

Analysis of the Streamline Upwind/Petrov Galerkin Method  
Applied to the Solution of Optimal Control Problems

S. Scott Collis  
Matthias Heinkenschloss

March 2002  
TR02-01

Department of Computational and Applied Mathematics - MS134  
Rice University  
6100 Main Street  
Houston, Texas 77251-1892  
USA  
<http://www.caam.rice.edu/>

# Analysis of the Streamline Upwind/Petrov Galerkin Method Applied to the Solution of Optimal Control Problems <sup>\*</sup>

S. Scott Collis <sup>†</sup>

Matthias Heinkenschloss <sup>‡</sup>

March 2002

## Abstract

We study the effect of the streamline upwind/Petrov Galerkin (SUPG) stabilized finite element method on the discretization of optimal control problems governed by linear advection-diffusion equations. We compare two approaches for the numerical solution of such optimal control problems. In the discretize-then-optimize approach the optimal control problem is first discretized, using the SUPG method for the discretization of the advection-diffusion equation, and then the resulting finite dimensional optimization problem is solved. In the optimize-then-discretize approach one first computes the infinite dimensional optimality system, involving the advection-diffusion equation as well as the adjoint advection-diffusion equation, and then discretizes this optimality system using the SUPG method for both the original and the adjoint equations. These approaches lead to different results. The main result of this paper is an estimates for the error between the solution of the infinite dimensional optimal control problem and their approximations computed using the previous approaches. For a class of problems prove that the optimize-then-discretize approach has better asymptotic convergence properties if finite elements of order greater than one are used. For linear finite elements our theoretical convergence results for both approaches are comparable, except in the zero diffusion limit where again the optimize-then-discretize approach seems favorable. Numerical examples are presented to illustrate some of the theoretical results.

**Key words** Optimal control, discretization, error estimates, stabilized finite elements.

**AMS subject classifications** 49M25, 49K20, 65N15, 65J10

## 1 Introduction

This paper is concerned with the accuracy of numerical solutions of optimal control problems governed by the advection-diffusion equation. Specifically, we are interested in the effect of the streamline upwind/Petrov Galerkin (SUPG) stabilized finite element method on the discretization of the optimal control problem. To be more precise, we consider the linear quadratic optimal control problem

$$\min \frac{1}{2} \int_{\Omega} (y(x) - \hat{y}(x))^2 dx + \frac{\omega}{2} \int_{\Omega} u^2(x) dx \quad (1.1)$$

---

<sup>\*</sup>This work was supported in part by TX-ATP grant 003604-0001-1999 and NSF grants DMS-0075731 and ACI-0121360.

<sup>†</sup>Department of Mechanical Engineering and Materials Science, MS-321, Rice University, 6100 Main Street, Houston, TX 77005-1892. E-mail: collis@rice.edu

<sup>‡</sup>Department of Computational and Applied Mathematics, MS-134, Rice University, 6100 Main Street, Houston, TX 77005-1892. E-mail: heinken@rice.edu

subject to

$$-\epsilon \Delta y(x) + \mathbf{c}(x) \cdot \nabla y(x) + r(x)y(x) = f(x) + u(x), \quad x \in \Omega, \quad (1.2a)$$

$$y(x) = d(x), \quad x \in \Gamma_d, \quad (1.2b)$$

$$\epsilon \frac{\partial}{\partial \mathbf{n}} y(x) = g(x), \quad x \in \Gamma_n, \quad (1.2c)$$

where  $\Gamma_d \cap \Gamma_n = \emptyset$ ,  $\Gamma_d \cup \Gamma_n = \partial\Omega$ ,  $\mathbf{c}, d, f, g, r, \hat{y}$  are given functions,  $\epsilon, \omega > 0$  are given scalars, and  $\mathbf{n}$  denotes the outward unit normal. Assumptions on these data that ensure the well-posedness of the problem will be given in the next section.

For advection dominated problems the standard Galerkin finite element method applied to the state equation (1.2) produces strongly oscillatory solutions, unless the mesh size  $h$  is chosen sufficiently small relative to  $\epsilon/\|\mathbf{c}(x)\|$ ,  $x \in \Omega$ . To produce better approximations to the solution of (1.2) for modest mesh sizes, various augmentations of the standard Galerkin finite element method have been proposed. For an overview see [11, 12, 13]. In this paper we focus on the streamline upwind/Petrov Galerkin (SUPG) method of Hughes and Brooks [2]. The SUPG method adds to the weak form of the state equation (1.2) a term with the properties that (a) the weak form of the modification has better stability properties than the bilinear form associated with (1.2) and (b) the added term evaluated at the exact solution of (1.2) vanishes. Because of these properties the SUPG method is called a strongly consistent stabilization method [12].

For the numerical solution of the optimal control problem there are at least two approaches. In the first approach, called the *optimize-then-discretize* approach, one first derives the optimality conditions for (1.1), (1.2). In Section (2.1) we will see that the optimality conditions consist of the state equation (1.2), the adjoint partial differential equation (PDE)

$$-\epsilon \Delta \lambda(x) - \mathbf{c}(x) \cdot \nabla \lambda(x) + (r(x) - \nabla \cdot \mathbf{c}(x))\lambda(x) = -(y(x) - \hat{y}(x)), \quad x \in \Omega, \quad (1.3a)$$

$$\lambda(x) = 0, \quad x \in \Gamma_d, \quad (1.3b)$$

$$\epsilon \frac{\partial}{\partial \mathbf{n}} \lambda(x) + \mathbf{c}(x) \cdot \mathbf{n}(x) \lambda(x) = 0, \quad x \in \Gamma_n \quad (1.3c)$$

and the gradient equation

$$\lambda(x) = \omega u(x) \quad x \in \Omega. \quad (1.4)$$

Then one discretizes each equation (1.2), (1.3) and (1.4), using possibly different discretization schemes for each one. Since the adjoint equation (1.3) is also an advection-diffusion equation, but with advection  $-\mathbf{c}$ , we discretize it using the SUPG method. If we proceed this way, the optimize-then-discretize approach leads to a discretization of the optimality system (1.2), (1.3), (1.4) that is strongly consistent. However, this discretization of the optimality system (1.2), (1.3), (1.4) leads to a nonsymmetric linear system, which implies that there is no finite dimensional optimization problem for which this discretization of (1.2), (1.3), (1.4) is the optimality system. The details of the optimize-then-discretize approach will be discussed in Section 2.4. In the other approach for the numerical solution of (1.1), (1.2) called the *discretize-then-optimize* approach, one first discretizes the state equation using SUPG and the objective function and then solves the resulting finite dimensional optimization problem. The optimality conditions of this finite dimensional optimization problem contain equations, which we call the discrete adjoint equation and the discrete gradient equation, that can be viewed as discretizations of (1.3) and (1.4), respectively. The SUPG stabilization term added to the state equation (1.2) produces a contribution to the discrete adjoint equation and to the

discrete gradient equation. This contribution to the discrete adjoint equation has a stabilizing effect, but the discrete adjoint equation is in general *not* a strongly consistent stabilization method for (1.3). We will give a detailed discussion of the optimize-then-discretize approach in Section 2.3. The main goal of this paper is to derive estimates of the error between the solution  $y, u, \lambda$  of the infinite dimensional optimality system (1.2), (1.3), (1.4) and their approximations computed using both, the discretize-then-optimize as well as the optimize-then-discretize approach. Such error estimates will be provided in Section 4. Section 5 contains a few numerical results that illustrate our theoretical findings.

We will see that even in our simple model problem (1.1), (1.2) differences can arise between the discretize-then-optimize and the optimize-then-discretize approach. It is important to understand and analyze these to better assess the implication of numerical solution approaches to much more complicated optimal control or optimal design problems that involve nonlinear state equations solved using stabilization techniques. We also note that the general issues described here for the SUPG stabilization also arise when other stabilizations are used, such as the Galerkin/Least-squares (GLS) method of Hughes, Franca and Hulbert [6] and the stabilization method of Franca, Frey and Hughes [4].

Throughout this paper we use the following notation for norms and inner products. We define  $\langle f, g \rangle_G = \int_G f(x)g(x)dx$ ,  $\|v\|_{0,\infty,G} = \text{ess sup}_{x \in G} |v(x)|$  or  $\|\mathbf{v}\|_{0,\infty,G} = \text{ess sup}_{x \in G} \sqrt{\sum_i v_i(x)^2}$  for vector valued  $\mathbf{v}$ , and

$$\|v\|_{k,G} = \left( \sum_{|\alpha| \leq k} \int_G (\partial^\alpha v(x))^2 dx \right)^{1/2}, \quad |v|_{k,G} = \left( \sum_{|\alpha|=k} \int_G (\partial^\alpha v(x))^2 dx \right)^{1/2},$$

where  $G \subset \Omega \subset \mathbb{R}^d$  or  $G \subset \partial\Omega$  and  $\alpha \in \mathbb{N}_0^d$  is a multi-index,  $|\alpha| = \sum_{i=1}^d \alpha_i$ , and  $\partial^\alpha = \partial^{\alpha_1} \dots \partial^{\alpha_d}$ . If  $G = \Omega$  we omit  $G$  and simply write  $\langle f, g \rangle$ , etc.

## 2 A Model Problem

### 2.1 Existence, Uniqueness and Characterization of Optimal Controls

We define the state and control space

$$Y = \{y \in H^1(\Omega) : y = d \text{ on } \Gamma_d\}, \quad U = L^2(\Omega) \quad (2.1)$$

and space of test functions

$$V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_d\}. \quad (2.2)$$

The weak form of the state equations (1.2) is given by

$$a(y, v) + b(u, v) = \langle f, v \rangle + \langle g, v \rangle_{\Gamma_n} \quad \forall v \in V, \quad (2.3)$$

where

$$a(y, v) = \int_\Omega \epsilon \nabla y(x) \cdot \nabla v(x) + \mathbf{c}(x) \cdot \nabla y(x) v(x) + r(x) y(x) v(x) dx, \quad (2.4)$$

$$b(u, v) = - \int_\Omega u(x) v(x) dx, \quad (2.5)$$

$$\langle f, v \rangle = \int_\Omega f(x) v(x) dx, \quad \langle g, v \rangle_{\Gamma_n} = \int_{\Gamma_n} g(x) v(x) dx. \quad (2.6)$$

We are interested in the solution of the optimal control problem

$$\text{minimize} \quad \frac{1}{2}\|y - \widehat{y}\|_0^2 + \frac{\omega}{2}\|u\|_0^2, \quad (2.7a)$$

$$\text{subject to} \quad \begin{aligned} a(y, v) + b(u, v) &= \langle f, v \rangle + \langle g, v \rangle_{\Gamma_n} \quad \forall v \in V, \\ y &\in Y, u \in U. \end{aligned} \quad (2.7b)$$

We assume that

$$f, \widehat{y} \in L^2(\Omega), \mathbf{c} \in (W^{1,\infty}(\Omega))^2, r \in L^\infty(\Omega), d \in H^{3/2}(\Gamma_d), g \in H^{1/2}(\Gamma_n), \omega > 0, \epsilon > 0, \quad (2.8a)$$

$$\Gamma_n \subset \{x \in \partial\Omega : \mathbf{c}(x) \cdot \mathbf{n}(x) \geq 0\} \quad (2.8b)$$

and

$$r(x) - \frac{1}{2}\nabla \cdot \mathbf{c}(x) \geq r_0 > 0 \text{ a.e. in } \Omega. \quad (2.8c)$$

If  $\Gamma_d \neq \emptyset$ , there exists  $\alpha > 0$  such that  $|y|_1 \leq \alpha\|y\|_1$  for all  $y \in V$  and (2.8c) can be replaced by

$$r(x) - \frac{1}{2}\nabla \cdot \mathbf{c}(x) \geq r_0 \geq 0 \text{ a.e. in } \Omega. \quad (2.8d)$$

For the well-posedness of the optimal control problem it is sufficient to impose fewer regularity requirements on the coefficient functions than those stated in (2.8a). We assume (2.8a) to establish convergence estimates for the SUPG finite element method.

Under the assumptions (2.8), the bilinear form  $a$  is continuous on  $V \times V$  and  $V$ -elliptic. In fact,  $a(y, y) \geq \epsilon\|\nabla y\|_0^2 + r_0\|y\|_0^2$  for all  $y \in V$  (e.g., [12, p. 165] or [11, Sec. 2.5]). Hence the theory in [10, Sec. II.1] guarantees the existence of a unique solution  $(y, u) \in Y \times U$  of (2.7).

**Theorem 2.1** *If (2.8) are satisfied, the optimal control problem (2.7) has a unique solution  $(y, u) \in Y \times U$ .*

The theory in [10, Sec. II.1] also provides necessary and sufficient optimality conditions, which can be best described using the Lagrangian

$$L(y, u, \lambda) = \frac{1}{2}\|y - \widehat{y}\|_0^2 + \frac{\omega}{2}\|u\|_0^2 + a(y, \lambda) + b(u, \lambda) - \langle f, \lambda \rangle - \langle g, v \rangle_{\Gamma_n}. \quad (2.9)$$

The necessary and, for our model problem, sufficient optimality conditions can be obtained by setting the partial Fréchet-derivatives of (2.9) with respect to states  $y$ , controls  $u$  and adjoints  $\lambda$  equal to zero. This gives the following system consisting of the adjoint equation

$$a(\psi, \lambda) = -\langle y - \widehat{y}, \psi \rangle \quad \forall \psi \in V, \quad (2.10a)$$

the gradient equation

$$b(w, \lambda) + \omega\langle u, w \rangle = 0 \quad \forall w \in U, \quad (2.10b)$$

and the state equation

$$a(y, v) + b(u, v) = \langle f, v \rangle + \langle g, v \rangle_{\Gamma_n} \quad \forall v \in V. \quad (2.10c)$$

The gradient equation (2.10b) simply means that  $\lambda(x) = \omega u(x)$ ,  $x \in \Omega$  (cf. (1.4)) and (2.10a) is the weak form of (1.3).

The adjoint equation (1.3) is also an advection-diffusion equation, but advection is now given by  $-\mathbf{c}$  and the reaction term is  $r - \nabla \cdot \mathbf{c}$ .

The convergence theory for SUPG methods requires that the solution  $y, u, \lambda$  is more regular than indicated by Theorem 2.1. This can be guaranteed if the problem data are such that the state equation (1.2) and adjoint equation (1.3) admit more regular solutions. This motivates our regularity assumptions (2.8a) on the data. The following result is an application of [5, Thm. 2.4.2.5] to (1.2) and (1.3).

**Theorem 2.2** *Let  $\Omega$  be a bounded open subset of  $\mathbb{R}^n$  with a  $C^{1,1}$  boundary and  $\Gamma_d = \partial\Omega$ . If the assumption (2.8a) is satisfied and  $r \geq r_0 > 0$  a.e., then the unique solution of the optimal control problem (2.7) and the associated adjoint satisfy  $y \in H^2(\Omega)$ ,  $u \in H^2(\Omega)$ ,  $\lambda \in H^2(\Omega)$ .*

## 2.2 Discretization of the State Equations

For the discretization of the state equation we use conforming finite elements. We let  $\{\mathcal{T}_h\}_{h>0}$  be a family of quasi-uniform triangulations of  $\Omega$  [3]. To approximate the state equation we use the spaces

$$\begin{aligned} Y_h &= \{y_h \in Y : y_h|_T \in P_k(T) \text{ for all } T \in \mathcal{T}_h\}, \\ V_h &= \{v_h \in V : v_h|_T \in P_k(T) \text{ for all } T \in \mathcal{T}_h\}, \quad k \geq 1. \end{aligned} \quad (2.11)$$

For advection dominated problems the standard Galerkin method applied to the state equation (2.3) produces strongly oscillatory approximations, unless the mesh size  $h$  is chosen sufficiently small relative to  $\epsilon/\|\mathbf{c}\|_{0,\infty}$ . To obtain approximate solutions of better quality on coarser meshes, various stabilization techniques have been proposed. For an overview see [12, Secs. 8.3.2,8.4] or [13, Sec.3.2]. We are interested in the streamline upwind/Petrov Galerkin (SUPG) method of Hughes and Brooks [2]. The SUPG method computes an approximation  $y_h \in Y_h$  of the solution  $y$  of the state equation (2.7b) by solving

$$a_h^s(y_h, v_h) + b_h^s(u_h, v_h) = \langle f, v_h \rangle_h^s + \langle g, v_h \rangle_{\Gamma_n} \quad \forall v_h \in V_h, \quad (2.12)$$

where

$$a_h^s(y, v_h) = a(y, v_h) + \sum_{T_e \in \mathcal{T}_h} \tau_e \langle -\epsilon \Delta y + \mathbf{c} \cdot \nabla y + ry, \mathbf{c} \cdot \nabla v_h \rangle_{T_e}, \quad (2.13a)$$

$$b_h^s(u, v_h) = - \int_{\Omega} u(x) v_h(x) dx - \sum_{T_e \in \mathcal{T}_h} \tau_e \langle u, \mathbf{c} \cdot \nabla v_h \rangle_{T_e}, \quad (2.13b)$$

$$\langle f, v_h \rangle_h^s = \langle f, v_h \rangle + \sum_{T_e \in \mathcal{T}_h} \tau_e \langle f, \mathbf{c} \cdot \nabla v_h \rangle_{T_e}. \quad (2.13c)$$

In (2.15b) the superscript  $s$  is used to indicate that the stabilization method is applied to the state equation, i.e., all parameters in the stabilization method are based on information from the state equation only. The reason for the additional superscript  $s$  will become apparent in Section 2.4. The addition of the term  $\sum_{T_e \in \mathcal{T}_h} \tau_e \langle -\epsilon \Delta y + \mathbf{c} \cdot \nabla y + ry, \mathbf{c} \cdot \nabla v_h \rangle_{T_e}$  to  $a(y, v_h)$  introduces additional element wise diffusion  $\langle \mathbf{c} \cdot \nabla y, \mathbf{c} \cdot \nabla v_h \rangle_{T_e}$  and enhances the stability properties of  $a(y, v_h)$  (see Lemma 4.1 below). The terms in (2.12) added to the standard Galerkin formulation are such that the exact solution  $y$  of (1.2) satisfies (2.12), provided  $y \in H^2(T_e)$ ,  $T_e \in \mathcal{T}_h$ . We will review the error estimates for the SUPG method in Section 4.1.

### 2.3 Discretization of the Optimization Problem

A frequently used approach for the numerical solution of an optimal control problem is to discretize the optimal control problem and to solve the resulting nonlinear programming problem using a suitable optimization algorithm. This is also called the *discretize-then-optimize* approach. In this scenario, the discretization of the optimal control problem typically follows discretization techniques used for the governing state equations.

In our problem we select the spaces (2.11) for the discretization of the state and

$$U_h = \{u_h \in U : u_h|_T \in P_m(T) \text{ for all } T \in \mathcal{T}_h\}, \quad m \geq 0, \quad (2.14)$$

for the control. To discretize the state equation, we apply the SUPG method. The discretized optimal control problem is given by

$$\text{minimize} \quad \frac{1}{2} \|y_h - \hat{y}\|_0^2 + \frac{\omega}{2} \|u_h\|_0^2, \quad (2.15a)$$

$$\begin{aligned} \text{subject to} \quad & a_h^s(y_h, v_h) + b_h^s(u_h, v_h) = \langle f, v_h \rangle_h^s + \langle g, v_h \rangle_{\Gamma_n} \quad \forall v_h \in V_h, \\ & y_h \in Y_h, \quad u_h \in U_h, \end{aligned} \quad (2.15b)$$

where  $a_h^s(y, v_h)$ ,  $b_h^s(u, v_h)$ ,  $\langle f, v_h \rangle_h^s$  are defined in (2.13).

The Lagrangian for the discretized problem (2.15) is given by

$$L_h(y_h, u_h, \lambda_h) = \frac{1}{2} \|y_h - \hat{y}\|_0^2 + \frac{\omega}{2} \|u_h\|_0^2 + a_h^s(y_h, \lambda_h) + b_h^s(u_h, \lambda_h) - \langle f, \lambda_h \rangle_h^s - \langle g, v \rangle_{\Gamma_n}, \quad (2.16)$$

where  $y_h \in Y_h$ ,  $u_h \in U_h$  and  $\lambda_h \in \Lambda_h \stackrel{\text{def}}{=} V_h$ . The necessary and sufficient optimality conditions for the discretized problem are obtained by setting the partial derivatives of (2.16) to zero. This gives the following system consisting of

the discrete adjoint equations

$$a_h^s(\psi_h, \lambda_h) = -\langle y_h - \hat{y}, \psi_h \rangle \quad \forall \psi_h \in V_h, \quad (2.17a)$$

the discrete gradient equations

$$b_h^s(w_h, \lambda_h) + \omega \langle u_h, w_h \rangle = 0 \quad \forall w_h \in U_h, \quad (2.17b)$$

and the discretized state equations

$$a_h^s(y_h, v_h) + b_h^s(u_h, v_h) = \langle f, v_h \rangle_h^s + \langle g, v \rangle_{\Gamma_n} \quad \forall v_h \in V_h. \quad (2.17c)$$

We use *discrete adjoint equations* and *discrete gradient equations* to mean that these are the adjoint and gradient equations for the discretized problem (2.15). We will use the phrases *discretized adjoint equations* and *discretized gradient equations* to refer to discretizations of the adjoint equation (1.3) and gradient equation (1.4), respectively. As we will see in the next section, there are significant differences between the discrete adjoint equations and the discretized adjoint equations as well as between the discrete gradient equations and the discretized gradient equations.

We notice that the discretized state equation (2.17c) is strongly consistent in the sense that (2.17c) is satisfied if  $u_h, y_h$  are replaced by the optimal control  $u$  and the corresponding optimal state  $y$ . However,

strong consistency is lost in the discrete adjoint equations (2.17a) and the discrete gradient equations (2.17b). Specifically,

$$a_h^s(\psi_h, \lambda_h) = a(\psi_h, \lambda_h) + \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle -\epsilon \Delta \psi_h + \mathbf{c} \cdot \nabla \psi_h + r \psi_h, \mathbf{c} \cdot \nabla \lambda_h \rangle_{T_e}.$$

The amount  $\sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle -\mathbf{c} \cdot \nabla \psi_h, -\mathbf{c} \cdot \nabla \lambda_h \rangle_{T_e}$  of streamline diffusion added to  $a(\psi_h, \lambda_h)$  in the adjoint equation appears to be right in the sense that this amount (although possibly with a different  $\tau_e^s$ ) would be added if the SUPG method had been applied to the adjoint equation (1.3). However, (2.17a) is not satisfied if  $y_h, \lambda_h$  are replaced by the optimal state  $y$  and corresponding adjoint  $\lambda$ . This lack of strong consistency is due to the fact that the discrete adjoint equation is not a method of weighted residuals for the continuous adjoint problem. In particular, the resulting stabilization term is not a weighted residual of the continuous adjoint equation on element interiors since the diffusion, reaction, and source terms are not accounted for. Similarly,  $b_h^s$  in the gradient equation contains terms that arise from the stabilization of the state equation and (2.17b) is not satisfied if  $u_h, \lambda_h$  are replaced by the optimal control  $u$  and corresponding adjoint  $\lambda$ .

## 2.4 Discretization of the Optimality Conditions

Alternatively to the discretize-then-optimize approach discussed in the previous section, one can obtain an approximate solution of the optimal control problem by tackling the optimality system (2.10) directly. This leads to the *optimize-then-discretize* approach. Here each equation in (2.10) is discretized using a potentially different scheme. In our case, we will use the same triangulation for all three equations and we will use the state space (2.11) and the control space (2.14) for the discretization of states and controls, respectively, and we will use

$$\Lambda_h = \{v_h \in V : v_h|_T \in P_\ell(T) \text{ for all } T \in \mathcal{T}_h\}, \quad \ell \geq 1, \quad (2.18)$$

for the discretization of the adjoints. It is possible to choose  $\ell \neq k$ . Now we take into account that the adjoint equation (1.3) is also an advection dominated problem, but with advective term  $-\mathbf{c} \cdot \nabla \lambda$ . We discretize (1.3) using the SUPG method. This leads to the discretized adjoint equations

$$a_h^a(\psi_h, \lambda_h) = -\langle y_h - \hat{y}, \psi_h \rangle_h^a \quad \forall \psi_h \in \Lambda_h, \quad (2.19a)$$

where

$$a_h^a(\psi_h, \lambda) = a(\psi_h, \lambda) + \sum_{T_e \in \mathcal{T}_h} \tau_e^a \langle -\epsilon \Delta \lambda - \mathbf{c} \cdot \nabla \lambda + (r - \nabla \cdot \mathbf{c}) \lambda, -\mathbf{c} \cdot \nabla \psi_h \rangle_{T_e}, \quad (2.19b)$$

$$\langle y - \hat{y}, \psi_h \rangle_h^a = \langle y - \hat{y}, \psi_h \rangle + \sum_{T_e \in \mathcal{T}_h} \tau_e^a \langle y - \hat{y}, -\mathbf{c} \cdot \nabla \psi_h \rangle_{T_e}. \quad (2.19c)$$

Here and in the following the superscript a is used to indicate that the SUPG method is applied to the adjoint equation, i.e., all parameters in the stabilization method applied to (1.3) are based on information from the adjoint equation (1.3).

The gradient equation (2.10b) is discretized using

$$b(w_h, \lambda_h) + \omega \langle u_h, w_h \rangle = 0 \quad \forall w_h \in U_h, \quad (2.19d)$$

and the discretization of the state equations is identical to the one used in the previous section, i.e.,

$$a_h^s(y_h, v_h) + b_h^s(u_h, v_h) = \langle f, v_h \rangle_h^s + \langle g, v_h \rangle_{\Gamma_n} \quad \forall v_h \in V_h. \quad (2.19e)$$

Unlike the discrete adjoint and gradient equations, the discretized state, adjoint and gradient equations are strongly consistent in the sense that if  $y, u, \lambda$  solve (2.10) and satisfy  $y, \lambda \in H^2(T_e)$ , for all  $T_e \in \mathcal{T}_h$ , then  $y, u, \lambda$  also satisfy (2.19).

Due to the occurrence of the SUPG terms in the right hand side of (2.19a) and in  $\tilde{\mathcal{K}}_h$ , the discretization (2.19) of the infinite dimensional optimality conditions leads to a nonsymmetric system for the computation of  $y_h, u_h, \lambda_h$ . This implies that (2.19) cannot be a system of optimality conditions for an optimization problem, e.g., a perturbation of (2.15).

### 3 Abstract Formulation

To analyze the error between the solution of the optimal control problem (2.7) and the solution of the discretized optimal control problem (2.15) we could apply the approximation theory for saddle point problems described, e.g., in [1]. However, the optimize-then-discretize approach leads to a non-symmetric system (2.19). Thus there is no optimization problem whose optimality system is given by (2.19) and the theory in [1] can not be applied to this situation. We prefer to use a framework that is common in Numerical Analysis for the estimation of the approximation error in operator equations. We give a brief review here and apply it in the following section to our problem.

The necessary and sufficient optimality conditions (2.10) can be viewed as an operator equation

$$\mathbf{K}\mathbf{x} = \mathbf{r} \quad (3.1)$$

in  $\mathcal{X}^*$ , where  $\mathcal{X}$  is a Banach space,  $\mathcal{X}^*$  is its dual and  $\mathbf{K} \in \mathcal{L}(\mathcal{X}, \mathcal{X}^*)$  is continuously invertible. In the following section we describe in detail how (3.1) relates to our problem. The discretized problem is described by the equation

$$\mathbf{K}_h \mathbf{x}_h = \mathbf{r}_h, \quad (3.2)$$

in  $\mathcal{X}_h^*$ , where  $\mathcal{X}_h$  is a finite dimensional Banach space with norm  $\|\cdot\|_h$  and  $\mathbf{K}_h \in \mathcal{L}(\mathcal{X}_h, \mathcal{X}_h^*)$  is continuously invertible.

To derive an error estimate we let  $\mathbf{R}_h : \mathcal{X} \rightarrow \mathcal{X}_h$  be a restriction operator and we consider the identity

$$\mathbf{K}_h(\mathbf{x}_h - \mathbf{R}_h(\mathbf{x})) = \mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x}).$$

We immediately obtain the estimate

$$\|\mathbf{x}_h - \mathbf{R}_h(\mathbf{x})\|_h \leq \|\mathbf{K}_h^{-1}\|_h \|\mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x})\|_h, \quad (3.3)$$

where  $\|\mathbf{K}_h^{-1}\|_h$  denotes the operator norm of  $\mathbf{K}_h^{-1}$  induced by  $\|\cdot\|_h$ . If there exists  $\kappa > 0$  independent of  $h$  such that the stability estimate

$$\|\mathbf{K}_h^{-1}\|_h \leq \kappa \quad \text{for all } h \quad (3.4)$$

is valid and if we can prove a consistency result of the form

$$\|\mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x})\|_h = O(h^p), \quad (3.5)$$

then we obtain  $\|\mathbf{x}_h - \mathbf{R}_h(\mathbf{x})\|_h = O(h^p)$ . If  $\|\cdot\|_h$  can be extended to define a norm on  $\mathcal{X}$ , and if we can prove

$$\|\mathbf{x} - \mathbf{R}_h(\mathbf{x})\|_h = O(h^q), \quad (3.6)$$

then a simple application of the triangle inequality shows

$$\|\mathbf{x}_h - \mathbf{x}\|_h \leq \|\mathbf{K}_h^{-1}\|_h \|\mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x})\|_h + \|\mathbf{R}_h(\mathbf{x}) - \mathbf{x}\|_h = O(h^{\min\{p,q\}}) \quad (3.7)$$

In (3.7) the error is measured in a norm that depends on  $h$ . This is certainly not problematic if there exists  $\eta > 0$  independent of  $h$  such that  $\|\mathbf{x}_h - \mathbf{x}\| \leq \eta \|\mathbf{x}_h - \mathbf{x}\|_h$  for all  $h$ , which will be true in our situation.

In our applications,  $\mathcal{X}_h = Y_h \times U_h \times \Lambda_h$  with some finite dimensional Banach spaces  $Y_h, U_h, \Lambda_h$  equipped with norms  $\|\cdot\|_{Y_h}, \|\cdot\|_{U_h}$  and  $\|\cdot\|_{\Lambda_h}$ , respectively. The norm  $\|\cdot\|_h$  on  $\mathcal{X}_h$  is defined as  $\|\mathbf{x}_h\|_h = \|y_h\|_{Y_h} + \|u_h\|_{U_h} + \|\lambda_h\|_{\Lambda_h}$ , where  $\mathbf{x} = (y_h, u_h, \lambda_h)$ . Furthermore, in our applications the operator  $\mathbf{K}_h$  is of the form

$$\mathbf{K}_h = \begin{pmatrix} H_h^{yy} & H_h^{yu} & \tilde{A}_h^* \\ H_h^{uy} & H_h^{uu} & \tilde{B}_h^* \\ A_h & B_h & 0 \end{pmatrix}. \quad (3.8)$$

The operator  $\mathbf{K}_h$  is not necessarily selfadjoint, i.e., we do *not* assume that  $A_h^* = \tilde{A}_h^*, B_h^* = \tilde{B}_h^*, (H_h^{yy})^* = H_h^{yy}, (H_h^{yu})^* = H_h^{uy},$  or  $(H_h^{uu})^* = H_h^{uu}$ . We assume, however, that  $A_h$  and  $\tilde{A}_h^*$  are invertible.

To estimate  $\|\mathbf{K}_h^{-1}\|_h$  we consider

$$\begin{pmatrix} H_h^{yy} & \tilde{A}_h^* & H_h^{yu} \\ A_h & 0 & B_h \\ H_h^{uy} & \tilde{B}_h^* & H_h^{uu} \end{pmatrix} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ \tilde{B}_h^*(\tilde{A}_h^*)^{-1} & (H_h^{uy} - \tilde{B}_h^*(\tilde{A}_h^*)^{-1}H_h^{yy})A_h^{-1} & I \end{pmatrix} \begin{pmatrix} H_h^{yy} & \tilde{A}_h^* & 0 \\ A_h & 0 & 0 \\ 0 & 0 & \hat{H}_h \end{pmatrix} \\ \times \begin{pmatrix} I & 0 & A_h^{-1}B_h \\ 0 & I & (\tilde{A}_h^*)^{-1}(H_h^{yu} - H_h^{yy}A_h^{-1}B_h) \\ 0 & 0 & I \end{pmatrix}, \quad (3.9)$$

where

$$\hat{H}_h = H_h^{uu} - \tilde{B}_h^*(\tilde{A}_h^*)^{-1}H_h^{yu} - H_h^{uy}A_h^{-1}B_h + \tilde{B}_h^*(\tilde{A}_h^*)^{-1}H_h^{yy}A_h^{-1}B_h \quad (3.10)$$

The operator on the left hand side in (3.9) is just a symmetric permutation of  $\mathbf{K}_h$ , which does not effect the invertibility of  $\mathbf{K}_h$  or the estimate for  $\|\mathbf{K}_h^{-1}\|_h$ . Under the assumption that  $A_h$  and  $\tilde{A}_h^*$  are invertible, (3.9) shows that  $\mathbf{K}_h$  is invertible if and only if  $\hat{H}_h$  is invertible. Using

$$\begin{pmatrix} H_h^{yy} & \tilde{A}_h^* \\ A_h & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & A_h^{-1} \\ (\tilde{A}_h^*)^{-1} & -(\tilde{A}_h^*)^{-1}H_h^{yy}A_h^{-1} \end{pmatrix} \quad (3.11)$$

and (3.9) we immediately obtain the following result.

**Lemma 3.1** *If  $\tilde{A}_h, \tilde{A}_h^*$  and  $\hat{H}_h$  are invertible for all  $h$  and if  $\|A_h^{-1}\|_{\mathcal{L}(Y_h^*, Y_h)}, \|(\tilde{A}_h^*)^{-1}\|_{\mathcal{L}(\Lambda_h^*, \Lambda_h)}, \|\hat{H}_h^{-1}\|_{\mathcal{L}(U_h, U_h^*)}, \|A_h^{-1}B_h\|_{\mathcal{L}(U_h, Y_h)}, \|\tilde{B}_h^*(\tilde{A}_h^*)^{-1}\|_{\mathcal{L}(Y_h^*, U_h^*)}, \|(H_h^{uy} - \tilde{B}_h^*(\tilde{A}_h^*)^{-1}H_h^{yy})A_h^{-1}\|_{\mathcal{L}(Y_h^*, U_h^*)}, \|(\tilde{A}_h^*)^{-1}(H_h^{yu} - H_h^{yy}A_h^{-1}B_h)\|_{\mathcal{L}(U_h, \Lambda_h)},$  and  $\|(\tilde{A}_h^*)^{-1}H_h^{yy}A_h^{-1}\|_{\mathcal{L}(Y_h, \Lambda_h)}$  are uniformly bounded, then there exists  $\kappa > 0$  such that (3.4) holds.*

## 4 Error Estimates for the SUPG Method

In this section we derive estimates for the error between the solution of the optimal control problem and the computed approximations using both, the discretize-then-optimize and the optimize-then-discretize approaches.

Before we apply the theory outlined in Section 3 to the optimization problem, we briefly review estimates for the error between the solution  $y$  of (2.3) and its approximation  $y_h$  by the SUPG method. Such estimates are given in the paper [8] and the books [9, 13]. See also [12]. We sketch the main points of the error analysis to recall some basic estimates needed in our analysis of the SUPG discretization for optimal control.

Throughout this section we assume that the Dirichlet boundary data are  $d = 0$ . This can always be achieved by a shift of the state.

### 4.1 Error Estimates for the State Equation

We define

$$\|v\|_{SD}^2 = \epsilon|v|_1^2 + r_0\|v\|_0^2 + \sum_{T_e \in \mathcal{T}_h} \tau_e \|\mathbf{c} \cdot \nabla v\|_{0,T_e}^2. \quad (4.1)$$

Recall that  $k \geq 1$  is the polynomial degree of the finite element spaces  $Y_h, V_h$  defined in (2.11). For  $y \in H^{k+1}(\Omega)$  we let  $y^I$  be its  $Y_h$ -interpolant. We recall the interpolation error estimate

$$|y - y^I|_{p,T_e} \leq \mu_{\text{int}} h_e^{k+1-p} |y|_{k+1,T_e} \quad \text{for } p = 0, 1, 2 \quad (4.2)$$

and the inverse inequalities

$$|v_h|_{1,T_e} \leq \mu_{\text{inv}} h_e^{-1} \|v_h\|_{0,T_e}, \quad \|\Delta v_h\|_{0,T_e} \leq \mu_{\text{inv}} h_e^{-1} \|\nabla v_h\|_{0,T_e}, \quad \forall v_h \in V_h, \quad (4.3)$$

see, e.g., [3, Thms. 16.2, 17.2]. Here  $h_e$  denotes the radius of the circumscribed circle of  $T_e$  and  $h = \max_{T_e \in \mathcal{T}_h} h_e$ . The following lemma can be found, e.g., in [13, L. 3.28] or in [9, pp. 325,326].

**Lemma 4.1** *If*

$$0 < \tau_e^s \leq \min \left\{ \frac{h_e^2}{\epsilon \mu_{\text{inv}}^2}, \frac{r_0}{\|r\|_{0,\infty,T_e}} \right\}, \quad (4.4)$$

*then*

$$a_h^s(v_h, v_h) \geq \frac{1}{2} \|v_h\|_{SD}^2 \quad \forall v_h \in V_h. \quad (4.5)$$

The following inequalities can be found in [13, p. 232] or in [9, pp. 327,328].

**Lemma 4.2** *Let  $y \in H^{k+1}(\Omega)$  with  $k \geq 1$ . There exists a constant  $C > 0$  dependent on  $\mu_{\text{int}}, \mathbf{c}, r$ , but independent of  $h_e, \tau_e$  such that*

$$\left| \epsilon \langle \nabla(y^I - y), \nabla v_h \rangle \right| \leq C \epsilon^{1/2} h^k |y|_{k+1} \|v_h\|_{SD} \quad (4.6)$$

$$\left| \langle \mathbf{c} \cdot \nabla(y^I - y) + r(y^I - y), v_h \rangle \right| \leq C h^k \left( \sum_{T_e \in \mathcal{T}_h} (1 + 1/\tau_e^s) h_e^2 |y|_{k+1,T_e} \right)^{1/2} \|v_h\|_{SD} \quad (4.7)$$

for all  $v_h \in V_h$ . Furthermore, if  $\tau_e$  satisfies (4.4), then

$$\begin{aligned} & \left| \sum_{T_e \in \mathcal{T}_h} \tau_e \langle -\epsilon \Delta(y^I - y) + \mathbf{c} \cdot \nabla(y^I - y) + r(y^I - y), \mathbf{c} \cdot v_h \rangle \right| \\ & \leq Ch^k \left( \sum_{T_e \in \mathcal{T}_h} (\epsilon + \tau_e^s) |y|_{k+1, T_e}^2 \right)^{1/2} \|v_h\|_{SD} \quad \text{for all } v_h \in V_h. \end{aligned} \quad (4.8)$$

The stability result (4.5), the estimates (4.6)-(4.8), and the identity  $a_h(y - y_h, v_h) = 0$  for all  $v_h \in V_h$ , yield

$$\begin{aligned} \frac{1}{2} \|y^I - y_h\|_{SD}^2 & \leq a_h(y^I - y_h, y^I - y_h) = a_h(y^I - y, y^I - y_h) \\ & \leq Ch^k \left( \sum_{T_e \in \mathcal{T}_h} (\epsilon + \tau_e^s + h_e^2/\tau_e^s + h_e^2) |y|_{k+1, T_e}^2 \right)^{1/2} \|y^I - y_h\|_{SD}. \end{aligned} \quad (4.9)$$

The stabilization parameter  $\tau_e$  is chosen to balance the terms in  $\epsilon + \tau_e^s + h_e^2/\tau_e^s + h_e^2$ . In particular, if

$$\tau_e^s = \begin{cases} \tau_1 \frac{h_e^2}{\epsilon}, & \text{Pe}_e \leq 1, \\ \tau_2 h_e, & \text{Pe}_e > 1, \end{cases} \quad (4.10)$$

where  $\tau_1, \tau_2 > 0$  are user specified constants and

$$\text{Pe}_e = \frac{\|\mathbf{c}\|_{0, \infty, T_e} h_e}{2\epsilon} \quad (4.11)$$

is the mesh Péclet number, then

$$\|y^I - y_h\|_{SD} \leq Ch^k (\epsilon^{1/2} + h^{1/2}) |y|_{k+1}. \quad (4.12)$$

An estimate of  $\|y^I - y\|_{SD}$  using inequalities similar to those in Lemma 4.2 and an application of the triangle inequality leads to the error estimate stated in the following theorem, see [13, Thm. 3.30] or [9, Thm. 9.3].

**Theorem 4.3** *Let (2.8) be valid and let the solution  $y$  of (2.7b) satisfy  $y \in H^{k+1}(\Omega)$  with  $k \geq 1$ . If  $\tau_e$  satisfies (4.4) and (4.10), then the solution  $y_h$  of (2.12) obeys*

$$\|y - y_h\|_{SD} \leq Ch^k (\epsilon^{1/2} + h^{1/2}) |y|_{k+1}. \quad (4.13)$$

## 4.2 Error Estimates for the Optimal Control Problem

We apply the framework of Section 3 to our problem. In this case,

$$\mathbf{x} = (y, u, \lambda), \quad \mathcal{X} = Y \times U \times \Lambda,$$

where  $Y, U$  are defined in (2.1) and  $\Lambda = V$  is specified in (2.2). Furthermore,

$$\mathcal{X}_h = Y_h \times U_h \times \Lambda_h,$$

where the discretized state and control spaces are given by  $Y_h, U_h$  are defined in (2.11) and (2.14), respectively. If we use the discretize-then-optimize approach, the discrete adjoints are in  $\Lambda_h = V_h$ , where  $V_h$  is defined in (2.11). If we use the optimize-then-discretize approach, then  $\Lambda_h$  is defined in (2.18). The discrete state and control spaces will be equipped with norms

$$\|y_h\|_{Y_h}^2 = \|y_h\|_{SD}^2 = \epsilon |y_h|_1^2 + r_0 \|y_h\|_0^2 + \sum_{T_e \in \mathcal{T}_h} \tau_e^s \|\mathbf{c} \cdot \nabla y_h\|_{0, T_e}^2$$

and  $\|\cdot\|_{U_h} = \|\cdot\|_0$ , respectively. If the discretize-then-optimize approach is used, then  $\|\cdot\|_{\Lambda_h} = \|\cdot\|_{Y_h} = \|\cdot\|_{SD}$ . If the optimize-then-discretize approach is used,

$$\|\lambda_h\|_{\Lambda_h}^2 = \|\lambda_h\|_{SD}^2 = \epsilon |\lambda_h|_1^2 + r_0 \|\lambda_h\|_0^2 + \sum_{T_e \in \mathcal{T}_h} \tau_e^a \|\mathbf{c} \cdot \nabla \lambda_h\|_{0, T_e}^2$$

Since stabilization parameters  $\tau_e^s$  and  $\tau_e^a$  might be different in the discretization of state and adjoint equation, it is not quite accurate to use  $\|\cdot\|_{SD}$  to denote both norms  $\|\cdot\|_{Y_h}$  and  $\|\cdot\|_{\Lambda_h}$ . However, we hope that its meaning is clear from the context. The space  $\mathcal{X}_h$  will be equipped with norm

$$\|\mathbf{x}_h\|_h = \|y_h\|_{SD} + \|u_h\|_0 + \|\lambda_h\|_{SD},$$

where  $\mathbf{x}_h = (y_h, u_h, \lambda_h)^T$ .

The equation (3.1) corresponds to the optimality conditions (2.10). Depending on whether the discretize-then-optimize approach or the optimize-then-discretize approach is used, the discrete equation (3.2) corresponds to (2.17) or (2.19), respectively.

As the restriction operator  $\mathbf{R}_h : \mathcal{X} \rightarrow \mathcal{X}_h$ , we choose

$$\mathbf{R}_h(\mathbf{x}) = \begin{pmatrix} y^I \\ Pu \\ \lambda^I \end{pmatrix},$$

where  $y^I, \lambda^I$  denote the interpolants of  $y, \lambda$  onto  $Y_h, \Lambda_h$  and where  $P : U \rightarrow U_h$  is the  $L^2$ -projection defined by

$$\langle Pu, w_h \rangle = \langle u, w_h \rangle \quad \forall w_h \in U_h. \quad (4.14)$$

If  $m \geq 1$  and if  $u \in H^{m+1}(\Omega)$ , then the optimality of the projection  $P$  and the interpolation estimate (4.2) imply that

$$\|u - Pu\|_0 \leq \|u - u^I\|_0 \leq \mu_{\text{int}} h^{m+1} |u|_{m+1}, \quad (4.15)$$

where  $u^I$  is the  $U_h$ -interpolant of  $u$ .

Recall that  $k, \ell \geq 1$  and  $m \geq 0$  are the polynomial degrees of the finite element spaces  $Y_h, \Lambda_h, U_h$  defined in (2.11), (2.18) and (2.14), respectively. If the discretize-then-optimize approach is used,  $\Lambda_h = V_h$  and we set  $\ell = k$ .

The following lemma provides an estimate for the term  $\|\mathbf{R}_h(\mathbf{x}) - \mathbf{x}\|_h$  in the abstract error estimate (3.7).

**Lemma 4.4** *Let  $\mathbf{x} = (y, u, \lambda)$  be the solution of (2.7). If  $k, \ell, m \geq 1$  and  $y \in H^{k+1}(\Omega)$ ,  $u \in H^{m+1}(\Omega)$  and  $\lambda \in H^{\ell+1}(\Omega)$ , then there exists a constant  $C$  depending on  $\mu_{\text{int}}, \mathbf{c}, r$  such that  $\|y - y^I\|_{SD} \leq Ch^k(\epsilon^{1/2} + h^{1/2})|y|_{k+1}$ ,  $\|\lambda - \lambda^I\|_{SD} \leq Ch^\ell(\epsilon^{1/2} + h^{1/2})|\lambda|_{\ell+1}$  and  $\|u - Pu\|_0 \leq Ch^{m+1}|u|_{m+1}$  for all  $h$ .*

**Proof:** The estimates for  $\|y - y^I\|_{SD}, \|\lambda - \lambda^I\|_{SD}$  follow from the interpolation estimate (4.2) using standard arguments, see [13, Thm. 3.30] or [9, Thm. 9.3]. The estimate for  $\|u - Pu\|_0$  is shown in (4.15).  $\square$

Note that  $u \in H^{m+1}(\Omega)$ ,  $\lambda \in H^{\ell+1}(\Omega)$  and the optimality condition (1.4) imply that  $\lambda = \omega u \in H^{\min\{\ell+1, m+1\}}(\Omega)$ . However, in Lemma 4.4 and in the following we prefer to impose the regularity assumption on  $\lambda$  and  $u$  separately, to better indicate where each is used.

### 4.3 Discretize-Then-Optimize

In the discretize-then-optimize approach the discrete equation (3.2) corresponds to (2.17). The components of  $\mathbf{K}_h$  in (3.8) are given by

$$\begin{aligned} \langle H_h^{yy} y_h, v_h \rangle_{V_h^* \times V_h} &= \langle y_h, v_h \rangle, & \langle H_h^{uu} u_h, w_h \rangle_{U_h^* \times U_h} &= \omega \langle u_h, w_h \rangle, & H_h^{uy} &= H_h^{yu} = 0 \\ \langle A_h y_h, v_h \rangle_{V_h^* \times V_h} &= a_h^s(y_h, v_h), & \langle B_h u_h, v_h \rangle_{V_h^* \times V_h} &= b_h^s(u_h, v_h), & \tilde{A}_h &= A_h, & \tilde{B}_h &= B_h. \end{aligned} \quad (4.16)$$

In particular,  $\mathbf{K}_h$  is selfadjoint.

The next result establishes a stability estimate for the optimal control problem.

**Lemma 4.5** *Let  $k \geq 1$  and suppose the solution  $\mathbf{x} = (y, u, \lambda)$  of (2.7) satisfies  $y \in H^{k+1}(\Omega)$  and  $\lambda \in H^{k+1}(\Omega)$ . If  $\tau_e^s$  satisfies (4.4) and (4.10), then there exists  $\kappa > 0$  such that (3.4) holds.*

**Proof:** We apply Lemma 3.1. By Lemma 4.1  $A_h$  is invertible and satisfies  $\|A_h^{-1}\|_{\mathcal{L}(Y_h^*, Y_h)} \leq 2$ . It is easy to see that there exists  $c > 0$  such that  $\|B_h\|_{\mathcal{L}(U_h, Y_h^*)} \leq c$  and  $\|H_h^{yy}\|_{\mathcal{L}(Y_h, Y_h^*)} \leq c$  for all  $h$ . Finally, since  $B_h^*(A_h^*)^{-1} H_h^{yy} A_h^{-1} B_h$  is positive semi definite,

$$\langle \hat{H}_h u_h, u_h \rangle_{U_h^* \times U_h} \geq \langle H_h^{uu} u_h, u_h \rangle_{U_h^* \times U_h} = \omega \|u_h\|_0^2,$$

which implies  $\|\hat{H}_h^{-1}\|_{\mathcal{L}(U_h^*, U_h)} \leq \omega^{-1}$ .  $\square$

Now we turn to the consistency, i.e., we want to find an estimate for  $\|\mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x})\|_h$  in our abstract error estimate (3.7). Let  $\mathbf{z}_h = (\psi_h, w_h, v_h) \in \mathcal{X}_h$ . The optimality conditions (2.17) of the discretized problem imply that

$$\langle \mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x}), \mathbf{z}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} = \begin{pmatrix} a_h^s(\psi_h, \lambda^I) + \langle y^I - \hat{y}, \psi_h \rangle \\ b_h^s(w_h, \lambda^I) + \omega \langle Pu, w_h \rangle \\ a_h^s(y^I, v_h) + b_h^s(Pu, v_h) - \langle f, v_h \rangle_h^s - \langle g, v_h \rangle_{\Gamma_n} \end{pmatrix}. \quad (4.17)$$

Since the solution  $\mathbf{x} = (y, u, \lambda)$  of (2.7) satisfies (2.10) and since  $y$  satisfies (1.2) on each  $T_e \in \mathcal{T}_h$ , we find that

$$\begin{aligned} a(\psi_h, \lambda) &= -\langle y - \hat{y}, \psi_h \rangle, \\ b(w_h, \lambda) + \omega \langle u, w_h \rangle &= 0, \\ a_h^s(y, v_h) + b_h^s(u, v_h) &= \langle f, v_h \rangle_h^s + \langle g, v_h \rangle_{\Gamma_n} \end{aligned}$$

for all  $\psi_h, v_h \in V_h$  and  $w_h \in U_h$ . With (4.14) this implies

$$\langle \mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x}), \mathbf{z}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} = \left( \begin{array}{c} \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle -\epsilon \Delta \psi_h + \mathbf{c} \cdot \nabla \psi_h + r \psi_h, \mathbf{c} \cdot \nabla \lambda^I \rangle_{0, T_e} + \langle y^I - y, \psi_h \rangle \\ \langle \lambda - \lambda^I, v_h \rangle - \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle w_h, \mathbf{c} \cdot \nabla \lambda^I \rangle_{0, T_e} \\ a_h^s(y^I - y, v_h) - \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle Pu - u, \mathbf{c} \cdot \nabla v_h \rangle_{0, T_e} \end{array} \right). \quad (4.18)$$

**Lemma 4.6** *Let  $k, m \geq 1$  and suppose the solution  $\mathbf{x} = (y, u, \lambda)$  of (2.7) satisfies  $y, \lambda \in H^{k+1}(\Omega)$  and  $u \in H^{m+1}(\Omega)$ . If  $\tau_e^s$  satisfies (4.4) and (4.10), then*

$$\|\mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x})\|_{\mathcal{X}'} \leq C \begin{cases} (\epsilon^{1/2} + h^{1/2})h^k |y|_{k+1} + h^{m+2}\epsilon^{-1/2}|u|_{m+1} \\ + h\epsilon^{-1/2}\|\nabla \lambda^I\|_0 + h^{k+1}(|y|_{k+1} + |\lambda|_{k+1}), & \text{Pe}_e \leq 1, \\ (\epsilon^{1/2} + h^{1/2})h^k |y|_{k+1} + h^{m+3/2}|u|_{m+1} \\ + (\epsilon^{1/2} + h^{1/2})\|\nabla \lambda^I\|_0 + h^{k+1}(|y|_{k+1} + |\lambda|_{k+1}), & \text{Pe}_e > 1. \end{cases} \quad (4.19)$$

**Proof:** The terms in (4.18) can be estimated as follows. Using the Hölder inequality and (4.3) gives

$$\begin{aligned} \left| \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle -\epsilon \Delta \psi_h, \mathbf{c} \cdot \nabla \lambda^I \rangle_{0, T_e} \right| &\leq \sum_{T_e \in \mathcal{T}_h} \tau_e^s \epsilon \|\Delta \psi_h\|_{0, T_e} \|\mathbf{c}\|_{0, \infty, T_e} \|\nabla \lambda^I\|_{0, T_e} \\ &\leq C \left( \sum_{T_e \in \mathcal{T}_h} \epsilon (\tau_e^s)^2 / h^2 \|\nabla \lambda^I\|_{0, T_e}^2 \right)^{1/2} \left( \sum_{T_e \in \mathcal{T}_h} \epsilon \|\nabla \psi_h\|_{0, T_e}^2 \right)^{1/2} \end{aligned} \quad (4.20)$$

Standard estimates give

$$\begin{aligned} \left| \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle \mathbf{c} \cdot \nabla \psi_h, \mathbf{c} \cdot \nabla \lambda^I \rangle_{0, T_e} \right| &\leq \sum_{T_e \in \mathcal{T}_h} \tau_e^s \|\mathbf{c} \cdot \nabla \psi_h\|_{0, T_e} \|\mathbf{c}\|_{0, \infty, T_e} \|\nabla \lambda^I\|_{0, T_e} \\ &\leq C \left( \sum_{T_e \in \mathcal{T}_h} \tau_e^s \|\nabla \lambda^I\|_{0, T_e}^2 \right)^{1/2} \left( \sum_{T_e \in \mathcal{T}_h} \tau_e^s \|\mathbf{c} \cdot \nabla \psi_h\|_{0, T_e}^2 \right)^{1/2} \end{aligned} \quad (4.21)$$

and

$$\begin{aligned} \left| \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle r \psi_h, \mathbf{c} \cdot \nabla \lambda^I \rangle_{0, T_e} \right| &\leq \sum_{T_e \in \mathcal{T}_h} \tau_e^s \|r\|_{0, \infty, T_e} \|\psi_h\|_{0, T_e} \|\mathbf{c}\|_{0, \infty, T_e} \|\nabla \lambda^I\|_{0, T_e} \\ &\leq C \left( \sum_{T_e \in \mathcal{T}_h} (\tau_e^s)^2 \|\nabla \lambda^I\|_{0, T_e}^2 \right)^{1/2} \|\psi_h\|_0. \end{aligned} \quad (4.22)$$

The estimate (4.2) implies

$$|\langle y^I - y, \psi_h \rangle| \leq \mu_{\text{int}} h^{k+1} |y|_{k+1} \|\psi_h\|_0. \quad (4.23)$$

Combining (4.20)-(4.23) gives

$$\begin{aligned} & \left| \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle -\epsilon \Delta \psi_h + \mathbf{c} \cdot \nabla \psi_h + r \psi_h, \mathbf{c} \cdot \nabla \lambda^I \rangle_{0, T_e} + \langle y^I - y, \psi_h \rangle \right| \\ & \leq C \left[ \left( \sum_{T_e \in \mathcal{T}_h} (\epsilon (\tau_e^s)^2 / h_e^2 + \tau_e^s + (\tau_e^s)^2) \|\nabla \lambda^I\|_{0, T_e}^2 \right)^{1/2} + h^{k+1} |y|_{k+1} \right] \|\psi_h\|_{SD}. \end{aligned} \quad (4.24)$$

Analogously to (4.22), (4.23) we obtain

$$\left| \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle w_h, \mathbf{c} \cdot \nabla \lambda^I \rangle_{0, T_e} \right| + \left| \langle \lambda^I - \lambda, w_h \rangle \right| \leq C \left[ \left( \sum_{T_e \in \mathcal{T}_h} (\tau_e^s)^2 \|\nabla \lambda^I\|_{0, T_e}^2 \right)^{1/2} + h^{k+1} |\lambda|_{k+1} \right] \|w_h\|_0 \quad (4.25)$$

Using the estimates in Lemma 4.2 we find that

$$|a_h^s(y^I - y, v_h)| \leq Ch^k \left( \sum_{T_e \in \mathcal{T}_h} (\epsilon + \tau_e^s + h_e^2 / \tau_e^s + h_e^2) |y|_{k+1, T_e}^2 \right)^{1/2} \|v_h\|_{SD}. \quad (4.26)$$

Finally, using standard estimates and (4.15) we obtain

$$\begin{aligned} \left| \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle Pu - u, \mathbf{c} \cdot \nabla v_h \rangle_{0, T_e} \right| & \leq \left( \sum_{T_e \in \mathcal{T}_h} \tau_e^s \|Pu - u\|_{0, T_e}^2 \right)^{1/2} \left( \sum_{T_e \in \mathcal{T}_h} \tau_e^s \|\mathbf{c} \cdot \nabla v_h\|_{0, T_e}^2 \right)^{1/2} \\ & \leq C (\max_{T_e \in \mathcal{T}_h} \tau_e^s)^{1/2} h^{m+1} |u|_{m+1} \|v_h\|_{SD}. \end{aligned} \quad (4.27)$$

The desired results now follows from (4.18), the estimates (4.24)-(4.27), the fact that

$$\epsilon + \tau_e^s + h_e^2 / \tau_e^s + h_e^2 \leq C(\epsilon + h_e), \quad \epsilon (\tau_e^s)^2 / h_e^2 + \tau_e^s + (\tau_e^s)^2 \leq C \begin{cases} h_e^2 / \epsilon, & \text{Pe}_e \leq 1, \\ \epsilon + h_e, & \text{Pe}_e > 1, \end{cases}$$

for  $\tau_e^s$  satisfying (4.10), and

$$\|\mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x})\|_{\mathcal{X}_h^*} = \sup_{\mathbf{z}_h \neq 0} \frac{\langle \mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x}), \mathbf{z}_h \rangle_{\mathcal{X}_h^*, \mathcal{X}_h}}{\|\mathbf{z}_h\|_{\mathcal{X}_h}}. \quad (4.28)$$

□

If we apply the abstract error estimate (3.7) and combine the results in Lemmas 4.4, 4.5 and 4.6 we obtain the following error estimate.

**Theorem 4.7** *Let  $k, m \geq 1$  and suppose the solution  $(y, u, \lambda)$  of (2.7) satisfies  $y, \lambda \in H^{k+1}(\Omega)$ ,  $u \in H^{m+1}(\Omega)$ . If  $\tau_e^s$  satisfies (4.4) and (4.10) and if  $\tau_e^a$  satisfies (4.44) and (4.10), then the error between the*

solution  $(y, u, \lambda)$  of (2.7) and the solution  $(y_h, u_h, \lambda_h)$  of the discretized problem (2.15) obeys

$$\begin{aligned} & \|y_h - y\|_{SD} + \|u_h - u\|_0 + \|\lambda_h - \lambda\|_{SD} \\ & \leq C \begin{cases} (\epsilon^{1/2} + h^{1/2})h^k |y|_{k+1} + h^{m+2}\epsilon^{-1/2}|u|_{m+1} \\ + h\epsilon^{-1/2}\|\nabla\lambda^I\|_0 + h^{k+1}(|y|_{k+1} + |\lambda|_{k+1}), & \text{Pe}_e \leq 1, \\ (\epsilon^{1/2} + h^{1/2})h^k |y|_{k+1} + h^{m+3/2}|u|_{m+1} \\ + (\epsilon^{1/2} + h^{1/2})\|\nabla\lambda^I\|_0 + h^{k+1}(|y|_{k+1} + |\lambda|_{k+1}), & \text{Pe}_e > 1. \end{cases} \end{aligned} \quad (4.29)$$

Theorem 4.7 gives an estimate for the states, controls and adjoints combined. Our numerical results in Section 5 show that this error estimate is often too conservative for the states and the controls. The reason for this is that while the error  $\lambda - \lambda_h$  in the  $\|\cdot\|_{SD}$  norm behaves as in (4.29), the error  $\lambda - \lambda_h$  measured in the  $L^2$ -norm is often much smaller. Because of the optimality conditions (1.4) and (2.17b) this tends to imply a smaller error  $\|u - u_h\|_0$  in the control than the one suggested by (4.29).  $L^2$  and  $L^\infty$  error estimates for the SUPG method are discussed, e.g., in [14, 15]. See also the overview in [13, Sec. 3.2.1]. However,  $L^2$  estimates for the error  $\lambda - \lambda_h$  in the optimal control context and  $L^2$  estimates for the error  $u - u_h$  have not yet been established.

If the error  $u - u_h$  in the control is smaller than the upper bound established in (4.29), we can obtain an improved estimate for the error in the states. This is stated in the next theorem.

**Theorem 4.8** *Let  $k \geq 1$  and suppose the solution  $y$  of (2.10c) satisfies  $y \in H^{k+1}(\Omega)$ . Furthermore, let  $y_h$  solve (2.17c). If  $\tau_e^s$  satisfies (4.4) and (4.10), then there exists  $C > 0$  such that*

$$\|y - y_h\|_{SD} \leq C \left( h^k (\epsilon^{1/2} + h^{1/2}) |y|_{k+1} + \|u_h - u\|_0 \right) \quad \forall h. \quad (4.30)$$

**Proof:** Let  $\tilde{y}_h \in Y_h$  be the solution of

$$a_h^s(\tilde{y}_h, v_h) + b_h^s(u, v_h) = \langle f, v_h \rangle_h^s + \langle g, v_h \rangle_{\Gamma_n} \quad \forall v_h \in V_h. \quad (4.31)$$

Theorem 4.3 implies

$$\|y - \tilde{y}_h\|_{SD} \leq Ch^k (\epsilon^{1/2} + h^{1/2}) |y|_{k+1}.$$

To estimate  $y_h - \tilde{y}_h$  we subtract (4.31) from (2.19e),

$$a_h^s(y_h - \tilde{y}_h, v_h) + b_h^s(u_h - u, v_h) = 0 \quad \forall v_h \in V_h,$$

set  $v_h = y_h - \tilde{y}_h$  and apply Lemma 4.4 to obtain

$$\frac{1}{2} \|y_h - \tilde{y}_h\|_{SD} \leq a_h^s(y_h - \tilde{y}_h, y_h - \tilde{y}_h) = -b_h^s(u_h - u, y_h - \tilde{y}_h) \leq \|u_h - u\|_0 \|y_h - \tilde{y}_h\|_{SD}.$$

This implies the desired estimate.  $\square$

The next result indicates that the estimate (4.29) for the error in the adjoints cannot be improved. Even if we solve the discrete adjoint (2.17a) with  $y_h$  replaced by the optimal state  $y$ , i.e., the solution of (2.7), we obtain an error estimate comparable to (4.29).

**Theorem 4.9** *Let  $y$  be the optimal state, i.e., the solution of (2.7). Let  $k \geq 1$  and suppose the solution  $\lambda$  of (2.10a) satisfies  $\lambda \in H^{k+1}(\Omega)$ . Furthermore, let  $\lambda_h$  solve*

$$a_h^s(\psi_h, \lambda_h) = -\langle y - \hat{y}, \psi_h \rangle \quad \forall \psi_h \in V_h. \quad (4.32)$$

If  $\tau_e^s$  satisfies (4.4) and (4.10), then

$$\|\lambda - \lambda_h\|_{SD} \leq C \begin{cases} h\epsilon^{-1/2} \|\nabla \lambda^I\|_0, & \text{Pe}_e \leq 1, \\ (\epsilon^{1/2} + h^{1/2}) \|\nabla \lambda^I\|_0, & \text{Pe}_e > 1. \end{cases} \quad (4.33)$$

**Proof:** This result follows from the stability result (4.5) and the consistency estimates (4.20)-(4.22). All other steps in the proof of this result are analogous to those in the proof of Theorem 4.3 given, e.g., [13, Thm. 3.30] or [9, Thm. 9.3].  $\square$

#### 4.4 Optimize-Then-Discretize

In the optimize-then-discretize approach the discrete equation (3.2) corresponds to (2.19). The components of  $\mathbf{K}_h$  in (3.8) are given by

$$\begin{aligned} \langle H_h^{yy} y_h, \psi_h \rangle_{\Lambda_h^* \times \Lambda_h} &= \langle y_h, \psi_h \rangle_h^a = \langle y_h, \psi_h \rangle + \sum_{T_e \in \mathcal{T}_h} \tau_e^a \langle y_h, -\mathbf{c} \cdot \psi_h \rangle_{T_e}, \\ \langle H_h^{uu} u_h, w_h \rangle_{U_h^* \times U_h} &= \omega \langle u_h, w_h \rangle, \quad H_h^{uy} = H_h^{yu} = 0, \\ \langle A_h y_h, v_h \rangle_{V_h^* \times V_h} &= a_h^s(y_h, v_h), \quad \langle B_h u_h, v_h \rangle_{V_h^* \times V_h} = b_h^s(u_h, v_h), \\ \langle \tilde{A}_h \psi_h, \lambda_h \rangle_{\Lambda_h^* \times \Lambda_h} &= a_h^a(\psi_h, \lambda_h), \quad \langle \tilde{B}_h u_h, \lambda_h \rangle_{\Lambda_h^* \times \Lambda_h} = b_h(u_h, \lambda_h). \end{aligned} \quad (4.34)$$

As we have pointed out at the end of Section 2.4, the discretization of the optimality system leads to a non-selfadjoint system. This makes the derivation of a stability result (3.4) more complicated than in the discretize-then-optimize approach. On the other hand, derivation of a consistency estimate is just a simple application of the standard SUPG consistency estimates reviewed in Section 4.1.

We first derive a stability result. The next lemma collects some preliminary results on the solution of the state and adjoint equations as well as their approximations computed using the SUPG method.

**Lemma 4.10** *i. Let  $u \in L^2(\Omega)$ . Suppose that the solution  $z(u) \in V$  of*

$$a(z, v) = b(u, v) \quad \forall v \in V \quad (4.35)$$

*satisfies  $z(u) \in H^2(\Omega)$  and that there exists  $C > 0$  independent of  $u$  such that*

$$\|z(u)\|_2 \leq C \|u\|_0. \quad (4.36)$$

*If  $z_h(u_h)$  and  $\tilde{z}_h(u_h)$  solve*

$$a_h^s(z_h, v_h) = b_h^s(u, v_h) \quad \forall v_h \in V_h \quad (4.37)$$

*and*

$$a_h^a(\tilde{z}_h, \psi_h) = b(u, \psi_h) \quad \forall \psi_h \in \Lambda_h, \quad (4.38)$$

respectively, and if  $\tau_e^s$  satisfies (4.4) and (4.10), then there exists  $C > 0$  independent of  $u_h$  such that

$$\|z_h(u)\|_{SD} \leq C\|u\|_0, \quad \|\tilde{z}_h(u_h)\|_{SD} \leq C\|u\|_0, \quad (4.39)$$

$$\|z_h(u) - z(u)\|_{SD} \leq Ch(\epsilon^{1/2} + h^{1/2})\|u\|_0. \quad (4.40)$$

ii. Let  $z \in L^2(\Omega)$ . Suppose that the solution  $\mu(z) \in V$  of

$$a(\psi, \mu) = \langle z, \psi \rangle \quad \forall \psi \in V, \quad (4.41)$$

satisfies  $\mu(z) \in H^2(\Omega)$  and that there exists  $C > 0$  independent of  $z$  such that

$$\|\mu(z)\|_2 \leq C\|z\|_0. \quad (4.42)$$

If  $\mu_h(z) \in \Lambda_h$  solves

$$a_h^a(\psi_h, \mu_h) = \langle z, \psi_h \rangle + \sum_{T_e \in \mathcal{T}_h} \tau_e^a \langle z, -\mathbf{c} \cdot \nabla \psi_h \rangle_{T_e} \quad \forall \psi_h \in \Lambda_h \quad (4.43)$$

and if  $\tau_e^a$  satisfies

$$0 < \tau_e^a \leq \min \left\{ \frac{h_e^2}{\epsilon \mu_{\text{inv}}^2}, \frac{r_0}{\|r - \nabla \cdot \mathbf{c}\|_{0, \infty, T_e}} \right\} \quad (4.44)$$

and (4.10) with  $\tau_e^s$  replaced by  $\tau_e^a$ , then

$$\|\mu_h(z) - \mu(z)\|_{SD} \leq Ch(\epsilon^{1/2} + h^{1/2})\|z\|_0. \quad (4.45)$$

**Proof:** The inequalities (4.39) follow from the SUPG stability estimate Lemma 4.1, inequalities (4.40), (4.45) follow from the SUPG convergence theory, cf. Theorem 4.3.  $\square$

**Lemma 4.11** *Let the assumptions of Lemma 4.10 be satisfied. Let  $k, \ell, m \geq 1$  and suppose the solution  $\mathbf{x} = (y, u, \lambda)$  of (2.7) satisfies  $y \in H^{k+1}(\Omega)$ ,  $u \in H^{m+1}(\Omega)$ ,  $\lambda \in H^{\ell+1}(\Omega)$ . If  $\tau_e^s$  satisfies (4.4) and (4.10) and if  $\tau_e^a$  satisfies (4.44) and (4.10) with  $\tau_e^s$  replaced by  $\tau_e^a$ , then there exist  $\bar{h} > 0$  and  $\kappa > 0$  such that  $\mathbf{K}_h$  is invertible for all  $h \geq \bar{h}$  and  $\|\mathbf{K}_h^{-1}\|_h \leq \kappa$  for all  $h \geq \bar{h}$ .*

**Proof:** We apply Lemma 3.1. By Lemma 4.1  $A_h$  and  $\tilde{A}_h$  are invertible and satisfy  $\|A_h^{-1}\|_{\mathcal{L}(Y_h^*, Y_h)} \leq 2$ ,  $\|\tilde{A}_h^{-1}\|_{\mathcal{L}(\Lambda_h^*, \Lambda_h)} \leq 2$ . It is straightforward to verify that there exists  $c > 0$  such that  $\|B_h\|_{\mathcal{L}(U_h, Y_h^*)} \leq c$ ,  $\|\tilde{B}_h\|_{\mathcal{L}(U_h, \Lambda_h^*)} \leq c$  and  $\|H_h^{yy}\|_{\mathcal{L}(Y_h, \Lambda_h^*)} \leq c$  for all  $h$ .

The proof of uniform boundedness of  $\hat{H}_h^{-1}$  is a little bit more involved than in Lemma 4.5 because  $\tilde{B}_h^*(\tilde{A}_h^*)^{-1}H_h^{yy}A_h^{-1}B_h$  is not symmetric.

By Definition 4.34 of  $A_h$ ,  $\tilde{A}_h$ ,  $B_h$  and  $\tilde{B}_h$  the vectors  $z_h = A_h^{-1}B_h u_h$  and  $\tilde{z}_h = \tilde{A}_h^{-1}\tilde{B}_h u_h$  solve (4.37) and (4.38), respectively, with  $u = u_h$ . Let  $z$  be the solution of (4.35) with  $u = u_h$ . From the Definition 3.10 of  $\hat{H}_h$  we obtain

$$\begin{aligned} & \langle \hat{H}_h u_h, u_h \rangle_{U_h^* \times U_h} \\ &= \langle H_h^{uu} u_h, u_h \rangle_{U_h^* \times U_h} + \langle \tilde{B}_h^* (\tilde{A}_h^*)^{-1} H_h^{yy} A_h^{-1} B_h u_h, u_h \rangle_{U_h^* \times U_h} \\ &= \langle H_h^{uu} u_h, u_h \rangle_{U_h^* \times U_h} + \langle H_h^{yy} A_h^{-1} B_h u_h, \tilde{A}_h^{-1} \tilde{B}_h u_h \rangle_{\Lambda_h^* \times \Lambda_h} \\ &= \omega \|u_h\|_0^2 + \langle H_h^{yy} z, z \rangle_{\Lambda_h^* \times \Lambda_h} + \langle H_h^{yy} (z_h - z), \tilde{z}_h \rangle_{\Lambda_h^* \times \Lambda_h} + \langle H_h^{yy} z, \tilde{z}_h - z \rangle_{\Lambda_h^* \times \Lambda_h}. \end{aligned} \quad (4.46)$$

Applying the definition (4.34) of  $H_h^{yy}$  and (4.36)-(4.40), we obtain

$$\begin{aligned} \langle H_h^{yy} z, z \rangle_{\Lambda_h^* \times \Lambda_h} &= \|z\|_0^2 + \sum_{T_e \in \mathcal{T}_h} \tau_e^a \langle z, -\mathbf{c} \cdot \nabla z \rangle_{T_e} \\ &\geq \|z\|_0^2 - \max_{T_e \in \mathcal{T}_h} \tau_e^a \|\mathbf{c}\|_{0,\infty} \|z\|_0 \|z\|_1 \\ &\geq \|z\|_0^2 - C \max_{T_e \in \mathcal{T}_h} \tau_e^a \|\mathbf{c}\|_{0,\infty} \|u_h\|_0^2 \end{aligned} \quad (4.47)$$

and

$$\begin{aligned} \langle H_h^{yy}(z_h - z), \tilde{z}_h \rangle_{\Lambda_h^* \times \Lambda_h} &\geq -\|z_h - z\|_0 \|\tilde{z}_h\|_0 - \max_{T_e \in \mathcal{T}_h} \tau_e^a \|\mathbf{c}\|_{0,\infty} \|z_h - z\|_0 \|\tilde{z}_h\|_1 \\ &\geq -Ch(\epsilon^{1/2} + h^{1/2}) \left( 1 + \|\mathbf{c}\|_{0,\infty} \max_{T_e \in \mathcal{T}_h} \tau_e^a \right) \|u_h\|_0^2. \end{aligned} \quad (4.48)$$

To estimate the last term in (4.46) we set we set  $\tilde{\mu} = (A^*)^{-1} H_h^{yy} z$  and  $\mu_h = (\tilde{A}^*)^{-1} H_h^{yy} z$ . The identity  $A^* \tilde{\mu} = H_h^{yy} z$  and the definitions of  $A$  and  $H_h^{yy}$  imply that

$$a(\psi, \tilde{\mu}) = \langle z, \psi \rangle + \sum_{T_e \in \mathcal{T}_h} \tau_e^a \langle z, -\mathbf{c} \cdot \nabla \psi \rangle_{T_e} \quad \forall \psi \in V.$$

Similarly,  $\mu_h$  solves (4.43). If  $\mu$  solves (4.41), then the Lipschitz continuity of the solution of the adjoint equation with respect to perturbations in the right hand side implies

$$\|\mu - \tilde{\mu}\|_1 \leq C \max_{T_e \in \mathcal{T}_h} \tau_e^a \|\mathbf{c}\|_{1,\infty} \|z\|_1 \leq C \max_{T_e \in \mathcal{T}_h} \tau_e^a \|\mathbf{c}\|_{1,\infty} \|u_h\|_0.$$

(To obtain the last inequality, recall that  $z$  solves (4.35) with  $u = u_h$  and that (4.36) holds with  $u = u_h$ .) If  $\mu$  solves (4.41), then (4.45) and (4.36) imply

$$\|\mu - \mu_h\|_{SD} \leq Ch(\epsilon^{1/2} + h^{1/2}) \|z\|_0 \leq Ch(\epsilon^{1/2} + h^{1/2}) \|u_h\|_0.$$

Hence

$$\begin{aligned} \langle H_h^{yy} z, \tilde{z}_h - z \rangle_{\Lambda_h^* \times \Lambda_h} &= \langle (A^*)^{-1} H_h^{yy} z - (\tilde{A}^*)^{-1} H_h^{yy} z, \tilde{B}_h u_h \rangle_{\Lambda_h \times \Lambda_h^*} \\ &= \langle \tilde{\mu} - \mu_h, \tilde{B}_h u_h \rangle_{\Lambda_h \times \Lambda_h^*} \geq \|\tilde{\mu} - \mu_h\|_{SD} \|\tilde{B}_h u_h\|_{\Lambda_h^*} \\ &\geq C \left( \max_{T_e \in \mathcal{T}_h} \tau_e^a + h(\epsilon^{1/2} + h^{1/2}) \right) \|u_h\|_0^2. \end{aligned} \quad (4.49)$$

Estimates (4.46)-(4.49) imply the existence of  $\bar{h} > 0$  such that  $\langle \hat{H}_h u_h, u_h \rangle_{U_h^* \times U_h} \geq \frac{1}{2} \omega \|u_h\|_0^2$ . This implies  $\|\hat{H}_h^{-1}\|_{\mathcal{L}(U_h^*, U_h)} \leq 2\omega^{-1}$ . The desired result now follows from Lemma 3.1.  $\square$

Now we turn to the consistency, i.e., we want to find an estimate for  $\|\mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x})\|$  in our abstract error estimate (3.7). The discretized optimality condition (2.19) imply that

$$\langle \mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x}), \mathbf{z} \rangle_{\mathcal{X}_h^* \times \mathcal{X}_h} = \begin{pmatrix} a_h^a(\psi_h, \lambda^I) + \langle y^I - \hat{y}, \psi_h \rangle_h^a \\ b(w_h, \lambda^I) + \omega \langle Pu, w_h \rangle \\ a_h^s(y^I, v_h) + b_h^s(Pu, v_h) - \langle f, v_h \rangle_h^s - \langle g, v_h \rangle_{\Gamma_n} \end{pmatrix}.$$

Since the solution  $\mathbf{x} = (y, u, \lambda)$  of (2.7) satisfies (2.10) and that  $y$  satisfies (1.2) on each  $T_e \in \mathcal{T}_h$ , we have

$$\begin{aligned} a_h^a(\psi_h, \lambda) &= -\langle y - \hat{y}, \psi_h \rangle_h^a, \\ b(w_h, \lambda) + \omega \langle u, w_h \rangle &= 0, \\ a_h^s(y, v_h) + b_h^s(u, v_h) &= \langle f, v_h \rangle_h^s + \langle g, v_h \rangle_{\Gamma_n} \end{aligned}$$

for all  $\psi_h, v_h \in V_h$  and  $w_h \in U_h$ . With (4.14) this implies

$$\langle \mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x}), \mathbf{z} \rangle_{\mathcal{X}_h^*, \mathcal{X}_h} = \begin{pmatrix} a_h^a(\psi_h, \lambda^I - \lambda) + \langle y^I - y, \psi_h \rangle_h^a \\ b(w_h, \lambda^I - \lambda) \\ a_h^s(y^I - y, v_h) - \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle Pu - u, \mathbf{c} \cdot v_h \rangle_{0, T_e} \end{pmatrix}.$$

**Lemma 4.12** *Let  $k, \ell, m \geq 1$  and suppose the solution  $\mathbf{x} = (y, u, \lambda)$  of (2.7) satisfies  $y \in H^{k+1}(\Omega)$ ,  $u \in H^{m+1}(\Omega)$ ,  $\lambda \in H^{\ell+1}(\Omega)$ . If  $\tau_e^s$  satisfies (4.4) and (4.10) and if  $\tau_e^a$  satisfies (4.44) and (4.10), then*

$$\|\mathbf{r}_h - \mathbf{K}_h \mathbf{R}_h(\mathbf{x})\|_{\mathcal{X}} \leq C \left( (\epsilon^{1/2} + h^{1/2})(h^k |y|_{k+1} + h^\ell |\lambda|_{\ell+1}) + h^{m+1} |u|_{m+1} \right). \quad (4.50)$$

**Proof:** Using the estimates in Lemma 4.2 we find that

$$\begin{aligned} a_h^s(y^I - y, v_h) &\leq Ch^k \left( \sum_{T_e \in \mathcal{T}_h} (\epsilon + \tau_e^s + h_e^2/\tau_e^s + h_e^2) |y|_{k+1, T_e}^2 \right)^{1/2} \|v_h\|_{SD}, \\ a_h^a(\psi_h, \lambda^I - \lambda) &\leq Ch^\ell \left( \sum_{T_e \in \mathcal{T}_h} (\epsilon + \tau_e^a + h_e^2/\tau_e^a + h_e^2) |\lambda|_{\ell+1, T_e}^2 \right)^{1/2} \|\psi_h\|_{SD}. \end{aligned}$$

Furthermore, (4.2) implies

$$b(w_h, \lambda^I - \lambda) \leq \mu_{\text{int}} h^{\ell+1} |\lambda|_{\ell+1} \|w_h\|_0.$$

By (4.27),

$$\left| \sum_{T_e \in \mathcal{T}_h} \tau_e^s \langle Pu - u, \mathbf{c} \cdot v_h \rangle_{0, T_e} \right| \leq C \max_{T_e \in \mathcal{T}_h} \tau_e^s h^{m+1} |u|_{m+1} \|v_h\|_{SD}.$$

Finally

$$\begin{aligned} \langle y^I - y, \psi_h \rangle_h^a &\leq \mu_{\text{int}} h^{k+1} |y|_{k+1} \|\psi_h\|_0 + \sum_{T_e \in \mathcal{T}_h} \tau_e^a \mu_{\text{int}} h_{T_e}^{k+1} |y|_{k+1, T_e} \|\mathbf{c} \cdot \nabla \psi_h\|_{0, T_e} \\ &\leq Ch^{k+1} |y|_{k+1} \|\psi_h\|_{SD}. \end{aligned}$$

The inequality (4.50) is obtained by combining the above estimates, using that (4.10) implies  $\epsilon + \tau_e + h_e^2/\tau_e + h_e^2 \leq C(\epsilon + h_e)$  and the identity (4.28).  $\square$

If we apply the abstract error estimate (3.7) and combine the results in Lemmas 4.4, 4.11 and 4.12 we obtain the following error estimate.

**Theorem 4.13** *Let  $k, \ell, m \geq 1$  and suppose the solution  $(y, u, \lambda)$  of (2.7) satisfies  $y \in H^{k+1}(\Omega)$ ,  $u \in H^{m+1}(\Omega)$ ,  $\lambda \in H^{\ell+1}(\Omega)$ . If  $\tau_e^s$  satisfies (4.4) and (4.10) and if  $\tau_e^a$  satisfies (4.44) and (4.10), then the error between the solution  $(y, u, \lambda)$  of (2.7) and the solution  $(y_h, u_h, \lambda_h)$  of the discretized optimality conditions (2.19) obeys*

$$\begin{aligned} & \|y_h - y\|_{SD} + \|u_h - u\|_0 + \|\lambda_h - \lambda\|_{SD} \\ & \leq C \left( (\epsilon^{1/2} + h^{1/2})(h^k |y|_{k+1} + h^\ell |\lambda|_{\ell+1}) + h^{m+1} |u|_{m+1} \right). \end{aligned} \quad (4.51)$$

As in the case of Theorem 4.7, Theorem 4.13 also gives an estimate for the states, controls and adjoints combined. This error estimate is sometimes too conservative for the controls for the same reasons sketched after Theorem 4.7. As in the discretize-then-optimize case,  $L^2$  estimates for the error  $\lambda - \lambda_h$  in the optimal control context and  $L^2$  estimates for the error  $u - u_h$  have not yet been established.

If the error  $u - u_h$  in the control is smaller than the upper bound established in (4.51), we can obtain an improved estimate for the error in the states and in the adjoints. This is stated in the next theorem. Note that if a better estimate for the error  $\lambda - \lambda_h$  can be obtained, this might also allow to further improve the error  $u - u_h$  in the control.

**Theorem 4.14** *Let  $k, \ell \geq 1$  and suppose the solution  $(y, u, \lambda)$  of (2.7) satisfies  $y \in H^{k+1}(\Omega)$ ,  $\lambda \in H^{\ell+1}(\Omega)$ . Furthermore, let  $y_h, \lambda_h$  solve (2.19a) and (2.19e), respectively. If  $\tau_e^s$  satisfies (4.4) and (4.10) and if  $\tau_e^a$  satisfies (4.44) and (4.10), then there exists  $C > 0$  such that*

$$\|y - y_h\|_{SD} \leq C \left( h^k (\epsilon^{1/2} + h^{1/2}) |y|_{k+1} + \|u_h - u\|_0 \right) \quad \forall h, \quad (4.52)$$

$$\|\lambda - \lambda_h\|_{SD} \leq C \left( h^\ell (\epsilon^{1/2} + h^{1/2}) |\lambda|_{\ell+1} + \|y_h - y\|_0 \right) \quad \forall h. \quad (4.53)$$

**Proof:** The proof is analogous to the proof of Theorem 4.8.  $\square$

A comparison of Theorems 4.7 and 4.13 indicates that the optimize-then-discretize approach leads to asymptotically better approximate solutions of the optimal control problem than the discretize-then-optimize approach, because the estimate (4.51) is dominated by the term  $h\epsilon^{-1/2} \|\nabla \lambda^I\|_0$  and  $(\epsilon^{1/2} + h^{1/2}) \|\nabla \lambda^I\|_0$ , respectively. The differences between the error bounds provided in Theorems 4.7 and 4.13 is small when piecewise linear polynomials are used for the discretization states, adjoints and controls, i.e., if  $k = \ell = m = 1$ . We also note that in the case of piecewise linear finite elements the contributions  $\langle -\epsilon \Delta y_h, \mathbf{c} \cdot \nabla v_h \rangle_{T_e}$  and  $\langle -\epsilon \Delta \lambda_h, -\mathbf{c} \cdot \nabla \psi_h \rangle_{T_e}$  of the SUPG method to the bilinear forms (2.13) and (2.19b) disappear and, hence, one source of difference between the discretize-then-optimize approach and the optimize-then-discretize approach is eliminated. This would not apply if reconstructions of second derivative terms had been used [7].

The differences between the discretize-then-optimize approach and the optimize-then-discretize approach are the greater the larger  $k, \ell$ , i.e., the higher the order of finite elements used for the states and the adjoints. Theorem 4.9 and Theorems 4.13, 4.14 show that there is a significant difference in quality of the adjoints computed by the discretize-then-optimize approach and the optimize-then-discretize approach. The latter leads to better adjoint approximations. This is confirmed by our numerical results reported in Section 5. However, our numerical results also show that this large difference in the quality of the adjoints does not necessarily implies a large difference in the quality of the controls. Often the observed error in

the controls computed by the discretize-then-optimize approach and the optimize-then-discretize approach is very similar, which by Theorem 4.8 and 4.14 leads to very similar errors in the computed states for both approaches.

## 5 Numerical Results

### 5.1 Example 1

Our first example is a one dimensional control problem on  $\Omega = (0, 1)$  with state equation

$$-\epsilon y''(x) + y'(x) = f(x) + u(x) \text{ on } (0, 1), \quad y(0) = y(1) = 0. \quad (5.1)$$

The functions  $f$  and  $\hat{y}$  are chosen so that the solution of the optimal control problem is

$$y_{\text{ex}}(x) = x - \frac{\exp((x-1)/\epsilon) - \exp(-1/\epsilon)}{1 - \exp(-1/\epsilon)}, \quad \lambda_{\text{ex}}(x) = \left(1 - x - \frac{\exp(-x/\epsilon) - \exp(-1/\epsilon)}{1 - \exp(-1/\epsilon)}\right)$$

and  $u_{\text{ex}} = \omega^{-1} \lambda_{\text{ex}}$ . This example is modeled after [13, pp. 2,3]. We set  $\epsilon = 0.0025$  and  $\omega = 1$ . In our numerical tests we use equidistant grids with mesh size  $h$ . If piecewise linear functions are used, the stabilization parameter for the state and adjoint equation is chosen to be

$$\tau_e = \begin{cases} h^2/(4\epsilon), & \text{Pe}_e \leq 1, \\ h/2 & \text{Pe}_e > 1. \end{cases} \quad (5.2)$$

For piecewise quadratic finite elements the stabilization parameter for the state and adjoint equation is given by (5.2) with  $h$  replaced by  $h/2$ .

Errors and estimated convergence order for the discretize-then-optimize approach as well as the optimize-then-discretize approach using linear ( $k = \ell = m = 1$ ) and quadratic ( $k = \ell = m = 2$ ) finite elements are given in Tables 5.1 and 5.2. In all examples we estimate the convergence order by taking the logarithm with base two of the quotient of the error at grid size  $h$  and the error at grid size  $h/2$ . In all examples the linear systems arising from the discretization of the discretized optimal control problem or from the discretization of the infinite dimensional optimality conditions are solved using a sparse LU-decomposition.

For this example  $\text{Pe}_e = 1$  for  $h = 5 \cdot 10^{-3}$ , i.e., half of the data in Tables 5.1 and 5.2 correspond to the case  $\text{Pe}_e \leq 1$ .

If linear finite elements are used, i.e., if  $k = m = \ell = 1$ , Theorems 4.7 and 4.13 predict that the error for states, controls and adjoints behaves like  $O(h)$  for  $\text{Pe}_e \leq 1$ . This is confirmed by the results in Table 5.1. Table 5.1 reveals few differences between the discretize-then-optimize and the optimize-then-discretize approach. If linear finite elements are used, both produce states and adjoints that are of the same quality. The controls computed using the discretize-then-optimize approach are slightly better than the controls obtained from the optimize-then-discretize approach. However, we have seen examples where the opposite is true.

The situation changes if quadratic finite elements are used, i.e., if  $k = m = \ell = 2$ . In this case Table 5.2 shows that convergence order for the adjoints computed using discretize-then-optimize approach is one, whereas the convergence order for the adjoints obtained from the optimize-then-discretize approach is two. This agrees with the theoretical results in Theorems 4.7, 4.9 and in Theorem 4.13, respectively. However, in both cases the observed convergence order for the  $L^2$  error in the adjoints is one higher than the convergence order for the SUPG-error. The  $L^2$  error for the controls is much smaller than the SUPG-error

in the states. In fact, the term  $h^k(\epsilon^{1/2} + h^{1/2})|y|_{k+1}$  dominates  $\|u_h - u_{\text{ex}}\|_0$  and Theorem 4.8 predicts that in the discretize-then-optimize approach the states converge with order two, instead of the pessimistic state error bound provided by Theorem 4.7. For the optimize-then-discretize approach Theorem 4.13 predicts that order of convergence in the SUPG error for both the states and the adjoints is two. The observed convergence order for the  $L^2$  error in the adjoints is one higher than the convergence order for the SUPG-error. Since the controls are multiples of the adjoints, the observed convergence order for the  $L^2$  error in the controls is three, one higher than the convergence order predicted by Theorem 4.13.

Table 5.1: Errors and estimated convergence order. Example 1,  $k = \ell = m = 1$ .

The discretize-then-optimize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
1.00e-1	1.86e-1		6.21e-1		7.76e-2		1.74e-1		6.20e-1	
5.00e-2	1.24e-1	0.59	7.47e-1	-0.27	4.45e-2	0.80	1.19e-1	0.54	7.47e-1	-0.27
2.50e-2	7.69e-2	0.69	1.11e+0	-0.57	2.18e-2	1.03	7.51e-2	0.67	1.11e+0	-0.57
1.25e-2	4.65e-2	0.73	9.86e-1	0.17	1.37e-2	0.67	4.58e-2	0.71	9.87e-1	0.17
6.25e-3	2.64e-2	0.82	6.42e-1	0.62	1.03e-2	0.41	2.61e-2	0.81	6.43e-1	0.62
3.13e-3	9.58e-3	1.46	3.01e-1	1.09	5.05e-3	1.03	9.50e-3	1.46	3.01e-1	1.09
1.56e-3	2.57e-3	1.90	1.35e-1	1.16	1.55e-3	1.70	2.55e-3	1.90	1.35e-1	1.16
7.81e-4	6.54e-4	1.97	6.48e-2	1.06	4.15e-4	1.90	6.49e-4	1.97	6.48e-2	1.06

The optimize-then-discretize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
1.00e-1	1.85e-1		6.35e-1		1.83e-1		1.83e-1		6.76e-1	
5.00e-2	1.23e-1	0.58	7.50e-1	-0.24	1.23e-1	0.58	1.23e-1	0.58	7.58e-1	-0.16
2.50e-2	7.66e-2	0.69	1.11e+0	-0.57	7.63e-2	0.69	7.63e-2	0.69	1.11e+0	-0.55
1.25e-2	4.63e-2	0.73	9.86e-1	0.17	4.61e-2	0.73	4.61e-2	0.73	9.86e-1	0.17
6.25e-3	2.63e-2	0.82	6.42e-1	0.62	2.62e-2	0.82	2.62e-2	0.82	6.41e-1	0.62
3.13e-3	9.56e-3	1.46	3.01e-1	1.09	9.52e-3	1.46	9.52e-3	1.46	3.01e-1	1.09
1.56e-3	2.56e-3	1.90	1.35e-1	1.16	2.55e-3	1.90	2.55e-3	1.90	1.35e-1	1.16
7.81e-4	6.52e-4	1.97	6.48e-2	1.06	6.50e-4	1.97	6.50e-4	1.97	6.47e-2	1.06

We have obtained qualitatively similar results when the choice (5.2) of the stabilization parameter is replaced by  $\tau_e = (|b|h/2)(\coth(\text{Pe}_e) + 1/\text{Pe}_e)$ , which applied to certain classes of state equations with fixed control gives approximations that are nodally exact [2], [13, p. 234].

Table 5.2: Errors and estimated convergence order. Example 1,  $k = \ell = m = 2$ .

The discretize-then-optimize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
1.00e-1	1.21e-1		6.11e-1		3.89e-2		1.21e-1		5.91e-1	
5.00e-2	7.61e-2	0.67	3.62e-1	0.76	2.00e-2	0.96	7.99e-2	0.60	3.55e-1	0.74
2.50e-2	3.89e-2	0.97	5.15e-1	-0.51	4.23e-3	2.24	4.70e-2	0.77	5.93e-1	-0.74
1.25e-2	1.73e-2	1.17	4.44e-1	0.21	6.53e-3	-0.63	2.92e-2	0.69	6.32e-1	-0.09
6.25e-3	3.71e-3	2.22	1.65e-1	1.43	3.83e-3	0.77	1.22e-2	1.26	3.85e-1	0.71
3.13e-3	4.23e-4	3.13	4.21e-2	1.97	1.29e-3	1.57	3.54e-3	1.78	1.92e-1	1.00
1.56e-3	4.56e-5	3.21	1.04e-2	2.02	3.62e-4	1.83	9.27e-4	1.93	9.59e-2	1.00
7.81e-4	5.33e-6	3.10	2.58e-3	2.01	9.45e-5	1.94	2.35e-4	1.98	4.79e-2	1.00

The optimize-then-discretize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
1.00e-1	1.21e-1		6.19e-1		1.20e-1		1.20e-1		6.40e-1	
5.00e-2	7.59e-2	0.67	3.64e-1	0.77	7.56e-2	0.67	7.56e-2	0.67	3.69e-1	0.79
2.50e-2	3.88e-2	0.97	5.14e-1	-0.50	3.87e-2	0.97	3.87e-2	0.97	5.13e-1	-0.48
1.25e-2	1.72e-2	1.17	4.44e-1	0.21	1.72e-2	1.17	1.72e-2	1.17	4.44e-1	0.21
6.25e-3	3.70e-3	2.22	1.65e-1	1.43	3.70e-3	2.22	3.70e-3	2.22	1.65e-1	1.43
3.13e-3	4.23e-4	3.13	4.21e-2	1.97	4.23e-4	3.13	4.23e-4	3.13	4.21e-2	1.97
1.56e-3	4.56e-5	3.21	1.04e-2	2.02	4.56e-5	3.21	4.56e-5	3.21	1.04e-2	2.02
7.81e-4	5.33e-6	3.10	2.58e-3	2.01	5.33e-6	3.10	5.33e-6	3.10	2.58e-3	2.01

## 5.2 Example 2

The second example is a two dimensional control problem on  $\Omega = (-1, 1) \times (0, 1)$ . We use the data

$$\Gamma_n = (0, 1) \times \{0\}, \quad \Gamma_d = \partial\Omega \setminus \Gamma_n, \quad \mathbf{c}(x) = \begin{pmatrix} 2x_2(1 - x_1^2) \\ -2x_1(1 - x_2^2) \end{pmatrix}, \quad r = 0,$$

$\epsilon = 10^{-5}$  and  $\omega = 10^{-2}$ . The functions  $f$ ,  $d$ ,  $g$  and  $\hat{y}$  are chosen so that the solution of the optimal control problem (1.1), (1.2) is given by

$$y_{\text{ex}}(x) = 1 + \tanh(1 - (2(x_1^2 + x_2^2)^{1/2} + 1)), \quad \lambda_{\text{ex}}(x) = (x_1^2 - 1)x_2^2(x_2 - 1)$$

and  $u_{\text{ex}} = \omega^{-1}\lambda_{\text{ex}}$ . This example is motivated by the first model problem in [11, pp. 9,10]. Our triangulation is computed by first subdividing  $\Omega$  into squares of size  $h \times h$  and then dividing each square into two triangles. If piecewise linear functions are used, the stabilization parameter for the state and adjoint equation is chosen to be

$$\tau_e = \begin{cases} h^2/(4\epsilon), & \text{Pe}_e \leq 1, \\ h/(2\|\mathbf{c}\|_{0,\infty,T_e}), & \text{Pe}_e > 1. \end{cases} \quad (5.3)$$

Table 5.3: Errors and estimated convergence order. Example 2,  $k = \ell = m = 1$ .

The discretize-then-optimize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
2.00e-1	1.69e-1		1.12e+0		8.27e-1		1.67e-2		5.32e-2	
1.00e-1	8.52e-2	0.99	6.11e-1	0.87	4.27e-1	0.95	9.18e-3	0.86	2.80e-2	0.93
5.00e-2	2.35e-2	1.86	2.58e-1	1.24	1.77e-1	1.27	4.60e-3	1.00	1.20e-2	1.22
2.50e-2	4.60e-3	2.35	9.50e-2	1.44	4.69e-2	1.92	2.21e-3	1.06	3.47e-3	1.79
1.25e-2	8.50e-4	2.44	3.35e-2	1.50	7.75e-3	2.60	1.09e-3	1.02	8.66e-4	2.00
6.25e-3	1.96e-4	2.12	1.18e-2	1.51	1.17e-3	2.73	5.48e-4	0.99	2.82e-4	1.62

The optimize-then-discretize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
2.00e-1	1.76e-1		1.12e+0		7.51e-1		7.51e-3		4.78e-2	
1.00e-1	8.64e-2	1.03	6.12e-1	0.87	4.14e-1	0.86	4.14e-3	0.86	2.59e-2	0.88
5.00e-2	2.38e-2	1.86	2.59e-1	1.24	1.74e-1	1.25	1.74e-3	1.25	1.13e-2	1.20
2.50e-2	4.64e-3	2.36	9.50e-2	1.45	4.66e-2	1.90	4.66e-4	1.90	3.13e-3	1.85
1.25e-2	8.55e-4	2.44	3.35e-2	1.50	7.72e-3	2.59	7.72e-5	2.59	6.75e-4	2.21
6.25e-3	1.97e-4	2.12	1.18e-2	1.51	1.17e-3	2.72	1.17e-5	2.72	2.04e-4	1.73

For piecewise quadratic finite elements the stabilization parameter for the state and adjoint equation is given by (5.2) with  $h$  replaced by  $h/2$ .

Errors and estimated convergence order for the discretize-then-optimize approach as well as the optimize-then-discretize approach using linear ( $k = \ell = m = 1$ ) and quadratic ( $k = \ell = m = 2$ ) finite elements are given in Tables 5.3 and 5.4. All the data in Tables 5.3 and 5.4 correspond to the case  $\text{Pe} > 1$ . The sizes of the smallest and largest systems (2.17) and (2.19) arising in our calculations are  $198 \times 198$  and  $155043 \times 155043$ , respectively. To avoid contamination of the convergence errors by the truncation of an iterative scheme, these systems were solved using a sparse LU-decomposition.

In this example our exact adjoints and controls are designed to be functions with small gradients. If linear finite element approximations are used, the observed convergence order for the SUPG error in the computed adjoints and the computed states is greater than one for both approaches. See Table 5.3. The observed convergence order for the  $L^2$ -error in the computed adjoints is only one for the discretize-then-optimize approach, while for the optimize-then-discretize approach the observed convergence order for the  $L^2$ -error in the computed adjoints is one higher than the observed convergence order for the SUPG-error. In this example the optimize-then-discretize approach produced better approximations.

If quadratic finite elements are used, i.e., if  $k = m = \ell = 2$ , then the optimize-then-discretize approach leads to superior results. See Table 5.4. For this approach, the observed convergence order for the SUPG error in the computed adjoints and the computed states is greater than the guaranteed convergence order of two. The observed convergence order for the  $L^2$ -error in the computed adjoints is one higher than the observed convergence order for the SUPG-error, which leads to an observed convergence order greater than three for the  $L^2$ -error in the controls. This is different for discretize-then-optimize approach. Initially the

Table 5.4: Errors and estimated convergence order. Example 2,  $k = \ell = m = 2$ .

The discretize-then-optimize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
2.00e-1	5.60e-2		6.64e-1		3.22e-1		8.34e-3		2.91e-2	
1.00e-1	1.43e-2	1.97	2.38e-1	1.48	9.68e-2	1.73	4.24e-3	0.98	1.07e-2	1.45
5.00e-2	1.92e-3	2.89	5.41e-2	2.14	1.62e-2	2.58	2.14e-3	0.99	2.77e-3	1.94
2.50e-2	3.52e-4	2.45	1.28e-2	2.08	2.78e-3	2.54	1.08e-3	0.98	1.33e-3	1.06
1.25e-2	1.27e-4	1.47	8.01e-3	0.67	2.75e-3	0.02	5.46e-4	0.99	1.99e-3	-0.58

The optimize-then-discretize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
2.00e-1	5.74e-2		6.66e-1		3.08e-1		3.08e-3		2.54e-2	
1.00e-1	1.45e-2	1.99	2.39e-1	1.48	9.43e-2	1.71	9.43e-4	1.71	8.97e-3	1.50
5.00e-2	1.92e-3	2.91	5.39e-2	2.15	1.53e-2	2.62	1.53e-4	2.62	1.59e-3	2.49
2.50e-2	3.26e-4	2.56	1.12e-2	2.26	1.26e-3	3.61	1.26e-5	3.61	1.49e-4	3.41
1.25e-2	5.66e-5	2.53	2.23e-3	2.33	1.23e-4	3.35	1.23e-6	3.35	1.97e-5	2.92

observed convergence orders for the SUPG-errors in states and adjoints as well as the  $L^2$ -error in control are comparable to those achieved by the optimize-then-discretize approach, but deteriorates subsequently.

We note that in this example the gradients of  $y_h$  are almost perpendicular to  $\mathbf{c}$ . Because of this feature, the standard Galerkin finite element method produced good solutions to the optimal control problem.

### 5.3 Example 3

In our third example we use  $\Omega = (0, 1)^2$  and

$$\Gamma_d = \partial\Omega, \quad \mathbf{c}(x) = (\cos(\theta) \sin(\theta))^T, \quad \theta = 45^\circ, \quad r = 0,$$

$\epsilon = 10^{-2}$  and  $\omega = 1$ . The functions  $f$ ,  $d$ ,  $g$  and  $\hat{y}$  are chosen so that the solution of the optimal control problem (1.1), (1.2) is given by

$$y_{\text{ex}}(x) = \eta(x_1)\eta(x_2), \quad \lambda_{\text{ex}}(x) = \mu(x_1)\mu(x_2),$$

where

$$\eta(z) = z - \frac{\exp((z-1)/\epsilon) - \exp(-1/\epsilon)}{1 - \exp(-1/\epsilon)}, \quad \mu(z) = \left(1 - z - \frac{\exp(-z/\epsilon) - \exp(-1/\epsilon)}{1 - \exp(-1/\epsilon)}\right)$$

and  $u_{\text{ex}} = \omega^{-1}\lambda_{\text{ex}}$ . Our triangulation is computed by first subdividing  $\Omega$  into squares of size  $h \times h$  and then dividing each square into two triangles. Our choice for the stabilization parameter is the same as in Section 5.2. Errors and estimated convergence order for the discretize-then-optimize approach as well as the optimize-then-discretize approach using linear ( $k = \ell = m = 1$ ) and quadratic ( $k = \ell = m = 2$ ) finite elements are given in Tables 5.5 and 5.6. All but the last row in Table 5.5 and all data in Table 5.6

correspond to the case  $Pe_\epsilon > 1$ . The sizes of the smallest and largest systems (2.17) and (2.19) arising in our calculations are  $363 \times 363$  and  $77763 \times 77763$ , respectively. These systems were solved using a sparse LU-decomposition.

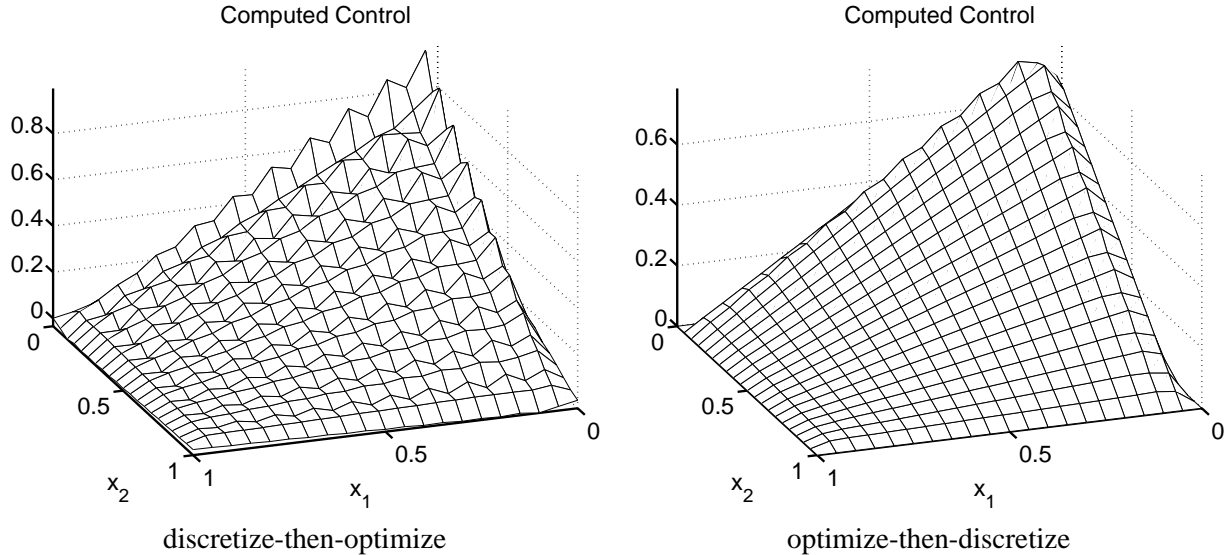


Figure 5.1: Computed controls, Example 3,  $k = \ell = m = 2$ ,  $h = 0.1$ .

The observations drawn from Tables 5.5 and 5.6 for this example are very similar to those of Example 1. However, when quadratic finite elements are used, we observed small node-to-node oscillations in the adjoints and controls computed by the discretize-then optimize approach. These are not present in the adjoints and controls computed by the optimize-then-discretize approach. See Figure 5.1. Such node-to-node oscillations in the adjoints and controls did not develop in either approach when linear finite elements are used as seen in Figure 5.2. For better visibility we show the results on a coarse grid, but the plots of our results on finer grid are qualitatively comparable to those in Figures 5.1 and 5.2.

We remark that for this example the standard Galerkin method produced poor results. For smaller diffusion  $\epsilon$  even SUPG using either approach did not produce satisfactory approximations to the optimal control, states and adjoints for coarser grids.

## 6 Conclusions

We have studied the effect of the SUPG finite element method on the discretization of optimal control problems governed by the linear advection-diffusion equation. Two approaches for the computation of approximate controls and corresponding states and adjoints were compared: The discretize-then-optimize approach and the optimize-then-discretize approach. Theoretical and numerical studies of the error between the exact solution of the control problem and its approximation were provided. Our theoretical results show that the optimize-then-discretize approach leads to asymptotically better approximate solutions than the discretize-then-optimize approach. The theoretical results also indicate that the differences in solution quality is small when piecewise linear polynomials are used for the discretization of states, adjoints and controls, but that they can be significant if higher-order finite elements are used for the states and the adjoints. There is always a significant difference in quality of the adjoints computed by the discretize-then-optimize

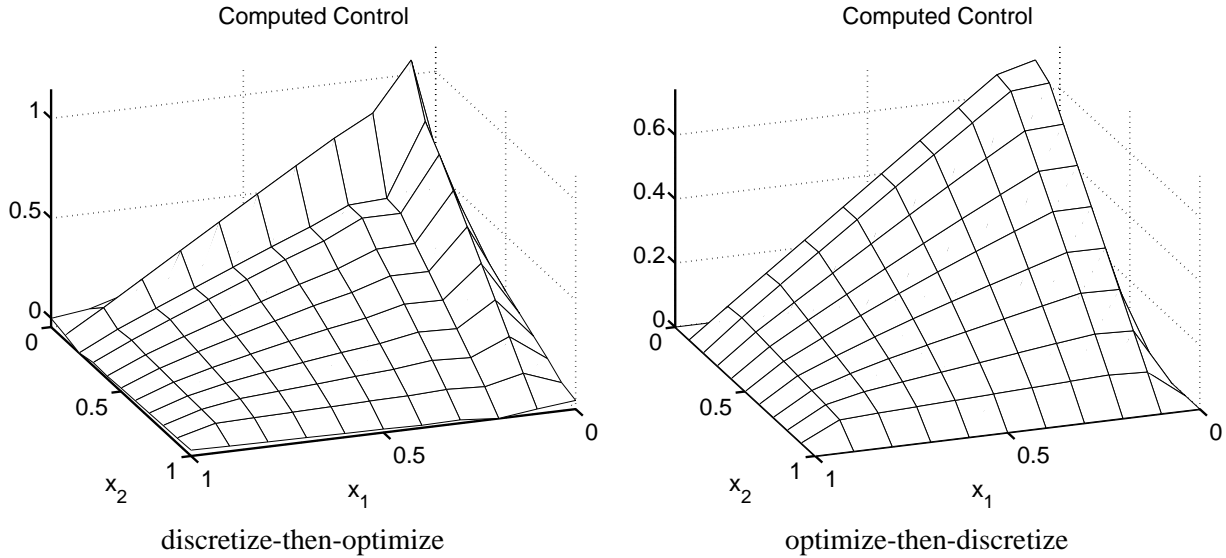


Figure 5.2: Computed controls, Example 3,  $k = \ell = m = 1$ ,  $h = 0.1$ .

approach and the optimize-then-discretize approach if finite element approximations with polynomial degree greater than one are used, and the optimize-then-discretize approach leads to better adjoint approximations. However, our numerical results have also shown that this large difference in the quality of adjoints does not necessarily imply a large difference in the quality of the controls.

Often the observed error in the controls computed by the discretize-then-optimize approach and the optimize-then-discretize approach is rather similar – even if the adjoints computed using both approaches are significantly different. This seems to be related to the fact that we consider distributed controls and that errors in the controls are measured in the  $L^2$  norm whereas errors in the adjoints are measured in the SUPG-norm. Since the distributed controls are multiples of the adjoints, our numerical results indicate that the  $L^2$ -error in the adjoints is much smaller than the error in the SUPG-norm. Whether these good convergence properties in the control also materialize if Neumann or Dirichlet boundary controls are used or if other objective functionals acting on the state are given is part of future studies. Another subject of future study is the influence of stabilization methods on the control of systems of advection diffusion equations, like the Navier-Stokes equations, where additional inconsistencies in the discretize-then-optimize approach can occur.

We conclude by reiterating that care is required when using the SUPG method for the solution of optimal control problems. If the discretize-then-optimize approach with SUPG is used, the order with which the computed solutions of the optimal control converge to the exact solution may be much lower than what one would expect from the solution of a single advection diffusion equation using SUPG. The asymptotic convergence behavior expected from the SUPG method applied to a single advection diffusion equation can be maintained for the optimal control problem if the optimize-then-discretize approach is used.

## Acknowledgements

The authors would like to thank Prof. M. Behr, Dept. of Mechanical Engineering, Rice University, for valuable discussions on stabilization methods.

Table 5.5: Errors and estimated convergence order. Example 3,  $k = \ell = m = 1$ .

The discretize-then-optimize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
1.00e-1	1.20e-1		1.25e+0		4.91e-2		1.09e-1		1.26e+0	
5.00e-2	7.43e-2	0.69	1.11e+0	0.17	2.48e-2	0.98	6.31e-2	0.79	1.09e+0	0.20
2.50e-2	4.17e-2	0.83	7.58e-1	0.55	1.36e-2	0.87	3.28e-2	0.94	7.22e-1	0.60
1.25e-2	1.46e-2	1.51	3.83e-1	0.99	5.88e-3	1.21	1.13e-2	1.54	3.69e-1	0.97
6.25e-3	3.86e-3	1.92	1.83e-1	1.07	1.68e-3	1.81	2.99e-3	1.92	1.81e-1	1.03

The optimize-then-discretize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
1.00e-1	1.22e-1		1.25e+0		1.18e-1		1.18e-1		1.25e+0	
5.00e-2	7.54e-2	0.70	1.11e+0	0.17	7.36e-2	0.69	7.36e-2	0.69	1.11e+0	0.17
2.50e-2	4.22e-2	0.84	7.59e-1	0.55	4.12e-2	0.83	4.12e-2	0.83	7.56e-1	0.55
1.25e-2	1.47e-2	1.52	3.83e-1	0.99	1.45e-2	1.51	1.45e-2	1.51	3.82e-1	0.98
6.25e-3	3.88e-3	1.92	1.83e-1	1.07	3.82e-3	1.92	3.82e-3	1.92	1.83e-1	1.06

## References

- [1] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Computational Mathematics, Vol. 15, Springer–Verlag, Berlin, 1991.
- [2] A. N. BROOKS AND T. J. R. HUGHES, *Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations*, *Comp. Meth. Appl. Mech. Engng.*, 32 (1982), pp. 199–259.
- [3] P. G. CIARLET, *Basic error estimates for elliptic problems*, in *Handbook of Numerical Analysis*, Vol.2: Finite Element Methods (Part 1), P. Ciarlet and J. Lions, eds., Elsevier/North–Holland, Amsterdam, New York, Oxford, Tokyo, 1991.
- [4] L. P. FRANCA, S. L. FREY, AND T. J. R. HUGHES, *Stabilized finite element methods. I. Application to the advective diffusive model*, *Comp. Meth. Appl. Mech. Engng.*, 95 (1992), pp. 253–276.
- [5] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, London, Melbourne, 1985.
- [6] T. J. R. HUGHES, L. P. FRANCA, AND G. M. HULBERT, *A new finite element formulation for computational fluid dynamics. VIII. the Galerkin/least squares methods for advective–diffusive systems*, *Comp. Meth. Appl. Mech. Engng.*, 73 (1989), pp. 173–189.
- [7] K. JANSEN, S. S. COLLIS, C. WHITING, AND F. SHAKIB, *A better consistency for low-order stabilized finite element methods*, *Comput. Methods Appl. Mech. Engrg.*, 174 (1999), pp. 153–170.
- [8] C. JOHNSON, U. NÄVERT, AND J. PITKÄRANTA, *Finite element methods for linear hyperbolic problems*, *Comp. Meth. Appl. Mech. Engng.*, 45 (1984), pp. 285–312.

Table 5.6: Errors and estimated convergence order. Example 3,  $k = \ell = m = 2$ .

The discretize-then-optimize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
2.00e-1	1.13e-1		6.56e-1		4.19e-2		1.09e-1		7.64e-1	
1.00e-1	6.07e-2	0.89	9.12e-1	-0.48	1.60e-2	1.39	5.96e-2	0.87	9.50e-1	-0.31
5.00e-2	2.73e-2	1.15	7.56e-1	0.27	1.29e-2	0.31	3.60e-2	0.73	1.04e+0	-0.13
2.50e-2	5.42e-3	2.33	2.92e-1	1.37	6.53e-3	0.98	1.47e-2	1.29	7.05e-1	0.56
1.25e-2	6.35e-4	3.09	8.09e-2	1.85	2.16e-3	1.60	4.29e-3	1.78	3.96e-1	0.83

The optimize-then-discretize approach										
$h$	$\ y_h - y_{\text{ex}}\ $		$\ u_h - u_{\text{ex}}\ $		$\ \lambda_h - \lambda_{\text{ex}}\ $					
	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order	$\ \cdot\ _0$	order	$\ \cdot\ _0$	order	$\ \cdot\ _{SD}$	order
2.00e-1	1.14e-1		6.54e-1		1.13e-1		1.13e-1		6.78e-1	
1.00e-1	6.13e-2	0.90	9.14e-1	-0.48	6.05e-2	0.90	6.05e-2	0.90	9.07e-1	-0.42
5.00e-2	2.76e-2	1.15	7.57e-1	0.27	2.72e-2	1.15	2.72e-2	1.15	7.54e-1	0.27
2.50e-2	5.43e-3	2.34	2.92e-1	1.38	5.42e-3	2.33	5.42e-3	2.33	2.92e-1	1.37
1.25e-2	6.35e-4	3.10	8.09e-2	1.85	6.35e-4	3.09	6.35e-4	3.09	8.09e-2	1.85

- [9] P. KNABNER AND L. ANGERMANN, *Numerik partieller Differentialgleichungen*, Springer-Verlag, Berlin, Heidelberg, New York, 2000.
- [10] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer Verlag, Berlin, Heidelberg, New York, 1971.
- [11] K. W. MORTON, *Numerical Solution of Convection-Diffusion Problems*, Chapman & Hall, London, Glasgow, New York, 1996.
- [12] A. QUARTERONI AND A. VALLI, *Numerical Approximation of Partial Differential Equations*, Springer, Berlin, Heidelberg, New York, 1994.
- [13] H. G. ROOS, M. STYNES, AND L. TOBISKA, *Numerical Methods for Singularly Perturbed Differential Equations*, Computational Mathematics, Vol. 24, Springer-Verlag, Berlin, 1996.
- [14] G. ZHOU, *How accurate is the streamline diffusion finite element method?*, Math. Comp., 66 (1997), pp. 31–44.
- [15] G. ZHOU AND R. RANNACHER, *Pointwise superconvergence of the streamline diffusion finite-element method*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 123–145.