_h - \nabla y_h\|_{L^2(T)}\|\psi_T v\|_{H^1(T)} \\
&\le \|\sigma_h - \nabla y_h\|_{L^2(T)} h_T^{-1}\|v\|_{L^2(T)}
\end{aligned}$$

where ν denotes an outward pointing unit normal on ∂T.

Now applying (4.19), using estimates (4.23), (4.22) and (4.20), we obtain

$$\begin{aligned}
c_e\|v\|^2_{L^2(T)} &\le \langle v, \psi_T v\rangle_{L^2(T)} \\
&= \mathcal{B}(\tilde{y}-y_h, \psi_T v) + \|\sigma_h - \nabla y_h\|_{L^2(T)} h_T^{-1}\|v\|_{L^2(T)} \\
&\quad + \langle d(y_h,u_h) - \Pi d(y_h,u_h), \psi_T v\rangle_{L^2(T)} \\
&\le \|\nabla(\tilde{y}-y_h)\|_{L^2(T)} h_T^{-1}\|v\|_{L^2(T)} + \|\sigma_h - \nabla y_h\|_{L^2(T)} h_T^{-1}\|v\|_{L^2(T)} \\
&\quad + \|d(y_h,u_h) - \Pi d(y_h,u_h)\|_{L^2(T)}\|\psi_T v\|_{L^2(T)} \\
&\le \|\nabla(\tilde{y}-y_h)\|_{L^2(T)} h_T^{-1}\|v\|_{L^2(T)} + \|\sigma_h - \nabla y_h\|_{L^2(T)} h_T^{-1}\|v\|_{L^2(T)} \\
&\quad + \|d(y_h,u_h) - \Pi d(y_h,u_h)\|_{L^2(T)}\|v\|_{L^2(T)}.
\end{aligned}$$

On dividing by $\|v\|_{L^2(T)}$, we obtain

$$\|v\|_{L^2(T)} \le c_e^{-1}h_T^{-1}\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} + c_e^{-1}h_T^{-1}\|\sigma_h - \nabla y_h\|_{L^2(T)} + c_e^{-1}\|d(y_h,u_h) - \Pi d(y_h,u_h)\|_{L^2(T)}.$$

Again since we can write $\|\sigma_h - \nabla y_h\|_{L^2(T)} \le \|\sigma_h - \nabla\tilde{y}\|_{L^2(T)} + \|\nabla(\tilde{y}-y_h)\|_{L^2(T)}$, it follows that

$$\begin{aligned}
\|v\|_{L^2(T)} \le c_e^{-1}h_T^{-1}\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} + c_e^{-1}h_T^{-1}\|\sigma_h - \nabla\tilde{y}\|_{L^2(T)} + c_e^{-1}h_T^{-1}\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} \\
+ c_e^{-1}\|d(y_h,u_h) - \Pi d(y_h,u_h)\|_{L^2(T)}. \quad (4.24)
\end{aligned}$$

Recall that $v = \mathrm{div}\sigma_h - \Pi d(y_h,u_h)$. Then (4.24) implies that the term $\pi^{-1}h_T\|\mathrm{div}\sigma_h - \Pi d(y_h,u_h)\|_{L^2(T)}$ in (4.18) can be estimated from the above by

$$\begin{aligned}
\pi^{-1}h_T\|\mathrm{div}\sigma_h - \Pi d(y_h,u_h)\|_{L^2(T)} &\le c_e^{-1}\pi^{-1}\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} + c_e^{-1}\pi^{-1}\|\sigma_h - \nabla\tilde{y}\|_{L^2(T)} \\
&\quad + c_e^{-1}\pi^{-1}\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} \\
&\quad + c_e^{-1}\pi^{-1}h_T\|d(y_h,u_h) - \Pi d(y_h,u_h)\|_{L^2(T)} \quad (4.25) \\
&= 2c_e^{-1}\pi^{-1}\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} + c_e^{-1}\pi^{-1}\|\sigma_h - \nabla\tilde{y}\|_{L^2(T)} \\
&\quad + c_e^{-1}\pi^{-1}h_T\|d(y_h,u_h) - \Pi d(y_h,u_h)\|_{L^2(T)}.
\end{aligned}$$

Finally using (4.25) in (4.18) gives

$$\begin{aligned}
\eta_{y,T} \le (2c_\pi+1)\|\nabla(\tilde{y}-y_h)\|_{L^2} + (c_\pi+1)\|\sigma_h - \nabla\tilde{y}\|_{L^2(T)} \\
+ \pi^{-1}(c_e^{-1}+1)h_T\|d(y_h,u_h) - \Pi d(y_h,u_h\|_{L^2(T)}
\end{aligned}$$

which is the desired result with $c_\pi = c_e^{-1}\pi^{-1}, C = 2c_\pi + 1, \hat{c} = c_\pi + 1, c = \pi^{-1}(c_e^{-1}+1)$. □

Similar estimate for the residual in the adjoint equation can be obtained after obvious modifications: for $\tau_h \in H(\mathrm{div})$ satisfying

$$(\mathrm{div}\tau_h,\, 1)_{L^2(T)} = (d'(y_h,u_h)p_h - g'(y_h),\, 1)_{L^2(T)} \quad \text{for all cells } T \in \mathcal{T}_h \quad (4.26)$$

and with the local error indicators defined by

$$\eta_{p,T} := 2\|\nabla p_h - \tau_h\|_{L^2(T)} + \pi^{-1} h_T \|d'(y_h, u_h)p_h - g'(y_h) - \mathrm{div}\tau_h\|_{L^2(T)}$$

we obtain the upper bound

$$\|-\Delta p_h + d'(y_h, u_h)p_h - g'(y_h)\|^2_{H^{-1}(\Omega)} \le (1 - I_2^2)^{-1} \sum_{T\in\mathcal{T}_h} \eta^2_{p,T} =: r_p^2 \tag{4.27}$$

as well as the lower bound

$$\begin{aligned}\eta_{p,T} \le C\|\nabla(\tilde{p} - p_h)\|_{L^2(T)} + \hat{c}\|\nabla\tilde{p} - \tau_h\|_{L^2(T)} \\ + c\ h_T\|(I - \Pi)(d'(y_h, u_h)p_h - g'(y_h))\|_{L^2(T)},\end{aligned} \tag{4.28}$$

where $\tilde{p} := \Delta^{-1}(d'(y_h, u_h)p_h - g'(y_h))$.

We remark that the upper bounds (4.6) and (4.27) are constant-free, making them explicitly computable. In our computations, we computed the functions σ_h and τ_h as a minimizer of the right-hand side in (4.6) and (4.27) (with constraints (4.4), (4.26)) respectively, using Raviart-Thomas elements for discretization of $H(\mathrm{div})$ space. This shows that the requirements of Assumption 7 on the computability of upper bounds on the residuals can be fulfilled.

Now let us argue that under the assumptions of Theorem 4.1 we also obtain lower bounds for the error, which proves efficiency of the error bound.

Theorem 4.6. *Let the assumptions of Theorem 4.1 be fulfilled. Let r_y and r_p be computed according to (4.6) and (4.27). Let $(\bar{y}, \bar{u}, \bar{p})$ be the local solution of (P) provided by Theorem 4.1. Then it holds*

$$\begin{aligned}\sum_{T\in\mathcal{T}_h} r_{y,T} \le \tilde{C}\Big(\|\bar{u} - u_h\|_U + \|\bar{y} - y_h\|_Y \\ + \|\nabla\bar{y} - \sigma_h\|_{L^2(\Omega)} + \sum_{T\in\mathcal{T}_h} h_T\|(I - \Pi)d(y_h, u_h)\|_{L^2(\Omega)}\Big),\end{aligned} \tag{4.29}$$

$$\begin{aligned}\sum_{T\in\mathcal{T}_h} r_{p,T} \le \tilde{C}\Big(\|\bar{u} - u_h\|_U + \|\bar{y} - y_h\|_Y + \|\bar{p} - p_h\|_Y \\ + \|\nabla\bar{p} - \tau_h\|_{L^2(\Omega)} + \sum_{T\in\mathcal{T}_h} h_T\|(I - \Pi)(d'(y_h, u_h)p_h - g'(y_h))\|_{L^2(\Omega)}\Big),\end{aligned} \tag{4.30}$$

where $\tilde{C} > 0$ depends only on the spatial dimension m, the shape regularity of the triangulation, and global bounds of derivatives d_y, d_u, d_{yy}, and d_{yu} of $d : Y \times U \to Y^*$ near (y_h, u_h).

Proof. The result of Theorem 4.5 gives $r_y^2 = \sum_{T\in\mathcal{T}_h} \eta_{y,T}^2$ with

$$\begin{aligned}\eta_{y,T} &\le C\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} + \hat{c}\|\sigma_h - \nabla\tilde{y}\|_{L^2(T)} + c\ h_T\|d(y_h,u_h) - \Pi d(y_h,u_h)\|_{L^2(T)} \\ &\le \hat{C}\Big(\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} + \|\sigma_h - \nabla\tilde{y}\|_{L^2(T)} + h_T\|d(y_h,u_h) - \Pi d(y_h,u_h)\|_{L^2(T)}\Big)\end{aligned} \tag{4.31}$$

where $\hat{C}$ is the maximum of the constants $C, \hat{c}$ and c. Let $\tilde{y}$ be given as $\tilde{y} := \Delta^{-1}d(y_h,u_h)$. Then we can estimate

$$\begin{aligned}\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} + \|\nabla\tilde{y}-\sigma_h\|_{L^2(T)} &\le 2\|\nabla(\tilde{y}-\bar{y})\|_{L^2(T)} + \|\nabla(\bar{y}-y_h)\|_{L^2(T)} + \|\nabla\bar{y}-\sigma_h\|_{L^2(T)} \\ &\le 2\|\nabla(\tilde{y}-\bar{y})\|_{L^2(T)} + \|\bar{y}-y_h\|_Y + \|\nabla\bar{y}-\sigma_h\|_{L^2(T)}.\end{aligned} \tag{4.32}$$

To estimate the first addend in the above, recall that the state $\bar{y}$ solves $-\Delta\bar{y} + d(\bar{y},\bar{u}) = 0$. Then with $\tilde{y} = \Delta^{-1}d(y_h,u_h)$ it follows that $-\Delta(\tilde{y}-\bar{y}) = d(y_h,u_h) - d(\bar{y},\bar{u})$ from which we obtain

$$\|\tilde{y}-\bar{y}\|_Y \le \delta^{-1}\|d(y_h,u_h) - d(\bar{y},\bar{u})\|_{L^2\Omega)}. \tag{4.33}$$

Now using (4.33) and Lipschitz continuity of d, we found

$$\begin{aligned}\|\nabla(\tilde{y}-\bar{y})\|_{L^2(\Omega)} \le \|\tilde{y}-\bar{y}\|_Y &\le \delta^{-1}\|d(y_h,u_h) - d(\bar{y},\bar{u})\|_{L^2(\Omega)} \\ &\le C_d(\|\bar{u}-u_h\|_U + \|\bar{y}-y_h\|_Y)\end{aligned} \tag{4.34}$$

with C_d depending on bounds of $\|d'\|_{\mathcal{L}(Y\times U,Y^*)}$ near (y_h,u_h). Hence (4.32) gives

$$\|\nabla(\tilde{y}-y_h)\|_{L^2(T)} + \|\nabla\tilde{y}-\sigma_h\|_{L^2(T)} \le \tilde{C}(\|\bar{u}-u_h\|_U + \|\bar{y}-y_h\|_Y + \|\nabla\bar{y}-\sigma_h\|_{L^2(T)}). \tag{4.35}$$

Finally, plugging (4.35) in (4.31), the result follows by writing $r_y = \sum_{T\in\mathcal{T}_h} r_{y,T}$. The estimate (4.30) can be obtained analogously. □

These lower bounds together with (4.2) and the local lower bounds (4.7) and (4.28) justify the use of the error indicators in an adaptive mesh-refinement procedure.

Remark 4.7. *Another possibility of constant-free a-posteriori error estimators based on $H(\mathrm{div})$-functions is described in [20]. There, fluxes across edges in a dual mesh are prescribed instead of the integrals on elements as in (4.4) and (4.26). In [20] it is proven that the resulting error estimate is reliable and efficient. Moreover, the terms $\|\nabla y_h - \sigma_h\|_{L^2(\Omega)}$ and $\|\nabla p_h - \tau_h\|_{L^2(\Omega)}$ do not appear in the lower error bound when compared to (4.29) and (4.30), respectively.*

Let us now extend the preceding analysis to operators in divergence form with bounded coefficients depending on the control u.

4.2.2 Problem class $E(y,u) = -\mathrm{div}(u\nabla y)$

In this case the elliptic operator E is given by $E(y,u) = -\mathrm{div}(a\nabla y) + b$ where the bounded coefficient $a = \sum_{i=1}^{n} \chi_i u_i$ and $b \in L^2(\Omega)$. The weak form of the elliptic operator $E(y,u) = 0$ reads: find $y \in H_0^1(\Omega)$ such that

$$\mathcal{B}(y,\phi) = \int_\Omega a\nabla y\nabla\phi = \langle -b,\phi\rangle_{L^2}$$

for every $\phi \in H_0^1(\Omega)$. In vector form this is equivalent to

$$\mathcal{B}_v(a\nabla y, a\nabla\phi) = \langle a^{\frac{1}{2}}\nabla y, a^{\frac{1}{2}}\nabla\phi\rangle_{L^2} = \langle -b,\phi\rangle_{L^2}$$

where the coefficient a is assumed to be positive and $\mathcal{B}_v(u,v) := \langle a^{-\frac{1}{2}}u, a^{-\frac{1}{2}}v\rangle_{L^2}$.

Let $a \in U_{ad}$ be given, then $y_h \in Y_h \subset H_0^1(\Omega)$ is a solution of the discretized equation if and only if

$$\mathcal{B}_v(y_h,\phi) = \int_\Omega a\nabla y_h\nabla\phi = \langle -b,\phi\rangle \tag{4.36}$$

for all test function $\phi \in Y_h$. Due to discrete Friedrich's inequality [57]

$$\|\phi\|^2_{H^1(\Omega)} \le C_F\|\nabla\phi\|^2_{L^2(\Omega)} \quad \forall\phi \in H_0^1(\Omega)$$

the semi-norm

$$|||\phi|||_a^2 := \mathcal{B}(\phi,\phi) = \langle a\nabla\phi,\nabla\phi\rangle = \|a^{\frac{1}{2}}\nabla\phi\|^2_{L^2}, \quad \phi \in H^1(\Omega)$$

becomes a norm on $H_0^1(\Omega)$. Similarly, through the bilinear form $\mathcal{B}_v$, we define the energy semi-norm for vectors $v \in L^2(\Omega)$ as

$$|||v|||^2_{a^{-1}} := \mathcal{B}_v(v,v) = \langle a^{-\frac{1}{2}}v, a^{-\frac{1}{2}}v\rangle_{L^2} = \|a^{-\frac{1}{2}}v\|^2_{L^2}. \tag{4.37}$$

Observe that setting $v = a\nabla y$ in (4.37) yields

$$|||a\nabla y|||_{a^{-1}} = \|a^{\frac{1}{2}}\nabla y\|_{L^2} = |||y|||_a. \tag{4.38}$$

Again thanks to the discrete Friedrich's inequality, the seminorm $|||\cdot|||_{a^{-1}}$ is equivalent to the full (weighted) norm on H^1 which is defined by

$$|||y|||^2_{H^1} := |||a\nabla y|||^2_{a^{-1}} + \|y\|^2_{L^2} = |||y|||^2_a + \|y\|^2_{L^2}. \tag{4.39}$$

Since we can estimate

$$|||a\nabla y|||_{a^{-1}}^2 = \|a^{-\frac{1}{2}} a\nabla y\|_{L^2}^2 = \|a^{\frac{1}{2}}\nabla y\|_{L^2}^2 \leq a_{\max}\|\nabla y\|_{L^2}^2,$$

then the weighted H^1-norm $|||\cdot|||_{H^1}$ is related to the standard H^1 Sobolev norm (cf. page 7) by

$$\begin{aligned} |||y|||_{H^1}^2 = |||a\nabla y|||_{a^{-1}}^2 + \|y\|_{L^2}^2 &\leq a_{\max}\|\nabla y\|_{L^2}^2 + \|y\|_{L^2}^2 \\ &\leq \max(a_{\max}, 1)\Big(\|\nabla y\|_{L^2}^2 + \|y\|_{L^2}^2\Big) \\ &= \max(a_{\max}, 1)\|y\|_{H^1} \\ &=: C_a\|y\|_{H^1}. \end{aligned} \tag{4.40}$$

Let us now proceed to computing the residual estimates r_y, r_p.

Theorem 4.8. *Let $y_h \in Y_h \subset H_0^1(\Omega)$, $a \in U_{ad}$ satisfy the discrete equation (4.36). Let $\sigma_h \in H(\mathrm{div})$ be given such that*

$$(\mathrm{div}\sigma_h,\, 1)_{L^2(T)} = (b,\, 1)_{L^2(T)} \quad \text{for all cells } T \in \mathcal{T}_h. \tag{4.41}$$

Let us define the cell-wise indicator $\eta_{y,T}$, $T \in \mathcal{T}_h$ as

$$\eta_{y,T} := 2|||a\nabla y_h - \sigma_h|||_{a^{-1},T} + \pi^{-1}h_T|||b - \mathrm{div}\sigma_h|||_{a^{-1},T}.$$

Then it holds

$$\|-\mathrm{div}(a\nabla y_h) + b\|_{H^{-1}(\Omega)}^2 \leq a_{max}\sum_{T\in\mathcal{T}_h}\eta_{y,T}^2 =: r_y^2 \tag{4.42}$$

where $a_{max} = \max_i(u_b(i))$. If moreover, $\mathcal{T}_h$ is shape-regular, then it holds

$$\eta_{y,T} \leq C|||a\nabla(\tilde{y} - y_h)|||_{a^{-1},T} + C|||\sigma_h - a\nabla\tilde{y}|||_{a^{-1},T} + c\ h_T\|b - \Pi b\|_{L^2(T)} \tag{4.43}$$

where $\tilde{y}$ is a solution of $-\mathrm{div}(a\nabla\tilde{y}) + b = 0$ and Π denotes the orthogonal L^2-projection onto Y_h. The constants C, c depend only on the spatial dimension m, the shape regularity of the triangulation, the constants a_{max} and $a_{min} = \min_i(u_a(i))$.

Proof. The proving technique is similar to that of Theorem 4.5. We proceed as follows. Firstly, using the definition of $\tilde{y}$ and (4.38) we estimate

$$\begin{aligned}
\|-\operatorname{div}(a\nabla y_h)+b\|^2_{H^{-1}(\Omega)} &\leq \|-\operatorname{div}(a\nabla y_h)+\operatorname{div}(a\nabla\tilde{y})\|^2_{H^{-1}(\Omega)} + \|-\operatorname{div}(a\nabla\tilde{y})+b\|^2_{H^{-1}(\Omega)} \\
&= \|-\operatorname{div}(a\nabla y_h)+\operatorname{div}(a\nabla\tilde{y})\|^2_{H^{-1}(\Omega)} \\
&\leq \|a\nabla y_h - a\nabla\tilde{y}\|^2_{L^2(\Omega)} \\
&\leq a_{\max}\|a^{\frac{1}{2}}\nabla(y_h-\tilde{y})\|^2_{L^2(\Omega)} \\
&= a_{\max}|||a\nabla(\tilde{y}-y_h)|||^2_{a^{-1}}.
\end{aligned} \tag{4.44}$$

Now we proceed to estimating (4.44) by writing

$$|||a\nabla(\tilde{y}-y_h)|||_{a^{-1}} \leq |||a\nabla\tilde{y}-\sigma_h|||_{a^{-1}} + |||\sigma_h - a\nabla y_h|||_{a^{-1}}. \tag{4.45}$$

Setting $v = a\nabla\tilde{y}, w = \sigma_h$ in Theorem 4.3 (which also hold for the scaled norm $|||\cdot|||_{a^{-1}}$, see [59, Theorem 3.1]) and for some vector $s \in H^1_0(\Omega), s \neq \tilde{y}$, $t = a\nabla s$ one obtains

$$\begin{aligned}
|||a\nabla\tilde{y}-\sigma_h|||_{a^{-1}} &\leq |||\sigma_h - a\nabla s|||_{a^{-1}} + \left|\mathcal{B}_v\left(a\nabla\tilde{y}-\sigma_h, \frac{a\nabla\tilde{y}-a\nabla s}{|||a\nabla\tilde{y}-a\nabla s|||_{a^{-1}}}\right)\right| \\
&\leq |||\sigma_h - a\nabla s|||_{a^{-1}} + |\mathcal{B}_v\left(a\nabla\tilde{y}-\sigma_h, a\nabla\phi\right)|
\end{aligned} \tag{4.46}$$

where $\phi = \frac{\tilde{y}-s}{|||\tilde{y}-s|||_a}$ (we have used again (4.38)). We have $|||\phi|||_a = 1$ and $\phi \in H^1_0(\Omega)$.

Using the definition $\mathcal{B}_v(u,v) = \langle a^{-\frac{1}{2}}u, a^{-\frac{1}{2}}v\rangle = \langle u, a^{-1}v\rangle$, we compute

$$\begin{aligned}
\mathcal{B}_v\left(a\nabla\tilde{y}-\sigma_h, a\nabla\phi\right) &= \int_\Omega (a\nabla\tilde{y}-\sigma_h)\nabla\phi \,\mathrm{d}x \\
&= \int_\Omega a\nabla\tilde{y}\nabla\phi \,\mathrm{d}x - \int_\Omega \sigma_h\nabla\phi \,\mathrm{d}x \\
&= \int_\Omega -b\phi \,\mathrm{d}x + \int_\Omega \operatorname{div}\sigma_h\phi \,\mathrm{d}x \\
&= \int_\Omega (\operatorname{div}\sigma_h - b)\phi \,\mathrm{d}x.
\end{aligned}$$

Now employing (4.41) and Cauchy-Schwarz inequality we estimate

$$\begin{aligned}
|\mathcal{B}_v\left(a\nabla\tilde{y}-\sigma_h, a\nabla\phi\right)| &= \left|\int_\Omega (\operatorname{div}\sigma_h - b)\phi \,\mathrm{d}x\right| \\
&= \left|\sum_{T\in\mathcal{T}_h}\int_T (\operatorname{div}\sigma_h - b)\phi \,\mathrm{d}x\right| \\
&= \left|\sum_{T\in\mathcal{T}_h}\int_T (\operatorname{div}\sigma_h - b)(\phi-\pi_0\phi) \,\mathrm{d}x\right| \\
&\leq \sum_{T\in\mathcal{T}_h} \|\operatorname{div}\sigma_h - b\|_{L^2(T)}\|\phi-\pi_0\phi\|_{L^2(T)}.
\end{aligned} \tag{4.47}$$

Using the fact that the coefficient a is piecewise constant on the elements, we continue our estimation of (4.47) by employing (4.10), (4.37) and (4.38) to obtain

$$\begin{aligned}\left|\mathcal{B}_v\left(a\nabla\tilde{y}-\sigma_h, a\nabla\phi\right)\right| &\leq \sum_{T\in\mathcal{T}_h} \|\mathrm{div}\sigma_h - b\|_{L^2(T)}\|\phi-\pi_0\phi\|_{L^2(T)}\\ &\leq \sum_{T\in\mathcal{T}_h} \|b-\mathrm{div}\sigma_h\|_{L^2(T)}\pi^{-1}h_T\|\nabla\phi\|_{L^2(T)}\\ &\leq \sum_{T\in\mathcal{T}_h} \|a^{-\frac{1}{2}}(b-\mathrm{div}\sigma_h)\|_{L^2(T)}\pi^{-1}h_T\|a^{\frac{1}{2}}\nabla\phi\|_{L^2(T)}\\ &= \sum_{T\in\mathcal{T}_h} |||b-\mathrm{div}\sigma_h|||_{a^{-1},T}\pi^{-1}h_T|||\phi|||_{a,T}.\end{aligned}$$

With a similar computation as in (4.14) on page 94, using the fact that $|||\phi|||_a = 1$ we obtain

$$\left|\mathcal{B}_v\left(a\nabla\tilde{y}-\sigma_h, a\nabla\phi\right)\right| \leq \sum_{T\in\mathcal{T}_h} \pi^{-1}h_T|||b-\mathrm{div}\sigma_h|||_{a^{-1},T}.$$

Successive substitution of the above estimate in (4.46) with $s = y_h$ and then (4.46) in (4.45) give

$$|||a\nabla\tilde{y}-a\nabla y_h|||_{a^{-1}} \leq \sum_{T\in\mathcal{T}_h} 2|||\sigma_h - a\nabla y_h|||_{a^{-1},T} + \pi^{-1}h_T|||b-\mathrm{div}\sigma_h|||_{a^{-1},T}. \tag{4.48}$$

Finally slotting (4.48) in (4.44) finishes the first part of the claim.

For the second part, since the source term $b \in L^2$ does not belong to the discretized space Y_h in general, using the projection Π we estimate

$$\begin{aligned}\eta_{y,T} &\leq 2|||a\nabla y_h - \sigma_h|||_{a^{-1},T} + \pi^{-1}h_T|||b-\mathrm{div}\sigma_h|||_{a^{-1},T}\\ &\leq 2\left(|||a\nabla(y_h-\tilde{y})|||_{a^{-1},T} + |||a\nabla\tilde{y}-\sigma_h|||_{a^{-1},T}\right) + \pi^{-1}h_T|||b-\Pi b|||_{a^{-1},T}\\ &\quad + \pi^{-1}h_T|||\Pi b-\mathrm{div}\sigma_h|||_{a^{-1},T}.\end{aligned} \tag{4.49}$$

Let us set $v = \mathrm{div}\sigma_h - \Pi b$. Now we will derive simultaneously, the estimates for $|||v|||_{a^{-1},T}$ and an upper bound for $\eta_{y,T}$. Since the bubble function $\psi_T = 0$ on the boundary ∂T and outside of element T, integration by parts gives

$$\begin{aligned}\mathcal{B}(\tilde{y}-y_h, \psi_T v) &= \int_T a\nabla\tilde{y}\nabla(\psi_T v) - \int_T a\nabla y_h\nabla(\psi_T v)\\ &= \int_T -g\psi_T v + \int_T \mathrm{div}(a\nabla y_h)\psi_T v\\ &= \langle \mathrm{div}\sigma_h - \Pi b, \psi_T v\rangle_{L^2(T)} - \int_T(\mathrm{div}\sigma_h - \mathrm{div}(a\nabla y_h))\psi_T v - \langle b-\Pi b, \psi_T v\rangle_{L^2(T)}\\ &= \langle v, \psi_T v\rangle_{L^2(T)} - \int_T(\mathrm{div}\sigma_h - \mathrm{div}(a\nabla y_h))\psi_T v - \langle b-\Pi b, \psi_T v\rangle_{L^2(T)}.\end{aligned} \tag{4.50}$$

Employing (4.38), (4.40) and (4.20) we derive

$$\begin{aligned}\mathcal{B}(\tilde{y}-y_h,\psi_T v) &= \int_T a\nabla(\tilde{y}-y_h)\nabla(\psi_T v)\\ &\le \|a^{\frac{1}{2}}\nabla(\tilde{y}-y_h)\|_{L^2(T)}\ \|a^{\frac{1}{2}}\nabla(\psi_T v)\|_{L^2(T)}\\ &= |\!|\!| a\nabla(\tilde{y}-y_h)|\!|\!|_{a^{-1},T}\ |\!|\!|\psi_T v|\!|\!|_{a,T}\\ &\le |\!|\!| a\nabla(\tilde{y}-y_h)|\!|\!|_{a^{-1},T}\ |\!|\!|\psi_T v|\!|\!|_{H^1(T)}\\ &\le |\!|\!| a\nabla(\tilde{y}-y_h)|\!|\!|_{a^{-1},T}\ C_a\|\psi_T v\|_{H^1(T)}\\ &\le |\!|\!| a\nabla(\tilde{y}-y_h)|\!|\!|_{a^{-1},T}\ C_a\ h_T^{-1}\|v\|_{L^2(T)}\end{aligned} \tag{4.51}$$

where the constant $C_a = \max(1, a_{\max})$. Similarly we estimate

$$\begin{aligned}\int_T(\mathrm{div}\sigma_h - \mathrm{div}(a\nabla y_h))\psi_T v &= \int_T(\sigma_h - a\nabla y_h)\nabla(\psi_T v)\\ &\le \|a^{-\frac{1}{2}}(\sigma_h - a\nabla y_h)\|_{L^2(T)}\ \|a^{\frac{1}{2}}\nabla(\psi_T v)\|_{L^2(T)}\\ &= |\!|\!|\sigma_h - a\nabla y_h|\!|\!|_{a^{-1},T}\ |\!|\!|\psi_T v|\!|\!|_{a,T}\\ &\le |\!|\!|\sigma_h - a\nabla y_h|\!|\!|_{a^{-1},T}\ C_a\ h_T^{-1}\|v\|_{L^2(T)}.\end{aligned} \tag{4.52}$$

Furthermore, using (4.21) we obtain

$$\begin{aligned}\langle b-\Pi b,\psi_T v\rangle_{L^2(T)} &\le \|b-\Pi b\|_{L^2(T)}\ \|\psi_T v\|_{L^2(T)}\\ &\le \|b-\Pi b\|_{L^2(T)}\ \|v\|_{L^2(T)}.\end{aligned} \tag{4.53}$$

Now using the relation (4.19) in (4.50) and applying the estimates (4.51)- (4.53) we obtain

$$\begin{aligned}c_e\|v\|^2_{L^2(T)} &\le \langle v,\psi_T v\rangle_{L^2(T)}\\ &= \mathcal{B}(\tilde{y}-y_h,\psi_T v) + \langle b-\Pi b,\psi_T v\rangle_{L^2(T)} + \int_T(\mathrm{div}\sigma_h - \mathrm{div}(a\nabla y_h))\psi_T v\\ &\le C_a|\!|\!| a\nabla(\tilde{y}-y_h)|\!|\!|_{a^{-1},T}\ h_T^{-1}\|v\|_{L^2(T)} + \|b-\Pi b\|_{L^2(T)}\ \|v\|_{L^2(T)}\\ &\quad + C_a|\!|\!|\sigma_h - a\nabla y_h|\!|\!|_{a^{-1},T}\ h_T^{-1}\|v\|_{L^2(T)}.\end{aligned} \tag{4.54}$$

On setting $v = \mathrm{div}\sigma_h - \Pi b$ we obtain

$$\begin{aligned}\|\mathrm{div}\sigma_h - \Pi b\|_{L^2(T)} \le C_a c_e^{-1}h_T^{-1}|\!|\!| a\nabla(\tilde{y}-y_h)|\!|\!|_{a^{-1},T} + c_e^{-1}\|b-\Pi b\|_{L^2(T)}\\ + C_a c_e^{-1}h_T^{-1}|\!|\!|\sigma_h - a\nabla y_h|\!|\!|_{a^{-1},T}.\end{aligned}$$

Finally recall that in (4.49), we need the estimate for $|\!|\!|\mathrm{div}\sigma_h - \Pi b|\!|\!|_{a^{-1},T}$. For that purpose we estimate

$$|\!|\!| v|\!|\!|^2_{a^{-1}} = \|a^{-\frac{1}{2}}v\|^2_{L^2} \le a_{\min}\|v\|^2_{L^2}$$

where $a_{\min} = \min_i(u_a(i))$. Hence it holds

$$\begin{aligned}|||\mathrm{div}\sigma_h - \Pi b|||_{a^{-1},T} \leq a_{\min} c_e^{-1} \Big(& C_a h_T^{-1} |||a\nabla(\tilde{y} - y_h)|||_{a^{-1},T} + \|b - \Pi b\|_{L^2(T)} \\ & + C_a h_T^{-1} |||\sigma_h - a\nabla y_h|||_{a^{-1},T} \Big)\end{aligned}$$

which we use in (4.49) to obtain (4.43) with $C = (2 + a_{\min} C_a c_e^{-1} \pi^{-1}), c = a_{\min} \pi^{-1}(1 + c_e^{-1})$. □

Remark 4.9. *Theorem 4.8 and its proof are also valid if a_{max}, a_{min} are replaced with*

$$\begin{aligned}\widetilde{a_{max}} &= \max_i u_h(i) \leq a_{max}, \\ \widetilde{a_{min}} &= \min_i u_h(i) \geq a_{min}\end{aligned}$$

so that the derived estimates remain valid for problems with $u_b = +\infty, u_a = -\infty$.

In a similar manner we compute the residual in the adjoint state. For a given $a \in U_{ad}$, and y_h satisfying (4.36), $p_h \in Y_h$ is the solution of the discrete adjoint equation if and only if

$$\int_\Omega a\nabla p_h \nabla\phi = \langle -g'(y_h), \phi\rangle \quad \forall \phi \in Y_h. \tag{4.55}$$

Theorem 4.10. *Let $p_h \in Y_h \subset H_0^1(\Omega)$, $a \in U_{ad}$ satisfy the discrete equation (4.55). Let $\tau_h \in H(\mathrm{div})$ be given such that*

$$(\mathrm{div}\tau_h,\, 1)_{L^2(T)} = (g'(y_h),\, 1)_{L^2(T)} \quad \textit{for all cells } T \in \mathcal{T}_h.$$

Let us define the cell-wise indicator $\eta_{p,T}$, $T \in \mathcal{T}_h$,

$$\eta_{p,T} := 2|||a\nabla p_h - \tau_h|||_{a^{-1},T} + \pi^{-1} h_T |||g'(y_h) - \mathrm{div}\tau_h|||_{a^{-1},T}.$$

Then it holds

$$\| - \mathrm{div}(a\nabla p_h) + g'(y_h)\|^2_{H^{-1}(\Omega)} \leq a_{max} \sum_{T\in\mathcal{T}_h} \eta_{p,T}^2 =: r_p^2. \tag{4.56}$$

If moreover, $\mathcal{T}_h$ is shape-regular, then it holds

$$\eta_{p,T} \leq C|||a\nabla(\tilde{p} - p_h)|||_{a^{-1},T} + C|||\tau_h - a\nabla\tilde{p}|||_{a^{-1},T} + c\; h_T \|g'(y_h) - \Pi g'(y_h)\|_{L^2(T)}$$

where $\tilde{y}$ is a solution of $-\mathrm{div}(a\nabla\tilde{p}) + g'(y_h) = 0$ and Π denotes the orthogonal L^2-projection onto Y_h. The constants C, c depend only on the spatial dimension m, the shape regularity of the triangulation and the set U_{ad} (via the constants a_{min}, a_{max}).

Proof. The proof is similar to that of Theorem 4.8 and therefore omitted. □

Finally as in the other class of problem, we obtain lower bounds for the error. Note that up to now, the norms $|||\cdot|||_a, |||\cdot|||_{a^{-1}}$ depend on a fixed $a \in U_{ad}$. In the following results, we will derive estimates not involving these weighted norms and choose $a = a_h = \sum_{i=1}^n \chi_i u_h(i)$.

Theorem 4.11. *Let the assumptions of Theorem 4.1 be fulfilled. Let r_y and r_p be computed according to (4.42) and (4.56). Furthermore let $(\bar y, \bar u, \bar p)$ be the local solution of (P) provided by Theorem 4.1 and $\bar a = \sum_{i=1}^n \chi_i \bar u_i$. Then it holds*

$$\sum_{T\in\mathcal{T}_h} r_{y,T} \le \tilde C\Big(\|\bar a - a_h\|_U + \|\bar y - y_h\|_Y + \|a_h\nabla\bar y - \sigma_h\|_{L^2(T)} + \sum_{T\in\mathcal{T}_h} h_T\|(I-\Pi)b\|_{L^2(\Omega)}\Big), \quad (4.57)$$

$$\sum_{T\in\mathcal{T}_h} r_{p,T} \le \tilde C\Big(\|\bar a - a_h\|_U + \|\bar y - y_h\|_Y + \|\bar p - p_h\|_Y + \|a_h\nabla\bar p - \tau_h\|_{L^2(T)} + \sum_{T\in\mathcal{T}_h} h_T\|(I-\Pi)g'(y_h)\|_{L^2(\Omega)}\Big), \quad (4.58)$$

where $\tilde C > 0$ depends only on the spatial dimension m, the shape regularity of the triangulation and the set U_{ad}.

Proof. Let us set $a = a_h$ in the results of Theorem 4.8. We deduce $\sum_{T\in\mathcal{T}_h} r_{y,T} = r_y \le \sum_{T\in\mathcal{T}_h} \eta_{y,T}$ where the error indicator $\eta_{y,T}$ is given by

$$\begin{aligned}\eta_{y,T} &= C|||a_h\nabla(\tilde y - y_h)|||_{a^{-1},T} + C|||\sigma_h - a_h\nabla\tilde y|||_{a^{-1},T} + c\,h_T\|b-\Pi b\|_{L^2(T)}\\ &\le C\Big(|||a_h\nabla(\tilde y-\bar y)|||_{a^{-1},T} + |||a_h\nabla(\bar y - y_h)|||_{a^{-1},T} + |||\sigma_h - a_h\nabla\bar y|||_{a^{-1},T} + |||a_h\nabla(\bar y-\tilde y)|||_{a^{-1},T}\Big)\\ &\quad + c\,h_T\|b-\Pi b\|_{L^2(T)}.\end{aligned} \quad (4.59)$$

Now we will estimate each of the addends in the above. By (4.38), (4.39) and (4.40) it holds

$$|||a_h\nabla(\tilde y-\bar y)|||_{a^{-1},T} = |||\tilde y - \bar y|||_a \le |||\tilde y-\bar y|||_{H^1} \le C_a\|\tilde y - \bar y\|_Y.$$

Similarly we estimate $|||a_h\nabla(\bar y - y_h)|||_{a^{-1},T} \le C_a\|\bar y - y_h\|_Y$. By definition (4.37) it holds

$$|||\sigma_h - a_h\nabla\bar y|||_{a^{-1},T} = \|a^{-\frac12}(\sigma_h - a_h\nabla\bar y)\|_{L^2} \le a_{\min}^{-\frac12}\|\sigma_h - a_h\nabla\bar y\|_{L^2}.$$

It then remains to estimate $\|\tilde y - \bar y\|_Y$. Recall by definition that $-\mathrm{div}(a_h\nabla\tilde y)+b = 0$. The difference $\tilde y - \bar y$ therefore solves

$$-\mathrm{div}(a_h\nabla(\tilde y-\bar y)) = -b + \mathrm{div}(a_h\nabla\bar y).$$

Since $(\bar{y}, \bar{a})$ is a solution of $-\mathrm{div}(\bar{a}\nabla\bar{y}) + b = 0$, it holds

$$-\mathrm{div}(a_h\nabla(\tilde{y}-\bar{y})) = -b + \mathrm{div}(a_h\nabla\bar{y}) = -\mathrm{div}(\bar{a}\nabla\bar{y}) + \mathrm{div}(a_h\nabla\bar{y}) = -\mathrm{div}((\bar{a}-a_h)\nabla\bar{y})$$

from which we obtain

$$\|\tilde{y}-\bar{y}\|_Y \leq \delta^{-1}\|\bar{a}-a_h\|_U\|\nabla\bar{y}\|_{L^2(\Omega)} \leq \delta^{-1}\|\bar{a}-a_h\|_U\|\bar{y}\|_Y. \tag{4.60}$$

The first part of the proof is complete on applying all the obtained estimates in (4.59). The estimate (4.58) can be proved in a similar manner. □

4.3 Adaptivity

In this section, we will compare the performance of adaptive mesh refinement using different strategies to mark elements for refinement. The first one, referred to as *'verified adaptive'*, is implemented as follows: in each step the verification procedure of Chapter 3 is carried out. If it confirms that the assumptions of Theorem 4.1 are satisfied, then the error indicator $\omega_y r_y + \omega_p r_p$ given by (4.2) is used to guide the mesh-refinement. If the requirements of Theorem 4.1 cannot be verified, then a uniform refinement step is carried out. Here, we expect that after a small number of uniform refinement steps the requirements of Theorem 4.1 are confirmed a-posteriori, which coincides with the numerical experiments done earlier in Chapter 3. After these initial uniform refinements steps, we expect that the method proceeds with adaptive steps.

A second strategy, called *'fully adaptive'*, which is frequently used in literature, omits the verification step, and simply uses $\omega_y r_y + \omega_p r_p$ from (4.2) without checking the validity of this bound.

Incorporation of eigenvalue errors

By exploiting the structure of the error matrix $\mathcal{E}$, the errors in the eigenvalue of the Hessians can also be incorporated in the estimate (4.2). Recall (4.2):

$$\|\bar{u}-u_h\|_U \leq \frac{2}{\alpha_h - \|\mathcal{E}\|_2}\left(\omega_y\, r_y + \omega_p\, r_p\right).$$

Let us suppose that α_h is positive, and localizable error bounds of the form

$$\|\mathcal{E}\|_2 \leq r_{\mathcal{E}} \leq \sum_{T\in\mathcal{T}_h} \eta_{\mathcal{E},T} \tag{4.61}$$

are available. If $\alpha_h - r_{\mathcal{E}} > 0$ then we can replace (4.2) with

$$(\alpha_h - r_{\mathcal{E}})\|\bar{u}-u_h\|_U \leq 2\left(\omega_y\, r_y + \omega_p\, r_p\right),$$

which is equivalent to

$$\alpha_h \|\bar{u} - u_h\|_U \leq 2\,(\omega_y\, r_y + \omega_p\, r_p) + \|\bar{u} - u_h\|_U r_{\mathcal{E}}. \tag{4.62}$$

Using the global bound of $\|\bar{u} - u_h\|_U$ provided by Theorem 4.1, the right-hand side of (4.62) is computable and yields a localizable error indicator, which takes error information of the eigenvalues of Hessians of (P) into account.

An obvious choice of $r_{\mathcal{E}}$ in (4.61) is the Frobenious norm of the error matrix $\|\mathcal{E}\|_F$ which is apparent from the relation (2.4). By definition,

$$\|\mathcal{E}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |\mathcal{E}_{ij}|^2.$$

Let us now derive the localized error bound $\eta_{\mathcal{E},T}$ as appeared in (4.61) above. Important here are the quantities $\epsilon_y, \epsilon_p, \epsilon_{z_i}$ given in Lemmas 3.6 and 3.19. We define the corresponding local quantities

$$\begin{aligned}
\epsilon_{y,T} &= \delta^{-1} r_{y,T},\\
\epsilon_{p,T} &= \delta^{-1}(c_{g'}\epsilon_{y,T} + r_{p,T} + c_{E_y}\epsilon_{y,T}\|p_h\|_Y),\\
\epsilon_{z_i,T} &= \delta^{-1}((c_{E_u} + c_{E_y}\|z_{i,h}\|_Y)\epsilon_{y,T} + r_{z_i,T}).
\end{aligned} \tag{4.63}$$

Here the residuals $r_{z_i,T}$ are computed using the method described on page 76. The entries $\mathcal{E}_{ij}$ of the error matrix $\mathcal{E}$ are given by

$$\mathcal{E}_{ij} = \epsilon_{g''_{i,j}} + \epsilon_{E''_{i,j}}$$

where the summands are given in Lemmas 3.21, 3.23 respectively as

$$\epsilon_{g''_{i,j}} = c_{g''} M_{z_i^h} M_{z_j^h} \epsilon_y + \|g''(y_h)\|_{(Y\times Y)^*} \left(M_{z_j^h}\epsilon_{z,i} + \|z_{i,h}\|_Y \epsilon_{z,j}\right), \tag{4.64}$$

$$\epsilon_{E''_{i,j}} := M_{d_{i,j}} c_{E''} \epsilon_y M_p + \|E''(y_h, u_h)\|_{\mathcal{B}(U\times Y, Y^*)} \left(\epsilon_{d_{i,j}} M_p + M_{d_{i,j,h}}\epsilon_p\right). \tag{4.65}$$

Therefore we estimate

$$\begin{aligned}
\|\mathcal{E}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |\mathcal{E}_{ij}|^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} |\epsilon_{g''_{i,j}} + \epsilon_{E''_{i,j}}|^2\\
&\leq 2\sum_{i=1}^{n}\sum_{j=1}^{n} \left(|\epsilon_{g''_{i,j}}|^2 + |\epsilon_{E''_{i,j}}|^2\right)\\
&= 2\left(\|\epsilon_{g''}\|_F^2 + \|\epsilon_{E''}\|_F^2\right).
\end{aligned} \tag{4.66}$$

Next we compute estimates of the terms on the right side of the above expression. By Corollary 3.20, the quantity $M_{z_i^h}$ in (4.64) is related to the localizable quantity $\epsilon_{z,i}$ through $M_{z_i^h} = \epsilon_{z,i} + \|z_{i,h}\|_Y$ which we rewrite as $M_{z_i^h} = \epsilon_{z_i} + \|z_{h_i}\|_Y$ for notational convenience. Let us denote $M_{z_i^h}$ restricted to element T by $M_{z_i^h,T} := \epsilon_{z_i,T} + \|z_{h_i}\|_Y$.

Lemma 4.12. *Let matrix $\epsilon_{g''} = (\epsilon_{g''_{i,j}})$ where $\epsilon_{g''_{i,j}}$ is given by (4.64). Then it holds*

$$\|\epsilon_{g''}\|_F^2 \leq \sum_T \eta_{\mathcal{E},T}^{(1)}$$

where

$$\eta_{\mathcal{E},T}^{(1)} = 2c_{g''}^2 \epsilon_{y,T}^2 \left(\sum_i^n M_{z_i^h,T}^2\right)^2 + 4\|g''(y_h)\|_{(Y\times Y)^*}^2 \sum_i^n \epsilon_{z_i,T}^2 \sum_i^n \left(\|z_{h_i}\|_Y^2 + M_{z_i^h,T}^2\right),\ i = 1,...,n.$$

Proof. From (4.64) we can estimate

$$\epsilon_{g''_{i,j}}^2 \leq 2\left(c_{g''}^2 M_{z_i^h}^2 M_{z_j^h}^2 \epsilon_y^2 + 2\|g''(y_h)\|_{(Y\times Y)^*}^2 (M_{z_j^h}^2 \epsilon_{z,i}^2 + \|z_{i,h}\|_Y^2 \epsilon_{z,j}^2)\right).$$

This implies

$$\begin{aligned}
\|\epsilon_{g''}\|_F^2 &\leq 2c_{g''}^2 \epsilon_y^2 \sum_i^n M_{z_i^h}^2 \sum_j^n M_{z_j^h}^2 + 4\|g''(y_h)\|_{(Y\times Y)^*}^2 \sum_j^n M_{z_j^h}^2 \sum_i^n \epsilon_{z_i}^2 \\
&\quad + 4\|g''(y_h)\|_{(Y\times Y)^*}^2 \sum_i^n \|z_{i,h}\|_Y^2 \sum_j^n \epsilon_{z_j}^2 \\
&\leq \sum_T \left[2c_{g''}^2 \epsilon_{y,T}^2 \left(\sum_i^n M_{z_i^h,T}^2\right)^2 + 4\|g''(y_h)\|_{(Y\times Y)^*}^2 \sum_i^n \epsilon_{z_i,T}^2 \sum_i^n \left(\|z_{i,h}\|_Y^2 + M_{z_i^h,T}^2\right)\right]
\end{aligned}$$

which yields the result. □

Lemma 4.13. *Let matrix $\epsilon_{E''} = (\epsilon_{E''_{i,j}})$ where $\epsilon_{E''_{i,j}}$ is given by (4.65). Then it holds*

$$\|\epsilon_{E''}\|_F^2 \leq \sum_T \eta_{\mathcal{E},T}^{(2)}$$

where

$$\eta_{\mathcal{E},T}^{(2)} = 2c_{E''}^2 M_p^2 \|M_d\|_{F,T}^2\ \epsilon_{y,T}^2 + 4\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}^2 \left(M_p^2\ \|\epsilon_d\|_{F,T}^2 + \|M_{d,h}\|_F^2\ \epsilon_{p,T}^2\right).$$

The quantities $\|M_d\|_{F,T}$, $\|\epsilon_d\|_{F,T}$ are specified in the course of the proof.

Proof. Similar computations as in the proof of Lemma 4.12 using (4.65) reveals that

$$\epsilon_{E''_{i,j}}^2 \leq 2\left(M_{d_{i,j}}^2 c_{E''}^2 M_p^2 \epsilon_y^2 + 2\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}^2 (\epsilon_{d_{i,j}}^2 M_p^2 + M_{d_{i,j,h}}^2 \epsilon_p^2)\right)$$

from which we obtain

$$\|\epsilon_{E''}\|_F^2 \leq 2c_{E''}^2 M_p^2 \ \|M_d\|_F^2 \ \epsilon_y^2 + 4\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}^2 \left(M_p^2 \ \|\epsilon_d\|_F^2 + \|M_{d,h}\|_F^2 \ \epsilon_p^2\right). \tag{4.67}$$

Now note that matrix $\epsilon_d = (\epsilon_{d_{i,j}})$ also depends on the localizable quantity ϵ_{z_j} through (3.80)

$$\epsilon_{d_{i,j}} = \epsilon_{z_i}(1 + \|z_{j,h}\|_Y) + \epsilon_{z_j}(1 + M_{z_i^h}).$$

Therefore we derive

$$\begin{aligned}
\|\epsilon_d\|_F^2 &\leq 2\sum_i \epsilon_{z_i}^2 \sum_j (1+\|z_{h_j}\|_Y)^2 + \sum_j \epsilon_{z_j}^2 \sum_i (1+M_{z_i^h})^2 \\
&= 2\sum_i \epsilon_{z_i}^2 \sum_i \left((1+\|z_{h_j}\|_Y)^2 + (1+M_{z_i^h})^2\right) \\
&= \sum_T \left[2\sum_i \epsilon_{z_i,T}^2 \sum_i \left((1+\|z_{j,h}\|_Y)^2 + (1+M_{z_i^h,T})^2\right)\right] \\
&=: \sum_T \|\epsilon_d\|_{F,T}^2.
\end{aligned} \tag{4.68}$$

In a similar manner, through (3.80) we also derive

$$\|M_d\|_{F,T}^2 = 4\sum_i (1+M_{z_i^h,T}^2) \sum_j (1+M_{z_j^h,T}^2).$$

We remark that the matrix $M_{d,h}$ appearing in (4.67) above, and which is as well given in (3.80), is independent of the localizable quantities $\epsilon_y, \epsilon_p, \epsilon_z$. Hence no further computation is required to estimate its Frobenious norm.

Altogether from (4.67) we obtain

$$\|\epsilon_{E''}\|_F^2 \leq \sum_T \left[2c_{E''}^2 M_p^2 \|M_d\|_{F,T}^2 \ \epsilon_{y,T}^2 + 4\|E''(y_h,u_h)\|_{\mathcal{B}(U\times Y,Y^*)}^2 \left(M_p^2 \ \|\epsilon_d\|_{F,T}^2 + \|M_{d,h}\|_F^2 \ \epsilon_{p,T}^2\right)\right],$$

which is the desired result. □

Let us now combine the results of the two preceding lemmas to obtain the estimate for $\eta_{\mathcal{E},T}$.

Lemma 4.14. *Let the quantities $\epsilon_{y,T}, \epsilon_{p,T}, \epsilon_{z_i,T}$ be given by (4.63). Furthermore let $\eta_{\mathcal{E},T}^{(1)}, \eta_{\mathcal{E},T}^{(2)}$ be as defined in Lemma 4.12 and Lemma 4.13 respectively. Then it holds*

$$\|\mathcal{E}\|_F \leq \sum_T \eta_{\mathcal{E},T}$$

where

$$\eta_{\mathcal{E},T}^2 = 2(\eta_{\mathcal{E},T}^{(1)} + \eta_{\mathcal{E},T}^{(2)}).$$

Proof. The claim follows on applying the results of Lemma 4.12 and Lemma 4.13 on (4.66). □

Now altogether, using the result of Lemma 4.14 in (4.62) yields a localizable error bound for the control:

$$\begin{aligned}\|u - u_h\|_U &\leq \frac{2}{\alpha_h - r_{\mathcal{E}}}(\omega_y r_y + \omega_p r_p) \\ &= \sum_{T \in \mathcal{T}_h} \eta_T\end{aligned}$$

where the local error indicator η_T is given by

$$\eta_T = \frac{2}{\alpha_h - \eta_{\mathcal{E},T}}(\omega_y \; r_{y,T} + \omega_p \; r_{p,T}).$$

4.4 Numerical results

Let us report about the outcome of the adaptive methods described at the beginning of Section 4.3. We considered selected examples, taken from Chapter 3. The functional J was chosen as

$$J(y,u) := \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\kappa}{2}\|u\|^2_{\mathbb{R}^n}.$$

Example 1

Here the nonlinear mapping E represents a semilinear elliptic equation given by

$$E(y,u) := -\Delta y + \sum_{k=1}^{n} u_k d_k(y) - b, \tag{4.69}$$

where the functions d_k are chosen as $d_1(y) = 1$, $d_j(y) = y|y|^{j-2}$ for $j = 2 \dots n$. This example is motivated by parameter identification: given a state y_d and source term b, find the set of coefficients u such that the resulting solution y of $E(y,u) = 0$ is as close as possible to y_d.

In order to make the operator E strongly monotone, we require positivity of the coefficients u_k, i.e. we set $U_{ad} = \{u \in \mathbb{R}^n : \; u_k \geq 0 \quad \forall k = 1 \dots n\}$. For the computations we used the following data: the source term $b = 10.0001$ and

$$\Omega = (0,1)^2, \; u_a = 0, \; u_b = 0.5, \; \kappa = 10^{-2}, \; y_d(x_1,x_2) = 0.5\sin(2\pi x_1 x_2), \; n = 4.$$

Let us remark, that the function d_3 is not of class C^2 globally. However, since b is non-negative, every solution y of (4.69) to $u \in U_{ad}$ will be non-negative. For non-negative functions y it holds $d_3(y) = y^2$, which is C^2, so the assumptions on E are satisfied. See also the discussion in Section 3.4.3.

We employed a discretization scheme described in Section 3.1.2. After the resulting nonlinear optimization problem is solved, the error indicators according to the chosen strategy are

computed. For an adaptive refinement, a subset $\tilde{\mathcal{T}} \subset \mathcal{T}$ of elements T with large local error contributions η_T that satisfies $\sum_{T\in\tilde{\mathcal{T}}} \eta_T^2 \geq \theta^2 \sum_{T\in\mathcal{T}} \eta_T^2$ with $\theta = 0.8$ were selected for refinement.

Let us now report on the outcome of the different adaptive strategies as described in Section 4.3. For all the methods, we compare the residual norms as given in (4.5), i.e. with the notation of that section

$$\epsilon_{\text{residual}} := \omega_y r_y + \omega_p r_p.$$

Moreover, we employed the verification procedure of Theorem 4.1 and report about the upper error bound

$$\epsilon_{\text{bound}} := \frac{2}{\alpha_h - \|\mathcal{E}\|_2} \left(\omega_y\, r_y + \omega_p\, r_p\right).$$

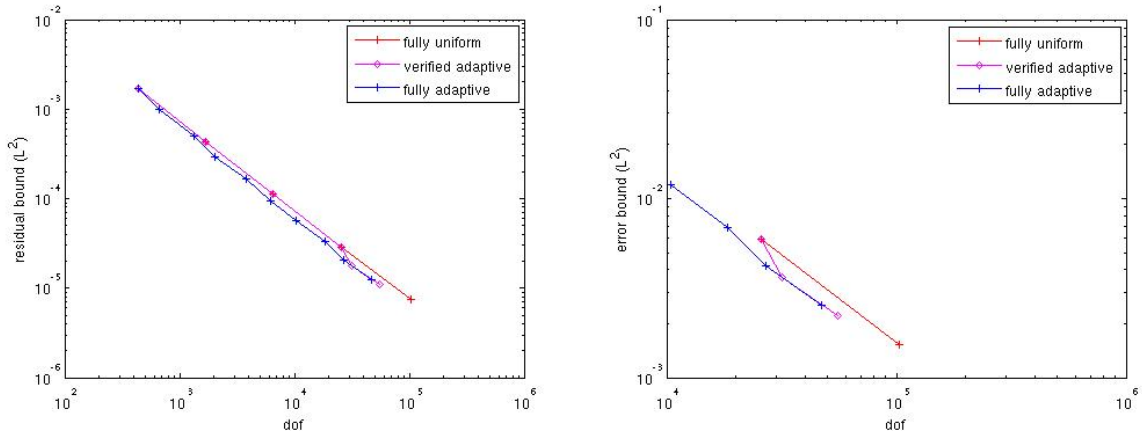

Figure 4.1: (a) Upper bound of residuals versus number of unknowns, (b) Verified error bound versus number of unknowns

As can be expected, the assumptions of Theorem 4.1 are only fulfilled on a sufficiently fine discretization. This is reflected by our numerical results.

Plots of $\epsilon_{\text{residual}}$ and ϵ_{bound} versus the number of degrees of freedom can be seen in Figure 4.1. For reference, we provided the numerical values in Tables 4.1 and 4.2. In the tables, L refers to refinement level, where $L = 0$ is the initial mesh, which is the same for all the different adaptive methods. Moreover, *dof* denotes the number of degrees of freedom.

Let us comment on the observed behavior of the verified adaptive methods for this problem. The conditions of Theorem 4.1 are fulfilled for the first time after three uniform refinement steps. The fourth and all further refinement levels were reached by using adaptive refinement according to the error indicator based on (4.5). The fully adaptive scheme, which refines according to the residuals in the optimality system, obtains verified error bounds as of level 7. After the verified adaptive methods actually start adaptive refinement, they quickly reach the same ratio of error bound versus number of degrees of freedom as the full adaptive method. That means, the early (unverified) adaptive refinements of the full adaptive methods does not seem to give this method an advantage over the verified method. The same observation also applies to the residual error versus number of degrees of freedom ratio, as can be seen in Figure 4.1 and Table 4.2.

fully uniform			*verified adaptive*			*fully adaptive*		
L	$\#\ dof$	ϵ_{error}	L	$\#\ dof$	ϵ_{error}	L	$\#\ dof$	ϵ_{error}
0	441	–	1	1681	–	5	6177	–
1	1681	–	2	6561	–	6	10341	–
2	6561	–	3	25921	$7.6226 \cdot 10^{-3}$	7	18427	$9.0724 \cdot 10^{-3}$
3	25921	$7.6226 \cdot 10^{-3}$	4	31491	$4.8745 \cdot 10^{-3}$	8	27155	$5.6376 \cdot 10^{-3}$
4	103041	$1.9183 \cdot 10^{-3}$	5	55061	$2.8728 \cdot 10^{-3}$	9	46979	$3.3167 \cdot 10^{-3}$

Table 4.1: Error bound estimates for Example 1

fully uniform			*verified adaptive*			*fully adaptive*		
L	$\#dof$	$\epsilon_{\text{residual}}$	L	$\#dof$	$\epsilon_{\text{residual}}$	L	$\#dof$	$\epsilon_{\text{residual}}$
1	1681	$5.4976 \cdot 10^{-4}$	1	1681	$5.4976 \cdot 10^{-4}$	2	1311	$6.6336 \cdot 10^{-4}$
2	6561	$1.4181 \cdot 10^{-4}$	2	6561	$1.4181 \cdot 10^{-4}$	5	6193	$1.2537 \cdot 10^{-4}$
3	25921	$3.6666 \cdot 10^{-5}$	3	25921	$3.6669 \cdot 10^{-5}$	7	18427	$4.3361 \cdot 10^{-5}$
			4	31491	$2.3746 \cdot 10^{-5}$	8	27155	$2.7370 \cdot 10^{-5}$
4	103041	$9.4714 \cdot 10^{-6}$	5	55061	$1.4121 \cdot 10^{-5}$	9	46979	$1.6271 \cdot 10^{-5}$

Table 4.2: Residual error bound estimates for Example 1

Example 2

The elliptic operator is given here by

$$E(y,u) := -\text{div}(a\nabla y) - b$$

with $a = \sum_{i=1}^{n} \chi_{\Omega_i} u_i$ and $b \in L^2(\Omega)$ is a source term. The computational domain is chosen as $\Omega = (0,1)^2$ which is further subdivided into four sub-domains

$$\Omega_1 = (0,0.5)^2,\ \Omega_2 = (0,0.5)\times(0.5,1),\ \Omega_3 = (0.5,1)\times(0,0.5),\ \Omega_4 = (0.5,1)^2.$$

The computational data were chosen as

$$u_a = 0.1,\ u_b = 0.9,\ \kappa = 9\times 10^{-1},\ y_d(x_1,x_2) = x_1x_2,\ n = 4,\ b = 10.0001.$$

We impose strict positivity of the coefficients u_k, i.e. we set
$U_{ad} = \{u \in \mathbb{R}^n :\ u_k > 0 \quad \forall k = 1\ldots n\}$.

As in the first example, the same discretization scheme is used. The marking strategy for elements with large error contributions is also the same with tolerance $\theta = 0.8$. Concerning the outcome of the different adaptive strategies, a similar behavior as in the preceding example is observed. However in this case, the fulfillment of Theorem 4.1 is obtained by the verified adaptive method only on a relatively finer discretization. The results of the numerical experiments are presented in Tables 4.3 and 4.4.

fully uniform			*verified adaptive*			*fully adaptive*		
L	$\#\ dof$	ϵ_{error}	L	$\#\ dof$	ϵ_{error}	L	$\#\ dof$	ϵ_{error}
0	25921	–	0	25921	–	1	28207	–
1	103041	–	1	103041	–	4	113761	–
2	410881	$1.5390 \cdot 10^{-3}$	2	410881	$1.5390 \cdot 10^{-3}$	6	331509	$1.2340 \cdot 10^{-3}$
			3	423581	$9.5811 \cdot 10^{-4}$	7	469095	$7.6125 \cdot 10^{-4}$
3	1640961	$3.9287 \cdot 10^{-4}$	4	645099	$5.9594 \cdot 10^{-4}$	8	873451	$4.7367 \cdot 10^{-4}$

Table 4.3: Error bound estimates for Example 2

fully uniform			*verified adaptive*			*fully adaptive*		
L	$\#dof$	$\epsilon_{\text{residual}}$	L	$\#dof$	$\epsilon_{\text{residual}}$	L	$\#\ dof$	ϵ_{error}
0	25921	$1.3579 \cdot 10^{-2}$	0	25921	$1.3579 \cdot 10^{-2}$	1	28207	$8.5488 \cdot 10^{-3}$
1	103041	$3.6125 \cdot 10^{-3}$	1	103041	$3.6125 \cdot 10^{-3}$	4	113761	$1.9972 \cdot 10^{-3}$
2	410881	$9.5549 \cdot 10^{-4}$	2	410881	$9.5549 \cdot 10^{-4}$	5	211109	$1.2490 \cdot 10^{-3}$
			3	423581	$6.0515 \cdot 10^{-4}$	6	331509	$7.7131 \cdot 10^{-4}$
3	1640961	$2.5203 \cdot 10^{-4}$	4	645099	$3.8048 \cdot 10^{-4}$	8	873451	$3.0344 \cdot 10^{-4}$

Table 4.4: Residual error bound estimates for Example 2

Chapter 5

5 Eigenvalue approximation in infinite dimensional spaces

In this chapter the control space is now the infinite dimensional space $U = L^2(\Omega)$. Our goal is to discuss possible challenges in extending the verification results of Chapter 3 to this infinite dimensional control space. The main result here is the H^2-regularity of the eigenfunctions of the associated eigenvalue problem.

Let Y be a real Banach space and the regularization parameter $\alpha > 0$. We will consider the following optimization problem.

Problem 1. *Minimize the functional*

$$J(y,u) := g(y) + \frac{\alpha}{2}\|u\|_U^2 \tag{5.1}$$

subject to the constraint

$$\begin{aligned} Ay + d(y) &= u \quad in \quad \Omega, \\ y &= 0 \quad on \quad \partial\Omega. \end{aligned} \tag{5.2}$$

We assume that the functional $g : Y \mapsto \mathbb{R}$ and function $d : Y \mapsto Y^*$ are twice continuously Fréchet differentiable. Furthermore d is assumed to be monotone. The space Y^* as before denotes the dual space of the state space Y. The domain $\Omega \subseteq \mathbb{R}^d, d = 2, 3$ is assumed to be of type $\mathcal{C}^{1,1}$ and the operator A is self-adjoint and uniformly elliptic in Ω with smooth coefficients. For the rest of the chapter, we assume $u \mapsto y$ is compact as a mapping from $U = L^2(\Omega)$ to $Y = H^1(\Omega)$.

Following the approach in Chapter 3, we proceed by writing the second-order sufficient condition as a generalized eigenvalue problem.

5.1 Second-order sufficient condition as eigenvalue problem

Let $(\bar{y}, \bar{u})$ be an optimal solution of problem (5.1)-(5.2). We define the Lagrange functional

$$\mathcal{L}(y,u,p) = g(y) + \frac{\alpha}{2}\|u\|_U^2 + \langle Ay + d(y) - u,\, p\rangle_{Y^*,Y}. \tag{5.3}$$

The second-order sufficient condition is then given as: there exists $\delta > 0$ such that

$$\mathcal{L}''(\bar{y}, \bar{u}, \bar{p})[z, v]^2 \geq \delta \|v\|_U^2 \ \forall v \in U \tag{5.4}$$

where z is the solution of a linearized state equation

$$\begin{aligned} Az + d'(\bar{y})z - v = 0 \quad &\text{in} \quad \Omega, \\ z = 0 \quad &\text{on} \quad \partial\Omega. \end{aligned} \tag{5.5}$$

Furthermore the adjoint state $\bar{p}$ is the unique solution of

$$\begin{aligned} A^*\bar{p} + d'(\bar{y})\bar{p} = -g'(\bar{y}) \ \ &\text{in} \ \ \Omega, \\ \bar{p} = 0 \quad &\text{on} \ \ \partial\Omega. \end{aligned} \tag{5.6}$$

The above second-order sufficient condition can equivalently be written as a generalized eigenvalue problem

$$\begin{pmatrix} \mathcal{L}_{uu}(\bar{y}, \bar{u}, \bar{p}) & \mathcal{L}_{uy}(\bar{y}, \bar{u}, \bar{p}) & \mathcal{L}_{up}(\bar{y}, \bar{u}, \bar{p}) \\ \mathcal{L}_{yu}(\bar{y}, \bar{u}, \bar{p}) & \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p}) & \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p}) \\ \mathcal{L}_{pu}(\bar{y}, \bar{u}, \bar{p}) & \mathcal{L}_{py}(\bar{y}, \bar{u}, \bar{p}) & 0 \end{pmatrix} \begin{pmatrix} v \\ z \\ q \end{pmatrix} = \lambda \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v \\ z \\ q \end{pmatrix}$$

where the eigenvalue $\lambda \in \mathbb{R}$ and $v \in U, z, q \in Y$ are the associated eigenfunctions to the control, state and the adjoint state respectively. Using (5.3) the derivatives of $\mathcal{L}$ are computed as

$$\mathcal{L}_{uu}(\bar{y}, \bar{u}, \bar{p}) = \alpha I, \quad \mathcal{L}_{uy}(\bar{y}, \bar{u}, \bar{p}) = \mathcal{L}_{yu}(\bar{y}, \bar{u}, \bar{p}) = 0, \quad \mathcal{L}_{up}(\bar{y}, \bar{u}, \bar{p}) = \mathcal{L}_{pu}(\bar{y}, \bar{u}, \bar{p}) = -I,$$

$$\mathcal{L}_{py}(\bar{y}, \bar{u}, \bar{p}) = A + d'(\bar{y}) =: \mathcal{L}_{py}, \quad \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p}) = g''(\bar{y}) + d''(\bar{y})\bar{p} =: \mathcal{L}_{yy}, \quad \mathcal{L}_{yp} = \mathcal{L}_{py}^*.$$

The eigenvalue problem is then: find $\lambda \in \mathbb{R}$ and $(v, z, q) \neq 0$ such that

$$\begin{pmatrix} \alpha I & 0 & -I \\ 0 & \mathcal{L}_{yy} & \mathcal{L}_{yp} \\ -I & \mathcal{L}_{py} & 0 \end{pmatrix} \begin{pmatrix} v \\ z \\ q \end{pmatrix} = \lambda \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v \\ z \\ q \end{pmatrix}. \tag{5.7}$$

The system (5.7) is equivalent to the set of equations

$$\alpha v - q = \lambda v, \tag{5.8}$$

$$A^*q + d'(\bar{y})q = -(d''(\bar{y})\bar{p} + g''(\bar{y}))z, \tag{5.9}$$

$$Az + d'(\bar{y})z = v \tag{5.10}$$

which can also be viewed as an eigenvalue problem associated to the second-order sufficient condition for problem (5.1)-(5.2).

Typical of eigenvalue problems for elliptic partial differential equations, system (5.7) is sometimes written in the operator form

$$T\mu = \lambda B\mu \tag{5.11}$$

where

$$T = \begin{pmatrix} \alpha I & 0 & -I \\ 0 & \mathcal{L}_{yy} & \mathcal{L}_{yp} \\ -I & \mathcal{L}_{py} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mu = (v, z, q)^T.$$

Observe that operator T is self-adjoint as $\mathcal{L}_{yp} = \mathcal{L}_{py}^*$. The major problem in the eigenvalue analysis of problem (5.11) arises due to the fact that both operators T and B are indefinite as well as any of their linear combinations. To the author's knowledge most of the existing theories on eigenvalue analysis of problems of the form (5.11) rely on the assumption of positive definiteness of both T and B or at least that of operator B see e.g. [38]. For the problem (5.1)-(5.2), this is unfortunately not the case as both B and T are indefinite. However, since operator T is self-adjoint we know that T possess a set of real eigenvalues [21, Section 9.1].

We can also write the system (5.8)-(5.10) as a standard eigenvalue problem. For that purpose let us define $S := (A + d'(\bar{y}))^{-1}$ and its adjoint $S^* := (A^* + d'(\bar{y})^*)^{-1}$. Then (5.8)-(5.10) can be written in the form

$$Tv = \lambda v$$

where

$$T = \alpha I + S^*(g''(\bar{y}) + d''(\bar{y})\bar{p})S. \tag{5.12}$$

By the compactness assumption on the mapping $u \mapsto y$, the solution operator S of the linearized equation is also compact. Hence, T is a scalar multiple of identity plus a compact operator. The following characterization of the eigenvalues of compact, self-adjoint operator is therefore applicable to the compact part $S^*(g''(\bar{y}) + d''(\bar{y})\bar{p})S$.

Theorem 5.1. *Let $T : H \to H$ be a self-adjoint and compact operator on a Hilbert space H. Then T has a finite or infinite sequence $\{\lambda_j\}_{j=1}^N, N \leq \infty$ of real eigenvalues $\lambda_j \neq 0$, and corresponding orthonormal sequence $\{v_j\}_{j=1}^N$ in H such that*

$$Tv_j = \lambda_j v_j$$

for all $1 \leq j \leq N$. If $N = \infty$ then

$$\lim_{j\to\infty} \lambda_j = 0$$

and 0 *is the only accumulation point of the eigenvalues* $\{\lambda_j\}$.

A proof of this theorem can be found in [21].

Let us now turn our attention to the eigenfunctions. Essential in the approximation theory for self-adjoint operators in Hilbert spaces is the regularity of eigenfunctions. Therefore in what follows, we present regularity results for the eigenfunctions of problem (5.7). For that purpose, we will switch back to consider the second-order sufficient condition as the system of equations (5.8)-(5.10).

5.2 Regularity of eigenfunctions

We begin the quest into the regularity result of the eigenfunctions with the following a-priori estimates for the state and the adjoint state variables.

Theorem 5.2. *Let the assumptions on the model problem 1 hold. Let* y *be a solution of* (5.2). *If* $u \in L^2(\Omega)$, *we have* $y \in H^1(\Omega)$ *and it holds*

$$\|y\|_{H^1(\Omega)} \leq c_A \|u\|_{L^2(\Omega)}.$$

If additionally the boundary $\partial\Omega$ *is of class* $\mathcal{C}^2$, *then* $y \in H^2(\Omega)$.

The result is standard. A proof can be found in e.g. [23, Section 8.4].

Similarly for the adjoint state we have

Theorem 5.3. *Let the assumptions of Theorem 5.2 hold and let* p *solves* (5.6). *If* $g'(\bar{y}) \in L^2(\Omega)$ *then we have* $p \in H^1(\Omega)$ *and it holds*

$$\|p\|_{H^1(\Omega)} \leq c_A \|g'(\bar{y})\|_{L^2(\Omega)}.$$

Furthermore if the boundary $\partial\Omega$ *is of class* $\mathcal{C}^2$, *then* $p \in H^2(\Omega)$.

Now using the above estimates, we obtain the following result.

Theorem 5.4. *Let the eigenfunctions* v, z, q *be defined by the system* (5.8)-(5.10) *and let the result of Theorem 5.3 holds. If* $v \in L^2(\Omega)$ *and* $\lambda \neq \alpha$, *then the eigenfunction* v *is* H^1-*regular.*

Proof. The eigenfunction z is H^1 regular by the application of Theorem 5.3 on (5.10). Furthermore for every $v \in L^2(\Omega)$ it holds $\|z\|_{H^1(\Omega)} \leq c_A \|v\|_{L^2(\Omega)}$. Hence from (5.9), using Theorem 5.3 again we obtain $q \in H^1(\Omega)$ and it holds

$$\begin{aligned}\|q\|_{H^1(\Omega)} &\leq c_A \Big(\|d''(\bar{y})\|_{\mathcal{L}(Y\times Y, Y^*)} \|p\|_{H^1(\Omega)} + \|g''(\bar{y})\|_{(Y\times Y)^*} \Big) \|z\|_{H^1(\Omega)} \\ &\leq c_A \Big(\|d''(\bar{y})\|_{\mathcal{L}(Y\times Y, Y^*)} c_A \|g'(\bar{y})\|_{L^2(\Omega)} + \|g''(\bar{y})\|_{(Y\times Y)^*} \Big) c_A \|v\|_{L^2(\Omega)}.\end{aligned}$$

Lastly as $\lambda \neq \alpha$ we obtain from (5.8)

$$(\alpha - \lambda) v = q$$

that $v \in H^1(\Omega)$. This completes the proof. □

If the boundary $\partial\Omega$ enjoys additional smoothness, we obtain higher regularity for the eigenfunction v.

Theorem 5.5. *Let the boundary $\partial\Omega$ be of class $\mathcal{C}^2$ and let the result of Theorem 5.3 holds. Furthermore let $v \in L^2(\Omega)$ and suppose $\lambda \neq \alpha$. Then we have $v \in H^2(\Omega)$.*

Proof. Using the second result of Theorem 5.3, we obtain from (5.10) $z \in H^2(\Omega)$. The rest of the proof is then similar to that of Theorem 5.4 and therefore omitted. □

Remark 5.6. *The above nice result has been obtained under the restriction $\lambda \neq \alpha$. The question on mind now is whether the case $\lambda = \alpha$ is even possible? If this is not the case, then the regularity result $v \in H^2(\Omega)$ will be valid for every eigenvalue $\lambda \in \mathbb{R}$.*

To answer the above question, let us have a look again at the eigenvalue system (5.7). Suppose $\lambda = \alpha$ is an eigenvalue corresponding to an eigenfunction $(v, z, q) \neq 0$. Then with $\lambda = \alpha$ in (5.7) we obtain

$$\begin{pmatrix} 0 & 0 & -I \\ 0 & \mathcal{L}_{yy} & \mathcal{L}_{yp} \\ -I & \mathcal{L}_{py} & 0 \end{pmatrix} \begin{pmatrix} v \\ z \\ q \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Solving this system successively gives a contradiction $(v, z, q) = 0$, showing that $\lambda = \alpha$ can not be an eigenvalue. Hence the regularity result for the eigenfunction v is indeed valid for every $\lambda \in \mathbb{R}$.

To obtain a nontrivial answer to the question posed in Remark 5.6, let us consider a slightly different problem where the control now acts only on a subdomain of the domain. We will again examine the regularity of the associated eigenfunction in this case.

Problem 2. *Let subdomain Ω' be an open subset of Ω and let $\chi_{\Omega'}$ denotes the characteristic function of Ω'. Consider the problem: Minimize (5.1) subject to the state equation*

$$\begin{aligned} Ay + d(y) &= \chi_{\Omega'} u \quad \text{in} \quad \Omega, \\ y &= 0 \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{5.13}$$

Let us define operator $B : L^2(\Omega') \to H^1(\Omega)$ such that $\langle Bu, \phi \rangle = -\int_\Omega u\phi, \ \forall \phi \in H_0^1(\Omega)$. In this case the second-order condition (5.4) is equivalent to the eigenvalue problem: find $\lambda \in \mathbb{R}$ such that

$$\begin{pmatrix} \alpha I & 0 & B^* \\ 0 & \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p}) & \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p}) \\ B & \mathcal{L}_{py}(\bar{y}, \bar{u}, \bar{p}) & 0 \end{pmatrix} \begin{pmatrix} v \\ z \\ q \end{pmatrix} = \lambda \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v \\ z \\ q \end{pmatrix} \tag{5.14}$$

holds for $(v, z, q)^T \neq 0$. The star notation in the above denotes the adjoint of an operator. Due to the structure of (5.13), we obtain only interior regularity for the eigenfunction v.

Lemma 5.7. *Let $v \in L^2(\Omega)$ and $\lambda \neq \alpha$. Then for any open subset $\Omega' \subset \Omega$, we have $v \in H^2(\Omega')$.*

Proof. As in the proof Theorem 5.4, for $v \in L^2(\Omega)$ we obtain $q \in H^1(\Omega)$ as a solution of

$$A^*q + d'(\bar{y})q = -(g''(\bar{y}) + d''(\bar{y})\bar{p})z.$$

Then the interior regularity result (see [22, Section 6.3.1, Theorem 2]) gives $q \in H^2(\Omega')$ and it holds

$$\|q\|_{H^2(\Omega')} \leq c_A\Big(\|g''(\bar{y})\|_{(Y\times Y)^*} + \|d''(\bar{y})\|_{\mathcal{L}(Y\times Y, Y^*)}c_A\|g'(\bar{y})\|_{L^2(\Omega)}\Big)c_A\|v\|_{L^2(\Omega)}.$$

Lastly, since $\lambda \neq \alpha$ the regularity $v \in H^2(\Omega')$ is obtained through

$$(\lambda - \alpha)v = B^*q.$$

The somewhat technical proof of the interior regularity can be found in [22, Section 6.3.1]. □

Let us now argue that the case $\lambda \neq \alpha$ does not occur for the present problem. We need the following. Let us define a different Lagrange function

$$\mathcal{L}_0(y, u, p) = g(y) + \langle Ay + d(y) + Bu,\, p\rangle_{Y^*,Y}.$$

We then assume

Assumption 11. *Let $(\bar{y}, \bar{u})$ be a solution of Problem 2 with associated adjoint $\bar{p}$. There exists $\delta > 0$ such that*

$$\mathcal{L}_0''(\bar{y}, \bar{u}, \bar{p})[z, v] \geq \delta\|z\|_Y^2 \quad \forall v \in U \tag{5.15}$$

where z is the solution of the linearized equation

$$\begin{aligned} Az + d'(\bar{y})z + Bv = 0 \quad &\text{in} \quad \Omega, \\ z = 0 \quad &\text{on} \quad \partial\Omega. \end{aligned} \tag{5.16}$$

The above assumption corresponds to the second-order sufficient conditions for bang-bang control problems in the absence of quadratic control term in the objective functional, i.e. the case $\alpha = 0$ (see [13]). Due to the structure of $\mathcal{L}_0$, (5.15) is the same as

$$\mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p})[z]^2 \geq \delta\|z\|_Y.$$

Lemma 5.8. *Let Assumption 11 hold. Then the regularity result of Lemma 5.7 holds for all eigenvalues $\lambda \in \mathbb{R}$.*

Proof. Suppose $\lambda = \alpha$ is an eigenvalue corresponding to an eigenfunction $(v, z, q) \neq 0$. Then from (5.14) it follows that

$$\begin{pmatrix} 0 & 0 & B^* \\ 0 & \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p}) & \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p}) \\ B & \mathcal{L}_{py}(\bar{y}, \bar{u}, \bar{p}) & 0 \end{pmatrix} \begin{pmatrix} v \\ z \\ q \end{pmatrix} = 0$$

which yields the system

$$\begin{pmatrix} B^* q \\ \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p})z + \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p})q \\ Bv + \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p})q \end{pmatrix} = 0. \tag{5.17}$$

Pre-multiplying the above equation by $(v,\ z,\ -q)$ gives

$$\begin{aligned} 0 &= (v,\ z,\ -q) \begin{pmatrix} B^* q \\ \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p})z + \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p})q \\ Bv + \mathcal{L}_{py}(\bar{y}, \bar{u}, \bar{p})z \end{pmatrix} \\ &= \langle v, B^* q \rangle + \langle \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p})z, z \rangle + \langle \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p})q, z \rangle - \langle Bv, q \rangle - \langle \mathcal{L}_{py}(\bar{y}, \bar{u}, \bar{p})z, q \rangle \\ &= \langle v, B^* q \rangle + \langle \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p})z, z \rangle + \langle \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p})q, z \rangle - \langle Bv, q \rangle - \langle \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p})^* z, q \rangle \\ &= \langle Bv, q \rangle + \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p})[z]^2 + \langle \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p})q, z \rangle - \langle Bv, q \rangle - \langle \mathcal{L}_{yp}(\bar{y}, \bar{u}, \bar{p})q, z \rangle \\ &= \mathcal{L}_{yy}(\bar{y}, \bar{u}, \bar{p})[z]^2 \\ &\geq \delta \|z\|_Y^2 \end{aligned}$$

by assumption 11. This implies $z = 0$ on Ω. Since the operator $\mathcal{L}_{yp} = A^* + d'(\bar{y})$ is boundedly invertible (cf. Proposition 2.7), then plugging $z = 0$ into the second row of (5.17) yields $q = 0$. Furthermore as operator B is injective as a mapping from the control space $U \to U$, using $q = 0$ in the last row of (5.17) we obtain $Bv = 0$ which then yields $v = 0$. Hence, we conclude as before that $\lambda = \alpha$ cannot be an eigenvalue of (5.14). □

Verification of infinite dimensional SSC versus the results of Chapter 3

The result of Lemma 5.5 is an essential preliminary in extending our verification result to infinite dimensional SSC. With the regularity of the eigenfunctions, approximation of the eigenfunctions by functions from the finite dimensional space can be computed. However, further research in deriving methods for the verification of infinite dimensional SSC and generalizing the results of Chapter 3 is still ongoing. Nevertheless to conclude this chapter, let us mention some obvious difficulties in extending the results of Chapter 3 to problems with infinite dimensional control space.

Due to the indefiniteness of matrices T and B in (5.11), the Courant-Fischer min-max repre-

sentation [16, Theorem 5.2] of eigenvalues, which was the basis of the proof of Theorem 3.25 is no longer valid. This arises due to the fact that it is possible for the denominator $u^T Bu$ of the corresponding Rayleigh quotient to vanish. Hence, we no longer have eigenvalue bound of the form $|\lambda - \lambda_h| \leq \|T - T_h\|$ which was the case in Theorem 3.25. However a somewhat similar spectrum perturbation result is available for operators in Hilbert spaces.

Theorem 5.9. *[35, Theorem 4.10, p 291] Let T be a self-adjoint operator on Hilbert space H. Let a self-adjoint operator $A \in \mathcal{B}(H)$, the set of bounded operators on H. Then $S = T + A$ is self-adjoint and*

$$\operatorname{dist}(\Sigma(S), \Sigma(T) \leq \|A\|_{\mathcal{L}(H)},$$

that is

$$\sup_{\lambda \in \Sigma(S)} \operatorname{dist}(\lambda, \Sigma(T) \leq \|A\|_{\mathcal{L}(H)}, \quad \sup_{\lambda \in \Sigma(T)} \operatorname{dist}(\lambda, \Sigma(S) \leq \|A\|_{\mathcal{L}(H)}$$

where dist *is defined as*

$$\operatorname{dist}(\lambda, \Sigma(S)) = \inf_{\mu \in \Sigma(S)} \|\lambda - \mu\|$$

and $\Sigma(T)$ denotes the spectrum of operator T.

Here, again in contrast with the finite dimensional control case, the estimation of the perturbation operator A is not easy to come by. In Chapter 3, where A was given as a difference of two matrices, it was relatively easier to compute a constant-free, computable upper bound for the norm of A, from which a lower bound for the eigenvalue of the continuous matrix was derived. To shed more light on this observation let us consider the following. For brevity let us set $w = g''(\bar{y}) + d''(\bar{y})\bar{p}$ in (5.12) which then allows us to write $T = \alpha I + S^*(w)S$ as a mapping from the control space U into itself. Let $T_h = \alpha I_h + S_h^*(w_h)S_h$ be the finite-dimensional discretization of T. Then the error matrix A in Theorem 5.9 above is given by

$$A = T - T_h = \alpha(I - I_h) + S^*(w)S - S_h^*(w_h)S_h. \tag{5.18}$$

Using the splitting

$$S^*(w)S - S_h^*(w_h)S_h = \Big((S^* - S_h^*)w_h + S^*(w - w_h)\Big)S_h + S^*(w)(S - S_h)$$

we can estimate

$$\begin{aligned}\|A\|_{\mathcal{L}(U,U^*)} \leq\ & \alpha\|I - I_h\|_{\mathcal{L}(U,U^*)} \\ & + \Big(\|S^* - S_h^*\|_{\mathcal{L}(Y,Y^*)}\|w_h\|_{(Y\times Y)^*} + \|S^*\|_{\mathcal{L}(Y,Y^*)}\|w - w_h\|_{(Y\times Y)^*}\Big)\|S_h\|_{\mathcal{L}(Y,Y^*)} \\ & + \|S^*\|_{\mathcal{L}(Y,Y^*)}\Big(\|w_h\|_{(Y\times Y)^*} + \|w - w_h\|_{(Y\times Y)^*}\Big)\|S - S_h\|_{\mathcal{L}(Y,Y^*)}.\end{aligned}$$

Note that on page 73 we used the residual norms in computing an estimate for the norm of the error matrix in that case. However for the present problem, instead of the residual norms one has to compute operator norms namely $\|S^* - S_h^*\|_{\mathcal{L}(Y,Y^*)}, \|S^*\|_{\mathcal{L}(Y,Y^*)}, \|S - S_h\|_{\mathcal{L}(Y,Y^*)}, \|S_h\|_{\mathcal{L}(Y,Y^*)}$. This is difficult computation-wise.

Moreover asides the difficulty of computing operator norms, with infinite dimensional control space things become more challenging as the perturbation of the associated eigenfunctions also has to be taken into consideration for any valid conclusion. More specifically, a considerably more effort would be needed to ensure that the eigenfunctions of the infinite dimensional Hessian are well approximated on the discretized control space. This is where the regularity result of Theorem 5.5 shall become useful.

6 Chapter 6
Conclusion and outlook

In this thesis we have developed a method to verify second-order sufficient optimality condition (SSC) for optimal control problems with finite dimensional control space. We have also derived as a side result, a-posteriori error bound for the control.

The introductory part of the thesis (Chapter 2) contains relevant basics of optimal control of partial differential equations. We proved existence result for the abstract optimal control problem under monotonicity assumption on the operator defining the state equation as well as compactness of the solution map $S : U \to Y$. Under the former assumption, the control-to-state map S turned out to be twice Fréchet differentiable which assisted in deriving optimality conditions for the problem. Furthermore by taking into account the strongly active constraints and the two-norm discrepancy that is typical of optimal control of nonlinear PDE, the sufficiency of the second-order condition was proved. We concluded the chapter by deriving a connection between the second-order sufficient conditions and the superlinear convergence of semi-smooth Newton method thereby confirming the importance of SSC in convergence proofs of numerical methods.

The major contribution of the thesis is contained in Chapter 3 which was devoted in its entirety to developing a method to verify SSC. By assuming a verifiable SSC at a discrete solution and through a careful eigenvalue and error analysis of the Hessian matrix associated with the SSC of the continuous problem, we derived conditions which allow the fulfillment of SSC to be deduced for the continuous problem. In a further step we obtained a computable upper bound for the error in the control in terms of the residuals of the optimality system. The main results are the Theorems 3.26 and 3.27. The numerical results supported the fact that the fulfillment of SSC for a discrete problem is only an indication of a similar result for the continuous counterpart, but not sufficient in general. The results are contained in [4].

Further contributions are made in Chapter 4 where we analyzed different adaptive methods based on the obtained error estimator for the control. Reliability and efficiency of the estimator were established with Theorems 4.5, 4.6, 4.8 and 4.11. The results of our numerical experiments illustrated the performance of the error estimator in different adaptive refinement procedures. Some of the results from this part are contained in [5].

The final chapter consists of preliminary results in extending our verification method to problems with infinite dimensional control space. In Theorem 5.5 we proved H^2 regularity for the eigenfunctions and highlighted expected challenges in the analysis of infinite dimensional control problems.

As final remarks, although the analysis presented in this thesis has been considered only for problems with finite dimensional control space, it however has a potential of being extended to

the general case of infinite dimensional SSC (the results of Theorem 5.9 is a step towards this direction). It is worth reiterating that in [47, 48], verification of infinite dimensional second-order conditions have been considered. However, the methods therein do not cover optimal control problems of the form (1.3). Therefore, subsequent research efforts would be devoted to extending the verification method of this thesis to infinite dimensional SSC, most especially to the problem type (1.3). The preliminary regularity result of Theorem 5.5 will play a vital role in such analysis. Finally, extension to problems with inequality constraints in the state variable would be an interesting endeavor.

Bibliography

[1] R.A. Adams. *Sobolev Spaces*. Academic Press, 1975.

[2] M. Ainsworth. A framework for obtaining guaranteed error bounds for finite element approximations. *J. Comput. Appl. Math.*, 234(9):2618–2632, 2010.

[3] M. Ainsworth and J.T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Pure and Applied Mathematics. John Wiley, 2000.

[4] S. Akindeinde and D. Wachsmuth. A-posteriori verification of optimality conditions for control problems with finite-dimensional control space. *Num. Funct. Anal. Opt.*, 33(5):473–523, 2012.

[5] S. Akindeinde and D. Wachsmuth. Adaptive methods for control problems with finite-dimensional control space. *The proceedings of the 25th IFIP TC7 Conference on system modelling and optimization*, 2012. to appear.

[6] N. Arada, E. Casas, and F. Tröltzsch. Error estimates for a semilinear elliptic control problem. *Computational Optimization and Application*, 23:201–229, 2002.

[7] N. Arada, J. Raymond, and F. Tröltzsch. On an augmented Lagrangian SQP method for a class of optimal control problems in Banach spaces. *Comput. Optim. Appl.*, 22(3):369–398, 2002.

[8] M. Bebendorf. A note on the Poincaré inequality for convex domains. *J. for Analysis and its Applications*, 22(4):751–756, 2003.

[9] R. Becker. Estimating the control error in discretized PDE-constrained optimization. *Journal of Numerical Mathematics*, 14:163–185, 2006.

[10] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numer.*, 10:1–102, 2001.

[11] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, third edition, 2007. Theory, fast solvers, and applications in elasticity theory.

[12] H. Brezis. *Analyse fonctionelle; théorie et applications*. Masson, Paris, 1983.

[13] E. Casas. Second order analysis for bang-bang control problems of PDEs. *SIAM J. Optim.*, 50(4):2355–2372, 2012.

[14] E. Casas, M. Mateos, and F. Tröltzsch. Error estimates for the numerical approximation of boundary semilinear elliptic control problems. *Comput. Optim. Appl.*, 31:193–219, 2005.

[15] E. Casas, F. Tröltzsch, and A. Unger. Second-order sufficient optimality conditions for a nonlinear elliptic control problem. *J. for Analysis and its Applications*, 15:687–707, 1996.

[16] J. W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[17] A. L. Dontchev. Local analysis of a Newton-type method based on partial linearization. In *The mathematics of numerical analysis (Park City, UT, 1995)*, volume 32 of *Lectures in Appl. Math.*, pages 295–306. Amer. Math. Soc., Providence, RI, 1996.

[18] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang. Optimality, stability, and convergence in nonlinear control. *Appl. Math. Optim.*, 31(3):297–326, 1995.

[19] J. C. Dunn. Second-order optimality conditions in sets of L^∞ functions with range in a polyhedron. *SIAM J. Control Optim.*, 33(5):1603–1635, 1995.

[20] L. El Alaoui, A. Ern, and M. Vohralík. Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems. *Comput. Methods Appl. Mech. Engrg.*, 200(37-40):2782–2795, 2011.

[21] K. Erwin. *Introductory functional analysis with applications.* Wiley, New York, NY, 1978.

[22] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, second edition, 2010.

[23] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order.* Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[24] G. H. Golub and C. F. Van Loan. *Matrix computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[25] A. Griesbaum, B. Kaltenbacher, and B. Vexler. Efficient computation of the Tikhonov regularization parameter by goal-oriented adaptive discretization. *Inverse Problems*, 24(2):025025, 20, 2008.

[26] R. Griesse. Parametric sensitivity analysis in optimal control of a reaction diffusion system. I. Solution differentiability. *Numer. Funct. Anal. Optim.*, 25(1-2):93–117, 2004.

[27] M. Hintermüller. A primal-dual active set algorithm for bilaterally control constrained optimal control problems. *Quarterly of Applied Mathematics*, 13:131–161, 2003.

[28] M. Hintermüller and M. Hinze. A SQP-semismooth Newton-type algorithm applied to control of the instationary Navier-Stokes system subject to control constraints. *SIAM J. Optim.*, 16(4):1177–1200, 2006.

[29] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth newton method. *SIAM J. Optim.*, 13(3):865–888, 2003.

[30] M. Hinze and K. Kunisch. Second order methods for optimal control of time-dependent fluid flow. *SIAM J. Control Optim.*, 40(3):925–946 (electronic), 2001.

[31] M. Hinze and K. Kunisch. Second order methods for boundary control of the instationary Navier-Stokes system. *ZAMM Z. Angew. Math. Mech.*, 84(3):171–187, 2004.

[32] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints.* Springer, 2008.

[33] K. Ito and K. Kunisch. The primal-dual active set method for nonlinear optimal control problems with bilateral constraints. *SIAM J. Optim.*, 43:357–376, 2004.

[34] K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[35] T. Kato. *Perturbation theory for linear operators.* Classics in Mathematics. Springer-Verlag,

Berlin, 1995. Reprint of the 1980 edition.

[36] K. Kunisch, W. Liu, Y. Chang, N. Yan, and R. Li. Adaptive finite element approximation for a class of parameter estimation problems. *J. Comput. Math.*, 28(5):645–675, 2010.

[37] K. Kunisch and A. Rösch. Primal-dual active set strategy for a general class of constrained optimal control problems. *SIAM J. Optim.*, 13(2):321–334, 2002.

[38] M. G. Larson. A posteriori and a priori error analysis for finite element approximations of self-adjoint elliptic eigenvalue problems. *SIAM J. Numer. Anal.*, 38(2):608–625 (electronic), 2000.

[39] K. Malanowski. Sensitivity analysis for parametric optimal control of semilinear parabolic equations. *J. Convex Anal.*, 9(2):543–561, 2002. Special issue on optimization (Montpellier, 2000).

[40] K. Malanowski and F. Tröltzsch. Lipschitz stability of solutions to parametric optimal control for elliptic equations. *Control Cybernet.*, 29(1):237–256, 2000.

[41] H. Maurer. First and second order sufficient optimality conditions in mathematical programming and optimal control. *Math. Programming Study*, 14:163–177, 1981.

[42] H. D. Mittelmann. Sufficient optimality for discretized parabolic and elliptic control problems. In *Fast solution of discretized optimization problems (Berlin, 2000)*, volume 138 of *Internat. Ser. Numer. Math.*, pages 184–196. Birkhäuser, Basel, 2001.

[43] H. D. Mittelmann. Verification of second-order sufficient optimality conditions for semilinear elliptic and parabolic control problems. *Comput. Optim. Appl.*, 20(1):93–110, 2001.

[44] L.E Payne and H.F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Rational Mech. Anal.*, 5:286–292, 1960.

[45] M. Plum. Computer-assisted enclosure methods for elliptic differential equations. *Linear Algebra Appl.*, 324(1-3):147–187, 2001.

[46] S. Repin. *A posteriori estimates for partial differential equations*, volume 4 of *Radon Series on Computational and Applied Mathematics*. Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[47] A. Rösch and D. Wachsmuth. Numerical verification of optimality conditions. *SIAM J. Control Optim.*, 47(5):2557–2581, 2008.

[48] A. Rösch and D. Wachsmuth. How to check numerically the sufficient optimality conditions for infinite-dimensional optimization problems. In *Control of Coupled Partial Differential Equations*, volume 158 of *Internat. Ser. Numer. Math.*, pages 297–317. Birkhäuser, Basel, 2009.

[49] E. Sachs and S. Volkwein. Augmented Lagrange-SQP methods with Lipschitz-continuous Lagrange multiplier updates. *SIAM J. Numer. Anal.*, 40(1):233–253, 2002.

[50] G. Stampacchia. Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus. *Ann. Inst. Fourier (Grenoble)*, 15(fasc. 1):189–258, 1965.

[51] F. Tröltzsch. On the Lagrange-Newton-SQP method for the optimal control of semilinear

parabolic equations. *SIAM J. Control Optim.*, 38(1):294–312 (electronic), 1999.

[52] F. Tröltzsch. *Optimal control of partial differential equations, theory, methods and applications*, volume 112 of *Graduate studies in Mathematics*. American Mathematical Society, 2010.

[53] M. Ulbrich. Semismooth newton methods for operator equations in function spaces. *SIAM J. Control Optim.*, 13:805–842, 2000.

[54] M. Ulbrich. *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces, Habilitationsschrift*. Fakultät für Mathematik, Technische Universität München, Germany, 2002.

[55] R. Verfürth. A note on constant-free a posteriori error estimates. *SIAM J. Numer. Anal.*, 47(4):3180–3194, 2009.

[56] B. Vexler and W. Wollner. Adaptive finite elements for elliptic optimization problems with control constraints. *SIAM J. Control Optim.*, 47(1):509–534, 2008.

[57] M. Vohralik. On the discrete Poincaré-Friedrichs inequalities for nonconforming approximations of the Sobolev space H^1. *Numer. Funct. Anal. Optim.*, 26(7-8):925–952, 2005.

[58] M. Vohralík. A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. *SIAM J. Numer. Anal.*, 45(4):1570–1599, 2007.

[59] M. Vohralík. Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. *Math. Comp.*, 79(272):2001–2032, 2010.

[60] D. Wachsmuth. Analysis of the SQP-method for optimal control problems governed by the nonstationary Navier-Stokes equations based on L^p-theory. *SIAM J. Control Optim.*, 46(3):1133–1153, 2007.

[61] D. Werner. *Funktionalanalysis*. Springer-Verlag, Berlin, extended edition, 2000.

[62] A. Wouk. *A course of applied functional analysis*. Wiley-Interscience [John Wiley & Sons], New York, 1979. Pure and Applied Mathematics.

[63] E. Zeidler. *Nonlinear functional analysis and its applications. II/B*. Springer-Verlag, New York, 1990. Nonlinear monotone operators.

[64] J. C. Ziems and S. Ulbrich. Adaptive multilevel inexact SQP methods for PDE-constrained optimization. *SIAM J. Optim.*, 21(1):1–40, 2011.