Investigation of Commutative Properties of Discontinuous Galerkin Methods in PDE Constrained Optimal Control Problems. *

Dmitriy Leykekhman †

December 27, 2011

Abstract

The aim of this paper is to investigate commutative properties of a large family of discontinuous Galerkin (DG) methods applied to optimal control problems governed by the advection-diffusion equations. To compute numerical solutions of PDE constrained optimal control problems there are two main approaches: optimize-then-discretize and discretize-then-optimize. These two approaches do not always coincide and may lead to substantially different numerical solutions. The methods for which these two approaches do coincide we call commutative. In the theory of single equations, there is a related notion of adjoint or dual consistency. In this paper we classify DG methods both in primary and mixed forms and derive necessary conditions that can be used to develop new commutative methods. Numerical examples reveal that in the context of PDE constrained optimal control problems a special care needs to be taken to compute the solutions. For example, choosing non-commutative methods and discretize-then-optimize approach may result in a badly behaved numerical solution.

Key words Optimal control, discontinuous Galerkin methods, discretization, error estimates, optimize-then-discretize, advection-diffusion.

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

1 Introduction

We start our investigation of commutative properties of discontinuous Galerkin methods in contents of the following model problem:

$$\min \frac{1}{2}\int_\Omega (y(x)-\widehat{y}(x))^2 dx + \frac{\alpha}{2}\int_\Omega u^2(x)dx \tag{1.1}$$

*This work was supported in part by the National Science Foundation (grant DMS-0811167).

†Department of Mathematics, University of Connecticut, 196 Auditorium Road, Unit 3009 Storrs, CT 06269-3009 E-mail: leykekhman@math.uconn.edu

subject to second order advection-diffusion equation

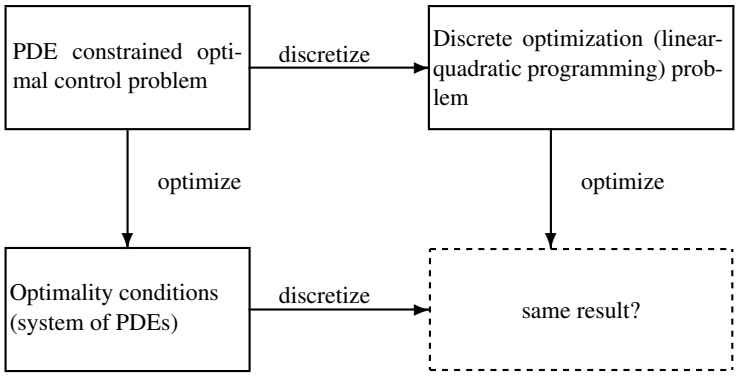
$$\nabla \cdot (-\varepsilon \nabla y(x) + \boldsymbol{\beta}(x) y(x)) + r(x) y(x) = f(x) + u(x), \qquad x \in \Omega, \tag{1.2a}$$

$$y(x) = g_D(x), \qquad x \in \Gamma_D, \tag{1.2b}$$

$$\varepsilon \frac{\partial}{\partial \mathbf{n}} y(x) = g_N(x), \qquad x \in \Gamma_N. \tag{1.2c}$$

Here $\Gamma_D \neq \emptyset$ is the Dirichlet and Γ_N is the Neumann part of the boundary, such that $\Gamma = \overline{\Gamma_D \cup \Gamma_N}$ and $\Gamma_D \cap \Gamma_N = \emptyset$; $\boldsymbol{\beta}, f, g_D, g_N, r, \widehat{y}$ are given functions, $\varepsilon > 0$, $\alpha > 0$ are given scalars, and $\mathbf{n}$ denotes the outward unit normal. Assumptions on these data that ensure that the problem is well-posed will be given in the next section.

For the numerical solution of the optimal control problems basically there are two approaches. In the *optimize-then-discretize* approach, one first derives the optimality conditions for (1.1)-(1.2) and then discretizes the resulting system. In the *discretize-then-optimize* approach, one first discretizes (1.1) and (1.2) and then solves the finite dimensional optimization problem. These two approaches do not always coincide even for our simple model problem and may lead to substantially different numerical solutions. This point was first illustrated in [15] for the streamline-diffusion method (SUPG). This result inspired interest for the search of commutative stabilization methods. First such method was analyzed in [5]. Later the ideas of this paper were generalized to optimal control problems constrained by Oseen equation [7].



Discontinuous Galerkin (DG) methods are attractive alternatives to other stabilized methods to solve advection-diffusion-reaction equations [2, 8, 12, 13, 18, 19, 23, 24, 31]. They provide greater flexibility to locally adapt the mesh or the polynomial degree of the basis functions which implies better ability to capture fine scales of the solution. The way DG methods treat Dirichlet boundary conditions has also positive effect on local error estimates (cf. [25]). Presently, there exist many DG methods for advection-diffusion problems and their number is constantly growing. Instead of checking the commutativity property for each individual method, following ideas of [9], we analyze a large family of existing DG methods and classify them. In addition we also provide conditions the new DG methods need to satisfy in order to be commutative.

Numerical results show that controls computed by non-commutative DG methods and *discretize-then-optimize* approach may have very low order of convergence in L^2 norm and not converge at all in H^1 norm. The quality of the solutions suffers as well. This is alarming since *discretize-then-optimize* approach is very popular in practice. It seems very natural to discretize the continuous problem and use available tools

to solve the resulting linear-quadratic programming problem. Therefore we believe that the commutative property is desirable and at least for in the case of a simple model problem, a "good" method should be independent of the approach one takes.

Also, we would like to mention that there is a high interest in dual or adjoint consistent DG method in the case of a single equation. Such methods have better convergence rates in the L^2 norm and allow double order of convergence in adaptivity for certain functional of interest [21, 28, 29]. In principle, for our simple model problem we could use the definition of adjoint consistency to investigate the commutative properties. However, for the PDE constrained optimal control problems in addition to the state and adjoint equations, one needs to look at the gradient equation as well. The role of this gradient equation is more subtle for nonlinear problems. Therefore we decided to use the analysis based on the minimization of the Lagrangian functional since it can be naturally extended to more complicated problems.

The rest of the paper is organized as follows. In Section 2, we introduce the notation and state the optimality conditions. In Section 3, first we describe the discontinuous Galerkin methods and then derive commutativity conditions when the state equation is just a Laplace equation, both in primal and mixed forms. We conclude this section with the commutativity conditions for a full advection-diffusion state problem. In Section 4 we provide error analysis for the optimal control problem for some DG methods in the energy and L^2 norms. Finally, in Section 5 we provide some numerical results that illustrate the dangers of using non-commutative methods and the *discretize-then-optimize* approach.

2 Optimal Control Problem and Optimality Conditions

In this section we collect some results on the existence, uniqueness and characterization of solutions of the optimal control problem (1.1), (1.2). We define the state and control spaces as

$$Y = \left\{ y \in H^1(\Omega) \,:\, y = g_D \text{ on } \Gamma_D \right\}, \quad U = L^2(\Omega) \tag{2.1a}$$

and the space of test functions as

$$V = \left\{ v \in H^1(\Omega) \,:\, v = 0 \text{ on } \Gamma_D \right\}. \tag{2.1b}$$

We also use the following inner product and (semi-)norms. Let $D \subset \overline{\Omega}$. For any $k > 0$ and multi-index α we define

$$(f,g)_D = \int_D fg, \qquad \|f\|_D^2 = \int_D f^2,$$
$$|f|_{k,D}^2 = \sum_{|\alpha|=k} \int_D |D^\alpha f|^2, \qquad \|f\|_{k,D}^2 = \sum_{|\alpha|\le k} \int_D |D^\alpha f|^2.$$

If $D = \Omega$, we will drop the subscripts.

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

$$a(y,v) + b(u,v) = (f,v) + \langle g_N, v \rangle_{\Gamma_N}, \quad \forall v \in V, \tag{2.2a}$$

where

$$a(y,v) = \int_\Omega \varepsilon \nabla y(x)\cdot\nabla v(x) + \nabla\cdot(\boldsymbol{\beta}(x)y(x))v(x) + r(x)y(x)v(x)dx, \tag{2.2b}$$

$$b(u,v) = -\int_\Omega u(x)v(x)dx, \tag{2.2c}$$

$$\langle g_N, v\rangle_{\Gamma_N} = \int_{\Gamma_N} g_N(x)v(x)dx. \tag{2.2d}$$

We are interested in the solution of the optimal control problem

$$\begin{aligned} &\underset{\{y,u\}\in Y\times U}{\text{minimize}} && \frac{1}{2}\|y-\widehat{y}\|^2 + \frac{\alpha}{2}\|u\|^2, && \text{(2.3a)}\\ &\text{subject to} && a(y,v)+b(u,v) = (f,v)+\langle g_N, v\rangle_{\Gamma_N}, \quad \forall v\in V. && \text{(2.3b)}\end{aligned}$$

We assume that

$$f,\widehat{y}\in L^2(\Omega), \boldsymbol{\beta}\in W^{1,\infty}(\Omega)^2, r\in L^\infty(\Omega), g_D\in H^{3/2}(\Gamma_D), g_N\in H^{1/2}(\Gamma_N), \alpha>0, \varepsilon>0, \tag{2.4}$$

$\Gamma_D\neq\emptyset$ is the Dirichlet and Γ_N is the Neumann part of the boundary, such that $\Gamma=\overline{\Gamma_D\cup\Gamma_N}$ and $\Gamma_D\cap\Gamma_N=\emptyset$.

Under the above assumptions, the bilinear form $a(\cdot,\cdot)$ is continuous on $V\times V$ and V-elliptic. Hence the theory in [26, Sec. II.1] guarantees the existence of a unique solution $(y,u)\in Y\times U$ of (2.3). The theory in [26, Sec. II.1] also provides necessary and sufficient optimality conditions, which can be best described using the Lagrangian functional

$$L(y,u,\lambda) = \frac{1}{2}\|y-\widehat{y}\|^2 + \frac{\alpha}{2}\|u\|^2 + a(y,\lambda)+b(u,\lambda)-(f,\lambda)-\langle g_N,\lambda\rangle_{\Gamma_N}. \tag{2.5}$$

The necessary and, for our model problem, sufficient optimality conditions can be obtained by setting the partial Fréchet-derivatives of (2.5) with respect to the state y, control u, and adjoint λ equal to zero. Accomplishing it we obtain, the adjoint equation

$$\frac{\partial L}{\partial y}\psi = a(\psi,\lambda)+(y-\hat{y},\psi)=0, \quad \forall\psi\in V, \tag{2.6a}$$

the gradient equation

$$\frac{\partial L}{\partial u}w = b(w,\lambda)+\alpha(u,w)=0, \quad \forall w\in U, \tag{2.6b}$$

and the state equation

$$\frac{\partial L}{\partial \lambda}v = a(y,v)+b(u,v)-(f,v)+\langle g,v\rangle_{\Gamma_N}=0, \quad \forall v\in V. \tag{2.6c}$$

Notice that the adjoint equation (2.6a) is also an advection-diffusion equation, with the strong form

$$\nabla \cdot (-\varepsilon \nabla \lambda(x) - \boldsymbol{\beta}(x)\lambda(x)) + (r(x) + \nabla \cdot \boldsymbol{\beta}(x))\lambda(x) = -(y(x) - \hat{y}(x)), \qquad x \in \Omega, \tag{2.7a}$$

$$\lambda(x) = 0, \qquad x \in \Gamma_D, \tag{2.7b}$$

$$\varepsilon \frac{\partial}{\partial \mathbf{n}} \lambda(x) + \boldsymbol{\beta}(x) \cdot \mathbf{n}(x)\ \lambda(x) = 0, \qquad x \in \Gamma_N. \tag{2.7c}$$

In contrast to the state equation, in the adjoint equation the advection field is $-\boldsymbol{\beta}$, the reaction term is $r + \nabla \cdot \boldsymbol{\beta}$, zero Dirichlet boundary condition on Γ_D, and zero mixed type boundary condition on Γ_N.

3 Discontinuous Galerkin Discretization

To set the DG framework we will need some notation. Let $T = \{T_h\}_h$ be a family of conforming triangulations such that $\overline{\Omega} = \cup_{\tau \in T_h} \overline{\tau}$, $\tau_i \cap \tau_j = \emptyset$ for $\tau_i, \tau_j \in T_h$, $i \neq j$. We set $\max_{\tau \in T_h} \operatorname{diam}(\tau) = \mathrm{h}$. The assumption that the triangulations are conforming can be relaxed in the formulation of the discontinuous Galerkin discretization.

Define $\mathcal{E}_h^0$ to be a set of interior edges of T_h and $\mathcal{E}_h^\partial$ to be a collection of the boundary edges. Hence the set of all edges is given by $\mathcal{E}_h = \mathcal{E}_h^\partial \cup \mathcal{E}_h^0$. We further decompose the boundary edges into $\mathcal{E}_h^\partial = \mathcal{E}_h^+ \cup \mathcal{E}_h^-$, where $\mathcal{E}_h^- \stackrel{\text{def}}{=} \left\{ e \in \mathcal{E}_h^\partial \,:\, e \subset \{x \in \partial\Omega \,:\, \boldsymbol{\beta}(x) \cdot \mathbf{n}(x) < 0\} \right\}$ and $\mathcal{E}_h^+ \stackrel{\text{def}}{=} \mathcal{E}_h^\partial \setminus \mathcal{E}_h^-$ are the sets of edges that corresponding to inflow and outflow parts of the boundary, respectively. For a given element $\tau \in T_h$, we decompose its boundary $\partial\tau$ into to inflow and outflow parts of the element boundary, $\partial_-\tau \stackrel{\text{def}}{=} \{x \in \partial\tau \,:\, \boldsymbol{\beta}(x) \cdot \mathbf{n}_\tau(x) < 0\}$ and $\partial_+\tau \stackrel{\text{def}}{=} \{x \in \partial\tau \,:\, \boldsymbol{\beta}(x) \cdot \mathbf{n}_\tau(x) \geq 0\}$, where $\mathbf{n}_\tau$ denotes the unit outward normal to τ.

Let τ_1 and τ_2 be two neighboring elements and let $\mathbf{n}^1$ and $\mathbf{n}^2$ be outward normal vectors at the boundary of elements τ_1 and τ_2 respectively. Let ϕ^i and $\boldsymbol{\varphi}^i$ be the restrictions to τ_i, $i = 1, 2$ respectively. We define the standard jump averages on the set of interior edges by

$$\{\phi\} = \frac{\phi^1 + \phi^2}{2}, \quad [\![\phi]\!] = \phi^1 \mathbf{n}^1 + \phi^2 \mathbf{n}^1, \tag{3.1}$$

$$\{\boldsymbol{\varphi}\} = \frac{\boldsymbol{\varphi}^1 + \boldsymbol{\varphi}^2}{2}, \quad [\![\boldsymbol{\varphi}]\!] = \boldsymbol{\varphi}^1 \cdot \mathbf{n}^1 + \boldsymbol{\varphi}^2 \cdot \mathbf{n}^1. \tag{3.2}$$

On the set of boundary edges we set

$$\{\phi\} = \phi, \quad [\![\phi]\!] = \phi \mathbf{n}, \quad \{\boldsymbol{\varphi}\} = \boldsymbol{\varphi}. \tag{3.3}$$

In the following analysis we will frequently use the identity,

$$\begin{aligned} \sum_{\tau \in T_h} (\boldsymbol{\varphi} \cdot \mathbf{n}, \phi)_{\partial\tau} &= \sum_{e \in \mathcal{E}_h} (\{\boldsymbol{\varphi}\}, [\![\phi]\!])_e + \sum_{e \in \mathcal{E}_h^0} ([\![\boldsymbol{\varphi}]\!], \{\phi\})_e \\ &= \sum_{e \in \mathcal{E}_h^0} (\{\boldsymbol{\varphi}\}, [\![\phi]\!])_e + ([\![\boldsymbol{\varphi}]\!], \{\phi\})_e + \sum_{e \in \mathcal{E}_h^\partial} (\boldsymbol{\varphi} \cdot \mathbf{n}, \phi)_e. \end{aligned} \tag{3.4}$$

We define the discrete state and control spaces to be

$$Y_h = V_h \stackrel{\text{def}}{=} \left\{ y \in L^2(\Omega) \,:\, y\,|_\tau \in \mathbb{P}^k(\tau) \quad \forall \tau \in T_h \right\}, \tag{3.5a}$$

$$U_h \stackrel{\text{def}}{=} \left\{ u \in L^2(\Omega) \,:\, u\,|_\tau \in \mathbb{P}^l(\tau) \quad \forall \tau \in T_h \right\}, \tag{3.5b}$$

respectively. $\mathbb{P}^k$ denotes the set of polynomials of order k. The orders $k, l \in \mathbb{N}$ of the finite element approximation can be different for the states and the controls. Note that since discontinuous Galerkin methods naturally impose boundary conditions weakly, the space Y_h of discrete states and the space of test functions V_h are identical. For the rest of the paper to avoid the unnecessary confusion the only finite dimensional space we will use is V_h.

3.1 Laplace equation. Primal formulation.

3.1.1 DG discretization of the state equation

For a clearer illustration of the ideas, first we consider in details the case when the state equation is just the Laplace equation, i.e. our optimal control problem is the following,

$$\min \frac{1}{2}\int_\Omega (y(x) - \widehat{y}(x))^2 dx + \frac{\alpha}{2}\int_\Omega u^2(x) dx \tag{3.6}$$

subject to

$$-\Delta y(x) = \nabla \cdot (-\nabla y(x)) = f(x) + u(x), \qquad x \in \Omega, \tag{3.7a}$$

$$y(x) = g_D(x), \qquad x \in \Gamma_D, \tag{3.7b}$$

$$\frac{\partial}{\partial \mathbf{n}} y(x) = g_N(x), \qquad x \in \Gamma_N. \tag{3.7c}$$

The optimality conditions for this problem can be obtained from (2.6) by setting $\varepsilon = 1$, $\boldsymbol{\beta} \equiv 0$, and $r \equiv 0$.

Following [8], we rewrite the state equation as

$$\begin{aligned}
-\Delta y(x) &= f(x) + u(x), && \text{in each } \tau \in T_h,\\
[\![y(x)]\!] &= 0, && \text{on each } e \in \mathcal{E}_h^0,\\
[\![-\nabla y(x)]\!] &= 0, && \text{on each } e \in \mathcal{E}_h^0,\\
y(x) &= g_D(x), && \text{on each } e \in \Gamma_D,\\
\frac{\partial}{\partial \mathbf{n}} y(x) &= g_N(x), && \text{on each } e \in \Gamma_N.
\end{aligned}$$

Assume we are given operators B_0, $\boldsymbol{B_1}$, B_2, B_D, and B_N. We use the usual convention. The bold letters denote vector valued operators and the regular letters denote the scalar valued operators.

We consider the following discrete problem: for a given $f, u \in L^2(\Omega)$ find $y \in V_h$ such that for any $v \in V_h$,

$$a_h(y,u;v) \stackrel{\text{def}}{=} \sum_{\tau\in T_h} (-\Delta y - f - u, B_0 v)_\tau + \sum_{e\in\mathcal{E}_h^0} (\llbracket y \rrbracket, \boldsymbol{B_1} v)_e - (\llbracket \nabla y \rrbracket, B_2 v)_e \tag{3.8}$$
$$+ \sum_{e\in\Gamma_D} (y - g_D, B_D v)_e + \sum_{e\in\Gamma_N} (\frac{\partial y}{\partial \mathbf{n}} - g_N, B_N v)_e = 0.$$

Example 3.1 *By taking*

$$B_0 v := v, \qquad \forall \tau \in T_h, \tag{3.9a}$$
$$\boldsymbol{B_1} v := -\{\nabla v\} + \frac{\sigma}{|e|}\llbracket v \rrbracket, \qquad \forall e \in \mathcal{E}_h^0, \tag{3.9b}$$
$$B_2 v := -\{v\}, \qquad \forall e \in \mathcal{E}_h^0, \tag{3.9c}$$
$$B_D v := -\frac{\partial v}{\partial \mathbf{n}} + \frac{\sigma}{|e|} v, \qquad \forall e \in \Gamma_D, \tag{3.9d}$$
$$B_N v := v, \qquad \forall e \in \Gamma_N, \tag{3.9e}$$

we obtain the usual symmetric interior penalty method (SIPG). To insure the stability, σ needs to be sufficiently large.

Example 3.2 *By keeping the same operators B_0, B_2, and B_N as in* (3.9a)*,* (3.9c)*, and* (3.9e)*, and taking*

$$\boldsymbol{B_1} v := \{\nabla v\} + \frac{\sigma}{|e|}\llbracket v \rrbracket, \qquad \forall e \in \mathcal{E}_h^0, \tag{3.10a}$$
$$B_D v := \frac{\partial v}{\partial \mathbf{n}} + \frac{\sigma}{|e|} v, \qquad \forall e \in \Gamma_D, \tag{3.10b}$$

we obtain the usual non-symmetric interior penalty method (NIPG). In this example σ can be any positive number.

Example 3.3 *Although we are primary interested in DG methods, the Continuous Interior Penalty (CIP) method [17] can be put in this framework as well by considering the continuous elements and taking*

$$B_0 v := v, \qquad \forall \tau \in T_h, \tag{3.11a}$$
$$\boldsymbol{B_1} v := \textit{arbitrary } (\textit{since } \llbracket u \rrbracket = 0), \qquad \forall e \in \mathcal{E}_h^0, \tag{3.11b}$$
$$B_2 v := -\{v\} - c_2 |e|^2 \llbracket \nabla v \rrbracket, \qquad \forall e \in \mathcal{E}_h^0, \tag{3.11c}$$
$$B_D v := -\frac{\partial v}{\partial \mathbf{n}} + \frac{\sigma}{|e|} v, \qquad \forall e \in \Gamma_D, \tag{3.11d}$$
$$B_N v := v, \qquad \forall e \in \Gamma_N. \tag{3.11e}$$

Remark 3.4 *There is some sign inconsistency with [8]. There $\llbracket \nabla y \rrbracket = 0$ was imposed instead of $\llbracket -\nabla y \rrbracket = 0$. In the formulation of the method, this choice only affects the sign in the definition of the operator B_2. Later, when we will consider the mixed form of the equation the choice of the sign is more important.*

3.1.2 Optimize-then-Discretize

Applying the above DG disretization to the optimality system (2.6), we obtain that a triplet $(y, u, \lambda) \in V_h \times V_h \times V_h$ is the unique solution of the system consisting of the discretized adjoint equation

$$\sum_{\tau\in T_h}(-\Delta\lambda - \hat{y} + y, B_0\phi)_\tau + \sum_{e\in\mathcal{E}_h^0}(\llbracket\lambda\rrbracket, \boldsymbol{B_1}\phi)_e - (\llbracket\nabla\lambda\rrbracket, B_2\phi)_e + \sum_{e\in\Gamma_D}(\lambda, B_D\phi)_e + \sum_{e\in\Gamma_N}(\frac{\partial\lambda}{\partial\mathbf{n}}, B_N\phi)_e = 0, \quad \forall\phi\in V_h, \tag{3.12a}$$

the discretized gradient equation

$$(\psi, \lambda)_\tau - \alpha(u, \psi)_\tau = 0, \quad \forall\psi\in V_h, \tag{3.12b}$$

and the discretized state equation

$$\sum_{\tau\in T_h}(-\Delta y - f - u, B_0\varphi)_\tau + \sum_{e\in\mathcal{E}_h^0}(\llbracket y\rrbracket, \boldsymbol{B_1}\varphi)_e - (\llbracket\nabla y\rrbracket, B_2\varphi)_e + \sum_{e\in\Gamma_D}(y - g_D, B_D\varphi)_e + \sum_{e\in\Gamma_N}(\frac{\partial y}{\partial\mathbf{n}} - g_N, B_N\varphi)_e = 0, \quad \forall\varphi\in V_h. \tag{3.12c}$$

3.1.3 Discretize-then-Optimize

Now we derive the optimality conditions for *discretize-then-optimize* approach when the optimal control problem is discretized by the method above. Thus we are solving

$$\begin{aligned} &\underset{\{y,u\}\in V_h\times V_h}{\text{minimize}} && \frac{1}{2}\|y-\widehat{y}\|^2 + \frac{\alpha}{2}\|u\|^2, && (3.13a)\\ &\text{subject to} && a_h(y,u;v) = 0, \quad \forall v\in V_h. && (3.13b)\end{aligned}$$

The discrete Lagrangian is now given by

$$L_h(y,u,\lambda) = \frac{1}{2}\|y-\widehat{y}\|^2 + \frac{\alpha}{2}\|u\|^2 + a_h(y,u;\lambda). \tag{3.14}$$

Again, the necessary and, for our model problem, sufficient optimality conditions can be obtained by setting the partial Fréchet-derivatives of (3.14) with respect to the discrete state y, discrete control u, and discrete adjoint λ equal to zero. Thus we obtain the system consisting of the discrete adjoint equation

$$\frac{\partial L_h}{\partial y}\phi = \sum_{\tau\in T_h}(-\Delta\phi, B_0\lambda)_\tau + (y-\hat{y},\phi)_\tau + \sum_{e\in\mathcal{E}_h^0}(\llbracket\phi\rrbracket, \boldsymbol{B_1}\lambda)_e - (\llbracket\nabla\phi\rrbracket, B_2\lambda)_e + \sum_{e\in\Gamma_D}(\phi, B_D\lambda)_e + \sum_{e\in\Gamma_N}(\frac{\partial\phi}{\partial\mathbf{n}}, B_N\lambda)_e = 0, \quad \forall\phi\in V_h, \tag{3.15a}$$

the discrete gradient equation

$$\frac{\partial L_h}{\partial u}\psi = (\psi, B_0\lambda)_\tau - \alpha(u,\psi)_\tau = 0, \quad \forall \psi \in V_h, \tag{3.15b}$$

and the discrete state equation

$$\begin{aligned}\frac{\partial L_h}{\partial \lambda}\varphi = \sum_{\tau\in T_h}(-\Delta y - f - u, B_0\varphi)_\tau + \sum_{e\in\mathcal{E}_h^0}([\![y]\!], \boldsymbol{B_1}\varphi)_e - ([\![\nabla y]\!], B_2\varphi)_e \\ + \sum_{e\in\Gamma_D}(y - g_D, B_D\varphi)_e + \sum_{e\in\Gamma_N}(\frac{\partial y}{\partial \mathbf{n}} - g_N, B_N\varphi)_e = 0, \quad \forall\varphi\in V_h.\end{aligned} \tag{3.15c}$$

3.1.4 Commutativity Conditions.

From the system of the discrete equations (3.12) and (3.15) we can easily identify the methods for which both approaches *optimize-then-discretize* and *discretize-then-optimize* commute. First of all comparing (3.12b) with (3.15b) we conclude that for the method to be commutative one needs

$$B_0 v := v. \tag{3.16}$$

This condition seems to hold for the most known DG methods. From now on and until the end of this section we assume (3.16).

Now we turn our attention to (3.12a) and (3.15a). Integrating $(-\Delta\phi, \lambda)_\tau$ by parts twice we obtain,

$$(-\Delta\phi,\lambda)_\tau = (\phi, -\Delta\lambda)_\tau - (\frac{\partial\phi}{\partial\mathbf{n}}, \lambda)_{\partial\tau} + (\phi, \frac{\partial\lambda}{\partial\mathbf{n}})_{\partial\tau}.$$

Summing over all elements τ and using (3.4), we can rewrite (3.15a) as

$$\begin{aligned}&\sum_{\tau\in T_h}(-\Delta\lambda - \hat{y} + y, \phi)_\tau \\ &\quad + \sum_{e\in\mathcal{E}_h^0}([\![\phi]\!], \boldsymbol{B_1}\lambda + \{\nabla\lambda\})_e - (\{\nabla\phi\}, [\![\lambda]\!])_e - ([\![\nabla\phi]\!], B_2\lambda + \{\lambda\})_e + (\{\phi\}, [\![\nabla\lambda]\!])_e \\ &\quad + \sum_{e\in\Gamma_D}(\phi, B_D\lambda + \frac{\partial\lambda}{\partial\mathbf{n}})_e - (\frac{\partial\phi}{\partial\mathbf{n}}, \lambda)_e + \sum_{e\in\Gamma_N}(\frac{\partial\phi}{\partial\mathbf{n}}, B_N\lambda - \lambda)_e + (\phi, \frac{\partial\lambda}{\partial\mathbf{n}})_e = 0.\end{aligned} \tag{3.17}$$

Now directly comparing (3.17) with (3.12a), in order to have a commutative method it is necessary

$$([\![\lambda]\!], \boldsymbol{B_1}\phi + \{\nabla\phi\})_e - ([\![\nabla\lambda]\!], B_2\phi + \{\phi\})_e = ([\![\phi]\!], \boldsymbol{B_1}\lambda + \{\nabla\lambda\})_e - ([\![\nabla\phi]\!], B_2\lambda + \{\lambda\})_e \tag{3.18}$$

on each interior edge. This condition can be satisfied for example by choosing

$$\begin{aligned}\boldsymbol{B_1}v &:= -\{\nabla v\} + c_1[\![v]\!], \\ B_2 v &:= -\{v\} + c_2[\![\nabla v]\!],\end{aligned}$$

with some parameters c_1 and c_2 that may depend on the edge e. For example the choice $c_1 = \frac{\sigma}{|e|}$ and $c_2 = 0$ leads to the SIPG method. However, in addition, the boundary operators B_D and B_N must also satisfy

$$(\phi, B_D\lambda + \frac{\partial \lambda}{\partial \mathbf{n}})_e = (\lambda, B_D\phi + \frac{\partial \phi}{\partial \mathbf{n}})_e \tag{3.19}$$

on each Dirichlet edge and

$$(\frac{\partial \phi}{\partial \mathbf{n}}, B_N\lambda - \lambda)_e = (\frac{\partial \lambda}{\partial \mathbf{n}}, B_N\phi - \phi)_e \tag{3.20}$$

on each Neumann edge. This leads to the following choice

$$B_D v := -\frac{\partial v}{\partial \mathbf{n}} + c_D v,$$
$$B_N v := v + c_N \frac{\partial v}{\partial \mathbf{n}},$$

with some parameters c_D and c_N that may depend on the edge e. Thus in particular for the SIPG and CIP methods both approaches coincide. We summarize our findings in the following proposition.

Proposition 3.5 *Assume that the optimal control problem* (3.6) *is discretized by a DG method that can be put in the form of* (3.8) *with given operators* $B_0, \boldsymbol{B_1}, B_2, B_D, B_N$ *. Then in order for the two approaches optimize-then-discretize and discretize-then-optimize to commute the operators must satisfy* (3.16) *in the interior of each triangle,* (3.18) *on each interior edge, and* (3.19) *and* (3.20) *on the boundary edges.*

In the Table 3.1 we list most popular DG methods and report whether they are commutative or not.

Table 3.1: Results for some common DG methods in primary form

Method	$B_0 v$	$\boldsymbol{B_1} v$	$B_2 v$	$B_D v$	$B_N v$	commutative
CIP [17]	v	$[\![v]\!] \equiv 0$	$-v + c_2[\![\nabla v]\!]$	$-\frac{\partial v}{\partial \mathbf{n}} + c_1 v$	v	yes
SIPG [1]	v	$-\{\nabla v\} + c_1[\![v]\!]$	$-\{v\}$	$-\frac{\partial v}{\partial \mathbf{n}} + c_1 v$	v	yes
NIPG [30]	v	$\{\nabla v\} + c_1[\![v]\!]$	$-\{v\}$	$\frac{\partial v}{\partial \mathbf{n}} + c_1 v$	v	no
B.O [4]	v	$\{\nabla v\}$	$-\{v\}$	$\frac{\partial v}{\partial \mathbf{n}}$	v	no
D.S.W. [16]	v	$c_1[\![v]\!]$	$-\{v\}$	$c_1 v$	v	no

3.2 Laplace equation. Mixed formulation

A larger class of DG methods that can be obtained from the mixed formulation of the problem. Following similar reasoning we can derive the commutativity conditions and as a result identify commutative methods in mixed form. Since the arguments are very similar to the previous section we will omit some details.

Following [8, sec. 2.2], we rewrite the state equation in the mixed form as

$$\begin{aligned}
\boldsymbol{\sigma}(x) &= -\nabla y(x), && \text{in each } \tau \in T_h,\\
\nabla \cdot \boldsymbol{\sigma}(x) &= f(x) + u(x), && \text{in each } \tau \in T_h,\\
[\![y(x)]\!] &= 0, && \text{on each } e \in \mathcal{E}_h^0,\\
[\![\boldsymbol{\sigma}(x)]\!] &= 0, && \text{on each } e \in \mathcal{E}_h^0,\\
y(x) &= g_D(x), && \text{on each } e \in \Gamma_D,\\
\boldsymbol{\sigma}(x) \cdot \mathbf{n}(x) &= g_N(x), && \text{on each } e \in \Gamma_N.
\end{aligned}$$

Assume we are given operators $\boldsymbol{B_{00}}$, $\boldsymbol{B_{01}}$, B_{02}, B_{10}, $\boldsymbol{B_{11}}$, B_{12}, B_{D_0}, B_{N_0}, B_{D_1}, and B_{N_1}. Then the mixed analog of (3.8) is a system

$$\begin{aligned}
&\sum_{\tau\in T_h} (\boldsymbol{\sigma} + \nabla y, \boldsymbol{B_{00}}\boldsymbol{\phi})_\tau + \sum_{e\in\mathcal{E}_h^0} ([\![y]\!], \boldsymbol{B_{01}}\boldsymbol{\phi})_e + ([\![\boldsymbol{\sigma}]\!], B_{02}\boldsymbol{\phi})_e \qquad (3.21a)\\
&\qquad + \sum_{e\in\Gamma_D} (y - g_D, B_{D_0}\boldsymbol{\phi})_e + \sum_{e\in\Gamma_N} (\boldsymbol{\sigma}\cdot\mathbf{n} - g_N, B_{N_0}\boldsymbol{\phi})_e = 0, \quad \forall \boldsymbol{\phi} \in H^{div}(T_h)\\
&\sum_{\tau\in T_h} (\nabla\cdot\boldsymbol{\sigma} - f - u, B_{10}\psi)_\tau + \sum_{e\in\mathcal{E}_h^0} ([\![y]\!], \boldsymbol{B_{11}}\psi)_e + ([\![\boldsymbol{\sigma}]\!], B_{12}\psi)_e \qquad (3.21b)\\
&\qquad + \sum_{e\in\Gamma_D} (y - g_D, B_{D_1}\psi)_e + \sum_{e\in\Gamma_N} (\boldsymbol{\sigma}\cdot\mathbf{n} - g_N, B_{N_1}\psi)_e = 0, \quad \forall \psi \in H^1(T_h),
\end{aligned}$$

where

$$H^{div}(T_h) = \{\boldsymbol{\phi} \in L^2(\Omega)^2 : \nabla\cdot\boldsymbol{\phi}_{|\tau} \in L^1(\tau),\ \forall \tau \in T_h\}$$

and

$$H^1(T_h) = \{\psi \in L^2(\Omega) : \psi \in H^1(\tau)\ \forall \tau \in T_h\}.$$

To complete the *optimize-then-discretize* system, we also need the gradient equation

$$(\alpha u, \psi)_\tau = (\lambda, \psi)_\tau, \quad \forall \tau \in T_h,\ \forall \psi, \tag{3.22}$$

and the adjoint equation in the mixed form

$$\begin{aligned}
\boldsymbol{p}(x) &= -\nabla\lambda(x) && x \in \Omega, && (3.23)\\
\nabla\cdot\boldsymbol{p}(x) &= \hat{y}(x) - y(x), && x \in \Omega, && (3.24)\\
\lambda(x) &= 0, && x \in \Gamma_D, && (3.25)\\
\boldsymbol{p}(x)\cdot\mathbf{n}(x) &= 0, && x \in \Gamma_N. && (3.26)
\end{aligned}$$

Discretizing the adjoint equation similarly to the state equation we obtain

$$\sum_{\tau\in T_h}(\boldsymbol{p}+\nabla\lambda, \boldsymbol{B_{00}}\boldsymbol{\phi})_\tau + \sum_{e\in\mathcal{E}_h^0}([\![\lambda]\!], \boldsymbol{B_{01}}\boldsymbol{\phi})_e + ([\![\boldsymbol{p}]\!], B_{02}\boldsymbol{\phi})_e + \sum_{e\in\Gamma_D}(\lambda, B_{D_0}\boldsymbol{\phi})_e + \sum_{e\in\Gamma_N}(\boldsymbol{p}\cdot\mathbf{n}, B_{N_0}\boldsymbol{\phi})_e = 0, \quad \forall \boldsymbol{\phi}\in V_h^2 \tag{3.27a}$$

$$\sum_{\tau\in T_h}(\nabla\cdot\boldsymbol{p}+y-\hat{y}, B_{10}\psi)_\tau + \sum_{e\in\mathcal{E}_h^0}([\![\lambda]\!], \boldsymbol{B_{11}}\psi)_e + ([\![\boldsymbol{p}]\!], B_{12}\psi)_e + \sum_{e\in\Gamma_D}(\lambda, B_{D_1}\psi)_e + \sum_{e\in\Gamma_N}(\boldsymbol{p}\cdot\mathbf{n}, B_{N_1}\psi)_e = 0, \quad \forall \psi\in V_h. \tag{3.27b}$$

The *discretize-then-optimize* system consists of the discrete state equation (3.21), the discrete gradient equation

$$(\alpha u, \psi)_\tau = (B_{10}\lambda, \psi)_\tau, \quad \forall \tau\in T_h, \tag{3.28}$$

and the discrete adjoint system

$$\sum_{\tau\in T_h}(\boldsymbol{\phi}, \boldsymbol{B_{00}}\boldsymbol{p})_\tau + (-\nabla\cdot\boldsymbol{\phi}, B_{10}\lambda)_\tau + \sum_{e\in\mathcal{E}_h^0}([\![\boldsymbol{\phi}]\!], B_{02}\boldsymbol{p})_e + ([\![\boldsymbol{\phi}]\!], B_{12}\lambda)_e + \sum_{e\in\Gamma_N}(\boldsymbol{p}\cdot\mathbf{n}, B_{N_0}\boldsymbol{p})_e + (\boldsymbol{p}\cdot\mathbf{n}, B_{N_1}\lambda)_e = 0, \tag{3.29a}$$

$$\sum_{\tau\in T_h}(-\nabla\psi, \boldsymbol{B_{00}}\boldsymbol{p})_\tau + (y-\hat{y}, \psi)_\tau + \sum_{e\in\mathcal{E}_h^0}([\![\psi]\!], \boldsymbol{B_{01}}\boldsymbol{p})_e + ([\![\psi]\!], \boldsymbol{B_{11}}\lambda)_e + \sum_{e\in\Gamma_D}(\psi, B_{D_0}\boldsymbol{p})_e + (\psi, B_{D_1}\lambda)_e = 0. \tag{3.29b}$$

Directly comparing (3.22) with (3.28) we immediately obtain that for the method to be commutative we must have

$$B_{10}v := v. \tag{3.30}$$

Next we compare (3.27) with (3.29). Looking at the terms over the elements τ and taking in consideration (3.30) we derive

$$\boldsymbol{B_{00}}v := -v. \tag{3.31}$$

As we can see from the Table 3.2, these choices are made in almost all DG methods.

Remark 3.6 *Again there is a sign inconsistency in the definition of operators $\boldsymbol{B_{00}}$, $\boldsymbol{B_{01}}$, and B_{02} in the present paper and [8]. The choice of the sign did not really matter in [8], since the equation* (2.38) *in that paper could be multiplied by negative one. However the choice of the sign in our paper is essential and can not be taken arbitrarily.*

Next we look at the terms over the interior edges. After some manipulation we derive that for the methods to be commutative, on each interior edge e we also must have

$$([\![\boldsymbol{\phi}]\!], B_{02}\boldsymbol{p}+B_{12}\lambda+\{\lambda\})_e = ([\![\lambda]\!], \boldsymbol{B_{01}}\boldsymbol{\phi}-\{\boldsymbol{\phi}\})_e + ([\![\boldsymbol{p}]\!], B_{02}\boldsymbol{\phi})_e, \quad \forall\boldsymbol{\phi}, \tag{3.32a}$$

$$([\![\psi]\!], \boldsymbol{B_{01}}\boldsymbol{p}+\boldsymbol{B_{11}}\lambda-\{\boldsymbol{p}\})_e = ([\![\boldsymbol{p}]\!], B_{12}\psi+\{\psi\})_e + ([\![\lambda]\!], \boldsymbol{B_{11}}\psi)_e, \quad \forall\psi. \tag{3.32b}$$

We omit the conditions for boundary edges, since the boundary conditions can always be modified to make the method commutative if necessary. We summarize the above result in the following proposition.

Proposition 3.7 *Assume that the optimal control problem* (3.6) *is discretized by a DG method that can be put in the form of* (3.21) *with given operators* $\boldsymbol{B_{00}}$, $\boldsymbol{B_{01}}$, B_{02}, B_{10}, $\boldsymbol{B_{11}}$, B_{12}, B_{D_0}, B_{N_0}, B_{D_1}, *and* B_{N_1}. *Then in order for the two approaches optimize-then-discretize and discretize-then-optimize to commute the operators must satisfy* (3.30) *and* (3.31) *in the interior of each triangle, and* (3.32) *on each interior edge.*

In the Table 3.2 we list the most common DG methods and report if they are commutative or not.

Table 3.2: Results for some common DG methods in mixed form

Method	$\boldsymbol{B_{00}\sigma}$	$\boldsymbol{B_{01}\sigma}$	$B_{02}\boldsymbol{\sigma}$	$B_{10}v$	$\boldsymbol{B_{11}}v$	$B_{12}v$	commutative
CIP [17]	$-\boldsymbol{\sigma}$	$[\![y]\!] \equiv 0$	0	v	$[\![y]\!] \equiv 0$	$-v + c_2[\![\nabla v]\!]$	yes
B.R. [3]	$-\boldsymbol{\sigma}$	$\{\boldsymbol{\sigma}\}$	0	v	0	$-\{v\}$	yes
LDG [14]	$-\boldsymbol{\sigma}$	$\{\boldsymbol{\sigma}\} + \boldsymbol{\gamma}[\![\boldsymbol{\sigma}]\!]$	0	v	$c_1[\![v]\!]$	$-\{v\} + \boldsymbol{\gamma} \cdot [\![v]\!]$	yes
C.C.P.S. [11]	$-\boldsymbol{\sigma}$	$\{\boldsymbol{\sigma}\} + \boldsymbol{\gamma}[\![\boldsymbol{\sigma}]\!]$	$c_2[\![\boldsymbol{\sigma}]\!]$	v	$c_1[\![v]\!]$	$-\{v\} + \boldsymbol{\gamma} \cdot [\![v]\!]$	yes
SIPG [1]	$-\boldsymbol{\sigma}$	$\{\boldsymbol{\sigma}\}$	0	v	$c_1[\![v]\!]$	$-\{v\}$	yes
NIPG [30]	$-\boldsymbol{\sigma}$	$-\{\boldsymbol{\sigma}\}$	0	v	$c_1[\![v]\!]$	$-\{v\}$	no
D.S.W. [30]	$-\boldsymbol{\sigma}$	0	0	v	$c_1[\![v]\!]$	$-\{v\}$	no
B.O. [4]	$-\boldsymbol{\sigma}$	$-\{\boldsymbol{\sigma}\}$	0	v	0	$-\{v\}$	no
[8, (2.56)-(2.57)]	$-\boldsymbol{\sigma}$	$\{\boldsymbol{\sigma}\}$	0	v	$c_0\{\nabla v\}$	$-\{v\}$	no
H.M. [27]	$-\boldsymbol{\sigma} + c_e\boldsymbol{\sigma}$	$\{\boldsymbol{\sigma}\}$	0	v	$\frac{c_e}{1-c_e}\{\nabla v\}$	$-\{v\}$	no

3.3 Advection-diffusion-reaction equation

Now we consider the problem with the advection-diffusion-reaction state equation (1.2). An additional difficulty lies in the fact that the advection field in the adjoint equation is the opposite of the advection field of the state equation. Since the operators B may depend on the advection field $\boldsymbol{\beta}$, for the adjoint equation we need a separate set of operators which we will denote B^* (see Remark 3.10). The optimality conditions for continuous problem are listed in (2.6). Following [2], we rewrite the state equation as

$$\begin{aligned}
\nabla \cdot (-\varepsilon \nabla y(x) + \boldsymbol{\beta} y(x)) + r(x)y(x) &= f(x) + u(x), && \text{in each } \tau \in T_h,\\
[\![y(x)]\!] &= 0, && \text{on each } e \in \mathcal{E}_h^0,\\
[\![-\varepsilon \nabla y(x) + \boldsymbol{\beta} y(x)]\!] &= 0, && \text{on each } e \in \mathcal{E}_h^0,\\
y(x) &= g_D(x), && \text{on each } e \in \Gamma_D,\\
\varepsilon \frac{\partial}{\partial \mathbf{n}} y(x) &= g_N(x), && \text{on each } e \in \Gamma_N.
\end{aligned}$$

Assume we have operators B_0, $\boldsymbol{B_1}$, B_2, B_D, and B_N. We consider the following discrete problem, find $y \in V_h$ such that for any $v \in V_h$

$$a_h(y,u;v) \stackrel{\text{def}}{=} \sum_{\tau\in T_h} (\nabla\cdot(-\varepsilon\nabla y + \boldsymbol{\beta} y) + ry - f - u, B_0 v)_\tau \tag{3.33}$$

$$+ \sum_{e\in\mathcal{E}_h^0} (\llbracket y\rrbracket, \boldsymbol{B_1} v)_e + (\llbracket -\varepsilon\nabla y + \boldsymbol{\beta} y\rrbracket, B_2 v)_e \tag{3.34}$$

$$+ \sum_{e\in\Gamma_D} (y - g_D, B_D v)_e + \sum_{e\in\Gamma_N} (\varepsilon\frac{\partial y}{\partial \mathbf{n}} - g_N, B_N v)_e.$$

Example 3.8 *By taking*

$$B_0 v := v, \qquad \forall \tau \in T_h, \tag{3.35a}$$

$$\boldsymbol{B_1} v := -\{\varepsilon\nabla v\} + \varepsilon\frac{\sigma}{|e|}\llbracket v\rrbracket + \frac{\mathbf{n}^+}{2}\llbracket \boldsymbol{\beta} v\rrbracket, \qquad \forall e \in \mathcal{E}_h^0, \tag{3.35b}$$

$$B_2 v := -\{v\}, \qquad \forall e \in \mathcal{E}_h^0, \tag{3.35c}$$

$$B_D v := -\varepsilon\frac{\partial v}{\partial \mathbf{n}} + \varepsilon\frac{\sigma}{|e|} v + |\boldsymbol{\beta}\cdot\mathbf{n}| v \chi_{\Gamma_D^-}, \qquad \forall e \in \Gamma_D, \tag{3.35d}$$

$$B_N v := v, \qquad \forall e \in \Gamma_N, \tag{3.35e}$$

we obtain the usual symmetric interior penalty method (SIPG) with upwinding. Here $\chi_{\Gamma_D^-}$ denotes the characteristic function on Γ_D^-. For this example to insure the stability one must take σ to be sufficiently large. Notice that in contrast to the Laplace equation, the advection-diffusion equation operators depend on the direction of the advection field $\boldsymbol{\beta}$.

Remark 3.9 *The Streamline Upwind Stabilized Petrov-Galerkin (SUPG) [10] or the Least-Squares method [6] can be put in this framework by considering continuous elements and choosing $B_0 v := v + \boldsymbol{\beta}\cdot\nabla v$ or $B_0 v := v + \nabla\cdot(-\varepsilon\nabla v(x) + \boldsymbol{\beta} v(x)) + r(x)v(x)$ respectively, with the appropriate choices of $\boldsymbol{B_1}$, B_2, B_D, and B_N. However, since the commutativity conditions always require $B_0 v := v$, these methods can never be commutative.*

3.3.1 Optimize-then-Discretize system

Applying the disretization above the optimality system (2.6), we obtain that a triplet $(y,u,\lambda) \in V_h \times V_h \times V_h$ is the unique solution of the following system consisting of the discretized adjoint equation

$$\sum_{\tau\in T_h} (\nabla\cdot(-\varepsilon\nabla\lambda - \boldsymbol{\beta}\lambda) + (r + \nabla\cdot\boldsymbol{\beta})\lambda - \hat{y} + y, B_0^*\phi)_\tau$$

$$+ \sum_{e\in\mathcal{E}_h^0} (\llbracket\lambda\rrbracket, \boldsymbol{B_1^*}\phi)_e + (\llbracket -\varepsilon\nabla\lambda - \boldsymbol{\beta}\lambda\rrbracket, B_2^*\phi)_e \tag{3.36a}$$

$$+ \sum_{e\in\Gamma_D} (\lambda, B_D^*\phi)_e + \sum_{e\in\Gamma_N} (\varepsilon\frac{\partial\lambda}{\partial\mathbf{n}} + \boldsymbol{\beta}\cdot\mathbf{n}\lambda, B_N^*\phi)_e = 0, \quad \forall \phi \in V_h,$$

the discretized gradient equation

$$(\psi, \lambda)_\tau - \alpha(u, \psi)_\tau = 0, \quad \forall \psi \in V_h, \tag{3.36b}$$

and the discretized state equation

$$\begin{aligned} &\sum_{\tau \in T_h} (\nabla \cdot (-\varepsilon \nabla y + \boldsymbol{\beta} y) + ry - f - u, B_0 \varphi)_\tau \\ &+ \sum_{e \in \mathcal{E}_h^0} ([\![y]\!], \boldsymbol{B_1} \varphi)_e + ([\![-\varepsilon \nabla y + \boldsymbol{\beta} y]\!], B_2 \varphi)_e \\ &+ \sum_{e \in \Gamma_D} (y - g_D, B_D \varphi)_e + \sum_{e \in \Gamma_N} (\varepsilon \frac{\partial y}{\partial \mathbf{n}} - g_N, B_N \varphi)_e = 0, \quad \forall \varphi \in V_h. \end{aligned} \tag{3.36c}$$

Remark 3.10 *Notice that the adjoint equation is again an advection-diffusion-reaction equation, but advection the field is $-\boldsymbol{\beta}$ instead of $\boldsymbol{\beta}$. Since the operators can depend on $\boldsymbol{\beta}$ and its direction, the choice of some operators B_0^*, $\boldsymbol{B_1^*}$, B_2^*, B_D^*, and B_N^* for the adjoint equation can differ from the choices of B_0, $\boldsymbol{B_1}$, B_2, B_D, and B_N for the state equation. Thus for example for the SIPG method we need*

$$B_D^* v := -\varepsilon \frac{\partial v}{\partial \mathbf{n}} + \varepsilon \frac{\sigma}{|e|} v + |\boldsymbol{\beta} \cdot \mathbf{n}| v \chi_{\Gamma_D^+}, \quad \forall e \in \Gamma_D.$$

3.3.2 Discretize-then-Optimize system

Now we derive the optimality conditions for *discretize-then-optimize* approach. The problem is

$$\begin{aligned} \underset{\{y,u\} \in V_h \times V_h}{\text{minimize}} \quad & \frac{1}{2}\|y - \widehat{y}\|^2 + \frac{\alpha}{2}\|u\|^2, && (3.37a) \\ \text{subject to} \quad & a_h(y, u; v) = 0, \quad \forall v \in V_h. && (3.37b) \end{aligned}$$

The discrete Lagrangian is given by

$$L_h(y, u, \lambda) = \frac{1}{2}\|y - \widehat{y}\|^2 + \frac{\alpha}{2}\|u\|^2 + a_h(y, u; \lambda). \tag{3.38}$$

The necessary and, for our model problem, sufficient optimality conditions again can be obtained by setting the partial Fréchet-derivatives of (3.38) with respect to the discrete state y, discrete control u, and discrete adjoint λ equal to zero. Thus we obtain the following system consisting of the discrete adjoint equation

$$\begin{aligned} \frac{\partial L_h}{\partial y}\phi = &\sum_{\tau \in T_h} (\nabla \cdot (-\varepsilon \nabla \phi - \boldsymbol{\beta}\phi) + r\phi, B_0 \lambda)_\tau + (y - \hat{y}, \phi)_\tau \\ &+ \sum_{e \in \mathcal{E}_h^0} ([\![\phi]\!], \boldsymbol{B_1} \lambda)_e + ([\![-\varepsilon \nabla \phi - \boldsymbol{\beta}\phi]\!], B_2 \lambda)_e \\ &+ \sum_{e \in \Gamma_D} (\phi, B_D \lambda)_e + \sum_{e \in \Gamma_N} (\varepsilon \frac{\partial \phi}{\partial \mathbf{n}}, B_N \lambda)_e = 0, \quad \forall \phi \in V_h, \end{aligned} \tag{3.39a}$$

the discrete gradient equation

$$\frac{\partial L_h}{\partial u}\psi = (\psi, B_0\lambda)_\tau - \alpha(u,\psi)_\tau = 0, \quad \forall \psi \in V_h, \tag{3.39b}$$

and the discrete state equation

$$\begin{aligned}\frac{\partial L_h}{\partial \lambda}\varphi = &\sum_{\tau\in T_h}(\nabla\cdot(-\varepsilon\nabla y + \boldsymbol{\beta} y) + ry - f - u, B_0\varphi)_\tau \\ &+ \sum_{e\in\mathcal{E}_h^0}(\llbracket y\rrbracket, \boldsymbol{B_1}\varphi)_e + (\llbracket -\varepsilon\nabla y + \boldsymbol{\beta} y\rrbracket, B_2\varphi)_e \\ &+ \sum_{e\in\Gamma_D}(y - g_D, B_D\varphi)_e + \sum_{e\in\Gamma_N}(\varepsilon\frac{\partial y}{\partial \mathbf{n}} - g_N, B_N\varphi)_e = 0, \quad \forall\varphi\in V_h.\end{aligned} \tag{3.39c}$$

3.3.3 Commutativity Conditions.

Comparing (3.39b) with (3.36b), similar to the Laplace equation, for the methods to be commutative it is required

$$B_0^* v = B_0 v := v. \tag{3.40}$$

From now on we assume that. Integrating $(\nabla\cdot(-\varepsilon\nabla\phi + \boldsymbol{\beta}\phi), \lambda)_\tau$ by parts and rearranging the terms we obtain

$$\begin{aligned}(\nabla\cdot(-\varepsilon\nabla\phi + \boldsymbol{\beta}\phi, \lambda)_\tau &= (\varepsilon\nabla\phi - \boldsymbol{\beta}\phi, \nabla\lambda)_\tau + (-\varepsilon\frac{\partial\phi}{\partial\mathbf{n}} + \boldsymbol{\beta}\cdot\mathbf{n}\phi, \lambda)_{\partial\tau} \\ &= (\phi, -\nabla\cdot(\varepsilon\nabla\lambda))_\tau + (\phi, \varepsilon\frac{\partial\lambda}{\partial\mathbf{n}})_{\partial\tau} + (\phi, -\boldsymbol{\beta}\cdot\nabla\lambda)_\tau + (-\varepsilon\frac{\partial\phi}{\partial\mathbf{n}} + \boldsymbol{\beta}\cdot\mathbf{n}\phi, \lambda)_{\partial\tau} \\ &= (\nabla\cdot(-\varepsilon\nabla\lambda - \boldsymbol{\beta}\lambda) + (\nabla\cdot\boldsymbol{\beta})\lambda, \phi)_\tau + (\phi, \varepsilon\frac{\partial\lambda}{\partial\mathbf{n}})_{\partial\tau} + (-\varepsilon\frac{\partial\phi}{\partial\mathbf{n}} + \boldsymbol{\beta}\cdot\mathbf{n}\phi, \lambda)_{\partial\tau}.\end{aligned}$$

Summing over all elements τ and using (3.4), we can rewrite (3.39a) as

$$\begin{aligned}&\sum_{\tau\in T_h}(\nabla\cdot(-\varepsilon\nabla\lambda - \boldsymbol{\beta}\lambda) + (r + \nabla\cdot\boldsymbol{\beta})\lambda, \phi)_\tau + (y - \hat{y}, \phi)_\tau \\ &+ \sum_{e\in\mathcal{E}_h^0}(\llbracket\varepsilon\nabla\lambda\rrbracket, \{\phi\})_e + (\{\varepsilon\nabla\lambda\}, \llbracket\phi\rrbracket)_e + (\llbracket -\varepsilon\nabla\phi + \boldsymbol{\beta}\phi\rrbracket, \{\lambda\})_e + (\{-\varepsilon\nabla\phi + \boldsymbol{\beta}\phi\}, \llbracket\lambda\rrbracket)_e \\ &+ \sum_{e\in\mathcal{E}_h^0}(\llbracket\phi\rrbracket, \boldsymbol{B_1}\lambda)_e + (\llbracket -\varepsilon\nabla\phi - \boldsymbol{\beta}\phi\rrbracket, B_2\lambda)_e \\ &+ \sum_{e\in\mathcal{E}_h^\partial}(\varepsilon\frac{\partial\lambda}{\partial\mathbf{n}}, \phi)_e + (\lambda, -\varepsilon\frac{\partial\phi}{\partial\mathbf{n}} + \boldsymbol{\beta}\cdot\mathbf{n}\phi)_e \\ &+ \sum_{e\in\Gamma_D}(\phi, B_D\lambda)_e + \sum_{e\in\Gamma_N}(\varepsilon\frac{\partial\phi}{\partial\mathbf{n}}, B_N\lambda)_e = 0, \quad \forall\phi\in V_h.\end{aligned} \tag{3.41}$$

Now directly comparing (3.41) with (3.36a) in order to have a commutative method we need

$$\begin{aligned}&([\![\lambda]\!], \boldsymbol{B_1^*}\phi + \{\varepsilon\nabla\phi - \boldsymbol{\beta}\phi\})_e - ([\![\varepsilon\nabla\lambda]\!], B_2^*\phi + \{\phi\})_e - ([\![\boldsymbol{\beta}\lambda]\!], B_2\phi)_e \\ &= ([\![\phi]\!], \boldsymbol{B_1}\lambda + \{\varepsilon\nabla\lambda\})_e - ([\![\varepsilon\nabla\phi]\!], B_2\lambda + \{\lambda\})_e + ([\![\boldsymbol{\beta}\phi]\!], B_2\lambda + \{\lambda\})_e,\end{aligned} \tag{3.42}$$

on each interior edge. In addition on the boundary edges we also need

$$(\lambda, B_D^*\phi + \varepsilon\frac{\partial\phi}{\partial\mathbf{n}} - \boldsymbol{\beta}\cdot\mathbf{n}\phi)_e = (\phi, B_D\lambda + \varepsilon\frac{\partial\lambda}{\partial\mathbf{n}})_e$$

on each Dirichlet edge and

$$(\varepsilon\frac{\partial\lambda}{\partial\mathbf{n}} + \boldsymbol{\beta}\cdot\mathbf{n}\lambda, B_N^*\phi - \phi)_e = (\varepsilon\frac{\partial\phi}{\partial\mathbf{n}}, B_N\lambda - \lambda)_e.$$

on each Neumann edge. Direct computations show that these conditions are satisfied for Example 3.8. Of course, if the diffusion part discretized with non-commutative method then the method for advection-diffusion equation can not be commutative. For example in [2] several new stable methods were proposed for a single equation. However none of the proposed methods is commutative.

4 Global error estimates for the SIPG methods

In this section we derive global error estimates for upwind SIPG method (cf. Example 3.8), which satisfies the symmetry condition. In the following analysis it is convenient to separate diffusion part from the advection-reaction part. Thus we define

$$a_h^d(y,v) := \varepsilon\sum_{\tau\in T_h}(\nabla y, \nabla v)_\tau + \varepsilon\sum_{e\in\mathcal{E}_h}\left(\frac{\sigma}{|e|}([\![y]\!],[\![v]\!])_e - (\{\nabla y\},[\![v]\!])_e - ([\![y]\!],\{\nabla v\})_e\right) \tag{4.1}$$

$$a_h^{ar}(y,v) := \sum_{\tau\in T_h}(\nabla\cdot(\boldsymbol{\beta}y) + ry, v)_\tau + \sum_{e\in\mathcal{E}_h^0}\left(|\mathbf{n}\cdot\boldsymbol{\beta}|(y^+ - y^-), v^+\right)_e + \sum_{e\in\mathcal{E}_h^-}\left(|\mathbf{n}\cdot\boldsymbol{\beta}|y^+, v^+\right). \tag{4.2}$$

The discontinuous Galerkin discretization of the state equation (1.2) for a fixed control u is now given as follows (cf. (2.2a)). Find $y\in V_h$ such that

$$a_h(y,v) - (u,v) = l_h(v) \qquad \forall v\in V_h, \tag{4.3}$$

where $a_h(y,v) = a_h^d(y,v) + a_h^{ar}(y,v)$ and for each $v\in V_h$

$$l_h(v) = \sum_{\tau\in T_h}(f,v)_\tau + \varepsilon\sum_{e\in\mathcal{E}_h^\partial}\left(\frac{\sigma}{|e|}(g_D,v)_e - (g_D,v)_e\right) + \sum_{e\in\mathcal{E}_h^-}\left(|\mathbf{n}\cdot\boldsymbol{\beta}|g_D, v^+\right)_e + \langle g_N, v\rangle_{\Gamma_N}. \tag{4.4}$$

4.1 Preliminaries

Here we list some of the known properties of the SIPG method. The presentation and the results we adapt from [2]. Since the state equation (1.2) for a smooth $\boldsymbol{\beta}$ can be rewritten as

$$-\Delta y(x) + \boldsymbol{\beta}(x)\cdot\nabla y(x) + (r(x) + \nabla\cdot\boldsymbol{\beta})y(x) = f(x) + u(x),$$

we introduce the "effective" reaction function $\rho(x)$ with the assumption

$$\rho(x) = r(x) + \tfrac{1}{2}\nabla\cdot\boldsymbol{\beta}(x) \geq \rho_0 \geq 0 \text{ a.e. in } \Omega. \tag{4.5}$$

Following [2] we make the following assumptions on the advective field $\boldsymbol{\beta}$.

Assumption 4.1 *$\boldsymbol{\beta}$ has no closed curves and stationary points. It implies (cf. [2, App. A]) that*

$$\exists\, \eta \in W^{k+1}_\infty(\Omega) \quad \textit{such that} \quad \boldsymbol{\beta}\cdot\nabla\eta \geq 2b_0 := 2\frac{\|\boldsymbol{\beta}\|_{0,\infty}}{L} \textit{ in } \Omega, \quad \textit{where } L = diam(\Omega). \tag{4.6}$$

Assumption 4.2 *We also assume that*

$$\exists\, c_\beta > 0 \quad \textit{such that} \quad |\boldsymbol{\beta}| \geq c_\beta \|\boldsymbol{\beta}\|_{1,\infty} \quad \forall x \in \Omega, \tag{4.7}$$

and for a given shape-regular family T_h of decomposition of Ω into triangles τ,

$$\exists\, c_\rho > 0 \quad \textit{such that} \quad \forall \tau \in T_h \quad \|\rho\|_{0,\infty,\tau} \leq c_\rho(\min_\tau \rho(x) + b_0). \tag{4.8}$$

For more detailed description of these assumption we refer to [2, App. 2.1].

Primary, we will be working with a norm

$$|||v|||^2 = |||v|||_d^2 + |||v|||_{ar}^2, \tag{4.9a}$$

where

$$|||v|||_d^2 := \varepsilon|v|_{1,h}^2 + \sum_{e\notin\Gamma_N} \frac{\varepsilon}{|e|}\|[\![v]\!]\|_e^2, \tag{4.9b}$$

$$|||v|||_{ar}^2 := \|(\overline{\rho} + b_0)^{1/2} v\|^2 + \sum_{e\in\mathcal{E}_h} \||\boldsymbol{\beta}\cdot\mathbf{n}|^{1/2}[\![v]\!]\|_e^2, \tag{4.9c}$$

where $b_0 = \|\boldsymbol{\beta}\|_\infty/L$ defined in (4.6) and $\overline{\rho}$ is the piecewise constant function defined as

$$\rho(x)\,|_\tau = \min_{x\in\tau}\rho(x) \quad \forall \tau \in T_h. \tag{4.10}$$

Since we treat the case of $\rho = 0$, to show a stability result in $|||\cdot|||$ norm we need to introduce the weight function $\chi = e^{-\eta}$, with η from (4.6). From the assumptions on η there exists positive constants χ_1^*, χ_2^*, and χ_3^*, such that

$$\chi_1^* \leq \chi \leq \chi_2^*, \quad |\nabla\chi| \leq \chi_3^*. \tag{4.11}$$

Following [2], we define a weight function φ by

$$\varphi = \chi + \kappa, \tag{4.12}$$

where κ is sufficiently large number to be specified later (cf. [2, eq. 4.15] for more details). The inclusion of κ in the definition of the weight function avoids the restriction of considering only advection-dominated problems. The important stability and continuity results we state in the following lemma.

Lemma 4.3 *There exists constants c_1, c_2, χ_4^*, and χ_5^* such that for $a_h^d(\cdot,\cdot)$ and $a_h^{ar}(\cdot,\cdot)$ defined in* (4.1) *and any $u, v \in V_h$,*

$$a_h(u,v) \le c_1 |||u||| |||v|||, \tag{4.13a}$$

$$a_h^d(v,\varphi v) \ge \frac{\chi_1^* + \kappa}{6} |||v|||_d^2, \tag{4.13b}$$

$$a_h^{ar}(v,\varphi v) \ge \frac{\chi_1^*}{2} |||v|||_{ar}^2, \tag{4.13c}$$

$$a_h^d(v,\varphi v - P_h(\varphi v)) \le \chi_4^* |||v|||_d^2, \tag{4.13d}$$

$$a_h^{ar}(v,\varphi v - P_h(\varphi v)) \le \chi_5^* (h/L)^{1/2} |||v|||_{ar}^2, \tag{4.13e}$$

$$|||P_h(\varphi v)||| \le c_2 |||v|||, \tag{4.13f}$$

where $P_h : L^2 \to V_h$ is the orthogonal L^2-projection defined by

$$(P_h u, v)_\tau = (u,v)_\tau, \quad \forall v \in V_h,\ \forall \tau \in T_h.$$

The proof of this lemma is given in [2, Lem. 4.1, Lem. 4.3] for a slightly different DG method. The arguments can easily be adapted for the SIPG method as well.

4.2 Energy error estimates

In this section we derive *a priori* error estimates for the control in the case of unconstrained problem. First we introduce two intermediate functions. Define $\widetilde{y}_h = \widetilde{y}_h(u) \in V_h$ for a given $u \in L^2$, to be the solution to

$$a_h(\widetilde{y}_h, v) = (u,v) + (f,v) + \langle g, v\rangle_{\Gamma_N}, \quad \forall v \in V_h. \tag{4.14}$$

Similarly we define $\widetilde{\lambda}_h = \widetilde{\lambda}_h(\widetilde{y}_h(u)) \in V_h$ to be the solution to the following equation

$$a_h(v, \widetilde{\lambda}_h) = (\widehat{y}, v) - (\widetilde{y}_h, v), \quad \forall v \in V_h. \tag{4.15}$$

Using the discrete and continuous gradient equations, we have

$$\begin{aligned} \alpha \|u - u_h\|^2 &= \alpha(u - u_h, u - u_h) = (\alpha u - \lambda, u - u_h) - (\alpha u_h - \lambda_h, u - u_h) + (\lambda - \lambda_h, u - u_h) \\ &= (\lambda - \widetilde{\lambda}_h, u - u_h) + (\widetilde{\lambda}_h - \lambda_h, u - u_h) := J_1 + J_2. \end{aligned} \tag{4.16}$$

Using the Cauchy-Schwarz and arithmetic-geometric mean inequalities we have,

$$J_1 = (\lambda - \widetilde{\lambda}_h, u - u_h) \le \frac{1}{2\alpha} \|\lambda - \widetilde{\lambda}_h\|^2 + \frac{\alpha}{2} \|u - u_h\|^2.$$

Next we will estimate $\|\lambda - \widetilde{\lambda}_h\|$. To accomplish this we will require the following two lemmas.

Lemma 4.4 *Let y be the exact solution to* (1.1)-(1.2) *and $\widetilde{y}_h$ be a solution of* (4.14) *with the exact control u. Assume $y \in H^s$, for some $s > 3/2$, then for h sufficiently small there exists a constant C_1 independent of u, λ, and y such that*

$$|||y - \widetilde{y}_h||| \le C_1 |||y - P_h y|||.$$

Proof: Since $y \in H^s$ for some $s > 3/2$ and the SIPG method is consistent, we have

$$a_h(\widetilde{y}_h - y, v) = 0 \quad \forall v \in V_h. \tag{4.17}$$

Put $\zeta_h = \widetilde{y}_h - P_h y$. Then from (4.13b) and (4.13c) we have,

$$\begin{aligned}\frac{\chi_1^* + \kappa}{6}|||\zeta_h|||_d^2 + \frac{\chi_1^*}{2}|||\zeta_h|||_{ar}^2 \le a_h(\zeta_h, \varphi\zeta_h) &= a_h(\zeta_h, \varphi\zeta_h - P_h(\varphi\zeta_h)) + a_h(\zeta_h, P_h(\varphi\zeta_h)) \\ &= a_h(\zeta_h, \varphi\zeta_h - P_h(\varphi\zeta_h)) + a_h(y - P_h y, P_h(\varphi\zeta_h)),\end{aligned} \tag{4.18}$$

where in the last step we used (4.17). From (4.13a), the Cauchy-Schwarz inequality, and (4.13f), we obtain

$$a_h(y - P_h y, P_h(\varphi\zeta_h)) \le c_1|||y - P_h y|||||P_h(\varphi\zeta_h)||| \le c_1 c_2|||y - P_h y|||||\zeta_h|||. \tag{4.19}$$

To estimate

$$a_h(\zeta_h, \varphi\zeta_h - P_h(\varphi\zeta_h)) = a_h^d(\zeta_h, \varphi\zeta_h - P_h(\varphi\zeta_h)) + a_h^{ar}(\zeta_h, \varphi\zeta_h - P_h(\varphi\zeta_h))$$

we use (4.13d) and (4.13e). Thus

$$a_h^d(\zeta_h, \varphi\zeta_h - P_h(\varphi\zeta_h)) + a_h^{ar}(\zeta_h, \varphi\zeta_h - P_h(\varphi\zeta_h)) \le \chi_4^*|||\zeta_h|||_d^2 + \chi_5^*(h/L)^{1/2}|||\zeta_h|||_{ar}^2.$$

Now if κ is so large that $\frac{\chi_1^*+\kappa}{12} \ge \chi_4^*$ and h is so small that $\chi_5^*(h/L)^{1/2} \le \frac{\chi_1^*}{4}$, then with $c_3 = \min\left(\frac{\chi_1^*+\kappa}{12}, \frac{\chi_1^*}{4}\right)$ and using (4.23), we have

$$c_3|||\zeta_h||| \le c_1 c_2|||y - P_h y|||. \tag{4.20}$$

Hence by the triangle inequality we have

$$|||y - \widetilde{y}_h||| \le \left(\frac{c_2 c_1}{c_3} + 1\right)|||y - P_h y|||,$$

which proves the lemma with $C_1 = \frac{c_2 c_1}{c_3} + 1$. □

Lemma 4.5 *Let λ be the exact adjoint and $\widetilde{\lambda}_h$ be a solution to* (4.15)*. Assume that $\lambda \in H^s(\Omega)$, $s > 3/2$ and h is sufficiently small. Then there exist constants C_2 and C_3, λ, y, and u such that*

$$|||\lambda - \widetilde{\lambda}_h||| \le C_2|||\lambda - P_h\lambda||| + C_3\|y - \widetilde{y}_h\|.$$

Proof: Proof is similar to the proof of Lemma 4.4. Since $y \in H^s$ for some $s > 3/2$ and the SIPG method is consistent, we have

$$a_h(v, \widetilde{\lambda}_h - \lambda) = (y - \widetilde{y}_h, v), \quad \forall v \in V_h. \tag{4.21}$$

Put $\zeta_h = \widetilde{\lambda}_h - P_h\lambda$. Then from (4.13b) and (4.13c) we have,

$$\begin{aligned}\frac{\chi_1^* + \kappa}{6}|||\zeta_h|||_d^2 + \frac{\chi_1^*}{2}|||\zeta_h|||_{ar}^2 \le a_h(\varphi\zeta_h, \zeta_h) &= a_h(\varphi\zeta_h - P_h(\varphi\zeta_h), \zeta_h) + a_h(P_h(\varphi\zeta_h), \zeta_h) \\ &= a_h(\varphi\zeta_h - P_h(\varphi\zeta_h), \zeta_h,) + a_h(P_h(\varphi\zeta_h), \lambda - P_h\lambda) + (y - \widetilde{y}_h, P_h(\varphi\zeta_h)),\end{aligned} \tag{4.22}$$

where in the last step we used (4.21). From (4.13a), the Cauchy-Schwarz inequality, and (4.13f), we obtain

$$\begin{aligned} a_h(P_h(\varphi\zeta_h), \lambda - P_h\lambda) + (y - \widetilde{y}_h, P_h(\varphi\zeta_h)) &\le c_1 |||\lambda - P_h\lambda||| |||P_h(\varphi\zeta_h)||| + \|y - \widetilde{y}_h\| \|P_h(\varphi\zeta_h)\| \\ &\le (c_6 |||\lambda - P_h\lambda||| + \|y - \widetilde{y}_h\|)\, c_5 |||\zeta_h|||. \end{aligned} \tag{4.23}$$

Similarly to the proof of the previous lemma, we have

$$c_4 |||\zeta_h||| \le c_5 \left(c_6 |||\lambda - P_h\lambda||| + \|y - \widetilde{y}_h\|\right). \tag{4.24}$$

Hence by the triangle inequality we have

$$|||\lambda - \widetilde{\lambda}_h||| \le \left(\frac{c_5 c_6}{c_4} + 1\right) |||\lambda - P_h\lambda||| + \frac{c_5}{c_4}\|y - \widetilde{y}_h\|,$$

which proves the lemma with $C_2 = \frac{c_5 c_6}{c_4} + 1$ and $C_3 = \frac{c_5}{c_4}$. □

Since $(\rho_0 + b_0)^{1/2}\|\lambda - \widetilde{\lambda}_h\| \le |||\lambda - \widetilde{\lambda}_h|||$, from Lemma 4.4 and Lemma 4.5, it follows that

$$J_1 \le \frac{C}{\alpha}\left(|||y - P_h y|||^2 + |||\lambda - P_h\lambda|||^2\right) + \frac{\alpha}{2}\|u - u_h\|^2. \tag{4.25}$$

Next we will show that $J_2 \le 0$. Using that $(y_h - \widetilde{y}_h) \in V_h$ and $(\lambda_h - \widetilde{\lambda}_h) \in V_h$ and the definitions of $\widetilde{y}_h$ and $\widetilde{\lambda}_h$ we have

$$J_2 = (\widetilde{\lambda}_h - \lambda_h, u - u_h) = a_h(\widetilde{y}_h - y_h, \widetilde{\lambda}_h - \lambda_h) = -(\widetilde{y}_h - y_h, \widetilde{y}_h - y_h) = -\|\widetilde{y}_h - y_h\|^2 \le 0. \tag{4.26}$$

From the above estimates we can derive the following error estimates for the optimal control problem.

Theorem 4.6 *Let y, u, λ be the state, control, and adjoint solutions to the optimal control system* (2.6)*, and let y_h, u_h, λ_h be the discrete solutions obtained by the SIPG method. Assume that $y \in H^s(\Omega)$ for some $s > 3/2$ and h is sufficiently small. Then, there exist a constant C independent of y, u, and λ such that*

$$|||y - y_h||| + \frac{1}{\alpha}|||\lambda - \lambda_h||| + \|u - u_h\| \le \frac{C}{\alpha}\left(|||\lambda - P_h\lambda||| + |||y - P_h y|||\right).$$

Proof: From the estimates (4.25) and (4.26) it follows that

$$\|u - u_h\| \le \frac{C}{\alpha}\left(|||\lambda - P_h\lambda||| + |||y - P_h y|||\right).$$

From the state equation we have $|||y - y_h||| \le C\|u - u_h\|$ and from the gradient equation we have $\alpha(u - u_h) = \lambda - \lambda_h$. Hence we have the theorem. □

Using the approximation theory of the L^2-projection, we can easily obtain

$$|||v - P_h v||| \le C h^k \left(\varepsilon^{1/2} + \|\boldsymbol{\beta}\|_\infty h^{1/2} + (\|\rho\|_\infty + b_0)^{1/2} h\right) |v|_{k+1},$$

hence we have the following result.

Corollary 4.7 *Let the solution y, u, λ of the optimal control problem to be in H^k. Then there is a constant C independent of y, u, and λ such that*

$$\begin{aligned}|||y - y_h||| + \frac{1}{\alpha}|||\lambda - P_h\lambda||| + \|u - u_h\| \\ \leq C\alpha^{-1}h^k\left(\varepsilon^{1/2} + \|\boldsymbol{\beta}\|_\infty h^{1/2} + (\|\rho\|_\infty + b_0)^{1/2}h\right)(|y|_{k+1} + |\lambda|_{k+1}).\end{aligned}$$

In summary,

$$\|u - u_h\| \leq \alpha^{-1}\begin{cases} O(h^k), & \textit{if diffusion dominated;} \\ O(h^{k+1/2}), & \textit{if advection dominated;} \\ O(h^{k+1}), & \textit{if reaction dominated.}\end{cases}$$

The above error estimate is optimal for advection and reaction dominated problems, but suboptimal for diffusion dominated problems.

4.3 L^2-error estimates

In this section we derive optimal error estimates in L^2 norm for diffusion dominated problem. For simplicity of the presentation we assume the constant advection field $\boldsymbol{\beta}$. For $e_y = y - y_h$ and $e_\lambda = \lambda - \lambda_h$ let z, v, p be a solution to a dual system

$$\nabla \cdot (-\varepsilon\nabla z - \boldsymbol{\beta} z) + rz + p = e_y \tag{4.27a}$$

$$\alpha v - z = 0 \tag{4.27b}$$

$$\nabla \cdot (-\varepsilon\nabla p + \boldsymbol{\beta} p) + rp - v = e_\lambda. \tag{4.27c}$$

The following theorem was shown in [22] for the optimal control problem (1.1)-(1.2).

Theorem 4.8 *Let Ω be a bounded open convex subset of $\mathbb{R}^n$, $\boldsymbol{\beta}$ be a constant vector, and $f, \widehat{y} \in L^2(\Omega)$. Then there exists a positive constant C independent of ε such that the unique solution of the optimal control problem* (1.1)-(1.2) *and the associated adjoint satisfy*

$$\varepsilon^{3/2}\left(\|y\|_2 + \|\lambda\|_2\right) \leq C\left(\|f\| + \|\widehat{y}\|\right).$$

Since the adjoint system is equivalent to the following ("dual") optimal control problem

$$\min_{p,v} \frac{1}{2}\|p - e_y\|^2 + \frac{\alpha}{2}\|v\|^2 \tag{4.28}$$

subject to second order advection-diffusion equation

$$\nabla \cdot (-\varepsilon\nabla p(x) + \boldsymbol{\beta}(x)p(x)) + r(x)p(x) = v(x) + e_\lambda(x), \qquad x \in \Omega, \tag{4.29a}$$

$$p(x) = 0, \qquad x \in \Gamma_D, \tag{4.29b}$$

$$\varepsilon\frac{\partial}{\partial \mathbf{n}}p(x) = 0, \qquad x \in \Gamma_N, \tag{4.29c}$$

similar argument give us the following regularity estimate for the adjoint system (4.27),

$$\varepsilon^{3/2}(\|p\|_2 + \|\lambda\|_2) \leq C\left(\|e_y\| + \|e_\lambda\|\right). \tag{4.30}$$

Define the discrete bilinear form for the original optimal control problem to be

$$\mathcal{A}_h(\{y,u,\lambda\},\{\phi,\varphi,\psi\}) = a_h(y,\phi) - (u,\phi) + \alpha(u,\varphi) - (\lambda,\varphi) + a_h(\psi,\lambda) + (y,\psi).$$

Since the SIPG method is consistent we have the following Galerkin orthogonality condition

$$\mathcal{A}_h(\{e_y,e_u,e_\lambda\},\{\phi,\varphi,\psi\}) = 0, \quad \forall\{\phi,\varphi,\psi\} \in V_h \times V_h \times V_h. \tag{4.31}$$

From the dual system (4.27) we have

$$\begin{aligned}\|e_y\|^2 + \|e_\lambda\|^2 = &(\nabla\cdot(-\varepsilon\nabla z - \boldsymbol{\beta} z) + rz + p, e_y) + (\alpha v - z, e_u)\\ &+ (\nabla\cdot(-\varepsilon\nabla p + \boldsymbol{\beta} p) + rp - v, e_\lambda).\end{aligned}$$

Writing the above expression as a sum over all elements and integrating by parts, we have

$$\begin{aligned}(\nabla\cdot(-\varepsilon\nabla z - \boldsymbol{\beta} z), e_y) &= \sum_\tau (\nabla\cdot(-\varepsilon\nabla z - \boldsymbol{\beta} z), e_y)_\tau\\ &= \sum_\tau \varepsilon(\nabla z, \nabla e_y)_\tau + (z, \boldsymbol{\beta}\cdot\nabla e_y)_\tau - \varepsilon(\nabla z\cdot\mathbf{n}, e_y)_{\partial\tau} - (\boldsymbol{\beta}\cdot\mathbf{n} z, e_y)_{\partial\tau}\end{aligned}$$

Using the fact that z is continuous, and as a result $[\![z]\!] = 0$, we have

$$\begin{aligned}(\nabla\cdot(-\varepsilon\nabla z - \boldsymbol{\beta} z) + rz, e_y) = &\sum_\tau \varepsilon(\nabla z, \nabla e_y)_\tau + (z, (r + \boldsymbol{\beta}\cdot\nabla) e_y)_\tau\\ &+ \sum_e -\varepsilon(\{\nabla z\}, [\![e_y]\!])_e + (|\boldsymbol{\beta}\cdot\mathbf{n}|(e_y^+ - e_y^-), z)_e = a_h(e_y, z).\end{aligned}$$

Similarly, we have

$$(\nabla\cdot(-\varepsilon\nabla p + \boldsymbol{\beta} p) + rp, e_\lambda) = a_h(p, e_\lambda).$$

Thus,

$$\|e_y\|^2 + \|e_\lambda\|^2 = a_h(e_y, z) + (p, e_y) + (\alpha v, e_u) - (z, e_u) + a_h(p, e_\lambda) + (v, e_\lambda).$$

Using the Galerkin orthogonality (4.31) we have

$$\|e_y\|^2 + \|e_\lambda\|^2 = a_h(e_y, z - I_h z) + (e_y, p - I_h p) + \alpha(e_u, v - I_h v) - (e_u, z - I_h z) + a_h(p - I_h p, e_\lambda) + (e_\lambda, v - I_h v), \tag{4.32}$$

where $I_h : C^0 \to S_h$ is the usual continuous interpolant on the space of continuous piecewise linear functions S_h. To show that

$$a_h(e_y, z - I_h z) \leq Ch|||e_y|||\|z\|_2,$$

we notice that since $z - I_h z$ is continuous, $[\![z - I_h z]\!] = 0$ and we have

$$\begin{aligned}a_h(e_y, z - I_h z) = &\sum_\tau \varepsilon(\nabla e_y, \nabla(z - I_h z))_\tau + (\boldsymbol{\beta}\cdot\nabla e_y, z - I_h z)_\tau\\ &+ \sum_e -\varepsilon([\![e_y]\!], \{\nabla(z - I_h z)\})_e + (|\boldsymbol{\beta}\cdot n|(e_y^+ - e_y^-), z - I_h z)_e = J_1 + J_2 + J_3 + J_4.\end{aligned}$$

By the Cauchy-Schwarz inequality and the standard approximation property of the interpolant I_h, we have

$$J_1 + J_2 + J_4 \leq Ch|||e_y|||\|z\|_2.$$

Using the trace inequality we obtain

$$J_3 = \sum_e \sqrt{\frac{\sigma}{h}}\|[\![e_y]\!]\|\sqrt{\frac{h}{\sigma}}\|\nabla(z - I_h z)\|_e \leq \left(\sum_e \frac{\sigma}{h}\|[\![e_y]\!]\|^2\right)^{1/2} C\sum_\tau \|\nabla(z - I_h z)\|_\tau^2 \leq Ch|||e_y|||\|z\|_2.$$

Similarly we can obtain

$$a_h(p - I_h p, e_\lambda) \leq Ch|||e_\lambda|||\ \|p\|_2,$$

and as a result

$$\|e_y\|^2 + \|e_\lambda\|^2 \leq Ch(\|z\|_2 + \|v\|_2 + \|p\|_2)(|||e_y||| + |||e_u||| + |||e_\lambda|||).$$

Using now the H^2 regularity (4.30) and the fact that $\alpha u = \lambda$ we obtain the following result.

Theorem 4.9 (L^2-error estimate) *Let y, u, λ be the state, control, and adjoint solutions to the optimal control system* (2.6)*, and let y_h, u_h, λ_h be the discrete solutions obtained by SIPG method. Assume the advection field $\boldsymbol{\beta}$ is constant and Ω is convex. Then for h is sufficiently small there exists a constant C independent of y, u, and λ such that*

$$\|y - y_h\| + \|u - u_h\| + \|\lambda - \lambda_h\| \leq C_\alpha h(|||y - y_h||| + |||u - u_h||| + |||\lambda - \lambda_h|||).$$

4.4 NIPG method

Examining the proof of Theorem 4.6, one can see that to the derive the error estimates in the energy norm the only properties of the SIPG we used are (4.13) and the consistency of the method for the state and the adjoint equations. Since the NIPG method with upwinding satisfies the same properties (cf. [2]), the result of Theorem 4.6 also holds for the NIPG method for the *optimize-then-discretize* approach. Of course since the NIPG method is not adjointly consistent, the L^2-error estimates are suboptimal even for a single equation. Our numerical examples in the next section illustrate that for the optimal control problems.

The situation with *discretize-then-optimize* approach for NIPG method is more peculiar. In this situation the adjoint equation is not consistent and as a result the Lemma 4.5 does not hold and we can not expect any convergence in the energy norm for the adjoint and control variable. This indeed is confirmed by the numerical experiments in the next section. However, the duality argument of Section 4.3 goes through since the dual problem for the adjoint equation is now consistent for the NIPG method. Thus, for diffusion dominated problems one can show the first order convergence for the adjoint and the control variable in the L^2 norm. This first order convergence is observed by the numerical experiments in the next section.

5 Numerical Results

In this section we provide several numerical examples that illustrate how the *optimize-then-discretize* and the *discretize-then-optimize* approaches may have substantially different numerical solutions for non-commutative methods.

5.1 Example 1

In the first example we show that the choice of the approach may have affect on the order of convergence. We illustrate it by considering a problem (1.1)-(1.2) with

$$\varepsilon = 1, \quad \alpha = 1, \quad \boldsymbol{\beta} = (\sqrt{2}/2, \sqrt{2}/2)^T,$$

and the exact solution

$$y(x,y) = \eta(x)\eta(y), \quad u(x,y) = \eta(1-x)\eta(1-y), \tag{5.1}$$

where

$$\eta(z) = z^3 - \frac{e^{z-1} - e^{-1}}{1 - e^{-1}}.$$

In Figures 5.1 we report the convergence rates with the SIPG solution for the state and control. As expected, the convergence rates are optimal. Recall that the SIPG method is commutative and both strategies *optimize-then-discretize* and *discretize-then-optimize* coincide.

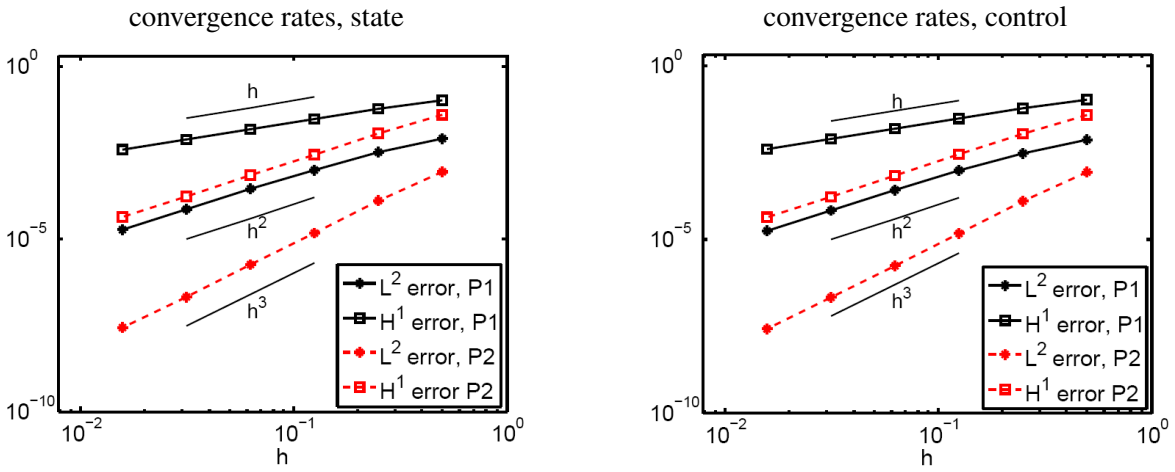


Figure 5.1: Results for Example 1. The left and the right plots show the convergence rates of the computed state and control, respectively, using the SIPG method with piecewise linear (P1) and piecewise quadratic (P2) elements on a uniform mesh.

In Figures 5.2 we report the convergence rates with the NIPG solution for the state and control for *optimize-then-discretize* strategy. Since the NIPG method is not adjointly consistent, similarly to a single equation, the convergence rates in L^2 norm for piecewise quadratic elements are suboptimal.

In Figures 5.3 we report the convergence rates with the NIPG solution for the state and the control for *discretize-then-optimize* strategy. Since the NIPG method is not commutative and as a result inconsistent for the adjoint equation the computed control fails to converge in H^1 norm for both piecewise linear and piecewise quadratic elements. As was expected from Section 4.4 we observe a first order convergence in L^2 norm.

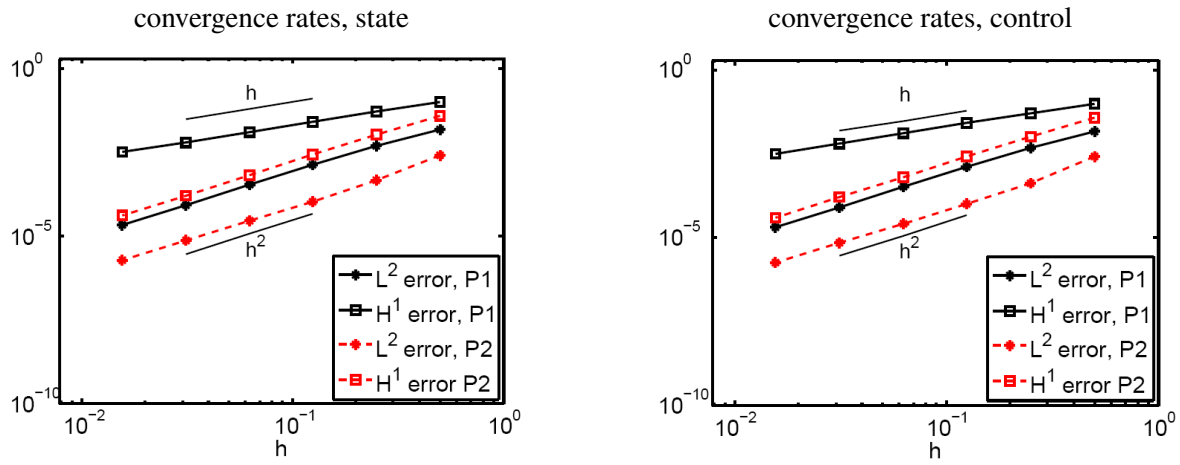


Figure 5.2: Results for Example 1. The left and the right plots show the convergence rates of the computed state and control, respectively, using optimize-then-discretize method with the NIPG piecewise linear (P1) and piecewise quadratic (P2) elements on a uniform mesh.

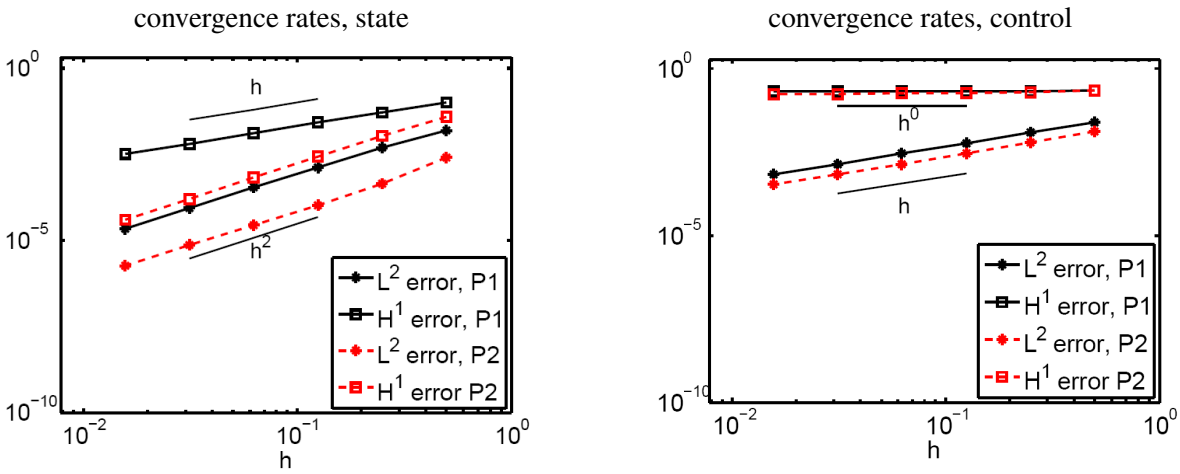


Figure 5.3: Results for Example 1. The left and the right plots show the convergence rates of the computed state and control, respectively, using discretize-then-optimize method with the NIPG piecewise linear (P1) and piecewise quadratic (P2) elements on a uniform mesh.

5.2 Example 2

In the second example we want to show that the quality of the solution may also be affected by the choice of the approach. We illustrate this by considering a problem (1.1)-(1.2) that has mild interior and boundary layers. We select

$$\varepsilon = 10^{-2}, \quad \alpha = 1, \quad \boldsymbol{\beta} = (\cos\theta, \sin\theta)^T, \ \theta = \pi/4, \quad f \equiv 0, \quad \hat{y} \equiv 1.$$

The boundary conditions for the state equation are displayed in Figure 5.4. The exact solution for this

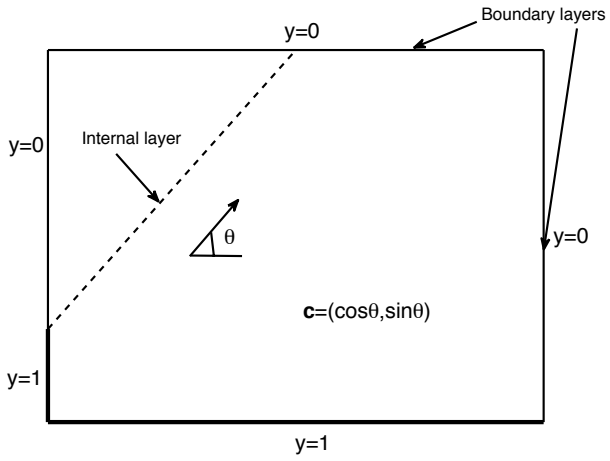


Figure 5.4: Problem set up for the state.

problem is not known. For small ε the state has interior layer along the line $y = 0.2 + x \tan\theta$ and boundary layers along the lines $y = 1$ and $x = 1$. In Figures 5.5, 5.6, and 5.7 we plot the SIPG and NIPG solutions with *optimize-then-discritize* strategy, and the NIPG solution with *discretize-then-optimize* approach for the state and control, respectively. One can see that SIPG solution is superior to the other two and essentially smooth. The NIPG *optimize-then-discritize* solution looks smooth, but has larger oscillations along the boundary layer. Finally, the NIPG *discretize-then-optimize* solution looks bad. The state has even larger oscillations at the boundary layer and the computed control looks very discontinuous. This kind of behavior for adjointly inconsistent methods was observed in [20, 28].

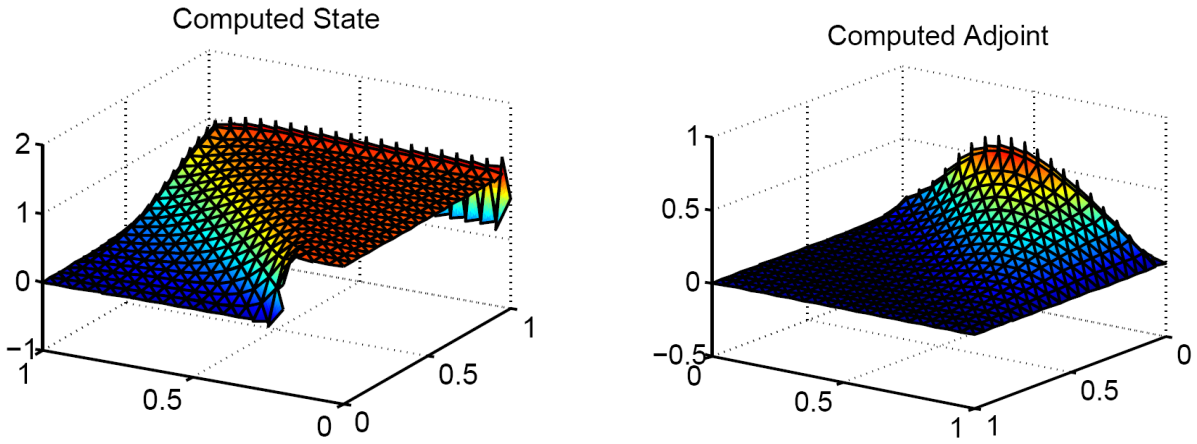


Figure 5.5: Numerical results for Example 2. The left and the right plots show the computed state and control, respectively, with the SIPG piecewise linear elements on a uniform mesh with 800 elements.

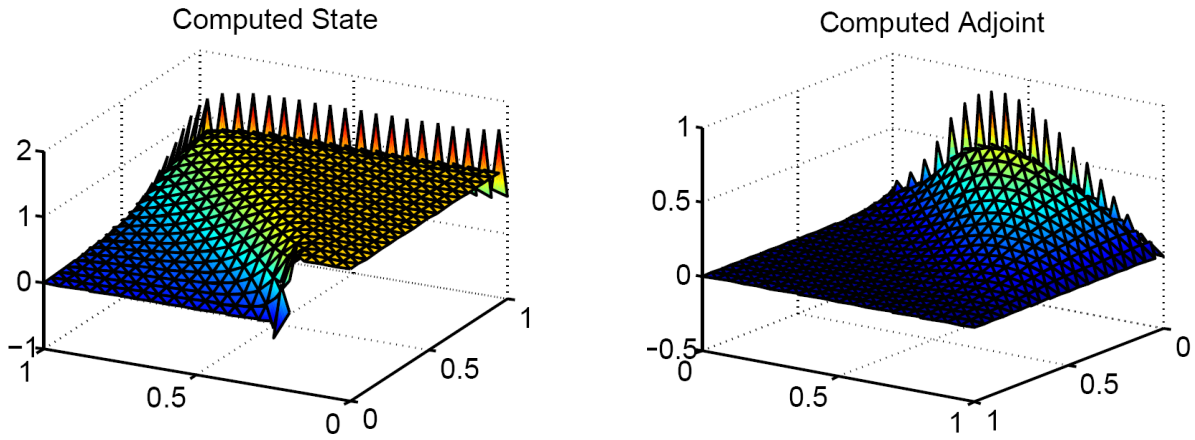


Figure 5.6: Numerical results for Example 2. The left and the right plots show the computed state and control, respectively, using optimize-then-discretize method with the NIPG piecewise linear elements on a uniform mesh with 800 elements.

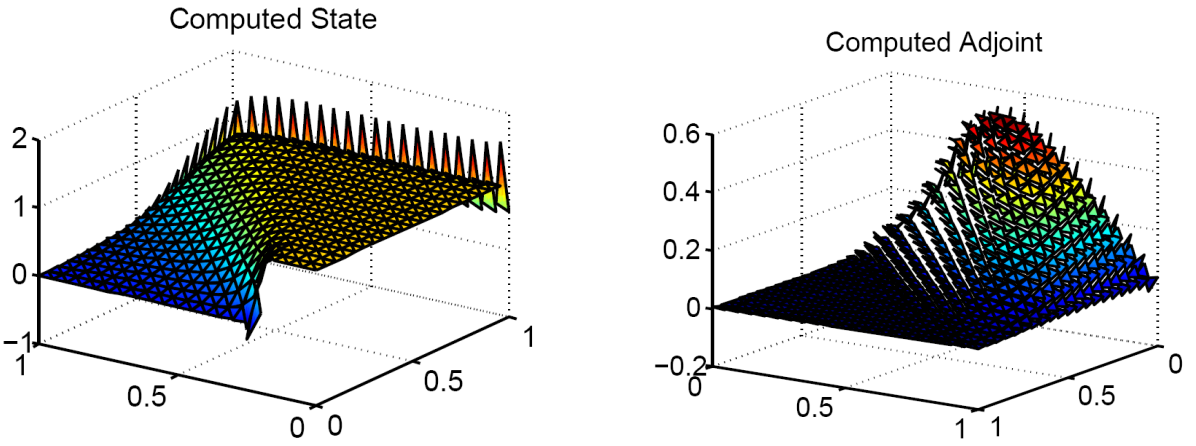


Figure 5.7: Numerical results for Example 2. The left and the right plots show the computed state and control, respectively, using discretize-then-optimize method with the NIPG piecewise linear elements on a uniform mesh with 800 elements.

6 Summary

In this paper we looked at the DG methods applied to a model optimal control problem governed by advection-diffusion equation. We derived the necessary symmetry conditions for a large class of DG methods both in primary and mixed forms and classified the most common ones. For the SIPG method we obtained optimal error estimates in the energy and the L^2-norm. However, the non-symmetric DG methods require extra care. For our simple model problem the analysis and the numerical experiments show that for the non-symmetric methods the *optimize-then-discretize* approach is preferable over the *discretize-then-*

optimize approach. For nonlinear problems the situation is less clear and needs to be further investigated.

Acknowledgements

We would like to thank the anonymous referee for suggestions that help improve the quality of the paper.

References

[1] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[2] B. AYUSO AND L. D. MARINI, *Discontinuous Galerkin methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 47 (2009), pp. 1391–1420.

[3] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations*, Journal of Computational Physics, 131 (1997), pp. 267–279.

[4] C. E. BAUMANN AND J. T. ODEN, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.

[5] R. BECKER AND B. VEXLER, *Optimal control of the convection-diffusion equation using stabilized finite element methods*, Numer. Math., 106 (2007), pp. 349–367.

[6] P. B. BOCHEV AND M. D. GUNZBURGER, *Least-squares finite element methods*, vol. 166 of Applied Mathematical Sciences, Springer, New York, 2009.

[7] M. BRAACK, *Optimal control in fluid mechanics by finite elements with symmetric stabilization*, SIAM J. Control Optim., 48 (2009), pp. 672–687.

[8] F. BREZZI, B. COCKBURN, L. D. MARINI, AND E. SÜLI, *Stabilization mechanisms in discontinuous Galerkin finite element methods*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3293–3310.

[9] F. BREZZI, L. D. MARINI, AND E. SÜLI, *Discontinuous Galerkin methods for first-order hyperbolic problems*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1893–1903.

[10] 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.

[11] P. CASTILLO, B. COCKBURN, I. PERGUGIA, AND D. SCHÖTZAU, *An a priori error analysis of the local discontinuous Galkerin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[12] B. COCKBURN, *Discontinuous Galerkin methods*, ZAMM Z. Angew. Math. Mech., 83 (2003), pp. 731–754.

[13] B. COCKBURN, B. DONG, AND J. GUZMÁN, *Optimal convergence of the original DG method for the transport-reaction equation on special meshes*, SIAM J. Numer. Anal., 46 (2008), pp. 1250–1265.

[14] B. COCKBURN AND C.-W. SHU, *The local discontinuous Galkerin method for time-dipendent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[15] S. S. COLLIS AND M. HEINKENSCHLOSS, *Analysis of the streamline upwind/petrov galerkin method applied to the solution of optimal control problems*, Tech. Report TR02–01, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005–1892, 2002. http://www.caam.rice.edu/∼heinken.

[16] C. DAWSON, S. SUN, AND M. F. WHEELER, *Compatible algorithms for coupled flow and transport*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 2565–2580.

[17] J. DOUGLAS, JR. AND T. DUPONT, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing methods in applied sciences (Second Internat. Sympos., Versailles, 1975), Springer, Berlin, 1976, pp. 207–216. Lecture Notes in Phys., Vol. 58.

[18] J. GOPALAKRISHNAN AND G. KANSCHAT, *A multilevel discontinuous Galerkin method*, Numer. Math., 95 (2003), pp. 527–550.

[19] J. GUZMÁN, *Local analysis of discontinuous Galerkin methods applied to singularly perturbed problems*, J. Numer. Math., 14 (2006), pp. 41–56.

[20] K. HARRIMAN, D. GAVAGHAN, AND E. SÜLI, *The importance of adjoint consistency in the approximation of linear functional using the discontinuous galerkin finite element method*, Tech. Report NA-04-18, University of Oxford, The Mathematical Institute, 2004. http://eprints.maths.ox.ac.uk/1172/.

[21] R. HARTMANN, *Adjoint consistency analysis of discontinuous Galerkin discretizations*, SIAM J. Numer. Anal., 45 (2007), pp. 2671–2696.

[22] M. HEINKENSCHLOSS AND D. LEYKEKHMAN, *Local error estimates for SUPG solutions of advection-dominated elliptic linear-quadratic optimal control problems*, SIAM J. Numer. Anal., 47 (2010), pp. 4607–4638.

[23] P. HOUSTON, C. SCHWAB, AND E. SÜLI, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163 (electronic).

[24] C. JOHNSON AND J. PITKÄRANTA, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[25] D. LEYKEKHMAN AND M. HEINKENSCHLOSS, *Local error analysis of discontinuous galerkin methods for advection-dominated elliptic linear-quadratic optimal control problems*, submitted, (2011).

[26] J.-L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer Verlag, Berlin, Heidelberg, New York, 1971.

[27] A. MASUD AND T. J. R. HUGHES, *A stabilized mixed finite element method for Darcy flow*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4341–4370.

[28] T. A. OLIVER AND D. L. DARMOFAL, *Analysis of dual consistency for discontinuous Galerkin discretizations of source terms*, SIAM J. Numer. Anal., 47 (2009), pp. 3507–3525.

[29] B. RIVIÈRE, *Discontinuous Galerkin methods for solving elliptic and parabolic equations*, vol. 35 of Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and implementation.

[30] B. RIVIERE, M. F. WHEELER, AND V. GIRAULT, *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems*, Computational Geosciences, 8 (1999), pp. 231–244.

[31] P. ZUNINO, *Discontinuous Galerkin methods based on weighted interior penalties for second order PDEs with non-smooth coefficients*, J. Sci. Comput., 38 (2009), pp. 99–126.