Adaptive goal–oriented error control for stabilized approximations of convection–dominated problems

Von der Fakultät für Maschinenbau
der Helmut–Schmidt–Universität/Universität der Bundeswehr
Hamburg
zur Erlangung des akademischen Grades einer Doktor–Ingenieurin
genehmigte

DISSERTATION

vorgelegt von

Kristina Schwegler
aus Nürnberg

Hamburg 2014

Erstgutachter:	Prof. Dr. Markus Bause
Zweitgutachter:	Prof. Dr. Florin Radu
Tag der mündlichen Prüfung:	09.09.2014

Acknowledgment

First, I would like to express my very great appreciation to my supervisor Markus Bause for placing the topic of this thesis. He supported me throughout my work with his knowledge and useful comments. I wish to thank Florin Radu for reviewing my thesis.

My special thanks are extended to my current and former colleagues for the friendly atmosphere. In particular, I would like to mention Wolfgang Zeuge who arranged such a great work environment that I was able to focus on my research. He is an excellent discussion partner regarding professional as well as personal subjects. Bastian Ebeling provided me computer support. By the way, he became a close friend at occasionally demanding times. Sharing the room with Uwe Köcher created a relaxed and funny working atmosphere. I wish to thank everyone who made a contribution to complete my thesis.

I am particularly grateful for the assistance in proof reading given by Markus Bause.

Last but not least I would like to thank my family for their ongoing support and especially Rodolphe for his love and everything else he gave to me.

Abstract

The approximate solution of convection–dominated transport problems obtained by standard finite element methods is characterized by unphysical oscillations. In general, there are two strategies in order to reduce this undesirable deviation from the exact solution. The first one is the calculation of the numerical solution on a highly refined grid, the second one is the use of stabilization techniques.

The drawback of a global refinement strategy is given by the increasing number of degrees of freedom. For that reason, it is common to use adaptive refinement strategies based on a–posteriori error estimates. The conventional error estimates for convection–diffusion–reaction equations often depend on undetermined constants or even the reciprocal of the small diffusion coefficient which leads to tremendous impracticality of these error bounds. However, stabilization techniques on their own are also not sufficient to completely reduce the oscillations. For that reason, we take advantage of the concept of adaptivity and of stabilization by combining a goal–oriented error representation with stabilization techniques of streamline upwind type (SUPG) and shock–capturing.

We present an error representation that is based on the dual weighted residual method that transfers information how to weight the residual terms with the solution of an associated adjoint linear problem which is convection–dominated as well. The error is given in terms of a user chosen quantity of interest such that point values or other application–related target quantities. The performance of the error representation which is exact except for negligible remainder terms and the corresponding adaptive strategy is presented for stationary nonlinear transport problems and different quantities of interest. The introduced method can be extended to transient convection–diffusion–reaction problems provided that a variational formulation in space as well as in time is available.

The numerical results for stationary and nonstationary test cases illustrate that the presented goal–oriented error representation is capable to control the adaptive refinement process in such a manner that unphysical effects are avoided or,

at least, clearly reduced in comparison to standard methods. The effectivity index, a measure for the quality of error estimates, tends to an optimal value with regard to reliability and precision. For that reason, the mesh adaption is carried out in an excellent way such that boundary layers, interior layers or sharp moving fronts are detected and the quantitative error in terms of the target functional decreases.

Zusammenfassung

Bei der näherungsweisen Berechnung von Lösungen von konvektionsdominanten Transportproblemen entstehen üblicherweise unphysikalische Oszillationen in der Lösung, wenn man diese mit dem Standardwerkzeug der Finite–Elemente–Methode behandelt. Im Allgemeinen gibt es zwei Herangehensweisen, um der unerwünschten Abweichung der numerischen Lösung von der exakten Lösung entgegen zu wirken. Zum einen können feine Netzschrittweiten Abhilfe schaffen, zum anderen werden Stabilisierungsverfahren eingesetzt.

Ein globales Verfeinern der Auflösung des Rechengitters führt allerdings zu einer wachsenden Anzahl von Freiheitsgraden im zu lösenden System. Deswegen verwendet man gewöhnlich adaptive Verfeinerungsstrategien auf der Grundlage von a–posteriori Fehlerschätzern. Üblicherweise hängen Fehlerschätzer für Konvektions–Diffusion–Reaktionsgleichungen von unbestimmten Konstanten oder sogar vom Kehrwert des sehr kleinen Diffusionskoeffizienten ab, was zu absoluter Unbrauchbarkeit der Fehlerschranken führt. Allerdings lösen auch Stabilisierungstechniken das Problem der Oszillationen nicht alleine. Darin liegt der Grund für unser Vorgehen, Stabilisierung vom Typ streamline upwind (SUPG) und shock–capturing mit Adaptivität basierend auf einer zielorientierten Fehlerdarstellung zu verbinden.

Dazu leiten wir mit Hilfe der dual–gewichteten Residuen–Methode (dual weighted residual method) eine Fehlerdarstellung her, die unter Berücksichtigung der Lösung eines ebenfalls konvektionsdominanten linearen dualen Problems die residualen Terme quantitativ gewichtet. Der Fehler wird hierbei bezüglich einer Messgröße dargestellt, die der Anwender nahezu beliebig wählen kann. Darunter kann man sich Punktgrößen, aber auch jedes andere Funktional, das für die Aufgabenstellung interessant ist, vorstellen. An verschiedenen Testbeispielen für stationäre nichtlineare Transportprobleme und einer Reihe von Zielgrößen zeigen wir den Nutzen, den die bis auf vernachlässigbare Restglieder exakte Fehlerdarstellung bringt, auf. Insofern wir für instationäre Konvektions–Diffusions–Reaktionsgleichungen eine variationelle Formulierung im Raum und auch in der Zeit vorliegen haben, kann die vorgestellte Methode

auf instationäre Problemstellungen erweitert werden.

An den numerischen Ergebnissen für den stationären wie für den instationären Fall veranschaulichen wir, dass die zielorientierte Fehlerdarstellung die adaptive Verfeinerungsstrategie so steuern kann, dass unphysikalische Effekte vermieden oder zumindest deutlich reduziert werden im Vergleich zu Standard–Verfahren. Der Effektivitätsindex stellt ein Maß für die Qualität eines Fehlerschätzers dar. In unseren Simulationen können wir feststellen, dass der Effektivitätsindex gegen einen optimalen Wert hinsichtlich Zuverlässigkeit und Exaktheit konvergiert. Darauf begründet ist die Fehlerdarstellung in der Lage, die Gitteradaptivität so zu steuern, dass Grenzschichten aufgespürt werden und der Fehler bezüglich des Zielfunktionals abnimmt.

Contents

1 Introduction **1**

1.1 Starting point and goal 1

1.2 An introduction to the dual weighted residual (DWR) method . 8

2 Notation & definitions **11**

2.1 Function spaces 11

2.2 Function spaces involving time 13

2.3 Aspects of functional analysis and finite element spaces 13

3 A SUPG stabilized dual weighted residual method **21**

3.1 A general framework 21

3.2 A $\mathcal{FSTD}$ method 23

3.3 A $\mathcal{FDTS}$ method 34

3.4 Both methods by numerical comparison 45

4 A SUPG and SOLD stabilized dual weighted residual method **53**

4.1 A nonlinear framework 53

4.2 A nonlinear $\mathcal{FDTS}$ method 55

4.3 Numerical studies for the nonlinear $\mathcal{FDTS}$ method 62

5 A nonstationary stabilized dual weighted residual method **81**

5.1 A nonstationary framework 82

5.2 A time–dependent $\mathcal{FDTS}$ method 86

5.3 Numerical studies for the time–dependent $\mathcal{FDTS}$ method . . . 95

6 Conclusions & outlook **109**

6.1 Summary & conclusions 109

6.2 Outlook 110

Chapter 1

Introduction

1.1 Starting point and goal

Research scientists in the scope of natural sciences and engineering have to address the often challenging task to ensure the reliability of their investigated models and methods. Flow dynamic phenomena are modeled by a coupled system of partial differential equations. As a prototype model for coupled flow and transport phenomena, we consider the dimensionless incompressible Navier–Stokes equations coupled with a system of n transport equations

$$\begin{aligned} \partial_t \boldsymbol{v} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} - \frac{1}{\mathrm{Re}} \Delta \boldsymbol{v} + \nabla p &= \boldsymbol{f}(t, \boldsymbol{u}) \,, \\ \nabla \cdot \boldsymbol{v} &= 0 \,, \\ \partial_t u_i + \boldsymbol{v} \cdot \nabla u_i - \nabla \cdot \left(\frac{1}{\mathrm{Re} \cdot c_i} \nabla u_i \right) &= g_i(t, \boldsymbol{u}) \,, \quad i = 1, \ldots, n \,. \end{aligned} \tag{1.1}$$

The first vector–valued equation is called the momentum equation which describes the fluid's momentum transport. Further, $\boldsymbol{v}$ denotes the fluid's velocity, p its pressure and Re is the Reynold's number defined by $\mathrm{Re} := \frac{\rho v_c l_c}{\eta}$ where ρ is the fluid's density, v_c and l_c are the characteristic velocity and length, respectively, and η denotes the dynamic viscosity. Moreover, f is an arbitrary right–hand side vector, for example electric body forces dependent on an ion concentration in the fluid or buoyant forces dependent on temperature differences in the fluid. The volume conservation law is given by the second equation. The last equation prescribes the motion and dissipation of a scalar arbitrary quantity u_i, e.g. a species' concentration or temperature. Further, c_i

is a model dependent constant which is usually larger than one, e.g. Schmidt number Sc or Prandtl number Pr. The functions g_i describe the right–hand side including reaction between the quantities u_i and u_j, for instance. Then, this equation describes convection–diffusion–reaction phenomena. If the Reynold's number is large which typically occurs in real–world applications, the Navier–Stokes equation consisting of the momentum equation and the volume conservation law is convection–dominated which may result in a turbulent flow field $\boldsymbol{v}$. Since c_i is usually in the magnitude of 10, the transport equation is even more convection–dominated in this case. In the given situation, the numerical solution is characterized by steep interior and boundary layers that standard discretization schemes are not capable to resolve. In the sequel, unphysical oscillations develop that will lead to wrong predictions. Since the convection–diffusion-reaction equation is more convection–dominated than the Navier–Stokes equation, it is important to consider carefully techniques that reduce the unphysical oscillations, especially for the transport problem.

There is a nearly endless number of applications that are modeled by the coupled system given above. We will discuss some of them. The model problem (1.1) is used for the design process of an air conditioning system. Especially car and plane manufacturer are in an endeavor to provide draft–free ventilation that starts up fast. It is very convenient to replace experiments with simulations as no prototype has to be built. In this case, we use the system (1.1) where $n = 1$ and u is the temperature.

Another area of application is the simulation of highly turbulent flows where it is not possible to resolve the smallest turbulent structures of the flow field. In this case, turbulence models are used to incorporate these phenomena. A common turbulence model is the two equation $k - \varepsilon$ model presented in [31]. Two additional transport equations are added to the Navier–Stokes equations to represent the turbulent properties of the flow. The first property is given by the turbulent kinetic energy k, the second one by the turbulent dissipation ε. This means for our model $n = 2$, $u_1 = k$, $u_2 = \varepsilon$ and appropriate functions g_1, g_2.

Electromobility requires high capacity batteries and capacitors. The transport of charge carriers inside these energy reservoirs is described by convection–dominated transport problems with nonlinear reaction terms. The velocity field is influenced by the motion of the charge carriers while moving from the anode to the cathode which in turn influences the motion of the charge carriers. The transport equations are defined by the Nernst-Planck equation that describes the concentration of charge carriers in a fluid. It is coupled with the Poisson equation that models the electric field in the energy storage device. The Poisson equation can be seen as specific variant of a transport equation. We choose $n = 2$, $u_1 = n_c$ to be the charge carrier's concentration and $u_2 = \varphi$ to be the electric potential in (1.1).

Even our daily weather forecast is based on this model problem. Actually in base models, the propagation of temperature and humidity has to be considered. In this case, we set $n = 2$, u_1 denotes the temperature and u_2 denotes the humidity. For special purposes, these forecast models can be extended by additional transport equations, for example after the volcanic eruption of the Eyjafjallajökull in Iceland, the climatologists were able to simulate the ash cloud propagation by means of a transport equation of the ash concentration in the atmosphere.

Another common area of application is the reactive transport which is involved in engine combustion, in drug delivery and reception, in chemical reactors and in fuel cells, for example. These applications have in common that they use two or more chemical species dissolved in a fluid which interact. If these reactions are highly endothermic or exothermic, the temperature and its propagation additionally have to be considered. In this case, $u_1, \ldots, u_{n-1}$ are the involved species' concentration and u_n optionally denotes the temperature.

Even though more complex or simplified models are used for the simulation of the flow field than given in (1.1), the transport equations are continued to be applied. The free surface Navier–Stokes equations are involved to simulate more complex flow phenomena, for example the heat distribution in a liquid tank under zero gravity conditions. At this, the temperature is simulated by

a first transport equation whereas the location of the fluid is given by a levelset function which is also propagated by a transport equation. Examples for a simplified flow model include ground water pollution and geothermics. In this connection, the Navier–Stokes equations are replaced by Darcy's problem. The pollutant's concentration is still described by a transport equation as well as the temperature for the geothermal application.

All applications have in common that they involve transport problems that contribute to a more complex physical model system. For the numerical treatment, the transport equation can be considered as a simplified prototype for the Navier–Stokes equations. For that reason, the convection–diffusion–reaction equation acts as an important prototype that is used to develop and evaluate numerical solution methods.

Most of the previous examples develop static or dynamic layers. A boundary layer is a harsh change of the value of the function u at a specific location. It is called static if the location does not change in the course of the time and dynamic otherwise. As mentioned earlier, these boundary layers provoke oscillations which will be reduced if the spatial discretization at the boundary layers' location is refined. For some static boundary layers, its position is known from scratch. For that reason, the discretization can be adapted accordingly in this case. If the position is not known from scratch or dynamic, the discretization has to be matched dynamically to the structure of the solution. This concerns current research topics such that the computational time is reduced and the quality of the solution is increased.

In the last decades, great efforts have been made to deal with the oscillatory phenomena due to the convection–dominated property of transport problems. [41] gives an overview of stabilization methods that are applied in order to reduce the undesirable oscillations. These introduced methods are of upwind type such as the variational multiscale method (VMS), the local projection stabilization (LPS), discontinuous Galerkin methods (dGFEM) or the streamline upwind Petrov–Galerkin (SUPG) method to name but a few of them. The SUPG method will be considered in this work. In a competitive study

in [4], this method is assessed as appropriate if sharpness and position of layers are important whereas remaining spurious oscillations can be tolerated. With respect to efficiency, the SUPG method outperforms many other considered approaches that are assessed in [4]. Loosely speaking, the SUPG method adds residual based stabilization terms in streamline direction. But remaining oscillations are not acceptable in view of wrong predictions. Thus, another stabilization term in crosswind direction is used; this kind of method is called shock–capturing, or to be more precise spurious oscillations at layers diminishing (SOLD) methods. The numerical results obtained by the SUPG method in combination with shock–capturing are also not perfect but another step in the right direction.

Nowadays, adaptive mesh refinement based on an a–posteriori error estimator is a standard tool in finite element codes. There exist error estimates for most every partial differential equation. Most of the error estimators are based on the $\mathcal{L}^2$ norm or the natural norm with respect to the applied discretization strategy which is not useful in the context of convection–dominated transport problems since this kind of error estimators is not able to detect the spurious oscillations. Another drawback of conventional error estimates can be seen in the dependence of often unknown constants. As a consequence, the error estimator does not really obtain information concerning the quantity of the error. Nevertheless, the error estimates usually are suited as basis of adaptive refinement techniques. With relation to convection–dominated transport problems, error estimates typically depend on the reciprocal of the small diffusive coefficient which leads to useless error bounds. In the literature, there exist only a few papers that address the design of error estimates with respect to dominating convection. In [25], a numerical study of a–posteriori error estimators for convection–diffusion equations is presented. It includes known error estimators starting with a simple gradient indicator to residual based error estimators in the $\mathcal{H}^1$ semi norm and energy norm through to an error estimator based on the solution of stabilized local Neumann problems. The result of this study is that there is no robust error estimator that works satisfactorily in all test cases as well as mesh adaption is not performed in an adequate way. In [18], an error estimate is derived without undetermined constants that is based on us-

ing $\boldsymbol{H}(\mathrm{div}, \Omega)$–conforming diffusive and convective flux reconstructions. Tests are performed with a diffusion coefficient of 10^{-2} and 10^{-4}. In these cases, the effectivity index achieves satisfactorily results. The mentioned method has not been compared to other error estimators and tested with smaller values of diffusion. Robust a–posteriori error estimates for stationary convection–diffusion equations are presented in [47]. They are based on local residuals or the solution of discrete local problems. Quite far away from our strategy for example in [20], there are attempts to use an improved unusual stabilized finite element method combined with an error estimator of the type just described and proposed in [48]. In [43], an almost–robust residual based a–posteriori error estimator in a special norm is proposed. The author accounts this error estimator to perform better than other proposed error estimators. Robust a–posteriori error estimators for the $\mathcal{L}^1(\Omega)$ and $\mathcal{L}^2(\Omega)$ norm are analyzed in [21] and [22]. At that, the variational multiscale theory is basically used. A recent work on the design of an robust a–posteriori error estimator for stationary convection–diffusion equations is presented in [29]. The error estimator for the SUPG finite element approximation is constructed in a norm typically used in the a–priori analysis of this method.

In addition to error estimators, there are further efforts to investigate methods that are able to damp spurious oscillations. In [46], an efficient preconditioning technique using a matrix reordering scheme is applied to solve the sparse linear systems arising from the nonlinear convection–dominated transport problem. Another approach in order to stabilize convection–dominated problems is based on algebraic flux–corrected transport (FCT) algorithms; cf. [34]. In [11], a modified Brezzi–Douglas–Marini (BDM_1) mixed finite element method is used for the solution of a system of convection–diffusion–reaction equations. Numerical simulations show that this nonstandard finite element method is of second order and is robust against small diffusion coefficients.

As mentioned above, the $\mathcal{L}^2$ norm is quite unsuitable as basis for error estimators because spurious oscillations of the numerical solution contribute little to the error in this norm. With regard to applications in the scope of engineering, one is not interested in a global error estimate but, e.g. in the drag coefficient

in the context of simulation of flows. This aim can be achieved by the concept of goal–oriented error estimates. The usual $\mathcal{L}^2$ error estimates fundamentally differ in the representation of the error in terms of an almost arbitrary user chosen quantity of interest. Point values of the solution as well as boundary integrals are possible quantities of interest in applications. In order to derive a goal–oriented error estimate, the dual weighted residual method proposed in [8] is often applied.

This work is organized as follows. The following introductory section of Chapter 1 presents the main principles of the dual weighted residual method. In Chapter 2, we introduce some necessary theory of partial differential equations as well as some global assumptions.

Chapter 3 addresses the derivation of a goal–oriented error representation by means of the dual weighted residual method for a linear transport problem. At that point, there are two approaches to achieve this aim with respect to the sequence of stabilizing the original problem with the SUPG method and taking the associated dual problem. We will compare the numerical results of both strategies that the adaptive refinement is based on.

According to the results in Chapter 3, we will follow the strategy to first take the associated dual problem and then separately stabilize both problems. In Chapter 4, we carry the introduced error representation over to a nonlinear convection–diffusion–reaction model problem. Since the original problem is nonlinear by itself, we use a further nonlinear stabilization technique called shock–capturing. The proposed error representation in terms of a user chosen target quantity is extensively tested by simulating different examples in combination with a variety of quantities of interest.

Based on a variational formulation in space as well as in time, we extend our stabilized goal–oriented error representation to nonstationary model problems in Chapter 5. The numerical results emphasize the benefit of the proposed method.

Finally, we top this work off with some conclusions concerning the applied method and some future thoughts.

1.2 An introduction to the dual weighted residual (DWR) method

This section illustrates the basic idea of the dual weighted residual (DWR) method. In order to keep this introduction simple, we will skip some mathematical details about weak solution theory.

We consider the weak formulation of Poisson's problem with an appropriate right–hand side f

$$A(u)(\varphi) := \langle \nabla u, \nabla \varphi \rangle_\Omega = \langle f, \varphi \rangle_\Omega =: F(\varphi)\,, \tag{1.2}$$

where $\langle \cdot, \cdot \rangle_\Omega$ denotes the usual $\mathcal{L}^2(\Omega)$ scalar product on a domain Ω. We call this problem the primal problem and u the exact primal solution of (1.2). Let u_h be a Galerkin approximate solution to u, e.g. obtained by a finite element method. Then define $e := u - u_h$ to be the error and $\rho(\varphi) := F(\varphi) - A(u_h)(\varphi)$ to be the residual of the primal problem. If φ is an element of the same space as u_h is an element of, then the residual will vanish. This property is called Galerkin orthogonality. The relation between the error e and the residual ρ is given by $A(e)(\varphi) = \rho(\varphi)$. Resulting from this identity, a solely residual based error estimator would clearly lack accuracy due to the uncertainty of the inverse operator related to the blilinear form A. For that reason, the structure of this error estimator is not useful. Additionally, we often are not only interested in the norm of the error but also in a practice–oriented functional involving the error. For that purpose, we are going to consider a goal–oriented error estimator.

In order to find a residual based error estimator which is capable to precisely quantify the error, the residual has to be weighted in an appropriate manner which is the task of the so–called dual solution of the following problem: Find z such that $A(\varphi, z) = \langle j, \varphi \rangle_\Omega =: \mathcal{J}(\varphi)$ where $j \in \mathcal{L}^2(\Omega)$ is user chosen and $\mathcal{J}$ is the corresponding quantity of interest. If we are interested in the solution

only on a certain sub–domain $\Omega_s \subset \Omega$ for example, we choose $j = \chi(\Omega_s)$, the characteristic function on Ω_s. By means of the definition of the dual problem, we get that

$$\begin{aligned} \mathcal{J}(e) = \langle j, e\rangle_\Omega = \langle j, u\rangle_\Omega - \langle j, u_h\rangle_\Omega = A(u)(z) - A(u_h)(z) \\ = F(z) - A(u_h)(z) \\ = \rho(z)\,. \end{aligned} \tag{1.3}$$

This error representation is described by the residual weighted with the dual solution. We note that the error representation involves the exact dual solution. If we were able to solve the dual problem in an exact way, we also could exactly solve the primal problem. As we will see later, it is sufficient to generate a higher order approximation to the dual solution instead of using the exact dual solution. Therefore, the method is feasible without loosing much quality.

In this introduction, we considered a linear partial differential equation in combination with a linear target quantity $\mathcal{J}$. To prove that the DWR technique still works in a more general situation where the problem and the quantity of interest are nonlinear, it is useful to reformulate the problem as a constraint optimization problem:

$$\mathcal{J}(u) = \min!\,, \quad A(u)(\varphi) = F(\varphi)\,. \tag{1.4}$$

We define the corresponding Lagrangian functional

$$\mathfrak{L}(u, z) := \mathcal{J}(u) + F(z) - A(u)(z)\,,$$

where the minimal solution u corresponds to the first component of a stationary point of $\mathfrak{L}(u, z)$ whereas the dual solution z corresponds to the second component. This representation is used to derive an error description for a nonlinear setting.

The approach shown above is the main idea of the DWR method that we will extend to nonlinear partial differential equations in the following sections. In this nonlinear case, the steps in (1.3) are more sophisticated.

Chapter 2

Notation & definitions

2.1 Function spaces

In view of the analytical sections of this work, we introduce some necessary theoretical framework. This includes the definition of function spaces and their associated norms. In these reflections, we follow [19]. We start with the definition of $\mathcal{L}^p$ spaces.

Definition 2.1 ($\mathcal{L}^p$ spaces, see *D.1. Banach spaces.* in [19])**.** *Let Ω be an open subset of $\mathbb{R}^n$ and $1 \leq p \leq \infty$. If $f : \Omega \to \mathbb{R}$ is measurable, we define*

$$\|f\|_{\mathcal{L}^p(\Omega)} := \begin{cases} \left(\int\limits_{\Omega} |f|^p \, d\boldsymbol{x} \right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty \\ \operatorname{ess\,sup}_{\Omega} |f| & \text{if } p = \infty \,. \end{cases}$$

We define $\mathcal{L}^p(\Omega)$ to be the linear space of all measurable functions $f : \Omega \to \mathbb{R}$ satisfying $\|f\|_{\mathcal{L}^p(\Omega)} < \infty$. Then $\mathcal{L}^p(\Omega)$ is a Banach space, provided we identify two functions which agree almost everywhere.

For details about the concept of measurability, we refer to [19, *E.2.*]. The space $\mathcal{L}^2(\Omega)$ is a Hilbert space endowed with an inner product

$$\langle f, g \rangle_{\Omega} := \int\limits_{\Omega} f g \, d\boldsymbol{x} \,,$$

which generates the norm

$$\|u\|_{\mathcal{L}^2(\Omega)} := \langle u, u \rangle_{\Omega}^{\frac{1}{2}} \,.$$

We next address the definition of Sobolev spaces. Therefor, we fix $1 \leq p \leq \infty$. Let k be a nonnegative integer. We define now certain function spaces whose members have weak derivatives of various orders lying in various $\mathcal{L}^p$ spaces.

Definition 2.2 (Sobolev spaces, see 5.2.2. *Definition of Sobolev spaces.* in [19])**.** *The Sobolev space*

$$\mathcal{W}^{k,p}(\Omega)$$

consists of all locally summable functions $u : \Omega \to \mathbb{R}$ such that for each multi-index α with $|\alpha| \leq k$, $\mathrm{D}^\alpha u$ exists in the weak sense and belongs to $\mathcal{L}^p(\Omega)$.

Remark 2.3. *If $p = 2$, we usually write*

$$\mathcal{H}^k(\Omega) = \mathcal{W}^{k,2}(\Omega) \quad (k = 0, 1, \ldots)\,.$$

$\mathcal{H}^k(\Omega)$ is a Hilbert space.

Definition 2.4 (Norm in Sobolev spaces, see *5.2.2. Definition of Sobolev spaces.* in [19])**.** *If $u \in \mathcal{W}^{k,p}(\Omega)$, we define its norm to be*

$$\|u\|_{\mathcal{W}^{k,p}(\Omega)} := \begin{cases} \left(\sum\limits_{|\alpha| \leq k} \int\limits_\Omega |\mathrm{D}^\alpha u|^p \, d\boldsymbol{x} \right)^{\frac{1}{p}} & (1 \leq p < \infty) \\ \sum\limits_{|\alpha| \leq k} \operatorname{ess\,sup}_\Omega |\mathrm{D}^\alpha u| & (p = \infty)\,. \end{cases}$$

Definition 2.5 (The space $\mathcal{W}_0^{k,p}(\Omega)$, see *5.2.2. Definition of Sobolev spaces.* in [19])**.** *We denote by*

$$\mathcal{W}_0^{k,p}(\Omega)$$

the closure of $\mathcal{C}_c^\infty(\Omega)$ in $\mathcal{W}^{k,p}(\Omega)$ where $\mathcal{C}_c^\infty(\Omega)$ denotes the space of infinitely differentiable functions $\phi : \Omega \to \mathbb{R}$ with compact support in Ω.

It is customary to write

$$\mathcal{H}_0^k(\Omega) = \mathcal{W}_0^{k,2}(\Omega)\,.$$

It is possible to characterize the space $\mathcal{H}_0^1(\Omega)$ by means of the trace operator; cf. [3] and Theorem 2.10. We will use the abbreviation $\mathcal{V} := \mathcal{H}_0^1(\Omega)$ together with its dual space $\mathcal{V}^*$.

Concerning the applied stabilization techniques in this work, we use a stabilized natural norm $|||\cdot|||$ which is given by

$$|||v||| := \left(\sum_{K \in \mathcal{T}_h} \|\sqrt{\varepsilon}\nabla v\|^2_{\boldsymbol{\mathcal{L}}^2(K)} + \|\sqrt{\alpha} v\|^2_{\mathcal{L}^2(K)} + \delta_K \|\boldsymbol{b} \cdot \nabla v\|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}}, \qquad (2.1)$$

where the parameters $\varepsilon\,, \alpha\,, \boldsymbol{b}$ and δ_K are specified in greater detail later.

2.2 Function spaces involving time

Let $\mathcal{X}$ denote a real Banach space, with norm $\| \|$.

Definition 2.6 ($\mathcal{L}^p$ spaces in time taking values in $\mathcal{X}$, see *5.9.2. Spaces involving time.* in [19])**.** *The space*

$$\mathcal{L}^p((0,T);\mathcal{X})$$

consists of all strongly measurable functions $u : [0,T] \to \mathcal{X}$ *with*

$$\|u\|_{\mathcal{L}^p((0,T);\mathcal{X})} := \left(\int_0^T \|u(t)\|^p \, dt \right)^{\frac{1}{p}} < \infty$$

for $1 \leq p < \infty$ *and*

$$\|u\|_{\mathcal{L}^\infty((0,T);\mathcal{X})} := \operatorname*{ess\,sup}_{0 \leq t \leq T} \|u(t)\| < \infty \, .$$

The space–time function space is defined by

$$\mathfrak{V} := \left\{ v \in \mathcal{L}^2\left((0,T);\mathcal{V}\right) \,|\, \partial_t v \in \mathcal{L}^2\left((0,T);\mathcal{V}^*\right) \right\} \, .$$

We note that $\mathfrak{V}$ is continuously embedded in $\mathcal{C}([0,T];\mathcal{L}^2(\Omega))$ see [19, *5.9.2. Spaces involving time. Theorem 3*].
We introduce the space–time scalar product

$$\langle\langle v, w \rangle\rangle := \int_0^T \langle v, w \rangle_\Omega \, dt \, . \tag{2.2}$$

2.3 Aspects of functional analysis and finite element spaces

For further details about the reflections on the *Fréchet derivative*, compare [49].

Definition 2.7 (Fréchet derivative)**.** *Assume* $f : U(u) \subseteq X \to Y$ *to be a given operator defined on an open neighborhood of the point* u*, where* X *and* Y *are Banach spaces over* $\mathbb{R}$*.*

a) The differential $df(u)$ of f at the point u exists iff there is a linear bounded operator denoted by

$$df(u) : X \to Y\,,$$

such that

$$f(u+h) - f(u) = df(u)(h) + \mathcal{O}\left(\|h\|\right)\,, \quad h \to 0\,,$$

holds for all $h \in X$ in some open neighborhood of $h = 0$ in X.
We also write $f'(u)$ instead of $df(u)$, equivalently, and we say that $f'(u)$ is the Fréchet derivative of f at the point u.

b) The second differential $d^2 f(u)$ of f at a point u exists iff there is a bilinear bounded operator denoted by

$$d^2 f(u) : X \times X \to Y\,,$$

such that

$$df(u+h)(k) - df(u)(k) = d^2 f(u)(h,k) + r\,,$$

where the "small" remainder r is defined by

$$\sup_{\|k\|\leq 1} \|r(u;h,k)\| = \mathcal{O}\left(\|h\|\right)\,, \quad h \to 0\,,$$

holds for all $k \in X$ and all h in some open neighborhood of $h = 0$ in X.
We also write $f''(u)$ instead of $df^2(u)$, equivalently, and we say that $f''(u)$ is the second Fréchet derivative of f at the point u.

Definition 2.8 (Partial Fréchet derivative)**.** *Assume that the map*

$$f : U(u,v) \subseteq X \times Y \to Z$$

is given on an open neighborhood of the point (u,v), where X, Y and Z are Banach spaces over $\mathbb{R}$.
Let v be fixed and set $g(w) := f(w,v)$. If the Fréchet derivative of g at the point u exists, then we define the partial Fréchet derivative $f_u(u,v)$ through

$$f_u(u,v) := g'(u)\,.$$

The partial Fréchet derivative $f_v(u,v)$ is defined similarly.

Proposition 2.9. *Let $f : U(u,v) \subseteq X \times Y \to Z$ be given in Definition 2.8. If f has a Fréchet derivative at the point (u,v), then the partial derivatives $f_u(u,v)\,, f_v(u,v)$ also exist and*

$$f'(u,v)(h,k) = f_u(u,v)(h) + f_v(u,v)(k) \quad \textit{for all } h \in X, k \in Y\,.$$

In the case of nonhomogeneous Dirichlet boundary conditions, we need a result about Sobolev spaces. Therefore, the *trace theorem* is quoted from [40, Theorem 1.3.1]. For more general results and the proof, we refer to [35], [36] or [1].

Theorem 2.10 (Trace theorem)**.** *Suppose Ω to be a bounded open set of $\mathbb{R}^d$ with Lipschitz continuous boundary $\partial\Omega$.*

a) There exists a unique linear continuous map $\gamma_0 : \mathcal{H}^1(\Omega) \to \mathcal{H}^{\frac{1}{2}}(\partial\Omega)$ such that

$$\gamma_0 v = v_{|\partial\Omega} \textit{ for each } v \in \mathcal{H}^1(\Omega) \cap \mathcal{C}^0(\bar{\Omega})\,.$$

b) There exists a linear continuous map $\mathcal{R}_0 : \mathcal{H}^{\frac{1}{2}}(\partial\Omega) \to \mathcal{H}^1(\Omega)$ such that

$$\gamma_0 \mathcal{R}_0 \varphi = \varphi \textit{ for each } \varphi \in \mathcal{H}^{\frac{1}{2}}(\partial\Omega)\,.$$

We make the following assumptions regarding the triangulation of the domain and the finite element space. When we talk about triangles in the following, we mean that the decomposition of $\Omega \subset \mathbb{R}^3$ into tetrahedrons is included. The following global assumptions are required throughout this work.

Assumption 2.11 (A1 Conforming triangulation and finite–dimensional space)**.** *Let $\mathcal{T}_h$ be a triangulation of domain Ω with a polygonal boundary and $\mathcal{V}_h$ a finite–dimensional space.*

a) We perform our simulations on a conforming triangulation $\mathcal{T}_h$ satisfying the following properties:

i) The computational domain Ω is divided into triangles K such that $\bar{\Omega} = \bigcup_{\bar{K} \in \mathcal{T}_h} K\,.$

ii) Arbitrary elements $K_i, K_j \in \mathcal{T}_h\,, i \neq j$ are disjoint, i.e. $K_i \cap K_j = \emptyset$, or share only a node or an edge.

b) The finite–dimensional subspace $\mathcal{V}_h \subset \mathcal{V}$ is defined by

$$\mathcal{V}_h = \mathrm{cG(p)} := \{v \in \mathcal{H}_0^1(\Omega) \cap \mathcal{C}(\bar{\Omega}) \,|\, v_{|K} \in \mathcal{P}_p(K)\ \forall K \in \mathcal{T}_h\}\,, \qquad (2.3)$$

with a triangulation $\mathcal{T}_h$ consisting of a finite number of triangles K as given in a), $\bar{\Omega}$ the closure of the domain and $\mathcal{P}_p(K)$ the usual function space of polynomials of degree at most p on K.

Remark 2.12. *According to Assumption A1, hanging nodes are not permitted. This feature is the reason why we use an adaptively defined grid consisting of triangle elements and not quadrilateral elements. If an arbitrary inner quadrilateral element is indicated to be refined, hanging nodes are added by the regular refinement strategy. Hanging nodes can be eliminated by inserting elements in the neighborhood of the hanging node in order to establish a conforming grid which leads to a higher number of unknowns than the indicator suggests. Another possibility of treating hanging nodes is to not consider the hanging nodes as unknowns. The value of the unknown that corresponds to a hanging node is generated from neighboring nodes by a suitable interpolation. So, we get less new degrees of freedom as indicated. The choice of triangles ensures that exactly the marked elements and no more than the direct neighbours are refined.*

We also need a theoretical result about the face values of functions in the Sobolev space $\mathcal{H}^1(\mathcal{T}_h)$. To this end, we present the definition of *Shape and contact regularity* and the *Continuous trace inequality*. The definition is quoted from [14, Definition 1.38], the result is quoted from [14, Lemma 1.49].

Definition 2.13 (Shape and contact regularity)**.** *We call a mesh sequence $\mathcal{T}_H$* shape– and contact–regular *if for all $h \in H$, $\mathcal{T}_h$ admits a matching simplicial submesh $\mathcal{S}_h$ such that*

a) The mesh sequence $\mathcal{S}_H$ is shape–regular in the usual sense of Ciarlet [13], meaning that there is a parameter $\rho_1 > 0$, independent of h, satisfying

$$\rho_1 h_{K'} \leq r_{K'}\,,$$

for all $K' \in \mathcal{S}_h$ where $h_{K'}$ denotes the diameter of K' and $r_{K'}$ the radius of the largest ball inscribed in K',

b) There is a parameter $\rho_2 > 0$, independent of h, such that, for all $K \in \mathcal{T}_h$ and for all $K' \in \mathcal{S}_K$,

$$\rho_2 h_K \leq h_{K'}.$$

Lemma 2.14 (Continuous trace inequality)**.** *Let $\mathcal{T}_H$ be a shape– and contact–regular mesh sequence. Then, for all $h \in H$, all $v \in \mathcal{H}^1(\mathcal{T}_h)$, all $K \in \mathcal{T}_h$, and all $F \in \mathcal{F}_K$ it holds that*

$$\|v\|^2_{\mathcal{L}^2(F)} \leq C_{cti}\big(2\|\nabla v\|_{\boldsymbol{\mathcal{L}}^2(K)} + dh_K^{-1}\|v\|_{\mathcal{L}^2(K)}\big)\|v\|_{\mathcal{L}^2(K)}\,, \tag{2.4}$$

with $C_{cti} := \rho_1^{-1}$ if $\mathcal{T}_h$ is matching and simplicial, while $C_{cti} := (1+d)(\rho_1\rho_2)^{-1}$ otherwise.

Remark 2.15 (Comments regarding the Continuous trace inequality)**.** *H denotes a countable subset of $\mathbb{R}_{>0} := \{x \in \mathbb{R} \,|\, x > 0\}$ having 0 as only accumulation point. The set $\mathcal{F}_K$ collects the mesh faces composing the boundary of K. Further, d denotes the space dimension. For the proof and details about the concepts of matching simplicial submeshes, and the definition of the mesh regularity parameters ρ_1 and ρ_2, we refer to [14].*

Furthermore, the following standard interpolation inequalities are derived from [12] including the proof.

Lemma 2.16 (Interpolation estimate)**.** *For any function $v \in \mathcal{H}^2(\Omega) \cap \mathcal{V}$ with $\mathcal{V} := \mathcal{H}^1_0(\Omega)$, it holds that*

$$\|v - \mathcal{I}_h v\|_{\mathcal{L}^2(K)} \leq C_1 h_K^2 \|\nabla^2 v\|_{\mathcal{L}^2(K)}\,, \tag{2.5}$$

$$\|\nabla(v - \mathcal{I}_h v)\|_{\boldsymbol{\mathcal{L}}^2(K)} \leq C_2 h_K \|\nabla^2 v\|_{\mathcal{L}^2(K)}\,, \tag{2.6}$$

for the standard Lagrange interpolation operator $\mathcal{I}_h v \in \mathcal{V}_h \subset \mathcal{V}$ of the function v satisfying $\mathcal{I}_h v(a_i) = v(a_i)$ for all Lagrange nodes a_i. The constants C_1 and C_2 are independent of the mesh size h_K.

In what follows, we present statements about existence and uniqueness of solutions of elliptic as well as parabolic partial differential equations. The following set of assumptions is quoted from [19] and [32]. For that, Ω is supposed to be an open, bounded subset of $\mathbb{R}^n$. We consider the differential operator

$$Lu = -\sum_{i,j=1}^{n} \left(a^{ij}(\boldsymbol{x})u_{x_i}\right)_{x_j} + \sum_{i=1}^{n} b^i(\boldsymbol{x})u_{x_i} + c(\boldsymbol{x})u\,, \tag{2.7}$$

where L has divergence form.

Assumption 2.17 (A2 Smoothness of the coefficients)**.** *Suppose that the following assumptions about the smoothness of the coefficients are satisfied:*

$$a^{ij} \in \mathcal{L}^\infty(\Omega), b^i \in \mathcal{W}^{1,\infty}(\Omega), c \in \mathcal{L}^\infty(\Omega) \quad (i,j = 1,\dots,n)\,. \tag{2.8a}$$

Moreover, the uniform ellipticity condition is required, i.e. there exists a constant $\theta > 0$ such that

$$\sum_{i,j=1}^{n} a^{ij}(\boldsymbol{x})\xi_i\xi_j \geq \theta|\boldsymbol{\xi}|^2\,, \tag{2.8b}$$

for almost every $\boldsymbol{x} \in \Omega$ and all $\boldsymbol{\xi} \in \mathbb{R}^n$. Further, assume that there is a constant $c_0 \geq 0$ satisfying

$$c(\boldsymbol{x}) - \frac{1}{2}\nabla \cdot \boldsymbol{b}(\boldsymbol{x}) \geq c_0\,, \tag{2.8c}$$

for almost every $\boldsymbol{x} \in \Omega$.

Assumption 2.18 (A3 Regularity of the coefficients)**.** *Suppose that the following assumptions about the smoothness of the coefficients are satisfied:*

$$a^{ij}, b^i \in \mathcal{W}^{1,\infty}(\Omega), c \in \mathcal{L}^\infty(\Omega) \quad (i,j = 1,\dots,n)\,. \tag{2.9}$$

Moreover, the uniform ellipticity condition (2.8b) is required. Further, we assume that condition (2.8c) is satisfied.

Now, we define a weak solution of a parabolic initial/boundary–value problem regarding [19]. Thus, we consider the parabolic initial/boundary–value problem

$$\begin{cases} \partial_t u + Lu &= f \quad \text{in } \Omega \times (0,T] \\ u &= 0 \quad \text{on } \partial\Omega \times (0,T] \\ u &= u_0 \quad \text{on } \Omega \times \{t = 0\}\,, \end{cases} \tag{2.10}$$

where L denotes for each time t a second–order partial differential operator, having the form

$$Lu = -\sum_{i,j=1}^{n} \left(a^{ij}(\boldsymbol{x},t)u_{x_i}\right)_{x_j} + \sum_{i=1}^{n} b^i(\boldsymbol{x},t)u_{x_i} + c(\boldsymbol{x},t)u\,.$$

Definition 2.19 (Weak solution)**.** *Suppose the assumptions in A2 to be fulfilled. Moreover, we assume that*

$$\partial\Omega \text{ is of class } \mathcal{C}^{0,1}\,, \tag{2.11}$$

$$f \in \mathcal{L}^2((0,T);\mathcal{V}^*)\,, \tag{2.12}$$

and

$$u_0 \in \mathcal{L}^2(\Omega)\,. \tag{2.13}$$

A function

$$u \in \mathcal{L}^2((0,T);\mathcal{H}_0^1(\Omega))\,,\ \textit{with } u' \in \mathcal{L}^2((0,T);\mathcal{H}^{-1}(\Omega))\,,$$

is called a weak solution of the parabolic initial/boundary–value problem (2.10) provided that

a) $\langle u', v\rangle_\Omega + \int\limits_\Omega \sum\limits_{i,j=1}^{n} a^{ij}(\boldsymbol{x},t)u_{x_i}v_{x_j} + \sum\limits_{i=1}^{n} b^i(\boldsymbol{x},t)u_{x_i}v + c(\boldsymbol{x},t)uv\,d\boldsymbol{x} = \langle f, v\rangle_\Omega\,,$
for each $v \in \mathcal{H}_0^1(\Omega)$ *and almost every* $0 \le t \le T$ *and*

b) $u(0) = u_0\,.$

Remark 2.20. *We can conclude that* $u \in \mathcal{C}([0,T];\mathcal{L}^2(\Omega))$*. Due to that fact, the equality b) makes sense.*

Remark 2.21. *The weak solution* $u \in \mathcal{H}_0^1(\Omega)$ *of the second–order elliptic boundary–value problem*

$$\left\{\begin{array}{rcll} Lu &=& f & \textit{in } \Omega \\ u &=& 0 & \textit{on } \partial\Omega\,, \end{array}\right. \tag{2.14}$$

with L *given in (2.7) can be defined analogously provided that*

$$f \in \mathcal{V}^*\,. \tag{2.15}$$

Theorem 2.22 (Elliptic boundary–value problem: Uniqueness and regularity)**.** *We sum up the statements about unique existence and regularity of solutions of an elliptic boundary–value problem.*

a) There exists a unique weak solution of (2.14) provided that the assumptions in A2 and the conditions (2.11) and (2.15) are fulfilled.

b) Let the assumptions in A3 be satisfied. Further, we assume that

$$\partial\Omega \textit{ is of class } \mathcal{C}^2\,,$$

and

$$f \in \mathcal{L}^2(\Omega)\,.$$

Then,

$$u \in \mathcal{H}^2(\Omega)\,,$$

and we have the estimate

$$\|u\|_{\mathcal{H}^2(\Omega)} \leq C\left(\|f\|_{\mathcal{L}^2(\Omega)} + \|u\|_{\mathcal{L}^2(\Omega)}\right)\,,$$

where the constant C only depends on Ω and the coefficients of L.

Theorem 2.23 (Parabolic initial/boundary–value problem: Uniqueness and regularity)**.** *We sum up the statements about unique existence and regularity of solutions of an parabolic initial/boundary–value problem.*

a) There exists a unique weak solution of (2.10) provided that the assumptions A2 and (2.11) – (2.13) are fulfilled.

b) Assume

$$u_0 \in \mathcal{H}_0^1(\Omega)\,, f \in \mathcal{L}^2\left((0,T);\mathcal{L}^2(\Omega)\right)\,.$$

Let $u \in \mathcal{L}^2\left((0,T);\mathcal{H}_0^1(\Omega)\right)$, with $u' \in \mathcal{L}^2\left((0,T);\mathcal{H}^{-1}(\Omega)\right)$, be the weak solution of (2.10). Then in fact

$$u \in \mathcal{L}^2\left((0,T);\mathcal{H}^2(\Omega)\right) \cap \mathcal{L}^\infty\left((0,T);\mathcal{H}_0^1(\Omega)\right)\,, u' \in \mathcal{L}^2\left((0,T);\mathcal{L}^2(\Omega)\right)\,.$$

For the proof of both previous theorems, the reader is referred to [19].

Chapter 3

A SUPG stabilized dual weighted residual method

3.1 A general framework

In this section, we provide a general framework to adapt and extend the well–known dual weighted residual method (DWR) (cf. [8]) to a more far–reaching approach which is capable of handling convection–dominated equations in the discrete setting. The aim is to combine the advantages of the DWR method with the stabilizing properties of the streamline upwind/Petrov–Galerkin method (SUPG) (cf. [10]).

We consider the linear convection–diffusion–reaction model problem

$$-\nabla \cdot (\varepsilon \nabla u) + \boldsymbol{b} \cdot \nabla u + \alpha u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega\,, \tag{3.1}$$

on a bounded Lipschitz domain $\Omega \subseteq \mathbb{R}^d, d \in \{2,3\}$. In (3.1) let $\alpha \in \mathbb{R}$, $\varepsilon \in \boldsymbol{\mathcal{L}}^\infty(\Omega)$, $\boldsymbol{b} \in \boldsymbol{\mathcal{H}}^1(\Omega) \cap \boldsymbol{\mathcal{L}}^\infty(\Omega)$ and $f \in \mathcal{L}^2(\Omega)$ be given with

$$\begin{aligned} &\alpha > 0\,, \\ &(\nabla \cdot \boldsymbol{b})(x) = 0 \text{ and } \varepsilon(x) \geq \varepsilon_0 > 0 \quad \text{almost everywhere in } \Omega\,. \end{aligned} \tag{3.2}$$

For the sake of simplicity, the partial differential equation in (3.1) is equipped with homogeneous Dirichlet boundary conditions. Remark 3.19 explains how to incorporate nonhomogeneous Dirichlet boundary conditions. With the function space $\mathcal{V} := \mathcal{H}_0^1(\Omega)$ the weak formulation of (3.1) reads as

Find $u \in \mathcal{V}$ such that

$$A(u)(\varphi) = F(\varphi) \qquad \forall \varphi \in \mathcal{V}\,, \tag{3.3}$$

with

$$\begin{aligned} A(v)(\psi) &:= \langle \varepsilon \nabla v, \nabla \psi \rangle_\Omega + \langle \boldsymbol{b} \cdot \nabla v, \psi \rangle_\Omega + \langle \alpha v, \psi \rangle_\Omega\,, \\ F(\psi) &:= \langle f, \psi \rangle_\Omega\,. \end{aligned}$$

$\langle \cdot, \cdot \rangle_\Omega$ denotes the inner product of $\mathcal{L}^2(\Omega)$. For the finite–dimensional subspace $\mathcal{V}_h \subset \mathcal{V}$ defined in (2.3), the Galerkin approximation u_h is set as the solution of the following discrete problem

Find $u_h \in \mathcal{V}_h$ such that

$$A(u_h)(\varphi_h) = F(\varphi_h) \qquad \forall \varphi_h \in \mathcal{V}_h\,. \tag{3.4}$$

Due to the approach we are going to introduce in the following, we usually use the space of continuous piece–wise linear polynomials to calculate an approximate solution u_h of equation (2.3).

The theoretical basis concept of the DWR method is embedded in the framework of optimal control. In order to derive an a–posteriori error estimation in terms of a given functional $\mathcal{J}(\cdot)$, we seek a solution $u \in \mathcal{V}$ that minimizes the target quantity $\mathcal{J}(u)$ under the constraint $A(u)(\varphi) = F(\varphi)$ for all $\varphi \in \mathcal{V}$. Based on the methods illustrated in [8] and [7], we can envisage two possible approaches:

- $\mathcal{FSTD}$: $\mathcal{F}$irst, we establish the SUPG $\mathcal{S}$tabilized formulation of the discrete equation (3.4), $\mathcal{T}$hen we take the corresponding $\mathcal{D}$ual problem.

- $\mathcal{FDTS}$: $\mathcal{F}$irst, we take the associated $\mathcal{D}$ual problem of equation (3.3), $\mathcal{T}$hen we add the SUPG $\mathcal{S}$tabilization terms to the resulting equation.

In the following two sections, the methods mentioned above are deduced.

3.2 A $\mathcal{FSTD}$ method

As pointed out before, we consider the discrete SUPG formulation of equation (3.4) which is written as

Find $u_h \in \mathcal{V}_h$ such that

$$A_S(u_h)(\varphi_h) = F(\varphi_h) \qquad \forall \varphi_h \in \mathcal{V}_h\,, \tag{3.5}$$

with

$$\begin{aligned} A_S(v_h)(\psi_h) &:= A(v_h)(\psi_h) + S(v_h)(\psi_h)\,, \\ S(v_h)(\psi_h) &:= \sum_{K\in\mathcal{T}_h} \delta_K \langle R(v_h), \boldsymbol{b}\cdot\nabla\psi_h\rangle_K\,, \\ R(v_h) &:= -\nabla\cdot(\varepsilon\nabla v_h) + \boldsymbol{b}\cdot\nabla v_h + \alpha v_h - f\,. \end{aligned}$$

Remark 3.1. *To facilitate matters, we suppose the diffusion coefficient ε to be constant. Failing that, formulation (3.5) has to be modified by inserting a projection operator $\Pi_K : \boldsymbol{\mathcal{L}}^2(K) \to \boldsymbol{\mathcal{P}}(K)$ to treat variable diffusion coefficients in the error analysis. For that purpose, the residual $R(v_h)$ is redefined by*

$$R(v_h) := -\nabla\cdot\Pi_K(\varepsilon\nabla v_h) + \boldsymbol{b}\cdot\nabla v_h + \alpha v_h - f\,.$$

Remark 3.2. *The SUPG tuning parameter δ_K has to be chosen in an optimal way which is the crucial point associated with this. Too large values of δ_K leads to a smearing of the solution whereas too small values do not reduce the unphysical oscillations. For example, [6] and [38] address the parameter design.*

Note that we additionally have to assume that

$$u \in \hat{\mathcal{V}} := \left\{ \hat{v} \in \mathcal{H}_0^1(\Omega) \mid (\nabla\cdot(\varepsilon\nabla\hat{v}))_{|K} \in \mathcal{L}^2(K)\,\forall K \in \mathcal{T}_h \right\}$$

in the nondiscrete case of (3.5). The existence of the solution of the variational problem (3.5) is guaranteed; see [38], [33], [24] and [6] for details.

Starting from the optimization problem (1.4), we introduce the Lagrangian functional

$$\mathfrak{L}_h(u,z) := \mathcal{J}(u) + F(z) - A_S(u)(z) \tag{3.6}$$

with the adjoint variable $z \in \mathcal{V}$ and a differentiable functional $\mathcal{J} : \hat{\mathcal{V}} \to \mathbb{R}$. Minimal solutions u solving the optimization problem correspond to stationary points $\{u, z\} \in \hat{\mathcal{V}} \times \mathcal{V}$ of the Lagrangian (3.6). Therefore, for the Fréchet derivatives of the Lagrangian functional, it holds that

$$\mathfrak{L}'_{h_z}(u,z)(\varphi) = F(\varphi) - A_S(u,\varphi) = 0 \qquad \forall \varphi \in \mathcal{V}\,, \tag{3.7a}$$

$$\mathfrak{L}'_{h_u}(u,z)(\zeta) = \mathcal{J}'(u)(\zeta) - A'_S(u)(\zeta,z) = 0 \qquad \forall \zeta \in \hat{\mathcal{V}}\,. \tag{3.7b}$$

Obviously, equation (3.7a) corresponds to the given variational problem, and equation (3.7b) turned out to be the adjoint problem:

Seek solutions $\{u, z\} \in \hat{\mathcal{V}} \times \mathcal{V}$ such that

$$A_S(u)(\varphi) = F(\varphi) \qquad \forall \varphi \in \mathcal{V}\,, \tag{3.8a}$$

$$A'_S(u)(\zeta,z) = \mathcal{J}'(u)(\zeta) \qquad \forall \zeta \in \hat{\mathcal{V}}\,, \tag{3.8b}$$

where we define the adjoint bilinear form

$$A'_S(v)(\xi,w) := A(\xi,w) + \sum_{K\in\mathcal{T}_h} \delta_K \langle -\nabla\cdot(\varepsilon\nabla\xi) + \boldsymbol{b}\cdot\nabla\xi + \alpha\xi, \boldsymbol{b}\cdot\nabla w\rangle_K\,. \tag{3.9}$$

In order to receive a discretization of the Euler–Lagrange system (3.8a) , (3.8b), we have to consider the following system of discrete equations

Find $\{u_h, z_h\} \in \mathcal{V}_h \times \mathcal{V}_h$ such that

$$A_S(u_h)(\varphi_h) = F(\varphi_h) \qquad \forall \varphi_h \in \mathcal{V}_h\,, \tag{3.10a}$$

$$A'_S(u_h)(\zeta_h, z_h) = \mathcal{J}'(u_h)(\zeta_h) \qquad \forall \zeta_h \in \mathcal{V}_h\,. \tag{3.10b}$$

Since the Lagrangian functional is affine with respect to z , the dual problem is always linear and the primal problem does not depend on z. For this reason, the operators of the system (3.10a), (3.10b) decouple. Hence, to guarantee unique solutions of the overall system it is adequate to proof existence and uniqueness for each single operator. To ensure the existence and uniqueness of the solution z_h for equation (3.10b), we produce the evidence that the proposition of the Lax-Milgram theorem can be transferred to our case.

Theorem 3.3 (Coercivity and boundedness of the adjoint bilinear form)**.** *Suppose that assumption (3.2) and the condition*

$$0 \leq \delta_K \leq \frac{1}{4} \min \left\{ \frac{h_K^2}{p_K^4 \mu_{\text{inv}}^2 \|\varepsilon\|_{\mathcal{L}^\infty(K)}} ; \frac{1}{\alpha} \right\} \tag{3.11}$$

are satisfied. Then, for the adjoint bilinear form (3.10b), there exist constants $\gamma, M > 0$ *such that*

$$A'_S(u_h)(\zeta_h, \zeta_h) \geq \gamma |||\zeta_h|||^2 \qquad \forall \zeta_h \in \mathcal{V}_h \,, \zeta_h \neq 0 \,, \tag{3.12a}$$

$$A'_S(u_h)(\zeta_h, z_h) \leq M |||\zeta_h||| \cdot |||z_h||| \qquad \forall z_h, \zeta_h \in \mathcal{V}_h \,, \zeta_h \neq 0 \,, \tag{3.12b}$$

with $|||\cdot|||$ *defined by (2.1).*

Proof. In order to keep this proof well–structured, some auxiliary calculations are attached below the proof. Using property (3.13) given below, we easily see that

$$\begin{aligned} A'_S(u_h)(\zeta_h, \zeta_h) = \sum_{K \in \mathcal{T}_h} \Big(& \left\| \sqrt{\varepsilon}\, \nabla \zeta_h \right\|^2_{\boldsymbol{\mathcal{L}}^2(K)} + \alpha \left\| \zeta_h \right\|^2_{\mathcal{L}^2(K)} + \delta_K \left\| \boldsymbol{b} \cdot \nabla \zeta_h \right\|^2_{\mathcal{L}^2(K)} \\ & - A_1 + A_2 \Big) \,, \end{aligned}$$

with

$$A_1 := \delta_K \langle \nabla \cdot (\varepsilon \nabla \zeta_h) \,, \boldsymbol{b} \cdot \nabla \zeta_h \rangle_K \text{ and } A_2 := \delta_K \langle \alpha \zeta_h, \boldsymbol{b} \cdot \nabla \zeta_h \rangle_K \,.$$

Assumption (3.11) and standard estimation techniques yield bounds for A_1 and A_2 :

$$\begin{aligned} A_1 &\leq \delta_K \left\| \nabla \cdot (\varepsilon \nabla \zeta_h) \right\|_{\mathcal{L}^2(K)} \left\| \boldsymbol{b} \cdot \nabla \zeta_h \right\|_{\mathcal{L}^2(K)} \\ &\leq \sqrt{\delta_K} \frac{\mu_{\text{inv}} p_K^2}{h_K} \left\| \varepsilon \right\|^{\frac{1}{2}}_{\mathcal{L}^\infty(K)} \left\| \sqrt{\varepsilon}\, \nabla \zeta_h \right\|_{\boldsymbol{\mathcal{L}}^2(K)} \sqrt{\delta_K} \left\| \boldsymbol{b} \cdot \nabla \zeta_h \right\|_{\mathcal{L}^2(K)} \\ &\leq \frac{1}{2} \left\| \sqrt{\varepsilon}\, \nabla \zeta_h \right\|_{\boldsymbol{\mathcal{L}}^2(K)} \sqrt{\delta_K} \left\| \boldsymbol{b} \cdot \nabla \zeta_h \right\|_{\mathcal{L}^2(K)} \,, \\ A_2 &\geq -\delta_K \alpha \left\| \zeta_h \right\|_{\mathcal{L}^2(K)} \left\| \boldsymbol{b} \cdot \nabla \zeta_h \right\|_{\mathcal{L}^2(K)} \\ &\geq -\frac{1}{2} \sqrt{\alpha} \left\| \zeta_h \right\|_{\mathcal{L}^2(K)} \sqrt{\delta_K} \left\| \boldsymbol{b} \cdot \nabla \zeta_h \right\|_{\mathcal{L}^2(K)} \,. \end{aligned}$$

Applying Young's inequality, we obtain that

$$\begin{aligned} A_1 &\leq \frac{1}{4} \left\| \sqrt{\varepsilon}\, \nabla \zeta_h \right\|^2_{\boldsymbol{\mathcal{L}}^2(K)} + \frac{\delta_K}{4} \left\| \boldsymbol{b} \cdot \nabla \zeta_h \right\|^2_{\mathcal{L}^2(K)} \,, \\ A_2 &\geq -\frac{1}{4} \alpha \left\| \zeta_h \right\|^2_{\mathcal{L}^2(K)} - \frac{\delta_K}{4} \left\| \boldsymbol{b} \cdot \nabla \zeta_h \right\|^2_{\mathcal{L}^2(K)} \,. \end{aligned}$$

Finally, we get that

$$A_S'(u_h)(\zeta_h,\zeta_h) \geq \frac{1}{2}|||\zeta_h|||^2 .$$

To prove the second assertion of the theorem, we find that

$$\begin{aligned} &A_S{}'(u_h)(\zeta_h, z_h) \\ &\quad \leq \sum_{K\in\mathcal{T}_h} \Big(\left\|\sqrt{\varepsilon}\,\nabla\zeta_h\right\|_{\boldsymbol{\mathcal{L}}^2(K)} \left\|\sqrt{\varepsilon}\,\nabla z_h\right\|_{\boldsymbol{\mathcal{L}}^2(K)} + \alpha \left\|\zeta_h\right\|_{\mathcal{L}^2(K)} \left\|z_h\right\|_{\mathcal{L}^2(K)} \\ &\qquad + \delta_K \left\|\boldsymbol{b}\cdot\nabla\zeta_h\right\|_{\mathcal{L}^2(K)} \left\|\boldsymbol{b}\cdot\nabla z_h\right\|_{\mathcal{L}^2(K)} + A_3 + A_4 + A_5 \Big) , \end{aligned}$$

with

$$\begin{aligned} A_3 &:= \delta_K \left\|\nabla\cdot(\varepsilon\nabla\zeta_h)\right\|_{\mathcal{L}^2(K)} \left\|\boldsymbol{b}\cdot\nabla z_h\right\|_{\mathcal{L}^2(K)} , \\ A_4 &:= \delta_K\alpha \left\|\zeta_h\right\|_{\mathcal{L}^2(K)} \left\|\boldsymbol{b}\cdot\nabla z_h\right\|_{\mathcal{L}^2(K)} \text{ and } A_5 := \langle \boldsymbol{b}\cdot\nabla\zeta_h, z_h\rangle_K . \end{aligned}$$

We repeat the calculus estimate of the terms A_1 and A_2 and have that

$$\begin{aligned} A_3 &\leq \frac{1}{2}\left\|\sqrt{\varepsilon}\,\nabla\zeta_h\right\|_{\boldsymbol{\mathcal{L}}^2(K)} \sqrt{\delta_K} \left\|\boldsymbol{b}\cdot\nabla z_h\right\|_{\mathcal{L}^2(K)} , \\ A_4 &\leq \frac{1}{2}\sqrt{\alpha} \left\|\zeta_h\right\|_{\mathcal{L}^2(K)} \sqrt{\delta_K} \left\|\boldsymbol{b}\cdot\nabla z_h\right\|_{\mathcal{L}^2(K)} . \end{aligned}$$

From (3.11), we can conclude that $-\dfrac{1}{\sqrt{\delta_K}} \leq -2\sqrt{\alpha}$. In combination with (3.14) determined below, we can write

$$A_5 = -\langle\zeta_h, \boldsymbol{b}\cdot\nabla z_h\rangle_K \leq 2\sqrt{\alpha} \left\|\zeta_h\right\|_{\mathcal{L}^2(K)} \sqrt{\delta_K} \left\|\boldsymbol{b}\cdot\nabla z_h\right\|_{\mathcal{L}^2(K)} .$$

(3.15) given below and estimation techniques lead to

$$\begin{aligned} &A_S'(u_h)(\zeta_h, z_h) \\ &\quad \leq \left(\sum_{K\in\mathcal{T}_h} \frac{5}{4}\left\|\sqrt{\varepsilon}\,\nabla\zeta_h\right\|^2_{\boldsymbol{\mathcal{L}}^2(K)^2} + \frac{7}{2}\alpha\left\|\zeta_h\right\|^2_{\mathcal{L}^2(K)} + \delta_K\left\|\boldsymbol{b}\cdot\nabla\zeta_h\right\|^2_{\mathcal{L}^2(K)^2} \right)^{\frac{1}{2}} \\ &\qquad \cdot \left(\sum_{K\in\mathcal{T}_h} \left\|\sqrt{\varepsilon}\,\nabla z_h\right\|^2_{\boldsymbol{\mathcal{L}}^2(K)^2} + \alpha\left\|z_h\right\|^2_{\mathcal{L}^2(K)} + \frac{9}{2}\delta_K\left\|\boldsymbol{b}\cdot\nabla z_h\right\|^2_{\mathcal{L}^2(K)^2} \right)^{\frac{1}{2}} , \end{aligned}$$

which proves the statement. □

Auxiliary calculation 3.4. *Owing to the assumption* $(\nabla\cdot\boldsymbol{b})(x) = 0$ *, we can observe that for all* $\zeta_h \in \mathcal{V}_h$

$$\begin{aligned} \langle\boldsymbol{b}\cdot\nabla\zeta_h, \zeta_h\rangle_\Omega &= \frac{1}{2}\langle\boldsymbol{b}\cdot\nabla\zeta_h, \zeta_h\rangle_\Omega + \frac{1}{2}\langle\boldsymbol{b}\cdot\nabla\zeta_h, \zeta_h\rangle_\Omega \\ &= \frac{1}{2}\langle\nabla\cdot(\zeta_h\boldsymbol{b}), \zeta_h\rangle_\Omega + \frac{1}{2}\langle\boldsymbol{b}\cdot\nabla\zeta_h, \zeta_h\rangle_\Omega \qquad (3.13) \\ &= -\frac{1}{2}\langle\boldsymbol{b}\zeta_h, \nabla\zeta_h\rangle_\Omega + \frac{1}{2}\langle\boldsymbol{b}\cdot\nabla\zeta_h, \zeta_h\rangle_\Omega = 0 . \end{aligned}$$

Furthermore, the following identity holds for all $\zeta, z \in \mathcal{V}$

$$\langle \boldsymbol{b} \cdot \nabla \zeta, z \rangle_K = \langle \nabla \cdot (\boldsymbol{b}\zeta), z \rangle_K = -\langle \boldsymbol{b}\zeta, \nabla z \rangle_K = -\langle \zeta, \boldsymbol{b} \cdot \nabla z \rangle_K . \tag{3.14}$$

A detailed estimate for the proof of Theorem 3.3 is given by the following calculation.

$$\begin{aligned} A_S'(u_h)(\zeta_h, z_h) \leq & \left(\sum_{K \in \mathcal{T}_h} \| \sqrt{\varepsilon} \nabla \zeta_h \|^2_{\boldsymbol{\mathcal{L}}^2(K)} \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{T}_h} \| \sqrt{\varepsilon} \nabla z_h \|^2_{\boldsymbol{\mathcal{L}}^2(K)} \right)^{\frac{1}{2}} \\ & + \left(\sum_{K \in \mathcal{T}_h} \alpha \| \zeta_h \|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{T}_h} \alpha \| z_h \|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}} \\ & + \left(\sum_{K \in \mathcal{T}_h} \delta_K \| \boldsymbol{b} \cdot \nabla \zeta_h \|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{T}_h} \delta_K \| \boldsymbol{b} \cdot \nabla z_h \|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}} \\ & + \left(\sum_{K \in \mathcal{T}_h} \frac{1}{4} \| \sqrt{\varepsilon} \nabla \zeta_h \|^2_{\boldsymbol{\mathcal{L}}^2(K)} \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{T}_h} \delta_K \| \boldsymbol{b} \cdot \nabla z_h \|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}} \\ & + \left(\sum_{K \in \mathcal{T}_h} \frac{1}{2} \alpha \| \zeta_h \|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{T}_h} \frac{1}{2} \delta_K \| \boldsymbol{b} \cdot \nabla z_h \|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}} \\ & + \left(\sum_{K \in \mathcal{T}_h} 2\alpha \| \zeta_h \|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{T}_h} 2\delta_K \| \boldsymbol{b} \cdot \nabla z_h \|^2_{\mathcal{L}^2(K)} \right)^{\frac{1}{2}} . \end{aligned} \tag{3.15}$$

Having shown the well–posedness of our model system (3.10a), (3.10b), we derive an error representation in terms of the quantity target $\mathcal{J}(\cdot)$. To this end, we have to provide a framework regarding the Galerkin approximation of stationary points.

Let $\mathfrak{L}_h(\cdot)$ be a differentiable functional on a function space $\mathcal{X}$. A stationary point $x \in \mathcal{X}$ of $\mathfrak{L}_h(\cdot)$ satisfies

$$\mathfrak{L}_h'(x)(y) = 0 \quad \forall y \in \mathcal{X} . \tag{3.16}$$

The discrete problem seeks a finite–dimensional Galerkin approximation $x_h \in \mathcal{X}_h \subset \mathcal{X}$ such that

$$\mathfrak{L}_h'(x_h)(y_h) = 0 \quad \forall y_h \in \mathcal{X}_h .$$

In the light of the above, we present an a–posteriori error representation of the error $\hat{e} := x - x_h$. For this, the reader is additionally referred to [8, Proposition 2.1].

Theorem 3.5 (Error representation in terms of the Lagrangian)**.** *For the Galerkin approximation of the variational problem (3.16), we obtain the a–posteriori error representation*

$$\mathfrak{L}_h(x) - \mathfrak{L}_h(x_h) = \frac{1}{2}\mathfrak{L}_h'(x_h)(x - y_h) + R\,, \tag{3.17a}$$

with an arbitrary $y_h \in \mathcal{X}_h$, and where the remainder term is defined by

$$R := \frac{1}{2}\int_0^1 \mathfrak{L}_h'''(x_h + s\hat{e})(\hat{e},\hat{e},\hat{e})s(s-1)\,ds\,. \tag{3.17b}$$

If $\mathfrak{L}_h(\cdot)$ is quadratic, the remainder term vanishes.

Proof. The following identity holds true

$$\mathfrak{L}_h(x) - \mathfrak{L}_h(x_h) = \int_0^1 \mathfrak{L}_h'(x_h + s\hat{e})(\hat{e})\,ds + \frac{1}{2}\mathfrak{L}_h'(x_h)(\hat{e}) - \frac{1}{2}\mathfrak{L}_h'(x_h)(\hat{e}) - \frac{1}{2}\mathfrak{L}_h'(x)(\hat{e})\,,$$

where $\mathfrak{L}_h'(x)(\hat{e}) = 0$. By definition of x_h, we get that

$$\begin{aligned}\mathfrak{L}_h(x) - \mathfrak{L}_h(x_h) = \int_0^1 \mathfrak{L}_h'(x_h + s\hat{e})(\hat{e})\,ds + \frac{1}{2}\mathfrak{L}'(x_h)(x - y_h) - \frac{1}{2}\mathfrak{L}_h'(x_h)(\hat{e}) \\ -\frac{1}{2}\mathfrak{L}'(x)(\hat{e})\,,\end{aligned}$$

with arbitrary $y_h \in \mathcal{X}_h$. Applying the well–known trapezoidal rule including the exact remainder term leads to

$$\mathfrak{L}_h(x) - \mathfrak{L}_h(x_h) = \frac{1}{2}\mathfrak{L}_h'(x_h)(x - y_h) + \frac{1}{2}\int_0^1 \mathfrak{L}_h'''(x_h + s\hat{e})(\hat{e},\hat{e},\hat{e})s(s-1)\,ds\,.$$

□

Now, we state an a–posteriori error representation in terms of the target quantity $\mathcal{J}(\cdot)$. On that point, we introduce the residual $\rho_S(u_h)(\cdot)$ and the adjoint residual $\rho_S^*(z_h)(\cdot)$ defined by

$$\rho_S(u_h)(\varphi) := F(\varphi) - A_S(u_h)(\varphi) \qquad \forall \varphi \in \mathcal{V}\,, \tag{3.18a}$$

$$\rho_S^*(z_h)(\zeta) := \mathcal{J}'(u_h)(\zeta) - A_S'(u_h)(\zeta, z_h) \qquad \forall \zeta \in \hat{\mathcal{V}}\,. \tag{3.18b}$$

By construction, both residuals vanish on $\mathcal{V}_h$ because of Galerkin orthogonality. The error associated to the solution $u_h \in \mathcal{V}_h$ and the adjoint error associated to the solution $z_h \in \mathcal{V}_h$ are given by $e := u - u_h$ and $e^* := z - z_h$, respectively.

Theorem 3.6 (Error representation in terms of the target functional)**.** *For the Galerkin approximation of system (3.8a), (3.8b), the a–posteriori error is represented by*

$$\mathcal{J}(u) - \mathcal{J}(u_h) = \frac{1}{2}\rho_S(u_h)(z - \varphi_h) + \frac{1}{2}\rho_S^*(z_h)(u - \zeta_h) + R\,, \tag{3.19a}$$

with arbitrary $\varphi_h \in \mathcal{V}_h\,, \zeta_h \in \mathcal{V}_h$. The residual $\rho_S(u_h)(\cdot)$ and the adjoint residual $\rho_S^*(z_h)(\cdot)$ are defined in (3.18a) and (3.18b), respectively. The remainder term is determined by

$$R := \frac{1}{2}\int_0^1 \Big(\mathcal{J}'''(u_h + se)(e,e,e)\Big) s(s-1)\,ds\,. \tag{3.19b}$$

Proof. We set

$$\mathcal{X} := \mathcal{V} \times \hat{\mathcal{V}}\,, \quad \mathcal{X}_h := \mathcal{V}_h \times \mathcal{V}_h\,, \quad x := \{u, z\} \text{ and } x_h := \{u_h, z_h\}\,.$$

Then, for the Fréchet derivative of the Lagrangian, it holds that

$$\mathfrak{L}_h'(u,z)(\zeta,\varphi) = \mathfrak{L}_{h_u}'(u,z)(\zeta) + \mathfrak{L}_{h_z}'(u,z)(\varphi) = 0 \quad \forall y := \{\zeta,\varphi\} \in \mathcal{V} \times \hat{\mathcal{V}}\,.$$

The solutions $\{u, z\}$ and $\{u_h, z_h\}$ of the systems (3.8a), (3.8b) and (3.10a), (3.10b) satisfy

$$\begin{aligned}\mathfrak{L}_h(u,z) - \mathfrak{L}_h(u_h,z_h) &= \mathcal{J}(u) + F(z) - A_S(u)(z) - \mathcal{J}(u_h)\\ &\quad -F(z_h) + A_S(u_h)(z_h)\\ &= \mathcal{J}(u) - \mathcal{J}(u_h)\,.\end{aligned}$$

Applying Theorem 3.5 to the described setting with $y_h = \{\zeta_h, \varphi_h\} \in \mathcal{X}_h$ leads to the following error representation:

$$\begin{aligned}\mathcal{J}(u) - \mathcal{J}(u_h) &= \frac{1}{2}\mathfrak{L}_h'(u_h,z_h)(u - \zeta_h, z - \varphi_h) + R\\ &= \frac{1}{2}\Big\{\mathcal{J}'(u_h)(u-\zeta_h) - A_S'(u_h)(u-\zeta_h, z_h) + F(z-\varphi_h)\\ &\quad -A_S(u_h)(z-\varphi_h)\Big\} + R\\ &= \frac{1}{2}\rho_S^*(z_h)(u-\zeta_h) + \frac{1}{2}\rho_S(u_h)(z-\varphi_h) + R\,.\end{aligned}$$

For the remainder term R, we consider the third Fréchet derivative of the Lagrangian $\mathfrak{L}_h$ consisting of the terms

$$\begin{aligned}\mathfrak{L}_h'''(x)(y,y,y) &= \mathfrak{L}_{huuu}'''(x)(\zeta,\zeta,\zeta) + 3\mathfrak{L}_{huuz}'''(x)(\zeta,\zeta,\varphi) + 3\mathfrak{L}_{huzz}'''(x)(\zeta,\varphi,\varphi) \\ &\quad + \mathfrak{L}_{hzzz}'''(x)(\varphi,\varphi,\varphi) \\ &= \mathcal{J}'''(u)(\zeta,\zeta,\zeta) - A_S'''(u)(\zeta,\zeta,\zeta,z) - 3A_S''(u)(\zeta,\zeta,\varphi)\,.\end{aligned}$$

If $A_S(\cdot)(\cdot)$ is linear and if, in addition, $\mathcal{J}(\cdot)$ is quadratic, the remainder term R completely vanishes. Otherwise, anything vanishes except the term

$$\mathcal{J}'''(u)(\zeta,\zeta,\zeta)\,.$$

Therefore, the remainder term R can be written as

$$R = \frac{1}{2}\int_0^1 \Big(\mathcal{J}'''(u_h + se)(e,e,e)\Big)s(s-1)\,ds\,,$$

which proves the assertion. □

Remark 3.7. *In the error representation (3.19a), ρ_S depends on the exact dual solution z whereas ρ_S^* involves the exact primal solution u. Since the exact solutions usually are not available, they have to be approximated by their discrete counterparts. In order to obtain an error representation that is as precise as possible, higher order approximations have to be generated. To reduce the complexity of the algorithm, we derive a relation between the primal and the dual residual such that a higher order approximation has to be calculated once only.*

Theorem 3.8. *For the primal and the dual residual, we find that*

$$\rho_S^*(z_h)(u-\zeta_h) = \rho_S(u_h)(z-\varphi_h) - \Delta\rho_{\mathcal{J}}\,, \tag{3.20a}$$

with arbitrary $\zeta_h, \varphi_h \in \mathcal{V}_h$. The difference between the residuals is given by

$$\Delta\rho_{\mathcal{J}} = \int_0^1 \mathcal{J}''(u_h + se)(e,e)\,ds\,. \tag{3.20b}$$

Proof. We set up the definition

$$k(s) := \mathcal{J}'(u_h + se)(u-\zeta_h) - A_S'(u_h + se)(u-\zeta_h, z_h + se^*)\,,$$

with its derivative

$$k'(s) = \mathcal{J}''(u_h + se)(e, u - \zeta_h) - A_S'(u_h + se)(u - \zeta_h, e^*)\,. \tag{3.21}$$

According to equation (3.8b), we have that

$$k(1) = \mathcal{J}'(u)(u - \zeta_h) - A_S'(u)(u - \zeta_h, z) = 0 \quad \forall \zeta_h \in \mathcal{V}_h\,. \tag{3.22}$$

Owing to the definition of the dual residual, the following identity holds true

$$k(0) = \mathcal{J}'(u_h)(u - \zeta_h) - A_S'(u_h)(u - \zeta_h, z_h) = \rho_S^*(z_h)(u - \zeta_h)\,. \tag{3.23}$$

From the fundamental theorem of calculus applied to (3.21), (3.22) and (3.23), we conclude that

$$\begin{aligned} &\rho_S^*(z_h)(u - \zeta_h) \\ &\quad = -\int_0^1 \Big(\mathcal{J}''(u_h + se)(e, u - \zeta_h) - A_S'(u_h + se)(u - \zeta_h, e^*)\Big)\, ds\,. \end{aligned} \tag{3.24}$$

By taking into consideration equation (3.24) and the definition of (3.20b), we get that

$$\begin{aligned} \rho_S^*(z_h)(u - \zeta_h) &= \rho_S^*(z_h)(u - u_h) \\ &= \int_0^1 \Big(A_S'(u_h + se)(u - u_h, e^*) - \mathcal{J}''(u_h + se)(e, u - u_h)\Big)\, ds \\ &= A_S(u)(e^*) - A_S(u_h)(e^*) - \Delta\rho_{\mathcal{J}} \\ &= \rho_S(u_h)(z - z_h) - \Delta\rho_{\mathcal{J}}\,. \end{aligned}$$

The claim in (3.20a) is confirmed by recalling that the residual $\rho_S(u_h)(\cdot)$ vanishes on $\mathcal{V}_h$ by construction

$$\rho_S^*(z_h)(u - \zeta_h) = \rho_S(u_h)(z - \varphi_h) - \Delta\rho_{\mathcal{J}}\,.$$

□

Remark 3.9. *Former studies of the dual weighted residual method, e.g. [8], have shown that it is reasonable to neglect the remainder terms R and $\Delta\rho_{\mathcal{J}}$ since they are cubic and quadratic, respectively, with respect to the error e which tends to zero as the grid size h tends to zero. As we will point out in our numerical studies, this approach is also justified by our numerical results.*

Theorem 3.10 (Local error representation for the $\mathcal{FSTD}$ method)**.** *For the finite element approximation of (3.8a), the cell–wise error representation reads*

$$\begin{aligned}\mathcal{J}(u) - \mathcal{J}(u_h) = \sum_{K\in\mathcal{T}_h} \Big\{ &\langle \mathcal{R}(u_h), z-\varphi_h\rangle_K \\ &+\delta_K \langle \mathcal{R}(u_h), \boldsymbol{b}\cdot\nabla(z-\varphi_h)\rangle_K \\ &-\langle \mathcal{E}(u_h), z-\varphi_h\rangle_{\partial K} \Big\}.\end{aligned} \tag{3.25a}$$

The cell and edge residuals $\mathcal{R}(u_h)$ and $\mathcal{E}(u_h)$, respectively, are given by

$$\mathcal{R}(u_h)_{|K} = f + \nabla\cdot(\varepsilon\nabla u_h) - \boldsymbol{b}\cdot\nabla u_h - \alpha u_h\,, \tag{3.25b}$$

$$\mathcal{E}(u_h)_{|\Gamma} = \begin{cases} \frac{1}{2}\boldsymbol{n}\cdot[\varepsilon\nabla u_h] & \text{if } \Gamma\subset\partial K\backslash\partial\Omega \\ 0 & \text{if } \Gamma\subset\partial\Omega\,, \end{cases} \tag{3.25c}$$

where $\boldsymbol{n}$ denotes the outer–pointing normal and $[\nabla u_h]$ defines the jump of ∇u_h over the inner edges Γ.

Proof. Combining the results of the two previous theorems provides an error representation which depends only on the primal residual

$$\mathcal{J}(u) - \mathcal{J}(u_h) = \rho_S(u_h)(z-\varphi_h)\,.$$

Cell–wise integration by parts offers a local description of the error with respect to the target quantity

$$\begin{aligned}&\mathcal{J}(u) - \mathcal{J}(u_h) \\ &\quad= \sum_{K\in\mathcal{T}_h} \Big\{ \langle f + \nabla\cdot(\varepsilon\nabla u_h) - \boldsymbol{b}\cdot\nabla u_h - \alpha u_h, z-\varphi_h + \delta_K\cdot\nabla(z-\varphi_h)\rangle_K \\ &\qquad -\langle \boldsymbol{n}\cdot\varepsilon\nabla u_h, z-\varphi_h\rangle_{\partial K} \Big\}.\end{aligned}$$

□

The quantities of the error representation (3.25a) that depend on u_h can easily be evaluated. To motivate the choice of φ_h, we present the following error estimate.

Theorem 3.11. *Let $\{u,z\}\in\hat{\mathcal{V}}\times\mathcal{V}$ be the solution of the system (3.8a), (3.8b) with $z\in\mathcal{H}^2(K)$. It holds that*

$$|\mathcal{J}(u) - \mathcal{J}(u_h)| \le \sum_{K\in\mathcal{T}_h} \rho_K\omega_K\,,$$

with

$$\rho_K := \|\mathcal{R}(u_h)\|_{\mathcal{L}^2(K)} + h_K^{-\frac{1}{2}} \|\mathcal{E}(u_h)\|_{\mathcal{L}^2(\partial K)} ,$$
$$\omega_K := C h_K^2 \|\nabla^2 z\|_{\mathcal{L}^2(K)} .$$

The cell and edge residuals $\mathcal{R}(u_h)$ and $\mathcal{E}(u_h)$ are given in (3.25b) and (3.25c), respectively.

Proof. From the error representation (3.25a), we can conclude that

$$\begin{aligned} |\mathcal{J}(u) - \mathcal{J}(u_h)| \leq \sum_{K \in \mathcal{T}_h} \Big(& \|\mathcal{R}(u_h)\|_{\mathcal{L}^2(K)} \|z - \varphi_h\|_{\mathcal{L}^2(K)} \\ & + \delta_K \|\mathcal{R}(u_h)\|_{\mathcal{L}^2(K)} \|\boldsymbol{b} \cdot \nabla (z - \varphi_h)\|_{\mathcal{L}^2(K)} \\ & + \|\mathcal{E}(u_h)\|_{\mathcal{L}^2(\partial K)} \|z - \varphi_h\|_{\mathcal{L}^2(\partial K)} \Big) \\ \leq \sum_{K \in \mathcal{T}_h} & \rho_K \tilde{\omega}_K , \end{aligned}$$

with

$$\rho_K = \|\mathcal{R}(u_h)\|_{\mathcal{L}^2(K)} + h_K^{-\frac{1}{2}} \|\mathcal{E}(u_h)\|_{\mathcal{L}^2(\partial K)} ,$$
$$\tilde{\omega}_K = \|z - \varphi_h\|_{\mathcal{L}^2(K)} + \delta_K \|\boldsymbol{b} \cdot \nabla (z - \varphi_h)\|_{\mathcal{L}^2(K)} + h_K^{\frac{1}{2}} \|z - \varphi_h\|_{\mathcal{L}^2(\partial K)} .$$

By choosing $\varphi_h = \mathcal{I}_h z \in \mathcal{V}_h$, we find that

$$\tilde{\omega}_K = \|z - \mathcal{I}_h z\|_{\mathcal{L}^2(K)} + \delta_K \|\boldsymbol{b} \cdot \nabla (z - \mathcal{I}_h z)\|_{\mathcal{L}^2(K)} + h_K^{\frac{1}{2}} \|z - \mathcal{I}_h z\|_{\mathcal{L}^2(\partial K)} .$$

Applying the *continuous trace inequality* (2.4) to the face values of the function $z - \mathcal{I}_h z \in \mathcal{V}$, we obtain that

$$\|z - \mathcal{I}_h z\|^2_{\mathcal{L}^2(\partial K)} \leq C \left(\|\nabla (z - \mathcal{I}_h z)\|_{\mathcal{L}^2(K)} + h_K^{-1} \|z - \mathcal{I}_h z\|_{\mathcal{L}^2(K)} \right) \|z - \mathcal{I}_h z\|_{\mathcal{L}^2(K)} .$$

Recalling the standard interpolation estimates (2.5) and (2.6) yields that

$$\|z - \mathcal{I}_h z\|^2_{\mathcal{L}^2(\partial K)} \leq C h_K^3 \|\nabla^2 z\|^2_{\mathcal{L}^2(K)} ,$$

and

$$\|z - \mathcal{I}_h z\|_{\mathcal{L}^2(\partial K)} \leq C h_K^{\frac{3}{2}} \|\nabla^2 z\|_{\mathcal{L}^2(K)} ,$$

respectively. Considering the second contribution to $\tilde{\omega}_K$, we get that

$$\delta_K \|\boldsymbol{b} \cdot \nabla (z - \mathcal{I}_h z)\|_{\mathcal{L}^2(K)} \leq \delta_K \|\boldsymbol{b}\|_{\mathcal{L}^\infty(K)} \|\nabla (z - \mathcal{I}_h z)\|_{\mathcal{L}^2(K)} .$$

Due to the assumption $\boldsymbol{b} \in \boldsymbol{\mathcal{H}}^1(\Omega) \cap \boldsymbol{\mathcal{L}}^\infty(\Omega)$, $\|\boldsymbol{b}\|_{\mathcal{L}^\infty(K)}$ is bounded. We notice that $\delta_K \leq Ch_K$ and conclude that

$$\delta_K \|\boldsymbol{b} \cdot \nabla(z - \mathcal{I}_h z)\|_{\mathcal{L}^2(K)} \leq Ch_K \|\boldsymbol{b} \cdot \nabla(z - \mathcal{I}_h z)\|_{\mathcal{L}^2(K)} .$$

The interpolation properties (2.5) and (2.6) prove the assertion

$$\tilde{\omega}_K \leq Ch_K^2 \|\nabla^2 z\|_{\mathcal{L}^2(K)} .$$

□

The a–posteriori error estimate analyzed in Theorem 3.11 has the same order as a standard $\mathcal{L}^2$ a–priori error estimate which is

$$\|u - u_h\|_{\mathcal{L}^2(\Omega)} \leq Ch^2 .$$

This proves that the DWR framework leads to a standard error estimate. Therefore, the dual weighted residual method is reasonable in the context of error estimation. However, we are not interested in error estimates. We focus on an exact error representation.

3.3 A $\mathcal{FDTS}$ method

In this section, we present our considerations with regard to the second possible approach how to combine the DWR method with stabilized finite element approximations. The $\mathcal{FDTS}$ method is inspired by [7].

Following this strategy, we consider the primal problem (3.1) together with the corresponding adjoint problem in its strong form

$$-\nabla \cdot (\varepsilon \nabla z) - \boldsymbol{b} \cdot \nabla z + \alpha z = j \quad \text{in } \Omega, \qquad z = 0 \quad \text{on } \partial\Omega . \tag{3.26}$$

We notice that the conditions of (3.2) hold and assume that $j \in \mathcal{L}^2(\Omega)$. Furthermore, we observe that primal and dual problem only differ in the negative sign of the vector field $\boldsymbol{b}$. The weak form of the primal problem is given by (3.3). $\mathcal{J} : \mathcal{V} \to \mathbb{R}$ is supposed to be a possibly nonlinear differentiable target quantity such that $\mathcal{J}'(\cdot)(\zeta) = \langle j(\cdot), \zeta \rangle_\Omega$. The definition of the Lagrangian functional

$$\mathfrak{L}(u, z) := \mathcal{J}(u) + F(z) - A(u)(z)$$

provides the Euler–Lagrange system

Find solutions $\{u, z\} \in \mathcal{V} \times \mathcal{V}$ such that

$$A(u)(\varphi) = F(\varphi) \qquad \forall \varphi \in \mathcal{V}\,, \tag{3.27a}$$
$$A(\zeta)(z) = \mathcal{J}'(u)(\zeta) \qquad \forall \zeta \in \mathcal{V}\,, \tag{3.27b}$$

with $\mathcal{V} := \mathcal{H}_0^1(\Omega)$ and

$$A(\xi)(w) := \langle \varepsilon \nabla w, \nabla \xi \rangle_\Omega - \langle \boldsymbol{b} \cdot \nabla w, \xi \rangle_\Omega + \langle \alpha w, \xi \rangle_\Omega\,,$$
$$\mathcal{J}'(u)(\xi) := \langle j, \xi \rangle_\Omega\,.$$

We observe that equation (3.27a) corresponds to the Fréchet derivative of the Lagrangian $\mathfrak{L}(u, z)$ with respect to the adjoint variable z whereas the calculation of the Fréchet derivative with respect to u offers the second equation. As we deal the case of convection–dominated equations, we solve the following discrete system of equations stabilized by the SUPG method

Seek solutions $\{u_h, z_h\} \in \mathcal{V}_h \times \mathcal{V}_h\,, \mathcal{V}_h \subset \mathcal{V}$, such that

$$A_S(u_h)(\varphi_h) = F(\varphi_h) \qquad \forall \varphi_h \in \mathcal{V}_h\,, \tag{3.28a}$$
$$A_{S^*}(\zeta_h)(z_h) = \mathcal{J}'(u_h)(\zeta_h) \qquad \forall \zeta_h \in \mathcal{V}_h\,, \tag{3.28b}$$

where

$$A_S(v_h)(\psi_h) := A(v_h)(\psi_h) + S(v_h)(\psi_h)\,,$$
$$S(v_h)(\psi_h) := \sum_{K \in \mathcal{T}_h} \delta_K \langle R(v_h), \boldsymbol{b} \cdot \nabla \psi_h \rangle_K\,,$$
$$R(v_h) := -\nabla \cdot (\varepsilon \nabla v_h) + \boldsymbol{b} \cdot \nabla v_h + \alpha v_h - f\,,$$

and

$$A_{S^*}(\xi_h)(w_h) := A(\xi_h)(w_h) + S^*(\xi_h)(w_h)\,,$$
$$S^*(\xi_h)(w_h) := \sum_{K \in \mathcal{T}_h} \delta_K^* \langle \nabla \cdot (\varepsilon \nabla w_h) + \boldsymbol{b} \cdot \nabla w_h - \alpha w_h + j, \boldsymbol{b} \cdot \nabla \xi_h \rangle_K\,.$$

Remark 3.12. *By construction of the stabilized dual problem, we observe that the dual tuning parameter δ_K^* for the SUPG stabilization of the dual problem can be chosen differently from the primal tuning parameter δ_K. Consequently, the stabilization can be better adapted to the respective structure of the equation.*

Obviously, the components of the system (3.28a), (3.28b) are decoupled. Hence, existence analysis is simply carried out for each single operator. As shown in [6], the primal problem possesses a unique discrete solution. In the following theorem, we present a statement concerning the existence and uniqueness of the solution of equation (3.28b).

Theorem 3.13 (Coercivity and boundedness of the adjoint bilinear form)**.** *Assume condition (3.2) and*

$$0 \leq \delta_K^* \leq \frac{1}{4} \min \left\{ \frac{h_K^2}{p_K^4 \mu_{\text{inv}}^2 \|\varepsilon\|_{\mathcal{L}^\infty(K)}}; \frac{1}{\alpha} \right\} \tag{3.29}$$

to be true. Then, we can indicate constants $\tilde{\gamma}, \tilde{M} > 0$ *such that*

$$\tilde{A}_{S^*}(\zeta_h)(\zeta_h) \geq \tilde{\gamma} |||\zeta_h|||^2 \qquad \forall \zeta_h \in \mathcal{V}_h \, , \zeta_h \neq 0 \, , \tag{3.30a}$$

$$\tilde{A}_{S^*}(\zeta_h)(z_h) \leq \tilde{M} |||\zeta_h||| \cdot |||z_h||| \qquad \forall z_h, \zeta_h \in \mathcal{V}_h \, , \varphi_h \neq 0 \, , \tag{3.30b}$$

with

$$\begin{aligned} \tilde{A}_{S^*}(\xi_h)(w_h) : &= A_{S^*}(\xi_h)(w_h) - \sum_{K \in \mathcal{T}_h} \langle j, \boldsymbol{b} \cdot \nabla \xi_h \rangle_K \\ &= \langle \varepsilon \nabla w_h, \nabla \xi_h \rangle_\Omega - \langle \boldsymbol{b} \cdot \nabla w_h, \xi_h \rangle_\Omega + \langle \alpha w_h, \xi_h \rangle_\Omega \\ &\quad + \sum_{K \in \mathcal{T}_h} \delta_K^* \langle \nabla \cdot (\varepsilon \nabla w_h) + \boldsymbol{b} \cdot w_h - \alpha w_h, \boldsymbol{b} \cdot \nabla \xi_h \rangle_K \, . \end{aligned}$$

Proof. For the first part of the theorem, we use (3.13) and easily see that

$$\begin{aligned} \tilde{A}_{S^*}(\zeta_h)(\zeta_h) = \sum_{K \in \mathcal{T}_h} \Big\{ &\|\sqrt{\varepsilon} \nabla \zeta_h\|_{\boldsymbol{\mathcal{L}}^2(K)}^2 + \alpha \|\zeta_h\|_{\mathcal{L}^2(K)}^2 + \delta_K^* \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)}^2 \\ &+ B_1 + B_2 \Big\} \, , \end{aligned}$$

where the terms B_1 and B_2 are given by

$$B_1 := \delta_K^* \langle \nabla \cdot (\varepsilon \nabla \zeta_h), \boldsymbol{b} \cdot \nabla \zeta_h \rangle_K \text{ and } B_2 := -\delta_K^* \langle \alpha \zeta_h, \boldsymbol{b} \cdot \nabla \zeta_h \rangle_K \, .$$

As the estimation of the previous terms is straightforward to follow the steps of the proof for the statement (3.12a) of Theorem 3.3, we keep the proof very brief at this point. We get that

$$\begin{aligned} B_1 &\geq -\delta_K^* \|\nabla \cdot (\varepsilon \nabla \zeta_h)\|_{\mathcal{L}^2(K)} \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)} \\ &\geq -\frac{1}{4} \|\sqrt{\varepsilon} \nabla \zeta_h\|_{\boldsymbol{\mathcal{L}}^2(K)}^2 - \frac{1}{4} \delta_K^* \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)}^2 \, , \end{aligned}$$

and

$$\begin{aligned} B_2 &\geq -\delta_K^* \alpha \|\zeta_h\|_{\mathcal{L}^2(K)} \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)} \\ &\geq -\frac{1}{4}\alpha \|\zeta_h\|_{\mathcal{L}^2(K)}^2 - \frac{1}{4}\delta_K^* \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)}^2 . \end{aligned}$$

All in all, we find that

$$\tilde{A}_{S^*}(\zeta_h)(\zeta_h) \geq \frac{1}{2} |||\zeta_h|||^2 ,$$

which proves assertion (3.30a). Now, we give proof for the second statement (3.30b) which also is very similar to the proof of Theorem 3.3. We have that

$$\begin{aligned} &\tilde{A}_{S^*}(\zeta_h)(z_h) \\ &\quad \leq \sum_{K \in \mathcal{T}_h} \Big(\left\| \sqrt{\varepsilon}\, \nabla z_h \right\|_{\boldsymbol{\mathcal{L}}^2(K)} \left\| \sqrt{\varepsilon}\, \nabla \zeta_h \right\|_{\boldsymbol{\mathcal{L}}^2(K)} + \alpha \left\| z_h \right\|_{\mathcal{L}^2(K)} \left\| \zeta_h \right\|_{\mathcal{L}^2(K)} \\ &\qquad + \delta_K \left\| \boldsymbol{b} \cdot \nabla z_h \right\|_{\mathcal{L}^2(K)} \left\| \boldsymbol{b} \cdot \nabla \zeta_h \right\|_{\mathcal{L}^2(K)} + B_3 + B_4 + B_5 \Big) , \end{aligned}$$

with

$$\begin{aligned} B_3 &:= \delta_K^* \|\nabla \cdot (\varepsilon \nabla z_h)\|_{\mathcal{L}^2(K)} \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)} , \\ B_4 &:= \delta_K^* \alpha \|z_h\|_{\mathcal{L}^2(K)} \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)} \text{ and } B_5 := -\langle \boldsymbol{b} \cdot \nabla z_h, \zeta_h \rangle_K . \end{aligned}$$

Assumption (3.29) and estimation techniques yield that

$$\begin{aligned} B_3 &\leq \frac{1}{2} \|\sqrt{\varepsilon} \nabla z_h\|_{\mathcal{L}^2(K)} \sqrt{\delta_K^*} \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)} , \\ B_4 &\leq \frac{1}{2} \sqrt{\alpha} \|z_h\|_{\mathcal{L}^2(K)} \sqrt{\delta_K^*} \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)} . \end{aligned}$$

Recalling (3.29), i.e. $-\frac{1}{\sqrt{\delta_K^*}} \leq -2\sqrt{\alpha}$, implies that

$$B_5 \leq 2\sqrt{\delta_K^*} \|\boldsymbol{b} \cdot \nabla z_h\|_{\mathcal{L}^2(K)} \sqrt{\alpha} \|\zeta_h\|_{\mathcal{L}^2(K)} .$$

For the sake of clarity, the reader is referred to (3.31) of the explanatory notes below this proof at this point. Finally, we get that

$$\begin{aligned} \tilde{A}_{S^*}(\zeta_h)(z_h) &\leq \left(\sum_{K \in \mathcal{T}_h} \frac{5}{4} \|\sqrt{\varepsilon} \nabla z_h\|_{\boldsymbol{\mathcal{L}}^2(K)}^2 + \frac{5}{4}\alpha \|z_h\|_{\mathcal{L}^2(K)}^2 + 5\delta_K^* \|\boldsymbol{b} \cdot \nabla z_h\|_{\mathcal{L}^2(K)}^2 \right)^{\frac{1}{2}} \\ &\qquad \cdot \left(\sum_{K \in \mathcal{T}_h} \|\sqrt{\varepsilon} \nabla \zeta_h\|_{\boldsymbol{\mathcal{L}}^2(K)}^2 + 2\alpha \|\zeta_h\|_{\mathcal{L}^2(K)}^2 + 3\delta_K^* \|\boldsymbol{b} \cdot \nabla \zeta_h\|_{\mathcal{L}^2(K)}^2 \right)^{\frac{1}{2}} \\ &\leq \tilde{M} |||z_h||| \cdot |||\zeta_h||| . \end{aligned}$$

□

Auxiliary calculation 3.14. *Supplementary estimate to the proof of Theorem 3.13.*

$$\begin{aligned}
\tilde{A}_{S^*}(\zeta_h)(z_h) \leq & \left(\sum_{K\in\mathcal{T}_h} \|\sqrt{\varepsilon}\nabla z_h\|^2_{\boldsymbol{\mathcal{L}}^2(K)}\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \|\sqrt{\varepsilon}\nabla \zeta_h\|^2_{\boldsymbol{\mathcal{L}}^2(K)}\right)^{\frac{1}{2}} \\
& + \left(\sum_{K\in\mathcal{T}_h} \alpha\|z_h\|^2_{\mathcal{L}^2(K)}\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \alpha\|\zeta_h\|^2_{\mathcal{L}^2(K)}\right)^{\frac{1}{2}} \\
& + \left(\sum_{K\in\mathcal{T}_h} \delta_K^*\|\boldsymbol{b}\cdot\nabla z_h\|^2_{\mathcal{L}^2(K)}\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \delta_K^*\|\boldsymbol{b}\cdot\nabla \zeta_h\|^2_{\mathcal{L}^2(K)}\right)^{\frac{1}{2}} \\
& + \left(\sum_{K\in\mathcal{T}_h} \frac{1}{4}\|\sqrt{\varepsilon}\nabla z_h\|^2_{\boldsymbol{\mathcal{L}}^2(K)}\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \delta_K^*\|\boldsymbol{b}\cdot\nabla \zeta_h\|^2_{\mathcal{L}^2(K)}\right)^{\frac{1}{2}} \\
& + \left(\sum_{K\in\mathcal{T}_h} \frac{1}{4}\alpha\|z_h\|^2_{\mathcal{L}^2(K)}\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \delta_K^*\|\boldsymbol{b}\cdot\nabla \zeta_h\|^2_{\mathcal{L}^2(K)}\right)^{\frac{1}{2}} \\
& + \left(\sum_{K\in\mathcal{T}_h} 4\delta_K^*\|\boldsymbol{b}\cdot\nabla z_h\|^2_{\mathcal{L}^2(K)}\right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \alpha\|\zeta_h\|^2_{\mathcal{L}^2(K)}\right)^{\frac{1}{2}} .
\end{aligned} \tag{3.31}$$

Now, we derive an error representation in terms of the Lagrangian functional. The general procedure as well as the following theorems are based on the approach to an a–posteriori error estimation for finite element discretizations of the Navier–Stokes equations suggested in [7]. For that purpose, we choose a differentiable functional $\mathfrak{L}(\cdot)$ on a function space $\mathcal{X} := \mathcal{V}\times\mathcal{V}$. The Lagrangian functional is defined by

$$\mathfrak{L}(u,z) := \mathcal{J}(u) + F(z) - A(u)(z)\,. \tag{3.32}$$

As usual, the derivative of $\mathfrak{L}$ with respect to z gives the original problem (3.27a) and its derivative with respect to u corresponds to the dual problem (3.27*b*). Let $x\in\mathcal{X}$ be a stationary point of $\mathfrak{L}$, i.e.

$$\mathfrak{L}'(x)(y) = 0 \quad \forall y\in\mathcal{X}\,. \tag{3.33}$$

We summarize the discrete equations (3.28a) and (3.28b)

$$\begin{aligned}
& F(\varphi_h) - A(u_h)(\varphi_z) - S(u_h)(\varphi_h) \\
& \qquad + \mathcal{J}'(u_h)(\zeta_h) - A(\zeta_h)(z_h) - S^*(\zeta_h)(z_h) = 0 \quad \forall\{\zeta_h,\varphi_h\}\in\mathcal{V}_h\times\mathcal{V}_h\,,
\end{aligned}$$

and find that

$$\begin{aligned}\mathfrak{L}'_u(u_h, z_h)(\zeta_h) + \mathfrak{L}'_z(u_h, z_h)(\varphi_h) \\ - S(u_h)(\varphi_h) - S^*(\zeta_h)(z_h) = 0 \quad \forall \{\zeta_h, \varphi_h\} \in \mathcal{V}_h \times \mathcal{V}_h \,.\end{aligned}$$

Setting

$$\begin{aligned} x_h &:= \{u_h, z_h\} \in \mathcal{V}_h \times \mathcal{V}_h =: \mathcal{X}_h \subset \mathcal{X} \,, \\ y_h &:= \{\zeta_h, \varphi_h\} \in \mathcal{V}_h \times \mathcal{V}_h \,, \end{aligned}$$

and

$$\mathcal{S}(x_h)(y_h) := S(u_h)(\varphi_h) + S^*(\zeta_h)(z_h) \,,$$

we obtain that

$$\mathfrak{L}'(x_h)(y_h) = \mathcal{S}(x_h)(y_h) \quad \forall y_h \in \mathcal{X}_h \,. \tag{3.34}$$

Theorem 3.15 (Error representation in terms of the Lagrangian functional)**.** *Let $x \in \mathcal{X}$ be a stationary point of $\mathfrak{L}$. Suppose that $x_h \in \mathcal{X}_h$ fulfills equation (3.34). For the error representation, we have that*

$$\mathfrak{L}(x) - \mathfrak{L}(x_h) = \frac{1}{2}\mathfrak{L}'(x_h)(x - y_h) + \frac{1}{2}\mathcal{S}(x_h)(y_h - x_h) + R \,, \tag{3.35a}$$

with an arbitrary $y_h \in \mathcal{X}_h$ and a remainder term defined by

$$R := \frac{1}{2}\int_0^1 \mathfrak{L}'''(x_h + s\hat{e})(\hat{e}, \hat{e}, \hat{e}) s(s-1)\, ds \,. \tag{3.35b}$$

Proof. Let $\hat{e} := x - x_h \in \mathcal{X}$. Elementary calculus gives that

$$\mathfrak{L}(x) - \mathfrak{L}(x_h) = \int_0^1 \mathfrak{L}'(x_h + s\hat{e})(\hat{e})\, ds \,.$$

We approximate this integral by the trapezoidal rule and get the result that

$$\mathfrak{L}(x) - \mathfrak{L}(x_h) = \frac{1}{2}\mathfrak{L}'(x_h)(x - x_h) + \frac{1}{2}\mathfrak{L}'(x)(x - x_h) + R \,,$$

with the remainder term R defined in (3.35b). Equation (3.33) requires that the second term of the previous equation vanishes. A calculation in combination with assumption (3.34) completes the proof

$$\begin{aligned}\mathfrak{L}(x) - \mathfrak{L}(x_h) &= \frac{1}{2}\mathfrak{L}'(x_h)(x - y_h) + \frac{1}{2}\mathfrak{L}'(x_h)(y_h - x_h) + R \\ &= \frac{1}{2}\mathfrak{L}'(x_h)(x - y_h) + \frac{1}{2}\mathcal{S}(x_h)(y_h - x_h) + R \,.\end{aligned}$$

□

For the subsequent assertion, we introduce the residual and the adjoint residual by

$$\rho(u_h)(\varphi) := F(\varphi) - A(u_h)(\varphi) \qquad \forall \varphi \in \mathcal{V}\,, \tag{3.36a}$$

$$\rho^*(z_h)(\zeta) := \mathcal{J}'(u_h)(\zeta) - A(\zeta)(z_h) \qquad \forall \zeta \in \mathcal{V}\,. \tag{3.36b}$$

Theorem 3.16 (Error representation in terms of the functional $\mathcal{J}$)**.** *For the error with respect to the target functional $\mathcal{J}(\cdot)$, there holds that*

$$\begin{aligned} \mathcal{J}(u) - \mathcal{J}(u_h) = \frac{1}{2}\rho(u_h)(z - \varphi_h) + \frac{1}{2}\rho^*(z_h)(u - \zeta_h) \\ + R_{\mathcal{S}} + R_{\mathcal{J}}\,, \end{aligned} \tag{3.37a}$$

with arbitrary $\varphi_h, \zeta_h \in \mathcal{V}_h$, the residual $\rho(u_h)(\cdot)$ given in (3.36a) and the adjoint residual $\rho^*(z_h)(\cdot)$ defined in (3.36b). The remainder terms are determined by

$$R_{\mathcal{S}} := \frac{1}{2}S(u_h)(\varphi_h + z_h) + \frac{1}{2}S^*(\zeta_h - u_h)(z_h)\,, \tag{3.37b}$$

and

$$R_{\mathcal{J}} := \frac{1}{2}\int_0^1 \mathcal{J}'''(u_h + se)(e,e,e)s(s-1)\,ds\,. \tag{3.37c}$$

The terms $R_{\mathcal{S}}$ and $R_{\mathcal{J}}$ are caused by the stabilization of the Euler–Lagrange system and the nonlinearity of the target functional, respectively.

Proof. We start with two identities

$$\begin{aligned} \mathfrak{L}(x) &= \mathcal{J}(u) + F(z) - A(u)(z) = \mathcal{J}(u)\,, \\ \mathfrak{L}(x_h) &= \mathcal{J}(u_h) + F(z_h) - A(u_h)(z_h) = \mathcal{J}(u_h) + S(u_h)(z_h)\,. \end{aligned}$$

By using these identities and applying the previous theorem, we find that

$$\begin{aligned} \mathcal{J}(u) - \mathcal{J}(u_h) &= \mathfrak{L}(x) - \mathfrak{L}(x_h) + S(u_h)(z_h) \\ &= \frac{1}{2}\mathfrak{L}'(x_h)(x - y_h) + \frac{1}{2}\mathcal{S}(x_h)(y_h - x_h) + S(u_h)(z_h) + R\,, \end{aligned}$$

where R denotes the remainder term given in (3.35b). For the Fréchet derivative of the Lagrangian functional, it holds that

$$\begin{aligned} \mathfrak{L}'(x_h)(x - y_h) &= \mathcal{J}'(u_h)(u - \zeta_h) - A(u - \zeta_h)(z_h) + F(z - \varphi_h) \\ &\quad - A(u_h)(z - \varphi_h) \\ &= \rho(u_h)(z - \varphi_h) + \rho^*(z_h)(u - \zeta_h)\,. \end{aligned}$$

The contributions with respect to the stabilization can be rewritten as

$$\frac{1}{2}\mathcal{S}(x_h)(y_h - x_h) + S(u_h)(z_h) = \frac{1}{2}S(u_h)(\varphi_h + z_h) + \frac{1}{2}S^*(\zeta_h - u_h)(z_h) =: R_{\mathcal{S}}\,.$$

The fact that, due to the bilinearity of $A(\cdot)(\cdot)$, all parts of the third derivative of $\mathfrak{L}$ vanish except the third derivative of $\mathcal{J}$, finishes the proof. □

Due to the fact that the error representation (3.37a) requires an improved approximation to the exact primal solution u in addition to the improved approximation to z, we derive the following theorem that shows a relation between the primal and dual residual.

Theorem 3.17. *The primal residual $\rho(u_h)(\cdot)$ and the adjoint residual $\rho^*(z_h)(\cdot)$ are related by*

$$\rho^*(z_h)(u - \zeta_h) = \rho(u_h)(z - \varphi_h) - \Delta\rho_{\mathcal{J}} + \Delta\rho_{\mathcal{S}}\,, \tag{3.38a}$$

where the nonlinearity of $\mathcal{J}$ affects the remainder $\Delta\rho_{\mathcal{J}}$ given by

$$\Delta\rho_{\mathcal{J}} := \int_0^1 \mathcal{J}''(u_h + se)(e, e)\,ds\,, \tag{3.38b}$$

and the stabilization affects the term $\Delta\rho_{\mathcal{S}}$ defined by

$$\Delta\rho_{\mathcal{S}} := S(u_h)(\varphi_h - z_h) - S^*(\zeta_h - u_h)(z_h)\,. \tag{3.38c}$$

Proof. We define

$$k(s) := \mathcal{J}'(u_h + se)(u - \zeta_h) - A(u - \zeta_h)(z_h + se^*)\,,$$

where $e^* := z - z_h$ denotes the adjoint error. The derivative of $k(\cdot)$ is given by

$$k'(s) = \mathcal{J}''(u_h + se)(e, u - \zeta_h) - A(u - \zeta_h)(e^*)\,. \tag{3.39}$$

The definition of z offers that

$$k(1) = \mathcal{J}'(u)(u - \zeta_h) - A(u - \zeta_h)(z) = 0 \quad \forall \zeta_h \in \mathcal{V}_h\,.$$

We evaluate $k(s)$ at $s = 0$ and find that

$$k(0) = \mathcal{J}'(u_h)(u - \zeta_h) - A(u - \zeta_h)(z_h) = \rho^*(z_h)(u - \zeta_h)\,. \tag{3.40}$$

By elementary calculus, we have that

$$k(0) = -\int_0^1 k'(s)\, ds\,. \tag{3.41}$$

Entering (3.39) and (3.40) into the expression (3.41) leads to

$$\rho^*(z_h)(u-\zeta_h) = \int_0^1 \Big(A(u-\zeta_h)(e^*) - \mathcal{J}''(u_h+se)(e, u-\zeta_h)\Big)\, ds\,. \tag{3.42}$$

By replacing $u-\zeta_h$ by $u-u_h$ in the dual residual, we get that

$$\begin{aligned}\rho^*(z_h)(u-\zeta_h) &= \mathcal{J}'(u_h)(u-\zeta_h) - A(u-\zeta_h)(z_h) + S^*(\zeta_h)(z_h) - S^*(\zeta_h)(z_h)\\ &\quad -\mathcal{J}'(u_h)(u_h) + A(u_h)(z_h) + S^*(u_h)(z_h)\\ &= \rho^*(z_h)(u-u_h) - S^*(\zeta_h - u_h)(z_h)\,.\end{aligned}$$

Owing to (3.42), we obtain that

$$\begin{aligned}\rho^*(z_h)(u-\zeta_h) &= \int_0^1 \Big(A(u-u_h)(e^*) - \mathcal{J}''(u_h+se)(e, u-u_h)\Big)\, ds\\ &\quad -S^*(\zeta_h - u_h)(z_h)\\ &= A(e)(e^*) - \int_0^1 \mathcal{J}''(u_h+se)(e,e)\, ds\\ &\quad -S^*(\zeta_h - u_h)(z_h)\,.\end{aligned} \tag{3.43}$$

By inserting

$$0 = A_S(u_h)(\varphi_h) - F(\varphi_h)\,,$$

we can rewrite $A(e)(e^*)$

$$\begin{aligned}A(u)(e^*) - A(u_h)(e^*) &= F(e^*) - A(u_h)(e^*)\\ &= F(z) - A(u_h)(z) + A(u_h)(z_h) + S(u_h)(z_h)\\ &\quad -F(z_h) - S(u_h)(z_h) + A(u_h)(\varphi_h)\\ &\quad +S(u_h)(\varphi_h) - F(\varphi_h)\\ &= F(z-\varphi_h) - A(u_h)(z-\varphi_h) + S(u_h)(\varphi_h - z_h)\\ &= \rho(u_h)(z-\varphi_h) + S(u_h)(\varphi_h - z_h)\,.\end{aligned} \tag{3.44}$$

Substituting formula (3.44) into (3.43) confirms the statement

$$\rho^*(z_h)(u-\zeta_h) = \rho(u_h)(z-\varphi_h) + \Delta\rho_{\mathcal{S}} - \Delta\rho_{\mathcal{J}}\,,$$

with the remainder terms $\Delta\rho_{\mathcal{J}}$ and $\Delta\rho_{\mathcal{S}}$ defined in (3.38b) and (3.38c), respectively. □

Since the remainder terms $R_{\mathcal{J}}$ and $\Delta\rho_{\mathcal{J}}$ are cubic and quadratic in e, respectively, they are neglectable as explained in Remark 3.9. Combining the results of the previous theorems leads to an element–wise description of the error in terms of the target quantity $\mathcal{J}$.

Theorem 3.18 (Cell–wise error representation for the $\mathcal{FDTS}$ method)**.** *For the stabilized finite element approximation of the model problem (3.1), the local error representation reads*

$$\begin{aligned}\mathcal{J}(u) - \mathcal{J}(u_h) = \sum_{K\in\mathcal{T}_h} \Big\{ &\langle \mathcal{R}(u_h), z-\varphi_h\rangle_K - \delta_K \langle \mathcal{R}(u_h), \boldsymbol{b}\cdot\nabla\varphi_h\rangle_K \\ &-\langle \mathcal{E}(u_h), z-\varphi_h\rangle_{\partial K}\Big\}\,.\end{aligned} \tag{3.45a}$$

The cell and edge residuals $\mathcal{R}(u_h)$ and $\mathcal{E}(u_h)$ are defined by

$$\mathcal{R}(u_h)_{|K} = f + \nabla\cdot(\varepsilon\nabla u_h) - \boldsymbol{b}\cdot\nabla u_h - \alpha u_h\,, \tag{3.45b}$$

$$\mathcal{E}(u_h)_{|\Gamma} = \begin{cases} \frac{1}{2}\boldsymbol{n}\cdot[\varepsilon\nabla u_h] & \text{if } \Gamma\subset\partial K\backslash\partial\Omega \\ 0 & \text{if } \Gamma\subset\partial\Omega\,, \end{cases} \tag{3.45c}$$

where $\boldsymbol{n}$ denotes the outer–pointing normal and $[\nabla u_h]$ defines the jump of ∇u_h over the inner edges Γ.

Proof. From the previous two theorems, we obtain that

$$\begin{aligned}\mathcal{J}(u) - \mathcal{J}(u_h) &= \rho(u_h)(z-\varphi_h) + \frac{1}{2}\Delta\rho_{\mathcal{S}} + \mathcal{R}_{\mathcal{S}} \\ &= \rho(u_h)(z-\varphi_h) + S(u_h)(\varphi_h)\,.\end{aligned}$$

By cell–wise integration by parts, we get that

$$\begin{aligned}&\mathcal{J}(u) - \mathcal{J}(u_h) \\ &\quad= \sum_{K\in\mathcal{T}_h} \Big\{ \langle f + \nabla\cdot(\varepsilon\nabla u_h) - \boldsymbol{b}\cdot\nabla u_h - \alpha u_h, z-\varphi_h\rangle_K \\ &\qquad -\langle \boldsymbol{n}\cdot\varepsilon\nabla u_h, z-\varphi_h\rangle_{\partial K}\Big\} + S(u_h)(\varphi_h)\,,\end{aligned}$$

which is equal to the expression (3.45a). □

Remark 3.19 (Nonhomogeneous Dirichlet boundary conditions; cf. [39])**.**
As addressed above we have to reflect on handling nonhomogeneous Dirichlet boundary conditions. The model problem reads

$$-\nabla\cdot(\varepsilon\nabla u)+\boldsymbol{b}\cdot\nabla u+\alpha u=f\quad \text{in } \Omega,\qquad u=g\quad \text{on } \partial\Omega\,,$$

where $g:\partial\Omega\mapsto\mathbb{R}$ is a given function. Additionally, g is supposed to be sufficiently smooth, or to be more exact $g\in\mathcal{H}^{\frac{1}{2}}(\partial\Omega)$ according to the trace theorem 2.10. Then, there exists a lifting u_g of g in $\mathcal{H}^1(\Omega)$

$$\mathcal{H}^1(\Omega)\ni u_g=g \text{ on } \partial\Omega\,,$$

cf. [17]. Enforcing this setting offers the weak formulation

Seek $u\in\mathcal{H}^1(\Omega)$ such that

$$\begin{aligned} u&=u_g+v\,, & v&\in\mathcal{H}^1_0(\Omega)\,,\\ A(v)(\varphi)&=\tilde{F}(\varphi) & \forall\varphi&\in\mathcal{H}^1_0(\Omega)\,, \end{aligned}\tag{3.46}$$

with

$$\tilde{F}(\varphi):=\langle f,\varphi\rangle_\Omega-\langle\varepsilon\nabla u_g,\nabla\varphi\rangle_\Omega-\langle\boldsymbol{b}\cdot\nabla u_g+\alpha u_g,\varphi\rangle_\Omega\,.$$

In the case of a linear target functional $\mathcal{J}(\cdot)$ the dual problem we consider is the following

Find $z\in\mathcal{H}^1_0(\Omega)$ such that

$$A(\zeta)(z)=\mathcal{J}(\zeta)\qquad\forall\zeta\in\mathcal{H}^1_0(\Omega)\,.\tag{3.47}$$

We assume that u_g is approximated by its interpolant g_h. Then, we get for the error in terms of the target quantity by elementary calculus

$$\begin{aligned}\mathcal{J}(e)&=\mathcal{J}\big(u-u_h-(u_g-g_h)\big)=A(e)(z)-A(u_g-g_h)(z)\\ &=\rho(u_h)(z-\varphi_h)-\langle(u_g-g_h),\boldsymbol{n}\cdot\varepsilon\nabla z\rangle_{\partial\Omega}\,,\end{aligned}\tag{3.48}$$

where the residual ρ is given by

$$\rho(u_h)(\cdot)=\tilde{F}(\cdot)-A(u_h)(\cdot)\,.$$

According to the fact that we assume the target functional $\mathcal{J}$ and the form A to be linear, we incorporate nonhomogeneous boundary conditions by adding the

boundary integral in (3.48) to the error representations (3.25a) and (3.45a), respectively.

The presented methods differ in the associated dual problem and resulting from it in the error representation. They are different in the way how to incorporate the SUPG stabilization technique into the DWR method. They vary significantly on the point of perspective: the $\mathcal{FDTS}$ method is based on a more application–related approach as we can indicate the strong form of the dual problem whereas the other strategy $\mathcal{FSTD}$ immediately follows the intention of the theoretical DWR framework. Furthermore, the dual stabilized problem (3.8b) with respect to the $\mathcal{FSTD}$ method is not consistent; that means that the stabilization terms do not vanish for the exact dual solution. The following section will demonstrate what the difference is from a numerical point of view.

3.4 Both methods by numerical comparison

In this section, we present numerical studies based on the above derived approaches. All simulations are performed with the software toolbox FEniCS [37]. Here, we investigate academic test problems whose solutions possess characteristic features of solution of convection–dominated equations.

First, we point out how the DWR method can be combined with the concepts of adaptivity. The following algorithm calculates the approximate solution on a hierarchy of successively refined meshes $\mathcal{M}_i\,, i \geq 1$ and corresponding finite element spaces $\mathcal{V}_h^i$. These finite element spaces are embedded in each other. Assume that the error with respect to a chosen target functional $\mathcal{J}$ can be represented by

$$\mathcal{J}(u) - \mathcal{J}(u_h) \approx \eta := \sum_{K \in \mathcal{T}_h} \eta_K^{\mathcal{FSTD}} \quad \text{and} \quad \mathcal{J}(u) - \mathcal{J}(u_h) \approx \eta := \sum_{K \in \mathcal{T}_h} \eta_K^{\mathcal{FDTS}}\,,$$

respectively, where $\eta_K^{\mathcal{FSTD}}$ and $\eta_K^{\mathcal{FDTS}}$ are defined below.

Adaptive solution algorithm

Initialization Set $i = 0$ and generate the initial finite element spaces.

Step 1 Solve the primal problem.

Find $u_h^i \in \mathcal{V}_h^i$ such that

$$A_S(u_h^i)(\varphi_h) = F(\varphi_h) \quad \forall \varphi_h \in \mathcal{V}_h^i \,.$$

Step 2 Solve the dual problem.

$\mathcal{FSTD}$ method

Find $z_H^i \in \mathcal{V}_H^i \supset \mathcal{V}_h^i$ such that

$$A_S'(u_h^i)(\varphi_H, z_H^i) = \mathcal{J}'(u_h^i)(\varphi_H) \quad \forall \varphi_H \in \mathcal{V}_H^i \,.$$

$\mathcal{FDTS}$ method

Find $z_H^i \in \mathcal{V}_H^i \supset \mathcal{V}_h^i$ such that

$$A_{S^*}(\varphi_H)(z_H^i) = \mathcal{J}'(u_h^i)(\varphi_H) \quad \forall \varphi_H \in \mathcal{V}_H^i \,.$$

$\mathcal{V}_H^i$ denotes the finite element space of higher order polynomials that corresponds to the refined mesh $\mathcal{M}_i$.

Step 3 Evaluate the a–posteriori error estimate.

$\mathcal{FSTD}$ method

$$\begin{aligned} \eta_K^{\mathcal{FSTD}} = {} & \langle \mathcal{R}(u_h^i), z_H^i - \mathcal{I}_h z_H^i \rangle_K + \delta_K \langle \mathcal{R}(u_h^i), \boldsymbol{b} \cdot \nabla (z_H^i - \mathcal{I} z_H^i) \rangle_K \\ & - \langle \mathcal{E}(u_h^i), z_H^i - \mathcal{I}_h z_H^i \rangle_{\partial K} \,, \end{aligned}$$

with the cell and edge residuals defined in (3.25b) and (3.25c).

$\mathcal{FDTS}$ method

$$\begin{aligned}\eta_K^{\mathcal{FDTS}} &= \langle \mathcal{R}(u_h^i), z_H^i - \mathcal{I}_h z_H^i \rangle_K - \delta_K \langle \mathcal{R}(u_h^i), \boldsymbol{b} \cdot \nabla \mathcal{I}_h z_H^i \rangle_K \\ &\quad - \langle \mathcal{E}(u_h^i), z_H^i - \mathcal{I}_h z_H^i \rangle_{\partial K} ,\end{aligned}$$

where the cell and edge residuals are given in (3.45b) and (3.45c). $I_h z_H^i \in \mathcal{V}_h^i$ is the linear interpolation of z_H^i.

Step 4 Histogram based refinement strategy.

Choose $\theta \in (0.25, 5)$. Set $\eta_{max} = \max_{K \in \mathcal{T}_h} \left|\eta_K^j\right| , j \in \{\mathcal{FSTD}, \mathcal{FDTS}\}$ and $\mu = \theta \dfrac{\sum_{K \in \mathcal{T}_h} \left|\eta_K^j\right|}{\#K}$.

```
while mu > eta_max:
    mu := mu/2.0
```

Mark the elements $\tilde{K}$ with $\left|\eta_{\tilde{K}}^j\right| > \mu$ to be refined. Generate a new mesh $\mathcal{M}_{i+1}$ by regular refinement.

Step 5 Check the exit condition.

If $\eta_{max} <$ TOL or $\eta <$ TOL is true, the *Adaptive solution algorithm* is completed; else increase i and go to Step 1.

Remark 3.20 (to Step 3)**.** *Note that setting $z_H = z_h$ would not simplify the error estimator but would cause that the estimator completely vanishes.*

Remark 3.21 (to Step 4)**.** *The performance of adaptive algorithms is enormously influenced by the choice of the marking strategy. Our presented error estimators combined with popular refinement strategies as the Dörfler marking (cf. [15]) or the marking of the elements with the largest local error indicators did not properly select the elements to be refined. We analyzed the distribution of the error indicators over the cells calculated by the error estimator and came up with a histogram based remeshing strategy. In our experience, a value of θ between 0.25 and 5 leads to satisfying results.*

Remark 3.22. *According to the adaptive solution algorithm presented above, we use the same mesh for solving the primal and the dual problem. In our numerical studies, we did not find out characterizing features of the dual solution that require an especially adapted mesh design. Figure 3.1 presents the dual solution and the used mesh obtained by simulating Example 3.23 according to the $\mathcal{FDTS}$ method.*

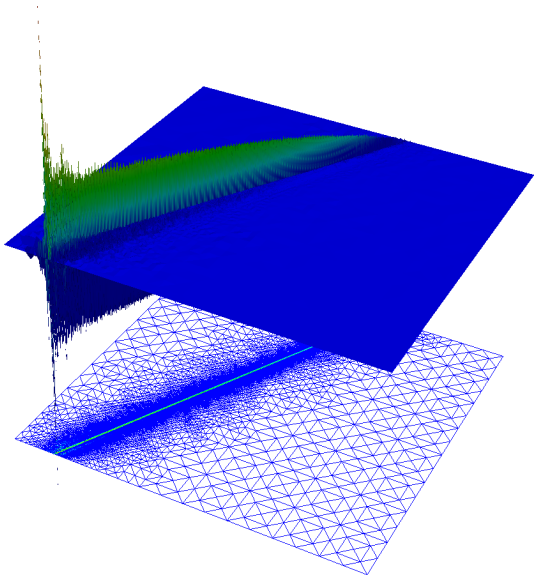


*Figure 3.1: (***Example 3.23***) Dual solution and mesh with 38831 nodes using quadratic finite elements.*

In the case of a given analytical solution, we can make a statement about the adaptive algorithm and its approximate solution in a quantitative manner. A standard measure of quality of an error estimator is the *effectivity index* which is defined by the quotient between the estimated error and the true error

$$\mathcal{I}_{eff} := \frac{\eta}{\mathcal{J}(u) - \mathcal{J}(u_h)} \,.$$

Obviously, the effectivity index of an well–constructed estimator asymptotically tends to one with respect to the number of degrees of freedom. As mentioned above, it is mandatory to use a higher order approximation to the exact dual solution. There are different possibilities to generate such an improved dual solution, e.g. a local higher order approximation obtained by a patch–wise higher order interpolation; cf. [8]. Since we try to develop an error

representation that is as exact as possible, the approximate solution z_h is calculated by polynomials of degree two even if it does not seem very economical to use a global higher order approximation. The simplicity of implementation and, by construction, the linearity of the dual problem let us achieve our goal to conceptionally investigate the interaction of the DWR method and stabilization techniques. Reflecting on issues like computational costs might be the next step in a future work.

Example 3.23. *Now, we focus on a test problem that is an adaption of [38, Example 4.2] and is often used as a benchmark problem for convection–dominated partial differential equations. The setting of this test case is illustrated in Figure 3.2. The solution is characterized by an interior layer of thickness* $\mathcal{O}(\sqrt{\varepsilon}|\ln\varepsilon|)$. *The problem is defined in* $\Omega = (0,1)^2$ *with* $\alpha = 1.0$ *and* $\boldsymbol{b} = \frac{1}{\sqrt{5}}(1,2)^\top$. *We choose the right–hand side* f *in such a way that*

$$u(\boldsymbol{x}) = \frac{1}{2}\left(1 - \tanh\frac{2x_1 - x_2 - 0.25}{\sqrt{5\varepsilon}}\right) \tag{3.49}$$

is the analytical solution of (3.1). The Dirichlet boundary condition is given by the exact solution.

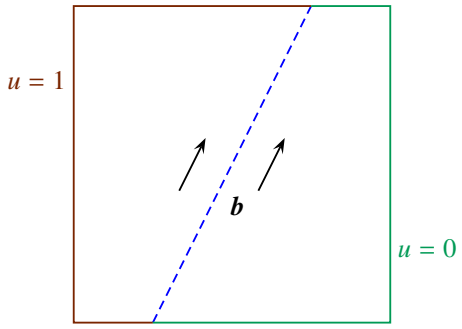


Figure 3.2: Configuration of Example 3.23, with an interior boundary layer (dashed line).

The quantity of interest is prescribed by

$$\mathcal{J}(\varphi) = \frac{1}{\|e\|_{\mathcal{L}^2(\Omega)}}\langle e, \varphi\rangle_\Omega\,,$$

which conforms to the $\mathcal{L}^2$ *norm.*

The Tables 3.1 and 3.2 show selected effectivity indices for different diffusion coefficients. ε is chosen in the range of 10^{-4} to 10^{-7}. The effectivity indices of

all refinement levels are presented in the left column of Figure 3.4.

$\varepsilon = 10^{-4}$		$\varepsilon = 10^{-6}$		$\varepsilon = 10^{-7}$			
dofs	$\mathcal{I}_{eff}$	dofs	$\mathcal{I}_{eff}$	dofs	$\mathcal{I}_{eff}$	dofs	$\mathcal{I}_{eff}$
1564	0.89	17932	0.81	52920	0.69	30905	0.72
2407	0.91	25053	0.83	74662	0.75	45645	1.03
3920	0.93	36391	0.93	106752	0.84	65676	0.88
6665	0.94	54107	0.96	155944	0.88	101112	1.00
12410	0.96	83676	0.98	235139	0.95	158258	0.97
23625	0.96	135009	0.98	367128	0.97	254942	1.00
48922	0.99	224146	0.99	581477	0.99	425945	0.99
$\mathcal{P}_1/\mathcal{P}_2$						$\mathcal{P}_1/\mathcal{P}_3$	

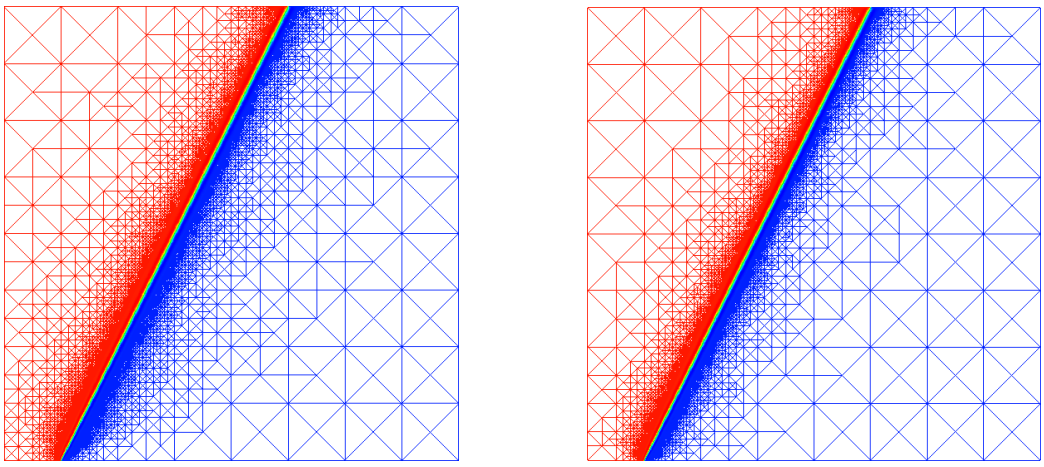
*Table 3.1: (***Example 3.23***) Selected effectivity indices of the $\mathcal{FSTD}$ method.*

$\varepsilon = 10^{-4}$		$\varepsilon = 10^{-6}$		$\varepsilon = 10^{-7}$			
dofs	$\mathcal{I}_{eff}$	dofs	$\mathcal{I}_{eff}$	dofs	$\mathcal{I}_{eff}$	dofs	$\mathcal{I}_{eff}$
1249	0.88	9728	0.92	34787	0.86	36819	0.75
1944	0.96	15191	0.90	53518	0.86	53798	0.87
3168	0.94	24455	0.97	83126	0.96	79820	0.87
5241	0.97	39792	0.96	129500	0.95	119620	0.96
8987	0.95	66142	0.99	207035	0.99	182681	0.96
16677	0.98	110747	0.98	328232	0.98	281116	0.99
31990	0.95	184891	0.99	530014	0.99	438228	0.98
$\mathcal{P}_1/\mathcal{P}_2$						$\mathcal{P}_1/\mathcal{P}_3$	

*Table 3.2: (***Example 3.23***) Selected effectivity indices of the $\mathcal{FDTS}$ method.*

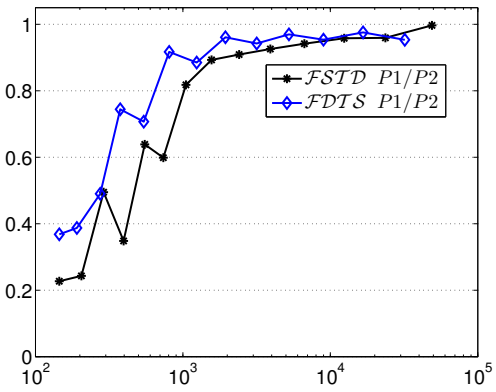
Obviously, the measure of quality tends to one whichever of the two alternative procedures is chosen. The effectivity index is also regardless of the amount of the diffusion coefficient ε. That emphasizes the quality of our method. In the left parts of the Tables 3.1 and 3.2, the computations were performed using linear finite elements for solving the primal problem and quadratic elements for solving the dual problem whereas in the right columns of the Tables 3.1 and 3.2, results using cubic finite elements for solving the dual problem are shown.

We observe that, by choosing cubic elements to solve the dual problem, the adaptive methods reach the stopping criterion including less degrees of freedom than using a quadratic finite element solution. This result is reasonable since a higher order approximation will be closer to the exact solution of the dual problem and, thus, the DWR method will give more accurate results because it is based on the exact solution of the dual problem. Nevertheless, the difference between $\mathcal{P}_2$ and $\mathcal{P}_3$ approximation of the dual solution is not significant, even less if we take into account the higher computational costs. The right part of Figure 3.4 aims at presenting the development of the error in terms of the target quantity $\mathcal{J}$. We notice that both introduced methods are capable to reduce the error in the adaptive process. Figure 3.3 compares the adaptively refined meshes the above mentioned strategies of adaptivity generate for $\varepsilon = 10^{-7}$ and the combination $\mathcal{P}_1/\mathcal{P}_2$ finite elements. They only vary very slightly.



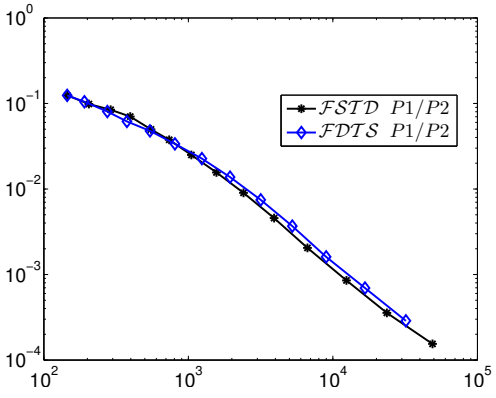
(a) 25992 dofs *(b) 22728 dofs*

*Figure 3.3: (***Example 3.23***) Adaptively refined grids using (a) the $\mathcal{FSTD}$ method and (b) the $\mathcal{FDTS}$ method.*



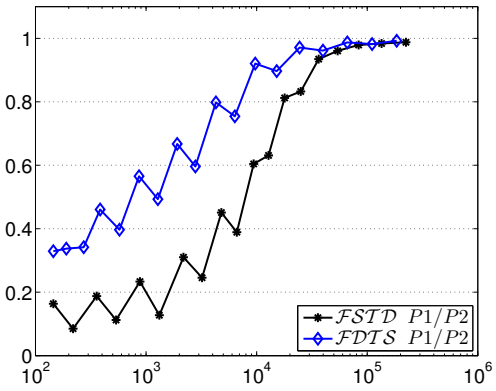
dofs vs. $\mathcal{I}_{\text{eff}}$

(a)



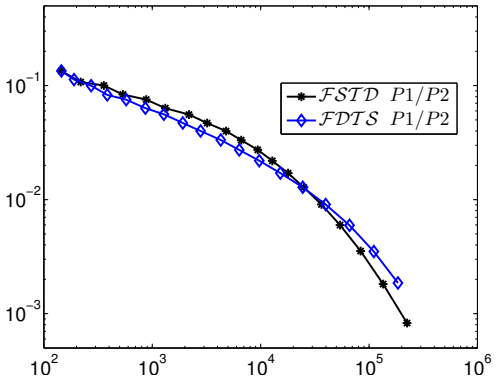
dofs vs. $\mathcal{J}(e)$

(b)



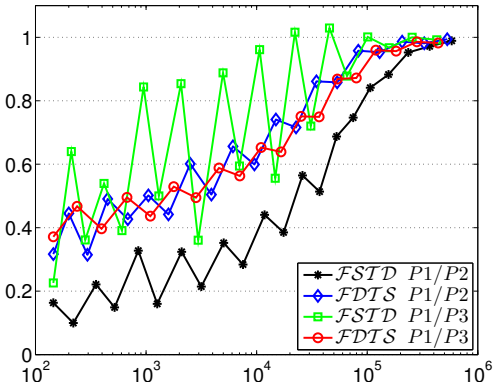
dofs vs. $\mathcal{I}_{\text{eff}}$

(c)



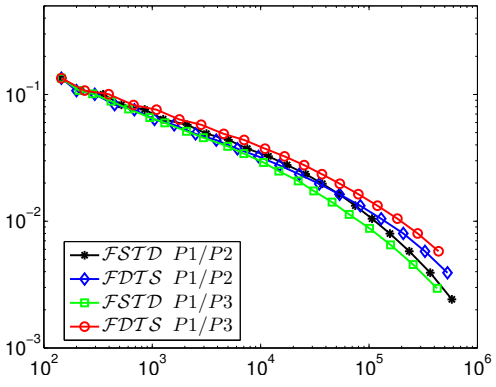
dofs vs. $\mathcal{J}(e)$

(d)



dofs vs. $\mathcal{I}_{\text{eff}}$

(e)



dofs vs. $\mathcal{J}(e)$

(f)

*Figure 3.4: (***Example 3.23***) Effectivity indices for (a) $\varepsilon = 10^{-4}$, (c) $\varepsilon = 10^{-6}$ and (e) $\varepsilon = 10^{-7}$ and errors of fact in terms of the target quantity $\mathcal{J}$ for (b) $\varepsilon = 10^{-4}$, (d) $\varepsilon = 10^{-6}$ and (f) $\varepsilon = 10^{-7}$.*

Chapter 4

A SUPG and SOLD stabilized dual weighted residual method

In contrast to the previous chapter, we now address a nonlinear convection–dominated model problem. We also pay attention to the remaining nonphysical oscillations the SUPG method cannot completely handle on its own. To this end and according to [26], we introduce a so–called spurious oscillations at layers diminishing (SOLD) method which adds terms to the SUPG discretization in order to obtain discrete solutions in which the local oscillations are suppressed.

4.1 A nonlinear framework

We consider a nonlinear adaption of our model problem in Chapter 3 and obtain the stationary scalar convection–diffusion–reaction equation

$$-\nabla \cdot (\varepsilon \nabla u) + \boldsymbol{b} \cdot \nabla u + \alpha u + r(u) = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega\,, \tag{4.1}$$

equipped with homogeneous boundary conditions in order to facilitate matters. We assume that Ω is a bounded Lipschitz domain in $\mathbb{R}^d\,, d \in \{2, 3\}$ and the conditions in (3.2) are valid. The reaction rate $r(\cdot)$ is supposed to be three times differentiable and represents a polynomial reaction rate or the Arrhenius law, for example. In order to ensure the existence of a weak solution of (4.1), we refer to [42] concerning necessary assumptions about the reaction rate.

We introduce the finite element space $\mathcal{V}_h$ as laid down in (2.3). The usual Galerkin finite element formulation of the given model problem (4.1) reads

Find $u_h \in \mathcal{V}_h$ such that

$$A(u_h)(\varphi_h) = F(\varphi_h) \qquad \forall \varphi_h \in \mathcal{V}_h\,, \tag{4.2a}$$

with

$$A(v_h)(\psi_h) := \langle \varepsilon \nabla v_h, \nabla \psi_h \rangle_\Omega + \langle \boldsymbol{b} \cdot \nabla v_h, \psi_h \rangle_\Omega + \langle \alpha v_h, \psi_h \rangle_\Omega + \langle r(v_h), \psi_h \rangle_\Omega\,, \tag{4.2b}$$

$$F(\psi_h) := \langle f, \psi_h \rangle_\Omega\,. \tag{4.2c}$$

It is well–known that this discrete formulation is unsuitable if convection dominates diffusion since then the discrete solution is usually stressed by spurious oscillations. Adding a SUPG stabilization term to the Galerkin discretization establishes an improvement; cf. [10]. This stabilization procedure leads to the following formulation

Find $u_h \in \mathcal{V}_h$ such that

$$A_S(u_h)(\varphi_h) = F(\varphi_h) \qquad \forall \varphi_h \in \mathcal{V}_h\,, \tag{4.3}$$

with

$$\begin{aligned} A_S(v_h)(\psi_h) &:= A(v_h)(\psi_h) + S(v_h)(\psi_h)\,, \\ S(v_h)(\psi_h) &:= \sum_{K \in \mathcal{T}_h} \delta_K \langle R(v_h), \boldsymbol{b} \cdot \nabla \psi_h \rangle_K\,, \\ R(v_h) &:= -\nabla \cdot (\varepsilon \nabla v_h) + \boldsymbol{b} \cdot \nabla v_h + \alpha v_h + r(v_h) - f\,. \end{aligned}$$

The shock–capturing method proposed in [33] adds a crosswind diffusion term to the left–hand side of (4.3). The amount of the artificial anisotropic diffusion depends on the unknown discrete solution u_h. Thus, the resulting approach is nonlinear. Since our model problem is nonlinear by itself, the arising nonlinearity of the improved discrete formulation does not disturb. The SUPG and shock–capturing stabilized method has the form

Find $u_h \in \mathcal{V}_h$ such that

$$A_{SC}(u_h)(\varphi_h) = F(\varphi_h) \qquad \forall \varphi_h \in \mathcal{V}_h\,, \tag{4.4}$$

with

$$A_{SC}(v_h)(\psi_h) := A(v_h)(\psi_h) + S(v_h)(\psi_h) + S_C(v_h)(\psi_h)\,, \tag{4.5a}$$

$$S_C(v_h)(\psi_h) := \sum_{K\in\mathcal{T}_h} \langle \tau_K(v_h)\boldsymbol{D}\nabla v_h, \nabla\psi_h\rangle_K\,, \tag{4.5b}$$

$$\tau_K(v_h) := l_K(v_h)\hat{R}_K(v_h) = \frac{l_K(v_h)\|R(v_h)\|_{\mathcal{L}^2(K)}}{\|v_h\|_{\mathcal{H}^1(K)} + \kappa_K}\,, \tag{4.5c}$$

$$l_K(v_h) := l_0 h_K \max\left\{0, \beta - \frac{2\|\varepsilon\|_{\mathcal{L}^\infty(K)}}{h_K\hat{R}_K(v_h)}\right\}\,, \tag{4.5d}$$

$$\boldsymbol{D} := \begin{cases} \boldsymbol{I} - \dfrac{\boldsymbol{b}\otimes\boldsymbol{b}}{|\boldsymbol{b}|^2}\,, & \boldsymbol{b}\neq\boldsymbol{0} \\ \boldsymbol{0}\,, & \boldsymbol{b}=\boldsymbol{0}\,, \end{cases} \tag{4.5e}$$

and nonnegative parameters κ_K, β and l_0 to be chosen by the user.

Remark 4.1 (Consistency of the stabilization method)**.** *The SUPG method as well as the shock–capturing stabilization technique are consistent as the stabilization terms completely vanish by substituting the exact solution into the stabilized discrete formulation (4.4).*

Having defined the discretization (4.4) of our model problem, there arises the question which one of the in Chapter 3 introduced strategies to pursue. Reflecting on the main steps of both methods, we reach the conclusion that the $\mathcal{FDTS}$ method is the procedure of choice that supplies the basis of our further work. The main reason behind this decision is that, by construction, we have to approximate the derivative of the shock–capturing term in the course of the $\mathcal{FSTD}$ method. This involves the danger that the quality of the resulting error estimator decreases.

4.2 A nonlinear $\mathcal{FDTS}$ method

As pointed out above, we follow the $\mathcal{FDTS}$ method presented in Chapter 3. Thereby, we $\mathcal{F}$irst take the associated $\mathcal{D}$ual problem of the continuous formulation of equation (4.2a), $\mathcal{T}$hen we add the SUPG and $\mathcal{S}$tabilization terms to the resulting equation. Since the dual problem is characterized by its linearity even if the primal problem is nonlinear, the dual problem is stabilized by SUPG terms only.

To this end, we define the Lagrangian functional

$$\mathfrak{L}(u,z) := \mathcal{J}(u) + F(z) - A(u)(z)\,, \tag{4.6}$$

with the quantity of interest $\mathcal{J}$, and F and A given in (4.2b) and (4.2c), respectively. We assume that $\mathcal{J}(\cdot)$ is differentiable and $\mathcal{J}'(\cdot)(\varphi) = \langle j(\cdot), \varphi\rangle_\Omega$ with $j \in \mathcal{L}^2(\Omega)$. The derivative of the Lagrangian functional with respect to u (4.6) offers the dual problem that corresponds to the continuous version of equation (4.2a)

Find $z \in \mathcal{V}$ such that

$$A'(u)(\zeta, z) = \mathcal{J}'(u)(\zeta) \qquad \forall \zeta \in \mathcal{V}\,, \tag{4.7}$$

with

$$A'(v)(\xi, w) := \langle \varepsilon \nabla w, \nabla \xi\rangle_\Omega - \langle \boldsymbol{b}\cdot\nabla w, \xi\rangle_\Omega + \langle \alpha w, \xi\rangle_\Omega + \langle r'(v)w, \xi\rangle_\Omega\,.$$

Then, the discrete and stabilized Euler–Lagrange system reads

Seek solutions $\{u_h, z_h\} \in \mathcal{V}_h \times \mathcal{V}_h$ such that

$$A_{SC}(u_h)(\varphi_h) = F(\varphi_h) \qquad \forall \varphi_h \in \mathcal{V}_h\,, \tag{4.8a}$$

$$A'_{S^*}(u_h)(\zeta_h, z_h) = \mathcal{J}'(u_h)(\zeta_h) \qquad \forall \zeta_h \in \mathcal{V}_h\,, \tag{4.8b}$$

where $A_{SC}(v_h)(\psi_h)$ is given in the scheme (4.5a) to (4.5e) and the dual form is defined by

$$\begin{aligned} A'_{S^*}(v_h)(\xi_h, w_h) &:= A'(v_h)(\xi_h, w_h) + S^*(v_h)(\xi_h, w_h)\,,\\ S^*(v_h)(\xi_h, w_h) &:= \sum_{K\in\mathcal{T}_h} \delta_K^* \langle \nabla\cdot(\varepsilon\nabla w_h) + \boldsymbol{b}\cdot\nabla w_h - \alpha w_h \\ &\qquad -r'(v_h)w_h + j, \boldsymbol{b}\cdot\nabla\xi_h\rangle_K\,. \end{aligned}$$

In the following, we are going to establish an error representation based on the dual weighted residual method. Therefore, we suppose $x \in \mathcal{X}$ to be a point that fulfills the stationary property (3.33), that is

$$\mathfrak{L}'(x)(y) = 0 \quad \forall y \in \mathcal{X}\,.$$

We add together the formulations (4.8a) and (4.8b)

$$F(\varphi_h) - A_{SC}(u_h)(\varphi_h) + \mathcal{J}'(u_h)(\zeta_h) - A'_{S^*}(u_h)(\zeta_h, z_h) \\ = 0 \quad \forall \{\zeta_h, \varphi_h\} \in \mathcal{V}_h \times \mathcal{V}_h \,,$$

and get that

$$\mathfrak{L}'_u(u_h, z_h)(\zeta_h) + \mathfrak{L}'_z(u_h, z_h)(\varphi_h) - S(u_h)(\varphi_h) \\ - S_C(u_h)(\varphi_h) - S^*(u_h)(\zeta_h, z_h) = 0 \quad \forall \{\zeta_h, \varphi_h\} \in \mathcal{V}_h \times \mathcal{V}_h \,.$$

Setting

$$x_h := \{u_h, z_h\} \in \mathcal{V}_h \times \mathcal{V}_h =: \mathcal{X}_h \subset \mathcal{X} \,, \\ y_h := \{\zeta_h, \varphi_h\} \in \mathcal{V}_h \times \mathcal{V}_h \,,$$

and

$$\mathcal{S}(x_h)(y_h) := S(u_h)(\varphi_h) + S_C(u_h)(\varphi_h) + S^*(u_h)(\zeta_h, z_h) \,,$$

we find that

$$\mathfrak{L}'(x_h)(y_h) = \mathcal{S}(x_h)(y_h) \quad \forall y_h \in \mathcal{X}_h \,. \tag{4.9}$$

Now that we have created the framework, we can state the following theorem. The approach is based on [7].

Theorem 4.2 (Error representation in terms of the Lagrangian functional)**.** *Suppose $x \in \mathcal{X}$ to be a stationary point of $\mathfrak{L}$ and $x_h \in \mathcal{X}_h$ to satisfy equation (4.9). Then, the error representation reads*

$$\mathfrak{L}(x) - \mathfrak{L}(x_h) = \mathfrak{L}'(x_h)(x - y_h) + \mathcal{S}(x_h)(y_h - x_h) + R \,, \tag{4.10a}$$

with an arbitrary $y_h \in \mathcal{X}_h$. The remainder term is defined by

$$R := \frac{1}{2} \int_0^1 \mathfrak{L}'''(x_h + s\hat{e})(\hat{e}, \hat{e}, \hat{e}) s(s-1) \, ds \,. \tag{4.10b}$$

Proof. The proof follows the proof of Theorem 3.15. □

For the following theorem, we recall the definitions of the residual $\rho(u_h)(\cdot)$ and the adjoint residual $\rho^*(z_h)(\cdot)$ (3.36a) and (3.36b), respectively,

$$\rho(u_h)(\varphi) = F(\varphi) - A(u_h)(\varphi) \qquad \forall \varphi \in \mathcal{V} \,, \\ \rho^*(z_h)(\zeta) = \mathcal{J}'(u_h)(\zeta) - A'(u_h)(\zeta, z_h) \qquad \forall \zeta \in \mathcal{V} \,.$$

Theorem 4.3 (Error representation with respect to the target quantity $\mathcal{J}$)**.** *For the error representation in terms of the functional $\mathcal{J}(\cdot)$, there holds that*

$$\begin{aligned}\mathcal{J}(u) - \mathcal{J}(u_h) = \frac{1}{2}\rho(u_h)(z-\varphi_h) + \frac{1}{2}\rho^*(z_h)(u-\zeta_h) \\ + R_{\mathcal{S}} + R_{\mathrm{nl}}\,,\end{aligned} \tag{4.11a}$$

with arbitrary $\varphi_h, \zeta_h \in \mathcal{V}_h$ where the residual $\rho(u_h)(\cdot)$ and the adjoint residual $\rho^*(z_h)(\cdot)$ are defined as above. The remainder terms are given by

$$R_{\mathcal{S}} := \frac{1}{2}S(u_h)(\varphi_h + z_h) + \frac{1}{2}S_C(u_h)(\varphi_h + z_h) + \frac{1}{2}S^*(u_h)(\zeta_h - u_h, z_h)\,, \tag{4.11b}$$

and

$$\begin{aligned}R_{\mathrm{nl}} := \frac{1}{2}\int_0^1 \Big\{ \mathcal{J}'''(u_h + se)(e,e,e) - \big\langle r'''(u_h+se)e^3, z_h + se^*\big\rangle_\Omega \\ -3\big\langle r''(u_h+se)e^2, e^*\big\rangle_\Omega \Big\} s(s-1)\,ds\,.\end{aligned} \tag{4.11c}$$

The terms (4.11b) and (4.11c) are caused by the stabilization terms of the system (4.8a), (4.8b), and by the nonlinear property of the target quantity $\mathcal{J}$ and the original problem, respectively.

Proof. By elementary calculus, we have the identities

$$\begin{aligned}\mathfrak{L}(x) &= \mathcal{J}(u) + F(z) - A(u)(z) = \mathcal{J}(u)\,,\\ \mathfrak{L}(x_h) &= \mathcal{J}(u_h) + F(z_h) - A(u_h)(z_h) = \mathcal{J}(u_h) + S(u_h)(z_h) + S_C(u_h)(z_h)\,.\end{aligned}$$

Consequently, we find a formulation for the error in the target quantity

$$\mathcal{J}(u) - \mathcal{J}(u_h) = \mathfrak{L}(x) - \mathfrak{L}(x_h) + S(u_h)(z_h) + S_C(u_h)(z_h)\,.$$

From the previous theorem, we conclude that

$$\begin{aligned}\mathcal{J}(u) - \mathcal{J}(u_h) = \frac{1}{2}\mathfrak{L}'(x_h)(x-y_h) + \frac{1}{2}\mathcal{S}(x_h)(y_h - x_h) + S(u_h)(z_h) \\ + S_C(u_h)(z_h) + R\,,\end{aligned}$$

with R defined in (4.10b). All terms of the third derivative of $\mathfrak{L}$ vanish except the three terms given in (4.11c). The Fréchet derivative of the Lagrangian functional results in

$$\begin{aligned}\mathfrak{L}'(x_h)(x-y_h) &= \mathfrak{L}'_u(u_h,z_h)(u-\zeta_h) + \mathfrak{L}'_z(u_h,z_h)(z-\varphi_h)\\ &= \mathcal{J}'(u_h)(u-\zeta_h) - A'(u_h)(u-\zeta_h, z_h) + F(z-\varphi_h)\\ &\quad -A(u_h)(z-\varphi_h)\\ &= \rho^*(z_h)(u-\zeta_h) + \rho(u_h)(z-\varphi_h)\,.\end{aligned}$$

For the stabilization terms, it holds that

$$\begin{aligned}
&\frac{1}{2}\mathcal{S}(x_h)(y_h - x_h) + S(u_h)(z_h) + S_C(u_h)(z_h) \\
&\quad = \frac{1}{2}\Big\{S(u_h)(\varphi_h - z_h) + S_C(u_h)(\varphi_h - z_h) + S^*(u_h)(\zeta_h - u_h, z_h)\Big\} \\
&\quad\quad + S(u_h)(z_h) + S_C(u_h)(z_h) \\
&\quad = \frac{1}{2}S(u_h)(\varphi_h + z_h) + \frac{1}{2}S_C(u_h)(\varphi_h + z_h) + \frac{1}{2}S^*(u_h)(\zeta_h - u_h, z_h) =: R_{\mathcal{S}}\,,
\end{aligned}$$

which proves the assertion. □

The error representation in (4.11a) would require the generation of improved approximations to z as well as to u as explained in Remark 3.7. Therefore, we derive a relation between the primal and the dual residual in order to remove the dependence on u. We observe that the primal and dual residuals coincide in the case of an unstabilized linear differential equation and a linearly chosen target quantity. In our case, there is a difference between the residuals shown by the following theorem.

Theorem 4.4. *The relation between the residual $\rho(u_h)(\cdot)$ and the adjoint residual $\rho^*(z_h)(\cdot)$ is described by*

$$\rho^*(z_h)(u - \zeta_h) = \rho(u_h)(z - \varphi_h) + \Delta\rho_{\mathrm{nl}} + \Delta\rho_{\mathcal{S}}\,, \tag{4.12a}$$

with arbitrary $\varphi_h, \zeta_h \in \mathcal{V}_h$. The nonlinear character of the target quantity $\mathcal{J}$ as well as the nonlinearity of the original problem itself induce the quadratic remainder term

$$\Delta\rho_{\mathrm{nl}} := \int_0^1 \Big\{A''(u_h + se)(e, e, z_h + se^*) - \mathcal{J}''(u_h + se)(e, e)\Big\}\, ds\,, \tag{4.12b}$$

and the stabilization of the discrete system (4.8a), (4.8b) induces the remainder term

$$\Delta\rho_{\mathcal{S}} := S(u_h)(\varphi_h - z_h) + S_C(u_h)(\varphi_h - z_h) - S^*(u_h)(\zeta_h - u_h, z_h)\,. \tag{4.12c}$$

Proof. We introduce

$$k(s) := \mathcal{J}'(u_h + se)(u - \zeta_h) - A'(u_h + se)(u - \zeta_h, z_h + se^*)\,,$$

together with its derivative

$$\begin{aligned} k'(s) :&= \mathcal{J}''(u_h + se)(e, u - \zeta_h) - A''(u_h + se)(e, u - \zeta_h, z_h + se^*) \\ &\quad - A'(u_h + se)(u - \zeta_h, e^*)\,. \end{aligned}$$

Due to the definition of z, it holds that

$$k(1) = \mathcal{J}'(u)(u - \zeta_h) - A'(u)(u - \zeta_h, z) = 0 \quad \forall \zeta_h \in \mathcal{V}_h\,.$$

Further, we find that

$$k(0) = \mathcal{J}'(u_h)(u - \zeta_h) - A'(u_h)(u - \zeta_h, z_h) = \rho^*(z_h)(u - \zeta_h)\,.$$

Hence, from the fundamental theorem of calculus, we can conclude that

$$\begin{aligned} \rho^*(z_h)(u - \zeta_h) = -\int_0^1 \Big\{ &\mathcal{J}''(u_h + se)(e, u - \zeta_h) \\ &- A''(u_h + se)(e, u - \zeta_h, z_h + se^*) \Big\}\, ds \\ &+ \int_0^1 A'(u_h + se)(u - \zeta_h, e^*)\, ds\,, \end{aligned} \tag{4.13}$$

with an arbitrary $\zeta_h \in \mathcal{V}_h$. Substituting $u - \zeta_h$ by e in the dual residual yields that

$$\begin{aligned} &\rho^*(z_h)(u - \zeta_h) \\ &\quad = \mathcal{J}'(u_h)(u - \zeta_h) - A'(u_h)(u - \zeta_h, z_h) + S^*(u_h)(\zeta_h, z_h) \\ &\qquad - S^*(u_h)(\zeta_h, z_h) - \mathcal{J}'(u_h)(u_h) + A'(u_h)(u_h, z_h) + S^*(u_h)(u_h, z_h) \\ &\quad = \rho^*(z_h)(u - u_h) - S^*(u_h)(\zeta_h - u_h, z_h)\,. \end{aligned}$$

By applying formulation (4.13) to the dual residual $\rho^*(z_h)(e)$, we get that

$$\begin{aligned} \rho^*(z_h)(u - u_h) = &\int_0^1 \Big\{ A''(u_h + se)(e, e, z_h + se^*) - \mathcal{J}''(u_h + se)(e, e) \Big\}\, ds \\ &+ \int_0^1 A'(u_h + se)(e, e^*)\, ds\,, \end{aligned}$$

The first integral corresponds to the remainder term $\Delta\rho_{\mathrm{nl}}$ defined in (4.12b). For the last term, we obtain that

$$\int_0^1 A'(u_h + se)(e, e^*)\, ds = A(u)(e^*) - A(u_h)(e^*) = F(e^*) - A(u_h)(e^*)\,.$$

Substituting e^* by $z - \varphi_h$ with an arbitrary $\varphi_h \in \mathcal{V}_h$ by using the relation $A_{SC}(u_h)(\varphi_h) - F(\varphi_h) + S(u_h)(z_h) - S(u_h)(z_h) = 0$, we observe that

$$\int_0^1 A'(u_h + se)(e, e^*)\, ds$$
$$= F(z - \varphi_h) - A(u_h)(z - \varphi_h) + S(u_h)(\varphi_h - z_h) + S_C(u_h)(\varphi_h - z_h)\,.$$

All together, it gives the following identity

$$\rho^*(u_h)(u - \zeta_h) = \rho(u_h)(z - \varphi_h) + \Delta\rho_{\mathrm{nl}} + \Delta\rho_{\mathcal{S}}\,,$$

with $\Delta\rho_{\mathcal{S}}$ given in (4.12c). This completes the proof. □

Based on the previous theorems, we now derive a cell–wise error representation.

Theorem 4.5 (Local error description for the nonlinear $\mathcal{FDTS}$ method)**.** *For the stabilized finite element approximation of the model problem (4.1), we have the element–wise error representation*

$$\begin{aligned}\mathcal{J}(u) - \mathcal{J}(u_h) = \sum_{K\in\mathcal{T}_h} \Big\{ &\langle \mathcal{R}(u_h), z - \varphi_h\rangle_K - \delta_K \langle \mathcal{R}(u_h), \boldsymbol{b}\cdot\nabla\varphi_h\rangle_K \\ &+ S_C(u_h)(\varphi_h) - \langle \mathcal{E}(u_h), z - \varphi_h\rangle_{\partial K}\Big\}\,.\end{aligned} \tag{4.14a}$$

The cell and edge residuals take the form

$$\mathcal{R}(u_h)_{|K} = f + \nabla\cdot(\varepsilon\nabla u_h) - \boldsymbol{b}\cdot\nabla u_h - \alpha u_h - r(u_h)\,, \tag{4.14b}$$

$$\mathcal{E}(u_h)_{|\Gamma} = \begin{cases} \frac{1}{2}\boldsymbol{n}\cdot[\varepsilon\nabla u_h] & \text{if } \Gamma \subset \partial K\backslash\partial\Omega \\ 0 & \text{if } \Gamma \subset \partial\Omega\,, \end{cases} \tag{4.14c}$$

where $\boldsymbol{n}$ denotes the outer–pointing normal and $[\nabla u_h]$ defines the jump of ∇u_h over the inner edges Γ.

Proof. From the Theorems 4.3 and 4.4, we conclude that

$$\mathcal{J}(u) - \mathcal{J}(u_h) = \rho(u_h)(z - \varphi_h) + \frac{1}{2}\Delta\rho_{\mathrm{nl}} + \frac{1}{2}\Delta\rho_{\mathcal{S}} + R_{\mathrm{nl}} + R_{\mathcal{S}}\,.$$

Since the parts $\Delta\rho_{\mathrm{nl}}$ defined in (4.12b) and R_{nl} given in (4.11c) are quadratic and cubic in e and e^*, respectively, they are neglectable. The remainder terms resulting from the stabilization technique are rearranged such that

$$\frac{1}{2}\Delta\rho_{\mathcal{S}} + R_{\mathcal{S}} = S(u_h)(\varphi_h) + S_C(u_h)(\varphi_h)\,.$$

Applying integration by parts to the residual confirms the assertion (4.14a).

□

4.3 Numerical studies for the nonlinear $\mathcal{FDTS}$ method

In this section, we present an assessment for our above developed nonlinear $\mathcal{FDTS}$ approach. Extensive tests are performed including different quantities of interest and various appropriate convection–dominated test examples. The adaptive process is controlled by the established error representation in terms of a chosen target quantity $\mathcal{J}$

$$\mathcal{J}(u) - \mathcal{J}(u_h) \approx \eta := \sum_{K \in \mathcal{T}_h} \eta_K \,.$$

We modify the *Adaptive solution algorithm* presented in section 3.4 for obtaining a hierarchy of sequently refined meshes $\mathcal{M}_i$, $i \geq 1$ and corresponding finite element spaces $\mathcal{V}_h^i$ in the view of the nonlinear problem.

Adaptive solution algorithm

Initialization Set $i = 0$ and generate the initial finite element spaces.

Step 1 Solve the primal problem.

Find $u_h^i \in \mathcal{V}_h^i$ such that

$$A_{SC}(u_h^i)(\varphi_h) = F(\varphi_h) \quad \forall \varphi_h \in \mathcal{V}_h^i \,.$$

Step 2 Solve the dual problem.

Find $z_H^i \in \mathcal{V}_H^i \supset \mathcal{V}_h^i$ such that

$$A'_{S^*}(u_h^i)(\varphi_H, z_H^i) = \mathcal{J}'(u_h^i)(\varphi_H) \quad \forall \varphi_H \in \mathcal{V}_H^i \,.$$

$\mathcal{V}_H^i$ is the finite element space of higher order polynomials that corresponds to the refined mesh $\mathcal{M}_i$.

Step 3 Evaluate the a–posteriori error estimate.

$$\eta_K = \langle \mathcal{R}(u_h^i), z_H^i - \mathcal{I}_h z_H^i \rangle_K - \delta_K \langle \mathcal{R}(u_h^i), \boldsymbol{b} \cdot \nabla \mathcal{I}_h z_H^i \rangle_K + \langle \tau_K(u_h^i) \boldsymbol{D} \nabla u_h, \nabla \mathcal{I}_h z_H^i \rangle_K - \langle \mathcal{E}(u_h^i), z_H^i - \mathcal{I}_h z_H^i \rangle_{\partial K} ,$$

where the cell and edge residuals are given in (4.14b) and (4.14c). $\boldsymbol{D}$ and τ_K are defined in the shock–capturing scheme (4.5a) – (4.5e). $I_h z_H^i \in \mathcal{V}_h^i$ is the linear interpolation of z_H^i.

Step 4 Histogram based refinement strategy.

Choose $\theta \in (0.25, 5)$. Set $\eta_{max} = \max_{K \in \mathcal{T}_h} |\eta_K|$, and $\mu = \theta \dfrac{\sum_{K \in \mathcal{T}_h} |\eta_K|}{\#K}$.

```
while mu > eta_max:
    mu := mu/2.0
```

Mark the elements $\tilde{K}$ with $|\eta_{\tilde{K}}| > \mu$ to be refined. Generate a new mesh $\mathcal{M}_{i+1}$ by regular refinement.

Step 5 Check the exit condition.

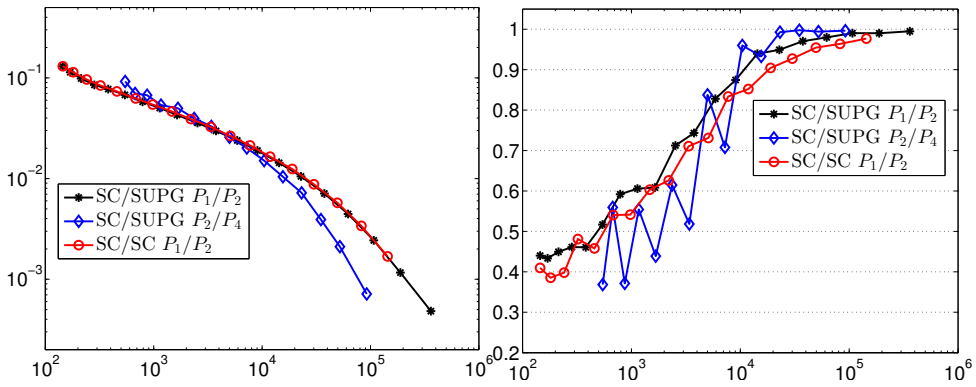
If $\eta_{max} <$ TOL or $\eta <$ TOL is true, the *Adaptive solution algorithm* is completed; else increase i and go to Step 1.

Now, we present the results of the adaptive algorithm to several test cases. As pointed out in section 3.4, we focus on the concept to what extent the DWR method and stabilization techniques interact. In our implementation, our priority is not efficiency but accuracy. The quality of our developed error representation is readily apparent from the *effectivity index* $\mathcal{I}_{\text{eff}}$, i.e. the ratio of the estimated error and the actual error. If these values nearly coincide, the effectivity index is close to one. To prevent side effects on the effectivity index caused for example by patch–wise higher order interpolation of the dual solution (as described in [8]), we use a global higher order approximation.

First of all, we want to justify our suggested approach to stabilize the linear dual problem by SUPG terms whereas the nonlinear primal problem is solved as shock–capturing formulation. For this purpose, we address ourselves to

Example 3.23 already introduced in section 3.4. All chosen parameters are maintained except for the diffusion coefficient that we set $\varepsilon = 10^{-6}$. The nonlinear reaction rate is set to a polynomial of degree two, $r(u) = u^2$. Figure 4.1 aims at showing the errors of fact $\mathcal{J}(e)$ of three different solution strategies. The chosen target quantity is the $\mathcal{L}^2$ norm, i.e.

$$\mathcal{J}(u) = \frac{1}{\|u\|_{\mathcal{L}^2(\Omega)}} \langle e, u \rangle_\Omega \,. \tag{4.15}$$



dofs vs. $\mathcal{J}(e)$ *(a)*

dofs vs. $\mathcal{I}_{\text{eff}}$ *(b)*

*Figure 4.1: (***Example 3.23***) (a) Errors $\mathcal{J}(e)$ and (b) effectivity indices $\mathcal{I}_{\text{eff}}$ for stabilization technique of the primal/dual problem with polynomial degree for solving the primal/dual problem – SC is the abbreviation of shock–capturing stabilization, SUPG means SUPG stabilization only.*

Figure 4.1 reinforces our strategy to take advantage of the linear character of the dual problem by not adding shock–capturing stabilization to the adjoint problem. As anticipated, the higher order approximation $\mathcal{P}_2/\mathcal{P}_4$ is able to reduce the error faster than a linear–quadratic finite element composition. One point worthy of note is the computational time that enormously increases when solving a nonlinear dual problem as well as using higher order finite elements.

Up to now, a residual based error estimator, as proposed in [48] for instance, would not have created results that are much different from the above presented ones. Using our method, we are able to not only focus on error bounds with

respect to the $\mathcal{L}^2$ norm but especially on error estimates in terms of physically relevant parameters. Possible quantities of interest are

$$\mathcal{J}_1(u) = \int_\Omega u \, d\boldsymbol{x} \quad \text{or} \quad \mathcal{J}_2(u) = u\,(\boldsymbol{x}_e)\ ,$$

with a chosen point $\boldsymbol{x}_e = \left(\frac{3}{16}, \frac{1}{8}\right)$. For reasons for the existence of a continuous solution, we have to regularize the functional $\mathcal{J}_2(u)$, for instance by setting

$$\mathcal{J}_r(u) = \frac{1}{|B_r|} \int_{B_r} u(\boldsymbol{x})\, d\boldsymbol{x}\ ,$$

where $B_r = \{\boldsymbol{x} \in \Omega |\, |\boldsymbol{x} - \boldsymbol{x}_e| < r\}$ with a small radius r. The adaptive solution process is guided by the error representation (4.14a).

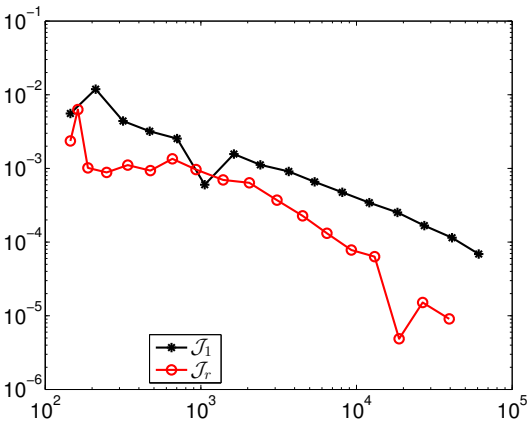


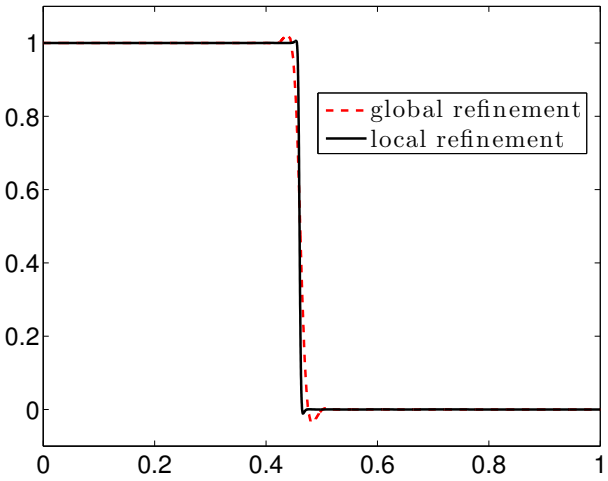
*Figure 4.2: (***Example 3.23***) Trends of errors in terms of the quantities of interest $\mathcal{J}_1$ and $\mathcal{J}_r$.*

$\mathcal{J}_1$		$\mathcal{J}_r$	
dofs	$\mathcal{I}_{eff}$	$\mathcal{I}_{eff}$	dofs
5383	0.45	0.03	4505
8105	0.44	0.07	6458
12081	0.45	0.14	9268
18321	0.57	0.22	13079
27276	0.71	3.42	18794
41073	0.76	1.25	26619
60957	0.83	2.27	39447

*Table 4.1: (***Example 3.23***) Effectivity indices with respect to the target quantities $\mathcal{J}_1$ and $\mathcal{J}_r$.*

The results in Figure 4.2 show that both investigated quantities of interest are up to controlling the adaptive solution process and, hence, reducing the errors of fact. In our simulation, the estimator slightly underestimates the exact error $\mathcal{J}_1(u-u_h)$ according to Table 4.1. Nevertheless, the values for the effectivity indices are satisfying as they tend to one, and even more, if we regard Figure 4.3.

The results of two simulations of Example 3.23 are presented in Figure 4.3. One solution is obtained by global refinement of the mesh, the comparative solution is obtained by adaptive refinement where the process of adaptivity is



Line plot

Figure 4.3: (**Example 3.23**) 27276 *adaptively refined nodes and* 33025 *uniform refined nodes in comparison. For a detailed view, see Figure 4.4.*

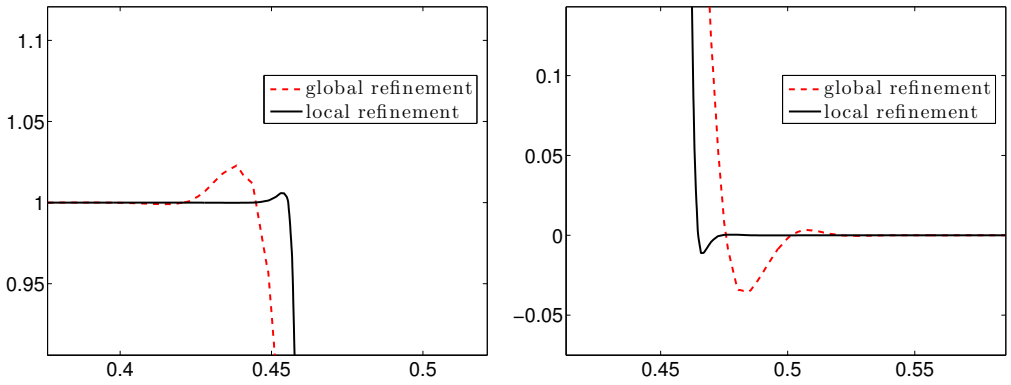
Global mesh refinement

dofs	$\|e\|_{\mathcal{L}^2(\Omega)}$	$\mathcal{I}_{eff}$
8321	0.056	0.58
33025	0.040	0.67

Local mesh refinement

dofs	$\mathcal{J}(e)$	$\mathcal{I}_{eff}$
22824	0.011	0.95
37555	0.007	0.97

Table 4.2: (**Example 3.23**) *Global and local mesh refinement in numbers.*



Upper edge

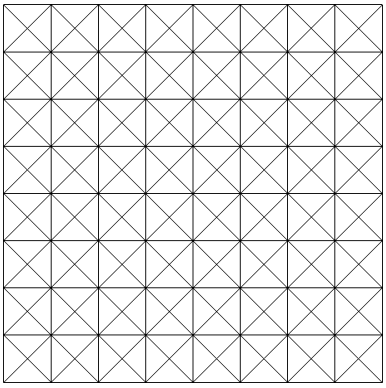
(a)

Lower edge

(b)

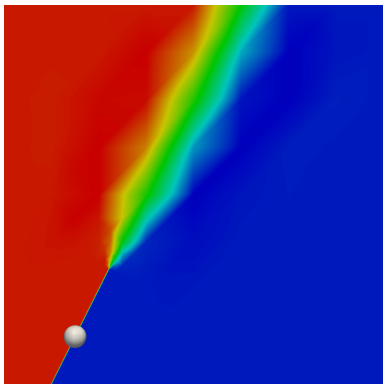
Figure 4.4: (**Example 3.23**) *Detailed view of the significant areas (a) on the left and (b) on the right of the layer.*

controlled by the quantity of interest $\mathcal{J}_1$. To ensure the comparable information, the adaptively refined mesh consists of 27276 nodes whereas the uniformly refined mesh consists of 33025 nodes. Observing the line plots in Figure 4.3, we notice that the undesirable over– and undershoots are considerably reduced



Initial triangular grid

(a)



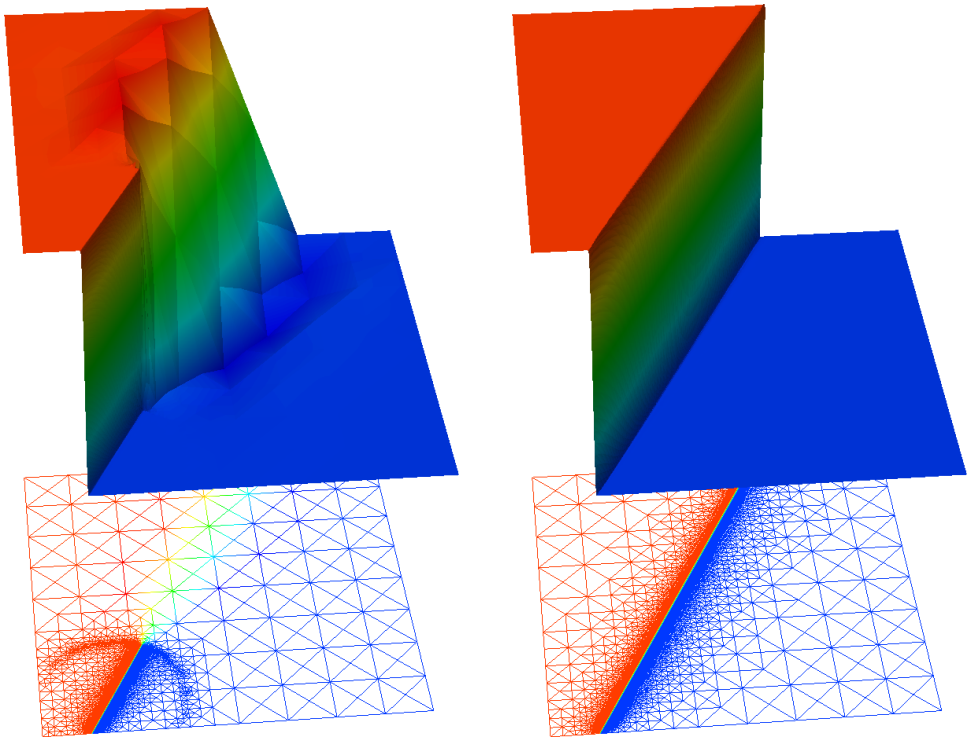
Position of $\boldsymbol{x}_e$

(b)

*Figure 4.5: (***Example 3.23***) (a) Initial triangulation of the domain* Ω *with* 145 *degrees of freedom and (b) position of the evaluation point of the target functional* $\mathcal{J}_r$.

as well as the steep layer converges better to the prescribed analytical solution. Figure 4.4 illustrates the improvement of the calculated solution in a detailed view on the significant upper and lower edge of the interior layer. Table 4.2 describes the comparison of local and global mesh refinement in numbers. The adaptive process is based on the $\mathcal{L}^2$ norm as target quantity for the sake of comparability to the error values of the global refinement. We see that local mesh refinement reduces the error $\mathcal{J}(e) = \|e\|_{\mathcal{L}^2(\Omega)}$ by a factor between 3.6 and 5.7.

If we are interested in the value of the solution at a user specified point, for instance a point along the interior layer of Example 3.23, we implement that by using the target functional $\mathcal{J}_r$. Its position is demonstrated in Figure 4.5 *(b)*. Figure 4.6 *(a)* illustrates the mechanism of action of the chosen point functional on the grid and the solution. Compared to the initial mesh in Figure 4.5 *(a)*, refinement of cells only takes place in the region around the point of interest. Since information transport is basically in streamline direction for vanishing diffusion, the functional takes the cells along the layer into consideration that range from the lower border to the point of interest. In order to clarify the different mechanism of actions of differently chosen quantities of interest, Figure 4.6 *(b)* shows the mesh refinement based on the target quantity



Adaptively governed by $\mathcal{J}_r$ | Adaptively governed by $\mathcal{J}_1$

(a) | *(b)*

*Figure 4.6: (***Example 3.23***) Solution with mesh with (a)* 39447 *and (b)* 60957 *degrees of freedom.*

$\mathcal{J}_1$. We observe that the cells along the whole layer are refined whereas cell refinement guided by $\mathcal{J}_r$ is limited to the region around the point of interest.

Now, we turn our attention to another test case that describes a solution with a circular internal layer; cf. [28].

Example 4.6. *We consider equation (4.1) with a chosen right–hand side f such that*

$$u(\boldsymbol{x}) = 16x_1(1-x_1)x_2(1-x_2) \cdot \left\{ \frac{1}{2} + \frac{\arctan\left(2\varepsilon^{-\frac{1}{2}}\left(r_0^2 - (x_1 - x_1^0)^2 - (x_2 - x_2^0)^2\right)\right)}{\pi} \right\},$$

is the exact solution of (4.1) with $r_0 = 0.25, x_1^0 = x_2^0 = 0.5$. We solve the problem in the domain $\Omega = (0,1)^2$ with $\varepsilon = 10^{-6}, \boldsymbol{b} = (2,3)^\top, \alpha = 1.0$ and

$r(u) = u^2$. The boundary conditions are also prescribed by the exact solution. Figure 4.7 illustrates the setting of the introduced example. In the shaded areas, especially behind the hump in the direction of the convection, there occur spurious undesirable oscillations in the case of a nonstabilized scheme. The

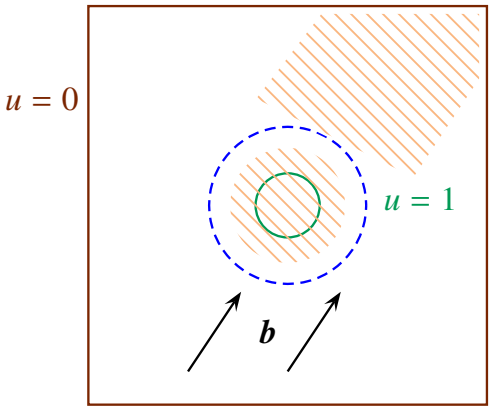


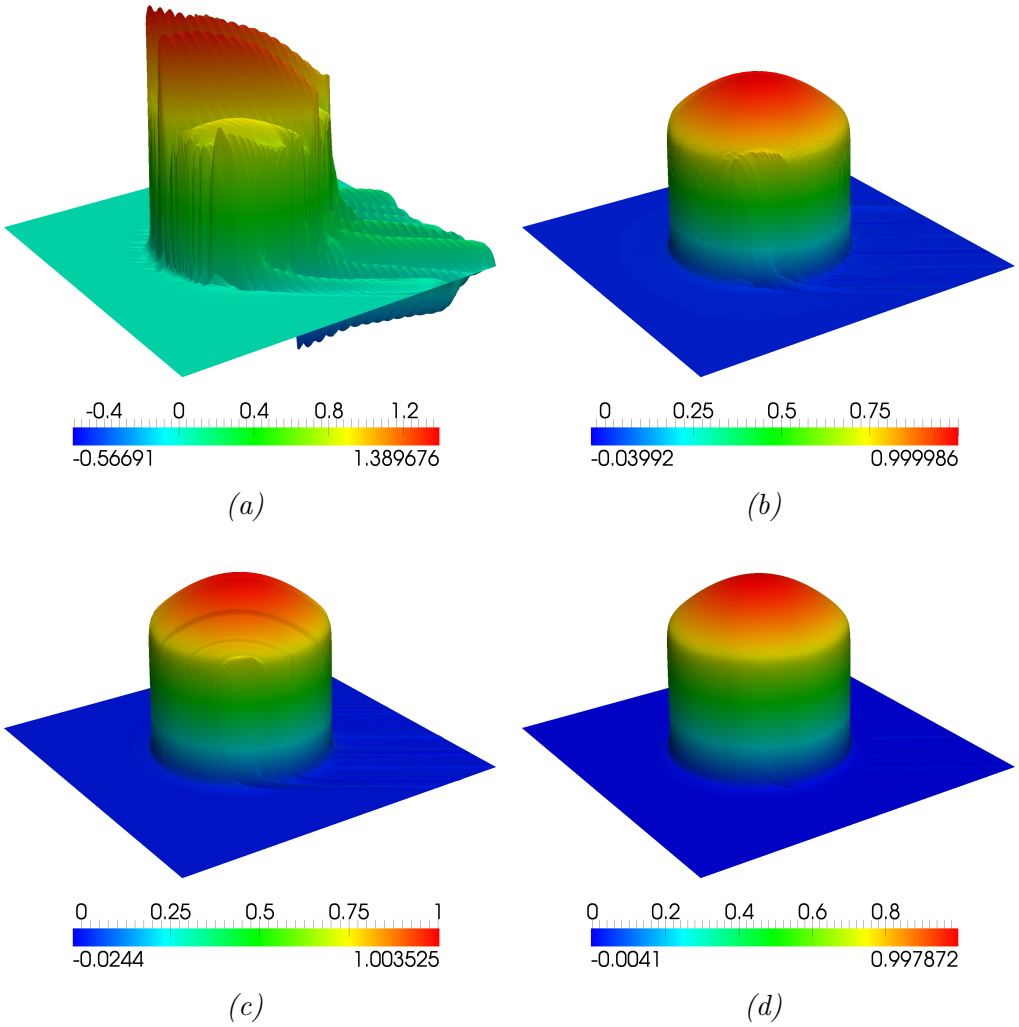
Figure 4.7: Configuration of Example 4.6 with a circular internal layer (dashed line). The hatched regions symbolize the areas where oscillations primarily occur.

quantities of interest we concentrate on in the following are the $\mathcal{L}^2$ norm and a target quantity with a more application–oriented backdrop

$$\mathcal{J}_3(u) = \int_\Omega r(u)\, d\boldsymbol{x}\,.$$

Since the great advantage of the above presented method lies in the fact that we are able to control more or less arbitrary application–related quantities of interest, we want to emphasize that we use the $\mathcal{L}^2$ norm defined in (4.15) for comparative purposes only. First of all, we compare the approximate solutions on a uniformly refined mesh and on a grid that is controlled by $\mathcal{J}_3$. Figure 4.8 gives us an impression that our method is capable to strongly reduce the oscillations that characterize a solution calculated on a uniformly refined mesh.

The solution obtained by a globally refined mesh with 33025 degrees of freedom is absolutely unusable whereas the comparative solution on a grid with 31053 nodes obtained by local mesh refinement is acceptable. Even a solution on a uniformly refined grid with a high resolution of 525313 nodes shows oscillations behind the hump in convection direction.



(a) *(b)*

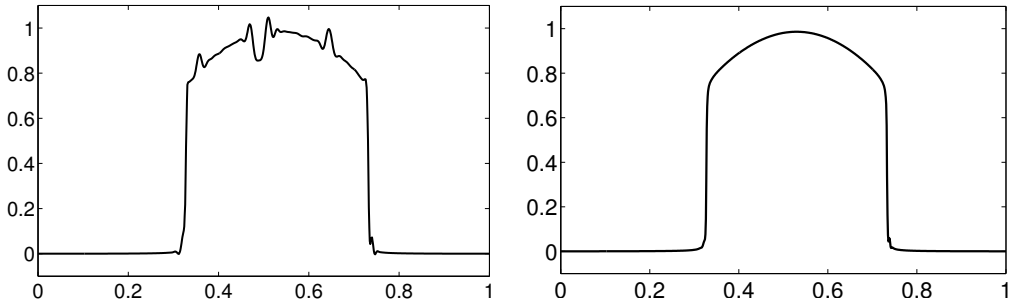
(c) *(d)*

Global refinement Local refinement

*Figure 4.8: (***Example 4.6***) Solution and value range on a globally refined mesh with (a)* 33025 *nodes and (c)* 525313 *nodes and on an adaptively refined mesh controlled by* $\mathcal{J}_3$ *with (b)* 31053 *nodes and (d)* 164477 *nodes.*

The line plots in Figure 4.9 emphasize the above described results. Table 4.3 presents comparative numbers, the error in the terms of the target functional $\mathcal{J}_3(u) - \mathcal{J}_3(u_h)$ in dependence of the degrees of freedom.

Using a mesh that is adapted to the quantity of interest $\mathcal{J}_3$ leads to an error in terms of this quantity of interest which is by a factor of at most 40 lower than it is by using a uniformly refined mesh. Table 4.4 emphasizes the accuracy of our introduced error representation since the illustrated effectivity indices tend to one. By including all these results of simulations of Example 4.6, we can state



Global refinement *(a)* Local refinement *(b)*

*Figure 4.9: (***Example 4.6***) Line plots of a solution (a) on a globally refined mesh with* 131585 *degrees of freedom and (b) on an adaptively refined grid with* 102647 *nodes.*

Global mesh refinement

dofs	$\mathcal{J}_3(u) - \mathcal{J}_3(u_h)$
8231	-0.03992
33025	-0.02507
131585	-0.00052
525313	-0.00025

Local mesh refinement

dofs	$\mathcal{J}_3(u) - \mathcal{J}_3(u_h)$
31053	-0.00063
44456	-0.00046
64657	-0.00032
102647	-0.00024
164477	-0.00015
277350	$-9.5\text{e}-5$

*Table 4.3: (***Example 4.6***) Comparison of the values* $\mathcal{J}_3(u) - \mathcal{J}_3(u_h) = \int_\Omega r(u) - r(u_h)\, d\boldsymbol{x}$.

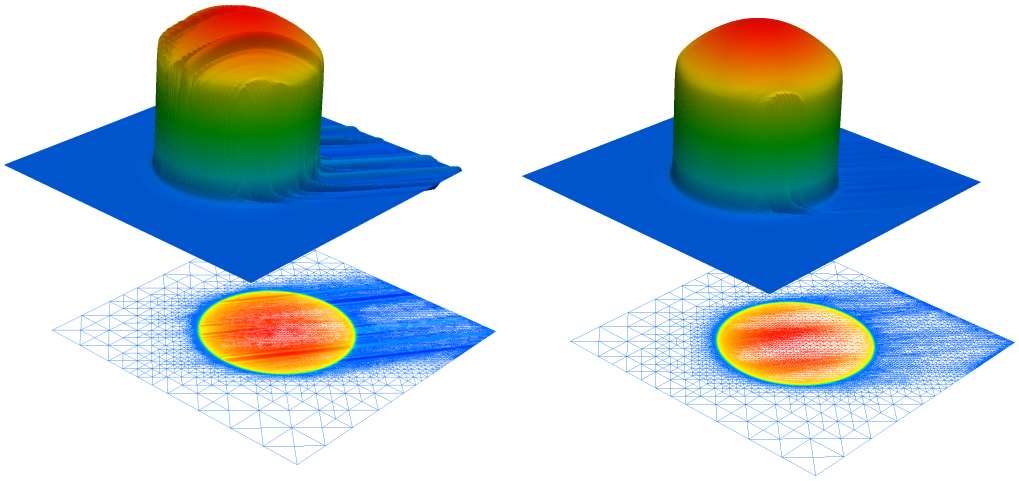
dofs	$\mathcal{I}_{\text{eff}}$	dofs	$\mathcal{I}_{\text{eff}}$
15029	1.19	64657	0.85
21497	1.29	102647	0.91
31053	1.01	164477	0.99
44456	0.95	277350	1.01

*Table 4.4: (***Example 4.6***) Effectivity indices with respect to the functional* $\mathcal{J}_3$.

that our method as well as the chosen quantity of interest are well–suited.

Now, we turn to test cases with unknown exact solutions.

Example 4.7. *Contrary to the vector fields in the previous test cases, the*

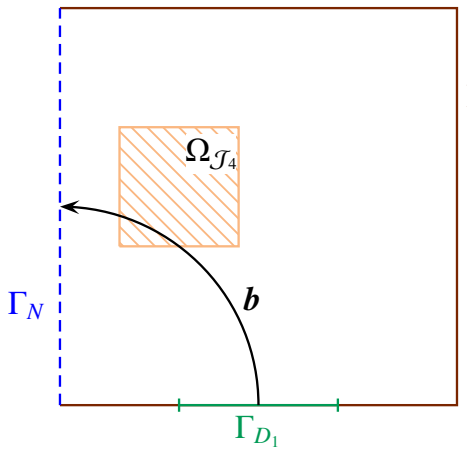


$\mathcal{J}(u) = \frac{1}{\|u\|_{\mathcal{L}^2(\Omega)}} \langle e, u \rangle_\Omega$ $\mathcal{J}_3(u) = \int_\Omega r(u)\, d\boldsymbol{x}$

(a) *(b)*

*Figure 4.10: (***Example 4.6***) Solution with mesh governed by (a) the $\mathcal{L}^2$ norm with 48691 degrees of freedom and (b) the functional $\mathcal{J}_3(u)$ with 44456 nodes.*

convection field is variable. Equation (4.1) is solved with $f \equiv 0$. We consider the computational domain $\Omega = (0,1)^2$ with $\varepsilon = 10^{-6}$, $\boldsymbol{b} = (-x_2, x_1)^\top$, $\alpha = 1.0$ and $r(u) = u^2$. The setting is illustrated in Figure 4.11.



$$\Gamma_N = \left\{(x_1, x_2) \in \Omega \,|\, x_1 = 0\right\},$$
$$\Gamma_{D_1} = \left\{(x_1, x_2) \in \Omega \,|\, 0.3 \leq x_1 \leq 0.7,\ x_2 = 0\right\},$$
$$\Gamma_{D_2} = \partial\Omega \backslash \left\{\Gamma_N \cup \Gamma_{D_1}\right\},$$
$$\Omega_{\mathcal{J}_4} = [0.15, 0.45] \times [0.4, 0.7].$$

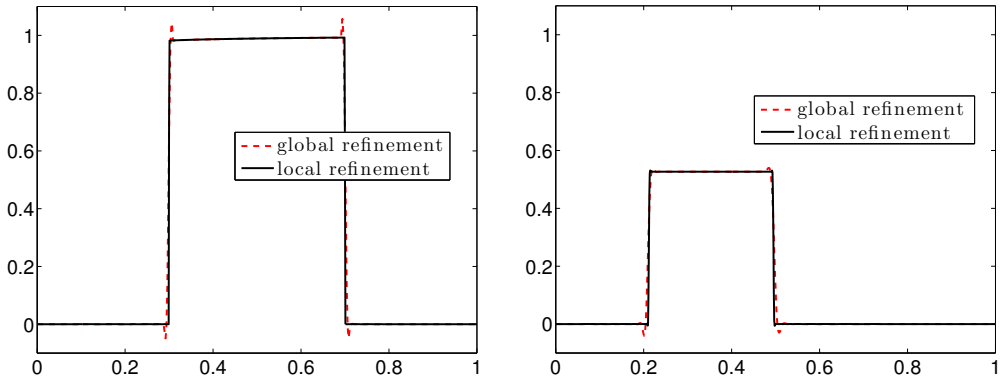
*Figure 4.11: (***Example 4.7***) Setting, definition of the boundary types and domain $\Omega_{\mathcal{J}_4}$ of the quantity of interest $\mathcal{J}_4$.*

The boundary conditions are given by $u = 1$ *on* Γ_{D_1} , $u = 0$ *on* Γ_{D_2} *and* $\frac{\partial u}{\partial \boldsymbol{n}} = 0$. *The target functional is as follows:*

$$\mathcal{J}_4(u) = \int_{\Omega_{\mathcal{J}_4}} u \, d\boldsymbol{x} \, .$$

In the context of application–related quantities of interest, we can imagine $u(\boldsymbol{x})$ *to be a species' concentration which we are interested in only on a certain domain* $\Omega_{\mathcal{J}_4}$.

We present simulations of Example 4.7 on a uniformly and on an adaptively refined grid in comparison. Figure 4.12 shows line plots along a straight line directly behind the inflow boundary as well as along the diagonal from left down to right up. A detailed view on the significant areas is given in Figure 4.13 and Figure 4.14.

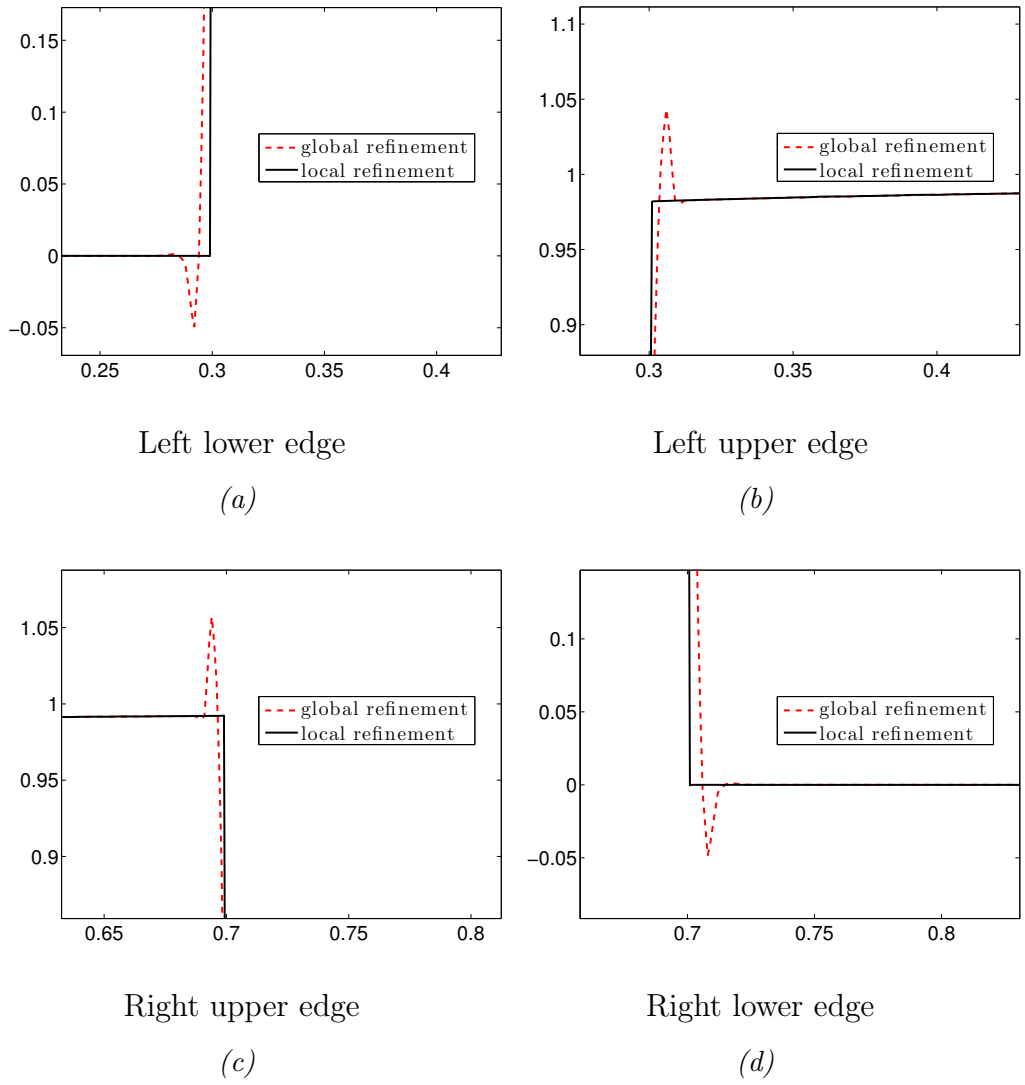


Line plot at $x_2 = 0.005$ *(a)*

Line plot at $x_2 = x_1$ *(b)*

*Figure 4.12: (***Example 4.7***)* 149283 *nodes on a adaptively refined mesh and* 131585 *nodes on a uniformly refined grid. For a detailed view, see Figure 4.13 and Figure 4.14.*

The line plots at both positions demonstrate that the adaptively generated mesh ensures a significant reduction of over– and undershoots. The inner as well as the outer layer are characterized by extremely small oscillations. Figure 4.15 illustrates that our presented method is capable to generate a solution that is very close to an exact solution whereas a highly resolved approximate solution on a uniform mesh still shows oscillating features.



Left lower edge

(a)

Left upper edge

(b)

Right upper edge

(c)

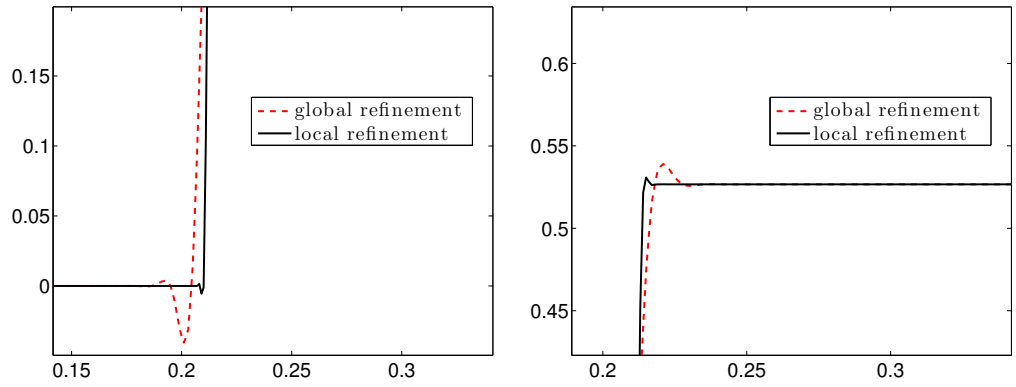
Right lower edge

(d)

*Figure 4.13: (***Example 4.7***) Detailed view of the significant areas at the inflow boundary.*

The accuracy and quality of our method is emphasized by Figure 4.16 that presents line plots at the outflow boundary $x_1 = 0.0$. A detailed view on the significant areas at the ouflow boundary is given in Figure 4.17.

Figure 4.18 aims at justifying our proposed approach. We repeat the simulation for Example 4.7 without shock–capturing stabilization. The line plot in Figure 4.18 illustrates that, besides adaptivity, shock–capturing stabilization is necessary in order to reduce oscillations especially near the inflow boundary. In comparison with the line plot in Figure 4.12 *(a)*, Figure 4.18 shows

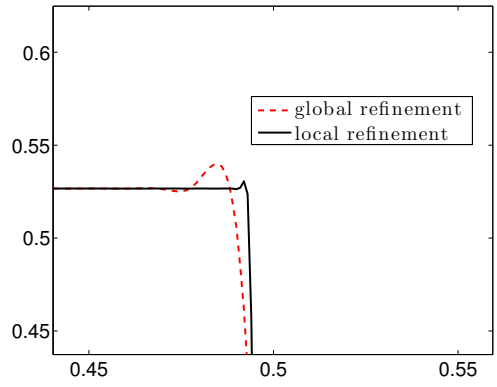


Left lower edge

(a)

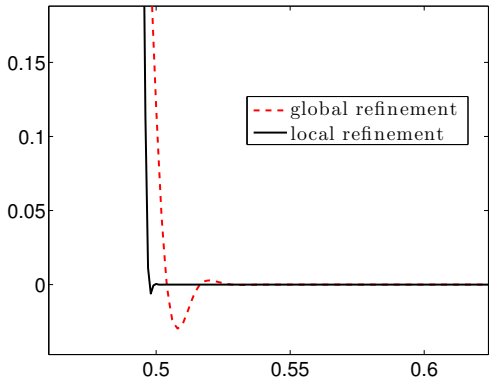
Left upper edge

(b)



Right upper edge

(c)



Right lower edge

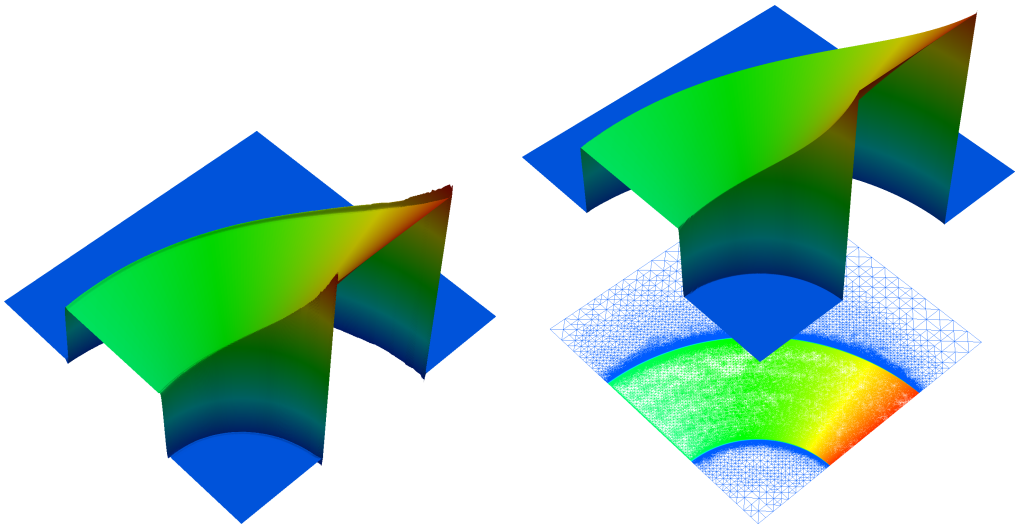
(d)

*Figure 4.14: (***Example 4.7***) Detailed view of the significant areas along the diagonal* $x_2 = x_1$.

that an adaptive method without the proper stabilization technique for the convection–dominated problem is not sufficient to yield a nearly perfect solution.

The following test case is discussed in detail in the context of a–posteriori error estimators in recent works; cf. [27] and [29]. The model problem was introduced first in [23] and is called the *Hemker problem*.

Example 4.8. *The linear convection–diffusion problem is solved, that is equa-*



Uniform refinement *(a)*

Local refinement *(b)*

*Figure 4.15: (***Example 4.7***) (a) Plot of the solution with* 525313 *degrees of freedom and (b) plot of the solution with mesh with* 520279 *degrees of freedom.*

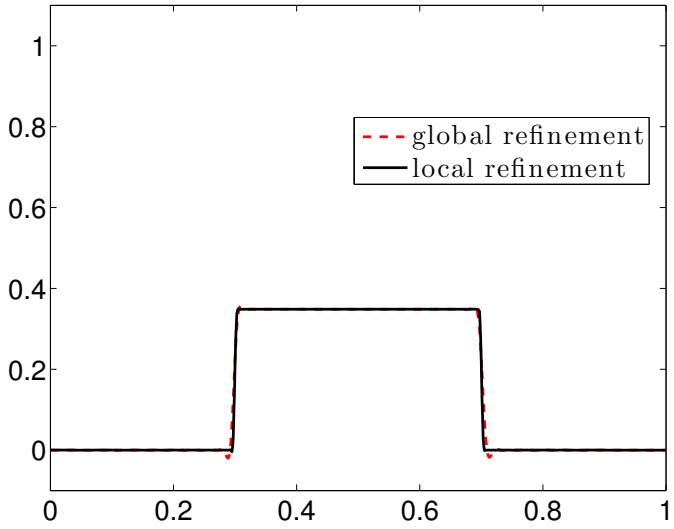
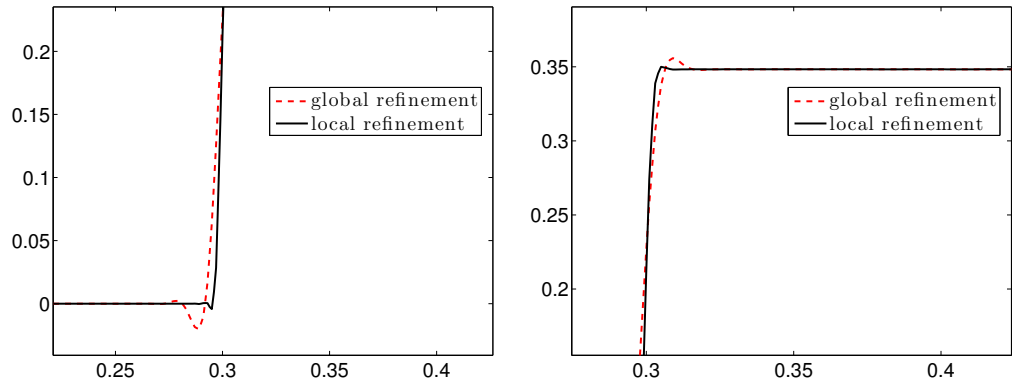


*Figure 4.16: (***Example 4.7***) Line plot at the outflow boundary, local refinement (*520279 *degrees of freedom) and global refinement (*525313 *degrees of freedom). For a detailed view, see Figure 4.17.*

tion 4.1 with $\alpha = 0 \equiv r(u), f \equiv 0$. *The reaction terms are set to zero to guarantee that the layer is completely transported through the domain. Con-*

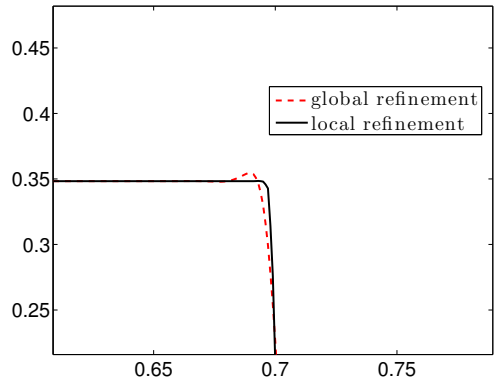


Left lower edge

(a)

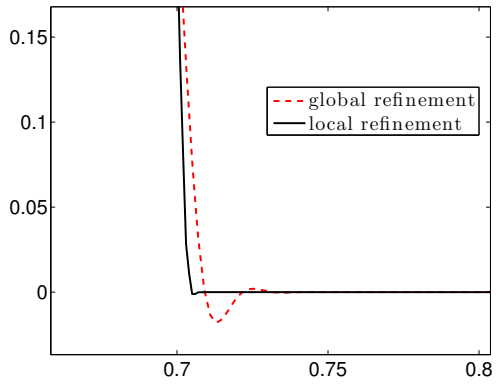
Left upper edge

(b)



Right upper edge

(c)



Right lower edge

(d)

*Figure 4.17: (***Example 4.7***) Detailed view of the significant areas at the outflow boundary.*

sequently, undesirable degradation of the species' concentration is due to numerical effects besides an insignificant amount of degradation caused by the diffusion term. Figure 4.19 illustrates the computational domain exterior of the unit circle as well as the boundary conditions. On Γ_{D_1} *and* Γ_{D_2}*, Dirichlet boundary conditions are prescribed. On all other parts of the boundary of* Ω*, homogeneous Neumann boundary conditions are defined. The diffusion and convection coefficients are given by* $\varepsilon = 10^{-6}$ *and* $\boldsymbol{b} = (1, 0)^\top$*. The quantity of*

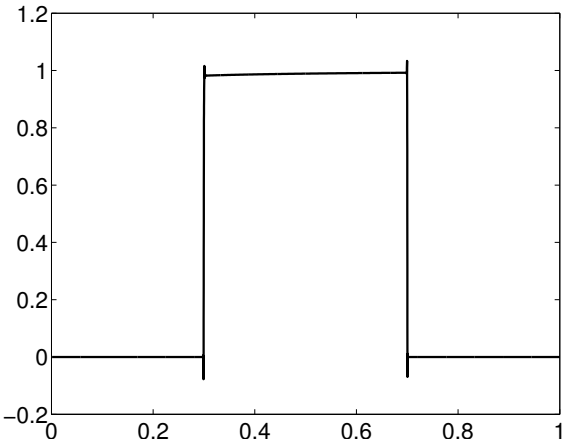
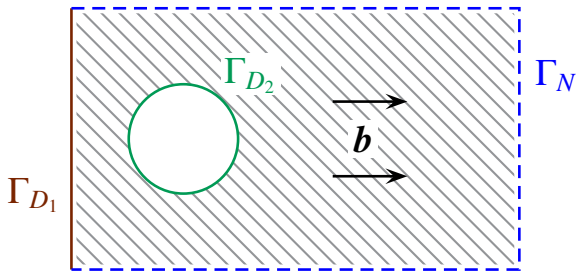


*Figure 4.18: (***Example 4.7***) Line plot of the solution stabilized only by the SUPG method with* 188527 *nodes on an adaptively refined mesh at* $x_2 = 0.005$.



$$\Omega = \left\{(-3,8)\times(-3,3)\right\} \setminus \left\{(x_1,x_2)\,|\,x_1^2+x_2^2 \le 1\right\},$$
$$u = 0 \text{ on } \Gamma_{D_1},$$
$$u = 1 \text{ on } \Gamma_{D_2},$$
$$\frac{\partial u}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma_N.$$

*Figure 4.19: (***Example 4.8***) Illustration of the setting of the Hemker problem.*

interest is

$$\mathcal{J}_1(u) = \int_\Omega u\, d\boldsymbol{x}\,.$$

A standard finite element method usually establishes a solution with wrong features especially negative oscillations in a region around the top and the bottom of the circular cutout. In [27], it is shown that the solution of the Hemker problem is often characterized by smeared interior layers in front of the circular boundary. Figure 4.21 presents results of the numerical solution for linear finite elements. For the corresponding dual problem, we use cubic finite elements in order to obtain a dual numerical solution that is close to the exact dual solution and, therefore, to increase the reliability of the error representation.

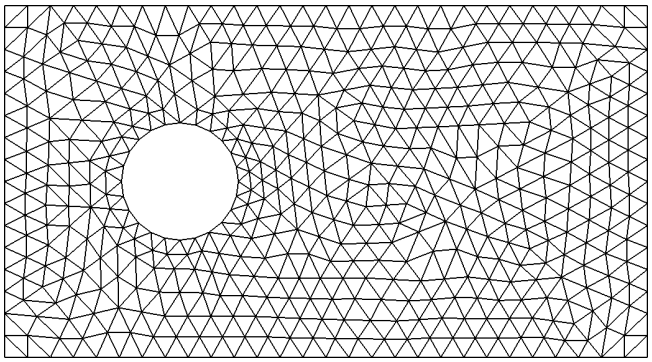


*Figure 4.20: (***Example 4.8***) Initial grid with* 528 *degrees of freedom.*

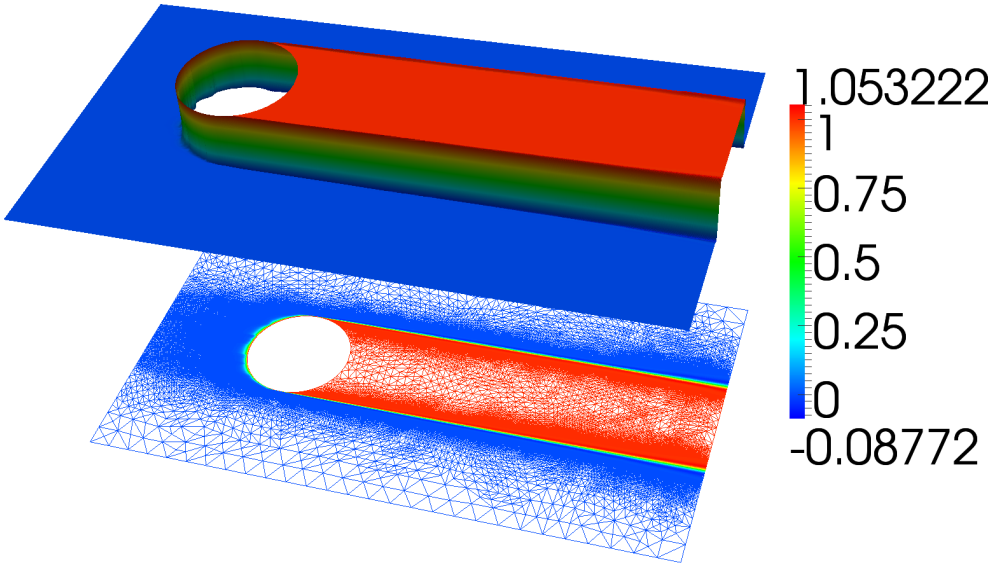


*Figure 4.21: (***Example 4.8***) Plot of the solution with mesh consisting of* 272252 *nodes.*

In [27], the optimization of the SUPG stabilizing parameters with respect to the minimization of a target functional is studied. The method the authors of [27] consider to perform best generates a solution with an undershoot of -0.24 using 47664 degrees of freedom and an undershoot of -0.13 using 425616 degrees of freedom. In our simulation, we achieve an undershoot of -0.088 using 272252 degrees of freedom. The comparison is quite rough since the number of degrees of freedom significantly differs.

Chapter 5

A nonstationary stabilized dual weighted residual method

In the preceding chapters, we have discussed the dual weighted residual (DWR) method and the numerical results using this method for stationary convection–dominated model problems. In the following, we are going to present a procedure that extends the concept of adaptivity to a nonstationary parabolic test case. For the sake of computing times, we restrict ourselves to a linear convection–diffusion–reaction model. As the DWR method strongly relies on the finite element Galerkin method, discretization in space as well as in time is done by means of Galerkin methods. Based on the notation in [16], we apply the cG(1)dG(0) method with conforming continuous discretization in space of order one and discontinuous discretization in time of order zero for the primal problem.

Considering this framework, we derive an a–posteriori error representation that governs the step size in time as well as the mesh refinement in space with respect to a user chosen target quantity. The development of an a–posteriori error estimator by means of the DWR method for nonstationary convection–dominated transport problems is mainly based on research works concerning nonstationary incompressible flow in [9], wave equations in [5] and parabolic equations in [44]. In contrast to our approach, their focus is on the separation of the temporal and spatial sources of error since this strategy provides computational advantages. However, we are not limited by necessity to separate the

temporal and spatial contributions to the error. It can even happen that we loose accuracy by separating the error contributions. Another main difference lies in the general design of the error estimation technique. We are going to present an error representation with a minimum of approximations during the development process whereas the methods cited above involve mappings of the computed solutions to approximations of the interpolation errors.

5.1 A nonstationary framework

In the following, we study the nonstationary linear convection–diffusion–reaction model problem

$$\begin{aligned} \partial_t u - \nabla \cdot (\varepsilon \nabla u) + \boldsymbol{b} \cdot \nabla u + \alpha u &= f && \text{in } \Omega \times (0,T]\,, \\ u(\boldsymbol{x},t) &= 0 && \text{on } \partial\Omega \times (0,T]\,, \\ u(\boldsymbol{x},0) &= u_0 && \text{in } \Omega\,, \end{aligned} \tag{5.1}$$

on a bounded Lipschitz domain $\Omega \subseteq \mathbb{R}^d, d \in \{2,3\}$ and in a time interval $I := (0,T]$. In view of a weak solution of (5.1), we assume $\alpha \in \mathbb{R}, \varepsilon \in \mathcal{L}^\infty(\Omega)$ and $\boldsymbol{b} \in \boldsymbol{\mathcal{H}}^1(\Omega) \cap \boldsymbol{\mathcal{L}}^\infty(\Omega)$ satisfying the conditions given in 3.2. Furthermore, we suppose that $f \in \mathcal{L}^2((0,T);\mathcal{V}^*)$ with $\mathcal{V} = \mathcal{H}_0^1(\Omega)$ and its dual space $\mathcal{V}^*$, see also Remark 5.1. The diffusion coefficient ε is supposed to be constant. The approach with respect to a variable diffusion coefficient is covered in Remark 3.1. Additionally, $u_0 \in \mathcal{L}^2(\Omega)$ is required.

We introduce the space

$$\mathfrak{V} = \left\{ v \in \mathcal{L}^2((0,T);\mathcal{V}) \,|\, \partial_t v \in \mathcal{L}^2((0,T);\mathcal{V}^*) \right\}\,,$$

and define the function space of test functions

$$\mathcal{W} := \mathcal{L}^2((0,T);\mathcal{V})\,.$$

Then, the weak formulation in space and time of (5.1) reads as:

Find $u \in \mathfrak{V}$ such that

$$A(u)(\varphi) = F(\varphi) \qquad \forall \varphi \in \mathcal{W}\,, \tag{5.2a}$$

$$u(\boldsymbol{x},0) = u_0\,, \tag{5.2b}$$

with

$$A(v)(\psi) := \langle\langle v', \psi\rangle\rangle + \langle\langle \varepsilon\nabla v, \nabla\psi\rangle\rangle + \langle\langle \boldsymbol{b}\cdot\nabla v, \psi\rangle\rangle + \langle\langle \alpha v, \psi\rangle\rangle\,, \tag{5.2c}$$

$$F(\psi) := \langle\langle f, \psi\rangle\rangle\,, \tag{5.2d}$$

where v' denotes the time derivative of v and $\langle\langle\cdot,\cdot\rangle\rangle$ denotes the space–time scalar product as defined in (2.2). On the foregoing conditions, there exists a unique solution of (5.2a) and (5.2b).

Remark 5.1. *To be precise, the $\mathcal{L}^2$ scalar product in (5.2c) and (5.2d) has to be taken as duality pairing $\langle\cdot,\cdot\rangle_{\mathcal{V}^*,\mathcal{V}}$. Since $\mathcal{V}$ is dense in $\mathcal{L}^2(\Omega)$ so that it holds that $\mathcal{V}\subset\mathcal{L}^2(\Omega)\subset\mathcal{V}^*$, the duality pairing can be viewed as an extension of the scalar product in $\mathcal{L}^2(\Omega)$. In particular, $\langle f(t), v^*\rangle_{\mathcal{V}^*,\mathcal{V}} = \langle f(t), v\rangle_{\mathcal{L}^2(\Omega)}$ whenever $f\in\mathcal{L}^2(0,T;\mathcal{L}^2(\Omega))$ for $v^*\in\mathcal{V}$ and $v\in\mathcal{L}^2(\Omega)$. The reader is referred to [17] for further details.*

To introduce the semi–discrete problem, we follow the presentation of the discontinuous Galerkin time stepping method in [45]. For the Galerkin discretization in time, we partition the time interval I in not necessarily equidistant sections $I_m := (t_{m-1}, t_m]$ by $0 = t_0 < t_1 < \ldots < t_{M-1} < t_M = T$ with step size $k_m = t_m - t_{m-1}$ and $k = \max\limits_m k_m$. We choose the space

$$\widetilde{\mathcal{W}} := \{w\in\mathcal{W}\,|\,w_{|I_m}\in\mathcal{C}(\bar{I}_m;\mathcal{V}) \text{ for } m = 1,\ldots,M\}\,,$$

that is required to derive the semi–discrete formulation of the problem (5.2a) and (5.2b). Applying integration by parts to the temporal derivative contribution of (5.2a) in each interval I_m which is allowed for smooth $\varphi\in\widetilde{\mathcal{W}}$ and setting $\varphi(T) = 0$, we obtain that

$$\langle\langle u', \varphi\rangle\rangle = -\langle\langle u, \varphi'\rangle\rangle - \langle u_0, \varphi(0)\rangle_\Omega\,.$$

Then we have to find a solution $u\in\mathfrak{V}$ satisfying

$$\begin{aligned} -\langle\langle u, \varphi'\rangle\rangle + \langle\langle \varepsilon\nabla u, \nabla\varphi\rangle\rangle + \langle\langle \boldsymbol{b}\cdot\nabla u, \varphi\rangle\rangle + \langle\langle \alpha u, \varphi\rangle\rangle \\ = \langle\langle f, \varphi\rangle\rangle + \langle u_0, \varphi(0)\rangle_\Omega\,, \end{aligned} \tag{5.3}$$

for all $\varphi\in\widetilde{\mathcal{W}}$. Now, we introduce the function space for discretization in time

$$\mathfrak{V}_k := \left\{v_k\in\mathcal{L}^2\left((0,T);\mathcal{V}\right)\,|\,v_{k|I_m}\in\mathcal{P}_0(I_m;\mathcal{V}), m = 1,\ldots,M\right\}\,, \tag{5.4}$$

where $\mathcal{P}_0(I_m;\mathcal{V})$ denotes the space of polynomials of degree zero on I_m with values in $\mathcal{V}$. We observe that the dG(0) method is nonconforming which means that $\mathfrak{V}_k \not\subset \mathfrak{V}$. This can easily be seen since elements of $\mathfrak{V}_k$ are continuous whereas elements of $\mathfrak{V}_k$ not necessarily have to be. We set $u_{k,m} := u_k(t_m)$, and the limit from above of $u_{k,m}$ at t_m is denoted by $u^+_{k,m}$. Substituting u in (5.3) by a time–discrete function $u_k \in \mathfrak{V}_k$ and integrating by parts in each interval I_m, we get for the contribution containing the time derivative that

$$\begin{aligned}
-\langle\langle u_k, \varphi'\rangle\rangle &= -\sum_{m=1}^{M} \int_{I_m} \langle u_k, \varphi'\rangle_\Omega \, dt = -\sum_{m=1}^{M} \left(\langle u_k, \varphi\rangle_\Omega \big|_{t_{m-1+0}}^{t_m} - \int_{I_m} \langle u_k', \varphi\rangle_\Omega \, dt \right) \\
&= \langle\langle u_k', \varphi\rangle\rangle + \sum_{m=2}^{M} \left(\left\langle u^+_{k,m-1} - u_{k,m-1}, \varphi(t_{m-1})\right\rangle_\Omega \right) + \left\langle u^+_{k,0}, \varphi(t_0)\right\rangle_\Omega \\
&\quad + \left\langle u^+_{k,M}, \varphi(t_M)\right\rangle_\Omega .
\end{aligned}$$

By setting $\varphi(t_M) = 0$ and defining the jump of $u_{k,m}$ at t_m by $[u_k]_m := u^+_{k,m} - u_{k,m}$, we can conclude that

$$-\langle\langle u_k, \varphi'\rangle\rangle = \langle\langle u_k', \varphi\rangle\rangle + \sum_{m=2}^{M} \left\langle [u_k]_{m-1}, \varphi(t_{m-1})\right\rangle_\Omega + \left\langle u^+_{k,0}, \varphi(t_0)\right\rangle_\Omega . \tag{5.5}$$

From (5.5), we conclude the following time–discrete formulation

Find $u_k \in \mathfrak{V}_k$ such that

$$A(u_k)(\varphi_k) = F(\varphi_k) \qquad \forall \varphi_k \in \mathfrak{V}_k , \tag{5.6a}$$

with

$$\begin{aligned}
A(v_k)(\psi_k) &:= \langle\langle v_k', \psi_k\rangle\rangle + \langle\langle \varepsilon \nabla v_k, \nabla \psi_k\rangle\rangle \\
&\quad + \langle\langle \boldsymbol{b} \cdot \nabla v_k, \psi_k\rangle\rangle + \langle\langle \alpha v_k, \psi_k\rangle\rangle \\
&\quad + \sum_{m=2}^{M} \left\langle [v_k]_{m-1}, \psi^+_{k,m-1}\right\rangle_\Omega + \left\langle v^+_{k,0}, \psi^+_{k,0}\right\rangle_\Omega , \\
F(\psi_k) &:= \langle\langle f, \psi_k\rangle\rangle + \left\langle v_0, \psi^+_{k,0}\right\rangle_\Omega .
\end{aligned} \tag{5.6b}$$

We observe that the dG(0) method in time (5.6a) is consistent with the continuous formulation (5.2a). That means that the exact solution $u \in \mathfrak{V}$ also fulfills (5.6a). By replacing $\mathcal{V}$ in the definition of the time–discrete space (5.4) by $\mathcal{V}_h^m$, we obtain the fully discrete function space

$$\mathfrak{V}_{kh} := \left\{ v_{kh} \in \mathfrak{V}_k \middle|\, v_{|I_m} \in \mathcal{P}_0(I_m; \mathcal{V}_h^m), m = 1, \ldots, M \right\} .$$

We note that the spatial finite element space $\mathcal{V}_h^m$ is allowed to be different in all intervals I_m which is natural in the context of discontinuous test and trial functions in time. The fully discrete method then has the form

Find $u_{kh} \in \mathfrak{V}_{kh}$ such that

$$A(u_{kh})(\varphi_{kh}) = F(\varphi_{kh}) \qquad \forall \varphi_{kh} \in \mathfrak{V}_{kh}\,, \tag{5.7a}$$

with

$$\begin{aligned} A(v_{kh})(\psi_{kh}) := {} & \langle\langle v'_{kh}, \psi_{kh}\rangle\rangle + \langle\langle \varepsilon \nabla v_{kh}, \nabla \psi_{kh}\rangle\rangle + \langle\langle \boldsymbol{b}\cdot\nabla v_{kh}, \psi_{kh}\rangle\rangle \\ & + \langle\langle \alpha v_{kh}, \psi_{kh}\rangle\rangle + \sum_{m=2}^{M} \left\langle [v_{kh}]_{m-1}\,, \psi^+_{kh,m-1}\right\rangle_\Omega \\ & + \left\langle v^+_{kh,0}, \psi^+_{kh,0}\right\rangle_\Omega\,, \\ F(\psi_k) := {} & \langle\langle f, \psi_{kh}\rangle\rangle + \left\langle v_0, \psi^+_{kh,0}\right\rangle_\Omega\,. \end{aligned} \tag{5.7b}$$

As the convection–dominated character of the equations usually leads to spurious oscillations in the approximate solution, we stabilize the fully discrete formulation (5.7a) by adding terms involving the residual in streamline direction. The SUPG stabilized fully discrete method reads as:

Find $u_{kh} \in \mathfrak{V}_{kh}$ such that

$$A_S(u_{kh})(\varphi_{kh}) = F(\varphi_{kh}) \qquad \forall \varphi_{kh} \in \mathfrak{V}_{kh}\,, \tag{5.8}$$

with

$$\begin{aligned} A_S(v_{kh})(\psi_{kh}) := {} & A(v_{kh})(\psi_{kh}) + S(v_{kh})(\psi_{kh})\,, \\ S(v_{kh})(\psi_{kh}) := {} & \int_0^T \sum_{K\in\mathcal{T}_h} \delta_K \langle R(v_{kh}), \boldsymbol{b}\cdot\nabla\psi_{kh}\rangle_K \, dt \\ & + \sum_{m=2}^{M} \sum_{K\in\mathcal{T}_h} \delta_K \left\langle [v_{kh}]_{m-1}\,, \boldsymbol{b}\cdot\nabla\psi^+_{kh,m-1}\right\rangle_K \\ & + \sum_{K\in\mathcal{T}_h} \delta_K \left\langle v^+_{kh,0} - v_0, \boldsymbol{b}\cdot\nabla\psi^+_{kh,0}\right\rangle_K\,, \\ R(v_h) := {} & v'_{kh} - \nabla\cdot(\varepsilon\nabla v_{kh}) + \boldsymbol{b}\cdot\nabla v_{kh} + \alpha v_{kh} - f\,. \end{aligned} \tag{5.9}$$

The SUPG method is also consistent with the continuous formulation (5.2a).

5.2 A time–dependent $\mathcal{FDTS}$ method

The general pattern of the methods developed in Chapter 3 and 4 can be directly transferred to nonstationary model problems by means of Galerkin discretization in time as well as in space. This emphasizes the rigorous property of the DWR method that it can be applied whenever the problem is based on a variational formulation.

The development of the time–dependent $\mathcal{FDTS}$ method is organized as follows: Starting with the introduction of a time–discrete Galerkin formulation of the model problem (5.1), we $\mathcal{F}$irst take the associated $\mathcal{D}$ual problem and $\mathcal{T}$hen add SUPG $\mathcal{S}$tabilization. Since we restrict ourselves to a linear nonstationary model problem, it is not reasonable to add nonlinear shock–capturing terms.

For that purpose, we introduce the Lagrangian functional

$$\mathfrak{L}(u,z) := \mathcal{J}(u) + F(z) - A(u)(z)\,, \tag{5.10}$$

with a user chosen target quantity $\mathcal{J}$, and F and A given in (5.7b). The functional $\mathcal{J}$ is supposed to be differentiable with $\mathcal{J}'(\cdot)(\varphi) = \langle\langle j(\cdot),\varphi\rangle\rangle$ and $j \in \mathcal{L}^2\left((0,T);\mathcal{L}^2(\Omega)\right)$. The derivative of $\mathfrak{L}$ with respect to z offers the original problem (5.2a). Differentiating $\mathfrak{L}$ with respect to u leads to the corresponding dual problem

$$\begin{aligned}\langle\langle\zeta', z\rangle\rangle + \langle\langle\varepsilon\nabla\zeta, \nabla z\rangle\rangle + \langle\langle \boldsymbol{b}\cdot\nabla\zeta, z\rangle\rangle + \langle\langle\alpha\zeta, z\rangle\rangle \\ + \sum_{m=2}^{M} \left\langle [\zeta]_{m-1}\,, z(t_{m-1})\right\rangle_{\Omega} + \langle\zeta(0), z(0)\rangle_{\Omega} = \langle\langle j,\zeta\rangle\rangle\,.\end{aligned} \tag{5.11}$$

Integration by parts in time applied to the first term of the left–hand side of (5.11) leads to

$$\langle\langle\zeta', z\rangle\rangle = -\langle\langle z',\zeta\rangle\rangle + \langle\zeta(T), z(T)\rangle_{\Omega} - \langle\zeta(0), z(0)\rangle_{\Omega}\,. \tag{5.12}$$

Using the identities (5.12) and (3.14) results in the time–discrete formulation of the dual problem

Find $z_k \in \mathfrak{V}_k$ such that

$$A(\zeta_k)(z_k) = \mathcal{J}'(u_k)(\zeta_k) \qquad \forall \zeta_k \in \mathfrak{V}_k\,, \tag{5.13a}$$

with

$$\begin{aligned} A(\xi_k)(w_k) := &-\langle\langle w_k', \xi_k\rangle\rangle + \langle\langle \varepsilon\nabla w_k, \nabla\xi_k\rangle\rangle - \langle\langle \boldsymbol{b}\cdot\nabla w_k, \xi_k\rangle\rangle \\ &+\langle\langle \alpha w_k, \xi_k\rangle\rangle + \sum_{m=2}^{M} \left\langle w_{k,m-1}^{+}, [\xi_k]_{m-1}\right\rangle_{\Omega} \\ &+ \langle w(T), \xi(T)\rangle_{\Omega} \,. \end{aligned} \tag{5.13b}$$

Based on the semi–discrete formulation (5.13a), we determine the dual weak solution.

Definition 5.2. *Assume that $\mathcal{J}$ is a functional on $\mathfrak{V}$ with $\mathcal{J}'(\cdot)(\zeta) = \langle\langle j(\cdot), \zeta\rangle\rangle$. We define $z \in \mathfrak{V} \subset \mathcal{W}$ to be the weak solution of*

$$A(\zeta)(z) = \mathcal{J}'(u)(\zeta) \qquad\qquad \forall \zeta \in \mathcal{W}\,, \tag{5.14a}$$

with

$$\begin{aligned} A(\xi)(w) := &-\langle\langle w', \xi\rangle\rangle + \langle\langle \varepsilon\nabla w, \nabla\xi\rangle\rangle - \langle\langle \boldsymbol{b}\cdot\nabla w, \xi\rangle\rangle \\ &+\langle\langle \alpha w, \xi\rangle\rangle + \langle w_0, \xi(T)\rangle_{\Omega} \,, \end{aligned} \tag{5.14b}$$

with a weakly incorporated terminal condition posed at the end time $t = T$. This variational formulation can be viewed as a convection–diffusion–reaction equation running backward in time with right–hand side j. We note that the jump terms in (5.13b) vanish for the weak solution z. According to standard existence theory of parabolic partial differential equations, we can formulate the following statement.

Lemma 5.3. *Assume that $j \in \mathcal{L}^2\left((0,T);\mathcal{L}^2(\Omega)\right)$ and $z_0 \in \mathcal{L}^2(\Omega)$. Then, problem (5.14a) has a unique weak solution.*

According to Lemma 5.3 and due to the inclusion $\mathfrak{V} \subset \mathcal{W}$, the dual solution $z \in \mathcal{W}$ satisfies in particular the following weak formulation of the dual problem

$$A(\zeta)(z) = \mathcal{J}'(u)(\zeta) \qquad\qquad \forall \zeta \in \mathfrak{V}\,, \tag{5.15}$$

where A is given in (5.14b).

From the fact that the adjoint problem is similar to the convection–dominated character of the original problem arises the necessity to stabilize the dual problem as well. The stabilized Euler–Lagrange system then reads

Seek solutions $\{u_{kh}, z_{kh}\} \in \mathfrak{V}_{kh} \times \mathfrak{V}_{kh}$ such that

$$A_S(u_{kh})(\varphi_{kh}) = F(\varphi_{kh}) \qquad \forall \varphi_{kh} \in \mathfrak{V}_{kh}\,, \tag{5.16a}$$

$$A_{S^*}(\zeta_{kh})(z_{kh}) = \mathcal{J}'(u_{kh})(\zeta_{kh}) \qquad \forall \zeta_{kh} \in \mathfrak{V}_{kh}\,, \tag{5.16b}$$

where $A_S(v_{kh})(\psi_{kh})$ is given in (5.9) and the dual form is defined by

$$A_{S^*}(\xi_{kh})(w_{kh}) := A(\xi_{kh})(w_{kh}) + S^*(\xi_{kh})(w_{kh})\,, \tag{5.16c}$$

with

$$\begin{aligned}
A(\xi_{kh})(w_{kh}) :=\ & -\langle\langle w'_{kh}, \xi_{kh}\rangle\rangle + \langle\langle \varepsilon \nabla w_{kh}, \nabla \xi_{kh}\rangle\rangle - \langle\langle \boldsymbol{b}\cdot\nabla w_{kh}, \xi_{kh}\rangle\rangle \\
& + \langle\langle \alpha w_{kh}, \xi_{kh}\rangle\rangle + \sum_{m=2}^{M} \left\langle w^+_{kh,m-1}, [\xi_{kh}]_{m-1}\right\rangle_\Omega + \langle w_0, \xi_{kh,M}\rangle_\Omega\,, \\
S^*(\xi_{kh})(w_{kh}) :=\ & \int_0^T \sum_{K\in\mathcal{T}_h} \delta_K^* \left\langle R^*(w_{kh}), \boldsymbol{b}\cdot\nabla\xi_{kh}\right\rangle_K dt \\
& - \sum_{m=2}^{M} \sum_{K\in\mathcal{T}_h} \delta_K^* \left\langle w^+_{kh,m-1}, \boldsymbol{b}\cdot\nabla\,[\xi_{kh}]_{m-1}\right\rangle_K \\
& - \sum_{K\in\mathcal{T}_h} \delta_K^* \left\langle w_0, \boldsymbol{b}\cdot\nabla\xi_{kh,M}\right\rangle_K\,, \\
R^*(w_{kh}) :=\ & w'_{kh} + \nabla\cdot(\varepsilon\nabla w_{kh}) + \boldsymbol{b}\cdot\nabla w_{kh} - \alpha w_{kh} + j\,.
\end{aligned}$$

Let $x \in \mathcal{X} := \mathfrak{V}\times\mathcal{W}$ be a stationary point of the Lagrangian functional (5.10), i.e.

$$\mathfrak{L}'(x)(y) = 0 \quad \forall y \in \mathcal{X}\,. \tag{5.17}$$

Summarizing the discrete formulations (5.16a) and (5.16b) leads to

$$\begin{aligned}
& F(\varphi_{kh}) - A_S(u_{kh})(\varphi_{kh}) \\
& \qquad + \mathcal{J}'(u_{kh})(\zeta_{kh}) - A_{S^*}(\zeta_{kh})(z_{kh}) = 0 \\
& \qquad\qquad\qquad\qquad \forall \{\zeta_{kh}, \varphi_{kh}\} \in \mathfrak{V}_{kh}\times\mathfrak{V}_{kh}\,,
\end{aligned}$$

and, continuing,

$$\begin{aligned}
& \mathfrak{L}'_u(u_{kh}, z_{kh})(\zeta_{kh}) + \mathfrak{L}'_z(u_{kh}, z_{kh})(\varphi_{kh}) \\
& \qquad - S(u_{kh})(\varphi_{kh}) - S^*(\zeta_{kh})(z_{kh}) = 0 \\
& \qquad\qquad\qquad\qquad \forall \{\zeta_{kh}, \varphi_{kh}\} \in \mathfrak{V}_{kh}\times\mathfrak{V}_{kh}\,.
\end{aligned}$$

Setting

$$x_{kh} := \{u_{kh}, z_{kh}\} \in \mathfrak{V}_{kh} \times \mathfrak{V}_{kh} =: \mathcal{X}_{kh}\,,$$
$$y_{kh} := \{\zeta_{kh}, \varphi_{kh}\} \in \mathfrak{V}_{kh} \times \mathfrak{V}_{kh}\,,$$

and

$$\mathcal{S}(x_h)(y_h) := S(u_{kh})(\varphi_{kh}) + S^*(\zeta_{kh})(z_{kh})\,,$$

we get that

$$\mathfrak{L}'(x_{kh})(y_{kh}) = \mathcal{S}(x_{kh})(y_{kh}) \quad \forall y_{kh} \in \mathcal{X}_{kh}\,. \tag{5.18}$$

We note that $\mathcal{X}_{kh}$ is not included in $\mathcal{X}$. For that reason, it is required to give a specific theorem that is adapted not only to the time dependence but also to this new condition. To this end, we define the function space

$$\hat{\mathcal{X}} := \mathcal{H} \times \mathcal{W} \quad \text{with } \mathcal{H} := \mathcal{L}^2\left((0,T); \mathcal{L}^2(\Omega)\right)\,.$$

At that point, the reader is referred to [5].

Theorem 5.4 (Error representation in terms of the Lagrangian functional)**.** *Let $\mathfrak{L} : \hat{\mathcal{X}} \to \mathbb{R}$ be a three times Fréchet differentiable functional. We seek a stationary point x of $\mathfrak{L}$ in $\mathcal{X} \subset \hat{\mathcal{X}}$. The Galerkin approximation $x_{kh} \in \mathcal{X}_{kh}$ fulfills (5.18) where $\mathcal{X}_{kh} \subset \hat{\mathcal{X}}$ but not necessarily $\mathcal{X}_{kh} \subset \mathcal{X}$. In addition, we assume that the Galerkin approximation satisfies*

$$\mathfrak{L}'(x)(x_{kh}) = 0\,. \tag{5.19a}$$

Then, the error in terms of the Lagrangian functional can be represented by

$$\mathfrak{L}(x) - \mathfrak{L}(x_{kh}) = \frac{1}{2}\mathfrak{L}'(x_{kh})(x - y_{kh}) + \frac{1}{2}\mathcal{S}(x_{kh})(y_{kh} - x_{kh}) + R\,, \tag{5.19b}$$

with an arbitrary $y_{kh} \in \mathcal{X}_{kh}$. The remainder term is given by

$$R := \frac{1}{2}\int_0^1 \mathfrak{L}'''(x_{kh} + s\hat{e})(\hat{e}, \hat{e}, \hat{e}) s(s-1)\, ds\,. \tag{5.19c}$$

Proof. Let $\hat{e} := x - x_{kh} \in \hat{\mathcal{X}}$. It holds that

$$\mathfrak{L}(x) - \mathfrak{L}(x_{kh}) = \int_0^1 \mathfrak{L}'(x_{kh} + s\hat{e})(\hat{e})\, ds\,.$$

Approximating the integral by the trapezoidal rule leads to

$$\mathfrak{L}(x) - \mathfrak{L}(x_{kh}) = \frac{1}{2}\mathfrak{L}'(x_{kh})(x - x_{kh}) + \frac{1}{2}\mathfrak{L}'(x)(x - x_{kh}) + R\,,$$

with the remainder term R determined in (5.19c). Condition (5.17) in combination with assumption (5.19a) requires that the second term of the previous equation vanishes. Due to condition (5.18), we find that

$$\begin{aligned}\mathfrak{L}(x) - \mathfrak{L}(x_{kh}) &= \frac{1}{2}\mathfrak{L}'(x_{kh})(x - y_{kh}) + \frac{1}{2}\mathfrak{L}'(x_{kh})(y_{kh} - x_{kh}) + R \\ &= \frac{1}{2}\mathfrak{L}'(x_{kh})(x - y_{kh}) + \frac{1}{2}\mathcal{S}(x_{kh})(y_{kh} - x_{kh}) + R\,,\end{aligned}$$

for arbitrary $y_{kh} \in \mathcal{X}_{kh}$. □

We define the primal residual $\rho(u_{kh})(\cdot)$ and the adjoint residual $\rho^*(z_{kh})(\cdot)$ by

$$\rho(u_{kh})(\varphi) := F(\varphi) - A(u_{kh})(\varphi) \qquad \forall \varphi \in \mathcal{W}\,, \tag{5.20a}$$

$$\rho^*(z_{kh})(\zeta) := \mathcal{J}'(u_{kh})(\zeta) - A(\zeta)(z_{kh}) \qquad \forall \zeta \in \mathfrak{V}\,, \tag{5.20b}$$

where the primal linear forms A and F are given in (5.6b), and the dual linear form is determined in (5.13b). The following theorem represents the error in terms of the target quantity $\mathcal{J}(\cdot)$.

Theorem 5.5 (Error representation with respect to the target quantity $\mathcal{J}$)**.** *Suppose $\{u, z\} \in \mathfrak{V} \times \mathcal{W}$ to be stationary points of $\mathfrak{L}$ and $\{u_{kh}, z_{kh}\} \in \mathfrak{V}_{kh} \times \mathfrak{V}_{kh}$ be their Galerkin approximations. Under the condition*

$$\mathcal{J}'(u)(u_{kh}) - A(u_{kh})(z) = 0\,, \tag{5.21a}$$

the error representation reads

$$\mathcal{J}(u) - \mathcal{J}(u_{kh}) = \frac{1}{2}\rho(u_{kh})(z - \varphi_{kh}) + \frac{1}{2}\rho^*(z_{kh})(u - \zeta_{kh}) + R_{\mathcal{S}} + R_{\mathcal{J}}\,, \tag{5.21b}$$

with arbitrary $\varphi_{kh}, \zeta_{kh} \in \mathfrak{V}_{kh}$. The primal and dual residuals are defined in (5.20a) and (5.20b). The remainder terms are determined by

$$R_{\mathcal{S}} := \frac{1}{2}S(u_{kh})(\varphi_{kh} + z_{kh}) + \frac{1}{2}S^*(\zeta_{kh} - u_{kh})(z_{kh})\,, \tag{5.21c}$$

and

$$R_{\mathcal{J}} := \frac{1}{2}\int_0^1 \mathcal{J}'''(u_{kh} + se)(e, e, e)s(s-1)\,ds\,. \tag{5.21d}$$

The stabilization of the primal and dual problem affects the term $R_{\mathcal{S}}$ whereas the nonlinear property of the quantity of interest $\mathcal{J}$ affects $R_{\mathcal{J}}$.

Proof. Owing to assumption (5.21a) and the fact that $z_{kh} \in \mathfrak{V}_{kh} \subset \mathcal{W}$, we find that

$$\begin{aligned}\mathfrak{L}'(x)(x_{kh}) &= \mathfrak{L}'(u,z)(u_{kh}) + \mathfrak{L}'(u,z)(z_{kh}) \\ &= \mathcal{J}'(u)(u_{kh}) - A(u_{kh})(z) + F(z_{kh}) - A(u)(z_{kh}) = 0\,.\end{aligned}$$

This, therefore, provides that condition (5.19a) of the previous theorem is satisfied. Using the identities

$$\begin{aligned}\mathfrak{L}(x) &= \mathcal{J}(u) + F(z) - A(u)(z) = \mathcal{J}(u) \\ \mathfrak{L}(x_{kh}) &= \mathcal{J}(u_{kh}) + F(z_{kh}) - A(u_{kh})(z_{kh}) = \mathcal{J}(u_{kh}) + S(u_{kh})(z_{kh})\,,\end{aligned}$$

leads to

$$\mathcal{J}(u) - \mathcal{J}(u_{kh}) = \mathfrak{L}(x) - \mathfrak{L}(x_{kh}) + S(u_{kh})(z_{kh})\,.$$

Applying the previous theorem to the error in terms of the Lagrangian functional yields that

$$\begin{aligned}\mathcal{J}(u) - \mathcal{J}(u_{kh}) = \frac{1}{2}\mathfrak{L}'(x_{kh})(x - y_{kh}) + \frac{1}{2}\mathcal{S}(x_{kh})(y_{kh} - x_{kh}) \\ + S(u_{kh})(z_{kh}) + R\,.\end{aligned} \tag{5.22}$$

The derivative of the Lagrangian functional $\mathfrak{L}$ can be rewritten to

$$\begin{aligned}&\mathfrak{L}'(u_{kh}, z_{kh})(u - \zeta_{kh}, z - \varphi_{kh}) \\ &\quad= \mathcal{J}'(u_{kh})(u - \zeta_{kh}) - A(u - \zeta_{kh})(z_{kh}) + F(z - \varphi_{kh}) - A(u_{kh})(z - \varphi_{kh}) \\ &\quad= \rho(u_{kh})(z - \varphi_{kh}) + \rho^*(z_{kh})(u - \zeta_{kh})\,.\end{aligned}$$

Substituting this result into formula (5.22) gives that

$$\mathcal{J}(u) - \mathcal{J}(u_{kh}) = \frac{1}{2}\rho(u_{kh})(z - \varphi_{kh}) + \frac{1}{2}\rho^*(z_{kh})(u - \zeta_{kh}) + R_{\mathcal{S}} + R\,,$$

with

$$\begin{aligned}&\frac{1}{2}\mathcal{S}(x_{kh})(y_{kh} - x_{kh}) + S(u_{kh})(z_{kh}) \\ &\qquad= \frac{1}{2}S(u_{kh})(\varphi_{kh} - z_{kh}) + \frac{1}{2}S^*(\zeta_{kh} - u_{kh})(z_{kh}) + S(u_{kh})(z_{kh}) \\ &\qquad= \frac{1}{2}S(u_{kh})(\varphi_{kh} + z_{kh}) + \frac{1}{2}S^*(\zeta_{kh} - u_{kh})(z_{kh}) := R_{\mathcal{S}}\,.\end{aligned}$$

We note that all parts of the third derivative of $\mathfrak{L}$ vanish except the third derivative of $\mathcal{J}$, which proves the assertion. □

The following theorem describes a relation between the primal and dual residual such that an approximate solution of higher order has to be generated once only. For further explanatory notes, see Remark 3.7.

Theorem 5.6. *Let u and z be the primal and dual solution, respectively, such that*

$$\mathcal{J}'(u)(u-\zeta_{kh}) - A(u-\zeta_{kh})(z) = 0 \tag{5.23a}$$

is fulfilled for all $\zeta_{kh} \in \mathfrak{V}_{kh}$. Then, we can state that

$$\rho^*(z_{kh})(u-\zeta_{kh}) = \rho(u_{kh})(z-\varphi_{kh}) - \Delta\rho_{\mathcal{J}} + \Delta\rho_{\mathcal{S}}\,. \tag{5.23b}$$

The remainder terms are given by

$$\Delta\rho_{\mathcal{J}} := \int_0^1 \mathcal{J}''(u_{kh}+se)(e,e)\,ds\,, \tag{5.23c}$$

and

$$\Delta\rho_{\mathcal{S}} := S(u_{kh})(\varphi_{kh}-z_{kh}) - S^*(\zeta_{kh}-u_{kh})(z_{kh})\,. \tag{5.23d}$$

Proof. Let $e^* := z - z_{kh}$ be the adjoint error. We introduce

$$k(s) := \mathcal{J}'(u_{kh}+se)(u-\zeta_{kh}) - A(u-\zeta_{kh})(z_{kh}+se^*)\,,$$

with its derivative

$$k'(s) = \mathcal{J}''(u_{kh}+se)(e,u-\zeta_{kh}) - A(u-\zeta_{kh})(e^*)\,.$$

The condition (5.23a) offers that $k(1) = 0$. Furthermore, it holds that

$$k(0) = \mathcal{J}'(u_{kh})(u-\zeta_{kh}) - A(u-\zeta_{kh})(z_{kh}) = \rho^*(z_{kh})(u-\zeta_{kh})\,.$$

By substituting $k(0)$ and $k(1)$ into the fundamental theorem of calculus

$$\int_0^1 k(s)\,ds = k(1) - k(0)\,,$$

we find that

$$\rho^*(z_{kh})(u-\zeta_{kh}) = \int_0^1 \left(A(u-\zeta_{kh})(e^*) - \mathcal{J}''(u_{kh}+se)(e,u-\zeta_{kh})\right)\,ds\,. \tag{5.24}$$

By using $u - u_{kh}$ instead of $u - \zeta_{kh}$, we get that

$$\begin{aligned}\rho^*(z_{kh})(u - \zeta_{kh}) &= \mathcal{J}'(u_{kh})(u - \zeta_{kh}) - A(u - \zeta_{kh})(z_{kh}) + S^*(\zeta_{kh})(z_{kh}) \\ &\quad -S^*(\zeta_{kh})(z_{kh}) - \mathcal{J}'(u_{kh})(u_{kh}) + A(u_{kh})(z_{kh}) + S^*(u_{kh})(z_{kh}) \\ &= \rho^*(z_{kh})(u - u_{kh}) - S^*(\zeta_{kh} - u_{kh})(z_{kh})\,.\end{aligned}$$

According to (5.24), we have considering the right–hand side of the previous equation that

$$\begin{aligned}\rho^*(z_{kh})(u - \zeta_{kh}) &= \int_0^1 \left((A(u - u_{kh})(e^*) - \mathcal{J}''(u_{kh} + se)(e, u - u_{kh})\right)\, ds \\ &\quad -S^*(\zeta_{kh} - u_{kh})(z_{kh}) \\ &= A(e)(e^*) - \int_0^1 \mathcal{J}''(u_{kh} + se)(e, e)\, ds - S^*(\zeta_{kh} - u_{kh})(z_{kh})\,.\end{aligned}$$

Inserting

$$0 = A_S(u_{kh})(\varphi_{kh}) - F(\varphi_{kh})\,,$$

yields that

$$\begin{aligned}A(e)(e^*) &= F(e^*) - A(u_{kh})(e^*) \\ &= F(z) - A(u_{kh})(z) + A(u_{kh})(z_{kh}) + S(u_{kh})(z_{kh}) - F(z_{kh}) \\ &\quad -S(u_{kh})(z_{kh}) + A(u_{kh})(\varphi_{kh}) + S(u_{kh})(\varphi_{kh}) - F(\varphi_{kh}) \\ &= F(z - \varphi_{kh}) - A(u_{kh})(z - \varphi_{kh}) + S(u_{kh})(\varphi_{kh} - z_{kh}) \\ &= \rho(u_{kh})(z - \varphi_{kh}) + S(u_{kh})(\varphi_{kh} - z_{kh})\,.\end{aligned}$$

We combine the last two results and confirm the claim

$$\begin{aligned}\rho^*(z_{kh})(u - \zeta_{kh}) &= \rho(u_{kh})(z - \varphi_{kh}) + S(u_{kh})(\varphi_{kh} - z_{kh}) - S^*(\zeta_{kh} - u_{kh})(z_{kh}) \\ &\quad - \int_0^1 \mathcal{J}''(u_{kh} + se)(e, e)\, ds \\ &=: \rho(u_{kh})(z - \varphi_{kh}) + \Delta\rho_{\mathcal{S}} - \Delta\rho_{\mathcal{J}}\,,\end{aligned}$$

with the remainder terms $\Delta\rho_{\mathcal{S}}$ and $\Delta\rho_{\mathcal{J}}$ given in (5.23c) and (5.23d), respectively. □

By means of Theorem 5.5 and Theorem 5.6, the final result is derived in the remainder of this section. This final result will give the exact appearance of the error representation for our model problem (5.1).

Theorem 5.7 (Local error description for the time–dependent $\mathcal{FDTS}$ method)**.** *For the stabilized finite element approximation of the model problem (5.1), we have the element–wise error representation*

$$\begin{aligned}\mathcal{J}(u)-\mathcal{J}(u_{kh}) = \int_0^T \sum_{K\in\mathcal{T}_h} \Big\{ & \langle \mathcal{R}(u_{kh}), z-\varphi_{kh}\rangle_K \\ & -\delta_K \langle \mathcal{R}(u_{kh}), \boldsymbol{b}\cdot\nabla\varphi_{kh}\rangle_K \\ & -\langle \mathcal{E}(u_{kh}), z-\varphi_{kh}\rangle_{\partial K}\Big\}\, dt \\ & -\left\langle u^+_{kh,0}-u_0, z(t_0)-\varphi^+_{kh,0}\right\rangle_\Omega \\ & -\sum_{m=2}^{M}\left\langle [u_{kh}]_{m-1}, z(t_{m-1})-\varphi^+_{kh,m-1}\right\rangle_\Omega \\ & +\sum_{K\in\mathcal{T}_h}\delta_K\left\langle u^+_{kh,0}-u_0, \boldsymbol{b}\cdot\nabla\varphi^+_{kh,0}\right\rangle_K \\ & +\sum_{m=2}^{M}\sum_{K\in\mathcal{T}_h}\delta_K\left\langle [u_{kh}]_{m-1}, \boldsymbol{b}\cdot\nabla\varphi^+_{kh,m-1}\right\rangle_K ,\end{aligned} \tag{5.25a}$$

The cell and edge residuals take the form

$$\mathcal{R}(u_{kh})_{|K} = f - u'_{kh} + \nabla\cdot(\varepsilon\nabla u_{kh}) - \boldsymbol{b}\cdot\nabla u_{kh} - \alpha u_{kh}\,, \tag{5.25b}$$

$$\mathcal{E}(u_{kh})_{|\Gamma} = \begin{cases} \frac{1}{2}\boldsymbol{n}\cdot[\varepsilon\nabla u_{kh}] & \text{if } \Gamma\subset\partial K\backslash\partial\Omega \\ 0 & \text{if } \Gamma\subset\partial\Omega\,, \end{cases} \tag{5.25c}$$

where $\boldsymbol{n}$ *denotes the outer–pointing normal and* $[\nabla u_{kh}]$ *defines the jump of* ∇u_{kh} *over the inner edges* Γ*.*

Proof. According to Theorem 5.5 and Theorem 5.6, the error is represented by

$$\mathcal{J}(u)-\mathcal{J}(u_{kh}) = \rho(u_{kh})(z-\varphi_{kh}) + \frac{1}{2}\Delta\rho_{\mathcal{J}} + \frac{1}{2}\Delta\rho_{\mathcal{S}} + R_{\mathcal{J}} + R_{\mathcal{S}}\,.$$

Since the contributions $\Delta\rho_{\mathcal{J}}$ determined in (5.23c) and $R_{\mathcal{J}}$ given in (5.21d) are of higher order with respect to the error e and the adjoint error e^*, respectively, they can be neglected. The remainder terms resulting from the stabilization technique are rearranged such that

$$\frac{1}{2}\Delta\rho_{\mathcal{S}} + R_{\mathcal{S}} = S(u_{kh})(\varphi_{kh})\,.$$

Finally, integration by parts applied to the cell residual proves the argument. □

As one can see, the structure of the terms of the error representation is similar to the stationary case. As mentioned earlier, only the dG jump terms resulting from the discretization of the time derivative are added to the error representation.

5.3 Numerical studies for the time–dependent $\mathcal{FDTS}$ method

In this section, we discuss the design of the algorithm including implementation aspects for the time–dependent case. Since we use dG(0) test functions in time, the primal and the dual problem decouple into a sequence of quasi stationary time stepping sub–problems which is an important issue concerning the performance of the complete algorithm.

Recalling the error representation (5.25a), we observe that, among approximate quantities like u_{kh} and φ_{kh}, continuous values of z contribute to the error representation. As pointed out before, the numerical approximation of z has to be generated by an improved solution in comparison to u_{kh}; otherwise the error representation would completely vanish. The primal problem is discretized in time by the dG(0) method which can be transferred into the implicit Euler time stepping scheme. This will be made evident below. To generate an improved dual solution, we choose continuous trial functions of order one in time which results in the Crank Nicolson scheme up to an integration error for nonlinear right–hand side functions. The space discretization is given by the cG(2) method. This approach corresponds to the cG(2)cG(1) method .

We recall the fully discrete stabilized formulation (5.8). Due to the temporal discontinuity of the test functions $\varphi_{kh} \in \mathfrak{V}_{kh}$, we set $\varphi_{kh|I_m} = 0 \; \forall \tilde{m} \neq m$ and $\varphi_{kh|I_m} \neq 0$. This leads to the decoupling of the time steps such that we seek a

solution $u_{kh} \in \mathcal{P}_0(I_m; \mathcal{V}_h^m)$ for $m = 1, \ldots, M$ satisfying

$$\begin{aligned}
&\int_{I_m} \langle u'_{kh}, \varphi_{kh}\rangle_\Omega + \langle \varepsilon \nabla u_{kh}, \nabla \varphi_{kh}\rangle_\Omega + \langle \boldsymbol{b}\cdot\nabla u_{kh}, \varphi_{kh}\rangle_\Omega + \langle \alpha u_{kh}, \varphi_{kh}\rangle_\Omega \, dt \\
&\quad + \left\langle [u_{kh}]_{m-1}, \varphi^+_{kh,m-1}\right\rangle_\Omega \\
&\quad + \int_{I_m} \sum_{K\in\mathcal{T}_h} \delta_K \langle u'_{kh} - \nabla\cdot(\varepsilon\nabla u_{kh}) + \boldsymbol{b}\cdot\nabla u_{kh} + \alpha u_{kh}, \boldsymbol{b}\cdot\nabla\varphi_{kh}\rangle_K \, dt \\
&\quad + \sum_{K\in\mathcal{T}_h} \delta_K \left\langle [u_{kh}]_{m-1}, \boldsymbol{b}\cdot\nabla\varphi^+_{kh,m-1}\right\rangle_K \\
&= \int_{I_m} \langle f, \varphi_{kh}\rangle \, dt + \int_{I_m} \sum_{K\in\mathcal{T}_h} \delta_K \langle f, \boldsymbol{b}\cdot\nabla\varphi_{kh}\rangle_K \, dt \,,
\end{aligned} \tag{5.26}$$

for all $\varphi_{kh} \in \mathcal{P}_0(I_m; \mathcal{V}_h^m)$ with

$$\begin{aligned}
&\left\langle u_0, \varphi^+_{kh,0}\right\rangle_\Omega + \sum_{K\in\mathcal{T}_h} \delta_k \left\langle u_0, \boldsymbol{b}\cdot\nabla\varphi^+_{kh,0}\right\rangle_K \\
&\qquad = \left\langle u^+_{kh,0}, \varphi^+_{kh,0}\right\rangle_\Omega + \sum_{K\in\mathcal{T}_h} \delta_k \left\langle u^+_{kh,0}\boldsymbol{b}\cdot\nabla\varphi^+_{kh,0}\right\rangle_K \,,
\end{aligned}$$

for $m = 1$. Setting $u_h^m := u_{kh|I_m} = \text{const}$ leads to $u'_{kh|I_m} \equiv 0$ and $[u_{kh}]_{m-1} = u^+_{kh,m-1} - u_{kh,m-1} = u_{kh,m} - u_{kh,m-1} = u_h^m - u_h^{m-1}$. Replacing the corresponding terms in (5.26) by these results, we obtain that

$$\begin{aligned}
&\int_{I_m} \langle \varepsilon \nabla u_h^m, \nabla \varphi_h\rangle_\Omega + \langle \boldsymbol{b}\cdot\nabla u_h^m, \varphi_h\rangle_\Omega + \langle \alpha u_h^m, \varphi_h\rangle_\Omega \, dt + \langle u_h^m - u_h^{m-1}, v_h\rangle_\Omega \\
&\quad + \int_{I_m} \sum_{K\in\mathcal{T}_h} \delta_K \langle -\nabla\cdot(\varepsilon\nabla u_h^m) + \boldsymbol{b}\cdot\nabla u_h^m + \alpha u_h^m, \boldsymbol{b}\cdot\nabla\varphi_h\rangle_K \, dt \\
&\quad + \sum_{K\in\mathcal{T}_h} \delta_K \langle u_h^m - u_h^{m-1}, \boldsymbol{b}\cdot\nabla\varphi_h\rangle_K \\
&= \int_{I_m} \langle f, \varphi_{kh}\rangle \, dt + \int_{I_m} \sum_{K\in\mathcal{T}_h} \delta_K \langle f, \boldsymbol{b}\cdot\nabla\varphi_{kh}\rangle_K \, dt \,.
\end{aligned}$$

Let $k_m = t_m - t_{m-1}$. The integrals on the right–hand side are approximated by the right endpoint rule. The remaining integrals can be calculated since

the integrands do not depend on the time. Rearranging the terms yields that

$$\begin{aligned}
&\langle u_h^m, \varphi_h\rangle_\Omega + k_m\langle \varepsilon\nabla u_h^m, \nabla\varphi_h\rangle_\Omega + k_m\langle \boldsymbol{b}\cdot\nabla u_h^m, \varphi_h\rangle_\Omega + k_m\langle \alpha u_h^m, \varphi_h\rangle_\Omega \\
&\quad + k_m \sum_{K\in\mathcal{T}_h} \delta_K\langle -\nabla\cdot(\varepsilon\nabla u_h^m + \boldsymbol{b}\cdot\nabla u_h^m + \alpha u_h^m, \boldsymbol{b}\cdot\nabla\varphi_h)\rangle_K \\
&\quad + \sum_{K\in\mathcal{T}_h} \delta_K\langle u_h^m, \boldsymbol{b}\cdot\nabla\varphi_h)\rangle_K \\
&= \langle u_h^{m-1}, \varphi_h\rangle_\Omega + \sum_{K\in\mathcal{T}_h} \delta_K\langle u_h^{m-1}, \boldsymbol{b}\cdot\nabla\varphi_h)\rangle_K + k_m\langle f(t_m), \varphi_h\rangle_\Omega \\
&\quad + k_m \sum_{K\in\mathcal{T}_h} \delta_K\langle f(t_m), \boldsymbol{b}\cdot\nabla\varphi_h\rangle_K \,.
\end{aligned} \tag{5.27}$$

From this result, it is evident that the stabilized dG(0)cG(1) method is equivalent to the stabilized implicit Euler method modulo quadrature error of nonlinear right–hand sides. This fact is used for the implementation of the primal problem.

By means of the error representation

$$\mathcal{J}(u) - \mathcal{J}(u_{kh}) \approx \eta := \sum_{m=1}^{M}\sum_{K\in\mathcal{T}_h} \eta_K^m \,,$$

with a user chosen target quantity $\mathcal{J}$ we design a solution algorithm that is adaptive in space as well as in time. Thereby, we obtain a hierarchy of sequently refined meshes $\mathcal{M}_i^m$, $i \geq 1$ starting with the initial mesh $\mathcal{M}_0^m$ which is identical for each time step m, i.e. $M_0^i = M_0^j \ \forall i, j = \{1, \ldots, M\}$. The corresponding finite element spaces are denoted by $\mathcal{V}_h^{m,i}$.

Adaptive solution algorithm

Initialization Set $i = 0$ and generate the initial finite element spaces.

Step 1 Solve the primal problem.

For $m = 1, \ldots, M$ find $u_h^{m,i} \in \mathcal{V}_h^{m,i}$ such that

$$A_S^{IE}(u_h^{m,i})(\varphi_h) = F^{IE}(\varphi_h) \quad \forall\varphi_h \in \mathcal{V}_h^{m,i} \,,$$

where A_S^{IE} and F^{IE} are the implicit Euler formulations of (5.8) defined in the left–hand side and right–hand side of (5.27), respectively.

Step 2 Solve the dual problem.

For $m = 1, \ldots, M$ find $z_H^{m,i} \in \mathcal{V}_H^{m,i} \supset \mathcal{V}_h^{m,i}$ such that

$$A_{S^*}^{CN}(\varphi_H, z_H^{m,i}) = \mathcal{J}'^{CN}(u_h^{m,i})(\varphi_H) \quad \forall \varphi_H \in \mathcal{V}_H^{m,i}\,.$$

$\mathcal{V}_H^{m,i}$ is the finite element space of higher order polynomials that corresponds to the refined mesh $\mathcal{M}_i^m$. $A_{S^*}^{CN}$ and $\mathcal{J}'^{CN}$ are the left–hand and right–hand side of the Crank Nicolson scheme applied to a stabilized and semi–discrete formulation in space of the dual problem (5.15). Note that the dual problem has to be converted into a forward problem.

Step 3 Evaluate the a–posteriori error estimate.

$$\begin{aligned}\eta_K^m = \int_{I_m} & \langle \mathcal{R}(u_h^{m,i}, z_H^{m,i} - \mathcal{I}_h z_H^{m,i}\rangle_K - \delta_K \langle \mathcal{R}(u_h^{m,i}, \boldsymbol{b}\cdot\nabla \mathcal{I}_h z_H^{m,i}\rangle_K \\ & - \langle \mathcal{E}(u_h^{m,i}), z_H^{m,i} - \mathcal{I}_h z_H^{m,i}\rangle_{\partial K}\, dt \\ & - \left\langle \left[u_h^i\right]_{m-1}, z_H^{m-1,i} - \mathcal{I}_h z_H^{m-1,i} \right\rangle_\Omega \\ & + \delta_K \left\langle \left[u_h^i\right]_{m-1}, \boldsymbol{b}\cdot\nabla I_h z_H^{m-1,i} \right\rangle_K ,\end{aligned}$$

where the cell and edge residuals are given in (5.25b) and (5.25c) and $[u_h^i]_0 := u_h^{1,i} - u_0$. $I_h z_H^{m,i} \in \mathcal{V}_h^{m,i}$ is the linear interpolation of $z_H^{m,i}$. The integrals over the time intervals I_m are approximated by the trapezoidal rule. We note this rule to be exact for the integrals to be calculated because $u_h^{m,i}$ is constant in time whereas $z_H^{m,i}$ is linear in time.

Step 4 Refinement strategy in time and space.

Compute $\eta^m := \sum_{K\in\mathcal{T}_h} |\eta_K^m|$ for each time step m. Mark the time intervals $\tilde{m}$ where $\tilde{m}$ belongs to the set of the time intervals m according to θ_t percent of the worst time indicators η^m ($\theta_t = 0.2$ is a useful value).
Choose $\theta_1 \leq \theta_2$ with $\theta_1, \theta_2 \in (1.0, 5.0)$.
For all time steps m:
Set $\eta_{max}^m = \max_{K\in\mathcal{T}_h} |\eta_K^m|$. If $m = \tilde{m}$ set $\mu^m = \theta_1 \frac{\sum_{K\in\mathcal{T}_h} |\eta_K^m|}{\#K^m}$ where $\#K^m$ denotes the number of elements in space in each time step m; otherwise, set $\mu^m = \theta_2 \frac{\sum_{K\in\mathcal{T}_h} |\eta_K^m|}{\#K^m}$.

```
while mu > eta_max:
    mu := mu/2.0
```

Mark the elements $\tilde{K}$ with $\left|\eta_{\tilde{K}}^m\right| > \mu^m$ to be refined. Generate a new mesh $\mathcal{M}_{i+1}^m$ by regular refinement. Half the length of the time intervals $\tilde{m}$.

Step 5 Check the exit condition.

If $\eta_{max} <$ TOL or $\eta <$ TOL is true, the *Adaptive solution algorithm* is completed; else increase i and go to Step 1.

Remark 5.8 (to Step 4)**.** *For θ_1, θ_2 chosen between* 1.0 *and* 5.0 *the resulting marking strategy is a good compromise between number of degrees of freedom and error reduction.*

In order to verify our implementation, we introduce a nonstationary counterpart of Example 4.6.

Example 5.9. *We study the given exact solution*

$$u(\boldsymbol{x}, t) = \frac{16}{\pi} \sin(\pi t) x_1(1-x_1)x_2(1-x_2) \cdot \left\{ \frac{\pi}{2} + \arctan\left(2\varepsilon^{-\frac{1}{2}}(r_0^2 - (x_1 - x_1^0)^2 - (x_2 - x_2^0)^2)\right)\right\},$$

to the model problem (5.1) on the space–time domain $\Omega \times I := (0,1)^2 \times (0, 0.5]$ with $r_0 = 0.25\,, x_1^0 = x_2^0 = 0.5$. We solve the equation by setting $\varepsilon = 10^{-6}\,, \boldsymbol{b} =$

$(2,3)^\top$ *without linear reaction contribution, i.e.* $\alpha = 0$. *The right–hand side* f *is calculated according to the given exact solution. The solution* u *is a hump changing its height in the course of time in the range of zero to a maximum of about* 0.997. *The boundary conditions as well as the initial condition are prescribed by the exact solution. The chosen target quantity is the* $\mathcal{L}^2$ *norm at the end time, i.e.*

$$\mathcal{J}(u) = \frac{1}{\|u(T)\|_{\mathcal{L}^2(\Omega)}} \langle e(T), u(T) \rangle_\Omega \,,$$

where u *and* $e = u - u_{kh}$ *are taken at the end time* T.

Due to the structure of the dual problem that is expressed by a behavior of $\mathrm{e}^{-\varepsilon C t}$ regarded from the viewpoint $t = T$ backwards in time, we expect the adaptive algorithm to refine the time intervals that are near to the end time $t = T$. The constant C depends on the eigenvalues of the corresponding operator. The more the space cells at a certain time are distant in time from the end time, the less refinement is to be expected in principle. The initial grid consists of 545 degrees of freedom in space which corresponds to a mesh size of $h = 0.0625$. The time interval $(0, T]$ is equally divided into 20 sub–intervals with a time step size of $k = 0.025$.

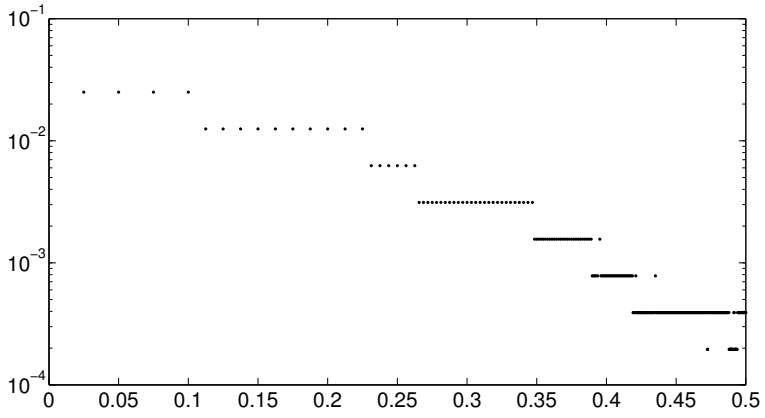


*Figure 5.1: (***Example 5.9***) Distribution of the temporal step size over the time* $(0, T]$.

Figure 5.1 and 5.2 present the distribution of the time cells and the degrees of freedom at each time step, respectively. We observe that our numerical computations confirm the expectations just described concerning the distribution of the refinement in time and in space governed by the derived error representation.

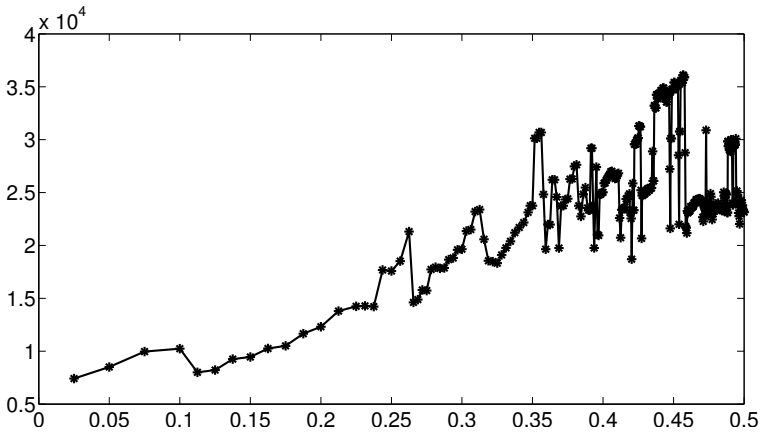


*Figure 5.2: (***Example 5.9***) Distribution of the degrees of freedom over the time* $(0, T]$.

The fact that our adaptive algorithm is capable to refine by preference in those temporal and spatial areas that have an influence on the quality of the approximation concerning the quantity of interest is the big advantage in comparison to global refinement in time and space. The mean value of the used degrees of freedom over the time $(0, 0.5]$ in the last adaptive iteration is 25368. For that reason, we compare the adaptively calculated solution by a numerical approximation with 33025 degrees of freedom at each time step. This comparison might be called more than fair. The stabilized solution on a uniformly refined mesh with 33025 is characterized by strong oscillations in terms of over– and undershoots as can be seen in Figure 5.3. The oscillations have a magnitude up to a maximum of 57 percent of the nominal value at the end time as well as over the whole time period for example at the midpoint. The exact maximum value of the solution u at $t = 0.25$ is 0.705.

We observe that the adaptively generated solution at mid–term with 15731 nodes also has characteristic oscillations. But in contrast to a solution on an uniformly refined mesh these oscillations are not necessarily undesirable because we are only interested in a quantity at the end time. Our error representation is able to organize the refinement in time and space in such a manner that the approximate solution at the end time $t = 0.5$ varies very slightly from the exact solution but the numerical solutions at time points that are not close to the end time are allowed to be of low resolution. This aspect emphasizes

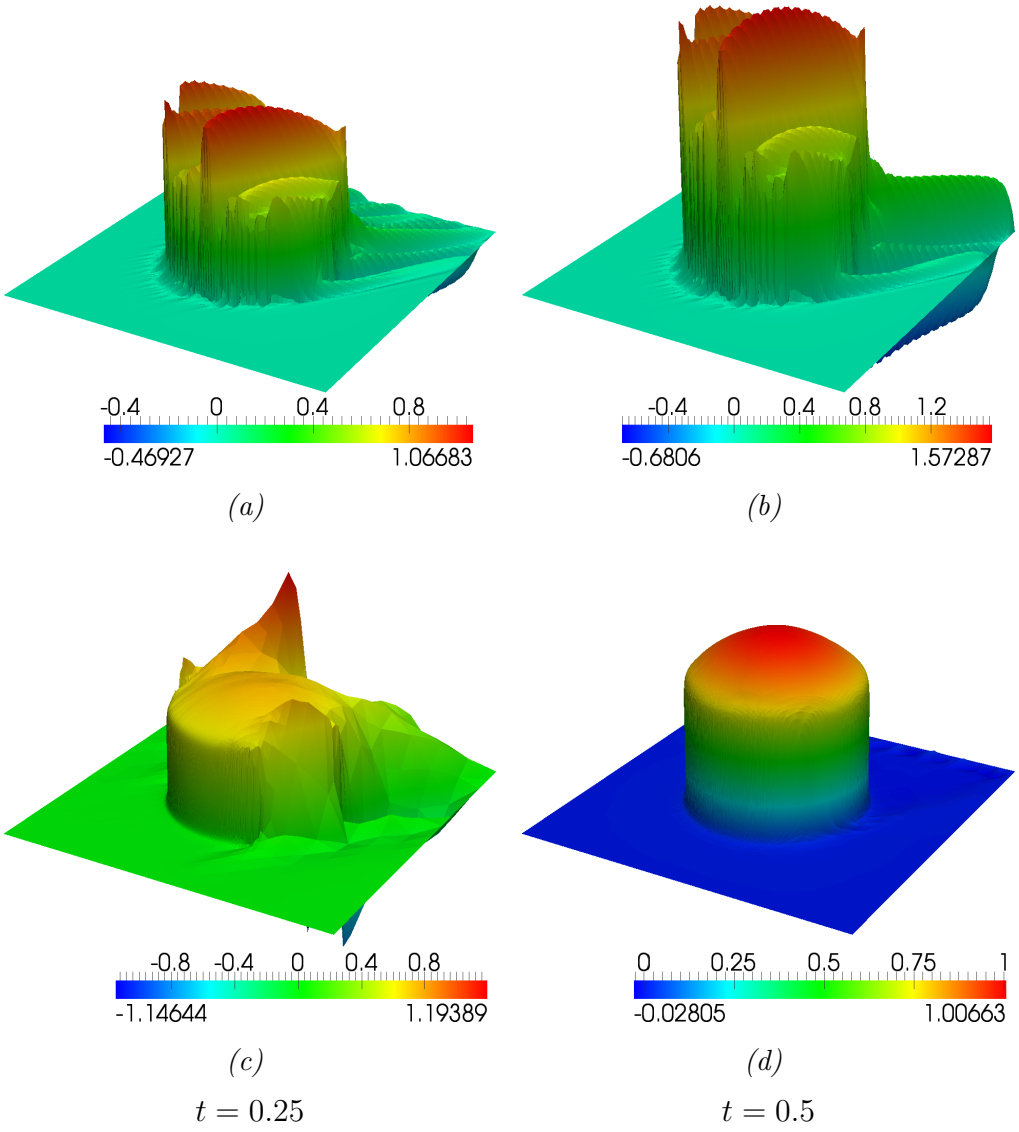
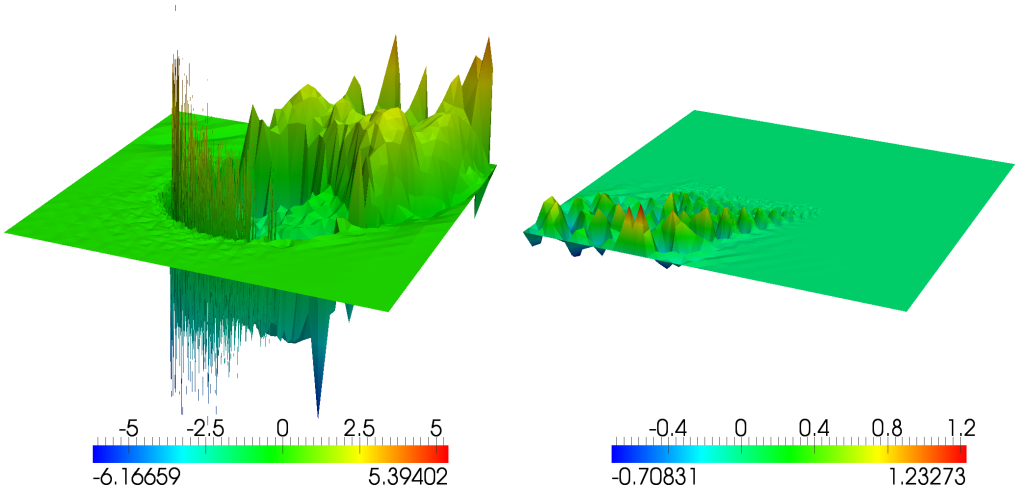


*Figure 5.3: (***Example 5.9***) Solution and value range on a globally refined mesh (a) and (c) with* 33025 *nodes and on an adaptively refined mesh controlled by* $\mathcal{J}$ *with (b)* 15731 *nodes and (d)* 23150 *nodes.*

the efficiency of our suggested method. The magnitude of the oscillations at $t = T$ lies in an area of 0.03 percent. As we note in the error representation (5.25a), the error description mainly consists of residual terms weighted by the dual solution. Considering Figure 5.4, we recognize that the weights are small in the beginning of the time interval whereas they make large contributions at later points in time. This is the reason why the adaptive algorithm affects refinement in time and in space at time points close to $t = 0.5$. This

behavior of the adaptive algorithm is confirmed in Figure 5.5. It illustrates the temporal development of the mesh during the adaptive refinement process. As expected, large values of the dual solution lead to intense cell refinement in the concerning areas whereas small values of the dual solution prevent refinement of these cells.



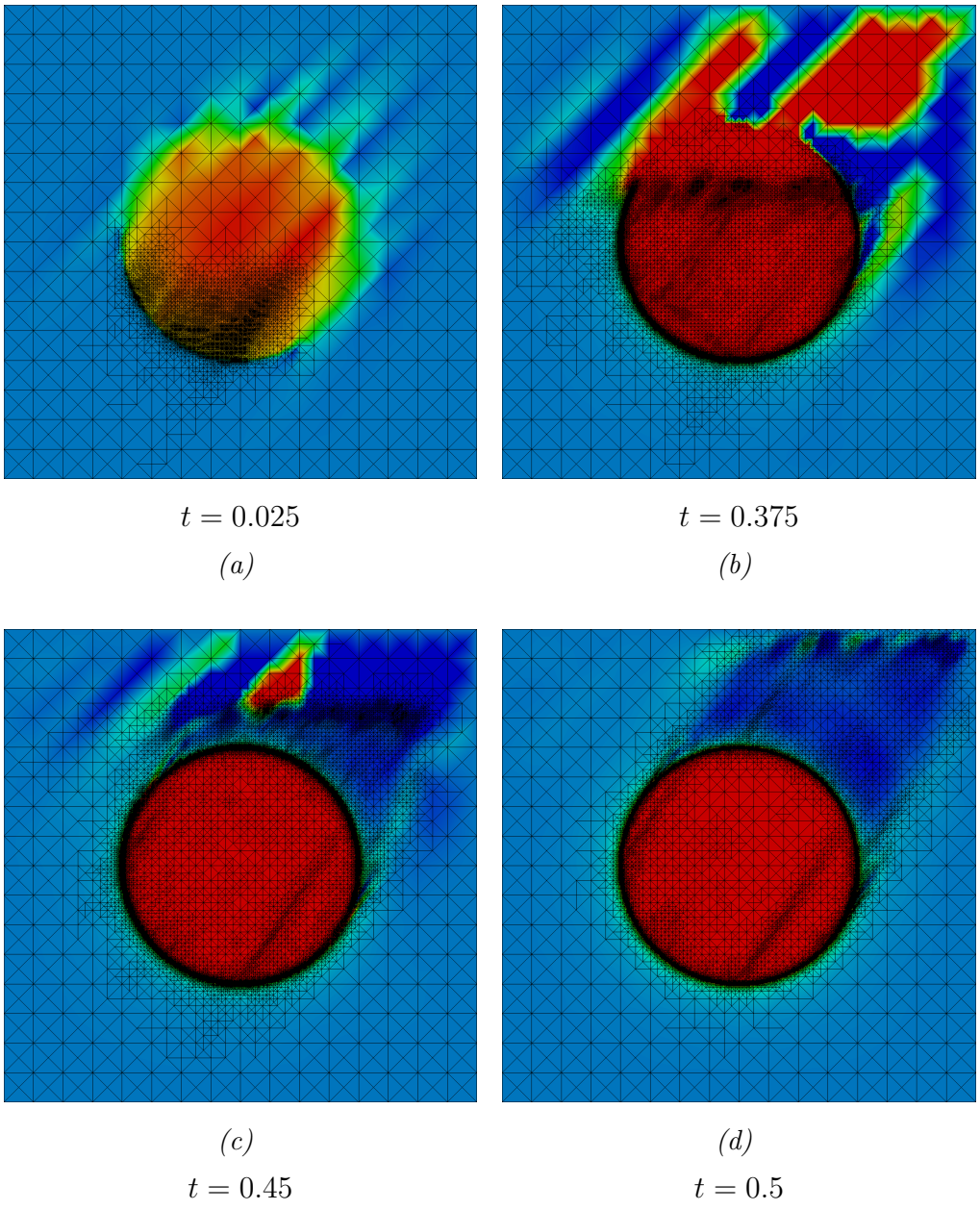
$t = T$ (a) $t = 0$ (b)

*Figure 5.4: (***Example 5.9***) Dual solution with exponential falling behavior from the viewpoint $t = T$.*

The results in Figure 5.6 show the development of the errors $\mathcal{J}(u) - \mathcal{J}(u_{kh})$ during an adaptive solution process. We observe that the error is continuously reduced in each iteration. In comparison to the solution on an uniformly refined grid with 33025 nodes, we obtain an error value $\mathcal{J}(e)$ of 0.1249 which performs rather badly compared to the use of the error representation generating an error value of 0.0067. In Table 5.1, we present the effectivity index during an adaptive run which is as usual defined by

$$\mathcal{I}_{\text{eff}} := \frac{\eta}{\mathcal{J}(u) - \mathcal{J}(u_{kh})} .$$

An effectivity index that tends to one means that the estimated error is an excellent approximation to the exact error. In Table 5.1, $\mathrm{M}_{\max}$ denotes the maximum number of time steps and $\mathrm{K}_{\max}$ is the maximum number of degrees of freedom that arise in the corresponding iteration.



$t = 0.025$ *(a)* $t = 0.375$ *(b)*

(c) $t = 0.45$ *(d)* $t = 0.5$

*Figure 5.5: (***Example 5.9***) Development of the mesh in the course of time.*

Example 5.10. *Now, we modify Example 5.9 and add a linear reaction term by setting* $\alpha = 1.0$. *Thus, the configuration of the* moving hump *is exactly the same as it is chosen in [2].*

The size of the spurious oscillations can be measured by

$$\mathrm{var}(t) := \max_{\boldsymbol{x}\in\Omega} u_{kh}(\boldsymbol{x}, t) - \min_{\boldsymbol{x}\in\Omega} u_{kh}(\boldsymbol{x}, t)\,,$$

where the maximum and minimum are taken only in the vertices of the mesh

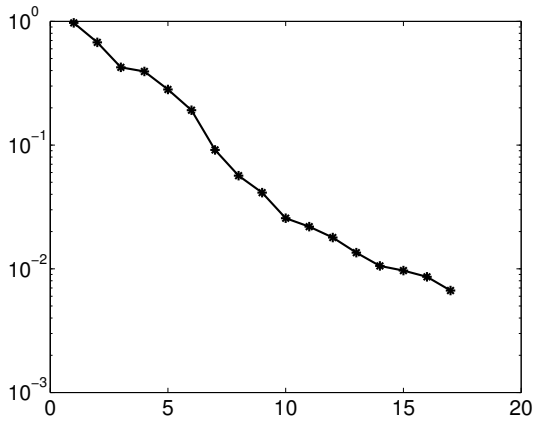


*Figure 5.6: (***Example 5.9***) DWR iterations vs. errors.*

it	M_{max}	K_{max}	$\mathcal{I}_{eff}$	it	M_{max}	K_{max}	$\mathcal{I}_{eff}$
1	25	735	0.55	9	95	6184	1.03
2	29	1002	0.63	10	113	7785	1.11
3	34	1255	0.78	11	135	9051	1.04
4	40	1786	1.07	12	161	12557	1.03
5	47	2300	1.11	13	193	16484	1.00
6	56	3067	1.13	14	231	20192	1.02
7	67	3857	1.26	15	277	25415	1.01
8	80	4611	1.23	16	332	36091	1.01

*Table 5.1: (***Example 5.9***) Effectivity indices with respect to $\mathcal{J}$ during the DWR iterations* it.

cells. In [2], the local projection stabilization method for the space discretization is combined with variational time discretizations of Galerkin–type. The use of continuous Galerkin–Petrov (cGP) methods and discontinuous Galerkin (dG) methods with underlying spaces of polynomials of different degrees are compared.

If we are interested in the quality of the approximate solution at the end time, we will regard the value of var at $t = 0.5$ where the adaptive solution process is governed by

$$\mathcal{J}_1(u) = \int_\Omega u(\boldsymbol{x}, T)\, d\boldsymbol{x}\,.$$

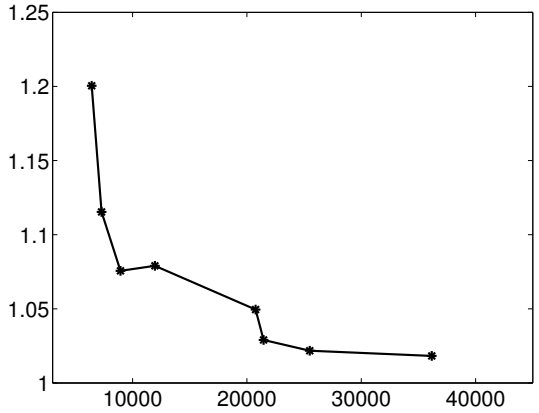


*Figure 5.7: (***Example 5.10***) Measure of the oscillations* var(0.5) *in dependence of the number of nodes at* $t = 0.5$.

dofs	var(0.5)	$\mathcal{I}_{eff}$	dofs	var(0.5)	$\mathcal{I}_{eff}$
545	9.392	1.03	6425	1.200	1.02
833	9.432	1.17	7283	1.115	1.00
1267	5.852	1.75	8939	1.076	0.96
1535	6.761	9.13	11954	1.079	0.97
2325	4.328	2.82	20757	1.050	0.99
2522	3.270	1.08	21455	1.030	1.03
3031	2.105	1.47	25492	1.022	1.05
3937	1.568	0.94	36167	1.018	1.07

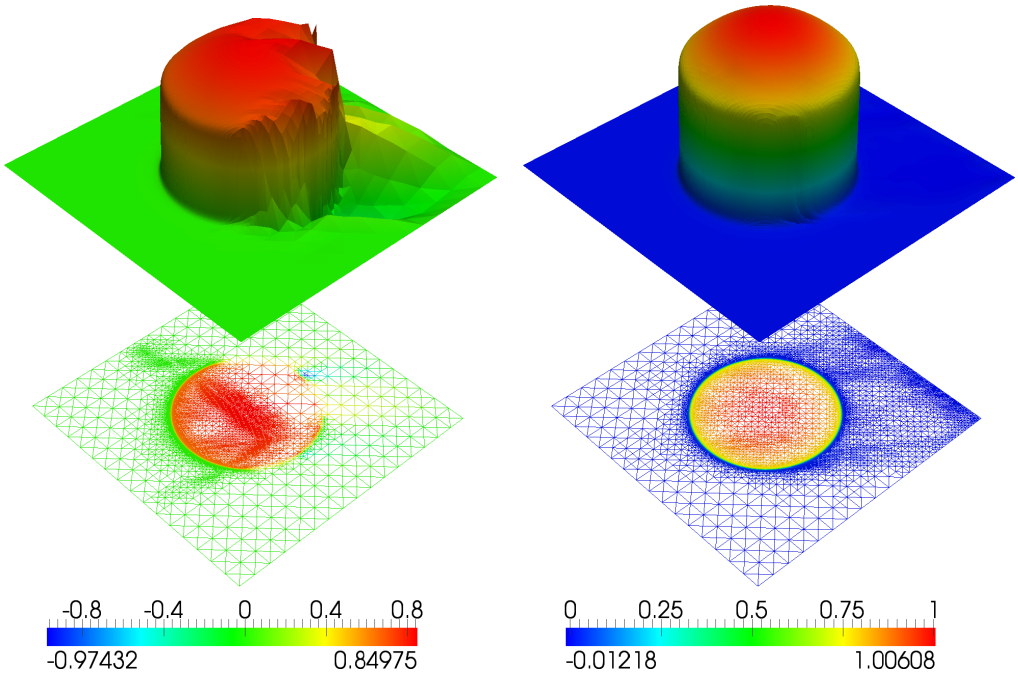
*Table 5.2: (***Example 5.10***) Degrees of freedom at* $t = 0.5$, var(0.5) *and effectivity index.*

The optimal value of var(0.5) is given by 0.997. Figure 5.7 presents the measured oscillations in dependence of the degrees of freedom at $t = 0.5$. In addition, Table 5.2 illustrates the effectivity indices with respect to the target functional $\mathcal{J}_1$. It aims at showing that the effectivity index tends to one and is not influenced by the additional linear reaction term.

Table 5.3 gives an overview of the quantity var(0.5) that is used in several publications to measure the size of the spurious oscillations of the approximate solution. The test setting of Example 5.10 is exactly the same in each publication. In [30], comparative studies of stabilized finite element methods

Method	Reference	var(0.5)	dofs	k
SUPG	[30]	1.3835	16641	10^{-3}
LPS	[30]	1.2007	32768	10^{-3}
SUPG	[6]	1.2504	33025	$2 \cdot 10^{-3}$
SUPG/SC	[6]	1.1946	33025	$2 \cdot 10^{-3}$
LPS/cGP(1)	[2]	1.0408	33025	10^{-3}
LPS/dG(1)	[2]	1.0408	33025	10^{-3}
SUPG/DWR	this work	1.0790	10900	$3.1 \cdot 10^{-3}$
SUPG/DWR	this work	1.0179	35931	$1.5 \cdot 10^{-3}$

*Table 5.3: (***Example 5.10***)* var(0.5) *for different stabilized approximation schemes.*



$t = 0.25$ $t = 0.5$

(a) *(b)*

*Figure 5.8: (***Example 5.10***) Solution with mesh governed by $\mathcal{J}_1$ (a) at half-time and (b) at end time.*

for solving time–dependent convection–diffusion–reaction equations with small diffusion is presented. These studies include the SUPG method and the local projection scheme, for example. The calculations were performed on a regular mesh with a number of degrees of freedom as indicated in Table 5.3. In ad-

dition to results obtained by the SUPG method, the application of additional shock–capturing stabilization is considered in [6]. These simulations were also performed on a regular mesh with a mesh size of $5.524 \cdot 10^{-3}$ which corresponds to 33025 degrees of freedom. In [30], [6] and [2], the time step size k is equidistantly chosen as indicated in Table 5.3. In our work, the time step size as well as the mesh size are adaptively refined. In Table 5.3, the average values of the step sizes in time and in space are presented. In [30] and [2], the performance of the stabilization methods is tested with respect to the design of the stabilization parameters. For the comparative Table 5.3, we always chose the best value that was reached with the particular method. Compared to all other results, our proposed method established the best results.

Figure 5.8 presents the approximate solution at two different time points. We observe that the error representation organizes the refinement process in such a way that the solution at the end time is very close to the exact solution whereas the solution at earlier time steps is allowed to possess oscillations that are not necessarily undesirable.

Chapter 6

Conclusions & outlook

To sum our work up, we sketch the main challenges of the previous chapters. Finally, we outline future work to be done.

6.1 Summary & conclusions

In this work, we developed and carefully analyzed a combined approach of stabilized finite element approximations and an a–posteriori error control with adaptive mesh refinement for convection–dominated stabilized nonlinear transport problems in the stationary and nonstationary case. We pointed out that our error representation based on the dual weighted residual method does not depend on undetermined constants which usually leads to absolute impracticality of the error bounds. The associated dual problem provides information about the weighting factors of the residual terms such that the error is quantifiable. Furthermore, our method contrasts strongly with conventional error estimates since the error is given in terms of a user chosen quantity of interest. This emphasizes the practice–oriented applicability of our proposed method.

An obvious and often mentioned disadvantage of the dual weighted residual method is that an additional dual problem has to be solved. As highlighted in the previous remarks, compared to solving the nonlinear primal problem on a grid with a sufficient number of degrees of freedom, it is cheap to compute the solution of the linear dual problem. We want to emphasize again that the dual problem is always linear even if the primal problem is nonlinear. For

that reason, especially in our case of a nonlinear model problem, our approach gains efficiency. We are aware of the fact that we invest much effort to achieve such an precise error representation. Departing from our approach of high accuracy might lead to tremendous loss of precision with regard to the sensitive convection–dominated transport problems. Compared to the bigger part of research related to the DWR method, this work develops an error representation that involves approximations as less as possible.

The proposed method is implemented using the FEniCS toolbox; cf. [37]. As illustrated in the previous chapters, the numerical results feature an effectivity index that tends to one. We also showed that the adaptive mesh refinement strategy based on the introduced error representation is well–adjusted such that the error is minimized with respect to a small number of degrees of freedom. The numerical results point out that it is justified to neglect the remainder terms and to use an higher order approximation to the dual solution. In addition, it turned out that the proposed method is robust against the choice of the stabilization parameters which often is discussed as crucial issue with respect to stabilization techniques.

6.2 Outlook

Even if the proposed method was presented very detailed, there is still space for further development. This includes the extension of the nonstationary framework to nonlinear partial differential equations as well as to systems of convection–diffusion–reaction equations. Additionally, the treatment of nonhomogeneous Dirichlet boundary conditions in the nonstationary case is not described. In this work, we did not discuss higher order stable variational time methods and the possible benefit of them nor alternative stabilization schemes. Finally, our method can be used beyond transport problems and might be extended to Navier–Stokes equations or even systems consisting of Navier–Stokes equations coupled with convection–diffusion–reaction models.

The main difficulty with respect to the nonlinear nonstationary equation can be seen in the rapidly growing compute time because in each time step sev-

eral linear solving steps are required. For that reason, this problem can be sorted out in the field of computational science. In this work, the linear systems are solved by a direct solver. Direct solvers perform very well for small problems especially in two dimensions. For three dimensional problems with a huge number of degrees of freedom, memory consumption and run time behavior which is $\mathcal{O}(dofs^2)$ are impractical. For that reason, iterative solvers are needed for more sophisticated applications. The development of problem–specific preconditioners is required to obtain a small iteration number of the iterative solver which is a challenging task. An even more challenging task is to construct preconditioners which additionally are able to run in parallel.

In view of systems of convection–diffusion–reaction equations as stated in (1.1), the proposed method can be extended to the systems straight forward. Some questions that may arise in the context of systems are how to deal with the different grids or with individual error representations. The dual weighted residual method is suitable especially for systems since the dual system is decoupled and, hence, can easily be solved.

Specific considerations will have to be taken if nonhomogeneous Dirichlet boundary conditions are included with respect to the nonstationary model problem. Additional terms have to be implemented in order to measure the impact of the nonhomogeneous boundary conditions on the error representation.

As explained in this work, the second order time discretization scheme is equivalent to the Crank Nicolson time stepping method modulo nonlinear right–hand side. This approach is an A–stable method but it is not strongly A–stable. This fact might lead to undesirable oscillations in time which are not damped in contrast to strongly A–stable or L–stable methods. For that reason, pairs of discretization spaces in time for the primal and dual problem have to be used such that the desired stability properties and orders are obtained.

The SUPG method is the most common stabilization scheme used in the field of convection–dominated transport problems. Of course, other stabilization

techniques like the local projection method (LPS) can be combined with the DWR method.

As mentioned in the introductory chapter, a lot of practice–related systems in the scope of flow dynamics are modeled by the Navier–Stokes equation coupled with a transport equation. Engineers are more and more interested in accurate and fast solution of such systems in order to understand the behavior of their devices. Therefore, the next step would be to apply the proposed method to this kind of systems. All the described considerations above have to be included when carrying over the method to coupled Navier–Stokes and transport problems.

Bibliography

[1] Adams, R. A.: *Sobolev Spaces*. New York: Academic Press, 1975

[2] Ahmed, N. ; Matthies, G.: Numerical studies of Galerkin–type time–discretizations applied to transient convection–diffusion–reaction equations. In: *World Academy of Science, Engineering and Technology* 6 (2012), pp. 549–556

[3] Alt, H. W.: *Lineare Funktionalanalysis*. Berlin: Springer, 1999

[4] Augustin, M. ; Caiazzo, A. ; Fiebach, A. ; Fuhrmann, J. ; John, V. ; Linke, A. ; Umla, R.: An assessment of discretizations for convection–dominated convection–diffusion equations. In: *Computer Methods in Applied Mechanics and Engineering* (2011), pp. 3395–3409

[5] Bangerth, W. ; Geiger, M. ; Rannacher, R.: Adaptive Galerkin Finite Element Methods for the Wave Equation. In: *Computational Methods in Applied Mathematics* 10 (2010), pp. 3–48

[6] Bause, M. ; Schwegler, K.: Analysis of stabilized higher–order finite element approximation of nonstationary and nonlinear convection–diffusion–reaction equations. In: *Computer Methods in Applied Mechanics and Engineering* 209–212 (2012), pp. 184–196

[7] Becker, R.: An optimal–control approach to a posteriori error estimation for finite element discretizations of the Navier–Stokes equations. In: *East–West Journal of Numerical Mathematics* 8 (2000), pp. 257–274

[8] Becker, R. ; Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. In: *Acta Numerica* (2001), pp. 1–102

[9] Besier, M. ; Rannacher, R.: Goal–oriented space–time adaptivity in the finite element Galerkin method for the computation of nonstationary incompressible flow. In: *International Journal for Numerical Methods in Fluids* 458 (2012), pp. 2735–2757

[10] Brooks, A. N. ; Hughes, T. J. R.: Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. In: *Computer Methods in Applied Mechanics and Engineering* 32 (1982), pp. 199–259

[11] Brunner, F. ; Radu, F. A. ; Bause, M. ; Knabner, P.: Optimal order convergence of a modified BDM_1 mixed finite element scheme for reactive transport in porous media. In: *Advances in Water Resources* 35 (2012), pp. 163–171

[12] Ciarlet, P. G.: *The Finite Element Methods for Elliptic Problems.* Amsterdam: North–Holland, 1978

[13] Ciarlet, P. G.: *The Finite Element Method for Elliptic Problems.* Philadelphia: Society for Industrial and Applied Mathematics (SIAM), 2002 (Classics in Applied Mathematics, Vol. 40)

[14] Di Pietro, D. A. ; Ern, A.: *Mathematical Aspects of Discontinuous Galerkin Methods.* Berlin: Springer, 2010 (Mathématiques et Applications)

[15] Dörfler, W.: A convergent adaptive algorithm for Poisson's equation. In: *SIAM Journal on Numerical Analysis* 33 (1996), pp. 1106–1124

[16] Eriksson, K. ; Johnson, C. ; Logg, A.: Adaptive Computational Methods for Parabolic Problems. In: *Encyclopedia of Computational Mechanics* (2004)

[17] Ern, A. ; Guermond, J.-L.: *Theory and Practice of Finite Elements.* 159. New York: Springer, 2004 (Applied Mathematical Sciences)

[18] Ern, A. ; Stephansen, A. ; Vohralík, M.: Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection–diffusion–

reaction problems. In: *Journal of Computational and Applied Mathematics* 234 (2010), pp. 114–130

[19] Evans, L. C.: *Partial Differential Equations*. Providence, Rhode Island: American Mathematical Society, 2010 (Graduate Studies in Mathematics, Vol. 19)

[20] Franca, L. P. ; Valentin, F.: On an improved unusual stabilized finite element method for the advective–reactive–diffusive equation. In: *Computer Methods in Applied Mechanics and Engineering* 190 (2000), pp. 1785–1800

[21] Hauke, G. ; Doweidar, M. H. ; Fuster, D. ; Gómez, A. ; Sayas, J.: Application of variational a–posteriori multiscale error estimation to higher–order elements. In: *Computational Mechanics* 38 (2006), pp. 356–389

[22] Hauke, G. ; Fuster, D. ; Doweidar, M. H.: Variational multiscale a–posteriori error estimation for multi–dimensional transport problems. In: *Computer Methods in Applied Mechanics and Engineering* 197 (2008), pp. 2701–2718

[23] Hemker, P. W.: A singularly perturbed model problem for numerical computation. In: *Journal of Computational and Applied Mathematics* 76 (1996), pp. 277–285

[24] Houston, P. ; Süli, E.: Stabilized hp-finite element approximation of partial differential equations with nonnegative characteristic form. In: *Computing 66* (2001), pp. 99–119

[25] John, V.: A numerical study of a posteriori error estimators for convection–diffusion equations. In: *Computer Methods in Applied Mechanics and Engineering* 190 (2000), pp. 757–781

[26] John, V. ; Knobloch, P.: On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. In: *Computer Methods in Applied Mechanics and Engineering* 196 (2007), pp. 2197–2215

[27] John, V. ; Knobloch, P. ; Savescu, S. B.: A posteriori optimization of parameters in stabilized methods for convection–diffusion problem – Part I. In: *Computer Methods in Applied Mechanics and Engineering* 200 (2011), pp. 2916–2929

[28] John, V. ; Maubach, J. M. ; Tobiska, L.: Nonconforming streamline–diffusion–finite–element–methods for convection–diffusion problems. In: *Journal of Numerical Mathematics* 78 (1997), pp. 165–188

[29] John, V. ; Novo, J.: A robust SUPG norm a posteriori error estimator for stationary convection–diffusion equations. In: *Computer Methods in Applied Mechanics and Engineering* 255 (2013), pp. 289–305

[30] John, V. ; Schmeyer, E.: Finite element methods for time–dependent convection–diffusion–reaction equations with small diffusion. In: *Computer Methods in Applied Mechanics and Engineering* 198 (2008), pp. 475–494

[31] Jones, W. P. ; Launder, B. E.: The prediction of laminarization with a two–equation model of turbulence. In: *International journal of heat and mass transfer* 15 (1972), pp. 301–314

[32] Knabner, P. ; Angermann, L.: *Numerik partieller Differentialgleichungen*. Berlin Heidelberg: Springer, 2000

[33] Knopp, T. ; Lube, G. ; Rapin, G.: Stabilized finite element methods with shock capturing for advection–diffusion problems. In: *Computer Methods in Applied Mechanics and Engineering* 191 (2002), pp. 2997–3013

[34] Kuzmin, D.: Explicit and implicit FEM-FCT algorithms with flux linearization. In: *Journal of Computational Physics* 228 (2009), pp. 2517–2534

[35] Lions, J.-L. ; Magenes, E.: *Problèmes aux Limites non Homogènes et Applications*. 1. Paris: Dunod, 1968

[36] Lions, J.-L. ; Magenes, E.: *Problèmes aux Limites non Homogènes et Applications*. 2. Paris: Dunod, 1968

[37] Logg, A. ; Mardal, K. A. ; Wells, G.: *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book*. Springer, 2012 (Lecture Notes in Computational Science and Engineering)

[38] Lube, G. ; Rapin, G.: Residual–based stabilized higher–order FEM for advection–dominated problems. In: *Computer Methods in Applied Mechanics and Engineering* 195 (2006), pp. 4124–4138

[39] Meidner, D. ; Rannacher, R. ; Vihharev, J.: Goal–oriented error control of the iterative solution of finite element equations. In: *Journal of Numerical Mathematics* 17 (2009), pp. 143–172

[40] Quarteroni, A. ; Valli, A.: *Numerical Approximation of Partial Differential Equations*. 23. Berlin Heidelberg: Springer, 2008 (Computational Mathematics)

[41] Roos, H.-G. ; Stynes, M. ; Tobiska, L.: *Robust numerical methods for singularly perturbed differential equations, Convection–Diffusion–Reaction and Flow Problems*. Berlin Heidelberg: Springer, 2008

[42] Roubíček, T.: *Nonlinear Partial Differential Equations with Applications*. Basel: Birkhäuser, 2005 (ISNM International Series of Numerical Mathematics)

[43] Sangalli, G.: Robust a–posteriori estimator for advection–diffusion–reaction problems. In: *Mathematics of Computation* 77 (2007), pp. 41–70

[44] Schmich, M. ; Vexler, B.: Adaptivity with Dynamic Meshes for Space–Time Finite Element Discretizations of Parabolic Equations. In: *Journal of Scientific Computing* 30 (2008), pp. 369–393

[45] Thomée, V.: *Galerkin Finite Element Methods for Parabolic Problems*. Berlin Heidelberg: Springer, 2006 (Springer Series in Computational Mathematics, Vol. 25)

[46] Uzunca, M. ; Karasözen, B. ; Manguoğlu, M.: Adaptive discontinuous Galerkin methods for non–linear diffusion–convection–reaction equations. In: *Computers & Chemical Engineering* 68 (2014), pp. 24–37

[47] Verfürt, R.: Robust a posteriori error estimates for stationary convection–diffusion equations. In: *SIAM Journal of Numerical Analysis* 43 (2005), pp. 1766–1782

[48] Verfürth, R.: A posteriori error estimators for convection–diffusion equations. In: *Numerische Mathematik* 80 (1998), pp. 641–663

[49] Zeidler, E.: *Applied Functional Analysis: Main Principles and Their Applications*. New York: Springer, 1995 (Applied Mathematical Sciences, Vol. 109)