

Eliminating the Gibbs Phenomenon

The Non-linear Petrov-Galerkin Method for the
Convection-Diffusion-Reaction Equation



Sarah Roggendorf

Supervisors: Dr. Kristoffer G. van der Zee

Prof. Paul Houston

School of Mathematical Sciences

University of Nottingham

This thesis is submitted for the degree of

Doctor of Philosophy

November 2019

It's a dangerous business, Frodo, going out your door. You step onto the road,
and if you don't keep your feet, there's no knowing where you might be swept off
to. - The Lord of the Rings, J.R.R. Tolkien [[163](#), Book One, Chapter 3]

Acknowledgements

I would like to use this opportunity to thank those people who have supported and encouraged me throughout my time as a PhD student at the University of Nottingham turning it into an enjoyable and hugely inspiring experience.

First and foremost, I would like to express my gratitude to my supervisors Dr Kris van der Zee and Professor Paul Houston, without whose support and guidance this thesis would not have been possible. Both have been a constant source of advice and encouragement. I would like to thank Kris for suggesting this challenging but exciting and inspiring research topic, for his unfailing enthusiasm, for always seeing the success even in unexpected results, for long discussions of ideas and complicated proofs, and for always pushing me to go one step further in my research. I would like to thank Paul for his constant and unfailing support, for his invaluable feedback, for always suggesting new things to try, for giving my supervision such a high priority even amidst his duties as Head of School, for his guidance regarding the implementation of the method, and for spotting every weakness in my research.

Next, I would like to thank my examiners Dr Max Jensen and Dr Matthew Hubbard for their feedback and comments on my thesis. Likewise, I thank Dr Donald Brown for the encouraging discussions of my research during my yearly reviews. Moreover, I would like to thank Matthew both for accompanying me throughout my PhD studies as my assessor and for supporting me as my teaching mentor ensuring that my debut as module convener could go smoothly.

Furthermore, I would like to thank my friends and colleagues in the School of Mathematical Sciences at the University of Nottingham. In particular I would like to thank Dr Robert Benkel, Dr Cerian Brewer, Dr Lee Curtin, Jonathan Marsh, Dr Emma McIvor, Dr Jakub Köry, Christopher Miles, Dr Sam Morley, Dr Laurence Shaw, Holly Smith, Dr Oliver Sutton and the Malt-Cross-Quiz-Team. I also express my gratitude to Dr David Hodge for being a good friend and an unofficial academic mentor always providing an outside perspective. Many thanks to Dave Parkin for

his excellent IT-support – especially regarding the installation of the finite element library Hermes2D on one of the servers.

Travelling, presenting at conferences and exchanging ideas is an essential part of research. Therefore, I would like to express my gratitude to the IMA for awarding me a Studentship to attend MAFELAP 2016 and to the Graduate School of the University of Nottingham and SIAM for travel support to attend the SIAM Annual Meeting 2018. Furthermore, I would like to thank Professor Carsten Carstensen, Professor Leszek Demkowicz and Professor Norbert Heuer for organizing an Oberwolfach seminar on DPG methods in 2017. The insights from this seminar as well as the discussions with my fellow participants have had a significant impact on my PhD project. Furthermore, I would like to thank Dr Gabriel Barrenechea, Dr Jesse Chan, Professor Leszek Demkowicz, Dr John Evans, Professor Jay Gopalakrishnan, Professor Norbert Heuer, Dr Brendan Keith, Professor Ignacio Muga and Professor Martin Stynes for fruitful discussions on the topic of this thesis.

My warmest thanks for being great company and good friends go to my housemates Eleanor Bellows, Louisa Cox, Rebecca Hock and Peter Jessop. Thanks to the Larsson family for taking me away on a holiday when I most needed it. Thanks to my church family at St. Mary's Wollaton Park, especially Eloise and Struan Ayre and Sarah Sim, for providing me with a family away from home. A very special thanks goes to my friends Mareike Lager, Rafaela Warkentin and Caila Lane. Moving to a different country without knowing anyone there would be so much harder without those friendships that outlast time and distance.

I also want to thank my family, especially my parents, who have supported me through every step of my life in every possible way always providing a safety net and a place to return home to.

Finally, I would like to express my heartfelt thanks to Jonathan Larsson for his love and support, for always pushing me to go on and never give up, for randomly turning up on my doorstep and never doubting me.

Abstract

In this thesis we consider the numerical approximation of the convection-diffusion-reaction equation. One of the main challenges of designing a numerical method for this problem is that boundary and interior layers typically present in the convection-dominated case can lead to non-physical oscillations in the numerical approximation, often referred to as Gibbs phenomena. The aim of this thesis is to develop a numerical method that eliminates Gibbs phenomena in the numerical approximation.

We consider a weak formulation of the partial differential equation of interest in L^q -type Sobolev spaces, with $1 < q < \infty$. We then apply a non-standard, non-linear Petrov-Galerkin discretization, that is applicable to reflexive Banach spaces such that the space itself and its dual are strictly convex. Similar to discontinuous Petrov-Galerkin methods, this method is based on minimizing the residual in a dual norm. By replacing the intractable dual norm by a suitable discrete dual norm gives rise to a non-linear inexact mixed method. This generalizes the Petrov-Galerkin framework developed in the context of discontinuous Petrov-Galerkin methods to more general Banach spaces, paving the way for designing finite element methods in non-standard function spaces. This, in particular, permits the selection of a solution space in which the best approximation of the solution has desirable properties.

For the convection-diffusion-reaction equation, we obtain a generalization of a similar approach from the L^2 -setting to the L^q -setting and discuss the choices we have made regarding the continuous and discrete test spaces and the corresponding norms. A key advantage of considering a more general Banach space setting is that, in certain cases, the oscillations in the numerical approximation vanish as q tends to 1, as we will demonstrate using a few simple numerical examples.

We furthermore demonstrate that the approximations obtained with our scheme qualitatively behave like the L^q -best approximation of the analytical solution in the same finite element space. We use this observation to study more closely in which cases the oscillations in the numerical approximation vanish as q tends to

1. To this end, we investigate Gibbs phenomena in the context of the L^q -best approximation of discontinuities in finite element spaces with $1 \leq q < \infty$. Using carefully selected examples, we show that on certain meshes the Gibbs phenomenon can be eliminated in the limit as q tends to 1. We then use these results to design the underlying meshes of the finite element spaces employed for our numerical scheme for the convection-diffusion-reaction equation such that Gibbs phenomena in the numerical approximation are eliminated.

While it is classical in the context of finite element methods to consider the solution of the convection-diffusion-reaction equation in the Hilbert space $H_0^1(\Omega)$, the Banach Sobolev space $W_0^{1,q}(\Omega)$, $1 < q < \infty$, has received very little attention in this context. However, it is more general allowing for less regular solutions and, moreover, it allows us to consider the non-linear Petrov-Galerkin method that forms the centre of this research. In this thesis, we therefore also present a well-posedness theory for the convection-diffusion-reaction equation in the $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ functional setting, $1/q + 1/q' = 1$. The theory is based on directly establishing the inf-sup conditions which are essential to the analysis of the non-linear Petrov-Galerkin method. Apart from a standard assumption on the advection and reaction coefficients, the other key assumption pertains to a subtle regularity requirement for the standard Laplace operator.

Contents

List of Figures	xiii
Notation	xv
1 Introduction	1
1.1 Summary of Results	4
1.1.1 The L^q -Best Approximation of Discontinuities in Finite Element Spaces	5
1.1.1.1 Problem Statement	5
1.1.1.2 Main Results	6
1.1.2 Well-posedness of the Convection-Diffusion-Reaction Equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$	7
1.1.3 The Non-linear Petrov-Galerkin Method for the Convection-Diffusion-Reaction Equation	9
1.1.3.1 Problem Statement	9
1.1.3.2 Summary of Results	10
1.2 Structure of the Thesis	11
2 Overview of Numerical Methods for the Convection-Diffusion-Reaction Equation	13
2.1 Finite Difference Methods	14
2.2 Finite Volume Methods	15
2.2.1 Numerical Flux Functions	17
2.2.2 High Resolution Methods	18
2.3 Stabilized Finite Element Methods	19
2.3.1 Standard Finite Element Approximation	20
2.3.2 Loss of Stability	21
2.3.3 Continuous Stabilized Finite Element Methods	22
2.3.4 Discontinuous Galerkin Finite Element Method	25

2.4	Discontinuous Petrov-Galerkin Methods	27
2.5	Methods for Approximating PDEs in L^1	28
3	Preliminaries	31
3.1	Banach Spaces and Hilbert Spaces	31
3.1.1	Banach spaces	32
3.1.2	Hilbert Spaces	33
3.2	Function Spaces	34
3.2.1	Continuously Differentiable Functions	34
3.2.2	Sobolev Spaces	35
3.3	Duality Mappings	37
3.3.1	Properties Duality Mappings	38
3.3.2	Examples of Duality Mappings on Sobolev Spaces	39
3.4	Characterization of L^q -Best Approximation	41
3.5	Closed Operators and the Closed Range Theorem	44
3.6	Graph Spaces and Trace Operators	47
4	Gibbs Phenomena in the L^q-Best Approximation of Discontinuities in Finite Element Spaces	51
4.1	Introduction	51
4.1.1	Motivation	51
4.1.2	Problem Statement	54
4.1.3	Summary of Results	55
4.1.4	Characterization of Best L^q -Approximation	62
4.1.5	Outline of the Chapter	63
4.2	Best Approximation of a Boundary Discontinuity in One Dimension	63
4.2.1	L^1 -Best Approximation	64
4.2.2	L^q -Best Approximation	66
4.2.3	Sufficient Conditions on General Meshes	67
4.2.4	A Three-Element Mesh in One Dimension	71
4.3	Over- and Undershoots at Jump Discontinuities	74
4.3.1	L^1 -Best Approximation without Over- or Undershoots	75
4.3.2	L^1 -Best Approximation with Over- and Undershoots	77
4.3.3	L^q -Best Approximation	79
4.4	Best Approximation of a Boundary Discontinuity in Two Dimensions	80
4.4.1	L^1 -Best Approximation	81
4.4.2	L^q -Best Approximation	85

4.4.3	L^q -Best Approximation on Meshes 2, 3 and 4	87
4.5	Numerical Examples	90
4.5.1	Refinement of the Mesh	91
4.5.1.1	Gibbs Phenomenon on Meshes in Two Dimensions	91
4.5.1.2	Gibbs Phenomenon on Meshes in One Dimension .	91
4.5.2	(Vanishing) Overshoot in One Dimension	92
4.5.3	(Vanishing) Overshoot in Two Dimensions	93
4.5.3.1	Overshoot on Meshes 2, 3 and 4 from Section 4.4.3	93
4.5.3.2	Overshoot on Unstructured Meshes	94
5	The Convection-Diffusion-Reaction Equation in Non-Hilbert So-	
	bolev Spaces: A Direct Proof of the Inf-Sup Condition	97
5.1	Introduction	97
5.1.1	Problem Statement	98
5.1.2	Assumptions	99
5.1.3	Main Results	100
5.1.4	Outline of the Chapter	101
5.2	The Poisson Equation and the Advection-Reaction Equation in $L^q(\Omega)$	102
5.2.1	The Poisson Problem in the $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ -setting . . .	102
5.2.2	Lower Order Terms	105
5.3	Proof of Theorem 5.4: Well-posedness of the Convection-Diffusion-	
	Reaction Equation in $W_0^{1,q}(\Omega)$	107
5.3.1	Continuity of the Bilinear Form	108
5.3.2	A Test Function for the Laplace Operator	109
5.3.3	A Test Function for the Convection-Reaction Operator . . .	110
5.3.4	Combining the Estimates	111
5.3.5	The Second Inf-Sup Condition	112
5.4	A Finite Element Method in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$	114
5.4.1	Stability of Galerkin's Method	115
5.4.2	Convergence Rates for a Simple Example in Two Dimensions	117
5.4.3	Examples with Rough Right-hand Sides	118
6	Minimum Residual Methods in Banach Spaces	121
6.1	An Ideal Petrov-Galerkin Method	122
6.2	Residual Minimization in Banach Spaces	123
6.3	The Optimal Test Norm	125
6.4	The Inexact Method	126

6.5	Breaking Spaces and Forms	127
6.5.1	Breaking Spaces	127
6.5.2	Breaking Variational Forms	130
6.5.3	Coding the DPG Method	133
6.6	(D)PG Methods for the Convection-Diffusion-Reaction Equation . .	134
6.6.1	Comparison of Several Petrov-Galerkin Methods	134
6.6.2	A DPG Method for the Ultra-Weak Formulation	137
7	The NPG Method for the Convection-Diffusion-Reaction Equation	141
7.1	Designing the NPG Method	142
7.1.1	Choices for the Test Norm	143
7.1.2	Weak Boundary Conditions on r	144
7.1.3	The Inexact Method for the Convection-Diffusion-Reaction Equation	145
7.1.4	The Limit Case $\varepsilon = 0$	145
7.1.4.1	Residual Minimization in $L^q(\Omega)$	146
7.1.4.2	Best- L^q Approximation	147
7.2	Numerical Examples	148
7.2.1	Convergence Tests	150
7.2.2	Convection-Diffusion in 1D	151
7.2.2.1	Comparing the Choices for the Test Norm and the Boundary Conditions	152
7.2.2.2	Vanishing oscillations as $q \rightarrow 1$	153
7.2.2.3	The Choice of the Space V_m	155
7.2.2.4	Robustness in ε	157
7.2.3	Vanishing Oscillations in Two Dimensions	157
7.2.3.1	The Eriksson-Johnson Model Problem	158
7.2.3.2	Boundary Layer in a Corner of the Domain	159
7.2.3.3	Example with an Interior Layer	160
7.3	Concluding Remarks	162
8	Conclusions and Future Directions	165
8.1	Accomplishments	165
8.2	Future Work	168
8.3	Extending the NPG Method to Non-linear Problems	171
	Bibliography	175

List of Figures

1.1	L^q -best approximation u_h of $u \equiv 1$ with $u_h(0) = 1$ and $u_h(1) = 0$. . .	6
4.1	L^q -best approximation u_h of $u \equiv 1$ with $u_h(0) = 1$ and $u_h(1) = 0$. . .	53
4.2	Possible approximations u_h of $u \equiv 1$ and x^q for $x \in [0, 1]$	54
4.3	L^q -best approximation: overshoot in one dimension.	58
4.4	Four different meshes on $(0, 1)^2$	60
4.5	L^q -best approximation: overshoot in two dimensions.	62
4.6	The space U_h for a two-element mesh on $(0, 1)$	64
4.7	The space U_h for a three-element mesh on $(0, 1)$	72
4.8	The space U_h for a four-element mesh on $(-1, 1)$	74
4.9	L^1 -best approximation of $\text{sgn}(x)$, cf., Theorem 4.5.	76
4.10	Four different meshes on $(0, 1)^2$, cf., Figure 4.4.	81
4.11	Criss-cross mesh: reference element, u_h and $\text{sgn}(u - u_h)$	82
4.12	Criss-cross mesh: uniform refinement.	85
4.13	L^2 -best approximation: maximal overshoot under mesh refinement.	91
4.14	L^q -best approximation in one dimension: maximal overshoot.	92
4.15	L^q -best approximations on two three-element meshes on $(0, 1)$	93
4.16	L^q -best approximation of a boundary discontinuity in two dimensions on Meshes 2,3 and 4, cf., Figure 4.10/4.4.	94
4.17	L^q -best approximation of a boundary discontinuity in two dimensions.	95
4.18	L^q -best approximation on an unstructured mesh.	95
4.19	Overshoot in the L^q -best approximation on unstructured meshes. See Figure 4.18 for Mesh A.	96
5.1	Convergence of Galerkin's method in $W_0^{1,q}(\Omega)$	117
5.2	2D test case with rough right hand side.	119
5.3	3D test case with rough right hand side.	119
7.1	Convergence of the NPG method: p -enrichment in the test space.	150

7.2	Convergence of the NPG method.	151
7.3	Convergence of the NPG method: the convection-dominated case. .	151
7.4	Choice of the test norm and boundary conditions.	154
7.5	Vanishing undershoot in one dimension.	155
7.6	p -enrichment in the test space.	156
7.7	h -refinement in the test space.	156
7.8	Robustness in ε	157
7.9	Four different meshes on the unit square	158
7.10	Vanishing overshoot in 2D.	158
7.11	Overshoot in 2D.	159
7.12	Comparison with the best-approximation of the analytical solution. .	160
7.13	Example with a boundary layer in the corner of the domain.	161
7.14	Modification of a structured mesh.	161
7.15	Example with an interior layer.	162
7.16	Example with interior layer: perturbation of the convection field. . .	163
8.1	The NPG method for Burgers' equation.	173

Notation

Acronyms

DGFEM	Discontinuous Galerkin finite element method.
DPG	Discontinuous Petrov-Galerkin.
FD	Finite difference.
FEM	Finite element method.
FV	Finite volume.
NPG	Non-linear Petrov-Galerkin.
PDE	Partial differential equation.
SUPG	Streamline upwind Petrov-Galerkin.
VMS	Variational multiscale.

General Symbols

$x, \boldsymbol{x}$	Scalar variable, vector valued variable (bold font)
$A^T, \boldsymbol{v}^T$	Transpose of a matrix A or a vector $\boldsymbol{v}$.
$A^\perp$	Orthogonal complement of $A \subset V$ in V .
$\overline{A}$	Closure of $A \subset V$ with respect to the norm on V .
B'	The adjoint of a linear operator B .
$\Omega \subset \mathbb{R}^d$	Domain in $\mathbb{R}^d$, $d \geq 1$.
$\partial\Omega = \Gamma$	Boundary of Ω .
$\boldsymbol{n}$	Unit outward normal to $\partial\Omega$.

f'	First derivative of a function f of one variable.
$\partial_x f = \frac{\partial f}{\partial x}$	Partial derivative of a function f with respect to the variable x .
∇	Gradient, Gâteaux derivative, Fréchet derivative.
$\operatorname{div}, \nabla \cdot$	Divergence.
Δ	Laplace operator.
$d\mathbf{x}; ds$	d -Dimensional Lebesgue measure for $\mathbf{x} \in \mathbb{R}^d$; surface measure on $\partial\Omega$.

Convection-Diffusion-Reaction Equation

$\mathbf{b}$	Advection field.
c	Reaction coefficient.
c_0	Constant such that $c - \frac{1}{q} \nabla \cdot \mathbf{b} \geq c_0$, $1 < q < \infty$.
ε	Diffusion parameter.
Γ_-	Inflow part of the boundary $\Gamma = \partial\Omega$
Γ_+	Outflow part of the boundary $\Gamma = \partial\Omega$
γ	Inf-sup constant of a bilinear form b .
M	Continuity constant of a bilinear form b .

Banach Spaces

V, V'	Banach space and its dual space, respectively.
$\ \cdot\ _V$	Norm on the Banach space V .
$\langle f, v \rangle_{V', V}$	$f(v)$, functional $f \in V'$ applied to $v \in V$, duality pairing.
∂f	Subdifferential of a functional $f : V \rightarrow \mathbb{R}$; $\partial f \subset V'$.
$\partial f(v)(w)$	$v'(w) = \langle v', w \rangle_{V', V}$ for an arbitrary $v' \in \partial f(v) \subset V'$, $v, w \in V$.

Function Spaces

$C^k(\Omega)$	Space of k times continuously differentiable functions on Ω , $0 \leq k \leq \infty$.
$C_c^k(\Omega)$	Space of functions in $C^k(\Omega)$ with compact support.

$\mathcal{D}(\Omega)$	The space $C_c^\infty(\Omega)$.
$\mathcal{D}'(\Omega)$	Space of distributions, dual space of $\mathcal{D}(\Omega)$.
q, q'	Dual exponents, $1 \leq q \leq \infty$, $1 = 1/q + 1/q'$.
q^*	Sobolev conjugate of q , $1/q^* = 1/q - 1/d$, where d denotes the space dimension.
$L^q(\Omega)$	Lebesgue space of q -integrable functions on Ω if $1 \leq q < \infty$, or functions with finite essential supremum on Ω if $q = \infty$.
$W^{1,q}(\Omega)$	Sobolev space. Subspace of $L^q(\Omega)$ containing all functions with gradient in $L^q(\Omega)^d$.
$W_0^{1,q}(\Omega)$	Subspace of $W^{1,q}(\Omega)$ containing all functions with zero trace on the boundary $\partial\Omega$.
$ \cdot _{W^{1,q}(\Omega)}$	Seminorm in $W^{1,q}(\Omega)$.
$\mathcal{T}_{\mathcal{L}}$	Trace operator corresponding to the graph space associated with the linear operator $\mathcal{L}$.
$W^q(\mathcal{L}, \Omega)$	Graph space in $L^q(\Omega)$ corresponding to the linear operator $\mathcal{L}$.
$W_0^q(\mathcal{L}, \Omega)$	Closure of $\mathcal{D}(\Omega)$ with respect to the graph space norm in $W^q(\mathcal{L}, \Omega)$.
$W^q(\mathbf{b}, \Omega)$	Graph space in $L^q(\Omega)$ corresponding to $\mathcal{L} = \mathbf{b} \cdot \nabla$.
$W^q(\text{div}, \Omega)$	Graph space in $L^q(\Omega)$ corresponding to $\mathcal{L} = \text{div}$.
$W_{0,\Gamma_+}^{1,q}(\Omega)$	Functions in $W^{1,q}(\Omega)$ with zero trace on the outflow boundary Γ_+ .
$H^1(\Omega)$	The space $W^{1,2}(\Omega)$.
$H_0^1(\Omega)$	The space $W_0^{1,2}(\Omega)$.
$H(\text{div}, \Omega)$	The space $W^2(\text{div}, \Omega)$.
$H^{1/2}(\partial\Omega)$	Space of traces of functions in $H^1(\Omega)$.
$H^{-1/2}(\partial\Omega)$	Space of traces of functions in $H(\text{div}, \Omega)$.
$B^{q',q',1-1/q'}$	Besov space.
$B^{q,q,-1/q}$	Dual space of $B^{q',q',1-1/q'}$.

Duality Mappings

$\mathcal{J}_V^\varphi$	Duality mapping on a Banach space V with weight φ .
$\mathcal{J}_V$	Normalized duality mapping on a Banach space V , i.e., duality mapping with weight $\varphi(t) = t$.
R_V	Riesz map on a Hilbert space V . Maps each $v \in V$ onto its Riesz representer.
J_q	Duality mapping on $L^q(\Omega)$ with weight $\varphi(t) = t^{q-1}$.
$\tilde{J}_q$	Normalized duality mapping on $L^q(\Omega)$.
$J_{W_0^{1,q}}$	Duality mapping on $W_0^{1,q}(\Omega)$ with respect to the norm $ \cdot _{W^{1,q}(\Omega)}$ with weight $\varphi(t) = t^{q-1}$.

Discretisation

h	Discretisation parameter: mesh or grid size.
U_h, U_n, V_m	Finite dimensional subspaces of U and V , respectively, where the subscript h usually refers to the mesh size of a finite element space and the subscripts n and m correspond to the dimension of the space U_n and V_m , respectively.
P_p	Space of polynomials of degree p .
p, p_n, p_m	Polynomial degrees. The subscripts n and m indicate polynomial degrees corresponding to U_n and V_m , respectively.
Ω_h	Triangulation of Ω , i.e., a subdivision of Ω into disjoint open elements κ such that $\overline{\Omega} = \bigcup_{\kappa \in \Omega_h} \overline{\kappa}$.
$\partial\Omega_h$	Collection of element boundaries $\partial\kappa$ for all $\kappa \in \Omega_h$.
$H^1(\Omega_h)$	Broken analogue of $H^1(\Omega)$, i.e., space of all functions on Ω_h that are in $H^1(\kappa)$ restricted to each $\kappa \in \Omega_h$.
$H(\text{div}, \Omega_h)$	Broken analogue of $H(\text{div}, \Omega)$.
$H^{1/2}(\Omega_h)$	Space of elementwise traces of functions in $H^1(\Omega)$.
$H^{-1/2}(\Omega_h)$	Space of elementwise traces of functions in $H(\text{div}, \Omega)$.
$\llbracket u_h \rrbracket, \{ \{ u_h \} \}$	Jump and average, respectively, across and interelement boundary.

Chapter 1

Introduction

The term Gibbs phenomenon originally refers to the effect that the partial sums of the Fourier series approximating a function with a jump discontinuity exhibit over- and undershoots near the discontinuity. The phenomenon is named after Willard Gibbs who described it in 1899 [74], though it had already been discovered earlier by Henry Wilbraham in 1848 [167]. It also occurs in the best approximation by spline functions in the L^2 -metric [147] and is one of the main challenges in the numerical approximation of partial differential equations (PDEs) whose solutions contain sharp features such as shocks or thin layers. In [150] it is shown in one dimension that the best approximation of jump discontinuities by polygonal lines on a uniform grid in one dimension does not lead to Gibbs phenomenon in L^1 . More precisely, it is shown that the overshoot of the best approximation in L^q , $1 < q < \infty$, is an increasing function of q that tends to 0 as q tends to 1.

The motivation of this thesis is to exploit the Gibbs-reducing property of L^q -type spaces in the numerical approximation of PDEs using finite element methods (FEMs). A similar idea has already been pursued for numerical methods in L^q by Guermond [82]. Indeed, in [82], Guermond points out that there are only very few attempts to approximate PDEs directly in L^1 , despite the fact that first-order PDEs and their non-linear generalizations have been extensively studied in L^1 . The existing numerical methods to achieve this include the ones outlined in the articles by Lavery [117–119], the re-weighted least-squares method of Jiang [104, 105] and the methods outlined in the series of articles by Guermond et al. [82, 83, 86, 84, 85]. In the context of approximating solutions containing discontinuities and underresolved interior and boundary layers, the numerical results for existing L^1 -methods suggest that such features can be approximated as sharply as a given mesh permits without exhibiting spurious over- or undershoots. This property clearly gives them an enormous advantage over traditional finite

element methods yielding approximations in subspaces of $L^2(\Omega)$. Indeed, it is well-known that even seemingly simple examples such as the transport equation or convection-dominated diffusion equations require extra care in the design of the method, with the standard Galerkin finite element method being unstable, and alternative methods often requiring so-called *stabilization* and/or *shock-capturing* techniques [e.g., 106–108, 148].

More recently, a novel approach to designing finite element methods in a very general Banach space setting has been introduced in [137]. This approach is rooted in the so-called Discontinuous Petrov-Galerkin (DPG) methodology [60] and extends the concept of optimal test norms and functions from Hilbert spaces to more general Banach spaces yielding a scheme that can be interpreted as a *non-linear* Petrov-Galerkin (NPG) method. At least in an abstract sense, this approach outlines how to design a numerical method that leads to a quasi-best approximation of the solution in a space of choice, provided the continuous problem is well-posed in a suitable sense. In this thesis we apply this abstract approach to the convection-diffusion-reaction equation in a $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ setting, where $1/q + 1/q' = 1$, and study the effect this has on the numerical approximation of boundary layers¹. We will see that as $q \rightarrow 1$, the Gibbs phenomenon can be eliminated entirely on certain types of meshes. We will also consider alternative choices of meshes for which this is not the case and consider how this can be overcome. In the context of this investigation, two further problems arise that are studied in this thesis: characterizing $L^q(\Omega)$ -best approximation of discontinuous functions in finite element spaces and well-posedness of the convection-diffusion-reaction equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$.

The problem of characterizing the $L^q(\Omega)$ -best approximation of discontinuous functions in finite element spaces naturally arises from the fact that the non-linear Petrov-Galerkin method considered in this thesis leads to a quasi-best approximation of the solution. Indeed, we will see that in the limit as the diffusion parameter ε approaches zero, our method yields a quasi-best approximation of the solution in $L^q(\Omega)$; moreover, the approximation using our scheme still closely resembles the $L^q(\Omega)$ -best approximation of the solution if $0 < \varepsilon \ll 1$. On coarse meshes, approximating a thin layer is numerically, at least, very similar to approximating a jump discontinuity in the analytical solution. Hence, we study the $L^q(\Omega)$ -best approximation of discontinuous functions in finite element spaces² using carefully selected examples in one and two dimensions in order to understand in which

¹The results related to this can also be found in the preprint [95].

²The results related to this can also be found in the preprint [94].

cases our scheme eliminates over- and undershoots in the approximation as $q \rightarrow 1$. For certain meshes in two dimensions and certain non-uniform meshes in one dimension, we show a result similar to [150] for the best approximation in L^q by piecewise linear functions on these meshes; i.e., we show that the maximal over- or undershoot is an increasing function of q that vanishes as $q \rightarrow 1$. However, there exist meshes both in one and two dimensions such that the (maximal) overshoot tends to some value $\alpha > 0$ as $q \rightarrow 1$. Nonetheless, we demonstrate that if the location of the discontinuity is known, a mesh can be constructed in such a way that the overshoot vanishes. In [135], it is shown that Gibbs phenomena occur in the best approximation of a function containing a jump discontinuity by a trigonometric polynomial in the L^1 -metric. This confirms that the choice of the finite dimensional approximation space crucially determines whether Gibbs phenomena can be eliminated by considering the best approximation in L^1 .

While it is classical to consider the convection-diffusion-reaction equation in the Hilbert space $H_0^1(\Omega)$ in the context of finite element methods, the more general $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ setting, where $1 < q < \infty$ and $1/q + 1/q' = 1$, has so far received little attention in this context. In this thesis we therefore present a well-posedness theory for the convection-diffusion-reaction equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$. In contrast to classical well-posedness results for elliptic differential operators in this setting, cf., e.g., [75], our theory is based on directly establishing the inf-sup conditions and explicitly determining an inf-sup constant. Apart from a standard assumption on the advection and reaction coefficients, the other key assumption pertains to a subtle regularity requirement for the standard Laplacian. In the context of finite element methods (FEMs), well-posedness of the convection-diffusion-reaction variational problem is traditionally established by proving coercivity and continuity of the underlying bilinear form in $H_0^1(\Omega)$. Due to the Lax-Milgram theorem this then implies well-posedness of the continuous problem and its discrete counterpart. The concept of coercivity, however, requires that the trial and test spaces should be the same.³ This is not always possible or desirable, e.g., in the context of Petrov-Galerkin methods or mixed methods, or if the continuous problem itself is stated in a non-symmetric manner such as in the $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ setting. A generalization of the Lax-Milgram theorem originally due to Nečas [140] replaces the coercivity requirement by two inf-sup conditions which can be formulated for Banach spaces with distinct test and trial spaces. A

³Moreover, the concept of coercivity is a notion relevant only in Hilbert spaces [69, Section A.2.4].

version of this theorem for approximations in finite dimensional subspaces is due to Babuška [9].

The remainder of this chapter is organized as follows: We give an overview of the results presented in this thesis in Section 1.1. To this end, we summarize the results regarding the L^q -best approximation of discontinuous functions in finite element spaces, the well-posedness of the convection-diffusion-reaction equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$, and the non-linear Petrov-Galerkin method for the convection-diffusion-reaction equation in Sections 1.1.1, 1.1.2, and 1.1.3, respectively. Finally, in Section 1.2, we give an overview of the structure of this thesis.

1.1 Summary of Results

Throughout this thesis, the convection-diffusion-reaction equation serves as a model problem that we study in detail in order to design the non-linear Petrov-Galerkin method in such a way that Gibbs phenomena can be eliminated in certain cases. This very simple example forms an ideal starting point for developing a method that addresses oscillations in the numerical approximation as it allows us to consider a linear equation with smooth right hand side while still encountering oscillatory approximations. Moreover, it paves the way for applying the results and ideas presented here to a much larger class of PDEs, i.e., so-called *singularly perturbed problems*.

Loosely speaking, singularly perturbed problems are differential equations that depend on a small positive parameter ε such that their solutions approach a discontinuous limit as $\varepsilon \rightarrow 0$, cf., [148]. In other words their solutions typically form thin layers. There is a wide range of problems that fit into this category and typically the equations obtained in the limit as $\varepsilon \rightarrow 0$ are interesting PDEs in their own right. Certainly the most prominent example of this are the Navier Stokes equations that reduce to the Euler equations in the limit as the viscosity parameter tends to zero. Burgers' equation is another excellent example. Indeed, Burgers himself originally [25] studied a version of the equation that included a second-order viscosity term in order to introduce a very simple model that captures two main features of gas dynamics: viscosity and a non-linear hyperbolic term.

1.1.1 The L^q -Best Approximation of Discontinuities in Finite Element Spaces

In this section, we give an overview of the results related to the L^q best approximation of discontinuous functions in finite element spaces presented in Chapter 4. To motivate the best approximation problem we analyse in Chapter 4, we consider the following simple convection-diffusion problem: find u such that

$$-\varepsilon u'' + u' = 0 \quad \text{in } (0, 1), \quad u(0) = 1, \quad u(1) = 0. \quad (1.1)$$

We seek an approximation of the analytical solution in a finite dimensional space that consists of continuous piecewise linear polynomials defined on a given mesh. If $\varepsilon \ll 1$, then the second-order term is completely dominated by the first-order term and away from the outflow boundary the solution is essentially given by the solution to the advection problem obtained by setting ε to zero. For the above problem this means that $u \approx 1$ away from the outflow boundary. Due to the Dirichlet boundary conditions, a boundary layer forms near the outflow boundary, located at $x = 1$. If the diameters of the elements near the boundary layer are large compared with ε , the layer is fully contained within these elements and, in the above problem, $u \approx 1$ in the rest of the domain. Numerically, this essentially means that we approximate the problem (1.1) with $\varepsilon = 0$ while still retaining the boundary conditions at both ends. Clearly, the analytical solution for the above problem with $\varepsilon = 0$ and the boundary condition only imposed on the inflow part of the boundary is $u \equiv 1$. This motivates us to consider the best approximations of $u \equiv 1$ by continuous piecewise linear finite element functions satisfying the boundary conditions given in (1.1).

Figure 1.1 shows the L^q -best approximation of $u \equiv 1$ by a continuous piecewise linear function u_h satisfying $u_h(0) = 1$ and $u_h(1) = 0$ on a uniform mesh consisting of four elements with $q = 2$ and $q = 1.2$. Here, we can see that in both cases over- and undershoots are present in the approximation, but the magnitude of these oscillations is significantly smaller for $q = 1.2$. This example illustrates the phenomenon of reducing oscillations in the approximation as $q \rightarrow 1$ that we shall investigate in Chapter 4 of this thesis.

1.1.1.1 Problem Statement

Let $u \equiv 1$ and consider the following (constrained) best approximation problem:

$$u_h = \arg \min_{v_h \in U_h} \|u - v_h\|_{L^q((0,1)^d)}, \quad d = 1, 2, \quad (1.2a)$$

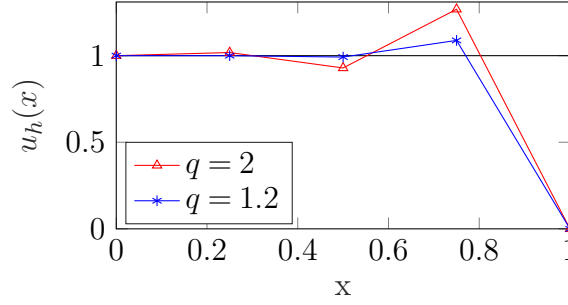


Figure 1.1 L^q -best approximation to $u \equiv 1$ by a piecewise linear function u_h satisfying $u_h(0) = 1$ and $u_h(1) = 0$ on a uniform mesh consisting of four elements with $q = 2$ and $q = 1.2$.

subject to

$$\begin{aligned} u_h(0) = 1, \quad u_h(1) &= 0 \quad \text{if } d = 1, \\ u_h(0, \cdot) = 1, \quad u_h(1, \cdot) &= 0 \quad \text{if } d = 2, \end{aligned} \tag{1.2b}$$

where U_h is a finite element space consisting of continuous piecewise linear polynomials on a given mesh. In one dimension, instead of just $u(x) \equiv 1$, we also consider the L^q -best approximation of $u(x) = \text{sgn}(x)$ on $(-1, 1)$ by a continuous piecewise linear function u_h satisfying $-u_h(-1) = u_h(1) = 1$. We use this example to establish the link between our work and [150].

1.1.1.2 Main Results

The main result of Chapter 4 consists of the precise analysis of very simple examples that illustrate the behaviour of L^q -best approximations of discontinuities by continuous piecewise linear polynomials on coarse meshes. We have furthermore included some numerical examples that confirm the theoretical analysis and illustrate how the observed behaviour in simple model examples applies to more general scenarios. In particular, we demonstrate that the over- and undershoots observed in L^q -best approximations for $1 < q < \infty$ decrease as $q \rightarrow 1$. Whether these oscillations disappear entirely depends on the mesh used to define the finite dimensional approximation space.

In one dimension, we consider the approximation problem (1.2) with $d = 1$, as well as $u(x) = \text{sgn}(x)$ on $\Omega = (-1, 1)$. For both examples, we consider a general two-element mesh on $(0, 1)$ which is mirrored at the origin to create a four-element mesh on $(-1, 1)$ for the second example. Hence, the four-element mesh is symmetric with respect to the origin. For any $q > 1$, we observe an overshoot in the best

approximation. For $q = 1$ it depends on the relative sizes of the two elements in $(0, 1)$ whether any overshoot is present. The difference between the two examples is that for the first example, i.e., $u \equiv 1$ on $(0, 1)$, the L^1 -best approximation is unique, whereas for the second example, i.e., $u = \text{sgn}(x)$ on $(-1, 1)$, there exists a whole family of L^1 -best approximations. This non-uniqueness of the best approximation in certain cases is due to the fact that $L^1(\Omega)$ is not strictly convex. Furthermore, we prove for the first example with an N element mesh that there exists an L^1 -best approximation with no over- or undershoot if either a grading-type mesh condition is satisfied or a stronger, simple element-size condition is fulfilled. A very useful consequence of the grading-type mesh condition is that it is possible to design a mesh in such a way near the discontinuity that the element-sizes are allowed to be arbitrary away from the discontinuity without leading to oscillations. To illustrate this, we consider two three element meshes and determine the corresponding L^1 -best approximations.

The final theoretical result concerns the solution to (1.2) with $d = 2$ on four different meshes. We show that on two of these meshes even the L^1 -best approximation contains overshoots, whereas on the remaining two meshes the over- and undershoots vanish as $q \rightarrow 1$. For one of the first two meshes we furthermore precisely characterize the L^q -best approximation for all $1 \leq q < \infty$.

We also include numerical experiments in Section 4.5 illustrating that the observations remain the same if u is a more general smooth function and that the over- and undershoots cannot be eliminated by refining the mesh for certain classes of meshes.

1.1.2 Well-posedness of the Convection-Diffusion-Reaction Equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$

In this section we state the main result of Chapter 5 regarding the well-posedness of the convection-diffusion-reaction equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$. In addition to the theoretical result stated in this section, we present numerical experiments for Galerkin's method in Chapter 5 considering the approximation of irregular solutions $u \notin H_0^1(\Omega)$. The article [93] contains a proof of the discrete analogue of the inf-sup condition under certain assumptions which implies that the discrete system stemming from Galerkin's method is well-posed.

We consider the following linear elliptic model problem: find u such that

$$\begin{aligned} -\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + cu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma = \partial\Omega, \end{aligned} \quad (1.3)$$

where $\varepsilon, \mathbf{b} : \Omega \rightarrow \mathbb{R}^d$, and $c : \Omega \rightarrow \mathbb{R}$ are the (positive) diffusion parameter, convection field and reaction coefficient, respectively, and $f : \Omega \rightarrow \mathbb{R}$ is a given source term. Multiplying by a smooth test function and integrating by parts, yields the following variational problem: find $u \in W_0^{1,q}(\Omega)$ such that

$$\mathcal{B}_\varepsilon(u, v) = \int_\Omega f v \, d\mathbf{x} \quad \forall v \in W_0^{1,q'}(\Omega), \quad (1.4)$$

where

$$\mathcal{B}_\varepsilon(u, v) = \varepsilon \int_\Omega \nabla u \cdot \nabla v \, d\mathbf{x} - \int_\Omega u \nabla \cdot (\mathbf{b}v) \, d\mathbf{x} + \int_\Omega c u v \, d\mathbf{x}. \quad (1.5)$$

In Chapter 5 we present a well-posedness theory for the above variational problem that relies on directly proving the inf-sup conditions. Our proof relies on two assumptions. One is the q -version of the standard Friedrich's positivity assumption,

$$c(\mathbf{x}) - \frac{1}{q} \nabla \cdot \mathbf{b}(\mathbf{x}) \geq c_0, \quad \text{a.e. in } \Omega,$$

cf., e.g., [33, 139]. The second assumption relates to the regularity of the solution to the standard $H_0^1(\Omega)$ -Poisson problem. It requires that the solution to the Poisson problem is in the Sobolev space $W_0^{1,r}(\Omega)$ with $r > 2$ if the right hand side is in the dual space of $W_0^{1,r'}(\Omega)$, where $1 = 1/r + 1/r'$. In other words, higher regularity in the right hand side leads to higher regularity in the solution. It can be proven to hold for any $r > 2$ for certain classes of Lipschitz domains, e.g., domains with C^1 boundary or convex Lipschitz domains. However, there exist bounded Lipschitz domains that only satisfy this condition for $r \in (2, r_0)$, where the exact value of $r_0 > 2$ depends on the domain.

Under the above assumptions, we prove the following result in Chapter 5:

Theorem 1.1. *Let $1 < q < \infty$, $U := W_0^{1,q}(\Omega)$ and $V := W_0^{1,q'}(\Omega)$. Under the above assumptions on Ω , $\mathbf{b}$ and c , the variational problem (1.4) is well-posed. In particular, the bilinear form $\mathcal{B}_\varepsilon$ defined in (1.5) is bounded, i.e.,*

$$\mathcal{B}_\varepsilon(w, v) \leq M \|v\|_V \|w\|_U$$

and satisfies the following inf-sup conditions: there exists $\gamma > 0$ such that

$$\inf_{w \in U} \sup_{v \in V} \frac{\mathcal{B}_\varepsilon(w, v)}{\|w\|_U \|v\|_V} \geq \gamma, \quad (1.6a)$$

and

$$\forall v \in V, \quad (\forall w \in W, \mathcal{B}_\varepsilon(w, v) = 0) \Rightarrow (v = 0). \quad (1.6b)$$

Both M and γ are positive constants that may depend on $\|c\|_{L^\infty(\Omega)}$, $\|\mathbf{b}\|_{L^\infty(\Omega)}$, ε and c_0 .

The bound for the inf-sup constant γ obtained in Chapter 5 is not robust in ε . However, it can be assumed that the estimate is not optimal, since the scaling in ε for $q = q' = 2$ is less favourable than in the standard proof when $q = q' = 2$.

1.1.3 The Non-linear Petrov-Galerkin Method for the Convection-Diffusion-Reaction Equation

The main goal of this thesis is to develop a non-linear Petrov-Galerkin method for the convection-diffusion-reaction equation that can be used to eliminate unwanted oscillations in the numerical approximation under certain circumstances. In this section, we summarize the results concerning the design of this method. The abstract theory used for this and related Petrov-Galerkin methods for the convection-diffusion equation are presented in detail in Chapter 6. In Chapter 7, the non-linear Petrov-Galerkin method for the convection-diffusion equation is introduced and certain choices in the design of the method underpinned by numerical experiments.

1.1.3.1 Problem Statement

The approach to generalizing the DPG framework to general Banach spaces introduced in [137] is based on residual minimization. To this end, the authors consider the abstract problem: find $u \in U$ such that

$$Bu = f \text{ in } V',$$

where U and V are Banach spaces, $B : U \rightarrow V'$ is a continuous, bounded-below linear operator, and the right hand side f is in the dual space V' . The associated residual minimization problem for a given finite dimensional subspace $U_n \subset U$, e.g.,

a finite element space, of dimension n is defined as follows:

$$u_n = \arg \min_{w_n \in U_n} \|f - Bw_n\|_{V'}.$$

It can be shown that the solution u_n is a quasi-best approximation of the analytical solution u , in the sense that

$$\|u - u_n\|_U \leq \frac{M}{\gamma} \inf_{w_n \in U_n} \|u - w_n\|_U,$$

where M is the continuity constant of B and γ the bounded-below constant of B .

In [137] it is shown that the minimization problem can be reformulated yielding the following non-linear mixed system: find $(r, u_n) \in V \times U_n$ such that

$$\begin{aligned} \langle J_V(r), v \rangle_{V', V} + \langle Bu_n, v \rangle_{V', V} &= \langle f, v \rangle_{V', V}, \quad \forall v \in V, \\ \langle Bw_n, r \rangle_{V', V} &= 0, \quad \forall w_n \in U_n, \end{aligned} \tag{1.7}$$

where J_V is a so-called duality mapping. Duality mappings will be introduced in detail in Chapter 3. They are a generalization of the Riesz map R_V to general Banach spaces with the main difference that they are non-linear mappings unless V is a Hilbert space. To turn the above mixed system into a practical method, V is additionally replaced by a finite dimensional subspace V_m of dimension m . As a result the minimization problem is no longer solved exactly. However, if the spaces U_n and V_m are chosen in a suitable way, one can obtain a well-posed fully discrete mixed system that retains the quasi-best approximation property of u_n with a modified constant.

The aim of this thesis is to apply this abstract framework to the convection-diffusion-reaction equation (1.3) with the corresponding bilinear form (1.5). The resulting method is a generalization to the $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ -setting of the method introduced for the $H_0^1(\Omega)$ -setting in [39], where the DPG framework is applied to the convection-diffusion-reaction equation without introducing broken test spaces. Related Petrov-Galerkin formulations are studied in [23, 24, 40, 47, 56].

1.1.3.2 Summary of Results

In Chapter 7 of this thesis we study how the mixed method (1.7) and its fully discrete counterpart can be applied to the convection-diffusion-reaction equation (1.3). In particular, we study the choice of the space V , the corresponding norm $\|\cdot\|_V$, and the discrete space V_m . The norm $\|\cdot\|_V$ on the space V determines the

operator J_V in (1.7) and thus crucially determines the resulting method. As in the special case $q = 2$ investigated in [39], we introduce weakly imposed boundary conditions on the inflow boundary in the space V_m to address robustness issues. We demonstrate in one dimension, how this choice of boundary conditions affects the approximation and compare this approach with the case when a weighted norm on V is employed.

The main focus of our numerical investigation is eliminating Gibbs phenomena in the finite element approximation. Indeed, we will see that the numerical approximation generated by our method qualitatively behaves like the $L^q(\Omega)$ -best approximation of the analytical solution. Thus, the Gibbs phenomenon can be eliminated by taking the limit $q \rightarrow 1$ provided that the $L^1(\Omega)$ -best approximation does not exhibit Gibbs phenomena. The results in Chapter 4 (see also [94]) show that this depends on the mesh that is chosen. In one dimension only certain non-uniform grids have the property that the $L^1(\Omega)$ -best approximation contains over- and undershoots, whereas in higher dimensions this can even occur on structured, uniform meshes. We will demonstrate for two-dimensional examples that it is possible to use the insights from Chapter 4 to modify the mesh in order to eliminate Gibbs phenomena as $q \rightarrow 1$.

1.2 Structure of the Thesis

The thesis is structured as follows. We start in Chapter 2 with an overview of existing numerical methods for the convection-diffusion-reaction equation. In Chapter 3, we continue with introducing a few concepts related to Banach and Hilbert spaces that are necessary to define the non-linear Petrov-Galerkin method for the convection-diffusion-reaction equation. In the same chapter we will also give precise definitions of the function spaces used in this thesis. In Chapter 4 we consider the L^q -best approximation of discontinuous functions in finite element spaces and determine L^q -best approximations for carefully selected examples both theoretically and computationally. Then, in Chapter 5, we prove a well-posedness result for the convection-diffusion-reaction equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$, state the corresponding well-posedness result for Galerkin's method and consider numerical experiments for Galerkin's method considering the approximation of irregular solutions $u \notin H_0^1(\Omega)$. Chapter 6 contains an introduction to the DPG method as well as its extension to more general Banach spaces introduced in [137]. Furthermore, Petrov-Galerkin methods for the convection-diffusion-reaction equation are discussed that are conceptually closely related to the non-linear Petrov-Galerkin

method considered in this thesis. The concepts introduced in Chapter 6 are then applied to the convection-diffusion-reaction equation in Chapter 7. Certain choices in the design of the method are underpinned by numerical experiments and further numerical experiments demonstrate how the scheme in conjunction with suitably designed meshes can be used to eliminate unphysical oscillations. Finally, in Chapter 8, we discuss the results presented in this thesis regarding their success, as well as their limitations and use this to point towards future research directions regarding this topic.

Chapter 2

Overview of Numerical Methods for the Convection-Diffusion-Reaction Equation

The convection-diffusion-reaction equation has been used time and again as a prototype for singularly perturbed problems in the development of numerical methods for PDEs modelling fluid dynamics. Even though it is a seemingly simple linear second-order PDE, its numerical approximation proves to be very challenging. If the diffusion parameter is very small, thin boundary layers may occur in the analytical solution even if the domain, the right hand side and the boundary conditions are smooth. Without introducing additional stabilization terms, standard discretization techniques typically lead to approximate solutions exhibiting unphysical oscillations in the vicinity of these layers. This is often linked to the fact that the discretization of the differential operator does not preserve certain properties of the operator, such as maximum and comparison principles or (inverse) monotonicity.

The aim of this section is to provide a brief overview of the most important discretization methods employed for the numerical approximation of the convection-diffusion(-reaction) equation and the strategies used to avoid oscillatory approximations. To this end, we first briefly discuss finite difference (FD) methods in Section 2.1 and finite volume (FV) methods in Section 2.2. We then cover a wide range of finite element techniques in Section 2.3 including streamline-upwind Petrov-Galerkin (SUPG) methods, discontinuous Galerkin finite element methods (DGFEM), discontinuous Petrov-Galerkin (DPG) methods and L^1 -methods. A very

detailed introduction to numerical methods for singularly perturbed problems can be found in the monograph [148], though DPG and L^1 -methods are not included.

We will see that the existing numerical methods essentially rely on two different approaches to eliminating unwanted oscillations: adding additional diffusion terms and employing flux limiters. L^1 -methods are an exception to this and instead rely on changing the metric used for the approximation. DPG methods also do not explicitly use either of the two strategies, but instead focus on optimizing the choice of discrete spaces. Our approach builds on ideas coming from L^1 -methods and DPG methods.

2.1 Finite Difference Methods

We start by summarizing important concepts related to FD methods for singularly perturbed problems based on several relevant chapters in [148] and Chapter 2 in [124]. FD methods are based on replacing the derivatives in the PDE by difference quotients, where the function is evaluated at the grid points on a given mesh. The key idea is to define the difference quotients in such a way that they converge to the exact derivatives as the mesh size tends to zero; this property of the FD approximation is referred to as consistency. The order of accuracy of the approximation can be obtained by employing Taylor's theorem. To achieve convergence of the method another property of the approximation is needed, namely stability. The discretization of the PDE by FDs defines a discrete differential operator. For the approximation to be stable, the inverse of this operator has to be bounded uniformly in the mesh size; otherwise the operator approaches a singular operator as the mesh size tends to zero.

For singularly perturbed problems, both stability and consistency are not necessarily easily achievable. On equidistant meshes, the consistency error may blow up in the vicinity of thin layers. Even in the absence of layers, stability may still be an issue. Indeed, the discretization of the simple one-dimensional problem $-\varepsilon u'' + u' = 0$ on $(0, 1)$ is unstable if the first-order derivative is approximated by a difference approximation centred around the point where the derivative is approximated, e.g.,

$$u'(x_i) \approx \frac{u(x_i + h) - u(x_i - h)}{2h},$$

where $0 = x_0 < x_1, \dots, < x_{N-1}, < x_N = 1$, $x_i - x_{i-1} = h$ for $i = 1, \dots, N$, and $h = 1/N$.

The simplest remedy is to approximate the first-order derivative by a difference approximation that is biased towards the upwind direction, i.e., in the direction that is upstream according to the flow direction determined by the dominant first-order term. The resulting schemes are called *upwind* schemes and the simplest (first-order) upwind approximation for the above PDE is given by

$$u'(x_i) \approx \frac{u(x_i) - u(x_i - h)}{h}.$$

Upwinding typically introduces a certain amount of artificial diffusion to the FD scheme. We will encounter the concept of upwinding again in the form of numerical fluxes both in the context of FV methods and DGFEM. More generally, the introduction of some amount of artificial diffusion in the discretization of the PDE is a common theme for most numerical schemes for singularly perturbed problems. Numerous upwind FD schemes have been proposed, cf., e.g., [1, 14, 78, 121, 122, 159]; however, some of these schemes still suffer from poor approximation properties in the vicinity of layers. This has led to the introduction of the concept of uniform convergence, i.e., convergence that is independent of the diffusion parameter ε . A uniformly stable and first-order uniformly convergent scheme that is second-order consistent for fixed ε is given by the Il'in-Allen-Southwell scheme, cf., [4, 100]. Uniformly convergent schemes of higher order can, for example, be constructed according to the HODIE (High Order Differences with Identity Expansion) framework, cf., [64, 127]. Alternatively, instead of equidistant meshes, certain types of graded meshes, cf., e.g., [10, 16, 113, 165], or certain piecewise equidistant meshes, so-called *Shishkin* meshes, cf., [131, 71], can be used to obtain uniform convergence results. The design of these meshes relies on a priori knowledge of the problem; a strategy to circumvent this is adaptive mesh generation, where information from the approximate solution on a given mesh is used to iteratively adapt the mesh to be more suitable for the problem at hand.

2.2 Finite Volume Methods

Since their introduction in the 1970s, numerous variants of FV methods have been developed for the approximate solution of PDEs predominantly stemming from applications in gas and fluid dynamics. In this section we will summarize some of the most important concepts related to FV methods applied to the convection-diffusion-reaction equation, cf., e.g., [30, 31, 128, 160, 120, 134, 148].

Very detailed introductions to FV methods for hyperbolic equations can be found in the monograph by LeVeque [123] and the review paper [133].

To illustrate the idea of FV methods for the convection-diffusion-reaction equation, we start with the PDE in *conservative* form, i.e.,

$$\nabla \cdot (-\varepsilon \nabla u + \mathbf{b}u) + cu = f \quad \text{in } \Omega, \quad u = g \quad \text{on } \partial\Omega. \quad (2.1)$$

FV methods are based on the integral form of the PDE and the application of the divergence theorem. We integrate (2.1) over a control volume $\omega \subset \Omega$ and obtain

$$\int_{\omega} \nabla \cdot (-\varepsilon \nabla u + \mathbf{b}u) d\mathbf{x} + \int_{\omega} cu d\mathbf{x} = \int_{\omega} f d\mathbf{x}.$$

This is similar to a weak formulation in the sense that it can be obtained by using a function that is 1 in ω and 0 in $\Omega \setminus \omega$ as a test function instead of $\varphi \in C^\infty(\Omega)$. Application of the divergence theorem yields

$$\int_{\partial\omega} \left(-\varepsilon \frac{\partial u}{\partial \mathbf{n}} + \mathbf{b} \cdot \mathbf{n} u \right) d\mathbf{x} + \int_{\omega} cu d\mathbf{x} = \int_{\omega} f d\mathbf{x}. \quad (2.2)$$

In order to define a FV method, we consider a subdivision Ω_h of Ω into disjoint cell domains κ such that $\bar{\Omega} = \bigcup_{\kappa \in \Omega_h} \bar{\kappa}$. The approximation space U_h is chosen as a set of functions that are smooth restricted to each cell $\kappa \in \Omega_h$. Usually, piecewise constant functions scaled with $|\kappa|^{-1}$ representing the approximate cell average of u are used. If we consider (2.2) with $\omega = \kappa$ for $\kappa \in \Omega_h$ and replace u by $u_h \in U_h$, we note that the terms in the boundary integral are in general not well-defined since u_h is allowed to be discontinuous across cell boundaries. Hence, $\partial_{\mathbf{n}} u$ is replaced by a finite difference approximation depending on the values of u_h in neighbouring cells and $\mathbf{b} \cdot \mathbf{n} u$ by a numerical flux function depending on the traces of the restriction of u_h to each cell. More details concerning the numerical fluxes follow below.

In one dimension the definition of a cell as some subinterval of the domain is very straightforward, whereas in two (or more) dimensions there are multiple ways to define a mesh or a cell. The review paper [133] focuses on quadrilateral and triangular meshes in two dimensions. In addition to the shape of the elements in a mesh, the points that are associated with the unknowns can be chosen in different ways and the cells over which the averages are taken. A FV method employing a secondary mesh to define the integration cells is called a *cell-centred* FV method. If instead the original grid is used to define the integration cell, it is called a *cell-vertex* method.

2.2.1 Numerical Flux Functions

Numerical flux functions appear in the context of FD methods (see e.g. [144]) and DGFEMs (e.g. [89]), as well as in the context of FV methods, cf., above. A numerical flux function is a function $\mathcal{H}(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n}_\kappa)$ that maps the interior state $\mathbf{u}^+$ of a cell/an element κ and the exterior state $\mathbf{u}^-$ to a flux in direction $\mathbf{n}_\kappa$, where $\mathbf{n}_\kappa$ usually denotes the outward unit normal vector to the boundary of κ .

If $\mathcal{F}(\mathbf{u}) = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_d)$, $d \geq 1$, such that each $\mathbf{f}_i : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is a $C^1(\Omega)$ -function, then the following two properties are generally required for a numerical flux function corresponding to the differential operator $\nabla \cdot \mathcal{F}$:

1. **Consistency:** A numerical flux function $\mathcal{H}(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n}_\kappa)$ is called *consistent* if

$$\mathcal{H}(\mathbf{u}, \mathbf{u}, \mathbf{n}_\kappa) = \mathcal{F}(\mathbf{u}) \cdot \mathbf{n}_\kappa.$$

2. **Conservation:** A numerical flux function $\mathcal{H}(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n}_\kappa)$ is called *conservative* if at any point on the boundary between two neighbouring cells or elements κ and κ' ,

$$\mathcal{H}(\mathbf{u}, \mathbf{v}, \mathbf{n}_\kappa) = -\mathcal{H}(\mathbf{v}, \mathbf{u}, -\mathbf{n}_\kappa) \quad \text{for all } \mathbf{x} \in \partial\kappa \cap \partial\kappa'.$$

Consistency ensures that the numerical flux and the true flux are identical for functions that are continuous across element boundaries and thus is exact for any continuous function. Conservation guarantees that there are no sinks or sources at inter-element boundaries.

In order to ensure stability of the approximation, it is usually required to employ a numerical flux function that is biased towards the upwind direction, i.e. towards the incoming characteristics. Again for $\mathcal{F}(u) = \mathbf{b}u$, the upwind flux is given by

$$\mathcal{H}(u^+, u^-, \mathbf{n}_\kappa) \Big|_{\partial\kappa} = \begin{cases} \mathbf{b} \cdot \mathbf{n}_\kappa u^- & \text{if } \mathbf{b} \cdot \mathbf{n}_\kappa < 0, \\ \mathbf{b} \cdot \mathbf{n}_\kappa u^+ & \text{otherwise.} \end{cases} \quad (2.3)$$

For general $\mathcal{F}$ the following two are examples of upwind fluxes:

1. The **Lax-Friedrichs Flux**

$$\mathcal{H}_{\text{LF}}(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n}) = \frac{1}{2} \left([\mathcal{F}(\mathbf{u}^+) + \mathcal{F}(\mathbf{u}^-)] \cdot \mathbf{n} - \alpha(\mathbf{u}^- - \mathbf{u}^+) \right),$$

where α is the largest eigenvalue in absolute value of the Jacobi matrix $A(\mathbf{u}, \mathbf{n})$ or an estimate thereof. Here,

$$A(\mathbf{u}, \mathbf{n}) := \sum_{i=1}^d n_i A_i(\mathbf{u}), \quad A_i(\mathbf{u}) = \nabla_{\mathbf{u}} \mathbf{f}_i(\mathbf{u}).$$

2. The Vijayasundaram Flux

$$\mathcal{H}_V(\mathbf{u}^+, \mathbf{u}^-, \mathbf{n}) = A^+(\hat{\mathbf{u}}, \mathbf{n}) \mathbf{u}^+ + A^-(\hat{\mathbf{u}}, \mathbf{n}) \mathbf{u}^-,$$

where the positive and the negative part of the Jacobian $A(\hat{\mathbf{u}}, \mathbf{n})$ is denoted by $A^+(\hat{\mathbf{u}}, \mathbf{n})$ and $A^-(\hat{\mathbf{u}}, \mathbf{n})$, respectively, evaluated at some average $\hat{\mathbf{u}}$ between the states $\mathbf{u}^+$ and $\mathbf{u}^-$.

Note that the Lax-Friedrichs flux is always consistent whereas the Vijayasundaram flux is only consistent for linear conservation laws and non-linearities that satisfy $\mathcal{F}(\mathbf{u}) = A(\mathbf{u}, \mathbf{n})\mathbf{u}$. Both fluxes reduce to the upwind flux (2.3) if $\mathcal{F}(u) = bu$. For more details on numerical flux functions we refer to [115], for example.

Numerical flux functions can also be considered as approximate Riemann solvers. The Riemann problem is the conservation law with a specific choice for the initial data, namely a piecewise constant function with one jump discontinuity at some point in space. A numerical flux function essentially seeks to approximate the solution of the Riemann problem with the states in the two neighbouring elements as the piecewise constant function with a jump discontinuity at the inter-element boundary.

2.2.2 High Resolution Methods

The order of accuracy of a FV method relies on the choice of numerical fluxes on the cell boundaries. While the upwind flux (2.3) can only lead to first-order accurate approximations, second-order accuracy can, for example, be achieved by employing the so-called *Lax-Wendroff* flux function. However, methods like this fail near discontinuities yielding non-physical oscillations.

To overcome this problem so-called TVD (total variation diminishing) schemes were introduced, cf. [88]. The total variation for a function u that assumes the constant value u_i restricted to the i th cell is given by

$$\text{TV}(u) = \sum_{i=1}^N |u_i - u_{i-1}|$$

and a method is TVD if

$$\mathrm{TV}(\bar{u}^n) \leq \mathrm{TV}(\bar{u}^{n+1}),$$

where $\bar{u}^n$ denotes an approximation at the n -th time step, $n \geq 0$. TVD methods have the property that they are monotonicity preserving, i.e., the scheme applied to a monotone function yields a function with the same monotonicity. TVD methods often rely on so-called slope or flux limiters which – as the name suggests – limit the magnitude of the flux in a suitable way. A strict imposition of the TVD condition may lead to a loss of accuracy at solution extrema. In fact, Gudonov’s barrier theorem states that linear monotonicity preserving schemes can be at most first-order accurate, cf., e.g., [77, 76]. So-called *ENO* (essentially non-oscillatory) schemes and *TVB* (total variation bounded) schemes, cf., [151] were developed to overcome this problem. These schemes allow the total variation to increase to a limited extent such that the order of accuracy is maintained uniformly throughout the domain. These methods rely on higher order recovery steps, i.e., instead of reconstructing a piecewise linear approximation to the solution from the cell averages of the neighbouring cells, a polynomial of a higher degree is chosen using the cell averages not only of the direct neighbours of the cell but a sequence of connected cells (stencil) to define the polynomial. TVB methods rely on a fixed choice for the stencil, cf., [151], while ENO methods use an adaptive strategy for choosing the stencil, i.e., the stencil yielding the polynomial with the least oscillation is chosen. An extension of ENO methods are *weighted ENO* (*WENO*) methods where a weighted average of the polynomials of a fixed degree obtained from all possible stencils is computed. The weights are determined with the help of a smoothness indicator. For a review of WENO methods we refer to [152] and for a recent survey of WENO methods in the context of finite volume methods and DG methods we refer to [153].

2.3 Stabilized Finite Element Methods

Finite element methods (FEM) are based on considering the restriction of the weak formulation of a PDE to a finite dimensional subspace consisting of piecewise polynomials defined on a mesh. More precisely, we seek a solution u_h of a variational problem posed in a finite dimensional function space, the trial space, and use test functions from (potentially) another finite dimensional function space, the test space. The resulting methods are called Galerkin methods if the test and the

trial space are identical and Petrov-Galerkin methods otherwise. If the finite dimensional space is a subspace of the original function space of the variational formulation, it is called a *conforming* FEM and a *non-conforming* FEM otherwise. DGFEM are in generally non-conforming and convergence is achieved by penalizing the deviation from a conforming solution in the discrete variational form as we will see later on. FEMs are particularly well suited for approximating elliptic PDEs (see, e.g., [19]). The key advantage of FEMs is their high degree of flexibility in terms of coping with complicated geometries, general boundary conditions and non-linearities. Furthermore, the formulation of the problem in a variational setting allows for the use of very powerful analytical tools to derive a priori and a posteriori error bounds.

One of the major challenges in the numerical approximation of PDEs is the efficient and accurate resolution of singularities, discontinuities, and sharp layers present in the analytical solution. Discontinuities commonly occur in solutions of non-linear hyperbolic equations or linear hyperbolic equations when the initial data is discontinuous. Sharp layers are typical for the solutions to singularly perturbed problems such as the convection-dominated convection-diffusion-reaction problems. To illustrate the key problems which may develop with a loss of ellipticity and how these can be addressed in the context of FEMs, we consider the convection-diffusion-reaction equation, cf., (1.3).

First we introduce the standard finite element approximation of the convection-diffusion-reaction equation in Section 2.3.1. In Section 2.3.2 we illustrate the need for stabilization for convection-dominated problems and in Section 2.3.3 we discuss a few stabilization techniques. The content of both sections can be found in [116, 148], for example. We will mainly focus on the Galerkin least squares method and the SUPG method, cf., [109]. For more stabilization techniques we refer to [148], for example.

2.3.1 Standard Finite Element Approximation

We recall the variational problem (1.4) with $q = 2$: find $u \in H_0^1(\Omega)$ such that

$$\mathcal{B}^\varepsilon(u, v) = \ell(v) \quad \text{for all } v \in H_0^1(\Omega), \quad (2.4)$$

where

$$\mathcal{B}^\varepsilon(u, v) := \varepsilon \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x} + \int_{\Omega} \mathbf{b} \cdot \nabla uv \, d\mathbf{x} + \int_{\Omega} cuv \, d\mathbf{x}, \quad \ell(v) := \int_{\Omega} f v \, d\mathbf{x},$$

for $\mathbf{b}$ and c as before and $f \in L^2(\Omega)$. This bilinear form is coercive under suitable assumptions on $\mathbf{b}$ and c . Hence, the Lax-Milgram theorem implies well posedness.

In order to define the finite element approximation of (2.4), we assume that a subdivision Ω_h of Ω into disjoint open element domains κ is given such that $\bar{\Omega} = \bigcup_{\kappa \in \Omega_h} \bar{\kappa}$. Furthermore, V_h is defined as the space of continuous piecewise polynomials on Ω_h , i.e., functions that are continuous and whose restriction to any element κ is a polynomial of degree $p_\kappa \geq 1$. The space $V_{h,0}$ is the subspace of V_h defined by,

$$V_{h,0} = \{v \in V_h : v|_\Gamma = 0\}.$$

We can now define the following finite dimensional problem: find $u_h \in V_{h,0}$ such that

$$\mathcal{B}^\varepsilon(u_h, v_h) = \ell(v_h) \quad \text{for all } v_h \in V_{h,0}. \quad (2.5)$$

2.3.2 Loss of Stability

The standard FEM yields good results even for relatively coarse meshes as long as diffusion remains dominant, i.e., as long as ε is sufficiently large relative to the size of $|\mathbf{b}|$. If, however, convection is strongly dominant (i.e., $\varepsilon \ll |\mathbf{b}|$), the solution of (2.4) typically possesses small subregions, containing interior and boundary layers, where the derivatives of the solution are very large and thus the solution varies rapidly. These layers may become significantly smaller than the mesh size for very small ε . In this case they cannot be resolved properly which may lead to spurious oscillations in the numerical solution. This causes a severe loss of accuracy and stability rendering the classical Galerkin method (2.5) inappropriate unless the mesh is sufficiently fine to fully resolve the layer. Choosing the mesh size sufficiently small to resolve the layers, however, is computationally expensive and becomes infeasible in higher dimensions and for systems of PDEs.

The loss of stability as ε decreases can be demonstrated by considering a stability estimate for the problem (1.4). By standard arguments, we obtain

$$\mathcal{B}^\varepsilon(u, u) \geq \varepsilon \|\nabla u\|_{L^2(\Omega)}^2.$$

If u is the solution of (1.4), we furthermore have

$$\mathcal{B}^\varepsilon(u, u) = \ell(u) \leq C \|f\|_{L^2(\Omega)} \|\nabla u\|_{L^2(\Omega)},$$

where we used the Cauchy-Schwarz inequality and the Poincaré inequality. Combining both finally yields the stability estimate

$$\varepsilon \|\nabla u\|_{L^2(\Omega)} \leq C \|f\|_{L^2(\Omega)}. \quad (2.6)$$

Thus, as ε approaches zero we lose control of the gradient of u and layers can occur.

2.3.3 Continuous Stabilized Finite Element Methods

To obtain the desired stability properties, we consider the problem from a different perspective: we seek a solution that minimizes the residual. To this end consider the abstract linear problem

$$Bu = f \quad \text{in } \Omega.$$

Measuring the residual $Bu - f$ in the $L^2(\Omega)$ -norm gives rise to the following minimization problem: find $u \in V$ such that

$$F(u) = \min_{v \in V} F(v),$$

where

$$F(v) := \|Bv - f\|_{L^2(\Omega)}^2.$$

The corresponding first-order optimality condition for this problem is

$$(Bu, Bv)_{L^2(\Omega)} = (f, Bv)_{L^2(\Omega)}.$$

Considering the least squares or energy norm $\|\cdot\|_E = \|B(\cdot)\|_{L^2(\Omega)}$, we see that the bilinear form (Bu, Bv) is always symmetric and coercive and we obtain the stability estimate $\|u\|_E \leq \|f\|_{L^2(\Omega)}$ which is much stronger than the stability estimate (2.6).

In order to combine the stability of the least squares approximation with the standard Galerkin FEM, we replace the test function in the Galerkin method by $v_h + \delta Bv_h$, where the parameter δ is chosen for maximal accuracy. This can be considered as a Petrov-Galerkin method where the test and the trial space are no longer identical. Let V_h again denote the finite element space and consider the

following problem: find $u_h \in V_h$ such that

$$(Bu_h, v_h + \delta Bv_h)_{L^2(\Omega)} = (f, v_h + \delta Bv_h)_{L^2(\Omega)} \quad \text{for all } v_h \in V_h,$$

where δ may not be constant as it is usually chosen elementwise to obtain maximal accuracy.

Returning to the original problem (2.4), we see that in this case we have restricted to an element κ

$$Bv_h = -\varepsilon \Delta v_h + \mathbf{b} \cdot \nabla v_h + cv_h \quad \text{in } \kappa.$$

For linear finite elements, i.e., $p_\kappa \equiv 1$ for all $\kappa \in \Omega_h$, the term $-\varepsilon \Delta v_h$ is zero in each element κ and if additionally $c = 0$ the stabilization term reduces to $\delta \mathbf{b} \cdot \nabla v_h$. This term adds an additional diffusion term, namely $(\mathbf{b} \cdot \nabla u_h, \delta \mathbf{b} \cdot \nabla v_h)_{L^2(\Omega)}$. Since a loss of diffusion led to the instability in the first place, it can be expected that this term by itself has a stabilizing effect. The method obtained by simply adding this term and considering the test functions $v_h + \delta \mathbf{b} \cdot \nabla v_h$ is called the streamline diffusion FEM or streamline upwind Petrov-Galerkin (SUPG) method, cf., [97, 109].

More generally, consider the following finite dimensional problem corresponding to (2.4): find $u_h \in V_h$ such that

$$\mathcal{B}^\varepsilon(u_h, v_h) + \sum_{\kappa \in \Omega_h} \delta_\kappa (L^\varepsilon u_h, \eta(v_h))_{L^2(\kappa)} = \ell(v) + \sum_{\kappa \in \Omega_h} \delta_\kappa (f, \eta(v_h))_{L^2(\kappa)}, \quad (2.7)$$

where L^ε is the differential operator corresponding to $\mathcal{B}^\varepsilon$, η is some chosen operator and it is assumed that the parameter δ restricted to an element κ takes the constant value δ_κ . For $\eta(v_h) = \mathbf{b} \cdot \nabla v_h$ we recover the SUPG method and for $\eta(v_h) = L^\varepsilon v_h$ the Galerkin least squares method. Other choices are also possible (see, e.g., [92, 148]). The choice of the parameter δ_κ is crucial for the performance of the SUPG method, see e.g., [90] for a hp -version of the method.

In more than one space dimension, the choice $\eta(v_h) = \mathbf{b} \cdot \nabla v_h$ in the SUPG method ensures that no diffusion is added in the crosswind direction, i.e., orthogonal to the streamline direction determined by $\mathbf{b}$, which allows for higher accuracy than adding diffusion in all directions. However, this may not be sufficient to avoid spurious oscillations along sharp layers. Indeed, it can be shown that the discrete maximum principle fails for the SUPG method and thus oscillations can be observed near sharp layers, [107]. Therefore, numerous techniques have been developed to overcome this problem, typically introducing another diffusion term, cf., [106–108].

These methods are sometimes referred to as discontinuity capturing or shock capturing methods. The crucial difficulty is to choose the direction and magnitude of the added diffusion in such a way that the accuracy of the resulting FEM is preserved while ensuring stability. Owing to Godunov's barrier theorem, a linear method satisfying a discrete maximum principle can be at most first-order accurate. Thus, shock capturing techniques often rely on non-linear stabilization terms.

Other strategies for selecting stabilizing terms include variational multiscale methods (VMS), residual-free bubbles, local projection stabilization and continuous interior penalty methods. Variational multiscale methods are based on splitting the infinite dimensional solution space into coarse (resolvable) and fine (unresolvable) scales, cf., [96, 99, 98]. The finite dimensional space is chosen to represent resolvable scales and its orthogonal complement represents the unresolvable scales. Eliminating the projection of the solution onto the fine scale space from the variational formulation yields an operator acting on the approximate solution u_h and an operator acting on the right hand side. The former operator defines the stabilizing term and the latter operator a modification to the right hand side. Except in some specific cases, these operators have to be approximated to yield a fully discrete method. The VMS framework can also be used to derive a class of stabilized methods, so-called *differentiated residual methods*, cf., [148], that coincides with the SUPG method in a special case, cf., [98, 162]. The residual-free bubble method is closely connected with the SUPG method and VMS methods. Similar to VMS methods it is based on considering the variational form in a larger or enriched space and eliminating additional unknowns. The difference is that this larger space is still a finite dimensional subspace and not the whole space. The SUPG method can be derived as a special case of the residual-free bubble method [11, 22, 149]. Local projection stabilization, cf., e.g., [148, 129], was introduced with the aim to remove terms from other stabilization techniques that introduce couplings of dependent variables without any physical counterpart and to avoid asymmetry that is not inherent to the PDE. The idea is to only add terms that improve the stability property of the equation and omit terms that are merely present for consistency of the formulation. The price for this is a consistency error that can in certain cases be improved to retain the same convergence rate as for the SUPG method by introducing a suitable projection of the stabilizing term, cf., [148]. The continuous interior penalty method relies on imposing boundary conditions weakly and adding stabilizing terms that penalize jumps in the gradient across inter-element boundaries, cf., e.g., [28, 26, 29].

Finally, it should be noted that convergence of FEMs and properties of the resulting approximation can be improved by adapting both the polynomial degrees of the finite element functions and the underlying meshes to a given problem. For example, it is possible to employ Shishkin meshes that were originally introduced in the context of FD methods. Furthermore, many adaptive methods relying on a posteriori estimators have been introduced, cf., e.g., [164, 3, 12].

2.3.4 Discontinuous Galerkin Finite Element Method

In contrast to the finite element approaches we have considered so far, DGFEMs rely on enforcing continuity weakly across inter-element boundaries. For hyperbolic problems this can be done using numerical flux functions (cf. Section 2.2.1). This has the advantage that some upwinding is already included in the formulation leading to better stability properties compared to continuous finite element approximations.

DGFEMs were initially introduced independently for first-order hyperbolic problems (cf. [146]) and for second-order elliptic problems (cf. [166, 7]). The review paper [46] provides an overview of the development of DGFEMs and [8] contains a unified analysis of several DGFEMs for second-order elliptic problems. The success of DGFEMs for singularly perturbed problems is a result of employing upwind fluxes for the first-order term. Therefore, we will focus on the formulation for this term as described in [15, 91, 89], for example. For an introduction to DGFEMs we refer to [63, 32].

The main idea of DGFEMs is to consider an elementwise variational formulation and impose boundary conditions weakly at the element boundaries. To make this precise, we consider the advection-reaction equation

$$\begin{aligned} \mathbf{b} \cdot \nabla u + cu &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \Gamma_-, \end{aligned}$$

where $\Gamma_- := \{\mathbf{x} \in \partial\Omega : \mathbf{b} \cdot \mathbf{n}(\mathbf{x}) \leq 0\}$ denotes the inflow part of the boundary.

We multiply this equation by a smooth test function v and integrate over an element $\kappa \in \Omega_h$. Finally, we obtain by integrating by parts:

$$\int_{\kappa} (\mathbf{b} \cdot \nabla u + cu)v \, d\mathbf{x} = - \int_{\kappa} u \nabla \cdot (\mathbf{b}v) \, d\mathbf{x} + \int_{\kappa} cuv \, d\mathbf{x} + \int_{\partial\kappa} \mathbf{b} \cdot \mathbf{n}_{\kappa} uv^+ \, ds = \int_{\kappa} f v \, d\mathbf{x}, \quad (2.9)$$

where $\mathbf{n}_{\kappa}$ again denotes the unit outward normal vector to $\partial\kappa$ and v^+ denotes the interior trace of v , i.e., the trace taken from within κ . If it is not clear from

the context which element $\kappa \in \Omega_h$ the quantity v^+ corresponds to, we will add a subscript denoting the element, e.g., v_κ^+ .

In order to obtain a discretization for the problem from the above equation, we first need to introduce the finite element space. In contrast to the standard FEM, we consider elementwise defined functions that are not necessarily continuous across inter-element boundaries, but when restricted to any element $\kappa \in \Omega_h$ the functions are polynomials of degree $p_\kappa \geq 0$. More precisely, we define the finite element space W_h by

$$W_h = \{v : v|_\kappa \in P_{p_\kappa}, \kappa \in \Omega_h\},$$

where P_{p_κ} denotes the space of polynomials of degree p_κ on κ . Note that this time we also include piecewise constant functions, whereas the lowest order basis functions for the continuous FEM are piecewise linear functions. W_h is a subspace of the infinite dimensional space

$$H^1(\Omega_h) := \{v \in L^2(\Omega) : \Omega \rightarrow \mathbb{R} : v|_\kappa \in H^1(\kappa) \text{ for all } \kappa \in \Omega_h\}.$$

Furthermore, we denote by v^- (or v_κ^- in case of ambiguity) the exterior trace of v , i.e., $v^-(\mathbf{x}) = v_{\kappa'}^+(\mathbf{x})$ if $\mathbf{x} \in \partial\kappa \cap \partial\kappa'$ and we replace $v^-(\mathbf{x})$ by appropriate boundary conditions if $\mathbf{x} \in \partial\kappa \cap \Gamma$.

We can now discretize (2.9) by replacing the analytical solution u by the finite element approximation $u_h \in W_h$ and the test function v by $v_h \in W_h$. Since the approximation u_h may be discontinuous across inter-element boundaries, we need to replace $\mathbf{b} \cdot \mathbf{n}_\kappa u$ by a numerical flux function $\mathcal{H}(u_h^+, u_h^-, \mathbf{n}_\kappa)$ (see Section 2.2.1). With the simple upwind flux (2.3), we obtain the following DG finite element approximation by summing over all elements $\kappa \in \Omega_h$: find $u_h \in W_h$ such that

$$\begin{aligned} & \sum_{\kappa \in \Omega_h} \left(\int_\kappa (\mathbf{b} \cdot \nabla u_h + c u_h) v_h \, d\mathbf{x} \right. \\ & \quad \left. - \int_{\partial_{-\kappa} \cap \partial\Omega} \mathbf{b} \cdot \mathbf{n}_\kappa u_h^+ v_h^+ \, ds + \int_{\partial_{-\kappa} \setminus \partial\Omega} \mathbf{b} \cdot \mathbf{n}_\kappa (u_h^+ - u_h^-) v_h^+ \, ds \right) \\ & = \int_\Omega f v_h \, d\mathbf{x} - \sum_{\kappa \in \Omega_h} \int_{\partial_{-\kappa} \cap \partial\Omega} \mathbf{b} \cdot \mathbf{n}_\kappa g v_h^+ \, ds \quad \forall v_h \in W_h, \end{aligned} \quad (2.10)$$

where $\partial_{-\kappa} := \{\mathbf{x} \in \partial\kappa : \mathbf{b} \cdot \mathbf{n}_\kappa(\mathbf{x}) \leq 0\}$ denotes the inflow part of the element boundary.

In order to derive the DGFEM formulation for second-order elliptic PDEs, the underlying PDE is typically rewritten as a first-order system by introducing

auxiliary variables. For this system an elemental weak formulation is then derived and appropriate numerical flux functions are introduced yielding the so-called flux formulation. Finally, the underlying primal formulation is obtained by eliminating the auxiliary variables. The interior penalty method for Poisson's problem can be formulated as follows: find $u_h \in W_h$ such that

$$\begin{aligned} \sum_{\kappa \in \Omega_h} \int_{\kappa} \nabla u_h \cdot \nabla v_h \, d\mathbf{x} + \sum_{F \in \mathcal{F}_h} \int_F (\vartheta \{\{\nabla u_h\}\} \cdot \llbracket v_h \rrbracket - \{\{\nabla v_h\}\} \cdot \llbracket u_h \rrbracket + \sigma \llbracket u_h \rrbracket \cdot \llbracket v_h \rrbracket) \, ds \\ = \int_{\Omega} f v_h \, d\mathbf{x}, \end{aligned}$$

where $\mathcal{F}_h$ denotes the set of faces on the mesh Ω_h and the average and jump operators on these faces are defined by

$$\{\{\nabla w_h\}\} := \frac{1}{2} (\nabla w_h^+ + \nabla w_h^-), \quad \llbracket w_h \rrbracket := w_h^+ \mathbf{n}_{\kappa}^+ + w_h^- \mathbf{n}_{\kappa}^-,$$

respectively, for a face shared by two adjacent elements κ^+ and κ^- with unit outward normals $\mathbf{n}_{\kappa}^+$ and $\mathbf{n}_{\kappa}^-$, respectively. If the face is contained in Γ and κ^+ is the adjacent element, u_h^- is replaced by the boundary conditions, $\{\{\nabla v_h\}\}$ is replaced by ∇v_h^+ and $\llbracket v_h \rrbracket$ by v_h^+ . Common choices for ϑ are -1 and 1 yielding the symmetric and the non-symmetric interior penalty method, respectively. Note that the formulation relies on penalizing jumps across inter-element boundaries and the penalty parameter σ has to be chosen sufficiently large to ensure convergence. Other DGFEM formulations are possible and we refer to [8] for more details. A DGFEM for the convection-diffusion-reaction equation can be obtained by adding the above weak form scaled by ε to (2.10).

Similar to the SUPG method, DGFEMs typically do not satisfy a discrete maximum principle. At very sharp layers, DGFEMs must therefore be modified by adding some dissipation in a way that does not reduce the formal order of accuracy. For this purpose limiters may be introduced to control the oscillations, see e.g., [45] or, for a brief overview of WENO limiters, [153]. Another possibility is using shock capturing terms in a similar fashion as for the SUPG method, cf., [44], for example. Furthermore, adaptive methods can also be applied to DGFEMs.

2.4 Discontinuous Petrov-Galerkin Methods

Within the last decade, cf., [60], the discontinuous Petrov-Galerkin (DPG) method has experienced a great deal of development, though the name was already used

in a similar sense a few years earlier [17, 18, 36, 37]. The idea behind the DPG methodology is to construct a test space for a given finite dimensional trial space that is optimal in the sense that it realizes the discrete inf-sup condition with the same constants as the continuous inf-sup condition. Computing the optimal test space relies on solving another infinite dimensional problem, more precisely it relies on inverting the Riesz map on the test space. In general, this can only be done approximately and to this end another finite dimensional space has to be introduced. This space is typically chosen to allow discontinuities across inter-element boundaries. Note, however, that this space is only needed for the approximate inversion of the Riesz map and the approximate solution to the PDE is still sought in a continuous finite element space. The DPG method has been applied to a wide range of problems including the convection-diffusion equation [23, 24, 40, 56], as well as Burgers' equation and the compressible Navier-Stokes equations [38]. Recently, a generalization to Banach spaces has been suggested [136]. In [40] a connection between the VMS framework and the DPG framework has been established implying a close connection between the DPG methodology and stabilized methods. There is, however, a remarkable achievement in the special case that the optimal test space can be computed exactly, that sets the DPG method apart from other stabilized methods: for diffusion-dominated equations the optimal L^2 -convergence rate under global h -refinement of a FEM is $\mathcal{O}(h^{p+1})$, where h is the mesh size and p the polynomial degree, whereas for convection dominated equations the optimal rate is only $\mathcal{O}(h^{p+1/2})$. Indeed, Peterson has demonstrated that this rate cannot be improved for DGFEM [143]. The so-called *ideal* DPG method can be shown to outperform DGFEM and SUPG methods for the transport equation at the improved rate $\mathcal{O}(h^{p+1})$, cf., [53]. In Chapter 6, we will consider the DPG methodology in more detail, as well as its extension to Banach spaces.

2.5 Methods for Approximating PDEs in L^1

In [82], Guermond points out that there are only very few attempts to approximate PDEs directly in L^1 despite the fact that first-order PDEs and their non-linear generalizations have been extensively studied in $L^1(\Omega)$. The existing numerical methods which seek an approximation directly in $L^1(\Omega)$ include the ones outlined in the articles by Lavery for a FV discretization [117–119], the reweighted least-squares method of Jiang [104, 105] and the methods outlined in the series of articles by Guermond et al. considering a finite element approach [82, 83, 86, 84, 85]. These

methods are derived in the context of first-order PDEs and the residual is minimized in an L^1 sense, though examples of singularly perturbed problems are considered as well. In the context of approximating solutions containing discontinuities and underresolved interior and boundary layers, the numerical results for existing L^1 -methods suggest such features can be approximated as sharply as a given mesh permits without exhibiting spurious over- or undershoots. This thesis expands on these methods by considering more general residual minimization problems in dual norms leading to an approach that is applicable to a wider range of PDEs including second-order PDEs.

Chapter 3

Preliminaries

This chapter contains fundamental definitions and results, mainly from functional analysis, that we will use throughout this thesis. The chapter is organized as follows: we start by introducing concepts related to Banach and Hilbert spaces in Section 3.1; in Section 3.2 we list precise definitions of the main function spaces we consider; Section 3.3 contains an introduction to duality mappings and Section 3.4 contains a characterization of best approximation in Banach spaces that is closely linked with duality mappings. In Section 3.5 we introduce the closed range theorem and some of its implications; Section 3.6, finally, provides a brief overview of a very general concept of graph spaces and trace operators.

3.1 Banach Spaces and Hilbert Spaces

In this section we define a few concepts from functional analysis and state some very useful results. For more details we refer to [6, 42, 41, 43, 21], for example. We restrict ourselves to real valued vector spaces and all definitions and results are only stated in this special case. In Section 3.3, we will use the concepts summarized here in order to introduce the duality mapping that we need to define the $L^q(\Omega)$ -version of the Petrov-Galerkin methods introduced in Chapter 6. We start with some basic concepts in Banach spaces in Section 3.1.1. Then we introduce Hilbert spaces in Section 3.1.2, as well as the Riesz Representation Theorem that provides us with the Hilbert space equivalent of the duality mapping, the Riesz map.

The concepts related to Banach and Hilbert spaces are certainly very useful beyond the definition of the duality map and the formulation of Petrov-Galerkin methods in $L^q(\Omega)$ -type spaces. In particular, FEMs are based on weak formulations of PDEs in Hilbert spaces.

3.1.1 Banach spaces

Definition 3.1 (Banach space). *A normed vector space that is complete is called a Banach space.*

Definition 3.2 (dual space and duality pairing). *Let V be a normed vector space with norm $\|\cdot\|_V$. Its dual space V' is defined as the set of all continuous linear operators from V to $\mathbb{R}$ and the dual norm $\|\cdot\|_{V'}$ is defined as*

$$\|\ell\|_{V'} := \sup_{v \in V, v \neq 0} \frac{|\ell(v)|}{\|v\|_V}.$$

Furthermore, the duality pairing $\langle \cdot, \cdot \rangle_{V', V}$ between V and V' is given by

$$\langle \ell, v \rangle_{V', V} := \ell(v) \quad \ell \in V', v \in V.$$

Note that the dual space V' equipped with the norm $\|\cdot\|_{V'}$ as defined above is always a Banach space, cf., [42, Theorem 3.2-4].

Definition 3.3 (Orthogonal complement). *Let V be a Banach space and let U be a subspace of V . The (topological) orthogonal complement of U , denoted by $U^\perp$, is defined as*

$$U^\perp := \{v' \in V' : \langle v', v \rangle_{V', V} = 0 \text{ for all } v \in U\}.$$

Definition 3.4 (uniformly convex spaces). *A normed space V is called uniformly convex if for any $\varepsilon \in (0, 2]$ there exists $\delta > 0$, such that for any $v, w \in V$, $v \neq w$, with $\|v\|_V = \|w\|_V = 1$ and $\|v - w\|_V \geq \varepsilon$ we have $\left\| \frac{1}{2}(v + w) \right\|_V \leq 1 - \delta$.*

Definition 3.5 (strictly convex space). *A normed space V is called strictly convex if for $v, w \in V$, $v \neq w$, $\|v\|_V = \|w\|_V = 1$, we have $\|\lambda v + (1 - \lambda)w\|_V < 1$ for all $\lambda \in (0, 1)$.*

Definition 3.6 (reflexive space). *Let V be a Banach space. We can define a linear, injective isometry $\mathcal{I}_V : V \rightarrow V'' := (V')'$ by*

$$\langle \mathcal{I}_V(v), v' \rangle_{V'', V'} := \langle v', v \rangle_{V', V} \quad \text{for } v \in V, v' \in V'.$$

$\mathcal{I}_V$ is referred to as the natural embedding of V into its bidual V'' and the space V is called reflexive if $\mathcal{I}_V$ is surjective.

Theorem 3.7 (cf., [41]). *Every uniformly convex space is strictly convex.*

Theorem 3.8 (Milman-Pettis, cf., [21, 43]). *A uniformly convex Banach space is reflexive.*

The following corollary that follows from the Hahn-Banach theorem is very useful.

Corollary 3.9 (cf., [21, Corollary 1.3]). *Let V be a Banach space. For every $v_0 \in V$ there exists $\ell_0 \in V'$ such that*

$$\|v_0\|_V = \|\ell_0\|_{V'} \text{ and } \langle \ell_0, v_0 \rangle_{V', V} = \|v_0\|_V^2.$$

Later on we will need several notions of differentiability in Banach spaces. Therefore we recall the following three definitions.

Definition 3.10 (Gâteaux derivative). *Let X and Y be two Banach spaces and let $D \subseteq X$ be open. The function $F : D \rightarrow Y$ is Gâteaux differentiable at $x_0 \in D$ if there exists a continuous linear operator $\nabla F(x_0) : X \rightarrow Y$ such that*

$$\lim_{t \rightarrow 0} \frac{F(x_0 + tx) - F(x_0)}{t} = \nabla F(x_0)x \quad \text{for every } x \in X.$$

Definition 3.11 (Fréchet derivative). *Let X and Y be two Banach spaces and let $D \subseteq X$ be open. The function $F : D \rightarrow Y$ is Fréchet differentiable at x_0 if it is Gâteaux differentiable at x_0 and*

$$\lim_{t \rightarrow 0} \sup_{\|x\|_X=1} \left\| \frac{F(x_0 + tx) - F(x_0)}{t} - \nabla F(x_0)x \right\|_Y = 0.$$

Definition 3.12 (subdifferential). *Let V be a Banach space. A function $F : V \rightarrow \mathbb{R}$ is called subdifferentiable at a point $v \in V$ if there exists $v' \in V'$, called sub-gradient of F at v , such that*

$$F(w) - F(v) \geq \langle v', w - v \rangle_{V', V} \quad \text{for all } w \in V.$$

The set of all sub-gradients of F at v is denoted by $\partial F(v)$ and the mapping $\partial F : V \rightarrow \mathcal{P}(V')$, where $\mathcal{P}(V')$ denotes the power set of V' , is called the subdifferential of F .

3.1.2 Hilbert Spaces

Definition 3.13 (Hilbert space). *An inner product space V with inner product $(\cdot, \cdot)_V$ is a Hilbert space if it is complete with respect to the norm $\|\cdot\|_V$ induced by*

the inner product, i.e., the norm that is defined as

$$\|v\|_V := \sqrt{(v, v)_V} \quad \text{for all } v \in V.$$

Theorem 3.14 (Riesz representation theorem). *Let V be a Hilbert space with inner product $(\cdot, \cdot)_V$ and the induced norm $\|\cdot\|_V$. For any continuous linear functional $\ell \in V'$, there exists a unique $w_\ell \in V$ such that*

$$\langle \ell, v \rangle_{V', V} = (v, w_\ell)_V \quad \text{for all } v \in V.$$

Furthermore,

$$\|\ell\|_{V'} = \|w_\ell\|_V,$$

and the Riesz map $R_V : V \rightarrow V'$ given by

$$R_V(v) := (\cdot, v)_V, \quad v \in V,$$

is a linear isometric isomorphism, i.e., R_V is linear, bijective, and

$$\|R_V(v)\|_{V'} = \|v\|_V.$$

3.2 Function Spaces

In this section we introduce the function spaces that will be used throughout this thesis; for more details we refer to [6], for example.

3.2.1 Continuously Differentiable Functions

Definition 3.15 (The spaces $C^k(\Omega)$). *Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$ be open. We denote by $C^0(\Omega)$ the set of all continuous functions on Ω . Furthermore, for any integer $k \geq 1$ we denote the space of all functions that are k times continuously differentiable over Ω by $C^k(\Omega)$. The space of infinitely continuously differentiable functions is given by*

$$C^\infty(\Omega) := \bigcap_{k=0}^{\infty} C^k(\Omega).$$

Definition 3.16 (support of a function). *The support of a real-valued function $f : \Omega \rightarrow \mathbb{R}$, denoted by $\text{supp } f$, is the closure of the set of points in $x \in \Omega$ such that $f(x)$ is non-zero, i.e.,*

$$\text{supp } f := \overline{\{x \in \Omega : f(x) \neq 0\}}.$$

Definition 3.17. *For $k \in \mathbb{N} \cup \{\infty\}$ we define the subspace $C_c^k(\Omega) \subset C^k(\Omega)$ of $C^k(\Omega)$ -functions with compact support in Ω as follows*

$$C_c^k(\Omega) := \{f \in C^k(\Omega) : \text{supp } f \subset \Omega \text{ is compact}\}.$$

For $k = \infty$, we also write

$$\mathcal{D}(\Omega) := C_c^\infty(\Omega).$$

Definition 3.18 (Space of distributions). *The dual space of $\mathcal{D}(\Omega)$ is called the space of distributions and is denoted by $\mathcal{D}'(\Omega)$.*

Equipped with the supremum norm,

$$\|f\|_{C^0(\bar{\Omega})} := \sup_{x \in \bar{\Omega}} |f(x)|,$$

the space $C^0(\bar{\Omega})$ is a Banach space.

3.2.2 Sobolev Spaces

Definition 3.19 ($L^q(\Omega)$ spaces). *We define the Lebesgue spaces $L^q(\Omega)$ for $1 \leq q \leq \infty$ as*

$$L^q(\Omega) := \left\{ v : \Omega \rightarrow \mathbb{R} : \|v\|_{L^q(\Omega)} < \infty \right\},$$

where

$$\begin{aligned} \|v\|_{L^q(\Omega)} &:= \left(\int_{\Omega} |v(\mathbf{x})|^q d\mathbf{x} \right)^{1/q} \quad \text{for } 1 \leq q < \infty, \\ \|v\|_{L^\infty(\Omega)} &:= \text{ess sup}_{\Omega} |v| \\ &= \inf \left\{ \sup_{\mathbf{x} \in \Omega \setminus N} |v(\mathbf{x})| : N \subset \Omega \text{ has Lebesgue measure } 0 \right\}. \end{aligned}$$

The spaces $L^q(\Omega)$ are Banach spaces for all $1 \leq q \leq \infty$ and the space $L^2(\Omega)$ is a Hilbert space with inner product

$$(v, w)_{L^2(\Omega)} = \int_{\Omega} v(\mathbf{x})w(\mathbf{x}) \, d\mathbf{x}.$$

Definition 3.20 (Sobolev spaces $W^{k,q}(\Omega)$). *We define the Sobolev spaces $W^{k,q}(\Omega)$ for $k = 0, 1, 2, \dots$ and $1 \leq q \leq \infty$, as the space of all real-valued functions such that all weak derivatives up to order k lie in $L^q(\Omega)$; more precisely, we have that*

$$W^{k,q}(\Omega) := \left\{ v : \Omega \rightarrow \mathbb{R} : \frac{\partial^{|\alpha|} v}{\partial \mathbf{x}^\alpha} \in L^q(\Omega), 0 \leq |\alpha| \leq k \right\},$$

where α denotes a multi-index of length $|\alpha|$. These spaces are Banach spaces equipped with the norms

$$\begin{aligned} \|v\|_{W^{k,q}(\Omega)} &:= \left(\sum_{|\alpha| \leq k} \left\| \frac{\partial^{|\alpha|} v}{\partial \mathbf{x}^\alpha} \right\|_{L^q(\Omega)}^q \right)^{1/q} \quad \text{for } 1 \leq q < \infty, \\ \|v\|_{W^{k,\infty}(\Omega)} &:= \max_{|\alpha| \leq k} \left\| \frac{\partial^{|\alpha|} v}{\partial \mathbf{x}^\alpha} \right\|_{L^\infty(\Omega)}. \end{aligned}$$

Furthermore, the associated semi-norms are given by

$$\begin{aligned} |v|_{W^{k,q}(\Omega)} &:= \left(\sum_{|\alpha|=k} \left\| \frac{\partial^{|\alpha|} v}{\partial \mathbf{x}^\alpha} \right\|_{L^q(\Omega)}^q \right)^{1/q} \quad \text{for } 1 \leq q < \infty, \\ |v|_{W^{k,\infty}(\Omega)} &:= \max_{|\alpha|=k} \left\| \frac{\partial^{|\alpha|} v}{\partial \mathbf{x}^\alpha} \right\|_{L^\infty(\Omega)}. \end{aligned}$$

We denote by $W_0^{1,q}(\Omega)$ the subspace of $W^{1,q}(\Omega)$ containing all functions in $W^{1,q}(\Omega)$ that are zero on the boundary of Ω in the sense that their trace is zero. Furthermore, we denote the dual space of $W_0^{1,q}(\Omega)$ by $W^{-1,q'}(\Omega)$, where $1/q + 1/q' = 1$.

We note that $W^{0,q}(\Omega) = L^q(\Omega)$.

Definition 3.21 (Sobolev spaces $H^k(\Omega)$, $H_0^1(\Omega)$ and $H^{-1}(\Omega)$). *For $q = 2$ we write*

$$H^k(\Omega) := W^{k,2}(\Omega), H_0^1(\Omega) := W_0^{1,2}(\Omega) \text{ and } H^{-1}(\Omega) = W^{-1,2}(\Omega).$$

The spaces $H^k(\Omega)$, $H_0^1(\Omega)$ and $H^{-1}(\Omega)$ are Hilbert spaces.

We conclude this section by quoting a Sobolev embedding theorem. There are multiple results concerning the embedding of Sobolev spaces into other function spaces with higher regularity or integrability, collectively referred to as Sobolev embeddings, cf., e.g., [6]. Here, we only state one special case that will be used for a crucial step in the proof of an inf-sup condition for the convection-diffusion-reaction equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$, cf. in Chapter 5.

Theorem 3.22 (Sobolev embedding). *Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$, be an open and bounded Lipschitz domain. Let $1 \leq q < \infty$, then the embedding*

$$\text{Id} : W^{1,q}(\Omega) \rightarrow L^r(\Omega)$$

is continuous if $r \leq q^*$, where q^* is the Sobolev conjugate of q defined by

$$\frac{1}{q^*} = \frac{1}{q} - \frac{1}{d}.$$

Moreover, the embedding is compact if $r < q^*$ and the same results hold with $W^{1,q}(\Omega)$ replaced by $W_0^{1,q}(\Omega)$.

3.3 Duality Mappings

One of the key ingredients used to extend concepts from Hilbert spaces to more general Banach spaces both in [136] and [93] is replacing the Riesz map with so-called duality mappings. In general, duality mappings are non-linear maps that share certain properties with the Riesz map. If the underlying space is a Hilbert space, any duality mapping is linear and the Riesz map is one possible choice for the duality mapping. In this section we will give a precise definition of duality mappings, summarize a range of very useful properties and list some relevant examples.

Definition 3.23.

1. *A weight function is a continuous and strictly increasing function $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\varphi(0) = 0$ and $\lim_{t \rightarrow \infty} \varphi(t) = +\infty$.*
2. *Let V be a Banach space and φ a weight function. Denote by $\mathcal{P}(V')$ the power set of V' . Then the multivalued map $\mathcal{J}_V^\varphi : V \rightarrow \mathcal{P}(V')$, defined by*

$$\mathcal{J}_V^\varphi(v) := \left\{ v' \in V' : \langle v', v \rangle_{V', V} = \|v\|_V \|v'\|_{V'}, \|v'\|_{V'} = \varphi(\|v\|_V) \right\}$$

is called a duality mapping of weight φ .

Due to Corollary 3.9 (cf., e.g., [21, Corollary 1.3]) the set $\mathcal{J}_V^\varphi(v)$ is non-empty. If we choose $\varphi(t) = t$ as a weight function we obtain the so-called *normalized* duality map. In many cases this choice is the simplest and most convenient one. In fact, some books only introduce the duality mapping for this choice of φ , cf., eg., [21, 51, 169]. As we will see later, however, $\varphi(t) = t^{q-1}$ is in some cases a more suitable choice for L^q -type spaces. If $q = 2$ this choice coincides with the normalized duality map. Furthermore, note that we have $\mathcal{J}_V(v) = \{R_V(v)\}$ if $\mathcal{J}_V$ is the normalized duality map and V is a Hilbert space due to the Riesz Representation theorem, cf. Theorem 3.14.

3.3.1 Properties Duality Mappings

In the following Proposition we summarize a few properties of the duality map for Banach spaces with particular structures.

Proposition 3.24. *Let V be a Banach space and denote by $\mathcal{J}_V^\varphi : V \rightarrow \mathcal{P}(V')$ the duality map of weight φ on V . Then the following statements are true:*

1. *V' is strictly convex if and only if $\mathcal{J}_V^\varphi$ is single valued, cf., [51, Prop. 12.3]. In this case we define the duality map $J_V^\varphi : V \rightarrow V'$ such that $\mathcal{J}_V^\varphi(v) = \{J_V^\varphi(v)\}$ for all $v \in V$.*
2. *If V is strictly convex, then $\mathcal{J}_V^\varphi(v) \cap \mathcal{J}_V^\varphi(w) = \emptyset$ for all $w \neq v$. In particular, $\mathcal{J}_V^\varphi$ is injective.*
3. *V is reflexive if and only if $\mathcal{J}_V^\varphi$ is surjective in the sense that for every $v' \in V'$ there is a $v \in V$ such that $v' \in \mathcal{J}_V^\varphi(v)$, cf., [43, Theorem 3.4, Chapter II].*
4. *If V is a reflexive Banach space and $\mathcal{J}_V^\varphi$ is a duality mapping of weight φ , then $(\mathcal{J}_V^\varphi)^{-1}$ is a duality mapping on V' of weight φ^{-1} , cf., [43, Cor. 3.5, Ch. II].*

The main implication of the above proposition is that in a strictly convex and reflexive space V with a strictly convex dual space V' , the duality mapping is bijective and its inverse can be identified with a duality mapping on the dual space V' with the inverse weight φ^{-1} .

The following theorem is a special case of Theorem 4.4 in [43, Chapter I] and states that the duality map on V can be characterized using the subdifferential of the norm on V . This is a key property of the duality map that will allow us to

derive the duality map for some specific Banach spaces in the special case that the subdifferential is essentially the Gâteaux or Fréchet derivative of the norm.

Theorem 3.25 (Asplund, cf., [43, Ch. I, Theorem 4.4]). *Let V be a Banach space and define $F_V^\varphi : V \rightarrow \mathbb{R}$ by $F_V^\varphi(\cdot) := \psi(\|\cdot\|_V)$, where $\psi(s) := \int_0^s \varphi(t) dt$ and φ is a weight function. Then for any $v \in V$, we have*

$$\mathcal{J}_V^\varphi(v) = \partial F_V^\varphi(v),$$

where $\partial F_V^\varphi(v)$ denotes the subdifferential of F_V^φ at v .

The result that finally allows us to compute duality maps is the following.

Proposition 3.26 (cf., [169, Proposition 47.19]). *Let V be a Banach space and define $F_V^\varphi : V \rightarrow \mathbb{R}$ by $F_V^\varphi(\cdot) := \psi(\|\cdot\|_V)$, where $\psi(s) := \int_0^s \varphi(t) dt$ and φ is a weight function.*

1. *If V' is strictly convex, then $\nabla F_V^\varphi(v)$ exists as a Gâteaux derivative and $\nabla F_V^\varphi(v) = J_V^\varphi(v)$ for all $v \in V$.*
2. *If V' is uniformly convex, then $\nabla F_V^\varphi(v)$ exists as a Fréchet derivative and $\nabla F_V^\varphi(v) = J_V^\varphi(v)$ for all $v \in V$.*

Note that due to Theorem 3.7 uniform convexity of V' implies strict convexity of V' and due to the Milman-Pettis theorem, cf. Theorem 3.8, V' and hence also V are reflexive in this case. Furthermore, if F_V^φ is Gâteaux differentiable, Theorem 3.25 guarantees that the duality map is single valued which implies strict convexity of the dual space V' due to Proposition 3.24.

3.3.2 Examples of Duality Mappings on Sobolev Spaces

First, consider $V = L^q(\Omega)$, $1 < q < \infty$ with the norm

$$\|v\|_{L^q(\Omega)} = \left(\int_{\Omega} |v|^q d\mathbf{x} \right)^{1/q}.$$

Let us denote the duality mapping on $L^q(\Omega)$ with weight $\varphi(t) = t^{q-1}$ by J_q . In this case $\psi(s) = \int_0^s \varphi(t) dt = \frac{1}{q}s^q$ and thus we can compute

$$\begin{aligned} \langle J_q(v), w \rangle_{L^{q'}(\Omega), L^q(\Omega)} &= \frac{d}{dt} \left(\frac{1}{q} \|v + tw\|_{L^q(\Omega)}^q \right) \Big|_{t=0} \\ &= \frac{d}{dt} \left(\frac{1}{q} \int_{\Omega} |v + tw|^q d\mathbf{x} \right) \Big|_{t=0} \\ &= \int_{\Omega} |v|^{q-1} \operatorname{sgn}(v) w d\mathbf{x}. \end{aligned}$$

Note that $\varphi^{-1}(t) = t^{q'-1}$ for q such that $1 = 1/q + 1/q'$ and therefore $J_{q'}^{-1} = J_q$. Moreover, we have for all $v \in L^q(\Omega)$

$$\|J_q(v)\|_{L^{q'}(\Omega)} = \|v\|_{L^q(\Omega)}^{q-1}, \quad \langle J_q(v), v \rangle_{L^{q'}(\Omega), L^q(\Omega)} = \|v\|_{L^q(\Omega)}^q.$$

Similarly, we can compute the normalized duality map $\tilde{J}_q$ on $L^q(\Omega)$, i.e., the duality map with weight $\varphi(t) = t$, and obtain

$$\langle \tilde{J}_q(v), w \rangle_{L^{q'}(\Omega), L^q(\Omega)} = \|v\|_{L^q(\Omega)}^{2-q} \int_{\Omega} |v|^{q-1} \operatorname{sgn}(v) w d\mathbf{x}.$$

Comparing the expressions for J_q and $\tilde{J}_q$, we can see that it may be useful, in particular for the computational implementation, to use J_q instead of the normalized duality mapping in order to avoid the additional scaling $\|v\|_{L^q(\Omega)}^{2-q}$.

As a second example consider the space $W_0^{1,q}(\Omega)$ equipped with the (semi-) norm $|v|_{W^{1,q}(\Omega)}$. In the same way as before we can compute the duality map of weight $\varphi(t) = t^{q-1}$ and obtain

$$\langle J_{W_0^{1,q}}(v), w \rangle_{L^{q'}(\Omega), L^q(\Omega)} = \int_{\Omega} \sum_{i=1}^d |\partial_i v|^{q-1} \operatorname{sgn}(\partial_i v) \partial_i w d\mathbf{x}.$$

The corresponding normalized duality map is given by

$$\tilde{J}_{W_0^{1,q}}(v) = |v|_{W^{1,q}(\Omega)}^{2-q} J_{W_0^{1,q}}(v).$$

Remark 3.27 (q-Laplacian). Let $1 \leq q < r < \infty$ and $\mathbf{x} \in \mathbb{R}^d$, $d \geq 1$. In this case we have the following norm equivalence on the finite dimensional space $\mathbb{R}^d$:

$$\begin{aligned} \|\mathbf{x}\|_{l^q} &= \left(\sum_{i=1}^d |x_i|^q \right)^{\frac{1}{q}} \leq d^{\frac{1}{q} - \frac{1}{r}} \left(\sum_{i=1}^d |x_i|^r \right)^{\frac{1}{r}} = d^{\frac{1}{q} - \frac{1}{r}} \|\mathbf{x}\|_{l^r}, \\ \|\mathbf{x}\|_{l^r} &\leq \|\mathbf{x}\|_{l^q}. \end{aligned}$$

As a result

$$\left(\int_{\Omega} \|\nabla v\|_{l^2}^q d\mathbf{x} \right)^{1/q}$$

defines a norm on $W_0^{1,q}(\Omega)$ that is equivalent to $|v|_{W^{1,q}(\Omega)}$ and the corresponding duality mapping for $q > 2$ is given by

$$\int_{\Omega} \|\nabla v\|_{l^2}^{q-2} \nabla v \cdot \nabla w d\mathbf{x},$$

which is also the weak form of the q -Laplacian.

3.4 Characterization of L^q -Best Approximation

In this section we describe a characterisation of best-approximation in Banach spaces and more specifically the Lebesgue spaces $L^q(\Omega)$, $1 \leq q < \infty$. This characterisation will be used in Chapter 4 to determine the best L^q -approximation for specific examples.

We now quote the following theorem, cf., [154, Theorem 1.1].

Theorem 3.28 (Characterisation of best approximation). *Let U be a Banach space, $U_h \subset U$ a closed subspace and $u \in U$. The following statements are equivalent:*

1. $u_h = \arg \min_{w_h \in U_h} \|u - w_h\|_U$.
2. *There exists a functional $r' \in \partial(\|\cdot\|_U)(u - u_h)$ which annihilates U_h , i.e.,*

$$\langle r', w_h \rangle_{U', U} = 0 \quad \text{for all } w_h \in U_h.$$

Remark 3.29. The subdifferential $\partial(\|\cdot\|_U)(\cdot)$ can be characterized as follows, cf., e.g., [43, Chapter 1, Proposition 3.4]. For any $w \in U$,

$$\partial(\|\cdot\|_U)(w) := \begin{cases} \{w' \in U' : \langle w', w \rangle_{U', U} = \|w\|_U, \|w'\|_{U'} = 1\} & \text{if } w \neq 0, \\ \{w' \in U' : \|w'\|_{U'} = 1\} & \text{if } w = 0. \end{cases} \quad (3.5)$$

This characterisation allows us to translate the above formulation of Theorem 3.28 directly into the formulation found in [154]. In [137], the same theorem is stated in terms of the so-called duality mapping, which can also be easily translated into the above formulation.

First we will use Theorem 3.28 to characterize best approximants in subspaces of $L^q(\Omega)$, $1 < q < \infty$. To this end, we determine the subdifferential $\partial(\|\cdot\|_{L^q(\Omega)})(w)$ for an arbitrary $w \in L^q(\Omega)$ and $1 < q < \infty$. Note that in this case the norm is Gâteaux differentiable; indeed, we can compute for $w \neq 0$:

$$\partial(\|\cdot\|_{L^q(\Omega)})(w)(v) = \left. \frac{d}{dt} \left(\int_{\Omega} |w + tv|^q d\mathbf{x} \right)^{\frac{1}{q}} \right|_{t=0} = \|w\|_{L^q(\Omega)}^{1-q} \int_{\Omega} \text{sgn}(w) |w|^{q-1} v d\mathbf{x},$$

where

$$\text{sgn}(w(x)) = \begin{cases} -1 & \text{if } w(\mathbf{x}) < 0, \\ 1 & \text{if } w(\mathbf{x}) > 0, \\ 0 & \text{if } w(\mathbf{x}) = 0, \end{cases}$$

hence $\partial(\|\cdot\|_{L^q(\Omega)})(w) = \|w\|_{L^q(\Omega)} \text{sgn} w |w|^{q-1}$ by the canonical identification of an element in the dual space of $L^q(\Omega)$ with a function in $L^{q'}(\Omega)$, where $1 = 1/q + 1/q'$.

The following Corollary is an immediate consequence of this by setting $w = u - u_h$.

Corollary 3.30. *Let $U := L^q(\Omega)$, $1 < q < \infty$, and $U_h \subset U$ a closed subspace. The function $u_h \in U_h$ is an L^q -best approximation of u if and only if*

$$\int_{\Omega} \text{sgn}(u - u_h) |u - u_h|^{q-1} v_h d\mathbf{x} = 0 \quad \forall v_h \in U_h.$$

Next we will use (3.5) to characterize best approximations in subspaces of $L^1(\Omega)$. Note that in this case the subdifferential $\partial(\|\cdot\|_{L^1(\Omega)})(w)$ is in general not single valued for an arbitrary $w \in L^1(\Omega)$. Since the dual space of $L^1(\Omega)$ is isomorphic with $L^\infty(\Omega)$, any $w' \in [L^1(\Omega)]'$ can be identified with some $\psi \in L^\infty(\Omega)$

such that

$$w'(v) = \int_{\Omega} \psi v d\mathbf{x}.$$

From (3.5), we deduce that $w' \in \partial(\|\cdot\|_{L^1(\Omega)})(w)$ if and only if the corresponding $\psi \in L^\infty(\Omega)$ satisfies the following properties

1. $\|\psi\|_{L^\infty(\Omega)} = 1$.
2. $\int_{\Omega} \psi w d\mathbf{x} = \|w\|_{L^1(\Omega)}$.

It is easy to see that any ψ such that $\psi = \text{sgn}(w)$ if $w \neq 0$ and $|\psi| \leq 1$ almost everywhere satisfies the above conditions. Conversely, the first property implies $|\psi(x)| \leq 1$ almost everywhere and the second property implies that $\psi(x) = 1$ almost everywhere on $\{w(\mathbf{x}) > 0\}$ and $\psi(x) = -1$ almost everywhere on $\{w(\mathbf{x}) < 0\}$ since

$$\|w\|_{L^1(\Omega)} = \int_{\Omega} |w| d\mathbf{x} = \int_{\Omega} \psi w d\mathbf{x} = \int_{\Omega \cap \{w(\mathbf{x}) > 0\}} \psi |w| d\mathbf{x} - \int_{\Omega \cap \{w(\mathbf{x}) < 0\}} \psi |w| d\mathbf{x}.$$

It is important to note, that the only condition on ψ on the set $\{w(x) = 0\}$ is that $|\psi| \leq 1$ almost everywhere. The following Corollary characterizing L^1 -best approximations is a direct consequence of this by setting $w = u - u_h$.

Corollary 3.31. *Let $U := L^1(\Omega)$ and $U_h \subset U$ a closed subspace. The function $u_h \in U_h$ is an L^1 -best approximation of u if and only if there exists a function $\psi_0 \in L^\infty(\Omega \cap \{u(\mathbf{x}) = u_h(\mathbf{x})\})$, $|\psi_0| \leq 1$, almost everywhere, such that for all $v_h \in U_h$,*

$$0 = \int_{\Omega \cap \{u(\mathbf{x}) > u_h(\mathbf{x})\}} v_h d\mathbf{x} - \int_{\Omega \cap \{u(\mathbf{x}) < u_h(\mathbf{x})\}} v_h d\mathbf{x} + \int_{\Omega \cap \{u(\mathbf{x}) = u_h(\mathbf{x})\}} \psi_0 v_h d\mathbf{x},$$

or, equivalently, for all $v_h \in U_h$,

$$0 = \int_{\Omega} \psi v_h d\mathbf{x},$$

where $\psi = \text{sgn}(u - u_h)$ on $\{u(\mathbf{x}) \neq u_h(\mathbf{x})\}$ and $\psi = \psi_0$ on $\{u(\mathbf{x}) = u_h(\mathbf{x})\}$.

Note that in the case that u and u_h only agree on a set of measure zero, the choice of $\psi_0 \in [-1, 1]$ becomes irrelevant.

3.5 Closed Operators and the Closed Range Theorem

Closed operators are a generalization of continuous operators. Even though we mostly work with continuous operators in the context of FEM, it is sometimes useful to consider a closed operator corresponding to a given PDE instead. The closed range theorem for closed operators is a powerful tool for proving well-posedness of weak formulations of linear PDEs. In this section, we will first introduce bounded and closed operators and then state the closed range theorem. Finally, we will state two theorems commonly used for the well-posedness analysis in the context of FEMs that are essentially a reformulation of the closed range theorem in a more specific setting. The first one can be applied to the bilinear form representing the weak formulation of a PDE in the infinite dimensional setting and the second can be applied to the discrete setting, i.e., the finite dimensional approximation of a variational problem in the context of FEM.

The concepts in this section are well-known and can be found in many books either on functional analysis or FEMs such as [142, 20, 6, 69].

Definition 3.32 (Domain and Graph of an Operator). *Let U and V be two normed spaces and let B be a linear operator $B : D(B) \rightarrow V$, where $D(B) \subset U$ is a subspace. We call $D(B)$ the domain of B and the Graph of B , denoted by $\mathcal{G}(B)$, is defined as*

$$\mathcal{G}(B) := \{(u, Bu) : u \in D(B)\} \subset U \times V.$$

Proposition 3.33 (Bounded Linear Operators). *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be two normed vector spaces and let $B : U \rightarrow V$ be a linear map. Then the following statements are equivalent*

1. *B is continuous.*
2. *B is continuous at 0.*
3. *There is a $C > 0$ such that $\|Bu\|_V \leq C\|u\|_U$ for all $u \in U$.*

Definition 3.34 (Closed Operator). *Let U and V be two normed spaces. A linear operator $B : D(B) \rightarrow V$ is called closed if its graph $\mathcal{G}(B)$ is a closed subspace of $U \times V$.*

Definition 3.35 (Adjoint Operator). *Let U and V be two normed spaces. The adjoint or transpose $B' : V' \rightarrow U'$ of a continuous linear operator B is defined by*

the identity

$$\langle v', Bu \rangle_{V',V} = \langle B'v', u \rangle_{U',U} \quad \forall u \in U, v' \in V'.$$

The above definition of the adjoint operator can be extended to closed operators $B : D(B) \rightarrow V$ such that $D(B)$ is dense in U , i.e., $\overline{D(B)} = U$.

If U and V are Hilbert spaces and we denote by R_U and R_V the Riesz maps on U and V , respectively, we can define an adjoint $B^* : V \rightarrow U$ as $B^* := R_U^{-1} \circ B' \circ R_V$. Furthermore, we obtain the following identity

$$\begin{aligned} (B^*v, u)_U &= ((R_U^{-1} \circ B' \circ R_V)v, u)_U = \langle (B' \circ R_V)v, u \rangle_{U',U} \\ &= \langle R_V v, Bu \rangle_{V',V} = (v, Bu)_V. \end{aligned}$$

Remark 3.36. The term *adjoint operator* is commonly used for both the adjoint B' in the context of Banach spaces and for B^* in the context of Hilbert spaces. To avoid confusion, we have introduced a different notation for the two different operators. In [142] this problem is solved by calling B' the transpose rather than the adjoint of B .

Theorem 3.37 (Closed Range Theorem for Closed Operators). *Let U and V be normed spaces and $B : U \subset D(B) \rightarrow V$, $\overline{D(B)} = U$ (i.e. the domain of B is dense in U), be linear and closed. The following conditions are equivalent to each other:*

- (i) $\mathcal{R}(B)$ is closed in V .
- (ii) $\mathcal{R}(B) = \mathcal{N}(B')^\perp$

The following theorem is a reformulation of the Closed Range Theorem for the continuous linear operator B induced by a given bilinear form $b : U \times V' \rightarrow \mathbb{R}$ via the definition

$$\langle Bu, v \rangle_{V',V} := b(u, v) \quad \forall v \in V,$$

if B is additionally injective.

Theorem 3.38 (Banach-Nečas-Babuška). *Let U and V be Banach spaces; additionally assume that V is reflexive. Let $b : U \times V \rightarrow \mathbb{R}$ be a continuous bilinear form and $\ell \in V'$. Furthermore, consider the problem: find $u \in U$ such that*

$$b(u, v) = \ell(v) \quad \forall v \in V. \tag{3.6}$$

Then, (3.6) is well-posed for all $\ell \in V'$ if and only if

$$\exists \gamma > 0, \quad \inf_{w \in U} \sup_{v \in V} \frac{b(w, v)}{\|w\|_U \|v\|_V} \geq \gamma, \quad (3.7a)$$

$$\forall v \in V, \quad (\forall w \in U, b(w, v) = 0) \Rightarrow (v = 0). \quad (3.7b)$$

Moreover, we have the a priori estimate:

$$\forall \ell \in V', \quad \|u\|_U \leq \frac{1}{\gamma} \|\ell\|_{V'}.$$

There are two steps to understanding how the above theorem is linked with the Closed Range Theorem. The less obvious step is to prove that boundedness below of B is equivalent to closedness of $\mathcal{R}(B)$ for continuous and injective operators. It is then easy to see that boundedness below of B is equivalent to the inf-sup-condition (3.7a). The second and more obvious step is to realize that, if V is reflexive, (3.7b) is equivalent to $\mathcal{N}(B') = 0$ implying surjectivity of B due to the Closed Range Theorem. Note that we only need reflexivity in this last step. The Closed Range Theorem itself can still be applied if V is not reflexive, e.g., $V = L^1(\Omega)$ or $V = L^\infty(\Omega)$.

Theorem 3.39 (Babuška, 1971¹). *Let U, V be Banach spaces with corresponding finite dimensional² subspaces U_h and V_h with $\dim U_h = \dim V_h$. Furthermore, assume that V is reflexive. Let $b : U \times V \rightarrow \mathbb{R}$ be a bilinear form that satisfies Theorem 3.38. Let b additionally satisfy the following discrete inf-sup-condition: there exists a constant $\gamma_h > 0$ such that*

$$\inf_{w_h \in U_h} \sup_{v_h \in V_h} \frac{b(w_h, v_h)}{\|w_h\|_U \|v_h\|_V} \geq \gamma_h > 0.$$

Then there is a unique $u_h \in U_h$ satisfying

$$b(u_h, v_h) = \ell(v_h) \quad \forall v_h \in V_h,$$

¹This result is originally due to Babuška, but is only stated for U, V Hilbert in the original paper [9]. A more general version that includes both Theorem 3.38 and Theorem 3.39 as special cases can be found in [142], for example, as “The Generalized Lax-Milgram Theorem”.

²The result can also be proven for *closed* subspaces U_h and V_h that are not necessarily finite dimensional, cf., [142, Theorem 6.6.1]. In this case an additional condition is necessary to ensure surjectivity of $B : U_h \rightarrow V_h$ instead of the condition $\dim U_h = \dim V_h$.

and

$$\|u - u_h\|_U \leq \left(1 + \frac{M}{\gamma_h}\right) \inf_{w_h \in U_h} \|u - w_h\|_U, \quad (3.8)$$

where M is a constant such that $|b(u, v)| \leq M\|u\|_U\|v\|_V$ for all $u \in U$, $v \in V$.

Remark 3.40. If U and V are Hilbert spaces, the estimate (3.8) can be improved by applying the following result: if U is a Hilbert space and $P : U \rightarrow U$ a linear projection (i.e., $P^2 = P$), then $\|I - P\| = \|P\|$, cf., [161]. The improved a priori estimate is then

$$\|u - u_h\|_U \leq \frac{M}{\gamma_h} \inf_{w_h \in U_h} \|u - w_h\|_U.$$

Remark 3.41. The result can also be sharpened in the more general Banach setting by introducing a geometric constant, cf., [136].

3.6 Graph Spaces and Trace Operators

In order to handle boundary conditions and function values on inter-element boundaries in the context of FEM, we need the concept of traces of functions in Sobolev spaces. To this end, we are interested in extending many concepts from $L^2(\Omega)$ -type spaces to $L^q(\Omega)$ -type spaces for an arbitrary $q \in (1, \infty)$. Therefore we introduce a very general concept of traces in this section, mainly based on the first chapter of [102].

Choose a pair $q, q' \in \mathbb{R}$ such that $1 < q < \infty$ and $q' = q/(q-1)$ and let $\mathcal{L} : L^q(\Omega)^m \rightarrow \mathcal{D}'(\Omega)^l$ be a first-order linear operator defined as follows:

$$(\mathcal{L}v)_i := \sum_{k=1}^d \partial_k \left(\sum_{j=1}^m B_{ijk} v_j \right) + \sum_{j=1}^m C_{ij} v_j, \quad i = 1, \dots, l,$$

for given $B \in W^{1,\infty}(\Omega)^{l \times m \times d}$ and $C \in L^\infty(\Omega)^{l \times m}$.

Definition 3.42 (Graph space). *The graph space of $\mathcal{L}$ is defined as*

$$W^q(\mathcal{L}, \Omega) := \{v \in L^q(\Omega)^m : \mathcal{L}v \in L^q(\Omega)^l\}.$$

We can then equip the graph space with the graph norm

$$\|v\|_{\mathcal{L},q} := \|v\|_{\mathcal{L},q,\Omega} := \sqrt[q]{\|v\|_{L^q(\Omega)^m}^q + \|\mathcal{L}v\|_{L^q(\Omega)^l}^q},$$

and the graph semi-norm

$$|v|_{\mathcal{L},q} := |v|_{\mathcal{L},q,\Omega} := \|\mathcal{L}v\|_{L^q(\Omega)^l}.$$

Example 3.43 ($W^{1,q}(\Omega)$). If we take $m = 1$, $l = d$, $B_{ijk} = \delta_{ik}$ and $C_{ij} = 0$, we obtain $\mathcal{L}v = \nabla v$ and recover the Sobolev space $W^{1,q}(\Omega)$ as a graph space with the usual norm and semi-norm.

Example 3.44 ($W^q(\operatorname{div}, \Omega)$). If $l = 1$, $m = d$ and $\mathcal{L} = \operatorname{div}$, then

$$W^q(\mathcal{L}, \Omega) = W^q(\operatorname{div}, \Omega) = \{v \in L^q(\Omega)^m : \operatorname{div} v \in L^q(\Omega)\}.$$

Example 3.45 (Scalar advection-reaction). If we choose $m = l = 1$, we can set $B = \mathbf{b} \in W^{1,\infty}(\Omega)^d$ and $C = c \in L^\infty(\Omega)$ and obtain $\mathcal{L}v = \nabla \cdot (\mathbf{b}v) + cv$. If we, alternatively, set $B = \mathbf{b}$ and $C = -\nabla \cdot \mathbf{b} + c$, we obtain³ $\mathcal{L}v = \mathbf{b} \cdot \nabla v + cv$. In the latter case we also identify $W^q(\mathcal{L}, \Omega)$ with the graph space

$$W^q(\mathbf{b}, \Omega) := \{v \in L^q(\Omega)^d : \mathbf{b} \cdot \nabla v \in L^q(\Omega)\}.$$

Definition 3.46 (Formal Adjoint). The formal adjoint $\mathcal{L}^*$ of the operator $\mathcal{L}$ can be defined as

$$\mathcal{L}^* : w \mapsto - \sum_{k=1}^d \sum_{j=1}^m B_{jik} \partial_k w_j + C_{ji} w_j.$$

Remark 3.47. In the above definition, we make sense of the expression by using the following convention

$$\begin{aligned} \sum_{k=1}^d \sum_{j=1}^m B_{ijk} \partial_k v_j + \sum_{j=1}^m C_{ij} v_j \\ = \sum_{k=1}^d \partial_k \left(\sum_{j=1}^m B_{ijk} v_j \right) + \sum_{j=1}^m \left(C_{ij} - \sum_{k=1}^d \partial_k B_{ijk} \right) v_j. \end{aligned}$$

In order to define traces, we introduce the Besov space⁴ $B^{q',q',1-1/q'}(\partial\Omega) \subset L^q(\partial\Omega)$ and its dual $B^{q,q,-1/q}(\partial\Omega)$, $q' \in (1, \infty)$. For Lipschitz domains, the space

³We have to be slightly careful with giving meaning to the expression $\mathbf{b} \cdot \nabla v$ for $v \in W^q(\mathcal{L}, \Omega)$, cf., [102, Remark 1], but this can easily be done by identifying $\mathbf{b} \cdot \nabla v + cv$ with $\nabla \cdot (\mathbf{b}v) + (-\nabla \cdot \mathbf{b} + c)v$.

⁴These Besov spaces are sometimes also called fractional-order Sobolev spaces with the notation $W^{1-1/q',q'}(\partial\Omega)$.

$B^{q',q',1-1/q'}(\partial\Omega)$ can be characterized as follows

$$B^{q',q',1-1/q'}(\partial\Omega) = \{v \in L^{q'}(\partial\Omega) : \exists w \in W^{1,q'}(\mathbb{R}^d) : v = w|_{\partial\Omega}\}.$$

If Ω is bounded with Lipschitz boundary, there exists an extension operator $\mathcal{E}_{\Omega,\mathbb{R}^d} : W^{1,q'}(\Omega) \rightarrow W^{1,q'}(\mathbb{R}^d)$ such that $(\mathcal{E}_{\Omega,\mathbb{R}^d}v)|_{\partial\Omega} = v$ [157]. We can thus define the surjective trace operator

$$\mathcal{T}_{W^{1,q'}(\Omega)} : W^{1,q'}(\Omega) \rightarrow B^{q',q',1-1/q'}(\partial\Omega), v \mapsto (\mathcal{E}_{\Omega,\mathbb{R}^d}v)|_{\partial\Omega}.$$

Now set $B(\mathbf{n})_i := (\sum_{j=1}^d B_{ij}\mathbf{n}_j)_i$. For $v \in W^{1,q}(\Omega)$ and $w \in W^{1,q'}(\Omega)^l$,

$$\langle B(\mathbf{n})v, w \rangle_{\partial\Omega} = \langle \mathcal{L}v, w \rangle_{\Omega} - \langle v, \mathcal{L}^*w \rangle_{\Omega},$$

where $\langle v, w \rangle_{\Omega} := \int_{\Omega} vw \, d\mathbf{x}$.

We can now define a continuous linear operator

$$\mathcal{T}_{\mathcal{L}} : W^{1,q}(\Omega) \rightarrow B^{q,q,-1/q}(\partial\Omega)^l, v \mapsto \langle B(\mathbf{n})v, \cdot \rangle_{\partial\Omega},$$

which can by density of $W^{1,q}(\Omega)$ in $W^q(\mathcal{L}, \Omega)$ be extended to a continuous linear operator $\mathcal{T}_{\mathcal{L}} : W^q(\mathcal{L}, \Omega) \rightarrow B^{q,q,-1/q}(\partial\Omega)^l$. This allows us to define the trace space of $W^q(\mathcal{L}, \Omega)$ as $W^q(\mathcal{T}_{\mathcal{L}}, \partial\Omega) := \text{Im } \mathcal{T}_{\mathcal{L}}$ and equip it with the minimal-extension norm

$$\|v\|_{\mathcal{T}_{\mathcal{L}},q} := \inf_{w \in W(\mathcal{L},\Omega)^q, \mathcal{T}_{\mathcal{L}}w=v} \|w\|_{\mathcal{L},q}.$$

By construction of the norm on $\|\cdot\|_{\mathcal{T}_{\mathcal{L}},q}$ the operator norm of $\mathcal{T}_{\mathcal{L}}$ is 1 if we consider $\mathcal{T}_{\mathcal{L}}$ as an operator from $W^q(\mathcal{L}, \Omega)$ to $W^q(\mathcal{T}_{\mathcal{L}}, \partial\Omega)$.

We can furthermore define the functions with zero trace as the closure of $\mathcal{D}(\Omega)$ with respect to the graph norm, i.e., $W_0^q(\mathcal{L}, \Omega) = \overline{\mathcal{D}(\Omega)}$ and indeed $W_0^q(\mathcal{L}, \Omega) = \ker \mathcal{T}_{\mathcal{L}}$, cf., [102, Theorem 9].

Chapter 4

Gibbs Phenomena in the L^q -Best Approximation of Discontinuities in Finite Element Spaces

4.1 Introduction

This chapter investigates the Gibbs phenomenon in the context of the L^q -best approximation of discontinuous functions in finite element spaces by considering a few carefully selected simple examples that can be analysed in detail. Saff and Tashev show in [150] that in one dimension the best approximation of a jump discontinuity by polygonal lines leads to Gibbs phenomena for all $1 < q < \infty$ that vanish as $q \rightarrow 1$; this is the starting point of our investigation.

We consider several meshes in one and two dimensions and show that on certain meshes the over- and undershoots can be eliminated in the limit $q \rightarrow 1$ as in the case considered by [150]. However, there exist meshes in both one and two dimensions that do not satisfy this property. The aim of this chapter is to illustrate which properties the underlying mesh must satisfy to allow us to predict whether the oscillations vanish in the L^q -best approximation of discontinuous functions as $q \rightarrow 1$.

4.1.1 Motivation

To motivate the best approximation problem we analyse in this chapter, we consider the following simple convection-diffusion problem: find u such that

$$-\varepsilon u'' + u' = 0 \quad \text{in } (0, 1), \quad u(0) = 1, \quad u(1) = 0. \quad (4.1)$$

The analytical solution to this problem is given by

$$u(x) = \frac{1 - e^{-\frac{1-x}{\varepsilon}}}{1 - e^{-\frac{1}{\varepsilon}}};$$

in particular, there is a boundary layer near $x = 1$ for small ε . In two dimensions, we consider a rather straightforward extension of the one-dimensional example to two dimensions: find u such that

$$-\varepsilon \Delta u + \partial_x u = 0 \quad \text{in } (0,1)^2, \quad u(0, \cdot) = 1, \quad u(1, \cdot) = 0, \quad \partial_{\mathbf{n}} u = 0 \text{ if } y = 0 \text{ or } y = 1, \quad (4.2)$$

where $\mathbf{n}$ denotes the unit outward normal vector on the boundary of the domain.

We seek an approximation of the analytical solution in a finite dimensional space that consists of continuous piecewise linear polynomials defined on a given mesh. In one dimension, we are interested both in uniform and non-uniform meshes. In two dimensions we consider predominantly uniform and structured meshes, although we include one example of an unstructured mesh.

If $\varepsilon \ll 1$, then the second-order term is completely dominated by the first-order term and away from the outflow boundary the solution is essentially given by the solution to the advection problem obtained by setting ε to zero. For the above problems this means that $u \approx 1$ away from the outflow boundary. Due to the Dirichlet boundary conditions, a boundary layer forms near the outflow boundary. If the diameter of the elements near the boundary layer is large compared with ε , the layer is fully contained within these elements and, in the above problems, $u \approx 1$ in the rest of the domain. Numerically, this essentially means that we approximate the problems (4.1)/(4.2) with $\varepsilon = 0$ while still keeping the boundary conditions at both ends. Clearly, the analytical solution for the above problems with $\varepsilon = 0$ and the boundary conditions only imposed on the inflow part of the boundary is $u \equiv 1$. This motivates us to consider the best approximations of $u \equiv 1$ by linear finite element functions satisfying the boundary conditions given in (4.1) and (4.2), respectively.

Figure 4.1 shows the L^q -best approximation of $u \equiv 1$ by a piecewise linear function u_h satisfying $u_h(0) = 1$ and $u_h(1) = 0$ on a uniform mesh consisting of four elements with $q = 2$ and $q = 1.2$. We can see that in both cases over- and undershoots are present in the approximation, but that the magnitude of these oscillations is significantly smaller for $q = 1.2$. This example illustrates the

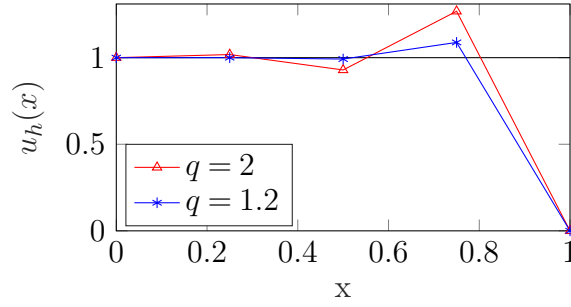


Figure 4.1 L^q -best approximation to $u \equiv 1$ by a piecewise linear function u_h satisfying $u_h(0) = 1$ and $u_h(1) = 0$ on a uniform mesh consisting of four elements with $q = 2$ and $q = 1.2$.

phenomenon of reducing oscillations in the approximation as $q \rightarrow 1$ that we shall investigate in this chapter.

Most of this chapter will be concerned with the precise and very technical analysis of specific situations. Before we delve into this, let us look at the simplest example in order to understand why the over- and undershoots in the L^q -best approximation of discontinuities reduce as $q \rightarrow 1$. To this end, we consider an approximation u_h to $u \equiv 1$, where u_h is a piecewise linear function on a two-element mesh on $(0, 1)$ satisfying $u_h(0) = 1$ and $u_h(1) = 0$. If h is the length of the second element, the free parameter in the approximation is $u_h(1 - h) = 1 + \delta$. Three different choices for δ are shown in Figure 4.2. Clearly, the error with $\delta < 0$ is always larger than the error with $\delta = 0$. Hence, we can assume $\delta \geq 0$. Whether $\delta = 0$ or some $\delta > 0$ yields the better approximation is, however, less obvious. Roughly speaking, replacing $\delta = 0$ with a small $\delta > 0$ increases the overall area that contains an error while at the same time decreasing the area where the pointwise error is close to 1. Therefore, we can expect that in certain situations there exists a $\delta > 0$ that yields an approximation with a smaller error than $\delta = 0$. This argument clearly fails for $\delta \geq 1$, hence we would always expect $\delta < 1$.

The error in the L^q -norm is determined by integrating $|u - u_h|^q$ over the whole interval. As $q \geq 1$ increases, areas containing larger pointwise errors are weighted increasingly more heavily compared to areas containing smaller pointwise errors. This can be seen by looking at the graph of x^q for different values of $q \geq 1$ as shown in Figure 4.2 on the right. As q increases the graph becomes flatter away from 1 such that the range of values for the pointwise error that contribute comparatively little to the L^q -error increases. At the same time the graph becomes steeper close to 1 implying that the contribution of pointwise errors close to 1 to the L^q -error increases. This suggests, that we can expect $\delta_2 > \delta_1$ if $u_{h,i} = 1 + \delta_i$, $i = 1, 2$, are the

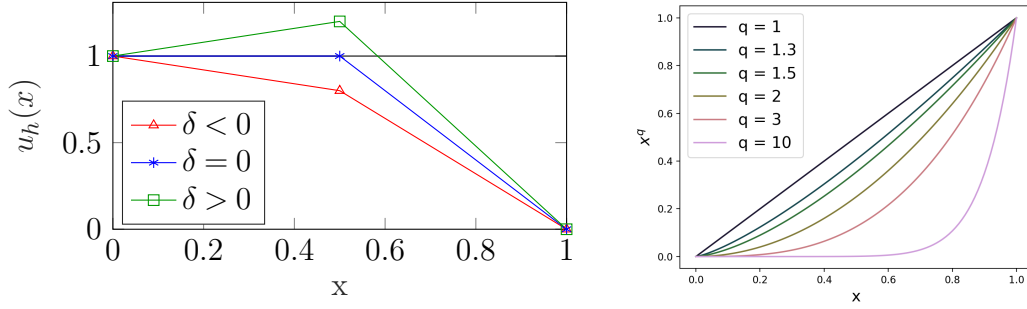


Figure 4.2 Left: Possible approximations to $u \equiv 1$ by a piecewise linear function u_h satisfying $u_h(0) = 1$ and $u_h(1) = 0$ on a uniform mesh consisting of two elements such that $u_h(0.5) = 1 + \delta$. Right: x^q for $x \in [0, 1]$ and several values of $q \geq 1$.

L^{q_i} -best approximations for some $1 \leq q_1 < q_2$. Indeed, we will later see that the overshoot is an increasing function in q . Moreover, for $q > 1$, $\delta = 0$ *never* yields an L^q -best approximation and for $q = 1$ only if the resulting area where $u_h < u$ is sufficiently small. Furthermore, the conjecture $\delta < 1$ will also be confirmed and a plot of the overshoot for large q suggests that $\delta \rightarrow 1$ as $q \rightarrow \infty$. In fact, it is easy to see that $\delta = 1$ yields an L^∞ -best approximation. Indeed, any $\delta \in [-1, 1]$ yields an L^∞ -best approximation since for $|\delta| \leq 1$ the maximal error never exceeds 1, but is always 1 independently of the precise value of δ due to the boundary condition at $x = 1$.

The observation that for larger q it is ‘better’ to commit small errors in the entire domain than a very large error in one element, is similar to the observations made by Guermond, Marpeau and Popov in [86] in the context of residual minimization in L^1 . The authors observe that the L^1 -minimizer commits a large error in one cell and no error in all other cells in contrast to the L^2 -minimizer that spreads the error over all cells. They furthermore observe that this corresponds to selecting a sparse residual vector in the discrete setting which reflects the sparsity property of discrete L^1 -minimizers [65, 66].

4.1.2 Problem Statement

We consider a subdivision Ω_h of the domain $\Omega = (0, 1)^d$, $d = 1, 2$, into n disjoint open simplicial elements (i.e., subintervals when $d = 1$ and triangles when $d = 2$) κ_i , $i = 1, \dots, n$, such that $\bar{\Omega} = \bigcup_{i=1}^n \bar{\kappa}_i$ and define U_h to be the standard finite element space consisting of continuous piecewise linear polynomials on the mesh Ω_h . Let

$u \equiv 1$ and consider the following (constrained) best approximation problem:

$$u_h = \arg \min_{v_h \in U_h} \|u - v_h\|_{L^q((0,1)^d)} \quad (4.3a)$$

subject to

$$\begin{aligned} u_h(0) = 1, \quad u_h(1) &= 0 & \text{if } d = 1, \\ u_h(0, \cdot) = 1, \quad u_h(1, \cdot) &= 0 & \text{if } d = 2. \end{aligned} \quad (4.3b)$$

Note that the constraint can be removed by using a Dirichlet lift argument as commonly employed in the context of finite element methods and restricting the space U_h to the functions that are zero on the part of the boundary where boundary conditions are employed.

In one dimension, instead of $u(x) \equiv 1$, we also consider the L^q -best approximation of $u(x) = \text{sgn}(x)$ on $(-1, 1)$ by a continuous piecewise linear function u_h satisfying $-u_h(-1) = u_h(1) = 1$. We use this example to establish the link between our work and [150].

There is a related body of literature studying the L^2 -projection onto finite element spaces, such as [13, 67, 48]. These works are mostly concerned with the stability of the projection operator in subspaces (e.g., $L^q(\Omega)$, $W^{1,q}(\Omega)$, $H_0^1(\Omega)$).

4.1.3 Summary of Results

The main result of this chapter consists of the precise analysis of very simple examples that illustrate the behaviour of L^q -best approximations of discontinuities by continuous piecewise linear polynomials on coarse meshes. We have furthermore included some numerical examples that confirm the theoretical analysis and illustrate how the observed behaviour in simple model examples applies to more general scenarios. In particular, we demonstrate that the over- and undershoots observed in L^q -best approximations for $1 < q < \infty$ decrease as $q \rightarrow 1$. Whether these oscillations disappear entirely depends on the mesh used to define the underlying finite dimensional approximation space. In one dimension, Gibbs phenomena can be eliminated on uniform meshes both for a boundary discontinuity and a jump discontinuity present in the interior of the domain. For non-uniform meshes it depends on the relative sizes of the elements. In two dimensions, we show that there exist uniform and structured meshes for which Gibbs phenomena are not eliminated. But, we also include examples of meshes in two dimensions on which the over- and undershoots vanish as $q \rightarrow 1$. Furthermore, we will illustrate that

there exist infinitely many L^1 -best approximations in certain cases which is due to the fact that $L^1(\Omega)$ is not strictly convex.

The first example we consider is the approximation problem (4.3) with $d = 1$. The following theorem precisely characterizes the L^q -best approximation for all $1 \leq q < \infty$ for any two-element mesh on $(0,1)$. Note that in this case, the approximation problem only has one degree of freedom due to the boundary conditions. Furthermore, we prove for an N -element mesh that there exists an L^1 -best approximation with no over- or undershoot if either a grading-type mesh condition is satisfied or a stronger, simple element-size condition. The precise result is given below.

Theorem 4.1 (L^q -best approximation of a boundary discontinuity).

1. Consider the mesh given by the subdivision of $(0,1)$ into the two intervals $(0, 1-h)$ and $(1-h, 1)$ with $h \in (0,1)$. For $1 \leq q < \infty$, the solution of the approximation problem (4.3) with $u \equiv 1$ and $d = 1$ is given by a continuous piecewise linear polynomial u_h that satisfies the boundary conditions and $u_h(1-h) = \alpha$, where α is defined as follows

(a) If $q = 1$,

$$\alpha = \begin{cases} 1 & \text{if } h \leq 0.5 \\ \sqrt{2h} & \text{if } h > 0.5. \end{cases} \quad (4.4)$$

(b) If $1 < q < \infty$, then $\alpha > 1$ and

$$0 = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q + 1)(\alpha-1)^q + h.$$

2. Let the mesh be given by a subdivision of the interval $(0,1)$ into $N \geq 2$ intervals (x_{i-1}, x_i) , $i = 1, \dots, N$, with $0 = x_0 < x_2 < \dots < x_{N-1} < x_N = 1$. The length h_i of the i th subinterval is given by $h_i = x_i - x_{i-1}$, $i = 1, \dots, N$. Define

$$\begin{aligned} \vartheta_N &:= 0, \\ \vartheta_i^2 &:= \frac{1}{2} \left(1 - \left(2(1 - \vartheta_{i+1})^2 - 1 \right) \frac{h_{i+1}}{h_i} \right), i = N-1, \dots, 1, \\ M &:= \max \left(\{0\} \cup \left\{ i \in \{1, \dots, N-1\} : \vartheta_i \geq 1 - \frac{1}{\sqrt{2}} \right\} \right). \end{aligned}$$

Then

$$h_i \geq (2(1 - \vartheta_{i+1})^2 - 1)h_{i+1}, \quad \text{for } i = M, M+1, \dots, N-1, \quad (4.5)$$

is a sufficient condition for the existence of an L^1 -best approximation u_h of $u \equiv 1$ with $u_h(0) = 1$ and $u_h(1) = 0$ satisfying $u_h(x_i) = 1$ for all $i = 1, \dots, N-1$. Hence, u_h contains no over- or undershoots.

3. Let the mesh be given by a subdivision of the interval $(0,1)$ into $N \geq 2$ intervals (x_{i-1}, x_i) , $i = 1, \dots, N$, with $0 = x_0 < x_1 < \dots < x_{N-1} < x_N = 1$. The length h_i of the i th subinterval is given by $h_i = x_i - x_{i-1}$, $i = 1, \dots, N$. Then $h_N \leq \min_{i=1, \dots, N-1} h_i$ is a sufficient condition for the existence of an L^1 -best approximation u_h of $u \equiv 1$ with $u_h(0) = 1$ and $u_h(1) = 0$ satisfying $u_h(x_i) = 1$ for all $i = 1, \dots, N-1$.

Remark 4.2. Note that in the second part of Theorem 4.1, condition (4.5) essentially states that elements cannot be too small compared to their neighbouring element closer to the discontinuity. Furthermore, there are no conditions on the size of the elements contained in $(0, x_{M-1})$ if $M > 0$. Moreover, it is always possible to ensure $M > 0$ by selecting h_M sufficiently large in comparison to h_{M+1} such that $\vartheta_M > 1 - 1/\sqrt{2}$. This means that the mesh can be designed in such a way that it is allowed to be arbitrary away from the discontinuity without leading to oscillations. This observation is particularly useful if more than one discontinuity is to be approximated.

Remark 4.3. With very similar arguments as in the proof of the final part of Theorem 4.1, it is easy to see that if $h_N > h_{N-1}$, but $h_{N-1} \leq h_i$ for all $i = 1, \dots, N-2$, then every L^1 -best approximation must contain over- or undershoots. Moreover, there exists an L^1 -best approximation with overshoot only at the node x_{N-1} and no further over- or undershoots, i.e., $u_h(x_i) = 1$ for $i = 1, \dots, N-2$ and

$$u_h(x_{N-1}) = \sqrt{\frac{2h_N}{h_N + h_{N-1}}}.$$

The value at $u_h(x_{N-1})$ follows from the first part of the theorem by rescaling the interval.

Figure 4.3 shows α specified in Theorem 4.1 for two different ranges of q and three different choices of h . The plot shows that $\alpha < 2$ for all $1 \leq q < \infty$ and that α decreases as $q \rightarrow 1$ for all three choices of h . Furthermore, we can see that the behaviour as $q \rightarrow \infty$ is very similar for all choices of h , but that there are clear

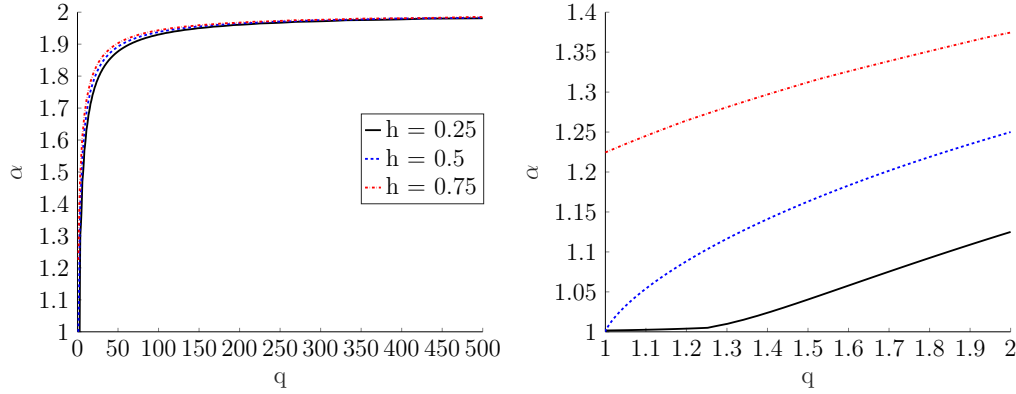


Figure 4.3 Theorem 4.1: Values for α for different ranges of q and three different choices of h .

differences as $q \rightarrow 1$. For $h = 0.25$ and $h = 0.5$, α approaches 1 as $q \rightarrow 1$, hence the overshoot vanishes as $q \rightarrow 1$, whereas for $h = 0.75$ it approaches $\sqrt{2h} \approx 1.2247$, hence the overshoot does not vanish. This is consistent with the results obtained for the L^1 -best approximation, cf., (4.4).

In order to illustrate the graded mesh condition in part two of Theorem 4.1, we consider the following example consisting of two three-element meshes violating the sufficient condition in part three of Theorem 4.1 such that one of the meshes satisfies (4.5), whereas the other mesh violates this condition as well.

Example 4.4 (Two three-element meshes). *Consider the following two three-element meshes on $(0,1)$: for the first one we choose $h_1 = 0.1$ and $h_2 = h_3 = 0.45$, i.e., the mesh consists of the subintervals $(0,0.1)$, $(0.1,0.55)$ and $(0.55,1)$; For the second mesh we choose $h_1 = 0.1$, $h_2 = 0.5$ and $h_3 = 0.4$, i.e., the mesh consisting of the subintervals $(0,0.1)$, $(0.1,0.6)$ and $(0.6,1)$.*

In the first case, (4.5) is violated. Moreover, there is no L^1 -best approximation without over- or undershoots and the unique L^1 -best approximation satisfies $u_h(0.1) \approx 0.9931$ and $u_h(0.55) \approx 1.0247$.

The second mesh satisfies (4.5). Hence, there exists an L^1 -best approximation with no over- or undershoots.

The corresponding proof is presented in Section 4.2.4. In Section 4.5.2, we furthermore include a numerical illustration of the behaviour of the L^q -best approximation. We will demonstrate that for the latter mesh the overshoot does indeed not vanish entirely as $q \rightarrow 1$.

The second best-approximation problem we analyse is the best approximation of $u(x) = \text{sgn}(x)$ on $(-1,1)$ on a mesh consisting of exactly four elements that is

symmetric with respect to $x = 0$. The main difference to the result in part one of Theorem 4.1 is that there exists a whole family of best approximations if $q = 1$. For $q > 1$, we observe the same behaviour as before.

Theorem 4.5 (L^q -best approximation of a jump discontinuity). *Consider the mesh given by the subdivision of $(-1, 1)$ into the four intervals $(-1, -h)$, $(-h, 0)$, $(0, h)$ and $(h, 1)$ with $h \in (0, 1)$. For $1 \leq q < \infty$, the L^q -best approximation of $u = \text{sgn}(x)$ on $(-1, 1)$ by a continuous piecewise linear function u_h on the above mesh such that $-u_h(-1) = u_h(1) = 1$ can be characterized as follows.*

1. *If $h \leq 0.5$, there exists an L^1 -best approximation for any $\beta \in [-1, 1]$ such that $u_h(0) = \beta$ and $-u_h(-h) = u_h(h) = 1$. Conversely, any L^1 -best approximation satisfies $-u_h(-h) = u_h(h) = 1$ and $u_h(0) \in [-1, 1]$.*
2. *If $h > 0.5$, then for any $\beta \in [-1, 1]$, u_h satisfies*

$$\begin{aligned} u_h(-h) &= \alpha := -\sqrt{2h} - \beta(\sqrt{2h} - 1), \\ u_h(0) &= \beta, \\ u_h(h) &= \gamma := \sqrt{2h} - \beta(\sqrt{2h} - 1), \end{aligned}$$

defining an L^1 -best approximation. Conversely, any L^1 -best approximation satisfies $u_h(0) \in [-1, 1]$ and is of the above form.

3. *The unique L^q -best approximation with $1 < q < \infty$ is given by $-u_h(-h) = u_h(h) = \alpha$ and $u_h(0) = 0$, where α satisfies*

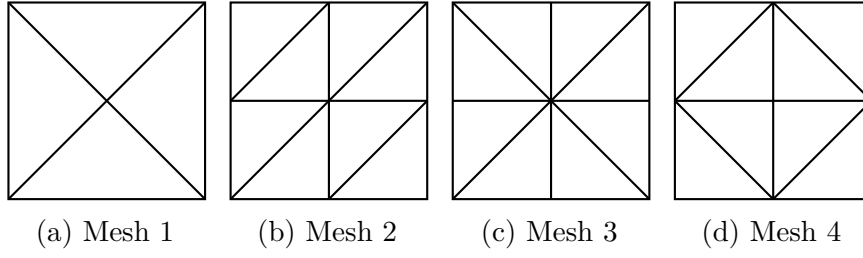
$$0 = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q + 1)(\alpha-1)^q + h,$$

and $\alpha > 1$.

4. *In the limit $q \rightarrow 1$ the L^q -best approximation converges to the L^1 -best approximation with $u_h(0) = 0$ for any $h \in (0, 1)$. The corresponding L^1 -best approximation is anti-symmetric and satisfies*

$$-u_h(-h) = u_h(h) = \begin{cases} 1 & \text{if } h \leq 0.5, \\ \sqrt{2h} & \text{if } h > 0.5. \end{cases}$$

We again observe that the presence of over- and undershoots in the L^1 -best approximation depends on the choice of mesh. Furthermore, there exists a whole family of L^1 -best approximations in this case which is possible since L^1 is not

Figure 4.4 Four different meshes on $(0,1)^2$.

strictly convex and therefore minimizers are not necessarily unique. We recover uniqueness if we define the minimizer as the limit as $q \rightarrow 1$ of the L^q -minimizer. Moreover, it follows from the proof of Theorem 4.5 that the L^1 -best approximation is unique if the subdivision of the interval is no longer symmetric, as we will see in Section 4.3.

In order to see how this result relates to the work in [150], it first has to be noted that there are two major differences between our investigation and [150]:

1. The interval in [150] is subdivided into $2n$ subintervals of equal length. In contrast to this, we only consider the special case that $(-1,1)$ is subdivided into 4 subintervals and instead allow the subdivision to be non-uniform but still symmetric with respect to the center of the interval.
2. We consider bounded domains with fixed boundary conditions, which are relevant to finite element approximations, whereas the investigation in [150] considers the limit $n \rightarrow \infty$ for the interval $[-nh, nh]$ (ergo essentially an infinite domain) with no boundary conditions.

In [150] it is shown that for a uniform subdivision of the interval $[-nh, nh]$, the over- and undershoots disappear as $n \rightarrow \infty$ and $q \rightarrow 1$. The last point in Theorem 4.5 shows that, on a fixed mesh, we recover the result that the over- and undershoots disappear as $q \rightarrow 1$ for $h \leq 0.5$, which includes the case of a uniform mesh. However, if $h > 0.5$, the over- and undershoots do not disappear as $q \rightarrow 1$.

The final theoretical result concerns the solution to (4.3) with $d = 2$ on the four meshes shown in Figure 4.4. Note that the discrete space U_h has only one degree of freedom on Mesh 1, corresponding to the value at the midpoint, and U_h has three degrees of freedom on the other meshes, corresponding to the values at the three nodes on the line $x = 0.5$. For the first mesh, we analyse the L^q -best approximation for all $1 \leq q < \infty$ and show that the solution contains an overshoot that does not disappear as $q \rightarrow 1$. For Mesh 2, we show that any L^1 -best approximation must contain over- or undershoots and characterize an L^1 -best approximation.

Furthermore, we prove that there exists an L^1 -best approximation on Meshes 3 and 4 without over- or undershoots. Finally, we demonstrate numerically in Section 4.5.3.1 that the L^q -best approximation on Meshes 2, 3 and 4 indeed approaches the L^1 -best approximation characterized in the theorem below.

Theorem 4.6 (L^q -best Approximation of a Boundary Discontinuity in Two Dimensions).

1. The unique solution to the best approximation problem (4.3) on Mesh 1 for $q = 1$ is given defined by $u_h((0.5, 0.5)) = \alpha$, where α satisfies

$$\alpha > 1 \quad \text{and} \quad 0 = 2\alpha^3 - 5\alpha + 2,$$

hence $\alpha \approx 1.3200$.

2. The unique solution to the best approximation problem (4.3) on Mesh 1 for $1 < q < \infty$ is defined by $u_h((0.5, 0.5)) = \alpha$, where α satisfies $\alpha > 1$ and

$$0 = (\alpha - 1)^{q-1} [4\alpha^3 q + 4(1 - q)\alpha^2 + (q - 6)\alpha + 2] - \alpha(q + 4) + 2.$$

3. If $q > 1$, the L^q -best approximation to (4.3) contains over- or undershoots on all four meshes.
4. If $q = 1$, there exists a solution to (4.3) on Mesh 2 such that $u_h((0.5, 1)) = u_h((0.5, 0.5)) = 1$ and $u_h((0.5, 0)) = \alpha$, where α satisfies

$$\alpha > 1 \quad \text{and} \quad 0 = -3\alpha^3 + 8\alpha - 4,$$

hence $\alpha \approx 1.2723$. Furthermore, $u_h((0.5, 1)) = u_h((0.5, 0.5)) = u_h((0.5, 0)) = 1$ does not define an L^1 -best approximation.

5. If $q = 1$, there exists a solution to (4.3) on Meshes 3 and 4 such that $u_h((0.5, 1)) = u_h((0.5, 0.5)) = u_h((0.5, 0)) = 1$.

Theorem 4.6 shows that, on Meshes 1 and 2, the L^q -best approximations exhibit an overshoot for all q , including $q = 1$, while on Meshes 3 and 4 the L^1 -best approximation does not contain any over- or undershoots.

Figure 4.5 shows the parameter α defining the L^q -best approximation on Mesh 1 for two different ranges of q . The plot shows that $\alpha < 2$ for all q and that α decreases as $q \rightarrow 1$, where it approaches 1.32. This is consistent with the result in Theorem 4.6 obtained for the L^1 -best approximation. To confirm the theoretical results,

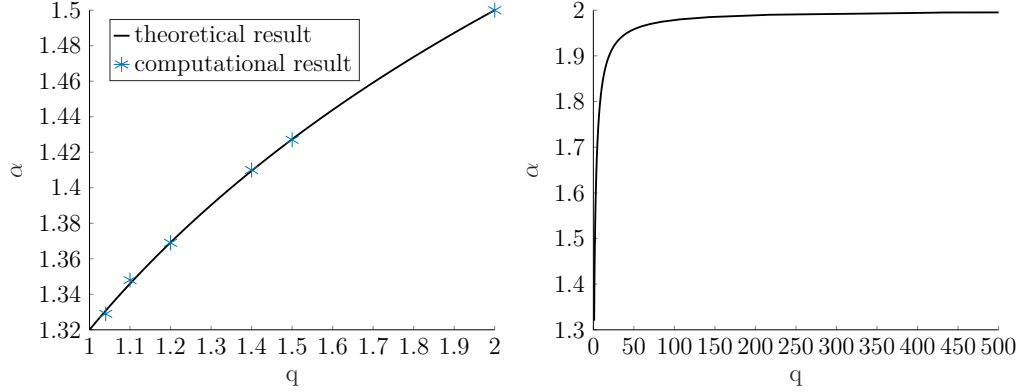


Figure 4.5 Theorem 4.6: Values for α for different ranges of q on Mesh 1.

we have also determined the L^q -best approximation numerically by implementing (4.29) using FEniCS [5]. The solution to the resulting non-linear system can be approximated using a Newton iteration if q is sufficiently close to 2. Note that this solver is not robust in q and stalls or diverges for q close to 1 and for $q \gg 2$. The left plot in Figure 4.5 shows numerically determined approximations of α for selected values of q which confirm the theoretical results.

We also include further numerical experiments in Section 4.5 illustrating that the observations remain the same if u is a more general smooth function and that the over- and undershoots cannot be eliminated by refining the mesh.

4.1.4 Characterization of Best L^q -Approximation

For convenience, we quote Corollaries 3.30 and 3.31 as stated in Section 3.4. These two Corollaries state a useful characterization of L^1 - and L^q -best approximations, respectively, in closed subspaces. Throughout this chapter we will use these characterizations to determine L^q -best approximations in finite element spaces.

Corollary 4.7. *Let $U := L^q(\Omega)$, $1 < q < \infty$, and $U_h \subset U$ a closed subspace. The function $u_h \in U_h$ is an L^q -best approximation of u if and only if*

$$\int_{\Omega} \operatorname{sgn}(u - u_h) |u - u_h|^{q-1} v_h \, d\mathbf{x} = 0 \quad \forall v_h \in U_h. \quad (4.7)$$

Corollary 4.8. *Let $U := L^1(\Omega)$ and $U_h \subset U$ a closed subspace. The function $u_h \in U_h$ is an L^1 -best approximation of u if and only if there exists a function $\psi_0 \in L^\infty(\Omega \cap \{u(\mathbf{x}) = u_h(\mathbf{x})\})$, $|\psi_0| \leq 1$, almost everywhere, such that for all*

$$v_h \in U_h,$$

$$0 = \int_{\Omega \cap \{u(\mathbf{x}) > u_h(\mathbf{x})\}} v_h d\mathbf{x} - \int_{\Omega \cap \{u(\mathbf{x}) < u_h(\mathbf{x})\}} v_h d\mathbf{x} + \int_{\Omega \cap \{u(\mathbf{x}) = u_h(\mathbf{x})\}} \psi_0 v_h d\mathbf{x}, \quad (4.8)$$

or, equivalently, for all $v_h \in U_h$,

$$0 = \int_{\Omega} \psi v_h d\mathbf{x},$$

where $\psi = \text{sgn}(u - u_h)$ on $\{u(\mathbf{x}) \neq u_h(\mathbf{x})\}$ and $\psi = \psi_0$ on $\{u(\mathbf{x}) = u_h(\mathbf{x})\}$.

Note that in the case that u and u_h only agree on a set of measure zero, the choice of $\psi_0 \in [-1, 1]$ becomes irrelevant.

4.1.5 Outline of the Chapter

The remainder of this chapter is organized as follows: Sections 4.2, 4.3 and 4.4 contain the proofs of Theorems 4.1, 4.5 and 4.6, respectively. We conclude with several numerical examples in Section 4.5 illustrating the effect of mesh refinement in one and two dimensions and showing the behaviour of the L^q -best approximation as $q \rightarrow 1$ in one dimension, as well as on structured and unstructured meshes in two dimensions.

4.2 Best Approximation of a Boundary Discontinuity in One Dimension

In this section we consider the best approximation problem (4.3) in one dimension and provide a proof of Theorem 4.1. We split this into three parts: Sections 4.2.1 and 4.2.2 contain the proof of the first part of the theorem, where the former addresses the case $q = 1$ and the latter the case $1 < q < \infty$; Section 4.2.3 contains the proof of the second and third part of the theorem. Finally, Section 4.2.4 contains the analysis of Example 4.4.

In the first part of Theorem 4.1, we consider the mesh consisting of the two subintervals $(0, 1 - h)$ and $(1 - h, 1)$. The best approximation u_h of $u \equiv 1$ by a continuous piecewise linear function satisfying the boundary conditions $u_h(0) = 1$ and $u_h(1) = 0$ is determined entirely by the value it takes at the point $x = 1 - h$.

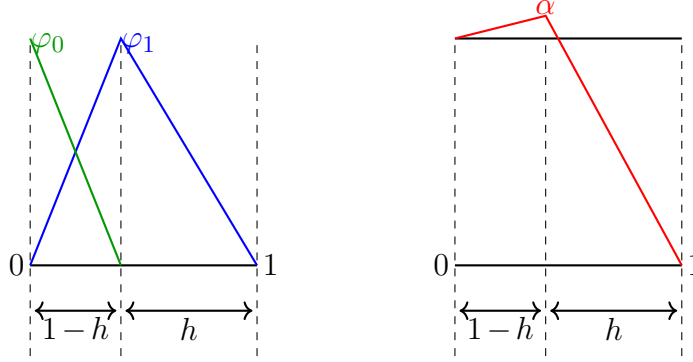


Figure 4.6 Proof of Theorem 4.1: Hat functions φ_0 and φ_1 (left). Approximation u_h with $\alpha > 1$ (right).

Thereby, we can write $u_h = \varphi_0 + \alpha\varphi_1$, where α is to be determined and

$$\varphi_0 = \begin{cases} \frac{(1-h)-x}{1-h} & \text{in } [0, 1-h], \\ 0 & \text{else,} \end{cases} \quad \varphi_1 = \begin{cases} \frac{x}{1-h} & \text{in } [0, 1-h], \\ \frac{1-x}{h} & \text{in } [1-h, 1]. \end{cases}$$

Figure 4.6 shows the two functions φ_0 and φ_1 as well as an approximation u_h of $u \equiv 1$ with $\alpha > 1$. To eliminate the constraint by introducing a Dirichlet lift in the best approximation problem (4.3), we could define the subspace U_h as the span of φ_1 and redefine $u = 1 - \varphi_0$ and $u_h = \alpha\varphi_1$. Note, however, that $u - u_h$ remains the same. The main consequence of this observation is, that (4.7) and (4.8) do not have to be satisfied for $w_h = \varphi_0$ due to the boundary condition constraint.

4.2.1 L^1 -Best Approximation

In this section we give a proof of the first part of Theorem 4.1 for the case $q = 1$. More precisely, we show that the L^1 -best approximation of $u \equiv 1$ on $(0, 1)$ by a continuous piecewise linear function u_h satisfying the boundary conditions $u_h(0) = 0$ and $u_h(1) = 1$ is given by $u_h = \varphi_0 + \alpha\varphi_1$, where

$$\alpha = \begin{cases} 1 & \text{if } h \leq 0.5, \\ \sqrt{2h} & \text{if } h > 0.5. \end{cases} \quad (4.9)$$

Proof. Using the characterization of the best approximation given in Corollary (4.8), to determine the best approximation, we distinguish between two cases:

1. The set $\{x \in (0, 1) : (u - u_h)(x) = 0\}$ has measure zero. (For continuous piecewise linear functions this set has to consist of a finite number of points).

This means that $\psi = \operatorname{sgn}(u - u_h)$ everywhere except on a set of measure zero and is thus uniquely defined almost everywhere.

2. The set $\{x \in (0, 1) : (u - u_h)(x) = 0\}$ has positive measure, i.e., the set contains an interval of positive length. This means that ψ is not uniquely defined on a set with positive measure.

Starting with the second case, we observe that this can only be true if $\alpha = 1$ and thus $u = u_h$ in $(0, 1 - h)$. In this case $\psi = \operatorname{sgn}(u - u_h) = 1$ in $(1 - h, 1)$ and we compute

$$\int_{1-h}^1 \psi \varphi_1 dx = \int_{1-h}^1 \frac{1-x}{h} dx = \frac{h}{2}.$$

Assuming $-h/(1-h) \in [-1, 1]$, we can choose $\psi_0 = -h/(1-h)$ in $(0, 1-h)$ and obtain

$$\int_0^{1-h} \psi \varphi_1 dx = -\frac{h}{1-h} \underbrace{\int_0^{1-h} (x+1) dx}_{=(1-h)/2} = -\frac{h}{1-h} \frac{1-h}{2} = -\frac{h}{2}.$$

Obviously, adding both integrals yields zero for this choice of ψ_0 . Therefore, by Corollary 4.8, $\alpha = 1$ is an L^1 -best approximation provided that $-h/(1-h) \in [-1, 1]$. This is the case if and only if $h \leq 1-h \Leftrightarrow h \leq 1/2$, hence proving (4.9) for $h \leq 1/2$.

If on the other hand $h > 1/2 \Leftrightarrow 1-h < h$, $\alpha = 1$ does not yield an L^1 -best approximation. Indeed,

$$\int_0^1 \psi \varphi_1 dx \geq \frac{h}{2} - \frac{1-h}{2} > 0,$$

since $\psi = 1$ in $(0, 1-h)$ and $\psi \geq -1$ in $(1-h, 1)$.

Next we consider the first case; this implies that $\alpha \neq 1$ and $\psi = \operatorname{sgn}(u - u_h)$ almost everywhere in $(0, 1)$. If $\alpha < 1$, we have $u - u_h > 0$ everywhere in $(0, 1)$ and since $\varphi_0 > 0$ everywhere in $(0, 1)$, we also have

$$\int_0^1 \operatorname{sgn}(u - u_h) \varphi_1 dx > 0.$$

Therefore, $u_h = \varphi_0 + \alpha\varphi_1$ cannot be an L^1 -best approximation if $\alpha < 1$. This leaves the case $\alpha > 1$. Now, $u - u_h < 0$ in $(0, \frac{\alpha-h}{\alpha})$ and $u - u_h > 0$ in $(\frac{\alpha-h}{\alpha}, 1)$; we compute

$$\begin{aligned} \int_0^1 \operatorname{sgn}(u - u_h) \varphi_1 dx &= - \int_0^{1-h} \frac{x}{1-h} dx - \int_{1-h}^{\frac{\alpha-h}{\alpha}} \frac{1-x}{h} dx + \int_{\frac{\alpha-h}{\alpha}}^1 \frac{1-x}{h} dx \\ &= -\frac{1-h}{2} - \frac{h(\alpha^2-1)}{2\alpha^2} + \frac{h}{2\alpha^2} = \frac{2h-\alpha^2}{2\alpha^2}. \end{aligned}$$

This integral becomes 0 for $\alpha = \sqrt{2h}$. Note that this only yields an L^1 -best approximation if $h > 1/2$. Indeed, if $h \leq 1/2$, then $\alpha = \sqrt{2h} \leq 1$, but we have assumed $\alpha > 1$.

We have therefore shown that $\alpha = 1$ is the only L^1 -best approximation if $h \leq 1/2$ and $\alpha = \sqrt{2h}$ is the only L^1 -best approximation if $h > 1/2$. $\square$

4.2.2 L^q -Best Approximation

In this section we give a proof of the first part of Theorem 4.1 for the case $1 < q < \infty$. More precisely, we show that the L^q -best approximation of $u \equiv 1$ on $(0, 1)$ by a continuous piecewise linear function u_h satisfying the boundary conditions $u_h(0) = 0$ and $u_h(1) = 1$ is given by $u_h = \varphi_0 + \alpha\varphi_1$, where $\alpha > 1$ and

$$0 = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q + 1)(\alpha-1)^q + h.$$

Proof. Corollary 4.7 implies that we seek α such that

$$\int_0^1 \operatorname{sgn}(u - u_h) |u - u_h|^{q-1} \varphi_1 dx = 0.$$

Again, we have to split the integral on each element into the parts where $u - u_h > 0$, $u - u_h < 0$ and $u - u_h = 0$. We consider three cases

$$(a) \ \alpha < 1, \quad (b) \ \alpha = 1, \quad (c) \ \alpha > 1.$$

If $\alpha < 1$, we have $u - u_h > 0$ everywhere in $(0, 1)$ and thus both $\operatorname{sgn}(u - u_h) |u - u_h|^{q-1} > 0$ and $\varphi_1 > 0$ in $(0, 1)$. Therefore, we have

$$\int_0^1 \operatorname{sgn}(u - u_h) |u - u_h|^{q-1} \varphi_1 dx > 0,$$

hence $\alpha < 1$ is not possible. If $\alpha = 1$, we have $u - u_h = 0$ in $(0, 1 - h)$ and $u - u_h > 0$ in $(1 - h, 1)$. Thus,

$$\int_0^1 \operatorname{sgn}(u - u_h) |u - u_h|^{q-1} \varphi_1 \, dx = \int_{1-h}^1 \operatorname{sgn}(u - u_h) |u - u_h|^{q-1} \varphi_1 \, dx > 0,$$

hence $\alpha = 1$ is not possible. We can therefore assume $\alpha > 1$. In this case $u - u_h < 0$ in $(0, \frac{\alpha-h}{\alpha})$ and $u - u_h > 0$ in $(\frac{\alpha-h}{\alpha}, 1)$. We compute

$$\begin{aligned} & \int_0^1 \operatorname{sgn}(u - u_h) |u - u_h|^{q-1} \varphi_1 \, dx \\ &= - \int_0^{1-h} \left(\frac{(\alpha-1)x}{1-h} \right)^{q-1} \frac{x}{1-h} \, dx \\ & \quad - \int_{1-h}^{\frac{\alpha-h}{\alpha}} \left(\frac{\alpha-h}{h} - \frac{\alpha x}{h} \right)^{q-1} \frac{1-x}{h} \, dx + \int_{\frac{\alpha-h}{\alpha}}^1 \left(\frac{\alpha x}{h} - \frac{\alpha-h}{h} \right)^{q-1} \frac{1-x}{h} \, dx \\ &= - \frac{(1-h)(\alpha-1)^{q-1}}{q+1} - \frac{h(\alpha q+1)(\alpha-1)^q}{\alpha^2 q(q+1)} + \frac{h}{\alpha^2 q(q+1)} \\ &= \frac{-(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q+1)(\alpha-1)^q + h}{\alpha^2 q(q+1)}. \end{aligned}$$

Hence, the L^q -best approximation can be determined by finding $\alpha > 1$ satisfying

$$0 = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q+1)(\alpha-1)^q + h,$$

existence of which is guaranteed since the L^q -best approximation always exists. $\square$

4.2.3 Sufficient Conditions on General Meshes

In this section we provide a proof of the second and third parts of Theorem 4.1. To this end, let the mesh be given by a subdivision of the interval $(0, 1)$ into $N \geq 2$ subintervals (x_{i-1}, x_i) , $i = 1, \dots, N$, with $0 = x_0 < x_2 < \dots < x_{N-1} < x_N = 1$. The length h_i of the i th subinterval is given by $h_i = x_i - x_{i-1}$, $i = 1, \dots, N$. In order to prove the second part of Theorem 4.1, we show that the following conditions are sufficient for the existence of an L^1 -best approximation u_h of $u \equiv 1$ with $u_h(0) = 1$ and $u_h(1) = 0$ satisfying $u_h(x_i) = 1$ for all $i = 1, \dots, N-1$:

$$h_i \geq (2(1 - \vartheta_{i+1})^2 - 1)h_{i+1} \quad \text{for } i = M, M+1, \dots, N-1, \quad (4.10)$$

where

$$\vartheta_N := 0, \quad (4.11)$$

$$\vartheta_i^2 := \frac{1}{2} \left(1 - \left(2(1 - \vartheta_{i+1})^2 - 1 \right) \frac{h_{i+1}}{h_i} \right), i = N-1, \dots, 1, \quad (4.12)$$

$$M := \max \left(\{0\} \cup \left\{ i \in \{1, \dots, N-1\} : \vartheta_i \geq 1 - \frac{1}{\sqrt{2}} \right\} \right). \quad (4.13)$$

We then show that the much simpler condition $h_N \leq \min_{i=1, \dots, N-1} h_i$ implies (4.10) which proves the third part of Theorem 4.1.

Proof. Define u_h such that $u_h(0) = 1$, $u_h(1) = 0$ and $u_h(x_i) = 1$ for all $i = 1, \dots, N-1$. Furthermore, denote by φ_i the hat function that is 1 at x_i and 0 at x_j , $j \neq i$, for $i = 1, \dots, N-1$.

For $\alpha \in (0, 1]$, define $\psi_\alpha(x)$ as follows:

$$\psi_\alpha(x) = \begin{cases} (-1)^{N-i+1} & x \in (x_{i-1}, x_{i-1} + \vartheta_i h_i), & \text{for all } i = M+1, \dots, N, \\ (-1)^{N-i} & x \in (x_{i-1} + \vartheta_i h_i, x_i) & \text{for all } i = M+1, \dots, N, \\ (-1)^{N-M+1} \alpha & x \in (x_{M-1}, x_{M-1} + \tilde{\vartheta}_M h_M) & \text{if } M > 0, \\ (-1)^{N-M} & x \in (x_{M-1} + \tilde{\vartheta}_M h_M, x_M) & \text{if } M > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\tilde{\vartheta}_M^2 = \frac{2}{1+\alpha} \vartheta_M^2 = \frac{1}{\alpha+1} \left(1 - \left(2(1 - \vartheta_{M+1})^2 - 1 \right) \frac{h_{M+1}}{h_M} \right).$$

We claim that there exists $\alpha \in (0, 1]$ such that

$$\int_0^1 \psi_\alpha(x) \varphi_i(x) dx = 0 \quad \forall i = 1, \dots, N-1, \quad (4.14)$$

if (4.10) is satisfied. First we show that ψ_α is well defined. For this we require ϑ_i to be well defined, i.e., we require $\vartheta_i \in [0, 1]$, for all $i = M, \dots, N-1$. This is trivially true for ϑ_N . Otherwise, for $i \geq M$,

$$\vartheta_i^2 \geq 0 \Leftrightarrow h_i \geq (2(1 - \vartheta_{i+1})^2 - 1) h_{i+1}.$$

Furthermore, $\vartheta_{i+1} < 1 - 1/\sqrt{2}$ for $i \geq M$ by definition of M and therefore

$$(1 - \vartheta_{i+1})^2 > \frac{1}{2} \Rightarrow (2(1 - \vartheta_{i+1})^2 - 1) > 0 \Rightarrow \vartheta_i^2 < \frac{1}{2} \Rightarrow \vartheta_i < 1.$$

Next, note that $\vartheta_N = 0$ implies $\psi_\alpha(x) = 1 = \text{sgn}(u - u_h)$ on (x_{N-1}, x_N) and that $\|\psi_\alpha\|_{L^\infty((0,1))} = 1$. Thus, proving (4.14) immediately implies that u_h is an L^1 -best approximation of $u \equiv 1$.

For all $i = 1, \dots, N-1$ it holds that

$$\int_0^1 \psi_\alpha(x) \varphi_i(x) dx = \int_{x_{i-1}}^{x_i} \psi_\alpha(x) \varphi_i(x) dx + \int_{x_i}^{x_{i+1}} \psi_\alpha(x) \varphi_i(x) dx. \quad (4.15)$$

With this in mind, we now consider (4.14) for $i > M$. In this case

$$\int_{x_{i-1}}^{x_i} \psi_\alpha(x) \varphi_i(x) dx = (-1)^{N-i+1} \frac{h_i}{2} (2\vartheta_i^2 - 1), \quad (4.16)$$

$$\int_{x_i}^{x_{i+1}} \psi_\alpha(x) \varphi_i(x) dx = (-1)^{N-i+2} \frac{h_{i+1}}{2} (1 - 2(1 - \vartheta_{i+1})^2), \quad (4.17)$$

Hence (4.15) becomes zero if and only if

$$\begin{aligned} 0 &= \frac{h_i}{2} (2\vartheta_i^2 - 1) - \frac{h_{i+1}}{2} (1 - 2(1 - \vartheta_{i+1})^2) \\ \Leftrightarrow \quad \vartheta_i^2 &= \frac{1}{2} \left(1 - (2(1 - \vartheta_{i+1})^2 - 1) \frac{h_{i+1}}{h_i} \right). \end{aligned}$$

which is the definition of ϑ_i .

Next, we consider $i = M$. We have already established $\vartheta_M^2 \leq 1/2$. With $\alpha > 0$, this implies $\tilde{\vartheta}_M < 1$. Furthermore, (4.17) still holds with $i = M$, whereas we obtain

$$\int_{x_{M-1}}^{x_M} \psi_\alpha \varphi_i dx = (-1)^{N-M+1} ((\alpha + 1)\vartheta_M^2 - 1) \frac{h_M}{2}.$$

Hence (4.15) becomes zero for $i = M$ if and only if

$$\begin{aligned} 0 &= \frac{h_M}{2} ((1 + \alpha)\tilde{\vartheta}_M^2 - 1) - \frac{h_{M+1}}{2} (2(1 - \vartheta_{M+1})^2 - 1) \\ \Leftrightarrow \quad \tilde{\vartheta}_M^2 &= \frac{1}{1 + \alpha} \left(1 - (2(1 - \vartheta_{M+1})^2 - 1) \frac{h_{M+1}}{h_M} \right), \end{aligned}$$

which is the definition of $\tilde{\vartheta}_M$. Finally, we have to show that there exists $\alpha \in (0, 1]$ such that

$$\begin{aligned} 0 &= \int_0^1 \psi_\alpha \varphi_{M-1} dx = \underbrace{\int_{x_{M-2}}^{x_{M-1}} \psi_\alpha \varphi_{M-1}}_{=0, \text{ since } \psi_\alpha(x)=0 \text{ for } x < x_{M-1}} + \int_{x_{M-1}}^{x_M} \psi_\alpha \varphi_{M-1} dx \\ &= (-1)^{N-M+1} (\alpha - (\alpha+1)(1-\tilde{\vartheta}_M)^2) \frac{h_M}{2}. \end{aligned}$$

The last expression becomes zero if and only if

$$(1 - \tilde{\vartheta}_M)^2 = \frac{\alpha}{1 + \alpha} \Leftrightarrow \tilde{\vartheta}_M = \frac{1}{\sqrt{1 + \alpha}} (\sqrt{1 + \alpha} - \sqrt{\alpha}).$$

Hence, we need α such that

$$(\sqrt{1 + \alpha} - \sqrt{\alpha}) = \sqrt{\left(1 - (2(1 - \vartheta_{M+1})^2 - 1) \frac{h_{M+1}}{h_M}\right)}. \quad (4.18)$$

For $\alpha > 0$, $g(\alpha) = \sqrt{1 + \alpha} - \sqrt{\alpha}$, is a strictly decreasing function of α and thus bijectively maps $(0, 1]$ onto $[\sqrt{2} - 1, 1)$. The equation (4.18) therefore has a unique solution $\alpha \in (0, 1]$ if and only if

$$\sqrt{\left(1 - (2(1 - \vartheta_{M+1})^2 - 1) \frac{h_{M+1}}{h_M}\right)} \in [\sqrt{2} - 1, 1) \Leftrightarrow \vartheta_M \in \left[1 - \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).$$

By the definition of M , we have $\vartheta_M > 1 - \frac{1}{\sqrt{2}}$ and $\vartheta_{M+1} < 1 - \frac{1}{\sqrt{2}} \Rightarrow \vartheta_M < \frac{1}{\sqrt{2}}$. This shows that (4.10) is indeed a sufficient condition for the existence of an L^1 -best approximation such that $u_h(x_i) = 1$ for all $i = 1, \dots, N-1$, which finishes the proof of part two of Theorem 4.1.

Next, we show that $h_N \leq \min_{i=1, \dots, N-1} h_i$ implies that (4.10) is satisfied. Let ϑ_i for $i = 1, \dots, N-1$ be defined as in Lemma 2. We first show that if ϑ_i is real and $\vartheta_i < 1 - 1/\sqrt{2}$ for some $k \geq 1$ and all $i \geq k$, then the following holds:

$$\frac{h_i}{2} (2(1 - \vartheta_i)^2 - 1) \leq \frac{h_{i+1}}{2} (2(1 - \vartheta_{i+1})^2 - 1) \quad \text{for all } i \geq k. \quad (4.19)$$

Indeed, by the definition of ϑ_i , we have

$$\frac{h_{i+1}}{2} (2(1 - \vartheta_{i+1})^2 - 1) = \frac{h_i}{2} (1 - 2\vartheta_i^2).$$

Hence, (4.19) is equivalent to

$$\frac{h_i}{2} (2(1 - \vartheta_i)^2 - 1) \leq \frac{h_i}{2} (1 - 2\vartheta_i^2) \Leftrightarrow \vartheta_i(\vartheta_i - 1) \leq 0,$$

which is true since $\vartheta_i \in (0, 1)$ for $i \geq k$. Next, note that if we apply (4.19) recursively, we obtain

$$\frac{h_k}{2} (2(1 - \vartheta_k)^2 - 1) \leq \frac{h_N}{2}. \quad (4.20)$$

In order to prove that (4.10) is satisfied, first note that (4.10) reduces to $h_{N-1} \geq h_N$ for $i = N - 1$ since $\vartheta_N = 0$. Now assume for the sake of contradiction that for some $j > 0$ and all $i = j, j + 1, \dots, N - 1$, $\vartheta_i < 1 - 1/\sqrt{2}$ and assume (4.10) holds for $i = j + 1, \dots, N - 1$ but not for $i = j$. In this case, we have

$$h_N \leq h_j < \left(2(1 - \vartheta_j)^2 - 1\right) h_{j+1} \stackrel{\text{by (4.20)}}{\leq} h_N,$$

which is a contradiction. $\square$

Remark 4.9. There is a more direct proof of the third part of Theorem 4.1 by defining $\psi(x) \equiv (-1)^j h_N / h_{N-j}$ on (x_{N-j-1}, x_{N-j}) for $j = 1, \dots, N - 1$ and showing that

$$\int_0^1 \psi \varphi_i dx = 0, \quad \text{for } i = 1, \dots, N - 1.$$

However, the proof given above shows that the condition $h_N \leq \min_{i=1, \dots, N-1} h_i$ is always stronger than (4.10).

4.2.4 A Three-Element Mesh in One Dimension

To illustrate the graded mesh condition in part two of Theorem 4.1 we consider Example 4.4. More precisely, we consider two three-element meshes on $(0, 1)$. For the first one we choose $h_1 = 0.1$ and $h_2 = h_3 = 0.45$, i.e., the mesh consists of the subintervals $(0, 0.1)$, $(0.1, 0.55)$ and $(0.55, 1)$. For the second one we choose $h_1 = 0.1$, $h_2 = 0.5$ and $h_3 = 0.4$, i.e., the mesh consisting of the subintervals $(0, 0.1)$, $(0.1, 0.6)$ and $(0.6, 1)$. We will check the condition (4.5) for both meshes. We will see that for the first mesh the condition is violated but it is satisfied for the second mesh. In the latter case, we already know that there exists an L^1 -best approximation without over- or undershoots. In the former case, it is a priori unknown whether or not such an L^1 -best approximation exists, since it is an open problem whether (4.5)

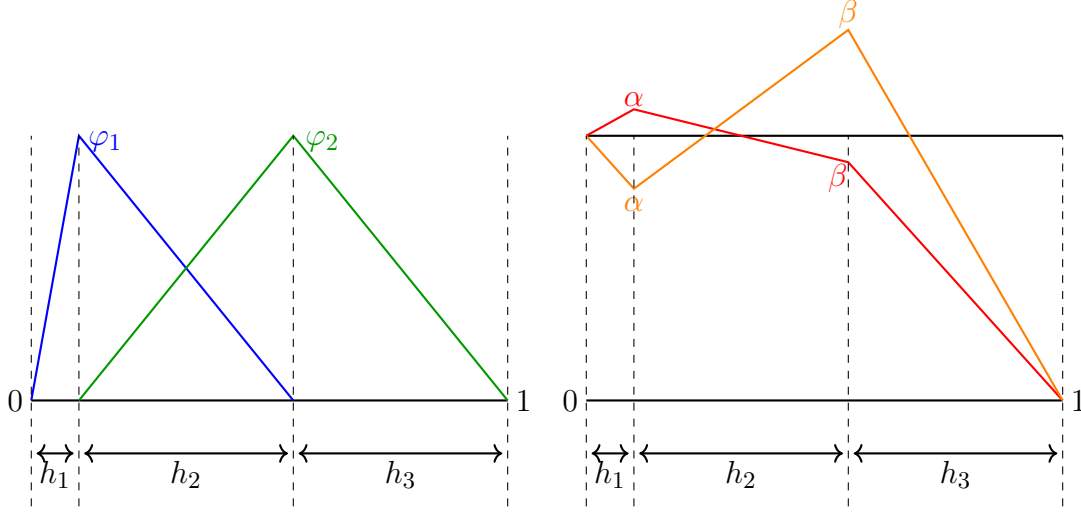


Figure 4.7 Example 4.4: Hat functions φ_0 and φ_1 (left). Possible approximations u_h (right).

is also a necessary condition. We show that in this case there is indeed no L^1 -best approximation without over- or undershoots and that the L^1 -best approximation is unique and satisfies $u_h(0.1) \approx 0.9931$ and $u_h(0.55) \approx 1.0247$.

Proof. First we consider the mesh with $h_2 = h_3 = 0.45$. Figure 4.7 shows the basis functions φ_1 and φ_2 in U_h associated with the two degrees of freedom of the best approximation problem on the left. The free parameters in the approximation problem are $u_h(h_1) = \alpha$ and $u_h(h_1 + h_2) = \beta$. Two potential approximants u_h are shown in Figure 4.7 on the right.

We obtain from (4.12) that $\vartheta_3 = \vartheta_2 = 0$, yielding the following sufficient conditions for the existence of an L^1 -best approximation without over- or undershoots: $h_2 \geq h_3$ and $h_1 \geq h_2$. The second condition is violated. Moreover, it is easy to show that there is indeed no L^1 -best approximation without over- or undershoots. Indeed, if $u_h(0.1) = 1$ and $u_h(0.55) = 1$, then $u - u_h > 0$ in $(0.55, 1)$ and zero everywhere else. In this case, the condition for u_h to be an L^1 -best approximation (4.8) with $v_h = \varphi_2$ for this choice of u_h can only be satisfied if $\psi \equiv -1$ in $(0.1, 0.55)$ since $h_2 = h_3$. However, if $\psi \equiv -1$ in $(0.1, 0.55)$, (4.8) with $v_h = \varphi_1$ is not satisfied for any ψ such that $|\psi| \leq 1$ in $(0, 0.1)$ since $h_1 < h_2$. Moreover, the same is true for any choice of h_i , $i = 1, 2, 3$, such that $h_2 = h_3$ and $h_1 < h_2$. Hence, $h_1 \geq h_2$ is indeed a necessary condition for the existence of an L^1 -best approximation without any over- or undershoots in this case.

The condition (4.8) with $v_h = \varphi_1$ implies that $u - u_h$ must change sign within $(0, 0.55)$. Hence, the only two possible options are

1. $\alpha < 1$ and $\beta > 1$,
2. $\alpha > 1$ and $\beta < 1$.

In the first case the approximation u_h intersects $u \equiv 1$ at the two points $h_1 + \lambda_1 h_2$ and $h_1 + h_2 + \lambda_2 h_3$, where $\lambda_i \in (0, 1)$, $i = 1, 2$. From (4.8) we obtain

$$\begin{aligned} \int_0^1 \operatorname{sgn}(u - u_h) \varphi_1 dx &= \frac{1}{2}(h_1 + h_2) - (1 - \lambda_1)^2 h_2, \\ \int_0^1 \operatorname{sgn}(u - u_h) \varphi_2 dx &= h_3(1 - \lambda_2)^2 + h_2 \lambda_1^2 - \frac{1}{2}(h_2 + h_3). \end{aligned}$$

Setting both integrals to 0, we can determine λ_1 and λ_2 ,

$$\begin{aligned} \lambda_1 &= 1 - \sqrt{\frac{1}{2} \frac{h_1 + h_2}{h_2}} = 1 - \frac{\sqrt{22}}{6}, \\ \lambda_2 &= 1 - \sqrt{\frac{1}{2} \frac{h_2 + h_3}{h_3} - \lambda_1^2 \frac{h_2}{h_3}} = 1 - \sqrt{\frac{6\sqrt{22} - 11}{18}}. \end{aligned}$$

This allows us to determine α and β :

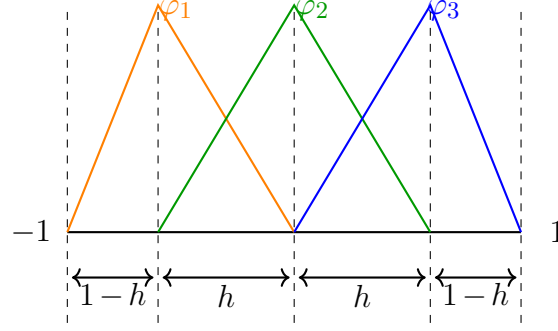
$$\beta = \frac{1}{1 - \lambda_2} = \sqrt{\frac{18}{6\sqrt{22} - 11}} \approx 1.0247, \quad \alpha = \frac{1 - \beta \lambda_1}{1 - \lambda_1} \approx 0.9931. \quad (4.21)$$

If $\alpha > 1$ and $\beta < 1$, u_h intersects $u \equiv 1$ at precisely one point $h_1 + \lambda_3 h_2$ with $\lambda_3 \in (0, 1)$. We can compute,

$$\begin{aligned} \int_0^1 \operatorname{sgn}(u - u_h) \varphi_1 dx &= -\frac{1}{2}(h_1 + h_2) + (1 - \lambda_3)^2 h_2, \\ \int_0^1 \operatorname{sgn}(u - u_h) \varphi_2 dx &= \frac{1}{2}(h_2 + h_3) - \lambda_3^2 h_2. \end{aligned}$$

Setting the second integral to zero yields $\lambda_3 = 1$. However, the first integral is non-zero for $\lambda_3 = 1$. Hence, it is impossible to satisfy (4.8) with $\alpha > 1$ and $\beta < 1$ on this mesh. The only L^1 -best approximation is therefore given by (4.21).

For the second mesh, we again have $\vartheta_3 = 0$ and (4.10) with $i = 2$ becomes $h_2 \geq h_3$, which holds for $h_2 = 0.5$ and $h_3 = 0.4$. For $i = 2$, we obtain from (4.12) $\vartheta_2^2 = 0.1$. Hence, $\vartheta_2 > 1 - \frac{1}{\sqrt{2}}$ and there is no condition on h_1 according to Theorem 4.1. Therefore, there exists an L^1 -best approximation without over- or undershoots. $\square$

Figure 4.8 Proof of Theorem 4.5: Basis for U_h

4.3 Over- and Undershoots at Jump Discontinuities

In this section we consider the L^q -best approximation of $u(x) = \text{sgn}(x)$ in $(-1, 1)$ as an example of a jump discontinuity in the interior of the domain and provide a proof of Theorem 4.5. We split this into three parts: in Section 4.3.1, we consider the first part of Theorem 4.5, i.e., we consider the case where the L^1 -best approximation does not exhibit Gibbs phenomena. In Section 4.3.2, we consider the case where the L^1 -best approximation does exhibit Gibbs phenomena (part two of Theorem 4.5); finally, in Section 4.3.3 we consider the L^q -best approximation for $1 < q < \infty$ and the limit as $q \rightarrow 1$ (parts three and four of Theorem 4.5).

For this example, $\Omega = (-1, 1)$ and $u(x) = \text{sgn}(x) \in L^1(\Omega)$. We seek an L^q -best approximation of this function by a continuous piecewise linear function on the mesh consisting of $(-1, -h)$, $(-h, 0)$, $(0, h)$ and $(h, 1)$. We fix the boundary conditions at -1 and 1 , i.e., $u_h(1) = u(1) = 1$ and $u_h(-1) = u(-1) = -1$. The finite dimensional approximation space U_h is given by the span of the hat functions φ_i , $i = 1, 2, 3$, depicted in Figure 4.8.

The condition for u_h to be an L^1 best-approximation in Corollary 4.8 can be written as follows: there exists $\psi : (-1, 1) \rightarrow [-1, 1]$ such that

$$\int_{-1}^1 \psi \varphi_i dx = 0, \quad \text{for } i = 1, 2, 3, \quad (4.22)$$

and $\psi(x) = \text{sgn}((u - u_h)(x))$ on $\{x \in (-1, 1) : u(x) \neq u_h(x)\}$.

4.3.1 L^1 -Best Approximation without Over- or Undershoots

In this section we provide a proof of the first part of Theorem 4.5. More precisely, if $h \leq 0.5$, a continuous piecewise linear function u_h on the mesh shown in Figure 4.8 such that $-u_h(-1) = u_h(1) = 1$ is an L^1 -best approximation of $u(x) = \text{sgn}(x)$ if and only if $u_h(0) = \beta$, with $\beta \in [-1, 1]$ arbitrary, and $-u_h(-h) = u_h(h) = 1$.

Proof. We first show that u_h must satisfy $-u_h(-h) = u_h(h) = 1$. For the sake of contradiction assume $u_h(-h) < -1$. In this case $u - u_h > 0$ on $(-1, -h)$ and $u - u_h > 0$ in $(-h, -h + \delta)$ with $h \geq \delta > 0$ and $u - u_h < 0$ in $(-h + \delta, 0)$. Thus, $\psi = \text{sgn}(u - u_h)$ almost everywhere and we obtain

$$\int_{-1}^1 \text{sgn}(u - u_h) \varphi_1 \, dx = \underbrace{\int_{-1}^{-h} \varphi_1 \, dx}_{=(1-h)/2} + \int_{-h}^0 \text{sgn}(u - u_h) \varphi_1 \, dx > \frac{1-h}{2} - \underbrace{\int_{-h}^0 \varphi_1 \, dx}_{=h/2} \geq 0,$$

since $h \leq 0.5 \Rightarrow 1 - h \geq h$. This is a contradiction since the condition (4.22) is violated. Note that we can use the same argument with opposite sign for $u_h(-h) > -1$. Hence, we have proven that $u_h(-h) = -1$. Due to the symmetry of the problem, it can easily be seen that the argument for $u_h(h) = 1$ is identical.

If $-u_h(-h) = u_h(h) = 1$, then $u_h(0) \notin [-1, 1]$ implies $\psi = \text{sgn}(u - u_h)$ and $\text{sgn}(u - u_h) = -\text{sgn}(u_h(0))$ on $(-h, h)$ and hence

$$\int_{-1}^1 \psi \varphi_2 \, dx = \int_{-h}^h \psi \varphi_2 \, dx = -\text{sgn}(u_h(0))h \neq 0.$$

Again, (4.22) is violated which implies $u_h(0) \in [-1, 1]$.

Next we establish, that $-u_h(-h) = u_h(h) = 1$ and $u_h(0) = \beta$ is indeed an L^1 -best approximation of $\text{sgn}(x)$ for any $\beta \in [-1, 1]$. To this end, we distinguish the following three cases:

$$(a) \, u_h(0) = 1, \quad (b) \, u_h(0) \in (-1, 1), \quad (c) \, u_h(0) = -1.$$

All three cases are shown in Figure 4.9a-c, where $u_h(0) = 0$ was chosen as an example for the second case. In the first case, $u - u_h = 0$ in $(-1, -h)$ and $u - u_h < 0$ in $(-h, 0)$. We compute

$$\int_{-1}^1 \psi \varphi_1 \, dx = \int_{-1}^{-h} \psi \varphi_1 \, dx - \underbrace{\int_{-h}^0 \varphi_1 \, dx}_{=h/2}.$$

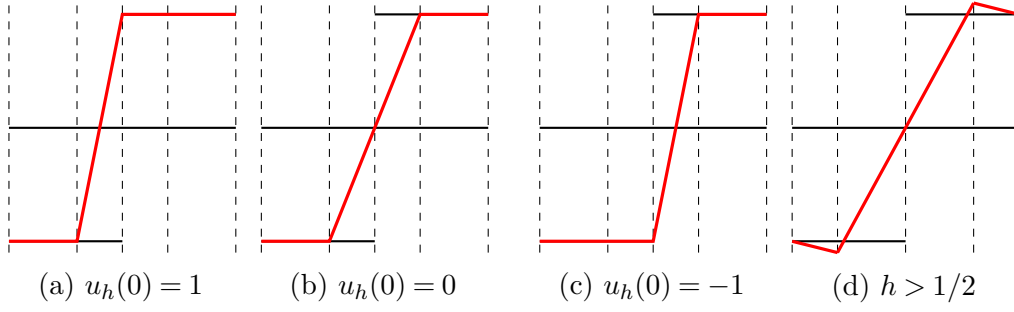


Figure 4.9 L^1 -best approximation of $\text{sgn}(x)$, cf., Theorem 4.5.

This integral is equal to zero for $\psi_0 \equiv h/(1-h)$ on $(-1, -h)$ which is a valid choice provided $h \leq 1-h$. Next, we compute

$$\int_{-1}^1 \psi \varphi_2 dx = - \underbrace{\int_{-h}^0 \varphi_2 dx}_{h/2} + \int_0^h \psi_0 \varphi_2 dx.$$

Note that $u - u_h = 0$ in $[0, h]$; it is easy to see that this integral becomes zero if and only if $\psi_0 \equiv 1$ in $[0, h]$. The remaining integral then becomes

$$\int_{-1}^1 \psi \varphi_3 dx = \underbrace{\int_0^h \varphi_3 dx}_{h/2} + \int_h^1 \psi_0 \varphi_3 dx.$$

This integral is zero for $\psi_0 \equiv -h/(1-h)$ in $[h, 1]$ which is again a valid choice if and only if $h \leq 1-h$. Therefore, we have that case (a) defines an L^1 -best approximation if $h \leq 1-h \Leftrightarrow h \leq 1/2$.

Due to the symmetry of the problem, it is easy to see that the third case, i.e., $u_h(0) = -1$ also defines an L^1 -best approximation if $h \leq 1-h$. This leaves the second case. As in the first case, $u - u_h = 0$ in $[-1, -h]$ and $u - u_h < 0$ in $[-h, 0]$. This implies that we have to require $h \leq 1-h$ in order to find ψ_0 such that the integral involving φ_1 becomes zero. Furthermore, note that – again due to the symmetry of the problem – the same applies to the integral involving φ_3 . Finally, since $u - u_h < 0$ in $[-h, 0]$ and $u - u_h > 0$ in $[0, h]$, we can compute

$$\int_{-1}^1 \psi \varphi_2 dx = - \int_{-h}^0 \varphi_2 dx + \int_0^h \varphi_2 dx = 0.$$

Thus, any choice of $u_h(0) \in [-1, 1]$ defines an L^1 -best approximation if $h \leq 1/2$. This completes the proof. $\square$

Remark 4.10. Note that we have shown that there is a whole family of L^1 -best approximations with no over- or undershoots for this particular example if $h \leq 1/2$. The situation is quite different if we instead consider a non-symmetric subdivision of the interval $(-1, 1)$ into $(-1, -h_1)$, $(-h_1, 0)$, $(0, h_2)$ and $(h_2, 1)$ with $h_1 \neq h_2$. The integral involving φ_2 then implies that the case $-1 < u_h(0) < 1$ does not yield an L^1 -best approximation; the case $u_h(0) = 1$ is an L^1 -best approximation if and only if $h_1 < h_2 \leq 1/2$, and the case $u_h(0) = -1$ is an L^1 -best approximation if and only if $h_2 < h_1 \leq 1/2$.

4.3.2 L^1 -Best Approximation with Over- and Undershoots

In this section we provide a proof of the second part of Theorem 4.5. More precisely, if $h > 0.5$, we show that a continuous piecewise linear function u_h on the mesh shown in Figure 4.8 such that $-u_h(-1) = u_h(1) = 1$ is an L^1 -best approximation of $u(x) = \text{sgn}(x)$ if and only if

$$u_h(-h) = \alpha := -\sqrt{2h} - \beta(\sqrt{2h} - 1), \quad (4.23a)$$

$$u_h(0) = \beta, \quad (4.23b)$$

$$u_h(h) = \gamma := \sqrt{2h} - \beta(\sqrt{2h} - 1), \quad (4.23c)$$

with $\beta \in [-1, 1]$ arbitrary.

Proof. We start by showing that there exists ψ as in Corollary 4.8, such that u_h as defined in (4.23) satisfies (4.8) with $v_h = \varphi_1$. In order to determine $\psi(x)$, we have to find the points where u_h intersects u . Note that u_h cannot intersect u in $(-1, -h)$ or $(h, 1)$ unless it is identical with u in all of $(-1, -h)$ or $(h, 1)$, respectively. Hence, if $\alpha \neq -1$, either $\psi \equiv 1$ or $\psi \equiv -1$ on $(-1, -h)$ and $u - u_h$ has to change sign within $(-h, 0)$ in order to obtain

$$\int_{-1}^1 \psi \varphi_1 dx = 0.$$

In other words, u and u_h intersect at a point $-h\vartheta$, $\vartheta \in (0, 1)$. In order to determine ϑ in the case that $\psi \equiv 1$ on $(-1, -h)$, we compute

$$\begin{aligned} \int_{-1}^1 \psi \varphi_1 dx &= \int_{-1}^{-h} \varphi_1 dx + \int_{-h}^{-\vartheta h} \varphi_1 dx - \int_{-\vartheta h}^0 \varphi_1 dx \\ &= \frac{1-h}{2} + \left(\frac{h}{2} - \frac{\vartheta^2 h}{2} \right) - \frac{\vartheta^2 h}{2} = \frac{1-2\vartheta^2 h}{2} = 0. \end{aligned} \quad (4.24)$$

Note that the above integral is zero if and only if $\vartheta = 1/\sqrt{2h}$. To consider the case $\psi \equiv -1$ on $(-1, -h)$, we simply multiply the right hand side of (4.24) by -1 and arrive at the same value for ϑ . Since two out of the three parameters ϑ , α and β uniquely determine the third, it is easy to see that, since ϑ is now uniquely determined, α and β must satisfy (4.23a).

So far we have considered $\alpha \neq -1$. If $\alpha = -1$, we claim that $\beta = -1$ as well, cf., (4.23a). For the sake of contradiction assume $\alpha = -1$ and $\beta > -1$. In this case $u - u_h = 0$ in $(-1, -h)$ and $u - u_h < 0$ in $(-h, 0)$. Since $|\psi| \leq 1$, we have

$$\int_{-1}^1 \psi \varphi_1 dx = \int_{-1}^{-h} \psi_0 \varphi_1 dx - \int_{-h}^0 \varphi_1 dx \leq \int_{-1}^{-h} \varphi_1 dx - \int_{-h}^0 \varphi_1 dx < 0,$$

since $h > 0.5$. This is a contradiction. The case $\beta < -1$ is analogous with opposite signs.

If $\beta = -1$, (4.24) implies that choosing

$$\psi_0 = \begin{cases} \pm 1 & \text{in } (-1, -\vartheta h), \\ \mp 1 & \text{in } (-\vartheta h, 0), \end{cases}$$

yields

$$\int_{-1}^0 \psi_0 \varphi_1 dx = 0.$$

In summary, we have so far shown that in order to satisfy (4.8) with $v_h = \varphi_1$, α and β must satisfy (4.23a). Conversely, we have also seen that for any α, β satisfying (4.23a), (4.8) is satisfied with $v_h = \varphi_1$.

Due to the symmetry of the problem, the proof that β and γ must satisfy (4.23c) and that for any such choice (4.8) is satisfied with $v_h = \varphi$ is completely analogous.

It remains to be shown that $|\beta| \leq 1$ and that (4.8) is also satisfied with $v_h = \varphi_2$. For the sake of contradiction, assume $\beta < -1$. We have already established that α , β and γ satisfy (4.23) and that u and u_h intersect at $-\vartheta h$ and ϑh . Thus, we obtain

$$\begin{aligned} \int_{-1}^1 \psi \varphi_2 dx &= - \int_{-h}^{-\vartheta h} \varphi_2 dx + \int_{-\vartheta h}^0 \varphi_2 dx + \int_0^{\vartheta h} \varphi_2 dx - \int_{\vartheta h}^h \varphi_2 dx \\ &= 2 \left(\int_0^{\vartheta h} \varphi_2 dx - \int_{\vartheta h}^h \varphi_2 dx \right) = h \left(1 - 2(1 - \vartheta)^2 \right) = 0 \\ &\Leftrightarrow (1 - \vartheta)^2 = \frac{1}{2} \Leftrightarrow \vartheta = 1 - \frac{1}{\sqrt{2}}. \end{aligned}$$

On the other hand, we require $\vartheta = 1/\sqrt{2h}$ for the other integrals to become zero. So,

$$\frac{1}{\sqrt{2h}} = 1 - \frac{1}{\sqrt{2}} \Leftrightarrow h = \left(\frac{1}{\sqrt{2}-1} \right)^2 > 1.$$

This is a contradiction, since $h \in (0, 1)$. For $\beta > 1$, the sign of $u - u_h$ on $(-h, h)$ is exactly opposite compared to the case $\beta < -1$; therefore, it is easy to see that $\beta > 1$ also leads to a contradiction.

If on the other hand $\beta \in [-1, 1]$, we obtain, using the symmetry of φ_2 ,

$$\int_{-1}^1 \psi \varphi_2 \, dx = \underbrace{\int_{-h}^{-\vartheta h} \varphi_2 \, dx - \int_{\vartheta h}^h \varphi_2 \, dx}_{=0} + \underbrace{\int_0^{\vartheta h} \varphi_2 \, dx - \int_{-\vartheta h}^0 \varphi_2 \, dx}_{=0} = 0.$$

Note that the above computation also applies to the cases $\beta = 1$ and $\beta = -1$ since it corresponds to a valid choice for ψ_0 in each case. $\square$

4.3.3 L^q -Best Approximation

In this section, we prove the third and fourth part of Theorem 4.5. More precisely, we show that a continuous piecewise linear function u_h on the mesh shown in Figure 4.8 such that $-u_h(-1) = u_h(1) = 1$ is an L^q -best approximation of $u(x) = \text{sgn}(x)$ for $1 < q < \infty$ if and only if $-u_h(-h) = u_h(h) = \alpha$ and $u_h = 0$, where α satisfies

$$0 = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q + 1)(\alpha-1)^q + h$$

and $\alpha > 1$. Furthermore, we show that in the limit $q \rightarrow 1$ the L^q -best approximation converges to the L^1 -best approximation as defined in (4.23) with $\beta = 0$, for any $h \in (0, 1)$, i.e., the corresponding L^1 -best approximation is anti-symmetric and satisfies

$$-u_h(-h) = u_h(h) = \begin{cases} 1 & \text{if } h \leq 0.5, \\ \sqrt{2h} & \text{if } h > 0.5. \end{cases}$$

Proof. We use the characterization of the L^q -best approximation in Corollary 4.7. Due to the uniqueness of the L^q -best approximation for $1 < q < \infty$ and the symmetry of the problem, we may assume that the L^q -best approximation is an odd function. This means that $u_h(0) = 0$ and $-u_h(-h) = u_h(h) = \alpha$ for some $\alpha \in \mathbb{R}$.

It is easy to see that

$$\int_{-1}^1 \operatorname{sgn}(u - u_h) |u - u_h|^{q-1} \varphi_2 \, dx = 0$$

for any choice of α and that

$$\int_{-1}^1 \operatorname{sgn}(u - u_h) |u - u_h|^{q-1} \varphi_3 \, dx = 0 \Leftrightarrow \int_{-1}^1 \operatorname{sgn}(u - u_h) |u - u_h|^{q-1} \varphi_1 \, dx = 0.$$

To determine for which α the latter two integrals become zero, note that this is the same situation as in the example presented in Sections 4.2.1 and 4.2.2, only mirrored. Therefore, we again obtain that α satisfies

$$0 = -(1-h)\alpha^2 q(\alpha-1)^{q-1} - h(\alpha q + 1)(\alpha-1)^q + h.$$

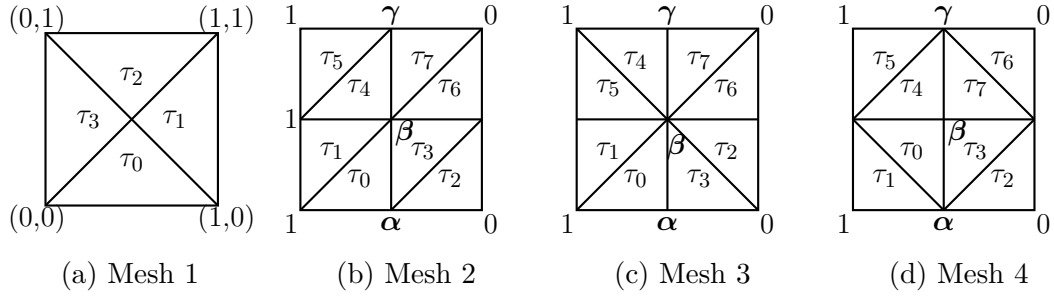
This completes the proof of the third part of Theorem 4.5.

Since u_h is an odd function, for any $1 < q < \infty$, the limit as $q \rightarrow 1$ must be an odd function as well and must therefore be zero at $x = 0$. Since the L^1 -best approximation is uniquely determined by the value it takes at zero according to the first two parts of Theorem 4.5, this completes the proof. Therefore, in the limit we obtain the solution in Figure 4.9b if $h \leq 1/2$. The corresponding L^1 -best approximation for $h > 1/2$ is shown in Figure 4.9d. $\square$

4.4 Best Approximation of a Boundary Discontinuity in Two Dimensions

In this section we consider the best approximation problem (4.3) with $d = 2$ and provide a proof of Theorem 4.6. Hence, we consider the function $u \equiv 1$ on $(0, 1)^2$. We consider the four meshes shown in Figure 4.10 and determine the best approximation of u by a continuous function u_h that is a linear polynomial on each of the triangles and takes the following values in the four corners: $u_h(0, 0) = u_h(0, 1) = 1$ and $u_h(1, 0) = u_h(1, 1) = 0$. For all meshes except the first one, we additionally fix the boundary conditions $u_h(0, 0.5) = 1$ and $u_h(1, 0.5) = 0$.

The free parameter of the best approximation problem for the first mesh is $\alpha = u_h(0.5, 0.5)$; there are three free parameters for each of the remaining meshes. For Meshes 2-4, we denote by v_1 the continuous piecewise linear function that is 1 at the node $(0.5, 0)$ and 0 at all other nodes; by v_2 the continuous piecewise linear function that is 1 at $(0.5, 0.5)$ and 0 at all other nodes; and by v_3 the continuous

Figure 4.10 Four different meshes on $(0,1)^2$, cf., Figure 4.4.

piecewise linear function that is 1 at $(0.5, 1)$ and zero at all other nodes. The coefficients defining the solution u_h are denoted as follows

$$u(0.5, 0) = \alpha, \quad u(0.5, 0.5) = \beta, \quad u(0.5, 1) = \gamma.$$

We split the proof of Theorem 4.6 into three parts: in Section 4.4.1, we prove the first part of the theorem, i.e., we consider Mesh 1 with $q = 1$; in Section 4.4.2 we continue with part two of the theorem and consider Mesh 1 with $1 < q < \infty$; Section 4.4.3 finally contains a proof of part three, four and five of the theorem, i.e., we show that the L^q -best approximation contains over- or undershoots on all four meshes if $q > 1$ and consider $q = 1$ for Meshes 2, 3 and 4.

4.4.1 L^1 -Best Approximation

In this section we prove the first part of Theorem 4.6. More precisely, we show that if we consider Mesh 1, the L^1 -best approximation is unique and $u_h((0.5, 0.5)) = \alpha$, where α satisfies

$$\alpha > 1 \quad \text{and} \quad 0 = 2\alpha^3 - 5\alpha + 2,$$

hence $\alpha \approx 1.3200$.

Proof. We again use the characterization of the L^1 -best approximation in Corollary 4.8. The space U_h is the span of the continuous function v_h that is a linear polynomial on each element, zero at the boundary of the domain and 1 at the centroid $(0.5, 0.5)$. We will use the reference triangle $\hat{\tau}$ as depicted on the left in Figure 4.11 for all computations. To this end, we define the affine transformations $\xi_i : \tau_i \rightarrow \hat{\tau}$, $i = 0, 1, 2, 3$ that are each composed of a scaling by 0.5, a rotation and a

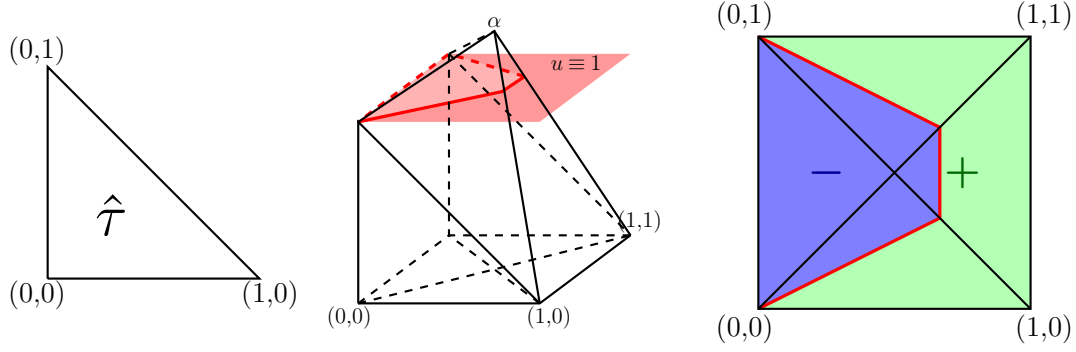


Figure 4.11 Proof of Theorem 4.6: Reference element $\hat{\tau}$ (left). The function u_h with $\alpha > 1$ and the intersection with u marked by red lines (centre). The mesh with the area where $(u - u_h) < 0$ coloured in blue and the area where $(u - u_h) > 0$ coloured in green (right).

translation. On $\hat{\tau}$ we define the basis functions $\hat{\varphi}_0$, $\hat{\varphi}_1$ and $\hat{\varphi}_2$ as

$$\hat{\varphi}_0 = 1 - x - y, \quad \hat{\varphi}_1 = x, \quad \hat{\varphi}_2 = y,$$

in the coordinates of the reference element $\hat{\tau}$. Note that $\xi_i(v_h|_{\tau_i}) = \hat{\varphi}_0$ for all $i = 0, 1, 2, 3$.

We consider the following two cases:

1. The set $\{(x, y) \in (0, 1)^2 : (u - u_h)(x, y) = 0\}$ has measure zero.
2. The set $\{(x, y) \in (0, 1)^2 : (u - u_h)(x, y) = 0\}$ has positive measure.

The second case is only possible if $\alpha = 1$. In this case we have $u_h = u$ in τ_3 and $u_h < u$ in τ_i , $i = 0, 1, 2$. Noting that the τ_i and $\hat{\tau}$ are similar and that the area of $\hat{\tau}$ is precisely twice the area of any τ_i , we obtain that

$$\begin{aligned} \sum_{i=0}^2 \int_{\tau_i} \psi v_h d\mathbf{x} &= \sum_{i=0}^2 \int_{\tau_i} \text{sgn}(u - u_h) v_h d\mathbf{x} = \frac{3}{2} \int_{\hat{\tau}} \hat{\varphi}_0 d\mathbf{x} \\ &= \frac{3}{2} \int_0^1 \int_0^{1-y} (1 - x - y) dx dy = \frac{1}{4}. \end{aligned}$$

On the other hand,

$$\int_{\tau_3} \psi v_h d\mathbf{x} \geq - \int_{\tau_3} v_h d\mathbf{x} = -\frac{1}{2} \int_{\hat{\tau}} \hat{\varphi}_0 d\mathbf{x} = -\frac{1}{12},$$

where $\psi(x, y)$ is arbitrary on τ_3 with $-1 \leq \psi(x, y) \leq 1$. Hence, for any choice of ψ , we have

$$\sum_{i=0}^3 \int_{\tau_i} \psi v_h d\mathbf{x} \geq \frac{1}{4} - \frac{1}{12} > 0.$$

This shows that $\alpha = 1$ does not yield a best approximation on this mesh.

We can furthermore rule out the case $\alpha < 1$; indeed, in this case $u - u_h > 0$ in the whole domain $(0, 1)^2$ and since $v_h > 0$ in $(0, 1)^2$, we have

$$\int_{(0,1)^2} \psi v_h d\mathbf{x} = \int_{(0,1)^2} \text{sgn}(u - u_h) v_h d\mathbf{x} > 0.$$

The only remaining case is $\alpha > 1$; in this setting we have $u - u_h < 0$ in τ_3 and thus

$$\int_{\tau_3} \psi v_h d\mathbf{x} = \int_{\tau_3} \text{sgn}(u - u_h) v_h d\mathbf{x} = -\frac{1}{2} \int_{\hat{\tau}} \hat{\varphi}_0 d\mathbf{x} = -\frac{1}{12}.$$

In the remaining three triangles, $u - u_h$ changes sign within the element. This is illustrated in Figure 4.11. Furthermore, $u - u_h = 0$ on a set of measure zero, hence $\psi = \text{sgn}(u - u_h)$ almost everywhere. In order to determine the sections of each element where $u - u_h$ is positive and negative, respectively, we will consider $\xi_i(u - u_h)$ for $i = 0, 1, 2$.

Due to the symmetry of the approximation problem, we can assume

$$\int_{\tau_0} \text{sgn}(u - u_h) v_h d\mathbf{x} = \int_{\tau_2} \text{sgn}(u - u_h) v_h d\mathbf{x}.$$

We compute

$$\xi_0(u - u_h) = 1 - \alpha \hat{\varphi}_0 - \hat{\varphi}_1 = -(\alpha - 1) + (\alpha - 1)x + \alpha y. \quad (4.25)$$

Thus,

$$\begin{aligned} \xi_0(u - u_h) &> 0 \text{ on } \left\{ (x, y) \in \hat{\tau} : y > \frac{\alpha - 1}{\alpha}(1 - x) \right\}, \\ \xi_0(u - u_h) &< 0 \text{ on } \left\{ (x, y) \in \hat{\tau} : y < \frac{\alpha - 1}{\alpha}(1 - x) \right\} \end{aligned} \quad (4.26)$$

and

$$\begin{aligned}
\int_{\tau_0} \operatorname{sgn}(u - u_h) v_h \, d\mathbf{x} &= \frac{1}{2} \int_{\hat{\tau}} \hat{\varphi}_0 \, d\mathbf{x} - \int_0^1 \int_0^{\frac{\alpha-1}{\alpha}(1-x)} \hat{\varphi}_0 \, dy \, dx \\
&= \frac{1}{12} - \int_0^1 \int_0^{\frac{\alpha-1}{\alpha}(1-x)} (1-x-y) \, dy \, dx \\
&= \frac{1}{12} - \frac{1}{6} \left(\frac{\alpha-1}{\alpha} \right) \left(\frac{\alpha+1}{\alpha} \right).
\end{aligned}$$

Similarly,

$$\xi_1(u - u_h) = 1 - \alpha \hat{\varphi}_0 = -(\alpha - 1) + \alpha x + \alpha y. \quad (4.27)$$

Thus,

$$\begin{aligned}
\xi_1(u - u_h) &> 0 \text{ on } \left\{ (x, y) \in \hat{\tau} : y > \frac{\alpha-1}{\alpha} - x \right\}, \\
\xi_1(u - u_h) &< 0 \text{ on } \left\{ (x, y) \in \hat{\tau} : y < \frac{\alpha-1}{\alpha} - x \right\}
\end{aligned} \quad (4.28)$$

and

$$\begin{aligned}
\int_{\tau_1} \operatorname{sgn}(u - u_h) v_h \, d\mathbf{x} &= \frac{1}{2} \int_{\hat{\tau}} \hat{\varphi}_0 \, d\mathbf{x} - \int_0^{\frac{\alpha-1}{\alpha}} \int_0^{\frac{\alpha-1}{\alpha}-x} \hat{\varphi}_0 \, dy \, dx \\
&= \frac{1}{12} - \int_0^{\frac{\alpha-1}{\alpha}} \int_0^{\frac{\alpha-1}{\alpha}-x} (1-x-y) \, dy \, dx \\
&= \frac{1}{12} - \frac{1}{6} \left(\frac{\alpha-1}{\alpha} \right)^2 \left(\frac{\alpha+2}{\alpha} \right).
\end{aligned}$$

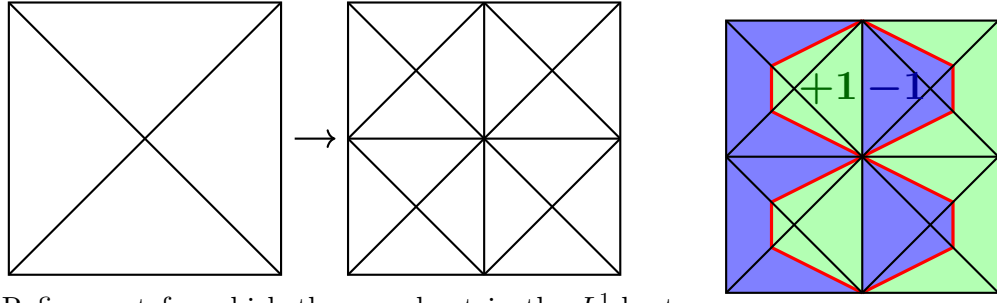
Finally,

$$\begin{aligned}
\int_{(0,1)^2} \operatorname{sgn}(u - u_h) v_h \, d\mathbf{x} &= \sum_{i=0}^3 \int_{\tau_i} \operatorname{sgn}(u - u_h) v_h \, d\mathbf{x} \\
&= \frac{1}{6} \left[1 - 2 \left(\frac{\alpha-1}{\alpha} \right) \left(\frac{\alpha+1}{\alpha} \right) - \left(\frac{\alpha-1}{\alpha} \right)^2 \left(\frac{\alpha+2}{\alpha} \right) \right] \\
&= -\frac{1}{6\alpha^3} (2\alpha^3 - 5\alpha + 2).
\end{aligned}$$

The polynomial $2\alpha^3 - 5\alpha + 2$ has three roots α_i , $i = 0, 1, 2$, where

$$\alpha_0 \approx -1.7623, \quad \alpha_1 \approx 0.43232, \quad \alpha_2 \approx 1.3200.$$

Only α_2 satisfies the condition $\alpha > 1$ and therefore $\alpha \approx 1.3200$ yields the only L^1 -best approximation. $\square$



(a) Refinement for which the overshoot in the L^1 -best approximation remains constant. (b) Possible choice for $\text{sgn}(0)$

Figure 4.12 Remark 4.11: Uniform refinement of the mesh preserving the structure.

Remark 4.11 (Uniform refinement). If the mesh is refined uniformly, keeping the same structure as shown in Figure 4.12a, it is easy to see that an L^1 -best approximation is given by $u(x_i, y_j) = \alpha$, with α as specified in Section 4.4.1, if the node (x_i, y_j) is connected with the boundary $x = 1$, and $u(x_i, y_j) = 1$ at the remaining interior nodes. Indeed, in this case we can choose $\psi = \psi_0$ on the set $\{x : u(x) - u_h(x) = 0\}$ as shown in Figure 4.12b. This shows that the overshoot in the L^1 -best approximation remains constant under this type of mesh refinement.

4.4.2 L^q -Best Approximation

In this section we prove the second part of Theorem 4.6. More precisely, we show that the unique solution to (4.3) on Mesh 1 for $1 < q < \infty$ is defined by $u_h((0.5, 0.5)) = \alpha$, where α satisfies

$$\alpha > 1 \quad \text{and} \quad 0 = (\alpha - 1)^{q-1} \left[4\alpha^3 q + 4(1 - q)\alpha^2 + (q - 6)\alpha + 2 \right] - \alpha(q + 4) + 2.$$

Proof. Corollary 4.7 implies that we seek α such that

$$\int_{(0,1)^2} \text{sgn}(u - u_h) |u - u_h|^{q-1} v_h \, d\mathbf{x} = 0. \quad (4.29)$$

Again, we have to split the integral on each element into the parts where $u - u_h > 0$, $u - u_h < 0$ and $u - u_h = 0$. We consider the three cases

$$(a) \ \alpha < 1, \quad (b) \ \alpha = 1, \quad (c) \ \alpha > 1.$$

If $\alpha < 1$, we have $u - u_h > 0$ everywhere in $(0, 1)^2$ and thus both $\text{sgn}(u - u_h)|u - u_h|^{q-1} > 0$ and $v_h > 0$ in $(0, 1)^2$. Therefore, we have

$$\int_{(0,1)^2} \text{sgn}(u - u_h)|u - u_h|^{q-1}v_h \, d\mathbf{x} > 0.$$

If $\alpha = 1$, we have $u - u_h = 0$ in τ_3 and $u - u_h > 0$ in τ_i , $i = 0, 1, 2$. Thus,

$$\int_{(0,1)^2} \text{sgn}(u - u_h)|u - u_h|^{q-1}v_h \, d\mathbf{x} = \sum_{i=0}^2 \int_{\tau_i} \text{sgn}(u - u_h)|u - u_h|^{q-1}v_h \, d\mathbf{x} > 0.$$

We can therefore again assume $\alpha > 1$; in this case $u - u_h < 0$ in τ_3 and we compute

$$\begin{aligned} \int_{\tau_3} \text{sgn}(u - u_h)|u - u_h|^{q-1}v_h \, d\mathbf{x} &= -\frac{1}{2} \int_{\hat{\tau}} (\xi_3(u_h - u))^{q-1} \hat{\varphi}_0 \, d\mathbf{x} \\ &= -\frac{1}{2} \int_{\hat{\tau}} (\alpha - 1)^{q-1} \hat{\varphi}_0^q \, d\mathbf{x} = -\frac{(\alpha - 1)^{q-1}}{2(q+1)(q+2)}. \end{aligned}$$

Next, consider τ_0 and τ_2 ; using the symmetry of the problem and equations (4.25) and (4.26), we obtain

$$\begin{aligned} &\int_{\tau_2} \text{sgn}(u - u_h)v_h \, d\mathbf{x} + \int_{\tau_0} \text{sgn}(u - u_h)v_h \, d\mathbf{x} \\ &= 2 \int_{\tau_0} \text{sgn}(u - u_h)v_h \, d\mathbf{x} \\ &= \int_{\hat{\tau}} \text{sgn}(\xi_0(u - u_h))|\xi_0(u - u_h)|^{q-1} \hat{\varphi}_0 \, d\mathbf{x} \\ &= \int_0^1 \int_{\frac{\alpha-1}{\alpha}(1-x)}^{1-x} (-(\alpha-1) + (\alpha-1)x + \alpha y)^{q-1} (1-x-y) \, dy \, dx \\ &\quad - \int_0^1 \int_0^{\frac{\alpha-1}{\alpha}(1-x)} ((\alpha-1) - (\alpha-1)x - \alpha y)^{q-1} (1-x-y) \, dy \, dx \\ &= \frac{1}{\alpha^2 q(q+1)(q+2)} - \frac{(\alpha-1)^q (\alpha q + 1)}{\alpha^2 q(q+1)(q+2)}. \end{aligned}$$

Finally, we consider τ_1 ; using equations (4.27) and (4.28), we obtain

$$\begin{aligned}
& \int_{\tau_1} \operatorname{sgn}(u - u_h) v_h \, d\mathbf{x} \\
&= \frac{1}{2} \int_{\hat{\tau}} \operatorname{sgn}(\xi_1(u - u_h)) |\xi_1(u - u_h)|^{q-1} \hat{\varphi}_0 \, d\mathbf{x} \\
&= \frac{1}{2} \int_0^{\frac{\alpha-1}{\alpha}} \int_{\frac{\alpha-1}{\alpha}-x}^{1-x} (-(\alpha-1) + \alpha x + \alpha y)^{q-1} (1-x-y) \, dy \, dx \\
&\quad + \frac{1}{2} \int_{\frac{\alpha-1}{\alpha}}^1 \int_0^{1-x} (-(\alpha-1) + \alpha x + \alpha y)^{q-1} (1-x-y) \, dy \, dx \\
&\quad - \frac{1}{2} \int_0^{\frac{\alpha-1}{\alpha}} \int_0^{\frac{\alpha-1}{\alpha}-x} ((\alpha-1) - \alpha x - \alpha y)^{q-1} (1-x-y) \, dy \, dx \\
&= \frac{\alpha-1}{2\alpha^3 q(q+1)} + \frac{1}{2\alpha^3(q+1)(q+2)} - \frac{(\alpha-1)^q (\alpha^2 q + (2-q)\alpha - 2)}{2\alpha^3 q(q+1)(q+2)}.
\end{aligned}$$

Combining all integrals, we obtain

$$\begin{aligned}
\int_{(0,1)^2} \operatorname{sgn}(u - u_h) v_h \, d\mathbf{x} &= \sum_{i=0}^3 \int_{\tau_i} \operatorname{sgn}(u - u_h) v_h \, d\mathbf{x} \\
&= - \frac{(\alpha-1)^{q-1} [4\alpha^3 q + 4(1-q)\alpha^2 + (q-6)\alpha + 2] - \alpha(q+4) + 2}{2\alpha^3 q(q+1)(q+2)}.
\end{aligned}$$

Hence, the L^q best approximation can be determined by finding $\alpha > 1$ satisfying

$$0 = (\alpha-1)^{q-1} [4\alpha^3 q + 4(1-q)\alpha^2 + (q-6)\alpha + 2] - \alpha(q+4) + 2.$$

□

4.4.3 L^q -Best Approximation on Meshes 2, 3 and 4

In this section we prove parts three, four and five of Theorem 4.6. More precisely, we show the following:

- If $q > 1$, the L^q -best approximation to (4.3) contains over- or undershoots on all four meshes.
- If $q = 1$, there exists a solution to (4.3) on Mesh 2 such that $u_h((0.5, 1)) = u_h((0.5, 0.5)) = 1$ and $u_h((0.5, 0)) = \alpha$, where α satisfies $\alpha > 1$ and $0 = -3\alpha^3 + 8\alpha - 4$. Furthermore, $u_h((0.5, 1)) = u_h((0.5, 0.5)) = u_h((0.5, 0)) = 1$ does not define an L^1 -best approximation.

- If $q = 1$, there exists a solution to (4.3) on Mesh 3 and Mesh 4 such that $u_h((0.5, 1)) = u_h((0.5, 0.5)) = u_h((0.5, 0)) = 1$.

Proof. To see that the first point (part three of Theorem 4.6) is true, note that if there are no over- or undershoots, i.e., $u_h((0.5, 1)) = u_h((0.5, 0.5)) = u_h((0.5, 0)) = 1$, we have $u - u_h = 0$ in $(0, 0.5) \times (0, 1)$ and $u - u_h > 0$ in $(0.5, 1) \times (0, 1)$ and hence

$$\int_{(0,1)^2} \text{sgn}(u - u_h) |u - u_h|^{q-1} v_i \, d\mathbf{x} > 0 \quad \text{for all } i = 1, 2, 3 \text{ and } 1 < q < \infty,$$

which contradicts Corollary 4.7.

Furthermore, it is easy to see that for the second and third mesh, $\alpha = \beta = \gamma = 1$ is an L^1 -best approximation, again using the characterization in Corollary 4.8. Indeed, we have $\psi = \text{sgn}(u - u_h)$ whenever $u(\mathbf{x}) \neq u_h(\mathbf{x})$ and we can choose $\psi = \psi_0 \equiv 1$ otherwise. In this case, for each of the three nodes $(0.5, 0)$, $(0.5, 0.5)$ and $(0.5, 1)$, we have that $\psi = -1$ on exactly half of the connected elements and $\psi = 1$ on the remaining connected elements. Note that

$$\int_{\tau_i} v_j \, d\mathbf{x} = \frac{1}{24} \quad \forall i, j,$$

and thus the terms with $\psi = -1$ and $\psi = 1$ cancel each other. This proves the third point (part five of Theorem 4.6).

The second point (part four of Theorem 4.6) is more interesting. The above argument could now only be applied to v_2 which is 1 at the node $(0.5, 0.5)$. First consider the node $(0.5, 1)$; the connected elements are τ_5 , τ_4 and τ_7 , cf., Figure 4.10b. If $\beta = \gamma = 1$, we have that $u - u_h = 0$ in τ_4 and τ_5 and $u - u_h > 0$ in τ_7 . If we choose $\psi \equiv -1$ on τ_4 and $\psi \equiv 0$ on τ_5 , we obtain

$$\int_{(0,1)^2} \psi v_3 \, d\mathbf{x} = \int_{\tau_7} v_3 \, d\mathbf{x} - \int_{\tau_4} v_3 \, d\mathbf{x} = \frac{1}{24} - \frac{1}{24} = 0. \quad (4.30)$$

Note that this is independent of α . Considering v_1 shows that there is no L^1 -best approximation with $\alpha = \beta = 1$. Indeed if $\alpha = \beta = 1$, $u - u_h > 0$ in τ_2 and τ_3 , and $u - u_h = 0$ in τ_0 . For any choice of ψ in τ_0 such that $|\psi| \leq 1$, we have

$$\begin{aligned} \int_{(0,1)^2} \psi v_1 \, d\mathbf{x} &\geq - \int_{\tau_0} v_1 \, d\mathbf{x} + \int_{\tau_2} v_1 \, d\mathbf{x} + \int_{\tau_3} v_1 \, d\mathbf{x} \\ &= -\frac{1}{24} + \frac{1}{24} + \frac{1}{24} = \frac{1}{24} > 0. \end{aligned}$$

We will now show that there is an L^1 -best approximation with $\beta = \gamma = 1$ and $\alpha > 1$. We have already established that (4.30) is independent of α ; this leaves

the integrals involving v_1 and v_2 . We have $u - u_h > 0$ in τ_6 and τ_7 , $u - u_h < 0$ in τ_0 , $u - u_h = 0$ in τ_1 and τ_4 , and $u - u_h$ changes sign within τ_2 and τ_3 . Therefore, $\psi = \text{sgn}(u - u_h)$ in τ_i , $i = 0, 2, 3, 6, 7$. Moreover, we have already fixed $\psi \equiv -1$ in τ_4 . In τ_1 , we now define ψ as follows

$$\psi \equiv -\frac{\int_{\tau_3} \text{sgn}((u - u_h)(\tilde{\mathbf{x}}))v_2(\tilde{\mathbf{x}})d\tilde{\mathbf{x}}}{\int_{\tau_1} v_2(\tilde{\mathbf{x}})d\tilde{\mathbf{x}}}.$$

Since $\int_{\tau_3} v_2 d\mathbf{x} = \int_{\tau_1} v_2 d\mathbf{x}$, ψ clearly satisfies $|\psi| \leq 1$ in τ_1 and the above choice is therefore a valid choice. On τ_1 , we then compute with this choice for ψ

$$\int_{\tau_1} \psi v_2 d\mathbf{x} = -\int_{\tau_1} \frac{\int_{\tau_3} \text{sgn}((u - u_h)(\tilde{\mathbf{x}}))v_2(\tilde{\mathbf{x}})d\tilde{\mathbf{x}}}{\int_{\tau_1} v_2(\tilde{\mathbf{x}})d\tilde{\mathbf{x}}} v_2(\mathbf{x}) d\mathbf{x} = -\int_{\tau_3} \text{sgn}(u - u_h)v_2 d\mathbf{x}.$$

Hence,

$$\begin{aligned} \int_{(0,1)^2} \psi v_2 d\mathbf{x} &= \sum_{i=6,7} \int_{\tau_i} v_2 d\mathbf{x} - \sum_{i=0,4} \int_{\tau_i} v_2 d\mathbf{x} \\ &\quad + \int_{\tau_3} \text{sgn}(u - u_h)v_2 d\mathbf{x} - \int_{\tau_3} \text{sgn}(u - u_h)v_2 d\mathbf{x} = 0; \end{aligned}$$

this is again independent of α . Finally, we consider v_1 to determine α . Let again ξ_i be the linear transformation that maps τ_i onto $\hat{\tau}$; we have $u - u_h > 0$ on τ_0 . Furthermore,

$$\xi_2(u - u_h) = 1 - \alpha y, \quad \xi_3(u - u_h) = (1 - \alpha)x + y.$$

Thus,

$$\begin{aligned} \xi_2(u - u_h) &> 0 \text{ on } \left\{ (x, y) \in \hat{\tau} : y < \frac{1}{\alpha} \right\}, \\ \xi_2(u - u_h) &< 0 \text{ on } \left\{ (x, y) \in \hat{\tau} : y > \frac{1}{\alpha} \right\} \end{aligned}$$

and

$$\begin{aligned} \xi_3(u - u_h) &> 0 \text{ on } \{ (x, y) \in \hat{\tau} : y > (\alpha - 1)x \}, \\ \xi_3(u - u_h) &< 0 \text{ on } \{ (x, y) \in \hat{\tau} : y < (\alpha - 1)x \}. \end{aligned}$$

Therefore, noting that the area of $\hat{\tau}$ is four times the area of τ_i for any $i = 0, \dots, 7$, we compute

$$\begin{aligned} \int_{\tau_2} \psi v_1 \, d\mathbf{x} &= \int_{\tau_2} \operatorname{sgn}(u - u_h) v_1 \, d\mathbf{x} = \frac{1}{4} \int_{\hat{\tau}} y \, dx \, dy - \frac{1}{2} \int_{\frac{1}{\alpha}}^1 \int_0^{1-y} y \, dx \, dy \\ &= \frac{-\alpha^3 + 6\alpha - 4}{24\alpha^3}, \end{aligned}$$

$$\begin{aligned} \int_{\tau_3} \psi v_1 \, d\mathbf{x} &= \int_{\tau_3} \operatorname{sgn}(u - u_h) v_1 \, d\mathbf{x} = -\frac{1}{4} \int_{\hat{\tau}} x \, dx \, dy + \frac{1}{2} \int_0^{\frac{1}{\alpha}} \int_{(1-\alpha)x}^{1-x} x \, dx \, dy \\ &= \frac{2 - \alpha^2}{24\alpha^2} \end{aligned}$$

and

$$\int_{\tau_0} \psi v_1 \, d\mathbf{x} = \int_{\tau_0} \operatorname{sgn}(u - u_h) v_1 \, d\mathbf{x} = -\frac{1}{24}.$$

Putting all three integrals together yields

$$\int_{(0,1)^2} \psi v_1 \, d\mathbf{x} = \frac{1}{24\alpha^3} (-3\alpha^3 + 8\alpha - 4).$$

Hence, $\alpha > 1$ has to satisfy the equation

$$0 = -3\alpha^3 + 8\alpha - 4.$$

Three roots of the above polynomial are $\alpha_0 \approx -1.8414$, $\alpha_1 \approx 0.56913$ and $\alpha_2 \approx 1.2723$. Only the third root satisfies $\alpha > 1$ and therefore defines an L^1 -best approximation.

□

4.5 Numerical Examples

In this section we consider selected examples of meshes for which we have determined the solution of the best approximation problem (4.3) numerically by interpreting the condition (4.7) as a variational problem that can be implemented using standard finite element techniques. Here, we have used FEniCS [5] for the implementation. In Section 4.5.1 we illustrate that the overshoot in the L^q -best approximation does not vanish if the mesh is refined and that these observations even apply if u is a more general smooth function.

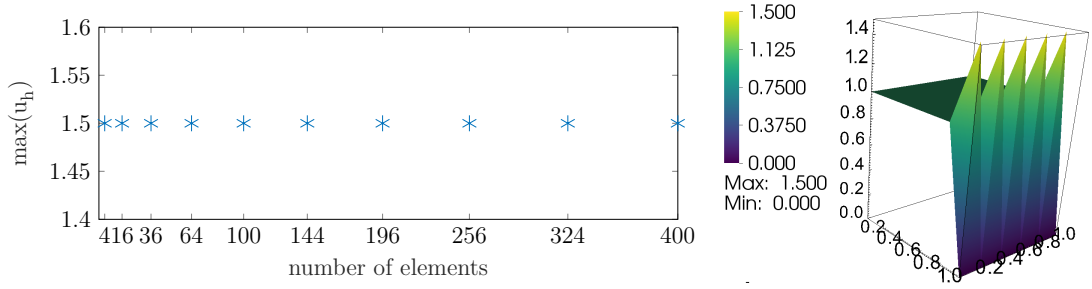


Figure 4.13 Left: $\max(u_h)$ for $q = 2$ and several refinements as shown in Figure 4.12a. Right: u_h with 100 elements.

In Section 4.5.3 we illustrate that the L^q -approximation on the three meshes considered in Section 4.4.3 (second half of Theorem 4.6) converges to the L^1 -best approximation characterized in the theorem. Furthermore, we show how the understanding of these special cases can be applied to predict the behaviour of the L^q -best approximation on a more general mesh.

4.5.1 Refinement of the Mesh

4.5.1.1 Gibbs Phenomenon on Meshes in Two Dimensions

We start with Mesh 1 depicted in Figure 4.10 and the refinement shown in Figure 4.12a that preserves the structure of the mesh. We have already shown in Remark 4.11 that for $q = 1$ there exists an L^1 -best approximation such that the overshoot remains constant as we refine the mesh. Indeed, Figure 4.13 shows the maximum value of u_h for this example with $q = 2$ and for several refinements of the mesh, as well as the approximation u_h for a mesh with this structure consisting of 100 elements. We can clearly see that the maximum value remains constant under this type of refinement which suggests that the maximum overshoot also remains constant for $q \neq 1$, as well as in the limit $q \rightarrow 1$.

4.5.1.2 Gibbs Phenomenon on Meshes in One Dimension

Next, we consider a one-dimensional example such that u is not piecewise linear and compute the L^q -best approximation numerically. Let $u(x) = 1 + 0.1 \sin(2\pi x)$ on $(0, 1)$ and consider the L^q -best approximation u_h with $u_h(0) = 1$ and $u_h(1) = 0$ on four different grids: two uniform grids with 5 and 100 elements, respectively, and two meshes where all elements are the same size except the last one which is twice the size of the others. Again we consider a mesh with 5 elements and one with 100 elements. Note that the latter two meshes violate the conditions

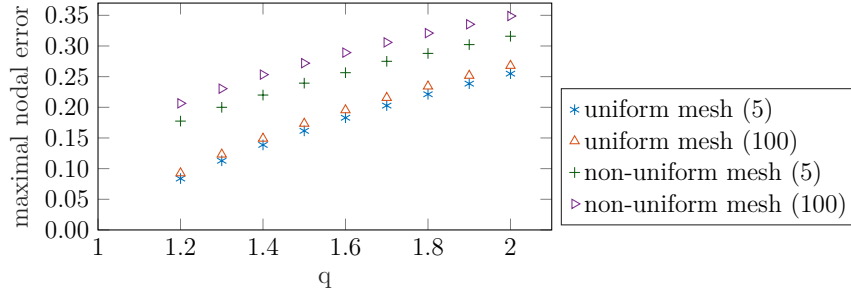


Figure 4.14 Maximal nodal error in the L^q -best approximation of $u(x) = 1 + 0.1\sin(2\pi x)$ with $u_h(1) = 0$ for different values of q and four different meshes. Two of the meshes are uniform meshes consisting of 5 and 100 elements, respectively. The remaining two meshes satisfy the condition in Remark 4.3 and again consist of 5 and 100 elements, respectively.

in parts two and three of Theorem 4.1, but satisfy the condition in Remark 4.3. We therefore expect the overshoot to vanish as $q \rightarrow 1$ in the first two cases and to decrease but still be present in the last two. Remark 4.3 and the observations for the previous example suggests that for $u \equiv 1$, we could expect the overshoot to be the same both when 5 and 100 elements are employed on both the uniform and the non-uniform meshes.

Figure 4.14 shows the maximum error at the nodes in all four cases for several values of q . We observe that the overshoot indeed decreases as $q \rightarrow 1$. Furthermore, we see that the overshoot is very similar for the coarse and fine meshes in both cases which confirms that the overshoot does not disappear under mesh refinement. However, the overshoot is not identical for 5 and for 100 elements in both cases which can be attributed to the fact that u is not constant. Furthermore, note that the overshoot for the non-uniform mesh is consistently larger than for the uniform mesh, which suggests that it does not disappear entirely as $q \rightarrow 1$. Note that on the non-uniform mesh when $u \equiv 1$ and $q = 1$, the overshoot would be $\frac{2}{3}\sqrt{3} - 1 \approx 0.15$; see Remark 4.3.

4.5.2 (Vanishing) Overshoot in One Dimension

To illustrate the graded mesh condition in part two of Theorem 4.1, we consider two three-element meshes on $(0,1)$. For the first one we choose $h_1 = 0.1$ and $h_2 = h_3 = 0.45$, i.e., the mesh consists of the subintervals $(0,0.1)$, $(0.1,0.55)$ and $(0.55,1)$, cf., Example 4.4. For the second one we choose $h_1 = 0.1$, $h_2 = 0.5$ and $h_3 = 0.4$, i.e., the mesh consisting of the subintervals $(0,0.1)$, $(0.1,0.6)$ and $(0.6,1)$, cf., Example 4.4. We have seen in Section 4.2.4 that for the first mesh condition

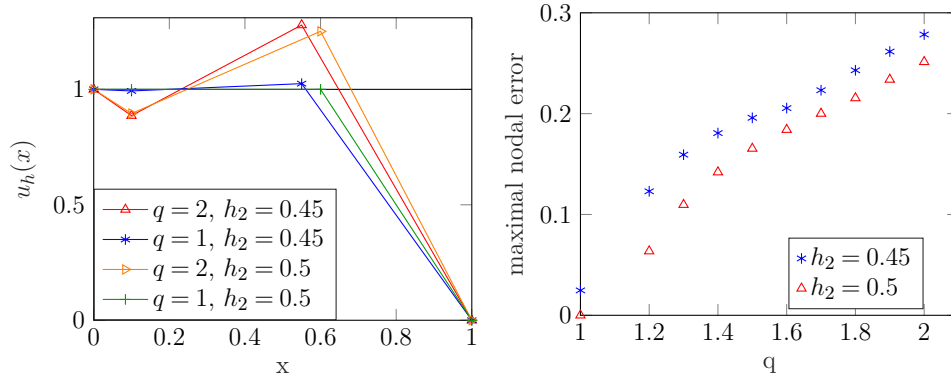


Figure 4.15 Example 4.4: L^q -best approximations on two three-element meshes on $(0, 1)$ with $h_1 = 0.1$ and two different choices for h_2 . Left: best approximation with $q = 2$ and $q = 1$. Right: maximal nodal error for several values of q .

(4.5) is violated but it is satisfied for the second mesh. In the latter case, there exists an L^1 -best approximation without over- or undershoots. In the former case, we have shown in Section 4.2.4 that there is no L^1 -best approximation without over- or undershoots. Moreover, we have shown that the L^1 -best approximation is unique and satisfies $u_h(0.1) \approx 0.9931$ and $u_h(0.55) \approx 1.0247$.

Figure 4.15 shows the L^q -best approximation on both meshes for $q = 2$ and $q = 1$ on the left and the maximal nodal error on both meshes for several values of q on the right. The approximations for $q > 1$ were again obtained using the implementation of the best approximation problem in FEniCS. We can clearly see, that the maximal overshoot is always larger on the first mesh. In both cases it decreases as $q \rightarrow 1$, but the overshoot only vanishes completely on the second mesh. However, even on the first mesh the maximal overshoot is very small for $q = 1$. Note that, if h_2 and h_1 as chosen for the first mesh were swapped, the maximal overshoot for $q = 1$ would be $u_h(0.55) - 1 = 0.2792$ according to Remark 4.3 and thus significantly larger than the overshoot we can observe. This shows that the effect of an element being too small and causing the L^1 -best approximation to contain over- and undershoots is much weaker away from the discontinuity than near the discontinuity.

4.5.3 (Vanishing) Overshoot in Two Dimensions

4.5.3.1 Overshoot on Meshes 2, 3 and 4 from Section 4.4.3

Figure 4.16 shows the best approximations for $q = 2$ and $q = 1.1$ for three of the meshes we have considered in Section 4.4.3. Even just a comparison of these two cases for each of the meshes illustrates clearly how the overshoot gradually vanishes

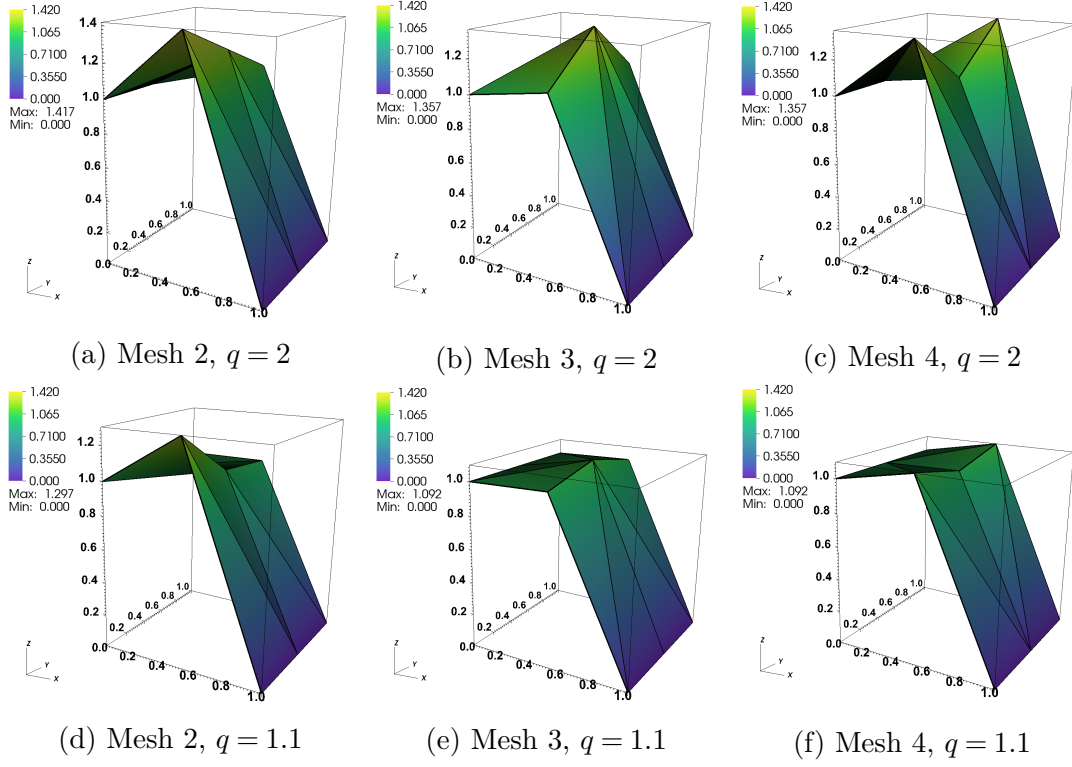


Figure 4.16 L^q -best approximation of a boundary discontinuity in two dimensions on Meshes 2,3 and 4, cf., Figure 4.10/4.4.

on Mesh 3 and Mesh 4. On Mesh 2, the overshoot vanishes away from the boundary $y = 0$; this is consistent with the L^1 -best approximation described above that only exhibits an overshoot at the node $(0.5, 0)$ and no overshoot at all other nodes.

Figure 4.17 shows the maximum overshoot for all three meshes for different values of q . The overshoot for $q = 1$ is taken from the theoretically determined L^1 -best approximations discussed in that section. All remaining values have been determined numerically with an implementation in FEniCS [5]. The plot shows that for the third and fourth mesh, the overshoot indeed disappears as $q \rightarrow 1$, whereas for the second mesh it decreases but does not vanish.

4.5.3.2 Overshoot on Unstructured Meshes

As a final example, we consider the unstructured mesh shown on the left in Figure 4.18. From the computations for the previous examples, we deduce that the L^1 -best approximation exhibits no overshoot if for every interior node (x_i, y_i) that is connected to the boundary $x = 1$ through one edge, the area of all triangles whose boundaries contain the node (x_i, y_i) and at least one node on the boundary $x = 1$ is smaller than the area of all remaining triangles whose boundaries contain the

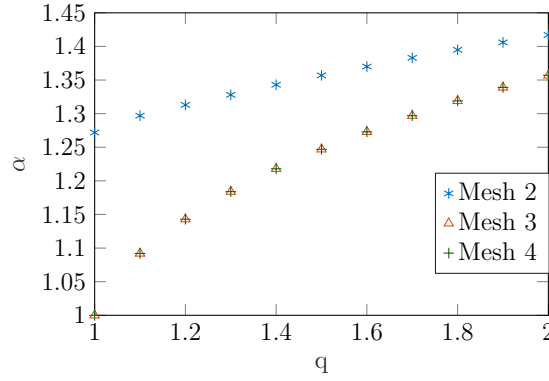


Figure 4.17 Values for α for different ranges of q on Meshes 2,3 and 4, cf., Figure 4.10/4.4.

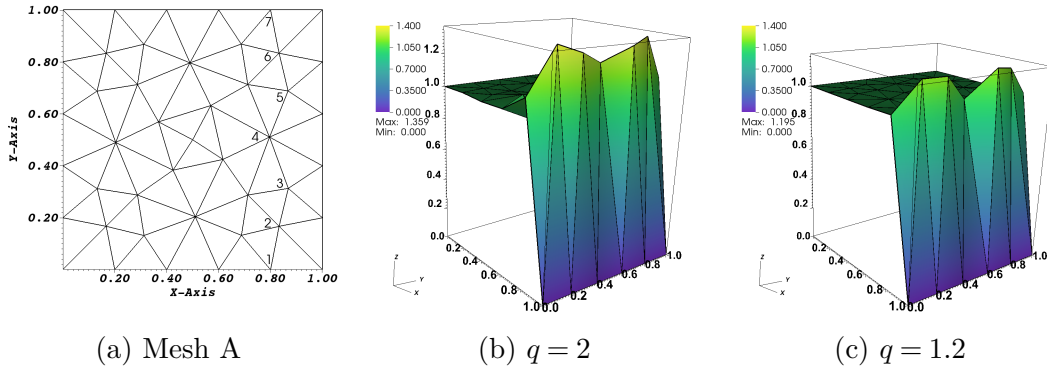


Figure 4.18 L^q -best approximation on an unstructured mesh.

node (x_i, y_i) . Furthermore, the numerics for Mesh 2 (cf., Figure 4.10) shown in Figure 4.16 suggest that the overshoot disappears for $q \rightarrow 1$ away from the nodes violating this condition on the area of connected elements.

The interior nodes connected to the boundary are labelled 1,2,...7 in Figure 4.18. It is easy to see that for the nodes 1,4 and 7 the total area of all triangles touching both the node and the boundary $x = 1$ is smaller than the total area of all remaining triangles touching the node, whereas this condition is violated for the nodes 2,3,5 and 6. Thus, we expect that the overshoot vanishes at the nodes 1,4 and 7 as $q \rightarrow 1$, while at the nodes 2,3,5 and 6 it reduces but does not disappear entirely. Figure 4.18 shows the L^q -best approximation on the unstructured mesh for $q = 2$ and $q = 1.2$ in the centre and on the right, respectively. Here, we clearly observe that the approximation for $q = 2$ exhibits overshoots at all nodes connected to the boundary $x = 1$ with larger overshoots at the nodes 2,3,5 and 6. At these nodes the overshoot is reduced but still clearly visible for $q = 1.2$. On the other hand at the nodes 1,4 and 7 the overshoot has nearly vanished for $q = 1.2$.

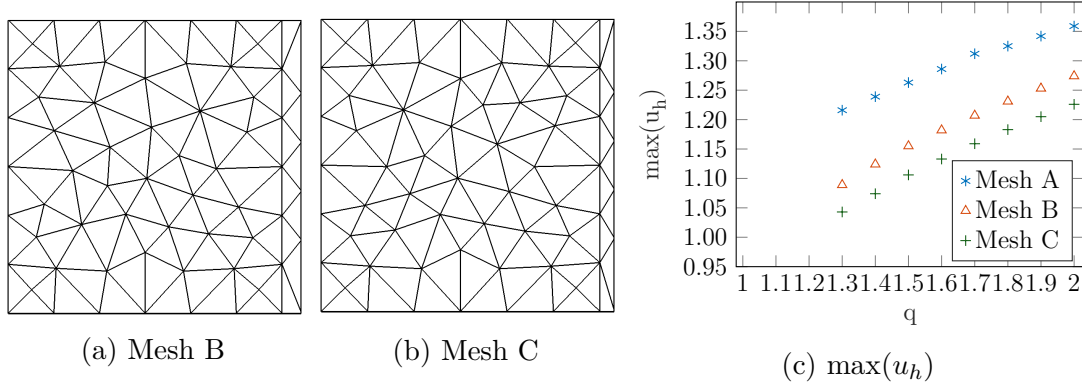


Figure 4.19 Overshoot in the L^q -best approximation on unstructured meshes. See Figure 4.18 for Mesh A.

Figure 4.19 shows two further unstructured meshes which have been designed in such a way that for every interior node (x_i, y_i) that is connected to the boundary $x = 1$ through one edge, the area of all triangles whose boundaries contain the node (x_i, y_i) and at least one node on the boundary $x = 1$ is smaller than the area of all remaining triangles whose boundaries contain the node (x_i, y_i) . The difference between the two meshes is that the distance between the boundary $x = 1$ and the vertical line containing all nodes connected to this boundary is smaller in Mesh C than in Mesh B. Figure 4.19c shows the maximum value of u_h for different q and Meshes A, B and C. This illustrates that the overshoot decreases on all three meshes as $q \rightarrow 1$. The overshoot on Mesh C is always smaller than the overshoot on the other two meshes and the overshoot on Mesh A is always larger than on the other two meshes. This illustrates that if the area of the elements connected to the boundary is decreased in comparison the area of the remaining elements, then the overshoot is reduced for any q and decreases more rapidly as $q \rightarrow 1$. This is consistent with the theoretical results in one dimension illustrated at the start of this chapter in Figure 4.3.

Chapter 5

The Convection-Diffusion-Reaction Equation in Non-Hilbert Sobolev Spaces: A Direct Proof of the Inf-Sup Condition

5.1 Introduction

In this chapter we prove the inf-sup conditions for the convection-diffusion-reaction problem in a non-symmetric Sobolev-space setting; in particular, we consider the $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ -setting, $1 < q < \infty$, $1/q + 1/q' = 1$. This setting is not only useful for the approach to eliminating unwanted oscillations near layers that is discussed in later chapters of this thesis, but it also allows more irregular solutions, if, e.g., the right-hand side is not in $H^{-1}(\Omega)$ but in $W^{-1,q'}(\Omega)$ for some $1 < q < 2$. Apart from a standard assumption on the advection and reaction coefficients, the other key assumption pertains to a subtle regularity requirement for the standard Laplacian. All results presented in this chapter are published as part of the article [\[93\]](#). A proof of an elementary stability result of Galerkin's method in that setting is furthermore given in [\[93\]](#).

In the context of FEMs, well-posedness of the convection-diffusion-reaction variational problem is traditionally established by proving coercivity and continuity of the underlying bilinear form in $H_0^1(\Omega)$. Due to the Lax-Milgram theorem this then implies well-posedness of the continuous problem and its discrete counterpart.

The concept of coercivity, however, requires that the trial and test spaces should be the same.¹ This is not always possible or desirable, e.g., in the context of Petrov-Galerkin methods or mixed methods, or if the continuous problem itself is stated in a non-symmetric manner such as in the $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ -setting.

A generalization of the Lax-Milgram theorem originally due to Nečas [140], cf., Theorem 3.38, replaces the coercivity requirement by two inf-sup conditions and can be formulated for Banach spaces with distinct test and trial spaces. A version of this theorem for approximations in finite dimensional subspaces is due to Babuška [9], cf., Theorem 3.39. It is important to note that the inf-sup conditions in the continuous formulation do not imply an inf-sup condition in the discrete setting. In fact, for the convection-diffusion-reaction equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$, discrete well-posedness is established under much stronger assumptions than in the continuous setting in [93].

In the remainder of the introduction we specify the weak formulation of the underlying problem, stipulate the assumptions and announce our main results.

5.1.1 Problem Statement

We consider the following model problem: find u such that

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + cu = f \quad \text{in } \Omega, \quad (5.1a)$$

$$u = 0 \quad \text{on } \Gamma = \partial\Omega, \quad (5.1b)$$

where $\varepsilon > 0$, $\mathbf{b} : \Omega \rightarrow \mathbb{R}^d$, and $c : \Omega \rightarrow \mathbb{R}$ are the (positive) diffusion parameter, convection field and reaction coefficient, respectively, and $f : \Omega \rightarrow \mathbb{R}$ is a given source.

Multiplying (5.1a) by a test function $v \in C_c^\infty(\Omega)$ and integrating by parts yields the bilinear form

$$\mathcal{B}_\varepsilon(u, v) = \varepsilon \int_\Omega \nabla u \cdot \nabla v \, d\mathbf{x} - \int_\Omega u \nabla \cdot (\mathbf{b}v) \, d\mathbf{x} + \int_\Omega cuv \, d\mathbf{x}. \quad (5.2)$$

This allows us to state the following variational problem for some $q \in (1, \infty)$: find $u \in U := W_0^{1,q}(\Omega)$ such that

$$\mathcal{B}_\varepsilon(u, v) = \langle f, v \rangle_{V', V} \quad \forall v \in V := W_0^{1,q'}(\Omega), \quad (5.3)$$

¹Moreover, the concept of coercivity is a notion relevant only in Hilbert spaces [69, Section A.2.4].

where f is allowed to be any element in $V' = W^{-1,q}(\Omega)$. Furthermore, we endow U and V with the following norms

$$\|u\|_U^q := c_0 \|u\|_{L^q(\Omega)}^q + \varepsilon \|\nabla u\|_{L^q(\Omega)}^q, \quad (5.4a)$$

$$\|v\|_V^{q'} := c_0 \|v\|_{L^{q'}(\Omega)}^{q'} + \varepsilon \|\nabla v\|_{L^{q'}(\Omega)}^{q'}, \quad (5.4b)$$

with c_0 given by Assumption 5.3 below.

The proof of well-posedness of this problem with $q = 2$ is well known, see, e.g., [69, Section 3.1]. Furthermore, well-posedness of the problem for $q \neq 2$ in smooth domains (e.g., domains with C^1 - or $C^{1,1}$ -boundary) is also well-established [75]. For $q \neq 2$ and general Lipschitz domains on the other hand, even simply proving well-posedness of the Poisson problem is far more challenging. It is no longer sufficient to require that Ω is a bounded Lipschitz domain. In fact, if $d \geq 3$, then for any $q > 3$ there exists a Lipschitz domain Ω and a right-hand side $f \in C^\infty(\bar{\Omega})$ such that the solution to the Poisson problem is not in $W^{1,q}(\Omega)$, cf., [103]; in two dimensions the same result holds for $q > 4$.

The novelty of our approach is that we provide a direct proof of the inf-sup conditions for the convection-diffusion-reaction equation illustrating that the so-called duality map, cf., Section 3.3, is invaluable as a replacement for the Riesz isometry in the context of Banach spaces.

Furthermore, our proof only relies on standard assumptions on $\mathbf{b}$ and c , and $W^{1,r}$ -regularity of the standard H_0^1 -Poisson problem for all r up to q' . The latter assumption can be interpreted as an assumption on the domain Ω for a given q , or, for a given bounded Lipschitz domain Ω it restricts the values that can be chosen for q . Another important technique in our proof (for the second inf-sup condition) is a bootstrapping argument involving repeated use of elliptic regularity and Sobolev embeddings to gain sufficient regularity. The proof of the *discrete* inf-sup condition in [93], as needed for the stability of Galerkin's method, furthermore assumes a diffusion-dominated case and the $W^{1,q'}$ -stability of the H_0^1 -projector.

5.1.2 Assumptions

We now present the assumptions in detail. One of the main assumptions we rely on is (adjoint-) regularity of the standard $H_0^1(\Omega)$ -Poisson problem.

Assumption 5.1 (Regularity for Laplacian). *Let $1 < q < 2$. For each $r \in (2, q']$, there is a stability constant $C_{r,\Omega}$, such that for all $g \in W^{-1,r}(\Omega) \subset H^{-1}(\Omega) :=$*

$[H_0^1(\Omega)]'$, the unique $z_g \in H_0^1(\Omega)$ that solves the problem

$$\int_{\Omega} \nabla w \cdot \nabla z_g \, d\mathbf{x} = \langle g, w \rangle_{W^{-1,r}(\Omega), W_0^{1,r'}(\Omega)} \quad \forall w \in H_0^1(\Omega),$$

satisfies the a priori estimate:

$$\|\nabla z_g\|_{L^r(\Omega)} \leq C_{r,\Omega} \|g\|_{W^{-1,r}(\Omega)}; \quad (5.5)$$

hence $z_g \in W_0^{1,r}(\Omega)$.

Remark 5.2. Assumption 5.1 holds for any $1 < q < 2$ if Ω has a C^1 boundary, cf., e.g., [75, 103, 81]. It is also true for any q if Ω is a convex Lipschitz domain, cf., [2, 73, 132]. The latter result can be extended to Lipschitz domains that satisfy an exterior ball condition [101]. For bounded Lipschitz domains, there exists $1 < q_0 < 2$ such that the assumption holds for all $q \geq q_0$, where $q_0 \leq 3/2$ if $d \geq 3$ and $q_0 \leq 4/3$ if $d = 2$. The precise value for q_0 depends on the domain Ω . It is possible to construct counterexamples for any $q < 3/2$ if $d \geq 3$ and for any $q < 4/3$ if $d = 2$. For more details we refer the reader to [103]. For $1 < q \leq 2$ it is possible to obtain higher regularity for the solution, i.e., $z_g \in W^{2,q}(\Omega)$, provided the right-hand side g is in $L^q(\Omega)$.

Similar to the analysis of the advection-reaction equation [33, 139], we also need some requirements on the advective field and the reaction coefficient.

Assumption 5.3 (Friedrich's positivity assumption). *Let $1 < q < \infty$. The advective field $\mathbf{b} \in L^\infty(\Omega)^d$ satisfies $\nabla \cdot \mathbf{b} \in L^\infty(\Omega)$, the reaction coefficient $c \in L^\infty(\Omega)$, and there exists a constant $c_0 > 0$ such that*

$$c(\mathbf{x}) - \frac{1}{q} \nabla \cdot \mathbf{b}(\mathbf{x}) \geq c_0, \quad \text{a.e. in } \Omega. \quad (5.6)$$

5.1.3 Main Results

We can now state the main result of this chapter, the proof of which is given in Section 5.3. We prove the inf-sup conditions in the continuous case under the assumptions stated in the previous section. To this end, we have the well-posedness in the continuous case.

Theorem 5.4. *Let $1 < q < \infty$. If Assumption 5.1 holds and additionally Assumption 5.3 is satisfied, then the variational problem (5.3) is well-posed. In particular,*

the bilinear form $\mathcal{B}_\varepsilon$ defined in (5.2) is bounded, i.e.,

$$\mathcal{B}_\varepsilon(w, v) \leq M \|v\|_V \|w\|_U \quad (5.7)$$

and satisfies the following inf-sup conditions: there exists $\gamma > 0$ such that

$$\inf_{w \in U} \sup_{v \in V} \frac{\mathcal{B}_\varepsilon(w, v)}{\|w\|_U \|v\|_V} \geq \gamma, \quad (5.8a)$$

and

$$\forall v \in V, \quad (\forall w \in U, \mathcal{B}_\varepsilon(w, v) = 0) \Rightarrow (v = 0). \quad (5.8b)$$

Both M and γ are positive constants that may depend on $\|\mathbf{b}\|_{L^\infty(\Omega)}$, $\|c\|_{L^\infty(\Omega)}$, c_0 , and ε .

Remark 5.5. The proof of (5.8a) given in Section 5.3 shows the following estimate for γ :

$$\gamma \geq \frac{\min\left(\frac{C_1^2}{2\varepsilon c_0^{2/q'}}, \frac{\varepsilon^{1-2/q'}}{2}\right)}{2C_{q',\Omega}(c_0 C_{F,q}^q + \varepsilon)^{1/q'} \varepsilon^{-1/q} + C_1^2 c_0^{-1} \varepsilon^{-1} 2^{\frac{1}{q} - \frac{1}{q'}}}. \quad (5.9)$$

Here, $C_{F,q}$ is the constant in the Poincaré-Friedrichs inequality in $W^{1,q}(\Omega)$, $C_{q',\Omega}$ is the constant given in Assumption 5.1 for $r = q'$ and

$$C_1 = C_{q',\Omega} \left(\|b\|_{L^\infty(\Omega)} + C_{F,q} \|c\|_{L^\infty(\Omega)} \right).$$

For very small ε the dominating term in the numerator scales like $\varepsilon^{1-\frac{2}{q'}}$. Note that this is equal to 1 for $q = q' = 2$. The denominator scales like ε^{-1} ; the overall scaling for very small ε is therefore $\varepsilon^{2-\frac{2}{q'}}$. For very small c_0 the lower bound for the inf-sup constant γ scales like c_0 . It can be assumed that this estimate is not optimal, since the scaling in ε and c_0 for $q = q' = 2$ is less favourable than in the standard proof when $q = q' = 2$.

5.1.4 Outline of the Chapter

The rest of the chapter is organized as follows. In Section 5.2, we present several necessary preliminaries, in particular, we present well-posedness results for the Poisson problem in the $W_0^{1,q}$ - $W_0^{1,q'}(\Omega)$ setting, fundamental properties of duality

mappings, and a brief discussion on difficulties in proving the continuous inf-sup conditions. We then present in Section 5.3 the proof of Theorem 5.4, relying heavily on the preliminaries in the preceding section. Section 5.3 discusses in the following order: the proof of the continuity of $\mathcal{B}_\varepsilon$, the first inf-sup condition, and finally, the second inf-sup condition. Finally, we state the discrete analogue of the inf-sup condition in Section 5.4 and present numerical experiments for Galerkin's method considering the approximation of irregular solutions $u \notin H_0^1(\Omega)$.

5.2 The Poisson Equation and the Advection-Reaction Equation in $L^q(\Omega)$

In the context of finite element methods the analysis of variational formulations of PDEs is traditionally undertaken in subspaces of $L^2(\Omega)$ such as $H_0^1(\Omega)$. Even though we are considering a formulation in more general Sobolev spaces that are no longer Hilbert spaces, we are still using many standard techniques that have been developed for the numerical analysis of finite element methods. In this section we focus on illustrating how concepts and techniques that rely on a Hilbert space setting with identical test and trial spaces can be extended to our more general setting. To this end, we first consider the Poisson problem in the $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ -setting as a simplified model problem, which will be crucial later when dealing with the full convection-diffusion-reaction problem.

As part of the proof of the inf-sup-conditions, we will construct a very specific functional in $W^{-1,q'}(\Omega)$ that is an example of a duality mapping, cf., Section 3.3. Duality mappings have proven to be a very useful concept for mimicking certain techniques that rely on properties of Hilbert spaces in more general Banach spaces. In some sense duality mappings are a suitable non-linear replacement for the Riesz map. This becomes particularly evident in the context of residual minimization problems; see, for example, [138].

Finally, we conclude this section by explaining how we intend to include the lower order terms in the convection-diffusion-advection equation. The next sections will then focus on the rigorous proof of Theorem 5.4.

5.2.1 The Poisson Problem in the $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ -setting

The goal is to extend the well-posedness proof of the Poisson problem to the convection-diffusion-reaction equation provided that Assumption 5.1 holds. Assumption 5.1 implies the following well-posedness result for the Poisson problem.

Proposition 5.6 (Well-posedness of the Poisson problem in Banach Sobolev spaces). *Let $1 < q < 2$. If Assumption 5.1 holds, then for every $f \in W^{-1,q}(\Omega)$ there exists a unique $u_f \in W_0^{1,q}(\Omega)$ that satisfies the Poisson problem:*

$$\int_{\Omega} \nabla u_f \cdot \nabla v \, d\mathbf{x} = \langle f, v \rangle_{W^{-1,q}(\Omega), W_0^{1,q'}(\Omega)} \quad \forall v \in W_0^{1,q'}(\Omega).$$

Furthermore, u_f satisfies the a priori estimate

$$\|\nabla u_f\|_{L^q(\Omega)} \leq C_{q',\Omega} \|f\|_{W^{-1,q}(\Omega)}.$$

Although the above well-posedness result is known (cf. [103]), a proof of the corresponding inf-sup conditions using the duality map is given in [93], where it is also demonstrated that when $d = 1$, Assumption 5.1 is not needed.

Combining Proposition 5.6 again with Assumption 5.1 immediately implies the following general regularity result; this will be required in the proof of Theorem 5.4 instead of Assumption 5.1 directly.

Corollary 5.7 (Elliptic regularity for Laplacian in Banach Sobolev spaces). *Let $1 < q < 2$ and $q \leq r \leq q'$. If Assumption 5.1 holds, then for all $g \in W^{-1,r}(\Omega) \subset W^{-1,q}(\Omega)$, the unique solution $u_g \in W_0^{1,q}(\Omega)$ that solves the problem*

$$\int_{\Omega} \nabla u_g \cdot \nabla v \, d\mathbf{x} = \langle g, v \rangle_{W^{-1,r}(\Omega), W_0^{1,r'}(\Omega)} \quad \forall v \in W_0^{1,q'}(\Omega)$$

satisfies the a priori estimate:

$$\|\nabla u_g\|_{L^r(\Omega)} \leq C_{r,\Omega} \|g\|_{W^{-1,r}(\Omega)};$$

hence $u_g \in W_0^{1,r}(\Omega)$. The constant $C_{r,\Omega} = C_{r',\Omega}$ ($r \neq 2$) is the same as in Assumption 5.1, and $C_{2,\Omega} = 1$.

Proof. This result is a direct consequence of Assumption 5.1 and Proposition 5.6. Suppose $r \in (q, 2)$, then apply Proposition 5.6 (replacing q by r and f by g) to obtain that $u_g \in W_0^{1,r}(\Omega) \subset W_0^{1,q}(\Omega)$; by duality $C_{r',\Omega} = C_{r,\Omega}$. Suppose that $r = 2$, then use the standard H_0^1 -Poisson problem with right-hand side $g \in H^{-1}(\Omega)$ to obtain that $u_g \in H_0^1(\Omega) \subset W_0^{1,q}(\Omega)$ and $\|\nabla u_g\|_{L^2(\Omega)} \leq \|g\|_{H^{-1}(\Omega)}$. Suppose $r \in (2, q')$, then apply Assumption 5.1 (replacing z_g by u_g) to obtain that $u_g \in W_0^{1,r}(\Omega) \subset H_0^1(\Omega) \subset W_0^{1,q}(\Omega)$. $\square$

Let us recall that for $q = 2$ well-posedness follows immediately from the Lax-Milgram Theorem since both continuity and coercivity are obvious in this case.

The same approach is obviously no longer applicable if $q \neq 2$ since it requires the trial and test spaces to be identical. We thus apply a generalized Lax-Milgram Theorem (or BNB Theorem [69, Chapter 2]) that is originally due to Nečas [140], cf., Theorem 3.38. A version of this theorem for approximations in finite dimensional subspaces was derived by Babuška [9], cf., Theorem 3.39.

For the Poisson problem with $q = 2$, i.e.,

$$b(u, v) := \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x}, \quad U = V := H_0^1(\Omega), \quad \|\cdot\|_U := |\cdot|_{H^1(\Omega)},$$

the inf-sup condition (3.7a) is immediate since in this case

$$\sup_{v \in V=U} \frac{b(u, v)}{\|v\|_V} \geq \frac{b(u, u)}{\|u\|_U} = \frac{\|u\|_U^2}{\|u\|_U} = \|u\|_U.$$

The key observation here is that $b(u, u)$ is equal to the square of the norm on U .

For $q < 2$, we can no longer use u itself as a test function. But if we suppose that there exists a v_u such that

$$b(u, v_u) = |u|_{W^{1,q}(\Omega)}^q, \tag{5.10a}$$

$$|v_u|_{W^{1,q'}(\Omega)} \leq C|u|_{W^{1,q}(\Omega)}^{q-1}, \tag{5.10b}$$

we would again immediately obtain an inf-sup condition.

Remark 5.8. Later we will instead use $\tilde{v}_u$ such that

$$b(u, \tilde{v}_u) = |u|_{W^{1,q}(\Omega)}^2,$$

$$|\tilde{v}_u|_{W^{1,q'}(\Omega)} \leq C|u|_{W^{1,q}(\Omega)}.$$

Note that if v_u satisfies (5.10a) and (5.10b), we can define $\tilde{v}_u = |u|_{W^{1,q}(\Omega)}^{2-q} v_u$.

The question now is whether it is possible to construct such a test function $v_u \in W_0^{1,q'}(\Omega)$ for any $u \in W_0^{1,q}(\Omega)$. Let us start with the right-hand side of (5.10a). As a first step we determine a $\boldsymbol{\xi} : \Omega \rightarrow \mathbb{R}^d$ such that

$$|u|_{W^{1,q}(\Omega)}^q = \int_{\Omega} \sum_{i=1}^d |\partial_i u|^q \, d\mathbf{x} = \int_{\Omega} \nabla u \cdot \boldsymbol{\xi} \, d\mathbf{x}.$$

If we divide each summand by $\partial_i u$, we can see that $\boldsymbol{\xi}$ has to be of the form

$$\xi_i = |\partial_i u|^{q-1} \operatorname{sgn}(\partial_i u).$$

Ideally, we would like to construct $v_u \in W_0^{1,q'}(\Omega)$ such that $\nabla v_u = \xi$, which, however, seems to be impossible in general. Instead, we identify ξ with $\ell_\xi \in W^{-1,q'}(\Omega)$ as follows

$$\langle \ell_\xi, w \rangle_{W^{-1,q'}(\Omega), W_0^{1,q}(\Omega)} = \int_{\Omega} \nabla w \cdot \xi \, d\mathbf{x}.$$

It is easy to check that we have $\xi_i \in L^{q'}(\Omega)$ and as a consequence $\ell_\xi \in W^{-1,q'}(\Omega)$. Furthermore, we can compute $\|\ell_\xi\|_{W^{-1,q'}(\Omega)} = |u|_{W^{1,q}(\Omega)}^{q-1}$ using the duality of the spaces $L^q(\Omega)$ and $L^{q'}(\Omega)$. We can now define v_u as the solution to the following Poisson problem:

$$\int_{\Omega} \nabla w \cdot \nabla v_u \, d\mathbf{x} = \langle \ell_\xi, w \rangle_{W^{-1,q'}(\Omega), W_0^{1,q}(\Omega)} \quad \forall w \in W_0^{1,q}(\Omega).$$

Note that the existence of $v_u \in H_0^1(\Omega)$ is due to the well-posedness of the Poisson problem in $H_0^1(\Omega)$. The higher regularity of v_u then follows from Assumption 5.1 and the a priori estimate in Assumption 5.1 implies (5.10b). Furthermore note that by definition $\ell_\xi(u) = |u|_{W^{1,q}(\Omega)}^q$, which implies (5.10a).

If we instead use

$$\tilde{\xi}_i = |u|_{W^{1,q}(\Omega)}^{2-q} |\partial_i u|^{q-1} \operatorname{sgn}(\partial_i u),$$

we can analogously define a linear functional $\ell_{\tilde{\xi}}$ and a test function $\tilde{v}_u$ that satisfies the conditions in Remark 5.8.

Let us recall the duality mapping with weight $\varphi(t) = t^{q-1}$ on the space $W_0^{1,q}(\Omega)$ from Section 3.3.2:

$$\langle J_{W_0^{1,q}}(v), w \rangle_{L^{q'}(\Omega), L^q(\Omega)} = \int_{\Omega} \sum_{i=1}^d |\partial_i v|^{q-1} \operatorname{sgn}(\partial_i v) \partial_i w \, d\mathbf{x}.$$

We can now clearly see that ℓ_ξ is nothing but the duality mapping $J_{W_0^{1,q}}$. Moreover, we have $\ell_{\tilde{\xi}} = \tilde{J}_{W_0^{1,q}}$, where $\tilde{J}_{W_0^{1,q}}(v) = |v|_{W^{1,q}(\Omega)}^{2-q} J_{W_0^{1,q}}(v)$ is the normalized duality mapping on $W_0^{1,q}(\Omega)$. This illustrates how duality mappings can serve as a useful tool in Banach spaces.

5.2.2 Lower Order Terms

We now aim to extend the result we have seen for the Poisson problem to the convection-diffusion-reaction equation. For $q = 2$, well-posedness again follows from

the Lax-Milgram Theorem due to the continuity and coercivity of the underlying bilinear form.

Let us take a brief look at the well-known proof of coercivity for the bilinear form $\mathcal{B}_\varepsilon$ under Assumption 5.3 with $q = 2$:

$$\begin{aligned}
 \mathcal{B}_\varepsilon(u, u) &= \varepsilon \int_{\Omega} \nabla u \cdot \nabla u \, d\mathbf{x} - \int_{\Omega} u \nabla \cdot (\mathbf{b}u) \, d\mathbf{x} + \int_{\Omega} cu^2 \, d\mathbf{x} \\
 &= \varepsilon \int_{\Omega} \nabla u \cdot \nabla u \, d\mathbf{x} - \frac{1}{2} \int_{\Omega} (\nabla \cdot \mathbf{b}) u^2 \, d\mathbf{x} + \int_{\Omega} cu^2 \, d\mathbf{x} \\
 &\geq \varepsilon \int_{\Omega} \nabla u \cdot \nabla u \, d\mathbf{x} + c_0 \int_{\Omega} u^2 \, d\mathbf{x} \\
 &= \varepsilon \|\nabla u\|_{L^2(\Omega)}^2 + c_0 \|u\|_{L^2(\Omega)}^2.
 \end{aligned} \tag{5.11}$$

As for the Poisson problem, coercivity also immediately implies the inf-sup condition (3.7a).

For $q \neq 2$, we have already seen how the Laplace operator can be treated; thereby, we will now focus on the lower order terms. In (5.11), there are two steps that are applied to the lower order terms: the first one is rearranging everything by integration by parts such that Assumption 5.3 can be applied and the second one is the rather trivial observation that the resulting term can be bounded by the square of the $L^2(\Omega)$ -norm of u . If $q \neq 2$, we again have the problem that we cannot simply test with u itself. The idea is to start from the end and mimic the steps in the proof for the Poisson problem by constructing a test function v such that

$$\int_{\Omega} uv \, d\mathbf{x} = \|u\|_{L^q(\Omega)}^2, \quad \|v\|_{L^{q'}(\Omega)} = \|u\|_{L^q(\Omega)}.$$

Due to the duality of the spaces $L^q(\Omega)$ and $L^{q'}(\Omega)$ this construction actually becomes easier than for the Poisson problem. Either by using the abstract concept of duality mappings or by explicitly constructing v , we can see immediately that v must be defined as

$$v := \tilde{J}_q(u) = \|u\|_{L^q(\Omega)}^{q-2} |u|^{q-1} \operatorname{sgn}(u),$$

cf., Section 3.3.2.

It is easy to verify that $v \in L^{q'}(\Omega)$. However, to use v as a test function we require $v \in W_0^{1,q'}(\Omega)$. It turns out, that this is not true for $q < 2$, but it is the case for $q > 2$ (a proof of this will be given in Section 5.3.3). This suggests to prove the

following inf-sup conditions on the adjoint instead:

$$\inf_{v \in V} \sup_{w \in U} \frac{\mathcal{B}_\varepsilon(w, v)}{\|w\|_U \|v\|_V} \geq \gamma, \quad (5.12a)$$

$$\forall w \in U, \quad (\forall v \in V, \mathcal{B}_\varepsilon(w, v) = 0) \Rightarrow (w = 0). \quad (5.12b)$$

This is equivalent to (5.8) (for an elementary proof, see, e.g., [130, Prop. A.2]). In particular, the inf-sup constant γ is the same.

Another observation is the following. We need to be able to mimic the steps in the proof of coercivity in the L^q -setting:

$$\mathbf{b} \cdot \nabla v |v|^{q'-1} \operatorname{sgn}(v) = \frac{1}{q'} \nabla \cdot (\mathbf{b} |v|^{q'}) - \frac{1}{q'} (\nabla \cdot \mathbf{b}) |v|^{q'}. \quad (5.13)$$

The last issue that has to be resolved is that we have two different test functions — one for the Laplace operator and one for the advection operator. In order to obtain an inf-sup condition for the full bilinear form we thus consider a linear combination of both functions and additionally estimate the Laplace term tested with the second test function and the advection-reaction term with the first. The rather technical details of this are presented in the next section.

5.3 Proof of Theorem 5.4: Well-posedness of the Convection-Diffusion-Reaction Equation in $W_0^{1,q}(\Omega)$

In this section we now present a rigorous proof of Theorem 5.4. We start by showing continuity of the bilinear form $\mathcal{B}_\varepsilon$ in Section 5.3.1. Sections 5.3.2 and 5.3.3 then contain all necessary estimates for the two test functions introduced in Sections 5.2.1 and 5.2.2, respectively. In Section 5.3.4 we consider a linear combination of the two test functions and combine all of the estimates to conclude the proof of the inf-sup condition (5.12a). We then finish the proof of Theorem 5.4 in Section 5.3.5 by proving the second inf-sup condition (5.12b). For this last step we employ a so-called bootstrap argument that is more commonly used in the PDE literature, cf., Remark 5.11 below.

Remark 5.9. If $q = 2$ it is possible to obtain ε -independent estimates for the inf-sup constant and the continuity constant if we choose a different norm on V , namely

$$\|v\|_{\tilde{V}} := \|v\|_V + \|\nabla \cdot (\mathbf{b}v)\|_{U'},$$

where we consider $\nabla \cdot (\mathbf{b}v) \in L^2(\Omega)$ as an element in $U' = H^{-1}(\Omega) \supset L^2(\Omega)$ with its norm on U' given by

$$\|\nabla \cdot (\mathbf{b}v)\|_{U'} = \sup_{u \in U} \frac{\langle \nabla \cdot (\mathbf{b}v), u \rangle_{U', U}}{\|u\|_U} \quad (5.14)$$

A proof of this result is presented in [164, Section 4.4.1]. There are essentially three steps to the proof: firstly, it is easy to prove that the continuity constant with respect to the norms $\|\cdot\|_U$ and $\|\cdot\|_{\tilde{V}}$ is independent of ε ; secondly, an ε -independent estimate for the inf-sup constant can be proven using the norm (5.4a) on V ; thirdly, it is then possible to bound the term (5.14) from above as well without losing robustness of the inf-sup constant. The first and the third step can easily be extended to $q \neq 2$. However, the generalization of the second step to Banach spaces remains an open problem.

5.3.1 Continuity of the Bilinear Form

In order to prove continuity of the bilinear form (5.7), we apply the Hölder inequality and obtain

$$\begin{aligned} \mathcal{B}_\varepsilon(w, v) &= \varepsilon \int_\Omega \nabla w \cdot \nabla v \, d\mathbf{x} - \int_\Omega w \nabla \cdot (\mathbf{b}v) \, d\mathbf{x} + \int_\Omega c w v \, d\mathbf{x} \\ &\leq \varepsilon \|\nabla w\|_{L^q(\Omega)} \|\nabla v\|_{L^{q'}(\Omega)} + \frac{\|c\|_{L^\infty(\Omega)}}{c_0} c_0 \|w\|_{L^q(\Omega)} \|v\|_{L^{q'}(\Omega)} \\ &\quad + \frac{\|\mathbf{b}\|_{L^\infty(\Omega)}}{c_0^{1/q'} \varepsilon^{1/q}} \varepsilon^{1/q} \|\nabla w\|_{L^q(\Omega)} c_0^{1/q'} \|v\|_{L^{q'}(\Omega)} \\ &\leq \left(\max \left(1, \frac{\|c\|_{L^\infty(\Omega)}}{c_0} \right) + \frac{\|\mathbf{b}\|_{L^\infty(\Omega)}}{c_0^{1/q'} \varepsilon^{1/q}} \right) \|w\|_U \|v\|_V. \end{aligned}$$

Here we chose to integrate the advection term by parts and then apply the Hölder inequality. This is a rather arbitrary choice, but note that some ε -dependence of the continuity constant cannot be avoided even if the advection term is estimated in its current form.

5.3.2 A Test Function for the Laplace Operator

We now start with proving the inf-sup condition (5.12a) by establishing estimates for a test function that is tailored for the diffusion part of the bilinear form. Let $u_v \in W_0^{1,q}(\Omega)$ be the unique solution to the problem

$$\int_{\Omega} \nabla u_v \cdot \nabla z \, d\mathbf{x} = \langle \tilde{J}_{W_0^{1,q'}}(v), z \rangle_{W^{-1,q}(\Omega), W_0^{1,q'}(\Omega)} \quad \forall z \in W_0^{1,q'}(\Omega).$$

Hence, by definition

$$\varepsilon \int_{\Omega} \nabla u_v \cdot \nabla v \, d\mathbf{x} = \varepsilon \langle \tilde{J}_{W_0^{1,q'}}(v), v \rangle_{W^{-1,q}(\Omega), W_0^{1,q'}(\Omega)} = \varepsilon \|\nabla v\|_{L^{q'}(\Omega)}^2. \quad (5.15)$$

Next we look at the advection-reaction term; to bound these terms we integrate by parts, and employ Hölder's inequality, the Poincaré-Friedrichs inequality:

$$\|w\|_{L^q(\Omega)} \leq C_{F,q} \|\nabla w\|_{L^q(\Omega)} \quad \forall w \in W_0^{1,q}(\Omega), \quad (5.16)$$

and the a priori estimate for ∇u_v (5.5). Thereby, we get

$$\begin{aligned} \left| -\int_{\Omega} u_v \nabla \cdot (\mathbf{b}v) \, d\mathbf{x} + \int_{\Omega} cu_v v \, d\mathbf{x} \right| &= \left| \int_{\Omega} (\mathbf{b} \cdot \nabla u_v) v \, d\mathbf{x} + \int_{\Omega} cu_v v \, d\mathbf{x} \right| \\ &\leq \|\mathbf{b}\|_{L^\infty(\Omega)} \|\nabla u_v\|_{L^q(\Omega)} \|v\|_{L^{q'}(\Omega)} + \|c\|_{L^\infty(\Omega)} \|u_v\|_{L^q(\Omega)} \|v\|_{L^{q'}(\Omega)} \\ &\leq \left(\|\mathbf{b}\|_{L^\infty(\Omega)} + C_{F,q} \|c\|_{L^\infty(\Omega)} \right) \|\nabla u_v\|_{L^q(\Omega)} \|v\|_{L^{q'}(\Omega)} \\ &\leq C_{q',\Omega} \left(\|\mathbf{b}\|_{L^\infty(\Omega)} + C_{F,q} \|c\|_{L^\infty(\Omega)} \right) \|\nabla v\|_{L^{q'}(\Omega)} \|v\|_{L^{q'}(\Omega)}. \end{aligned}$$

To simplify the notation we define

$$C_1 = C_{q',\Omega} \left(\|\mathbf{b}\|_{L^\infty(\Omega)} + C_{F,q} \|c\|_{L^\infty(\Omega)} \right).$$

Using Young's inequality gives

$$\left| -\int_{\Omega} u_v \nabla \cdot (\mathbf{b}v) \, d\mathbf{x} + \int_{\Omega} cu_v v \, d\mathbf{x} \right| \leq \left(\frac{\varepsilon}{2} \|\nabla v\|_{L^{q'}(\Omega)}^2 + \frac{C_1^2}{2\varepsilon} \|v\|_{L^{q'}(\Omega)}^2 \right). \quad (5.17)$$

Remark 5.10. The above definition of the test function u_v relies on the fact that ε is a constant scalar. The generalization to more general elliptic operators is not entirely trivial and would require a generalization of Assumption 5.1 and Proposition 5.6 such that both apply to the more general operator. With this, it should then be possible to extend the proof to the more general setting by adjusting (5.15) such that the more general elliptic operator replaces ε on the left hand side

and a suitable constant, most likely the corresponding ellipticity constant, replaces ε on the right hand side.

5.3.3 A Test Function for the Convection-Reaction Operator

We now consider the second test function introduced in Section 5.2.2 to obtain a lower bound for the advection-reaction term. As we have already mentioned, we require $W_0^{1,q}(\Omega)$ -regularity for the test function and hence we first have to prove $\tilde{J}_{q'}(v) \in W_0^{1,q}(\Omega)$. By definition $\tilde{J}_{q'}(v) \in L^q(\Omega)$; we will now show that $\nabla \tilde{J}_{q'}(v) = \|v\|_{L^{q'}(\Omega)}^{2-q'} |v|^{q'-2} \nabla v \in [L^q(\Omega)]^d$. Indeed, since $v \in W_0^{1,q'}(\Omega)$, we deduce that

$$\begin{aligned}
 \|\nabla \tilde{J}_{q'}\|_{L^q(\Omega)}^q &= \|v\|_{L^{q'}(\Omega)}^{(2-q')q} \int_{\Omega} |v|^{q'-2} |\nabla v|^q \, d\mathbf{x} \\
 &= \|v\|_{L^{q'}(\Omega)}^{(2-q')q} \int_{\Omega} |v|^{(q'-2)q} |\nabla v|^q \, d\mathbf{x} \\
 &\leq \|v\|_{L^{q'}(\Omega)}^{(2-q')q} \|v\|_{L^{\frac{q'-1}{q'-2}}(\Omega)}^{(q'-2)q} \|\nabla v\|_{L^{q'-1}(\Omega)}^q \\
 &= \|v\|_{L^{q'}(\Omega)}^{(2-q')q} \left(\int_{\Omega} |v|^{q'} \, d\mathbf{x} \right)^{\frac{q'-2}{(q'-1)}} \left(\int_{\Omega} |\nabla v|^{q'} \, d\mathbf{x} \right)^{\frac{1}{(q'-1)}} \\
 &= \|v\|_{L^{q'}(\Omega)}^{(2-q')q} \|v\|_{L^{q'}(\Omega)}^{(q'-2)q} \|\nabla v\|_{L^{q'}(\Omega)}^q = \|\nabla v\|_{L^{q'}(\Omega)}^q < \infty,
 \end{aligned} \tag{5.18}$$

where it is essential that $q' > 2$. Now we can apply (5.13) to obtain

$$\begin{aligned}
 & - \int_{\Omega} \tilde{J}_{q'}(v) \nabla \cdot (\mathbf{b}v) \, d\mathbf{x} + \int_{\Omega} c \tilde{J}_{q'}(v) v \, d\mathbf{x} \\
 &= \|v\|_{L^{q'}(\Omega)}^{2-q'} \int_{\Omega} |v|^{q'-1} \operatorname{sgn}(v) (cv - \nabla \cdot (\mathbf{b}v)) \, d\mathbf{x} \\
 &= \|v\|_{L^{q'}(\Omega)}^{2-q'} \int_{\Omega} |v|^{q'-1} \operatorname{sgn}(v) (cv - (\nabla \cdot \mathbf{b})v - \mathbf{b} \cdot \nabla v) \, d\mathbf{x} \\
 &= \|v\|_{L^{q'}(\Omega)}^{2-q'} \int_{\Omega} |v|^{q'} \left(c - \frac{1}{q} \nabla \cdot \mathbf{b} \right) - \frac{1}{q'} \nabla \cdot (\mathbf{b}|v|^{q'}) \, d\mathbf{x} \\
 &\geq c_0 \|v\|_{L^{q'}(\Omega)}^2 - \frac{\|v\|_{L^{q'}(\Omega)}^{2-q'}}{q'} \int_{\partial\Omega} (\mathbf{b} \cdot \mathbf{n}) |v|^{q'} \, ds = c_0 \|v\|_{L^{q'}(\Omega)}^2.
 \end{aligned} \tag{5.19}$$

Next, we have to bound the diffusion term when testing with $\tilde{J}_{q'}(v)$. Here, we observe that

$$\int_{\Omega} \nabla(\tilde{J}_{q'}(v)) \cdot \nabla v \, d\mathbf{x} = \|v\|_{L^{q'}(\Omega)}^{2-q'} \int_{\Omega} |v|^{q'-2} \nabla v \cdot \nabla v \, d\mathbf{x} \geq 0; \quad (5.20)$$

thereby, we can disregard this term in the inf-sup analysis.

5.3.4 Combining the Estimates

We now combine all the above estimates and prove a lower bound for the bilinear form when testing with a linear combination of the two test functions above, i.e., we test with

$$w_v := u_v + \beta \tilde{J}_{q'}(v),$$

where β is a constant to be chosen. Combining the estimates (5.19), (5.20), (5.15) and (5.17), we obtain

$$\begin{aligned} \mathcal{B}_{\varepsilon}(w_v, v) \geq & \varepsilon \|\nabla v\|_{L^{q'}(\Omega)}^2 + \beta c_0 \|v\|_{L^{q'}(\Omega)}^2 \\ & - \left(\frac{\varepsilon}{2} \|\nabla v\|_{L^{q'}(\Omega)}^2 + \frac{C_1^2}{2\varepsilon} \|v\|_{L^{q'}(\Omega)}^2 \right). \end{aligned}$$

If we now choose $\beta = C_1^2 c_0^{-1} \varepsilon^{-1}$, we have

$$\begin{aligned} \mathcal{B}_{\varepsilon}(w_v, v) & \geq \frac{C_1^2}{2\varepsilon} \|v\|_{L^{q'}(\Omega)}^2 + \frac{\varepsilon}{2} \|\nabla v\|_{L^{q'}(\Omega)}^2 \\ & \geq \min \left(\frac{C_1^2}{2\varepsilon c_0^{2/q'}}, \frac{\varepsilon^{1-2/q'}}{2} \right) \left(c_0^{2/q'} \|v\|_{L^{q'}(\Omega)}^2 + \varepsilon^{2/q'} \|\nabla v\|_{L^{q'}(\Omega)}^2 \right) \\ & \geq \frac{1}{2} \min \left(\frac{C_1^2}{2\varepsilon c_0^{2/q'}}, \frac{\varepsilon^{1-2/q'}}{2} \right) \|v\|_V^2. \end{aligned} \quad (5.21)$$

To conclude, we now only need to estimate $\|w_v\|_U$. According to Proposition 5.6 we have

$$\|\nabla u_v\|_{L^q(\Omega)} \leq C_{q',\Omega} \|\tilde{J}_{W_0^{1,q'}}(v)\|_{W^{-1,q}(\Omega)} = C_{q',\Omega} \|\nabla v\|_{L^{q'}(\Omega)}.$$

Thus, recalling the Poincaré-Friedrichs inequality (5.16), we obtain

$$\begin{aligned}
 \|u_v\|_U &= \left(c_0 \|u_v\|_{L^q(\Omega)}^q + \varepsilon \|\nabla u_v\|_{L^q(\Omega)}^q \right)^{1/q} \\
 &\leq \left(c_0 C_{F,q}^q + \varepsilon \right)^{1/q} \|\nabla u_v\|_{L^q(\Omega)} \\
 &\leq C_{q',\Omega} \left(c_0 C_{F,q}^q + \varepsilon \right)^{1/q} \|\nabla v\|_{L^{q'}(\Omega)} \\
 &\leq C_{q',\Omega} \left(c_0 C_{F,q}^q + \varepsilon \right)^{1/q} \varepsilon^{-1/q'} \|v\|_V.
 \end{aligned}$$

Similarly, using (5.18), gives

$$\begin{aligned}
 \|\beta J_{q'}(v)\|_U &= \beta \left(c_0 \|\tilde{J}_{q'}(v)\|_{L^q(\Omega)}^q + \varepsilon \|\nabla \tilde{J}_{q'}(v)\|_{L^q(\Omega)}^q \right)^{1/q} \\
 &\leq \beta \left(c_0 \|v\|_{L^{q'}(\Omega)}^q + \varepsilon \|\nabla v\|_{L^q(\Omega)}^q \right)^{1/q} \\
 &\leq C_1^2 c_0^{-1} \varepsilon^{-1} 2^{1/q-1/q'} \|v\|_V.
 \end{aligned}$$

Dividing (5.21) by $\|w_v\|_U$ and using the above estimates finally yields the inf-sup condition (5.12a). We refer the reader to (5.9) for the final estimate of the inf-sup constant γ .

5.3.5 The Second Inf-Sup Condition

The final step is the proof of the second inf-sup condition (5.12b). To this end, assume there exists $0 \neq w \in W_0^{1,q}(\Omega)$ such that

$$\begin{aligned}
 0 = \mathcal{B}_\varepsilon(w, v) &= \varepsilon \int_\Omega \nabla w \cdot \nabla v \, d\mathbf{x} - \int_\Omega w \nabla \cdot (\mathbf{b}v) \, d\mathbf{x} + \int_\Omega c w v \, d\mathbf{x} \\
 &= \varepsilon \int_\Omega \nabla w \cdot \nabla v \, d\mathbf{x} + \int_\Omega \mathbf{b} \cdot \nabla w v \, d\mathbf{x} + \int_\Omega c w v \, d\mathbf{x} \quad \forall v \in W_0^{1,q'}(\Omega).
 \end{aligned}$$

The idea is to use the above equation to gain sufficient regularity in order to test with w itself. A priori this is not possible since $W_0^{1,q}(\Omega) \not\subset W_0^{1,q'}(\Omega)$ for $q' > q$. We will consider w as a solution to a Poisson problem and then use the regularity of the right-hand side in order to employ Corollary 5.7 and the Sobolev embedding theorem, cf., Theorem 3.22, to gain higher regularity for w . We then iterate this until we have gained sufficient regularity.

Given that $\mathbf{b} \cdot \nabla w + cw \in L^q(\Omega)$, we can consider w as the solution of a Poisson problem with right hand side $g = \mathbf{b} \cdot \nabla w + cw$. Denote by q^* the Sobolev conjugate

of q given by

$$\frac{1}{q^*} = \frac{1}{q} - \frac{1}{d}.$$

Furthermore, let s be the dual exponent to q^* , i.e., $1 = 1/q^* + 1/s$. Then q' is the Sobolev conjugate to s . Indeed,

$$\frac{1}{q^*} = \frac{1}{q} - \frac{1}{d} \Leftrightarrow 1 - \frac{1}{s} = 1 - \frac{1}{q'} - \frac{1}{d} \Leftrightarrow \frac{1}{q'} = \frac{1}{s} - \frac{1}{d}.$$

Thus we obtain from the Sobolev embedding theorem that $W_0^{1,s}(\Omega) \subset L^{q'}(\Omega)$ and therefore $L^q(\Omega) = [L^{q'}(\Omega)]' \subset [W_0^{1,s}(\Omega)]' = W^{-1,q^*}(\Omega)$.

Suppose that $q^* \geq q'$, then Corollary 5.7 with $r = q'$ can be used (indeed $L^q(\Omega) \subset W^{-1,q^*}(\Omega) \subset W^{-1,q'}(\Omega)$) to obtain $w \in W_0^{1,q'}(\Omega)$, and we can select $v = w$ as a test function.

Otherwise, for $q^* < q'$, by Corollary 5.7 with $r = q^*$, we obtain $w \in W_0^{1,q^*}(\Omega)$. We subsequently observe that w satisfies a Poisson problem with a right-hand side $\mathbf{b} \cdot \nabla w + cw \in L^{q^*}(\Omega)$ and we can define iteratively

$$\frac{1}{r_{i+1}} = \max \left(\frac{1}{q'}, \frac{1}{r_i} - \frac{1}{d} \right), \quad r_1 := q^*.$$

We then obtain iteratively $w \in W_0^{1, \min\{r_i, q'\}}(\Omega)$ and $\mathbf{b} \cdot \nabla w + cw \in L^{\min\{r_i, q'\}}(\Omega)$. Note that

$$\frac{1}{r_i} = \max \left(\frac{1}{q'}, \frac{1}{q} - \frac{i}{d} \right).$$

Thus if we choose an integer $i \geq d/q - d/q'$, we observe that $w \in W_0^{1,q'}(\Omega)$ and hence we can choose $v = w$ as a test function.

Therefore, $0 = \mathcal{B}_\varepsilon(w, w) = \varepsilon \|\nabla w\|_{L^2(\Omega)}^2 + c_0 \|w\|_{L^2(\Omega)}^2$ and thus $w = 0$, which is a contradiction.

Remark 5.11. The idea of exploiting the Sobolev embedding multiple times iteratively is often referred to as a bootstrapping argument. These types of arguments are commonly used to obtain improved integrability or regularity in the context of elliptic partial differential equations. One example is the proof of Lemma 9.16 in [75]. There the same argument is used locally to improve a previously obtained L^q -estimate (cf., [75, Theorem 9.13]) to an $L^{q'}$ -estimate. The fact that we can apply this argument globally, i.e., on the whole domain Ω , in our case is due to the

regularity assumption on the domain (Assumption 5.1). For sufficiently smooth boundaries (e.g., C^1 or $C^{1,1}$), local estimates can often be turned into global estimates by combining them with boundary estimates (see, e.g., [75, Chapter 6] in the context of Schauder estimates).

Remark 5.12 (The case $q > 2$). Theorem 5.4 can be analogously proven for $q > 2$. Indeed, instead of proving the inf-sup condition on the adjoint, we can prove the inf-sup condition directly. The steps of the proof in Sections 5.3.1 and 5.3.2 are identical with q and q' swapped and the constant C_1 has to be adjusted slightly to account for the fact that the advection term is not symmetric. To imitate the argument in Section 5.3.3, note that this time $\tilde{J}_q(w) \in W^{1,q'}(\Omega)$ for any $w \in W^{1,q}(\Omega)$, since $q > 2$ by the same argument as before and we can thus again apply (5.13) to obtain

$$\begin{aligned} - \int_{\Omega} w \nabla \cdot (\mathbf{b} \tilde{J}_q(w)) \, d\mathbf{x} + \int_{\Omega} c \tilde{J}_q(w) w \, d\mathbf{x} &= \int_{\Omega} (\mathbf{b} \cdot \nabla w) \tilde{J}_q(w) \, d\mathbf{x} + \int_{\Omega} c \tilde{J}_{q'}(w) w \, d\mathbf{x} \\ &= \|w\|_{L^q(\Omega)}^{2-q} \int_{\Omega} \left(c|w|^q - \frac{1}{q} (\nabla \cdot \mathbf{b}) |w|^q + \frac{1}{q} \nabla \cdot (\mathbf{b} |w|^q) \right) \, d\mathbf{x} \\ &\geq c_0 \|w\|_{L^q(\Omega)}^2 + \frac{\|w\|_{L^q(\Omega)}^{2-q}}{q} \int_{\partial\Omega} (\mathbf{b} \cdot \mathbf{n}) |w|^q \, ds = c_0 \|w\|_{L^q(\Omega)}^2. \end{aligned}$$

To prove the second inf-sup condition, we observe that for $v \in W^{1,q'}(\Omega)$ we have $cv - \nabla \cdot (\mathbf{b}v) \in L^{q'}(\Omega)$ and we can apply the same bootstrap argument as in Section 5.3.5 after swapping q and q' .

5.4 A Finite Element Method in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$

In this section, we consider the (Bubnov-) Galerkin approximation for a (finite-element) space $U_h \subset W_0^{1,\infty}(\Omega)$: find $u_h \in U_h$ such that

$$\mathcal{B}_{\varepsilon}(u_h, v_h) = \langle f, v_h \rangle_{W^{-1,q}(\Omega), W_0^{1,q'}(\Omega)} \quad \forall v_h \in U_h. \quad (5.22)$$

In Section 5.4.1 we quote a well-posedness result presented in [93] and the corresponding additional assumption concerning the stability of the $H_0^1(\Omega)$ -projector. We then consider illustrative numerical examples in Sections 5.4.2 and 5.4.3. To this end, we first consider an example with smooth right-hand side on a square domain with a smooth solution to show optimal convergence rates in the $W^{1,q}(\Omega)$ norm. Secondly, we will illustrate why it is useful to consider a non-Hilbert setting for the convection-diffusion equation by considering examples in two and three

dimensions with right-hand sides with very low regularity. Note that Assumptions 5.1 and 5.13 are satisfied since all domains are convex Lipschitz domains.

5.4.1 Stability of Galerkin's Method

It is well-known that the stability of the (Bubnov)-Galerkin approximation (5.22) with $q = 2$ degenerates as $\varepsilon \rightarrow 0$, cf., Section 2. Similarly, for $q \neq 2$, stability is only proven in the diffusion dominated case in [93]. The proof in [93] relies on the assumption that the $H_0^1(\Omega)$ -projector onto the chosen h -parametrized family of finite-dimensional subspaces is stable in the following sense.

Assumption 5.13 (Stability of $H_0^1(\Omega)$ -projector). *Let $q' > 2$ and $U_h \subset W_0^{1,\infty}(\Omega)$ be a finite-dimensional subspace. The $H_0^1(\Omega)$ -projector $P_h : H_0^1(\Omega) \rightarrow U_h$ is uniformly stable in $W^{1,q'}(\Omega)$, i.e., there is a stability constant $C_P \geq 1$, independent of h , such that for all $z \in W_0^{1,q'}(\Omega) \subset H_0^1(\Omega)$, the unique $H_0^1(\Omega)$ -projection $P_h z \in U_h$ that solves the discrete problem*

$$\int_{\Omega} \nabla v_h \cdot \nabla (P_h z) \, d\mathbf{x} = \int_{\Omega} \nabla v_h \cdot \nabla z \, d\mathbf{x} \quad \forall v_h \in U_h,$$

satisfies the a priori bound:

$$|P_h z|_{W^{1,q'}(\Omega)} \leq C_P |z|_{W^{1,q'}(\Omega)}.$$

The validity of Assumption 5.13 has been proven in the following settings: If Ω is a bounded interval in $\mathbb{R}$, a convex polygonal domain in $\mathbb{R}^2$, or a convex polyhedral domain in $\mathbb{R}^3$, Assumption 5.13 is known to hold for all $q' > 2$ for standard finite element spaces on quasi-uniform meshes, see [145, 87] and [20, Chapter 8], and certain graded meshes, see [61, 125]. Furthermore, the assumption is valid for all $q' > 2$ if U_h is a spectral space of d-variate polynomials of fixed degree on a (finite union of) star-shaped domain(s), see [62].

Under the above assumption the following result was established in [93].

Theorem 5.14. *Let $1 < q < 2$, $\mathbf{b} \in L^\infty(\Omega)^d$, $\nabla \cdot \mathbf{b} \in L^\infty(\Omega)$ and $c \in L^\infty(\Omega)$. If Assumption 5.1 and Assumption 5.13 hold, then the following discrete inf-sup condition holds true:*

$$\sup_{v_h \in U_h} \frac{\mathcal{B}_\varepsilon(w_h, v_h)}{|v_h|_{W^{1,q'}(\Omega)}} \geq \hat{\gamma} |w_h|_{W^{1,q}(\Omega)} \quad \text{for all } w_h \in U_h,$$

where

$$\hat{\gamma} \geq \left(\varepsilon C_P^{-1} C_{q',\Omega}^{-1} - C_{F,q'} (\|\mathbf{b}\|_\infty + C_{F,q} \|c\|_\infty) \right), \quad (5.23)$$

with $C_{F,q}$ and $C_{F,q'}$ the Poincaré-Friedrichs constants for $W_0^{1,q}(\Omega)$ and $W_0^{1,q'}(\Omega)$, respectively (see (5.16)).

Note that it can be guaranteed that the constant $\hat{\gamma}$ in (5.23) is strictly positive. Indeed, this is true if the convection-diffusion-reaction problem is sufficiently *diffusion-dominated*, i.e., the advection and the reaction coefficients are sufficiently small compared to the diffusion parameter ε . From this perspective, Theorem 5.14 is an elementary result: It is reasonable to believe that (5.23) is a suboptimal result, and to conjecture that the discrete inf-sup condition holds true for any choice of the parameters (although not robustly).

Recall from (5.4a) that $\|\cdot\|_U$ is a norm on $W_0^{1,q}(\Omega)$ equivalent to, e.g., $|\cdot|_{W^{1,q}(\Omega)}$. By standard arguments (see, e.g., [69, Chapter 2]), Theorem 5.14 implies that in the diffusion-dominated case, Galerkin's method is a stable and convergent method in the $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ -setting. In particular, the following a priori bound holds

$$\|u_h\|_U \leq \frac{1}{\hat{\gamma}} \|f\|_{W^{-1,q}(\Omega)}, \quad (5.24)$$

as well as the following a priori error estimate

$$\|u - u_h\|_U \leq \left(1 + \frac{M}{\hat{\gamma}}\right) \inf_{w_h \in U_h} \|u - w_h\|_U.$$

This last result can be sharpened, by invoking the error estimate due to Stern [158]:

$$\|u - u_h\|_U \leq \min \left\{ 2^{\lfloor \frac{2}{q} - 1 \rfloor} \frac{M}{\hat{\gamma}}, 1 + \frac{M}{\hat{\gamma}} \right\} \inf_{w_h \in U_h} \|u - w_h\|_U.$$

To illustrate the results for Galerkin's method, we verify the bound in (5.24) by performing several numerical experiments in two- and three-dimensional domains on convection-diffusion problems, for which the solution is not in $H_0^1(\Omega)$ but in $W_0^{1,r}(\Omega)$ for suitable $r < 2$. These numerical experiments show that Galerkin's method is indeed stable for the anticipated values $q \leq r$.

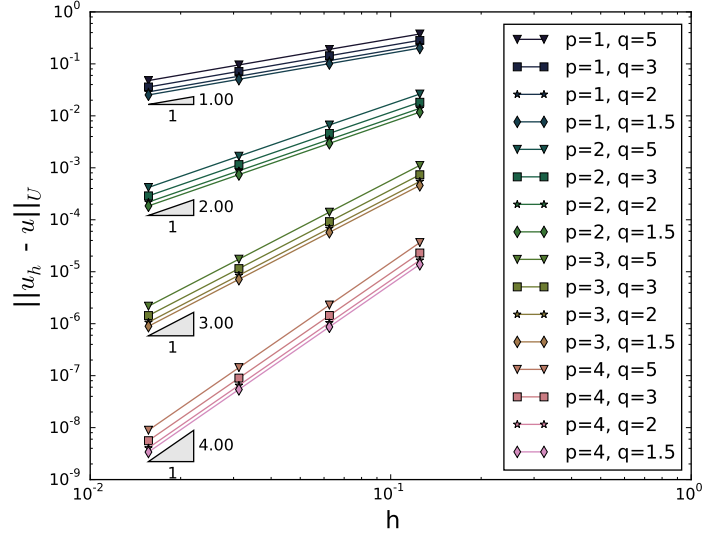


Figure 5.1 Eriksson-Johnson model problem with $\varepsilon = 1$: error in the $W^{1,q}(\Omega)$ -norm for $q = 1.5, 2, 3, 5$, with the polynomial degree chosen uniformly as $p = 1, 2, 3, 4$, on a uniform triangulation of the domain.

5.4.2 Convergence Rates for a Simple Example in Two Dimensions

In order to illustrate the quasi-optimality estimates given at the end of Section 5.1.3, we consider the following simple example (which is essentially the Eriksson–Johnson test case, but with a reaction term):

$$\begin{aligned} \frac{\partial u}{\partial x} - \varepsilon \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + u &= f(x, y) \quad \text{in } (0, 1)^2, \\ u &= 0 \text{ on } x = 1, y = 0, 1, \quad u = \sin(\pi y) \text{ on } x = 0, \end{aligned}$$

where

$$f(x, y) = \frac{\exp(r_1(x-1)) - \exp(r_2(x-1))}{\exp(-r_1) - \exp(-r_2)} \sin(\pi y),$$

with

$$r_1 = \frac{1 + \sqrt{1 + 4\pi^2\varepsilon^2}}{2\varepsilon}, \quad r_2 = \frac{1 - \sqrt{1 + 4\pi^2\varepsilon^2}}{2\varepsilon}.$$

Note that the exact solution of this problem is also given by $f(x, y)$. Figure 5.1 shows the error in the $W^{1,q}(\Omega)$ -norm for $q = 1.5, 2, 3, 5$, $\varepsilon = 1$, and with the polynomial degree chosen uniformly as $p = 1, 2, 3, 4$, on a uniform triangulation of

the domain. Here, we observe the optimal rate of convergence $\mathcal{O}(h^p)$, as the mesh is uniformly refined for each fixed p . This illustrates that in the diffusion-dominated case and sufficiently smooth right-hand side the underlying finite element method performs similarly to the case when $q = 2$.

5.4.3 Examples with Rough Right-hand Sides

To motivate looking at the $W^{1,q}(\Omega)$ - $W^{1,q'}(\Omega)$ -setting instead of the standard $H_0^1(\Omega)$ -setting, we consider examples with rough right-hand sides, viz. Dirac delta distributions, both in two and three dimensions.

To this end, consider the problem

$$\begin{aligned} -\Delta u + \partial_{x_1} u + u &= \delta_0 \quad \text{in } \Omega = [-0.5, 0.5]^d, \quad d = 2, 3, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

For the weak formulation to be well-defined, we require that $\delta_0 \in [W^{1,q'}(\Omega)]'$. In other words, we require $W^{1,q'}(\Omega) \subset C^0(\Omega)$. According to the Sobolev embedding result (see, e.g., [6]), we have for any bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ that $W^{1,q'}(\Omega) \subset C^0(\Omega)$ if $q' > d$. Thus we require $q' > 2$ for $d = 2$ and $q' > 3$ for $d = 3$ or, equivalently, $q < 2$ for $d = 2$ and $q < 1.5$ for $d = 3$. Furthermore, the solution u is not in $W^{1,q}(\Omega)$ for $q \geq 2$ in two dimensions and $q \geq 1.5$ in three dimensions.

Indeed, it is well-known, see, e.g., [70], that the fundamental solution of the Poisson problem contains a singularity of the form $1/|x|$ in three dimensions and $\ln(|x|)$ in two dimensions. Thus the fundamental solution is not contained in $W^{1,q}(\Omega)$ for $q \geq 1.5$ in three dimensions and for $q \geq 2$ in two dimensions. The same applies to the convection-diffusion equation; near the origin the singularity of the fundamental solution behaves like $\mathcal{O}(1/|x|)$ for $d = 3$ and like $\mathcal{O}(\ln(|x|))$ for $d = 2$, cf. [72, 68, 114].

We can thus only expect convergence of $|u_h|_{W^{1,q}(\Omega)}$ to a finite value if $q < 2$ for $\Omega \subset \mathbb{R}^2$ and if $q < 1.5$ for $\Omega \subset \mathbb{R}^3$ and should observe divergence otherwise. This is illustrated for the 2D-case in Figure 5.2 and for the 3D-case in Figure 5.3. These figures plot $|u_h|_{W^{1,q}(\Omega)}$ for the finite element method using linear finite elements on a mesh (triangles in 2D, or tetrahedra in 3D) of mesh size h . In two dimensions, we can observe that $|u_h|_{W^{1,q}(\Omega)}$ diverges for $q = 2, 3, 5$. For $q = 2$, i.e., the borderline case, divergence is very slow; the values converge for $q = 1.5$. Similarly, in three dimensions, we clearly observe divergence for $q = 1.5, 1.7, 2$, while the values converge for $q = 1.3$.

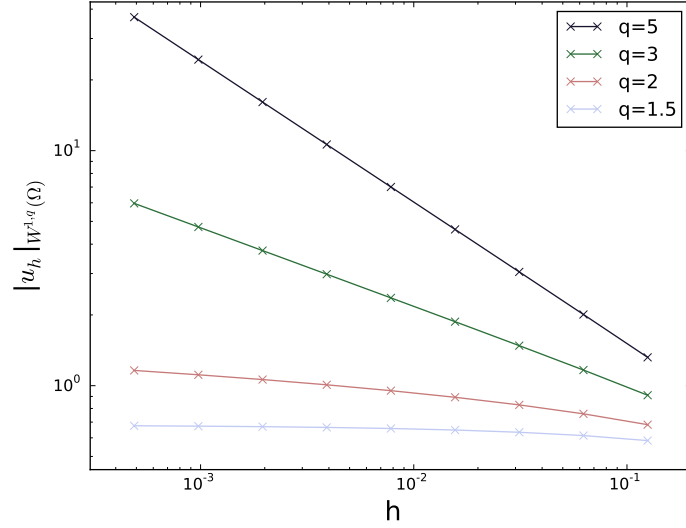


Figure 5.2 Two-dimensional test case: Convergence ($q = 1.5$) and divergence ($q = 2, 3, 5$) of $|u_h|_{W^{1,q}(\Omega)}$ for linear finite-element approximations u_h . The irregular exact solution u is not in $W_0^{1,r}(\Omega)$ for any $r \geq 2$.

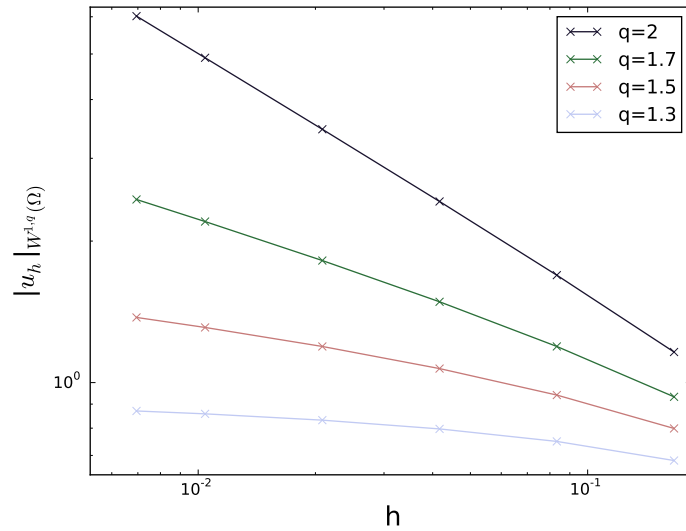


Figure 5.3 Three-dimensional test case: Convergence ($q = 1.3$) and divergence ($q = 1.5, 1.7, 2$) of $|u_h|_{W^{1,q}(\Omega)}$ for linear finite-element approximations u_h . The irregular exact solution u is not in $W_0^{1,r}(\Omega)$ for any $r \geq 1.5$.

Chapter 6

Minimum Residual Methods in Banach Spaces

The key observation that makes it possible to extend the DPG methodology to general Banach spaces is the equivalence with a residual minimization problem, cf., [39] and [55] for the DPG method. In this chapter, we will see that, in the context of Banach spaces, the Riesz map can be replaced by a duality mapping, cf., Section 3.3, but due to the non-linearity of the duality mapping we lose the concept of an optimal test space which is used in the context of DPG methods, cf., [59]. The extension to Banach spaces was first introduced in [136] and we will repeat the main concepts in this chapter. We will furthermore describe the DPG framework in the classical Hilbert space setting based on [34] and the very concise summary of the method in [55]. For more details we refer to [79, 60, 55]. As mentioned in the introduction to this thesis, the method for the convection-diffusion-reaction equation that we will consider in Chapter 7 is a generalization of the scheme discussed in [39]. Conceptually, this method is closely related to DPG methods with the main difference that only continuous finite element spaces are considered in [39].

This chapter is structured as follows: we start by introducing the ideal Petrov-Galerkin method in the Hilbert space setting in Section 6.1, as well as its equivalent formulation as a residual minimization problem. We then continue with the residual minimization problem in a more general Banach space setting in Section 6.2; in Section 6.3 we discuss the concept of the optimal test norms both in Hilbert and in Banach spaces. Then, in Section 6.4 we introduce a fully discrete or *practical* version of the ideal Petrov-Galerkin method in Section 6.1 and the residual minimization method in Section 6.2. Section 6.5 focusses on the choice

of broken test spaces in the practical Petrov-Galerkin approximation. Finally, we conclude in Section 6.6 by presenting an overview of Petrov-Galerkin methods applied to the convection-diffusion-reaction equation that have been developed within this framework.

6.1 An Ideal Petrov-Galerkin Method

Let U and V be two Hilbert spaces and $b : U \times V \rightarrow \mathbb{R}$ a continuous bilinear form that satisfies the following inf-sup conditions:

$$\inf_{w \in U} \sup_{v \in V} \frac{b(w, v)}{\|w\|_U \|v\|_V} = \gamma > 0, \quad (6.1a)$$

$$\{v \in V : b(w, v) = 0, \forall w \in U\} = \{0\}. \quad (6.1b)$$

For a given right-hand side $\ell \in V'$, we consider the problem: find $u \in U$ such that

$$b(u, v) = \ell(v) \quad \forall v \in V.$$

Introducing two finite dimensional subspaces $U_n \subset U$ and $V_n \subset V$ with $\dim U_n = \dim V_n$, we obtain the finite dimensional problem: find $u_n \in U_n$ such that

$$b(u_n, v_n) = \ell(v_n) \quad \forall v_n \in V_n,$$

which is well-posed if b satisfies the discrete inf-sup condition,

$$\inf_{w_n \in U_n} \sup_{v_n \in V_n} \frac{b(w_n, v_n)}{\|w_n\|_{U_n} \|v_n\|_{V_n}} \geq \gamma_n > 0, \quad (6.2)$$

cf., [9, 142].

The main idea of the DPG method is to choose a test space $V_n \subset V$ for a given trial space $U_n \subset U$ that guarantees that the continuous inf-sup condition (6.1a) implies its discrete version (6.2). This is obviously not true in general since the supremum is taken over a smaller space in the discrete inf-sup condition. However, if we choose the test space in such a way that it contains test functions that, for a given $u_n \in U_n$, realize the supremum taken over the whole space V , the discrete inf-sup condition follows immediately.

Indeed, if we define $B : U \rightarrow V'$ as the linear operator corresponding to b , i.e., $\langle Bu, v \rangle_{V', V} = b(u, v)$, $u \in U$, $v \in V$, and let $R_V : V \rightarrow V'$ denote the Riesz map

corresponding to the inner product on V , we can define $V_n = R_V^{-1}BU_n$ and obtain

$$\begin{aligned} \gamma \|u_n\|_U &\leq \sup_{v \in V} \frac{b(u_n, v)}{\|v\|_V} = \|Bu_n\|_{V'} = \|R_V^{-1}Bu_n\|_V \\ &= \frac{(R_V^{-1}Bu_n, R_V^{-1}Bu_n)}{\|R_V^{-1}Bu_n\|_V} = \frac{\langle Bu_n, R_V^{-1}Bu_n \rangle_{V', V}}{\|R_V^{-1}Bu_n\|_V} \\ &= \frac{b(u_n, R_V^{-1}Bu_n)}{\|R_V^{-1}Bu_n\|_V} = \sup_{v_n \in V_n} \frac{b(u_n, v_n)}{\|v_n\|_V}. \end{aligned}$$

In this sense $V_n = R_V^{-1}BU_n$ is the *optimal* test space. We can also recover V_n by starting with a residual minimization problem, i.e.,

$$\begin{aligned} u_n &= \arg \min_{w_n \in U_n} H_V(w_n), \\ H_V(w_n) &:= \frac{1}{2} \|Bw_n - \ell\|_{V'}^2 = \frac{1}{2} \|R_V^{-1}(Bw_n - \ell)\|_V^2. \end{aligned} \tag{6.3}$$

It can be shown that this minimization problem is equivalent to $\nabla H_V(u_n) = 0$,

$$(R_V^{-1}(Bu_n - \ell), R_V^{-1}Bw_n) = 0 \quad \forall w_n \in U_n. \tag{6.4}$$

We thus obtain the Petrov-Galerkin scheme with optimal test functions by rewriting the above as

$$\langle Bu_n, v_n \rangle_{V', V} = \ell(v), \quad v_n := R_V^{-1}w_n, \quad w_n \in U_n. \tag{6.5}$$

Alternatively, we define $r := R_V^{-1}(\ell - Bu_n)$ and rewrite (6.4) as a mixed problem: find $(r, u_n) \in V \times U_n$ such that

$$(r, v)_V + b(u_n, v) = \ell(v) \quad v \in V, \tag{6.6a}$$

$$b(w_n, r) = 0 \quad w_n \in U_n. \tag{6.6b}$$

In Section 6.4, we will see that this formulation can be used to implement the practical method both in Banach and Hilbert spaces.

6.2 Residual Minimization in Banach Spaces

In this section, we consider the residual minimization approach in more general Banach spaces and demonstrate how this can be used to extend the ideal Petrov-Galerkin method to Banach spaces. This idea was first introduced in [136]. In this

context, a duality mapping replaces the Riesz map and we lose the concept of an ‘optimal’ test space since this is a non-linear map unless the problem is set in Hilbert spaces. Note that a best approximation problem can be formulated in any Banach space. However, we need strict convexity for uniqueness of minimizers, cf., e.g., [156, 136]. The closed range theorem is still applicable in general Banach spaces and provides conditions for well-posedness of linear problems. Its reformulation in terms of inf-sup conditions only requires reflexivity of the test space to describe the kernel of B' by (6.1b).

If we now assume that U and V are Banach spaces, we can again consider a residual minimization problem similar to (6.3),

$$u_n = \arg \min_{w_n \in U_n} H_V(w_n),$$

$$H_V(w_n) := \psi(\|Bw_n - \ell\|_{V'}),$$

where $\psi(s) = \int_0^s \varphi(t) dt$ and φ is a weight function as defined in Section 3.3. A typical choice would be $\psi(t) = t^q/q$, $1 < q < \infty$. If V' is strictly convex, we know that the duality mapping $J_{V'}^\varphi$ is single valued and $J_{V'}^\varphi = \nabla(v' \mapsto \psi(\|v'\|_{V'}))$.

In this setting, i.e., if V' is strictly convex, it can be shown that the residual minimization problem is equivalent to $\nabla H_V(u_n) = J_{V'}^\varphi(\ell - Bu_n) = 0$. If V is additionally reflexive, as well as being strictly convex, we obtain the following non-linear Petrov Galerkin formulation: find $u_n \in U_n$ such that

$$\begin{aligned} \langle J_{V'}^\varphi(\ell - Bu_n), Bw_n \rangle_{V'', V} \\ = \left\langle Bw_n, \left(J_V^{\varphi^{-1}} \right)^{-1} (\ell - Bu_n) \right\rangle_{V', V} = 0 \quad \forall w_n \in U_n. \end{aligned} \quad (6.7)$$

Here, we used $J_{V'}^\varphi = \left(J_V^{\varphi^{-1}} \right)^{-1}$ by means of canonical identification, cf., Proposition 3.24. Introducing an auxiliary variable $r = \left(J_V^{\varphi^{-1}} \right)^{-1} (\ell - Bu_n)$, this can be reformulated as a mixed method: find $u_n \in U_n$ and $r \in V$ such that

$$\langle J_V^{\varphi^{-1}}(r), v \rangle_{V', V} + \langle Bu_n, v \rangle_{V', V} = \langle \ell, v \rangle_{V', V} \quad \forall v \in V, \quad (6.8a)$$

$$\langle Bw_n, r \rangle_{V', V} = 0 \quad \forall w_n \in U_n. \quad (6.8b)$$

Note that $J_V^{\varphi^{-1}}$ is a non-linear map unless V is a Hilbert space. If V is a Hilbert space and we select $\psi(t) = t^2/2$, we have that $J_V^{\varphi^{-1}} = R_V$ and recover the mixed method (6.6).

6.3 The Optimal Test Norm

In applications, we are often more interested in minimizing the error in the approximation with respect to the norm on U rather than minimizing the residual. In other words, we usually want to choose the norm on V in such a way that we ultimately control the error $\|u - u_n\|_U$. With this in mind, employing the inf-sup condition (6.1a) and continuity of B , we deduce that

$$\|u - u_n\|_U \leq \frac{1}{\gamma} \sup_{v \in V} \frac{b(u - u_n, v)}{\|v\|_V} = \frac{1}{\gamma} \|\ell - Bu_n\|_{V'} \leq \frac{M}{\gamma} \|u - u_n\|_U,$$

where M is the continuity constant of the bilinear form b . Hence, in order to control $\|u - u_n\|_U$, we require γ and M as close to one as possible and independent of certain problem specific parameters (the scaling of the diffusion term, for example, in case of the convection-diffusion-reaction equation). The *optimal* test norm is a concept introduced in the context of DPG methods [170], but unlike the concept of optimal test functions and spaces it can easily be extended to Banach spaces. The optimal test norm is defined as the norm on V such that $M = \gamma = 1$ and is given by

$$\|v\|_{\text{opt}} := \sup_{u \in U} \frac{b(u, v)}{\|u\|_U} = \|B'v\|_{U'}.$$

Indeed,

$$\sup_{v \in V} \frac{b(w, v)}{\|v\|_{\text{opt}}} = \sup_{v \in V} \frac{(B'v)(w)}{\|B'v\|_{U'}} = \sup_{g \in U'} \frac{g(w)}{\|g\|_{U'}} = \|w\|_U.$$

Conversely, we have the optimal constants $M = \gamma = 1$ for any given test norm, if we endow U with the so-called energy norm,

$$\|u\|_E := \sup_{v \in V} \frac{b(u, v)}{\|v\|_V}.$$

Ideally, we want to select the norm on U in order to approximate the solution in the desired norm and then work with the optimal test-norm. However, the optimal test-norm is a dual norm which is, in general, not computable. Thus, we have to replace the optimal test-norm with an equivalent norm that is computable. The difficulty here is to obtain equivalence constants, i.e., constants γ and M , that are independent of problem parameters.

In the context of the convection-diffusion-reaction equation, Broersen and Stevenson work directly with the optimal test norm in [23], whereas the analysis presented in [56] and [40] relies on robust estimates for γ and M by looking at the adjoint problem. Extending any of the robust estimates from Hilbert subspaces of $L^2(\Omega)$ to Banach spaces $L^q(\Omega)$, $1 < q < \infty$, is highly non-trivial and to the best of our knowledge remains an open problem.

6.4 The Inexact Method

So far, we have not yet obtained a method that can be implemented. The mixed methods (6.6) and (6.8) are still infinite dimensional problems since they rely on the whole test space V . In the Petrov-Galerkin formulation (6.5) this is hidden in the inversion of the Riesz-map R_V . This can in general only be done approximately.

The inversion of the Riesz operator corresponds to solving the following auxiliary variational problem: given $w_n \in U_n$, find $v \in V$ such that

$$(v, z)_V = b(w_n, z) \quad \forall z \in V.$$

Choosing a finite dimensional subspace $V_m \subset V$ with $\dim V_m \geq \dim U_n$, we can approximate the solution of the above problem by solving the following finite dimensional problem: find $v_m \in V_m$ such that

$$(v_m, z_m)_V = b(w_n, z_m) \quad \forall z_m \in V_m.$$

Again, we can equivalently consider a mixed formulation of the method and then generalize it to Banach spaces. More precisely, we obtain the following fully discrete mixed problem: find $(r_m, u_n) \in V_m \times U_n$ such that

$$\langle J_V^{\varphi^{-1}}(r_m), v_m \rangle_{V', V} + \langle B u_n, v_m \rangle_{V', V} = \langle \ell, v_m \rangle_{V', V} \quad \forall v_m \in V_m, \quad (6.9a)$$

$$\langle B w_n, r_m \rangle_{V', V} = 0 \quad \forall w_n \in U_n, \quad (6.9b)$$

where $J_V^{\varphi^{-1}} = R_V$ if V is a Hilbert space.

Replacing V by a finite-dimensional subspace we obviously lose the best approximation properties. The question of quantifying this ‘loss’ was addressed in [80] for Hilbert spaces by introducing the concept of a Fortin-operator and was later extended to Banach spaces in [136]. For well-posedness of the inexact method

we require the existence of a bounded projection operator $\Pi : V \rightarrow V_m$ such that

$$\langle Bw_n, v - \Pi v \rangle_{V', V} = 0 \quad \forall w_n \in U_n, v \in V. \quad (6.10)$$

The norm of Π then enters into the a priori estimate and quantifies the ‘loss’ due to the inexactness, i.e., we can obtain an a priori estimate of the form

$$\|u - u_n\| \leq C \inf_{w_n \in U_n} \|u - w_n\|, \quad C = \frac{(\|\Pi\| + \|I - \Pi\|)M}{\gamma}.$$

Remark 6.1. The constant C can be improved both in the Hilbert setting, where we can obtain $C = \|\Pi\|M/\gamma$, cf., [79], and in the Banach setting, where we would have to introduce geometric constants for the improved estimate, cf., [136].

6.5 Breaking Spaces and Forms

So far we have only considered an abstract bilinear form defined on two given Banach or Hilbert spaces without specifying the type of spaces we typically encounter. Another key concept of the DPG method is to consider a broken test space V , e.g., the space of functions that are in $H^1(\kappa)$ restricted to each element $\kappa \in \Omega_h$, where Ω_h is a subdivision of the domain Ω into a finite number of elements κ such that $\overline{\Omega} = \bigcup_{\kappa \in \Omega_h} \bar{\kappa}$. The idea is that if the inner product can be localized to elements, the inversion of the Riesz map can be done elementwise. This step justifies the ‘D’ in the name of the method. Note that in contrast to DGFEMs, the ‘D’ does not refer to approximating the solution in a discontinuous trial space, but only to choosing a discontinuous test space. This section is based on [34] and we will see under which circumstances the broken formulation inherits the well-posedness from its unbroken analogue. We mainly consider the Hilbert space setting, but we will comment on the extension to Banach spaces — or more precisely $L^q(\Omega)$ -type Sobolev spaces — whenever possible.

6.5.1 Breaking Spaces

Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$ be open and bounded with Lipschitz boundary. Denote by Ω_h a disjoint partitioning of Ω into open elements κ with Lipschitz boundaries $\partial\kappa$ such that $\overline{\Omega} = \bigcup_{\kappa \in \Omega_h} \bar{\kappa}$. Furthermore, denote by $\partial\Omega_h$ the collection of element boundaries $\partial\kappa$ for all $\kappa \in \Omega_h$. We define the broken Sobolev spaces $H^1(\Omega_h)$ and

$H(\operatorname{div}, \Omega_h)$ by

$$\begin{aligned} H^1(\Omega_h) &= \{u \in L^2(\Omega) : u|_\kappa \in H^1(\kappa), \kappa \in \Omega_h\} = \prod_{\kappa \in \Omega_h} H^1(\kappa), \\ H(\operatorname{div}, \Omega_h) &= \{\boldsymbol{\sigma} \in L^2(\Omega)^d : \boldsymbol{\sigma}|_\kappa \in H(\operatorname{div}, \kappa), \kappa \in \Omega_h\} = \prod_{\kappa \in \Omega_h} H(\operatorname{div}, \kappa). \end{aligned}$$

On each element we have trace operators $\mathcal{T}_{\operatorname{grad}}^\kappa$ and $\mathcal{T}_{\operatorname{div}}^\kappa$ such that

$$\begin{aligned} \mathcal{T}_{\operatorname{grad}}^\kappa u &= u|_{\partial\kappa} & \forall u \in H^1(\kappa), \\ \mathcal{T}_{\operatorname{div}}^\kappa \boldsymbol{\sigma} &= \boldsymbol{\sigma}|_{\partial\kappa} \cdot \mathbf{n}_\kappa & \forall \boldsymbol{\sigma} \in H(\operatorname{div}, \kappa). \end{aligned}$$

Here $\mathbf{n}_\kappa$ refers to the unit outward normal to $\partial\kappa$.

Remark 6.2. We have to be slightly careful what we mean by this notation, since $\boldsymbol{\sigma}|_{\partial\kappa}$ as such is not well-defined. For more details on trace operators we refer to Section 3.6 of this thesis and [102], for example.

We can use these operators to define continuous linear maps $\mathcal{T}_{\operatorname{grad}}^h : H^1(\Omega_h) \rightarrow \prod_{\kappa \in \Omega_h} H^{1/2}(\partial\kappa)$ and $\mathcal{T}_{\operatorname{div}}^h : H(\operatorname{div}, \Omega_h) \rightarrow \prod_{\kappa \in \Omega_h} H^{-1/2}(\partial\kappa)$ by performing the above operation elementwise. Both maps are continuous and surjective.

These trace operators can now be used to define interface functions. Note that functions in the range of $\mathcal{T}_{\operatorname{grad}}^h$ and $\mathcal{T}_{\operatorname{div}}^h$ are generally multivalued on parts of the interface $f \subset \partial\kappa^+ \cap \partial\kappa^-$ that are shared by two elements $\kappa^\pm \in \Omega_h$. In order to avoid this problem, we define the interface functions as the image of $\mathcal{T}_{\operatorname{grad}}^h|_{H^1(\Omega)}$ and $\mathcal{T}_{\operatorname{div}}^h|_{H(\operatorname{div}, \Omega)}$, i.e.,

$$\begin{aligned} H^{1/2}(\partial\Omega_h) &:= \left\{ \hat{u} \in \prod_{\kappa \in \Omega_h} H^{1/2}(\partial\kappa) : \exists u \in H^1(\Omega) : \mathcal{T}_{\operatorname{grad}}^h u = \hat{u} \right\}, \\ H^{-1/2}(\partial\Omega_h) &:= \left\{ \hat{\sigma} \in \prod_{\kappa \in \Omega_h} H^{-1/2}(\partial\kappa) : \exists \boldsymbol{\sigma} \in H(\operatorname{div}, \Omega) : \mathcal{T}_{\operatorname{div}}^h \boldsymbol{\sigma} = \hat{\sigma} \right\}. \end{aligned}$$

We then equip these spaces with the quotient norms,

$$\begin{aligned} \|\hat{u}\|_{H^{1/2}(\Omega_h)} &:= \inf_{u \in H^1(\Omega) \cap (\mathcal{T}_{\operatorname{grad}}^h)^{-1} \hat{u}} \|u\|_{H^1(\Omega)}, \\ \|\hat{\sigma}\|_{H^{-1/2}(\Omega_h)} &:= \inf_{\boldsymbol{\sigma} \in H(\operatorname{div}, \Omega) \cap (\mathcal{T}_{\operatorname{div}}^h)^{-1} \hat{\sigma}} \|\boldsymbol{\sigma}\|_{H(\operatorname{div}, \Omega)}. \end{aligned}$$

Note that $\ker \mathcal{T}_{\text{grad}}^\Omega = H_0^1(\Omega)$ and $\ker \mathcal{T}_{\text{div}}^\Omega = H_0(\text{div}, \Omega)$, where $H_0(\text{div}, \Omega)$ is the closure of $C_c^\infty(\Omega)$ with respect to the $H(\text{div}, \Omega)$ -norm. With this in mind one can show the following equivalences:

$$v \in H_0^1(\Omega) \Leftrightarrow \langle \hat{\sigma}, v \rangle_h = 0 \quad \forall \hat{\sigma} \in H^{-1/2}(\partial\Omega_h), \quad (6.11)$$

$$\sigma \in H_0(\text{div}, \Omega) \Leftrightarrow \langle \sigma \cdot \mathbf{n}, \hat{u} \rangle_h = 0 \quad \forall \hat{u} \in H^{1/2}(\partial\Omega_h). \quad (6.12)$$

Remark 6.3. So far all concepts can be extended to subspaces of $L^q(\Omega)$, $1 < q < \infty$, instead of $L^2(\Omega)$. We simply have to choose the right subspace of $L^q(\partial\kappa)$ in order to replace $H^{1/2}(\partial\kappa)$. For details regarding the definition of trace operators and spaces in a very general setting we refer to Section 3.6 of this thesis or, for more details, to [102].

Remark 6.4. One very useful property of the two trace spaces $H^{1/2}(\partial\kappa)$ and $H^{-1/2}(\partial\kappa)$ is that they are dual to each other. This in particular implies the duality of the norms, i.e.,

$$\begin{aligned} \|\hat{\sigma}\|_{H^{-1/2}(\partial\kappa)} &= \sup_{u \in H^1(\kappa)} \frac{\langle \hat{\sigma}, \mathcal{T}_{\text{grad}} u \rangle_{H^{-1/2}(\partial\kappa), H^{1/2}(\partial\kappa)}}{\|u\|_{H^1(\kappa)}} \\ &= \sup_{\hat{u} \in H^{1/2}(\partial\kappa)} \frac{\langle \hat{\sigma}, \hat{u} \rangle_{H^{-1/2}(\partial\kappa), H^{1/2}(\partial\kappa)}}{\|\hat{u}\|_{H^{1/2}(\partial\kappa)}}, \\ \|\hat{u}\|_{H^{1/2}(\partial\kappa)} &= \sup_{\sigma \in H(\text{div}, \kappa)} \frac{\langle \mathcal{T}_{\text{div}} \sigma, \hat{u} \rangle_{H^{-1/2}(\partial\kappa), H^{1/2}(\partial\kappa)}}{\|\sigma\|_{H(\text{div}, \kappa)}} \\ &= \sup_{\hat{\sigma} \in H^{-1/2}(\partial\kappa)} \frac{\langle \hat{\sigma}, \hat{u} \rangle_{H^{-1/2}(\partial\kappa), H^{1/2}(\partial\kappa)}}{\|\hat{\sigma}\|_{H^{-1/2}(\partial\kappa)}}. \end{aligned}$$

Moreover, even though $H^{1/2}(\partial\Omega_h)$ and $H^{-1/2}(\partial\Omega_h)$ are no longer dual to each other in general, they inherit the duality of the norms in the following sense

$$\|\hat{\sigma}\|_{H^{-1/2}(\partial\Omega_h)} = \sup_{u \in H^1(\Omega_h)} \frac{\langle \hat{\sigma}, \mathcal{T}_{\text{grad}}^h u \rangle_{H^{-1/2}(\partial\Omega_h), H^{1/2}(\partial\Omega_h)}}{\|u\|_{H^1(\Omega_h)}}, \quad (6.13)$$

$$\|\hat{u}\|_{H^{1/2}(\partial\Omega_h)} = \sup_{\sigma \in H(\text{div}, \Omega_h)} \frac{\langle \mathcal{T}_{\text{div}}^h \sigma, \hat{u} \rangle_{H^{-1/2}(\partial\Omega_h), H^{1/2}(\partial\Omega_h)}}{\|\sigma\|_{H(\text{div}, \Omega_h)}}. \quad (6.14)$$

In order to extend this to $q \neq 1$, consider the trace operators $\mathcal{T}_{\text{grad}}^{q'} : W^{1, q'}(\Omega) \rightarrow B^{q', q', 1-1/q'}(\partial\Omega)$ and $\mathcal{T}_{\text{div}}^q : W^q(\text{div}, \Omega) \rightarrow B^{q, q, -1/q}(\partial\Omega)$. Now, $\mathcal{T}_{\text{grad}}^{q'}$ surjectively maps onto $B^{q', q', 1-1/q'}(\partial\Omega)$ for any $1 < q' < \infty$, cf., e.g., [102, Example 6]. Since $B^{q, q, -1/q}(\partial\Omega)$ is the dual of $B^{q', q', 1-1/q'}(\partial\Omega)$, we require $\mathcal{T}_{\text{div}}^q$ to be surjective onto

$B^{q,q,-1/q}(\partial\Omega)$ in order for the two trace spaces to be dual to each other. The proof of the required result for $q = 2$ follows from the existence of a solution to the Neumann problem for Laplace's equation, cf., [50, Theorem 1, Chapter IX]. In general Lipschitz domains, the Neumann problem is solvable for boundary data in $L^q(\Omega)$ if $q \in (1, q_0)$ for some $q_0 > 2$, where the exact value of q_0 depends on the regularity of the boundary. This result is sharp, cf., [111, 112]. However, if Ω is a bounded convex or semiconvex Lipschitz domain, it can be shown that $q_0 = \infty$, cf., [168].

6.5.2 Breaking Variational Forms

This section is based on Section 3 in [34]. We will see that none of the concepts rely on Hilbert spaces and we can thus state everything in Banach spaces. Let U_0 and V denote two Banach spaces; let V be reflexive and let V_0 be a closed subspace of V . In all the applications V will be a broken space and V_0 its unbroken analogue. Furthermore, assume that the norm on V_0 is *localizable*, i.e., we can define a norm on V such that the following norm equivalence holds:

$$c_1 \|v\|_V \leq \|v\|_{V_0} \leq c_2 \|v\|_V \quad \forall v \in V_0. \quad (6.15)$$

We consider a continuous bilinear form¹ $b_0 : U_0 \times V \rightarrow \mathbb{R}$ with continuity constant $\|b_0\|$ satisfying the inf-sup condition

$$c_0 \|w\|_{U_0} \leq \sup_{v \in V_0} \frac{|b_0(w, v)|}{\|v\|_{V_0}} \leq \sup_{v \in V_0} \frac{|b_0(w, v)|}{c_1 \|v\|_V} \quad \forall w \in U_0, \quad (6.16)$$

where the second inequality is a direct consequence of the first due to the localizability of the norm.

If furthermore,

$$Z_0 = \{v \in V_0 : b_0(w, v) = 0 \text{ for all } w \in U_0\} = \{0\}, \quad (6.17)$$

we obtain well-posedness of the problem: find $u \in U_0$ such that

$$b_0(u, v) = \ell(v) \quad \forall v \in V_0,$$

¹To be more precise, we usually consider a bilinear form $b_0 : U_0 \times V_0 \rightarrow \mathbb{R}$ and a localized version $b_0^h : U_0 \times V \rightarrow \mathbb{R}$ with $b_0^h|_{U_0} = b_0$ which is normally given by an elementwise evaluation of the bilinear form b_0 .

for a given $\ell \in V_0'$. Existence of a solution can still be obtained if $Z_0 \neq \{0\}$ provided that the data ℓ satisfies the compatibility condition $\ell(z) = 0$ for all $w \in Z_0$.

In order to obtain a ‘broken’ version of the bilinear form, we have to introduce another Banach space $\hat{U}$ and a bilinear form $\hat{b} : \hat{U} \times V \rightarrow \mathbb{R}$. With respect to applications, we can think of $\hat{U}$ as an interface-flux space and V a broken Sobolev space.

The bilinear form $b : (U := U_0 \times \hat{U}) \times V \rightarrow \mathbb{R}$,

$$b((w, \hat{w}), v) := b_0(w, v) + \hat{b}(\hat{w}, v),$$

inherits the continuity from the bilinear forms b_0 and $\hat{b}$, if we equip U with the following norm:

$$\|(u, \hat{u})\|_U = \sqrt[q]{\|u\|_{U_0}^q + \|\hat{u}\|_{\hat{U}}^q}, \quad (6.18)$$

where $1 < q < \infty$. We now consider the problem: find $(u, \hat{u}) \in U$ such that

$$b((u, \hat{u}), v) = \ell(v) \quad \forall v \in V.$$

We furthermore assume

$$V_0 = \{v \in V : \hat{b}(\hat{w}, v) = 0 \text{ for all } \hat{w} \in \hat{U}\} \quad (6.19)$$

and there is a positive constant $\hat{c}$ such that

$$\hat{c}\|\hat{w}\|_{\hat{U}} \leq \sup_{v \in V} \frac{\hat{b}(\hat{w}, v)}{\|v\|_V} \quad \forall \hat{w} \in \hat{U}. \quad (6.20)$$

Remark 6.5. Note that the last assumption is often difficult to prove for applications.

Theorem 6.6. *Under the above assumptions, we obtain*

$$c_q \|(w, \hat{w})\| \leq \sup_{v \in V} \frac{|b((w, \hat{w}), v)|}{\|v\|_V},$$

where c_q is defined by

$$\frac{1}{c_q^q} = \frac{1}{(c_0 c_1)^q} + \frac{1}{\hat{c}^q} \left(\frac{\|b_0\|}{c_0 c_1} + 1 \right)^q.$$

Here q is the same q as in (6.18). Moreover, $Z_0 = \{v \in V : b((w, \hat{w}), v) = 0 \quad \forall w \in U_0, \hat{w} \in \hat{U}\}$.

Before we prove the above theorem, let us look at a simple example that has been analysed in detail in [54] and is also included in [34].

Example 6.7 (Poisson problem). *Consider the bilinear form $b_0 : U_0 \times V_0 \rightarrow \mathbb{R}$, where $U_0 = V_0 = H_0^1(\Omega)$,*

$$b_0(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x} = \sum_{\kappa \in \Omega_h} \int_{\kappa} \nabla u \cdot \nabla v \, d\mathbf{x}.$$

We now equip the broken test space $V := H^1(\Omega_h)$ with the norm

$$\|v\|_V = \sqrt{\sum_{\kappa \in \Omega_h} \|v\|_{H^1(\kappa)}^2} = \sqrt{\sum_{\kappa \in \Omega_h} \left(\int_{\kappa} v^2 \, d\mathbf{x} + \int_{\kappa} \nabla v \cdot \nabla v \, d\mathbf{x} \right)}.$$

This is just the $H^1(\Omega)$ -norm evaluated elementwise and hence (6.15) is trivially satisfied with constants $c_1 = c_2 = 1$. Note that b_0 is well defined on all of V if we choose the second representation. Coercivity of the bilinear form b_0 implies both (6.16) and (6.17). Furthermore, we define $\hat{U} := H^{-1/2}(\partial\Omega_h)$ and

$$\hat{b}(\hat{u}, v) = -\langle \hat{u}, v \rangle_{\partial\Omega_h}.$$

Then, (6.11) implies (6.19) and (6.13) implies (6.20) with $\hat{c} = 1$.

Proof of Theorem 6.6. The proof in [34] can easily be extended to Banach spaces. To make this obvious, we repeat the details here. First, we have to bound $\|w\|_{U_0}$ and $\|\hat{w}\|_{\hat{U}}$. The inf-sup conditions (6.16) and (6.17) imply

$$\begin{aligned} c_0 c_1 \|w\|_{U_0} &\leq \sup_{v \in V_0} \frac{|b_0(w, v)|}{\|v\|_V} \\ &\leq \sup_{v \in V_0} \frac{|b_0(w, v) + \hat{b}(\hat{w}, v)|}{\|v\|_V} \\ &\leq \sup_{v \in V} \frac{|b((w, \hat{w}), v)|}{\|v\|_V}, \end{aligned}$$

where we used $V_0 \subset V$ in the last step. Next, (6.20) implies

$$\begin{aligned} \hat{c} \|\hat{w}\|_{\hat{U}} &\leq \sup_{v \in V} \frac{|\hat{b}(\hat{w}, v)|}{\|v\|_V} \leq \sup_{v \in V} \frac{|b((w, \hat{w}), v) - b_0(w, v)|}{\|v\|_V} \\ &\leq \|b_0\| \|w\|_{U_0} + \sup_{v \in V} \frac{|b((w, \hat{w}), v)|}{\|v\|_V}. \end{aligned}$$

Now we can combine both estimates and obtain

$$\begin{aligned} \|(w, \hat{w})\|_U^q &= \|w\|_{U_0}^q + \|\hat{w}\|_{\hat{U}}^q \\ &\leq \left(\frac{1}{c_0 c_1} \sup_{v \in V} \frac{|b((w, \hat{w}), v)|}{\|v\|_V} \right)^q + \left(\frac{1}{\hat{c}} \left(\frac{\|b_0\|}{c_1 c_0} + 1 \right) \sup_{v \in V} \frac{|b((w, \hat{w}), v)|}{\|v\|_V} \right)^q. \end{aligned}$$

Finally, for $Z := \{v \in V : b((w, \hat{w}), v) = 0 \text{ for all } w \in U_0 \text{ and } \hat{w} \in \hat{U}\}$, we have

$$\begin{aligned} v \in Z &\iff b_0(w, v) = 0 \text{ for all } w \in U_0 \text{ and } \hat{b}(\hat{w}, v) = 0 \text{ for all } \hat{w} \in \hat{U} \\ &\iff b_0(w, v) = 0 \text{ for all } w \in U_0 \text{ and } v \in V_0 \iff v \in Z_0. \end{aligned}$$

Here we used the definition of Z_0 as well as (6.19). $\square$

Remark 6.8. If V is defined as some broken space of the form $V = V(\Omega_h) = \prod_{\kappa \in \Omega_h} V(\kappa)$, $\mathcal{T} : V \rightarrow \mathcal{T}V$ is a trace operator with $\ker \mathcal{T} = V_0$, and $\hat{b}(\hat{w}, v) = \langle \hat{w}, \mathcal{T}v \rangle_{(\mathcal{T}V)', \mathcal{T}V}$, then (6.20) is trivially satisfied with $\hat{c} = 1$ if we select $\hat{U} = (\mathcal{T}V)'$ with the dual norm.

6.5.3 Coding the DPG Method

We will now look at some details of the implementation of the DPG method. For more details we refer to [110]. Consider the inexact version of the mixed formulation (6.6): find $(r_m, u_n) \in V_m \times U_n$ such that

$$(r_m, v_m)_V + b(u_n, v_m) = \ell(v_m) \quad v_m \in V_m, \quad (6.21a)$$

$$b(w_n, r_m) = 0 \quad w_n \in U_n, \quad (6.21b)$$

where U_n and V_m are finite element spaces that are chosen such that a Fortin-projector exists. The easiest way to implement the DPG method is to use this formulation and implement it directly. This has one major disadvantage compared with standard finite element methods or even DGFEMs: the introduction of the auxiliary unknown r_m leads to a substantial increase in degrees of freedom.

The mixed problem (6.21) generates a matrix equation of the form:

$$\begin{pmatrix} G & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} R \\ U \end{pmatrix} = \begin{pmatrix} F \\ 0 \end{pmatrix}, \quad (6.22)$$

which yields the following system for U

$$B^T G^{-1} B U = B^T G^{-1} F. \quad (6.23)$$

Solving the above system for U is considerably cheaper than solving (6.22) provided that inverting G is cheap. If V_m is a discontinuous space, the matrix G generated by $(r_m, v_m)_V$ becomes block-diagonal and inverting G is equivalent to solving a series of small, local problems. Implementing (6.23) thus becomes equivalent to computing the approximation of the optimal test functions by solving local problems and then constructing the finite element matrices using these test functions.

Remark 6.9. In the $L^q(\Omega)$ -setting, we have to solve a series of small, local *non-linear* problems instead, provided we choose the ‘correct’ weight for the duality mapping. This will become clear in the next chapter, where we will look at the convection-diffusion equation both in the Hilbert and the Banach setting in detail.

6.6 Petrov-Galerkin Methods for the Convection-Diffusion-Reaction Equation

In this section we consider the Petrov-Galerkin methods for the convection-diffusion(-reaction) equation in [23, 24, 39, 40, 47, 56]. The approaches rely on different strong and weak formulations of the PDE, as well as on different choices of the trial and test spaces and the corresponding norms that are employed on the continuous and discrete level. We will first summarize the different formulations of the convection-diffusion(-reaction) equations and comment on the choices made in each of the aforementioned articles. We will then consider the formulation in [40] as an example in more detail and comment on the potential to extend this approach to an $L^q(\Omega)$ -setting. The approach in [39] will be considered in more detail in Chapter 7, where it is extended to L^q -type Banach spaces.

6.6.1 Comparison of Several Petrov-Galerkin Methods

Despite being a seemingly simple equation, there are several formulations of the convection-diffusion-reaction equation that differ both in their strong form and their weak form. One choice we have to make is whether to consider the advective part in *conservative* form $(\nabla \cdot (\mathbf{b}u))$ or *non-conservative* form $(\mathbf{b} \cdot \nabla u)$. Furthermore, we have to choose the boundary conditions on the domain boundary $\partial\Omega$.

In this section we consider two strong versions of the convection-diffusion-reaction equation. The first one is in conservative form with a flux boundary condition on the inflow boundary $\Gamma_- := \{\mathbf{x} \in \partial\Omega : \mathbf{b} \cdot \mathbf{n}(\mathbf{x}) \leq 0\}$ and Dirichlet boundary conditions on the outflow boundary $\Gamma_+ := \partial\Omega \setminus \Gamma_-$:

$$-\varepsilon \Delta u + \nabla \cdot (\mathbf{b}u) + cu = f \quad \text{in } \Omega, \quad (6.24a)$$

$$u = u_0 \quad \text{on } \Gamma_+, \quad (6.24b)$$

$$(\mathbf{b}u - \varepsilon \nabla u) \cdot \mathbf{n} = \sigma_0 \quad \text{on } \Gamma_-. \quad (6.24c)$$

The reason for choosing the above boundary condition is to obtain robust estimates, cf., [40]. In the earlier paper [56], robust estimates were achieved for the same equation with Dirichlet boundary conditions everywhere by introducing a weighted test norm with non-constant, problem specific weighting.

The second version of the convection-diffusion-reaction equation is the same as in the previous section, i.e., in non-conservative form with Dirichlet boundary conditions everywhere:

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + cu = f \quad \text{in } \Omega, \quad (6.25a)$$

$$u = u_0 \quad \text{on } \partial\Omega. \quad (6.25b)$$

This formulation is used in [47, 23, 24, 39].

In the context of DPG methods, the second-order convection-diffusion-reaction equation is often reformulated as a first-order system. It is then possible to obtain various weak formulations from the first-order system by multiplying the system by test functions and integrating none, both or one of the equations by parts.

In order to simplify things a little, we now assume $u_0 \equiv 0$ and $c \equiv 0$. Introducing $\boldsymbol{\sigma} = \varepsilon \nabla u$, (6.24) can be reformulated as the following first-order system:

$$\varepsilon^{-1} \boldsymbol{\sigma} - \nabla u = \mathbf{0} \quad \text{in } \Omega, \quad (6.26a)$$

$$\nabla \cdot (\mathbf{b}u - \boldsymbol{\sigma}) = f \quad \text{in } \Omega, \quad (6.26b)$$

$$u = 0 \quad \text{on } \Gamma_+, \quad (6.26c)$$

$$(\mathbf{b}u - \boldsymbol{\sigma}) \cdot \mathbf{n} = \sigma_0 \quad \text{on } \Gamma_-. \quad (6.26d)$$

There is some arbitrariness in the definition of $\boldsymbol{\sigma}$. Broersen and Stevenson [23, 24] prefer $\boldsymbol{\sigma} := \sqrt{\varepsilon} \nabla u$ due to the fact that — ignoring boundary conditions — the first-order system then reduces to the advection-reaction equation in the limit $\varepsilon \rightarrow 0$.

Starting with either the first-order system (6.26) or its analogue derived from (6.25) with $\sigma = \sqrt{\varepsilon} \nabla u$, several weak formulations can be obtained by multiplying the two equations by test functions and integrating none, both or one of the equations by parts. If none of the equations are integrated by parts, we obtain the so-called *strong* or *trivial* form of the equation, and if both are integrated by parts, we obtain the *ultra-weak* formulation. The underlying operators of these two formulations are formally adjoint to each other and thus they are simultaneously well- or ill-posed, cf., [52]. With respect to well-posedness, both formulations only differ in allowing for different regularity in the solution and the right hand side.

If we only integrate (6.26b) by parts, the first-order system can be reduced to the classical or *primal* weak formulation that we have seen in the previous section by eliminating σ again. In this case, the weak formulation for σ and u is referred to as the (*primal*) *mixed* formulation and, in [24], the *mild-weak* formulation. If instead only (6.26a) is integrated by parts we obtain the *dual mixed* formulation. In [52] it is shown that all of the weak formulations mentioned here are simultaneously well- or ill-posed using the theory for closed operators, cf., Section 3.5. Moreover, proving well-posedness for any of the formulations is essentially reduced to well-posedness of the classical weak formulation. The same idea is applied to the Maxwell equations in [34].

The Petrov-Galerkin methods in [23, 56, 40] rely on an ultra-weak formulation of the convection-diffusion equation, the method in [24] relies on the primal mixed or mild-weak formulation, and the methods in [47, 39] rely on the classical weak formulation. On the infinite dimensional level, trial and test spaces are chosen as subspaces of $L^2(\Omega)$ with the minimal regularity that is required for the weak formulation to be well-defined. The norms are chosen canonically according to the spaces employed, though various weightings of the individual terms are used.

In [39, 47] only continuous finite element functions are considered; in [24, 56, 40] discontinuous functions are chosen in the test space and continuous functions in the trial space; finally, in [23] both test and trial functions are allowed to be discontinuous.

All of the approaches aim to achieve robustness in the diffusion parameter ε employing the following strategies some of which have already been mentioned.

1. A weak formulation that reduces to a weak formulation of the convection-diffusion equation in the limit as $\varepsilon \rightarrow 0$, cf., [23, 24].

2. Using the optimal test norm, cf., [24]. Note that in [24] the optimal test functions are computed analytically in certain special cases, whereas in other cases an approximation of the optimal test norm is introduced.
3. A weighted test-norm with non-constant, problem specific weighting, cf., [56, 39].
4. Modified boundary conditions, cf., [39, 40, 47].

6.6.2 A DPG Method for the Ultra-Weak Formulation

Let us now look at the DPG method in [40] in more detail. We again assume $\Omega \subset \mathbb{R}^d$ to be open and bounded with Lipschitz boundary. Denote by Ω_h a disjoint partitioning of Ω , $d \geq 1$, into open elements κ with Lipschitz boundaries $\partial\kappa$ such that $\overline{\Omega} = \bigcup_{\kappa \in \Omega_h} \overline{\kappa}$. Furthermore, denote by $\partial\Omega_h$ the collection of element boundaries $\partial\kappa$ for all $\kappa \in \Omega_h$. We multiply (6.26) by test functions $\boldsymbol{\tau}$ and v and, assuming sufficient regularity, integrate by parts on each element yielding

$$\begin{aligned} & \sum_{\kappa \in \Omega_h} \int_{\kappa} u(\nabla \cdot \boldsymbol{\tau} - \mathbf{b} \cdot \nabla v) d\mathbf{x} + \sum_{\kappa \in \Omega_h} \int_{\kappa} \boldsymbol{\sigma} \cdot (\varepsilon^{-1} \boldsymbol{\tau} + \nabla v) d\mathbf{x} \\ & + \sum_{\kappa \in \Omega_h} \int_{\partial\kappa \setminus \Gamma_+} \boldsymbol{\tau} \cdot \mathbf{n} u ds + \sum_{\kappa \in \Omega_h} \int_{\partial\kappa \setminus \Gamma_-} \boldsymbol{\sigma} \cdot v \mathbf{n} ds = \ell((v, \boldsymbol{\tau})). \end{aligned} \quad (6.27)$$

The right hand side is now given by

$$\ell((v, \boldsymbol{\tau})) = \int_{\Omega} f v d\mathbf{x} - \int_{\Gamma_-} \sigma_0 v ds.$$

Ignoring the boundary integrals in (6.27), we need the following regularity assumptions:

$$u \in L^q(\Omega), \quad \boldsymbol{\sigma} \in [L^q(\Omega)]^d, \quad v \in W^{1,q'}(\Omega_h), \quad \boldsymbol{\tau} \in W^{q'}(\text{div}, \Omega_h).$$

Note that we are using broken test spaces. We then need to replace u and $\boldsymbol{\sigma}$ on the interface by additional unknowns $\hat{u}$ and $\hat{\boldsymbol{\sigma}}$ that are taken from the trace space of the graph space of the operator A corresponding to the first-order system (6.24) subject to the boundary conditions. The operator A is given by

$$A(u, \boldsymbol{\sigma}) = \begin{pmatrix} \nabla \cdot (\mathbf{b}u - \boldsymbol{\sigma}) \\ \varepsilon^{-1} \boldsymbol{\sigma} - \nabla u \end{pmatrix}.$$

For $q = 2$, the space for $(\hat{u}, \hat{\sigma})$ is a subspace of $H^{1/2}(\partial\Omega_h) \times H^{-1/2}(\partial\Omega_h)$, cf., Section 6.5.1. We thus obtain the bilinear form

$$\begin{aligned} \mathcal{B}_\varepsilon^U((u, \boldsymbol{\sigma}, \hat{u}, \hat{\sigma}), (v, \boldsymbol{\tau})) := & \sum_{\kappa \in \Omega_h} \int_{\kappa} u(\nabla \cdot \boldsymbol{\tau} - \mathbf{b} \cdot \nabla v) \, d\mathbf{x} \\ & + \sum_{\kappa \in \Omega_h} \int_{\kappa} \boldsymbol{\sigma} \cdot (\varepsilon^{-1} \boldsymbol{\tau} + \nabla v) \, d\mathbf{x} \\ & + \sum_{\kappa \in \Omega_h} \int_{\partial\kappa \setminus \Gamma_+} \boldsymbol{\tau} \cdot \mathbf{n} \hat{u} \, ds + \sum_{\kappa \in \Omega_h} \int_{\partial\kappa \setminus \Gamma_-} \hat{\sigma} \cdot v \mathbf{n} \, ds. \end{aligned}$$

In order to define the inexact mixed method, we now only need to choose the test norm on $V = W^{1,q'}(\Omega_h) \times W^{q'}(\text{div}, \Omega_h)$. In [40] the following mesh dependent norm is suggested for $q = q' = 2$:

$$\begin{aligned} \|(v, \boldsymbol{\tau})\|_V^2 = & \sum_{\kappa \in \Omega_h} \left(\min\left(\frac{\varepsilon}{|\kappa|}, 1\right) \|v\|_{L^2(\kappa)}^2 + \varepsilon \|\nabla v\|_{L^2(\kappa)}^2 + \|\mathbf{b} \cdot \nabla v\|_{L^2(\kappa)}^2 \right) \\ & + \sum_{\kappa \in \Omega_h} \left(\|\nabla \cdot \boldsymbol{\tau}\|_{L^2(\kappa)}^2 + \min\left(\frac{1}{\varepsilon}, \frac{1}{|\kappa|}\right) \|\boldsymbol{\tau}\|_{L^2(\kappa)}^2 \right). \end{aligned} \quad (6.28)$$

This norm can trivially be generalized for $q' \geq 2$,

$$\begin{aligned} \|(v, \boldsymbol{\tau})\|_V^{q'} = & \sum_{\kappa \in \Omega_h} \left(\min\left(\frac{\varepsilon}{|\kappa|}, 1\right) \|v\|_{L^{q'}(\kappa)}^{q'} + \varepsilon \|\nabla v\|_{L^{q'}(\kappa)}^{q'} + \|\mathbf{b} \cdot \nabla v\|_{L^{q'}(\kappa)}^{q'} \right) \\ & + \sum_{\kappa \in \Omega_h} \left(\|\nabla \cdot \boldsymbol{\tau}\|_{L^{q'}(\kappa)}^{q'} + \min\left(\frac{1}{\varepsilon}, \frac{1}{|\kappa|}\right) \|\boldsymbol{\tau}\|_{L^{q'}(\kappa)}^{q'} \right), \end{aligned}$$

and we can compute the corresponding duality mapping with weight $\varphi(t) = t^{q'-1}$:

$$\begin{aligned} \langle J_V^\varphi(r, \boldsymbol{\rho}), (v, \boldsymbol{\tau}) \rangle_{V', V} = & \sum_{\kappa \in \Omega_h} \left(\min\left(\frac{\varepsilon}{|\kappa|}, 1\right) \int_{\kappa} |r|^{q'-2} r v \, d\mathbf{x} \right) \\ & + \sum_{\kappa \in \Omega_h} \varepsilon \int_{\kappa} \sum_{i=1}^d |\partial_i r|^{q'-2} \partial_i r \partial_i v \, d\mathbf{x} \\ & + \sum_{\kappa \in \Omega_h} \int_{\kappa} |\mathbf{b} \cdot \nabla r|^{q'-2} (\mathbf{b} \cdot \nabla r) (\mathbf{b} \cdot \nabla v) \, d\mathbf{x} \\ & + \sum_{\kappa \in \Omega_h} \int_{\kappa} |\nabla \cdot \boldsymbol{\rho}|^{q'-2} (\nabla \cdot \boldsymbol{\rho}) (\nabla \cdot \boldsymbol{\tau}) \, d\mathbf{x} \\ & + \sum_{\kappa \in \Omega_h} \min\left(\frac{1}{\varepsilon}, \frac{1}{|\kappa|}\right) \int_{\kappa} |\boldsymbol{\rho}|^{q'-2} \boldsymbol{\rho} \cdot \boldsymbol{\tau} \, d\mathbf{x}, \end{aligned}$$

which reduces to the inner product corresponding to the norm (6.28) if $q' = 2$. For suitably chosen discrete spaces V_m and U_n , the mixed method (6.9) is then given as: find $(u_n, \boldsymbol{\sigma}_n, \hat{u}_n, \hat{\sigma}_n) \in U_n$ and $(r_m, \boldsymbol{\rho}_m) \in V_m$ such that for all $(w_n, \boldsymbol{\zeta}_n, \hat{w}_n, \hat{\zeta}_n) \in U_n$ and all $(v_m, \boldsymbol{\tau}_m) \in V_m$

$$\begin{aligned} \langle J_V^\varphi(r_m, \boldsymbol{\rho}_m), (v_m, \boldsymbol{\tau}_m) \rangle_{V', V} + \mathcal{B}_\varepsilon^U((u_n, \boldsymbol{\sigma}_n, \hat{u}_n, \hat{\sigma}_n), (v_m, \boldsymbol{\tau}_m)) &= \ell(v_m, \boldsymbol{\tau}_m), \\ \mathcal{B}_\varepsilon^U((w_n, \boldsymbol{\zeta}_n, \hat{w}_n, \hat{\zeta}_n), (r_m, \boldsymbol{\rho}_m)) &= 0. \end{aligned}$$

Since the spaces for u_n and $\boldsymbol{\sigma}_n$ only need to be $L^q(\Omega)$ -conforming, we can choose piecewise polynomials of degree p_n that are not continuous across inter-element boundaries. For $q = 2$, the trace variables $\hat{u}_n$ and $\hat{\sigma}_n$ can be chosen as traces of $H^1(\Omega)$ - and $H(\text{div}, \Omega)$ -conforming elements, respectively, i.e., we can choose the traces of piecewise polynomials of degree p_n that are continuous across inter-element boundaries for $\hat{u}_n$ and the traces of Raviart-Thomas functions of degree p_n (RT_{p_n}) for $\hat{\sigma}_n$. For V_m we can use an enriched space, i.e., we can use spaces of piecewise polynomials of degree $p_m = p_n + \Delta p$ with $\Delta p \geq 1$. For r_m we can choose $H^1(\Omega_h)$ -conforming elements and $H(\text{div}, \Omega_h)$ -conforming elements for $\boldsymbol{\rho}_m$. The choice of the discrete spaces is adapted to the $L^2(\Omega)$ -setting, but —at least as far as the implementation is concerned— can also be used for $L^q(\Omega)$ -setting. However, to the best of our knowledge, the discrete stability is an open problem.

Chapter 7

The Non-linear Petrov-Galerkin Method for the Convection-Diffusion-Reaction Equation

Consider the following minimization problem: find $u_n \in U_n$ such

$$u_n = \arg \min_{w_n \in U_n} \|f - Bw_n\|_{V'},$$

where U and V are Banach spaces, $U_n \subset U$ a finite dimensional subspace, $B : U \rightarrow V'$ an invertible and bounded linear operator. In Chapter 6, we have seen that this is equivalent to the following non-linear mixed system: find $(r, u_n) \in V \times U_n$ such that

$$\begin{aligned} \langle J_V^{\varphi^{-1}}(r), v \rangle_{V', V} + \langle Bu_n, v \rangle_{V', V} &= \langle f, v \rangle_{V', V}, \quad \forall v \in V, \\ \langle Bw_n, r \rangle_{V', V} &= 0, \quad \forall w_n \in U_n, \end{aligned} \tag{7.1}$$

where $J_V^{\varphi^{-1}}$ is a duality mapping of weight φ^{-1} on V . The above system depends on the whole space V , but can be turned into a fully discrete system by replacing V with a suitable finite dimensional subspace V_m . It can be shown that under certain assumptions, the solution u_n is a quasi-best approximation of the analytical solution u , in the sense that

$$\|u - u_n\|_U \leq C \inf_{w_n \in U_n} \|u - w_n\|_U,$$

where the constant C depends on the continuity constant of B , the inf-sup constant of B , geometric constants corresponding to the spaces U and V and a constant stemming from a Fortin-type condition on the spaces U_n and V_m .

In this chapter we study how the mixed method (7.1) and its fully discrete counterpart can be applied to the convection-diffusion-reaction equation. In particular, we study the choice of the space V , the corresponding norm $\|\cdot\|_V$, and the discrete space V_m . The norm $\|\cdot\|_V$ on the space V determines the operator $J_V^{\varphi^{-1}}$ in (7.1) and thus crucially determines the resulting method. As in the special case $q = 2$ investigated in [39], we introduce weakly imposed boundary conditions on the inflow boundary in the space V_m to address robustness issues. We demonstrate in one dimension, how this choice of boundary conditions affects the approximation and compare this approach with the case when a weighted norm on V is employed.

The main focus of our numerical investigation is eliminating Gibbs phenomena in the finite element approximation. Indeed, we will see that the numerical approximation generated by our method qualitatively behaves like the $L^q(\Omega)$ -best approximation of the analytical solution. Thus, the Gibbs phenomenon can be eliminated by taking the limit $q \rightarrow 1$ provided that the $L^1(\Omega)$ -best approximation does not exhibit Gibbs phenomena. The results in Chapter 4 (see also [94]) show that this depends on the mesh that is selected. In one dimension only certain non-uniform grids have the property that the $L^1(\Omega)$ -best approximation contains over- and undershoots, whereas in higher dimensions this can even occur on structured, uniform meshes. We will demonstrate for two two-dimensional examples that it is possible to use the insights from Chapter 4 to modify the mesh in order to eliminate Gibbs phenomena as $q \rightarrow 1$.

7.1 Designing the Non-linear Petrov-Galerkin Method for the Convection-Diffusion-Reaction Equation

The main focus of this chapter is to apply the abstract approach described in Sections 6.2-6.4 to the scalar convection-diffusion-reaction equation in a $W^{1,q}(\Omega)$ - $W^{1,q'}(\Omega)$ setting. To this end, consider the following model problem:

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + cu = f \quad \text{in } \Omega, \quad (7.2a)$$

$$u = 0 \quad \text{on } \Gamma = \partial\Omega, \quad (7.2b)$$

where ε , $\mathbf{b} : \Omega \rightarrow \mathbb{R}^d$ and $c : \Omega \rightarrow \mathbb{R}$ are the (positive and constant) diffusion parameter, the convection field and the reaction coefficient, respectively, and $f : \Omega \rightarrow \mathbb{R}$ is a given source term. Multiplying by a test function $v \in C_c^\infty(\Omega)$ and integrating by parts yields the bilinear form

$$\mathcal{B}_\varepsilon(u, v) = \varepsilon \int_\Omega \nabla u \cdot \nabla v \, d\mathbf{x} - \int_\Omega u \nabla \cdot (\mathbf{b}v) \, d\mathbf{x} + \int_\Omega cuv \, d\mathbf{x}. \quad (7.3)$$

This allows us to state the following variational problem: find $u \in U := W_0^{1,q}(\Omega)$ such that

$$\mathcal{B}_\varepsilon(u, v) = \langle f, v \rangle_{V', V} \quad \forall v \in V := W_0^{1,q'}(\Omega). \quad (7.4)$$

In Chapter 5 (see also [93]), a proof of an inf-sup condition is given, where the norms on U and V are weighted versions of the standard norms $W^{1,q}(\Omega)$ and $W^{1,q'}(\Omega)$, respectively. Furthermore, this proof requires certain regularity assumptions on the solution to the Poisson problem and that the convection field $\mathbf{b}$ and the reaction coefficient c satisfy $c - \nabla \cdot \frac{1}{q} \mathbf{b} \geq c_0 > 0$. The continuity constant and the inf-sup constant that are established for the bilinear $\mathcal{B}_\varepsilon$ in Chapter 5 depend on the problem specific parameters. It should be noted that the estimates in Chapter 5 can be expected to be sub-optimal since sharper bounds are known for $q = 2$.

7.1.1 Choices for the Test Norm

The choice of the norm for the space V crucially defines the method described in Section 6.4. To this end, we endow V with the following norm

$$\|v\|_V^{q'} := \alpha \|v\|_{L^{q'}(\Omega)}^{q'} + \varepsilon \|\nabla v\|_{L^{q'}(\Omega)}^{q'} + \frac{|\Omega|^{1/2}}{\|\mathbf{b}\|_{L^\infty(\Omega)}} \|(\omega(x))^{1/q'} \mathbf{b} \cdot \nabla v\|_{L^{q'}(\Omega)}^{q'}, \quad (7.5)$$

where $\alpha \geq 0$ is a constant $\omega(x)$ is a positive and smooth weighting function. It is well known that for $q = 2$ and $\omega(x) \equiv \text{const} \geq 0$ the above choice of norm on V does not yield a robust formulation, i.e., M/γ depends on the problem parameters. Here, M denotes the continuity constant of the bilinear form and γ its inf-sup constant. The inf-sup constant for $1 < q < \infty$ obtained in Chapter 5 corresponds to the choice $\omega(x) \equiv 0$ and $\alpha = c_0$, where c_0 is the constant in the Friedrichs positivity assumption, $c - \frac{1}{q} \nabla \cdot \mathbf{b} \geq c_0$. As mentioned above, the continuity and inf-sup constants are not parameter independent and can be expected to be suboptimal.

One approach to address the robustness issue is to introduce a non-constant weighting function, cf., [56, 39]. For example, we may require $\omega(x)$ to be of

magnitude $\mathcal{O}(\varepsilon)$ near the inflow boundary but $\mathcal{O}(1)$ elsewhere. We will show the effect of this by considering a simple one-dimensional example on $\Omega = (0, 1)$ with the inflow boundary at 0. For this example $\omega(x) = x + \varepsilon$ has the desired properties; for comparison, we will also consider $\omega(x) \equiv 0$ and $\omega(x) \equiv 1$.

7.1.2 Weak Boundary Conditions on r

Alternatively, the robustness issue can be addressed by changing the boundary conditions on r ; this has been considered in different ways in [47, 39]. In [40], the boundary conditions on u were modified instead. We will now present the approach in [39] and extend it for $1 < q < 2$. The idea is to relax the boundary condition on the test space on the inflow part of the boundary. The reasoning behind this is that in the mixed method we essentially approximate the adjoint equation in the test space in order to approximate the residual or the optimal test functions. Underresolved layers at the inflow boundary—which is the outflow boundary for the adjoint equation—then pollute the solution to the primal problem in the inexact method.

Instead of $W_0^{1,q'}(\Omega)$, we consider the modified test space $V = W_{0,\Gamma_+}^{1,q'}(\Omega)$, where $q' = q/(q-1)$, and

$$W_{0,\Gamma_+}^{1,q'}(\Omega) := \left\{ v \in W^{1,q'}(\Omega) : v|_{\Gamma_+} = 0 \right\},$$

Here,

$$\Gamma_- := \{ \mathbf{x} \in \partial\Omega : \mathbf{b} \cdot \mathbf{n}(\mathbf{x}) \leq 0 \}, \quad \Gamma_+ := \partial\Omega \setminus \Gamma_-,$$

where $\mathbf{n}(x)$ denotes the unit outward normal at a point on the boundary $\partial\Omega$. The modified bilinear form (cf., [39]) is given by

$$\tilde{\mathcal{B}}_\varepsilon(u, v) := \varepsilon \int_\Omega \nabla u \cdot \nabla v \, d\mathbf{x} + \int_\Omega (\mathbf{b} \cdot \nabla u) v \, d\mathbf{x} - \varepsilon \int_{\Gamma_-} \frac{\partial u}{\partial \mathbf{n}} v \, ds. \quad (7.6)$$

The boundary term is merely the term that is picked up from the integration by parts if v is non-zero on the inflow boundary. Note, however, that the term $\varepsilon \int_{\Gamma_-} \frac{\partial u}{\partial \mathbf{n}} v \, ds$ is a variational crime if we assume $u \in W_0^{1,q}(\Omega)$. As noted in [39], the correct way of including this term would be to introduce it as an additional unknown on the boundary and the inclusion of $\varepsilon \int_{\Gamma_-} \frac{\partial u}{\partial \mathbf{n}} v \, ds$ can be viewed as a discrete elimination of this unknown, cf., [39, 23].

7.1.3 The Inexact Method for the Convection-Diffusion-Reaction Equation

The key step for implementing the inexact method (6.9) for any specific problem is determining the duality mapping $J_V^{\varphi^{-1}}$. We choose the weight function $\varphi^{-1}(t) = t^{q'-1}$ and compute the duality mapping, similar to Section 3.3.2; thereby, we get

$$\begin{aligned} \langle J_V^{\varphi^{-1}}(v), w \rangle_{V', V} &= \int_{\Omega} |v|^{q'-1} \operatorname{sgn}(v) w \, d\mathbf{x} \\ &\quad + \varepsilon \int_{\Omega} \sum_{i=1}^d |\partial_i v|^{q'-1} \operatorname{sgn}(\partial_i v) \partial_i w \, d\mathbf{x} \\ &\quad + \int_{\Omega} \omega(x) |\mathbf{b} \cdot \nabla v|^{q'-1} \operatorname{sgn}(\mathbf{b} \cdot \nabla v) \mathbf{b} \cdot \nabla w \, d\mathbf{x}. \end{aligned}$$

For a given right hand side $\ell \in V'$, we solve the following non-linear system: find $(u_n, r_m) \in U_n \times V_m \subset U \times V$ such that

$$\langle J_V^{\varphi^{-1}}(r_m), v_m \rangle_{V', V} + \mathcal{B}_{\varepsilon}(u_n, v_m) = \ell(v) \quad \text{for all } v_m \in V_m, \quad (7.7a)$$

$$\mathcal{B}_{\varepsilon}(w_n, r_m) = 0 \quad \text{for all } w_n \in U_n. \quad (7.7b)$$

One can easily implement (7.7) in, e.g., FEniCS [5, 126] using standard $H^1(\Omega)$ -conforming Lagrange finite elements for both U_n and V_m . The spaces U_n and V_m are chosen over a common mesh. For U_n a global polynomial degree p_n is chosen and for V_m we choose an enriched finite element space with polynomial degree $p_m = p_n + \Delta p$, $\Delta p \geq 1$. To consider the weak boundary conditions on r introduced in the previous section, we simply replace $\mathcal{B}_{\varepsilon}$ with $\tilde{\mathcal{B}}_{\varepsilon}$ and adjust the space V_m to only satisfy Dirichlet boundary conditions on the outflow boundary.

7.1.4 The Limit Case $\varepsilon = 0$

Since we are interested in the convection-dominated case, i.e., $\varepsilon \ll \|\mathbf{b}\|_{L^\infty(\Omega)}$, it makes sense to consider the limit $\varepsilon \rightarrow 0$. Simply setting $\varepsilon = 0$ in (7.3) does not yield a well-posed problem unless we only consider boundary conditions on the inflow boundary Γ_- . However, even the ill-posed problem with Dirichlet boundary conditions on the whole boundary $\partial\Omega$ is of interest, since this is essentially the problem that is approximated numerically on coarse meshes when $\varepsilon \ll \|\mathbf{b}\|_{L^\infty(\Omega)}$, cf., the discussion in [82]. On a discrete level we can think of imposing boundary conditions for the approximation u_n on Γ_+ simply as considering an approximation problem in a smaller subspace of U .

When $\varepsilon = 0$, there are two different weak formulations of the *convection-reaction* equation depending on whether $\int_{\Omega}(\mathbf{b} \cdot \nabla u)v \, d\mathbf{x}$ is integrated by parts or not. The two cases differ in the regularity of the trial and test spaces which is reflected in the norms chosen on U and V . Thus, in each case the residual is measured in a different norm and (quasi-)best approximation of the analytical solution is achieved in a different space. Both weak formulations can be extended to weak formulations of the convection-diffusion-equation by adding the diffusion term and accounting for different boundary conditions and regularity requirements. If $\varepsilon > 0$, we require $u \in W_0^{1,q}(\Omega)$ in both cases and the two weak formulations are formally equivalent. However, the differences observed in the limit case $\varepsilon = 0$ can be reflected in the choice of the norm on $W_0^{1,q}(\Omega)$ by choosing the weighting of the terms $\|w\|_{L^q(\Omega)}$, $\|\nabla w\|_{L^q(\Omega)}$ and $\|\nabla \cdot \mathbf{b}w\|_{L^q(\Omega)}$ accordingly.

We show that for the first choice the exact mixed method is equivalent to the formulation used in [82] and that in the second case a quasi-best approximation in $L^q(\Omega)$ can be computed by determining the optimal test norm. We use this to interpret certain choices of the test norm for the convection-diffusion-reaction equation.

7.1.4.1 Residual Minimization in $L^q(\Omega)$

Both weak formulations for the convection-reaction equation are obtained by multiplying the convection-diffusion-reaction equation (7.2) with $\varepsilon = 0$ by a smooth test function and integrating over the domain Ω . This immediately yields the first possible choice for the weak formulation: find $u \in W_{0,\Gamma_-}^q(\mathbf{b}, \Omega)$ such that

$$\hat{\mathcal{B}}_0(u, v) = \int_{\Omega}(\mathbf{b} \cdot \nabla u)v \, d\mathbf{x} + \int_{\Omega} cuv \, d\mathbf{x} \quad \forall v \in L^{q'}(\Omega).$$

Here, the bilinear form $\hat{\mathcal{B}}_0$ is well defined for $v \in L^{q'}(\Omega)$ and u in the graph space

$$W_{0,\Gamma_-}^q(\mathbf{b}, \Omega) := \{w \in L^q(\Omega) : \mathbf{b} \cdot \nabla w \in L^q(\Omega) \text{ and } w = 0 \text{ on } \Gamma_-\}$$

endowed with the norm

$$\|u\|_{W_{0,\Gamma_-}^q(\mathbf{b}, \Omega)}^q = \|u\|_{L^q(\Omega)}^q + \|\mathbf{b} \cdot \nabla u\|_{L^q(\Omega)}^q.$$

The associated residual minimization problem is

$$u_n = \arg \min_{w_n \in U_n} \|\hat{\mathcal{B}}_0(w_n, \cdot) - f\|_{L^q(\Omega)};$$

this was considered in [82]. In general, method (6.9) yields a formulation for solving the residual minimization problem inexactly. However, since in this case $J_{V'}^\varphi = J_q$, we can avoid the inexactness by directly implementing (6.7); this is exactly the approach considered in [82] where a regularization is introduced to compute (6.7). The norm on $V = L^{q'}(\Omega)$ corresponds to the choice of the test-norm for the convection-diffusion-reaction equation given in (7.5) with $\omega(x) \equiv 0$ and $\alpha = 1$ since this yields the $L^{q'}(\Omega)$ -norm if ε is set to zero. This suggests that this choice of $\omega(x)$ and α corresponds to the natural extension of this bilinear form to the case $\varepsilon > 0$. Moreover, the equivalence of the exact minimum residual method and the approach in [82] for $\varepsilon = 0$ together with ε weighting of the norm of the gradient suggest that for this choice of the test norm, our scheme closely resembles the approach in [82] for $0 < \varepsilon \ll 1$.

7.1.4.2 Best- L^q Approximation

Next, we can integrate the convection term by parts in order to obtain the second weak formulation. To this end, we note that

$$\int_{\Omega} (\mathbf{b} \cdot \nabla u) v \, d\mathbf{x} = - \int_{\Omega} \nabla \cdot (\mathbf{b} v) u \, d\mathbf{x} + \int_{\partial\Omega} (\mathbf{b} \cdot \mathbf{n}) u v \, d\mathbf{s}.$$

The boundary term on Γ_- vanishes since $u = 0$ on Γ_- (for non-zero boundary conditions this term would be absorbed into the right hand side by inserting the boundary condition). For v in the graph space

$$W_{0,\Gamma_+}^{q'}(\mathbf{b}, \Omega) := \{w \in L^{q'}(\Omega) : \mathbf{b} \cdot \nabla w \in L^{q'}(\Omega) \text{ and } w = 0 \text{ on } \Gamma_+\},$$

the boundary term also vanishes on Γ_+ and we obtain the bilinear form

$$\mathcal{B}_0(u, v) = - \int_{\Omega} u \nabla \cdot (\mathbf{b} v) \, d\mathbf{x} + \int_{\Omega} c u v \, d\mathbf{x},$$

which is well-defined for $u \in L^q(\Omega)$. In this case we can compute the optimal test norm

$$\|v\|_{\text{opt}} := \sup_{u \in L^q(\Omega)} \frac{\mathcal{B}_0(u, v)}{\|u\|_{L^q(\Omega)}} = \|\mathcal{B}_0(\cdot, v)\|_{(L^q(\Omega))'} = \|-\nabla \cdot (\mathbf{b} v) + cv\|_{L^{q'}(\Omega)}.$$

Choosing the optimal test norm, allows us to obtain the L^q -best approximation in a given finite dimensional space U_n . It is easy to see that the optimal test norm is

bounded from above up to a constant by the graph norm

$$\|w\|_{W_{0,\Gamma+}^{q'}}^{q'} := \|w\|_{L^{q'}(\Omega)}^{q'} + \|\mathbf{b} \cdot \nabla w\|_{L^{q'}(\Omega)}^{q'}.$$

In [139] this formulation of the convection-reaction equation is analysed in detail and an inf-sup condition is established assuming $c - \frac{1}{q} \nabla \cdot \mathbf{b} \geq c_0 > 0$. This implies that the optimal test norm is up to a constant also bounded from below by the graph norm. In other words, the graph norm is equivalent to the optimal test norm. The graph norm can be obtained from (7.5) by choosing $\omega(x) \equiv 1$, $\alpha = 1$ and setting ε to zero. This suggests that this choice of α and $\omega(x)$ yields the natural extension of this formulation to the convection-diffusion-reaction equation. The dependence of the inf-sup constant in [139] on c_0 suggests that the equivalence constants can be improved by choosing $\alpha = c_0$.

7.2 Numerical Examples

In this section we consider a range of numerical test cases to illustrate the performance of our proposed numerical scheme. To this end, in Section 7.2.1 we demonstrate that in the diffusion-dominated regime optimal convergence rates are achieved and moreover that in the convection-dominated regime the convergence rate is as expected. We also study the effect of different choices of the discrete test space V_m on the convergence rates. In Section 7.2.2, we then compare different choices for the test norm on V and the boundary conditions as described in Sections 7.1.2 and 7.1.1 for a one-dimensional example. Furthermore, for the same example we study for which choices of V , $\|\cdot\|_V$ and V_m we can observe vanishing over- and undershoots as $q \rightarrow 1$ and whether the method is robust in ε . In Section 7.2.3 we consider three two-dimensional examples, where for certain selected meshes the overshoots disappear as $q \rightarrow 1$; this behaviour can be predicted by considering L^1 -best approximations of discontinuities.

The four examples we are using in this section are given below. They consist of one simple one-dimensional example and three two-dimensional examples each with solutions containing boundary layers for small ε . The solution to the last example additionally contains an interior layer.

Example 7.1.

$$-\varepsilon u'' + u' = 0 \quad \text{in } (0, 1), \quad u(0) = 0, \quad u(1) = 1.$$

The analytical solution is given by

$$u(x) = \frac{\exp\left(-\frac{1}{\varepsilon}\right) - \exp\left(\frac{x-1}{\varepsilon}\right)}{\exp\left(-\frac{1}{\varepsilon}\right) - 1}.$$

Example 7.2 (Eriksson-Johnson model problem).

$$\begin{aligned} \frac{\partial u}{\partial x} - \varepsilon \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) &= 0 \quad \text{in } (0,1)^2, \\ u &= 0 \quad \text{if } x = 1, y = 0, 1, \quad u = \sin(\pi y) \quad \text{if } x = 0. \end{aligned}$$

The analytical solution is given by

$$u(x, y) = \frac{\exp(r_1(x-1)) - \exp(r_2(x-1))}{\exp(r_1) - \exp(r_2)} \sin(\pi y),$$

where

$$r_1 = \frac{1 + \sqrt{1 + 4\pi^2\varepsilon^2}}{2\varepsilon}, \quad r_2 = \frac{1 - \sqrt{1 + 4\pi^2\varepsilon^2}}{2\varepsilon}.$$

Example 7.3 (Boundary Layer in the Corner of the Domain).

$$\begin{aligned} \mathbf{b} \cdot \nabla u - \varepsilon \Delta u &= b_2 h_1(x) + b_1 h_2(y) \quad \text{in } (0,1)^2, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

with $\mathbf{b} = (b_1, b_2)^T$, $b_{1,2} > 0$ and

$$h_1(x) = x - \frac{1 - \exp\left(\frac{b_1 x}{\varepsilon}\right)}{1 - \exp\left(\frac{b_1}{\varepsilon}\right)}, \quad h_2(y) = y - \frac{1 - \exp\left(\frac{b_2 y}{\varepsilon}\right)}{1 - \exp\left(\frac{b_2}{\varepsilon}\right)}.$$

The analytical solution is given by

$$u(x, y) = h_1(x)h_2(y).$$

Since $b_{1,2} > 0$, the outflow boundary is defined by the two lines $x = 1$ and $y = 1$. For small ε we can observe a boundary layer at the outflow boundary and in particular near the corner $(x, y) = (1, 1)$.

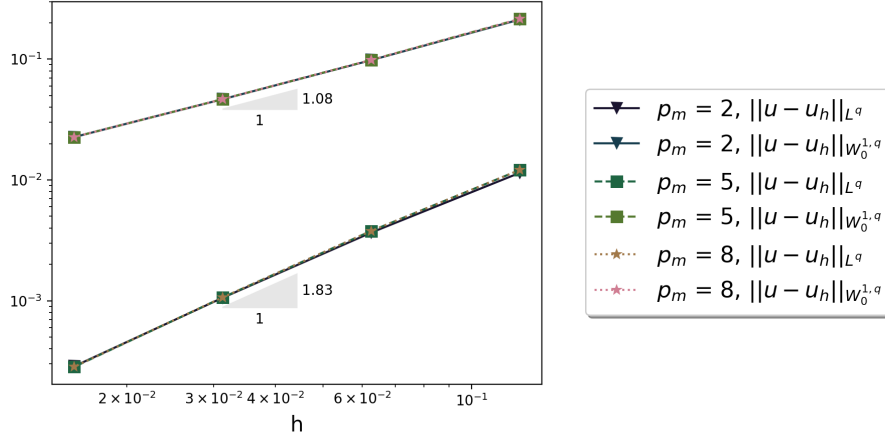


Figure 7.1 Example 7.2 with $\varepsilon = 1$ and $q = 1.2$. Convergence for $p_n = 1$ and $\Delta p = 1, 4, 7$.

Example 7.4 (Interior and Boundary Layer).

$$\begin{aligned} \mathbf{b} \cdot \nabla u - \varepsilon \Delta u &= 0 && \text{in } (0, 1)^2, \\ u &= \frac{(\delta - x)^+}{\delta} && \text{on } \partial\Omega, \end{aligned}$$

with $(\delta - x)^+ = \max(0, \delta - x)$, $\delta \ll 1$ and $\mathbf{b} = (2, 1)^T$.

For this example, a boundary layer develops at $\Gamma_+ \cap \{y > 0.5\}$ and an interior layer along the line $y = 0.5x$ if δ is smaller than the mesh size.

7.2.1 Convergence Tests

We start with investigating the convergence of the proposed method with weak boundary conditions on the inflow boundary in the space V_m , i.e., formulation (7.4) with $\tilde{\mathcal{B}}_\varepsilon$ from (7.6), and the norm (7.5) with $\omega(x) \equiv 1$ and $\alpha = 1$. To this end, we first consider the convergence of the method for Example 7.2 with $\varepsilon = 1$. We consider a uniform mesh with element size h . The space U_n consists of piecewise polynomials of degree p_n , while space V_m consists of piecewise polynomials of degree $p_m = p_n + \Delta p$ on the same mesh with $\Delta p \geq 1$. Figure 7.1 shows that for $q = 1.2$ the error is essentially independent of the choice of Δp and hence there is no benefit to enriching the space of the residual r_m beyond $p_m = p_n + 1$. Figure 7.2 shows that we obtain the expected optimal convergence rates for different polynomial degrees p_n under uniform h -refinement.

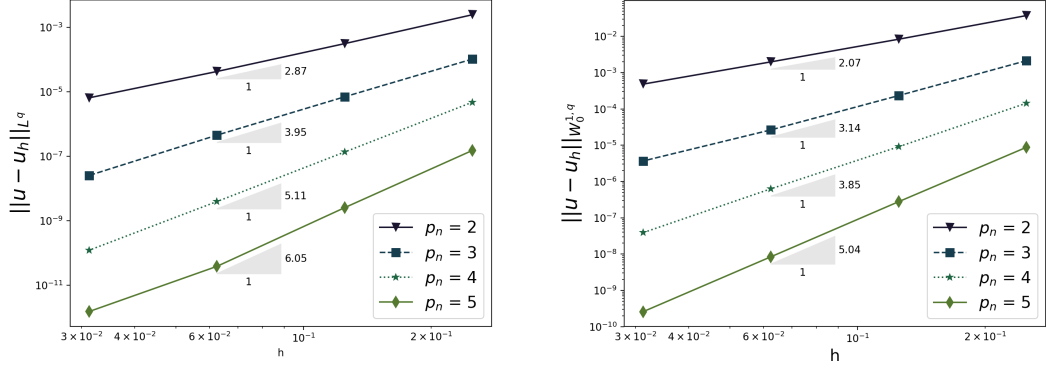


Figure 7.2 Example 7.2 with $\varepsilon = 1$ and $q = 1.2$, for $p_n = 2, 3, 4, 5$ and $\Delta p = 2$. Left: $L^q(\Omega)$ -norm, Right: $W^{1,q}(\Omega)$ -norm

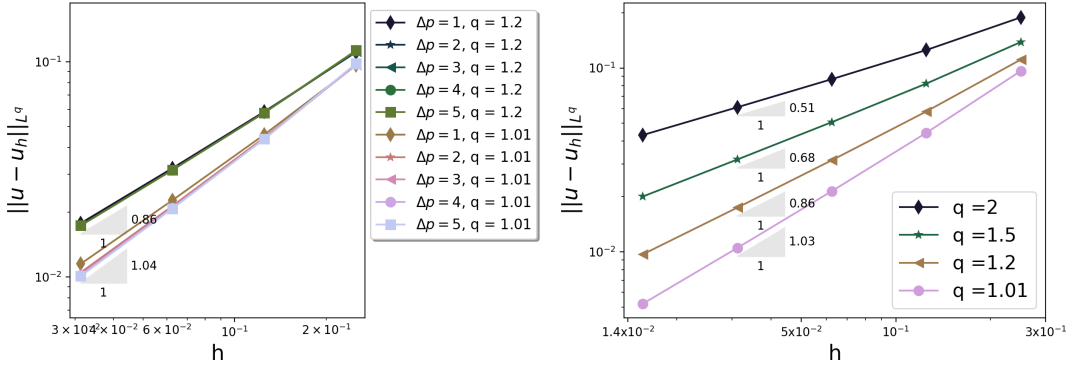


Figure 7.3 $L^q(\Omega)$ -error for Example 7.2 with $\varepsilon = 10^{-4}$ and $p_n = 1$.

Next, we consider the same example with the same finite element spaces for $\varepsilon = 10^{-4}$. Figure 7.3 (left) shows that again both for $q = 1.01$ and $q = 1.2$ the error does not improve as Δp is increased beyond $\Delta p = 2$; for $q = 1.01$, however, we can observe that the error is larger for $\Delta p = 1$. Note that we did not observe this in the diffusive regime. The plot on the right shows that in the convection-dominated regime we obtain a convergence rate of approximately $\mathcal{O}(h^{\frac{1}{q}})$ as h tends to zero; this is consistent with the approximation error bound of the piecewise linear interpolant of a jump discontinuity.

7.2.2 Convection-Diffusion in 1D

In this section we consider different choices for V , $\|\cdot\|_V$ and V_m and show how this affects the approximation to the solution u of Example 7.1. To this end, in Section 7.2.2.1, we investigate how the choice of α , $\omega(x)$ and the boundary conditions in V_m affect the approximation u_n . In Section 7.2.2.2 we consider one specific choice

for the test norm and show that the oscillations vanish entirely as $q \rightarrow 1$. Next, in Section 7.2.2.3, we investigate the choice of the finite element space V_m more closely. We have seen that increasing Δp has no significant effect on the convergence rates; however, we will see that this does indeed affect certain qualitative properties of the solution. Finally, we demonstrate that the method is robust in ε in Section 7.2.2.4

7.2.2.1 Comparing the Choices for the Test Norm and the Boundary Conditions

We start by comparing different versions of the non-linear Petrov-Galerkin method. In order to ensure that the space V_m is sufficiently large we use piecewise polynomials of degree $p_m = 10$ as the basis, while U_n consists of piecewise linear polynomials. We split the interval $(0, 1)$ uniformly into 8 elements. We consider three different choices for the weighting function $\omega(x)$ as mentioned in Section 7.1.1, namely $\omega(x) = x + \varepsilon$, $\omega(x) \equiv 1$ and $\omega(x) \equiv 0$. We combine this with two different choices of boundary conditions on r , i.e., zero Dirichlet boundary conditions on the whole boundary and weak boundary conditions on the inflow boundary as described in Section 7.1.2. This creates six test cases; we first consider these six test cases for $\varepsilon = 10^{-3}$ with $\alpha = 1$ and $q = 2, 1.01$. Secondly, we consider all six test cases with $\varepsilon = 10^{-6}$, $\alpha = 0, 1$ and $q = 2, 1.01$. The solution u_n for each of the cases is shown in Figure 7.4.

We can see in Figures 7.4a, 7.4c and 7.4e that for $q = 2$ and Dirichlet boundary conditions on the whole boundary for r , the approximation u_n only resembles the analytical solution if $\omega(x) = x + \varepsilon$. Introducing weak boundary conditions on r on the inflow boundary resolves this issue for $\omega(x) \equiv 1$ but the approximation with $\omega(x) \equiv 0$ only shows improvement if $\alpha = 0$. These observations are the same for $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-6}$. This indicates that we can circumvent constructing a non-constant function $\omega(x)$ – which can be challenging for complicated geometries – by introducing weak boundary conditions on r on the inflow boundary. These observations are consistent with the results for a very similar problem studied in [39]. Furthermore, note that $\omega(x) = 0$ and $\alpha = 1$ most closely resembles the method introduced in [82]. In [82] it is demonstrated for a different (two-dimensional) example that the approximation for $q = 2$ can be very inaccurate.

If we now consider $q = 1.01$, cf., Figures 7.4b, 7.4d and 7.4f, we observe a different behaviour of the approximation. If $\varepsilon = 10^{-3}$, the approximations are all very similar and much closer to the analytical solution with close to no undershoots;

we will investigate the phenomenon of vanishing undershoots more closely both in one dimension and two dimensions later on. If $\varepsilon = 10^{-6}$, the approximation with Dirichlet boundary conditions on the whole boundary and $\omega(x) \equiv 0$ or $\omega(x) \equiv 1$ again leads to a very poor approximation to the analytical solution. This again points to robustness issues that for $q = 1.01$ are only visible for much smaller ε than for $q = 2$. An interesting observation is that for $\alpha = 1$, the combination of weak inflow boundary conditions on r and $\omega(x) \equiv 0$ is not distinguishable from other choices of $\omega(x)$ for $q = 1.01$ whereas, for $q = 2$ it is significantly different. The improvement of the approximation as $q \rightarrow 1$ for $\omega(x) \equiv 0$ and $\alpha = 1$ resembles the results presented in [82].

In conclusion, choosing the boundary conditions as described in Section 7.1.2 allows us to avoid constructing a non-constant weighting function $\omega(x)$. Close to $q = 1$, the results are very similar for $\omega(x) \equiv 1$ and $\omega(x) \equiv 0$, whereas close to $q = 2$, $\omega(x) \equiv 1$ clearly yields a significantly better approximation of the solution.

The choice of α does not seem to have a big impact on the approximation except in the case $\omega(x) \equiv 0$ and $q = 2$, where choosing $\alpha = 0$ improves the approximation compared to $\alpha = 1$. If $\omega(x) \equiv 1$, then both for $\alpha = 1$ and $\alpha = 0$, the dominant term in the norm on V is essentially $\|\nabla u\|_{L^{q'}(\Omega)}$ since $\|\mathbf{b} \cdot \nabla v\|_{L^{q'}(\Omega)}$ and $\|\nabla v\|_{L^{q'}(\Omega)}$ are the same up to a constant in one dimension. If $\omega(x) \equiv 0$, then $\alpha = 0$ implies that V is endowed with the $W_0^{1,q'}(\Omega)$ -norm, whereas for $\alpha = 1$ the $L^{q'}(\Omega)$ -term is the dominant term in the norm on V due to the ε -weighting of the gradient. Since $\alpha = 0$ yields a better approximation in this case, this suggests that for $q = 2$ it is favourable to choose a stronger norm on V that is either dominated by the norm of the gradient or the streamline term $\|\mathbf{b} \cdot \nabla v\|_{L^{q'}(\Omega)}$.

7.2.2.2 Vanishing oscillations as $q \rightarrow 1$

The examples in Figure 7.4 already illustrate that the undershoot in the approximate solution nearly vanishes for $q = 1.01$. We will now investigate in more detail how the undershoot depends on q if we choose $\varepsilon = 10^{-5}$, $\omega(x) \equiv 1$, $\alpha = 1$ and impose weak inflow boundary conditions on r_m . To this end, we choose a large space V_m with polynomial degree $p_m = 10$ and piecewise linear polynomials for U_n . We split the interval into 8 elements and compute the approximate solutions (r_m, u_n) for several choices of q . Figure 7.5 shows u_n on the left and r_m on the right. A close-up of the undershoot and a plot $\min(u_n)$ vs. q is shown in the centre of the figure. We can see that the undershoot decreases monotonically as q approaches 1. In contrast to this, we can see that the oscillations in the residual r_m increase as

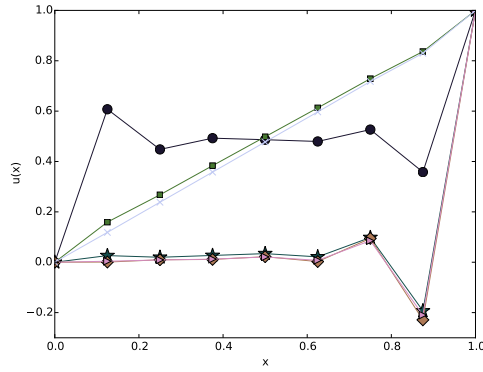
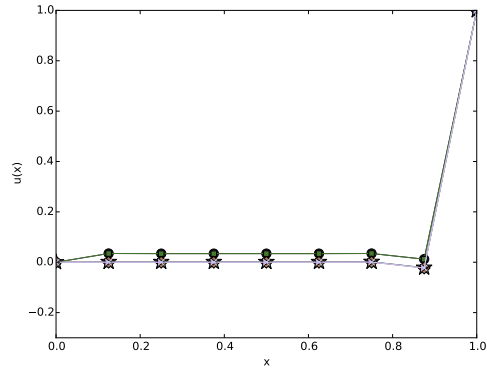
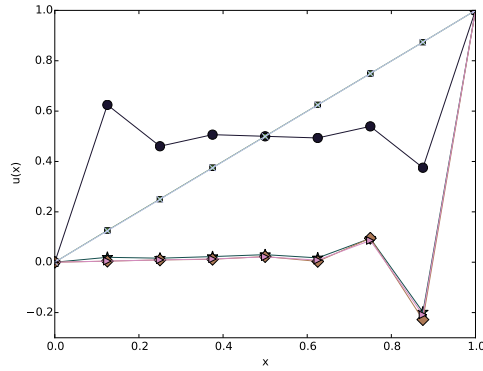
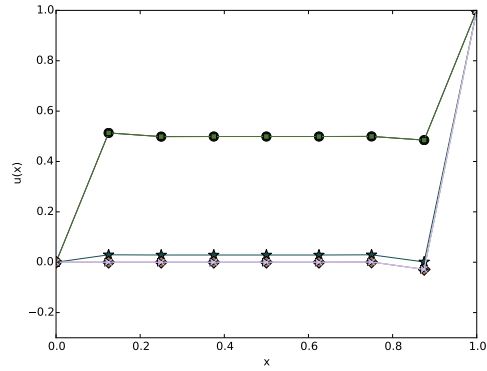
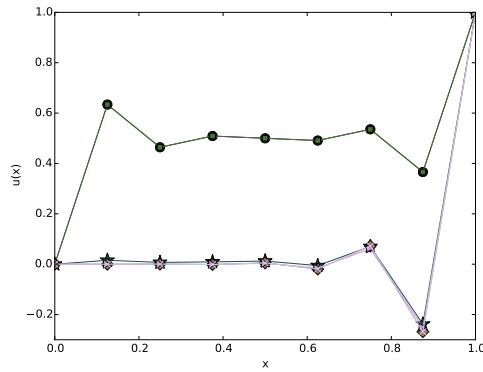
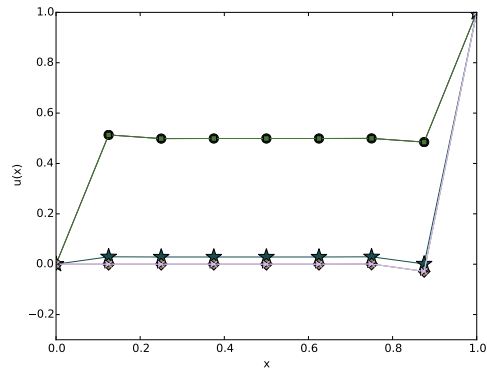
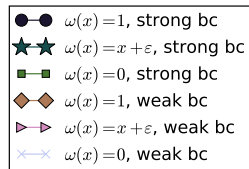
(a) $\varepsilon = 10^{-3}, q = 2, \alpha = 1$ (b) $\varepsilon = 10^{-3}, q = 1.01, \alpha = 1$ (c) $\varepsilon = 10^{-6}, q = 2, \alpha = 1$ (d) $\varepsilon = 10^{-6}, q = 1.01, \alpha = 1$ (e) $\varepsilon = 10^{-6}, q = 2, \alpha = 0$ (f) $\varepsilon = 10^{-6}, q = 1.01, \alpha = 0$ 

Figure 7.4 Example 7.1: Solution u_n for $q = 2$ and $q = 1.01$ using different norms, i.e., different weighting functions $\omega(x)$, and either Dirichlet boundary conditions for r on the whole boundary (strong bc) or weak boundary conditions on r on the inflow boundary and Dirichlet conditions on the outflow boundary (weak bc).

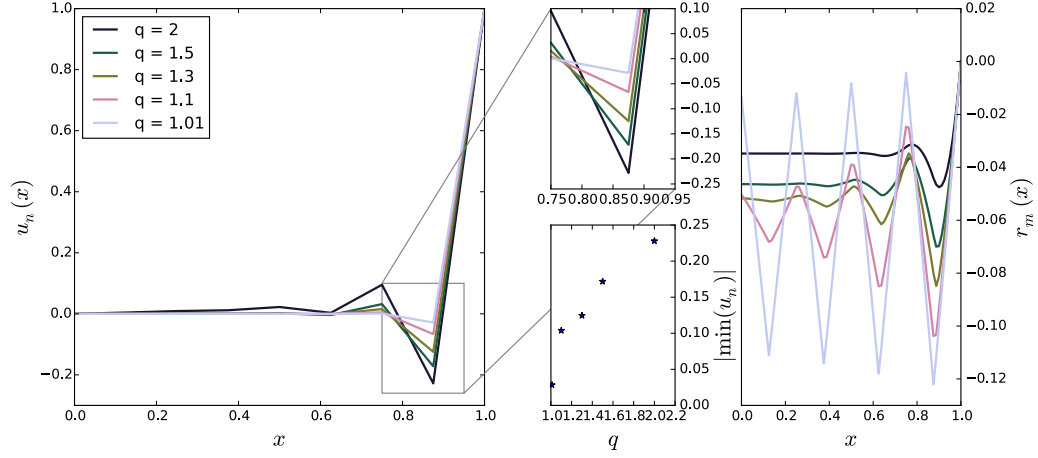


Figure 7.5 Example 7.1 with $\varepsilon = 10^{-5}$: Numerical approximations u_n (left) and r_m (right) with a uniform mesh consisting of 8 intervals using piecewise linear polynomials for u_n and polynomials of degree $p_m = 10$ for r_m for varying values of q . Top centre: Zoom-in showing over- and undershoots. Bottom centre: $|\min(u_n)|$ vs q .

$q' \rightarrow \infty$ which suggests that a large space V_m may be necessary. This is the next aspect of the method we will investigate.

7.2.2.3 The Choice of the Space V_m

In [137], it has been shown that the inexact method is well-posed if a Fortin projector, cf., (6.10) exists. It is easy to see that $\dim(V_m) \geq \dim(U_n)$ is a necessary condition. Finding a sufficient condition or in other words finding a compatible pair (U_n, V_m) is highly non-trivial. In [139], certain special cases for the convection-reaction equation are considered. The observations in Section 7.2.1 suggest that for $\Delta p \geq 2$, the Fortin condition is typically satisfied. We now investigate how the choice of V_m affects the undershoot for q close to 1. To this end, we consider two different strategies of enlarging V_m : global p -enrichment and uniform h -refinement. We again consider $\varepsilon = 10^{-5}$, $\omega(x) \equiv 1$, $\alpha = 1$ and impose weak inflow boundary conditions on r_m .

p-enrichment: In terms of implementation, the simplest choice for the space V_m such that $\dim(V_m) > \dim(U_n)$ is a finite element space over the same grid with the polynomial degree increased globally by some integer Δp . Figure 7.6 shows how the overshoot reduces as we increase Δp . The difference in r_m for varying p_m is barely visible even though we have used a finer mesh for plotting in order to capture the behaviour of the higher modes and yet it has a significant effect on the approximation u_n .

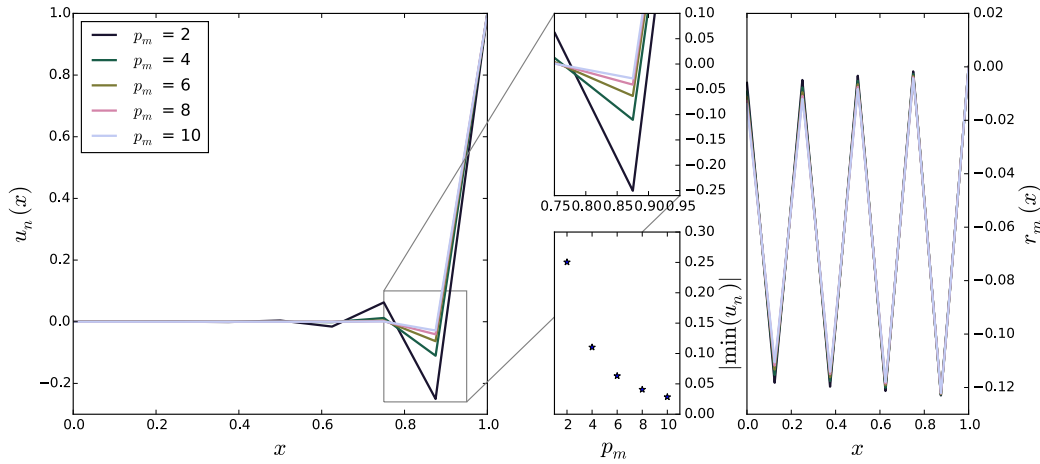


Figure 7.6 Example 7.1 with $\varepsilon = 10^{-6}$ and $q = 1.01$, $q' = 101$: Numerical approximations u_n (left) and r_m (right) with a uniform mesh consisting of 8 intervals using piecewise linear polynomials for u_n and varying polynomial degrees p_m for r_m . Top centre: Zoom-in showing over- and undershoots. Bottom centre: $|\min(u_n)|$ vs p_m . Right: Projection of r_m onto the space of piecewise linear functions with a uniform mesh consisting of 16 intervals using piecewise linear polynomials for u_n and varying polynomial degrees p_m for r_m

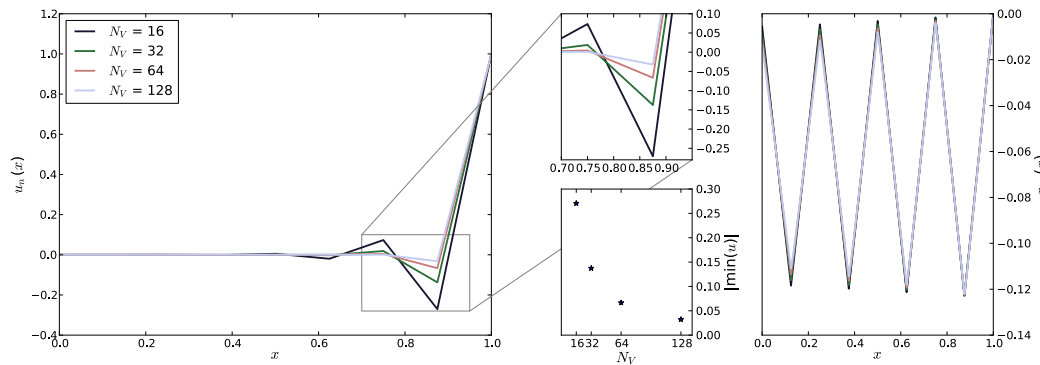


Figure 7.7 Example 7.1 with $q = 1.01$, $q' = 101$: Numerical approximations u_n (left) and r_m (right) with a uniform mesh consisting of 8 intervals using piecewise linear polynomials for u_n and piecewise linear polynomials for r_m on a refined mesh consisting of N_V intervals. Top centre: Zoom-in showing over- and undershoots. Bottom centre: $|\min(u_n)|$ vs N_V .

h-refinement: This time both U_n and V_m consist of piecewise linear polynomials. To ensure that $\dim(V_m) > \dim(U_n)$, we choose a refinement of the underlying mesh of the space U_n to construct the space V_m . This is the only numerical experiment in this chapter that cannot be implemented in FEniCS; instead, e.g., the C++ library Hermes2D [155] can be used. Figure 7.7 shows how the overshoot reduces as we refine the mesh for V_m ; the difference in r_m is again barely visible despite the significant effect on the approximation u_n .

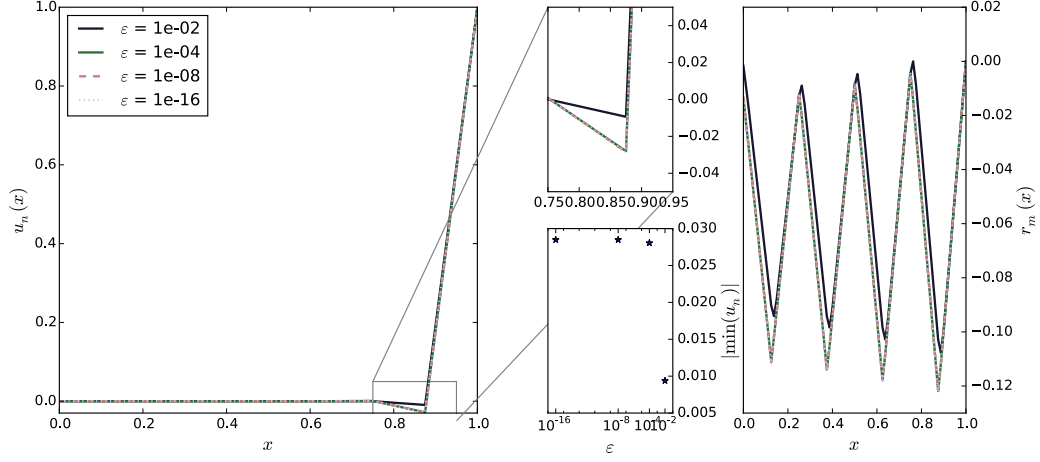


Figure 7.8 Example 7.1 with $q = 1.01$, $q' = 101$: Numerical approximations u_n (left) and r_m (right) with a uniform mesh consisting of 8 intervals using piecewise linear polynomials for u_n and polynomials of degree $p_m = 10$ for r_m for varying values of ε . Top centre: Zoom-in showing over- and undershoots. Bottom centre: $|\min(u_n)|$ vs ε .

7.2.2.4 Robustness in ε

As a final experiment in one dimension, we study how ε affects the undershoot in the approximation. We continue with the same setting as in the previous section, but keep V_m fixed as the space of piecewise polynomials of degree $p_m = 10$ on the same mesh as used for U_n and vary ε instead. Figure 7.8 shows that, although the undershoot is not the same for all ε leading to an under resolved layer, the undershoot seems to be approximately the same for all $\varepsilon \leq \varepsilon_0$ for some ε_0 . This can be traced back to the inflow boundary conditions on r_m — which are scaled with ε — if we look at r_m on the right in Figure 7.8. The method therefore seems to be robust in ε .

7.2.3 Vanishing Oscillations in Two Dimensions

In this section we explore whether the under- and overshoots in the approximation still vanish as $q \rightarrow 1$ if we apply the method to two-dimensional problems. To this end, we consider the Examples 7.2, 7.3 and 7.4 on different meshes for $\varepsilon = 10^{-6}$. We consider $\omega(x) \equiv 1$, $\alpha = 1$ and impose weak boundary conditions on r_m as described in Section 7.1.2. In one case we will also consider $\alpha = 0$ to compare the approximations for both choices of α . We will see that the overshoots disappear on certain meshes, but remain present on others. We will furthermore demonstrate that the approximation qualitatively behaves like the $L^q(\Omega)$ -best approximation of

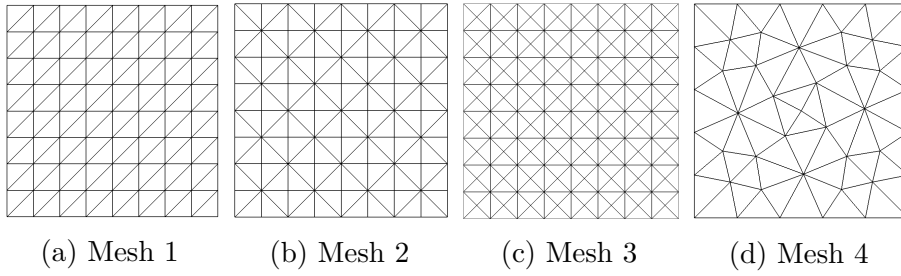
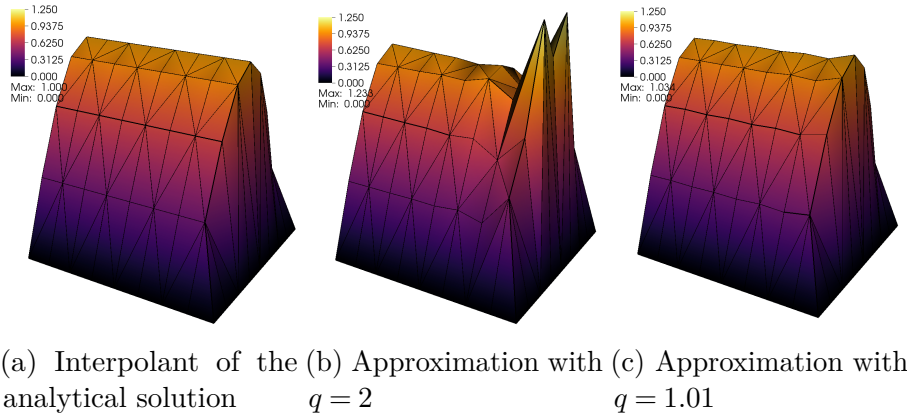


Figure 7.9 Four different meshes on the unit square

Figure 7.10 Approximation of the analytical solution of Example 7.2 (Eriksson-Johnson model problem) on Mesh 2 (see Figure 7.9) for $\varepsilon = 10^{-6}$.

the analytical solution and thus we can apply the observations in [94] to (a) predict on which meshes the overshoot will disappear as $q \rightarrow 1$ and (b) design meshes that have this property in specific situations.

7.2.3.1 The Eriksson-Johnson Model Problem

We consider the Eriksson-Johnson Model Problem (Example 7.2) on the four different meshes depicted in Figure 7.9. Note that these are the same meshes that are investigated in Chapter 4 (see also [94]).

First, we compare the approximations on Mesh 2 and Mesh 3 for $q = 2$ and $q = 1.01$ with the interpolant of the analytical solution in U_n . Figure 7.10 shows the approximations on Mesh 2; for $q = 2$ we clearly observe overshoots near the boundary layer (centre) compared to the interpolant of the analytical solution (left). On the other hand, for $q = 1.01$ we can see that the approximation approaches the interpolant of the analytical solution. Figure 7.11 shows the approximations on Mesh 3; in this case we can see that the overshoot is nearly the same for $q = 2$ and $q = 1.01$. From Chapter 4 we note that this is the qualitative behaviour that the L^q -best approximation exhibits.

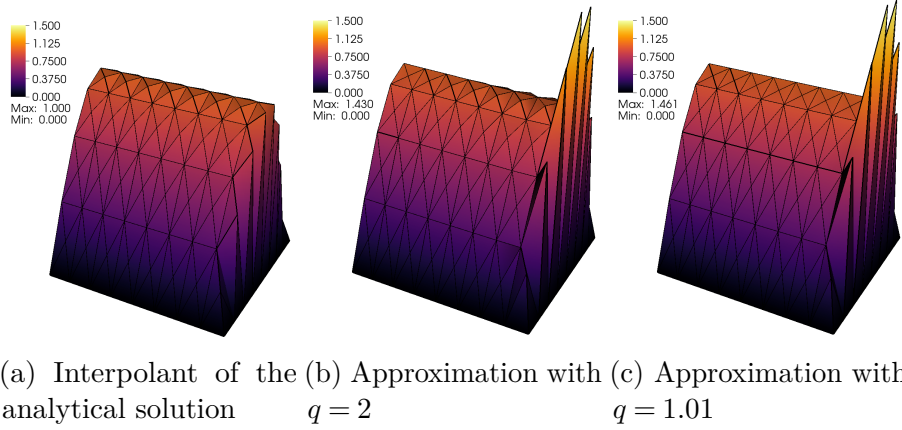


Figure 7.11 Approximation of the analytical solution of Example 7.2 (Eriksson-Johnson model problem) on Mesh 3 (see Figure 7.9) for $\varepsilon = 10^{-6}$.

With this in mind, we have numerically computed the $L^q(\Omega)$ -best approximation of the analytical solution for several choices of q and have also computed the finite element approximations with $\omega(x) = 1$ and both, $\alpha = 1$ and $\alpha = 0$, with weak boundary conditions imposed on r_m on the inflow boundary for the same values of q . Figure 7.12 shows the maximal error between the interpolant of the analytical solution and the approximation for the $L^q(\Omega)$ -best approximation as well as the error between the interpolant and the numerical approximations for $\alpha = 1$ and $\alpha = 0$. We can clearly see that the overshoot observed for our proposed scheme is very similar to the overshoot in the $L^q(\Omega)$ -best approximation. Therefore, understanding for which meshes discontinuities/layers are captured as sharply as the grid allows by the $L^1(\Omega)$ -best approximation can serve as an indicator regarding whether the over- and undershoots disappear as $q \rightarrow 1$. This is the subject of Chapter 4, where the L^q -best approximation of discontinuous functions is analysed in detail in certain cases.

7.2.3.2 Boundary Layer in a Corner of the Domain

In this section we consider Example 7.3 with $\mathbf{b} = (2, 1)^T$. We will see that on Mesh 2 in this case the overshoot does not disappear as $q \rightarrow 1$ for our method. This is due to the layer not only appearing along an edge of the unit square but also near the corner $(x, y) = (1, 1)$. With this in mind, we suggest an alternative mesh that is only a slight modification of Mesh 2 for which the overshoot disappears. Figure 7.13 shows the two versions of Mesh 2 that we have used and the corresponding approximations for $q = 2$ and $q = 1.01$ along the line through the interior nodes closest to the boundary $y = 1$. We can see that on Mesh 2, the overshoot near

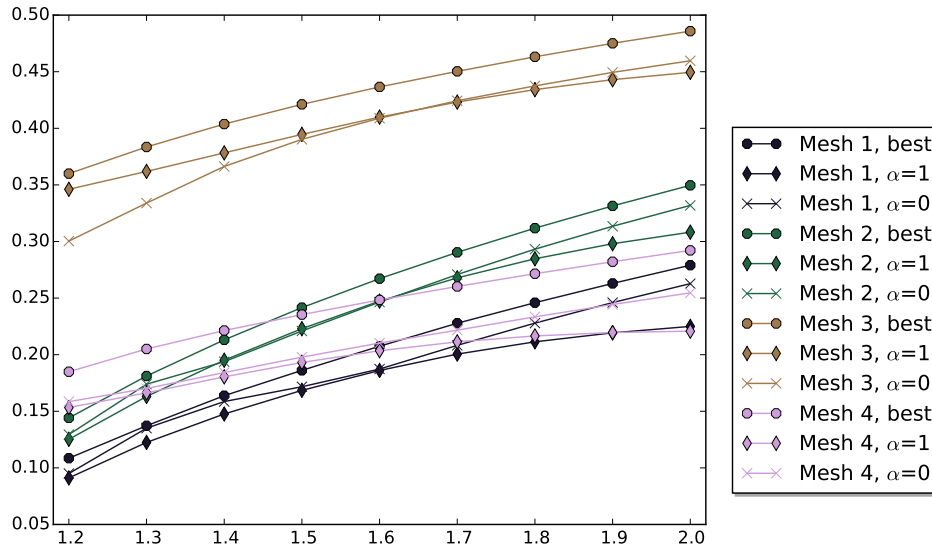


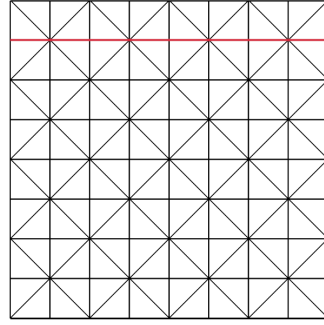
Figure 7.12 Example 7.2: $\max(u_n - \hat{u})$, where $\hat{u}$ is the interpolant of the analytical solution u in the space U_n , for the $L^q(\Omega)$ -best approximation and the finite element approximations with $\alpha = 1$ and $\alpha = 0$. See Figure 7.9 for Meshes 1-4.

the corner does not disappear as $q \rightarrow 1$, whereas it does reduce significantly away from the corner. For the modified version of Mesh 2, the overshoot disappears everywhere as $q \rightarrow 1$.

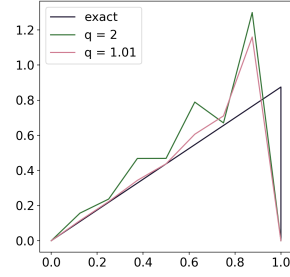
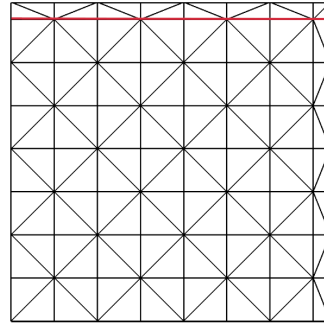
To understand the improvement when using this modified mesh, we focus on the interior node closest to the corner $(x, y) = (1, 1)$. Figure 7.14 shows all elements connected to this node. The elements with either a node or an edge on the boundary are marked in green; the elements separated from the boundary are blue. From [94] we can infer that the L^1 -best approximation does not exhibit overshoots if the volume of the green area is smaller or equal to the volume of the blue area. This is obviously violated on Mesh 2 with the green area being three times as large as the blue area. The modified version of the mesh is designed to satisfy this condition by changing the mesh as indicated in Figure 7.14.

7.2.3.3 Example with an Interior Layer

In this section we consider Example 7.4 and demonstrate that a mesh can be constructed such that the over- and undershoots both near the boundary layer and the interior layer can be eliminated as $q \rightarrow 1$. To this end, Figure 7.15 shows the approximation to Example 7.4 for $q = 2$ and $q = 1.1$ on two different meshes. The first mesh is designed such that the overshoot at the boundary layer disappears



(a) Mesh 2

(b) Exact solution u and approximation u_n along red line as indicated in Figure 7.13a for $q = 2$ and $q = 1.01$ 

(c) Modified version of Mesh 2

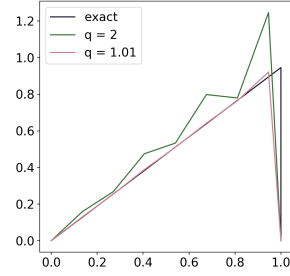
(d) Exact solution u and approximation u_n along red line as indicated in Figure 7.13c for $q = 2$ and $q = 1.01$

Figure 7.13 Example 7.3 with $\mathbf{b} = (2, 1)^T$: Mesh and approximation for $\varepsilon = 10^{-6}$ with $p_n = 1$ and $p_m = 8$.

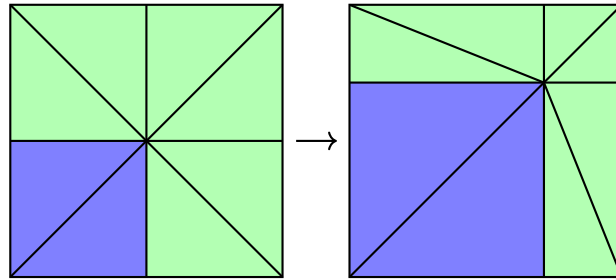
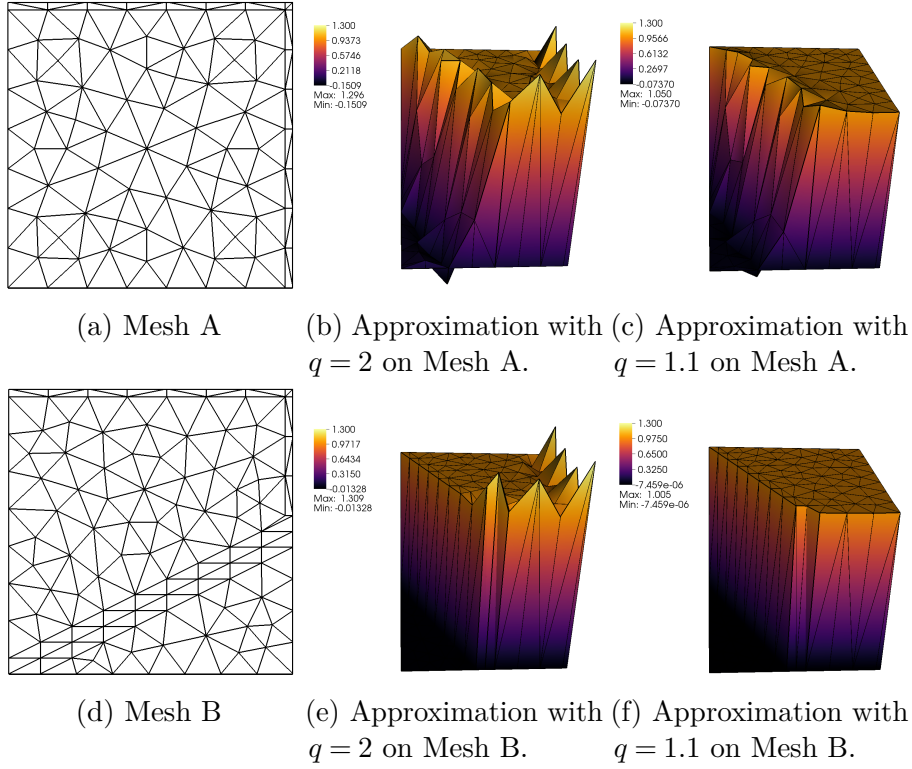


Figure 7.14 Modification of Mesh 2: Elements connected to the boundary are marked in green, elements separated from the boundary in blue.

as $q \rightarrow 1$; this is already clearly visible for $q = 1.1$. Along the interior layer, the overshoot also reduces significantly, but some over- and undershoots are still present for $q = 1.1$. The second mesh is additionally designed to align with the interior layer in such a way that we can expect the over- and undershoots to disappear as $q \rightarrow 1$. We can see, that this mesh even nearly eliminates the overshoot along the

Figure 7.15 Example 7.4 with $\varepsilon = 10^{-6}$.

interior layer for $q = 2$ and for $q = 1.1$ all over- and undershoots along both layers have essentially vanished.

Figure 7.16 shows the approximations for $q = 2$ and $q = 1.1$ if $\mathbf{b}$ is not perfectly aligned with the mesh, i.e., $\mathbf{b} = (2, 1.2)^T$ and $\mathbf{b} = (2, 1.06)^T$. The difference between these two choices for $\mathbf{b}$ is that in the latter case the layer is still contained in between the two lines parallel to the line $(x, y) = t(2, 1)$, whereas in the former case it is not. In both cases, we clearly observe overshoots along the interior layer for $q = 2$ which are reduced for $q = 1.1$. Both, for $q = 2$ and $q = 1.1$ the overshoot reduces slowly the closer $\mathbf{b}$ is to $(2, 1)^T$. For $\mathbf{b} = (2, 1.06)^T$ and $q = 1.1$ the overshoot is barely visible in Figure 7.16, but comparing the maximum and minimum values of the approximations reveals that the over- and undershoots are similar to the case $\mathbf{b} = (2, 1.2)$ in magnitude.

7.3 Concluding Remarks

The numerical results presented in this chapter demonstrate that our proposed numerical scheme can be used to eliminate Gibbs phenomena in the numerical approximation in the limit as $q \rightarrow 1$ provided the mesh is suitably designed. In

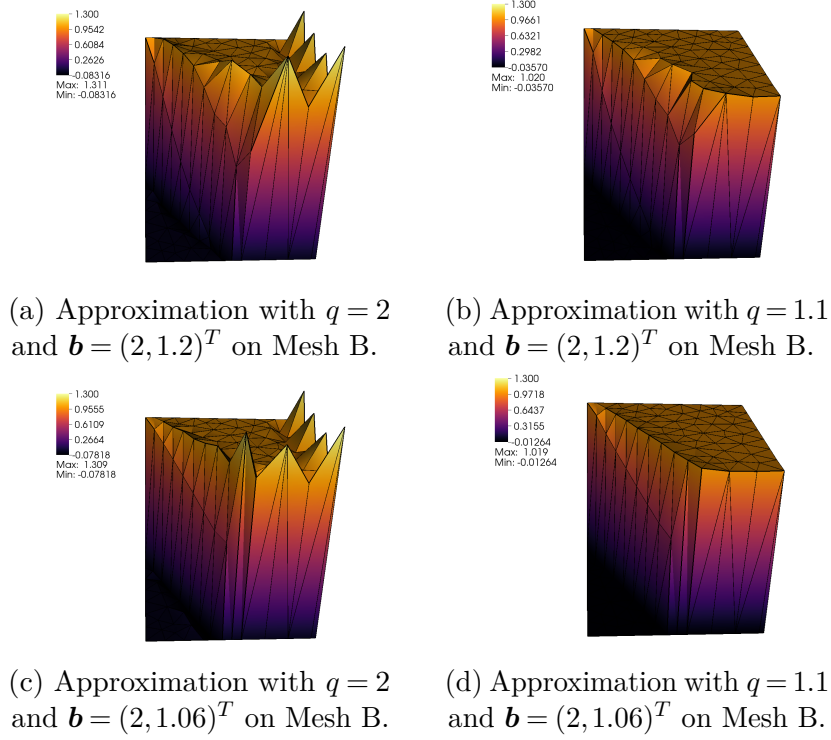


Figure 7.16 Example 7.4 with $\varepsilon = 10^{-6}$, $\mathbf{b} = (2, 1.2)^T$ (top) and $\mathbf{b} = (2, 1.06)^T$ (bottom) on Mesh B, cf., Figure 7.15.

practice one would therefore aim to apply this method either for a fixed q sufficiently close to 1 such that over- and undershoots become negligible or to apply this method directly with $q = 1$. In the latter case some of the theory is no longer applicable, in particular minimizers are not necessarily unique. The crucial ingredient to make this method feasible in practice is an efficient non-linear solver that is either directly designed for the case $q = 1$ or robust in q as $q \rightarrow 1$. Since uniqueness is in general lost as $q \rightarrow 1$, it can be assumed that such a solver incorporates a selection principle for choosing the ‘right’ solution. Generally speaking, the closer q is to 2 the easier it is to solve the non-linear system. So another possible strategy would be to use adaptive mesh refinement together with suitable mesh design strategies with the aim of creating a mesh that allows for the oscillations to become negligible for q much closer to 2.

Another challenge in terms of computational feasibility is the increased number of degrees of freedom. The DPG methodology addresses this problem by using broken test spaces in order to reduce inverting the Riesz map to approximately inverting the Riesz map locally on each element. A similar strategy could be applied to our proposed scheme, though extra care must be taken due to the added complexity stemming from the non-linearity of the duality mapping. Furthermore,

the number of additional degrees of freedom can likely be reduced considerably by adaptively constructing the test space. Some strategies of addressing the above mentioned challenges are also outlined as areas of future research in Section 8.2.

It should be noted that it is well known that higher-order monotonicity preserving methods can be expected to be non-linear, cf., e.g., [27]; thus, the non-linear nature of our approach by no means diminishes its potential. In [27], Burman and Ern analyse a non-linear shock-capturing scheme for the convection-diffusion-reaction equation that relies on adding two stabilization terms – a linear least-squares term and a non-linear shock-capturing term – to the standard Galerkin finite element formulation. They are able to rigorously establish a discrete maximum principle guaranteeing the absence of unphysical oscillations in the numerical approximation. Hence, our proposed scheme and the shock capturing scheme in [27] aim to provide a remedy for the same problem and agree in the fact that in both cases a non-linear scheme is deemed necessary. However, they differ fundamentally in the approach to designing a method. Our scheme relies on a very general residual minimization approach, that is in principle applicable to any (linear) variational problem – provided a robust formulation in suitable spaces is known. In Section 8.3 we will furthermore demonstrate its potential for non-linear hyperbolic equations for which an L^1 -setting is very natural for the analysis of the underlying PDE problem. The key advantage of this approach is that the scheme naturally follows from a robust formulation of the variational problem. In the case of the convection-diffusion-reaction equation, the robust formulation is an open problem and hence the design of the scheme had to be guided by numerical experiments instead. The approach in [27] on the other hand follows a long tradition of designing problem-specific stabilization and shock-capturing terms and has the clear advantage that computational feasibility is achieved more easily, while for our scheme this remains an open problem. The difficulty of this approach lies in designing the appropriate problem-specific stabilization terms relying on user-specific constants.

Chapter 8

Conclusions and Future Directions

The goal of this work has been to design a non-linear Petrov-Galerkin method for the convection-diffusion-reaction equation that can eliminate unphysical oscillations. To this end, we have considered a generalization of the DPG framework based on residual minimization that was originally introduced in [137]. At least in an abstract sense, this approach outlines how to design a numerical method that leads to a quasi-best approximation of the solution in a space of choice, provided the continuous problem is well-posed in a suitable sense. In this thesis we have focussed on applying this method to the convection-diffusion-reaction equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$ addressing questions regarding the inf-sup conditions and the well-posedness of the continuous problem (cf., Chapter 5), the L^q -best approximation of discontinuities in finite element spaces (cf., Chapter 4), the extension of the DPG framework to more general Banach spaces (cf., Chapter 6), and, finally the design of the method for the convection-diffusion-reaction equation (cf., Chapter 7). In this chapter we summarize the achievements presented in this thesis in Section 8.1. Then, in Section 8.2 we will outline several research questions that naturally arise from the results presented in this thesis. Finally, in Section 8.3, we present some preliminary results for one of these research questions, namely the extension of the method to non-linear PDEs.

8.1 Accomplishments

- **Characterization of L^q -best approximations:** In Chapter 4, we have investigated Gibbs phenomena in the L^q -best approximation of discontinuities

within finite element spaces. Using selected examples, we have proven that the Gibbs phenomenon can be eliminated as $q \rightarrow 1$ on certain meshes. However, we have seen that there exist non-uniform meshes in one dimension that lead to Gibbs phenomena even if $q = 1$. In two dimensions, even some uniform, structured meshes lead to Gibbs phenomena if $q = 1$. Nonetheless, the magnitude of the oscillations decreases as $q \rightarrow 1$ on all meshes.

The computational examples presented in Chapter 4 confirm the theoretical results. Moreover, we have seen that similar observations can be made for more general examples. Furthermore, we have demonstrated that the Gibbs phenomenon cannot be eliminated on certain meshes under mesh refinement. For the final computational example, we have been able to establish a link between the structure of the mesh near the discontinuity and the magnitude of the overshoot at the nodes. This observation suggests that the oscillations can be eliminated in the limit as q tends to 1 if the mesh structure near the discontinuity is suitably adjusted. Indeed, this has been used to design meshes for the non-linear Petrov- Galerkin method for the convection-diffusion-reaction equation presented in Chapter 7.

- **Well-posedness of the convection-diffusion-reaction equation:** In Chapter 5 we have presented a well-posedness theory for the convection-diffusion-reaction equation in $W_0^{1,q}(\Omega)$ - $W_0^{1,q'}(\Omega)$, $1 < q < \infty$, based on directly proving the inf-sup conditions. The significance of this for the non-linear Petrov-Galerkin method is that the continuity constant and the inf-sup constant associated with the variational problem determine the scaling of the constant in the quasi-best approximation result that has been proven for the method in [137]. Hence, estimating these constants is an important step in the analysis of the method. Even though the estimates obtained in Chapter 5 can be expected to be suboptimal, they form an important first step towards establishing sharp estimates.
- **Generalization to Banach spaces of concepts related to broken spaces and broken variational forms:** While the generalization of the minimum residual framework of the DPG method to Banach spaces was first introduced in [137], this work does not include considerations regarding broken test spaces and broken variational forms as described in Section 6.5.1. Hence, the generalization to Banach spaces of the variational formulation in *broken* spaces and the generalization of the related result showing under which circumstances the broken variational formulation inherits well-

posedness from its unbroken counterpart are original to this thesis, laying important groundwork for future research regarding the use of discontinuous test spaces.

- **Design of a method for the convection-diffusion-reaction equation:**

In Chapter 7, we have demonstrated how to design a finite element method for the convection-diffusion-reaction problem that in the convection-dominated case yields solutions that qualitatively behave like the $L^q(\Omega)$ -best approximation of the analytical solution. The choice of the norm on the test space, in our case $W_0^{1,q'}(\Omega)$, determines the corresponding duality mapping on the test space and thus crucially determines the method. Hence, we have considered various choices for the test norm. Moreover, we have considered both Dirichlet boundary conditions in the test space as well as weakly imposed boundary conditions following the strategy suggested in [39] to improve robustness of the method. In one dimension, we have compared the different choices for the test norm, as well as the two choices for the boundary conditions regarding robustness in the diffusion parameter and the Gibbs-reducing property of the method as $q \rightarrow 1$. This led to a scheme that in the convection-dominated case yields solutions that qualitatively behave like the $L^q(\Omega)$ -best approximation of the analytical solution. For this scheme, we have demonstrated convergence at the optimal rates numerically and compared different choices of the discrete test space. Furthermore, we have discussed the limit case where the diffusion parameter is zero and established the connection to the method for approximating first-order PDEs by Guermond, cf., [82].

An important implication of the connection to the $L^q(\Omega)$ -best approximation is that on meshes where the $L^1(\Omega)$ -best approximation does not contain any over- and undershoots, these can also be avoided by taking $q \rightarrow 1$ in our proposed method. Indeed, the final two examples presented in Chapter 7 demonstrate that it is possible to adjust meshes to specific situations in order to eliminate over- and undershoots. Hence, we can use the insights from Chapter 4 to design suitable meshes. Note that even in one dimension the over- and undershoots will not disappear for certain non-uniform meshes. Roughly speaking, meshes where the elements near the discontinuity or layer present in the analytical solution are smaller than at a distance from those features are more favourable. Additionally, it is desirable that all interior nodes that are closest to such a feature are aligned roughly in parallel with the feature; for further details, we refer to Chapter 4. This is similar to the

observations used to design the modified version of the computational mesh in the final two examples. This suggests that if the mesh is refined in a suitable way near the layer, we can eliminate over- and undershoots. Note, however, that even though we modified the mesh in the final example in a way that resembles refinement towards the boundary layer, the boundary layer clearly remains underresolved, even on the modified mesh. It is therefore not at all necessary to resolve the layer to eliminate the oscillations in this case.

8.2 Future Work

It is the nature of research that every discovery and every achievement opens a door to more unanswered questions. In this section we outline several areas of work that very naturally arise from the research presented in this thesis.

- **A robust inf-sup condition:** We point out in Remark 5.5 that the bound for the inf-sup constant in Chapter 5 is not robust in the diffusion parameter ε . Moreover, the dependence on ε , for $\varepsilon \ll 1$, is less favourable than in the bound for the case $q = 2$ relying on the standard coercivity proof.

Indeed, cf., Remark 5.9, if $q = 2$ it is possible to obtain ε -independent estimates for the inf-sup constant and the continuity constant if we choose the following norms for the trial space $U = H_0^1(\Omega)$ and the test space $V = H_0^1(\Omega)$

$$\begin{aligned} \|u\|_U &:= c_0 \|u\|_{L^2(\Omega)} + \varepsilon |u|_{H^1(\Omega)} \\ \|v\|_V &:= c_0 \|v\|_{L^2(\Omega)} + \varepsilon |v|_{H^1(\Omega)} + \|\nabla \cdot (\mathbf{b}v)\|_{U'}, \end{aligned}$$

respectively, where we consider $\nabla \cdot (\mathbf{b}v) \in L^2(\Omega)$ as an element in $U' = H^{-1}(\Omega) \supset L^2(\Omega)$ with its norm in U' given by

$$\|\nabla \cdot (\mathbf{b}v)\|_{U'} = \sup_{u \in U} \frac{\langle \nabla \cdot (\mathbf{b}v), u \rangle_{U', U}}{\|u\|} = \frac{\int_{\Omega} \nabla \cdot (\mathbf{b}v) u \, d\mathbf{x}}{\|u\|_U}.$$

A proof of this result is presented in [164, Section 4.4.1]. There are essentially three steps to the proof: firstly, it is easy to prove that the continuity constant with respect to the norms $\|\cdot\|_U$ and $\|\cdot\|_V$ is independent of ε ; secondly, an ε -independent estimate for the inf-sup constant can be proven using the norm $\|\cdot\|_U$ on V ; thirdly, this estimate can be used to obtain an ε -independent estimate for the inf-sup constant with respect to the norm $\|\cdot\|_V$ as well. The first and the third step can easily be extended to $q \neq 2$. Generalizing

the second step, however, is more challenging since this step relies on the standard estimate for the coercivity constant. In Chapter 5 we have been able to obtain an estimate for this inf-sup constant for $q \neq 2$; however, the resulting constant is ε dependent and suboptimal even for $q = 2$. Hence, it remains an open problem to improve this estimate of the inf-sup constant such that it is optimal for $q = 2$.

- **Non-linear solvers:** As noted in Remark 3.27, the duality mapping $J_{W_0^{1,q'}}$ essentially defines the q' -Laplace operator. More generally, the duality mappings on subspaces of $L^{q'}(\Omega)$ have non-linear structures similar to the q' -Laplace operator. Hence, inverting the duality mapping approximately becomes more difficult as $q' \rightarrow \infty$, i.e., when $q \rightarrow 1$, which is the scenario that we have been considering in this thesis. Developing efficient non-linear solvers for these operators with large q' is therefore crucial for making the method feasible in practice.
- **Adaptivity in the trial space:** Our method is given as a non-linear mixed system of the following form: find $(r_m, u_n) \in V_m \times U_n$ such that

$$\begin{aligned} \langle J_V(r_m), v_m \rangle_{V',V} + \langle Bu_n, v_m \rangle_{V',V} &= \langle f, v_m \rangle_{V',V}, \quad \forall v_m \in V_m, \\ \langle Bw_n, r_m \rangle_{V',V} &= 0, \quad \forall w_n \in U_n, \end{aligned}$$

where J_V is a so-called duality mapping. If V is a Hilbert space, then J_V is simply the Riesz map. Solving the first line for r_m is essentially equivalent to approximately inverting J_V yielding $r_m \approx r = J_V^{-1}(f - Bu_n)$. Thus, r_m represents the error in the residual. In the Hilbert space setting, the variable r_m has already successfully been employed as an error estimator to drive adaptivity, cf., e.g., [57, 39]. This approach can be generalized to the more general Banach space setting if the global error represented by r_m can be related to a local error in a suitable way.

- **(Adaptive) mesh design:** In Chapter 7 we have seen that the structure of the mesh near discontinuities and thin layers crucially determines whether unphysical oscillations can be eliminated as $q \rightarrow 1$ using our method. Hence, strategies have to be developed for designing suitable meshes and automating this process. In Chapter 7 we have used a priori knowledge of the specific problem, i.e., the location of the layer, to design a mesh. Generally, the location of the layer may not be known a priori, or, in the case of a non-linear time dependent problems, may change with time. This suggests incorporating

requirements on the structure of the mesh near layers into adaptive refinement strategies in order to eliminate Gibbs phenomena in more general situations.

- **More general characterizations of L^q -best approximations:** We have been able to employ the results regarding L^q -best approximations presented in Chapter 4 in order to design meshes in Chapter 7 that lead to a finite element approximation with vanishing oscillations as $q \rightarrow 1$. Hence, obtaining generalizations of the results in Chapter 4 would be useful for developing strategies to design meshes that ensure that our method can be used to eliminate oscillations. In one dimension we have been able to formulate a very general sufficient condition on the mesh guaranteeing vanishing oscillations, cf., Theorem 4.1. To the best of our knowledge, it remains an open problem whether this condition is also necessary. Furthermore, the results in two dimensions are restricted to very specific examples.
- **Choosing the test space adaptively:** In Chapter 7, we have seen that the choice of the discrete test space is not only crucial for well-posedness of the discrete problem, but also for eliminating over- and undershoots in the limit $q \rightarrow 1$. We usually choose a test space V_m for a fixed trial space U_n such that $U_n \subset V_m$, where V_m is obtained by increasing the polynomial degree on each element by $\Delta p \geq 1$. In one dimension, we have shown that similar results can be achieved by uniformly refining the mesh to obtain V_m from U_n .

In the context of *linear* Petrov-Galerkin methods for transport or convection-diffusion equations, various strategies have been employed to construct V_m from U_n , including global p -enrichment (e.g., [39]), Shishkin type sub-meshes (cf., [141]) or combinations of p -enrichment and h -refinement (cf., e.g., [49]). Cohen, Dahmen and Welper, cf. [47], go a step further by introducing a double adaptivity approach where the trial space is refined adaptively and within each refinement step the space V_m is constructed adaptively. The main advantage of this approach is that the inner adaptivity loop circumvents the necessity of a priori satisfying a Fortin condition, cf., [58], which is in general difficult to guarantee.

The numerical examples presented in Section 7.2.2.3 furthermore suggest that the Gibbs-reducing property of our scheme usually requires a larger space V_m than is required for satisfying a Fortin-condition when $q = 2$, potentially leading to a substantially larger number of degrees of freedom. Hence, it

is essential to develop methods for constructing discrete spaces V_m with a minimal number of degrees of freedom without losing the Gibbs-reducing property of the method. It can be expected that the ideal choice for V_m is highly dependent on the specific problem which suggests using adaptive strategies for designing the space, such as a generalization of the double adaptivity approach (cf., [47, 58]) to Banach spaces.

- **Discontinuous test space:** As obtained in Chapter 6, the test space V and its discrete counterpart V_m are chosen such that they contain functions that are discontinuous across inter-element boundaries in the context of DPG methods. This warrants the introduction of additional unknowns on the interface in the variational formulation and yields a well-posed discrete system if the spaces satisfy suitable conditions. The main advantage of this is that the matrix associated with approximately inverting the Riesz map becomes block-diagonal and hence cheap to invert, cf., Section 6.5.3. Formally, it is easily possible to extend this approach to L^q -type Banach spaces and, as pointed out in Section 6.5.3, approximately inverting the duality mapping reduces to solving a series of local non-linear problems in this case which would increase the efficiency of the method immensely. Hence, this is a very natural extension of the work presented in this thesis.
- **NPG for non-linear equations:** The results in Chapter 5 illustrate the potential of our scheme for the approximation of more challenging, non-linear PDEs with analytical solutions that contain sharp layers or discontinuities. In Section 8.3, we will present a few preliminary results concerning an extension of the method to non-linear operators.

8.3 Extending the Non-linear Petrov-Galerkin Method to Non-linear Problems

In this section, we present preliminary results for an extension to non-linear problems that is again based on residual minimization. For some results regarding the Hilbert-space DPG method for non-linear equations, see [35].

Let U and V again be two reflexive and strictly convex Banach spaces. We start with the following abstract non-linear variational problem: find $u \in U$ such

that

$$\mathcal{N}(u, v) = \langle \ell, v \rangle_{V', V} \quad \forall v \in V, \quad (8.1)$$

where $\mathcal{N} : U \times V \rightarrow \mathbb{R}$ is a semi-linear form such that $\mathcal{N}(w, \cdot) : V \rightarrow \mathbb{R}$ is linear for all $w \in U$. We can therefore define a non-linear operator $N : U \rightarrow V'$ by

$$\langle N(w), v \rangle_{V', V} := \mathcal{N}(w, v), \quad v \in V, w \in U.$$

Thus, (8.1) is equivalent to the following problem: find $u \in U$ such that

$$N(u) = \ell \quad \text{in } V'.$$

Given a finite dimensional subspace $U_n \subset U$ we define the following minimization problem: find $u_n \in U_n$ such that

$$u_n = \arg \min_{w_n \in U_n} \|\ell - N(w_n)\|_{V'}.$$

This is of course equivalent to minimizing $H_V : U \rightarrow \mathbb{R}$ given by

$$H_V(w) = \psi(\|\ell - N(w)\|_{V'}),$$

where $\psi(t) = \int_0^t \varphi(s) ds$ for a weight function φ . Setting the derivative of H_V to zero yields the following fully non-linear Petrov-Galerkin method:

$$\begin{aligned} 0 = H'_V(u_n)(w_n) &= \langle J_{V'}^\varphi(\ell - N(u_n)), -N'(u_n)(w_n) \rangle_{V'', V'} \\ &= \langle -N'(u_n)(w_n), \left(J_V^{\varphi^{-1}}\right)^{-1}(\ell - N(u_n)) \rangle_{V', V} \quad \forall w_n \in U_n, \end{aligned}$$

where $J_{V'}^\varphi : V' \rightarrow V''$ again denotes the duality mapping on V' with weight φ . Introducing an auxiliary variable $r = \ell - N(u_n)$ and restating the problem as a mixed problem we obtain the following: find $(r, u_n) \in V \times U_n$ such that

$$\begin{aligned} \langle J_V^{\varphi^{-1}}(r), v \rangle_{V', V} + \langle N(u_n), v \rangle_{V', V} &= \langle \ell, v \rangle_{V', V} & \forall v \in V, \\ \langle N'(u_n)(w_n), r \rangle_{V', V} &= 0 & \forall w_n \in U_n. \end{aligned}$$

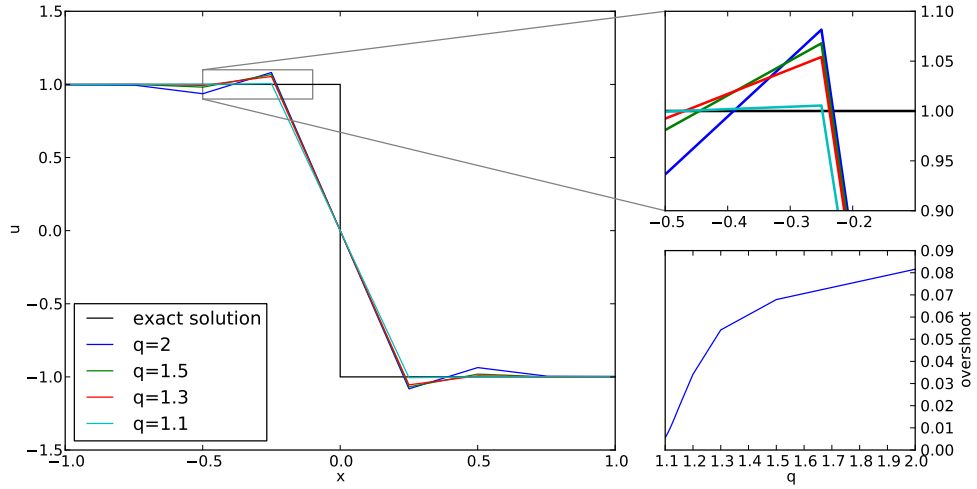


Figure 8.1 Numerical approximation of (8.4) for varying values of q .

Considering a finite dimensional subspace $V_m \subset V$ yields the corresponding inexact mixed method: find $(r_m, u_n) \in V_m \times U_n$ such that

$$\langle J_V^{\varphi^{-1}}(r_m), v_m \rangle_{V', V} + \langle N(u_n), v_m \rangle_{V', V} = \langle \ell, v_m \rangle_{V', V} \quad \forall v_m \in V_m, \quad (8.3a)$$

$$\langle N'(u_n)(w_n), r_m \rangle_{V', V} = 0 \quad \forall w_n \in U_n. \quad (8.3b)$$

Note that both equation (8.3a) and equation (8.3b) are now non-linear in (r_m, u_n) .

As a first example we will consider the following version of Burgers' equation:

$$\partial_x \left(\frac{1}{2} u^2 \right) = 0 \quad \text{in } \Omega = (-1, 1), \quad u(-1) = 1, \quad u(1) = -1, \quad (8.4)$$

and the corresponding variational problem: find $u \in L^{2q}(\Omega)$ such that

$$-\int_{\Omega} \frac{1}{2} u^2 \partial_x v \, dx + \int_{\partial\Omega} \mathcal{H}(u^+, u^-, \mathbf{n}) v \, ds = 0 \quad \forall v \in W^{1, q'}(\Omega),$$

where $\mathcal{H}(u^+, u^-, \mathbf{n})$ is a numerical flux function such as the Lax-Friedrichs flux, cf., Section 2.2.1. Note that for this problem a steady state solution (i.e., a solution such that $\partial_t u = 0$ which we have already omitted in the above formulation) only exists for certain choices of the boundary conditions including the boundary conditions given above. A first numerical test is shown in Figure 8.1. We can see that the oscillations disappear as q approaches 1, which illustrates the potential of our method for non-linear problems.

Bibliography

- [1] L. R. Abrahamsson, H. B. Keller, and H. O. Kreiss. Difference approximations for singular perturbations of systems of ordinary differential equations. *Numer. Math.*, 22:367–391, 1974. doi:[10.1007/BF01436920](https://doi.org/10.1007/BF01436920).
- [2] V. Adolfsson. L^p -integrability of the second order derivatives of Green potentials in convex domains. *Pacific J. Math.*, 159(2):201–225, 1993. doi:[10.2140/pjm.1993.159.201](https://doi.org/10.2140/pjm.1993.159.201).
- [3] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.
- [4] D. N. d. G. Allen and R. V. Southwell. Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. *Quart. J. Mech. Appl. Math.*, 8:129–145, 1955. doi:[10.1093/qjmam/8.2.129](https://doi.org/10.1093/qjmam/8.2.129).
- [5] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. The FEniCS project version 1.5. *Archive of Numerical Software*, 3(100), 2015. doi:[10.11588/ans.2015.100.20553](https://doi.org/10.11588/ans.2015.100.20553).
- [6] H. W. Alt. *Linear functional analysis*. Universitext. Springer-Verlag London, Ltd., London, 2016. ISBN 978-1-4471-7279-6; 978-1-4471-7280-2. doi:[10.1007/978-1-4471-7280-2](https://doi.org/10.1007/978-1-4471-7280-2). An application-oriented introduction, Translated from the German edition by Robert Nürnberg.
- [7] D. N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19(4):742–760, 1982. doi:[10.1137/0719052](https://doi.org/10.1137/0719052).
- [8] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2001/02. doi:[10.1137/S0036142901384162](https://doi.org/10.1137/S0036142901384162).
- [9] I. Babuška. Error-bounds for finite element method. *Numer. Math.*, 16: 322–333, 1970/1971. doi:[10.1007/BF02165003](https://doi.org/10.1007/BF02165003).
- [10] N. S. Bahvalov. On the optimization of the methods for solving boundary value problems in the presence of a boundary layer. *Ž. Vyčisl. Mat i Mat. Fiz.*, 9:841–859, 1969. doi:[10.1016/0041-5553\(69\)90038-X](https://doi.org/10.1016/0041-5553(69)90038-X). In Russian.
- [11] C. Baiocchi, F. Brezzi, and L. P. Franca. Virtual bubbles and Galerkin-least-squares type methods (Ga.L.S.). *Comput. Methods Appl. Mech. Engrg.*, 105 (1):125–141, 1993. doi:[10.1016/0045-7825\(93\)90119-I](https://doi.org/10.1016/0045-7825(93)90119-I).

- [12] W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2003.
- [13] R. E. Bank and H. Yserentant. On the H^1 -stability of the L_2 -projection onto finite element spaces. *Numer. Math.*, 126(2):361–381, 2014. doi:[10.1007/s00211-013-0562-4](https://doi.org/10.1007/s00211-013-0562-4).
- [14] A. E. Berger, J. M. Solomon, M. Ciment, S. H. Leventhal, and B. C. Weinberg. Generalized OCI schemes for boundary layer problems. *Math. Comp.*, 35(151):695–731, 1980. doi:[10.2307/2006188](https://doi.org/10.2307/2006188).
- [15] K. S. Bey and J. T. Oden. hp -version discontinuous Galerkin methods for hyperbolic conservation laws. *Comput. Methods Appl. Mech. Engrg.*, 133(3-4):259–286, 1996. doi:[10.1016/0045-7825\(95\)00944-2](https://doi.org/10.1016/0045-7825(95)00944-2).
- [16] I. P. Boglaev. An approximate solution of a nonlinear boundary value problem with a small parameter multiplying the highest derivative. *Zh. Vychisl. Mat. i Mat. Fiz.*, 24(11):1649–1656, 1758, 1984. In Russian.
- [17] C. L. Bottasso, S. Micheletti, and R. Sacco. The discontinuous Petrov-Galerkin method for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 191(31):3391–3409, 2002. doi:[10.1016/S0045-7825\(02\)00254-2](https://doi.org/10.1016/S0045-7825(02)00254-2).
- [18] C. L. Bottasso, S. Micheletti, and R. Sacco. A multiscale formulation of the discontinuous Petrov-Galerkin method for advective-diffusive problems. *Comput. Methods Appl. Mech. Engrg.*, 194(25-26):2819–2838, 2005. doi:[10.1016/j.cma.2004.07.024](https://doi.org/10.1016/j.cma.2004.07.024).
- [19] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, third edition, 2007. doi:[10.1017/CBO9780511618635](https://doi.org/10.1017/CBO9780511618635).
- [20] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002. doi:[10.1007/978-1-4757-3658-8](https://doi.org/10.1007/978-1-4757-3658-8).
- [21] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.
- [22] F. Brezzi and A. Russo. Choosing bubbles for advection-diffusion problems. *Mathematical Models and Methods in Applied Sciences*, 4(4):571–587, 1994. doi:[10.1142/S0218202594000327](https://doi.org/10.1142/S0218202594000327).
- [23] D. Broersen and R. Stevenson. A robust Petrov-Galerkin discretisation of convection-diffusion equations. *Comput. Math. Appl.*, 68(11):1605–1618, 2014. doi:[10.1016/j.camwa.2014.06.019](https://doi.org/10.1016/j.camwa.2014.06.019).
- [24] D. Broersen and R. P. Stevenson. A Petrov-Galerkin discretization with optimal test space of a mild-weak formulation of convection-diffusion equations in mixed form. *IMA J. Numer. Anal.*, 35(1):39–73, 2015.
- [25] J. M. Burgers. A mathematical model illustrating the theory of turbulence. In *Advances in Applied Mechanics*, pages 171–199. Academic Press, Inc., New York, N. Y., 1948. edited by Richard von Mises and Theodore von Kármán,.

- [26] E. Burman. A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty. *SIAM J. Numer. Anal.*, 43(5): 2012–2033, 2005. doi:[10.1137/S0036142903437374](https://doi.org/10.1137/S0036142903437374).
- [27] E. Burman and A. Ern. Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. *Math. Comp.*, 74(252):1637–1652, 2005. doi:[10.1090/S0025-5718-05-01761-8](https://doi.org/10.1090/S0025-5718-05-01761-8).
- [28] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193(15-16):1437–1453, 2004. doi:[10.1016/j.cma.2003.12.032](https://doi.org/10.1016/j.cma.2003.12.032).
- [29] E. Burman, M. A. Fernández, and P. Hansbo. Continuous interior penalty finite element method for Oseen’s equations. *SIAM J. Numer. Anal.*, 44(3): 1248–1274, 2006. doi:[10.1137/040617686](https://doi.org/10.1137/040617686).
- [30] Z. Q. Cai and S. McCormick. On the accuracy of the finite volume element method for diffusion equations on composite grids. *SIAM J. Numer. Anal.*, 27(3):636–655, 1990. doi:[10.1137/0727039](https://doi.org/10.1137/0727039).
- [31] Z. Q. Cai, J. Mandel, and S. McCormick. The finite volume element method for diffusion equations on general triangulations. *SIAM J. Numer. Anal.*, 28(2):392–402, 1991. doi:[10.1137/0728022](https://doi.org/10.1137/0728022).
- [32] A. Cangiani, Z. Dong, E. H. Georgoulis, and P. Houston. *hp-Version Discontinuous Galerkin Methods on Polygonal and Polyhedral Meshes*. SpringerBriefs in Mathematics. Springer, 2017.
- [33] P. Cantin. Well-posedness of the scalar and the vector advection-reaction problems in Banach graph spaces. *C. R. Math. Acad. Sci. Paris*, 355(8): 892–902, 2017. doi:[10.1016/j.crma.2017.07.009](https://doi.org/10.1016/j.crma.2017.07.009).
- [34] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. Breaking spaces and forms for the DPG method and applications including Maxwell equations. *Comput. Math. Appl.*, 72(3):494–522, 2016. doi:[10.1016/j.camwa.2016.05.004](https://doi.org/10.1016/j.camwa.2016.05.004).
- [35] C. Carstensen, P. Bringmann, F. Hellwig, and P. Wriggers. Nonlinear discontinuous Petrov-Galerkin methods. *Numer. Math.*, 139(3):529–561, 2018. doi:[10.1007/s00211-018-0947-5](https://doi.org/10.1007/s00211-018-0947-5).
- [36] P. Causin and R. Sacco. A discontinuous Petrov-Galerkin method with Lagrangian multipliers for second order elliptic problems. *SIAM J. Numer. Anal.*, 43(1):280–302, 2005. doi:[10.1137/S0036142903427871](https://doi.org/10.1137/S0036142903427871).
- [37] P. Causin, R. Sacco, and C. L. Bottasso. Flux-upwind stabilization of the discontinuous Petrov-Galerkin formulation with Lagrange multipliers for advection-diffusion problems. *M2AN Math. Model. Numer. Anal.*, 39(6): 1087–1114, 2005. doi:[10.1051/m2an:2005050](https://doi.org/10.1051/m2an:2005050).
- [38] J. Chan, L. Demkowicz, and R. Moser. A DPG method for steady viscous compressible flow. *Comput. & Fluids*, 98:69–90, 2014.
- [39] J. Chan, J. A. Evans, and W. Qiu. A dual Petrov-Galerkin finite element method for the convection-diffusion equation. *Comput. Math. Appl.*, 68(11), 2014. doi:[10.1016/j.camwa.2014.07.008](https://doi.org/10.1016/j.camwa.2014.07.008).

- [40] J. Chan, N. Heuer, T. Bui-Thanh, and L. Demkowicz. A robust DPG method for convection-dominated diffusion problems II: adjoint boundary conditions and mesh-dependent test norms. *Comput. Math. Appl.*, 67(4):771–795, 2014. doi:[10.1016/j.camwa.2013.06.010](https://doi.org/10.1016/j.camwa.2013.06.010).
- [41] C. Chidume. *Geometric properties of Banach spaces and nonlinear iterations*, volume 1965 of *Lecture Notes in Mathematics*. Springer-Verlag London, Ltd., London, 2009.
- [42] P. G. Ciarlet. *Linear and nonlinear functional analysis with applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.
- [43] I. Cioranescu. *Geometry of Banach spaces, duality mappings and nonlinear problems*, volume 62 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1990. doi:[10.1007/978-94-009-2121-4](https://doi.org/10.1007/978-94-009-2121-4).
- [44] B. Cockburn and P.-A. Gremaud. Error estimates for finite element methods for scalar conservation laws. *SIAM J. Numer. Anal.*, 33(2):522–554, 1996. doi:[10.1137/0733028](https://doi.org/10.1137/0733028).
- [45] B. Cockburn and C.-W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws. V. Multidimensional systems. *J. Comput. Phys.*, 141(2):199–224, 1998. doi:[10.1006/jcph.1998.5892](https://doi.org/10.1006/jcph.1998.5892).
- [46] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. The development of discontinuous Galerkin methods. In *Discontinuous Galerkin methods (Newport, RI, 1999)*, volume 11 of *Lect. Notes Comput. Sci. Eng.*, pages 3–50. Springer, Berlin, 2000. doi:[10.1007/978-3-642-59721-3_1](https://doi.org/10.1007/978-3-642-59721-3_1).
- [47] A. Cohen, W. Dahmen, and G. Welper. Adaptivity and variational stabilization for convection-diffusion equations. *ESAIM Math. Model. Numer. Anal.*, 46(5):1247–1273, 2012. doi:[10.1051/m2an/2012003](https://doi.org/10.1051/m2an/2012003).
- [48] M. Crouzeix and V. Thomée. The stability in L_p and W_p^1 of the L_2 -projection onto finite element function spaces. *Math. Comp.*, 48(178):521–532, 1987. doi:[10.2307/2007825](https://doi.org/10.2307/2007825).
- [49] W. Dahmen, C. Huang, C. Schwab, and G. Welper. Adaptive Petrov-Galerkin methods for first order transport equations. *SIAM J. Numer. Anal.*, 50(5):2420–2445, 2012. doi:[10.1137/110823158](https://doi.org/10.1137/110823158).
- [50] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology. Vol. 3*. Springer-Verlag, Berlin, 1990. Spectral theory and applications, With the collaboration of Michel Artola and Michel Cessenat, Translated from the French by John C. Amson.
- [51] K. Deimling. *Nonlinear functional analysis*. Springer-Verlag, Berlin, 1985. doi:[10.1007/978-3-662-00547-7](https://doi.org/10.1007/978-3-662-00547-7).
- [52] L. Demkowicz. Various variational formulations and closed range theorem. *ICES Report*, pages 15–03, 2015.
- [53] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: the transport equation. *Comput. Methods Appl. Mech. Engrg.*, 199(23-24):1558–1572, 2010. doi:[10.1016/j.cma.2010.01.003](https://doi.org/10.1016/j.cma.2010.01.003).

- [54] L. Demkowicz and J. Gopalakrishnan. A primal DPG method without a first-order reformulation. *Comput. Math. Appl.*, 66(6):1058–1064, 2013. doi:[10.1016/j.camwa.2013.06.029](https://doi.org/10.1016/j.camwa.2013.06.029).
- [55] L. Demkowicz and J. Gopalakrishnan. Discontinuous Petrov-Galerkin (DPG) method. *ICES Report*, (15–20), 2015.
- [56] L. Demkowicz and N. Heuer. Robust DPG method for convection-dominated diffusion problems. *SIAM J. Numer. Anal.*, 51(5):2514–2537, 2013. doi:[10.1137/120862065](https://doi.org/10.1137/120862065).
- [57] L. Demkowicz, J. Gopalakrishnan, and A. H. Niemi. A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity. *Appl. Numer. Math.*, 62(4):396–427, 2012. doi:[10.1016/j.apnum.2011.09.002](https://doi.org/10.1016/j.apnum.2011.09.002).
- [58] L. Demkowicz, T. F urer, N. Heuer, and X. Tian. The Double Adaptivity Paradigm (How to circumvent the discrete inf-sup conditions of Babu ska and Brezzi). *ICES Report*, (19–07), 2019.
- [59] L. F. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. II. Optimal test functions. *Numer. Methods Partial Differential Equations*, 27(1):70–105, 2011. doi:[10.1002/num.20640](https://doi.org/10.1002/num.20640).
- [60] L. F. Demkowicz and J. Gopalakrishnan. An overview of the discontinuous Petrov Galerkin method. In *Recent developments in discontinuous Galerkin finite element methods for partial differential equations*, volume 157 of *IMA Vol. Math. Appl.*, pages 149–180. Springer, Cham, 2014. doi:[10.1007/978-3-319-01818-8_6](https://doi.org/10.1007/978-3-319-01818-8_6).
- [61] A. Demlow, D. Leykekhman, A. H. Schatz, and L. B. Wahlbin. Best approximation property in the W_∞^1 norm for finite element methods on graded meshes. *Math. Comp.*, 81(278):743–764, 2012. ISSN 0025-5718. doi:[10.1090/S0025-5718-2011-02546-9](https://doi.org/10.1090/S0025-5718-2011-02546-9).
- [62] D. A. Di Pietro and J. Droniou. $W^{s,p}$ -approximation properties of elliptic projectors on polynomial spaces, with application to the error analysis of a hybrid high-order discretisation of Leray-Lions problems. *Math. Models Methods Appl. Sci.*, 27(5):879–908, 2017. ISSN 0218-2025. doi:[10.1142/S0218202517500191](https://doi.org/10.1142/S0218202517500191).
- [63] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Math matiques   Applications (Berlin) [Mathematics   Applications]*. Springer, Heidelberg, 2012. doi:[10.1007/978-3-642-22980-0](https://doi.org/10.1007/978-3-642-22980-0).
- [64] E. J. Doedel. The construction of finite difference approximations to ordinary differential equations. *SIAM J. Numer. Anal.*, 15(3):450–465, 1978. doi:[10.1137/0715029](https://doi.org/10.1137/0715029).
- [65] D. L. Donoho. For most large underdetermined systems of equations, the minimal l_1 -norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.*, 59(7):907–934, 2006. doi:[10.1002/cpa.20131](https://doi.org/10.1002/cpa.20131).

- [66] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l_1 -norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829, 2006. doi:[10.1002/cpa.20132](https://doi.org/10.1002/cpa.20132).
- [67] J. Douglas, Jr., T. Dupont, and L. Wahlbin. Optimal L_∞ error estimates for Galerkin approximations to solutions of two-point boundary value problems. *Math. Comp.*, 29:475–483, 1975. doi:[10.2307/2005565](https://doi.org/10.2307/2005565).
- [68] K. Eriksson. Improved accuracy by adapted mesh-refinements in the finite element method. *Math. Comp.*, 44(170):321–343, 1985. ISSN 0025-5718. doi:[10.2307/2007955](https://doi.org/10.2307/2007955).
- [69] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004. doi:[10.1007/978-1-4757-4355-5](https://doi.org/10.1007/978-1-4757-4355-5).
- [70] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010. doi:[10.1090/gsm/019](https://doi.org/10.1090/gsm/019).
- [71] P. A. Farrell, A. F. Hegarty, J. J. H. Miller, E. O’Riordan, and G. I. Shishkin. *Robust computational techniques for boundary layers*, volume 16 of *Applied Mathematics (Boca Raton)*. Chapman & Hall/CRC, Boca Raton, FL, 2000.
- [72] S. Franz and N. Kopteva. On the sharpness of Green’s function estimates for a convection-diffusion problem. *arXiv preprint arXiv:1102.4520*, 2011. URL <https://arxiv.org/pdf/1102.4520.pdf>.
- [73] S. J. Fromm. Potential space estimates for Green potentials in convex domains. *Proc. Amer. Math. Soc.*, 119(1):225–233, 1993. doi:[10.2307/2159846](https://doi.org/10.2307/2159846).
- [74] J. W. Gibbs. Fourier’s series. *Nature*, 59(1539):606, 1899.
- [75] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*, volume 224 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 1983. doi:[10.1007/978-3-642-61798-0](https://doi.org/10.1007/978-3-642-61798-0).
- [76] E. Godlewski and P.-A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1996.
- [77] S. K. Godunov, A. V. Zabrodin, and G. P. Prokopov. A difference scheme for two-dimensional nonstationary problems of gas dynamics and calculation of the flow with a receding shock wave. *Ž. Vyčisl. Mat i Mat. Fiz.*, 1:1020–1050, 1961.
- [78] A. L. Goncharov and I. V. Fryazinov. Difference schemes on a nine-point “cross” pattern for solving Navier-Stokes equations. *Zh. Vychisl. Mat. i Mat. Fiz.*, 28(6):867–878, 958, 1988.
- [79] J. Gopalakrishnan. Five lectures on DPG methods. *arXiv preprint arXiv:1306.0557*, 2013. URL <https://arxiv.org/pdf/1306.0557.pdf>.

- [80] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Mathematics of Computation*, 83(286):537–552, 2014. doi:[10.1090/S0025-5718-2013-02721-4](https://doi.org/10.1090/S0025-5718-2013-02721-4).
- [81] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.
- [82] J.-L. Guermond. A finite element technique for solving first-order PDEs in Lp. *SIAM Journal on Numerical Analysis*, 42(2):714–737, 2004. doi:[10.1137/S0036142902417054](https://doi.org/10.1137/S0036142902417054).
- [83] J.-L. Guermond and B. Popov. Linear advection with ill-posed boundary conditions via L^1 -minimization. *Int. J. Numer. Anal. Model.*, 4(1):39–47, 2007.
- [84] J.-L. Guermond and B. Popov. L^1 -approximation of stationary Hamilton-Jacobi equations. *SIAM Journal on Numerical Analysis*, 47(1):339–362, 2008/09. doi:[10.1137/070681922](https://doi.org/10.1137/070681922).
- [85] J.-L. Guermond and B. Popov. An optimal L^1 -minimization algorithm for stationary Hamilton-Jacobi equations. *Commun. Math. Sci.*, 7(1):211–238, 2009. doi:[10.4310/CMS.2009.v7.n1.a11](https://doi.org/10.4310/CMS.2009.v7.n1.a11).
- [86] J.-L. Guermond, F. Marpeau, and B. Popov. A fast algorithm for solving first-order PDEs by L^1 -minimization. *Commun. Math. Sci.*, 6(1):199–216, 2008. doi:[10.4310/CMS.2008.v6.n1.a10](https://doi.org/10.4310/CMS.2008.v6.n1.a10).
- [87] J. Guzmán, D. Leykekhman, J. Rossmann, and A. H. Schatz. Hölder estimates for Green’s functions on convex polyhedral domains and their applications to finite element methods. *Numer. Math.*, 112(2):221–243, 2009. ISSN 0029-599X. doi:[10.1007/s00211-009-0213-y](https://doi.org/10.1007/s00211-009-0213-y).
- [88] A. Harten. On a class of high resolution total-variation-stable finite-difference schemes. *SIAM J. Numer. Anal.*, 21(1):1–23, 1984. doi:[10.1137/0721001](https://doi.org/10.1137/0721001). With an appendix by Peter D. Lax.
- [89] R. Hartmann and P. Houston. Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws. *SIAM J. Sci. Comput.*, 24(3):979–1004 (electronic), 2002. doi:[10.1137/S1064827501389084](https://doi.org/10.1137/S1064827501389084).
- [90] P. Houston and E. Süli. Stabilised hp -finite element approximation of partial differential equations with nonnegative characteristic form. *Computing*, 66(2):99–119, 2001. doi:[10.1007/s006070170030](https://doi.org/10.1007/s006070170030). Archives for scientific computing. Numerical methods for transport-dominated and related problems (Magdeburg, 1999).
- [91] P. Houston and E. Süli. hp -adaptive discontinuous Galerkin finite element methods for first-order hyperbolic problems. *SIAM J. Sci. Comput.*, 23(4):1226–1252, 2001. doi:[10.1137/S1064827500378799](https://doi.org/10.1137/S1064827500378799).
- [92] P. Houston, R. Rannacher, and E. Süli. A posteriori error analysis for stabilised finite element approximations of transport problems. *Comput. Methods Appl. Mech. Engrg.*, 190(11-12):1483–1508, 2000. doi:[10.1016/S0045-7825\(00\)00174-2](https://doi.org/10.1016/S0045-7825(00)00174-2).

- [93] P. Houston, I. Muga, S. Roggendorf, and K. G. van der Zee. The Convection-Diffusion-Reaction Equation in Non-Hilbert Sobolev Spaces: A Direct Proof of the Inf-Sup Condition and Stability of Galerkin's Method. *Comput. Methods Appl. Math.*, 19(3):503–522, 2019. doi:[10.1515/cmam-2018-0198](https://doi.org/10.1515/cmam-2018-0198).
- [94] P. Houston, S. Roggendorf, and K. G. van der Zee. Gibbs Phenomena in the L^q -Best Approximation of Discontinuities in Finite Element Spaces – Some Examples. *arXiv preprint arXiv:1909.00658*, 2019. URL <https://arxiv.org/pdf/1909.00658.pdf>.
- [95] P. Houston, S. Roggendorf, and K. G. van der Zee. Eliminating Gibbs Phenomena: A Non-linear Petrov-Galerkin Method for the Convection-Diffusion-Reaction Equation. *arXiv preprint arXiv:1908.00996*, 2019. URL <https://arxiv.org/pdf/1908.00996.pdf>.
- [96] T. J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Comput. Methods Appl. Mech. Engrg.*, 127(1-4):387–401, 1995. doi:[10.1016/0045-7825\(95\)00844-9](https://doi.org/10.1016/0045-7825(95)00844-9).
- [97] T. J. R. Hughes and A. Brooks. A multidimensional upwind scheme with no crosswind diffusion. In *Finite element methods for convection dominated flows (Papers, Winter Ann. Meeting Amer. Soc. Mech. Engrs., New York, 1979)*, volume 34 of *AMD*, pages 19–35. Amer. Soc. Mech. Engrs. (ASME), New York, 1979.
- [98] T. J. R. Hughes and G. Sangalli. Variational multiscale analysis: the fine-scale Green's function, projection, optimization, localization, and stabilized methods. *SIAM J. Numer. Anal.*, 45(2):539–557, 2007. doi:[10.1137/050645646](https://doi.org/10.1137/050645646).
- [99] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy. The variational multiscale method—a paradigm for computational mechanics. *Comput. Methods Appl. Mech. Engrg.*, 166(1-2):3–24, 1998. doi:[10.1016/S0045-7825\(98\)00079-6](https://doi.org/10.1016/S0045-7825(98)00079-6).
- [100] A. M. Il'in. Differencing scheme for a differential equation with a small parameter affecting the highest derivative. *Mathematical Notes of the Academy of Sciences of the USSR*, 6(2):237–248, 1969. doi:[10.1007/BF01093706](https://doi.org/10.1007/BF01093706).
- [101] T. Jakab, I. Mitrea, and M. Mitrea. Sobolev estimates for the Green potential associated with the Robin-Laplacian in Lipschitz domains satisfying a uniform exterior ball condition. In *Sobolev spaces in mathematics. II*, volume 9 of *Int. Math. Ser. (N. Y.)*, pages 227–260. Springer, New York, 2009.
- [102] M. Jensen. *Discontinuous Galerkin methods for Friedrichs systems with irregular solutions*. PhD thesis, University of Oxford, 2005.
- [103] D. Jerison and C. E. Kenig. The inhomogeneous Dirichlet problem in Lipschitz domains. *J. Funct. Anal.*, 130(1):161–219, 1995. doi:[10.1006/jfan.1995.1067](https://doi.org/10.1006/jfan.1995.1067).
- [104] B.-N. Jiang. Non-oscillatory and non-diffusive solution of convection problems by the iteratively reweighted least-squares finite element method. *Journal of Computational Physics*, 105(1):108–121, 1993. doi:[10.1006/jcph.1993.1057](https://doi.org/10.1006/jcph.1993.1057).

- [105] B.-N. Jiang. *The least-squares finite element method: theory and applications in computational fluid dynamics and electromagnetics*. Springer Science & Business Media, 1998. doi:[10.1007/978-3-662-03740-9](https://doi.org/10.1007/978-3-662-03740-9).
- [106] V. John and P. Knobloch. On the performance of SOLD methods for convection-diffusion problems with interior layers. *Int. J. Comput. Sci. Math.*, 1(2-4):245–258, 2007. doi:[10.1504/IJCSM.2007.016534](https://doi.org/10.1504/IJCSM.2007.016534).
- [107] V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I–A review. *Computer Methods in Applied Mechanics and Engineering*, 196(17):2197–2215, 2007. doi:[10.1016/j.cma.2006.11.013](https://doi.org/10.1016/j.cma.2006.11.013).
- [108] V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II–analysis for P1 and Q1 finite elements. *Computer Methods in Applied Mechanics and Engineering*, 197(21):1997–2014, 2008. doi:[10.1016/j.cma.2007.12.019](https://doi.org/10.1016/j.cma.2007.12.019).
- [109] C. Johnson, U. Nävert, and J. Pitkänta. Finite element methods for linear hyperbolic problems. *Comput. Methods Appl. Mech. Engrg.*, 45(1-3):285–312, 1984. doi:[10.1016/0045-7825\(84\)90158-0](https://doi.org/10.1016/0045-7825(84)90158-0).
- [110] B. Keith, S. Petrides, F. Fuentes, and D. Leszek. Discrete least-squares finite element methods. *ICES Report*, pages 17–09, 2017.
- [111] C. E. Kenig. Elliptic boundary value problems on Lipschitz domains. In *Beijing lectures in harmonic analysis (Beijing, 1984)*, volume 112 of *Ann. of Math. Stud.*, pages 131–183. Princeton Univ. Press, Princeton, NJ, 1986.
- [112] C. E. Kenig. *Harmonic analysis techniques for second order elliptic boundary value problems*, volume 83 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 1994. doi:[10.1090/cbms/083](https://doi.org/10.1090/cbms/083).
- [113] N. V. Kopteva. On the convergence, uniform with respect to the small parameter, of a scheme with central difference on refined grids. *Zh. Vychisl. Mat. Mat. Fiz.*, 39(10):1662–1678, 1999.
- [114] J. P. Krasovskii. Isolation of singularities of the Green’s function. *Mathematics of the USSR-Izvestiya*, 1(5):935, 1967.
- [115] D. Kröner. *Numerical schemes for conservation laws*. Wiley-Teubner Series Advances in Numerical Mathematics. John Wiley & Sons, Ltd., Chichester; B. G. Teubner, Stuttgart, 1997.
- [116] M. G. Larson and F. Bengzon. *The finite element method: theory, implementation, and applications*, volume 10 of *Texts in Computational Science and Engineering*. Springer, Heidelberg, 2013. doi:[10.1007/978-3-642-33287-6](https://doi.org/10.1007/978-3-642-33287-6).
- [117] J. E. Lavery. Nonoscillatory solution of the steady-state inviscid Burgers’ equation by mathematical programming. *J. Comput. Phys.*, 79(2):436–448, 1988. doi:[10.1016/0021-9991\(88\)90024-1](https://doi.org/10.1016/0021-9991(88)90024-1).

- [118] J. E. Lavery. Solution of steady-state one-dimensional conservation laws by mathematical programming. *SIAM J. Numer. Anal.*, 26(5):1081–1089, 1989. doi:[10.1137/0726060](https://doi.org/10.1137/0726060).
- [119] J. E. Lavery. Solution of steady-state, two-dimensional conservation laws by mathematical programming. *SIAM J. Numer. Anal.*, 28(1):141–155, 1991. doi:[10.1137/0728007](https://doi.org/10.1137/0728007).
- [120] R. D. Lazarov, I. D. Mishev, and P. S. Vassilevski. Finite volume methods for convection-diffusion problems. *SIAM J. Numer. Anal.*, 33(1):31–55, 1996. doi:[10.1137/0733003](https://doi.org/10.1137/0733003).
- [121] B. Leonard. A survey of finite differences of opinion on numerical modelling of the incompressible diffusion convection equation. *Finite element methods for convection dominated flows*, pages 1–30, 1979.
- [122] B. P. Leonard. A stable and accurate convective modelling procedure based on quadratic upstream interpolation. *Computer methods in applied mechanics and engineering*, 19(1):59–98, 1979. doi:[10.1016/0045-7825\(79\)90034-3](https://doi.org/10.1016/0045-7825(79)90034-3).
- [123] R. J. LeVeque. *Finite volume methods for hyperbolic problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2002. doi:[10.1017/CBO9780511791253](https://doi.org/10.1017/CBO9780511791253).
- [124] R. J. LeVeque. *Finite difference methods for ordinary and partial differential equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007. Steady-state and time-dependent problems.
- [125] H. Li. The W_p^1 stability of the Ritz projection on graded meshes. *Math. Comp.*, 86(303):49–74, 2017. ISSN 0025-5718. doi:[10.1090/mcom/3101](https://doi.org/10.1090/mcom/3101).
- [126] A. Logg, K.-A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012. doi:[10.1007/978-3-642-23099-8](https://doi.org/10.1007/978-3-642-23099-8).
- [127] R. E. Lynch and J. R. Rice. A high-order difference method for differential equations. *Math. Comp.*, 34(150):333–372, 1980. doi:[10.2307/2006090](https://doi.org/10.2307/2006090).
- [128] J. A. Mackenzie and K. W. Morton. Finite volume solutions of convection-diffusion test problems. *Mathematics of Computation*, 60(201):189–220, 1993. doi:[10.2307/2153161](https://doi.org/10.2307/2153161).
- [129] G. Matthies, P. Skrzypacz, and L. Tobiska. Stabilization of local projection type applied to convection-diffusion problems with mixed boundary conditions. *Electron. Trans. Numer. Anal.*, 32:90–105, 2008.
- [130] J. M. Melenk and C. Schwab. An hp finite element method for convection-diffusion problems. Tech. Rep. 97-05, Eidgenössische Technische Hochschule Zürich (ETH). Seminar für Angewandte Mathematik, 1997.
- [131] J. J. H. Miller, E. O’Riordan, and G. I. Shishkin. *Fitted numerical methods for singular perturbation problems*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, revised edition, 2012. Error estimates in the maximum norm for linear problems in one and two dimensions.

- [132] D. Mitrea, M. Mitrea, and L. Yan. Boundary value problems for the Laplacian in convex and semiconvex domains. *J. Funct. Anal.*, 258:2507–2585, 2010. doi:[10.1016/j.jfa.2010.01.012](https://doi.org/10.1016/j.jfa.2010.01.012).
- [133] K. W. Morton and T. Sonar. Finite volume methods for hyperbolic conservation laws. *Acta Numer.*, 16:155–238, 2007. doi:[10.1017/S0962492906300013](https://doi.org/10.1017/S0962492906300013).
- [134] K. W. Morton, M. Stynes, and E. Süli. Analysis of a cell-vertex finite volume method for convection-diffusion problems. *Math. Comp.*, 66(220):1389–1406, 1997. doi:[10.1090/S0025-5718-97-00886-7](https://doi.org/10.1090/S0025-5718-97-00886-7).
- [135] E. Moskona, P. Petrushev, and E. Saff. The Gibbs phenomenon for best L1-trigonometric polynomial approximation. *Constructive Approximation*, 11(3):391–416, 1995. doi:[10.1007/BF01208562](https://doi.org/10.1007/BF01208562).
- [136] I. Muga and K. G. van der Zee. Discretization of Linear Problems in Banach Spaces: Residual Minimization, Nonlinear Petrov-Galerkin, and Monotone Mixed Methods. *arXiv preprint arXiv:1511.04400v1*, 2015. URL <https://arxiv.org/pdf/1511.04400v1.pdf>.
- [137] I. Muga and K. G. van der Zee. Discretization of Linear Problems in Banach Spaces: Residual Minimization, Nonlinear Petrov-Galerkin, and Monotone Mixed Methods. *arXiv preprint arXiv:1511.04400v2*, 2017. URL <https://arxiv.org/pdf/1511.04400v2.pdf>.
- [138] I. Muga and K. G. van der Zee. Discretization of Linear Problems in Banach Spaces: Residual Minimization, Nonlinear Petrov-Galerkin, and Monotone Mixed Methods. *arXiv preprint arXiv:1511.04400v3*, 2018. URL <https://arxiv.org/pdf/1511.04400v3.pdf>.
- [139] I. Muga, M. J. Tyler, and K. G. van der Zee. The discrete-dual minimal-residual method (DDMRes) for weak advection-reaction problems in Banach spaces. *Computational Methods in Applied Mathematics*, 19(3):557–579, 2019. doi:[10.1515/cmam-2018-0199](https://doi.org/10.1515/cmam-2018-0199).
- [140] J. Nečas. Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 16(4):305–326, 1962.
- [141] A. H. Niemi, N. O. Collier, and V. M. Calo. Discontinuous Petrov-Galerkin method based on the optimal test space norm for one-dimensional transport problems. *Procedia Computer Science*, 4:1862–1869, 2011. doi:[10.1016/j.procs.2011.04.202](https://doi.org/10.1016/j.procs.2011.04.202).
- [142] J. T. Oden and L. F. Demkowicz. *Applied functional analysis*. CRC Press, Boca Raton, FL, second edition, 2010.
- [143] T. E. Peterson. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM J. Numer. Anal.*, 28(1):133–140, 1991. doi:[10.1137/0728006](https://doi.org/10.1137/0728006).
- [144] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer-Verlag, Berlin, second edition, 2007. doi:[10.1007/b98885](https://doi.org/10.1007/b98885).

- [145] R. Rannacher and R. Scott. Some optimal error estimates for piecewise linear finite element approximations. *Math. Comp.*, 38(158):437–445, 1982. ISSN 0025-5718. doi:[10.2307/2007280](https://doi.org/10.2307/2007280).
- [146] W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. *Technical Report LA-UR-73-479*, 1973.
- [147] F. Richards. A Gibbs phenomenon for spline functions. *Journal of approximation theory*, 66(3):334–351, 1991. doi:[10.1016/0021-9045\(91\)90034-8](https://doi.org/10.1016/0021-9045(91)90034-8).
- [148] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations*, volume 24 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2008. Convection-diffusion-reaction and flow problems.
- [149] A. Russo. Streamline-upwind Petrov/Galerkin method (SUPG) vs residual-free bubbles (RFB). *Comput. Methods Appl. Mech. Engrg.*, 195(13-16):1608–1620, 2006. doi:[10.1016/j.cma.2005.05.031](https://doi.org/10.1016/j.cma.2005.05.031).
- [150] E. B. Saff and S. Tashev. Gibbs phenomenon for best L_p approximation by polygonal lines. *East J. Approx.*, 5(2):235–251, 1999.
- [151] C.-W. Shu. TVB uniformly high-order schemes for conservation laws. *Math. Comp.*, 49(179):105–121, 1987. doi:[10.2307/2008252](https://doi.org/10.2307/2008252).
- [152] C.-W. Shu. High order weighted essentially nonoscillatory schemes for convection dominated problems. *SIAM Rev.*, 51(1):82–126, 2009. doi:[10.1137/070679065](https://doi.org/10.1137/070679065).
- [153] C.-W. Shu. High order WENO and DG methods for time-dependent convection-dominated PDEs: A brief survey of several recent developments. *Journal of Computational Physics*, 316:598–613, 2016. doi:[10.1016/j.jcp.2016.04.030](https://doi.org/10.1016/j.jcp.2016.04.030).
- [154] I. Singer. *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, volume 171 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag Berlin Heidelberg, 1970. doi:[10.1007/978-3-662-41583-2](https://doi.org/10.1007/978-3-662-41583-2). Translated by R. Georgescu.
- [155] P. Solin, L. Korous, and P. Kus. Hermes2D, a C++ library for rapid development of adaptive hp -FEM and hp -DG solvers. *J. Comput. Appl. Math.*, 270:152–165, 2014.
- [156] I. Stakgold and M. J. Holst. *Green’s functions and boundary value problems*, volume 99. John Wiley & Sons, 2011.
- [157] E. M. Stein. *Singular integrals and differentiability properties of functions*. Princeton Mathematical Series, No. 30. Princeton University Press, Princeton, N.J., 1970.
- [158] A. Stern. Banach space projections and Petrov–Galerkin estimates. *Numer. Math.*, 130:125–133, 2015. doi:[10.1007/s00211-014-0658-5](https://doi.org/10.1007/s00211-014-0658-5).
- [159] G. Stoyan. Monotone difference schemes for diffusion-convection problems. *Z. Angew. Math. Mech.*, 59(8):361–372, 1979. doi:[10.1002/zamm.19790590805](https://doi.org/10.1002/zamm.19790590805).

- [160] M. Stynes. Finite volume methods for convection-diffusion problems. *J. Comput. Appl. Math.*, 63(1-3):83–90, 1995. doi:[10.1016/0377-0427\(95\)00056-9](https://doi.org/10.1016/0377-0427(95)00056-9). International Symposium on Mathematical Modelling and Computational Methods Modelling 94 (Prague, 1994).
- [161] D. B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numer. Algorithms*, 42(3-4):309–323, 2006. doi:[10.1007/s11075-006-9046-2](https://doi.org/10.1007/s11075-006-9046-2).
- [162] L. Tobiska. Analysis of a new stabilized higher order finite element method for advection-diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 196(1-3):538–550, 2006. doi:[10.1016/j.cma.2006.05.009](https://doi.org/10.1016/j.cma.2006.05.009).
- [163] J. R. R. Tolkien. *The Lord of the Rings: One Volume*. Houghton Mifflin Harcourt, 2012.
- [164] R. Verfürth. *A posteriori error estimation techniques for finite element methods*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013. doi:[10.1093/acprof:oso/9780199679423.001.0001](https://doi.org/10.1093/acprof:oso/9780199679423.001.0001).
- [165] R. Vulcanović. On a numerical solution of a type of singularly perturbed boundary value problem by using a special discretization mesh. *Univ. u Novom Sadu Zb. Rad. Prirod.-Mat. Fak. Ser. Mat.*, 13:187–201, 1983.
- [166] M. F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.*, 15(1):152–161, 1978. doi:[10.1137/0715010](https://doi.org/10.1137/0715010).
- [167] H. Wilbraham. On a certain periodic function. *Cambridge and Dublin Math. J.*, 3(198):1848, 1848.
- [168] S. Yang. The Neumann problem of Laplace’s equation in semiconvex domains. *Nonlinear Anal.*, 133:275–291, 2016. doi:[10.1016/j.na.2015.12.017](https://doi.org/10.1016/j.na.2015.12.017).
- [169] E. Zeidler. *Nonlinear functional analysis and its applications. III*. Springer-Verlag, New York, 1985. Variational methods and optimization, Translated from the German by Leo F. Boron.
- [170] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V. M. Calo. A class of discontinuous Petrov-Galerkin methods. Part IV: the optimal test norm and time-harmonic wave propagation in 1D. *J. Comput. Phys.*, 230(7):2406–2432, 2011. doi:[10.1016/j.jcp.2010.12.001](https://doi.org/10.1016/j.jcp.2010.12.001).

